Enquête « Elfe » :

Pondérations des enquêtes nationales

Thierry Siméon – Première version. Décembre 2019

Sommaire

PONDERATIONS DES ENQUETES NATIONALES.

La présente note a pour but de décrire les méthodes de pondération utilisées pour les données de l'enquête « Elfe » : une première méthode mise en œuvre pour calculer les pondérations jusqu'aux enquêtes réalisées aux 2 ans de l'enfant et une nouvelle méthode utilisée à partir des enquêtes réalisées aux 3 ans ½ de l'enfant.

Cette note reprend tout d'abord l'ensemble de éléments nécessaires à la compréhension de ce changement de méthodologie en rappelant les grandes lignes de la pondération aux différents temps d'enquêtes précédents (maternité, 2 mois, 1 an, et 2 ans) permettant d'inférer à la population les estimations produites à partir des données fournies par les seuls répondants.

Pour plus de détail sur les résultats de la méthode mise en place aux temps précédents, on se reportera aux différentes notes:

- Pondération de l'enquête « Elfe » en maternité – Juillard, Thierry, Razafindratsima, Bringe, Lanoë.
- Pondération de l'enquête « Elfe » au temps 1 (2 mois de l'enfant) – Juillard.
- Enquête nationale 1 an. Pondérations au temps 2 (1 an de l'enfant) – Pilorin.
- Enquête nationale 2 ans. Pondérations au temps 3 (2 ans de l'enfant) – Pilorin.

La présente note propose ensuite les grands principes de pondération mis en place pour la suite des enquêtes, à savoir aux enquêtes réalisées aux 3 ans ½ de l'enfant et suivantes.

Pour plus de détail sur la méthode mise en place aux 3 ans ½ de l'enfant, on se reportera à:

- Enquête nationale. Pondérations aux 3 ans ½ de l'enfant- Siméon

Toutes ces notes sont disponibles sur les pages correspondantes aux différents temps d'enquête de la plateforme Pandora.

Remarque : La nouvelle méthode permettant de s'affranchir de connaissances spécifiques aux non répondants en estimant les poids par calage directs sur les seuls répondants, il est possible de mettre en place à la demande de nouveaux jeux de pondération sur un sous échantillon représentatif de la population.

Si vous analysez des résultats issus des enquêtes Elfe, n'hésitez pas à contacter Thierry Siméon (thierry.simeon@ined.fr) qui pourra vous guider et générer pour vous les pondérations vous permettant d'inférer les analyses calculées sur votre sous population spécifique à la population cible.

Méthode de pondération utilisée dans les enquêtes jusqu'aux 2 ans de l'enfant

Pondération enquête maternité :

La méthode de pondération en maternité est décrite précisément dans le document « Pondération de l'enquête ELFE en maternité ». Rappelons-en simplement ici les grands principes :

Les nourrissons inclus dans la cohorte sont sélectionnés ainsi : leur date de naissance fait partie d'une sélection de jours de l'année 2011, et leur lieu de naissance appartient à un échantillon de maternités en France métropolitaine.

Dans le cadre de l'enquête ELFE, le tirage se compose :

- D'une base de sondage des maternités, constituée de la liste des maternités (publiques et privées) de France métropolitaine en 2008 : 544 maternités ont été répertoriées. La variable de stratification est la taille de la maternité. 5 strates ont été construites à effectifs égaux avec allocation proportionnelle au nombre de naissances. 349 maternités sont présentes dans l'échantillon.

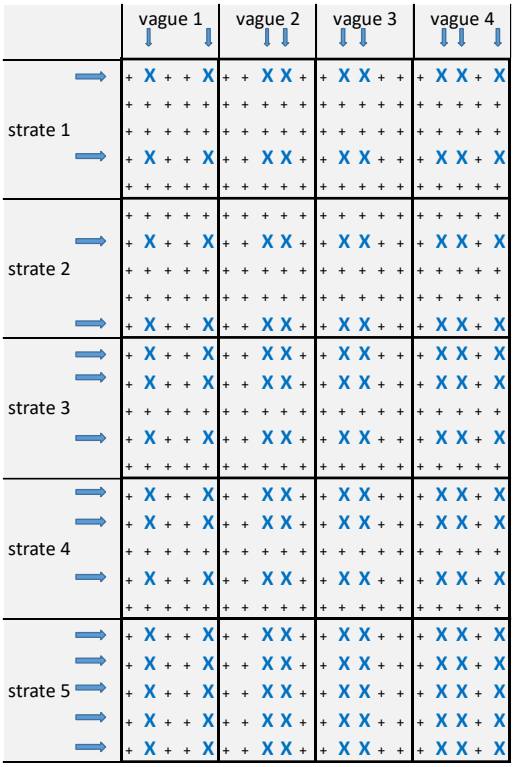
Strates g	Nb d'accouchements par maternité en 2008	Taille dans la population N_g	Taille de l'échantillon n_g
1	[145-699[	108	28
2	[700-1009[	108	47
3	[1010-1418[	109	66
4	[1422-2187[	108	97
5	[2197-5215[	111	111
TOTAL		544	349

- D'une base de jours (l'ensemble des jours de l'année 2011). Pour représenter chaque saison, les 25 jours d'enquête (4, 6, 7 et 8 jours) ont été répartis en vague. Ces jours ne sont pas issus d'un tirage aléatoire pour raisons logistiques et ont été fixés : 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.

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

La sélection des jours est le même pour chaque maternité tirée (ou vice versa, l'échantillon des maternités étant le même pour chaque jour choisi). L'échantillon final se forme au croisement de lieux sélectionnés et de temps choisis. On peut donc schématiser le tirage mis en œuvre dans le cadre de l'enquête ELFE ainsi :

PONDERATIONS DES ENQUETES NATIONALES.



Par l'indépendance des tirages des lignes et des colonnes, chaque unité « maternité x jour » de l'échantillon de nourrissons se voit simplement attribuer un poids initial qui correspond à l'effet plan de sondage $\frac{N_g M_h}{n_g m_h}$, soit :

		Vague *h*			
		1	2	3	4
Strate *g*	1	86,79	58,50	50,69	44,36
	2	51,70	34,85	30,20	26,43
	3	37,16	25,05	21,71	18,99
	4	25,05	16,89	14,63	12,80
	5	22,50	15,17	13,14	11,50

Poids initial issu du plan de sondage

Au niveau maternité et jour, 2 causes majeures de non réponses ont ensuite été analysées.

