Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes: l'adaptation à un monde en évolution

Étude du plan de sondage "produit", à partir de l'exemple de l'enquête Elfe

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

Résumé

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

Mots-clés: Estimation de variance, indépendance, plan à deux phases, plan à plusieurs degrés.

1. Motivation initiale

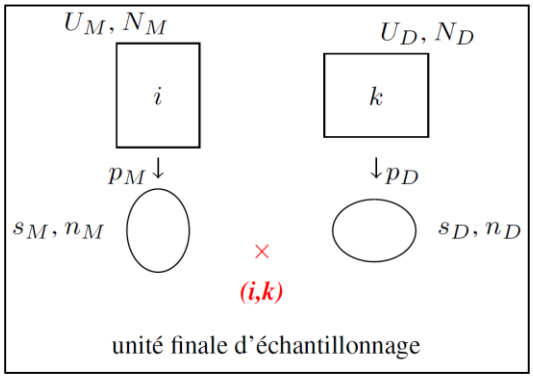
Elfe est une enquête longitudinale de type cohorte, comprenant 18 300 nourrissons à l'inclusion. Elle est consacrée au suivi des enfants, de la naissance à l'âge adulte, et aborde les multiples aspects de la vie de l'enfant sous l'angle des sciences sociales, de la santé et de la santé-environnement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

[1] Guillaume Chauvet, Ecole Nationale de la Statistique et de l'Analyse de l'Information (ENSAI), Campus de Ker-Lann, 35710 BRUZ cedex, France, chauvet@ensai.fr
[2] Hélène Juillard, Institut National d'Etudes Démographiques (INED), 133, boulevard Davout - 75020 Paris, France, helene.juillard@ined.fr
[3] Anne Ruiz-Gazen, Toulouse School of Economics (TSE), 21, allée de Brienne - 31015 Toulouse Cedex 6, France, anne.ruiz-gazen@tse-fr.eu

2. Echantillonnage produit indépendant

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

Figure 1
Échantillonnage dans la population produit



L'unité finale d'échantillonnage qui nous intéresse est caractérisée par un couple d'éléments (i,k), avec $i \in U_M$ et $k \in U_D$ (voir Figure 1). On s'intéresse à une variable Y prenant la valeur Y_{ik} dans la maternité i, le jour k. Dans l'enquête Elfe, on se référera donc à l'élément (i,k) comme *à* la « grappe des nourrissons nés dans la même maternité i, le même jour k ».

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

$$U = U_M \times U_D.$$

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

$$s = s_M \times s_D.$$

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

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

Remarquons que cette **hypothèse d'indépendance** entre deux tirages est analogue à l'hypothèse d'invariance utilisée dans l'échantillonnage classique à deux degrés (Särndal, Swensson et Wretman, 1992, page 134). On retrouve ce plan de sondage dans Vos (1964), comme un cas particulier des plans espace-temps.[4]

Sous ces conditions, on peut alors facilement calculer les probabilités d'inclusion d'ordres un et deux et les

[4] Les auteurs remercient les participants du Symposium de leur avoir notifié cette référence.

covariances relatives au plan produit à partir de celles de chaque plan source. Pour toutes les unités $i, j \in U_M$ et $k, l \in U_D$:

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

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

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

$$\hat{t}_Y = \sum_{i\in S_M}\sum_{k\in S_D}\frac{Y_{ik}}{\pi_i^M \pi_k^D} = \sum_{i\in S_M}\frac{\hat{Y}_{i\bullet}}{\pi_i^M} = \sum_{k\in S_D}\frac{\hat{Y}_{\bullet k}}{\pi_k^D} \tag{2}$$

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

$$V_{prod}\left(\hat{t}_Y\right) = \sum_{i,j\in U_M}\sum_{k,l\in U_D}\Gamma_{ijkl}\frac{Y_{ik}}{\pi_i^M \pi_k^D}\frac{Y_{jl}}{\pi_j^M \pi_l^D}. \tag{3}$$

On considère le cas particulier du plan SI $\times$ SI où $p_D(\cdot)$ est un sondage aléatoire simple sans remise (SI) de taille n_D au sein de la population U_D de taille N_D, et où $p_M(\cdot)$ est un sondage aléatoire simple sans remise de taille n_M au sein de la population U_M de taille N_M. En utilisant l'identité (1) donnant Γ_{ijkl}, la variance donnée en (3) peut se réécrire sous la forme:

$$\begin{aligned}
V_{prod}\left(\hat{t}_Y\right) &= (N_D)^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)S^2_{Y_{\bullet\circ}} + (N_M)^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)S^2_{Y_{\circ\bullet}} \\
&+ (N_D)^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)(N_M)^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)S^2
\end{aligned} \tag{4}$$

