DOCUMENTS DE TRAVAIL 226

Estimation de la variance pour l'enquête Elfe

Rapport méthodologique pour l'utilisateur

Hélène Juillard

ined
INSTITUT NATIONAL D'ÉTUDES DÉMOGRAPHIQUES

Juin 2015

ESTIMATION DE LA VARIANCE POUR L'ENQUETE ELFE
Rapport méthodologique pour l'utilisateur

Hélène Juillard(⋆) *

(⋆) dans le cadre de ses travaux de doctorat financés par l'Ined
sous la direction d'Anne Ruiz-Gazen et de Guillaume Chauvet

Document actualisé le 2 mai 2016

Travail en collaboration avec Marie Cheminat, administrateur bases de données Elfe.
Ce document est destiné aux utilisateurs des données issues de la cohorte Elfe et propose un estimateur de variance prenant en compte le plan de sondage utilisé pour l'étude Elfe.

Chaque échantillonnage conduit à une variance dite d'échantillonnage. 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 (tirage exhaustif), 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 suivi de plusieurs phases de non-réponse. Le calcul de l'estimateur de variance n'est pas directement disponible dans la littérature et fait l'objet d'un travail de recherche de la part de l'auteur de cette note. Ce document propose de détailler l'application au cas spécifique de l'enquête Elfe.

Une modélisation du plan de sondage est proposée, avec prise en compte de la non-réponse et du calage et les estimateurs de variance associés sont dérivés. L'estimateur de variance sans biais n'étant a priori programmé dans aucun logiciel, plusieurs estimateurs simplifiés prenant en compte les procédures logicielles déjà existantes (R / SAS / Stata) sont décrits et illustrés sur données Elfe. Après comparaison, un unique estimateur simplifié est recommandé aux utilisateurs des données Elfe.

Les détails des calculs présentés dans ce document sont donnés dans Chauvet, Juillard, et Ruiz-Gazen (2016).

Il est conseillé à l'utilisateur de bien lire la première section. Un résumé de la suite du document est proposé en page 19.

*L'auteur remercie pour leurs conseils, Emmanuel Gros et Stéphane Legleye.

Table des matières

1 Echantillonnage et variance

Dans les enquêtes, on s'intéresse à des populations de tailles finies, dans lesquelles on choisit parfois de sélectionner un échantillon : on parle alors d'enquête par sondage. On s'intéresse à un paramètre θ inconnu (calculable seulement sur toute la population) que l'on estime à partir d'un échantillon par $\hat{\theta}$ (l'accent circonflexe symbolise l'estimateur). On veut inférer les résultats de l'échantillon à la population. Par exemple, on veut estimer le nombre total de naissances sous césarienne qui ont eu lieu durant l'année 2011 en enquêtant seulement quelques maternités durant quelques jours.

Ce qui nous intéresse c'est de savoir si le $\hat{\theta}$ obtenu à partir de notre échantillon sélectionné est proche de θ. Si l'on avait sélectionné un autre échantillon, aurait-on obtenu le même $\hat{\theta}$? Et si l'on en avait sélectionné un autre ? Ou encore un autre ? C'est en imaginant toutes ces différentes valeurs de $\hat{\theta}$ que l'on peut se représenter la **variance dite d'échantillonnage** $\mathbf{V}\left(\hat{\theta}\right)$ (voir la Figure 1).

En pratique, la variance $\mathbf{V}\left(\hat{\theta}\right)$ est inconnue mais estimée sur l'échantillon par $\hat{\mathbf{V}}\left(\hat{\theta}\right)$.

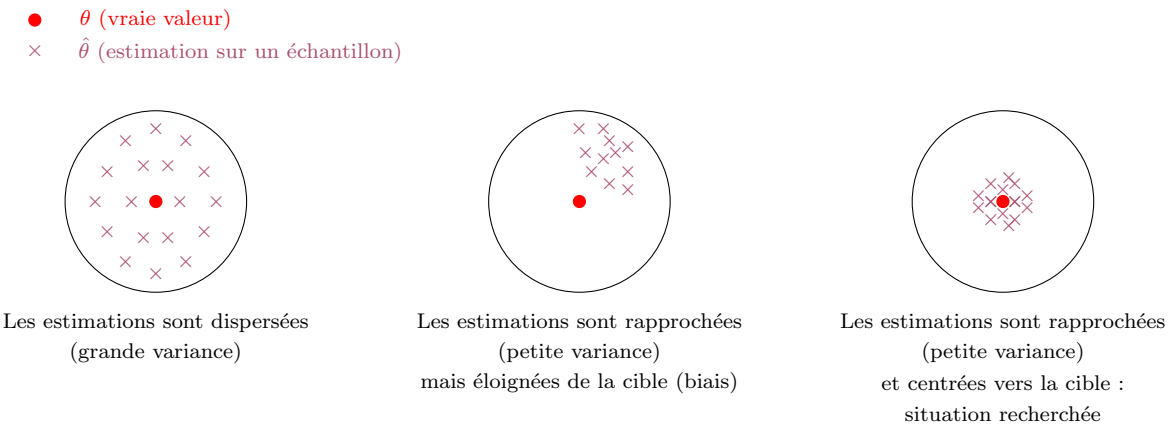

FIGURE 1 – Biais et variances

En théorie des sondages, l'aléa provient de la sélection de l'échantillon. Pour comparaison, en statistique classique, la variable d'intérêt y est une variable aléatoire, alors qu'en statistique d'enquête elle est fixée : c'est l'indicatrice d'appartenance à l'échantillon qui est aléatoire.

La méthode de tirage de l'échantillon est importante : les calculs de $\hat{\theta}$ et $\hat{\mathbf{V}}(\hat{\theta})$ vont en dépendre. Par exemple, si pour connaître le nombre de césariennes en 2011, on sélectionnait certaines régions de France, puis à l'intérieur de ces régions, certaines maternités, les formules associées à $\hat{\theta}$ et $\hat{\mathbf{V}}(\hat{\theta})$ ne seraient pas les mêmes que si l'on sélectionnait directement certaines maternités parmi toutes les maternités de France. Autrement dit, une formule de variance d'échantillonnage doit refléter la structure du plan de sondage.

La variance est donc une **mesure d'incertitude dépendant du plan de sondage** : rien ne nous dit que la valeur issue de notre échantillon est exacte, on l'espère seulement rapprochée de θ. Elle va par exemple permettre à notre estimateur $\hat{\theta}$ d'être associé à un intervalle de confiance, c'est-à-dire deux valeurs entre lesquelles θ aura 95 % (ou 90 %, ou 99 %...) de chance d'être compris. Si à partir d'une enquête, on estime le taux de césariennes à 20 %, ce chiffre est-il sûr, précis ? L'intervalle de confiance estimé construit autour de cette valeur est-il [19 %, 21 %] ? Ou plutôt [15 %, 25 %] ? **L'enquête effectuée, la précision issue du plan de sondage ne se choisit plus, c'est avant** que les choix sont faits concernant le plan d'échantillonnage et la précision qui en découle. Toutefois, il existe différentes méthodes post-échantillonnage usant d'informations auxiliaires, permettant d'améliorer la précision des estimateurs, comme le calage que nous verrons dans ce document.

2 L'enquête Elfe : contexte et modélisation du plan de sondage

La population d'inférence est celle des nourrissons nés durant l'année 2011 en France métropolitaine, issus d'un accouchement au plus gémellaire, hors grands prématurés, ayant une mère majeure, 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. Toutes les familles sélectionnées ont été enquêtées peu de temps après l'accouchement dans certaines maternités métropolitaines et durant certains jours de l'année (voir Figure 2).

Le plan de sondage pour les maternités est un plan probabiliste. Concernant les jours, 25 ont été choisis durant quatre périodes (appelées vagues) couvrant les quatre saisons de l'année [1] (dont la moitié devait coïncider avec l'échantillon démographique permanent E.D.P.). Notons que les deux échantillons (maternités et jours) ont été sélectionnés indépendamment.

FIGURE 2 – 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 d'accouchements recensés en 2008. Il s'agissait d'un tirage systématique avec pour variables 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 (tirage aléatoire simple sans remise) dans chaque strate : plan STSI.

L'échantillonnage des jours n'est pas aléatoire, d'où la nécessité de le modéliser. Une modélisation est proposée dans la suite de ce document et deux autres modélisations possibles sont développées dans Juillard, Chauvet et Ruiz-Gazen (2015). La modélisation proposée consiste en un plan STSI avec quatre strates (voir Figure 3) que nous nommerons pour simplifier "saisons" dans la suite de ce document 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).

