DOCUMENTS DE TRAVAIL 256

# Compléments à l'estimation de la variance pour l'enquête Elfe

Thierry Siméon

Septembre 2020

ined
INSTITUT NATIONAL D'ÉTUDES DÉMOGRAPHIQUES

**Thierry Siméon, 2020, *Compléments à l'estimation de la variance pour l'enquête Elfe*, Paris, Ined, Document de travail, 256**

# COMPLEMENTS A L'ESTIMATION DE LA VARIANCE POUR L'ENQUETE ELFE

*Simplification du plan de sondage, impact sur le calcul de variance et préconisations aux utilisateurs*

Ce document est destiné aux utilisateurs des données issues de la cohorte Elfe. Il se veut être un complément au document « Documents de travail 226 - INED – Estimation de la variance pour l'enquête Elfe - Rapport méthodologique pour l'utilisateur - Hélène Juillard », (https://www.ined.fr/fr/publications/editions/document-travail/estimation-variance-enquete-elfe/) **dont on conseille fortement la lecture**.

Ce document suppose donc connus les principaux concepts d'échantillonnage et de variance, ainsi que tous éléments de contextes sur l'enquête Elfe.

Ainsi, après avoir repris les principales conclusions du document de travail INED 226, on discutera un élément important sur la simplification du plan de sondage en expliquant pourquoi cette simplification nous semble pertinente. **Contrairement aux simplifications proposées et analysées dans le document de travail INED 226, il ne s'agit pas là directement d'une simplification de la méthode de calcul de l'estimateur de variance mais d'une simplification plus conceptuelle sur le plan de sondage mis en œuvre dans le cas de l'enquête Elfe et de l'analyse de ses implications dans le cadre du calcul de variance.**

Sous l'hypothèse de simplification proposée, on analysera ainsi comment les différentes étapes de constitution de l'échantillon influent sur l'estimation de la variance d'une variable d'intérêt en comparant les différents éléments constituant la variance dans le cadre de l'enquête Elfe avec celle du plan de sondage aléatoire simple.

On simulera ces différents calculs sur une cinquantaine de variables issues des enquêtes réalisées en maternité et aux 2 ans de l'enfant. On montrera alors que ce plan de sondage simplifié peut être approché avec quelques précautions par un plan de sondage aléatoire simple pour lequel les procédures logicielles classiques peuvent être utilisées.

On réalisera ensuite ces mêmes analyses sur quelques variables issues des enquêtes réalisées aux 3 ans ½ pour se convaincre de la constance de l'analyse au cours du temps.

Ce document s'appuiera autant que possible sur des simulations réalisées avec le logiciel SAS 9.4 (SAS Institue Inc., 2013).

## RESUME

L'étude Elfe est une enquête nationale dont l'objet est de suivre environ 18000 enfants nés en France en 2011 jusqu'à l'âge adulte. Le plan d'échantillonnage utilisé pour sélectionner les nourrissons en maternité est complexe. L'analyse de la précision des résultats issue des différentes phases d'enquêtes également. Les procédures classiques des logiciels statistiques standards ne permettent pas de calculer cette précision.

Ce travail fait suite au « Document de travail INED 226 - Estimation de la variance pour l'enquête ELFE Hélène Juillard » publié en 2016 (https://www.ined.fr/fr/publications/editions/document-travail/estimation-variance-enquete-elfe/) et propose une simplification du concept même du plan de sondage mis en œuvre. L'étude de cette simplification, de son impact dans le cadre du calcul de variance, ainsi que l'estimation de la précision d'une cinquantaine d'indicateurs permettent de proposer des préconisations pour estimer avec des procédures simplifiées les variances dans le cadre de l'analyse des résultats issus de l'enquête Elfe, et donc d'en approcher simplement l'incertitude.

Les résultats sont illustrés grâce au logiciel SAS.



## MOTS CLES

ELFE (enquête longitudinale française depuis l'enfance), précision, variance



## ABSTRACT

The Elfe study is a national survey which aims to follow around 18 000 children born in France in 2011 until adulthood. The sampling plan used to select infants in maternity hospitals is complex. The analysis of the precision of the results also. The standard procedures of standard statistical softwares do not allow this precision to be calculated.

This work follows the " Document de travail INED 226 - Estimation de la variance pour l'enquête ELFE Hélène Juillard " published in 2016 (https://www.ined.fr/fr/publications/editions/document-travail/estimation-variance-enquete-elfe/) and proposes a simplification of the concept of the sampling plan implemented. The study of this simplification, of its impact in the context of variance calculation, as well as the estimation of the precision of around fifty indicators allows us to propose recommendations for estimating the variances with simplified procedures in the context of analysis of the results from the Elfe survey, and therefore simply approaching uncertainty.

The results are illustrated using SAS software.



## KEYWORDS

ELFE (French longitudinal survey from childhood), precision, variance

# Table des matières

# 1. Principaux résultats relatifs au document de travail INED 226[1]

Le plan de sondage 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 et d'une phase de calage.

**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, nés dans une maternité métropolitaine et dont les parents ne résidaient pas en métropole seulement de façon temporaire**. Toutes les familles sélectionnées ont été enquêtées peu de temps après l'accouchement dans certaines maternités métropolitaines et durant certains jours de l'année.

Le **plan de sondage pour les maternités** est un plan probabiliste correspondant à un plan stratifié avec constitution de cinq strates à effectifs égaux et tirages à allocation proportionnelle au nombre d'accouchements recensés en 2008.

Concernant les jours, 25 ont été choisis durant quatre périodes (appelées vagues) couvrant les quatre saisons de l'année. Les jours n'ont pas été sélectionnés de manière aléatoire mais fixés a priori (la moitié devait coïncider avec l'échantillon démographique permanent E.D.P.).

Les deux échantillons (maternités et jours) ont été sélectionnés indépendamment.

| Strates $g$ | Nb d'accouchements par maternité en 2008 | Taille dans la population $N_g$ | Taille de l'échantillon $n_g$ |
|---|---|---|---|
| 1 | [145-699[ | 108 | 28 |
| 2 | [700-1009[ | 108 | 47 |
| 3 | [1010-1418[ | 109 | 66 |
| 4 | [1422-2187[ | 108 | 97 |
| 5 | [2197-5215[ | 111 | 111 |
| TOTAL | | 544 | 349 |

| Vague $h$ | Taille dans la population $M_h$ | Taille de l'échantillon $m_h$ |
|---|---|---|
| 1 | 90 | 4 |
| 2 | 91 | 6 |
| 3 | 92 | 7 |
| 4 | 92 | 8 |
| TOTAL | 365 | 25 |

*Figure 1 – tailles des strates et des vagues*
*'Population' réfère respectivement à l'ensemble de maternités métropolitaine et à l'ensemble des jours de l'année*

Durant l'enquête Elfe, 29 maternités parmi les 349 sélectionnées n'ont pas participé à l'enquête. De plus, 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. 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 mais traitées en ajustant simplement les probabilités d'inclusion et donc les poids de chaque nourrisson.

Sous ces hypothèses, H. Juillard montre qu'un estimateur sans biais de la variance issue du plan de sondage Elfe peut être décomposé en 3 éléments :

$$\widehat{V_{prod}} = \widehat{V_D} + \widehat{V_M} - \widehat{V_E}, \text{ avec}$$

$\widehat{V_D}$ l'estimation de la variance due au tirage stratifié de jours ;

$\widehat{V_M}$ l'estimation de la variance due au tirage stratifié de maternités ;

$\widehat{V_E}$ un « effet croisé » dû au fait que jours et maternités sont identiques (Les jours enquêtés sont les mêmes pour chaque maternité tirée, ou vice versa).

Il existe également dans l'enquête Elfe une phase importante de non-réponse au niveau nourrisson : 49 % des 36 000 familles approchées n'ont pas souhaité participer. Il faut bien sûr considérer cet élément dans le cadre du calcul de la variance. Pour tenir compte de la non réponse totale, on considère que la décision de répondre

[1] L'ensemble de ce chapitre est entièrement repris du document de travail INED 226 – Hélène Juillard.

est aléatoire. Cela signifie que, dans les mêmes conditions (par exemple d'âge, de revenu, de nationalité, ..), le processus de réponse ne renverra pas toujours le même résultat. Certains individus accepteront de répondre, d'autres non. L'échantillon de répondants est donc issu d'un tirage en (n+1) phases. Les *n* premières phases sont des phases d'échantillonnage (le nourrisson est sélectionné), la dernière est une phase d'acceptation (les parents du nourrisson acceptent ou non de participer). L'estimation de la variance devient alors :

$$\widehat{V_{ELFE}} = \widehat{V_D} + \widehat{V_M} - \widehat{V_E} + \widehat{V_{NR}} \quad (1)$$

Enfin, un processus de calage est réalisé. Pour calculer l'estimateur de variance lorsqu'un tel processus est mis en œuvre, on réalise la régression pondérée de la variable d'intérêt analysée sur les variables de calage, et on calcule la variance en appliquant la formule (1) non plus à la variable d'intérêt mais aux résidus de la régression.

Rappelons également que les parties précédentes concernent l'estimation de la variance de l'estimation d'un total. Pour d'autres paramètres (ratio, moyenne, ...), la méthode de linéarisation peut être utilisée afin de pouvoir estimer les variances. Prenons l'exemple comme dans toute la suite de ce document d'un ratio $R = \frac{t_Y}{t_X}$ (total d'une variable *Y* sur le total d'une variable *X*). Pour l'estimation de la variance de l'estimation du ratio $\hat{R}$, il est nécessaire d'estimer les totaux $\widehat{t_x}$, $\widehat{t_y}$, puis $\hat{R} = \frac{\widehat{t_Y}}{\widehat{t_X}}$ et enfin de calculer pour chaque individu *k* la linéarisée du paramètre, définie par :

$$lin_k = \frac{1}{\widehat{t_X}} (y_k - \hat{R}.x_k). \quad (2)$$

Pour prendre en compte l'étape de calage, il convient alors de réaliser la régression de $lin_k$ sur les variables de calage et d'utiliser les résidus $\varepsilon_k$ de cette régression comme variable de la formule (1).

## 2. Que cherchons-nous à mesurer ?

Il faut avant tout préciser ce qu'on cherche à estimer. En effet, dans le cas de l'enquête Elfe, on ne cherche pas à estimer un nombre moyen d'individus ou de naissances **par jour** (nombre quotidien de nourrissons avec une caractéristique donnée par exemple), mais bien un total **annuel** ou un ratio (le nombre total annuel de nourrissons avec telle ou telle caractéristique, la part de nourrissons présentant telle caractéristique,… ), une moyenne (taille moyenne par nourrisson par exemple), un score ou une répartition que l'on veut quantifier, comparer à un instant *t*, ou suivre dans le temps.

On ne cherche donc à calculer des variances que sur des éléments dont l'unité statistique est le nourrisson. Pour obtenir un échantillon suffisant, il a bien évidement été nécessaire de réaliser des enquêtes en maternité pendant plusieurs jours, ces jours étant répartis dans l'année pour faciliter le travail des enquêteurs. **Mais ce n'est pas parce qu'on a enquêté des jours différents que le jour doit être considéré comme un élément du plan de sondage.**

Il est également évident que le jour de naissance sert à sélectionner des individus, mais lors des enquêtes suivantes, le jour de naissance n'est plus un élément constituant le plan de sondage. On enquête tous les nourrissons à la même période (en répartissant éventuellement les enquêtes selon la vague de naissance pour assurer que l'âge des enfants est relativement comparable) et encore une fois sur des jours différents pour faciliter le travail des enquêteurs. Le jour de naissance n'est plus alors un élément entrant en compte dans la sélection ou l'interrogation des familles.

De plus, le principe de la stratification, lorsqu'on réalise une enquête, est de constituer a priori des groupes homogènes d'individus, au sein desquels on réalise un tirage aléatoire. Ainsi, les estimations obtenues dans chacun des groupes (et donc dans l'échantillon final) seront moins susceptibles de dépendre du hasard. Plus les groupes sont homogènes par rapport à la variables d'intérêt étudiée et plus ils sont hétérogènes entre eux, plus la stratification améliore la précision de l'enquête.

Une autre justification pour la stratification peut être trouvée si on s'intéresse à une population particulière pour laquelle on souhaite obtenir une précision suffisante. On va alors surreprésenter cette sous population dans notre tirage en appliquant des taux de sondage très différents d'un groupe à l'autre.

Aucune des 2 raisons ne trouve de justification lorsqu'on considère le jour de naissance. Les jours des différentes vagues sont identiques. Il n'y a aucune sur-représentativité d'un trimestre nécessaire et les caractéristiques des nourrissons ne dépendent pas de la vague pendant laquelle ils ont été enquêtés (cela ne veut pas dire que tous les jours sont identiques, mais que la vague n'a pas de sens théorique pour calculer des variances).

Afin de s'en convaincre, on trace ici les parts de nourrissons avec certaines caractéristiques selon le jour de naissance. Aucune relation ne peut être établie. Les fluctuations les plus importantes (type d'accouchement) sont dues à la typologie de jour (samedi, dimanche, semaine) et non à la vague.


*Figure 2 - part des nourrissons avec certaines caractéristiques selon le jour de naissance*

Ainsi, on propose de considérer le jour comme un élément permettant d'obtenir un échantillon suffisant, de faciliter le recrutement et le travail des enquêteurs, mais ne constituant pas en tant que tel un élément du plan de sondage. **On travaillera donc sur un simple plan stratifié de maternités.**

Bien évidemment, cette simplification n'aurait pas de sens sur le tirage des maternités. On n'enquête pas toutes les maternités (il faut donc bien mettre en évidence que les maternités sont issues d'un tirage aléatoire), les nourrissons sont très différents d'une maternité à l'autre, les poids et les taux de sondage par strate sont très différents (il faut donc bien garder la stratification pour améliorer les estimations et les calculs de précisions).