où

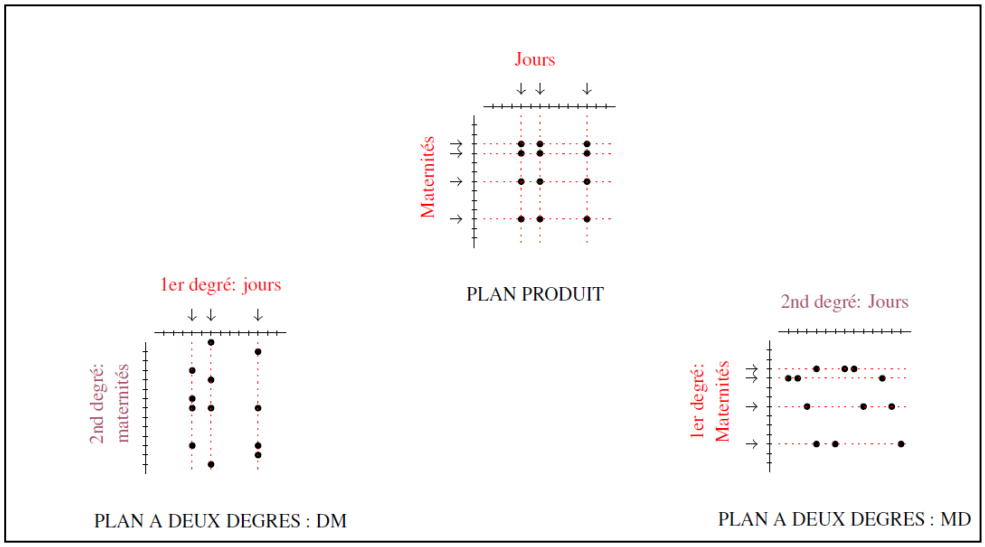
$$\begin{aligned}
S^2_{Y_{\bullet\circ}} &= \frac{1}{N_D-1}\sum_{k\in U_D}\left(Y_{\bullet k} - \frac{1}{N_D}\sum_{l\in U_D} Y_{\bullet l}\right)^2, \\
S^2_{Y_{\circ\bullet}} &= \frac{1}{N_M-1}\sum_{i\in U_M}\left(Y_{i\bullet} - \frac{1}{N_M}\sum_{j\in U_M} Y_{j\bullet}\right)^2, \\
S^2 &= \frac{1}{N_D-1}\frac{1}{N_M-1}\sum_{k\in U_D}\sum_{i\in U_M}\left(Y_{ik} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet k} + \bar{\bar{Y}}_{\bullet\bullet}\right)^2
\end{aligned}$$

avec $\bar{Y}_{i\bullet} = \frac{1}{N_D}\sum_{l\in U_D} Y_{il}$, $\bar{Y}_{\bullet k} = \frac{1}{N_M}\sum_{j\in U_M} Y_{jk}$ et $\bar{\bar{Y}}_{\bullet\bullet} = \frac{1}{N_D}\frac{1}{N_M}\sum_{l\in U_D}\sum_{j\in U_M} Y_{jl}$.

3. Comparaison entre plan produit et plan à deux degrés

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

Figure 2
Echantillonnage produit dans la population de maternités-jours, plan de sondage à deux degrés avec tirage de jours au premier degré, plan de sondage à deux degrés avec tirage de maternités au premier degré



Dans le cas d'un plan de sondage à deux degrés noté MD (voir Figure 2), 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. Tous les échantillons S_i sont sélectionnés avec la même taille n_D. On note V_{MD}, la variance correspondant à ce plan de sondage. Dans le cas d'un sondage aléatoire simple sans remise à chaque degré, que l'on note {SI,SI}, on obtient:

$$V_{MD}\left(\hat{t}_Y\right) = N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)S_{Y_{\bullet\circ}}^2 + \frac{N_M}{n_M}N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)\sum_{i\in U_M}S_{Y_{i\circ}}^2 \tag{5}$$

où

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

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:

$$V_{DM}\left(\hat{t}_Y\right) = N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)S_{Y_{\circ\bullet}}^2 + \frac{N_D}{n_D}N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)\sum_{i\in U_D}S_{Y_{\circ k}}^2 \tag{6}$$

où

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

La différence entre la variance issue d'un plan produit SI $\times$ SI donnée en formule (4), d'une part, et la variance issue d'un plan {SI,SI} donnée en formule (5) ou (6), d'autre part, n'est pas nécessairement positive. Nous considérons le modèle de comportement

$$m : Y_{ik} = \mu + \sigma_1 U_i + \sigma_2 V_k + \sigma_3 W_{ik} \tag{7}$$

avec $U_i, V_k, W_{ik} \square N(0,1)$ et $\sigma_1, \sigma_2, \sigma_3 \in \square^+$, et où σ_1 représente un effet « maternité », σ_2 un effet « jour », et σ_3 un effet résiduel. Sous le modèle (7), on peut montrer que la variance anticipée du plan produit est toujours plus grande que celle du plan à deux degrés considéré. En effet:

