

Théorie des sondages : cours 4

Méthodes de redressement

Camelia GOGA

LMB, Université de Franche-Comté
e-mail : camelia.goga@univ-fcomte.fr

Master 2 Modélisation Statistique Besançon-2017

Techniques de redressement

1. l'estimateur par le ratio
2. poststratification

Question : comment peut-on faire pour améliorer l'estimateur de Horvitz-Thompson

$$\hat{t}_{\pi} = \sum_s \frac{y_k}{\pi_k}?$$

L'estimateur d'HT ne contient pas d'information auxiliaire à part les probabilités d'inclusion π_k .

Réponse : en prenant en compte l'information auxiliaire.

Information auxiliaire : des variables $\mathcal{X}_1, \dots, \mathcal{X}_J$ prenant les valeurs $x_{1k}, \dots, x_{Jk}$ pour le k -ème individu. Deux situations :

1. les valeurs x_{jk} de la variable $\mathcal{X}_j$ sont connues pour chaque individu $k \in U$ de la population avant la mise en place de l'enquête ou
2. le total de $\mathcal{X}_j$, $\sum_U x_{jk}$ est connu sans avoir accès aux valeurs individuelles x_{jk} .

On note $\mathbf{x}_k = (x_{1k}, \dots, x_{Jk})'$ pour tous $k \in U$.

Prise en compte de l'information auxiliaire : estimateur par le ratio et poststratifié

Information auxiliaire : une caractéristique de la théorie des sondages.

Ce sont des variables disponibles dans la base de sondage.

Ces variables peuvent servir de deux façons pour améliorer la qualité des estimations :

- ▶ lors de la conception du plan de sondage : le plan stratifié, à probabilités inégales, par grappes ou à plusieurs degrés ; le plan équilibré (pas dans ce cours)
- ▶ lors de la conception d'un nouvel estimateur : poststratifié, l'estimateur par le ratio, par la régression, par le calage (pas dans ce cours).

Comment améliorer l'estimation après le tirage de l'échantillon ?

- ▶ Le plan stratifié ou le plan proportionnel à la taille permettent d'améliorer l'estimateur d'Horvitz-Thompson.
- ▶ Ces plans de sondages demandent que l'information auxiliaire soit connue **avant l'enquête** pour chaque individu de la population. Parfois, on n'a que les valeurs agrégées de cette information : le nombre total de filles ou garçons.
Si on pense que notre information est liée à notre variable d'intérêt, alors comment peut-on l'utiliser ?
Il s'agit de construire **à postériori** (après l'enquête), un estimateur plus efficace que l'estimateur d'Horvitz-Thompson.
- ▶ L'estimateur **par le ratio** est l'exemple le plus simple d'un tel estimateur. Il a été proposé pour la première fois par Laplace (1802).
- ▶ Il est possible également de combiner les deux procédés, par exemple le plan stratifié et l'estimateur par le ratio.

Exemple

- ▶ Considérons l'exemple d'une enquête agricole américaine de 1992 : on veut estimer la superficie totale d'une région dédiée aux fermes ;
- ▶ On ne dispose pas de l'information nécessaire pour stratifier la population de fermes en fonction de leurs tailles, alors il est décidé de faire une enquête par SAS ;
- ▶ On dispose par ailleurs de la superficie totale en 1987. La superficie d'une ferme en 1992 est très liée à celle de 1987, donc on aimerait utiliser cette information, mais comment ?
- ▶ **Réponse** : on utilise l'estimateur par le ratio.

L'estimateur par le ratio : exemple de Laplace

Objectif : estimer la population de la France vers 1800.

