Formation calage sur marges - Redressements

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

INSEE - Division Sondages

29 avril 2015

Sommaire I

Sommaire II

Chapitre 1

L'estimateur par le ratio

Exemple introductif

On cherche à estimer le nombre d'habitants d'une région comportant $N = 2536$ villages. On tire un échantillon de $n = 127$ villages par sondage aléatoire simple. On observe une taille moyenne de $\bar{y} = 377.2$ habitants sur l'échantillon.

Pour chacun des 2536 villages, on connaît la population au dernier recensement, organisé trois ans auparavant.

Exemple introductif

Taille moyenne	Ensemble des villages de la région	Échantillon de villages
Au moment de l'enquête	?	377,2
Au moment du recensement	345,1	341,7

(extrait de Manuel de sondages, Applications aux pays en développement, R. Clairin et Ph. Brion, Documents et Manuels du CEPED numéro 3)

Partie 1

Principe

Principe

Au moment de l'enquête, on recueille deux types d'informations :

1. sur la variable d'intérêt Y
2. sur une variable auxiliaire X dont le total sur la population est connu.

Principe

Les estimateurs d'Horvitz-Thompson pour les totaux de X et Y sont :

$$\hat{T}_{Y\pi} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k$$

$$\hat{T}_{X\pi} = \sum_{k \in s} \frac{x_k}{\pi_k} = \sum_{k \in s} d_k x_k$$

$$\text{où : } d_k = \frac{1}{\pi_k} = \text{poids de sondage}$$

Principe

En général, $\hat{T}_{X\pi}$ est différent du vrai total connu $T(X)$. On se sert de la connaissance de $T(X)$ pour modifier l'estimateur de $T(Y)$:

- On suppose que la variable Y est, même approximativement, proportionnelle à la variable X : $y_k \approx R \cdot x_k$
- On a donc également : $T(Y) \approx R \cdot T(X)$
- on utilise l'échantillon pour estimer R : $\hat{R} = \dfrac{\hat{T}_{Y\pi}}{\hat{T}_{X\pi}}$
- On estime $T(Y)$ par $\hat{T}_{Y,ratio} = \hat{R} \cdot T(X)$

Partie 2

Définition

Définition

Définition (Estimateur par le ratio d'un total)

L'estimateur par le ratio (ou par le quotient) du total $T(Y)$ se définit par :

$$\hat{T}_{Y,ratio} = \frac{\hat{T}_{Y\pi}}{\hat{T}_{X\pi}} T(X) = \hat{R} \cdot T(X)$$

Définition

Remarque : On peut aussi écrire $\hat{T}_{Y,ratio} = \frac{T(X)}{\hat{T}_{X\pi}} \hat{T}_{Y\pi}$, et voir l'estimateur par le ratio comme l'estimateur d'Horvitz-Thompson corrigé du facteur $\frac{T(X)}{\hat{T}_{X\pi}}$, qui mesure l'écart entre l'estimation de X et sa vraie valeur : règle de trois !

Définition

Définition (Estimateur par le ratio d'une moyenne)

L'estimateur par le ratio (ou par le quotient) de la moyenne $\bar{Y}$ se définit par :

$$\bar{Y}_{ratio} = \frac{1}{N}\hat{T}_{Y,ratio} = \frac{\hat{T}_{Y\pi}}{\hat{T}_{X\pi}}\bar{X} = \frac{\bar{Y}_{\pi}}{\bar{X}_{\pi}} \cdot \bar{X}$$

Partie 3

Propriétés

Estimateur pondéré

Propriété

*L'estimateur par le ratio est un estimateur linéaire homogène (**pondéré**). Pour toute variable* $Z = (z_k)_{k \in [[1,n]]}$*, les poids :*

$$w_k = \frac{T(X)}{\hat{T}_{X\pi}} \frac{1}{\pi_k}$$

permettent de construire l'estimateur par le ratio sur la variable X *:* $\hat{T}_{Z,ratio} = \sum_{k \in s} w_k z_k$

Estimateur pondéré

Démonstration.

$$\begin{aligned}\hat{T}_{Y,ratio} &= \frac{T(X)}{\hat{T}_{X\pi}}\hat{T}_{Y\pi} \\ &= \sum_{k\in s}\frac{T(X)}{\hat{T}_{X\pi}}\frac{y_k}{\pi_k} \\ &= \sum_{k\in s} w_k y_k\end{aligned}$$

□

Propriété de calage

Propriété (de calage)

L'estimateur par le ratio sur la variable X estime parfaitement le total de X.

