Construction d'échantillons astreints à des conditions de recouvrement par rapport à un échantillon antérieur et à des conditions d'équilibrage par rapport à des variables courantes :

aspects théoriques et mise en œuvre dans le cadre du renouvellement des échantillons des enquêtes d'évaluation des élèves

Marc CHRISTINE[1], Thierry ROCHER[2]

***Document de travail, version provisoire ; ne pas citer*.**

[1] Insee, DCSRI
[2] DEPP

1 Résumé.

L'objet de ce papier est de fournir un cadre théorique et des méthodes de résolution d'un problème d'échantillonnage qui intervient dès lors qu'on a construit un 1er échantillon d'enquête à une certaine date et que, postérieurement, on souhaite tirer un 2nd échantillon de caractéristiques fixées (taille, probabilités d'inclusion, conditions d'équilibrage) et présentant des conditions d'articulation avec le 1er (en un sens précisé ci-dessous), mais sans pouvoir agir sur le tirage déjà effectué du 1er échantillon.

A l'origine de ce problème repose la problématique des enquêtes PISA[3] : différentes vagues d'enquête portent sur une même matière dominante (exemple : mathématiques en 2003 et 2012). Pour faire des comparaisons pertinentes entre les deux périodes, il est préférable d'assurer un recouvrement entre les échantillons d'écoles aux deux dates. Mais il est aussi important d'avoir la meilleure « représentativité» de l'échantillon 2012 par rapport aux caractéristiques actuelles de l'univers, ce que l'on traduira par des conditions d'équilibrage.

Plus généralement, les cas les plus usuels relevant de cette approche sont : les échantillons en deux phases (2ème échantillon inclus dans le 1er), les échantillons avec disjonction lors des tirages successifs (2ème échantillon disjoint du 1er), les échantillons avec recouvrement de l'un par rapport à l'autre, les échantillons où l'on impose des conditions de « représentativité » lorsqu'on travaille soit sur la réunion des deux, soit sur le second seulement…

Le cadre proposé de réflexion et de résolution peut s'appliquer à tous les cas où les unités sont tirées *à probabilités inégales* (souvent conditionnées par un facteur de taille) : enquêtes entreprises, tirage d'unités primaires géographiques dans une enquête ménages, tirage d'établissements scolaires pour les enquêtes d'évaluation des élèves…

L'approche théorique repose sur la notion d'**échantillonnages séquentiels conditionnels** et sur la technique **des échantillons équilibrés**. On montrera dans ce papier que l'ensemble des contraintes astreintes au 2ème échantillon sont en général incompatibles, ce qui conduira à **rechercher des solutions approchées**, obtenues en relâchant la contrainte relative aux probabilités d'inclusion finales du 2ème échantillon. Ces solutions approchées posent des problèmes de calcul explicite des solutions (en général seulement possible de manière numérique) et de propriétés statistiques des estimateurs en dérivant.

[3] Programme for international student assessment

Une mise en œuvre de ces méthodes a été ensuite expérimentée à partir des bases d'établissements scolaires françaises. L'objectif est de simuler le tirage d'un échantillon dans la base 2009 avec des conditions de recouvrement par rapport à un échantillon tiré dans la base 2000 et des conditions d'équilibrage par rapport à l'environnement de 2009. Pour simplifier, les bases ont été apurées de façon que les univers qu'elles représentent soient rigoureusement identiques (les établissements nouveaux ou disparus sont mis hors base ; en revanche, un même établissement peut avoir des caractéristiques différentes entre les deux années de référence, voire appartenir à des strates différentes).

Dans un premier temps, on simule des tirages d'échantillons qui représenteront l'échantillon tiré en 2000. Pour chacun de ces échantillons, on calcule les probabilités approchées d'inclusion de l'échantillon 2009 rendant le problème soluble. On étudie empiriquement les propriétés de ces nouvelles probabilités et, notamment, leur écart par rapport aux valeurs des probabilités de référence que l'on souhaitait imposer a priori à ce 2nd échantillon.

Des variantes seront étudiées, selon les modalités de tirage du 1er échantillon (avec ou sans équilibrage), selon le type de distance utilisée pour définir la proximité entre les probabilités d'inclusion de référence du 2nd échantillon et les probabilités approchées et, enfin, selon les valeurs du taux de recouvrement entre le 1er et le 2nd échantillon.

Dans un second temps, pour chaque tirage du 1er échantillon, on simulera des tirages du 2nd échantillon avec des probabilités conditionnelles de tirages adaptées en fonction du calcul des nouvelles probabilités finales. Des estimateurs d'un certain nombre de totaux de variables d'intérêt seront calculés en utilisant ces nouvelles probabilités d'inclusion. On étudiera empiriquement le biais et la précision de ces estimateurs, en comparant leurs résultats aux vraies valeurs connues dans la base de sondage.

2 Introduction.

L'objet de ce papier est de fournir un cadre théorique et des méthodes de résolution d'un problème d'échantillonnage qui intervient dès lors qu'on a construit un 1er échantillon d'enquête à une certaine date et que, postérieurement, on souhaite tirer un 2nd échantillon présentant des conditions d'articulation avec le 1er (en un sens qui sera précisé dans le corps du papier), mais sans pouvoir agir sur le tirage déjà effectué du 1er échantillon.

Les cas les plus usuels relevant de cette approche sont : les échantillons en deux phases, les échantillons avec disjonction lors des tirages successifs, les échantillons avec recouvrement de l'un par rapport à l'autre, les échantillons où l'on impose des conditions de « représentativité » lorsqu'on travaille soit sur la réunion des deux, soit sur le second seulement, mais avec l'une des contraintes citées précédemment.

Ce problème prend tout son intérêt lorsque les unités sont tirées *à probabilités inégales*. De ce fait, le cadre proposé de réflexion et de résolution peut s'appliquer aux enquêtes entreprises, au tirage d'unités primaires géographiques dans une enquête ménages ou entreprises, au tirage d'établissements scolaires comme c'est le cas pour l'enquête PISA, toutes enquêtes dont les unités statistiques ont en général un facteur de taille qui conditionne leurs probabilités de tirage.

On verra que les outils principaux de cette approche sont d'une part la notion d'**échantillonnages séquentiels conditionnels**, d'autre part la **technique des échantillons équilibrés**. C'est d'ailleurs lorsque l'on cherche à imposer des conditions d'équilibrage ou lorsque les contraintes peuvent se traduire en termes d'équilibrage que le problème devient compliqué. On verra de surcroît que **seules des solutions approchées peuvent être obtenues**, non seulement en termes statistiques mais aussi en termes de calcul explicite des solutions, seulement possible de manière numérique.

3 Cadre général et notations.

Univers = population de référence = base de sondage : U .
Individus (unités statistiques) : notés i .
La taille de l'univers (nombre d'individus) est notée N .
Variables d'intérêt (**non aléatoires**) définies sur chaque individu i , notées génériquement Y_i , de total $T(Y)$ dans l'univers.

Un premier échantillon S_1 (**sans remise**) a été tiré dans U , caractérisé par :

- sa taille (éventuellement aléatoire) : $n_1 = n(S_1)$
- des probabilités d'inclusion des individus : $\boxed{\mathrm{P}\{i \in S_1\} = \pi_i^1 \in [0, 1]}$.

  Ces probabilités sont définies ex-ante, avant tout tirage, pour chacun des individus i .

  Si l'échantillon est de taille fixe n_1 , on a la relation : $\sum_{i \in U} \pi_i^1 = n_1$.

  Le cas où : $\pi_i^1 = 0$ peut correspondre à une restriction du champ de l'enquête (on tire l'échantillon seulement dans la partie de l'univers où : $\pi_i^1 > 0$).

  Les unités pour lesquelles : $\pi_i^1 = 1$ sont retenues d'office, elles forment une partie de l'univers baptisée *strate exhaustive*. Il s'agit en général de « grosses » unités.

- d'éventuelles conditions d'équilibrage sur des variables notées génériquement X_i, éventuellement vectorielles, qui s'écrivent : $\boxed{\sum_{i \in S_1} \frac{X_i}{\pi_i^1} = T(X)}$.

  L'équilibrage permet d'obtenir un échantillon « représentatif » vis-à-vis des variables X, c'est-à-dire réalisant un modèle réduit de l'univers, conforme à celui-ci relativement aux totaux de ces variables : **l'estimation de ces totaux à partir de l'échantillon fournit une valeur numérique égale aux vraies valeurs des totaux dans l'univers**.

  On se restreint ici à des estimateurs à coefficients fixes, non aléatoires (estimateur de HORVITZ-THOMSON).

  Dans l'équation d'équilibrage, le terme de droite représente à la fois :

  - la *somme* $\sum_{i \in U} X_i$
  - l'*espérance* $E\left(\sum_{i \in S_1} \frac{X_i}{\pi_i^1} \right)$.

  Exemples :

  - l'équilibrage sur π_i^1 équivaut à imposer un échantillonnage de taille fixe.
  - l'équilibrage sur la constante 1 équivaut à imposer la condition naturelle :

    somme des poids de sondage = taille de l'univers.

  - avec des probabilités proportionnelles à une variable de taille T_i, on a naturellement un équilibrage vis-à-vis du total de cette variable.
  - l'équilibrage sur $\pi_i^1 \mathbf{1}_{i \in U_h}$ équivaut à imposer une stratification avec une taille d'échantillon fixée dans une sous-population U_h.
  - l'équilibrage est linéaire : s'il y a équilibrage sur X_i, il y a alors équilibrage sur λX_i. S'il y a équilibrage à la fois sur X_i et sur Y_i, il y a alors équilibrage sur $X_i + Y_i$.

  Pour que l'équilibrage soit efficace, les variables d'équilibrage doivent être corrélées aux variables d'intérêt observées dans les enquêtes. Elles peuvent être *qualitatives*, l'équilibrage équivalant alors à une stratification, ou *quantitatives* (âge, revenu, chiffre d'affaires, nombre de salariés ou d'élèves dans un établissement …).

Equilibrage par rapport à une moyenne

L'équation $\sum_{i\in S_1} \frac{X_i}{\pi_i^1} = T(X)$ peut aussi s'écrire : $\frac{1}{N}\sum_{i\in S_1} \frac{X_i}{\pi_i^1} = \frac{T(X)}{N} = \overline{X}$. Elle exprime alors que l'estimateur de HORVITZ-THOMSON de la moyenne de X prend exactement la valeur de la vraie moyenne $\overline{X}$ dans l'univers.

On peut aussi utiliser l'estimateur de HAJEK : $\frac{\sum_{i\in S_1} \frac{X_i}{\pi_i^1}}{\sum_{i\in S_1} \frac{1}{\pi_i^1}}$. Dans ce cas, cet estimateur prendra une valeur égale à la vraie moyenne $\overline{X}$ dès lors que : $\begin{cases} \sum_{i\in S_1} \frac{X_i}{\pi_i^1} = T(X) \\ \sum_{i\in S_1} \frac{1}{\pi_i^1} = N \end{cases}$, conditions exprimant l'équilibrage à la fois sur le total de X et sur celui de la constante 1 (cf. interprétation, supra).

4 Problématique.

L'échantillon S_1 a été tiré une fois pour toutes dans le passé et ses caractéristiques ne peuvent plus être modifiées. On dispose d'une réalisation de cet échantillon, soit s_1.

On cherche alors à réaliser le tirage d'un second échantillon, S_2. Si ce second échantillon ne présente aucune articulation d'aucune sorte avec le 1er, le plus simple est alors de le tirer de manière indépendante du 1er. Ce cas est sans intérêt vis-à-vis de la problématique que l'on va développer.

Si, au contraire, le 2nd échantillon présente une articulation par rapport au 1er, quelle qu'elle soit (conditions de recouvrement, de disjonction, utilisation conjointe des deux échantillons, équilibrage global …), alors le tirage du 2nd échantillon doit tenir compte des résultats du tirage du 1er. Dans ces cas plus complexes que l'on va étudier ici, **ce 2nd tirage doit donc se comprendre comme conditionnel au 1er tirage (c'est-à-dire conditionnel à la réalisation de celui-ci).**

Plus concrètement, les circonstances où la présente approche va s'appliquer correspondent à celles où l'une des conditions suivantes est requise :

- la base de sondage de S_2 dépend de S_1
- on cherche à mettre en œuvre des conditions d'équilibrage impliquant simultanément S_1 et S_2
- on cherche à assurer un certain recouvrement par rapport à S_1 .

Champs d'application :

- échantillon additionnel, complémentaire : échantillon de réserve, extension régionale, augmentation de la taille d'un échantillon tiré antérieurement

- échantillon d'actualisation, panélisation.

4.1 Notations

Le tirage du 2nd échantillon conditionnellement au 1er (et **sans remise**) est régi par les paramètres suivants :

- l'univers dans lequel on tire peut dépendre de S_1 : on le notera $U(S_1)$, de taille $N_2(S_1)$. $U(S_1)$ représente la base de sondage de S_2.
- sa taille, également potentiellement dépendante de S_1 : $n_2(S_1)$.
- des probabilités d'***inclusion conditionnelle*** des individus : $\boxed{\text{P}\{i \in S_2 / S_1\} = \pi_i^{2/S_1} \in [0, 1]}$.

  Les unités pour lesquelles : $\pi_i^{2/S_1} = 1$ sont retenues d'office dans S_2, elles forment une partie de l'univers baptisée *strate exhaustive*.

  Nota : distinguer P $\{i \in S_2 / S_1\}$ et P $\{i \in S_2 / i \in S_1\}$, cf. infra.

