# 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

*Marc CHRISTINE (Insee, DCSRI)*

*Thierry ROCHER (Depp)*

XI èmes JMS Insee
Paris
24 - 26 Janvier 2012

# Plan de la présentation

1. Fondements du problème
2. Cadre Général
3. Cas particulier des enquêtes d'évaluation
4. Une méthode proposée pour trouver des solutions approchées : « *l'équilibrage inverse* »
5. Etudes par simulations dans le cadre des enquêtes d'évaluation de la Depp
6. Remarques finales, conclusion et travaux à venir

# 1. Fondements du problème.

- A l'origine : une expertise sur les échantillons PISA.
- PISA : Programme for International Student Assessment.
- Depuis 2000, une évaluation tous les 3 ans des élèves de 15 ans en lecture, mathématiques et sciences.
- En 2009, 470 000 élèves échantillonnés dans 65 pays ou régions.
- Problème pour 2012 : PISA 2012 portera sur la même matière principale qu'en 2003 (mathématiques).

=> On cherchera à faire des ***comparaisons de résultats entre les deux enquêtes***.

- Pour cela, il est souhaitable d'enquêter une partie des mêmes écoles qu'en 2003 ...

- … avec un nombre ou une proportion donnés d'écoles appartenant aux deux échantillons

  ⇔ ***Conditions de recouvrement.***

- Mais ***le nouvel échantillon doit aussi satisfaire des propriétés fixées :*** taille, probabilités d'inclusion, « représentativité » (i.e., être un « modèle réduit » de la population de référence 2012).

- Le problème devient : comment tirer un nouvel échantillon :
  - dans une base de sondage mise à jour,
  - respectant des conditions telles que données ci-dessus
  - sans avoir la possibilité de modifier le tirage du 1er échantillon ?

- **Un cadre général a été développé pour résoudre ce problème…**

- … mais il s'agit d'une approche expérimentale qui n'a pas encore été mise en œuvre dans des situations réelles.

- Il peut s'appliquer à une large palette de situations : sondage en 2 phases, coordination *négative* ou *positive*, échantillons de réserve, échantillons pour des extensions régionales…

- Il couvre le champ :

  - des enquêtes Entreprises
  - du tirage d'Unités Primaires géographiques pour les enquêtes Ménages
  - d'autres enquêtes tirées avec des probabilités proportionnelles à la taille (évaluation des élèves…).

- Les principaux outils :
  - Echantillons conditionnels successifs
  - Echantillonnage équilibré (méthode du CUBE)

- L'objectif est de montrer que :

  - Il est difficile de satisfaire simultanément l'intégralité des contraintes
  - et il n'existe pas de solution exacte…
  - … mais que l'on peut obtenir des solutions *approchées*.

# 2. Cadre général

Univers = population de référence = base de sondage : $U$ .
Individus (unités statistiques) : notés $i$ .
Taille de l'univers = $N$ .
Variables d'intérêt (**non aléatoires**) définies sur chaque individu $i$ : $Y_i$, 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 $i$ .

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

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

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

Le 2nd échantillon sera tiré conditionnellement au 1er avec les caractéristiques suivantes :

- **sans remise**
- l'univers dans lequel on tire peut dépendre de $S_1$ : $U(S_1)$, de taille $N_2(S_1)$
- $U(S_1)$ représente la base de sondage de $S_2$
- taille, également potentiellement dépendante de $S_1$ : $n_2(S_1)$.

## Principaux cas de figure.

| **Base de sondage pour $S_2$ ↓** / **Utilisation de $S_2$ →** | **$S_2$ seul**<br>**Conditions d'équilibrage sur $S_2$**<br>**Eventuellement : $S_2$ ou $S_1 \cap S_2$ de taille fixe** | **Utilisation conjointe avec $S_1$ : $S = S_1 \cup S_2$ .**<br>**Conditions d'équilibrage sur $S$**<br>**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 $S_1$<br>Echantillon d'actualisation, panel avec conditions de recouvrement (cas PISA), coordination *positive* | Echantillon complémentaire avec recouvrement |

# 3. Cas particulier des enquêtes d'évaluation.

- $U(S_1) = U$
- $S_2$ est **utilisé seul.**

- avec des probabilités d'inclusion (inconditionnelles) fixées, notées $\pi_i^2$.

- satisfaisant des conditions d'équilibrage sur des variables $V_i$ données (éventuellement vectorielles) :

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

- il y a des conditions de **recouvrement** par rapport à $S_1$.

