PLAN DE SONDAGE DES ENQUETES GENERATION : UTILISATION D'UN CALAGE POUR SURECHANTILLONNER LES EXTENSIONS

Christophe Barret, Mady Cissé, Christophe Dzikowski (*)

(*) Céreq, Centre d'Etudes et de REcherches sur les Qualifications

christophe.barret@cereq.fr, mady.cisse@cereq.fr, christophe.dzikowski@insee.fr

Mots-clés : Échantillonnage, calage, extension

Résumé

Les enquêtes Génération du Céreq permettent d'étudier l'insertion professionnelle des jeunes à l'issue du système éducatif. Il s'agit d'une enquête sur la France entière et sur tous les niveaux et spécialités de formations. De nombreux acteurs publics partenaires sollicitent le Céreq pour mener des extensions d'échantillon sur leur population d'intérêt. Il s'agit par exemple de ministères qui souhaitent une précision accrue sur les formations qui relèvent de leur compétence ou bien des Régions qui souhaitent étudier l'insertion des jeunes sorties de formation sur leur territoire.

Cet article présente les différentes étapes de l'échantillonnage de l'enquête 2016 auprès de la Génération 2013. La méthodologie retenue permet de prendre en compte trois difficultés importantes pour la constitution de l'échantillon.

En effet, le Céreq constitue lui-même la base de sondage des élèves sortants du système éducatif en 2012-2013. Cette base de sondage présente un défaut de sous-couverture lié à la non-réponse de certains établissements et un défaut de sur-couverture lié à la présence dans la base d'individus hors champ de l'enquête. Enfin, le plan de sondage tient compte des demandes d'extension des partenaires. La difficulté principale provient du fait que certaines extensions se croisent. Une solution innovante, par calage sur marges sur les cibles d'extension, est proposée pour éviter d'aboutir à des échantillons dans lesquels le nombre d'individus échantillonnés au sein des intersections entre extensions soit trop important.

La méthodologie peut se décomposer en deux phases. La première phase consiste à effectuer le calcul des probabilités individuelles de tirage qui tiennent compte du taux de couverture de la base de sondage, des probabilités estimées d'être dans le champ de l'enquête et des probabilités anticipés de réponse. Dans un premier temps, les probabilités de tirage sont déterminées en l'absence d'extension pour atteindre le nombre de questionnaires financés par le Céreq. Pour satisfaire les besoins des extensions, des coefficients de dilatation sont calculés pour atteindre les cibles des extensions et appliqués pour obtenir les probabilités de tirage finales. La deuxième phase consiste à effectuer un unique tirage équilibré de l'échantillon sur la base de ces dernières probabilités de tirage.

En l'absence d'intersection entre les extensions, le calcul des coefficients de dilatation pourrait se faire séquentiellement extension par extension. En revanche, en présence d'intersections entre les extensions, le traitement séquentiel des calculs des coefficients de dilatation pose plusieurs problèmes. Le calcul serait en effet dépendant de l'ordre dans lesquels seraient considérés chaque extension et les intersections seraient de plus artificiellement sur-échantillonnées. La méthode de calage pour calculer les coefficients de dilatation proposée ici permet à la fois de corriger le problème de dépendance à l'ordre et le problème de sur-échantillonnage des intersections.

Cette stratégie d'échantillonnage, en un unique tirage, sera comparée à une stratégie d'échantillonnage plus classique qui consiste à réaliser un premier tirage de l'échantillon national

(sans extension) puis un second tirage dans le complémentaire du premier échantillon pour satisfaire les besoins des extensions. Les avantages et inconvénients des deux méthodes seront discutés.

Abstract

The Generation surveys main objective is to facilitate a regular evaluation of the education-to-work transition over the first three years of the young people's working lives. The survey methodology must deal with several issues. In the first hand, there are coverage issues in the sample frame. In the second hand, partner institutions' requests for sample extensions involve complexity in the sampling methodology. The major risk is to overestimate the population which belongs to several extensions. This article introduces a method to fix the inclusion probabilities by a calibration method. This method is compared with a two-step sampling.

Introduction

Les enquêtes Génération du Centre d'études et de recherche sur les qualifications (Céreq) permettent d'étudier l'insertion professionnelle des jeunes à l'issue du système éducatif. Il s'agit d'une enquête sur la France entière et sur tous les niveaux et spécialités de formations. De nombreux acteurs publics partenaires sollicitent le Céreq pour mener des extensions d'échantillon sur leur population d'intérêt. Il s'agit par exemple de ministères qui souhaitent une précision accrue sur les formations qui relèvent de leur compétence ou bien des Régions qui souhaitent étudier l'insertion des jeunes sortis de formation sur leur territoire.

Le plan de sondage tient compte de ces demandes d'extension. La difficulté principale provient du fait que certaines extensions se croisent. Une solution originale de calcul de probabilité de tirage, par calage sur marges sur les cibles d'extension, est proposée pour éviter d'aboutir à des échantillons dans lesquels le nombre d'individus échantillonnés dans les intersections entre extensions soit trop important.

Afin de contextualiser au mieux cette méthode, l'ensemble des étapes de l'échantillonnage de l'enquête 2016 auprès de la Génération 2013 sera présenté. La méthodologie générale retenue permet de prendre en compte deux autres difficultés importantes pour la constitution de l'échantillon, en plus de celle sur la prise en compte des extensions avec une intersection non vide.

En effet, le Céreq constitue lui-même la base de sondage des élèves sortants du système éducatif en 2012-2013. Cette base de sondage présente un défaut de couverture lié à la non-réponse de certains établissements et un défaut de sur-couverture lié à la présence dans la base d'individus hors champ de l'enquête.

La méthodologie générale peut se décomposer en deux phases, qui elles-mêmes se décomposent en différentes étapes :

- Phase A : le calcul des probabilités individuelles de tirage
    - Étape A1 : le taux de couverture
    - Étape A2 : détermination des probabilités de tirage en l'absence d'extension
    - Étape A3 : prise en compte des extensions dans le calcul des probabilités de tirage
- Phase B : le tirage équilibré de l'échantillon
    - Étape B1 : Tirage de l'échantillon global (principal + réserve)
    - Étape B2 : Tirage de la réserve et de l'échantillon principal

L'article met l'accent sur l'utilisation d'un calage pour calculer les probabilités de tirage des extensions (étape A3). Les autres phases de la méthodologie seront présentées en annexe.
Cette stratégie d'échantillonnage, en un unique tirage, sera comparée à une stratégie d'échantillonnage plus classique qui consiste à réaliser un premier tirage de l'échantillon national (sans extension) puis un second tirage dans le complémentaire du premier échantillon pour satisfaire les besoins des extensions. Les avantages et inconvénients des deux méthodes seront discutés en conclusion.