- ▶ Laplace sélectionne un échantillon de taille 30 de communes en France et il obtient une estimation de $\hat{t}_{habitants,30com} = 2037615$ habitants ;
- ▶ il peut disposer du nombre de naissances dans ces 30 communes et il estime le nombre total de naissances par $\hat{t}_{naissances,30com} = 71866$.
Donc, il y a

$$\text{une naissance pour } \frac{2037615}{71866} = 28.35 \text{ personnes}$$

- ▶ il connaît également **le nombre total de naissances** en 1802 et il fait le raisonnement que les communes avec beaucoup d'habitants vont avoir plus de naissances, donc il propose d'estimer le nombre total d'habitants en France par

$$\hat{t}_{habitants} = t_{naissances} \times \frac{\hat{t}_{habitants,30com}}{\hat{t}_{naissances,30com}}$$

Dans l'exemple de Laplace :

- ▶ le nombre de naissances est **l'information auxiliaire** qui est très liée à la variable nombre d'habitants ;
- ▶ on n'a besoin que du nombre total de naissances, t_x , et du nombre de naissances dans chaque ville sélectionnée : x_k pour $k \in s$.
- ▶ L'estimateur proposé par Laplace est beaucoup utilisé en pratique pour estimer des totaux ;
- ▶ Il s'appelle **l'estimateur par le ratio** :

$$\hat{t}_{y, ratio} = t_x \times \frac{\hat{t}_y}{\hat{t}_x} = t_x \times \hat{R}$$

où $\hat{t}_y = \sum_{k \in s} \frac{y_k}{\pi_k}$ et $\hat{t}_x = \sum_{k \in s} \frac{x_k}{\pi_k}$;

- ▶ La variable peut être aussi qualitative (ex. $x_k = 1$ si l'individu k est une fille et zéro sinon).

Propriétés

- ▶ L'estimateur par le ratio est une somme pondérée des valeurs de y sur l'échantillon :

$$\hat{t}_{y, ratio} = \sum_{k \in s} \underbrace{\frac{t_x}{\hat{t}_x} \frac{1}{\pi_k}}_{w_{ks}} y_k = \sum_{k \in s} w_{ks} y_k$$

- ▶ Ces poids vérifient la propriété suivante :

$$\sum_{k \in s} w_{ks} x_k = \sum_{k \in U} x_k$$

autrement dit, on estime parfaitement le total de la variable auxiliaire quel que soit l'échantillon s :

$$\hat{t}_{x, ratio} = t_x.$$

- ▶ En général, cet estimateur est biaisé mais pour des échantillons de grandes tailles, ce biais est très proche de zéro.

Exemple (Lohr, 1999)

Supposons que nous voulons estimer le nombre de poissons pêchés au long d'un lac. On considère un plan aléatoire simple sans remise de taille $n = 2$ dans une population de taille $N = 4$:

- ▶ nous avons $N = 4$ endroits pour pêcher ;
- ▶ on connaît le nombre x_i de filets qui se trouvent à chaque endroit pour pêcher.

Les valeurs pour toute la population se trouvent dans le tableau suivant :

endroit, i	1	2	3	4	total
filet, x_i	4	5	8	5	$t_x = 22$
nombre poissons, y_i	200	300	500	400	$t_y = 1400$

L'estimateur par le ratio est $\hat{t}_{y_{rat}} = t_x \times \frac{\sum_s y_i}{\sum_s x_i}$ où s est un des 6 échantillons possibles de taille 2. Par exemple, si $s = (1, 2)$

$$\hat{t}_{y_{rat}} = 22 \times \frac{200+300}{4+5} = 1222.$$

échantillon	$\hat{t}_{y_{rat}}$	$(\hat{t}_{y_{rat}} - t_y)^2$
(1, 2)	1222	31684
(1, 3)	1283	13689
(1, 4)	1467	4489
(2, 3)	1354	2116
(2, 4)	1540	19600
(3, 4)	1523	15129

L'espérance de $\hat{t}_{y_{rat}}$ est la moyenne arithmétique des valeurs possibles de $\hat{t}_{y_{rat}}$ pour chaque échantillon, $E(\hat{t}_{y_{rat}}) = 1398.17$ et le biais est

$$1398.17 - 1400 = -1.83$$

L'erreur quadratique moyenne est donnée par

$$EQR(\hat{t}_{y_{rat}}) = E(\hat{t}_{y_{rat}} - t_y)^2 = \frac{1}{6} \sum_s (\hat{t}_{y_{rat}} - t_y)^2 = 14451,2$$

Considérons maintenant l'estimateur par les valeurs dilatées qui ne prend pas en compte l'information auxiliaire : $\hat{t}_y = \frac{N}{n} \sum_s y_i$:

échantillon	$\hat{t}_y$	$(\hat{t}_y - t_y)^2$
(1, 2)	$2 \cdot (200 + 300) = 1000$	160000
(1, 3)	1400	0
(1, 4)	1200	40000
(2, 3)	1600	40000
(2, 4)	1400	0
(3, 4)	1800	160000

Le biais de cet estimateur est 0 mais sa variance est 66 667. Alors, l'estimateur par le ratio est légèrement biaisé mais son EQR est nettement inférieure à celle de l'estimateur par les valeurs dilatées qui est sans biais.

