Pondération de l'enquête ELFE au temps 0 (maternité)*

Hélène Juillard - avril 2015

Deux pondérations sont proposées : une pondération 'nourrisson' et une pondération 'famille' (information relative aux jumeaux non doublée). Les résultats ci-après sont ceux de la pondération 'nourrisson', la méthode employée étant identique pour la famille.

L'enquête Elfe (Etude longitudinale française depuis l'enfance), a pour objectif de recueillir des informations sur la totalité des nouveau-nés d'un échantillon aléatoire de 349 maternités en France métropolitaine, sur 25 jours répartis tout au long de l'année 2011 en 4 temps d'enquête.

L'échantillon final constitué d'environ **18 300 nourrissons**, soit 1/42ème des naissances métropolitaines, est issu d'un plan de sondage à probabilités d'inclusion inégales.

Une pondération, si elle est utilisée dans les analyses, doit permettre d'obtenir des résultats généralisables à l'ensemble de la population cible (et pas seulement à l'échantillon). Elle consiste à assigner à chacun des 18 300 nourrissons, un poids statistique qui correspond au nombre d'enfants qu'il représente dans la population visée (764 000 nourrissons, respectivement un total estimé de 753 500 familles). La population d'inférence est celle **des nourrissons nés en 2011 dans une maternité métropolitaine, issus d'un accouchement au plus gémellaire, hors grands prématurés, ayant une mère majeure, en mesure de donner un consentement éclairé notamment dans l'une des langues proposées (français, anglais, arabe ou turc), nés dans une maternité métropolitaine et dont les parents ne résident pas temporairement en métropole.**
La partie II donne les grandes lignes de cette pondération et la partie III s'adresse aux utilisateurs et aux non-utilisateurs de cette pondération.

I – CONTEXTE

L'échantillon est issu d'un plan de sondage** **à plusieurs phases d'échantillonnage** : une phase concernant les maternités, une autre les jours, et la dernière, celle exhaustive des nourrissons. Les maternités sélectionnées aléatoirement sont issues d'un **plan de sondage stratifié** avec allocations proportionnelles à leurs tailles. Pour représenter chaque saison, quatre périodes de l'année ont été sélectionnées : 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 : 25 jours au total.