1. Calcul des probabilités de tirage pour simuler un échantillon socle

Un des objectifs de cette méthodologie d'échantillonnage est de garantir que la structure des répondants se rapproche le plus possible de la structure de la population d'intérêt. Pour cela, les probabilités de tirage intègrent des coefficients pour corriger d'une part les défauts de couverture de la base de sondage construite par le Céreq et d'autre part anticiper les forts taux de non réponse de certaines sous populations. Les étapes méthodologiques correspondantes sont en annexe 1.
Par conséquent, les probabilités de tirage sont construites pour sur-représenter les individus dont les taux de réponse attendus sont plus faibles ainsi que ceux appartenant à un type d'établissement mal couvert dans la base de sondage. Ainsi, la structure des répondants sera proche de la structure théorique de la population cible (du moins sur les variables utilisées pour le calcul des poids de couverture et des probabilités de réponse). La justification de cette propriété est en annexe.
Ainsi, les probabilités de tirage π_{i1}sont de la forme :

$$\pi_{i1} = \frac{poids\ de\ couverture_i}{probabilité\ de\ répondre_i} * coeff1 = \frac{pcouv_i}{pr_i} * coeff1$$

Où $coeff1$ est un coefficient de dilatation identique pour tous les individus de la base de sondage. Ce coefficient est calculé pour simuler 8 500 répondants dans le champ à partir de la relation suivante :

$$\sum_{i \epsilon U} pr_i^C * \pi_{i\,1} = \mathrm{E}(\mathrm{nr}^C)$$

Avec pr_i^C la probabilité de i de répondre dans le champ sachant qu'il est tiré dans l'échantillon
nr^C l'estimation du nombre de répondants dans le champ

Soit

$$coeff1 = 8500 \Big/ \sum_{i \epsilon U} \left(pr_i^C * \frac{pcouv_i}{pr_i} \right)$$

La valeur du coefficient $coeff1$ est environ 0,011.

2. Prise en compte des cibles d'extension dans le calcul des probabilités de tirage (étape A3)

Pour cette enquête, cinq partenaires publics financent des extensions d'échantillon sur une ou plusieurs populations d'intérêt. À ces demandes externes, certaines populations font également l'objet d'extension pour les besoins propres du Céreq (collège, bacs généraux et techno, bac pro, cap, …).
Pour chacune des populations d'intérêt, une cible de questionnaires (ie. une cible d'individus répondants dans le champ) a été déterminée. Le tableau 1, ci-après, précise pour un extrait des sous-population d'intérêt, la cible souhaitée ainsi que le nombre de questionnaires espéré à partir de l'échantillon socle (après l'étape précédente de calcul des probabilités de tirage pour obtenir 8500 questionnaires).
L'objectif de cette étape est de déterminer, à partir de calcul des probabilités de tirage de l'échantillon socle, les coefficients de suppléments de tirage nécessaires pour atteindre les cibles sur chaque population d'intérêt.
La principale difficulté est liée au recoupement des extensions. Le risque est d'aboutir à une structure de l'échantillon qui soit atypique du point de vue des extensions, par la sur-représentation des intersections, notamment dans le cas particulier d'une inclusion d'une extension dans une autre.

Tableau 1 : effectifs cibles par sous-populations d'extensions (extrait)

Sous population d'extension	codage des sous populations	cible (effectif extension +réserve)	nombre de questionnaire à partir de l'échantillon socle
Collège	c1	903	404
Baccalauréat général et technologique	c2	1 068	570
Baccalauréat professionnel	c3	2 391	1 374
CAP	c4	1 373	860
…	…	…	…
Ensemble	c38	24 700	8 490

2.1. Explication de la méthode par un exemple

Supposons deux sous-populations d'intérêt A et B. Les cibles de questionnaires fixées sont respectivement C_A et C_B.

Sur l'échantillon socle (visant 8500 questionnaires), on estime le nombre de questionnaires de A, noté $nsoc_A$, et de B, noté $nsoc_B$.

1er cas : supposons $A \cap B = \emptyset$

Dans ce cas, les coefficients de dilatation, notés dil_A et dil_B, sont immédiats et donnés par :

$$dil_A = \frac{C_A}{nsoc_A} \text{ et } dil_B = \frac{C_B}{nsoc_B}$$

Et les probabilités de tirage tenant compte des extensions sont donnés par :

$$\text{pour i}\epsilon A, \quad \pi_{i2} = \pi_{i1} \times dil_A = \pi_{i1} \times \frac{C_A}{nsoc_A}$$

$$\text{pour i}\epsilon B, \quad \pi_{i2} = \pi_{i1} \times dil_B = \pi_{i1} \times \frac{C_B}{nsoc_B}$$

Prenons les données fictives suivantes :

Les cibles sont fixés à $C_A = 1000$ et $C_B = 1300$, dans toute la suite.

Supposons que $nsoc_A = 100$ et $nsoc_B = 200$ (sans oublier que $nsoc_{A\cap B} = 0$)

On a donc $dil_A = \frac{1000}{100} = 10$ et $dil_B = \frac{1300}{200} = 6{,}5$

Autrement dit la probabilité de tirage des individus de A sera multiplié uniformément par 10, pour ceux de B par 6,5.

2e cas : supposons $A \cap B \neq \emptyset$

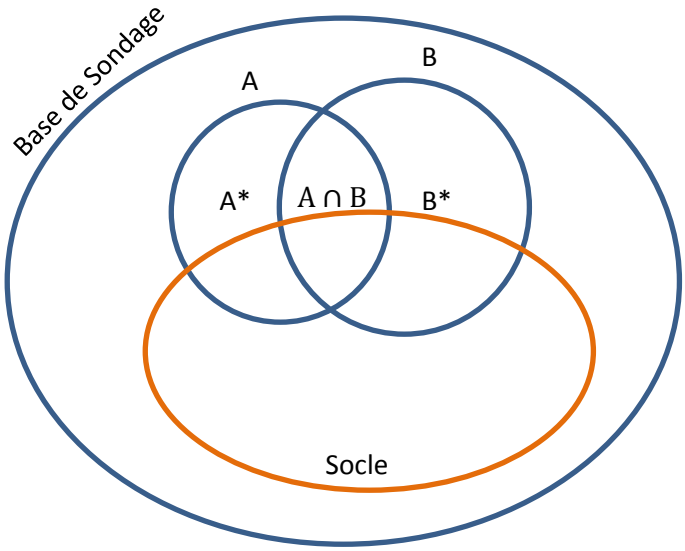
La principale difficulté de la prise en compte des extensions est liée au fait que certaines sous-populations d'intérêt se croisent.

Introduisons les notations supplémentaires suivantes :

$A^* = A \backslash (A \cap B)$

$B^* = B \backslash (A \cap B)$

$Z = SOCLE \backslash (A \cup B)$



Supposons toujours que $nsoc_A = 100$ et $nsoc_B = 200$

On considère que $nsoc_{A \cap B} = 50$ et donc que $nsoc_{A^*} = 50$ et $nsoc_{B^*} = 150$

La taille de l'échantillon socle est de 8500 on a donc $nsoc_Z = 8500 - 50 - 50 - 150 = 8250$

Une première méthode naturelle consiste à traiter séquentiellement chaque extension. On dilate les probabilités de tirage des individus de A (avec dil_A=1000/100). Comme l'intersection est non vide avec B, cette étape va conduire à gonfler le nombre d'individus de B présents dans l'échantillon (et répondants dans le champ). On recalcule dil_B avec les nouveaux effectifs (dil_B=1300/650=2), ce qui revient à doubler les poids de tirage de B. Le tableau ci-dessous détaille les calculs effectués pour cette première méthode et sa variante qui consiste à faire la même chose en commençant par B.

