Formation calage sur marges - Calages simultanés

Emmanuel Gros, Antoine Rebecq
emmanuel.gros@insee.fr, antoine.rebecq@insee.fr

INSEE - Division Sondages

29 avril 2015

Sommaire I

Chapitre 1

Population et échantillons

Partie 1

Population

Population

Population de ménages (ou entreprises) : $\mathcal{U}_{men} = [1 \dots m \dots N_{men}]$

Population d'individus (ou salariés) vivant dans les ménages (ou travaillant dans les entreprises) : $\mathcal{U}_{ind} = [1 \dots i \dots N_{ind}]$

Un ménage (ou entreprise) m est constitué de n_m individus (ou salariés) : $men_m = [(m,1) \dots (m,i) \dots (m,n_m)]$

$$\mathcal{U}_{ind} = \bigcup_{m \in \mathcal{U}_{men}} men_m$$

$$N_{ind} = \sum_{m \in \mathcal{U}_{men}} n_m$$

Partie 2

Échantillons : sondage à plusieurs degrés

Échantillons : sondage à plusieurs degrés

- **Premier degré de tirage :** Échantillon de ménages (ou entreprises) s_{men} de taille n_{men}, tiré dans $\mathcal{U}_{men}$ selon un plan p, avec des probabilités d'inclusion π_{men} :

$$d_{men} = \frac{1}{\pi_{men}} = \text{poids du ménage } men$$

- **Second degré de tirage :** Les individus (ou salariés) tirés selon :
  - Un sondage en grappes
  - Un sondage à deux degrés
  - Cas de l'individu Kish

Sondage en grappes

Échantillon d'individus s_{ind} de taille n_{ind} , constitué de tous les individus appartenant aux ménages sélectionnés :

$$s_{ind} = \bigcup_{m \in s_{men}} men_m$$

$$n_{ind} = \sum_{m \in s_{men}} n_m$$

Probabilités de tirage des individus : $\forall i \in men_m,\ \pi_{mi} = \pi_m$

Pondérations : $\forall i \in men_m,\ d_{mi} = d_m$

Sondage à deux degrés

Dans chaque entreprise m de l'échantillon s_{men}, on tire un échantillon de salariés sm, avec des probabilités conditionnelles de tirage $\pi_{i|m}$

Probabilités de tirage des individus : $\forall i \in men_m,\ \pi_{mi} = \pi_m \cdot \pi_{i|m}$

Pondérations : $\forall i \in men_m,\ d_{mi} = d_m \cdot d_{i|m}$

où $d_{i|m} = \dfrac{1}{\pi_{i|m}}$ est le poids conditionnel de tirage du salarié i travaillant dans l'entreprise m (ou de l'individu i appartenant au ménage m).

Cas de l'individu Kish

Cas des enquêtes ménages / individus. En plus de l'échantillon s_{ind} constitué de tous les individus des ménages sélectionnés (cas a), on sélectionne un échantillon d'individus s_{kisj} de la façon suivante :

un individu k_m dans chaque ménage m est tiré selon un sondage aléatoire simple, parmi les e_m individus "éligibles".

Probabilité de tirage de l'individu Kish du ménage m :
$\pi_{km} = \pi_m \cdot \frac{1}{e_m}$

Pondération de l'individu Kish du ménage m : $d_{km} = d_m \cdot e_m$

Partie 3

Exemples

Exemples

- Cas de sondages en grappe :
  - Enquête Patrimoine,
  - Enquête Emploi en continu,
  - Dispositif SRCV (ressources et conditions de vie)
  - Enquête FQP (formation qualification professionnelle)

  Cas de sondages en deux degrés :
  - Enquête COI-TIC,
  - Enquête ECMOSS,
  - Dispositif d'enquêtes permanentes des Conditions de vie / EPCV

  Cas des individus Kish :
  - Enquête Victimation / cadre de vie et sécurité (CVS)

Chapitre 2

Intérêt du calage simultané

Partie 1

Propriétés avant calage

Propriétés avant calage du sondage en grappe

1. Possibilité d'établir des statistiques sur les ménages à partir des fichiers individus.
   Égalité entre les poids des individus d'un même ménage et le poids ménage : $\forall i \in m, d_{mi} = d_m$
2. Obtenir des estimations sur des nombres d'individus identiques à partir des échantillons ménages et individus.
   Par exemple : estimation du nombre de femmes de professions intermédiaires dans la population :
   - on note y_{mi} l'indicatrice "être une femme de professions intermédiaires"
   - et y_m le nombre de femmes de professions intermédiaires du ménage m

Propriétés avant calage du sondage en grappe

Propriété 2 :

$$\begin{aligned}
\hat{T}_Y^{(ind)} &= \sum_{m \in s_{men}} \sum_{i \in men_m} d_{mi} y_{mi} \\
&= \sum_{m \in s_{men}} d_m \sum_{i \in men_m} y_{mi} \\
&= \sum_{m \in s_{men}} d_m y_m \\
&= \hat{T}_Y^{(men)}
\end{aligned}$$