$$E_m \left[V_{prod}\left(\hat{t}_Y\right) - V_{MD}\left(\hat{t}_Y\right) \right] = N_M^2 N_D^2 \frac{n_M - 1}{n_M} \left(\frac{1}{n_D} - \frac{1}{N_D} \right) \sigma_2^{\,2}, \tag{8}$$

$$E_m \left[V_{prod}\left(\hat{t}_Y\right) - V_{DM}\left(\hat{t}_Y\right) \right] = N_M^2 N_D^2 \frac{n_D - 1}{n_D} \left(\frac{1}{n_M} - \frac{1}{N_M} \right) \sigma_1^{\,2}, \tag{9}$$

en notant E_m l'espérance sous le modèle (7). 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 entre unités secondaires $\sigma_2^{\,2}$ est grande. De façon analogue, la variance est d'autant plus grande devant V_{DM} que la variabilité entre unités secondaires $\sigma_1^{\,2}$ est grande.

4. Estimation de variance

L'estimateur Hotvitz-Thompson de $V_{prod}\left(\hat{t}_Y\right)$ est:

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

Cet estimateur est sans biais si tous les π^M_{ij} et tous les π^D_{kl} sont strictement positifs, pour tous $(i,j) \in U_M{}^2$, $(k,l) \in U_D{}^2$. Dans le cas d'un plan produit SI $\times$ SI, cet estimateur peut s'écrire sous une forme symétrique par rapport aux échantillons S_M et S_D :

$$\begin{aligned} \hat{V}_{HT}\left(\hat{t}_Y\right) \;=\; & N_M^2 \left(\frac{1}{n_M} - \frac{1}{N_M} \right) s^2_{\hat{Y}_{\circ\bullet}} + N_D^2 \left(\frac{1}{n_D} - \frac{1}{N_D} \right) s^2_{\hat{Y}_{\bullet\circ}} \\ & - \; N_D^2 \left(\frac{1}{n_D} - \frac{1}{N_D} \right) N_M^2 \left(\frac{1}{n_M} - \frac{1}{N_M} \right) s^2 \end{aligned} \tag{10}$$

où

$$
\begin{aligned}
s^2_{\hat{Y}_{\bullet\circ}} &= \frac{1}{n_D - 1}\sum_{k\in S_D}\left(\hat{Y}_{\bullet k} - \frac{1}{n_D}\sum_{l\in S_D}\hat{Y}_{\bullet l}\right)^2, \\
s^2_{\hat{Y}_{\circ\bullet}} &= \frac{1}{n_M - 1}\sum_{i\in S_M}\left(\hat{Y}_{i\bullet} - \frac{1}{n_M}\sum_{j\in S_M}\hat{Y}_{j\bullet}\right)^2, \\
s^2 &= \frac{1}{n_D - 1}\frac{1}{n_M - 1}\sum_{k\in S_D}\sum_{i\in S_M}\left(Y_{ik} - \frac{1}{n_D}\sum_{l\in S_D}Y_{il} - \frac{1}{n_M}\sum_{j\in S_M}Y_{jk} + \frac{1}{n_D n_M}\sum_{l\in S_D}\sum_{j\in S_M}Y_{jl}\right)^2.
\end{aligned}
$$

Notons que cet estimateur n'est pas sans biais terme à terme pour la forme de variance donnée en (4). En particulier, le troisième terme de (10) est négatif. Si ce dernier terme s'avère plus grand en valeur absolue que la somme des deux premiers, l'estimateur de variance $\hat{V}_{HT}\left(\hat{t}_Y\right)$ peut prendre des valeurs négatives. Ceci est en particulier possible lorsque n_M et n_D sont faibles.

À notre connaissance, cet estimateur n'est proposé dans aucune procédure logicielle. Nous étudions des **estimateurs simplifiés** afin de proposer un outil simple d'accès aux estimations de variance, pour l'utilisateur. Nous présentons trois de ces estimateurs simplifiés dans le cas du plan SI $\times$ SI:

$$\hat{V}_{SIMP1}\left(\hat{t}_Y\right) = N_M^2\left(\frac{1}{n_M} - \frac{1}{N_M}\right)s^2_{\hat{Y}_{\circ\bullet}}, \tag{11}$$

$$\hat{V}_{SIMP2}\left(\hat{t}_Y\right) = N_D^2\left(\frac{1}{n_D} - \frac{1}{N_D}\right)s^2_{\hat{Y}_{\bullet\circ}}, \tag{12}$$

$$\hat{V}_{SIMP3}\left(\hat{t}_Y\right) = N_M^2\left(\frac{1}{n_M} - \frac{1}{N_M}\right)s^2_{\hat{Y}_{\circ\bullet}} + N_D^2\left(\frac{1}{n_D} - \frac{1}{N_D}\right)s^2_{\hat{Y}_{\bullet\circ}}. \tag{13}$$