  Elles peuvent être écrites avec un **taux de recouvrement** :

$$\text{Card}\,(S_1 \cap S_2) = \alpha\ \text{Card}\,(S_1)$$

  Une fraction $\alpha$ de l'échantillon $S_1$ se retrouve sélectionnée dans $S_2$.

- éventuellement, de taille fixe imposée $n_2 \Rightarrow \sum_{i \in U} \pi_i^2 = n_2$.

# Comment tirer $s_2$ ?

**On définit :**

- des probabilités d'***inclusion conditionnelle*** pour chaque unité $i$ :

$$\mathrm{P}\{i \in S_2 / S_1\} = \pi_i^{2/S_1} \in [0, 1].$$

Si l'échantillon $S_2$ est de taille fixe $n_2$, ***non aléatoire*** :

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

- d'éventuelles **conditions d'équilibrage** sur des variables $Z_i$ :

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

# Comment tirer $s_2$ ?

**On a la latitude de choisir n'importe quelle probabilité d'inclusion conditionnelle.**

On prend ici la formulation suivante :

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

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

Relation fondamentale :

$$\pi_i^2 = E\pi_i^{2/S_1}$$

$$\Leftrightarrow$$

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

# Les difficultés :

- On va montrer que les contraintes imposées à $S_2$ ne peuvent être satisfaites ….,
- … à moins que des conditions spécifiques n'aient été imposées sur $S_1$ lors de son tirage…
- … et que cela implique des relations entre les paramètres.

## a) Comment satisfaire simultanément les conditions d'équilibrage sur $S_2$ et le respect des probabilités d'inclusion finales ?

- **Pour le tirage conditionnel de $S_2$,** on stratifie les unités selon leur appartenance ou pas à $S_1$, et on tire deux sous-échantillons équilibrés, l'un dans $S_1$, l'autre dans $CS_1$, avec :
  - équilibrage sur des variables $Z_{1,i}$ dans $S_1$
  - équilibrage sur des variables $Z_{2,i}$ dans $CS_1$ **(seront définies plus tard)**.

Conditions d'équilibrage :

$$\begin{cases} \sum_{i \in S_2 \cap S_1} \frac{Z_{1,i}}{b_i} = \sum_{i \in S_1} Z_{1,i} \\ \sum_{i \in S_2 \cap CS_1} \frac{Z_{2,i}}{a_i} = \sum_{i \in CS_1} Z_{2,i} \end{cases},$$

$$\Leftrightarrow$$

$$\boxed{\begin{array}{c} \sum_{i \in S_2 \cap S_1} \frac{Z_{1,i}}{b_i} = \sum_{i \in S_1} Z_{1,i} \\ \sum_{i \in S_2 \cap S_1} \frac{Z_{2,i}}{a_i} = \sum_{i \in S_2} \frac{Z_{2,i}}{a_i} + \sum_{i \in S_1} Z_{2,i} - \sum_{i \in U} Z_{2,i} \end{array}}.$$

- Le but est d'obtenir une **condition d'équilibrage global sur** $S_2$, de la forme :

$$\sum_{i \in S_2} \frac{V_i}{\pi_i^2} = \sum_{i \in U} V_i \, .$$

→ Pour cela, on choisit : $\boxed{\frac{Z_{1,i}}{b_i} = \frac{Z_{2,i}}{a_i} = \frac{V_i}{\pi_i^2}}$.

Les conditions d'équilibrage ci-dessus relatives à $Z_{1,i}$ et $Z_{2,i}$ entraînent la relation :

$$\boxed{\sum_{i \in S_1} \frac{V_i}{\pi_i^2} = \sum_{i \in S_2} \frac{Z_{1,i}}{b_i} = \sum_{i \in S_1} (1 - \frac{a_i}{b_i}) Z_{1,i} + \sum_{i \in U} \frac{a_i}{b_i} Z_{1,i}}.$$

Cette relation est équivalente à : $\sum_{i \in S_2} \frac{V_i}{\pi_i^2} = \sum_{i \in U} V_i$ , si et seulement si :

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

Cette condition met en jeu $S_1$ et elle est *aléatoire*.

**Elle doit être satisfaite quel que soit $S_1$.**

$\Leftrightarrow$ **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** :

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

[cette dernière condition est **satisfaite** du fait de la relation : $\pi_i^2 = b_i \pi_i^1 + a_i(1-\pi_i^1)$ ]

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 est équilibré** 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 $S_1$, est équilibré 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 $CS_1$, est équilibré sur les variables $Z_{2,i} = \frac{a_i}{b_i}Z_{1,i} = \frac{a_i}{\pi_i^2}V_i$.