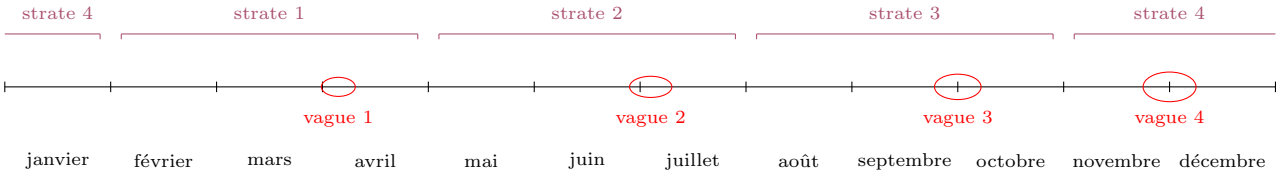

FIGURE 3 – Exemple de découpage de l'année 2011 en 4 strates et sélection de jours

Le plan de sondage final utilisé pour l'enquête Elfe résulte du croisement indépendant de ces deux plans de sondage (l'un dans la population des maternités, l'autre dans celle des jours) et est appelé *plan produit* (ou encore *cross-classified sampling*, voir Ohlsson, 1996 ou Slinner, 2015). L'échantillonnage bien particulier utilisé pour l'enquête Elfe est comparé dans la section suivante à d'autres plans de sondage, afin d'en comprendre les spécificités.

1. 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.

3 Plan produit et autres plans (que l'on ne peut confondre)

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 n_M et un plan de sondage p_D dans la population U_D menant à un échantillon S_D de taille n_D. Pour un plan produit (le cas de l'enquête Elfe), ces deux plans sont indépendants (voir Figure 4). L'unité finale d'échantillonnage qui nous intéresse est caractérisée par un couple maternité × jour (i, k), avec $i \in U_M$ et $k \in U_D$.

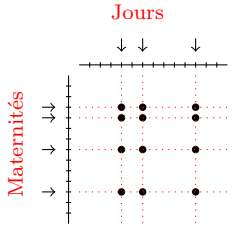

FIGURE 4 – Echantillonnage de maternités et de jours pour un plan produit

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 dans la Figure 5.

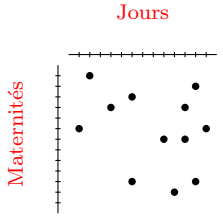

FIGURE 5 – Echantillonnage de maternités et de jours pour un tirage direct dans la population produit

Pour l'enquête Elfe, on distingue deux phases d'échantillonnage : celle sur les jours et celle sur les maternités. Néanmoins le plan Elfe ne peut 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 6). Il ne peut non plus symétriquement être considéré comme un plan classique à deux degrés avec au premier degré un échantillonnage des jours et au second degré un échantillonnage des maternités.

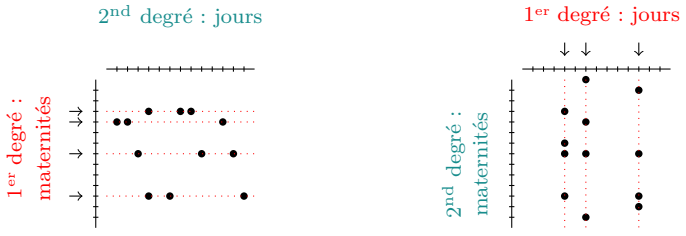

FIGURE 6 – Echantillonnage de maternités et de jours pour un plan à deux degrés, avec tirage de maternités au premier degré (à gauche) ou tirage de jours au premier degré (à droite)

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).

Des estimateurs de variance pour un plan produit quelconque sont dérivés dans Chauvet, Juillard, et Ruiz-Gazen (2016). Une comparaison a été effectuée entre la variance issue d'un plan produit et celle issue d'un plan à deux degrés. La comparaison entre ces deux plans est aussi détaillée dans Juillard (2016) avec possibilité de mettre en pratique sur les logiciels R, SAS et Stata, les étapes d'échantillonnage et d'estimation. Dans la suite de ce document, nous proposons des estimateurs pour le cas particulier de l'enquête Elfe.

4 Estimation de la variance issue du plan Elfe

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 le Tableau 2). 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

$$\hat{t}_Y = \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{i \in S_{Mg}} \sum_{k \in S_{Dh}} \frac{N_{Dh}}{n_{Dh}} \frac{N_{Mg}}{n_{Mg}} Y_{ik} = \sum_{g=1}^{G} \sum_{i \in S_{Mg}} \frac{N_{Mg}}{n_{Mg}} \hat{Y}_{i\bullet} = \sum_{h=1}^{H} \sum_{k \in S_{Dh}} \frac{N_{Dh}}{n_{Dh}} \hat{Y}_{\bullet k} \tag{1}$$

avec $\hat{Y}_{i\bullet}$, l'estimateur d'Horvitz-Thompson du total sur la maternité i et $\hat{Y}_{\bullet k}$, l'estimateur d'Horvitz-Thompson du total sur le jour k. Un estimateur sans biais de la variance de $\hat{t}_Y$ est donné par :

$$\hat{\mathbf{V}}_{prod}\left(\hat{t}_Y\right) = \hat{\mathbf{V}}_D + \hat{\mathbf{V}}_M - \hat{\mathbf{V}}_E \tag{2}$$

avec

$$\hat{\mathbf{V}}_D\left(\hat{t}_Y\right) = \sum_{h=1}^{H} N_{Dh}^2 \left(\frac{1}{n_{Dh}} - \frac{1}{N_{Dh}}\right) s^2_{\hat{Y}_{\bullet\circ,h}}, \tag{3}$$

$$\hat{\mathbf{V}}_M\left(\hat{t}_Y\right) = \sum_{g=1}^{G} N_{Mg}^2 \left(\frac{1}{n_{Mg}} - \frac{1}{N_{Mg}}\right) s^2_{\hat{Y}_{\circ\bullet,g}}, \tag{4}$$

$$\hat{\mathbf{V}}_E\left(\hat{t}_Y\right) = \sum_{g=1}^{G} N_{Mg}^2 \left(\frac{1}{n_{Mg}} - \frac{1}{N_{Mg}}\right) \sum_{h=1}^{H} N_{Dh}^2 \left(\frac{1}{n_{Dh}} - \frac{1}{N_{Dh}}\right) \frac{1}{(n_{Mg}-1)(n_{Dh}-1)} s_{E,hg}, \tag{5}$$

où

$$s^2_{\hat{Y}_{\bullet\circ,h}} = \frac{1}{n_{Dh}-1} \sum_{k \in S_{Dh}} \left(\hat{Y}_{\bullet k} - \frac{1}{n_{Dh}} \sum_{l \in S_{Dh}} \hat{Y}_{\bullet l}\right)^2, \tag{6}$$

$$s^2_{\hat{Y}_{\circ\bullet,g}} = \frac{1}{n_{Mg}-1} \sum_{i \in S_{Mg}} \left(\hat{Y}_{i\bullet} - \frac{1}{n_{Mg}} \sum_{j \in S_{Mg}} \hat{Y}_{j\bullet}\right)^2, \tag{7}$$

$$s_{E,hg} = \sum_{i \in S_{Mg}} \sum_{k \in S_{Dh}} \left[Y_{ik} - \frac{1}{n_{Mg}} \sum_{j \in S_{Mg}} Y_{jk} - \frac{1}{n_{Dh}} \sum_{l \in S_{Dh}} Y_{il} + \frac{1}{n_{Mg}} \frac{1}{n_{Dh}} \sum_{j \in S_{Mg}} \sum_{l \in S_{Dh}} Y_{jl}\right]^2. \tag{8}$$

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.

Dans le Tableau 2 sont présentées les tailles de strates et des échantillons dans chaque strate pour une modélisation STSI du plan de sondage sur les jours.

Strate g	Taille de la strate N_{Mg}	Taille de l'échantillon n_{Mg}	Critère de stratification Nombre d'accouchements en 2008
1	108	28	[145 ; 699]
2	108	47	[700 ; 1009]
3	109	66	[1010 ; 1418]
4	108	97	[1422 ; 2187]
5	111	111	[2197 ; 5215]

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

Strate h	Taille de la strate N_{Dh}	Taille de l'échantillon n_{Dh}	Critère de stratification Saison
1	91	4	Printemps
2	91	6	Ete
3	91	7	Automne
4	91	8	Hiver (fin automne)

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

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

Le traitement de la non-réponse au niveau du biais de l'estimateur est présenté ici succinctement (voir Juillard *et al.*, 2015). L'estimateur de variance prenant en compte l'échantillonnage produit mais aussi la non-réponse est calculé.

5.1 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 : 15 maternités n'ont pas participé au trimestre 1, 8 au trimestre 2, 9 au trimestre 3 et 11 au trimestre 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 trimestre 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 (voir la Figure 7) : 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 (âge de la mère, groupe de région d'habitation, statut immigré de la mère, état matrimonial de la mère, primiparité et niveau d'étude de la mère). 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 pour l'instant.