Précision de l'estimateur par le ratio

- L'estimateur par le ratio $\hat{t}_{rat} = t_x \hat{R}$ est un estimateur non-linéaire et il est linéarisé par :

$$\hat{t}_{rat} \simeq t_y + \sum_s \frac{u_k}{\pi_k},$$

avec u_k la variable linéarisée de $\hat{t}_{rat}$ donnée par $u_k = y_k - R x_k$

- Le biais est donné par

$$E(\hat{t}_{rat}) - t_y = -\text{Cov}(\hat{R}, \hat{t}_x)$$

- Pour n grand, $\text{Var}(\hat{t}_{rat}) \simeq \text{Var} \left(\sum_s \frac{u_k}{\pi_k} \right)$

Comparaison entre l'estimateur de HT et l'estimateur par le ratio pour un plan SRS

- ▶ Nous voulons estimer le total $t_y = \sum_{k \in U} y_k$;
- ▶ On considère un échantillon SRS de taille n ;
 - ▶ **sans inf. aux.** et en utilisant l'estimateur par les valeurs dilatées :

$$\hat{t}_y = \frac{N \sum_s y_k}{n}$$

avec une variance estimée égale à

$$N^2 \frac{1-f}{n} S_{ys}^2 = N^2 \frac{1-f}{n} \frac{1}{n-1} \sum_s (y_k - \bar{y}_s)^2;$$

- ▶ **avec l'inf.aux.** et l'estimateur par le ratio :

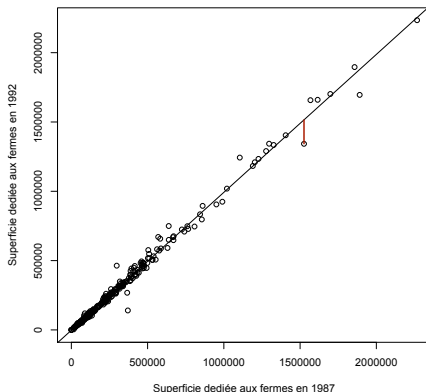
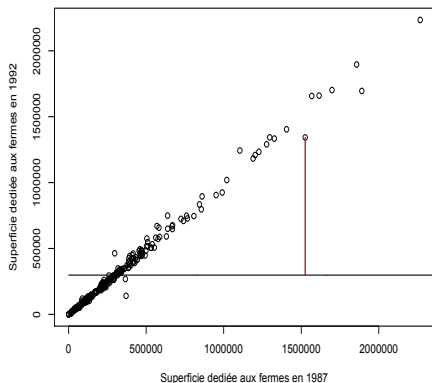
$$\hat{t}_{y_{rat}} = t_x \cdot \frac{\hat{t}_y}{\hat{t}_x} = t_x \cdot \frac{\sum_s y_k}{\sum_s x_k}$$

avec une variance estimée égale à

$$N^2 \frac{1-f}{n} S_{y-\hat{R}x,s}^2 = N^2 \frac{1-f}{n} \frac{1}{n-1} \sum_s (y_k - \hat{R}x_k)^2$$

car $\bar{y}_s - \hat{R}\bar{x}_s = 0$.