Tableau 2 : méthode séquentielle de calcul des suppléments de tirage

		METHODE 1 : d'abord A puis B		METHODE 1bis : d'abord B puis A	
		d'abord A (on applique dilA=1000/100 à la population A)	puis B (on applique dilB=1300/650 à la population B)	d'abord B (on applique dilB=1300/200 à la population A)	puis A (on applique dilA=1000/375 à la population B)
	nombre de répondants estimé dans l'échantillon socle	nombre de répondants estimé dans l'échantillon intermédiaire	nombre de répondants estimé dans l'échantillon final méthode 1 (F1)	nombre de répondants estimé dans l'échantillon intermédiaire	nombre de répondants estimé dans l'échantillon final méthode 1 (F1bis)
A*	50	500	500	50	133
B*	150	150	300	975	975
A∩B	50	500	1000	325	867
Z	8250	8250	8250	8250	8250
Total	8500	9400	10050	9600	10225
A	100	1000	1500	375	1000
B	200	650	1300	1300	1842

Il y a un double inconvénient à cette méthode « intuitive » :

- Les résultats dépendent de l'ordre dans lequel on traite les extensions
- Dans chaque méthode, on obtient le bon nombre de questionnaires (par rapport à la cible fixée) uniquement pour l'une des deux populations (celle traitée en dernier). Pour l'autre, en revanche, le nombre de questionnaires obtenus sera sensiblement supérieur à la cible fixée. Dans F1, il y aura 1500 répondants de A (au lieu des 1000 fixés). Dans F1bis, il y aura 1842 répondants de B (au lieu des 1300 fixés).

Une deuxième méthode, basée sur la précédente, permet de pallier le problème de dépendance à l'ordre de traitement des extensions. Il suffit de cumuler les coefficients multiplicatifs initiaux. En effet, on peut définir

$$\mathrm{dil}_A = 1 + \alpha_A \text{ et } \mathrm{dil}_B = 1 + \alpha_B$$

Puis on dilate les probabilités comme suit :

$$\begin{aligned} \text{pour i}\epsilon A^*, &\quad \pi_{i2} = \pi_{i1} \times (1 + \alpha_A) \\ \text{pour i}\epsilon B^*, &\quad \pi_{i2} = \pi_{i1} \times (1 + \alpha_B) \\ \text{pour i}\epsilon A \cap B, &\quad \pi_{i2} = \pi_{i1} \times (1 + \alpha_A + \alpha_B) \end{aligned}$$

Tableau 3 : méthode simultanée de calcul des suppléments de tirage

		METHODE 2 : A et B simultanément	
	nombre de répondants estimé dans l'échantillon socle	coefficient de dilatation	nombre de répondants estimé dans l'échantillon final méthode 2 (F2)
A*	50	10	500
B*	150	6,5	975
A∩B	50	15,5	775
Z	8250	1	8250
total	8500		10500
A	100		1275
B	200		1750

Cette méthode n'est plus dépendante de l'ordre des extensions. En revanche, les deux cibles sur les populations A et B sont dépassées.

Au-delà des inconvénients déjà cités, le principal problème des 3 méthodes présentées provient du fait d'aboutir à des échantillons qui vont exagérément surreprésenter les individus de l'intersection des deux extensions, comme le montre le tableau suivant.

Tableau 4 : poids des intersections

	socle	F1	F1bis	F2
poids des répondants de A∩B parmi ceux de A	50%	67%	87%	61%
poids des répondants de A∩B parmi ceux de B	25%	77%	47%	44%

Dans l'échantillon socle, on s'approche du poids « naturel » de l'intersection dans chacune des sous-populations. Les trois méthodes conduisent à très largement surreprésenter les individus de l'intersection.

2.2. Méthode de calcul des coefficients multiplicatifs par calage.

Pour pallier les inconvénients précédents (dépendance de l'ordre des extensions, dépassement des cibles, surreprésentation exagérée des intersections), le choix a été fait de procéder à des calculs de coefficients multiplicatifs de tirage par une méthode de calage sur un tableau de données particulier.

L'idée de base de cette méthode est d'itérer les méthodes M1 et M1bis jusqu'à convergence. C'est la raison pour laquelle nous nous sommes orientés vers une méthode de calage décrite ci-après.

2.2.1. Les marges de calage

Les marges de calage sont issues du *Tableau 1 : effectifs cibles par sous-populations d'extensions*. Pour que l'algorithme de calage se déroule normalement, les niveaux généraux (CGDD (c5), santé social (c19), Ensemble (c38)) ont été retirés (pour éviter les colinéarités entre les variables de calages, les effectifs des niveaux généraux étant le total de leurs sous spécialités). D'autre part, des contraintes facilement atteintes en pratique (DGESIP licence (c16) et Quartier prioritaire de la ville (c18)) et une population tirée exhaustivement dans l'échantillon global (thèse (c36)) ont été retirés. Il y a donc au final 32 marges de calage. La table SAS des marges pour l'application de la macro calmar est la suivante. Dans Mar1 sont enregistrées les cibles à atteindre pour chaque sous population. Dans Mar2 se trouvent le complémentaire de Mar1 pour atteindre 24700 observations.

Tableau 5 : Les 32 marges de calage

var	N	MAR1	MAR2
c1	2	903	23797
c2	2	1068	23632
c3	2	2391	22309
c4	2	1373	23327
c6	2	842	23858
c7	2	817	23883
c8	2	567	24133
c9	2	748	23952
c10	2	416	24284
c11	2	7026	17674
c12	2	1959	22741
c13	2	972	23728
c14	2	771	23929
c15	2	350	24350
c17	2	1509	23191
c20	2	1180	23520

var	N	MAR1	MAR2
c21	2	349	24351
c22	2	454	24246
c23	2	208	24492
c24	2	324	24376
c25	2	474	24226
c26	2	62	24638
c27	2	1191	23509
c28	2	427	24273
c29	2	461	24239
c30	2	78	24622
c31	2	31	24669
c32	2	140	24560
c33	2	84	24616
c34	2	306	24394
c35	2	1678	23022
c37	2	112	24588

2.2.2. Données auxquelles est appliquée la méthode de calage.

Le calage est appliqué sur la table obtenue en croisant les 32 indicatrices d'extension retenues. Ce croisement d'indicatrices d'extension crée une partition de la base de sondage. Le croisement des 32 indicatrices donne une table avec 94 observations, autrement dit, il s'agit d'une partition en 94 parties qui représentent chacune une combinaison unique d'indicatrices d'extension. Pour chacune de ces parties, on estime le nombre d'enquêtes réalisées, à partir de la première simulation (correspondant à l'échantillon socle de 8500). Cette table est appelée *PARTITION*.
Le tableau suivant est un extrait de la table *PARTITION* sur laquelle est appliqué l'algorithme de calage. La colonne *Effectif Agrégé* donne, pour chaque partie de la partition, les effectifs anticipés de répondants dans le champ Céreq élargi[1]. C'est cette variable *Effectif Agrégé* que l'on va caler sur les marges des totaux souhaités.

Tableau 6 : extrait de la table PARTITION

Parties	c1	c2	c3	c4	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	c17	c20	c21	c22	c23	c24	c25	c26	c27	c28	c29	c30	c31	c32	c33	c34	c35	c37	Effectif agrégé
P1	1	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2,33
P2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	401,51
P3	2	1	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	16,07
…																																	…
P93	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	154,14
P94	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	203,66

*les indicatrices C5, C16, C18, C19, C36, C38 ont été retirées pour la création de cette table.
Note de lecture : l'observation 3 de ce tableau correspond à la partie n°3 qui est composée des individus appartenant aux sous-populations C2 et C9 sans appartenir à aucune autre sous-population. À partir de l'échantillon socle simulé, on estime le nombre de répondants dans le champ à 16,07 pour la partie 3.

