Formation calage sur marges - Introduction

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

INSEE - Division Sondages

29 avril 2015

Sommaire I

1. Rappels de théorie des sondages
   - Notations
   - Plan de sondage
   - Estimation
   - Qualité - Précision
   - Divers plans de sondage
     - Le sondage aléatoire simple
     - Autres plans de sondage

2. Redressements
   - Notations et objectifs
   - Plan de la formation

Chapitre 1

Rappels de théorie des sondages

Partie 1

Notations

Notations

- Population $\mathcal{U} = u_1, \ldots, u_k, \ldots, u_N$ de taille N
- Échantillon $s \subset \mathcal{U}$, de taille (fixe) n
- Variable d'intérêt Y, qui prend la valeur y_k pour l'individu k.
- On s'intéresse à l'estimation de totaux ou de moyennes. Les estimateurs sont notés : $\hat{T}(Y)$, $\hat{\bar{Y}}$
- Valeurs calculées à partir de l'échantillon notées en minuscules : $\bar{Y} = \frac{1}{N}\sum_{k \in \mathcal{U}} y_k$, $\bar{y} = \frac{1}{n}\sum_{k \in s} y_k$

Notations

- Espérance d'un estimateur $\mathbb{E}(\hat{\Phi}) = \sum_s p(s)\hat{\Phi}(s)$
- Biais d'un estimateur $\hat{\Phi}$: $B(\hat{\Phi}) = \mathbb{E}(\hat{\Phi}) - \Phi$
- $\mathrm{Var}(\hat{\Phi}) = \sum_{\mathrm{s}} \mathrm{p(s)} \cdot \left[\mathbb{E}(\hat{\Phi}) - \hat{\Phi}(\mathrm{s})\right]^2$
- $\sigma(\hat{\Phi}) = \sqrt{\mathrm{Var}(\hat{\Phi})}$, écart-type
- $CV(\hat{\Phi}) = \dfrac{\sigma(\hat{\Phi})}{\mathbb{E}(\hat{\Phi})}$, coefficient de variation
- $EQM(\hat{\Phi}) = \sum_s p(s) \cdot \left[\Phi - \hat{\Phi}(s)\right]^2 = \mathrm{Var}(\hat{\Phi}) + \mathrm{B}(\hat{\Phi})^2$

Partie 2

Plan de sondage

Plan de sondage sans remise - définition

On note $\mathcal{S}$ l'ensemble des parties de $\mathcal{U}$.
Le plan de sondage p est une loi de probabilité sur $\mathcal{S}$ tel que :

$$\forall s \in \mathcal{S},\ p(s) \geq 0$$

$$\sum_{s \in \mathcal{S}} p(s) = 1$$

Plan de sondage sans remise - exemple

Soit $\mathcal{U} = \{1, 2, 3\}$. On a alors :
$\mathcal{S} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$

On peut définir un plan de sondage p par :

$$p(\{1\}) = 0 \quad p(\{1, 2\}) = \frac{1}{2} \quad p(\{1, 2, 3\}) = 0$$

$$p(\{2\}) = 0 \quad p(\{1, 3\}) = \frac{1}{3}$$

$$p(\{3\}) = 0 \quad p(\{2, 3\}) = \frac{1}{6}$$

Probabilités d'inclusion π_k et π_{kl}

En pratique, p est peu utile. On utilise plutôt les probabilités d'inclusion de premier et de second degré : pour $k \in \mathcal{U}$,

$$\pi_k = \mathbb{P}(k \in s) = \mathbb{P}(\delta_k = 1) = \sum_{s \ni k} p(s)$$

$$\pi_{kl} = \mathbb{P}(k, l \in s) = \mathbb{P}(\delta_k \delta_l = 1) = \sum_{s \ni k,l} p(s)$$

(où δ_k est l'indicatrice d'appartenance de k à $\mathcal{S}$, appelée aussi variable de Cornfield)

Probabilités d'inclusion π_k et π_{kl} - Exemple

On reprend l'exemple de la slide 9 :

$$\pi_1 = \frac{5}{6}$$
$$\pi_2 = \frac{2}{3}$$
$$\pi_3 = \frac{1}{2}$$

$$\pi_{1,2} = \frac{1}{2}$$
$$\pi_{1,3} = \frac{1}{3}$$
$$\pi_{2,3} = \frac{1}{6}$$