Propriétés avant calage du sondage à deux degrés

1. Obtenir des estimations identiques sur des effectifs de regroupements de ménages (entreprises) à partir des 2 échantillons

Via l'échantillon de ménages :

$$\hat{n}_{G,men} = \sum_{m \in s_{men}} d_m n_m \mathbb{1}_{m \in G}$$

Propriétés avant calage du sondage à deux degrés

Via l'échantillon d'individus :

$$\hat{n}_{G,ind} = \sum_{m \in s_{men}} \sum_{i \in s_m} d_{mi} \mathbb{1}_{m \in G} = \sum_{m \in s_{men}} \sum_{i \in s_m} d_m d_{i|m} \mathbb{1}_{m \in G}$$
$$= \sum_{m \in s_{men}} d_m \mathbb{1}_{m \in G} \sum_{i \in s_m} d_{i|m}$$

Ainsi, dès que $\sum_{i \in s_m} d_{i|m} = n_m$, l'égalité entre les deux termes sera vérifiée.

Partie 2

Pourquoi pas des calages indépendants ?

Objectif du calage : Retrouver des marges ménages et des marges individus.

Procéder par deux calages, un sur le fichier ménages, puis un sur le fichier individu n'aurait aucune raison d'assurer les propriétés vues précédemment ("le second calage détruit le premier").

Pour assurer les propriétés précédentes avec les poids calés il faudrait :

- Pour le sondage en grappes : $w_{mi} = w_m$
- Pour le sondage à deux degrés $\sum_{i \in s_m} w_{i|m} = n_m$

Ce qui revient à **conserver les poids conditionnels de tirage**.

Conserver les poids conditionnels de tirage

$$\begin{aligned} d_{i|m} &= \frac{d_{mi}}{d_m} \\ &= \frac{w_{mi}}{w_m} \\ &= w_{i|m} \end{aligned}$$

Équations de calage - niveau ménages

Informations auxiliaires : $x_m =$ vecteur des variables auxiliaires pour le ménage m de s_{men}

$T_X = \sum_{m \in \mathcal{U}_{men}} x_m =$ vecteur des totaux sur la population $\mathcal{U}_{men}$, connus.

Équations de calage :

$$\sum_{m \in s_{men}} d_m F(x_m' \lambda) x_m = \sum_{m \in s_{men}} w_m x_m = T_X$$

Équations de calage - niveau individus

Informations auxiliaires : z_{mi} = vecteur des variables auxiliaires pour l'individu (m, i) de s_{ind}

$T_Z = \sum_{(m,i) \in \mathcal{U}_{ind}} z_{mi}$ = vecteur des totaux sur la population $\mathcal{U}_{ind}$, connus.

Équations de calage :

$$\sum_{(m,i) \in s_{ind}} d_{mi} F(z'_{mi}\lambda) z_{mi} = \sum_{(m,i) \in s_{ind}} w_{mi} z_{mi} = T_Z$$

Chapitre 3

Calages simultanés pour plusieurs plans de sondages

Partie 1

Sondage en grappes

Principe : Réaliser le calage au niveau ménage. On construit les totaux par ménage des variables de calage de niveau individu :

$$z_m = \sum_{i=1}^{n_m} z_{mi}$$

Les totaux de ces variables sur la population des ménages $\mathcal{U}_{men}$ valent :

$$\sum_{m \in \mathcal{U}_{men}} z_m = \sum_{m \in \mathcal{U}_{men}} \sum_{i=1}^{n_m} z_{mi} = \sum_{(m,i) \in \mathcal{U}_{ind}} z_{mi} = T_Z$$

Les équations de calage donnent les poids w :

$$\sum_{m \in s_{men}} d_m F(x'_m \lambda + z'_m \mu)(x_m, z_m) = \sum_{m \in s_{men}} w_m (x_m, z_m) = (T_X, T_Z)$$

On obtient :

- w_m = pondération du ménage m dans s_{men}
- $w_{mi} = w_m$ = pondération de l'individu (m, i) du ménage m dans s_{ind}

L'échantillon est bien calé sur les totaux T_X et T_Z :

$$\sum_{m \in s_{men}} w_m x_m = T_X$$

$$\sum_{i \in s_{ind}} w_{mi} z_{mi} = \sum_{m \in s_{men}} w_m \left(\sum_{(m,i) \in men_m} z_{mi} \right) = \sum_{m \in s_{men}} w_m z_m = T_Z$$

Exemple : enquête sur la consommation alimentaire

Au niveau ménage : On cale l'échantillon de ménages sur des structures issues de l'enquête-emploi, pour les variables catégorielles suivantes :

- nombre de personnes du ménage (6 modalités)
- catégorie socioprofessionnelle du chef de ménage (7 modalités)
- tranche d'âge du chef de ménage (7 modalités)
- catégorie de commune (6 modalités)

Exemple : enquête sur la consommation alimentaire

Au niveau individu : On cale l'échantillon d'individus sur la structure par sexe et tranche d'âge ($2 \times 4 = 8$ modalités) issue de l'enquête emploi. Les variables de calage proprement dites sont donc les 8 variables indicatrices associées à cette variable catégorielle.