À l'issue de l'étape de calage, pour chaque observation p de la table PARTITION, le coefficient $coeff_p$ est calculé. Celui-ci correspond au rapport entre les valeurs finales et initiales de la variable *Effectif Agrégé*. Pour préciser les variables mise en jeu, le système sur lequel est appliqué l'algorithme est le suivant :

Pour chacune des 32 sous-populations *j*, (les 37 indicatrices de sous populations moins les 5 retraits) l'algorithme cherche à résoudre le problème suivant :

$$min_{somme\ calée_p} \sum_{p \in PARTITION} eff_p * G\left(\frac{eff\ calé_p}{eff_p}\right)$$

Sous la contrainte :

$$\sum_{p \in PARTITION} eff\ calé_p * \begin{pmatrix} 1_{pj1} \\ 1_{pj2} \end{pmatrix} = t_{Xj}$$

$$\text{où } t_{Xj} = \begin{pmatrix} Mar1_j \\ Mar2_j \end{pmatrix}$$

[1] Lorsque l'on considère le champ plus vaste « champ Céreq + champ spécifique (post-initiaux) des extensions sport et santé », la simulation conduit à 8900 questionnaires et non à 8500 comme anticipé avec le champ Céreq uniquement. Et c'est sur la base de ce nombre de questionnaire réalisé dans le cadre de ce champ élargi que sont faites les étapes suivantes d'ajustement des probabilités de tirage.

- Le système de poids initial considéré ici, est le nombre de répondants obtenus sur chacune des parties p, à savoir $effectif\ agrégé_p$ (eff_p). Les poids finaux des observations p sont enregistrées dans la variable $eff\ calé_p$. C'est le nombre de répondants souhaité dans la partie p.
- L'indice p parcours les données du tableau PARTITION. Ce dernier contient 94 observations, une pour chaque partie de la partition.
- Dans le vecteur $\begin{pmatrix}1_{pj1}\\1_{pj2}\end{pmatrix}$,
    - 1_{pj1} est l'indicatrice valant 1 si la case de la $p^{ième}$ ligne et de la $j^{ième}$ colonne de PARTITION vaut 1. Et 0 sinon.
    - De même 1_{pj2} est l'indicatrice valant 1 si la case de la $p^{ième}$ ligne et de la $j^{ième}$ colonne de PARTITION vaut 2. Et 0 sinon.
- Le vecteur t_{Xj} a pour composantes les marges définies par la $j^{ième}$ ligne du tableau des marges de calage, c'est-à-dire le *Tableau 5*. La première composante étant Mar1, la deuxième Mar2.
- Le coefficient $coeff_p$ est alors le ratio $eff\ calé_p/eff_p$
- Avec G une fonction de distance. Dans le cas présent, il s'agit de la distance de la méthode linéaire dans la macro Calmar.

Après la réalisation du calage, les probabilités de tirage π_{i2} sont de la forme :

$$\pi_{i2} = \pi_{i1} * coeff_p$$

Ici on fait l'hypothèse qu'il y a une relation approximativement proportionnelle entre les probabilités de tirage et le nombre de répondants. L'approximation vient du fait que certaines probabilités dépassent 1 et que celles-ci sont ramenés à ce seuil maximal.

$$\pi_{i2} = \min(\pi_{i1} * coeff_p\ , 1)$$

Cette étape de calage ne fait pas tout, mais réalise tout de même l'essentiel du travail de détermination des coefficients de suppléments de tirage $coeff_p$. En effet, en plus du contrôle des valeurs des probabilités de tirage (inférieures ou égales à 1), on contrôle les valeurs des probabilités de tirage π_{i2} par rapport à leur valeur initiale π_{i1} . En effet, dans la démarche retenue, les probabilités de tirage π_{i2} doivent être supérieures aux premières probabilités de tirage π_{i1} . Or, les contraintes sur les cibles d'extension ne prennent pas en compte ici de contraintes sur les probabilités de tirage π_{i1}. Il est donc nécessaire de corriger les probabilités d'inclusion π_{i2} en forçant qu'elles soient supérieures ou égales à π_{i1}.

Le calage permet d'aboutir au résultat souhaité. Cependant les corrections avales provoquent un écart entre le résultat obtenu et le résultat souhaité sans que cela soit dommageable quant aux objectifs de cet échantillon.

Les probabilités de tirage obtenues conduisent à un échantillon dit global qui contient assez d'individus échantillonnés pour être scindé en un échantillon principal pour la collecte et un échantillon de réserve. Ce dernier est une sécurité, en dernier recours, en phase de collecte si les objectifs sur certaines populations risquent de ne pas être atteint. La méthode pour tirer l'échantillon principal et l'échantillon de réserve est mise en annexe 4.

3. Vérification de la méthode d'échantillonnage

Une simulation du nombre de répondants a été réalisée à partir de l'échantillon principal. Le tableau suivant montre les effectifs de répondants anticipés à partir de l'échantillon global S. Pour chaque sous population *J* le nombre de répondant souhaité à partir de l'échantillon global est estimé par :

$$\sum_{i \epsilon\, S \cap J} p(i\ repond\ dans\ le\ champ\ ) = \text{E(nombre de répondant dans le champ pour J)}$$

Il s'agit de l'espérance du nombre de questionnaire calculée sur les individus échantillonnés et faisant partie de la sous population J. Le même calcul est utilisé pour les effectifs estimés pour l'échantillon principal en restreignant cette fois à $S\ prin \cap J$. Ce tableau sur les estimations des effectifs répondant dans le champ à partir de l'échantillon principal est fourni plus loin.

Tableau 7 : écart entre échantillon global et cibles

Sous population d'extension	Effectif répondant à partir de l'échantillon global	cible (effectif extension +réserve)	différence entre échantillon global et les cibles
Collège	895	903	-8
Baccalauréat général et technologique	1 059	1 068	-9
Baccalauréat professionnel	2 403	2 391	12
CAP	1 372	1 373	-1
CGDD	3 372	3 349	23
CGDD niveau 1	831	842	-11
CGDD niveau 2	826	817	9
CGDD niveau 3	564	567	-3
CGDD niveau 4	744	748	-4
CGDD niveau 5	407	416	-9
DGESIP	7 350	7 026	324
DGESIP autre	2 044	1 959	85
DGESIP BTS	974	972	2
DGESIP Grandes écoles	802	771	31
DGESIP IUT	348	350	-2
DGESIP Licence	1 666	1 464	202
DGESIP Master	1 516	1 509	7
Quartier priorité de la ville	2 145	2 287	-142
Santé social	5 971	5 981	-10
Santé social aide-soignante	1 171	1 180	-9
Santé social assistant social	395	349	46
Santé social auxiliaire puéricultrice	510	454	56
Santé social conseiller économie familiale	207	208	-1
Santé social éducateur jeunes	323	324	-1
Santé social éducateur spécialisé	531	474	57
Santé social ergothérapeute	62	62	0
Santé social infirmier	1 190	1 191	-1
Santé social masseur kinésithérapeute	425	427	-2
Santé social moniteur éducateur	521	461	60
Santé social orthophoniste	77	78	-1
Santé social orthoptiste	31	31	0
Santé social podologue pédicure	140	140	0
Santé social psychomotricien	84	84	0
Santé social sage femmes	304	306	-2
Sport	2 605	1 678	927
Thèse	2 486	2 446	40
Docteur en santé	110	112	-2
Ensemble	26 734	24 700	2 034

Les cibles visées sont respectées à l'issue du tirage de l'échantillon. Le nombre d'observations échantillonnées garantit systématiquement les cibles des conventions. En effet, Les faibles écarts négatifs (un nombre de répondant inférieur aux effectifs cibles) ne constituent pas un risque, les cibles de l'échantillon principal ayant été définies avec une légère sécurité.