  Il faut noter qu'imposer que l'échantillon S_2 soit de taille fixe n_2, ***non aléatoire***, conduit à introduire cette condition lors du tirage conditionnel du 2nd échantillon par rapport au 1er et implique que les probabilités d'inclusion conditionnelle vérifient la relation :

$$\boxed{\sum_{i \in U(S_1)} \pi_i^{2/S_1} = n_2}.$$

- d'éventuelles conditions d'équilibrage sur des variables notées génériquement Z_i, s'écrivant donc :

$$\boxed{\sum_{i \in S_2} \frac{Z_i}{\pi_i^{2/S_1}} = \sum_{i \in U(S_1)} Z_i}. \textbf{ (1)}$$

  Cette condition (qui peut être réalisée sur le plan algorithmique par la méthode du CUBE[4]) correspond bien au tirage d'un échantillon S_2 dans un univers $U(S_1)$, où S_1 est supposé fixé. L'algorithme fonctionne indépendamment du caractère aléatoire de S_1, pourvu que S_1 soit fixé et parfaitement déterminé au moment du tirage de S_2.

4.2 Discussion sur la signification de $U(S_1)$.

On peut distinguer plusieurs cas :

- cas $U(S_1) = U$: pas de restriction du champ du tirage de S_2.
- cas $U(S_1) = S_1$: cas du tirage en *deux phases*.
- cas $U(S_1) = \mathcal{C}S_1$: problème de la *disjonction* : échantillon S_2 tiré dans le complémentaire de S_1, coordination négative.

[4] Inventée et développée par Jean-Claude DEVILLE et Yves TILLÉ.

- cas $U(S_1) = U_0$: tirage dans une sous-population définie de manière exogène, non aléatoire : domaine, région géographique, strate exhaustive ou non-exhaustive…
- cas $U(S_1) = S_1 \cap U_0$: tirage en deux phases mais seulement dans une sous-population U_0.
- cas $U(S_1) = \complement S_1 \cap U_0$: tirage d'un échantillon disjoint du 1er, mais ne couvrant qu'une sous-population U_0.
- des cas plus compliqués peuvent survenir : par exemple, tirage à l'intérieur ou à l'extérieur du sous-échantillon constitué des *k* plus grandes valeurs d'une variable d'intérêt observée sur S_1, ou parmi les unités (ou leur complémentaire) dont les valeurs observées de cette variable sont supérieures (ou inférieures) à un certain seuil … Autres exemples ?

Remarque :

On suppose implicitement ici que l'univers de référence U ne change pas entre les tirages des deux échantillons. Cette hypothèse peut être infirmée si l'univers est soumis à une dynamique temporelle, avec disparition ou apparition d'unités.

Toutefois, on peut traiter le cas des disparitions en se restreignant à une partie de l'univers (les « vivants ») dans laquelle le 2ème échantillon sera tiré et introduire un univers complémentaire $U_{N,2}$ pour traiter le cas des nouvelles unités. Développements spécifiques à faire.

On peut relier $U(S_1)$ et les probabilités d'inclusion conditionnelle π_i^{2/S_1} : dire que $\pi_i^{2/S_1} = 0$ signifie que, conditionnellement à la réalisation du tirage de S_1, l'individu i ne peut être tiré dans S_2. Tout se passe donc comme s'il n'était pas dans $U(S_1)$. On peut ainsi convenir pour simplifier que :

$$\boxed{U(S_1) = \{ i \in U / \pi_i^{2/S_1} > 0 \}}.$$

Ceci ne constitue pas une définition de $U(S_1)$ mais plutôt une propriété des probabilités d'inclusion conditionnelle pour les unités de $U(S_1)$: $\boxed{i \notin U(S_1) \Leftrightarrow \pi_i^{2/S_1} = 0}$.

4.3 Modalités d'utilisation de S_2.

On a en général deux optiques :

- soit on utilise S_2 conjointement avec S_1 (problématique d'échantillons complémentaires, de réserve, d'extension sur un domaine ou un territoire géographique …). L'échantillon d'intérêt sera alors $S_1 \cup S_2$. Toutefois, il faut prendre garde au problème d'éventuelles unités tirées deux fois.

- soit on utilise S_2 tout seul.

Dans les deux cas, on doit tenir compte du caractère aléatoire de S_1 pour définir les propriétés probabilistes, *inconditionnelles*, de $S_1 \cup S_2$ ou de S_2.

Tableau des principaux cas de figure possibles à étudier.

Base de sondage pour S_2 ↓ / **Utilisation de S_2 →**	S_2 **seul** **Conditions d'équilibrage sur S_2** **Eventuellement : S_2 ou $S_1 \cap S_2$ de taille fixe**	**Utilisation conjointe avec S_1 : $S = S_1 \cup S_2$.** **Conditions d'équilibrage sur S** **Eventuellement : S ou $S_1 \cap S_2$ de taille fixe**
S_1	Echantillon en 2 phases	*Sans intérêt*
CS_1	Échantillon disjoint du 1er, coordination *négative*	Echantillon de réserve, complémentaire, additionnel
U	Echantillon indépendant de S_1 Echantillon d'actualisation, panel avec conditions de recouvrement (cas PISA), coordination *positive*	Echantillon complémentaire avec recouvrement

5 Cas d'utilisation seule de S_2.

Le problème à résoudre est le suivant : comment tirer un échantillon S_2 conditionnellement à S_1, de manière que cet échantillon :

- ait des probabilités d'inclusion (inconditionnelles) fixées, notées π_i^2.
- satisfaisant des conditions d'équilibrage sur des variables V_i données (éventuellement vectorielles), qui s'écriront :

$$\boxed{\sum_{i \in S_2} \frac{V_i}{\pi_i^2} = T(V) = \sum_{i \in U} V_i} \text{. (2)}$$

- éventuellement, de taille fixe imposée n_2. Cette dernière condition impose la condition de compatibilité $\boxed{\sum_{i \in U} \pi_i^2 = n_2}$

Les seuls paramètres sur lesquels on peut jouer sont ceux du tirage conditionnel de S_2.

5.1 Relations.

- $\pi_i^2 = \mathrm{P}\{i \in S_2\} = \mathrm{E}\, \mathbf{1}_{i \in S_2} = \mathrm{E}\, \mathrm{E}^*(\mathbf{1}_{i \in S_2} / S_1) = \mathrm{E}[\mathrm{P}\{i \in S_2 / S_1\}]$, d'où :
$$\boxed{\pi_i^2 = \mathrm{E}\, \pi_i^{2/S_1}} \text{. (3)}$$
- $\mathrm{P}\{i \in S_1 \cap S_2\} = \mathrm{E}[\mathbf{1}_{i \in S_1} \mathbf{1}_{i \in S_2}] = \mathrm{E}\, \mathrm{E}^*(\mathbf{1}_{i \in S_1} \mathbf{1}_{i \in S_2} / S_1) = \mathrm{E}[\mathbf{1}_{i \in S_1} \mathrm{E}^*(\mathbf{1}_{i \in S_2} / S_1)]$, d'où : $\boxed{\mathrm{P}\{i \in S_1 \cap S_2\} = \mathrm{E}[\mathbf{1}_{i \in S_1}\, \pi_i^{2/S_1}]}$.

- $\mathrm{P}\{i \in S_2 / i \in S_1\} = \dfrac{P\{i \in S_1 \cap S_2\}}{P\{i \in S_1\}}$, d'où : $\boxed{\mathrm{P}\{i \in S_2 / i \in S_1\} = \dfrac{E[1_{i \in S_1} \pi_i^{2/S_1}]}{\pi_i^1}}$.

- $\mathrm{P}\{i \in S_2 / i \notin S_1\} = \dfrac{P\{i \in CS_1 \cap S_2\}}{P\{i \notin S_1\}}$.

  Or :
  $\mathrm{P}\{i \in CS_1 \cap S_2\} = \mathrm{E}[\mathbf{1}_{i \notin S_1} \mathbf{1}_{i \in S_2}] = \mathrm{E}\,\mathrm{E}^*(\mathbf{1}_{i \notin S_1} \mathbf{1}_{i \in S_2} / S_1) = \mathrm{E}[\mathbf{1}_{i \notin S_1} \mathrm{E}^*(\mathbf{1}_{i \in S_2} / S_1)]$,
  d'où :

$$\boxed{\mathrm{P}\{i \in S_2 / i \notin S_1\} = \frac{E[1_{i \notin S_1} \pi_i^{2/S_1}]}{1-\pi_i^1}}.$$

5.2 1er cas : $\boxed{U(S_1) = S_1}$.

[pour mémoire : échantillonnage en deux phases]

5.3 2ème cas : $\boxed{U(S_1) = CS_1}$.

On va prendre des probabilités d'inclusion conditionnelles de la forme : $\pi_i^{2/S_1} = a_i \ \mathbf{1}_{i \notin S_1}$. L'équation **(3)** donne alors : $\pi_i^2 = a_i (1-\pi_i^1)$. **Si l'on suppose que le 1er échantillon n'a pas fait apparaître de strate exhaustive (soit :** $\pi_i^1 \neq 1$), on en déduit : $\boxed{a_i = \dfrac{\pi_i^2}{1-\pi_i^1}}$, d'où :

$$\boxed{\pi_i^{2/S_1} = \frac{\pi_i^2}{1-\pi_i^1} \ \mathbf{1}_{i \notin S_1}}.$$

Cette valeur n'est licite que si : $\pi_i^1 \neq 1$ et $\dfrac{\pi_i^2}{1-\pi_i^1} \leq 1$, soit : $\boxed{\pi_i^2 \leq 1-\pi_i^1}$.

Le cas où c**ertaines unités ont** $\pi_i^1 = 1$ est traité plus bas.

On en déduit alors que :

$$\boxed{\mathrm{P}\{i \in S_2 / i \in S_1\} = 0}.$$

$$\boxed{\mathrm{P}\{i \in S_2 / i \notin S_1\} = a_i = \frac{\pi_i^2}{1-\pi_i^1}}.$$

Pour atteindre une condition d'équilibrage de la forme **(2)**, on ne peut utiliser que des conditions d'équilibrage sur le tirage conditionnel, de la forme **(1)**, qui va s'écrire ici :

$$\sum_{i \in S_2} Z_i \frac{1-\pi_i^1}{\pi_i^2} = \sum_{i \in CS_1} Z_i = \sum_{i \in U} Z_i - \sum_{i \in S_1} Z_i .$$

Si l'on définit les variables Z_i telles que : $V_i = Z_i(1-\pi_i^1)$, soit : $Z_i = \frac{V_i}{1-\pi_i^1}$, l'équation ci-dessus s'écrira : $\sum_{i \in S_2} \frac{V_i}{\pi_i^2} = \sum_{i \in U} \frac{V_i}{1-\pi_i^1} - \sum_{i \in S_1} \frac{V_i}{1-\pi_i^1}$.

Or on souhaite obtenir : $\sum_{i \in S_2} \frac{V_i}{\pi_i^2} = \sum_{i \in U} V_i$. Il convient donc de faire en sorte que :

$$\sum_{i \in U} \frac{V_i}{1-\pi_i^1} - \sum_{i \in S_1} \frac{V_i}{1-\pi_i^1} = \sum_{i \in U} V_i \text{ , soit : } \sum_{i \in S_1} \frac{V_i}{1-\pi_i^1} = \sum_{i \in U} \frac{V_i \pi_i^1}{1-\pi_i^1} \text{ , ou encore :}$$

$$\boxed{\sum_{i \in S_1} \frac{1}{\pi_i^1} \frac{V_i \pi_i^1}{1-\pi_i^1} = \sum_{i \in U} \frac{V_i \pi_i^1}{1-\pi_i^1}}.$$

Cette dernière condition s'interprète comme un équilibrage, lors du tirage du 1er échantillon, sur les variables $\frac{V_i \pi_i^1}{1-\pi_i^1}$.

Par suite, pour ce choix des probabilités d'inclusion conditionnelles, le 2ème échantillon sera équilibré sur les variables V_i si et seulement si :

- le 1er échantillon l'est sur les variables $\frac{V_i \pi_i^1}{1-\pi_i^1}$
- et le tirage conditionnel l'est sur les variables $Z_i = \frac{V_i}{1-\pi_i^1}$.

On voit que ce résultat impose une contrainte d'équilibrage sur le 1er échantillon qui, si elle n'a pas été prévue à l'avance, pose un problème pour l'équilibrage du 2nd. On verra plus loin (§ 5) comment y trouver une solution approchée.

Cas particuliers :

- Si le 1er échantillon est à probabilités égales ($\pi_i^1 = \frac{n_1}{N}$), le 2ème échantillon sera équilibré sur les variables V_i si et seulement si le 1er échantillon *ainsi que le tirage conditionnel* sont équilibrés sur les mêmes variables.
- Si l'on veut que l'échantillon S_2 soit de taille fixe n_2, il doit être équilbré sur les variables π_i^2, donc le 1er échantillon doit l'être sur les variables $\frac{\pi_i^1 \pi_i^2}{1-\pi_i^1}$.

On peut vérifier que cette condition est bien suffisante.

En effet, on doit vérifier que : $n_2 = \sum_{i \notin S_1} \pi_i^{2/S_1} = \sum_{i \notin S_1} \frac{\pi_i^2}{1-\pi_i^1}$ **, qui est la condition pour que le tirage conditionnel soit de taille fixe.**.