Probabilités d'inclusion π_k et π_{kl} - Propriétés

On note $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

$$\mathbb{E}(\delta_k) = \pi_k \qquad \mathbb{E}(\delta_k \delta_l) = \pi_{kl}$$

$$\mathrm{Var}(\delta_\mathrm{k}) = \pi_\mathrm{k}(1 - \pi_\mathrm{k}) \qquad \mathrm{Cov}(\delta_\mathrm{k}\delta_\mathrm{l}) = \Delta_\mathrm{kl}$$

Probabilités d'inclusion π_k et π_{kl} - Propriétés

Pour un plan à **taille fixe** n, on a :

$$\sum_{k \in \mathcal{U}} \pi_k = n$$

$$\sum_{\substack{k,l \in \mathcal{U} \\ k \neq l}} \sum \pi_{kl} = n(n-1)$$

$$\sum_{\substack{l \in \mathcal{U} \\ l \neq k}} \pi_{kl} = \pi_k (n-1)$$

Partie 3

Estimation

Définition

Définition

L'estimateur d'Horvitz-Thompson (ou π-estimateur) est défini :

$$\textit{pour un total : } \hat{T}_{y\pi} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

$$\textit{pour une moyenne : } \hat{\bar{y}}_{\pi} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k}$$

C'est donc un **estimateur pondéré** *utilisant les poids* $w_k = \frac{1}{\pi_k}$

Estimation sans biais

Théorème

Si $\forall k \in \mathcal{U}, \pi_k > 0$, *alors l'estimateur d'Horvitz-Thompson est* **sans biais** *pour le total et la moyenne.*

Estimation sans biais

Démonstration.

$$\begin{aligned}
\mathbb{E}[\hat{T}_{y\pi}] &= \mathbb{E}\left[\sum_{k\in s}\frac{y_k}{\pi_k}\right] \\
&= \mathbb{E}\left[\sum_{k\in \mathcal{U}}\frac{y_k\delta_k}{\pi_k}\right] \\
&= \sum_{k\in\mathcal{U}}\frac{y_k\mathbb{E}[\delta_k]}{\pi_k} \\
&= \sum_{k\in\mathcal{U}} y_k \\
&= T(y)
\end{aligned}$$

Partie 4

Qualité - Précision

Variance de l'estimateur de Horvitz-Thompson

Propriété

La variance de l'estimateur de Horvitz-Thompson s'écrit :

$$\mathrm{Var}[\hat{\mathrm{T}}_{\mathrm{y}\pi}] = \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \frac{y_k y_l}{\pi_k \pi_l} \Delta_{kl}$$

(où : $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ *)*

Variance de l'estimateur de Horvitz-Thompson

Démonstration.

$$\begin{aligned}
\mathrm{Var}(\hat{\mathrm{t}}_{\mathrm{y}\pi}) &= \mathrm{Var}\Big(\sum_{\mathrm{k}\in\mathrm{U}} \frac{\mathrm{y_k}}{\pi_\mathrm{k}}\delta_\mathrm{k}\Big) \\
&= \sum_{k\in U} \frac{y_k^2}{\pi_k^2}\mathrm{Var}(\delta_\mathrm{k}) + \sum_{\mathrm{k}\in\mathrm{U}}\sum_{\mathrm{l}\in\mathrm{U},\mathrm{l}\neq\mathrm{k}} \frac{\mathrm{y_k}}{\pi_\mathrm{k}}\frac{\mathrm{y_l}}{\pi_\mathrm{l}}\mathrm{Cov}(\delta_\mathrm{k},\delta_\mathrm{l}) \\
&= \sum_{k\in U} \frac{y_k^2}{\pi_k^2}\pi_k\cdot(1-\pi_k) + \sum_{k\in U}\sum_{l\in U,l\neq k} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}(\pi_{kl}-\pi_k\pi_l) \\
&= \sum_{k,l\in U} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\Delta_{kl}
\end{aligned}$$

□

Variance pour un plan de taille fixe

Propriété

Si le plan de sondage est de taille fixe (formule de Yates-Grundy) :

$$\mathrm{Var}(\hat{\mathrm{t}}_{\mathrm{y}\pi}) = -\frac{1}{2}\sum_{\mathrm{k}\in\mathrm{U}}\sum_{\mathrm{l}\in\mathrm{U},\mathrm{l}\neq\mathrm{k}}(\frac{\mathrm{y_k}}{\pi_\mathrm{k}} - \frac{\mathrm{y_l}}{\pi_\mathrm{l}})^2\Delta_{\mathrm{kl}}$$

Variance de l'estimateur de Horvitz-Thompson

Démonstration.

