Théorie des sondages : cours 5

Camelia Goga

IMB, Université de Bourgogne
e-mail : camelia.goga@u-bourgogne.fr

Master Besançon-2010

Chapitre 5 : Techniques de redressement

1. poststratification
2. l'estimateur par le ratio
3. l'estimateur par la régression

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 $\mathcal{X}_1, \ldots, \mathcal{X}_J$ prenant les valeurs $x_{1k}, \ldots, x_{Jk}$ pour le k-ème individu et connues avant la mise en place d'une enquête. Deux situations :

1. une variable $\mathcal{X}_j$ est connue pour chaque individus de la population ou
2. le total de $\mathcal{X}_j$, $\sum_U x_{jk}$ est connu sans avoir accès aux valeurs individuelles x_{jk}.

On note $\mathbf{x}_k = (x_{1k}, \ldots, x_{Jk})'$ 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 : la 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).

L'estimateur par le ratio

Laplace (1802) a été le premier a avoir proposé faire un sondage au lieu d'une énumération exhaustive ;

Objectif : estimer la population de la France.

Méthode : l'estimateur par ratio (une "règle de trois") :

- il prend un échantillon de taille 30 de communes de toute la France et il obtient un total de $\hat{t}_{habitants,30com} = 2037615$ habitants ;
- il connaît également le nombre de naissances dans ces 30 communes : $\hat{t}_{naissances,30com} = 71866$ ce qui implique que il y a

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

- il connaît le nombre total de naissances par an en 1802 et il fait le raisonnement que les communes avec beaucoup d'habitants vont avoir plus de naissances, il estime le nombre total d'habitants comme

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

Dans l'exemple de Laplace : le nombre de naissances est l'information auxiliaire qui est bien corrélée avec la variable nombre d'habitants. (Il doit connaître que t_x et x_k pour $k \in s$.)
L'estimateur proposé par Laplace est beaucoup utilisé pour estimer des totaux : et il s'appelle **l'estimateur par ratio** :

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

Propriétés : en général, cet estimateur est biaisé mais pour des échantillons de grande taille, ce biais est très proche de zéro. Si l'information auxiliaire est très corrélée avec la variable d'intérêt alors, l'erreur quadratique moyenne de $\hat{t}_{y,ratio}$ est plus petite que si on n'avait pas utilisé d'information auxiliaire.

Exemple (Lohr, 1999)

Supposons que nous voulons estimer le nombre de poissons péchés 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 pécher ;
- on connaît le nombre x_i de filets qui se trouvent à chaque endroit pour pécher.

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 $\hat{t}_{yrat} = t_x \times \frac{\sum_s y_i}{\sum_s x_i}$ où s est un des 6 échantillons possibles de taille 2. Par exemple, si $s = (1, 2)$ $\hat{t}_{yrat} = 22 \times \frac{200+300}{4+5} = 1222$.

échantillon	$\hat{t}_{yrat}$	$(\hat{t}_{yrat} - t_y)^2$
$(1, 2)$	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 $\hat{t}_{yrat}$ est la moyenne arithmétique des valeurs possibles de $\hat{t}_{yrat}$ pour chaque échantillon, $E(\hat{t}_{yrat}) = 1398.17$ et le biais est

$$1398.17 - 1400 = -1.83$$

L'erreur quadratique moyenne est donnée par

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

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 $u_k = y_k - Rx_k$

- 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 :

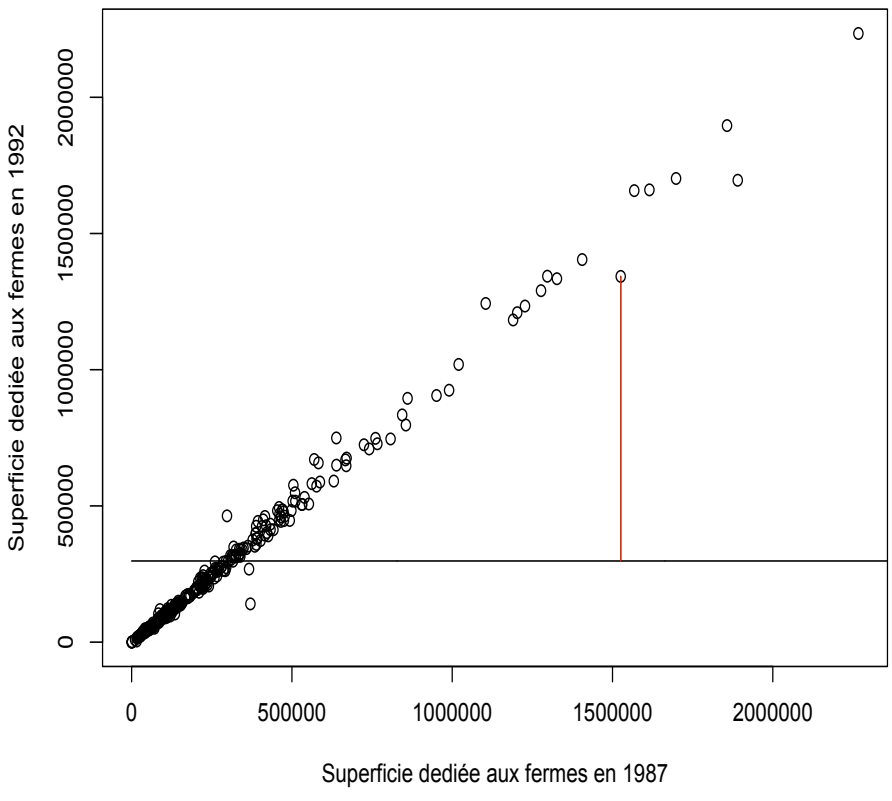
$$\hat{t}_{yrat} = t_x \cdot \frac{\hat{t}_y}{\hat{t}_x} = t_x \cdot \frac{\sum_s y_k}{\sum_s x_k}$$