Sous des conditions standard de régularité, $\hat{V}_{SIMP1}\left(\hat{t}_Y\right)$ est approximativement sans biais quand $n_D \to \infty$ et quand n_M est borné. De façon symétrique, $\hat{V}_{SIMP2}\left(\hat{t}_Y\right)$ est approximativement sans biais quand $n_M \to \infty$ et quand n_D est borné. Finalement, l'estimateur simplifié $\hat{V}_{SIMP3}\left(\hat{t}_Y\right)$ est approximativement sans biais quand $n_D \to \infty$ ou $n_M \to \infty$. Ces trois estimateurs ont l'avantage de présenter des valeurs toujours positives et d'être déjà programmés dans des logiciels tels que R, SAS ou Stata.

5. Etude par simulations

Une étude par simulations permet d'illustrer les résultats présentés. Le modèle proposé en (7) est utilisé pour générer une population produit de $N_M = 1000$ maternités et de $N_D = 1000$ jours. Nous utilisons les paramètres $\sigma_1 = \sigma_2 = \sigma_3 = 5$ et $\mu = 200$. On répète $B = 10\,000$ fois la sélection d'un échantillon produit selon un plan de sondage SI $\times$ SI de taille $n_M \times n_D$. Dans les simulations, on a fait varier les tailles d'échantillon n_M et n_D entre 2 et 300. Pour chaque estimateur de variance $\hat{V}$, on calcule le Biais Relatif de Monte Carlo:

$$\%RB_{MC}(\hat{V}) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{V}^{(b)} - V}{V},$$

où la vraie variance V est approchée à l'aide d'un ensemble de 50 000 simulations indépendantes. Le nombre de

valeurs négatives # NEGATIFS que prend l'estimateur sans biais $\hat{V}_{HT}$ est également calculé.

Tableau 1
Biais relatif de quatre estimateurs de variance et nombre de valeurs négatives pour l'estimateur sans biais de variance

n_M	2	10	300	300
n_D	2	300	10	300
$\%RB_{MC}\left(\hat{V}_{HT}\right)$	-0.71	0.51	0.02	-0.00
# NEGATIFS	1309	0	0	0
$\%RB_{MC}\left(\hat{V}_{SIMP1}\right)$	-41.10	-98.12	-2.42	-49.64
$\%RB_{MC}\left(\hat{V}_{SIMP2}\right)$	-37.38	-2.23	-97.63	-50.79
$\%RB_{MC}\left(\hat{V}_{SIMP3}\right)$	21.52	-0.54	-0.05	-0.44

Les résultats sont présentés dans le tableau 1. On remarque la présence de 1309 valeurs négatives sur les 10 000 simulations lorsque n_D et n_M sont égaux à 2. Comme attendu, $\hat{V}_{SIMP1}$ est approximativement non biaisé lorsque n_M est grand et n_D petit (-2.42 % de biais) et inversement pour $\hat{V}_{SIMP2}$. Avec notre jeu de simulations, le biais du dernier estimateur simplifié $\hat{V}_{SIMP3}$ est négligeable dès que n_M est grand ou que n_D est grand (ici, égal à 300).

6. Conclusion

D'autres estimateurs simplifiés ont été étudiés, l'un des objectifs poursuivis étant d'aiguiller l'utilisateur parmi les procédures existant dans les logiciels, en fonction de ses données et des hypothèses requises.

Nous avons traité plus en détail le cas du plan SI $\times$ SI, mais le cadre dans lequel nous avons défini le plan produit est plus général et s'applique à des plans de sondage $p_D(\cdot)$ et $p_M(\cdot)$ quelconques. Une étude est en cours sur différents estimateurs de type Yates-Grundy, avec application dans le cas du tirage de Poisson conditionnel à la taille. Nous travaillons en ce moment sur la prise en compte d'une phase éventuelle de non-réponse dans l'estimation de variance issue d'un plan produit. Enfin, nous étudions l'estimation de variance par linéarisation ou par bootstrap pour des paramètres plus complexes qu'un total.

Bibliographie

Pirus, C., Bois, C., Dufourg, M.-N., Lanoë, J.-L., Vandentorren, S., Leridon, H. et l'équipe Elfe (2010), "La construction d'une cohorte: l'expérience du projet français Elfe", Population 65(4): p.637-670.

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

Vos, J. W. E. (1964), "Sampling in space and time", Review of the International Statistical Institute, Vol. 32, No. 3: p.226-241.