Certaines maternités tirées n'ont en réalité pas participé à l'enquête pour différentes raisons. Ainsi, parmi les 349 maternités de l'échantillon, 25 n'ont participé à aucune vague. De plus, 4 maternités tirées au sort n'ont finalement pas été invitées. Ce qui fait donc un total de 29 maternités non-participantes à prendre en compte dans cette phase de non-réponse.

D'autres maternités n'ont participé qu'à certaines vagues. Ainsi, sur les 320x25 = 8000 « maternité x jour » attendus, seulement 7741 ont été réalisés.

Les maternités ayant refusées de prendre part à l'enquête se répartissent ainsi :

PONDERATIONS DES ENQUETES NATIONALES.

Strates *g*	Nb d'accouchements par maternité en 2008	Taille dans la population N_g	Taille de l'échantillon n_g	Nombre de maternités participantes
1	[145-699[	108	28	25
2	[700-1009[	108	47	44
3	[1010-1418[	109	66	62
4	[1422-2187[	108	97	88
5	[2197-5215[	111	111	101
TOTAL		544	349	320

Nous disposons, outre la strate de la maternité, de 3 variables pour caractériser les maternités participantes et les non-participantes : sa région, son niveau de médicalisation et son statut juridique.

	Participation			Probabilité de participation
	NON	OUI	Total	
Groupe_region4				
Ilde de France, Centre, Picardie	17	84	101	83,17%
Sud Est	7	62	69	89,86%
autre	5	174	179	97,21%

	Participation			Probabilité de participation
	0	1	Total	
Statut_juridiq(Statut_juridiq)				
privé non lucratif	5	25	30	83,33%
privé lucratif	9	86	95	90,53%
public	15	209	224	93,30%

	Participation			Probabilité de participation
	0	1	Total	
Autorisation(Autorisation)				
niveau 1	11	114	125	91,20%
niveau 2	16	145	161	90,06%
niveau 3	2	61	63	96,83%

La non réponse est importante en Ile de France (15 maternités sur les 29 non répondantes sont en Ile de France) et pour les maternités à statut privé non lucratif (5 non répondants sur les 30 maternités ayant ce statut).

Même si sur l'ensemble des maternités françaises, l'hypothèse d'indépendance entre le mécanisme de réponse et les variables « groupe de région » et « statut juridique » peut être rejetée, on a souhaité conserver l'ensemble des informations disponibles pour caractériser la non réponse au niveau maternité (on peut en effet se demander si, par exemple, le niveau de médicalisation n'est pas important pour caractériser des enfants et leur futur développement). Pour ce faire, la méthode des scores avec taux de réponse pondérés par les poids initiaux a été retenue : 10 groupes de réponse à effectif égaux ont été constitués avec les variables strate, région, niveau de médicalisation et statut juridique. Les 10 groupes retenus sont modélisés avec des probabilités de participation des maternités allant de 71 à 100%. Les groupes extrêmes ont été regroupés (les 2 avec le plus faible score d'un côté, les 2 avec le plus fort score de l'autre). 8 groupes de réponses homogènes sont donc considérés.

De plus, même parmi les maternités ayant acceptées de répondre, certaines d'entre elles n'ont pu être enquêtées pendant les 4 vagues pour raisons logistiques. Afin de remédier à cela, on a simplement choisi de redresser, par vague et par strate, en fonction du nombre de maternités réellement enquêtées par rapport aux nombres de maternités participantes.

Malgré tout ce qui a été écrit jusqu'ici, il faut rappeler que ce ne sont en réalité pas des unités « maternité x jour » qui sont enquêtées dans l'enquête ELFE. L'unité de sondage est le nourrisson. Il était simplement prévu de sélectionner tous les nourrissons appartenant à une maternité tirée et nés pendant les jours où l'enquête a été réalisée (tirage en grappes de nourrissons).

PONDERATIONS DES ENQUETES NATIONALES.

Les mères souhaitant participer à l'enquête ont répondu à un questionnaire en face-à-face. De plus, plusieurs informations ont pu être récoltées pour les mères non-répondantes au travers d'un « dossier refus ». Ces variables sont communes aux mères répondantes et non-répondantes : c'est ce qui permettra d'effectuer une repondération en fonction de la non-réponse prenant en compte les caractéristiques des mères.

La non réponse des nourrissons est en fait dû à 2 effets.

Le 1er effet est le défaut de couverture: certaines mères éligibles n'ont pas été approchées. Dans les faits, il était parfois impossible pour les enquêteurs d'aborder toutes les mères lorsqu'il y avait plusieurs naissances en même temps ou lorsque la mère quittait trop tôt la maternité. On parle alors de sous-couverture, des individus de la population cible étant absents de la base de sondage. Or, le nombre de naissances éligibles par maternité est connu, il a été récolté en salle d'accouchement (il est approximatif et très certainement surestimé). Afin de corriger ce défaut, un coefficient a été calculé par région (nombre de nourrissons éligibles / nombres de nourrissons enquêtés). On affecte donc ce coefficient, légèrement supérieur à 1, à chaque nourrisson afin de rectifier l'erreur de sous-couverture.

Le second effet est bien évidemment plus important : la non réponse du fait du refus simple de la mère de participer à l'enquête. Après correction de la non réponse partielle, un modèle logistique a été réalisé pour constituer de groupes de réponses homogènes à partir de la méthode des scores non pondérés. Les variables prises en compte sont :

- Strate, statut juridique et niveau de médicalisation de la maternité ;
- Age de la mère : [18,22], [23 ; 24], [25 ; 29], [30 ; 34], [35 ; 39], plus de 40 ans ;
- Age gestationnel (en semaines): [33 ; 37], [38 ; 40], plus de 40 semaines ;
- Département d'habitation de la mère regroupe par région, puis par groupe de régions : Ile-de-France, Centre, Picardie, Nord-Est, Nord-Ouest, Sud-Est, Sud-Ouest ;
- PCS (professions et catégories socioprofessionnelles) inspirée de la nomenclature en 8 postes : Agriculteurs exploitants, Artisans, commerçants et chefs d'entreprise, Cadres et professions intellectuelles supérieures, Professions Intermédiaires, Employés, Ouvriers, Sans profession, Ne peut classer la profession ;
- Activité au moment de la grossesse : oui/non ;
- Indicatrice gémellaire : a eu des jumeaux ou naissance unique ;
- Primiparité (fait d'être pour la première fois parent) : oui ou non ;

L'hypothèse d'indépendance entre le mécanisme de réponse et de nombreuses variables peut être rejetée.