## b) Comment prendre en compte les conditions de recouvrement entre les deux échantillons ?

- Avec la contrainte : Card $(S_1 \cap S_2) = \alpha$ Card $(S_1)$, on doit avoir :

$$\sum_{i \in S_1} \pi_i^{2/S_1} = \alpha n_1 \text{ , soit :}$$

$$\boxed{\sum_{i \in S_1} b_i = \alpha n_1} \text{ . (C2)}$$

$$\Leftrightarrow \sum_{i \in S_1} \frac{1}{\pi_i^1} b_i \pi_i^1 = \alpha n_1 \text{ , } \forall \ S_1 \text{ .}$$

- Cette égalité peut s'interpréter comme une condition d'équilibrage de l'échantillon $S_1$
  - sur les variables $b_i \pi_i^1$
  - sous la condition : $\boxed{\sum_{i \in U} b_i \pi_i^1 = \alpha n_1}$.

## c) Peut-on assurer la condition supplémentaire que l'échantillon $S_2$ soit de taille fixe $n_2$ ?

- Avec la contrainte : $Card\,(S_2 \cap CS_1) = n_2 - \alpha n_1$, on doit avoir :

$$\sum_{i \notin S_1} \pi_i^{2/S_1} = n_2 - \alpha n_1 \text{ , soit :}$$

$$\boxed{\sum_{i \in S_1} a_i = \sum_{i \in U} a_i - n_2 + \alpha n_1,\ \forall\ S_1} \text{ . (C3)}$$

- Cette égalité peut s'interpréter comme une condition d'équilibrage de l'échantillon $S_1$
  - sur les variables $a_i \pi_i^1$
  - sous la condition : $\sum_{i \in U} a_i \pi_i^1 = \sum_{i \in U} a_i - n_2 + \alpha n_1$ , soit :

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

➢ Cette dernière condition est satisfaite à cause des relations :

- $\sum_{i \in U} a_i (1 - \pi_i^1) = \sum_{i \in U} (\pi_i^2 - b_i \pi_i^1)$
- $\sum_{i \in U} \pi_i^2 = n_2$
- $\sum_{i \in U} b_i \pi_i^1 = \alpha n_1$ .

**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 :**

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

# Synthèse

| 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 $\pi_i^2$** | | $\pi_i^2 = a_i(1-\pi_i^1) + b_i\pi_i^1$ | | |
| **Equilibrage sur des variables $V_i$** | Equilibrage sur les variables $(b_i - a_i)\frac{\pi_i^1}{\pi_i^2}V_i$ | | Equilibrage conditionnel sur les variables $\frac{b_i}{\pi_i^2}V_i$ | Equilibrage conditionnel sur les variables $\frac{a_i}{\pi_i^2}V_i$. |
| **Taux de recouvrement $\alpha$ 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 $\alpha n_1$ $\Leftrightarrow$ 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 $a_i\pi_i^1$ | $\sum_{i\in U} \pi_i^2 = n_2$ | | |

## Est-il possible de trouver des solutions ?

- Etant données les $\pi_i^2$ :
  - On doit déterminer les paramètres inconnus $a_i$ et $b_i$ (à valeurs dans [0,1])
  - satisfaisant différentes contraintes linéaires.

➢ Des solutions peuvent être trouvées si et seulement si l'échantillon $S_1$ a été tiré antérieurement avec des conditions d'équilibrage appropriées,

➢ … ***ce qui n'est en général pas le cas***.

# 4. Une méthode proposée pour trouver des solutions approchées : « *l'équilibrage inverse* »

- La principale difficulté dans ce qui précède est que le 1er échantillon doit satisfaire des contraintes d'équilibrage spécifiques.

- Si celles-ci n'ont pas été prévues, les contraintes imposées au 2nd échantillon ne pourront pas être satisfaites.

- L'idée est de *changer les probabilités d'inclusion finales*…
  - …et de trouver de nouvelles probabilités d'inclusion finales $\tilde{\pi}_i^2(S_1)$
  - … proches des $\pi_i^2$
  - … et conservant toutes les conditions d'équilibrage sur le 1er échantillon.