Propriété de calage

Démonstration.

$$\begin{aligned}\hat{T}_{X,ratio} &= \sum_{k\in s} w_k x_k \\ &= \sum_{k\in s} \frac{T(X)}{\hat{T}_{X\pi}} \frac{1}{\pi_k} x_k \\ &= \frac{T(X)}{\hat{T}_{X\pi}} \hat{T}_{X\pi} \\ &= T(X)\end{aligned}$$

□

Biais

Propriété (Estimateur asymptotiquement sans biais)

$$B(\hat{T}_{Y,ratio}) \underset{n\to+\infty}{\sim} \frac{C}{n}, C \in \mathbb{R}$$

En particulier :

$$B(\hat{T}_{Y,ratio}) \underset{n\to+\infty}{\to} 0$$

Biais

Démonstration.

On écrit :

$$\begin{aligned}\hat{T}_{Y,ratio} - \hat{T}_{Y\pi} &= T(X) \cdot \frac{\hat{T}_{Y\pi} - r\hat{T}_{X\pi}}{\hat{T}_{X\pi}} \\ &= \frac{\hat{T}_{Y\pi} - R\hat{T}_{X\pi}}{1+\epsilon} \text{ , avec : } \epsilon = \frac{\hat{T}_{X\pi} - T(X)}{T(X)}\end{aligned}$$

En faisant un développement limité à l'ordre 1 en ϵ, on obtient :

$$\begin{aligned}\hat{T}_{Y,ratio} - \hat{T}_{Y\pi} &\approx (\hat{T}_{Y\pi} - R\hat{T}_{X\pi})(1-\epsilon) \\ &\approx (\hat{T}_{Y\pi} - R\hat{T}_{X\pi})(1 - \frac{\hat{T}_{X\pi} - T(X)}{T(X)})\end{aligned}$$

Soit :

$$\mathbb{E}(\hat{T}_{Y,ratio} - \hat{T}_{Y\pi}) \approx \mathbb{E}\left[(\hat{T}_{Y\pi} - R\hat{T}_{X\pi})(1 - \frac{\hat{T}_{X\pi} - T(X)}{T(X)})\right]$$

Biais

Démonstration.

$$\approx \mathbb{E}(\hat{T}_{Y\pi} - R\hat{T}_{X\pi}) - \frac{\mathbb{E}(\hat{T}_{Y\pi}(\hat{T}_{X\pi} - T(X))) - R\mathbb{E}(\hat{T}_{X\pi}(\hat{T}_{X\pi} - T(X)))}{T(X)}$$

$$\approx \frac{R\mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{X}\pi}) - \mathrm{Cov}(\hat{\mathrm{T}}_{\mathrm{X}\pi}, \hat{\mathrm{T}}_{\mathrm{Y}\pi})}{T(X)}$$

$$\approx \frac{N(N-n)}{n}\frac{RS_x^2 - S_{xy}}{T(X)}$$

□

Précision

Propriété

$$\mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{Y,ratio}}) \approx \mathrm{Var}(\sum_{\mathrm{k}\in\mathrm{s}} \frac{1}{\pi_{\mathrm{k}}}(\mathrm{y_k} - \mathrm{Rx_k}))$$
$$= \mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{U}\pi})$$
$$\text{où : } u_k = y_k - Rx_k \text{ ("résidus")}$$

Précision

Démonstration.

On néglige le biais asymptotique, de sorte que $\mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{Y,ratio}}) \approx \mathbb{E}\left[(\hat{\mathrm{T}}_{\mathrm{Y,ratio}} - \hat{\mathrm{T}}_{\mathrm{Y}\pi})^2\right]$. Alors :

$$\begin{aligned}
\mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{Y,ratio}}) &\approx \mathbb{E}\left[(\hat{T}_{Y\pi} - R\hat{T}_{X\pi})^2\right] \\
&\approx \mathbb{E}\left[(\hat{T}_{Y\pi} - T(Y)) - R(\hat{T}_{X\pi} - T(X))^2\right] \\
&\approx \mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{Y}\pi}) + \mathrm{R}^2\mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{X}\pi}) - 2\mathrm{RCov}(\hat{\mathrm{T}}_{\mathrm{X}\pi}, \hat{\mathrm{T}}_{\mathrm{Y}\pi}) \\
&\approx \frac{N(N-n)}{n}(S_y^2 + R^2 S_x^2 - 2RS_{xy}^2) = \mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{U}\pi})
\end{aligned}$$

□

Précision

Cela signifie que si Y et X sont bien proportionnelles, alors les u_k seront petits. On aura donc une variance de l'estimateur d'Horvitz-Thompson plus petite pour u que pour Y, et l'estimation par le ratio apportera un gain en variance par rapport à l'estimation par Horvitz-Thompson.