D'autre part, on observe un léger écart sur la taille de la population totale suite aux corrections manuelles apportées (on passe en effet d'un effectif calé de 24700 à un effectif de 26734 observations après les ajustements manuels des probabilités de tirage).

Cet écart s'explique en bonne partie par la décision d'envoyer en production l'intégralité des diplômés DRJSCS. En effet, il s'agit d'une population difficile à enquêter par expérience. Les fichiers que nous avons reçus étant en outre de qualité moindre (absence de numéro de téléphone plus fréquente, moins d'informations identifiantes), nous nous attendons à des probabilités de réponses inférieures à celles observées sur l'enquête précédente. Cette contrainte aurait pu être intégrée initialement dans les marges de calage mais l'ajustement sur les effectifs du sport a été fait à l'issue du calage.

Le reste de la différence sur le total de la population se fait sur le complémentaire des indicatrices d'extension. Il s'agit ici en l'occurrence des mentions complémentaires, les formations à l'issue du CAP ou du Baccalauréat professionnel.

Par ailleurs plusieurs cibles sont ici indicatives. Il s'agit par exemple des cibles sur les sous populations des formations de santé social. En effet, celles-ci se rapprochent des cibles que les partenaires auraient souhaitées idéalement. Cependant une étude de faisabilité a montré que certains effectifs de sous populations ne pouvaient pas être atteints. La convention mentionne donc deux objectifs pour l'ensemble des formations de santé d'une part et l'ensemble des formations du social d'autre part. Les sous-populations qui ne pouvaient être atteintes ont été envoyées exhaustivement. L'échantillon devrait par ailleurs permettre d'atteindre les deux cibles générales.

Dans la dernière étape les individus sont affectés à l'échantillon principal. Les estimations du nombre de répondants sont données dans le tableau suivant :

Tableau 8 : écart entre échantillon principal et cible

Sous population d'extension	Effectif répondant à partir de l'échantillon principal	Cibles échantillon principal	Différence entre échantillon principal et les cibles
Collège	895	903	-8
Baccalauréat général et technologique	1 059	1 068	-9
Baccalauréat professionnel	2 403	2 391	12
CAP	1 372	1 373	-1
CGDD	3 163	3 080	83
CGDD niveau 1	753	765	-12
CGDD niveau 2	750	742	8
CGDD niveau 3	509	515	-6
CGDD niveau 4	744	680	64
CGDD niveau 5	407	378	29
DGESIP	6 665	6 387	278
DGESIP autre	1 840	1 781	59
DGESIP BTS	887	884	3
DGESIP Grandes écoles	730	701	29
DGESIP IUT	318	318	-1
DGESIP Licence	1 518	1 331	186
DGESIP Master	1 373	1 372	1
Quartier priorité de la ville	1 980	2 079	-99
Santé social	5 595	5 438	158
Santé social aide-soignante	1 071	1 073	-1
Santé social assistant social	359	318	41
Santé social auxiliaire puéricultrice	461	413	48
Santé social conseiller économie familiale	207	208	-1
Santé social éducateur jeunes	323	324	-1
Santé social éducateur spécialisé	488	431	57
Santé social ergothérapeute	62	62	0
Santé social infirmier	1 085	1 083	2
Santé social masseur kinésithérapeute	425	427	-2
Santé social moniteur éducateur	477	419	58
Santé social orthophoniste	77	78	0
Santé social orthoptiste	31	31	0
Santé social podologue pédicure	140	140	0
Santé social psychomotricien	84	84	0
Santé social sage femmes	304	306	-2
Sport	2 360	1 525	835
Thèse	1 863	1 500	363
Docteur en santé	100	100	0
Ensemble	24 570	23 000	1 570

La différence sur les docteurs (thèse) s'explique par le fait que la convention mentionne une cible de diplômés et non de sortants. Une variable indiquant l'obtention du diplôme existe dans la base de sondage pour certaines formations sans que nous ayons de certitude quant à la qualité de cette variable. De ce fait, le nombre d'observations tirées, qui concernent ici tous les sortants, est légèrement supérieur à la cible sachant que l'on estime une part de diplômés entre 70% et 80%.

4. Une méthode alternative plus classique pour gérer les extensions

Une autre méthode pour échantillonner les individus des extensions consiste à effectuer, un tirage en plusieurs étapes : un premier tirage pour définir l'échantillon socle S_{socle} , puis un sur-échantillon S_{ext} des extensions tiré dans le complémentaire de l'échantillon socle. Dans un premier temps, l'échantillon socle est tiré avec les probabilités d'inclusion π_{i1}. En notant $\{E^j\}_{j\leq M}$ l'ensemble des M extensions. Notons n^j la cible pour l'extension E^j et n^j_{socle} le nombre d'individus de l'extension E^j dans l'échantillon socle. Si l'échantillon socle est équilibré sur les effectifs des extensions, n^j_{socle} est fixe et est connu à l'avance : $n^j_{socle} = \sum_{i\in E^j} \pi_{i1}$

Il s'agit alors, pour chaque extension *j* indépendamment, de tirer un échantillon supplémentaire pour satisfaire les besoins des partenaires d'extension de taille $n^j_{supp} = n^j - n^j_{socle}$. Il faut alors déterminer des probabilités de tirage π^j_i pour cette étape qui respectent la cible du sur-échantillonnage n^j_{supp} , ainsi que la forme des probabilités de tirage corrigeant les défauts de couverture et anticipant les taux de non-réponse.

Pour cela, il faut imposer comme probabilité de tirage dans le complémentaire de S_{prin} pour chaque extension d'échantillon E^j,

$$\pi^j_i = \frac{\pi_{i1} \cdot n^j_{supp}}{\sum_{k\epsilon E^j/S_{socle}} \pi_{k1}}$$

De cette manière, les probabilités de tirage sont proportionnelles à celles du tirage de l'échantillon socle, et le nombre cible d'individus de l'extension est atteint.

π^j_i est la probabilité de chaque individu d'être tiré pour l'extension E^j. Pour un individu appartenant au croisement de deux extensions $E^j \cap E^k$, π^j_i et π^k_i sont potentiellement différents.

Pour chaque extension E^j un tirage équilibré garantit l'échantillonnage de n^j_{supp} individus de l'extension E^j avec probabilité π^j_i. A la fin du tirage de ces extensions, pour l'extension E^j, il y a n^j_{socle} individus échantillonnés issus du tirage socle, n^j_{supp} individus issus du tirage de l'extension E^j, mais également des individus issus des tirages des autres extensions qui sont à l'intersection de l'extension E^j et d'une ou plusieurs autres au nombre de n^j_{autre} .

En effectuant des estimations sur l'ensemble des individus échantillonnés appartenant à l'extension E^j, avec les poids inverses des probabilités de tirage, il y aurait surestimation, à la fois du nombre d'individus de cette extension, mais aussi de la proportion d'individus qui font partie des autres extensions.