# Formalisation.

**<u>Le problème</u> : trouver les $\tilde{\pi}_i^2$ (et les $\tilde{a}_i$ et $\tilde{b}_i$ en découlant)**
**satisfaisant toutes les contraintes ainsi que les équations d'équilibrage relatives à $S_1$.**

- <u>Les contraintes sur les paramètres</u> :

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

$$\sum_{i \in U} \pi_i^2 = n_2$$

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

$$a_i \text{ et } b_i \in [0, 1]$$

- <u>Les équations d'équilibrage à satisfaire sur $S_1$</u> :

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

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

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

- Et **minimiser une fonction objectif** : $\sum_{i \in U} d^2(\tilde{\pi}_i^2, \pi_i^2)$, où $d$ désigne une distance

# Un exemple particulier où l'on peut trouver une solution explicite.

- Les seules conditions sur $S_2$ : **taille fixe** et **taux de recouvrement** par rapport à $S_1$ .
- Pas d'autre condition d'équilibrage sur $S_2$ .
- $S_1$ a une **taille fixe** $n_1$ .

$\Rightarrow$ On peut prendre : $\boxed{b_i = \alpha \in [0,1],\ \forall\ i \in U}$,

qui satisfait la relation : $\sum_{i \in U} b_i \pi_i^1 = \alpha n_1$ .

Cela conduit à un **sondage aléatoire simple** de $S_2$ au sein de $S_1$ avec un taux $\alpha$ .

- Alors : $\boxed{a_i = \dfrac{\pi_i^2 - b_i \pi_i^1}{1 - \pi_i^1} = \dfrac{\pi_i^2 - \alpha \pi_i^1}{1 - \pi_i^1}\ (si\ \pi_i^1 \neq 1)}$.

- $S_1$ doit satisfaire une condition d'équilibrage sur les variables :

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

# Comment prendre en compte le cas des strates exhaustives ?

## Si certains $\pi_i^1$ valent 1 :

- On définit : $U_j = \left\{ i \in U \; ; \; \pi_i^j = 1 \right\}$ pour $j = 1 \; ou \; 2$.

- On posera alors :

$$\pi_i^{2/S_1} = \begin{vmatrix} 1 \; si \; i \in U_2 \\ \pi_i^2 \; si \; i \in CU_2 \cap U_1 \\ a_i 1_{i \notin S_1} \; si \; i \in CU_2 \cap CU_1 \end{vmatrix}.$$

- 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, et en utilisant les spécifications correspondantes de $\pi_i^{2/S_1}$.

Problème :

Programme d'optimisation :

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

sous les contraintes :

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

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

Solution :

- Lagrangien :

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

- D'où :

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

- Puis :

$$\boxed{\begin{cases} \tilde{b}_i = \alpha \\ \tilde{a}_i = \dfrac{\tilde{\pi}_i^2 - \alpha \pi_i^1}{1-\pi_i^1} \end{cases}}.$$

# Difficultés rencontrées dans cette approche (retour au cas général) :

- Existence des solutions.
- Ces solutions sont-elles licites ? Cela implique des ***contraintes*** sur $\alpha$.
- Difficultés numériques, programmes d'optimisation complexes.
- Les solutions $\tilde{\pi}_i^2(S_1)$ dépendent du 1er échantillon.
- Les vraies probabilités finales d'inclusion du 2nd échantillon sont :

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

- … difficiles à calculer.

# Choix d'estimateurs :

**Pour estimer un total $T(Y)$, 4 estimateurs possibles :**

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

(estimateur de HORVITZ-THOMSON naïf)

Utilise les valeurs des probabilités-cibles $\pi_i^2$.
Mais $S_2$ n'a pas été tiré avec ces probabilités $\Rightarrow$ estimateur potentiellement **biaisé**.

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

Utilise les valeurs effectives des probabilités d'inclusion, $\tilde{\pi}_i^2(S_1)$.
Celles-ci sont *aléatoires* (ce n'est pas un estimateur du type HORVITZ-THOMSON habituel) $\Rightarrow$ estimateur potentiellement **biaisé.**

c) $\boxed{\hat{T}_3(Y) = \sum_{i \in S_2 \cap S_1} \frac{Y_i}{\tilde{b}_i(S_1)} + \sum_{i \in S_2 \cap \overline{CS_1}} \frac{Y_i}{\tilde{a}_i(S_1)}}$,

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.**

d) $\boxed{\hat{T}_4(Y) = \sum_{i \in S_2} \frac{Y_i}{\pi_i^{*2}}}$, avec : $\boxed{\pi_i^{*2} = \frac{1}{1-\pi_i^1} E\left(\tilde{\pi}_i^2(S_1) 1_{i \notin S_1}\right)}$.