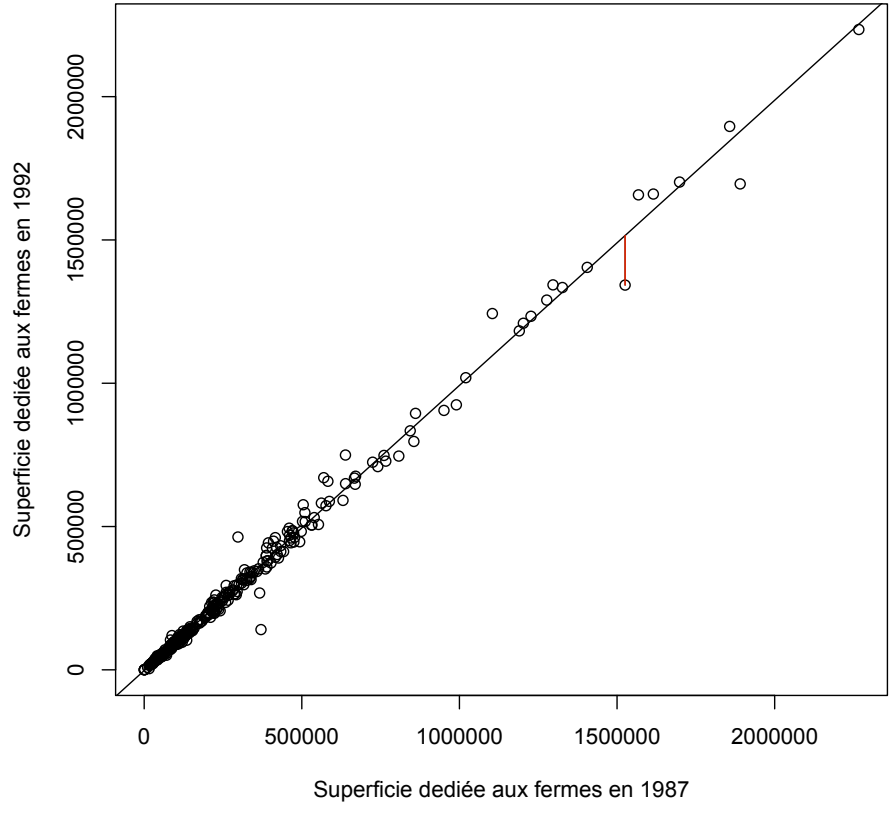
avec une variance estimée égale à

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

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 passante par l'origine et la variance de $\mathcal{Y}$ est proportionnelle avec $\mathcal{X}$, $var(y_k) = \sigma^2 x_k$;
2. Dans ces conditions, le ratio $\hat{R}$ est la pente de la droite de régression obtenue par des moindres carrées pondérés :

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

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 :

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

Poststratification

Considérons l'exemple suivant (Lohr, 1999) :

- on veut estimer le nombre total d'étudiants qui veulent être professeurs après leurs études dans une population de 4000 étudiants ;
- on prend un échantillon aléatoire simple sans remise de taille $n = 400$;
- on sait par ailleurs que dans la population totale on a 2700 femmes et 1300 hommes ;
- dans notre échantillon, il y a 240 femmes dont 84 veulent être des professeurs et 160 hommes dont 40 veulent être des professeurs

On utilise cette information et on obtient l'estimation suivante :

$$\frac{84}{240} \times 2700 + \frac{40}{160} \times 1300 = 1270$$

Il s'agît d'une règle de trois à l'intérieur de chaque groupe !