Pour diminuer le poids des intersections par rapport au reste des individus, un partage des poids est effectué : un individu non tiré dans l'échantillon socle, appartenant à p extensions E^{j_1},E^{j_2}..., E^{j_p} peut être tiré dans l'extension E^{j_1} avec probabilité $\pi^{j_1}_i$, dans l'extension E^{j_2} avec probabilité $\pi^{j_2}_i$... dans l'extension E^{j_p} avec probabilité $\pi^{j_p}_i$, et a donc p façons d'être tiré. Pour que dans les cas où il est tiré plusieurs fois dans les différentes extensions, son poids ne soit pas trop élevé par rapport aux autres personnes appartenant à une unique extension, le partage des poids (Favre-Martinoz, Gros, 2017) va normaliser son poids. Au lieu d'être l'inverse de la probabilité d'inclusion, celui-ci va devenir

$$W_{i,ext} = \frac{1}{p} \cdot \sum_{k=1}^{p} \frac{L^{j_k}_i}{\pi^{j_k}_i}$$

Avec $L^{j_k}_i$ l'indicatrice que l'individu i est tiré pour l'extension E^{j_k}

À ce stade, plus de n^j individus ont été échantillonnés pour chaque extension E^j . Cependant, il y a des poids pour les individus tirés pour l'échantillon socle, et d'autres pour ceux venant des extensions. $W_{i,ext}$ ne prend pas en compte la probabilité d'être tiré pour l'échantillon socle, et $W_{i,socle} = \frac{1}{\pi_{i1}}$, qui est le poids des individus échantillonnés pour l'échantillon socle, ne prend pas en compte la probabilité d'être tiré pour les extensions. Le jeu de pondération $W_{i,socle}$permet de réaliser des estimations sur l'ensemble de la population. Cependant, la non-prise en compte des individus échantillonnés pour les extensions pour les estimations globales est dommageable puisque ces extensions représentent une partie importante des questionnaires récoltés sur les sous-populations d'extensions.

Pour pouvoir inclure ces individus afin réaliser des estimations de la population sur l'ensemble des individus échantillonnés, il faut réaliser une étape supplémentaire pour corriger les poids, afin que la somme des $W_{i,socle}$ et des $W_{i,ext}$ soit égale au total connu de l'extension, par calage par exemple.

Même si le problème de sur-pondération des intersections des extensions a été résolu, il y a toujours plus de questionnaires que nécessaire dans les extensions (les n^j_{autre} individus pour l'extension E^j) ; l'objectif principal de réduction du nombre de questionnaires n'est donc pas atteint.

Conclusion

En conclusion, quelle que soit la méthodologie retenue, la multiplicité des demandes d'extensions des partenaires entraine des complexifications techniques.
Il semble que la méthode de tirage en deux étapes soit plus naturelle dans un premier temps puisqu'elle permet d'identifier clairement les échantillons et leur fonction. Une conséquence pratique qui en découle, dans le cadre d'une enquête par panel, où les besoins des partenaires ne sont pas nécessairement reconduits d'une interrogation à une autre, est de pouvoir retenir ou supprimer simplement les extensions d'échantillon pour une ré-interrogation.
Cependant le risque d'avoir des effectifs interrogés plus importants dans les intersections n'est pas pris en compte dans cette approche en deux temps. Le partage de poids est alors nécessaire afin de ne pas avoir de biais dû à la sur-représentation des intersections d'extensions en diminuant la pondération des individus liés à plusieurs extensions.
Bien que cette approche donne satisfaction dans le fait de mener à bien les différentes contraintes imposées à cet échantillon, Il semble que le contexte idéal d'application de la méthode de partage des poids se rencontre davantage lorsque les bases de tirage sont réellement séparées. Dans notre cas en effet, le fait d'avoir une base unifiée, permet de mettre en œuvre des optimisations qui ne peuvent pas se poser dans la situation où les bases sont distinctes.
Le tirage en une étape proposé dans cet article permet quant à lui d'inclure la population échantillonnée pour les extensions aux estimations de la population globale facilement, puisque les jeux de poids sont donnés directement : ce sont les inverses des probabilités d'inclusions calculées par le calage. L'avantage principal de cette méthode est qu'elle permet d'obtenir le nombre exact de questionnaires souhaités pour chaque extension, sans pour autant sur-échantillonner les intersections entre extensions.

Annexe 1 : Le calcul des probabilités individuelles de tirage (phase A)

Le taux de couverture (étape A1)

L'échantillon Génération est tiré dans une base de sondage construite par le Céreq. Le tableau suivant donne par grands types d'établissement :

- les effectifs présents dans la base de sondage (après opération de dédoublonnage)
- le taux de couverture moyen (en %)
- le poids de couverture moyen
- l'effectif estimé de la base de sondage en l'absence de défaut de couverture

Pour un type d'établissement donné, le taux de couverture est estimé en utilisant la meilleure information externe disponible (par ordre de priorité, nombre de sortants nationaux, nombre de diplômés, nombre d'élèves inscrits). En l'absence d'informations externes, le taux de couverture est estimé par le taux de réponse des établissements pour un type donné.

Pour certains types d'établissement, plusieurs sources ou données détaillées (notamment sur des effectifs par formation) peuvent être mobilisées. Cela explique, qu'au sein d'un même type d'établissement, il y ait plusieurs taux de couverture.

Par exemple, pour le type d'établissement « universités », on calcule le taux de couverture en rapportant les effectifs de sortants présents dans la base de sondage aux effectifs de sortants présents dans la base ministérielle SISE (Système d'information sur le suivi de l'étudiant). Ce calcul s'effectue au niveau du croisement des variables Région, Niveau de formation et disciplines de formation. Ainsi, pour la région « Midi-Pyrénées », le niveau de formation « M2 » et la discipline « sciences humaines et sociales », la base de sondage, issue de la collecte auprès des universités, contient 765 individus présumés sortants. Dans la source SISE, il y a un effectif de 772 sortants. Ainsi le taux de couverture, à ce niveau de croisement, est de 765/772=99 %.

On attribue ensuite à chaque individu présent dans la base de sondage, le poids de couverture qui le concerne. En reprenant l'exemple, le poids de couverture des individus concernés (universités de Midi-Pyrénées, M2, SHS) est donc de 1/0,99=1,01.

Pour les notations suivantes, le poids de couverture individuel sera noté $\boldsymbol{pcouv_i}$

Tableau 9 : la sous-couverture de la Base de Sondage par grands types d'établissements

Type d'établissement	Effectif base de sondage	Taux de couverture moyen (en %)**	Poids de couverture moyen	Effectif théorique corrigé du défaut de couverture*
Lycées et collèges MEN	548 922	100	1,00	548 922
Universités	332 514	90	1,11	367 745
Centre de Formation des Apprentis	177 490	100	1,00	177 490
Lycées agricoles	55 742	100	1,00	55 742
Écoles professions sociales	47 106	49	2,04	96 171
Écoles de commerce	28 716	39	2,55	73 357
Autre universités	18 939	33	3,17	56 777
Écoles ingénieurs	18 338	48	2,10	38 578
Écoles ministère de la culture	12 036	37	2,68	32 197
DRJS	11 697	70	1,43	16 785
Classe préparatoire et autres étab.	4 902	25	4,08	19 949
Écoles secteur service	4 664	27	3,64	16 962
DGAFP	3 644	41	2,44	8 903
Écoles secteur industriel	2 154	69	1,44	3 111
IEP	1 847	85	1,18	2 177
CIFRE	943	95	1,05	994
Écoles formations agricoles	756	46	2,19	1 646
Centres privés d'enseignement	693	55	1,82	1 263
Écoles administrations publiques	467	39	2,56	1 197
Écoles Normales Supérieures	437	36	2,78	1 214
Ensemble	1 272 007	84	1,21	1 521 180

** l'effectif théorique corrigé du défaut de couverture est égale à la somme des poids de couverture individuels.*
***Le taux de couverture moyen est obtenu en faisant le ratio « effectif base de sondage » sur « effectif théorique corrigé du poids de couverture ». Cet indicateur a été préféré à la moyenne des taux de couverture individuels. Les écarts entre les deux indicateurs sont faibles.*

Le taux de couverture de l'ensemble de la base de sondage est ainsi estimé à 84 %. La qualité de la couverture n'est pas uniforme selon l'origine des données : La couverture est quasiment exhaustive lorsque les fichiers sont centralisés[2]. A l'inverse, les taux de couverture sont moins bons pour les types d'établissements concernés par la collecte auprès des établissements. Dans l'ensemble, près de 90% des individus de la base de sondage sont issus d'une réception de fichiers centralisés ce qui représente environ les trois quarts des individus de notre champ.