Notre variable d'intérêt prend la valeur $y_{a_{ik}}$ pour le nourrisson a de la maternité i du jour k. Le total t_Y peut alors s'écrire

$$t_Y = \sum_{i \in U_M} \sum_{k \in U_D} Y_{ik} \quad \text{avec} \quad Y_{ik} = \sum_{a \in U_{ik}} y_a, \tag{9}$$

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

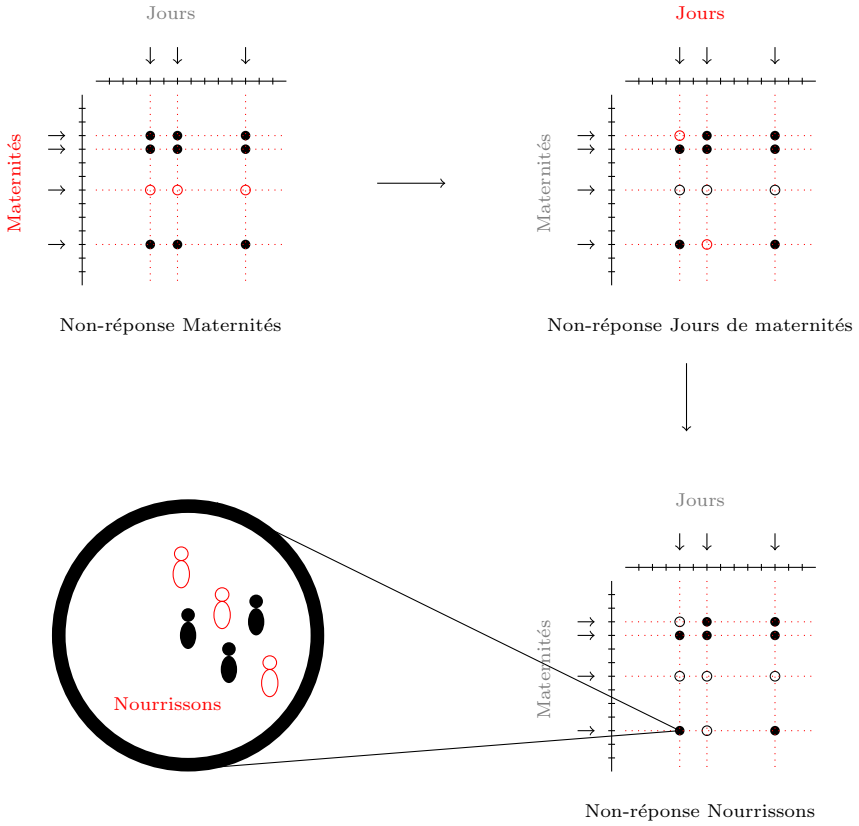

FIGURE 7 – Non-réponse au niveau maternité, puis jour de maternité, puis nourrisson

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

$$\begin{aligned}
\hat{t}_{Y\star} &= \sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{i\in S_{Mg}}\sum_{k\in S_{Dh}} \frac{N_{Dh}}{n_{Dh}}\frac{N_{Mg}}{n_{Mg}}\hat{Y}_{ik} \quad \text{avec} \quad \hat{Y}_{ik} = \sum_{a\in S_{R_{ik}}} \frac{y_a}{\hat{p}_a}, \\
&= \sum_{g=1}^{G}\sum_{i\in S_{Mg}} \frac{N_{Mg}}{n_{Mg}}\hat{\hat{Y}}_{i\bullet} \quad \text{avec} \quad \hat{\hat{Y}}_{i\bullet} = \sum_{h=1}^{H}\sum_{k\in S_{Dh}} \frac{N_{Dh}}{n_{Dh}}\hat{Y}_{ik}, \\
&= \sum_{h=1}^{H}\sum_{k\in S_{Dh}} \frac{N_{Dh}}{n_{Dh}}\hat{\hat{Y}}_{\bullet k} \quad \text{avec} \quad \hat{\hat{Y}}_{\bullet k} = \sum_{g=1}^{G}\sum_{i\in S_{Mg}} \frac{N_{Mg}}{n_{Mg}}\hat{Y}_{ik}.
\end{aligned} \tag{10}$$

5.2 Estimation de la 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). Dans le cas particulier de l'enquête Elfe cela conduit à :

$$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right) = \hat{\mathbf{V}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{V}}_{\text{NR}}\left(\hat{t}_{Y\star}\right) \tag{11}$$

où

$$\hat{\mathbf{V}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \hat{\mathbf{V}}_{\text{ech1}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) - \hat{\mathbf{V}}_{\text{ech2}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) \tag{12}$$

(13)

$$\hat{\mathbf{V}}_{\text{ech1}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \hat{\mathbf{V}}_{D}^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{V}}_{M}^{\text{NR}}\left(\hat{t}_{Y\star}\right) - \hat{\mathbf{V}}_{E}^{\text{NR}}\left(\hat{t}_{Y\star}\right) \tag{14}$$

$$\hat{\mathbf{V}}_{\text{ech2}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{g=1}^{G}\left(\frac{N_{Mg}}{n_{Mg}}\right)^2 \sum_{h=1}^{H}\left(\frac{N_{Dh}}{n_{Dh}}\right)^2 \sum_{f=1}^{F}\sum_{a\in S_{R_f}}\left(1-\frac{n_{Mg}n_{Dh}}{N_{Mg}N_{Dh}}\right)\frac{1-\hat{p}_f}{\hat{p}_f^2}{y_a}^2 \tag{15}$$

$$\hat{\mathbf{V}}_{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{f=1}^{F}\sum_{a\in S_{R_f}}\frac{1-\hat{p}_f}{\hat{p}_f^2}\left(\check{y}_a - \frac{1}{n_{R_f}}\sum_{b\in S_{R_f}}\check{y}_b\right)^2 \quad \text{avec} \quad \check{y}_a = \frac{y_a}{\pi_i^M \pi_k^D}, \tag{16}$$

avec

$$\hat{\mathbf{V}}_{D}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{h=1}^{H}(N_{Dh})^2\left(\frac{1}{n_{Dh}}-\frac{1}{N_{Dh}}\right)s^2_{\hat{\hat{Y}}_{\bullet\circ,h}} \tag{17}$$

$$\hat{\mathbf{V}}_{M}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{g=1}^{G}(N_{Mg})^2\left(\frac{1}{n_{Mg}}-\frac{1}{N_{Mg}}\right)s^2_{\hat{\hat{Y}}_{\circ\bullet,g}} \tag{18}$$

$$\hat{\mathbf{V}}_{E}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{g=1}^{G}(N_{Mg})^2\left(\frac{1}{n_{Mg}}-\frac{1}{N_{Mg}}\right)\sum_{h=1}^{H}(N_{Dh})^2\left(\frac{1}{n_{Dh}}-\frac{1}{N_{Dh}}\right)\frac{1}{(n_{Mg}-1)(n_{Dh}-1)}s_{\hat{E},hg} \tag{19}$$

et

$$s^2_{\hat{\hat{Y}}_{\bullet\circ,h}} = \frac{1}{n_{Dh}-1}\sum_{k\in S_{Dh}}\left(\hat{\hat{Y}}_{\bullet k} - \frac{1}{n_{Dh}}\sum_{l\in S_{Dh}}\hat{\hat{Y}}_{\bullet l}\right)^2, \tag{20}$$

$$s^2_{\hat{\hat{Y}}_{\circ\bullet,g}} = \frac{1}{n_{Mg}-1}\sum_{i\in S_{Mg}}\left(\hat{\hat{Y}}_{i\bullet} - \frac{1}{n_{Mg}}\sum_{j\in S_{Mg}}\hat{\hat{Y}}_{j\bullet}\right)^2, \tag{21}$$

$$s_{\hat{E},hg} = \sum_{i\in S_{Mg}}\sum_{k\in S_{Dh}}\left[\hat{Y}_{ik} - \frac{1}{n_{Mg}}\sum_{j\in S_{Mg}}\hat{Y}_{jk} - \frac{1}{n_{Dh}}\sum_{l\in S_{Dh}}\hat{Y}_{il} + \frac{1}{n_{Mg}}\frac{1}{n_{Dh}}\sum_{j\in S_{Mg}}\sum_{l\in S_{Dh}}\hat{Y}_{jl}\right]^2. \tag{22}$$

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 avec probabilités de réponse estimées 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. 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 (2) (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.

6 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 étudiés pour différentes raisons :

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

6.1 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 (18),

$$\hat{\mathbf{V}}_{\text{SIMP1}} \equiv \hat{\mathbf{V}}_M^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{g=1}^{G}(N_{Mg})^2\left(\frac{1}{n_{Mg}} - \frac{1}{N_{Mg}}\right)s^2_{\hat{\hat{Y}}_{\circ\bullet,g}}, \tag{23}$$

- le deuxième estimateur correspond à une partie de l'estimateur sans biais, représentant la variance estimée inter-jours en formule (17),

$$\hat{\mathbf{V}}_{\text{SIMP2}} \equiv \hat{\mathbf{V}}_D^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{h=1}^{H}(N_{Dh})^2\left(\frac{1}{n_{Dh}} - \frac{1}{N_{Dh}}\right)s^2_{\hat{\hat{Y}}_{\bullet\circ,h}}, \tag{24}$$

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

$$\begin{aligned}\hat{\mathbf{V}}_{\text{SIMP3}} &\equiv \hat{\mathbf{V}}_{\text{SIMP1}} + \hat{\mathbf{V}}_{\text{SIMP2}} \\ &= \hat{\mathbf{V}}_D^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{V}}_M^{\text{NR}}\left(\hat{t}_{Y\star}\right),\end{aligned} \tag{25}$$