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.

  variation totale=variation inter-groupes+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.

Notations et résultats

Considérons les notations utilisées lors du plan stratifié : 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 la sous-population 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 :

$$\hat{t}_{ypost} = \sum_{g=1}^{G} N_g \overline{y}_{s_g}$$

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

L'estimateur post-stratifié : 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 $S_{yU_g}^2 = \frac{1}{N_g - 1} \sum_{U_g} (y_k - \overline{y}_{U_g})^2$.

Remarque : Supposons que les sous-populations U_g sont des strates et on utilise l'allocation proportionnelle, alors la variance du π-estimateur est le premier terme de l'expression de $AV(\hat{t}_{ypost})$.

- L'estimateur de la variance est

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

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

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.

Exemple : On veut estimer la somme moyenne dédiée à la nourriture/mois dans les ménages américaines. 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.
Mais on peut faire une poststratification.

L'estimateur par la régression du total

Si la relation entre $\mathcal{Y}$ et $\mathcal{X}$ est proche d'une droite avec intercept, alors

$$y_k = \beta_0 + \beta_1 x_k = \mathbf{X}_k' \beta$$

où $\beta = (\beta_0, \beta_1)'$ et $\mathbf{X}_k = (1, x_k)'$ peut donner une bonne approximation du nuage des points.
L'estimateur par la regression pour le total $t_y = \sum_U y_k$ est

$$\begin{aligned} \hat{t}_{yr} &= \sum_s \frac{y_k - \hat{y}_k}{\pi_k} + \sum_U \hat{y}_k \\ &= \sum_s \frac{y_k - \mathbf{x}_k' \widehat{\mathbf{B}}}{\pi_k} + \sum_U \mathbf{x}_k' \widehat{\mathbf{B}} \end{aligned}$$

$$\hat{y}_k = \mathbf{x}_k' \widehat{\mathbf{B}}$$

et $\widehat{\mathbf{B}} = (\widehat{B}_0, \widehat{B}_1)'$ l'estimateur de β dans l'échantillon s.

L'estimateur par la regression pour un plan SRS

Le coefficient de regression β est estimé par moindres carrés dans la population par

$$\begin{pmatrix} B_0 \\ B_1 \end{pmatrix} = \begin{pmatrix} \overline{y}_U - B_1 \overline{x}_U \\ \frac{\sum_U (x_k - \overline{x}_U)(y_k - \overline{y}_U)}{\sum_U (x_k - \overline{x}_U)^2} \end{pmatrix}$$

qui est ensuite estimé dans l'échantillon par

$$\begin{pmatrix} \hat{B}_0 \\ \hat{B}_1 \end{pmatrix} = \begin{pmatrix} \overline{y}_s - \hat{B}_1 \overline{x}_s \\ \frac{\sum_s (x_k - \overline{x}_s)(y_k - \overline{y}_s)}{\sum_s (x_k - \overline{x}_s)^2} \end{pmatrix}$$

et $\hat{B}_1 = \frac{S_{xy,s}}{S_x^2}$.
L'estimateur par regression est

$$\hat{t}_{yr} = N \left(\overline{y}_s + \hat{B}_1 (\overline{x}_U - \overline{x}_s) \right)$$

Variance et estimateur de la variance pour un plan général

L'estimateur par régression $\hat{t}_{yr} = \hat{t}_{y\pi} + (t_{\mathbf{x}} - t_{\mathbf{x}\pi})'\widehat{\mathbf{B}}$ est approximativement sans biais pour le total $t_y = \sum_U y_k$ est sa variance approximative est donné par

$$AV(\hat{t}_{yr}) = \sum_U \sum_U \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}$$

pour $E_k = y_k - \mathbf{x}_k'\mathbf{B}$ et l'estimateur de la variance est

$$\widehat{V}(\hat{t}_{yr}) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_{ks}}{\pi_k} \frac{e_{ls}}{\pi_l}.$$

pour $e_{ks} = y_k - \mathbf{x}_k'\widehat{\mathbf{B}}$.

Variance pour un plan SRS

$$
\begin{aligned}
AV(\hat{t}_{yr}) &= N^2 \frac{1-f}{n} S^2_{E,U} = N^2 \frac{1-f}{n} \frac{1}{N-1} \sum_U (E_k - \overline{E}_U)^2 \\
&= N^2 \frac{1-f}{n} S^2_{y,U}(1-R^2)
\end{aligned}
$$

où R^2 est le coefficient de corrélation entre $\mathcal{X}$ et $\mathcal{Y}$ dans la population. Donc, cette variance est petite si :

- n est grand
- n/N grand
- S_y petit
- $R = \pm 1$ ou proche.