Chapitre 2

L'estimateur par la régression

Cadre

- On se limite au SAS (Sondage Aléatoire Simple). On a donc : $\hat{\bar{Y}}_\pi = \bar{y}$, $\hat{\bar{X}}_\pi = \bar{x}$
- On suppose qu'il existe une relation linéaire approchée (voire très approchée) entre X et Y : $Y \approx aX + b$

Partie 1

Principe

Détermination de a et b sur la population

On écrit $\forall k \in \mathcal{U}$, $y_k = ax_k + b + E_k$ et on impose que le résidus E_k soient "petits" et vérifient $\sum_{k \in \mathcal{U}} E_k = 0$.

Il existe une infinité de solutions (par exemple $a = \frac{\bar{Y}}{\bar{X}}$ et $b = 0$).

On cherche une solution "optimale" par la méthode des moindres carrés.

Détermination de a et b sur la population

Méthode des moindres carrés :

$$\min_{a,b} \sum_{k \in \mathcal{U}} (y_k - a x_k - b)^2$$

Ce qui donne :

$$\tilde{a} = \frac{\sum_{k \in \mathcal{U}} (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k \in \mathcal{U}} (x_k - \bar{X})^2} = \frac{S_{xy}}{S_x^2}$$

$$\tilde{b} = \bar{Y} - \tilde{a}\bar{X}$$

Principe de l'estimation par la régression

On a donc : $\forall k \in \mathcal{U}\ , y_k = \tilde{a}x_k + \tilde{b} + E_k$.

Donc, dans $\mathcal{U}$: $\bar{Y} = \tilde{a}\bar{X} + \tilde{b} + \bar{E} = \tilde{a}\bar{X} + \tilde{b}$

Puis, dans s : $\bar{y} = \tilde{a}\bar{x} + \tilde{b} + \bar{e}$, avec $\bar{e} = \frac{1}{n}\sum_{k \in s} E_k$

D'où : $\bar{Y} - \bar{y} = \tilde{a}(\bar{X} - \bar{x}) - \bar{e}$

Principe de l'estimation par la régression

On fait **l'hypothèse :** $\bar{e} \approx 0$, ce qui donne :

$$\bar{Y} \approx \bar{y} + \tilde{a}(\bar{X} - \bar{x})$$

Principe de l'estimation par la régression

Généralement, on ne connaît pas $\tilde{a}$. On peut l'estimer grâce à :

$$\hat{a} = \frac{\sum_{k \in s} (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k \in s} (x_k - \bar{X})^2} = \frac{s_{xy}}{s_x^2}$$

Partie 2

Définition

Définition

Définition (Estimateur par la régression d'une moyenne)

L'estimateur par la régression d'une moyenne est défini par :

$$\hat{\bar{Y}}_{reg} = \bar{y} + \hat{a}(\bar{X} - \bar{x})$$

Définition

Interprétation : $\hat{\bar{Y}}_{reg} = \bar{y} +$ terme correctif fonction de :

- $\bar{X} - \bar{x}$
- Valeur de $\hat{a}$
- Signe de $\hat{a}$

Définition

Définition (Estimateur par la régression d'un total)

L'estimateur par la régression d'un total est défini par :

$$\hat{T}_{Y,reg} = N\left[\bar{y} + \hat{a}(\bar{X} - \bar{x})\right]$$

Remarque : Cela nécessite la connaissance de N, la taille de la population.

Partie 3

Propriétés

Estimateur pondéré

Propriété

*L'estimateur par la régression est un estimateur linéaire homogène (**pondéré**). Pour toute variable* $Z = (z_k)_{k\in[[1,n]]}$*, les poids :*

$$w_k = N\left[\frac{1}{n} + (\bar{X} - \bar{x})\frac{(x_k - \bar{x})}{\sum_{k\in s}(x_k - \bar{x})^2}\right]$$

permettent de construire l'estimateur par la régression sur la variable X *du total* $T(Z)$ *:* $\hat{T}_{Z,reg} = \sum_{k\in s} w_k z_k$

Estimateur pondéré

Démonstration.

On remarque que : $s_{xy} = \frac{1}{n-1}\sum_{k\in s}(x_k - \hat{\bar{x}})y_k$, qui s'écrit sous une forme linéaire homogène par rapport à y_k. On a donc :

$\hat{T}_{Y,reg} = N\left[\bar{y} + \hat{a}(\bar{X} - \bar{x})\right] = \hat{T}_{Y,reg} = N\left[\bar{y} + \frac{s_{xy}}{s_x^2}(\bar{X} - \bar{x})\right]$, qui est linéaire homogène comme somme de constantes et d'estimateurs linéaires homogènes. □

Propriété de calage

Propriété (de calage)

L'estimateur par la régression, sur la variable X estime parfaitement N la taille de la population, ainsi que le total de X.

Propriété de calage

Démonstration.