Les données de contexte, sociales, d'accouchement ou encore de santé dépendent fortement des maternités. Tout cela pousse bien à centrer l'ensemble de nos calculs sur cet élément. Comme on peut le voir, les parts de nourrissons avec certaines caractéristiques selon la maternité de naissance sont bien plus variables :

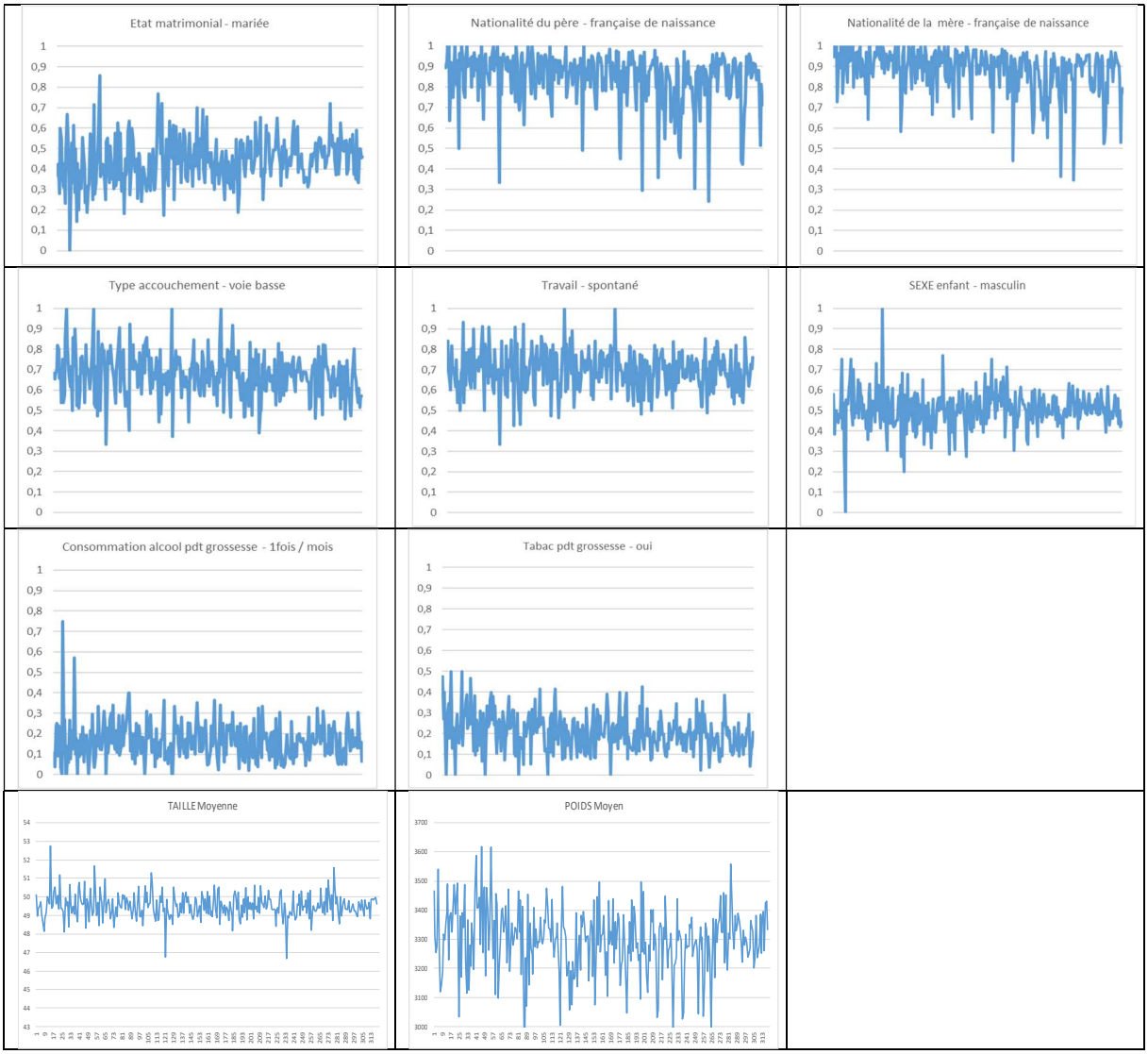


*Figure 3 - part des nourrissons avec certaines caractéristiques selon la maternité de naissance*

Ainsi, si on accepte la simplification forte proposée ici, on pourra schématiser le plan de sondage de l'enquête Elfe comme ceci :

- Stratification de maternités : les maternités sont réparties en 5 strates selon leur taille ;
- Tirage aléatoire de maternités de chacune des strates (allocation proportionnelle au nombre d'accouchements recensés en 2008) ;
- Interrogation de tous les nourrissons nés dans ces maternités ;
- Processus de non réponse à chacun de ces tirages (négligeable au niveau des maternités, sensible au niveau des nourrissons).

## 3. Quantification de la simplification du plan de sondage

L'objectif de ce paragraphe est de mesurer **l'effet de la simplification** proposée au paragraphe précédent et de vérifier que cette simplification se base sur un effet inter jours négligeable par rapport à l'effet inter maternités. Après avoir présenté les éléments permettant de mesurer ces effets, on propose de les calculer sur une sélection importante de variables.

Ainsi, on définit les éléments suivants :

Soit une variable $y_k$ = 1 si le nourrisson *k* possède la caractéristique étudiée, 0 sinon (par exemple $y_k$ = 1 si lieu de naissance mère = France métropolitaine, 0 sinon).

Soit une variable $x_k$ = 1 si la variable analysée est renseignée pour le nourrisson *k* (simplement pour estimer une proportion hors non réponse, par exemple $x_k$ = 1 si lieu de naissance mère est connu, 0 sinon).

On s'intéresse au rapport pondéré par la variable *poids* $\hat{R} = \frac{\widehat{t_Y}}{\widehat{t_X}}$ (proportion d'individu avec la caractéristique étudiée parmi ceux dont, par exemple, le lieu de naissance est connu). On génère alors, pour chaque individu *k* le linéarisé $lin_k$ par la formule (2).

On estime ensuite[2]:

La part de variance due au tirage de maternités, estimée par le calcul de la variance dans le cadre classique d'un tirage stratifié (stratification par taille des maternités) en grappes (cluster de maternités).

$effetMAT$: $\widehat{V_M}$

```
proc surveymeans total= degreM;
cluster maternite; strata strate;
var lin;
weight poids;
run;
```

Où « degreM » représente le nombre de maternités par strate.

La part de variance due au tirage des jours, estimée par le calcul de la variance dans le cadre classique d'un tirage stratifié (stratification par vague) en grappes (cluster de jours).

$effetDAY$: $\widehat{V_D}$

```
proc surveymeans total= degreJ;
cluster jour; strata vague;
var lin;
weight poids;
run;
```

Où « degreJ » représente le nombre de jours par vague.

La part de variance due à la non réponse « Nourrisson », cette non réponse revenant à ajouter un degré au tirage de l'enquête Elfe. On a sélectionné par une procédure quelconque des couples « maternité x jour » ayant accepté de participer. Ce nouveau degré de tirage est un plan de Poisson (parmi l'ensemble des nourrissons appartenant aux « maternité et jour » acceptant de répondre, on réalise un tirage qui sélectionne les nourrissons avec probabilité $\emptyset_k$ et qui rejette les nourrissons avec probabilité $1 - \emptyset_k$). En utilisant la décomposition de la variance, on peut montrer que l'estimation sans biais de la variance due à la non réponse est donnée par :

effetNR: $\widehat{V_{NR}}$ $\sum_{\substack{sur\ les\ seuls \\ "nourrissons" \\ répondants}} \left(\frac{{lin_k}^2}{{\pi_k}^2} \frac{1-\emptyset_k}{{\emptyset^2}_k}\right)$ , avec $\pi_k$ la probabilité que le nourrisson *k* soit sélectionné, et $\emptyset_k$ l'estimation de la probabilité qu'il accepte de participer à l'enquête.

On peut ainsi quantifier :

$$variance\ ELFE[3] = effetMAT + effetDAY + effetNR$$

En plus de ces calculs nécessaires à l'estimation de la variance d'un ratio dans le cas du tirage réalisé dans l'enquête Elfe, on doit pour comprendre la justification de la simplification proposée au paragraphe précédent évaluer deux autre éléments importants. On va calculer ainsi, pour chaque maternité *i* le ratio de nourrissons avec la caractéristique mesurée ainsi que le ratio moyen, soit :

$$Ratio_{maternité\ i} = \sum_{\text{nourrissons k de la maternité i}} (y_k) \Big/ \sum_{\text{nourrisso k de la maternité i}} (x_k)$$

$$\overline{Ratio}_{maternité} = \left(\frac{1}{320}\right) \sum_{\text{maternité i}} Ratio_{maternité\ i}$$

[2] Cf. là aussi le document de travail INED 226 – Hélène Juillard – Mai 2015.

[3] Comme dans le document de travail INED 226, l'effet croisé est négligé ici.

Puis :

$$\text{dispersionMAT} = \left(\frac{1}{320}\right) \sum_{\text{maternités}} \left(Ratio_{maternité\ i} - \overline{Ratio}_{maternité}\right)^2$$

On calcule de manière analogue pour chaque jour *j* le ratio de nourrissons avec la caractéristique mesurée ainsi que le ratio moyen, puis :

$$\text{dispersionDAY} = \left(\frac{1}{25}\right) \sum_{\text{jours}} \left(Ratio_{jour\ j} - \overline{Ratio}_{jour}\right)^2$$

Ces différents effets sont présentés ci-dessous pour 53 variables recueillies de la maternité (préfixe M00) aux 2 ans de l'enfant (préfixe A02). Ces variables reprennent aussi bien des données sociodémographiques (lieu de naissance des parents, nationalité, état matrimonial, situation par rapport à l'emploi, ...), que des variables de santé (tabagisme, alcool, …), des variables sur la grossesse et l'accouchement (diabète, hypertension, type d'accouchement, ...) ou encore des variables sur les occupations de l'enfant (dessin, puzzle, …). L'ensemble des 53 variables sont listées en annexe 1.

Pour estimer ces effets, on calcule la part de nourrissons avec la modalité « 1 » à chacune de ces variables, ainsi que l'ensemble des éléments définis plus haut. Les données sont pondérées par la pondération transversale « Enfant » du temps d'enquête correspondant à la variable analysée.

La quantification simple des différents éléments constituant la variance montre ainsi la part prépondérante de l'effet jour dans son estimation. Cette part s'élève à environ 48%, contre 23% pour l'effet maternité et 29% pour l'effet NR (la part précise de chaque élément est donnée en annexe 2).

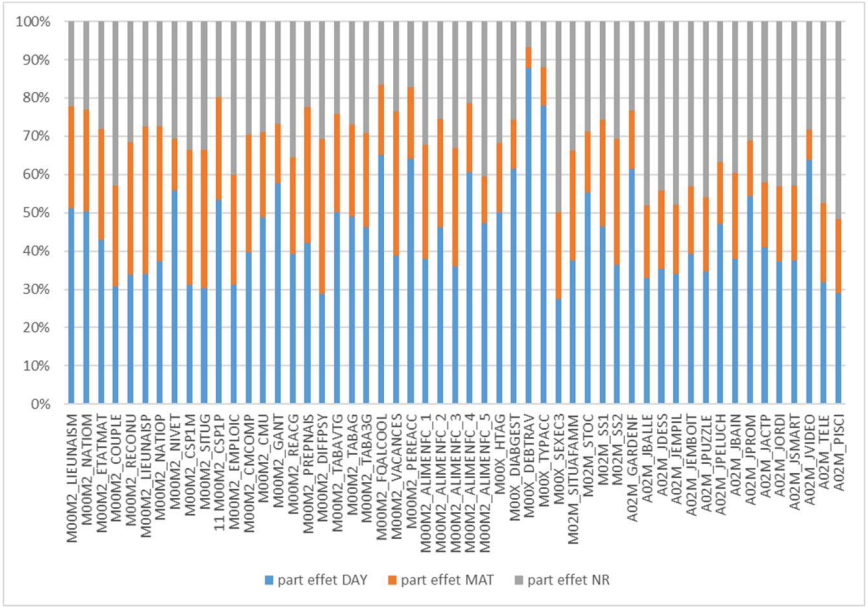


*Figure 4 - part de l'effet Jour / effet Maternité / effet NR dans la variance théorique*

*Note de lecture : La figure 4 représente, pour chaque indicateur, la part de l'effet Maternité (orange), l'effet jour (bleu) et l'effet non réponse (gris) issue de la formule: $variance\ ELFE = effetMAT + effetDAY + effetNR$. On note ainsi la part prépondérante de l'effet jour (40 à 60% environ), contre 20 à 30% pour l'effet maternité et 20 à 30% pour l'effet NR.*

On signale évidement que la part de l'effet jour est la plus importante pour les variables pour lesquelles on a noté les fluctuations journalières les plus importantes (données sur l'accouchement notamment). A l'inverse, cet effet jour est le plus faible pour les variables les plus indépendantes du jour (sexe, situation par rapport à l'emploi par exemple). Il est donc bien clair que la simplification de plan de sondage proposée au paragraphe précédent a un effet important sur l'estimation de la variance théorique prenant en compte le plan de sondage complet mis en œuvre dans l'enquête Elfe.

Cependant, il convient bien de clarifier les choses. Les effets jour et maternité reposent sur des calculs de variances dans le cas de 2 sondages stratifiés. Ces variances dépendent donc de 2 éléments : le taux de sondage par strate et la dispersion moyenne des variables au sein des strates. Mais les taux de sondage sont trop différents pour qu'on n'analyse pas plus en détail la variabilité des données (on enquête entre 1 jour sur 11 et 1 jour sur 22 selon le trimestre, alors qu'on enquête entre ¼ et 9/10ème des maternités selon les strates).

Il faut donc analyser la précision de notre enquête en analysant un autre critère que le « simple » rapport de variances : la dispersion moyenne inter jours est négligeable par rapport à la dispersion moyenne inter maternités. Ainsi, si la part de la variance due au tirage des jours est environ 2,5 fois supérieure à la part de variance due au tirage des maternités, la **dispersion moyenne inter jours est environ 35 fois inférieure à la dispersion inter maternités** (la part précise de chaque élément est donnée en annexe 2).