- 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),

$$\begin{aligned}\hat{\mathbf{V}}_{\text{SIMP4}} &\equiv \hat{\mathbf{V}}_M^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \sum_{g=1}^{G}\frac{N_{Mg}}{n_{Mg}}\sum_{i\in S_{Mg}}\sum_{h=1}^{H}N_{Dh}^2\left(\frac{1}{n_{Dh}} - \frac{1}{N_{Dh}}\right)s^2_{\hat{Y}_{i\circ,h}} \\ \text{avec} \quad s^2_{Y_{i\circ,h}} &= \frac{1}{n_{Dh}-1}\sum_{k\in S_{Dh}}(\hat{Y}_{ik} - \frac{1}{n_{Dh}}\sum_{l\in S_{Dh}}\hat{Y}_{il})^2,\end{aligned} \tag{26}$$

- 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,

$$\begin{aligned}\hat{\mathbf{V}}_{\text{SIMP5}} &\equiv \hat{\mathbf{V}}_D^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \sum_{h=1}^{H}\frac{N_{Dh}}{n_{Dh}}\sum_{k\in S_{Dh}}\sum_{g=1}^{G}N_{Mg}^2\left(\frac{1}{n_{Mg}} - \frac{1}{N_{Mg}}\right)s^2_{\hat{Y}_{\circ k,g}} \\ \text{avec} \quad s^2_{Y_{\circ k,g}} &= \frac{1}{n_{Mg}-1}\sum_{i\in S_{Mg}}(\hat{Y}_{ik} - \frac{1}{n_{Mg}}\sum_{j\in S_{Mg}}\hat{Y}_{jk})^2.\end{aligned} \tag{27}$$

Ces estimateurs simplifiés sont positifs et calculables à partir de procédures déjà programmées (voir Tableau 3) mais ne sont **pas sans biais**. Dans le paragraphe suivant, ces cinq estimateurs sont comparés.

Estimateur simplifié	Logiciels
$\hat{\mathbf{V}}_{\text{SIMP1}}$	R/SAS/Stata
$\hat{\mathbf{V}}_{\text{SIMP2}}$	R/SAS/Stata
$\hat{\mathbf{V}}_{\text{SIMP3}}$	R/SAS/Stata
$\hat{\mathbf{V}}_{\text{SIMP4}}$	R/Stata
$\hat{\mathbf{V}}_{\text{SIMP5}}$	R/Stata

TABLEAU 3 – Procédures logicielles R/SAS/Stata et estimateurs simplifiés

6.2 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 4, pour chacune des variables Elfe choisie, on calcule le total $\hat{t}_{Y\star}$ donné en formule (10), sa variance estimée $\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)$ donnée en formule (11), ainsi que chaque partie qui la compose : $\hat{\mathbf{V}}_{\text{ech1}}^{\text{NR}}\left(\hat{t}_{Y\star}\right)$ en (14), $\hat{\mathbf{V}}_{\text{ech2}}^{\text{NR}}\left(\hat{t}_{Y\star}\right)$ en (15) et $\hat{\mathbf{V}}_{\text{NR}}\left(\hat{t}_{Y\star}\right)$ en (16). Il est à noter que les résultats affichés ne tiennent pas compte de l'étape de calage qui est prise en compte dans les poids livrés aux utilisateurs. L'utilisateur ne peut donc retomber (exactement) sur les mêmes résultats à partir des variables de sa base de données. Le calage sera pris

en compte dans la prochaine section.

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

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

On constate dans le Tableau 4 que la part de variance estimée due à la non-réponse $\hat{\mathbf{V}}_{\text{NR}}$ est faible comparée à celle d'échantillonnage $\hat{\mathbf{V}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right)$. 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 $\hat{\mathbf{V}}_{\text{SIMP3}}$, 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 $\hat{\mathbf{V}}_{\text{SIMP3}}$ 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 $\hat{\mathbf{V}}_{\text{SIMP5}}$ 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 $\hat{\mathbf{V}}_{\text{SIMP1}}$ et $\hat{\mathbf{V}}_{\text{SIMP4}}$ s'avèrent inacceptables avec jusqu'à -95 % d'erreur relative.

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

Modélisation STSI × STSI, NR	Nombre de naissances	Nombre de naissances sous césarienne en début du travail	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
$\hat{t}_{Y\star}$	7.5×10^5	7.4×10^4	9.8×10^4	3.3×10^5	3.3×10^5
$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right) = \hat{\mathbf{v}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{v}}_{\text{NR}}\left(\hat{t}_{Y\star}\right)$	2.9×10^8	7.1×10^7	1.9×10^7	6.2×10^7	8.9×10^7
$\hat{\mathbf{V}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right)$	2.8×10^8	6.7×10^7	1.3×10^7	5.2×10^7	7.8×10^7
$\hat{\mathbf{V}}_{\text{NR}}\left(\hat{t}_{Y\star}\right)$	8.1×10^6	4.2×10^6	6.1×10^6	1.0×10^7	1.1×10^7
$\hat{\mathbf{V}}_{\text{SIMP1}}\left(\hat{t}_{Y\star}\right)$ (ER)	8.9×10^7 (-69 %)	3.7×10^6 (-95 %)	1.4×10^7 (-29 %)	2.3×10^7 (-63 %)	2.4×10^7 (-73 %)
$\hat{\mathbf{V}}_{\text{SIMP2}}\left(\hat{t}_{Y\star}\right)$ (ER)	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 %)
$\hat{\mathbf{V}}_{\text{SIMP3}}\left(\hat{t}_{Y\star}\right)$ (ER)	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 %)
$\hat{\mathbf{V}}_{\text{SIMP4}}\left(\hat{t}_{Y\star}\right)$ (ER)	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 %)
$\hat{\mathbf{V}}_{\text{SIMP5}}\left(\hat{t}_{Y\star}\right)$ (ER)	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 préparation	Nombre de nourrissons ayant une mère étrangère ou apatride	Nombre de jumeaux
$\hat{t}_{Y\star}$	1.2×10^5	8.1×10^4	3.6×10^5	9.2×10^4	2.4×10^4
$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)$ $\hat{\mathbf{v}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{v}}_{\text{NR}}\left(\hat{t}_{Y\star}\right)$	2.1×10^7	1.6×10^7	6.4×10^7	2.5×10^7	4.5×10^6
$\hat{\mathbf{V}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right)$	1.4×10^7	1.1×10^7	5.6×10^7	1.8×10^7	3.4×10^6
$\hat{\mathbf{V}}_{\text{NR}}\left(\hat{t}_{Y\star}\right)$	7.0×10^6	5.0×10^6	7.9×10^6	6.6×10^6	1.1×10^6
$\hat{\mathbf{V}}_{\text{SIMP1}}\left(\hat{t}_{Y\star}\right)$ (ER)	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 %)
$\hat{\mathbf{V}}_{\text{SIMP2}}\left(\hat{t}_{Y\star}\right)$ (ER)	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 %)
$\hat{\mathbf{V}}_{\text{SIMP3}}\left(\hat{t}_{Y\star}\right)$ (ER)	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 %)
$\hat{\mathbf{V}}_{\text{SIMP4}}\left(\hat{t}_{Y\star}\right)$ (ER)	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 %)
$\hat{\mathbf{V}}_{\text{SIMP5}}\left(\hat{t}_{Y\star}\right)$ (ER)	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 4 – Comparaison entre différents estimateurs simplifiés et l'estimateur sans biais.

Lecture du Tableau 4 : On estime à 7.5×10^5 le nombre total de naissance dans la population (définie) avec une variance estimée à 2.9×10^8. L'estimateur de variance simplifié préconisé vaut 3.3×10^8, soit une sur-estimation de 14 % par rapport à l'estimateur sans biais.

7 Estimation de la variance avec prise en compte du calage

Le calage est une méthode permettant d'intégrer des informations connues sur l'ensemble de la population, après que l'enquête ait eu lieu. La méthode consiste à modifier les poids de sondage en utilisant certaines équations respectant certaines contraintes (méthode détaillée dans Deville et Särndal, 1992). Cette modification impacte les estimateurs (paramètre **et** variance), cela peut diminuer le biais et améliorer la précision si la variable d'intérêt est liée aux variables de calage.

7.1 Calage dans Elfe

Le choix du vecteur de variables de calage s'est porté sur *Classe d'âge, Groupe de régions, Etat matrimonial, Statut immigré, Niveau d'étude* et *Primiparité*, permettant un calage caractérisant la situation familiale, géographique et socio-demographique. Pour plus de détails concernant le découpage de ces variables catégorielles, voir Juillard *et al.*, 2015. Notons que le calage ne prenant pas en compte les données manquantes, les variables de calage avaient été imputées (à petits taux) et que les poids calés ont subi une phase de troncature afin de limiter leur dispersion. Les sources de calage pour l'enquête Elfe sont l'état civil et l'enquête nationale périnatale (ENP) 2010.
On note w_a le poids calé associé à l'individu a de l'échantillon des répondants S_R. Le total t_Y est estimé approximativement sans biais par

$$\hat{t}_{Y_c} = \sum_{a \in S_R} w_a y_a, \tag{28}$$

L'idée du calage est de réduire la variance associée aux estimateurs calés : plus la variable d'intérêt y sera corrélée aux variables de calage, meilleure sera la précision.