Or : $\sum_{i \notin S_1} \frac{\pi_i^2}{1-\pi_i^1} = \sum_{i \in U} \frac{\pi_i^2}{1-\pi_i^1} - \sum_{i \in S_1} \frac{\pi_i^2}{1-\pi_i^1}$

$$= \sum_{i \in U} \frac{\pi_i^2}{1-\pi_i^1} - \sum_{i \in S_1} \frac{1}{\pi_i^1} \frac{\pi_i^2 \pi_i^1}{1-\pi_i^1} .$$

Mais l'équilibrage du 1er échantillon sur les variables $\frac{\pi_i^1 \pi_i^2}{1-\pi_i^1}$ se traduit par la condition : $\sum_{i \in S_1} \frac{1}{\pi_i^1} \frac{\pi_i^2 \pi_i^1}{1-\pi_i^1} = \sum_{i \in U} \frac{\pi_i^2 \pi_i^1}{1-\pi_i^1}$ et l'on en déduit :

$$\sum_{i \notin S_1} \frac{\pi_i^2}{1-\pi_i^1} = \sum_{i \in U} \frac{\pi_i^2}{1-\pi_i^1} - \sum_{i \in U} \frac{\pi_i^2 \pi_i^1}{1-\pi_i^1} = \sum_{i \in U} \pi_i^2 .$$

Si les π_i^2 ont été choisies de manière à vérifier la condition : $\sum_{i \in U} \pi_i^2 = n_2$, on aura bien :

$$\sum_{i \notin S_1} \frac{\pi_i^2}{1-\pi_i^1} = n_2 . \blacksquare$$

Conclusion :

Sous la condition de compatibilité $\sum_{i \in U} \pi_i^2 = n_2$, on ne peut, **avec ce choix des probabilités d'inclusion conditionnelles** π_i^{2/S_1}, assurer simultanément le respect des probabilités finales d'inclusion π_i^2 et celui de la taille fixe n_2 pour l'échantillon S_2 (et ce, même en l'absence d'autres conditions d'équilibrage pour ce dernier) que si l'échantillon S_1 est équilibré sur les variables $\frac{\pi_i^1 \pi_i^2}{1-\pi_i^1}$.

Cas particuliers : $\pi_i^1 = \pi_i^2$ ou l'un des deux tirages est un SAS

Aperçu sur le cas des strates exhaustives.

Celles-ci peuvent apparaître à deux niveaux : à celui du tirage de S_1 et à celui du tirage de S_2. Dans ce cas, il n'y a plus de compatibilité entre l'existence de telles strates et la contrainte de tirer S_2 dans le complémentaire de S_1 : si une unité i a été retenue d'office dans S_1, elle ne pourrait plus être tirée dans S_2 ; inversement, retenir d'office une unité dans S_2 doit pouvoir se faire sans prendre en compte l'appartenance ou non de cette unité à S_1.

Le mieux dans ce cas est de *stratifier l'univers*. On définit : $U_j = \left\{ i \in U \ ; \ \pi_i^j = 1 \right\}$ pour $j = 1$ *ou* 2.

On posera alors : $\boxed{\pi_i^{2/S_1} = \left| \begin{array}{c} 1 \ si \ i \in U_2 \\ \pi_i^2 \ si \ i \in CU_2 \cap U_1 \\ a_i 1_{i \notin S_1} \ \ si \ i \in CU_2 \cap CU_1 \end{array} \right.}$.

Toutes les équations d'équilibrage (y compris les contraintes de taille) devront être écrites en faisant apparaître la décomposition de l'univers selon les trois strates définies ci-dessus, caractérisées par la spécification de π_i^{2/S_1}.

Discussion sur la contrainte de taille fixe pour le 2nd échantillon.

On a vu ci-dessus que l'échantillon S_2 peut ne pas être de taille fixe imposée n_2, même avec des probabilités π_i^2 choisies de manière adéquate (i.e. vérifiant la condition $\sum_{i \in U} \pi_i^2 = n_2$), si l'on n'a pas imposé au préalable, dès le tirage de l'échantillon S_1, des conditions d'équilibrage appropriées.

Ceci peut paraître gênant, pour une contrainte (taille fixe) qui peut paraître assez banale. On va donc voir s'il n'est pas possible de contourner la contrainte d'équilibrage sur S_1, ce qui va imposer en fait d'agir sur le seul levier restant : la forme des probabilités d'inclusion conditionnelle π_i^{2/S_1}.

Peut-on choisir les π_i^{2/S_1} de manière à assurer simultènement :

- **des probabilités d'inclusion finales dans S_2, π_i^2, données**
- **un échantillon S_2 de taille fixe n_2 ?**

On va tester ici deux formulations.

a) Formulation additive.

On va chercher une formulation : $\boxed{\pi_i^{2/S_1} = [a_i + \mathrm{f}(S_1)] \ \mathbf{1}_{i \notin S_1}}$.

Alors : $E\pi_i^{2/S_1} = a_i(1 - \pi_i^1) + E[f(S_1)1_{i \notin S_1}]$.

On veut un 2nd échantillon de taille fixe : $\sum_{i \in CS_1} \pi_i^{2/S_1} = n_2$, d'où : $\sum_{i \in CS_1} [a_i + f(S_1)] = n_2$, soit :

$\sum_{i \in CS_1} a_i + [N - n(S_1)] f(S_1) = n_2$, d'où : $\boxed{f(S_1) = \dfrac{n_2 - \sum_{i \in CS_1} a_i}{N - n(S_1)}}$.

Pour imposer au 2nd échantillon des probabilités d'inclusion finales données π_i^2, on doit ensuite résoudre : $\pi_i^2 = a_i(1-\pi_i^1) + E[f(S_1)1_{i \notin S_1}]$.

Or : $E[f(S_1)1_{i \notin S_1}] = E\left[\dfrac{n_2 - \sum_{j \in CS_1} a_j}{N - n(S_1)} 1_{i \notin S_1}\right]$.

On va supposer, pour simplifier, que la taille $n(S_1)$ n'est pas aléatoire, soit n_1.

Alors : $E\left[\dfrac{n_2 - \sum_{j \in CS_1} a_j}{N - n(S_1)} 1_{i \notin S_1}\right] = \dfrac{E\left[[n_2 - \sum_{j \in CS_1} a_j] 1_{i \notin S_1}\right]}{N - n_1}$

$$= \frac{n_2(1-\pi_i^1) - \sum_{j \in U} a_j E(1_{j \notin S_1} 1_{i \notin S_1})}{N - n_1}.$$

Or : $E(1_{j \notin S_1} 1_{i \notin S_1}) = \mathrm{P}\{j \notin S_1 \ et \ i \notin S_1\} = \begin{cases} 1-\pi_i^1 \ si \ i = j \\ 1-\pi_i^1 - \pi_j^1 + \pi_{i,j}^1 \ si \ i \neq j \end{cases}$,

en notant $\pi_{i,j}^1$ les probabilités d'inclusion doubles lors du tirage de S_1, soit :

$$\pi_{i,j}^1 = \mathrm{P}\{i \in S_1 \ et \ j \in S_1\}.$$

On obtient donc :

$$\sum_{j \in U} a_j E(1_{j \notin S_1} 1_{i \notin S_1}) = a_i(1-\pi_i^1) + \sum_{\substack{j \in U \\ j \neq i}} a_j (1 - \pi_i^1 - \pi_j^1 + \pi_{i,j}^1)$$

$$= a_i(1-\pi_i^1) + (1-\pi_i^1) \sum_{\substack{j \in U \\ j \neq i}} a_j - \sum_{\substack{j \in U \\ j \neq i}} a_j \pi_j^1 + \sum_{\substack{j \in U \\ j \neq i}} a_j \pi_{i,j}^1$$

$$= a_i(1-\pi_i^1) + (1-\pi_i^1)(\sum_{j \in U} a_j - a_i) - (\sum_{j \in U} a_j \pi_j^1 - a_i \pi_i^1) + \sum_{\substack{j \in U \\ j \neq i}} a_j \pi_{i,j}^1$$

$$= (1-\pi_i^1)(\sum_{j \in U} a_j) - \sum_{j \in U} a_j \pi_j^1 + a_i \pi_i^1 + \sum_{\substack{j \in U \\ j \neq i}} a_j \pi_{i,j}^1 .$$

On en déduit que :

$$\pi_i^2 = a_i(1-\pi_i^1) + \frac{n_2(1-\pi_i^1) - (1-\pi_i^1)(\sum_{j\in U} a_j) + \sum_{j\in U} a_j \pi_j^1 - a_i \pi_i^1 - \sum_{\substack{j\in U \\ j\neq i}} a_j \pi_{i,j}^1}{N-n_1}$$

$$= a_i\left(1-\pi_i^1 \frac{N-n_1+1}{N-n_1}\right) + \frac{n_2(1-\pi_i^1)}{N-n_1} - \frac{1-\pi_i^1}{N-n_1}\left(\sum_{j\in U} a_j\right) + \frac{\sum_{j\in U} a_j \pi_j^1 - \sum_{\substack{j\in U \\ j\neq i}} a_j \pi_{i,j}^1}{N-n_1}.$$

On obtient un système linéaire d'équations, d'inconnues a_i pour toute unité i de l'univers, relativement complexe. Ce système présente les difficultés suivantes :

- il dépend des probabilités d'inclusion doubles relatives à S_1, $\pi_{i,j}^1$, souvent inconnues.
- il n'est pas soluble analytiquement mais peut s'écrire matriciellement et doit pouvoir être résolu par des voies numériques.
- l'existence de solutions reste à prouver
- les solutions doivent respecter des contraintes de cohérence logique : $a_i + f(S_1) \in [0, 1]$.

b) Formulation multiplicative.

Prenons ici : $\boxed{\pi_i^{2/S_1} = f(S_1)\, a_i\, \mathbf{1}_{i\notin S_1}}$.

La condition : $\sum_{i\notin S_1} \pi_i^{2/S_1} = n_2$ s'écrit : $f(S_1)\sum_{i\notin S_1} a_i = n_2$, ce qui détermine : $f(S_1) = \frac{n_2}{\sum_{i\notin S_1} a_i}$, d'où :

$$\pi_i^{2/S_1} = \frac{n_2}{\sum_{j\notin S_1} a_j} a_i\, \mathbf{1}_{i\notin S_1}.$$

La condition : $E\pi_i^{2/S_1} = \pi_i^2$ devient alors : $\boxed{\pi_i^2 = n_2\, a_i\, E\left(\frac{1_{i\notin S_1}}{\sum_{j\notin S_1} a_j}\right)}$.

Cet ensemble de relations (valables pour tout i) constitue un système d'équations dont les inconnues sont les a_i. Malheureusement, ce système est insoluble, d'une part parce qu'il nécessiterait de connaître la loi complète de S_1 et pas seulement les probabilités d'inclusion relatives à cet échantillon (π_i^1), et d'autre part parce qu'il conduit à des équations non linéaires : il pourrait au mieux y avoir des solutions numériques.

Nota : même si S_1 est aléatoire simple, le système est insoluble analytiquement.

Un essai de solution approchée.

On peut écrire : $\sum_{j\notin S_1} a_j = \sum_{j\in U} a_j - \sum_{j\in S_1} a_j = \sum_{j\in U} a_j - \sum_{j\in S_1} \frac{a_j \pi_j^1}{\pi_j^1}$

La somme $\sum_{j\in S_1} \frac{a_j \pi_j^1}{\pi_j^1}$ est un estimateur sans biais de $\sum_{j\in U} a_j \pi_j^1$. Pour de « grands » échantillons S_1, on pourrait admettre l'approximation : $\sum_{j\in S_1} \frac{a_j \pi_j^1}{\pi_j^1} \approx \sum_{j\in U} a_j \pi_j^1$. On aurait alors :

$$\boxed{\sum_{j\notin S_1} a_j \approx \sum_{j\in U} a_j (1-\pi_j^1)}.$$

Par suite, l'équation encadrée deviendrait : $\pi_i^2 \approx \frac{n_2\, a_i}{\sum_{j\in U} a_j (1-\pi_j^1)} E(1_{i\notin S_1}) = \frac{n_2\, a_i}{\sum_{j\in U} a_j (1-\pi_j^1)} (1-\pi_i^1)$.

Les a_i vérifient donc le système : $a_i \approx \frac{\pi_i^2}{n_2 (1-\pi_i^1)} \sum_{j\in U} a_j (1-\pi_j^1)$.

En posant : $\sigma = \sum_{j\in U} a_j (1-\pi_j^1)$, on obtient : $a_i \approx \sigma \frac{\pi_i^2}{n_2 (1-\pi_i^1)}$, d'où :

$$\sum_{i\in U} a_i (1-\pi_i^1) \approx \frac{\sigma}{n_2} \sum_{i\in U} \frac{\pi_i^2}{1-\pi_i^1}.$$

Cette solution ne marche que si : $\boxed{\sum_{i\in U} \frac{\pi_i^2}{1-\pi_i^1} = n_2}$.