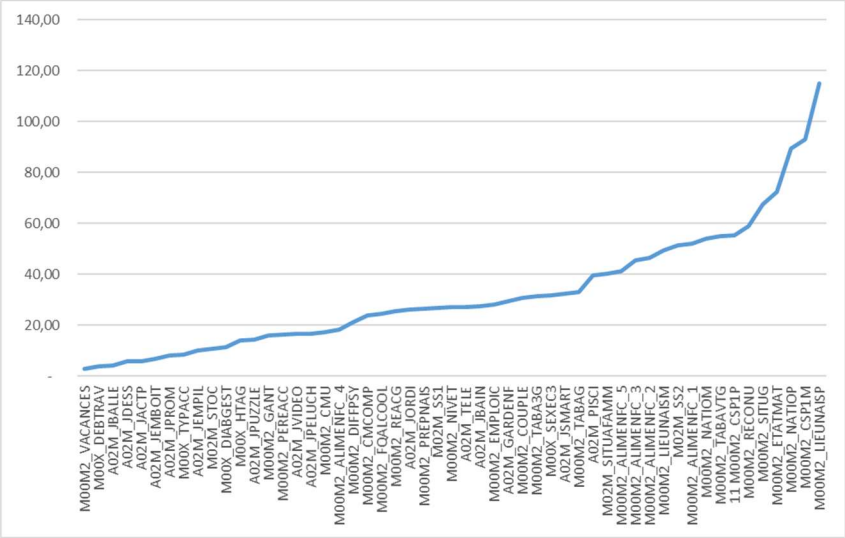


*Figure 5 - Rapport entre dispersion moyenne Maternité / dispersion moyenne Jour*

C'est donc bien la différence importante de taux de sondage et sa prise en compte dans le calcul de la variance, qui entraine le déséquilibre important entre effet jour et effet maternité, et non la fluctuation de la donnée. On aurait enquêté un autre jour, les résultats auraient été les mêmes ou presque. Ainsi, même si la théorie des sondages peut prendre en compte un effet jour, il ne semble pas illogique de ne pas prendre en compte cet élément dans le calcul de précision des analyses menées grâce aux données issues de l'enquête Elfe : à la fois parce que le jour de naissance peut être vu comme ne constituant pas en tant que tel un élément du plan de sondage, mais aussi parce que la précision des résultats n'en est que peu affectée.

On propose donc dans la suite de ce document de schématiser l'analyse de la variance de l'enquête Elfe ainsi :

- Stratification de maternités : les maternités sont réparties en 5 strates selon leur taille (effet stratification), et tirage aléatoire de maternités de chacune des strates;
- Interrogation théorique de tous les nourrissons nés dans ces maternités, mais prise en compte d'un processus de non réponse nourrissons – intégration d'un second degré de tirage pour n'interroger que ceux acceptant de répondre (effet de grappe dû au second degré + effet non réponse);
- Calage (effet calage).

# 4. Définition des différents éléments constituant l'estimation de la variance

L'objectif de ce paragraphe est de présenter les éléments qui permettent de quantifier maintenant les différents effets du plan de sondage simplifié (**effet de grappe**, **effet de la stratification**, **effet de la NR** et **effet du calage**) sur l'estimation sans biais de la variance de la proportion de nourrissons avec telle ou telle caractéristique.

Ainsi, on définit comme au paragraphe précédent les éléments suivants :

Soit une variable $y_k$ = 1 si le nourrisson *k* possède la caractéristique étudiée, 0 sinon (par exemple $y_k$ = 1 si lieu de naissance mère = France métropolitaine, 0 sinon).

Soit une variable $x_k$ = 1 si la variable analysée est renseignée pour le nourrisson *k* (simplement pour estimer une proportion hors non réponse, par exemple $x_k$ = 1 si lieu de naissance mère est connu, 0 sinon).

On génère alors, pour chaque individu *k* le linéarisé $lin_k$ par la formule (2).

On calcule également les résidus $\varepsilon_k$ issus de la régression pondérée de la variable $lin_k$ sur les variables de calage.

On estime ensuite :

$\hat{V}_{SAS}(lin_k)$, qui correspond à l'estimation de la variance dans le cadre classique d'un sondage aléatoire simple avec lequel on aurait cherché à estimer un ratio en absence de tout autre élément constitutif du plan de sondage:

```
proc surveymeans total=764000;
var lin;
weight poids;
run;
```

A noter que la procédure Surveymeans propose directement une option pour calculer un ratio. Le calcul précédent est donc équivalent à la commande :

```
proc surveymeans total=764000;
ratio Y/X;
weight poids;
run;
```

Le tirage en grappes de maternités est pris en compte par le calcul de $\hat{V}_{GR}(lin_k)$, qui correspond à l'estimation de la variance dans le cadre classique d'un tirage à 2 degrés (cluster de maternités).

```
proc surveymeans total= 544;
cluster maternite;
var lin;
weight poids;
run;
```

Cette procédure est là encore équivalente à :

```
proc surveymeans total=544;
cluster maternite;
ratio Y/X;
weight poids;
run;
```

Le tirage stratifié de maternités est pris en compte par le calcul de $\hat{V}_{GR\_ST}(lin_k)$, qui correspond à l'estimation de la variance dans le cadre classique d'un tirage stratifié (stratification par taille des maternités) à 2 degrés (cluster de maternités).

```
proc surveymeans total= degreM;
cluster maternite; strata strate;
var lin;
weight poids;
run;
```

Cette procédure est là encore équivalente à :

```
proc surveymeans total=degreM;
cluster maternite; strata strate;
ratio Y/X;
weight poids;
run;
```

Où « degreM » comprend le nombre de maternités par strate.

Remarque : dans toutes les procédures précédentes, la variable 'TOTAL' permet de prendre en compte la taille de la population. Cette variable vaut donc 764000 lorsqu'on parle de nourrissons (764000 représentant le nombre total d'enfants nés en France métropolitaine éligibles à l'enquête Elfe en 2011), ou 544 lorsqu'on travaille sur les maternités.

La prise en compte la non réponse « Nourrisson » revient à ajouter :

$\hat{V}_{NR}(lin_k) = \sum_{\substack{sur\ les\ seuls \\ "nourrissons" \\ répondants}} \left(\frac{{lin_k}^2}{{\pi_k}^2} \frac{1-\emptyset_k}{{\emptyset^2}_k}\right)$ , avec $\pi_k$ la probabilité que le nourrisson *k* soit sélectionné, et $\emptyset_k$ l'estimation de la probabilité qu'il accepte de participer à l'enquête.

On peut ainsi estimer chacun des éléments constituant le plan de sondage simplifié Elfe par :

**Effet de grappe** : $\hat{V}_{GR}(lin_k) \ / \hat{V}_{SAS}(lin_k)$

**Effet de stratification** : $\hat{V}_{GR\_ST}(lin_k) \ / \hat{V}_{GR}(lin_k)$

**Effet NR :** $\hat{V}_{GR\_ST}(lin_k) + \hat{V}_{NR}(lin_k) \ / \hat{V}_{GR\_ST}(lin_k)$

➔ **Effet du plan de sondage** : $\hat{V}_{GR\_ST}(lin_k) + \hat{V}_{NR}(lin_k) / \hat{V}_{SAS}(lin_k)$

L'effet du plan de sondage exprime la « différence de qualité » entre le plan de sondage mis en œuvre dans « Elfe » et un sondage aléatoire simple. Si ce coefficient est supérieur à 1, le plan de sondage fait « perdre » de la précision. Il regroupe 2 effets plutôt antagonistes : l'impact de la stratification (on améliore la précision si la variable analysée est assez homogène au sein des maternités de même taille), et l'impact de la NR (on dégrade la précision puisqu'on ajoute un degré de sondage, surtout si la variable étudiée est liée à la NR et si ceux qui ont une probabilité de répondre estimée plus faible ont un comportement atypique).

➔ **Effet calage :** $\hat{V}_{GR\_ST}(\varepsilon_k) + \hat{V}_{NR}(\varepsilon_k) / \hat{V}_{GR\_ST}(lin_k) + \hat{V}_{NR}(lin_k)$ (rapport entre variance totale après calage et variance totale si ce calage n'avait pas eu lieu)

L'effet de calage permet de mesurer, pour chaque variable, à quel point le calage améliore la précision des estimateurs. Globalement, un calage améliore toujours la précision d'un calcul, puisqu'il réduit l'incertitude en limitant certains aléas en « bloquant » les marges. Si la variable étudiée dépend au moins un peu des variables de calage, l'aléa diminue forcément. Si la variable est totalement indépendante des variables de calage, on n'a rien amélioré, mais rien perdu non plus.

L'effet complet du processus mis en œuvre dans le cadre de l'enquête Elfe est donc :

➔ $\mathbf{Effet\ ELFE = effet\ plan\ de\ sondage \times effet\ calage}$

On a estimé comme au paragraphe précédent ces différents effets sur 53 variables recueillies de la maternité aux 2 ans de l'enfant. Pour rappel, l'ensemble des 53 variables sont listées en annexe 1.

On a calculé la part de nourrissons avec la modalité « 1 » à chacune de ces variables, et mesuré par les procédures listées précédemment l'ensemble de ces éléments. Les données sont pondérées par la pondération « Enfant » du temps d'enquête correspondant à la variable analysée.

## 5. Quantification des différents éléments constituant l'estimation de la variance

L'**effet de grappes** mesure en quoi le tirage de maternités puis de nourrissons joue sur la précision du sondage. Dans la plupart des cas, il traduit un phénomène de perte de précision due à l'existence de similarités entre les nourrissons d'une même maternité. Enquêter des nourrissons d'une même maternité apporte alors moins de « connaissance » que si on avait enquêté des nourrissons de manière totalement aléatoire. Plus l'effet grappe est grand (plus la variance prenant en compte le tirage par grappe est grande par rapport à la variance d'un tirage aléatoire simple), plus la perte de précision est grande.

A titre d'exemple, supposons qu'on ait mesuré un effet de grappes de 9. Cela signifie que la variance de notre analyse est 9 fois plus élevée que si on avait enquêté le même nombre d'individus par tirage aléatoire simple. Ainsi, pour éviter cette perte de précision due au plan de sondage, il aurait fallu enquêter 3 fois plus de nourrissons.

Pour essayer de simplifier la perception de l'effet de grappes, on peut l'exprimer plus facilement sous certaines conditions. Ainsi, même si ce n'est pas le cas pour Elfe, on peut montrer que lorsque le sondage est un sondage aléatoire simple de maternités et un sondage aléatoire simple de nourrissons, alors l'effet de grappes s'écrit simplement :
$\frac{\hat{V}_{GR}(y_k)}{\hat{V}_{SAS}(y_k)} = 1 + \rho(\bar{n} - 1)$, avec $\bar{n}$ le nombre moyen de nourrissons enquêtés par maternité et $\rho$ proportionnel à $\sum_{maternités\ i} \sum_{j=1}^{taille\ mater} \sum_{\substack{k=1 \\ k \neq j}}^{taille\ mater} (y_{ij} - \bar{y})(y_{ik} - \bar{y})$ .

Cela permet d'identifier 2 éléments importants de l'effet de grappes : il dépend donc du nombre moyen de nourrissons enquêtés par maternité (il est évident que si on enquêtait un seul enfant par maternité, il n'y aurait aucun effet de grappe et le tirage serait un tirage aléatoire simple), et de la dispersion de la variable étudiée pour les nourrissons d'une même maternité par rapport à la moyenne générale de tous les enquêtés (si les enfants d'une même maternité sont semblables par rapport à la moyenne générale, $\rho$ sera une somme de valeurs positives alors que si certains nourrissons d'une même maternité sont supérieurs et d'autres inférieurs à la moyenne générale, $\rho$ sera proportionnel à une somme de valeurs positives et négatives et sera donc plus faible, voire même négatif. Ainsi, plus les enfants d'une même maternité sont semblables, moins le fait d'enquêter un nouvel individu apporte d'informations. L'effet de grappes est alors plus grand, ce qui dégrade la précision.

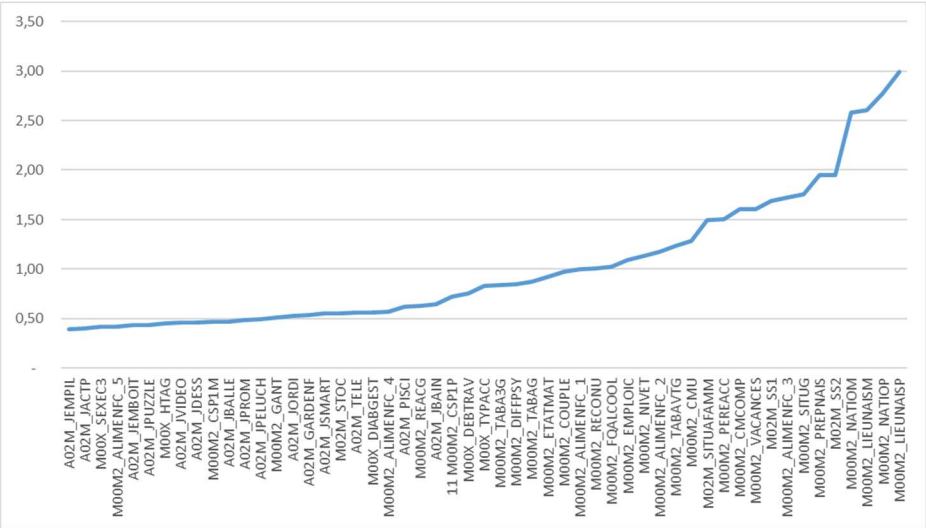

*Figure 6 - effet de grappe*

Prenons des cas extrêmes : les enfants nés de mères françaises (M00M2_NATIOM) sont plus regroupés dans certaines maternités que dans d'autres (cf figure 10 plus bas – la part des mères françaises atteints régulièrement 100% dans une maternités donnée). Enquêter plusieurs enfants dans ce type de maternité n'apporte donc pas autant d'information qu'attendue. L'effet de grappes est élevé, la précision dégradée.

Inversement, les garçons (M00X_SEXE3) n'ont aucune raison d'être regroupés dans certaines maternités (cf figure 8 plus bas). Enquêter un nouveau nourrisson dans une maternité ou dans une autre apporte donc toujours autant d'informations. La part des garçons ne subissant d'ailleurs que très peu de fluctuations d'une maternité à l'autre, on peut même mieux estimer la moyenne de garçons par maternité grâce au tirages en grappes en dispersant moins notre échantillon que si on avait tiré aléatoirement de enfants dans toutes les maternités. Notre estimation globale de la part des garçons n'en est que meilleure. L'effet de grappes est faible (et même inférieur à 1), et la précision augmente par rapport au tirage aléatoire simple.