7.2 Estimation de la variance après calage

Pour calculer l'estimateur de variance, on s'intéresse aux résidus estimés e de la régression (pondérée) de notre variable d'intérêt y sur les variables de calage x :

$$e_a = y_a - \hat{b} x_a \tag{29}$$

où $\hat{b} = \left(\sum_{a \in S_R} d_a x_a x_a^{\mathrm{T}}\right)^{-1} \sum_{a \in S_R} d_a x_a y_a$ avec d_a, les poids de sondage corrigés de la non-réponse avant calage. L'estimateur de la variance après calage est obtenu en remplaçant chaque y_a par son résidu associé e_a dans la formule (11). On comprend alors pourquoi la variance estimée sera d'autant plus faible que les variables y et x sont liées.

Dans le Tableau 5, afin de comparer les résultats avant et après calage, l'estimateur avant calage a été ajusté sur le nombre total de naissance (764000). Le calage permet bien de diminuer l'estimation de la variance et ceci d'autant plus que les variables d'intérêt sont corrélées aux variables de calage : ceci est flagrant pour les variables *Nombre de nourrissons ayant une mère primipare*, *Nombre de nourrissons ayant une mère mariée ou remariée* et *Nombre de nourrissons ayant une mère âgée entre 18 et 25 ans* qui correspondent respectivement à des modalités des variables de calage *Primiparité*, *Etat matrimonial* et *Classe d'âge*. Notons cependant qu'il ne s'agit pas exactement des variables de calage, qui avaient été imputées, mais des variables brutes. Les véritables variables de calage donnent une estimation de la variance approximativement nulle.

Lecture du Tableau 5 : Sans calage, on estime à 9.8×10^4 le nombre total de nourrissons ayant une mère suivie par une sage-femme dans la population (définie) avec une variance estimée à 2.2×10^7. Après calage, on estime cette variance à 1.2×10^7 et l'estimateur de variance simplifié préconisé vaut 1.3×10^7, soit une sur-estimation de 11 % par rapport à l'estimateur sans biais calé.

8 Estimation de la variance pour une statistique plus complexe

Les parties précédentes concernent l'estimation de la variance d'un total estimé. Pour d'autres paramètres, tels qu'un ratio ou un coefficient de corrélation, la méthode de linéarisation (Deville, 1999) peut être utilisée afin de pouvoir estimer leurs variances.

	Nombre de naissances	Nombre de naissances sous césarienne en début de travail	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
$\hat{t}_{Y\star}$	7.6×10^5	7.5×10^4	9.8×10^4	3.4×10^5	3.4×10^5
$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)$	$3{,}1 \times 10^8$	$7{,}4 \times 10^7$	$2{,}2 \times 10^7$	$6{,}8\text{E} \times 10^7$	$9{,}7 \times 10^7$
$\hat{t}_{Y_c}$	7.6×10^5	7.5×10^4	9.6×10^4	3.2×10^5	3.3×10^5
$\hat{\mathbf{V}}\left(\hat{t}_{Y_c}\right)$		$4{,}7 \times 10^7$	$1{,}2 \times 10^7$	$4{,}2 \times 10^5$	$6{,}6 \times 10^5$
$\hat{\mathbf{V}}_{\text{SIMP1}}\left(\hat{t}_{Y_c}\right)$ (ER)		$1{,}7 \times 10^6$ (-96 %)	$9{,}4 \times 10^6$ (-21 %)	$1{,}4 \times 10^5$ (-66 %)	$2{,}4 \times 10^5$ (-64 %)
$\hat{\mathbf{V}}_{\text{SIMP2}}\left(\hat{t}_{Y_c}\right)$ (ER)		$4{,}6 \times 10^7$ (-01 %)	$3{,}8 \times 10^6$ (-68 %)	$3{,}7 \times 10^5$ (-11 %)	$5{,}6 \times 10^5$ (-16 %)
$\hat{\mathbf{V}}_{\text{SIMP3}}\left(\hat{t}_{Y_c}\right)$ (ER)		$4{,}8 \times 10^7$ (02 %)	$1{,}3 \times 10^7$ (11 %)	$5{,}2 \times 10^5$ (22 %)	$8{,}0 \times 10^5$ (20 %)
$\hat{\mathbf{V}}_{\text{SIMP4}}\left(\hat{t}_{Y_c}\right)$ (ER)		$5{,}7 \times 10^6$ (-88 %)	$1{,}4 \times 10^7$ (17 %)	$6{,}2 \times 10^5$ (47 %)	$7{,}1 \times 10^5$ (07 %)
$\hat{\mathbf{V}}_{\text{SIMP5}}\left(\hat{t}_{Y_c}\right)$ (ER)		$4{,}7 \times 10^7$ (01 %)	$4{,}9 \times 10^6$ (-59 %)	$4{,}3 \times 10^5$ (03 %)	$6{,}9 \times 10^5$ (04 %)

	Nombre de nourrissons ayant une mère âgée entre 18 et 25 ans	Nombre de nourrissons ayant une mère avec un IMC supérieur à 30	Nombre de nourrissons ayant mère ayant suivi des séances de préparation	Nombre de nourrissons ayant une mère étrangère ou apatride	Nombre de jumeaux
$\hat{t}_{Y\star}$	1.2×10^5	8.2×10^4	3.7×10^5	9.6×10^4	2.5×10^4
$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)$	$2{,}1 \times 10^7$	$1{,}8 \times 10^7$	$6{,}8 \times 10^7$	$2{,}7 \times 10^7$	$4{,}6 \times 10^6$
$\hat{t}_{Y_c}$	1.1×10^5	8.2×10^4	3.7×10^5	9.8×10^4	2.5×10^4
$\hat{\mathbf{V}}\left(\hat{t}_{Y_c}\right)$	$2{,}4 \times 10^4$	$5{,}7 \times 10^6$	$1{,}4 \times 10^7$	$3{,}7 \times 10^6$	$3{,}3 \times 10^6$
$\hat{\mathbf{V}}_{\text{SIMP1}}\left(\hat{t}_{Y_c}\right)$ (ER)	$9{,}3 \times 10^3$ (-62 %)	$2{,}0 \times 10^6$ (-65 %)	$5{,}6 \times 10^6$ (-59 %)	$7{,}5 \times 10^5$ (-80 %)	$1{,}3 \times 10^6$ (-61 %)
$\hat{\mathbf{V}}_{\text{SIMP2}}\left(\hat{t}_{Y_c}\right)$ (ER)	$2{,}3 \times 10^4$ (-05 %)	$5{,}0 \times 10^6$ (-12 %)	$1{,}1 \times 10^7$ (-24 %)	$3{,}5 \times 10^6$ (-05 %)	$2{,}8 \times 10^6$ (-16 %)
$\hat{\mathbf{V}}_{\text{SIMP3}}\left(\hat{t}_{Y_c}\right)$ (ER)	$3{,}3 \times 10^4$ (33 %)	$7{,}0 \times 10^6$ (23 %)	$1{,}6 \times 10^7$ (17 %)	$4{,}3 \times 10^6$ (15 %)	$4{,}1 \times 10^6$ (23 %)
$\hat{\mathbf{V}}_{\text{SIMP4}}\left(\hat{t}_{Y_c}\right)$ (ER)	$3{,}5 \times 10^4$ (43 %)	$5{,}9 \times 10^6$ (02 %)	$1{,}2 \times 10^7$ (-13 %)	$3{,}0 \times 10^6$ (-18 %)	$3{,}7 \times 10^6$ (12 %)
$\hat{\mathbf{V}}_{\text{SIMP5}}\left(\hat{t}_{Y_c}\right)$ (ER)	$2{,}7 \times 10^4$ (10 %)	$6{,}0 \times 10^6$ (06 %)	$1{,}2 \times 10^7$ (-12 %)	$3{,}9 \times 10^6$ (05 %)	$3{,}3 \times 10^6$ (-02 %)

TABLEAU 5 – Comparaison entre différents estimateurs simplifiés et l'estimateur sans biais avant et après calage.

Prenons l'exemple d'un ratio $R = t_Y/t_X$, il peut simplement s'estimer par $\hat{R}_c = \hat{t}_{Y_c}/\hat{t}_{X_c}$. Pour l'estimation de sa variance, il est nécessaire d'estimer la linéarisée l_a de notre paramètre ratio (les calculs pour plusieurs paramètres sont expliqués dans Dell *et al.*, 2002) :

$$\hat{l}_a = \frac{1}{\hat{t}_{X_c}} \left(y_a - \hat{R}_c x_a \right). \tag{30}$$

Ensuite, pour prendre en compte l'étape de calage, il faut (tout comme dans la section précédente) estimer les résidus e_a de la régression des $\hat{l}_a$ sur les variables de calage et de nouveau remplacer les y_a de la formule (11) par ces résidus.