Découle de la formule de Horwitz-Thompson quand le plan de sondage est de taille fixe. Pour démontrer la formule, il vaut mieux procéder à rebours :

$$
\begin{aligned}
&-\frac{1}{2}\sum_{k\in U}\sum_{l\in U,l\neq k}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\Delta_{kl}\\
&=\frac{1}{2}\sum_{k\in U}\sum_{l\in U,l\neq k}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2(\pi_k\pi_l-\pi_{kl})\\
&=\frac{1}{2}\sum_{k\in U}\sum_{l\in U,l\neq k}\left(\frac{y_k^2}{\pi_k^2}+\frac{y_l^2}{\pi_l^2}-2\frac{y_k y_l}{\pi_k\pi_l}\right)(\pi_k\pi_l-\pi_{kl})\\
&=\sum_{k\in U}\sum_{l\in U,l\neq k}\frac{y_k^2}{\pi_k^2}(\pi_k\pi_l-\pi_{kl})-\sum_{k\in U}\sum_{l\in U,l\neq k}y_k y_l\left(1-\frac{\pi_{kl}}{\pi_k\pi_l}\right)\\
&=\sum_{k\in U}\frac{y_k^2}{\pi_k}\left(\sum_{l\in U,l\neq k}\pi_l-\frac{1}{\pi_k}\frac{y_k^2}{\pi_k}\pi_{kl}\right)-\sum_{k\in U}\sum_{l\in U,l\neq k}y_k y_l\left(1-\frac{\pi_{kl}}{\pi_k\pi_l}\right)
\end{aligned}
$$

Variance de l'estimateur de Horvitz-Thompson

Démonstration.

...

Or, d'après le cours 1, on a dans le cas taille fixe : $\sum_{k\in U}\pi_k = n$ et $\sum_{l\in U, l\neq k}\pi_{kl} = \pi_k(n-1)$, cela donne :

$$
\begin{aligned}
&-\frac{1}{2}\sum_{k\in U}\sum_{l\in U, l\neq k}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\Delta_{kl}\\
&=\sum_{k\in U}\frac{y_k^2}{\pi_k}\left(n-\pi_k-\frac{\pi_k(n-1)}{\pi_k}\right)-\sum_{k\in U}\sum_{l\in U, l\neq k}y_k y_l\left(1-\frac{\pi_{kl}}{\pi_k\pi_l}\right)\\
&=\sum_{k\in U}\frac{y_k^2}{\pi_k^2}\pi_k(1-\pi_k)-\sum_{k\in U}\sum_{l\in U, l\neq k}\frac{y_k y_l}{\pi_k\pi_l}\left(\pi_k\pi_l-\pi_{kl}\right)\\
&=\sum_{k,l\in U}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\Delta_{kl}
\end{aligned}
$$

Et on retombe bien sur la formule d'Horvitz-Thompson. □

Estimation de variance

Les quantités précédentes sont les **vraies variances**. On peut utiliser les estimateurs suivants, qui sont sans biais dès lors que $\forall k, l, \pi_{kl} > 0$:

$$\hat{\mathrm{Var}}(\hat{t}_{y\pi}) = \sum_{k \in s} \frac{y_k^2}{\pi_k^2}(1 - \pi_k) - \sum_{k \in s} \sum_{l \in s, l \neq k} \frac{y_k y_l}{\pi_k \pi_l \pi_{kl}}(\pi_k \pi_l - \pi_{kl})$$

Pour un plan de taille fixe :

$$\hat{\mathrm{Var}}(\hat{t}_{y\pi}) = -\frac{1}{2} \sum_{k \in s} \sum_{l \in s, l \neq k} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}}$$

Construction d'un intervalle de confiance

On fait **l'hypothèse** : $\hat{\Phi}(s) \sim \mathcal{N}(\Phi, \mathrm{Var}(\Phi))$

L'intervalle de confiance à 95% est défini par :

$$IC_{95\%} = \left[\hat{\Phi} - 2\sigma(\hat{\Phi}); \hat{\Phi} + 2\sigma(\hat{\Phi})\right]$$

L'intervalle de confiance **estimé** est défini par :

$$\hat{IC}_{95\%} = \left[\hat{\Phi} - 2\hat{\sigma}(\hat{\Phi}); \hat{\Phi} + 2\hat{\sigma}(\hat{\Phi})\right]$$