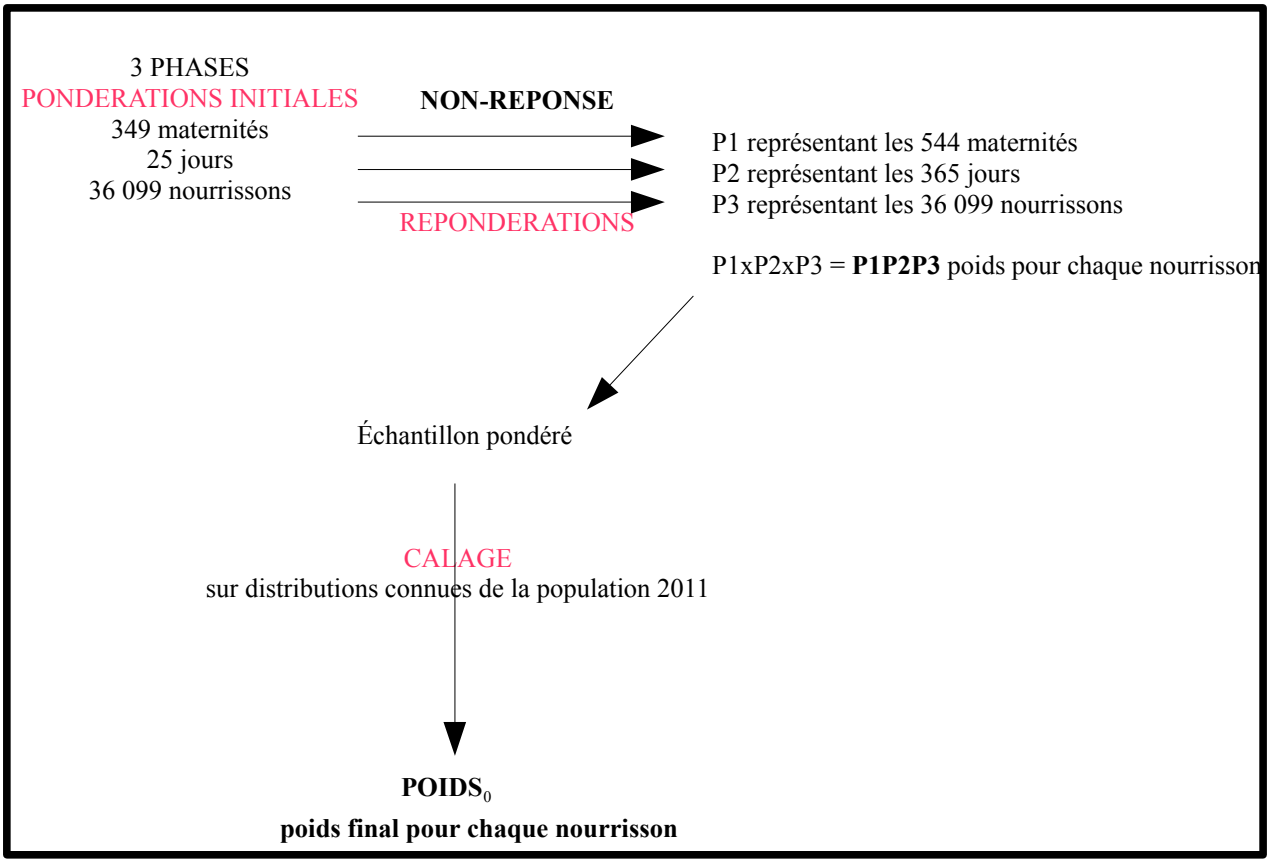
Après avoir pris en compte les poids initiaux dus au plan de sondage stratifié, les poids sont ajustés pour rendre compte de la non-réponse qui s'observe à différents niveaux : lors des enquêtes en maternité, une partie des maternités n'a pas participé, de même qu'une partie des mères ayant accouché les jours d'enquête. On discerne deux types de **non-réponse pour les maternités** : celle des maternités n'ayant pas du tout participé et celles ayant participé partiellement (quelques jours). On a en notre possession les données de variables communes aux maternités participantes et à celles n'ayant pas

* Existence d'une note plus détaillée

** Plan de sondage construit par Nicolas Razafindratsima (INED) et Hélène Sarter (InVS)

participé. On dispose aussi, pour traiter la **non-réponse des mères**, de quelques informations, communes aux mères ayant refusé de participer et aux mères volontaires. En second lieu, un calage a été effectué sur les marges de l'état civil et de l'ENP 2010 (Enquête Nationale Périnatale), permettant à notre échantillon pondéré d'être cohérent géographiquement et sur la situation socio-démographique de la mère.

Pour construire cette pondération, le plan de sondage a été schématisé par 3 phases : maternités, jours, nourrissons.



Les 36 099 nourrissons sont ceux approchés dans les maternités participantes et durant les jours auxquels elles ont participé.

II – LA PONDERATION

1) Phase 1, les maternités

L'échantillon de cette phase est construite selon une stratification avec allocations proportionnelles à la taille. Notons donc, des poids initiaux inégaux $\frac{1}{\pi_i}$ selon la **taille de la maternité** (c'est-à-dire son nombre d'accouchements).

STRATES h	Nombre d'accouchements par maternité en 2008	Taille des strates N_h	Taille de l'échantillon n_h	Probabilité d'inclusion π_h	**poids-initial** $\frac{1}{\pi_h}$
1	[145, 699]	108	28	0.26	3,86
2	[700, 1009]	108	47	0.44	2,3
3	[1010, 1418]	109	66	0.60	1,65
4	[1422, 2187]	108	97	0.90	1,1
5	[2197, 5215]	111	111	1	1
Total		**544**	**349**		**544**

Taille de la population : 544 maternités — Taille de l'échantillon : 349 maternités

Nombre de maternités ayant accepté de participer : 320 — Taux de participation : 91,69%

Nous disposons de 4 variables communes aux répondants et aux non-répondants : la **région**, le **statut juridique**, la **strate** et le **niveau de médicalisation** de la maternité.