On prend $Z = 1$:

$$\sum_{k \in s} w_k = N$$

La taille de la population est donc parfaitement estimée.
On prend ensuite $Z = X$:

$$\sum_{k \in s} w_k x_k = N\bar{x} + N(\bar{X} - \bar{x}) = X$$

On est donc calé sur les totaux N et X. □

Biais

Propriété (Estimateur asymptotiquement sans biais)

L'estimateur par la régression est asymptotiquement sans biais :

$$B(\hat{T}_{Y,reg}) \underset{n\to+\infty}{\sim} \frac{C}{n}, C \in \mathbb{R}$$

En particulier :

$$B(\hat{T}_{Y,reg}) \underset{n\to+\infty}{\to} 0$$

Précision

Propriété (Variance de l'estimateur par la régression)

$$\mathrm{Var}(\hat{\bar{Y}}_{\mathrm{reg}}) \approx \frac{1-f}{n} S_E^2$$

$$où : S_E^2 = \frac{1}{N-1} \sum_{k \in \mathcal{U}} E_k = S_y^2 (1 - r^2)$$

$$et : r^2 = \frac{S_{XY}}{S_X S_Y}, \textit{ coefficient de corrélation linéaire}$$

entre X et Y dans $\mathcal{U}$

Précision

Démonstration.

On néglige le biais asymptotique, ce qui donne : $\mathrm{Var}(\hat{\bar{Y}}_{\mathrm{reg}}) \approx \mathrm{EQM}(\hat{\bar{Y}}_{\mathrm{reg}})$:

$$\begin{aligned}
\mathrm{Var}(\hat{\bar{Y}}_{\mathrm{reg}}) &\approx \mathbb{E}\left[\hat{T}_{Y\pi} + \hat{a}(T(X) - \hat{T}_{X\pi}) - T(Y)\right]^2 \\
&\approx \mathbb{E}\left[\hat{T}_{Y\pi} + \tilde{a}(T(X) - \hat{T}_{X\pi}) - T(Y)\right]^2 \\
&\approx \frac{N(N-n)}{n}(S_y^2 - 2\tilde{a}S_{xy} + \tilde{a}^2 S_x^2) \\
&= N^2 \frac{1-f}{n} S_E^2
\end{aligned}$$

Le passage de la première à la seconde ligne s'effectue par développement de Taylor (on linéarise la variance, et on montre que les termes liés à $\hat{a} - \tilde{a}$ sont d'ordre deux) : voir Tillé, *Théorie des sondages* (Dunod, 2001), chapitre 12. □

Précision

Théorème (Gain en précision)

$$\frac{\mathrm{Var}(\hat{\bar{\mathrm{Y}}}_{\mathrm{reg}})}{\mathrm{Var}(\bar{\mathrm{y}})} = 1 - r^2$$

Précision

Démonstration.

$$\begin{aligned}\mathrm{Var}(\hat{\bar{Y}}_{\mathrm{reg}}) &\approx \frac{N(N-n)}{n}(S_y^2 - 2\tilde{a}S_{xy} + \tilde{a}^2 S_x^2) \\ &\approx \frac{N(N-n)}{n} S_y^2(1-r^2) \text{ avec : } r = \frac{S_{xy}}{S_x S_y}\end{aligned}$$

□

Précision

La précision est d'autant plus grande que la corrélation linéaire entre X et Y est forte, ou que les résidus sont faibles.

Estimation de variance

La variance peut être estimée par :

$$\hat{\mathrm{Var}}(\hat{\bar{Y}}_{reg}) = \frac{1-f}{n} S_e^2$$

$$\text{où : } S_e^2 = \frac{1}{n-1} \sum_{k \in s} e_k^2$$

Les $e_k = y_k - \hat{a}x_k - \hat{b}$ désignent les résidus de la régression **dans l'échantillon**

Estimation de variance

On a encore :

$$\frac{\hat{\mathrm{Var}}(\hat{\bar{Y}}_{reg})}{\hat{\mathrm{Var}}(\bar{y})} = 1 - r_s^2$$

$$\text{où : } r_s = \frac{s_{XY}}{s_X s_Y}$$

Comparaison avec l'estimation par le ratio

On a :

$$\mathrm{Var}(\hat{\bar{\mathrm{Y}}}_{\mathrm{ratio}}) \geq \mathrm{Var}(\hat{\bar{\mathrm{Y}}}_{\mathrm{reg}})$$

Avec égalité quand : $\frac{\bar{Y}}{\bar{X}} = \frac{S_{XY}}{S_X^2} = \tilde{a}$ (c'est-à-dire $\tilde{b} = 0$). $\hat{\bar{Y}}_{reg}$ est préférable à $\hat{\bar{Y}}_{ratio}$ dès que la droite de régression de Y sur X dans $\mathcal{U}$ ne passe pas par l'origine.