Partie 5

Divers plans de sondage

Paragraphe 1

Le sondage aléatoire simple

Définition

Sondage aléatoire simple sans remise (SAS) de taille n : plan de sondage sans remise de taille fixe n tel que tous les échantillons de taille n ont la même probabilité d'être tirés. Cette probabilité vaut :

$$\begin{aligned} p(s) &= \frac{1}{\binom{n}{N}} && \text{si } |s| = \text{n} \\ &= 0 && \text{sinon.} \end{aligned}$$

On note le taux de sondage : $f = \frac{n}{N}$

Probabilités d'inclusion

$$\forall k \in \mathcal{U}, \pi_k = \mathbb{P}(k \in s) = \frac{n}{N} = f$$

$$\forall k \neq l \in \mathcal{U}, \pi_{k,l} = \mathbb{P}(k \wedge l \in s) = \frac{n(n-1)}{N(N-1)}$$

Estimateur d'Horvitz-Thompson

L'estimateur d'Horvitz-Thompson pour le total et la moyenne s'écrit :

$$T(\hat{Y}) = \sum_{k \in s} \frac{1}{\pi_k} y_k = \frac{N}{n} \sum_{k \in s} y_k = N\bar{y}$$

$$\hat{\bar{Y}} = \frac{1}{N} \sum_{k \in s} \frac{1}{\pi_k} y_k = \bar{y}$$

Poids de sondage

Les poids pour l'estimation par Horvitz-Thompson sont :

$$w_k = \frac{1}{\pi_k} = \frac{N}{n}$$

On peut dire que l'individu k "représente" $w_k = \frac{N}{n}$ individus de la population $\mathcal{U}$.

Attention, w_k n'est pas un effectif (en particulier, w_k n'est pas forcément entier !)

Précision

Théorème

En utilisant la formule de Yates-Grundy, la **vraie** *variance des estimateurs d'Horvitz-Thompson s'écrit :*

$$\mathrm{Var}(\bar{\mathrm{y}}) = (1-f)\frac{S^2}{n}$$

$$\mathrm{Var}(\mathrm{T}(\hat{\mathrm{Y}})) = N^2(1-f)\frac{S^2}{n}$$

Précision

Démonstration.

$$
\begin{aligned}
\mathrm{Var}[\hat{\bar{\mathrm{Y}}}] &= \frac{1}{N^2}\mathrm{Var}[\mathrm{T}(\hat{\mathrm{Y}})] \\
&= \frac{-1}{2N^2}\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U},l\neq k}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\Delta_{kl} \\
&= \frac{1}{2N^2}\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U},l\neq k}\left(\frac{y_k N}{n}-\frac{y_l N}{n}\right)^2\frac{n(N-n)}{N^2(N-1)} \\
&= \frac{N-n}{nN}\frac{1}{2N(N-1)}\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U},l\neq k}(y_k-y_l)^2 \\
&= \frac{N-n}{nN}S^2 \\
&= (1-f)\frac{S^2}{n}
\end{aligned}
$$

Estimation de la précision

Théorème

La variance empirique (ou dispersion) dans l'échantillon

$s^2 = \frac{1}{n-1}\sum_{k\in s}(y_k - \bar{y})^2$ *est un estimateur sans biais de*

$S^2 = \frac{1}{N-1}\sum_{k\in \mathcal{U}}(Y_k - \bar{Y})^2$

Estimation de la précision

Démonstration.

$$
\begin{aligned}
\mathbb{E}[s^2] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{k\in s}(y_k-\bar{y})^2\right] \\
&= \mathbb{E}\left[\frac{1}{2n(n-1)}\sum_{k\in s}\sum_{l\in s, l\neq k}(y_k-y_l)^2\right] \\
&= \frac{1}{2n(n-1)}\sum_{k\in \mathcal{U}}\sum_{l\in \mathcal{U}, l\neq k}(y_k-y_l)^2\mathbb{E}(\delta_k\delta_l) \\
&= \frac{1}{2n(n-1)}\sum_{k\in \mathcal{U}}\sum_{l\in \mathcal{U}, l\neq k}(y_k-y_l)^2\frac{n(n-1)}{N(N-1)} \\
&= \frac{1}{2N(N-1)}\sum_{k\in \mathcal{U}}\sum_{l\in \mathcal{U}, l\neq k}(y_k-y_l)^2 \\
&= S^2
\end{aligned}
$$