Comparaison des maternités répondantes et des maternités non-répondantes sur 4 variables communes	Nombre de maternités	Nombre de maternités qui n'ont participé à aucune vague (NON-réponse totale)	**Taux de NON-réponse**	Test d'indépendance du Khi-deux ou test exact de Fisher (p-value)
Total	**349**	**29**	**8,31%**	
Région				Effectifs par classe trop petits
Île-de-France	77	15	19,5%	
Champagne-Ardenne	7	0	0%	
Picardie	11	0	0%	
Haute Normandie	8	0	0%	
Centre	13	2	15,4%	
Basse Normandie	8	0	0%	
Bourgogne	10	0	0%	
Nord pas de Calais	24	1	4,2%	
Lorraine	11	0	0%	
Alsace	12	0	0%	
Franche-Comté	6	0	0%	
Pays de la Loire	15	0	0%	
Bretagne	21	2	9,5%	
Poitou Charentes	11	1	9,1%	
Aquitaine	15	0	0%	
Midi-Pyrénées	13	1	7,7%	
Limousin	2	0	0%	
Rhône-Alpes	37	6	16,2%	
Auvergne	3	0	0%	
Languedoc-Roussillon	16	0	0%	
PACA	28	1	3,6%	
Corse	1	0	0%	

Taille (nombre d'accouchements en 2008)				0.8989
[145, 699]	28	3	10,7%	
[700, 1009]	47	3	6,4%	
[1010, 1418]	66	4	6,1%	
[1422, 2187]	97	9	9,3%	
[2197, 5215]	111	10	9,0%	
Groupe de régions				2.703e-13
Ile-de-France, Centre, Picardie	101	17	16,8%	
Sud-Est	69	7	10,1%	
Autres	179	5	2,8%	
Autorisation				0.1889
niveau 1	125	11	8,8%	
niveau 2	161	16	9,9%	
niveau 3	63	2	3,2%	
Statut juridique				0.0969
privé non lucratif	30	5	16,7%	
privé lucratif	95	9	9,5%	
public	224	15	6,7%	

La méthode des scores* pondérés** (utilisant les 4 variables ci-dessus) a été utilisée pour compenser la non-réponse des 29 maternités non-répondantes (poids nul) en rehaussant le poids des 320 maternités répondantes par un facteur d'ajustement $\frac{1}{\hat{p}_i}$.

$$P1_i = \frac{1}{\pi_i}\frac{1}{\hat{p}_i}$$

2) Phase 2, les jours

Au total, sur les 320*25=8000 jours d'enquête attendus chez les maternités participantes, 7741 ont été effectués, soit 96,76 %. Chaque nourrissons a été pondéré en fonction de sa vague v de naissance (saison). La première vague contenant 4 jours d'inclusion, la 2nde 6 jours, la 3ème 7 jours et la dernière 8 jours :

$$P2_v = \frac{91}{\text{NB de jours vague}_v}\frac{\text{NB de maternités participant à au moins une vague}}{\text{NB de maternités participant à la vague}_v}$$

Ceci a été effectué au sein de chaque strate de maternités.

3) Phase 3, les nourrissons

Au sein de chaque maternité participante et pendant les jours d'enquête auxquels elle participe, les 36 099 nourrissons éligibles sont invités à faire partie de la cohorte. Pour plus de 18300 enfants, les parents ont consenti au suivi de leur(s) enfant(s) par l'étude ELFE.

Nombre d'individus abordés : 36 099 nourrissons

Taux de participation : 50,8%

* Méthode utilisant la régression logistique pour créer des groupes de réponses homogènes desquels seront issues les probabilités moyennes de réponse $\hat{p}_i$.