Partie 4

Estimation par la régression généralisée

Cadre

- On suppose qu'il existe une relation linéaire approchée entre Y et les X_j : $Y \approx \sum_{j=1}^{J} b_j X_j$
- Plus de restriction sur le plan de sondage

Principe

Le principe est le même que pour la régression simple. On commence par calculer le coefficient $\tilde{b} = (\tilde{b}_1, \ldots, \tilde{b}_j, \ldots, \tilde{b}_J)'$ qui minimise la quantité : $\sum_{k \in \mathcal{U}} \left(y_k - \sum_{j=1}^{J} b_j x_{jk} \right)^2 = \sum_{k \in \mathcal{U}} (y_k - b' x_k)^2$

Principe

$$\tilde{b} = (\sum_{k \in \mathcal{U}} x_k x_k')^{-1} (\sum_{k \in \mathcal{U}} x_k y_k) = A^{-1}\theta$$

A est une matrice de terme général $a_{jj'} = \sum_{k \in \mathcal{U}} x_{jk} x_{j'k}$ inconnu,

estimé sans biais par : $\hat{A} = \sum_{k \in s} \frac{x_k x_k'}{\pi_k}$, de terme général

$\hat{a}_{jj'} = \sum_{k \in s} \frac{x_{jk} x_{jk}'}{\pi_k}$

Principe

θ, vecteur de composantes $\theta_j = \sum_{k \in \mathcal{U}} x_{jk} y_k$ inconnues, peut être estimé par $\hat{\theta} = \sum_{k \in s} \frac{x_k y_k}{\pi_k}$, de composantes $\hat{\theta}_j = \sum_{k \in s} \frac{x_{jk} y_k}{\pi_k}$. $\tilde{b}$ peut être alors estimé par :

$$\hat{b} = \hat{A}^{-1} \hat{\theta} = \left(\sum_{k \in s} \frac{x_k x_k'}{\pi_k} \right)^{-1} \left(\sum_{k \in s} \frac{x_k y_k}{\pi_k} \right)$$

Remarque : $\hat{b}$ est biaisé, mais on peut montrer que son biais est asymptotiquement nul.

Définition

Définition (Estimateur par la régression généralisée d'un total)

L'estimateur par la régression d'un total est défini par :

$$\hat{T}_{Y,greg} = \hat{T}_{Y,\pi} + \sum_{j=1}^{J} \hat{b}_j (T_{X_j} - \hat{T}_{X_{j\pi}}) = \hat{T}_{Y,\pi} + \hat{b}'(T_X - \hat{T}_{X_\pi})$$

Estimateur "pondéré"

Propriété

L'estimateur par la régression d'un total est linéaire homogène ("pondéré"). Pour toute variable $Z = (z_k)_{k\in[[1,n]]}$*, les poids :*

$$w_k = \frac{1 + (T_X - \hat{T}_{X_\pi})' \hat{A}^{-1} x_k}{\pi_k}$$

permettent de construire l'estimateur par la régression généralisée sur les variables auxiliaires X *:* $\hat{T}_{Z,greg} = \sum_{k\in s} w_k z_k$

Propriété de calage

Propriété (de calage)

L'estimateur par la régression généralisée sur les variables auxiliaires X estime parfaitement les X_j.

Propriété de calage

Démonstration.

Le coefficient de régression pour la variable auxiliaire x_j vaut :

$$\hat{a} = (\sum_{k \in s} \frac{c_k x_k x_k'}{\pi_k})^{-1} \sum_{k \in s} \frac{c_k x_k x_{kj}}{\pi_k} = (0 \dots 010 \dots 0)'$$

L'estimateur GREG vaut alors :

$$\hat{T}_{x_j,greg} = \hat{T}_{x_j,\pi} + (T(X) - \hat{T}_{X\pi})'\hat{a} = \hat{T}_{x_j,\pi} + (T(x_j) - \hat{T}_{x_j,\pi}) = T(x_j)$$

Et l'estimateur est donc calé sur le total x_j □

Biais

Propriété (Estimateur asymptotiquement sans biais)

L'estimateur par la régression généralisée est asymptotiquement sans biais :

$$B(\hat{T}_{Y,greg}) \underset{n\to+\infty}{\sim} \frac{C}{n}, C \in \mathbb{R}$$

En particulier :

$$B(\hat{T}_{Y,greg}) \underset{n\to+\infty}{\to} 0$$

Précision

Propriété (Variance de l'estimateur par la régression)

$$\mathrm{Var}(\hat{\bar{\mathrm{Y}}}_{\mathrm{greg}}) \approx \mathrm{Var}(\hat{\mathrm{E}}_{\pi}) = \mathrm{Var}(\sum_{\mathrm{k}\in\mathrm{s}} \frac{\mathrm{E_k}}{\pi_{\mathrm{k}}})$$

$$\hat{\mathrm{Var}}(\hat{\bar{Y}}_{greg}) = \hat{\mathrm{Var}}(\hat{e}_{\pi}) = \hat{\mathrm{Var}}(\sum_{k\in s} \frac{e_k}{\pi_k})$$

$$avec : E_k = y_k - \tilde{b}' x_k$$

$$et : e_k = y_k - \hat{b}' x_k$$

Chapitre 3

La post-stratification

Exemple introductif

On cherche à estimer le nombre d'habitants d'une région comportant 2536 villages. On tire un échantillon de 127 villages par sondage aléatoire simple. On observe que les villages de l'échantillon ont une taille moyenne de 377,2 habitants.