Si on compare les deux formules, dans la première situation nous avons la somme des carrés des écarts de y_k à la moyenne $\bar{y}_s$ et dans la deuxième situation, nous avons la somme des carrés des écarts de y_k à la droite de régression $y = \hat{R}_x$



En conclusion

1. L'estimateur par le ratio est conseillé lorsque la relation entre $\mathcal{Y}$ et $\mathcal{X}$ est proche d'une droite passant par l'origine et la variance de $\mathcal{Y}$ est proportionnelle avec $\mathcal{X}$, $\text{var}(y_k) = \sigma^2 x_k$;
2. Dans ces conditions, le ratio $\hat{R}$ est la pente de la droite de régression obtenue par des moindres carrés pondérés :

$$\hat{R} = \operatorname{argmin}_R \sum_s \frac{(y_k - R x_k)^2}{x_k}$$

Application de l'estimateur par le ratio : estimation dans un domaine quand la taille du domaine est connue

- Soit $U_d \subset U$ un domaine et nous voulons estimer le total d'une variable $\mathcal{Y}$ dans le domaine U_d ,

$$t_d = \sum_{U_d} y_k$$

- Soit $s \subset U$ un échantillon SRS de taille n et $s_d = s \cap U_d$.
- L'estimateur de HT de t_d (voir TD 1) est $\hat{t}_\pi = \frac{N}{n} \sum_{s_d} y_k$.
- Si on connaît la taille du domaine, $N_d = \sum_U \mathbf{1}_{k \in U_d}$ on peut utiliser l'estimateur par le ratio qui est plus efficace :

$$\hat{t}_{d, \text{rat}} = N_d \times \frac{\sum_{s_d} y_k}{\sum_{s_d} \mathbf{1}_{k \in U_d}} = N_d \bar{y}_{s_d}$$

L'estimateur poststratifié : exemple 1

Considérons l'exemple suivant :

- ▶ on veut estimer le nombre total d'étudiants qui veulent être professeurs après leurs études dans une population de $N = 4000$ étudiants :

$$t_y = \sum_{k \in U} y_k$$

avec

$$y_k = \begin{cases} 1 & \text{si l'individu } k \text{ veut être prof.} \\ 0 & \text{sinon} \end{cases}$$

- ▶ on sait par ailleurs que dans la population totale on a $N_F = 2400$ femmes (60%) et $N_H = 1600$ hommes (40%) et on pense que la population de femmes (ou d'hommes) est relativement homogène vis-à-vis de la volonté d'être professeur ; on désire donc, utiliser cette information auxiliaire ;
- ▶ il n'est pas possible de stratifier (avant l'enquête) la population en deux strates : hommes et femmes. Donc, on prend un échantillon aléatoire simple sans remise de taille $n = 400$; on utilisera l'information N_F et N_H après le tirage de l'échantillon.

- ▶ dans notre échantillon, il y a $n_F = 320$ femmes (80% dans l'échantillon), dont 120 veulent être des professeurs, et $n_H = 80$ (20% dans l'échantillon) hommes dont 40 veulent être des professeurs.
- ▶ Avec l'estimateur d'Horvitz-Thompson, tous les individus de l'échantillon (H ou F) ont le même poids : $w_k = \frac{N}{n} = 10$ et

$$\hat{t}_\pi = \sum_{k \in s} w_k y_k = 10 \times 160 = 1600$$

- ▶ On construit l'estimateur poststratifié : on augmente le poids des hommes de l'échantillon (sous-représentés) et on baisse le poids des femmes (sur-représentées) :

$$w_{kH} = \frac{N_H}{n_H} = 20 \quad \text{et} \quad w_{kF} = \frac{N_F}{n_F} = \frac{60}{8}$$

$$\hat{t}_{post} = \sum_{k \in s_H} w_{kH} y_k + \sum_{k \in s_F} w_{kF} y_k = 20 \times 40 + \frac{60}{8} \times 120 = 1700$$

- ▶ Cet estimateur peut être vu comme la somme de deux estimateurs par le ratio par la taille de chaque groupe (**règle de trois** à l'intérieur de chaque groupe) !

L'estimateur poststratifié : exemple 2

Considérons l'exemple suivant :

On veut estimer la somme moyenne dédiée à la nourriture/mois dans les ménages américains. Nous avons la distribution de la taille des ménages comme suit :

taille du ménage	pourcentage
1	25.75
2	31.17
3	17.50
4	15.58
5+	10.00

Malheureusement, la base de sondage ne contient pas l'information concernant la taille de chaque ménage, par conséquent on ne peut pas faire une stratification.

La consommation est fortement liée à la taille du ménage, donc on aimerait bien utiliser cette information, mais comment ?