C'est le véritable estimateur de HORVITZ-THOMSON (à coefficients *fixes*).
Utilise les vraies probabilités d'inclusion $\pi_i^{*2}$.
Malheureusement celles-ci sont **très difficiles à calculer explicitement**.

# 5. Etudes par simulations dans le cadre des enquêtes d'évaluation de la Depp

- Deux degrés : 1-écoles et 2-élèves (un nombre fixé d'élèves par école)
- Stratification des écoles en 5 strates (3 types d'écoles + 2 strates pour les petites écoles)

- Tirage des écoles avec des *probabilités proportionnelles à leur taille*
- **Sondage aléatoire simple d'élèves** dans chaque école tirée (32)
- Taille de l'échantillon d'élèves : 4500 (reprise du cadre PISA)

# Simulation

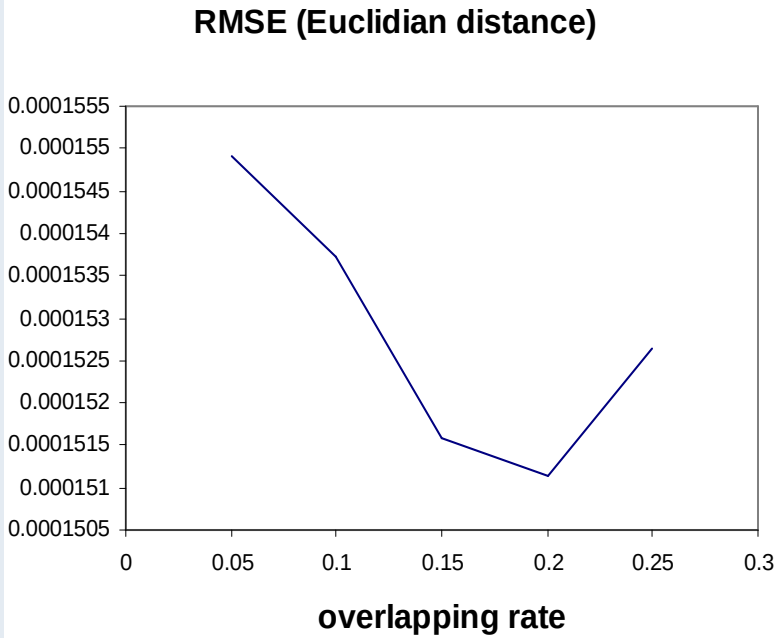
- Bases de données :
  - On utilise les bases d'établissements français ("écoles") pour 2000 et 2009
  - 1ère restriction : seulement les écoles présentes en 2000 ***ET*** 2009
  - 2ème restriction : seulement les écoles ***DANS LA MÊME STRATE*** en 2000 et 2009.
  - 6672 écoles (568 456 élèves en 2009).

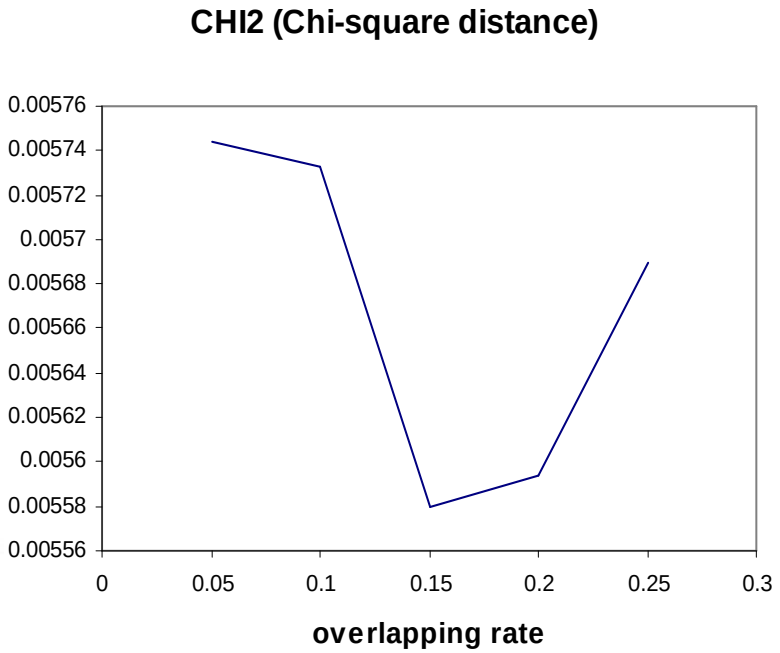
- Principe :
  - On n'impose à $S_2$ (2009) que des conditions de **recouvrement** et de **taille fixe** (pas d'autres conditions d'équilibrage).