La région comporte deux zones, nord et sud.

Exemple introductif

Zone	Nombre de villages	Nombre de villages dans l'échantillon	Taille moyenne des villages de l'échantillon
Nord	1421	65	402,8
Sud	1115	62	350,4
Ensemble	2536	127	377,2

(extrait de Manuel de sondages, Applications aux pays en développement, R. Clairin et Ph. Brion, Documents et Manuels du CEPED numéro 3)

Partie 1

Principe

Principe

On définit après l'enquête des groupes d'individus, appelés post-strates, et on suppose connue la répartition de la population selon ces post-strates.

Cette répartition n'a aucune raison de coïncider exactement avec la répartition dans l'échantillon. L'estimation va utiliser les proportions connues des groupes dans la population.

Différence fondamentale avec la stratification

La répartition de l'échantillon par post-strate n'est pas contrôlée : les effectifs des post-strates dans l'échantillon ne sont connus qu'après enquête (ce sont des variables aléatoires, dépendant de l'échantillon tiré).
On se limite dans la suite du chapitre au cas du sondage aléatoire simple sans remise (donc $\forall k \in \mathcal{U}\ \pi_k = \frac{n}{N}$).

Partie 2

Notations

La post-stratification

La population $\mathcal{U}$ est partitionnée en H post-strates disjointes $(\mathcal{U}_h)_{h\in[[1,H]]}$. Pour chaque post-strate on note :

- $N_h =$ effectif de la post-strate
- $T_{Yh} = \sum_{k\in\mathcal{U}_h} y_k =$ total de Y
- $\bar{Y}_h = \frac{1}{N_h}\sum_{k\in\mathcal{U}_h} y_k =$ moyenne de Y
- $S_h^2 = \frac{1}{N_h-1}\sum_{k\in\mathcal{U}_h}(y_k-\bar{Y}_h)^2 =$ dispersion/variance empirique de Y

La post-stratification

$s_{(h)} = s \cap \mathcal{U}_h$ désigne la partie de l'échantillon s incluse dans la post-strate h. On note :

- $n_h =$ effectif $s_{(h)}$
- $\bar{y}_h = \frac{1}{n_h} \sum_{k \in s_{(h)}} y_k =$ moyenne de Y dans $s_{(h)}$
- $s_h^2 = \frac{1}{n_h - 1} \sum_{k \in s_{(h)}} (y_k - \bar{Y}_h)^2 =$ dispersion/variance empirique de Y dans $s_{(h)}$

Partie 3

Définition

Définition

Définition (Estimateur post-stratifié d'une moyenne)

L'estimateur post-stratifié d'une moyenne est défini par :

$$\hat{\bar{Y}}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} y_h$$

Définition

Définition (Estimateur post-stratifié d'un total)

L'estimateur post-stratifié d'un total est défini par :

$$\hat{T}_{Y,post} = \sum_{h=1}^{H} N_h y_h$$

Définition

Remarque : Dans le cas général (pas seulement SAS), l'estimateur post-stratifié du total s'écrit :

$$\hat{T}_{Y,post} = \sum_{h=1}^{H} N_h \frac{\hat{T}_{Y\pi h}}{\hat{N}_{\pi h}}$$

Partie 4

Propriétés

Estimateur "pondéré"

Propriété

L'estimateur post-stratifié est linéaire homogène ("pondéré"). Pour toute variable $Z = (z_k)_{k \in [[1,n]]}$*, les poids :*

$$w_k = \frac{N_h}{n_h}$$

où h désigne la post-strate à laquelle appartient l'unité k

permettent de construire l'estimateur post-stratifié du total Z :

$$\hat{T}_{Z,post} = \sum_{k \in s} w_k z_k$$

Estimateur "pondéré"

Remarque : L'estimateur post-stratifié pour le total de Z ne sera efficace que si le choix de strates convient pour Z.

Propriété de calage

Propriété

L'estimateur $\hat{T}_{Y,post}$ est calé sur les tailles des post-strates N_h.

Propriété de calage

Démonstration.