Analyse des effets Type 3			
Effet	DDL	Khi-2 de Wald	Pr > Khi-2
grp_3regb	2	173.1769	<.0001
Age	6	65.9868	<.0001
Act	2	22.1372	<.0001
CSP_corr	6	2386.1134	<.0001
Id_gem	2	14.0053	0.0009
age_gesta	3	0.3148	0.9572
Ind_enf	2	7.1868	0.0275
Strate	4	12.1682	0.0161
Statut_juridiq	2	13.1712	0.0014
Autorisation	2	6.9373	0.0312

50 groupes de réponses homogènes ont été constitués avec toutes les variables retenues. Les groupes extrêmes sont regroupés (les 10 avec les scores les plus faibles et les 5 avec les scores les plus importants). 35 groupes sont finalement conservés avec des probabilités de participation modélisées entre 13 et 75%.

Enfin, un processus de calage a été réalisé. L'Enquête Nationale Périnatale (ENP) a lieu régulièrement (1995, 1998, 2003, 2010) en France. Elle vise à connaitre l'état de santé et les soins périnatals des enfants, des mères, leurs caractéristiques, les facteurs à risques et par sa répétition, permet de suivre les évolutions entre enquêtes. L'ENP 2010 a eu lieu du 15 au 21 mars 2010 dans toutes les maternités de métropole ainsi que trois départements d'outre-mer Réunion). Il nous a été possible de travailler sur le sous-échantillon respectant les critères d'éligibilité ELFE : 14 492 nourrissons (filtre sur l'âge gestationnel, l'âge de la mère, l'indicatrice gémellaire et les naissances métropolitaines).

Pour les variables de calage, le choix s'est porte sur les variables : Age, Région, état matrimonial, statut immigré, niveau d'étude et Primiparité. L'âge et la région sont regroupés en 5 et 6 modalités, le niveau d'étude en 3, et les autres variables sont binaires.

Pour éviter une trop grande dispersion des poids, ces derniers sont tronqués à 200 (1% des poids concernés). L'ensemble est enfin légèrement redressé pour conserver le total de la population.

En résumé, la « pondération maternité » est donc la suivante :

- Effet du plan de sondage

- Ce poids est corrigé pour prendre en compte la non réponse « maternité x jour » (non-participation complète d'une maternité ou non réponse pour certaines unités « maternité x jour »).
  Variables prises en compte : la taille de la maternité (sa strate d'appartenance), sa vague, son statut juridique, son niveau de médicalisation, son groupe de région.

- Ce poids est à nouveau corrigé pour prendre en compte la non réponse « Nourrisson ».
  Variables prises en compte : Strate, statut juridique et niveau de médicalisation de la maternité, Age de la mère, Age gestationnel, Département d'habitation de la mère par groupe de régions, PCS, Activité au moment de la grossesse, Indicatrice gémellaire, Primiparité.

- Enfin, ce poids nourrisson est calé sur les données connues sur la population, puis tronqué à 200. L'ensemble est redressé pour conserver le total de nourrissons nés en 2011 cible de l'enquête ELFE
  Variables prises en compte: Age, Région, état matrimonial, statut immigré, niveau d'étude et Primiparité

Pondérations enquêtes suivantes : 2 mois, 1 an, 2 ans :

La méthode actuelle de pondération au temps suivants est décrite précisément dans les documents « Pondérations de l'enquête ELFE au temps 1 (2 mois de l'enfant) », « Enquête nationale 1 an. Pondérations au temps 2 (1 an de l'enfant) » et « Enquête nationale 2 ans. Pondérations au temps 3 (2 ans de l'enfant) ».

Le principe retenu est identique à chacune de ces pondérations. On ajuste la pondération maternité décrite précédemment pour traiter la non réponse au temps de l'enquête (on repart de la pondération maternité pour éviter une trop grande dispersion des poids et leur cohérence longitudinale, et non de la dernière pondération connue).

Ainsi, on modélise à chaque nouveau temps d'enquête la participation ou non à cette vague par procédure logistique. Signalons dès à présent que les variables utilisées pour modéliser cette non réponse sont les variables

connues grâce à l'enquête maternité. La méthode retenue est celle des groupes de réponses homogènes créés à partir des probabilités estimées par la régression logistique. A partir des scores triés résultants de cette régression, un certain nombre de groupes (un quinzaine) est créé. Au sein de chacun de ces groupes, on estime une probabilité de réponse par la simple proportion de répondants au sein de chaque groupe. Les variables utilisées sont les suivantes :

Enquête 2 mois :

Age mère, groupe de régions, PCS mère, Situation de couple, statut immigré du couple, séances de préparation à la naissance, avoir pris des vacances pendant la grossesse

Enquête 1 an :

Variables utilisées à l'enquête 2 mois + niveau d'étude de la mère, PCS père, consommation de tabac avant la grossesse, IMC de la mère.

Enquête 2 ans :

Variables utilisées à l'enquête 2 mois + niveau d'étude de la mère, PCS père, consommation de tabac avant la grossesse, IMC de la mère.

De plus, lorsque la pondération concerne le parent non référents, 2 variables ont été ajouté aux modèles permettant d'estimer la non réponse : père assiste à l'accouchement, activité du père au moment l'accouchement

Les poids de chaque nourrisson sont alors redressés d'un coefficient d'ajustement égal à l'inverse de cette probabilité de réponse.