  - Simulation 1 : tirage de $S_1$ (année 2000), calcul de $\tilde{\pi}_i^2$ et comparaison en termes de distance avec $\pi_i^2$ . Influence sur la distance du taux de recouvrement et du type de distance (Euclidienne ou Chi-2).
  - Simulation 2 : tirage de $S_1$, tirage de $S_2$ / $S_1$, estimation du total de variables connues selon 3 types d'estimateurs. Qualité des estimateurs selon le taux de recouvrement et le type de distance.

# Simulation 1 - résultats

- Remarque : le taux de recouvrement est ***borné*** (à environ 25 % dans notre cas).



# Simulation 2 - résultats

- 100 réplications de $S_1$, 100 réplications de $S_2$ pour chaque $S_1$

| | | EUCLIDIAN - overlapping rate 10% | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| variable | POP | ESTIM1 | | | ESTIM2 | | | ESTIM3 | | |
| | | E() | s.e. | B | E() | s.e. | B | E() | s.e. | B |
| girls | 274501 | 274543 | (82) | 0.0% | 274531 | (82) | 0.0% | 274521 | (84) | 0.0% |
| foreign | 19790 | 19824 | (48) | 0.2% | 19827 | (48) | 0.2% | 19782 | (49) | 0.0% |
| grade9 | 157871 | 157922 | (67) | 0.0% | 157912 | (68) | 0.0% | 157876 | (70) | 0.0% |
| grade10 | 351164 | 351154 | (66) | 0.0% | 351154 | (66) | 0.0% | 351208 | (73) | 0.0% |
| SES1 | 209998 | 210024 | (83) | 0.0% | 210016 | (83) | 0.0% | 210062 | (85) | 0.0% |
| SES2 | 135900 | 135907 | (62) | 0.0% | 135909 | (62) | 0.0% | 135927 | (63) | 0.0% |
| SES3 | 74391 | 74404 | (50) | 0.0% | 74404 | (50) | 0.0% | 74392 | (52) | 0.0% |
| SES4 | 122978 | 122936 | (78) | 0.0% | 122929 | (78) | 0.0% | 122924 | (79) | 0.0% |

# Simulation 2 - résultats

| variable | POP | EUCLIDIAN - overlapping rate 15% | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ESTIM1 | | | ESTIM2 | | | ESTIM3 | | |
| | | E() | s.e. | B | E() | s.e. | B | E() | s.e. | B |
| girls | 274501 | 274480 | (83) | 0.0% | 274721 | (83) | 0.1% | 274471 | (86) | 0.0% |
| foreign | 19790 | 19842 | (50) | 0.3% | 19874 | (50) | 0.4% | 19809 | (51) | 0.1% |
| grade9 | 157871 | 157819 | (67) | 0.0% | 158029 | (68) | 0.1% | 157842 | (72) | 0.0% |
| grade10 | 351164 | 351190 | (65) | 0.0% | 351438 | (66) | 0.1% | 351117 | (72) | 0.0% |
| SES1 | 209998 | 210080 | (82) | 0.0% | 210306 | (83) | 0.1% | 209918 | (86) | 0.0% |
| SES2 | 135900 | 135910 | (62) | 0.0% | 136040 | (62) | 0.1% | 135916 | (64) | 0.0% |
| SES3 | 74391 | 74456 | (51) | 0.1% | 74522 | (51) | 0.2% | 74397 | (52) | 0.0% |
| SES4 | 122978 | 122851 | (77) | -0.1% | 122925 | (77) | 0.0% | 122988 | (77) | 0.0% |

# Simulation 2 - résultats

| variable | POP | EUCLIDIAN - overlapping rate 20% | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ESTIM1 | | | ESTIM2 | | | ESTIM3 | | |
| | | E() | s.e. | B | E() | s.e. | B | E() | s.e. | B |
| girls | 274501 | 274587 | (82) | 0.0% | 274639 | (82) | 0.1% | 274466 | (87) | 0.0% |
| foreign | 19790 | 19856 | (51) | 0.3% | 19858 | (51) | 0.3% | 19794 | (54) | 0.0% |
| grade9 | 157871 | 157875 | (67) | 0.0% | 157814 | (69) | 0.0% | 157920 | (77) | 0.0% |
| grade10 | 351164 | 351116 | (65) | 0.0% | 351309 | (67) | 0.0% | 351048 | (74) | 0.0% |
| SES1 | 209998 | 209917 | (82) | 0.0% | 209952 | (83) | 0.0% | 209970 | (87) | 0.0% |
| SES2 | 135900 | 135974 | (62) | 0.1% | 136021 | (62) | 0.1% | 135911 | (65) | 0.0% |
| SES3 | 74391 | 74384 | (51) | 0.0% | 74406 | (51) | 0.0% | 74379 | (54) | 0.0% |
| SES4 | 122978 | 122939 | (77) | 0.0% | 122966 | (77) | 0.0% | 122940 | (79) | 0.0% |