Estimation de la précision

On peut estimer sans biais la variance de l'estimateur d'Horvitz-Thompson par :

$$\hat{\mathrm{Var}}(\bar{y}) = (1-f)\frac{s^2}{n}$$

$$\hat{\mathrm{Var}}(T(\hat{Y})) = N^2(1-f)\frac{s^2}{n}$$

Paragraphe 2

Autres plans de sondage

Divers plans de sondage

Plan de sondage	Information auxiliaire	Difficulté du tirage et estimation	Précision	Coût terrain
Sondage aléatoire simple	non	★	+	$$$
Sondage stratifié, allocation proportionnelle	oui	★ ★	+++	$$$
Sondage stratifié, allocation optimale	oui	★ ★ ★	++++	$$$

Divers plans de sondage

Plan de sondage	Information auxiliaire	Difficulté du tirage et estimation	Précision	Coût terrain
Sondage à plusieurs degrés quelconque	oui	★ ★ ★	–	$$
Sondage en grappes	oui	★	– –	$
Sondage à probabilités inégales	oui	★	Si Y est proportionnel à X : ++++ sinon : –	$$$

Divers plans de sondage

Plan de sondage	Information auxiliaire	Difficulté du tirage et estimation	Précision	Coût terrain
Sondage équilibré	oui	★ ★ ★ ★	++++	$$$
Sondage par quotas ou unités-types	non	★ ★	NC	$

Chapitre 2

Redressements

Partie 1

Notations et objectifs

Redressements

On suppose que l'on dispose de J variables auxiliaires : $X_1, \ldots, X_j, \ldots, X_J$, qui sont mesurées pour tout individu de l'échantillon, et dont on connaît les totaux sur la population : $T(X_j) = \sum_{k \in \mathcal{U}} x_{jk}$.

L'échantillon fournit des estimateurs des X_j : $\hat{T}(X_j)_\pi = \sum_{k \in s} \frac{x_{jk}}{\pi_k}$,

qui n'ont aucune raison de coïncider avec les $T(X_j)$ connus, et auront en général une variance non nulle.

Redressements

Deux objectifs :

1. Prendre en compte l'information auxiliaire pour essayer de rendre l'estimation des totaux des variables d'intérêt plus précise
2. Faire en sorte que les estimations des totaux des variables auxiliaires soient exactes (cohérence des données publiées)

Remarque

Ici, l'information auxiliaire est utilisée non pas au stade du tirage de l'échantillon, mais à celui de l'estimation, c'est-à-dire une fois que l'échantillon est tiré et que l'enquête est réalisée.

Ceci est parfois rendu nécessaire parce que les variables auxiliaires ne sont pas connues au niveau individuel dans la base de sondage, mais seulement au moment de la collecte des données.

Partie 2

Plan de la formation

Définition

Définition (Estimateur linéaire homogène (pondéré))

Soit $\hat{T}_{Yl}$ un estimateur du total $T(Y)$ utilisant les valeurs de l'échantillon s. On dit que $\hat{T}_{Yl}$ est **linéaire homogène** *(ou "pondéré") s'il s'écrit sous la forme :*

$$\hat{T}_{Yl} = \sum_{k \in s} w_k(S) y_k$$

En particulier, $w_k(S)$ ne doit pas dépendre de Y.

Estimateur linéaire homogène

Intérêt : travailler avec une colonne de poids

ident	Sexe	Département	Salaire	Poids
1	f	13	1200	470
2	h	75	1500	150
3	f	59	3000	1250
		⋮		
19999	h	18	1500	500
20000	f	69	2100	815

Définition

Définition (Estimateur calé)

Soit $\hat{T}_{Yc}$ un estimateur du total $T(Y)$ utilisant les valeurs de l'échantillon s, et soit X une variable auxiliaire connue sur s et dont le total $T(X)$ (sur $\mathcal{U}$) est également connu. $\hat{T}_{Yc}$ est dit **calé** *si :*

$$\hat{T}_{Xc} = T(X)$$

c'est-à-dire que $\hat{T}_{Xc}$ estime parfaitement $T(X)$ (variance nulle).

Déroulement de la formation

1. Partie théorique (avec un peu de pratique)
2. Partie pratique (avec un peu de théorie)

Partie théorique

Logique de la partie théorique : Avant d'aboutir au calage sur marges, présenter différents estimateurs "historiques" (redressements pour des variables auxiliaires quantitatives, puis qualitatives) pour lesquels on essayera de déterminer s'ils possèdent des propriétés intéressantes :

- Estimateurs linéaires homogènes
- Estimateurs calés
- En termes de biais
- En termes d'augmentation de la précision

Partie pratique

- Deux exercices avec données simples, pour comprendre l'estimation par redressement
- Deux exercices sur données réelles