Approximation plus développée.

$$\frac{1}{\sum_{j\notin S_1} a_j} = \frac{1}{\sum_{j\notin S_1} a_j - \sigma + \sigma} = \frac{1}{\sigma} \frac{1}{1+\frac{\sum_{j\notin S_1} a_j - \sigma}{\sigma}} \approx \frac{1}{\sigma}\left(1 - \frac{\sum_{j\notin S_1} a_j - \sigma}{\sigma}\right).$$

Par suite :

$$E\left(\frac{1_{i\notin S_1}}{\sum_{j\notin S_1} a_j}\right) \approx \frac{1}{\sigma} E\left(1_{i\notin S_1} [1 - \frac{\sum_{j\notin S_1} a_j - \sigma}{\sigma}]\right) = \frac{1}{\sigma} [2(1-\pi_i^1) - \frac{1}{\sigma} E(1_{i\notin S_1} \sum_{j\notin S_1} a_j)] .$$

Or : $E(1_{i\notin S_1} \sum_{j\notin S_1} a_j) = E(1_{i\notin S_1} \sum_{j\in U} a_j 1_{j\notin S_1}) = \sum_{j\in U} a_j E(1_{i\notin S_1} 1_{j\notin S_1})$

$$= (1-\pi_i^1)(\sum_{j\in U} a_j) - \sum_{j\in U} a_j \pi_j^1 + a_i \pi_i^1 + \sum_{\substack{j\in U \\ j\neq i}} a_j \pi_{i,j}^1 .$$

On obtient donc un système qui s écrirait :

$$\pi_i^2 \approx n_2 a_i \frac{1}{\sigma}\left[2(1-\pi_i^1) - \frac{1}{\sigma}[(1-\pi_i^1)(\sum_{j\in U} a_j) - \sum_{j\in U} a_j \pi_j^1 + a_i \pi_i^1 + \sum_{\substack{j\in U \\ j\neq i}} a_j \pi_{i,j}^1] \right].$$
$$avec : \sigma = \sum_{j\in U} a_j (1-\pi_j^1)$$

Là encore, on obtient un système complexe d'équations linéaires d'inconnues a_i, présentant les mêmes difficultés que celui obtenu dans la formulation additive.

5.4 3ème cas : $U(S_1) = U$.

Dans ce cas, la base de sondage de S_2 ne dépend pas de S_1 et, si l'on ne cherche qu'à exploiter l'échantillon S_2 seul, d'éventuelles conditions d'équilibrage sur S_2 ne devraient pas faire intervenir S_1. En revanche, la difficulté se présente si l'on vise une **articulation avec S_1 en termes de recouvrement**.

Cela correspond au cas de l'enquête PISA : on veut un nouvel échantillon avec des probabilités d'inclusion données, éventuellement une taille imposée, des conditions d'équilibrage exprimant sa « représentativité »[5] et la possibilité d'une exploitation en panel avec une partie de l'échantillon S_1, ce qui va conduire à imposer qu'un certain nombre d'unités soient communes aux deux échantillons.

Concernant cette dernière condition, on peut adopter trois formulations :

- une formulation en *effectifs* : Card ($S_1 \cap S_2$) = $n_{1,2}$
- une formulation en *taux de recouvrement* : Card ($S_1 \cap S_2$) = α Card (S_1) : une fraction α de l'échantillon S_1 se retrouve sélectionnée dans S_2.

  Les deux formulations ci-dessus sont équivalentes dès lors que S_1 est de taille fixe n_1.

- une formulation en *espérance individuelle* : $\forall i \in U$: P $\{i \in S_1 \cap S_2\}$ = γ_i .

On va prendre ici des probabilités d'inclusion conditionnelles de la forme :

[5] Par exemple : répartition par sexe et niveau d'études de la population dans le champ de l'enquête, zone géographique, degré d'urbanisation, taille des établissements scolaires, éventuellement caractéristiques socio-démographiques des familles auxquelles appartiennent les élèves du champ…

$$\boxed{\pi_i^{2/S_1} = \begin{vmatrix} a_i \ si\ i \notin S_1 \\ b_i \ si\ i \in S_1 \end{vmatrix} = a_i 1_{i \notin S_1} + b_i 1_{i \in S_1}},$$

avec : $\boxed{a_i \ et\ b_i \in [0,1]}$.

On en déduit alors que :

$$\boxed{\mathrm{P}\{i \in S_2 / i \in S_1\} = b_i}.$$
$$\boxed{\mathrm{P}\{i \in S_2 / i \notin S_1\} = a_i}.$$

On notera que la condition $a_i = b_i$ conduit à des probabilités conditionnelles indépendantes de S_1.

5.4.1 Pourquoi une solution « naïve » ne marche-t-elle pas ?

On pourrait imaginer simplement de tirer S_2 dans U avec des probabilités π_i^2 données et d'imposer d'éventuelles conditions d'équilibrage : $\sum_{i \in S_2} \frac{V_i}{\pi_i^2} = \sum_{i \in U} V_i$. Mais, pour imposer le recouvrement avec S_1 sous la forme, par exemple, d'un nombre fixé $n_{1,2}$ d'unités de S_2 tirées dans S_1 (voir supra pour d'autres formulations du recouvrement), on doit avoir la relation : $\sum_{i \in S_1} \pi_i^2 = n_{1,2}$. Or cette condition fait intervenir S_1 et est aléatoire de surcroît. *Elle devrait être satisfaite quelle que soit la réalisation de* S_1, ce qui impose que S_1 ait été tiré *d'une certaine manière*.

Concrètement, la relation $\sum_{i \in S_1} \pi_i^2 = n_{1,2}$ s'écrit aussi : $\sum_{i \in S_1} \frac{1}{\pi_i^1} \pi_i^2 \pi_i^1 = n_{1,2}$, **ce qui exprime que** S_1 **est équilibré sur les variables** $\pi_i^2 \pi_i^1$**, à condition que :** $\sum_{i \in U} \pi_i^2 \pi_i^1 = n_{1,2}$**.**

On voit donc que cela ne marche que si S_1 **a été astreint à une certaine relation d'équilibrage et qu'il existe une relation entre les paramètres du problème (ici :** π_i^1**,** π_i^2 **et** $n_{1,2}$**).**

Pourquoi interpréter une condition de la forme $\sum_{i \in S_1} \alpha_i = T$ en termes d'équilibrage ? La condition $\sum_{i \in S_1} \alpha_i = T$ s'écrit : $\sum_{i \in S_1} \frac{1}{\pi_i^1} \alpha_i \pi_i^1 = T$ et, comme : $E\left(\sum_{i \in S_1} \frac{1}{\pi_i^1} \alpha_i \pi_i^1 \right) = \sum_{i \in U} \alpha_i \pi_i^1$, on doit s'attendre à ce que $\sum_{i \in S_1} \frac{1}{\pi_i^1} \alpha_i \pi_i^1$ soit peu différent de $\sum_{i \in U} \alpha_i \pi_i^1$. Par suite, un échantillon S_1 quelconque tiré avec des probabilités d'inclusion π_i^1 ne peut pas nécessairement satisfaire, même de manière approchée, la condition $\sum_{i \in S_1} \alpha_i = T$.

Ce n'est possible que si : $\sum_{i \in U} \alpha_i \pi_i^1 = T$ **et la condition cherchée s'interprète alors comme une condition d'équilibrage de** S_1 **sur les variables** $\alpha_i \pi_i^1$.

$\rightarrow$ Revenons alors à la formulation générale proposée ci-dessus.

5.4.2 Regardons tout d'abord comment satisfaire simultanément les conditions d'équilibrage sur S_2 et le respect des probabilités d'inclusion finales.

L'équation **(3)** donne : $\boxed{\pi_i^2 = a_i (1 - \pi_i^1) + b_i \pi_i^1}$.

$\rightarrow$ *En particulier*, si $a_i = b_i = \pi_i^{2/S_1}$, on aura : $\boxed{\pi_i^2 = \pi_i^{2/S_1}}$.

Par ailleurs, pour atteindre une condition d'équilibrage de la forme **(2)**, on n'utilise à nouveau que des conditions d'équilibrage sur le tirage conditionnel, de la forme **(1)**, qui va s'écrire ici :

$$\sum_{i \in S_2 \cap S_1} \frac{Z_i}{b_i} + \sum_{i \in S_2 \cap CS_1} \frac{Z_i}{a_i} = \sum_{i \in U} Z_i \ .$$

Avec : $\sum_{i \in S_2 \cap CS_1} \frac{Z_i}{a_i} = \sum_{i \in S_2} \frac{Z_i}{a_i} - \sum_{i \in S_2 \cap S_1} \frac{Z_i}{a_i}$, on obtient la condition :

$$\boxed{\sum_{i \in S_2 \cap S_1} Z_i \left(\frac{1}{b_i} - \frac{1}{a_i}\right) + \sum_{i \in S_2} \frac{Z_i}{a_i} = \sum_{i \in U} Z_i}.$$

Cette condition est *globale*. En réalité, on est capable, lors du tirage conditionnel de S_2, de faire une stratification selon l'appartenance ou pas à S_1, et de réaliser le tirage de deux sous-échantillons équilibrés, l'un dans S_1, l'autre dans CS_1 :

- équilibrage sur des variables $Z_{1,i}$ dans S_1
- équilibrage sur des variables $Z_{2,i}$ dans CS_1.

On assurera ainsi des conditions d'équilibrage de la forme : $\begin{cases} \sum_{i \in S_2 \cap S_1} \frac{Z_{1,i}}{b_i} = \sum_{i \in S_1} Z_{1,i} \\ \sum_{i \in S_2 \cap CS_1} \frac{Z_{2,i}}{a_i} = \sum_{i \in CS_1} Z_{2,i} \end{cases}$.

On note que la 2ème équation peut s'écrire : $\sum_{i \in S_2} \frac{Z_{2,i}}{a_i} - \sum_{i \in S_2 \cap S_1} \frac{Z_{2,i}}{a_i} = \sum_{i \in U} Z_{2,i} - \sum_{i \in S_1} Z_{2,i}$, d'où finalement le système : $\boxed{\begin{array}{c} \sum_{i \in S_2 \cap S_1} \frac{Z_{1,i}}{b_i} = \sum_{i \in S_1} Z_{1,i} \\ \sum_{i \in S_2 \cap S_1} \frac{Z_{2,i}}{a_i} = \sum_{i \in S_2} \frac{Z_{2,i}}{a_i} + \sum_{i \in S_1} Z_{2,i} - \sum_{i \in U} Z_{2,i} \end{array}}$.

On souhaite atteindre une condition d'équilibrage global sur S_2, de la forme : $\sum_{i \in S_2} \frac{V_i}{\pi_i^2} = \sum_{i \in U} V_i$.

Cette condition ne faisant pas intervenir $S_1 \cap S_2$, on se débarrasse des équations où apparaissent des sommes portant sur $S_1 \cap S_2$ dans le système précédent : **on s'arrange pour que :** $\boxed{\frac{Z_{1,i}}{b_i} = \frac{Z_{2,i}}{a_i}}$

et on élimine le terme $\sum_{i \in S_2 \cap S_1} \frac{Z_{2,i}}{a_i}$ **entre les deux équations de ce système**. On obtient alors :

$$\sum_{i \in S_1} Z_{1,i} = \sum_{i \in S_2} \frac{Z_{1,i}}{b_i} + \sum_{i \in S_1} \frac{a_i}{b_i} Z_{1,i} - \sum_{i \in U} \frac{a_i}{b_i} Z_{1,i} \text{ , soit :}$$

$$\boxed{\sum_{i \in S_2} \frac{Z_{1,i}}{b_i} = \sum_{i \in S_1} (1 - \frac{a_i}{b_i}) Z_{1,i} + \sum_{i \in U} \frac{a_i}{b_i} Z_{1,i}}.$$

Par suite, pour que cette équation soit équivalente à : $\sum_{i \in S_2} \frac{V_i}{\pi_i^2} = \sum_{i \in U} V_i$, il suffit de définir :

$$Z_{1,i} = \frac{b_i}{\pi_i^2} V_i$$

et de s'arranger pour que : $\sum_{i \in S_1} (1 - \frac{a_i}{b_i}) \frac{b_i}{\pi_i^2} V_i + \sum_{i \in U} \frac{a_i}{b_i} \frac{b_i}{\pi_i^2} V_i = \sum_{i \in U} V_i$, soit :

$$\sum_{i \in S_1} (1 - \frac{a_i}{b_i}) \frac{b_i}{\pi_i^2} V_i = \sum_{i \in U} (1 - \frac{a_i}{\pi_i^2}) V_i \text{ .}$$

Ecrivant cette dernière relation sous la forme : $\sum_{i \in S_1} \frac{1}{\pi_i^1} (1 - \frac{a_i}{b_i}) \frac{b_i}{\pi_i^2} V_i \pi_i^1 = \sum_{i \in U} (1 - \frac{a_i}{\pi_i^2}) V_i$, on voit **qu'elle s'interprète comme une condition d'équilibrage, lors du tirage de** S_1**, sur la variable** $(1 - \frac{a_i}{b_i}) \frac{b_i}{\pi_i^2} V_i \pi_i^1 = (b_i - a_i) \frac{\pi_i^1}{\pi_i^2} V_i$, **sous la réserve que** :

$$\sum_{i \in U} (b_i - a_i) \frac{\pi_i^1}{\pi_i^2} V_i = \sum_{i \in U} (1 - \frac{a_i}{\pi_i^2}) V_i \text{ .}$$

Or cette dernière condition est satisfaite dès que : $(b_i - a_i)\frac{\pi_i^1}{\pi_i^2} V_i = (1-\frac{a_i}{\pi_i^2}) V_i$, ce qui s'écrit : $(b_i - a_i)\pi_i^1 = \pi_i^2 - a_i$, soit : $\pi_i^2 = b_i\pi_i^1 + a_i(1-\pi_i^1)$. Cette condition est précisément celle de l'équation **(3)** : $\pi_i^2 = E\pi_i^{2/S_1}$.

En conclusion, pour ce choix des probabilités d'inclusion conditionnelles, le 2ème échantillon sera équilibré sur les variables V_i si et seulement si :

- le 1er échantillon l'est sur les variables $(b_i - a_i)\frac{\pi_i^1}{\pi_i^2} V_i$ **(C1)**
- le tirage conditionnel, pour la partie de l'échantillon S_2 puisée dans S_1, l'est sur les variables

$$Z_{1,i} = \frac{b_i}{\pi_i^2} V_i$$

- le tirage conditionnel, pour la partie de l'échantillon S_2 puisée dans CS_1, l'est sur les variables $Z_{2,i} = \frac{a_i}{b_i} Z_{1,i} = \frac{a_i}{\pi_i^2} V_i$.