[2] BEA pour les bases rectorales lorsqu'il s'agit des collèges et des lycées. Fichier Agri pour les lycées agricole. SISE pour les Universités. SIFA pour les apprentis. Pour ces différentes sources la couverture est quasi exhaustive.

Détermination des probabilités de tirage en l'absence d'extension (étape A2)

Cette étape consiste à déterminer les probabilités de tirage qui simulent un échantillon national donnant lieu à 8500 répondants dans le champ Céreq[3]. Pour ce calcul, il est nécessaire de faire intervenir le poids de couverture, la probabilité de réponse anticipée ainsi que la probabilité d'appartenir au champ de l'enquête.

En effet, la deuxième difficulté liée à la base de sondage provient de la présence (importante) d'individus hors champ. Il s'agit essentiellement de poursuivants l'année scolaire suivante (non repérés au préalable) mais également d'individus qui ont déjà interrompu leur scolarité dans le passé (post-initiaux). Sur l'ensemble de la base de sondage, le taux de hors champ est proche de 50% (avec une forte variabilité qui dépend essentiellement du type d'établissement). L'effectif théorique de la base de sondage corrigé du défaut de couverture étant de 1 500 000 individus, la population théorique du champ Céreq est approximativement de l'ordre de 750 000 individus.

Définitions - Propriétés

Probabilité de réponse anticipée

Les probabilités de réponses anticipées individuelles sont calculées à l'aide de modèles logistiques sur les données de l'enquête précédente, à savoir l'enquête Génération 2010 à 3 ans. Les modèles font intervenir les variables explicatives : Age, appartenance à une Zus, région de l'établissement de formation, indicatrice valant 1 s'il s'agit d'un niveau terminal (classe permettant d'obtenir un diplôme à la fin de l'année scolaire), présence d'au moins un numéro de téléphone, et strate de sortie. Ici, nous utilisons (historiquement) de manière inappropriée le terme de « strate » pour désigner une variable agrégeant des classes de sortie sans qu'elle constitue une variable de stratification au sens propre pour le tirage de l'échantillon. Cette variable a été calculée à partir d'une classification des classes de sortie concernant leur homogénéité sur les taux de réponse, taux de hors champ. Cette variable « strate » est donc utilisée comme variable explicative du taux de réponse avec les autres variables citées.

Par la suite, la probabilité de réponse anticipée d'un individu i, est notée :

$$pr_i = P(i\ accepte\ de\ répondre \mid i \in s)$$

Probabilité de réponse dans le champ anticipé

La présence d'hors champ dans la base de sondage nous impose d'introduire, en plus de la probabilité de réponse, la notion de probabilité de répondre dans le champ. En effet, globalement, un individu sur deux de la base de sondage ne fait pas partie du champ de l'enquête. Ce taux d'appartenance au champ Céreq, varie sensiblement selon les classes de sortie. Nous définissons donc une probabilité de répondre dans le champ qui est l'évènement « a répondu et est dans le champ ». Cet évènement est modélisé à partir de l'enquête précédente et avec les mêmes variables explicatives (issues uniquement de la base de sondage) que précédemment.

La probabilité de réponse dans le champ d'un individu i, est notée :

$$pr_i^C = P(\{i\ accepte\ de\ répondre\ \cap i\ dans\ le\ Champ\} \mid i \in s)$$

[3] Le Céreq finance pour ses besoins propres 10 000 questionnaires afin d'obtenir un échantillon national « représentatif » de 8 500 répondants et une extension d'échantillon ciblée sur certains niveaux de sortie de 1 500 répondants.

Estimation de la taille de l'échantillon à partir des probabilités de tirage

Soient

- U la base de sondage,
- s un échantillon
- $\pi_i = P(i \in s)$ la probabilité que l'échantillon s contienne l'individu i (probabilité de tirage de *i*)

L'estimation de la taille de l'échantillon, n, est donnée par :

$$\sum_{i \in U} \pi_i = E(n)$$

La somme des probabilités de tirage sur toute la base de sondage U est égale à l'espérance de la taille de l'échantillon.

Dans le cadre d'un sondage équilibré sur les probabilités d'inclusion, on obtient une égalité entre la taille de l'échantillon et la somme des probabilités de tirage sur U si le tirage est parfaitement équilibré (phase de vol uniquement).

$$\sum_{i \in U} \pi_i = n$$

Estimation du nombre de répondants dans le champ

L'estimation du nombre de répondants dans le champ, nr^C, est donnée par :

$$\sum_{i \in U} P(\{i \, répond \cap i \, dans \, le \, Champ\} \mid i \in s) * \pi_i = \mathrm{E}(\text{nombre de répondants dans le champ})$$

Soit,

$$\sum_{i \in U} pr_i^C * \pi_i = \mathrm{E}(\mathrm{nr}^C)$$

On calcule ainsi l'espérance du nombre de répondants dans le champ en faisant cette somme sur l'intégralité de la base de sondage U.

Annexe 2 : La structure de la population théorique respectée

Dans cette partie, on montre qu'avec la forme de la probabilité de tirage la structure des répondants est proche de la structure de notre population d'intérêt. On repart de l'expression donnée en début de chapitre :

$$\pi_{i1} = \frac{\text{pcouv}_i}{\text{pr}_i} * \text{coeff1}$$

En rappelant les notations :

Le poids de couverture de i est noté pcouv_i

La probabilité de répondre de i est notée pr_i

La probabilité de i de répondre dans le champ sachant qu'il est tiré dans l'échantillon est notée pr_i^C

La probabilité d'avoir un répondant dans le champ est donc égale à l'expression suivante :

$$pr_i^C * \pi_{i1} = pr_i^C * \frac{\text{pcouv}_i}{\text{pr}_i} * coeff1$$

On peut approximer que la probabilité de répondre dans le champ est le produit de la probabilité de répondre par le taux d'individus dans le champ pour une strate considérée. C'est-à-dire que pour une strate donnée les comportements de réponse des individus dans le champ et hors champ sont identiques. Ce qui revient à écrire :

$$pr_i^C = \text{pr}_i * \text{taux\_dans\_champ}_i$$

Donc on obtient $pr_i^C * \pi_{i1} = \text{taux\_dans\_champ}_i * \text{pcouv}_i * \text{coeff1}$

C'est-à-dire que la probabilité d'avoir un répondant dans le champ est le taux d'individus dans le champ de sa strate multiplié par la correction de défaut de couverture correspond au poids associé et multiplié par le coefficient de dilatation identique à tous les individus. On a donc une probabilité d'avoir un individu dans le champ corrigée des différences de comportements de réponses et des différences de couverture.

Comme dit précédemment, la structure de la population obtenue est normalement proche de celle de la population théorique sur les variables qui ont été mobilisées pour le calcul des probabilités de réponse.