Ces données sont ensuite, comme au niveau maternité, calées sur les mêmes marges que celle décrite précédemment. Les poids extrêmes (plus de 250) sont ensuite tronqués et le tout légèrement redressé pour obtenir le total de nourrissons souhaité.

Pourquoi cette méthode pose problème si on la poursuit pour les enquêtes suivantes

Le processus de base consiste comme on l'a vu à enchainer « repondération » pour traiter la non réponse (NR) puis calage. Un individu passe donc d'un poids initial « d» à un poids « d. F(X) / R », avec R sa probabilité de réponse (estimée) et F(X) une fonction dépendant des variables de calage X.

Le processus est « simple » à comprendre, mais nécessite de connaitre sur les non répondants toute l'information nécessaire à modéliser la NR, alors que l'information nécessaire ne correspond plus forcément aux données recueillies sur les non répondants en maternité (activité, couple, fratrie, …). De plus, si on a trop peu de répondants dans certaines « classes » des variables de calage (ou plus exactement sur certains croisements de modalités des variables de calage), les poids vont être disproportionnés (risque de non convergence).

Ainsi, lors d'enquête longitudinale, conserver la méthode actuelle et l'appliquer à des temps futurs pose 2 questions :

- La réalisation de nouvelles pondérations, en calculant à nouveau des groupes de réponses homogènes entre l'enquête maternité et les différents temps de l'enquête puis en multipliant « simplement » les poids maternité par l'inverse de ces probabilités de réponse risque d'engendrer très vite des poids « exponentiels », dont la troncature déformera à terme fortement le calage retenu. On risque donc à la fois de faire porter des poids trop importants sur un petit groupe d'individus ou, en tronquant ces poids, de rendre le redressement, et donc la prise en compte de la non réponse, potentiellement inefficace.

- Modéliser des probabilités de réponse pour des nourrissons à des temps futurs (de plus en plus éloignés de la naissance), obligent à connaitre, pour ces non répondants, les variables permettant de modéliser ce phénomène de non réponse. Hors, par définition, on ne connait pas ces variables, souvent socio démographiques, pour les non répondants. On doit donc se baser sur des variables datant de la dernière enquête avec réponse, ce qui peut très vite s'avérer « bancal » (garder des variables comme vie en couple, activités, …. qui datent de plusieurs années pour modéliser la non réponse aux temps futurs est assez risqué).

On peut alors se demander si le calage sur plusieurs variables correctement choisies, effectué directement depuis l'échantillon de répondants ne va pas également traiter la NR. Cela revient à passer du processus de base à un processus dit de calage simultané. C'est directement le poids de tirage qui est considéré pour le calage (redressé par le taux de réponse global pour des questions de convergence).

Remarque: les 2 méthodes sont strictement équivalentes si la probabilité de réponse est totalement expliquée par les variables de calage uniquement

Ce processus est « facile » à mettre en œuvre, mais attention : les variables explicatives qui modélisent la non réponse sont aussi celles qui assurent le calage. On suppose donc que la NR nourrissons peut être correctement expliquée par des variables dont on connait le vrai total.

De plus, il existe un avantage certain de la méthode simultanée pour des enquêtes sur lesquelles la donnée « vieillit ». En effet, si on sélectionne correctement les variables qui expliquent la NR « globale » (et pas forcément à un temps donné), avec des variables dont on connait le total, on n'a pas besoin de connaitre chacune de ces modalités pour les non répondants. En effet, on va directement caler les seuls répondants pour obtenir les totaux cherchés. Cela suppose que les variables expliquant la NR ne dépendent pas du « temps » de l'enquête au-delà de la maternité (2 ans, 3 ans, 5 ans ,….).

On estime alors *a posteriori* un équivalent de probabilité de réponse pour chaque nourrisson, par le rapport poids avant calage / poids après calage. Contrairement à la pondération maternité, la probabilité de réponse est donc « individuelle » (chaque nourrisson a une probabilité de réponse qui lui est propre), et non commune aux nourrissons dans un même groupe de réponse homogène.

Méthode de pondération retenue : le calage simultané

Comme précisé dans le paragraphe précédent, la qualité de la méthode de calage simultané dépend fortement des variables prises en compte, ces variables devant autant que possible expliquer le phénomène de NR.

On a vu que les variables expliquant la NR aux 2 mois, 1 an et 2 ans de l'enfant sont sensiblement les mêmes. On dispose de plus à ce sujet d'une étude complète sur l'attrition de l'enquête ELFE : « ELFE Attrition 2011 – 2016 ». Cette étude propose d'analyser le phénomène d'attrition selon de nombreux axes. Après de nombreuses analyses descriptives, la recherche d'un meilleur modèle expliquant l'attrition au cours du temps a été réalisée par une régression pas à pas. Après imputation des valeurs manquantes à petits taux (moins de 7%), on obtient

un modèle à 10 variables avec par ordre d'importance : PCS de la mère, PCS du père, acceptation de la transmission de données de l'enfant, séance de préparation à la naissance, consommation d'alcool, âge du père, vacances, niveau d'étude de la mère, activité du père, activité de la mère.

Il est intéressant de noter que toutes ces études proposent toujours de modéliser la non réponse ou l'attrition par les mêmes variables portant sur des variables socio démographiques, des variables de santé et des variables d'implication du couple pendant la grossesse.

Afin de ne pas multiplier les variables de calage, on a privilégié les variables expliquant la non réponse en maternité (taux de non réponse avoisinant les 50%, contre 10 à 20% à chaque temps suivants). On a également dû privilégier les variables avec des données disponibles sur une population comparable au champs Elfe dans l'ENP ou celles avec des données disponibles sur l'état civil.

Pour ce qui est de l'activité, la PCS étant parfois mal codifiée (déclaratif avec des modalités parfois incomprises, PCS du père renseignée par la mère, ..), on a préféré conserver les variables les plus fiables, croisant niveaux d'études pour la mère, tranche d'âge et statut face à l'emploi (en activité ou non) pour les 2 parents.

On propose donc d'utiliser la méthode de calage simultané sur 13 variables.