On voit à nouveau que ce résultat impose une contrainte d'équilibrage sur le 1er échantillon qui, si elle n'a pas été prévue à l'avance, pose un problème pour l'équilibrage du 2nd. On verra plus loin (§ 5) comment y trouver une solution approchée.

A ce stade, les constantes a_i et b_i restent à déterminer et sont seulement liées par la condition :

$$\pi_i^2 = b_i\pi_i^1 + a_i(1-\pi_i^1).$$

5.4.3 Comment prendre en compte les conditions de recouvrement entre les deux échantillons ?

Formulation en effectifs ou en taux.

On va écrire la contrainte sous la forme : Card ($S_1 \cap S_2$) = α Card (S_1). Ceci peut se traduire en disant que la partie de l'échantillon conditionnel qui est tirée dans S_1 est de taille fixe αn_1 (on suppose ici S_1 lui-même de taille fixe n_1). A voir : généralisation ?

On doit donc avoir la condition : $\sum_{i \in S_1} \pi_i^{2/S_1} = \alpha n_1$, soit : $\boxed{\sum_{i \in S_1} b_i = \alpha n_1}$. **(C2)**

Cette condition aléatoire doit être satisfaite quel que soit S_1. Elle peut s'écrire :

$$\sum_{i \in S_1} \frac{1}{\pi_i^1} b_i \pi_i^1 = \alpha n_1$$

et cette égalité peut s'interpréter comme une condition d'équilibrage de l'échantillon S_1 sur les variables $b_i\pi_i^1$, à condition que : $\boxed{\sum_{i \in U} b_i \pi_i^1 = \alpha n_1}$.

On obtient donc une nouvelle condition d'équilibrage sur le 1er échantillon, doublée d'une condition supplémentaire sur les inconnues b_i .

Dans ce cas, le tirage conditionnel de la partie de S_2 puisée dans S_1 se fait à taille fixe αn_1, c'est--à-dire avec un équilibrage sur les variables b_i .

5.4.4 Peut-on assurer la condition supplémentaire que l'échantillon S_2 soit de taille fixe n_2 ?

Cette condition équivaut à : $Card\,(S_2 \cap CS_1) = n_2 - \alpha n_1$ et on ne peut l'assurer que si :
$\sum_{i \notin S_1} \pi_i^{2/S_1} = n_2 - \alpha n_1$, soit : $\sum_{i \notin S_1} a_i = n_2 - \alpha n_1$, ou : $\sum_{i \in S_1} a_i = \sum_{i \in U} a_i - n_2 + \alpha n_1$.

En récrivant cette condition sous la forme : $\sum_{i \in S_1} \frac{1}{\pi_i^1} a_i \pi_i^1 = \sum_{i \in U} a_i - n_2 + \alpha n_1$, on voit qu'elle peut s'interpréter comme une condition d'équilibrage de l'échantillon S_1 sur la variable $a_i \pi_i^1$, à condition que les variables a_i vérifient la condition : $\sum_{i \in U} a_i \pi_i^1 = \sum_{i \in U} a_i - n_2 + \alpha n_1$, soit :

$$\boxed{\sum_{i \in U} a_i (1 - \pi_i^1) = n_2 - \alpha n_1}.$$

Or cette dernière condition est vérifiée à cause de la relation : $\sum_{i \in U} a_i (1 - \pi_i^1) = \sum_{i \in U} (\pi_i^2 - b_i \pi_i^1)$, de la condition de compatibilité $\sum_{i \in U} \pi_i^2 = n_2$ et de celle sur les b_i : $\sum_{i \in U} b_i \pi_i^1 = \alpha n_1$.

Au final, l'échantillon S_2 est de taille fixe n_2 si et seulement si on rajoute aux conditions précédentes (C1 et C2) la condition d'équilibrage du 1er échantillon sur les variables :

$$a_i \pi_i^1 = \frac{\pi_i^1 (\pi_i^2 - b_i \pi_i^1)}{1 - \pi_i^1} \quad \textbf{(C3)}.$$

Il faut alors réaliser le tirage conditionnel de la partie de S_2 puisée dans CS_1 avec une taille fixe $n_2 - \alpha n_1$, c'est-à-dire avec un équilibrage sur les variables a_i .

Nota :

La condition de taille fixe pour S_2 peut se déduire directement de l'équilibrage de S_2 sur π_i^2 , qui implique l'équilibrage de S_1 sur $(b_i - a_i)\pi_i^1$ (cf. supra, §§ b)).

Cette dernière approche est bien compatible avec le résultat obtenu ci-dessus. En effet, la condition de recouvrement se traduisant par une condition d'équilibrage de l'échantillon S_1 sur les variables

$b_i\pi_i^1$, si celle-ci est vérifiée, **il y alors équivalence entre l'équilibrage de** S_1 **sur** $a_i\,\pi_i^1\,[=\frac{\pi_i^1(\pi_i^2 - b_i\pi_i^1)}{1-\pi_i^1}]$ **et sur** $(b_i - a_i)\,\pi_i^1$.

Le résultat se déduit en effet par linéarité des conditions d'équilibrage (s'il y a équilibrage sur $a_i\pi_i^1$ et sur $b_i\pi_i^1$, alors il y a équilibrage sur $b_i\,\pi_i^1 - a_i\pi_i^1 = \pi_i^2 - a_i$. ■

Synthèse.

Sous la condition de de taille fixe n_1 pour l'échantillon S_1 :

Contraintes imposées à S_2	**Equilibrage nécessaire en résultant sur S_1**	**Relations nécessaires entre les paramètres**	**Modalités de tirage conditionnel de S_2**	
			Partie puisée dans S_1 $\pi_i^{2/S_1} = b_i$	**Partie puisée dans $\mathcal{C}S_1$** $\pi_i^{2/S_1} = a_i$
Respect des probabilités d'inclusion finales π_i^2		$\pi_i^2 = a_i\,(1-\pi_i^1) + b_i\pi_i^1$		
Equilibrage sur des variables V_i	Equilibrage sur les variables $(b_i - a_i)\dfrac{\pi_i^1}{\pi_i^2}V_i$		Equilibrage conditionnel sur les variables $\dfrac{b_i}{\pi_i^2}V_i$	Equilibrage conditionnel sur les variables $\dfrac{a_i}{\pi_i^2}V_i$.
Taux de recouvrement α par rapport à S_1	Equilibrage sur les variables $b_i\,\pi_i^1$	$\sum_{i\in U} b_i\,\pi_i^1 = \alpha n_1$	Taille fixe αn_1 $\Leftrightarrow$ Equilibrage conditionnel sur les variables b_i	Taille fixe $n_2 - \alpha n_1$ $\Leftrightarrow$ Equilibrage conditionnel sur les variables a_i
Taille fixe n_2	Equilibrage sur les variables $a_i\pi_i^1$	$\sum_{i\in U}\pi_i^2 = n_2$		

Nota :

Sous la condition de compatibilité (taille fixe de S_2) : $\sum_{i\in U}\pi_i^2 = n_2$:

Equilibrage de S_1 sur $b_i\pi_i^1$ et sur $a_i\pi_i^1$ avec la relation : $\boxed{\sum_{i\in U} b_i\,\pi_i^1 = \alpha n_1}$

$\Leftrightarrow$

Equilibrage de S_1 sur $a_i\pi_i^1$ et sur $\pi_i^2 - a_i$ avec la relation : $\boxed{\sum_{i\in U} a_i(1-\pi_i^1) = n_2 - \alpha n_1}$.

5.4.5 Solutions possibles.

Les inconnues à déterminer sont les a_i et les b_i (à valeurs dans [0, 1]), qui représentent les probabilités d'inclusion conditionnelles π_i^{2/S_1} selon que $i \notin S_1$ ou $i \in S_1$. Elles sont en nombre $2N$ et astreintes à différentes relations linéaires (en nombre $N+1$). La stratégie consistera alors à tirer deux sous-échantillons selon ces probabilités conditionnelles, l'un dans CS_1 (avec probabilité b_i), l'autre dans S_1 (avec probabilité a_i).

On peut choisir de mettre en majeur le respect des différentes contraintes de la colonne de gauche du tableau ci-dessus. Les deux sous-échantillons tirés conditionnellement à S_1 seront de tailles fixes (respectivement αn_1 dans S_1 et $n_2 - \alpha n_1$ dans CS_1.). Il reste encore des degrés de liberté et plusieurs solutions sont envisageables, **mais elles ne fonctionnent rigoureusement que si l'échantillon S_1 a été tiré avec des conditions d'équilibrage adéquates**.

- S'il existe au moins une variable d'équilibrage X_0 (*à valeurs strictement positives*) pour le 1er échantillon, on peut s'arranger pour que : $b_i \pi_i^1 = \lambda X_{0,i}$, soit : $b_i = \lambda \dfrac{X_{0,i}}{\pi_i^1}$.

  La constante λ est déterminée par la condition $\sum_{i \in U} b_i \pi_i^1 = \alpha n_1$ qui s'écrit alors :

$$\lambda \sum_{i \in U} X_{0,i} = \alpha n_1, \text{ soit : } \lambda^* = \frac{\alpha n_1}{\sum_{i \in U} X_{0,i}} = \frac{\alpha n_1}{T(X_0)}.$$

  Sous ces hypothèses , les constantes a_i et b_i seront parfaitement déterminées :

$$\boxed{\begin{aligned} b_i &= \lambda^* \frac{X_{0,i}}{\pi_i^1} \\ a_i &= \frac{\pi_i^2 - b_i \pi_i^1}{1 - \pi_i^1} = \frac{\pi_i^2 - \lambda^* X_{0,i}}{1 - \pi_i^1} \end{aligned}},$$

  Cette solution est licite si et seulement si les conditions suivantes sont satisfaites :

  - $b_i \leq 1$, ce qui impose : $\alpha \dfrac{X_{0,i}}{T(X_0)} \dfrac{n_1}{\pi_i^1} \leq 1$
  - $a_i \in [0, 1]$
  - $\pi_i^1 \neq 1$. Voir le cas de strate exhaustive pour le 1er échantillon.

  L'avantage de cette solution est qu'elle permet d'incorporer une condition d'équilibrage existante sur S_1, sans requérir d'imposer de nouvelle condition d'équilibrage. Il reste toutefois à satisfaire aussi la condition d'équilibrage de S_1 sur les $a_i \pi_i^1$.

- En particulier, avec un échantillon S_1 **de taille fixe** n_1, il y a automatiquement équilibrage de S_1 sur les variables π_i^1. On obtient alors : $b_i = \lambda^*$, avec : $\lambda^* = \frac{\alpha n_1}{T(\pi^1)} = \alpha$, soit :

$$\boxed{b_i = \alpha \in [0, 1], \forall i \in U}.$$

Cela conduit à faire simplement un tirage aléatoire simple au taux α dans S_1.

On en déduit ensuite : $\boxed{a_i = \frac{\pi_i^2 - b_i \pi_i^1}{1 - \pi_i^1} = \frac{\pi_i^2 - \alpha \pi_i^1}{1 - \pi_i^1}}$.

Cette condition n'est admissible que si $a_i \in [0, 1]$, ce qui conduit aux conditions : $\forall i \in U : 1 - \frac{1 - \pi_i^2}{\pi_i^1} \leq \alpha \leq \frac{\pi_i^2}{\pi_i^1}$, soit : $\boxed{\underset{i \in U}{Max}\left(1 - \frac{1 - \pi_i^2}{\pi_i^1}\right) \leq \alpha \leq \underset{i \in U}{Min}\left(\frac{\pi_i^2}{\pi_i^1}\right)}$.

Toutes les valeurs du taux de renouvellement α ne sont donc pas admissibles.

C'est notamment le cas pour des grandes valeurs de π_i^1 et π_i^2, telles que : $\pi_i^1 + \pi_i^2 > 1$ (auquel cas la borne inférieure est « active ») ou lorsque π_i^2 est sensiblement inférieur à π_i^1 (auquel cas la borne supérieure est « active »).

A nouveau, la condition de taille fixe pour S_2 impose que S_1 soit équilibré sur les variables

$$\boxed{a_i \pi_i^1 = \pi_i^1 \frac{\pi_i^2 - \alpha \pi_i^1}{1 - \pi_i^1}}.$$

On peut alors récrire la solution et la stratégie de tirage sous forme synthétique en introduisant ces valeurs dans le tableau ci-dessus.

- Une autre solution consiste à prendre : $b_i \pi_i^1 = \alpha \frac{n_1}{N}$, soit : $\boxed{b_i = \alpha \frac{n_1}{N} \frac{1}{\pi_i^1}}$, sous réserve que : $b_i \leq 1$, soit : $\alpha \frac{1}{N} \frac{n_1}{\pi_i^1} \leq 1$ (condition de même type que ci-dessus, avec : $X_{0,i} = 1$).

On en déduit ensuite que : $\boxed{a_i = \frac{\pi_i^2 - b_i \pi_i^1}{1 - \pi_i^1} = \frac{\pi_i^2 - \alpha \frac{n_1}{N}}{1 - \pi_i^1}}$, sous réserve toujours de conditions de compatibilité et d'existence : $a_i \in [0, 1]$ et : $\pi_i^1 \neq 1$.
Voir le cas de strate exhaustive pour le 1er échantillon.

Dans ce cas, **l'équilibrage requis sur S_1 pour les variables $b_i \pi_i^1$ équivaut à un équilibrage sur une constante**, ce qui donne la condition assez usuelle et relativement peu contraignante :

$$\boxed{\textbf{somme des poids de sondage de } S_1 \textbf{ = taille de l'univers} \text{ (cf. supra)}}.$$

L'équilibrage nécessaire sur les variables $a_i \pi_i^1$ ne présente pas d'interprétation particulière.