Le Tableau 6 permet de comparer les résultats avant et après calage pour différents ratios estimés. Tout comme observé pour le paramètre total dans la section précédente, on remarque que la variance estimée du ratio diminue en prenant en compte le calage, et ceci d'autant plus que la variable d'intérêt est liée aux variables de calage. Concernant les différents estimateurs simplifiés, les mêmes commentaires faits pour le cas d'un total en sous-section 6.2 peuvent s'appliquer au ratio. Seul $\hat{\mathbf{V}}_{\text{SIMP3}}$ s'avère intéressant avec des ER relativement faibles et peu variables.

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

Lecture du Tableau 6 : Sans calage, on estime à 0.44 le pourcentage de nourrissons ayant une mère suivie par une sage-femme dans la population (définie) avec une variance estimée à $3.3 \times 10^{-}5$. Après calage, on estime cette variance à $7.2 \times 10^{-}7$ et l'estimateur de variance simplifié préconisé vaut $8.8 \times 10^{-}7$, soit une sur-estimation de 22 % par rapport à l'estimateur sans biais calé.

9 Procédures logicielles (SAS/R/Stata)

Les codes proposés concernent trois logiciels couramment utilisés : R 3.2.2 (R Core Team, 2015), SAS 9.4 (SAS Institute Inc., 2013), Stata 13.1 (StataCorp., 2013). Le logiciel R est disponible à partir du CRAN (Comprehensive R Archive Network, `http://CRAN.R-project.org/`).
L'estimation se déroule en trois étapes. En premier lieu, il est nécessaire d'estimer la linéarisée de votre paramètre. Si votre paramètre est un total, vous pouvez passer directement à l'étape 2. Sinon, il vous faut coder

	% de naissances	% de naissances sous césarienne en début de travail	% de nourrissons ayant une mère suivie par sage-femme	% de nourrissons ayant une mère primipare	% de nourrissons ayant une mère mariée ou remariée
$\hat{R}_\star$	1,000	0,098	0,128	0,440	0,444
$\hat{\mathbf{V}}\left(\hat{R}_\star\right)$		$8{,}8 \times 10^{-5}$	$2{,}9 \times 10^{-5}$	$3{,}3 \times 10^{-5}$	$4{,}4 \times 10^{-5}$
$\hat{R}_c$		0,098	0,126	0,427	0,439
$\hat{\mathbf{V}}\left(\hat{R}_c\right)$		$8{,}0 \times 10^{-5}$	$2{,}1 \times 10^{-5}$	$7{,}2 \times 10^{-7}$	$1{,}1 \times 10^{-6}$
$\hat{\mathbf{V}}_{\text{SIMP1}}\left(\hat{R}_c\right)$ (ER)		$3{,}0 \times 10^{-6}$ (-96 %)	$1, \times 10^{-5}$ (-21 %)	$2{,}4 \times 10^{-7}$ (-66 %)	$4{,}1 \times 10^{-7}$ (-64 %)
$\hat{\mathbf{V}}_{\text{SIMP2}}\left(\hat{R}_c\right)$ (ER)		$7{,}9 \times 10^{-5}$ (-01 %)	$6{,}6 \times 10^{-6}$ (-68 %)	$6{,}4 \times 10^{-7}$ (-11 %)	$9{,}5 \times 10^{-7}$ (-16 %)
$\hat{\mathbf{V}}_{\text{SIMP3}}\left(\hat{R}_c\right)$ (ER)		$8{,}2 \times 10^{-5}$ (02 %)	$2{,}3 \times 10^{-5}$ (11 %)	$8{,}8 \times 10^{-7}$ (22 %)	$1{,}4 \times 10^{-6}$ (20 %)
$\hat{\mathbf{V}}_{\text{SIMP4}}\left(\hat{R}_c\right)$ (ER)		$9{,}7 \times 10^{-6}$ (-88 %)	$2{,}4 \times 10^{-5}$ (17 %)	$1{,}1 \times 10^{-6}$ (47 %)	$1{,}2 \times 10^{-6}$ (07 %)
$\hat{\mathbf{V}}_{\text{SIMP5}}\left(\hat{R}_c\right)$ (ER)		$8{,}0 \times 10^{-5}$ (01 %)	$8{,}5 \times 10^{-6}$ (-59 %)	$7{,}4 \times 10^{-7}$ (03 %)	$1{,}2 \times 10^{-6}$ (04 %)
	% de nourrissons ayant une mère âgée entre 18 à 25 ans	% de nourrissons ayant une mère avec un IMC supérieur à 30	% de nourrissons ayant mère ayant suivi des séances de préparation	% de nourrissons ayant une mère étrangère ou apatride	% de jumeaux
$\hat{R}_\star$	0,153	0,107	0,481	0,125	0,032
$\hat{\mathbf{V}}\left(\hat{R}_\star\right)$	$2{,}1 \times 10^{-5}$	$1{,}7 \times 10^{-5}$	$4{,}8 \times 10^{-5}$	$3{,}5 \times 10^{-5}$	$6{,}3 \times 10^{-6}$
$\hat{R}_c$	0,139	0,108	0,480	0,128	0,033
$\hat{\mathbf{V}}\left(\hat{R}_c\right)$	$4{,}2 \times 10^{-8}$	$9{,}8 \times 10^{-6}$	$2{,}4 \times 10^{-5}$	$6{,}4 \times 10^{-6}$	$5{,}7 \times 10^{-6}$
$\hat{\mathbf{V}}_{\text{SIMP1}}\left(\hat{R}_c\right)$ (ER)	$1{,}6 \times 10^{-8}$ (-62 %)	$3{,}4 \times 10^{-6}$ (-65 %)	$9{,}6 \times 10^{-6}$ (-59 %)	$1{,}3 \times 10^{-6}$ (-80 %)	$2{,}2 \times 10^{-6}$ (-61 %)
$\hat{\mathbf{V}}_{\text{SIMP2}}\left(\hat{R}_c\right)$ (ER)	$4{,}0 \times 10^{-8}$ (-05 %)	$8{,}6 \times 10^{-6}$ (-12 %)	$1{,}8 \times 10^{-5}$ (-24 %)	$6{,}1 \times 10^{-6}$ (-05 %)	$4{,}8 \times 10^{-6}$ (-16 %)
$\hat{\mathbf{V}}_{\text{SIMP3}}\left(\hat{R}_c\right)$ (ER)	$5{,}6 \times 10^{-8}$ (33 %)	$1{,}2 \times 10^{-5}$ (23 %)	$2{,}8 \times 10^{-5}$ (17 %)	$7{,}4 \times 10^{-6}$ (15 %)	$7{,}0 \times 10^{-6}$ (23 %)
$\hat{\mathbf{V}}_{\text{SIMP4}}\left(\hat{R}_c\right)$ (ER)	$6{,}0 \times 10^{-8}$ (43 %)	$1{,}0 \times 10^{-5}$ (02 %)	$2{,}1 \times 10^{-5}$ (-13 %)	$5{,}2 \times 10^{-6}$ (-18 %)	$6{,}4 \times 10^{-6}$ (12 %)
$\hat{\mathbf{V}}_{\text{SIMP5}}\left(\hat{R}_c\right)$ (ER)	$4{,}6 \times 10^{-8}$ (10 %)	$1{,}0 \times 10^{-5}$ (06 %)	$2{,}1 \times 10^{-5}$ (-12 %)	$6{,}7 \times 10^{-6}$ (05 %)	$5{,}6 \times 10^{-6}$ (-02 %)

TABLEAU 6 – Comparaison entre différents estimateurs simplifiés de variance d'un ratio estimé et l'estimateur sans biais avant et après calage.

la formule de la linéarisée, vous trouverez par exemple dans Dell *et al.* (2002) la formule pour l'indice de Gini et dans la suite du document sera traité le cas du ratio (moyenne, proportion...). Ensuite, il vous faudra effectuer une régression de cette linéarisée sur les variables de calage et récupérer les résidus : ceci permet de prendre en compte l'étape de calage censée diminuer la variance si votre variable est corrélée aux variables de calage. En dernier lieu, insérer ces résidus dans les procédures logicielles proposées pour estimer la variance (il s'agit de l'estimateur simplifié $\hat{\mathbf{V}}_{\text{SIMP3}}$ décrit précédemment).
Notons que pour des calculs corrects, il faut prendre en compte tous les croisements jour × maternité existants, c'est-à-dire 25 jours × 320 maternités = 8000 croisements. Puisque certains croisements n'apparaissent pas dans la base de données (puisqu'il n'y a pas eu de naissances Elfe ce jour-là dans cette maternité-là), il faut ajouter une ligne pour chacun de ces croisements avec un 0 pour la variable d'intérêt. En effet, ces 0 modifient les estimations de variance.