On définit $\mathbb{1}_h(k)$ l'indicatrice d'appartenance à la post-strate h. Alors l'estimateur post-stratifié du total de $\mathbb{1}_h(k)$ vaut :

$$\begin{aligned}\hat{N}_{h,post} &= \sum_{k \in s} w_k \mathbb{1}_h(k) \\ &= \sum_{k \in s_{(h)}} \frac{N_h}{n_h} \\ &= \frac{N_h}{n_h} n_h \\ &= N_h\end{aligned}$$

□

Biais

Propriété (Estimateur sans biais)

L'estimateur post-stratifié $\hat{T}_{Y,post}$ *estime sans biais* $T(Y)$ *:*

$$\mathbb{E}(\hat{T}_{Y,post}) = T(Y)$$

à condition d'utiliser les valeurs exactes des N_h. *Sinon, l'estimateur est biaisé et le biais ne décroît pas avec la taille de l'échantillon.*

Biais

Démonstration.

Voir Tillé, *Théorie des Sondages* (Dunod, 2001), p.191 □

Précision

Propriété

$$\mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{Y,post}}) \approx \mathrm{N}^2 \left[\frac{1-\mathrm{f}}{\mathrm{n}} \sum_{\mathrm{h}=1}^{\mathrm{H}} \frac{\mathrm{N_h}}{\mathrm{N}} \mathrm{S_h^2} + \frac{1-\mathrm{f}}{\mathrm{n}^2} \sum_{\mathrm{h}=1}^{\mathrm{H}} \frac{\mathrm{N}-\mathrm{N_h}}{\mathrm{N}} \mathrm{S_h^2} \right]$$

Précision

Démonstration.

Voir Tillé, *Théorie des Sondages* (Dunod, 2001), p.193 □

Précision

La variance se décompose en deux termes :

- le premier représente la variance obtenue dans le cas d'un sondage stratifié (avec allocation proportionnelle)
- le second est la perte de précision due au fait de ne prendre en considération la stratification qu'a posteriori. Pour un gros échantillon, le second terme (terme correctif en $\frac{1}{n^2}$) devient négligeable devant le premier.

Variance estimée

$$\hat{\mathrm{Var}}(\hat{T}_{Y,post}) \approx N^2 \left[\frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} s_h^2 + \frac{1-f}{n^2} \sum_{h=1}^{H} \frac{N-N_h}{N} s_h^2 \right]$$

Comparaison avec l'estimateur d'Horvitz-Thompson

Horvitz-Thompson

$$\hat{\bar{Y}}_\pi = \bar{y} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h$$

Post-stratifié

$$\hat{\bar{Y}}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

En général $\frac{n_h}{n} \neq \frac{N_h}{N}$, les estimateurs ne coïncident donc pas.

Comparaison avec l'estimateur d'Horvitz-Thompson

On a :

$$\mathrm{Var}(\hat{\bar{\mathrm{Y}}}_{\pi}) = N^2 \frac{1-f}{n} S^2 = N^2 \frac{1-f}{n} \frac{1}{N-1} \sum_{k \in \mathcal{U}} (y_k - \bar{Y})^2$$

$$\mathrm{Var}(\hat{\bar{\mathrm{Y}}}_{\mathrm{post}}) \approx N^2 \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} S_h^2$$

Équation de décomposition de la variance :

$$(N-1)S^2 = \sum_{h=1}^{H} (N_h - 1) S_h^2 + \sum_{h=1}^{H} N_h (\bar{Y}_h - \bar{Y})^2$$

Comparaison avec l'estimateur d'Horvitz-Thompson

D'où :

$$\frac{\mathrm{Var}(\hat{\bar{\mathrm{Y}}}_{\mathrm{post}})}{\mathrm{Var}(\hat{\bar{\mathrm{Y}}}_{\pi})} \approx 1 - \frac{\sum_{h=1}^{H} N_h(\bar{Y}_h - \bar{Y})^2}{(N-1)S^2}$$
$$= 1 - R^2$$

où R^2 est le coefficient de détermination du modèle d'analyse de la variance "expliquant" la variable Y par la variable de stratification.

Comparaison avec l'estimateur d'Horvitz-Thompson

La diminution de variance est d'autant plus importante que les $\bar{Y}_h$ sont peu dispersées, ou, corrélativement, que la variable Y est peu dispersée dans chaque post-strate. Finalement :

Théorème

Post-stratifier est efficace si la variable de post-stratification explique bien la variable Y.

