Techniques avancées d'échantillonnage
Estimation de variance

Guillaume Chauvet

École Nationale de la Statistique et de l'Analyse de l'Information

28 janvier 2014

1. Estimation de la variance due à l'échantillonnage
2. Estimation d'un paramètre complexe
3. Estimation de variance : prise en compte du calage
4. Application : Enquête Logement 2006
5. Méthodes de rééchantillonnage

Estimation de la variance due à l'échantillonnage

Cas d'un total : estimateur de Horvitz-Thompson

Estimateur de Horvitz-Thompson

On se place dans le cadre d'une population finie $U = \{1, \ldots, k, \ldots, N\}$. On utilise un plan de sondage $p(\cdot)$ respectant des probabilités d'inclusion $\pi_k > 0$ choisies. On peut estimer sans biais le total t_y par

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Pour un plan de sondage quelconque :

$$V_p\left[\hat{t}_{y\pi}\right] = \sum_{k,l \in U} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \Delta_{kl} \tag{1}$$

avec $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. Cette variance peut être estimée sans biais par

$$v_{HT}\left[\hat{t}_{y\pi}\right] = \sum_{k,l \in S} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}} \tag{2}$$

si tous les π_{kl} sont strictement positifs.

Plan de Poisson

Chaque unité k est tirée avec une probabilité π_k, indépendamment des autres unités. La taille de l'échantillon est aléatoire.

On a $\delta_{kl} = 0$ pour tous $k \neq l \in U$. La variance s'obtient à partir de la formule de Horvitz-Thompson :

$$V_p\left[\hat{t}_{y\pi}\right] = \sum_{k\in U}\left(\frac{y_k}{\pi_k}\right)^2 \pi_k(1-\pi_k). \quad (3)$$

On l'estime sans biais par

$$v_{HT}\left[\hat{t}_{y\pi}\right] = \sum_{k\in S}\left(\frac{y_k}{\pi_k}\right)^2 (1-\pi_k). \quad (4)$$

Variance pour un plan de taille fixe

Pour un plan de sondage de taille fixe, la variance du π-estimateur peut s'écrire sous la forme

$$V_p\left[\hat{t}_{y\pi}\right] = -\frac{1}{2}\sum_{k\neq l\in U}\left[\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right]^2 \Delta_{kl}. \tag{5}$$

On l'estime sans biais par

$$v_{YG}\left[\hat{t}_{y\pi}\right] = -\frac{1}{2}\sum_{k\neq l\in S}\left[\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right]^2 \frac{\Delta_{kl}}{\pi_{kl}} \tag{6}$$

si tous les π_{kl} sont strictement positifs.

Si le plan de sondage vérifie les **conditions de Yates-Grundy** :
$\forall k \neq l \in U \quad \Delta_{kl} \leq 0$, cet estimateur de variance est toujours positif.

Sondage aléatoire simple

Il s'agit du plan qui donne la même probabilité à tous les échantillons de taille n d'être sélectionnés. L'estimateur de Horvitz-Thompson du total est donné par

$$\hat{t}_{y\pi} = N\ \bar{y} \quad \text{avec} \quad \bar{y} = \frac{1}{n}\sum_{k\in S} y_k.$$

Sa variance s'obtient à partir de la formule de Sen-Yates-Grundy :

$$V_p[\hat{t}_{y\pi}] = N^2 \frac{1-f}{n} S_y^2 \quad \text{avec} \quad S_y^2 = \frac{1}{N-1}\sum_{k\in U}(y_k - \mu_y)^2. \qquad (7)$$

On l'estime sans biais par

$$v_{YG}(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} s_y^2 \quad \text{avec} \quad s_y^2 = \frac{1}{n-1}\sum_{k\in S}(y_k - \bar{y})^2. \qquad (8)$$

Plans à un degré : sondage stratifié, tirage réjectif

Sondage aléatoire simple stratifié

Un plan de sondage stratifié consiste à découper la population en groupes (si possible) homogènes en intra par rapport à la variable d'intérêt, et à tirer des échantillons indépendants dans chaque strate.