# 6. Remarques finales, conclusion et travaux à venir

- Il est difficile d'imposer des conditions à un échantillon à tirer prenant en compte un échantillon tiré antérieurement. ***Il n'y a pas de solutions exactes, seulement des solutions approchées.***

- *On peut en déduire une ligne directrice pour l'avenir :* un échantillon tiré à une date donnée **doit anticiper les tirages ultérieurs et les contraintes qui s'y appliqueront, de manière à les intégrer sous forme de conditions d'équilibrage appropriées.**

- Les simulations montrent qu'il est possible de mettre en œuvre l'approche de l'"équilibrage inverse" et de vérifier ainsi les propriétés de la méthode.

- De nouvelles simulations devraient être faites (cas plus généraux, mise en œuvre de méthodes de résolution numériques).

- Approfondissement de cas particuliers :
  - Stratification de l'univers
  - Cas où l'on a une strate exhaustive pour $S_1$ ou $S_2$
  - Cas où l'univers de référence est modifié entre les deux périodes.

- Questions théoriques à résoudre :
  - Calcul des ***vraies*** probabilités d'inclusion finale dans $S_2$
  - Utilisation d'estimateurs et de conditions d'équilibrage avec des ***coefficients aléatoires***, dépendant de $S_1$.

- Une question majeure : *calcul et estimation de la* ***variance des estimateurs***.

- Mise en œuvre en situation réelle.

*Merci pour votre attention*

*et vos remarques !*

*marc.christine@insee.fr*

*thierry.rocher@education.gouv.fr*

# To remind about Balanced sampling

# A. Objects

- Increase the « representativeness » of a sample…
- …according to particular variables…
- …which are assumed to be highly correlated with the variables of interest…
- …in order to improve the precision of estimates of the total of those variables.

# What does "representativeness" mean ?

- The aim is the sample to be the best « scale model » of the whole population.

- A very simple method to obtain it : stratification with *proportional allocation.* In that case, the sample has the same structure as the whole population according to some categories.

=> balanced sampling generalizes this approach (and allows to consider the case of continuous variables).

## B. Mathematical definition

A sample s is balanced on a variable X if and only if :

The HORVITZ-THOMSON estimator for the total of X (called ***balancing variable***) takes the same value as the true total (***known***) of X over the whole population.

➢ A sampling design is balanced if and only if any sample S drawn according to this design is balanced :

$$\forall s \quad : \quad \sum_{k \in s} \frac{x_k}{\pi_k} = \sum_{k \in U} x_k$$

This is called "*balancing condition*".

➢ Nota : X may be a vector of variables.

➢ Generally, with any other variable Y, the above equality is only true with *expectancy*, when one tries to obtain unbiased estimate for the total of Y (Horvitz-Thompson estimator).

# Examples

- To balance on the variable *equal to 1* is equivalent to the "adjustment" of the size of the sample to the size N of the population (***the sum of sample weights is N***).

- To balance on the variable "*inclusion probability*" is equivalent to ensure a ***fixed size sampling design***.

- To balance on the *variables indicating the belonging to different strata* is equivalent to ensure a ***stratified sampling with a given size of the sample in each stratum***.

# Important remarks

- Balanced sampling can be implemented *with any given set of inclusion probabilities.*
- The sample weights remain unchanged (inverse of the inclusion probabilities)…
- … in opposition to *calibration procedures* where original weights are changed.
- It is necessary *to know auxiliary information on the whole population* considered as sampling frame (totals of X).

# C. A fictional example of balanced sampling

Population = 10 individuals.

Sample of size 3.

Sampling with equal probabilities, value $\frac{3}{10}$.

| Individuals | Balancing variable X | Inclusion probabilty |
|---|---|---|
| | | |
| 1 | 30 | 3/10 |
| 2 | 45 | 3/10 |
| 3 | 30 | 3/10 |
| 4 | 15 | 3/10 |
| 5 | 40 | 3/10 |
| 6 | 35 | 3/10 |
| 7 | 35 | 3/10 |
| 8 | 20 | 3/10 |
| 9 | 50 | 3/10 |
| 10 | 50 | 3/10 |
| **TOTAL** | **350** | |

Sample ramdom sampling.