Finalement, pour ce choix des probabilités d'inclusion conditionnelles, le 2ème échantillon **présentera un taux de recouvrement α par rapport au 1er échantillon et sera équilibré sur les variables V_i** si et seulement si :

- le 1er échantillon l'est sur les variables :

$$1,\ \pi_i^1 \frac{\pi_i^2 - \alpha \frac{n_1}{N}}{1-\pi_i^1} \text{ et } \frac{\alpha \frac{n_1}{N}\frac{1}{\pi_i^1} - \pi_i^2}{1-\pi_i^1}\frac{\pi_i^1}{\pi_i^2} V_i = \frac{\alpha \frac{n_1}{N}\frac{1}{\pi_i^2} - \pi_i^1}{1-\pi_i^1} V_i$$

- le tirage conditionnel, pour la partie de l'échantillon S_2 puisée dans S_1, l'est sur les variables $\alpha \frac{n_1}{N}\frac{1}{\pi_i^1}$ et : $Z_{1,i} = \alpha \frac{n_1}{N}\frac{1}{\pi_i^1 \pi_i^2} V_i$

  (ou, plus simplement, sur les variables $\frac{1}{\pi_i^1}$ et : $Z_{1,i}^* = \frac{V_i}{\pi_i^1 \pi_i^2}$)

- le tirage conditionnel, pour la partie de l'échantillon S_2 puisée dans CS_1, l'est sur les variables $a_i = \frac{\pi_i^2 - \alpha \frac{n_1}{N}}{1-\pi_i^1}$ et $Z_{2,i} = \frac{a_i}{\pi_i^2} V_i$.

Cas particuliers : $\pi_i^1 = \pi_i^2$ ou l'un des deux tirages est un SAS

Formulation en probabilité d'appartenance aux deux échantillons

Cette formulation en termes de probabilité est plus précise d'un certain point de vue puisqu'elle porte sur l'assignation individuelle de chaque unité i aux deux échantillons et non sur la taille totale de $S_1 \cap S_2$; mais elle ne s'exprime qu'en espérance et la réalisation peut différer de la cible.

On note en effet que le nombre d'unités communes aux deux échantillons est :

$$n_{1,2}(S_1, S_2) = \sum_{i \in U} 1_{i \in S_1} 1_{i \in S_2} \text{ , d'où :}$$

$$E\, n_{1,2}(S_1, S_2) = \sum_{i \in U} E(1_{i \in S_1} 1_{i \in S_2}) = \sum_{i \in U} \mathrm{P}\{i \in S_1 \cap S_2\} = \sum_{i \in U} \gamma_i \ .$$

On peut par exemple imposer la contrainte : $\frac{E\, n_{1,2}(S_1, S_2)}{n_1} = \alpha$ (taux de recouvrement moyen, avec un échantillon S_1 de taille fixe n_1), soit : $\boxed{\sum_{i \in U} \gamma_i = \alpha n_1}$.

Avec le choix des probabilités d'inclusion conditionnelles définies dans le § 3.3, on obtient :

$$\gamma_i = \mathrm{P}\{i \in S_1 \cap S_2\} = \mathrm{E}[\mathbf{1}_{i\in S_1}\ \pi_i^{2/S_1}] = \mathrm{E}[b_i\ \mathbf{1}_{i\in S_1}] = b_i\,\pi_i^1 .$$

On doit donc satisfaire la condition : $\boxed{\sum_{i\in U} b_i\,\pi_i^1 = \alpha\, n_1}$.

On obtient exactement la même condition sur les b_i que celle obtenue dans le cas où l'on fixe la taille de $S_1 \cap S_2$. Mais la différence essentielle est que **l'on n'introduit pas de condition d'équilibrage supplémentaire sur** S_1 : cette solution est donc moins contraignante, en revanche, **elle n'assure la valeur du taux de recouvrement qu'en espérance.**

Des solutions particulières peuvent être obtenues en imposant des conditions simples aux γ_i, à partir desquelles on déterminera les a_i et b_i :

- γ_i constants : $\boxed{\gamma_i = \alpha\dfrac{n_1}{N}}$
- ou proportionnels à une caractéristique individuelle K_i (en général, un facteur de taille) :

$$\boxed{\gamma_i = \alpha\frac{n_1}{N\,\overline{K}}K_i}.$$

- on peut aussi imposer des contraintes sur la probabilité conditionnelle :

$$\mathrm{P}\{i \in S_1 \cap S_2 / i \in S_1\} = \frac{P\{i \in S_1 \cap S_2\}}{P\{i \in S_1\}} = \frac{\gamma_i}{\pi_i^1} = b_i .$$

  Intuitivement, on pourra être amené à choisir b_i comme fonction croissante d'un facteur de taille, par exemple proportionnel à une variable K_i. Compte tenu des contraintes sur les b_i, on obtient : $\boxed{b_i = \alpha\, n_1 \dfrac{K_i}{\sum_{j\in U} K_j \pi_j^1}}$.

6 Cas d'utilisation conjointe de S_1 et S_2.

Cette partie à développer ultérieurement concerne la problématique des échantillons complémentaires ou d'extension.

7 Une solution approchée : l'équilibrage « inverse[6] ».

7.1 Problématique et principe général.

On a vu dans les développements précédents que le principal obstacle au tirage d'un échantillon S_2 devant satisfaire des contraintes de taille, d'équilibrage ou de recouvrement par rapport à S_1 et dont les probabilités d'inclusion finales sont données alors que le tirage se fait conditionnellement à S_1, résidait dans les contraintes d'équilibrage imposées ex-ante à S_1. Comme il n'est pas possible de revenir sur le tirage de S_1, si ces contraintes n'ont pas été anticipées, les propriétés assignées à S_2 ne pourront pas être satisfaites.

On va chercher des solutions approchées en s'astreignant à conserver toutes les conditions d'équilibrage à imposer sur S_1 qui permettront, via le jeu de probabilités conditionnelles adéquates, d'assurer les équilibrages cherchés sur S_2, mais en relâchant la contrainte sur les probabilités finales d'inclusion de S_2.

On va donc chercher de nouvelles probabilités d'inclusion finales, notées $\tilde{\pi}_i^2$, proches des π_i^2 au sens d'une certaine distance, tout en respectant les équations d'équilibrage imposées à S_1, qui, grâce aux équilibrages appropriés lors du tirage conditionnel, permettront d'atteindre les conditions globales d'équilibrage souhaitées pour S_2.

7.2 Formalisation.

- Les contraintes sur les paramètres :

$$\pi_i^2 = a_i\,(1-\pi_i^1) + b_i\pi_i^1$$
$$\sum_{i\in U}\pi_i^2 = n_2$$
$$\sum_{i\in U} b_i\,\pi_i^1 = \alpha\, n_1$$
$$a_i \;et\; b_i \in [0,1]$$

- Les équations d'équilibrage à satisfaire sur S_1 :

$$\sum_{i\in S_1}\frac{1}{\pi_I^1}a_i\pi_i^1 = \sum_{i\in U} a_i\pi_i^1$$

[6] Dans la procédure classique d'équilibrage, on peut tirer n'importe quel échantillon avec des probabilités d'inclusion données en astreignant chaque échantillon à vérifier des contraintes d'équilibrage. Dans la présente procédure, on va procéder de manière « inverse » : on détermine les probabilités d'inclusion qui assureront les équations d'équilibrage.

$$\sum_{i\in S_1}\frac{1}{\pi_I^1}b_i\pi_i^1=\sum_{i\in U}b_i\pi_i^1$$

$$\sum_{i\in S_1}\frac{1}{\pi_I^1}(b_i-a_i)\frac{\pi_i^1}{\pi_i^2}V_i=\sum_{i\in U}(b_i-a_i)\frac{\pi_i^1}{\pi_i^2}V_i$$

- La fonction objectif à minimiser : $\sum_{i\in U}d^2(\tilde{\pi}_i^2,\pi_i^2)$, où d désigne une distance (distance euclidienne ou du χ^2 en pratique).

- Le problème : trouver les $\tilde{\pi}_i^2$ (donc les $\tilde{a}_i$ et $\tilde{b}_i$ en découlant) réalisant le minimum de la fonction objectif et satisfaisant toutes les contraintes et équations d'équilibrage relatives à S_1.

La résolution de ce programme va conduire à déterminer de nouvelles probabilités d'inclusion finales $\tilde{\pi}_i^2$ proches des probabilités imposées initialement π_i^2, **mais la solution va dépendre de** S_1 : on devrait noter les nouvelles probabilités obtenues $\tilde{\pi}_i^2(S_1)$. A partir de là, on déterminera de nouvelles probabilités d'inclusion conditionnelles $\tilde{\pi}_i^{2/S_1}$ et, en tirant l'échantillon S_2 conditionnellement à S_1 selon ces probabilités, avec les conditions d'équilibrage appropriées, on pourra obtenir une « pseudo-condition d'équilibrage » sur S_2 utilisant les $\tilde{\pi}_i^2$ au lieu des π_i^2. ***D'où le nom « d'équilibrage inverse » que l'on peut attribuer à la procédure.***

Les « vraies » probabilités d'inclusion de S_2 seront alors : $\pi_i^{*2}=E\tilde{\pi}_i^{2/S_1}$. Elles ne seront pas égales aux π_i^2, mais la procédure devrait assurer qu'elles en soient proches.

Cette approche pose toutefois la question de l'existence de solutions admissibles (avec tous les $\tilde{\pi}_i^2$ dans]0, 1[) et peut présenter des difficultés de résolution numérique de programmes mathématiques d'optimisation complexes.

Finalement, on sera confronté à l'alternative suivante :

- soit faire un tirage équilibré de S_2 avec des conditions de taille et de recouvrement mais sans respecter exactement des probabilités d'inclusion fixées et en acceptant une modification des probabilités de sélection finale des unités.

- soit renoncer aux contraintes d'équilibrage et de recouvrement pour S_2, tout en procédant à un tirage à probabilités finales données.

7.3 1er Exemple de résolution.

On suppose que l'on ne cherche pas de condition d'équilibrage sur S_2 autre que celles assurant la taille fixe et le recouvrement par rapport à S_1 et l'on prend une **distance euclidienne**.

On prend la solution du § 3.3 e), avec $\boxed{b_i = \alpha \text{ , } a_i = \frac{\pi_i^2 - \alpha \pi_i^1}{1 - \pi_i^1}}$.

Dans ce cas, la condition d'équilibrage de S_1 sur les variables $b_i \pi_i^1$ est automatiquement satisfaite puisqu'elle s'écrit : $\sum_{i \in S_1} \frac{1}{\pi_i^1} b_i \pi_i^1 = \sum_{i \in U} b_i \pi_i^1$, soit : $\sum_{i \in S_1} \alpha = \alpha \underbrace{\sum_{i \in U} \pi_i^1}_{=n_1}$, soit : $\alpha n_1 = \alpha n_1$.

Il faut donc seulement assurer la condition d'équilibrage de S_1 sur les variables :

$$a_i \pi_i^1 = \pi_i^1 \frac{\pi_i^2 - \alpha \pi_i^1}{1 - \pi_i^1}.$$

On note que :

$$\begin{aligned} \sum_{i \in U} a_i \pi_i^1 &= \sum_{i \in U} a_i (\pi_i^1 - 1) + \sum_{i \in U} a_i \\ &= -\sum_{i \in U} \pi_i^2 + \alpha \sum_{i \in U} \pi_i^1 + \sum_{i \in U} a_i \\ &= \sum_{i \in U} \frac{\pi_i^2 - \alpha \pi_i^1}{1 - \pi_i^1} - (n_2 - \alpha n_1). \end{aligned}$$

On doit donc résoudre le programme d'optimisation :

$$Min \sum_{i \in U} (\tilde{\pi}_i^2 - \pi_i^2)^2 \text{ ,}$$

$$\text{sous les contraintes : } \begin{cases} \sum_{i \in U} \tilde{\pi}_i^2 = n_2 \\ \sum_{i \in S_1} \frac{1}{\pi_i^1} \pi_i^1 \frac{\tilde{\pi}_i^2 - \alpha \pi_i^1}{1 - \pi_i^1} = \sum_{i \in U} \frac{\tilde{\pi}_i^2 - \alpha \pi_i^1}{1 - \pi_i^1} - (n_2 - \alpha n_1) \end{cases}.$$

On peut récrire la 2ème contrainte sous la forme :

$$\sum_{i \in S_1} \frac{\tilde{\pi}_i^2}{1 - \pi_i^1} - \alpha \sum_{i \in S_1} \frac{\pi_i^1}{1 - \pi_i^1} = \sum_{i \in U} \frac{\tilde{\pi}_i^2}{1 - \pi_i^1} - \alpha \sum_{i \in U} \frac{\pi_i^1}{1 - \pi_i^1} - (n_2 - \alpha n_1) \text{ , soit :}$$

$$\boxed{\sum_{i \notin S_1} \frac{\tilde{\pi}_i^2}{1 - \pi_i^1} - \alpha \sum_{i \notin S_1} \frac{\pi_i^1}{1 - \pi_i^1} = n_2 - \alpha n_1}.$$

$\Rightarrow$ Lagrangien :

$$L = \sum_{i\in U}(\tilde{\pi}_i^2 - \pi_i^2)^2 - \lambda\left(\sum_{i\notin S_1}\frac{\tilde{\pi}_i^2}{1-\pi_i^1} - \alpha\sum_{i\notin S_1}\frac{\pi_i^1}{1-\pi_i^1} - (n_2 - \alpha n_1)\right) - \mu(\sum_{i\in U}\tilde{\pi}_i^2 - n_2)\,.$$

D'où :

$$\frac{\partial L}{\partial \tilde{\pi}_i^2} = 2(\tilde{\pi}_i^2 - \pi_i^2) - \lambda\frac{1}{1-\pi_i^1}1_{i\notin S_1} - \mu\,.$$

Les conditions du 1er ordre donnent donc : $\boxed{\tilde{\pi}_i^2 = \pi_i^2 + \frac{\lambda}{2}\frac{1}{1-\pi_i^1}1_{i\notin S_1} + \frac{\mu}{2}}$.