Exemple : enquête sur la consommation alimentaire

Le calage simultané ménages - individus s'effectue au niveau ménage.

Les variables z_m sont les totaux par ménage des variables indicatrices de la variable $sexe \times tranche - d'age$: elles contiennent, pour chaque ménage m, le nombre d'hommes de moins de 15 ans dans le ménage, le nombre d'hommes de 15 à 34 ans, etc.

Partie 2

Sondage à deux degrés

On réalise le calage au niveau ménage / entreprise.

On construit les totaux estimés par ménage / entreprise des variables de calage de niveau individus / salariés :

$$\hat{z}_m = \sum_{i \in s_m} d_{i|m} z_{mi}$$

Les équations de calage donnent les poids w :

$$\sum_{m \in s_{men}} d_m F(x'_m \lambda + \hat{z}'_m \mu)(x_m, \hat{z}_m) = \sum_{m \in s_{men}} w_m(x_m, \hat{z}_m) = (T_X, T_Z)$$

On obtient :

- w_m = pondération du ménage m dans s_{men}
- $w_{mi} = d_{i|m} \cdot w_m$ = pondération de l'individu (m, i) du ménage m dans s_{ind} (ie $w_{i|m} = d_{i|m}$)

L'échantillon est bien calé sur les totaux T_X et T_Z :

$$\sum_{m \in s_{men}} w_m x_m = T_X$$

$$\sum_{i \in s_{ind}} w_{mi} z_{mi} = \sum_{m \in s_{men}} w_m \left(\sum_{(m,i) \in men_m} d_{i|m} z_{mi} \right) = \sum_{m \in s_{men}} w_m \hat{z}_m = T_Z$$

Partie 3

Cas particulier de l'individu Kish

Il s'agit d'effectuer un calage simultané ménages/individus/individus Kish.

V_{k_m} = vecteur de variables auxiliaires pour l'individu-Kish k_m de s_{Kish}

$T_V = \sum_{i \in \mathcal{U}^e_{ind}} v_i$ = vecteur des totaux sur la population des éligibles $\mathcal{U}^e_{ind}$, connus.

On réalise le calage au niveau ménage / entreprise.

On calcule pour chaque ménage m les totaux des variables-individus ainsi que les totaux estimés des variables-individus Kish :

$$z_m = \sum_{(m,i) \in men_m} z_{mi}$$

$$\hat{v}_m = e_m v_{k_m}$$

Vecteur de variables de calage pour le ménage m : $(x_m, z_m, \hat{v}_m)$.
Vecteur des totaux : (T_X, T_Z, T_V)

Les équations de calage donnent les poids w :

$$\sum_{m \in s_{men}} d_m F(x'_m \lambda + z'_m \mu + \hat{v}'_m \gamma)(x_m, z_m, \hat{v}_m) = \sum_{m \in s_{men}} w_m (x_m, z_m, \hat{v}_m) = (T_X, T_Z, T_V)$$

On obtient :

- w_m = pondération du ménage m dans s_{men}
- $w_{mi} = w_m$ = pondération de l'individu (m, i) du ménage m dans s_{ind}
- $w_{k_m} = e_m w_m$ = pondération de l'individu Kish k_m du ménage m dans s_{Kish}

L'échantillon est bien calé sur les totaux T_X, T_Z et T_V :

$$\sum_{m \in s_{men}} w_m x_m = T_X$$

$$\sum_{i \in s_{ind}} w_{mi} z_{mi} = \sum_{m \in s_{men}} w_m \left(\sum_{(m,i) \in men_m} z_{mi} \right) = \sum_{m \in s_{men}} w_m z_m = T_Z$$

$$\sum_{k_m \in s_{Kish}} w_{k_m} v_{k_m} = \sum_{k_m \in s_{Kish}} w_m e_m v_{k_m} = \sum_{k_m \in s_{Kish}} w_m \hat{v}_m = T_V$$

Avec Calmar 2 ?

Calmar 2 permet de réaliser de tels calages simultanés.

L'utilisateur doit fournir les différentes tables en entrée et les tables des marges correspondantes : le programme réalise toutes les opérations nécessaires pour se ramener à un calage unique, et affecte les pondérations adéquates dans les différentes tables.

Chapitre 4

Non-réponse et calages simultanés

L'existence de non-réponse a un impact sur les propriétés de cohérence.

Calage au niveau ménages, méthode INSEE

- On remonte les variables individus (salariés)
- On cale au niveau ménages (entreprises)
- On conserve les poids conditionnels de tirage

Différentes méthodes existent : voir Estevao & Särndal [2006], Cordier-Villoing & Sautory [JMS 2012]