** Pondéré par le poids initial de la strate.

Parmi les 18 329 participants à l'enquête, 71 sont nés hors des 25 jours d'enquête : ils sont donc hors champ d'enquête et ôtés de l'échantillon auquel correspondra la pondération. De plus, les 51 individus ont demandé la destruction de leurs données (entre l'enquête et le moment où la pondération est construite).

Note : il se peut donc que la base de données qui vous est livrée contienne moins d'individus que ceux dénombrés dans ce document (le nombre de suppressions n'est pas un nombre fixe mais croissant dans le temps, le phénomène est supposé rare et l'impact sur la pondération, minime).

Taille de l'échantillon des répondants et non-répondants : 36 028 nourrissons

Taille du sous-échantillon des répondants : 18 207 nourrissons

Nous disposons de 11 variables communes aux répondants et aux non-répondants : l'**âge de la mère**, son **département**, sa **PCS**, l'indicatrice de son **activité** au moment de la grossesse, l'indicatrice de sa **primiparité**, l'indicatrice **gémellaire**, l'**âge gestationnel** du nourrisson et les **4 variables** caractérisant la maternité d'accouchement.

Note : les informations de l'enquête 2 mois ont été utilisées pour diminuer les taux de valeurs manquantes des variables Age, Activité et CSP de l'enquête en maternité.

Comparaison des caractéristiques des nourrissons répondants et non répondants sur les 11 variables communes (sur les données non manquantes)*	Nombre de nourrissons dans l'échantillon	Part des nourrissons dans l'échantillon	Nombre de nourrissons NON-répondants	**Taux de NON-réponse des nourrissons**	Test du chi-deux (p-value)
Total	36028	100,00%	17800	49,40%	
Vague					0.0003
-1	5300	14,7%	2510	47,3%	
-2	8913	24,8%	4321	48,3%	
-3	10344	28,7%	5197	50,1%	
-4	11471	31,8%	5772	50,3%	
MERES					
Naissance					0.0020
- unique	34659	96,5%	17068	49,2%	
- multiple	1240	3,5%	666	53,6%	
Activité au moment de la grossesse					<.0001
- oui	24946	72,1%	9802	39,3%	
- non	9638	27,9%	6978	72,4%	
Est primipare					<.0001
- oui	15633	44,2%	7416	47,3%	
- non	19731	55,8%	9940	50,3%	
Age					<.0001
- moins de 22 ans	2597	7,3%	1564	60,2%	
- [23 ; 24]	2566	7,2%	1432	55,7%	
- [25 ; 29]	11361	31,9%	5738	50,4%	
- [30 ; 34]	11699	32,9%	5308	45,3%	
- [35 ; 39]	5798	16,3%	2645	45,5%	
- plus de 40 ans	1537	4,3%	776	50,3%	