Par suite :

- $\underbrace{\sum_{i\in U}\tilde{\pi}_i^2}_{=n_2} = \underbrace{\sum_{i\in U}\pi_i^2}_{=n_2} + \frac{\lambda}{2}\sum_{i\in U}\frac{1}{1-\pi_i^1}1_{i\notin S_1} + \frac{\mu}{2}N$, soit : $\lambda\sum_{i\in U}\frac{1}{1-\pi_i^1}1_{i\notin S_1} + \mu N = 0$, ou :

$$\boxed{\mu = -\frac{\lambda}{N}\sum_{i\notin S_1}\frac{1}{1-\pi_i^1}}.$$

- $\sum_{i\notin S_1}\frac{\tilde{\pi}_i^2}{1-\pi_i^1} = \sum_{i\notin S_1}\frac{\pi_i^2}{1-\pi_i^1} + \frac{\lambda}{2}\sum_{i\notin S_1}\frac{1}{(1-\pi_i^1)^2} + \frac{\mu}{2}\sum_{i\notin S_1}\frac{1}{1-\pi_i^1}$.

  En reportant cette valeur dans la contrainte et en utilisant la 1ère relation entre λ et μ, on obtient :

  $\sum_{i\notin S_1}\frac{\pi_i^2}{1-\pi_i^1} + \frac{\lambda}{2}\left[\sum_{i\notin S_1}\frac{1}{(1-\pi_i^1)^2} - \frac{1}{N}\left(\sum_{i\notin S_1}\frac{1}{1-\pi_i^1}\right)^2\right] - \alpha\sum_{i\notin S_1}\frac{\pi_i^1}{1-\pi_i^1} = n_2 - \alpha n_1$, d'où :

$$\frac{\lambda}{2} = \frac{n_2 - \alpha n_1 - \sum_{i\notin S_1}\frac{\pi_i^2 - \alpha\pi_i^1}{1-\pi_i^1}}{\sum_{i\notin S_1}\frac{1}{(1-\pi_i^1)^2} - \frac{1}{N}\left(\sum_{i\notin S_1}\frac{1}{1-\pi_i^1}\right)^2}\,.$$

Finalement :

$$\boxed{\tilde{\pi}_i^2(S_1) = \pi_i^2 + \left(\frac{1_{i\notin S_1}}{1-\pi_i^1} - \frac{1}{N}\sum_{j\notin S_1}\frac{1}{1-\pi_j^1}\right)\frac{n_2 - \alpha n_1 - \sum_{j\notin S_1}\frac{\pi_j^2 - \alpha\pi_j^1}{1-\pi_j^1}}{\sum_{j\notin S_1}\frac{1}{(1-\pi_j^1)^2} - \frac{1}{N}\left(\sum_{j\notin S_1}\frac{1}{1-\pi_j^1}\right)^2}}.$$

Analyse de cette formule :

Les probabilités π_i^2 sont translatées d'une quantité fixe pour toutes les unités de S_1. Pour celles hors S_1, cette translation fixe est modifiée par une autre translation en sens inverse qui dépend de π_i^1 (via le facteur $\frac{1}{1-\pi_i^1}$).

Le signe de ces translations n'est pas prédéterminé. En effet :

- Le dénominateur $\sum_{j\notin S_1} \frac{1}{(1-\pi_j^1)^2} - \frac{1}{N}\left(\sum_{j\notin S_1}\frac{1}{1-\pi_j^1}\right)^2$ est **positif**.

  En effet, d'après l'inégalité de SCHWARZ :

$$\left(\sum_{j\notin S_1}\frac{1}{1-\pi_j^1}\right)^2 \le \left(\sum_{j\notin S_1} 1\right)\left(\sum_{j\notin S_1}\frac{1}{(1-\pi_j^1)^2}\right) = (N-n_1)\left(\sum_{j\notin S_1}\frac{1}{(1-\pi_j^1)^2}\right) \le N\left(\sum_{j\notin S_1}\frac{1}{(1-\pi_j^1)^2}\right).$$

- Mais le numérateur $n_2 - \alpha n_1 - \sum_{j\notin S_1}\frac{\pi_j^2-\alpha\pi_j^1}{1-\pi_j^1}$ est **d'espérance nulle** car :

$$E\left(\sum_{j\notin S_1}\frac{\pi_j^2-\alpha\pi_j^1}{1-\pi_j^1}\right) = \sum_{j\in U}\frac{\pi_j^2-\alpha\pi_j^1}{1-\pi_j^1}(1-\pi_j^1) = \sum_{j\in U}(\pi_j^2-\alpha\pi_j^1) = \sum_{j\in U}\pi_j^2 - \alpha\sum_{j\in U}\pi_j^1 = n_2-\alpha n_1.$$

  Il peut donc prendre des valeurs positives ou négatives selon les tirages de S_1.

La question de savoir si ces perturbations sont « *marginales* » (c'est-à-dire si les $\tilde{\pi}_i^2(S_1)$ sont proches des π_i^2 et dans quel sens elles en diffèrent) est difficile à vérifier analytiquement et sera étudiée empiriquement au moyen de simulations sur des exemples réels.

On notera que $\tilde{\pi}_i^2(S_1) = \pi_i^2,\ \forall i \in U$ si et seulement si : $n_2 - \alpha n_1 - \sum_{j\notin S_1}\frac{\pi_j^2-\alpha\pi_j^1}{1-\pi_j^1} = 0$, ce qui fournit une valeur α « optimale » (dépendant de S_1) : $\alpha^*(S_1) = -\frac{C(S_1)}{B(S_1)}$, soit :

$$\alpha^*(S_1) = -\frac{n_2 - \sum_{j\notin S_1}\frac{\pi_j^2}{1-\pi_j^1}}{-n_1 + \sum_{j\notin S_1}\frac{\pi_j^1}{1-\pi_j^1}} = \frac{n_2 - \sum_{j\notin S_1}\frac{\pi_j^2}{1-\pi_j^1}}{n_1 - \sum_{j\notin S_1}\frac{\pi_j^1}{1-\pi_j^1}}$$

(sous réserve que cette valeur soit dans]0, 1[).

Naturellement, les solutions $\tilde{\pi}_i^2(S_1)$, qui sont des probabilités, doivent appartenir à [0, 1]. Si ces contraintes à l'inégalité ne sont pas explicitement introduites dans le programme de minimisation, elles doivent être vérifiées ex-post. Compte tenu du caractère aléatoire de l'écart $\tilde{\pi}_i^2(S_1) - \pi_i^2$, la violation de ces contraintes peut survenir selon les réalisations du tirage de S_1. **Il en résulte des contraintes (aléatoires) sur la plage de variation admissible de α, qu'il faudra déterminer par itérations**.

Calcul de la distance résultante.

La formule donnant $\tilde{\pi}_i^2(S_1)$ est de la forme :

$$\tilde{\pi}_i^2(S_1) = \pi_i^2 + A_i(S_1)\frac{\alpha B(S_1) + C(S_1)}{D(S_1)}.$$

D'où :

$$[\tilde{\pi}_i^2(S_1) - \pi_i^2]^2 = [A_i(S_1)]^2\left[\frac{\alpha B(S_1) + C(S_1)}{D(S_1)}\right]^2$$

et :

$$\boxed{\sum_{i\in U}[\tilde{\pi}_i^2(S_1) - \pi_i^2]^2 = \left[\frac{\alpha B(S_1) + C(S_1)}{D(S_1)}\right]^2 \sum_{i\in U}[A_i(S_1)]^2}.$$

On notera le caratère quadratique de cette distance, considérée comme fonction de α. Pour une réalisation de S_1, il existe une valeur « optimale » (dépendant de S_1) qui annule cette distance (cf. supra). SI l'on prend l'espérance par rapport au tirage aléatoire de S_1, on obtiendra :

$$E\left[\sum_{i\in U}[\tilde{\pi}_i^2(S_1) - \pi_i^2]^2\right] =$$
$$\alpha^2 E\left[\frac{B^2(S_1)}{D^2(S_1)}\sum_{i\in U}[A_i(S_1)]^2\right] + 2\alpha E\left[\frac{B(S_1)C(S_1)}{D(S_1)}\sum_{i\in U}[A_i(S_1)]^2\right] + E\left[\frac{C^2(S_1)}{D^2(S_1)}\sum_{i\in U}[A_i(S_1)]^2\right],$$

ce qui met en évidence **l'existence d'une valeur de α réalisant le minimum de la distance « moyenne »** $E\left[\frac{1}{N}\sum_{i\in U}[\tilde{\pi}_i^2(S_1) - \pi_i^2]^2\right]$**.**

Dépendance de $\tilde{\pi}_i^2(S_1)$ par rapport à α.

On note que $\tilde{\pi}_i^2(S_1)$ est une *fonction linéaire* de α. Le coefficient de dépendance (aléatoire) vaut :

$$\frac{\partial\tilde{\pi}_i^2(S_1)}{\partial\alpha} = \sum_{j\notin S_1}\frac{\pi_j^1}{1-\pi_j^1} - n_1.$$

Il est **d'espérance nulle** puisque :

$$E\left(\sum_{j\notin S_1}\frac{\pi_j^1}{1-\pi_j^1}\right)=\sum_{j\in U}\frac{\pi_j^1}{1-\pi_j^1}(1-\pi_j^1)=\sum_{j\in U}\pi_j^1=n_1.$$

Son signe peut donc varier selon les réalisations du tirage de S_1.

La détermination de $\tilde{\pi}_i^2(S_1)$ permet ensuite d'en déduire : $\begin{cases}\tilde{b}_i=\alpha\\ \tilde{a}_i=\dfrac{\tilde{\pi}_i^2(S_1)-\alpha\pi_i^1}{1-\pi_i^1}\end{cases}$, d'où :

$$\boxed{\tilde{\pi}_i^{2/S_1}=\alpha 1_{i\in S_1}+\frac{\tilde{\pi}_i^2(S_1)-\alpha\pi_i^1}{1-\pi_i^1}1_{i\notin S_1}.}$$

Il conviendra de vérifier, à nouveau empiriquement, que les coefficients obtenus $\tilde{a}_i$ sont bien admissibles (à valeurs dans [0, 1]), ce qui conduira à de nouvelles contraintes sur α.

Enfin, les vraies probabilités finales d'inclusion dans S_2 seront :

$$\begin{aligned}\pi_i^{*2}&=E\,\tilde{\pi}_i^{2/S_1}\\&=\alpha E(1_{i\in S_1})+E\left(\frac{\tilde{\pi}_i^2(S_1)-\alpha\pi_i^1}{1-\pi_i^1}1_{i\notin S_1}\right)\\&=\alpha\pi_i^1-\frac{\alpha\pi_i^1}{1-\pi_i^1}(1-\pi_i^1)+E\left(\frac{\tilde{\pi}_i^2(S_1)}{1-\pi_i^1}1_{i\notin S_1}\right),\end{aligned}$$

soit :

$$\boxed{\pi_i^{*2}=E\left(\frac{\tilde{\pi}_i^2(S_1)}{1-\pi_i^1}1_{i\notin S_1}\right)=\frac{1}{1-\pi_i^1}E\left(\tilde{\pi}_i^2(S_1)1_{i\notin S_1}\right).}$$

Cette quantité est difficile à calculer explicitement.

On peut l'écrire dans le cas présent :

$$\boxed{\pi_i^{*2}(S_1)=\pi_i^2+\frac{1}{1-\pi_i^1}E\left[1_{i\notin S_1}\left(\frac{1}{1-\pi_i^1}-\frac{1}{N}\sum_{j\notin S_1}\frac{1}{1-\pi_j^1}\right)\frac{n_2-\alpha n_1-\sum_{j\notin S_1}\dfrac{\pi_j^2-\alpha\pi_j^1}{1-\pi_j^1}}{\sum_{j\notin S_1}\dfrac{1}{(1-\pi_j^1)^2}-\dfrac{1}{N}\left(\sum_{j\notin S_1}\dfrac{1}{1-\pi_j^1}\right)^2}\right].}$$

7.4 2ème exemple de résolution.

On se place sous les mêmes hypothèses que dans le § 5.3 mais on prend la **distance du χ^2 .**

On doit donc résoudre le programme d'optimisation :

$$Min \sum_{i\in U} \frac{(\tilde{\pi}_i^2 - \pi_i^2)^2}{\pi_i^2},$$

sous les mêmes contraintes que précédemment : $\boxed{\begin{cases} \sum_{i\in U} \tilde{\pi}_i^2 = n_2 \\ \sum_{i\notin S_1} \frac{\tilde{\pi}_i^2}{1-\pi_i^1} - \alpha \sum_{i\notin S_1} \frac{\pi_i^1}{1-\pi_i^1} = n_2 - \alpha n_1 \end{cases}}$.