Variables prises en compte :

6 variables dites « de contexte » (variables déjà prise en compte dans le principe de calage depuis l'enquête maternité) :

- Groupe de région du domicile en 5 groupes (Ile de France, Centre, Picardie / Nord-Est / Nord-Ouest / Sud-Est / Sud-Ouest)
- Mère primipare (oui /non)
- Etat matrimonial (né dans le mariage / hors mariage)
- Age de la mère (18-24 / 25-29 / 30-34 / 35 et +)
- Niveau études de la mère (non scolarisée, primaire, CAP, BEP / 2nd, 1ère, terminale / études supérieures)
- Statut mère immigrée (oui / non)

Auxquelles on ajoute les variables suivantes permettant de prendre en compte l'attrition :

- Séances de préparation à l'accouchement (oui / non)
- Activité du père au moment de l'accouchement (en emploi / autre)
- Age du père (18-24 / 25-29 / 30-34 / 35 et +)
- Mère vivant en couple à la naissance (oui / non)
- Consommation d'alcool pendant la grossesse (oui / non)
- Naissance gémellaire (oui / non)
- Activité de la mère au moment de l'accouchement (en emploi / autre)

Signalons que pour ces nouvelles variables, les données manquantes (5% pour l'âge du père, moins de 2% pour les autres) ont été imputées avant le calage. Comme pour les méthodes précédentes, l'ensemble des totaux sur la population Elfe sont issues de l'état civil ou de l'ENP de 2010.

Comparaison de la nouvelle méthode avec les résultats déjà obtenus

Afin de s'assurer de la légitimité de cette méthode et de mesurer son impact sur les résultats déjà publiés, on a appliqué cette méthode aux nourrissons ayant répondu à l'enquête maternité, à l'enquête 1 an et à l'enquête 2

ans. On peut donc à la fois comparer, selon les 2 méthodes, le poids obtenu pour un même individu aux différents horizons et les résultats obtenus grâce aux répondants aux différentes enquêtes.

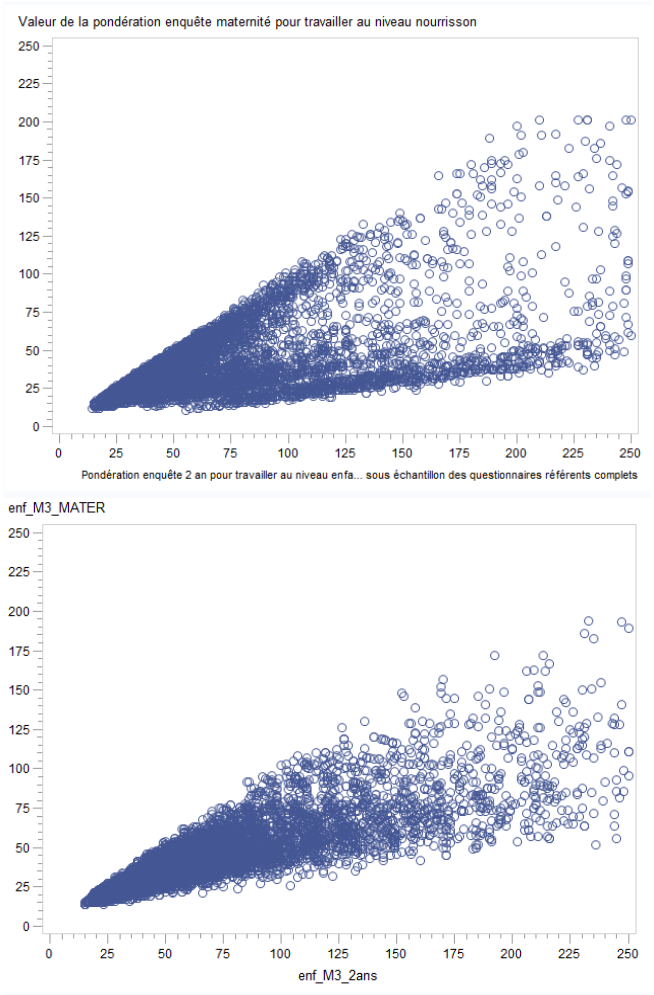
Les statistiques de base sont équivalentes.

Variable	Libellé	N	Moyenne	Maximum	Minimum	Intervalle	10ème ctl	25ème ctl	50ème ctl	75ème ctl	90ème ctl
MATER	ancienne méthode	18 201	41,96	201,29	10,64	190,65	19,79	23,85	31,34	45,45	74,89
	méthode calage simultané	18 201	41,98	200,51	14,64	185,87	21,48	26,01	34,14	49,02	70,78
1 AN	ancienne méthode	14 031	54,43	252,66	13,24	239,41	23,24	28,68	38,75	58,16	103,05
	méthode calage simultané	14 031	54,45	252,69	15,98	236,71	23,85	29,71	41,35	63,31	101,67
2 ANS	ancienne méthode	12 904	59,20	254,73	14,10	240,63	22,78	28,22	40,04	64,93	123,02
	méthode calage simultané	12 904	59,21	254,93	15,08	239,85	24,74	31,09	43,93	68,48	114,88

Toutefois, la nouvelle méthode présente un premier avantage important. Les poids sont plus compacts. En effet, même si leur étendue est similaire, de même que leur intervalle inter quartile, les poids extrêmes sont plus diffus avec la méthode actuelle. Cela entraine une variance des poids bien plus forte (20 à 25% supérieure).

ancienne méthode				méthode calage simultané			
Mesures statistiques de base				**Mesures statistiques de base**			
Location		**Variabilité**		**Location**		**Variabilité**	
Moyenne	41,96	**Ecart-type**	32,55	**Moyenne**	41,98	**Ecart-type**	25,14
Médiane	31,34	**Variance**	1 059,00	**Médiane**	34,14	**Variance**	631,88
Mode	201,29	**Intervalle**	190,65	**Mode**	200,51	**Intervalle**	185,87
		Ecart interquartile	21,60			**Ecart interquartile**	23,01
Mesures statistiques de base				**Mesures statistiques de base**			
Location		**Variabilité**		**Location**		**Variabilité**	
Moyenne	54,43	**Ecart-type**	46,14	**Moyenne**	54,45	**Ecart-type**	39,55
Médiane	38,75	**Variance**	2 129,00	**Médiane**	41,35	**Variance**	1 565,00
Mode	252,66	**Intervalle**	239,41	**Mode**	252,69	**Intervalle**	236,71
		Ecart interquartile	29,47			**Ecart interquartile**	33,60
Mesures statistiques de base				**Mesures statistiques de base**			
Location		**Variabilité**		**Location**		**Variabilité**	
Moyenne	59,20	**Ecart-type**	51,80	**Moyenne**	59,21	**Ecart-type**	44,69
Médiane	40,04	**Variance**	2 683,00	**Médiane**	43,93	**Variance**	1 997,00
Mode	254,73	**Intervalle**	240,63	**Mode**	254,93	**Intervalle**	239,85
		Ecart interquartile	36,72			**Ecart interquartile**	37,39