Pour chacun des codes proposés, en Tableau 7 sont listés les noms des variables de la base de données Elfe :

VARIABLES A NE PAS OUBLIER LORS DE LA DEMANDE D'ACCES	
Pondération en maternité	M00E_PONDVALC2 ou M00F_PONDVALC2
Strates pour le plan sur les jours	M00M1_VAGUE
Identifiant du jour	M00M2_JNAISSEALEA
Strates pour le plan sur les maternités	M00M1_MATSTRATEC1
Identifiant de la maternité	M00M1_IDGROUPNAMEALEAC1 ou M00M1_IDGROUPNAMEALEAC1B
Variables de calage	CAL1 CAL2 CAL3 CAL4 CAL5 CAL6

TABLEAU 7 – Liste des variables du plan de sondage Elfe nécessaires pour estimer la variance

9.1 Logiciel R

9.1.1 1ère étape : linéarisation du paramètre

Pour le cas d'un ratio t_{num}/t_{den} (moyenne, proportion...), le code permet de récupérer la linéarisée notée *lin* qui sera utilisée à la deuxième étape :

Listing 1 – Code R

```
########################################
#DESCRIPION: estimates the linearized variable of a ratio Y2/Y1
#USAGE: LINratio(w,Y1,Y2)
#ARGUMENTS:
#                Y2       vector of the numerator variable
#                Y1       vector of the denominator variable
#                w        vector of the weights
#VALUE: the function returns a vector
LINratio <- function(w,Y1,Y2)
        {rpi=EstTOTAL(w,Y2)/EstTOTAL(w,Y1)
        txpi=EstTOTAL(w,Y1)
        upi = (Y2 - Y1*rpi) / txpi ; return(upi)
        }

########################################
#DESCRIPION: estimates a population total
#USAGE: EstTOTAL(w,Y1)
#ARGUMENTS:
#                Y1       vector of the interest variable
#                w        vector of the weights
#VALUE: the function returns a numeric
EstTOTAL <- function(w,Y1)
                {t=sum(Y1*w) ; return(t)}

lin=LINratio(w=M00E_PONDVALC2,Y1=denominator,Y2=numerator)
```

9.1.2 2ème étape : régression sur les variables de calage

Pour cette étape, il suffit de prendre la linéarisée *lin* issue de la première étape et de l'injecter dans la régression (pondérée) sur les variables de calage. On récupère les résidus *res* de cette régression pour la prochaine étape.

Listing 2 – Code R

```
########################################
#DESCRIPION: estimates a population total
#USAGE: REScalib(w,Y,X1,X2,X3,X4,X5,X6)
#ARGUMENTS:
#                Y1       vectors of the auxiliary variables
#                w        vector of the weights before calibration
#VALUE: the function returns a vector
REScalib <- function(w,Y,X1,X2,X3,X4,X5,X6)
        {modele=lm(Y ~ X1+X2+X3+X4+X5+X6,weights=w,na.action=na.exclude)
        e=residuals(modele)
        return(e)
        }

#pour le ratio
res=REScalib(w=M00E_PONDVALC2,Y=lin,X1=CAL1,X2=CAL2,X3=CAL3,X4=CAL4,X5=CAL5,X6=CAL6)

# Remplacer lin par votre variable d'intérêt si vous voulez estimer un total.
```

9.1.3 3ème étape : estimation de la variance

Dans cette section sont décrites les procédures logicielles à sommer pour pouvoir calculer $\hat{\mathbf{V}}_{\text{SIMP3}}$. L'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ en formule (25), recommandé aux utilisateurs, est la somme de deux termes calculables séparément. Les résidus *res* calculés dans l'étape précédente sont insérés dans la formule de l'estimateur de variance d'un total. Attention, si la somme des deux termes correspond bien à l'estimation de la variance, le total estimé, lui, correspond au total estimé des résidus. Il faut donc calculer séparément votre paramètre (total, ratio...). Pour le logiciel R, différents packages sont possibles, voici les procédures utilisant le package *survey* (Lumley, 2014) :

Listing 3 – Code R

```
library(survey)

fpcm=c()
fpcm[M00M1_MATSTRATEC1==1] <- 108 ; fpcm[M00M1_MATSTRATEC1==2] <- 108
```

```
fpcm[M00M1_MATSTRATEC1==3] <- 109 ; fpcm[M00M1_MATSTRATEC1==4] <- 108
fpcm[M00M1_MATSTRATEC1==5] <- 111

infoplan<-svydesign(id=~M00M1_IDGROUPNAMEALEAC1, strata =~M00M1_MATSTRATEC1, fpc=~fpcm, weights=M00E_
    PONDVALC2)
infoplan

(Result <- svytotal(~res , infoplan, na.rm=TRUE))
vcov(Result)
SE(Result)^2

# Si vous utilisez la fonction confint(), l'intervalle de confiance n'est pas correct.
# Cette première fonction calcule le premier terme de l'estimateur de variance.

M00M2_JNAISSEALEA[M00M1_VAGUE==2 & M00M2_JNAISSEALEA==3] <- 7 ; M00M2_JNAISSEALEA[M00M1_VAGUE==2 & M00M2_
    JNAISSEALEA==4] <- 8 ; M00M2_JNAISSEALEA[M00M1_VAGUE==2 & M00M2_JNAISSEALEA==5] <- 9 ; M00M2_
    JNAISSEALEA[M00M1_VAGUE==2 & M00M2_JNAISSEALEA==6] <- 10 ; M00M2_JNAISSEALEA[M00M1_VAGUE==2 & M00M2_
    JNAISSEALEA==1] <- 5 ; M00M2_JNAISSEALEA[M00M1_VAGUE==2 & M00M2_JNAISSEALEA==2] <- 6 ;
M00M2_JNAISSEALEA[M00M1_VAGUE==3 & M00M2_JNAISSEALEA==1] <- 11 ; M00M2_JNAISSEALEA[M00M1_VAGUE==3 & M00M2_
    JNAISSEALEA==2] <- 12 ; M00M2_JNAISSEALEA[M00M1_VAGUE==3 & M00M2_JNAISSEALEA==3] <- 13 ; M00M2_
    JNAISSEALEA[M00M1_VAGUE==3 & M00M2_JNAISSEALEA==4] <- 14 ; M00M2_JNAISSEALEA[M00M1_VAGUE==3 & M00M2_
    JNAISSEALEA==5] <- 15 ; M00M2_JNAISSEALEA[M00M1_VAGUE==3 & M00M2_JNAISSEALEA==6] <- 16 ; M00M2_
    JNAISSEALEA[M00M1_VAGUE==3 & M00M2_JNAISSEALEA==7] <- 17 ;
M00M2_JNAISSEALEA[M00M1_VAGUE==4 & M00M2_JNAISSEALEA==1] <- 18 ; M00M2_JNAISSEALEA[M00M1_VAGUE==4 & M00M2_
    JNAISSEALEA==2] <- 19 ; M00M2_JNAISSEALEA[M00M1_VAGUE==4 & M00M2_JNAISSEALEA==3] <- 20 ; M00M2_
    JNAISSEALEA[M00M1_VAGUE==4 & M00M2_JNAISSEALEA==4] <- 21 ; M00M2_JNAISSEALEA[M00M1_VAGUE==4 & M00M2_
    JNAISSEALEA==5] <- 22 ; M00M2_JNAISSEALEA[M00M1_VAGUE==4 & M00M2_JNAISSEALEA==6] <- 23 ; M00M2_
    JNAISSEALEA[M00M1_VAGUE==4 & M00M2_JNAISSEALEA==7] <- 24 ; M00M2_JNAISSEALEA[M00M1_VAGUE==4 & M00M2_
    JNAISSEALEA==8] <- 25 ;

fpcd=c()
fpcd[M00M1_VAGUE==1] <- 91 ; fpcd[M00M1_VAGUE==2] <- 91
fpcd[M00M1_VAGUE==3] <- 91 ; fpcd[M00M1_VAGUE==4] <- 91

infoplan2<-svydesign(id=~M00M2_JNAISSEALEA, strata =~M00M1_VAGUE, fpc=~fpcd, weights=M00E_PONDVALC2)
infoplan2

(Result <- svytotal(~res , infoplan2, na.rm=TRUE))
vcov(Result)
SE(Result)^2

# Si vous utilisez la fonction confint(), l'intervalle de confiance n'est pas correct.
# Cette seconde fonction calcule le second terme de l'estimateur de variance.

# !! Les deux termes doivent être sommés afin d'obtenir l'estimateur recommandé V_SIMP3.
```

9.2 Logiciel SAS

9.2.1 1ère étape : linéarisation du paramètre

Pour le cas d'un ratio t_{num}/t_{den} (moyenne, proportion...), le code permet de récupérer la linéarisée notée *lin* qui sera utilisée à la deuxième étape :

Listing 4 – Code SAS

```
PROC MEANS DATA=table sum ;
VAR numerator denominator ;
WEIGHT M00E_PONDVALC2;
OUTPUT OUT = stat sum= totnum totden ;  /* récupère les totaux estimés du numérateur et du dénominateur */
RUN ;

DATA _null_; SET stat;
call symput('tnum', totnum);
call symput('tden', totden);
RUN;

DATA table ; SET table ;
lin = (numerator - (&tnum/&tden)*denominator) * (1/&tden) ;     /* formule de la linéarisée */
RUN;
```

9.2.2 2ème étape : régression sur les variables de calage

Pour cette étape, il suffit de prendre la linéarisée *lin* issue de la première étape et de l'injecter dans la régression (pondérée) sur les variables de calage. On récupère les résidus *res* de cette régression pour la prochaine étape.