On peut faire une **poststratification**. Cela revient à considérer la somme des estimateurs par le ratio à l'intérieur de chaque groupe (type de ménage).

Notations et résultats

Considérons G groupes ou poststrata $U_1, \dots, U_g, \dots, U_G$. Un échantillon s de taille n est sélectionné dans U selon un plan SAS (le plus utilisé mais un autre type de plan de sondage peut être envisagé).

Soit $s_g = s \cap U_g$ de taille aléatoire n_{s_g} la partie de s se trouvant dans le groupe U_g ,

$$s = \bigcup_{g=1}^G s_g, \quad n = \sum_{g=1}^G n_{s_g}.$$

L'estimateur post-stratifié du total est la somme des G estimateurs par le ratio considérés à l'intérieur de chaque groupe :

$$\hat{t}_{y\text{post}} = \sum_{g=1}^G N_g \bar{y}_{s_g}$$

avec $\bar{y}_{s_g} = \sum_{s_g} y_k / n_{s_g}$ pour $n_{s_g} \neq 0$.

Propriétés

- ▶ La méthode d'estimation utilisée est appelée *poststratification* car utilise l'information auxiliaire (le fait d'appartenir à un certain groupe) après le tirage de l'échantillon.
- ▶ Dans la plupart des cas, l'estimateur poststratifié est meilleur que l'estimateur par les valeurs dilatées et il est équivalent au plan stratifié avec l'allocation proportionnelle.

$$\text{variation totale} = \text{variation inter-groupes} + \text{variation intra-groupes}$$

La poststratification réduit la variance si la variation entre groupes est grande.

- ▶ Dans l'exemple donné, l'information auxiliaire est le fait que la population est plus homogène à l'intérieur de chaque groupe que dans la population totale.

Propriétés

- ▶ La méthode d'estimation utilisée est appelée *poststratification* car utilise l'information auxiliaire (le fait d'appartenir à un certain groupe)

$$X_{kg} = \mathbf{1}_{kg} = 1 \quad \text{si } k \text{ appartient au groupe}$$

après le tirage de l'échantillon.

- ▶ L'estimateur poststratifié peut s'écrire comme un estimateur pondéré

$$\hat{t}_{post} = \sum_{k \in s} w_{ks} y_k$$

et les poids w_{ks} vérifient :

$$\sum_{k \in s} w_{ks} \mathbf{1}_{kg} = \sum_{k \in U} \mathbf{1}_{kg} = N_g, \quad g = 1, \dots, G$$

c'est à dire, on estime parfaitement les effectifs $N_1, \dots, N_G$ quel que soit l'échantillon s .

Variance et estimateur de la variance

- La variance approximative est donnée par

$$AV(\hat{t}_{ypost}) = N^2 \frac{1-f}{n} \sum_{g=1}^G W_g S_{yU_g}^2 + N^2 \frac{1-f}{n^2} \sum_{g=1}^G (1 - W_g) S_{yU_g}^2$$

pour $W_g = \frac{N_g}{N}$ et $S_{yU_g}^2 = \frac{1}{N_g - 1} \sum_{U_g} (y_k - \bar{y}_{U_g})^2$.

- La poststratification réduit la variance si la variation entre groupes est grande.
- La variance de l'estimateur post-stratifié est approximativement égale à la variance de l'estimateur d'HT sous le plan stratifié avec allocation proportionnelle (est le premier terme de l'expression de $AV(\hat{t}_{ypost})$).
- L'estimateur de la variance est

$$\hat{V}(\hat{t}_{ypost}) \simeq N^2 \frac{1-f}{n} \sum_{g=1}^G \frac{N_g}{N} S_{ys_g}^2$$

pour $S_{ys_g}^2 = \frac{1}{n_{s_g} - 1} \sum_{s_g} (y_k - \bar{y}_{s_g})^2$.

Quand utiliser la poststratification ?

1. dans le cas des sondages à plusieurs items : on peut utiliser différentes variables auxiliaires pour différentes variables d'intérêt.
2. dans le traitement de la nonréponse ;
3. l'appartenance d'un individu à un certain groupe n'est pas connue au moment de la mise en place de l'enquête.