Annexe 3 : codage des extensions

Tableau 10 : effectifs cibles par sous-populations d'extensions

Sous population d'extension	codage des sous-populations
Collège	c1
Baccalauréat général et technologique	c2
Baccalauréat professionnel	c3
CAP	c4
CGDD	c5
CGDD niveau 1	c6
CGDD niveau 2	c7
CGDD niveau 3	c8
CGDD niveau 4	c9
CGDD niveau 5	c10
DGESIP	c11
DGESIP autre	c12
DGESIP BTS	c13
DGESIP Grandes écoles	c14
DGESIP IUT	c15
DGESIP Licence	c16
DGESIP Master	c17
Quartier priorité de la ville	c18
Santé social	c19
Santé social aide-soignante	c20
Santé social assistant social	c21
Santé social auxiliaire puéricultrice	c22
Santé social conseiller économie familiale	c23
Santé social éducateur jeunes	c24
Santé social éducateur spécialisé	c25
Santé social ergothérapeute	c26
Santé social infirmier	c27
Santé social masseur kinésithérapeute	c28
Santé social moniteur éducateur	c29
Santé social orthophoniste	c30
Santé social orthoptiste	c31
Santé social podologue pédicure	c32
Santé social psychomotricien	c33
Santé social sage femmes	c34
Sport	c35
Thèse	c36
Docteur en santé	c37
Ensemble	c38

Annexe 4 : Le tirage équilibré de l'échantillon (phase B)

Tirage de l'échantillon global (principal + réserve) (étape B1)

Trente contraintes d'équilibrage, toutes construites de la même manière, ont été utilisées pour contrôler les effectifs envoyés en production des sous-populations les plus sensibles (il s'agit essentiellement de contraintes d'équilibrage pour garantir les objectifs de questionnaires pour les partenaires d'extension). Les sous populations sur lesquelles des contraintes d'équilibrage ont portées sont les suivantes :

Sous population faisant l'objet d'une contrainte d'équilibrage	
Baccalauréat professionnel	Santé social conseiller économie familiale
CAP	Santé social éducateur jeunes
CGDD niveau 1	Santé social éducateur spécialisé
CGDD niveau 2	Santé social ergothérapeute
CGDD niveau 3	Santé social infirmier
CGDD niveau 4	Santé social masseur kinésithérapeute
CGDD niveau 5	Santé social moniteur éducateur
DGESIP BTS	Santé social orthophoniste
DGESIP Grandes écoles	Santé social orthoptiste
DGESIP IUT	Santé social podologue pédicure
DGESIP Licence	Santé social psychologue
DGESIP Master	Santé social sage femmes
Santé social aide-soignante	Sport
Santé social assistant social	Thèse
Santé social auxiliaire puéricultrice	

On équilibre sur la variable probabilité de tirage pour chaque sous population *j*.

Les contraintes d'équilibrage sont ainsi données par :

$$\sum_{i \in S} \frac{1_{i \in J} * \pi_{i2}}{\pi_{i2}} = \sum_{i \in U} 1_{i \in J} * \pi_{i2}$$

Soit

$$\sum_{i \in S} 1_{i \in J} = \sum_{i \in U} 1_{i \in J} * \pi_{i2}$$

Soit

$$n_{j=} \sum_{i \in U} 1_{i \in J} * \pi_{i2}$$

Où n_j est la taille de la sous population J dans l'échantillon et $1_{i \in J}$ est l'indicatrice d'appartenance à l'extension *J*. En équilibrant sur les probabilités de tirage correspondant à une extension particulière (délimitée par une indicatrice), on a un sous échantillon de taille fixe pour la sous population J. on s'assure donc d'envoyer en production un nombre de personnes à enquêter adapté au nombre de répondants que l'on souhaite sur cette sous-population.

Ici on utilise la loi des grands nombres, bien que les probabilités d'obtenir un questionnaire varient selon les individus, en moyenne le nombre de questionnaire espéré est assez bien connu. Les probabilités de tirage ont été déterminées pour définir une taille d'échantillon adaptée pour atteindre une cible donnée de questionnaire pour la sous population J. Ces probabilités d'inclusion sont respectées lors de la phase de vol du tirage équilibré[4] . En équilibrant sur les probabilités d'inclusion, on fixe une taille de sous échantillon adéquate pour atteindre les objectifs sur la sous population *J*.

Un échantillon global, contenant la réserve et l'échantillon principal, a été tiré de manière équilibrée avec le package sampling de R.

Tirage de l'échantillon principal (étape B2)

Une contrainte d'équilibrage a été utilisée pour affecter les individus soit à l'échantillon principal soit à la réserve à partir de l'échantillon global. Le coefficient $coeffprin_i$ correspond à la probabilité que l'individu *i* échantillonné se retrouve dans l'échantillon principal. Cette probabilité est de l'ordre de 0.9 pour les sous populations disposant d'une réserve, et elle est égale à 1 si la sous-population est envoyée intégralement en production dans l'échantillon principal (c'est-à-dire que l'individu est automatiquement affecté à l'échantillon principal et qu'il n'y a pas de réserve pour cette sous-population). Cette variation du coefficient $coeffprin_i$ est due aux fluctuations des nombres de questionnaires anticipés suite aux corrections manuelles nécessaires après la phase de calage. La contrainte d'équilibrage s'écrit de la manière suivante :

$$\sum_{i \in S\ prin} \frac{coeffprin_i}{coeffprin_i} = \sum_{i \in S} coeffprin_i$$

Soit

$$\sum_{i \in S\ prin} 1 = \sum_{i \in S} coeffprin_i$$

Soit

$$n_{principal=} \sum_{i \in S} coeffprin_i$$

Où $Sprin$ désigne l'échantillon principal et S l'échantillon global.

Cette contrainte assure que le nombre d'observations comprises dans l'échantillon principal respecte globalement les probabilités d'affectation. Les probabilités d'inclusion finales dans l'échantillon principal sont donc :

$$\pi_{i3} = \pi_{i2} * coeffprin_i$$

Les individus non sélectionnés sont dans la réserve.

Pour l'enquête Génération 2013 à 3 ans, la réserve n'a pas été mobilisée, ce qui était déjà le cas pour la précédente Génération. Si toutefois celle-ci aurait dû être mobilisée, nous aurions procédé de la sorte : En cours de production, si les projections du nombre de répondants d'une sous-population montrent que la cible ne sera pas atteinte avant la fin du plateau, alors toute la réserve de cette sous-population particulière serait débloquée.

[4] Le tirage de l'échantillon est quasiment parfaitement équilibré. A l'issue de la phase de vol, l'algorithme a du statuer sur 22 observations (sur 170 000 au regard de l'échantillon global). Le calcul d'un échantillon de 22 observations qui s'éloigne le moins possible des contraintes d'équilibrage était trop gourmand en temps de calcul. L'option qui a été choisie pour la phase d'atterrissage a donc été de supprimer les contraintes (method= 2 dans la fonction samplecube du package sampling).

Bibliographie

[1] Deville J.C., Tillé Y., « Efficient balanced sampling: the cube method». *Biometrika*, vol 91, n°4, pp 893-912, 2004.
[2] Deville J.-C., Särndal C.-E., Sautory O., « Generalized raking procedures in survey sampling ». *Journal of the American Statistical Association*, vol 88, n°423, pp. 1013-1020, 1993.
[3] Favre-Martinoz C., Gros E., « La méthode du partage des poids ». Insee, Département des méthodes statistiques, Version n°1, 2017