Signalons également que l'effet de grappes est dépendant de la temporalité des enquêtes. Il est bien évidemment beaucoup plus faible pour les mesures aux 2 ans de l'enfant, à la fois parce qu'on enquête moins d'enfants par maternité (12000 répondants au total contre 18000 en maternité), et surtout parce que les enfants nés dans une même maternité ont moins de raisons de se ressembler lorsqu'ils grandissent. Par sa définition même, l'effet de grappe sera certainement encore plus faible au temps futurs de l'enquête.

L'**effet strate** mesure en quoi la stratification améliore la précision de l'enquête Elfe. La stratification peut diminuer la variance d'un ratio grâce à 2 éléments : en comparant la dispersion du taux de nourrissons avec telle ou telle modalité dans une maternité par rapport au taux moyen des maternités de la même strate et donc d'une taille équivalente (et non au taux moyen de l'ensemble de maternités), et en jouant sur le taux de sondage de chaque strate.

En effet, dans le cas d'un tirage stratifié, on estime la variance d'une moyenne par $\hat{V}_{GR\_ST}(\hat{\bar{y}}) = \sum_{strate\ h} \left(\frac{N_h}{N}\right)^2 (1 - f_h).\ {s_h}^2 / n_h$. Le tirage stratifié sera d'autant plus précis si les maternités de chaque strate sont le plus semblables possible ( $s_h$, la dispersion de notre variable d'intérêt calculée au sein de chaque strate *h* sera alors petit) ou si on enquête beaucoup de maternités là où la dispersion est la plus mauvaise ($n_h$, nombre de maternités enquêtées dans la strate *h* sera alors important et le terme en $1/n_h$ pourra compenser la forte dispersion).

Dans le cas de l'enquête Elfe, le taux de sondage est très différent selon les strates, allant de 23% des maternités enquêtées dans la strate 1 à plus de 80% des maternités enquêtées dans les strates 4 et même 90% dans la strate 5. L'effet de la stratification sera donc d'autant plus important que la variabilité est forte dans les maternités de grande taille de la strate 5. On compensera alors cette variabilité par un taux de sondage important, diminuant fortement la variance intra strate et donc la variance globale. Si au contraire la variabilité ne dépend pas de la taille de la maternité ou si elle est plus importante pour les maternités de petite taille, la stratification n'aura que peu d'effet.

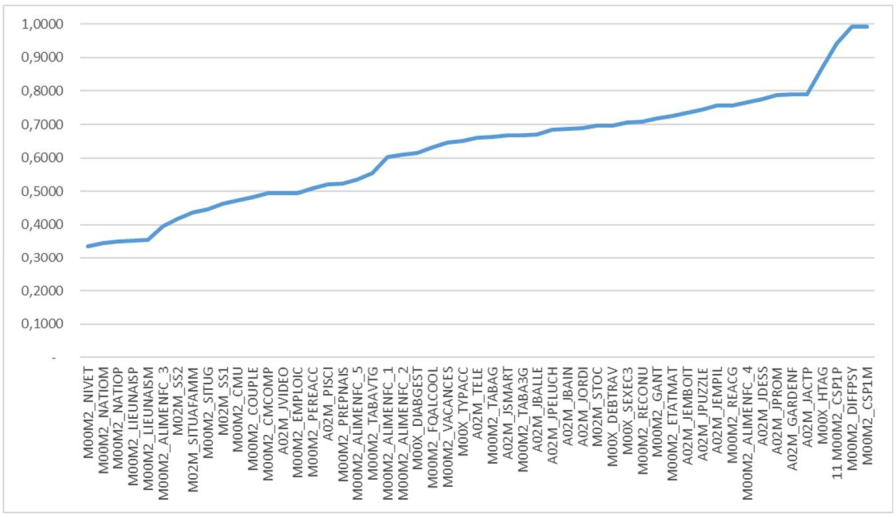


*Figure 7 - effet stratification*

Il faut signaler avant tout que l'effet strate est relativement constant, surtout comparativement aux autres effets étudiés (entre 0.6 et 0.8 pour la grande majorité des variables analysées). On peut simplement noter que l'effet strate est plus important pour toutes les variables reprenant les données sociodémographiques des parents (lieu de naissance, nationalité, niveau d'étude, couple, ...), ou pour les variables sur les notions de couverture sociale par exemple. Il est au contraire bien plus faible pour les variables d'occupation à 2 ans, de sexe, de tabagisme ou d'état matrimonial.

On peut vérifier que ces taux sont bien dépendants de la stratification. Dans tous les graphiques suivants, chaque variable est calculée par maternité et présentée dans l'ordre de taille de ces dernières. Les maternités d'une même strate sont représentées par la même couleur.

Ainsi, en traçant la part des mères non fumeuses ou la part des garçons par maternité, on ne note aucun lien avec la strate. La dispersion parait même plus grande pour les petites maternités. Appliquer des taux de sondage plus faibles là où les ratios sont les plus dispersés rend négligeable l'effet stratification.

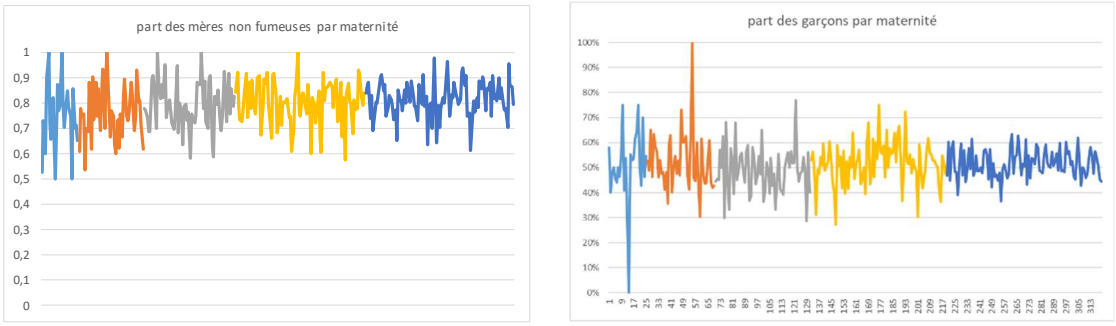


*Figure 8 – quelques indicateurs par strate*

Cette conclusion est la même lorsqu'on analyse la part de mères sans diabète gestationnel, ou la part d'enfants prenant un bain tous les jours à 2 ans.

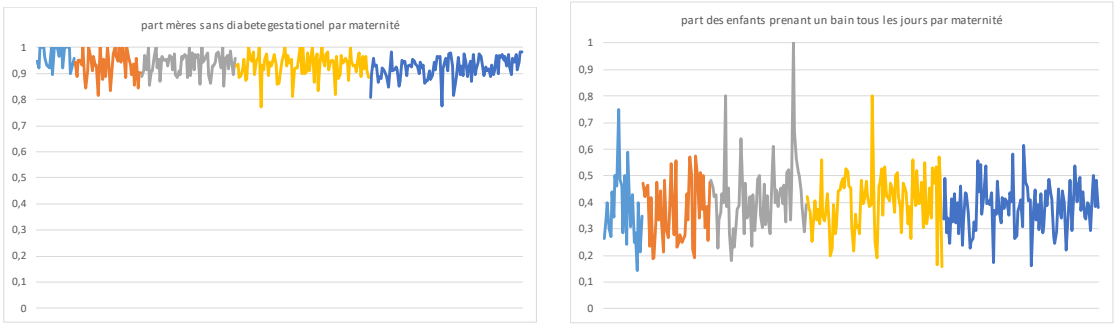


*Figure 9 – quelques indicateurs par strate*

A l'inverse, la part des mères françaises ou des mères occupant un emploi par maternité montre une dispersion plus importante pour les maternités des strates 4 et 5. La stratification apporte donc une amélioration importante dans la précision de l'analyse.

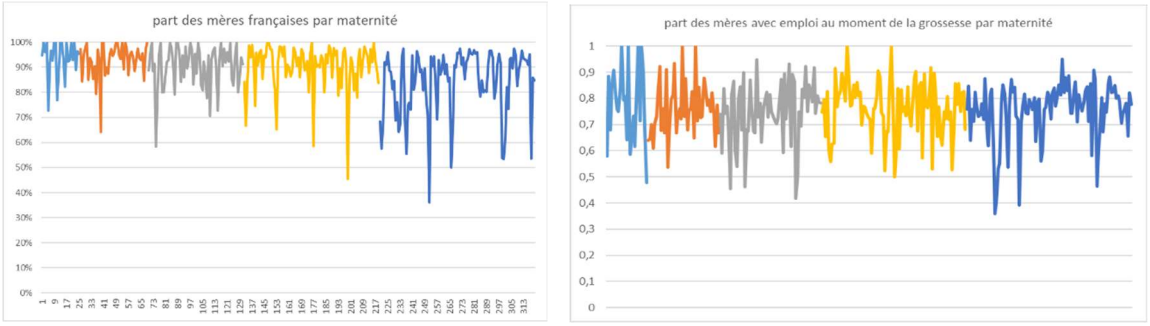


*Figure 10 – quelques indicateurs par strate*

L'**effet non réponse** mesure en quoi la non réponse dégrade la précision de l'enquête Elfe. Par sa définition même, on note que l'effet non réponse revient à ajouter à la variance un terme proportionnel à ${lin_k}^2$ (et donc proportionnel à $(y_k - \hat{R})^2$) et inversement proportionnel à la probabilité de répondre ${\emptyset^2}_k$.

Ainsi, si les individus avec les probabilités de réponse les plus faibles sont atypiques (loin du ratio global $\hat{R}$), l'ajout de la variance due à la non réponse est important. En effet, si par exemple le ratio estimé est inférieur à ½ (il y a donc moins d'enfants avec la caractéristique analysée que sans), les individus avec $y_k = 1$ (donc avec la caractéristique analysée) sont atypiques et sont les plus « loin » du ratio. Si ces individus ont une probabilité de réponse faible, la part de variance due à la NR sera importante.

Cet effet est bien évidemment un peu plus important pour les mesures aux 2 ans de l'enfant (12000 répondants contre 18000 en maternité).

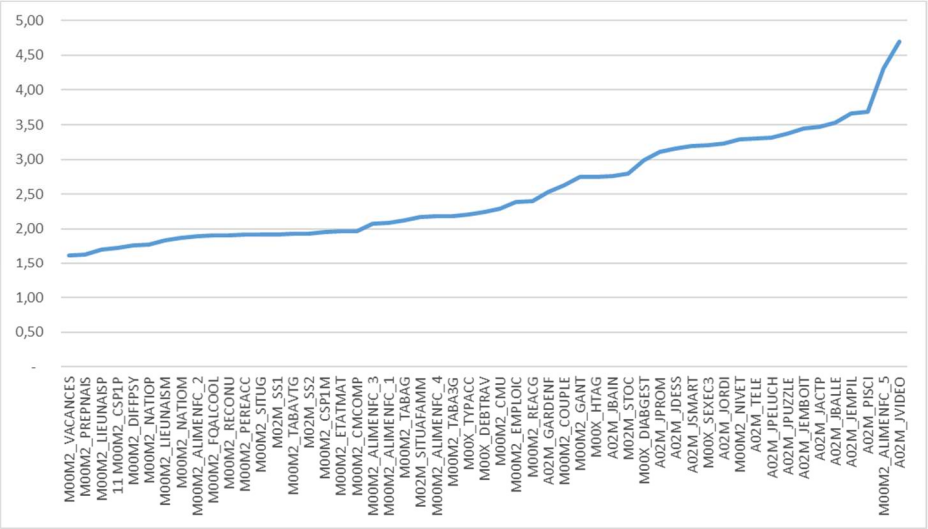


*Figure 11 – effet Non Réponse*

L'**effet plan** mesure en quoi la variance estimée avec le plan de sondage mis en œuvre dans l'enquête Elfe diffère d'une variance estimée par sondage aléatoire simple.

Il regroupe les 3 effets précédents, ces différents éléments le constituant étant relativement fluctuants. Comme on l'a déjà pressenti, cette évolution contrastée est surtout très marquée lorsqu'on compare effet non réponse et effet de grappe : ces 2 effets évoluent de manière inverse en fonction du nombre d'enfants enquêtés. En effet, moins on enquête de personnes, plus la perte de précision due à la non réponse est importante (la probabilité de réponse sera plus petite et donc l'effet non réponse plus grand, la précision se dégrade), mais moins l'effet de grappes est fort (le nombre moyen d'enfants enquêtés par maternité est plus faible, l'effet de grappes est donc plus faible, la perte de précision est moindre).

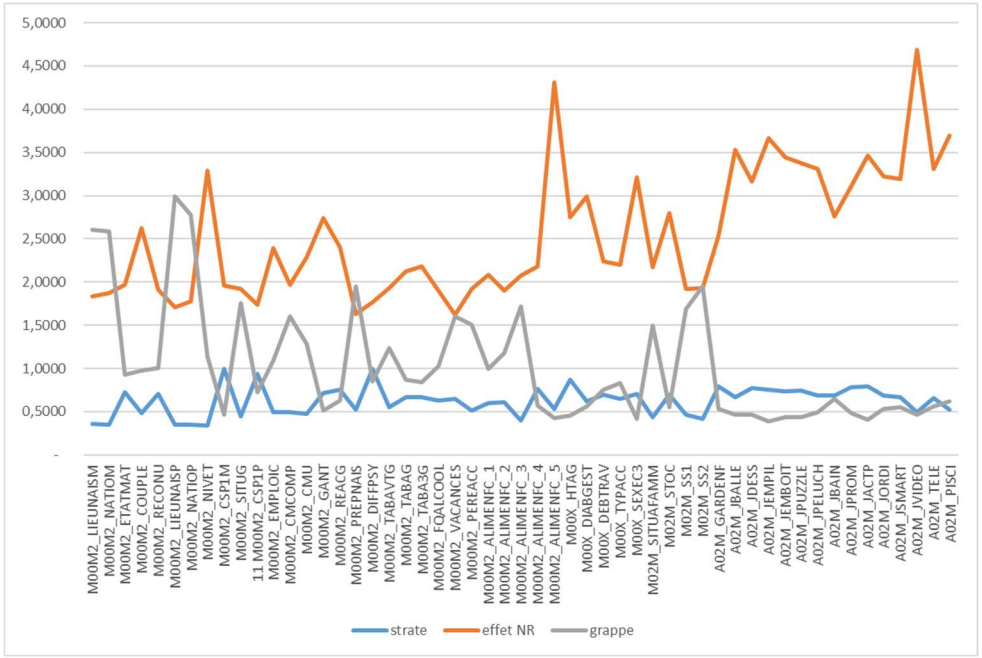


*Figure 12 – analyse des différents effets constituant le plan de sondage Elfe*

Au global, on peut donc noter une relative stabilité de l'effet plan entre 1 et 1,4. Il est logiquement un peu plus important lorsque, pour les variables étudiées, les non répondants ont un caractère atypique (nationalité des parents, lieu de naissance, …), l'effet NR étant dans ce cas supérieur et non compensé par les autres effets (on a

vu que l'impact simple d'une forte probabilité de réponse entrainait un effet NR important qui était partiellement compensé par un effet grappe faible).