* Tableau construit pour en février 2014

Distributions (sur les données non manquantes)	Nombre de nourrissons dans l'échantillon	Part des nourrissons dans l'échantillon	Nombre de nourrissons NON-répondants	**Taux de NON-réponse des nourrissons**	Test du chi-deux (p-value)
Age gestationnel					0.0557
- [33 ; 37]	4370	12,2%	2249	51,4%	
- [38 ; 40]	24913	69,8%	12370	49,5%	
- plus de 40 semaines	6383	17,9%	3152	49,2%	
5 groupes de régions (mères)					<.0001
- Ile-de-France, Centre, Picardie	10202	28,3%	5287	51,7%	
- Nord-Est	7774	21,6%	3428	43,9%	
- Nord-Ouest	5986	16,6%	3039	50,7%	
- Sud-Est	6721	18,7%	3346	49,7%	
- Sud-Ouest	5316	14,8%	2671	50,2%	
3 groupes de régions (mères)					<.0001
- Ile-de-France, Centre, Picardie	10202	28,4%	5292	51,7%	
- Nord-Est	7774	21,6%	3428	43,9%	
- et le reste	18023	50,0%	9056	50,2%	
PCS brute					<.0001
- 1 Agriculteurs, exploitants	98	0,3%	42	42,9%	
- 2 Artisans, commerçants et chefs d'entreprise	971	2,9%	406	41,8%	
- 3 Cadres et professions intellectuelles supérieures	4105	12,2%	1198	28,2%	
- 4 Professions Intermédiaires	6132	18,4%	2462	40,1%	
- 5 Employés	13534	41,3%	6063	44,8%	
- 6 Ouvriers	823	3,1%	505	61,4%	
- 7 Sans profession	6711	21,8%	5860	87,3%	
- 9 Ne peut classer la profession	3407	9,5%	1235	36,2%	
MATERNITES					
Strate					0.0005
- 1	863	2,4%	402	46,5%	
- 2	2451	6,8%	1136	46,2%	
- 3	4750	13,2%	2422	50,8%	
- 4	9850	27,4%	4939	50,0%	
- 5	18085	50,2%	8872	49,0%	
Statut juridique					<.0001
- privé non lucratif	3166	8,8%	1403	44,2%	
- privé lucratif	8929	24,8%	4458	49,9%	
- public	23904	66,4%	11910	49,7%	
Autorisation					0.5734
- niveau 1	8191	22,8%	4015	48,8%	
- niveau 2	17159	47,7%	8520	49,5%	
- niveau 3	10649	29,5%	5236	49,1%	

La méthode des scores* non pondérés (utilisant les 11 variables ci-dessus) a été utilisée pour compenser la non-réponse des 17 800 nourrissons (poids nul) en rehaussant le poids des 18 207 nourrissons répondants par un facteur d'ajustement $\frac{1}{\hat{q}_j}$.

Il existe un défaut de **sous-couverture** : certaines mères éligibles n'ont pas été approchées (en moyenne, 4%). Or, le nombre de naissances éligibles par maternité est connu. Afin de corriger ce défaut, un coefficient a été calculé par région (nombre de nourrissons éligibles / nombres de nourrissons enquêtés) et affecté à chaque nourrisson.

$$P3_j = \frac{1}{\hat{q}_j} \, coeff_{sous-couverture}$$

* Méthode utilisant la régression logistique pour créer des groupes de réponses homogènes desquels seront issues les probabilités moyennes de réponse $\hat{q}_j$.

4) Calage

Chaque nourrisson j se voit donc affecté du poids corrigé de la maternité i dans laquelle il est né, du poids temps associé à celle-ci et de son poids corrigé en fonction des caractéristiques des mères des nourrissons non participants.

$$P1P2P3_j = P1_i \; P2_v \; P3_j$$

Afin d'être cohérent avec quelques informations choisies et disponibles sur toute la population, on effectue un calage sur des distributions provenant de l'état civil et de l'ENP. Ce calage va modifier les poids $P1P2P3_j$.

Le calage sur l'***Age*** (état civil filtré sur métropole et mères majeures) va permettre d'augmenter les poids des mères très jeunes ou âgées, qui malgré la repondération ajustée de la non-réponse restaient encore sous-représentées. De la même façon, le calage sur les groupes de ***régions*** (état civil filtré sur métropole) assure une bonne représentation géographique. La ***Primiparité*** (ENP) et le ***Statut matrimonial*** (état civil filtré sur métropole) sont des variables permettant de caractériser la situation familiale, dimension importante dans cette enquête.
On choisit de caler sur le ***Niveau d'étude*** (ENP) qui est une caractéristique peu susceptible de changer après accouchement. La population des immigrées étant plus vaste que celle des étrangères (le fait d'acquérir la nationalité française concerne une sous-population), il a été décidé de caler sur le ***Statut immigré*** (état civil filtré sur métropole).