Second enseignement : si on compare, pour un même individu, le poids qu'il représente en maternité, à 1 an ou à 2 ans, on note une plus grande dérive de cette relation pour la méthode de pondération actuelle. Par exemple ; on lit dans les graphiques suivants que les nourrissons avec une pondération à 2 ans de 100 (en abscisse) pouvaient avoir un poids en maternité entre 15 et 100. Certains voient donc leur poids multiplier par 6, d'autres n'en changent pas. Avec la nouvelle méthode, cette fourchette se situe entre 30 et 100. On conserve donc des poids plus similaires au cours du temps. Ce fait est d'autant plus important que la dérive au cours des enquêtes futures risque d'être encore plus importante. Cela pourrait entrainer, à termes, que des résultats calculés en maternité seraient très différents si on les analysait avec un sous échantillon de répondants au temps futurs.



Comparaison des poids nourrisson maternité - 2 ans. Méthode actuelle (en haut) / nouvelle (en bas)

A noter enfin qu'il y a que très peu de différences entre les prévalences calculées avec la méthode actuelle de pondération et la nouvelle méthode appliquée en maternité. On trouvera par exemple à la suite des distributions avec la méthode utilisée jusqu'aux 2 ans de l'enfant (nommées Elfe_) et ce que ces distributions seraient si on avait utilisé la méthode de calage simultanée (nommée Nle_), avec les pondérations maternités (MATER), à 1 an et à 2 ans.

Remarque : les écarts observés entre certaines variables de calage et les distributions finales proviennent de la troncature réalisée après le calage.

PONDERATIONS DES ENQUETES NATIONALES.

Distributions de variables sur la mère et le père:

	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_LIEUNAISM(Lieu de naissance mère)						
1-En France	81,55	81,39	81,79	81,79	82,19	81,84
2-Dans un autre pays	18,45	18,61	18,21	18,21	17,81	18,16
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_NATIOM(Nationalité mère)						
1-Française de naissance (y compris par réintégration)	82,95	83,07	83,32	83,65	83,5	83,48
2-Française par acquisition (naturalisation, mariage, déclaration, ou option à la majorité)	4,1	4,81	4,6	4,94	4,88	5,11
3-Etrangère	12,9	12,08	12,06	11,39	11,58	11,38
4-Apatride	0,04	0,04	0,02	0,02	0,04	0,03
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_ETATMAT(Etat matrimonial mère)						
1-Mariée ou remariée (y compris séparée légalement)	45,08	44,99	44,78	45,13	44,94	45,26
2-Pacsée	12,7	12,91	13,79	13,75	14,01	14,13
3-Divorcée	1,28	1,29	1,28	1,32	1,22	1,27
4-Célibataire	40,76	40,67	40,08	39,74	39,74	39,26
5-Veuve	0,18	0,14	0,06	0,06	0,09	0,07
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_COUPLE(La mère vit en couple)						
0-non	7,63	7,31	7,15	7,07	6,5	6,82
1-oui	92,37	92,69	92,85	92,93	93,5	93,18
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_NIVET(Niveau d'études mère)						
1-Ecole primaire	1,33	1,04	0,85	0,74	0,72	0,78
2-Collège (classes de la 6e à la 3e)	7,58	6,71	6,3	6,01	5,6	5,91
3-Classes préparant à un CAP ou à un BEP	17,98	19,53	18,84	20,23	17,86	19,77
4-Classes de seconde, première ou terminale générales	8,26	7,92	8,01	7,86	8,3	7,9
5-Classes de seconde, première ou terminale techniques	3,4	3,5	3,58	3,62	3,69	3,53
6-Classes de seconde, première ou terminale professionnelles	8,36	8,5	8,27	8,44	8,81	8,51
7-Etudes supérieures (facultés, IUT, etc,)	52,53	52,38	53,89	52,81	54,69	53,26
8-Vous n'avez jamais été scolarisée	0,56	0,43	0,26	0,28	0,32	0,34
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_PROFESS(Catégorie profession mère)						
1-Agriculteur, exploitant	0,31	0,33	0,39	0,39	0,38	0,4
2-Artisan, commerçant ou chef d'entreprise	2,81	3,21	3,07	3,24	3,26	3,29
3-Cadre ou profession intellectuelle supérieure	11,54	13,57	12,15	13,88	12,1	13,97
4-Profession intermédiaire (instituteur, infirmier, technicien, contremaître...)	16,41	17,05	17,7	17,64	17,75	17,87
5-Employé	37,4	41,7	41,58	42,28	43,09	42,67
6-Ouvrier	1,91	2,15	2,31	2,23	2,34	2,19
7-Sans profession	8,96	6,51	6,92	6,16	6,33	5,96
9-Ne peut classer la profession	20,66	15,49	15,88	14,19	14,74	13,64

PONDERATIONS DES ENQUETES NATIONALES.