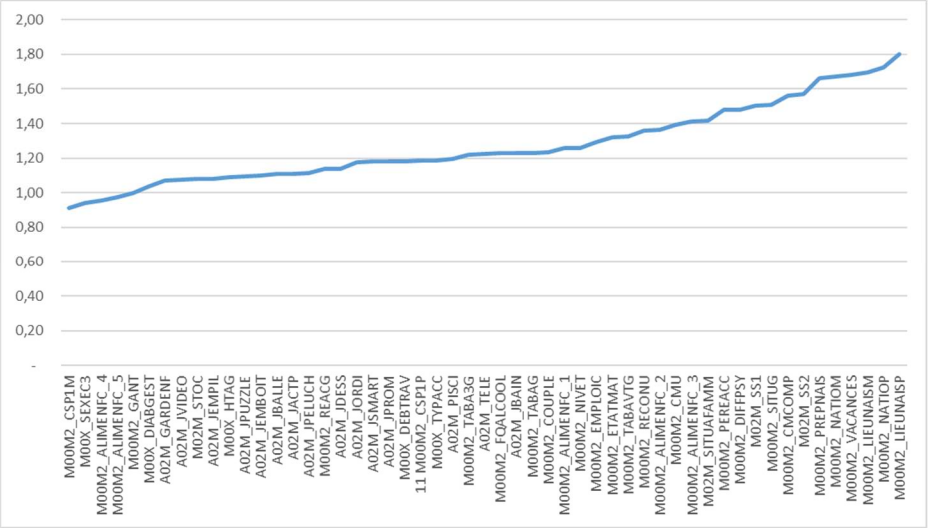


*Figure 13 - effet plan de sondage*

Enfin, l'**effet de calage** permet de mesurer, pour chaque variable d'intérêt, à quel point le calage améliore la précision des estimateurs. Globalement, un calage améliore toujours la précision d'un calcul, puisqu'il réduit l'incertitude en limitant certains aléas en « bloquant » les marges. Si la variable étudiée dépend (même un petit peu) des variables de calage, l'aléa dû au tirage aléatoire d'individus diminue forcément. A l'extrême, lorsqu'on mesure un taux de nourrissons avec une caractéristique dépendant directement des variables de calage, il est évident que la donnée ne subit plus aucun aléa, puisque la part de nourrissons avec une caractéristique donnée est « fixée » à l'avance. C'est pourquoi les variances après calage tombent à 0 pour les variables de lieu de naissance et de nationalité par exemple.

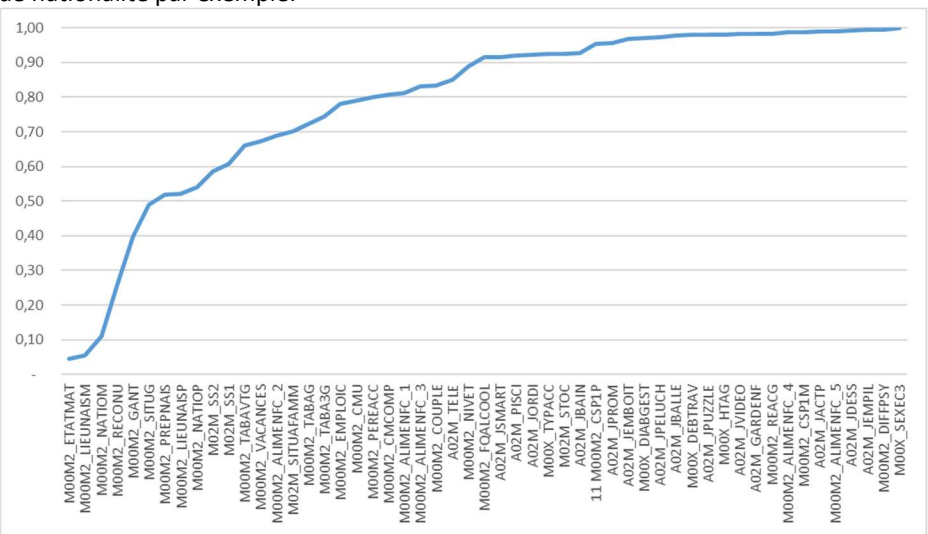


*Figure 14 - effet calage*

Il est intéressant de comparer l'effet plan et l'effet calage. Rappelons en effet que les variables de calage sont liées fortement aux variables expliquant la non réponse (c'est vrai jusqu'aux enquêtes 2 ans, cela l'est encore plus à partir des 3 ans ½ de l'enfant[4]). Par exemple, lors de l'enquête maternité, les variables expliquant la non réponse portent sur l'âge de la mère, l'âge gestationnel, la région d'habitation, la PCS de la mère, l'activité au

[4] Cf. note **Description de la méthode de pondération mise en œuvre aux différents temps d'enquêtes Elfe** : https://pandora.vjf.inserm.fr/doc/Pond%C3%A9rations_enqu%C3%AAtes_nationales.pdf

moment de la grossesse, l'indicatrice gémellaire et la primiparité. Les variables de calage sont l'âge, la région, l'état matrimonial, le statut immigré, le niveau d'étude et la primiparité, auxquels on peut ajouter l'indicatrice gémellaire puisqu'on réalise 2 calages distinguant les familles et les enfants. On retrouve donc dans les 2 ensembles des variables communes ou extrêmement corrélées (le niveau d'étude explique la PCS et l'activité, ...).

Ainsi, si la variable étudiée n'a aucun lien avec les variables de calage, il n'y a pas de raison que les répondants aient un caractère différent des non répondants, ou qu'une caractéristique particulière soit sous ou sur représentée. Le plan de sondage n'a que peu d'effet sur la précision de la variable analysée, et l'effet du calage est limité. Si, à l'inverse, la variable analysée est corrélée plus fortement avec les variables de calage, cela revient à dire que la variable analysée est liée au moins partiellement avec les variables expliquant la non réponse. On aura donc un effet plan plutôt important. Mais, par la mécanique même du calage qui diminue l'aléa sur la partie de non réponse expliquée par les variables de calage, l'effet calage sera bien meilleur.

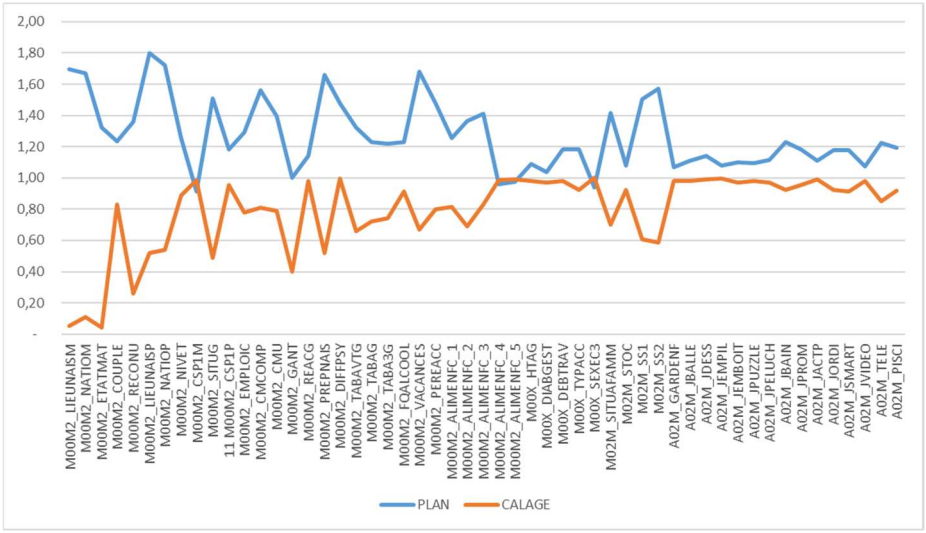


*Figure 15 – comparaison entre effet plan / effet calage*

Au final, **l'effet Elfe** compare la variance estimée en prenant en compte le plan de sondage simplifié et un sondage aléatoire simple. On notera que cet effet est très souvent inférieur à 1.2.

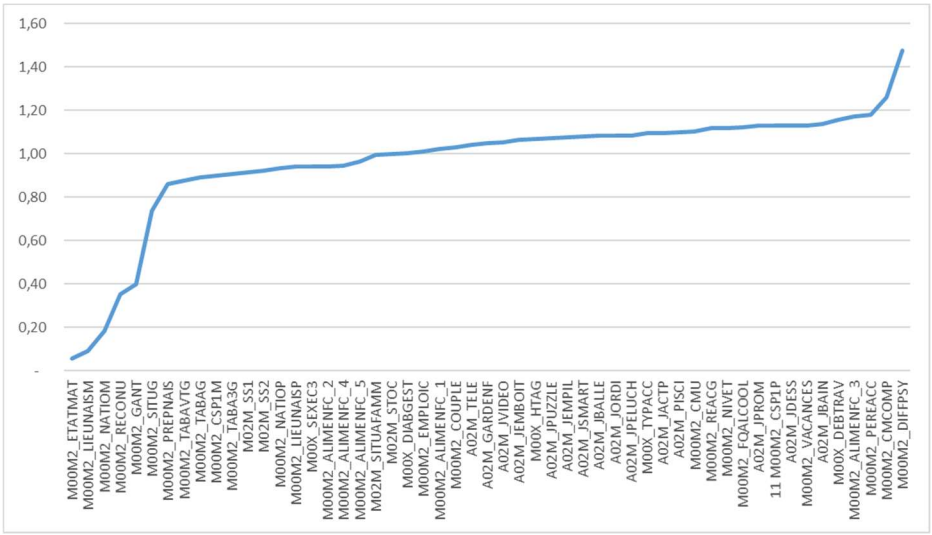


*Figure 16 - effet global Elfe : rapport entre variance Elfe simplifié / variance sondage aléatoire simple*

Lorsqu'on mesure ce rapport sur les écarts types plutôt que sur les variances, et donc sur la taille des intervalles de confiance, on obtient des rapports inférieurs à 1.1

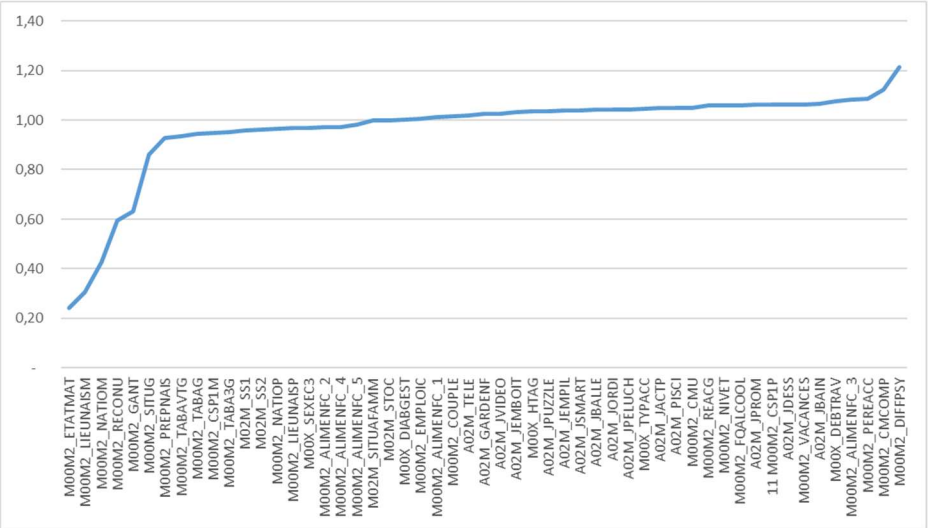


*Figure 17 - rapport entre Ecart type Elfe simplifié / Ecart type sondage aléatoire simple*

## 6. Quelques constatations aux 3ans ½ de l'enfant

Pour finir cette analyse des différents éléments constituant le plan de sondage Elfe, on a réalisé ces mêmes calculs sur quelques données recueillies aux 3 ans de l'enfant. Ces variables portent aussi bien sur les activités pratiquées, que sur les soins reçus ou les utilisations d'appareils électroniques.

Pour réaliser ces calculs, on a utilisé la pondération transversale « Enfant » issue de la méthode de calage simultané[5]. Pour son application aux calculs de variance, la principale différence avec la méthode de pondération utilisée jusqu'aux 2 ans est que les variables de calage sont plus nombreuses (les 6 variables utilisées jusque-là + 7 variables), et que ces nouvelles variables reprennent exactement les variables expliquant la non réponse. Il n'y a pas de calcul à proprement parler de la probabilité qu'un nourrisson participe à l'enquête, cette dernière étant estimée « a posteriori » en comparant le poids final après calage et le poids issu du plan de sondage.

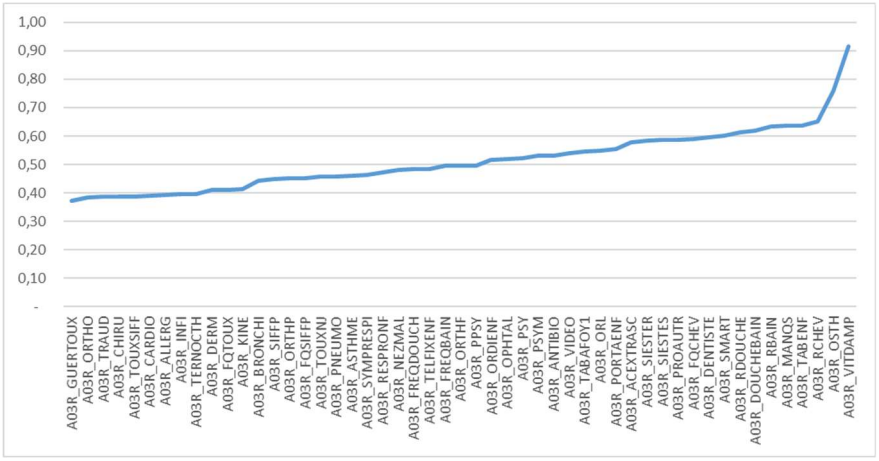


*Figure 18 – effet de grappes- enquête 3 ans 1/2*

Comme dans les enquêtes réalisées aux 2 ans de l'enfant, l'effet de grappe est concentré entre 0,4 et 0,6 pour les analyses réalisées à 3 ans ½. Le nombre de répondants par maternité (11700 répondants) est équivalent aux répondants à 2 ans, et les enfants nés dans une même maternité ont encore moins de raisons de se ressembler. Il n'y a plus aucun effet de grappe supérieur à 1. Quelle que soit la variable analysée, la variabilité entre maternités du taux de nourrissons avec la caractéristique recherchée est forcément plus faible que la variabilité de cette caractéristique mesurée directement sur les nourrissons.