There are $C_{10}^{3} = 120$ possible samples, each of them with the probability $\frac{1}{120}$.

*For instance* :

The sample (2, 5, 9) gives the estimate for the total : $\frac{10}{3}(45+40+50) = 450.$

Balanced sampling on the total of X with equal probabilities.

The balanced samples are those for which : $\sum_{i \in s} \frac{X_i}{3/10} = 350$, that is : $\boxed{\sum_{i \in s} X_i = 105}$.

It is possible to show that there are **only 12 balanced samples** :

(1, 2, 3) (1, 5, 6) (1, 5, 7) (2, 5, 8) (3, 5, 6) (3, 5, 7)
(4, 5, 9) (4, 5, 10) (6, 8, 9) (6, 8, 10) (7, 8, 9) (7, 8, 10)

| Individuals | Balancing variable X | Inclusion probabilty |
|---|---|---|
| | | |
| 1 | 30 | 3/10 |
| 2 | 45 | 3/10 |
| 3 | 30 | 3/10 |
| 4 | 15 | 3/10 |
| 5 | 40 | 3/10 |
| 6 | 35 | 3/10 |
| 7 | 35 | 3/10 |
| 8 | 20 | 3/10 |
| 9 | 50 | 3/10 |
| 10 | 50 | 3/10 |
| **TOTAL** | **350** | |

$\Rightarrow$ To draw a balanced sample on the variable X, keeping the same inclusion probabilities, ***given and fixed for any individual in the population***, leads to define a new sampling design, that is ***a new random distribution on the set of samples of size 3***.

This new distribution :

- assigns a null probability to all samples but those in the above list
- assigns a probability to the samples of this list in order that any first order inclusion probabilty keeps the value $\frac{3}{10}$.

If one wants to calculate the probabilities to be assigned to each of those 12 samples, one might have to solve a system with 10 equations (defining the inclusion probabilities for each individual) with 12 unkown variables.

There might be no solution or several ones !

Let us modify the initial data as following :

| Individuals | Balancing variable X | Inclusion probabilty |
|---|---|---|
| | | |
| 1 | 30 | 3/10 |
| 2 | 45 | 3/10 |
| 3 | 30 | 3/10 |
| 4 | 15 | 3/10 |
| 5 | 40 | 3/10 |
| 6 | 35 | 3/10 |
| 7 | 35 | 3/10 |
| 8 | 20 | 3/10 |
| 9 | 50 | 3/10 |
| 10 | 51 | 3/10 |
| **TOTAL** | **351** | |

The balanced samples are those for which : $\sum_{i \in s} \frac{X_i}{3/10} = 351$, that is : $\boxed{\sum_{i \in s} X_i = 105{,}3}$.

**It can be seen that there is no balanced sample on this total with those inclusion probabilities**.

However, we can find *approached solutions* : all samples giving an estimated value for the total of X very close to 351 : the 12 above samples give the approached solution.

# D. A new method : the « CUBE »

- Method devised and implemented by J-C. DEVILLE and Y. TILLE (2000).
- Further developments by G. CHAUVET (« fast CUBE »).
- Macro CUBE written in SAS 8.
- Macro and guidelines for users are available and free on the Website of Insee :

http : //www.insee.fr/fr/nom_def_met/outils_stat/macro.htm

## E. Balance sheet of the method

ADVANTAGES

- It allows to keep any given set of inclusion probabilities.
- It allows to use balancing conditions on a set of different variables (age, gender, income…) and therefore increases the representativeness of the sample.
- It allows to build unbiased estimators for any total (and not only for the balancing variables), since inclusion probabilities are not changed.
- It reduces the variance of those estimators since balancing variables are correlated with variables of interest.

## DIFFICULTIES

- It is not always possible to obtain exact balanced samples => to be satisfied with approached solutions.
- Balancing constraints are more or less satisfied at last => not to demand too much (if too many constraints, the selection of samples might become deterministic).
- Do not forget that a fixed size sample is a particular balancing condition.
- If the sampling frame size is too large, it may take a long time for computation (but a new procedure called « FAST CUBE » has been developed recently).

## Main relations.

$$\pi_i^2 = \mathrm{E}\ \pi_i^{2/S_1}.$$

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

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

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