	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_LIEUNAISP(Lieu de naissance père)						
1-En France	81.48	82.35	81.96	83.84	82.54	83.87
2-Dans un autre pays	18.52	17.65	18.04	16.16	17.46	16.13
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_NATIOP(Nationalité père)						
1-Française de naissance (y compris par réintégration)	82.56	83.26	82.99	84.75	83.46	84.61
2-Française par acquisition (naturalisation, mariage, déclaration, ou option à la majorité)	5.30	5.07	5.26	4.77	5.06	4.93
3-Etrangère	11.84	11.40	11.54	10.28	11.21	10.19
4-Ne sait pas	0.30	0.27	0.21	0.21	0.27	0.27
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_EMPLOIC(Situation professionnelle père)						
1-A un emploi	88.35	87.48	89.48	87.93	90.30	88.27
2-Est homme au foyer	0.40	0.45	0.35	0.40	0.32	0.38
3-Est élève, étudiant ou en formation	1.41	1.52	1.55	1.77	1.61	1.85
4-Est au chômage	6.60	7.13	5.78	6.64	5.33	6.44
5-Est en congé parental	0.11	0.13	0.10	0.11	0.11	0.12
6-Est retraité	0.16	0.18	0.13	0.15	0.11	0.17
7-Est dans une autre situation	2.98	3.11	2.62	2.99	2.22	2.77
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_PEREACC(Le père a assisté à l'accouchement)						
0-non	21.77	21.12	20.12	19.79	20.33	19.27
1-oui	78.23	78.88	79.88	80.21	79.67	80.73

Distributions de variables sur la grossesse :

	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_TABAG(Tabagisme pendant la grossesse)						
0-Non	78,38	78,05	77,79	78,74	78,84	79,22
1-Oui	21,62	21,95	22,21	21,26	21,16	20,78
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_TABA3G(Tabagisme pendant le 3e trimestre)						
0-Non	15,01	15,3	16,67	16,46	15,92	15,85
1-Oui	81,31	81,03	80,3	80,18	81,18	81,22
9-Non renseigné	3,68	3,67	3,03	3,36	2,89	2,93
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_FQALCOOL(Consommation d'alcool)						
0-Jamais	77,88	76,57	76,36	76,45	76,32	76,4
1-1 fois par mois ou moins souvent, ou lors d'occasions particulières comme les fêtes	15,16	16,15	16,44	16,49	16,6	16,64
2-2 à 4 fois par mois	1,57	1,6	1,55	1,48	1,59	1,55
3-2 à 3 fois par semaine	0,22	0,29	0,24	0,28	0,18	0,21
4-4 fois par semaine ou plus, mais pas tous les jours	0,04	0,05	0,02	0,02	0,05	0,03
5-Tous les jours	0,09	0,05	0	0	0,03	0,02
6-Seulement avant de se savoir enceinte	4,98	5,23	5,35	5,22	5,17	5,11
7-Ne souhaite pas répondre	0,06	0,07	0,04	0,06	0,05	0,04
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00X_HTAG(Hypertension artérielle pendant la grossesse)						
0-Non	96,14	96,34	96,4	96,35	96,1	96,23
1-Oui avec protéinurie (?0,3g/l ou par 24h)	1,69	1,55	1,51	1,54	1,74	1,72
2-Oui sans protéinurie	2,16	2,11	2,09	2,11	2,16	2,05
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00X_DIABGEST(Diabète gestationnel)						
0-Non	92,33	92,52	92,46	92,51	92,25	92,53
1-Oui	7,67	7,48	7,54	7,49	7,75	7,47

PONDERATIONS DES ENQUETES NATIONALES.

Distributions de variables sur l'accouchement :

	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00X_DEBTRAV(Début du travail)						
1-Travail spontané	70,1	70,46	70,61	70,53	69,66	70,24
2-Déclenchement (y compris maturation du col seul)	19,95	19,76	19,72	19,6	20,34	19,8
3-Césarienne avant le début du travail	9,95	9,79	9,67	9,87	10	9,97
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00X_TYPACC(Accouchement)						
1-Voie basse spontanée	67,62	67,81	67,91	67,84	67,96	68,13
2-Forceps, spatules, ventouses	11,66	11,61	11,37	11,23	11,19	11,08
3-Césarienne	18,79	18,52	18,69	18,79	18,88	18,74
9-Ne sait pas	1,93	2,06	2,03	2,15	1,97	2,05
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00X_SEXEC3(Sexe)						
1-Masculin	51,33	51,32	51,14	50,91	50,41	50,4
2-Féminin	48,53	48,53	48,75	48,96	49,44	49,45
9-Ne sait pas	0,14	0,15	0,11	0,13	0,16	0,15

Annexe 1 : Procédure logicielle SAS

Le code permettant de générer une pondération est proposé, utilisant le logiciel SAS 9.4 (SAS Institute Inc, 2013).

Cette procédure nécessite l'emploi de la macro SAS CALMAR (CALage sur MARges) permettant de redresser un échantillon provenant d'une enquête par sondage, par repondération des individus, en utilisant une information auxiliaire disponible sur un certain nombre de variables, appelées variables de calage : https://www.insee.fr/fr/information/2021902.

Après création des marges, cette procédure nécessite de fournir :

table_selection=nom de la table où sont les données à pondérer. Attention, cette table doit comprendre **uniquement** ces nourrissons. Il convient donc que l'utilisateur génère précédemment cette table par une étape DATA classique.
Cette table doit contenir a minima les champs suivants:

- *identifiant* = champs identifiant les enregistrements (par défaut ID...) ;
- *strate* = champs qui identifie la strate de la maternité (par défaut M00M1_MATSTRATEC1) ;
- *vague* = champs qui identifie la vague de l'enquête (par défaut M00M1_VAGUE) ;
- les variables de calage *CS_1* à *CS_13.*

Tsortie = nom de la table en sortie
Psortie = nom du champ de la *Tsortie* avec le poids
poidsMAX = poids max pour troncature
Tmarge = nom de la table avec les marges. Par défaut margesM3
Label = Label de *Psortie*
tronc = 1 si les poids doivent être tronqués à *poidsMAX*. 0 sinon

A titre d'exemple, la table SAS « *pond3ans* » contient les données des 11706 enfants ayant participé à l'enquête 3 ans ½. La pondération a été générée par la commande suivante :

```
%CALAGE
(table_selection=pond3ans,
strate=M00M1_MATSTRATEC1,
vague=M00M1_VAGUE,
identifiant=IDXX_XX,
Tsortie=pond_3ans,
Psortie=enf_3ans,
poidsMAX=250,
Tmarge=margesM3,
Label=enf_3ans,
tronc=1);
```

Cette procédure génère la table « *pond_3ans* », qui comprend uniquement les 2 variables *IDXX_XX* et *enf_3ans*.