Chapitre 4

Post-stratification sur plusieurs critères et "raking-ratio"

Partie 1

Principe

Principe

Il s'agit d'une généralisation de la post-stratification au cas où il y a plus d'une variable auxiliaire qualitative. Deux cas de figure peuvent se présenter :

- soit on connaît la structure de la population sur chacun des croisements des critères de post-stratification, et dans ce cas on se ramène à une simple post-stratification
- soit on dispose des effectifs par catégorie uniquement sur les variables de post-stratification prises isolément, et dans ce cas il faut utiliser un algorithme spécifique pour obtenir un redressement simultané sur chaque modalité de chaque variable

Exemple - calage sur les marges de variables catégorielles

- $X =$ catégorie socioprofessionnelle
- $Y =$ âge

On note les effectifs estimés sur l'échantillon et les effectifs connus sur la population dans un tableau.

$$\hat{N}_{ij} = \sum_{k \in s, X=i, Y=j} \frac{N}{n} = \frac{N}{n} n_{ij}$$

$$\hat{N}_{i+} = \frac{N}{n} \sum_{j} n_{ij}$$

$$\hat{N}_{+j} = \frac{N}{n} \sum_{i} n_{ij}$$

Exemple

	15-24 ans	...	35-44 ans	...	Plus de 75 ans	Marges
Agriculteurs						$\hat{N}_{1+}/N_{1+}$
...						
Cadres supérieurs			$\hat{N}_{ij}/N_{ij}$			$\hat{N}_{1+}/N_{1+}$
...						
Indépendants						$\hat{N}_{I+}/N_{I+}$
Marges	$\hat{N}_{+1}/N_{+1}$		$\hat{N}_{+j}/N_{+j}$		$\hat{N}_{+J}/N_{+J}$	$\hat{N}/N$

Principe

On cale l'échantillon sur les distributions marginales des variables dans la population ; on utilise comme information auxiliaire les valeurs $N_{i+}, \ldots, N_{+j}$, c'est-à-dire les marges du tableau de contingence croisant les deux variables. D'où le nom de **calage sur marges**.
On ne cale pas sur les effectifs N_{ij} correspondant aux croisements des modalités, qui peuvent ne pas être connus.

Objectif

Objectif : transformer les poids initiaux $d_k = \frac{N}{n}$ en poids w_k^{RR} permettant d'estimer parfaitement les effectifs marginaux connus sur la population, c'est-à-dire tels que :

$$\forall i, \hat{N}_{i+}^{RR} = \sum_{k \in s, X=i} w_k^{RR} = N_{i+}$$

$$\forall i, \hat{N}_{+j}^{RR} = \sum_{k \in s, Y=j} w_k^{RR} = N_{+j}$$

Fonctionnement

Principe du raking-ratio : enchaînement de redressements par règle de trois, de façon à caler la structure de l'échantillon une fois sur une marge, une fois sur l'autre. C'est un algorithme itératif procédant par succession d'étapes.

Algorithme du raking ratio

Situation initiale :

$$\hat{N}_{i+}^{(0)} = \frac{N}{n} n_{i+} \neq N_{i+}$$
$$\hat{N}_{+j}^{(0)} = \frac{N}{n} n_{+j} \neq N_{+j}$$

Première étape : Calage sur les marges de la variable X

Pour tout i, on multiplie les poids initiaux par $\frac{N_{i+}}{\hat{N}_{i+}^{(0)}}$. Cela définit des nouveaux poids : $\forall k / X = i, d_k^{(1)} = \frac{N_{i+}}{n_{i+}}$, qui, par construction, sont calés sur les totaux N_{i+}.

Algorithme du raking ratio

Deuxième étape : Calage sur les marges de la variable Y

On a alors :

$$\hat{N}_{+j}^{(1)} = \sum_{k \in s, Y_k = j} d_k^{(1)} = \sum_i \frac{N_{i+}}{n_{i+}} \sum_{k \in s, X_k = i, Y_k = j} 1 = \sum_i \frac{N_{i+}}{n_{i+}} n_{ij} \neq N_{+j}$$

Pour tout j, on multiplie les poids initiaux par $\frac{N_{+j}}{\hat{N}_{+j}^{(1)}}$. Cela définit des nouveaux poids : $\forall k / X = i, d_k^{(2)} = d_k^{(1)} \frac{N_{+j}}{\hat{N}_{+j}^{(1)}}$, qui, par construction, sont calés sur les totaux N_{+j}.

Par contre, le calage sur X n'est plus assuré :

$$\hat{N}_{i+}^{(2)} = \sum_{k \in s, X_k = i} d_k^{(2)} \neq N_{i+}$$

Algorithme du raking ratio

Troisième étape : recommencer
On recommence les calages successifs sur les marges X et Y, jusqu'à ce que la différence des poids entre deux itérations soit inférieure à un certain seuil.

En pratique, dès lors que les marges sont cohérentes, l'algorithme converge assez rapidement vers une situation raisonnablement stable.

Partie 2

Propriétés

Propriétés

- Estimateur linéaire homogène : par essence
- Totaux calés : par essence
- Estimateur asymptotiquement sans biais (c'est un cas particulier du calage sur marges)
- Gain de précision pour les variables liées aux marges (voir aussi calage sur marges)