$$\text{Calage}(P1P2P3_j) = POIDS_{0j}$$

Variables de calage	Distribution avant pondération P1P2P3	Distribution après pondération P1P2P3 et avant calage	Source
âge de la mère			Etat civil (filtre sur métropole et mères majeures)
- [18, 22]	- 05,71%	- 07,86%	- 06,86%
- [23, 24]	- 06,27%	- 07,45%	- 07,10%
- [25, 29]	- 31,07%	- 32,15%	- 31,22%
- [30, 34]	- 35,32%	- 32,50%	- 33,25%
- [35, 39]	- 17,42%	- 15,88%	- 16,90%
- + de 40 ans	- 04,21%	- 04,17%	- 04,67%
Groupe de régions de domicile			Etat civil (filtre sur métropole)
- Ile de France/Centre/Picardie	- 26,96%	- 30,4%	- 29,96%
- Nord-Est	- 23,84%	- 19,6%	- 19,15%
- Nord-Ouest	- 16,17%	- 14,3%	- 15,42%
- Sud-Est	- 18,52%	- 20,0%	- 19,03%
- Sud-Ouest	- 14,51%	- 15,7%	- 15,54%
statut mère immigrée			Etat civil (filtre sur métropole)
- née en France	- 86,66%	- 82,2%	- 81,25%
- immigrée	- 13,34%	- 17,8%	- 18,75%
état matrimonial			Etat civil (filtre sur métropole)
- né dans le mariage	- 46,17%	- 45,8%	- 45%
- né hors mariage	- 53,83%	- 54,2%	- 55%
mère primipare			ENP (champ ELFE)
- oui	- 45,65%	- 44,8%	- 43,1%
- non	- 54,35%	- 55,2%	- 56,9%
niveau d'étude de la mère			ENP (champ ELFE)
- non scolarisée/école primaire/collège/CAP ou BEP	- 18,91%	- 23,7%	- 27,79%
- 2nde/1ère/terminale générale, technique ou professionnelle	- 20,93%	- 23,9%	- 19,88%
- études supérieures	- 60,16%	- 52,3%	- 52,33%

Le calage assure donc des distributions de l'échantillon pondéré ELFE identiques à celles de la colonne 'Source'.
La méthode utilisée est celle du raking ratio.

5) Descriptif du poids

Afin de limiter la dispersion, la variance des poids (puisqu'elle impactera la variance de nos estimations), certains poids ont été tronqués à 200.

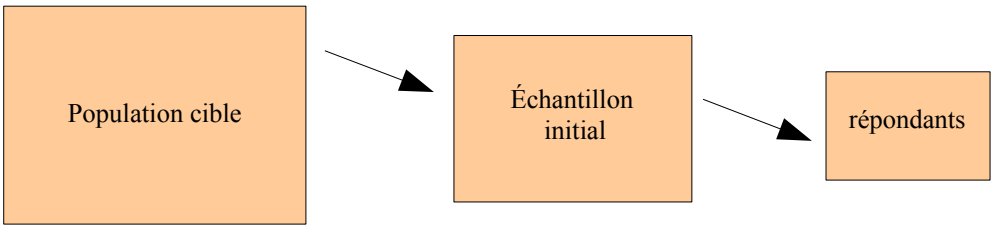
Min	P5	P10	P90	P95	Max	Max/min	Somme des poids	Moyenne	écart-type	Coefficient de variation (%)
11	18	20	75	114	201	15	764000	42	32	77,6

III – UTILISATION DES POIDS ?

Cette partie présente quelques éléments qui nécessitent réflexion avant de choisir d'utiliser une pondération **ou** de choisir de ne pas l'utiliser.

La pluridisciplinarité de l'enquête ELFE amène à travailler sur des variables de type santé, environnement, sociologie, démographie... Des variables qui ne seront donc pas toujours particulièrement corrélées entre elles. La pondération étant unique (il n'existe pas un jeu de poids par variable ou par domaine), elle sera meilleure en terme de biais et/ou de variance pour certaines variables et moins adaptée pour d'autres.

Pourquoi construire une pondération ?