Cette procédure fournit également quelques statistiques sur les poids générés (avant et après troncature).

Variable d'analyse : Temp_poids enf_3ans												
N	Moyenne	Maximum	Minimum	Intervalle	Somme	5ème ctl	10ème ctl	25ème ctl	50ème ctl	75ème ctl	90ème ctl	95ème ctl	99ème ctl
11706	65.2656757	1154.88	15.5508769	1139.33	764000.00	22.2652972	25.1097985	31.5196500	45.0881653	72.7323350	125.5486840	176.6201885	324.5000692

Variable d'analyse : enf_3ans enf_3ans												
N	Moyenne	Maximum	Minimum	Intervalle	Somme	5ème ctl	10ème ctl	25ème ctl	50ème ctl	75ème ctl	90ème ctl	95ème ctl	99ème ctl
11706	65.2656757	259.9092983	16.1672700	243.7420284	764000.00	23.1478311	26.1050805	32.7690005	46.8753337	75.6152406	130.5250814	183.6209171	259.9092983

Code SAS :

```
data margesM3;
input var $ n mar1-mar6;
cards;
CS_1 5 29.96 19.15 15.42 19.93 15.54 .
CS_2 2 43.1 56.9 . . . .
CS_3 2 45 55 . . . .
CS_4 4 13.96 31.22 33.25 21.57 . .
CS_5 3 27.8 19.9 52.3 . . .
CS_6 2 81.25 18.75 . . . .
CS_7 2 50.2 49.8 . . . .
CS_8 2 87.36 12.64 . . . .
CS_9 2 7.3 92.7 . . . .
CS_10 4 6.66 22.65 32.94 37.75 . .
CS_11 2 76.6 23.4 . . . .
CS_12 2 97.25 2.75 . . . .
CS_13 2 69.7 30.3 . . . .
;


%MACRO CALAGE (table_selection, identifiant, strate, vague, Tsortie, Psortie, poidsMAX,
Tmarge,Label, tronc );
title  ' ';


%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;
/* table avec uniquement les variables de calage nécessaires */
data calageSIMU (keep=&identifiant
&strate  &vague
&listeCALAGE
);
set &table_selection;
run;


data calageSIMU; set calageSIMU;
if &strate = 1 then _TOTALMAT_=108;
else if &strate = 2 then _TOTALMAT_=108;
else if &strate = 3 then _TOTALMAT_=109;
else if &strate = 4 then _TOTALMAT_=108;
else if &strate = 5 then _TOTALMAT_=111;

if &strate = 1 then _MATsel_=25;
else if &strate = 2 then _MATsel_=44;
else if &strate = 3 then _MATsel_=62;
else if &strate = 4 then  MATsel =88;
else if &strate = 5 then _MATsel_=101;


if &vague = 1 then _TOTALJOUR_=90;
else if &vague = 2 then _TOTALJOUR_=91;
else if &vague = 3 then  TOTALJOUR =92;
else if &vague = 4 then _TOTALJOUR_=92;

if &vague = 1 then _JOURsel_=4;
else if &vague = 2 then _JOURsel_=6;
else if &vague = 3 then _JOURsel_=7;
else if &vague = 4 then _JOURsel_=8;

pondAVANT_calage = (_TOTALMAT_ * _TOTALJOUR_) / (_MATsel_ * _JOURsel_) ;
run;

proc sort data=calageSIMU; by &identifiant; run;
```

```
*  écrit en dur NB=36028 = nb de nourrissons si aucune NR;
%let NB = 36028;

/* stocker dans NBP le total nourrissons à traiter = présents dans la table*/
proc sql;
create table totalP
as select count(*) as NBP from calageSIMU ;
quit;

data _null_;set totalP;
CALL SYMPUT('NBP', NBP);
run;

/* on redresse la pond avant calage par le taux de réponse global. NB/NBP.
c'est juste pour améliorer la convergence de calmar*/
data calageSIMU;
set calageSIMU;
pondAVANT_calageR=pondAVANT_calage*&NB/&NBP;
run;
proc delete data= totalP;


%let Ttemp=Tmp_&Tsortie;

%CALMAR (data= calageSIMU , poids=pondAVANT_calageR , ident=&identifiant ,
                datamar=&Tmarge , m=2 , /*editpoi=oui, */  /* 3 : logit */
               /* 2 : raking ratio */
               /* LO=0.6 , UP=1.3 ,*/
                datapoi=&Ttemp , poidsfin=Temp_poids , labelpoi=&Label,
                PCT=oui , EFFPOP=&TOTPOP);


/* si besoin, pour tronquer poids calés */
data Temp_sortie;
set &Ttemp ;
if (&tronc=1 and Temp_poids>&poidsMAX) then Temp_poids_T=&poidsMAX;
else Temp_poids_T=Temp_poids;
run;

proc sql ;
create table tronque
as select sum(Temp_poids_T) as tt from Temp_sortie;
quit;

data _null_;set tronque;
CALL SYMPUT('total_Tronque', tt);
run;
proc delete data= tronque; run;

DATA Temp_sortie ;
set Temp_sortie ;
Temp_poids_T = Temp_poids_T *&TOTPOP/ &total_Tronque;
run;

* génération de la table finale;
data &Tsortie (keep=&identifiant &Psortie);
set Temp_sortie;
&Psortie= Temp_poids_T;
label &Psortie=&Label;
run;

proc sort data=&Tsortie; by &identifiant;run;


title 'statistiques poids AVANT troncature';
*données avant/apres troncature;
```

PONDERATIONS DES ENQUETES NATIONALES.

```
proc means data=Temp_sortie  n mean max min range sum p5 p10 p25 p50 p75 p90 p95 p99;
var Temp_poids; run;

title  'statistiques poids APRES troncature';
*données avant/apres troncature;
proc means data=&Tsortie  n mean max min range sum p5 p10 p25 p50 p75 p90 p95 p99;
var &Psortie; run;

proc delete data= &Ttemp;
proc delete data= Temp_sortie;
proc delete data= calageSIMU;

title ' ';
%mend;
```