[5] Cf. là encore la note « Pondérations des enquêtes nationales aux 3 ans ½ de l'enfant et au-delà »

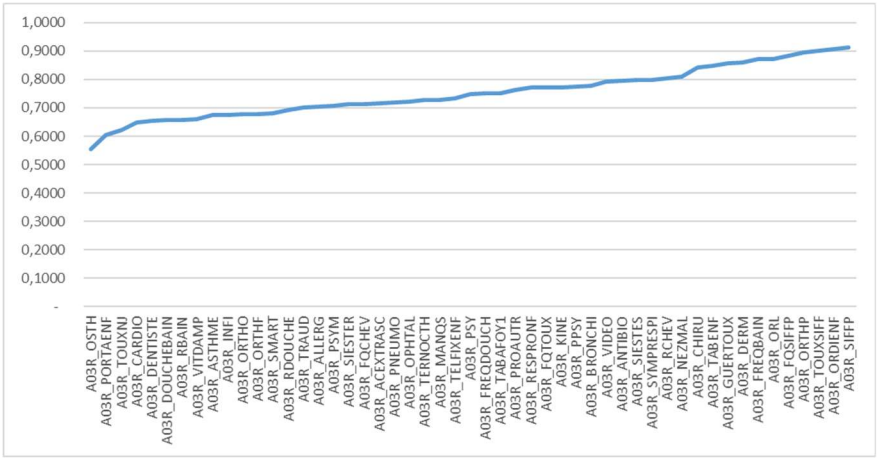


*Figure 19 – effet stratification- enquête 3 ans 1/2*

L'effet strate devient stable. Il y a de moins en moins de raisons pour que la taille de la maternité influe sur les analyses. On gagne donc simplement un effet dû au taux de sondage important dans les maternités de plus grande taille (et donc là où on a le plus enquêté).

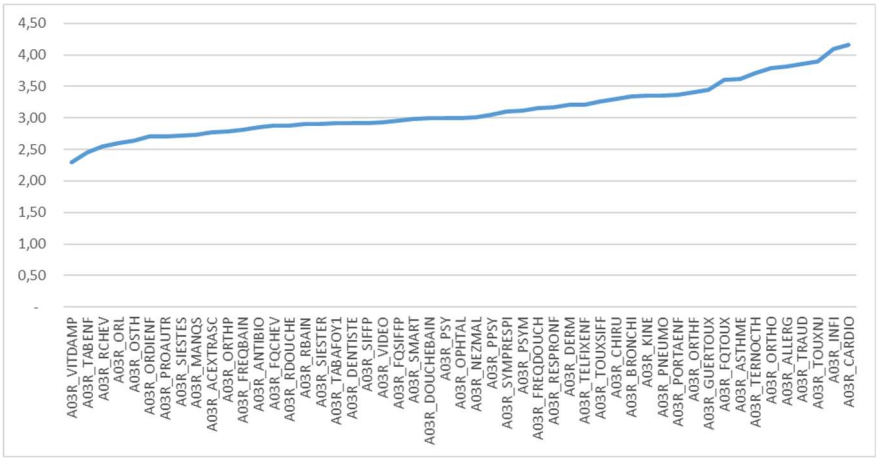


*Figure 20 - effet Non Réponse- enquête 3 ans 1/2*

Comme attendu, l'effet non réponse est important. Il est équivalent à l'effet non réponse obtenu sur les enquêtes 2 ans (autour de 3 voire plus), ce qui rend l'effet plan constant autour de 1,2 (ce n'est finalement que la résultante d'effets qui se stabilisent puisque le plan de sondage retenu en maternité - qui se concentre sur le fait que ce sont des nourrissons tirées dans des maternités elle mêmes réparties en strates - a de moins en moins d'impact global sur la précision des résultats, et donc d'un effet grappes autour de 0,5, effet strate vers 0,7, effet NR vers 3,5 soit effet plan = 0,5 x 0,7 x 3,5 = 1,2)

Mais là encore, il faut mettre en relation effet plan / effet de calage pour avoir l'effet global du plan de sondage mis en place dans l'enquête Elfe. On a choisi volontairement des variables ne dépendant pas ou peu des variables de calage (on a montré que si les variables analysées dépendaient de variables de calage, par construction même l'effet calage était très important et rendait la précision de l'analyse très bonne et très inférieure au sondage aléatoire simple). L'effet calage est donc relativement faible (entre 0,8 et 1).

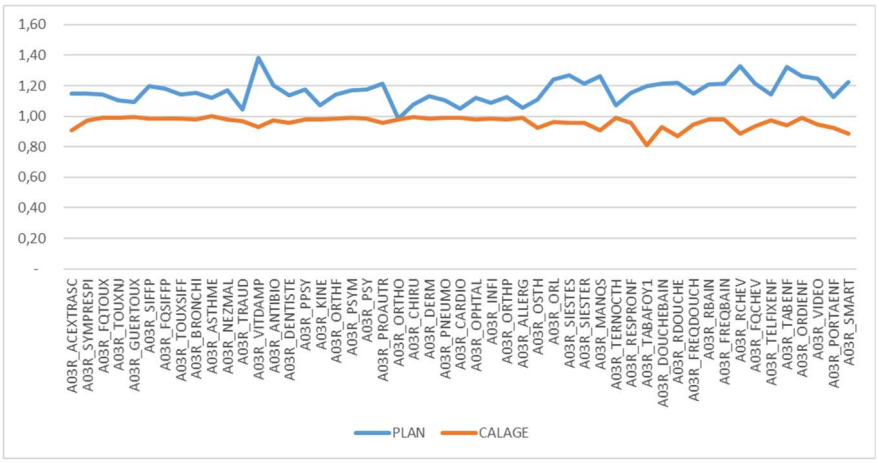


*Figure 21 – comparaison entre effet plan / effet calage- enquête 3 ans 1/2*

Au final, **l'effet Elfe** comparant la variance estimée en prenant en compte le plan de sondage simplifié et un sondage aléatoire simple est là encore situé entre 1 et 1,2. Lorsqu'on mesure ce rapport sur les écarts types plutôt que sur les variances, et donc sur la taille des intervalles de confiance, on obtient toujours des rapports inférieurs à 1,1.

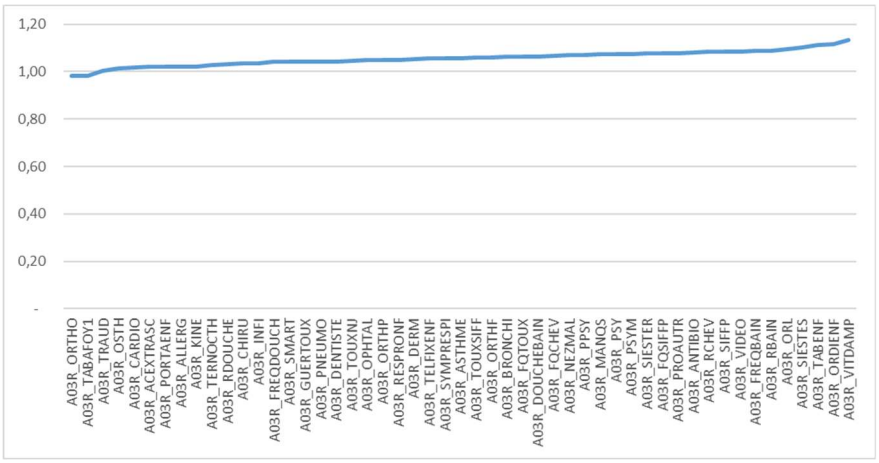


*Figure 22 - rapport entre Ecart type Elfe simplifié / Ecart type sondage aléatoire simple- enquête 3 ans 1/2*

# 7. Préconisations pour les utilisateurs de l'enquête Elfe

On a montré dans ce document que le plan de sondage complet mis en œuvre dans le cas de l'enquête Elfe pouvait se simplifier en négligeant un élément constitutif du plan (l'effet jour) sans pour autant que la précision des analyses ne puisse être remise en cause, grâce à la variabilité négligeable des données selon le jour d'enquête comparativement aux autres éléments constituant ce plan de sondage.

On a montré ensuite que, sous l'hypothèse de simplification proposée, en comparaison avec la précision obtenue dans le cas d'un sondage aléatoire simple, la précision du plan de sondage simplifié dépendait essentiellement de la taille de l'échantillon obtenu.

En effet, une baisse significative du nombre de répondants (et donc de la probabilité de participation des enfants) entraine mécaniquement une perte de précision due à la hausse de l'effet non réponse, mais que cet effet était compensé à la fois par la baisse du nombre de répondants par maternité pris en compte dans l'effet de grappes et à la fois par le principe même du calage pour lequel les variables retenues assuraient que la perte de précision due à la non réponse était en partie contrebalancée par l'effet calage.

On a également montré que ces effets se stabilisaient au cours du temps. L'impact de la méthode de tirage réalisé en maternité sur des données mesurées aux 2 ans puis aux 3ans ½ de l'enfant diminue fortement.

**Au final, on a vérifié sur une centaine de variables que la précision obtenue avec le plan de sondage simplifié est comparable à la précision du plan de sondage aléatoire simple** (sous-estimation de 10% des écarts types et souvent bien moins).

**On pourra donc conseiller à l'utilisateur de l'enquête Elfe d'utiliser les procédures SAS classiques, en étant éventuellement un peu conservateur sur les seuils des tests réalisés (par exemple réaliser des tests au seuil de significativité de 3% plutôt que les 5% habituellement utilisés), à savoir :**

Pour estimer **moyenne, proportion, fréquence**

- PROC SURVEYMEANS- variables continues
- PROC SURVEYFREQ- variables discrètes

Pour réaliser **une régression linéaire**

- PROC SURVEYREG- régression linéaire, test d'égalité.

Pour réaliser **une régression logistique**

- PROC SURVEYLOGISTIC- régression logistique

Attention enfin à bien utiliser les procédures Survey…. de SAS, et non les *proc means*, *freq*, *reg* par exemple qui sous-estiment très fortement les variances en estimant les éléments nécessaires aux tests comme si la table sur laquelle on estime les paramètres contenait l'ensemble de la population et non un échantillon issu d'une enquête réalisée par sondage. L'estimation de moyennes, totaux, ratios sera identique, mais l'estimation de variances, et donc la significativité des résultats peut être très différente.

# Annexe 1 : liste des variables analysées

| nom | |
|---|---|
| M00M2_LIEUNAISM | Lieu de naissance mère |
| M00M2_NATIOM | Nationalité mère |
| M00M2_ETATMAT | Etat matrimonial mère |
| M00M2_COUPLE | La mère vit en couple |
| M00M2_RECONU | Le père a reconnu l'enfant |
| M00M2_LIEUNAISP | Lieu de naissance père |
| M00M2_NATIOP | Nationalité père |
| M00M2_NIVET | Niveau d'études mère |
| M00M2_CSP1M | Recodage : profession et catégorie sociale de la mère |
| M00M2_SITUG | Emploi mère au moment de la grossesse |
| M00M2_CSP1P | Recodage : profession et catégorie sociale du père |
| M00M2_EMPLOIC | Situation professionnelle père |
| M00M2_CMCOMP | Couverture maladie complémentaire |
| M00M2_CMU | CMU complémentaire |
| M00M2_GANT | Grossesse(s) antérieure(s) |
| M00M2_REACG | Réaction à la découverte de la grossesse |
| M00M2_PREPNAIS | Séances de préparation à la naissance |
| M00M2_DIFFPSY | Difficultés psy pendant la grossesse |
| M00M2_TABAVTG | Tabagisme avant la grossesse |
| M00M2_TABAG | Tabagisme pendant la grossesse |
| M00M2_TABA3G | Tabagisme pendant le 3e trimestre |
| M00M2_FQALCOOL | Consommation d'alcool |
| M00M2_VACANCES | Vacances pendant la grossesse |
| M00M2_PEREACC | Le père a assisté à l'accouchement |
| M00M2_ALIMENFC_1 | Alimentation de l'enfant : lait maternel uniquement |
| M00M2_ALIMENFC_2 | Alimentation de l'enfant : lait 1er âge uniquement |
| M00M2_ALIMENFC_3 | Alimentation de l'enfant : allaitement mixte |
| M00M2_ALIMENFC_4 | Alimentation de l'enfant : NSP |
| M00M2_ALIMENFC_5 | Alimentation de l'enfant : autre |
| M00X_HTAG | Hypertension artérielle pendant la grossesse |
| M00X_DIABGEST | Diabète gestationnel |
| M00X_DEBTRAV | Début du travail |
| M00X_TYPACC | Accouchement |
| M00X_SEXEC3 | Sexe |
| M02M_SITUAFAMM | Situation familiale de la mère |
| M02M_STOC | Situation du ménage par rapport au logement |
| M02M_SS1 | Régime de sécurité sociale |
| M02M_SS2 | Couverture maladie complémentaire |
| A02M_GARDENF | Mode de garde principal semaine |
| A02M_JBALLE | Balle |
| A02M_JDESS | Dessin ou peinture |
| A02M_JEMPIL | Empiler |
| A02M_JEMBOIT | Emboîter |
| A02M_JPUZZLE | Puzzles |
| A02M_JPELUCH | Peluches |
| A02M_JBAIN | Jeux de bains ou jeux d'eau |
| A02M_JPROM | Promenades |
| A02M_JACTP | Jeux / activités physiques |
| A02M_JORDI | Jm_ordi |
| A02M_JSMART | Smartphone |
| A02M_JVIDEO | Jeux vidéo |
| A02M_TELE | Télévision |
| A02M_PISCI | Piscine |