On s'intéresse aux pondérations lorsque l'inférence cherchée est à la population cible et non à l'échantillon.

Deux phases principales sont à l'origine de cette pondération : une première phase de sélection de l'échantillon initial (36 028) et une deuxième phase de réponse des familles (18 207).

La phase 1 correspond surtout au plan de sondage à probabilités inégales : les **maternités** de très petite taille (strate 1) ont été incluses à 25 % tandis que les très grandes maternités (strate 5) ont été incluses à 100 %. Pour prendre en

compte ces différences, la pondération affecte un poids égal à 4 aux petites maternités et un poids égal à 1 aux grandes. Ainsi, au travers de leurs poids, les 25 % de petites maternités sont censées représenter, outre elles-mêmes, les 75 % petites maternités non-enquêtées. De façon analogue, les enfants nés en **avril** (4 jours d'inclusion) ont été inclus deux fois moins que ceux nés en **décembre** (8 jours d'inclusion). Ainsi, leurs poids sont en moyenne deux fois plus grands que ceux correspond aux enfants nés en décembre.

Il est alors conseillé de pondérer si la variable d'intérêt (celle sur laquelle on souhaite travailler) **est liée à la strate de la maternité** (c'est-à-dire si cette variable prend des valeurs différentes suivant qu'il s'agit d'une grande ou d'une petite maternité), ne pas pondérer pourrait alors engendrer un biais. De façon analogue, **si la variable d'intérêt est lié à la saison, il est conseillé de pondérer.**

La phase 2 correspond à la 'sélection' des familles en terme de participation : cette participation s'est révélée être liée à certaines caractéristiques. Par exemple, les mères ayant déjà eu un enfant ont moins souhaité participer à l'enquête que les mères primipares. Alors, pour compenser ce manque d'information, les poids (statistiques) des nourrissons des mères ayant déjà eu un enfant ont été augmentés. La pondération a utilisé les 11 variables présentées précédemment au degré 3. De plus, en terminant par un calage, on assure des distributions exactes sur les 6 variables utilisées.

Si la variable d'intérêt est corrélée à l'une de ces variables, pour une réduction des biais il est conseillé d'utiliser la pondération.

Certaines variables, notamment celles liées à la santé, ne sont pas toujours sensibles à la pondération. Ceci peut s'expliquer par une sélection en phase 2 caractérisée par des variables plus sociales que médicales. Il se peut malgré tout qu'il y ait d'autres effets de sélection non mesurables (ainsi la sélection des variables de pondération et de calage reste tributaire des données disponibles).

Exemples :

Variables d'intérêt	Échantillon sans pondération	Échantillon avec pondération	Source de comparaison
			État civil 2011
Nationalité de la mère			
- française (par naissance ou par acquisition)	- 91,72 %	- 87,01 %	- 86,6%
- étrangère ou apatride	- 08,28 %	- 12,99 %	- 13,4%
			ENP (échantillon)
Accouchement : début du travail			
- travail spontané	- 70,82 %	- 70,01 %	- 66,63 %
- déclenchement	- 19,51 %	- 19,78 %	- 22,44 %
- césarienne avant début du travail	- 09,67 %	- 10,20 %	- 10,93 %

Pourquoi prendre en compte les mécanismes d'échantillonnage et de non-réponse ?

L'utilisation de ces poids dans les logiciels, lorsqu'elle est nécessaire, assurera le caractère (presque) sans **biais** des

variables d'intérêt. Pour la **variance**, c'est plus compliqué : utiliser les procédures standards (sans préciser qu'il s'agit d'un plan à plusieurs phases, stratifié, avec de la non-réponse et un calage), entraînera des estimations d'écarts-type vraisemblablement sous-estimées. Les tests statistiques rejetteront alors l'hypothèse nulle plus fréquemment qu'ils ne le devraient. Il faudrait donc utiliser des procédures appropriées pour estimer correctement les variances. Ce travail fait l'objet d'une autre note mise à disposition des utilisateurs en 2015.