On note que : $\sum_{i\in U} \frac{(\tilde{\pi}_i^2 - \pi_i^2)^2}{\pi_i^2} = \sum_{i\in U} \frac{(\tilde{\pi}_i^2)^2}{\pi_i^2} - 2\sum_{i\in U} \frac{\tilde{\pi}_i^2 \pi_i^2}{\pi_i^2} + \sum_{i\in U} \frac{(\pi_i^2)^2}{\pi_i^2}$

$$= \sum_{i\in U} \frac{(\tilde{\pi}_i^2)^2}{\pi_i^2} - 2\underbrace{\sum_{i\in U} \tilde{\pi}_i^2}_{=n_2} + \underbrace{\sum_{i\in U} \pi_i^2}_{=n_2}$$

$$= \sum_{i\in U} \frac{(\tilde{\pi}_i^2)^2}{\pi_i^2} - n_2 .$$

La fonction à minimiser sera donc : $\sum_{i\in U} \frac{(\tilde{\pi}_i^2)^2}{\pi_i^2}$.

$\Rightarrow$ Lagrangien : $L = \sum_{i\in U} \frac{(\tilde{\pi}_i^2)^2}{\pi_i^2} - \lambda\left(\sum_{i\notin S_1} \frac{\tilde{\pi}_i^2}{1-\pi_i^1} - \alpha \sum_{i\notin S_1} \frac{\pi_i^1}{1-\pi_i^1} - (n_2 - \alpha n_1) \right) - \mu(\sum_{i\in U} \tilde{\pi}_i^2 - n_2)$.

D'où :

$$\frac{\partial L}{\partial \tilde{\pi}_i^2} = 2\frac{\tilde{\pi}_i^2}{\pi_i^2} - \lambda \frac{1}{1-\pi_i^1} 1_{i\notin S_1} - \mu .$$

Les conditions du 1er ordre donnent donc : $\boxed{\tilde{\pi}_i^2 = \pi_i^2 \left(\frac{\lambda}{2} \frac{1}{1-\pi_i^1} 1_{i\notin S_1} + \frac{\mu}{2}\right)}$.

Par suite :

- $\underbrace{\sum_{i\in U} \tilde{\pi}_i^2}_{=n_2} = \frac{\lambda}{2} \sum_{i\in U} \frac{\pi_i^2}{1-\pi_i^1} 1_{i\notin S_1} + \frac{\mu}{2} \underbrace{\sum_{i\in U} \pi_i^2}_{=n_2}$, soit : $n_2 = \frac{\lambda}{2} \sum_{i\in U} \frac{\pi_i^2}{1-\pi_i^1} 1_{i\notin S_1} + \frac{\mu}{2} n_2$, ou :

$$\boxed{\frac{\mu}{2} = 1 - \frac{\lambda}{2n_2} \sum_{i \notin S_1} \frac{\pi_i^2}{1-\pi_i^1}}.$$

- $$\sum_{i \notin S_1} \frac{\tilde{\pi}_i^2}{1-\pi_i^1} = \frac{\lambda}{2} \sum_{i \notin S_1} \frac{\pi_i^2}{(1-\pi_i^1)^2} + \frac{\mu}{2} \sum_{i \notin S_1} \frac{\pi_i^2}{1-\pi_i^1}.$$

  En reportant cette valeur dans la 2ème contrainte et en utilisant la 1ère relation obtenue entre λ et μ, on obtient :

  $$\frac{\lambda}{2} \sum_{i \notin S_1} \frac{\pi_i^2}{(1-\pi_i^1)^2} + (1 - \frac{\lambda}{2n_2} \sum_{i \notin S_1} \frac{\pi_i^2}{1-\pi_i^1})(\sum_{i \notin S_1} \frac{\pi_i^2}{1-\pi_i^1}) - \alpha \sum_{i \notin S_1} \frac{\pi_i^1}{1-\pi_i^1} = n_2 - \alpha n_1, \text{ soit :}$$

  $$\frac{\lambda}{2} \left[\sum_{i \notin S_1} \frac{\pi_i^2}{(1-\pi_i^1)^2} - \frac{1}{n_2} \left(\sum_{i \notin S_1} \frac{\pi_i^2}{1-\pi_i^1} \right)^2 \right] + \sum_{i \notin S_1} \frac{\pi_i^2}{1-\pi_i^1} - \alpha \sum_{i \notin S_1} \frac{\pi_i^1}{1-\pi_i^1} = n_2 - \alpha n_1, \text{ soit :}$$

  $$\frac{\lambda}{2} = \frac{n_2 - \alpha n_1 - \sum_{i \notin S_1} \frac{\pi_i^2 - \alpha \pi_i^1}{1-\pi_i^1}}{\sum_{i \notin S_1} \frac{\pi_i^2}{(1-\pi_i^1)^2} - \frac{1}{n_2} \left(\sum_{i \notin S_1} \frac{\pi_i^2}{1-\pi_i^1} \right)^2}.$$

Finalement :

$$\begin{aligned} \tilde{\pi}_i^2(S_1) &= \pi_i^2 \left(\frac{\lambda}{2} \frac{1}{1-\pi_i^1} 1_{i \notin S_1} + 1 - \frac{\lambda}{2n_2} \sum_{j \notin S_1} \frac{\pi_j^2}{1-\pi_j^1} \right) \\ &= \pi_i^2 \left(1 + \frac{\lambda}{2} \left(\frac{1_{i \notin S_1}}{1-\pi_i^1} - \frac{1}{n_2} \sum_{j \notin S_1} \frac{\pi_j^2}{1-\pi_j^1} \right) \right), \end{aligned}$$

soit :

$$\boxed{\tilde{\pi}_i^2(S_1) = \pi_i^2 \left[1 + \frac{n_2 - \alpha n_1 - \sum_{j \notin S_1} \frac{\pi_j^2 - \alpha \pi_j^1}{1-\pi_j^1}}{\sum_{j \notin S_1} \frac{\pi_j^2}{(1-\pi_j^1)^2} - \frac{1}{n_2} \left(\sum_{j \notin S_1} \frac{\pi_j^2}{1-\pi_j^1} \right)^2} \left(\frac{1_{i \notin S_1}}{1-\pi_i^1} - \frac{1}{n_2} \sum_{j \notin S_1} \frac{\pi_j^2}{1-\pi_j^1} \right) \right]}.$$

A nouveau, il faudra vérifier empiriquement si les probabilités approchées $\tilde{\pi}_i^2$ sont proches des probabilités de référence π_i^2.

On en déduit ensuite, comme précédemment, les valeurs des probabilités conditionnelles de tirage $\tilde{a}_i$ et $\tilde{b}_i$ et il conviendra de vérifier à nouveau empiriquement que les coefficients obtenus $\tilde{a}_i$ sont bien admissibles (à valeurs dans [0, 1]), ce qui peut entraîner des contraintes sur α .

Exemple : cas où S_1 est aléatoire simple.

7.5 Généralisation.

Le cas où l'on introduit d'autres conditions d'équilibrage ne permet plus de déterminer de solutions analytiques. En effet, pour un équilibrage sur des variables V_i, on rajoute aux contraintes précédentes la relation : $\sum_{i\in S_1} \frac{1}{\pi_i^1}(b_i - a_i)\frac{\pi_i^1}{\pi_i^2}V_i = \sum_{i\in U} (b_i - a_i)\frac{\pi_i^1}{\pi_i^2}V_i$, soit :

$$\sum_{i\in S_1} \frac{1}{\pi_i^1}\frac{\alpha-\pi_i^2}{1-\pi_i^1}\frac{\pi_i^1}{\pi_i^2}V_i = \sum_{i\in U} \frac{\alpha-\pi_i^2}{1-\pi_i^1}\frac{\pi_i^1}{\pi_i^2}V_i \text{, soit :}$$

$$\alpha\sum_{i\in S_1} \frac{V_i}{(1-\pi_i^1)\pi_i^2} - \sum_{i\in S_1} \frac{V_i}{(1-\pi_i^1)} = \alpha\sum_{i\in U} \frac{\pi_i^1 V_i}{(1-\pi_i^1)\pi_i^2} - \sum_{i\in U} \frac{\pi_i^1 V_i}{(1-\pi_i^1)}.$$

La présence d'un terme π_i^2 au dénominateur (remplacé par $\tilde{\pi}_i^2$ dans le programme d'optimisation) fera apparaître un terme en $\frac{1}{(\tilde{\pi}_i^2)^2}$ dans les conditions du 1er ordre issues de la dérivation du lagrangien, ce qui conduira à une équation du 3ème degré en $\tilde{\pi}_i^2$. A moins d'employer les formules de CARDAN, seules des solutions numériques, par itérations successives, pourront être obtenues.

7.6 Calcul d'estimateurs.

On revient au cas général.

Une fois réalisé le tirage de S_2 conditionnellement à S_1, on peut proposer, pour estimer un total $T(Y)$, les estimateurs suivants :

a) $\boxed{\hat{T}_1(Y) = \sum_{i\in S_2} \frac{Y_i}{\pi_i^2}}$.

Cet estimateur est l'estimateur de HORVITZ-THOMSON naïf, utilisant les valeurs des probabilités-cibles π_i^2. Mais, du fait que l'échantillon S_2 n'a pas été tiré avec ces probabilités, cet estimateur est très certainement **biaisé**.

b) $\boxed{\hat{T}_2(Y) = \sum_{i \in S_2} \frac{Y_i}{\tilde{\pi}_i^2(S_1)}}$.

Cet estimateur est un estimateur de type HORVITZ-THOMSON, mais utilisant les valeurs effectives des probabilités d'inclusion, $\tilde{\pi}_i^2(S_1)$. Celles-ci sont *aléatoires* ; de ce fait, les propriétés habituelles des estimateurs de HORVITZ-THOMSON ne se retrouvent plus et l'estimateur est très certainement **biaisé**.

c) $\boxed{\hat{T}_3(Y) = \sum_{i \in S_2} \frac{Y_i}{\pi_i^{*2}}}$, avec : $\boxed{\pi_i^{*2} = \frac{1}{1-\pi_i^1} E\left(\tilde{\pi}_i^2(S_1) 1_{i \notin S_1}\right)}$.

Cet estimateur est l'estimateur de HORVITZ-THOMSON habituel (à coefficients *fixes*) utilisant les vraies probabilités d'inclusion π_i^{*2}. Malheureusement celles-ci sont très difficiles à calculer explicitement.

d) $\boxed{\hat{T}_4(Y) = \sum_{i \in S_2 \cap S_1} \frac{Y_i}{\tilde{b}_i(S_1)} + \sum_{i \in S_2 \cap CS_1} \frac{Y_i}{\tilde{a}_i(S_1)}}$,

avec : $\tilde{\pi}_i^{2/S_1}(S_1) = \tilde{a}_i(S_1) 1_{i \notin S_1} + \tilde{b}_i(S_1) 1_{i \in S_1}$.

Cet estimateur (à coefficients *aléatoires*) sera sans biais.

En effet :

$$E[\hat{T}_4(Y)/S_1] = E\left(\sum_{i \in S_2 \cap S_1} \frac{Y_i}{\tilde{b}_i(S_1)} + \sum_{i \in S_2 \cap CS_1} \frac{Y_i}{\tilde{a}_i(S_1)} \right) / S_1$$

$$= E\left(\sum_{i \in S_1} \frac{Y_i 1_{i \in S_2}}{\tilde{b}_i(S_1)} + \sum_{i \in CS_1} \frac{Y_i 1_{i \in S_2}}{\tilde{a}_i(S_1)} \right) / S_1$$

$$= \sum_{i \in S_1} \frac{Y_i}{\tilde{b}_i(S_1)} \underbrace{E(1_{i \in S_2}/S_1)}_{=\tilde{b}_i(S_1) \ pour \ i \in S_1} + \sum_{i \in CS_1} \frac{Y_i}{\tilde{a}_i(S_1)} \underbrace{E(1_{i \in S_2}/S_1)}_{=\tilde{a}_i(S_1) \ pour \ i \in CS_1}$$

$$= \sum_{i \in S_1} Y_i + \sum_{i \notin S_1} Y_i$$

$$= T(Y).$$

Cet estimateur est sans biais conditionnellement à S_1, quelle que soit la valeur de S_1.

On en déduit :

$$\boxed{E\,\hat{T}_4(Y) = E\,E[\hat{T}_4(Y)/S_1] = E\,T(Y) = T(Y)}.$$

8 Conclusion et futurs travaux.

- difficulté à imposer des conditions à un échantillon donné prenant en compte un échantillon tiré antérieurement. Pas de solution exacte, des solutions approchées possibles.

- nécessité et utilité de tester toutes ces procédures, y compris dans leurs aspects numériques

- en déduire des lignes directrices pour le futur : un échantillon tiré à un moment donné doit anticiper d'éventuels tirages ultérieurs et les contraintes qui s'y appliqueront, pour les intégrer, via des conditions d'équilibrage appropriées.

- des extensions de la méthode à étudier dans un certain nombre de cas :

    - stratification de l'univers de référence
    - cas des strates exhaustives.
    - prise en compte de modifications temporelles de l'univers de référence.

- vérifier les propriétés théoriques de ces solutions approchées (notamment, calculer les vraies probabilités d'inclusion finales du 2nd échantillon lorsque l'on les approche par des probabilités dépendant du 1er échantillon).

- tester d'autres formulations des probabilités d'inclusion conditionnelles, avec des facteurs aléatoires dépendant de l'échantillon S_1.

- calculs de variance des estimateurs

- extension à des estimateurs et des équations d'équilibrage à coefficients aléatoires.

La réalisation de simulations à partir des bases d'établissements de la Depp pour le tirage d'échantillons en vue de la réalisation d'enquêtes d'évaluation de type PISA, permettra de tester empiriquement la validité de la méthode, les difficultés numériques de mise en œuvre du calcul des solutions ainsi que les propriétés statistiques des différents estimateurs proposés.

■ ■ ■ ■ ■ ■ ■