# Annexe 2 : part des effets Jour, Maternité, Non réponse dans l'estimation de variance théorique avant calage et analyse des dispersions moyennes

| nom | part effet DAY | part effet MAT | part effet NR |
|---|---|---|---|
| M00M2_LIEUNAISM | 51% | 27% | 22% |
| M00M2_NATIOM | 50% | 27% | 23% |
| M00M2_ETATMAT | 43% | 29% | 28% |
| M00M2_COUPLE | 31% | 26% | 43% |
| M00M2_RECONU | 34% | 35% | 32% |
| M00M2_LIEUNAISP | 34% | 39% | 27% |
| M00M2_NATIOP | 37% | 35% | 27% |
| M00M2_NIVET | 56% | 13% | 31% |
| M00M2_CSP1M | 31% | 35% | 34% |
| M00M2_SITUG | 30% | 36% | 33% |
| 11 M00M2_CSP1P | 53% | 27% | 20% |
| M00M2_EMPLOIC | 31% | 29% | 40% |
| M00M2_CMCOMP | 40% | 31% | 30% |
| M00M2_CMU | 49% | 22% | 29% |
| M00M2_GANT | 58% | 15% | 27% |
| M00M2_REACG | 39% | 25% | 35% |
| M00M2_PREPNAIS | 42% | 36% | 22% |
| M00M2_DIFFPSY | 29% | 40% | 31% |
| M00M2_TABAVTG | 50% | 26% | 24% |
| M00M2_TABAG | 49% | 24% | 27% |
| M00M2_TABA3G | 46% | 25% | 29% |
| M00M2_FQALCOOL | 65% | 18% | 17% |
| M00M2_VACANCES | 39% | 38% | 23% |
| M00M2_PEREACC | 64% | 19% | 17% |
| M00M2_ALIMENFC_1 | 38% | 30% | 32% |
| M00M2_ALIMENFC_2 | 46% | 28% | 25% |
| M00M2_ALIMENFC_3 | 36% | 31% | 33% |
| M00M2_ALIMENFC_4 | 61% | 18% | 21% |
| M00M2_ALIMENFC_5 | 47% | 12% | 41% |
| M00X_HTAG | 50% | 18% | 32% |
| M00X_DIABGEST | 61% | 13% | 26% |
| M00X_DEBTRAV | 88% | 5% | 7% |
| M00X_TYPACC | 78% | 10% | 12% |
| M00X_SEXEC3 | 28% | 23% | 50% |
| M02M_SITUAFAMM | 37% | 29% | 34% |
| M02M_STOC | 55% | 16% | 29% |
| M02M_SS1 | 46% | 28% | 26% |
| M02M_SS2 | 36% | 33% | 31% |
| A02M_GARDENF | 62% | 15% | 23% |
| A02M_JBALLE | 33% | 19% | 48% |
| A02M_JDESS | 35% | 20% | 44% |
| A02M_JEMPIL | 34% | 18% | 48% |
| A02M_JEMBOIT | 39% | 18% | 43% |
| A02M_JPUZZLE | 35% | 19% | 46% |
| A02M_JPELUCH | 47% | 16% | 37% |
| A02M_JBAIN | 38% | 23% | 40% |
| A02M_JPROM | 54% | 15% | 31% |
| A02M_JACTP | 41% | 17% | 42% |
| A02M_JORDI | 37% | 19% | 43% |
| A02M_JSMART | 37% | 20% | 43% |
| A02M_JVIDEO | 64% | 8% | 28% |
| A02M_TELE | 32% | 21% | 48% |
| A02M_PISCI | 29% | 19% | 52% |

| nom | rapport effet DAY / effet MAT | rapport dispersion MAT / dispersion DAY |
|---|---|---|
| M00M2_LIEUNAISM | 1,92 | 49,54 |
| M00M2_NATIOM | 1,89 | 53,96 |
| M00M2_ETATMAT | 1,48 | 72,37 |
| M00M2_COUPLE | 1,16 | 30,75 |
| M00M2_RECONU | 0,97 | 58,91 |
| M00M2_LIEUNAISP | 0,88 | 115,04 |
| M00M2_NATIOP | 1,06 | 89,39 |
| M00M2_NIVET | 4,17 | 27,00 |
| M00M2_CSP1M | 0,89 | 92,83 |
| M00M2_SITUG | 0,84 | 67,29 |
| 11 M00M2_CSP1P | 1,99 | 55,37 |
| M00M2_EMPLOIC | 1,09 | 28,15 |
| M00M2_CMCOMP | 1,29 | 23,77 |
| M00M2_CMU | 2,18 | 17,24 |
| M00M2_GANT | 3,76 | 15,94 |
| M00M2_REACG | 1,56 | 25,33 |
| M00M2_PREPNAIS | 1,18 | 26,55 |
| M00M2_DIFFPSY | 0,72 | 21,09 |
| M00M2_TABAVTG | 1,94 | 55,00 |
| M00M2_TABAG | 2,05 | 32,87 |
| M00M2_TABA3G | 1,87 | 31,51 |
| M00M2_FQALCOOL | 3,57 | 24,59 |
| M00M2_VACANCES | 1,03 | 2,80 |
| M00M2_PEREACC | 3,43 | 16,38 |
| M00M2_ALIMENFC_1 | 1,28 | 51,91 |
| M00M2_ALIMENFC_2 | 1,64 | 46,51 |
| M00M2_ALIMENFC_3 | 1,17 | 45,53 |
| M00M2_ALIMENFC_4 | 3,36 | 18,18 |
| M00M2_ALIMENFC_5 | 3,86 | 41,37 |
| M00X_HTAG | 2,77 | 14,02 |
| M00X_DIABGEST | 4,78 | 11,41 |
| M00X_DEBTRAV | 16,16 | 3,70 |
| M00X_TYPACC | 7,86 | 8,31 |
| M00X_SEXEC3 | 1,22 | 31,80 |
| M02M_SITUAFAMM | 1,30 | 40,18 |
| M02M_STOC | 3,48 | 10,62 |
| M02M_SS1 | 1,67 | 26,78 |
| M02M_SS2 | 1,11 | 51,45 |
| A02M_GARDENF | 4,05 | 29,35 |
| A02M_JBALLE | 1,73 | 4,27 |
| A02M_JDESS | 1,72 | 5,94 |
| A02M_JEMPIL | 1,89 | 10,07 |
| A02M_JEMBOIT | 2,22 | 6,79 |
| A02M_JPUZZLE | 1,80 | 14,24 |
| A02M_JPELUCH | 2,96 | 16,60 |
| A02M_JBAIN | 1,68 | 27,48 |
| A02M_JPROM | 3,68 | 8,14 |
| A02M_JACTP | 2,41 | 5,97 |
| A02M_JORDI | 1,93 | 26,26 |
| A02M_JSMART | 1,91 | 32,28 |
| A02M_JVIDEO | 8,31 | 16,53 |
| A02M_TELE | 1,54 | 27,11 |
| A02M_PISCI | 1,51 | 39,42 |

# Annexe 3 : Quantification des différents éléments constituant l'estimation de la variance

| nom | ratio | $\hat{V}_{SAS}(lin_k)$ | $\hat{V}_{GR}(lin_k)$ | $\hat{V}_{GR\_ST}(lin_k)$ | $\hat{V}_{NR}(lin_k)$ | $\hat{V}_{GR\_ST}(lin_k)+\hat{V}_{NR}(lin_k)$ | $\hat{V}_{GR\_ST}(\varepsilon_k)+\hat{V}_{NR}(\varepsilon_k)$ |
|---|---|---|---|---|---|---|---|
| M00M2_LIEUNAISM | 80,2% | 1,830E-05 | 4,768E-05 | 1,690E-05 | 1,410E-05 | 3,099E-05 | 1,693E-06 |
| M00M2_NATIOM | 81,8% | 1,744E-05 | 4,504E-05 | 1,556E-05 | 1,355E-05 | 2,911E-05 | 3,189E-06 |
| M00M2_ETATMAT | 43,9% | 2,115E-05 | 1,960E-05 | 1,423E-05 | 1,373E-05 | 2,796E-05 | 1,236E-06 |
| M00M2_COUPLE | 91,3% | 9,623E-06 | 9,391E-06 | 4,524E-06 | 7,353E-06 | 1,188E-05 | 9,889E-06 |
| M00M2_RECONU | 48,0% | 2,144E-05 | 2,159E-05 | 1,528E-05 | 1,385E-05 | 2,913E-05 | 7,557E-06 |
| M00M2_LIEUNAISP | 77,5% | 1,880E-05 | 5,628E-05 | 1,980E-05 | 1,405E-05 | 3,385E-05 | 1,767E-05 |
| M00M2_NATIOP | 79,1% | 1,802E-05 | 5,004E-05 | 1,750E-05 | 1,353E-05 | 3,103E-05 | 1,679E-05 |
| M00M2_NIVET | 1,3% | 2,216E-06 | 2,518E-06 | 8,480E-07 | 1,939E-06 | 2,787E-06 | 2,479E-06 |
| M00M2_CSP1M | 0,3% | 2,083E-07 | 9,770E-08 | 9,696E-08 | 9,280E-08 | 1,898E-07 | 1,874E-07 |
| M00M2_SITUG | 64,0% | 2,443E-05 | 4,291E-05 | 1,914E-05 | 1,767E-05 | 3,681E-05 | 1,805E-05 |
| 11 M00M2_CSP1P | 1,5% | 1,136E-06 | 8,244E-07 | 7,759E-07 | 5,694E-07 | 1,345E-06 | 1,281E-06 |
| M00M2_EMPLOIC | 84,3% | 1,551E-05 | 1,692E-05 | 8,385E-06 | 1,165E-05 | 2,004E-05 | 1,565E-05 |
| M00M2_CMCOMP | 84,8% | 1,542E-05 | 2,477E-05 | 1,223E-05 | 1,185E-05 | 2,408E-05 | 1,943E-05 |
| M00M2_CMU | 7,5% | 9,034E-06 | 1,161E-05 | 5,503E-06 | 7,089E-06 | 1,259E-05 | 9,947E-06 |
| M00M2_GANT | 67,8% | 1,814E-05 | 9,225E-06 | 6,616E-06 | 1,154E-05 | 1,815E-05 | 7,234E-06 |
| M00M2_REACG | 73,9% | 1,731E-05 | 1,087E-05 | 8,208E-06 | 1,151E-05 | 1,972E-05 | 1,936E-05 |
| M00M2_PREPNAIS | 48,2% | 2,118E-05 | 4,120E-05 | 2,158E-05 | 1,360E-05 | 3,518E-05 | 1,825E-05 |
| M00M2_DIFFPSY | 12,9% | 9,967E-06 | 8,454E-06 | 8,374E-06 | 6,381E-06 | 1,476E-05 | 1,469E-05 |
| M00M2_TABAVTG | 41,4% | 2,065E-05 | 2,558E-05 | 1,418E-05 | 1,320E-05 | 2,737E-05 | 1,809E-05 |
| M00M2_TABAG | 21,3% | 1,524E-05 | 1,329E-05 | 8,822E-06 | 9,945E-06 | 1,877E-05 | 1,358E-05 |
| M00M2_TABA3G | 17,3% | 1,326E-05 | 1,110E-05 | 7,416E-06 | 8,752E-06 | 1,617E-05 | 1,202E-05 |
| M00M2_FQALCOOL | 15,0% | 9,686E-06 | 9,892E-06 | 6,239E-06 | 5,645E-06 | 1,188E-05 | 1,086E-05 |
| M00M2_VACANCES | 47,8% | 2,121E-05 | 3,409E-05 | 2,200E-05 | 1,366E-05 | 3,566E-05 | 2,396E-05 |
| M00M2_PEREACC | 76,6% | 1,839E-05 | 2,772E-05 | 1,413E-05 | 1,302E-05 | 2,715E-05 | 2,170E-05 |
| M00M2_ALIMENFC_1 | 57,5% | 2,148E-05 | 2,146E-05 | 1,292E-05 | 1,405E-05 | 2,698E-05 | 2,192E-05 |
| M00M2_ALIMENFC_2 | 32,0% | 1,906E-05 | 2,243E-05 | 1,369E-05 | 1,232E-05 | 2,601E-05 | 1,795E-05 |
| M00M2_ALIMENFC_3 | 9,2% | 8,367E-06 | 1,440E-05 | 5,698E-06 | 6,122E-06 | 1,182E-05 | 9,812E-06 |
| M00M2_ALIMENFC_4 | 0,3% | 2,138E-07 | 1,227E-07 | 9,395E-08 | 1,108E-07 | 2,047E-07 | 2,019E-07 |
| M00M2_ALIMENFC_5 | 0,4% | 4,237E-07 | 1,792E-07 | 9,590E-08 | 3,175E-07 | 4,134E-07 | 4,092E-07 |
| M00X_HTAG | 1,7% | 1,813E-06 | 8,272E-07 | 7,196E-07 | 1,259E-06 | 1,978E-06 | 1,938E-06 |
| M00X_DIABGEST | 7,3% | 6,198E-06 | 3,485E-06 | 2,143E-06 | 4,273E-06 | 6,416E-06 | 6,217E-06 |
| M00X_DEBTRAV | 69,0% | 1,868E-05 | 1,416E-05 | 9,857E-06 | 1,221E-05 | 2,207E-05 | 2,160E-05 |
| M00X_TYPACC | 67,2% | 1,873E-05 | 1,550E-05 | 1,008E-05 | 1,212E-05 | 2,221E-05 | 2,052E-05 |
| M00X_SEXEC3 | 51,0% | 2,151E-05 | 8,943E-06 | 6,303E-06 | 1,394E-05 | 2,024E-05 | 2,022E-05 |
| M02M_SITUAFAMM | 76,9% | 2,125E-05 | 3,179E-05 | 1,390E-05 | 1,621E-05 | 3,011E-05 | 2,114E-05 |
| M02M_STOC | 7,3% | 6,598E-06 | 3,660E-06 | 2,547E-06 | 4,576E-06 | 7,123E-06 | 6,586E-06 |
| M02M_SS1 | 66,5% | 2,371E-05 | 4,009E-05 | 1,854E-05 | 1,715E-05 | 3,569E-05 | 2,170E-05 |
| M02M_SS2 | 72,6% | 2,345E-05 | 4,564E-05 | 1,903E-05 | 1,779E-05 | 3,682E-05 | 2,160E-05 |
| A02M_GARDENF | 18,9% | 1,333E-05 | 7,131E-06 | 5,629E-06 | 8,629E-06 | 1,426E-05 | 1,399E-05 |
| A02M_JBALLE | 27,8% | 2,891E-05 | 1,357E-05 | 9,085E-06 | 2,296E-05 | 3,204E-05 | 3,135E-05 |
| A02M_JDESS | 23,0% | 2,355E-05 | 1,095E-05 | 8,488E-06 | 1,835E-05 | 2,683E-05 | 2,659E-05 |
| A02M_JEMPIL | 18,7% | 2,068E-05 | 8,077E-06 | 6,099E-06 | 1,626E-05 | 2,236E-05 | 2,225E-05 |
| A02M_JEMBOIT | 21,6% | 2,218E-05 | 9,617E-06 | 7,072E-06 | 1,728E-05 | 2,435E-05 | 2,358E-05 |
| A02M_JPUZZLE | 7,5% | 8,691E-06 | 3,775E-06 | 2,813E-06 | 6,682E-06 | 9,495E-06 | 9,300E-06 |
| A02M_JPELUCH | 36,0% | 3,058E-05 | 1,503E-05 | 1,030E-05 | 2,381E-05 | 3,411E-05 | 3,318E-05 |
| A02M_JBAIN | 52,6% | 3,365E-05 | 2,187E-05 | 1,500E-05 | 2,636E-05 | 4,136E-05 | 3,829E-05 |
| A02M_JPROM | 27,3% | 2,875E-05 | 1,390E-05 | 1,094E-05 | 2,302E-05 | 3,396E-05 | 3,242E-05 |
| A02M_JACTP | 21,9% | 2,419E-05 | 9,809E-06 | 7,744E-06 | 1,908E-05 | 2,683E-05 | 2,653E-05 |
| A02M_JORDI | 11,4% | 1,505E-05 | 7,964E-06 | 5,487E-06 | 1,221E-05 | 1,770E-05 | 1,632E-05 |
| A02M_JSMART | 9,6% | 1,284E-05 | 7,101E-06 | 4,741E-06 | 1,040E-05 | 1,514E-05 | 1,383E-05 |
| A02M_JVIDEO | 0,5% | 9,126E-07 | 4,229E-07 | 2,090E-07 | 7,714E-07 | 9,804E-07 | 9,622E-07 |
| A02M_TELE | 17,1% | 2,655E-05 | 1,491E-05 | 9,826E-06 | 2,266E-05 | 3,249E-05 | 2,762E-05 |
| A02M_PISCI | 9,4% | 1,641E-05 | 1,021E-05 | 5,310E-06 | 1,429E-05 | 1,960E-05 | 1,804E-05 |