Listing 5 – Code SAS

```
PROC GLM data=table noprint ;
```

```
CLASS CAL1 CAL2 CAL3 CAL4 CAL5 CAL6 ;
MODEL lin = CAL1 CAL2 CAL3 CAL4 CAL5 CAL6 ;
WEIGHT M00E_PONDVALC2 ;
OUTPUT OUT = table r=res ;
quit ;

/* Remplacer lin par votre variable d'intérêt si vous souhaitez estimer un total */
```

9.2.3 3ème étape : estimation de la variance

Dans cette section sont décrites les procédures logicielles à sommer pour pouvoir calculer $\hat{\mathbf{V}}_{\text{SIMP3}}$. L'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ en formule (25), recommandé aux utilisateurs, est la somme de deux termes calculables séparément. Les résidus *res* calculés dans l'étape précédente sont insérés dans la formule de l'estimateur de variance d'un total. Attention, si la somme des deux termes correspond bien à l'estimation de la variance, le total estimé, lui, correspond au total estimé des résidus. Il faut donc calculer séparément votre paramètre (total, ratio...). Les deux morceaux de l'estimateur peuvent se calculer à partir de la procédure *surveymeans* du logiciel SAS :

Listing 6 – Code SAS

```
DATA NBDegreM ;                          /* Tailles des strates de maternités */
input M00M1_MATSTRATEC1 _TOTAL_;
datalines ;
1 108
2 108
3 109
4 108
5 111
;

proc SURVEYMEANS data=table TOTAL=NBDegreM sum mean var varsum missing;
CLUSTER M00M1_IDGROUPNAMEALEAC1;
STRATA M00M1_MATSTRATEC1 ;
VAR res ;
WEIGHT M00E_PONDVALC2 ;
run;

/* L'intervalle de confiance que peut produire cette procédure n'est pas correct.
/* Cette 1ère procédure calcule le 1er terme de l'estimateur de variance dans la colonne "Var of Sum". */

DATA NBDegreD ;                          /* Tailles des strates de jours */
input M00M1_VAGUE _TOTAL_;
datalines ;
1 91
2 91
3 91
4 91
;

proc SURVEYMEANS data=table TOTAL=NBDegreD sum mean var varsum missing;
CLUSTER M00M2_JNAISSEALEA;
STRATA M00M1_VAGUE ;
VAR  res ;
WEIGHT M00E_PONDVALC2 ;
run;

/* L'intervalle de confiance que peut produire cette procédure n'est pas correct.
/* Cette 2nde procédure calcule le 2nd terme de l'estimateur de variance dans la colonne "Var of Sum". */


/* !! Les deux termes doivent être sommés afin d'obtenir l'estimateur recommandé V_SIMP3. */
```

9.3 Logiciel Stata

9.3.1 1ère étape : linéarisation du paramètre

Pour le cas d'un ratio t_{num}/t_{den} (moyenne, proportion...), le code permet de récupérer la linéarisée notée *lin* qui sera utilisée à la deuxième étape :

Listing 7 – Code Stata

```
. clear
* total estimé du numérateur
. egen tnum= sum(numerator*M00E_PONDVALC2)
* total estimé du dénominateur
. egen tden= sum(denominator*M00E_PONDVALC2)
* formule de la linéarisée lin
```

```
. gen lin = (numerator - (tnum/tden)*denominator) / tden
```

9.3.2 2ème étape : régression sur les variables de calage

Pour cette étape, il suffit de prendre la linéarisée *lin* issue de la première étape et de l'injecter dans la régression (pondérée) sur les variables de calage. On récupère les résidus *res* de cette régression pour la prochaine étape.

Listing 8 – Code Stata

```
. reg lin CAL1 CAL2 CAL3 CAL4 CAL5 CAL6 [weight=M00E_PONDVALC2]
* Récupérer les résidus res de la régression
. predict res, residuals

* Remplacer lin par votre variable d'intérêt si vous voulez estimer un total.
```

9.3.3 3ème étape : estimation de la variance

Dans cette section sont décrites les procédures logicielles à sommer pour pouvoir calculer $\hat{\mathbf{V}}_{\text{SIMP3}}$. L'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ en formule (25), recommandé aux utilisateurs, est la somme de deux termes calculables séparément. Les résidus *res* calculés dans l'étape précédente sont insérés dans la formule de l'estimateur de variance d'un total. Attention, si la somme des deux termes correspond bien à l'estimation de la variance, le total estimé, lui, correspond au total estimé des résidus. Il faut donc calculer séparément votre paramètre (total, ratio...).

Listing 9 – Code Stata

```
* Tailles des strates de maternités
. gen fpcm=108 if M00M1_MATSTRATEC1==1
. replace fpcm=108 if M00M1_MATSTRATEC1==2
. replace fpcm=109 if M00M1_MATSTRATEC1==3
. replace fpcm=108 if M00M1_MATSTRATEC1==4
. replace fpcm=111 if M00M1_MATSTRATEC1==5

. svyset M00M1_IDGROUPNAMEALEAC1 [pweight=M00E_PONDVALC2], strata(M00M1_MATSTRATEC1) fpc(fpcm)

. svy: total res

* L'intervalle de confiance n'est pas correct.
* Il faut prendre la valeur du "Std. Err." et la mettre au carré.

* Tailles des strates de jours
. gen fpcd=91 if M00M1_VAGUE==1
. replace fpcd=91 if M00M1_VAGUE==2
. replace fpcd=91 if M00M1_VAGUE==3
. replace fpcd=91 if M00M1_VAGUE==4

. svyset M00M2_JNAISSEALEA [pweight=M00E_PONDVALC2], strata(M00M1_VAGUE) fpc(fpcd)

. svy: total res

* L'intervalle de confiance n'est pas correct.
* Il faut prendre la valeur du "Std. Err." et la mettre au carré.

* !! Les deux termes doivent être sommés afin d'obtenir l'estimateur recommandé V_SIMP3.
```

A retenir :

La **population d'inférence** est celle des nourrissons nés durant l'année 2011 en maternité de France métropolitaine, issus d'un accouchement au plus gémellaire, hors grands prématurés, ayant une mère majeure, en mesure de donner un consentement éclairé notamment dans l'une des langues proposées et dont les parents ne résidaient pas temporairement en métropole.

Le plan de sondage utilisé par l'enquête Elfe est appelé **plan d'échantillonnage produit** (cross-classified sampling design) et résulte du croisement indépendant d'un plan de sondage sur la population des maternités et d'un plan de sondage sur la population des jours. Le plan sur les maternités est un plan stratifié (cinq strates relatives à la taille des maternités) et le plan sur les jours peut être modélisé par un plan stratifié (sur les saisons).

Un **estimateur sans biais** de la variance pour l'enquête Elfe a été présenté en formule (11).

Comme sa forme (complexe) demande une programmation spécifique, plusieurs **estimateurs simplifiés** et calculables avec des procédures logicielles déjà existantes ont été proposés et comparés : les résultats montrent que l'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ peut être recommandé aux utilisateurs pour une estimation de la variance simple, et peu biaisée. Les **codes R, SAS, Stata** pour cet estimateur sont proposés en section 9 ainsi que la démarche à suivre en trois étapes.

Les **variables des bases de données Elfe** nécessaires pour estimer la variance sont listées dans le Tableau 7.

Un travail est en cours concernant l'estimation de la variance dans le longitudinal (pour un panel avec processus de non-réponse monotone dans le temps).

Références

G. Chauvet, H. Juillard et A. Ruiz-Gazen : Estimation under cross-classified sampling with application to a chilhood survey. *A paraître dans Journal of the American Statistical Association*, 2016.

F. Dell, X. D'Haultfoeuille, P. Février et E. Massé : Mise en œuvre du calcul de variance par linéarisation. *Insee-Méthodes : Actes des Journées de Méthodologie Statistique*, 2002.

J.-C. Deville : Variance estimation for complex statistics and estimators : Linearization and residual techniques. *Survey Methodology*, 25(2):193–203, 1999.

J.-C. Deville et C.-E Särndal : Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382, 1992.

J. L. Eltinge et I. S. Yansaneh : 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, 1997.

H. Juillard : Two-dimensional sampling in practice. *En révision*, 2016.

H. Juillard, X. Thierry, N. Razafindratsima, A. Bringé et J.L. Lanoë : Pondérations de l'enquête Elfe en maternité. Rapport technique, 2015. URL https://pandora.vjf.inserm.fr/public/docs/ELFE_NoteDet0.pdf.

J. K. Kim et J. J. Kim : Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35:501–514, 2007.

T. Lumley : survey : analysis of complex survey samples, 2014. R package version 3.30.

E. Ohlsson : Cross-classified sampling. *Journal of Official Statistics*, 12(3):241–251, 1996.

R Core Team : *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/.

C.-E. Särndal, B. Swensson et J.H. Wretman : *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992.

SAS Institute Inc. : *SAS/STAT® 14.1 User's Guide*. Cary, NC, 2013. URL http://www.sas.com/.

C. J. Slinner : Cross-classified sampling : some estimation theory. *Statistics and Probability Letters*, 104:163–168, 2015.

StataCorp. : *Stata Statistical Software : Release 13*. StataCorp LP, College Station, TX, 2013. URL http://www.stata.com/.