# Annexe 4 : liste des variables analysées aux 3 ans ½ de l'enfant

| nom | |
|---|---|
| A03R_ACEXTRASC | Cette année, [enfant elfe] pratique-t-il/elle régulièrement une activité de loisir dans un club ou association, comme par exemple du judo, du dessin ou de la musique (en dehors de l'école et du centre de loisir) ? |
| A03R_SYMPRESPI | [ enfant elfe] a-t-il/elle déjà eu, dans les 12 derniers mois une toux, une gêne respiratoire ou un épisode de sifflements ? |
| A03R_FQTOUX | Ces épisodes de toux surviennent-ils ? |
| A03R_TOUXNJ | [ enfant elfe] tousse-t-il/elle ? |
| A03R_GUERTOUX | Entre les épisodes de toux, [enfant elfe] est-il/elle complètement guéri(e) ? |
| A03R_SIFFP | [ enfant elfe] a-t-il/elle déjà eu, au cours des 12 derniers mois, au moins un épisode de sifflements dans la poitrine ? |
| A03R_FQSIFFP | Ces épisodes de sifflements surviennent-ils ? |
| A03R_TOUXSIFF | Ces sifflements accompagnent-ils toujours les épisodes de toux ? |
| A03R_BRONCHI | [ enfant elfe] a-t-il/elle fait une bronchiolite depuis l'âge de 2 ans ? |
| A03R_ASTHME | [ enfant elfe] a-t-il/elle eu des crises d'asthme au cours des 12 derniers mois ? |
| A03R_NEZMAL | Considérez-vous que [enfant elfe] a souvent le nez bouché ou le nez qui coule ? |
| A03R_TRAUD | [ enfant elfe] est-il/elle suivi pour un trouble de l'audition ? |
| A03R_VITDAMP | Au cours des 12 derniers mois, [enfant elfe] a-t-il/elle pris de la vitamine d sous forme d'ampoule (zymad, vitamine d3 bon, uvedose) ou en doses quotidiennes (zymad, zymaduo, uvesterol…) ? |
| A03R_ANTIBIO | Au cours des 12 derniers mois [enfant elfe] a-t-il/elle reçu un traitement antibiotique (clamoxyl, hiconcil, agram, amoxicilline, augmentin, ciblor, orelox, penicilline g, oroken, bristopen, bactrim, rocephine, josacine, zythromax, pediazole, pyostacine) |
| A03R_DENTISTE | Dentiste |
| A03R_PPSY | Pédopsychiatre |
| A03R_KINE | Kinésithérapeute |
| A03R_ORTHF | Orthophoniste |
| A03R_PSYM | Psychomotricien |
| A03R_PSY | Psychologue |
| A03R_PROAUTR | Un ou d'autres professionnels de santé spécialisés |
| A03R_ORTHO | Orthopédiste |
| A03R_CHIRU | Un chirurgien autre qu'orthopédiste |
| A03R_DERM | Dermatologue |
| A03R_PNEUMO | Pneumologue |
| A03R_CARDIO | Cardiologue |
| A03R_OPHTAL | Ophtalmologiste |
| A03R_INFI | Infirmière |
| A03R_ORTHP | Orthoptiste |
| A03R_ALLERG | Allergologue |
| A03R_OSTH | Ostéopathe |
| A03R_ORL | Orl |
| A03R_SIESTES | [ enfant elfe] fait-il/elle la sieste en semaine ? |
| A03R_SIESTER | Fait-il/elle la sieste le week-end, en vacances ? |
| A03R_MANQS | Selon vous [enfant elfe] manque-t-il/elle de sommeil ? |
| A03R_TERNOCTH | Lui arrive-t-il de se réveiller la nuit en criant, en étant confus(e), impossible à approcher, sans s'en souvenir le matin ? |
| A03R_RESPRONF | Lorsque [enfant elfe] n'est pas enrhumé(e), à quelle fréquence ronfle-t-il/elle ? |
| A03R_TABAFOY1 | Oui, un fumeur |
| A03R_DOUCHEBAIN | [ enfant elfe] prend-il/elle : |
| A03R_RDOUCHE | Indiquer un nombre de fois |
| A03R_FREQDOUCH | A quelle fréquence [enfant elfe] prend-t-il/elle une douche sans compter les bains ? |
| A03R_RBAIN | Indiquer un nombre de fois |
| A03R_FREQBAIN | Freqbain |
| A03R_RCHEV | Indiquer un nombre de fois |
| A03R_FQCHEV | En général à quel rythme [enfant elfe] a-t-il/elle les cheveux lavés ? |
| A03R_TELFIXENF | Arrive-t-il à [enfant elfe] de parle-t-il/elle au téléphone sans fil au moins une fois par semaine ? |
| A03R_TABENF | [ enfant elfe] utilise-t-il/elle une tablette au domicile au moins une fois par semaine ? |
| A03R_ORDIENF | [ enfant elfe] utilise-t-il/elle un ordinateur au domicile au moins une fois par semaine ? |
| A03R_VIDEO | [ enfant elfe] joue-t-il/elle actuellement à des jeux vidéo sur une console (wii, psp, xbox, ds, …) au moins une fois par semaine ? |
| A03R_PORTAENF | Arrive-t-il à [enfant elfe] de parle-t-il/elle au téléphone portable au moins une fois par semaine ? |
| A03R_SMART | [ enfant elfe] joue-t-il/elle actuellement sur un téléphone portable au moins une fois par semaine ? |

# Annexe 5 : Code SAS utilisé pour générer les différents éléments constituant l'estimation de la variance

Soit une table contenant au minimum les champs suivants:

```
Table =nom de la table où sont les données.

-       Les variables de calage CS_1 à CS_13
-       identifiant = champs identifiant les enregistrements ID...
-       strate = champ qui identifie la strate de la maternité. Par défaut M00M1_MATSTRATEC1
-       vague = champ qui identifie la vague de l'enquête. Par défaut M00M1_VAGUE
-       mater = champ qui identifie la maternité. Par défaut M00M1_IDGROUPNAMEALEAC1
-       jour = champ qui identifie le jour. Par défaut M00M2_JNAISSEALEA

-       poids = nom de la variable de pondération. Par exemple M00E_PONDVALC2

-       variable = nom de la variable pour laquelle on va calculer le TOTAL
        OU
-       variable1 = nom de la variable pour laquelle on va calculer le NUMERATEUR
-       variable2 = nom de la variable pour laquelle on va calculer le DENOMINATEUR

-       méthode=1 si enquête AVANT 3 ans 1/2, méthode = 2 si enquête 3 ans 1/2 et suivantes

data degreM;
input strateN _TOTAL_;
datalines;
1 108
2 108
3 109
4 108
5 111
;
data degreJ;
input vagueN _TOTAL_;
datalines;
1 90
2 91
3 92
4 92
;

%if &methode=1 %then %let listeCALAGE= CS_1 CS_2 CS_3 CS_4 CS_5 CS_6 ;
%if &methode=2 %then %let listeCALAGE= CS_1 CS_2 CS_3 CS_4 CS_5 CS_6 CS_7 CS_8 CS_9 CS_10
CS_11 CS_12 CS_13;
```

Pour un TOTAL :

```
/* régression pondérée par poids de la variable d'intérêt sur les variables de calage*/
proc glm data=table noprint ;
        class &listeCALAGE ;
        model &variable = &listeCALAGE ;
        weight &poids;
        output out=residus RESIDUAL = res;
run;
```

Pour un RATIO :

```
/* estimation du total de la variable1 pour le numérateur, total variable2 pour le
dénominateur et du ratio */
proc sql;
create table ESTIMATION_TOTAL as
select N as estim_NUM, D as estim_DEN, N/D as estimateur from
(select sum (&variable1 * &poids) as N, sum (&variable2 * &poids) as D
from  table) as t;quit;

data _null_;
set ESTIMATION_TOTAL;
CALL SYMPUT('numerateur',estim_NUM);
CALL SYMPUT('denominateur',estim_DEN);
CALL SYMPUT('ratio',estimateur);run;

/*création de lin*/
```

```
data table;
set table;
Ratio_i = (1/&denominateur)*(&variable1 - &ratio*&variable2);run;

/* régression pondérée par poids de la linéarisée sur les variables de calage*/
proc glm data=table noprint ;
        class &listeCALAGE ;
        model Ratio_i = &listeCALAGE ;
        weight &poids;
        output out=residus RESIDUAL = res;
run;
```

Puis calculer :

```
data residus (keep= &identifiant res); set residus; run;
proc sort data= table; by &identifiant; run;
proc sort data= residus; by &identifiant; run;

data tableRES; merge table residus; by &identifiant; run;

/* effet MAT*/
proc surveymeans data=tableRES total=degreM mean clm stderr var sum clsum std varsum;
weight &poids;
cluster &mater;
strata & strateN;
var res;
ods output Statistics=StatRATMAT;run;
/* effet JOUR*/
proc surveymeans data=tableRES total=degreJ mean clm stderr var sum clsum std varsum;
weight &poids;
cluster &jour;
strata &vagueN;
var res;
ods output Statistics=StatRATJOUR;run;
/* effet NR*/
proc sql;
create table NR_calage as
select sum (res*res*(1-probaR)/(probaR*probaR*(1/pondAVANT_calage)*(1/pondAVANT_calage))) as
effetNR from tableRES;run;quit;

data _null_;
set  StatRATMAT;
CALL SYMPUT('var_calage_effetMAT',varsum);run;
data _null_;
set  StatRATJOUR;
CALL SYMPUT('var_calage_effetJOUR',varsum);run;
data _null_;
set  NR_calage;
CALL SYMPUT('NR_calage',effetNR);run;
```

Pour calculer la variance exacte, on sommera `var_calage_effetMAT + var_calage_effetJOUR + NR_calage`

Pour calculer la variance du plan simplifié, on sommera `var_calage_effetMAT + NR_calage`

Pour estimer plus simplement la variance par un sondage aléatoire simple

```
* SAS TOTAL;
proc surveymeans data=table total=764000;
var &variable;
weight & poids;
ods output Statistics=SAS;run;

data _null_; set  SAS;
CALL SYMPUT('V_SAS',varsum);run;


* SAS RATIO;
proc surveymeans data= table total=764000;
ratio &variable1/&variable2;
weight & poids;
ods output Statistics=SAS;run;

data _null_; set  SAS;
CALL SYMPUT('V_SAS',var);run;
```

## REFERENCES


Hélène Juillard *(2016) : Méthodes d'estimation et d'estimation de variance pour une enquête longitudinale - Application aux données de l'Etude Longitudinale Française depuis l'Enfance (Elfe)* – Université de Toulouse – Thèse

Hélène Juillard *(2016) : Estimation de la variance pour l'enquête ELFE* - Collection : Documents de travail INED n° 226

Fabien DELL, Xavier d'HAULTFOEUILLE, Philippe FÉVRIER, Emmanuel MASSÉ (**) (2002) – *Mise en œuvre du calcul de variance par linéarisation* - Insee-Méthodes : Actes des Journées de Méthodologie Statistique.

Jean-Claude Deville (1999): *Variance estimation for complex statistics and estimators : Linearization and residual techniques.Survey Methodology*

Jean-Claude Deville, Carl-Erik Sarndal (1992): *Calibration Estimators in Survey Sampling* - Journal of the American Statistical Association, Vol. 87, No. 418, pp.376-382

Skinner, C. J. (2015) : *Cross-classified sampling: some estimation theory*. Statistics and Probability Letters. ISSN 0167-7152