Dans le cas d'un SRS stratifié, l'estimateur de Horvitz-Thompson du total est donné par

$$\hat{t}_{y\pi} = \sum_{h=1}^{H} N_h \, \bar{y}_h,$$

avec

$$\bar{y}_h = \frac{1}{n_h} \sum_{k \in S_h} y_k$$

l'estimateur de la moyenne μ_{yh} dans la strate h.

Sondage aléatoire simple stratifié (2)

Il s'agit d'un plan de taille fixe. Sa variance s'obtient donc à partir de la formule de Sen-Yates-Grundy :

$$V_p\left[\hat{t}_{y\pi}\right] = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{yh}^2. \tag{9}$$

On l'estime sans biais par

$$v_{YG}\left[\hat{t}_{y\pi}\right] = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} s_{yh}^2. \tag{10}$$

La taille globale d'échantillon doit être allouée de façon à tirer des échantillons plus gros dans les strates présentant une plus forte dispersion (allocation de Neyman).

Tirage réjectif

Le plan de sondage réjectif est obtenu :

- en tirant un échantillon selon un plan de Poisson de probabilités d'inclusion $p_k,\ k \in U$, avec $\sum_{k \in U} p_k = n$;
- en rejetant l'échantillon tant qu'il n'est pas de la taille voulue n.

Les probabilités p_k sont choisies pour que les probabilités d'inclusion effectives soient égales à π_k (fixées à l'avance).

Le tirage est de taille fixe par construction. On peut toujours estimer sans biais la variance par

$$v_{YG}\left[\hat{t}_{y\pi}\right] = \frac{1}{2} \sum_{k \neq l \in S} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}}.$$

Tirage réjectif (2)

En utilisant l'approximation des probabilités d'inclusion d'ordre 2

$$\pi_k\pi_l - \pi_{kl} = \frac{\pi_k(1-\pi_k)\pi_l(1-\pi_l)}{\sum_{j\in U}\pi_j(1-\pi_j)}[1+o(1)] \quad \text{pour} \quad k \neq l \in U,$$

proposée par Hajek, on obtient l'estimateur de variance (voir aussi Deville, 1993)

$$v_{dev}\left[\hat{t}_{y\pi}\right] = \frac{n}{n-1}\sum_{k\in S}(1-\pi_k)\left(\frac{y_k}{\pi_k} - \widehat{R}\right)^2 \text{ où } \widehat{R} = \frac{\sum_{k\in S}\frac{y_k}{\pi_k}(1-\pi_k)}{\sum_{k\in S}(1-\pi_k)}. \quad (11)$$

Cet estimateur est couramment utilisé dans les enquêtes Insee (Caron, 1998). Dans le cas de probabilités d'inclusion égales, on retrouve l'estimateur de variance pour le sondage aléatoire simple.

Plans à un degré : échantillonnage équilibré

Echantillonnage équilibré

On suppose qu'un vecteur $\mathbf{x}_k$ est disponible pour tout $k \in U$. Un échantillon s est dit *équilibré* sur les totaux $t_{\mathbf{x}}$ si

$$\hat{t}_{\mathbf{x}\pi}(s) = t_{\mathbf{x}}.$$

Un plan de sondage est dit équilibré sur les totaux $t_{\mathbf{x}}$ si seuls les échantillons équilibrés sur $\mathbf{x}$ ont une probabilité non nulle d'être sélectionnés.

La méthode du Cube (Deville et Tillé, 2004) permet de tirer des échantillons (approximativement équilibrés) :

- la phase de vol permet de sélectionner/écarter des individus en respectant exactement les contraintes d'équilibrage,
- la phase d'atterrissage permet de terminer l'échantillonnage en relâchant les contraintes d'équilibrage,

Cas particulier : la méthode du pivot

La méthode a été proposée par Deville et Tillé (1998). On l'obtient en appliquant la méthode du Cube avec la seule contrainte de taille fixe.

Basée sur des duels entre unités. A l'étape 1, les unités 1 et 2 s'affrontent :

- si $\pi_1 + \pi_2 \leq 1$, une des unités est éliminée et l'autre survit :

$$(\pi_1, \pi_2) = \begin{cases} (\pi_1 + \pi_2, 0) & \text{avec proba } \frac{\pi_1}{\pi_1+\pi_2}, \\ (0, \pi_1 + \pi_2) & \text{avec proba } \frac{\pi_2}{\pi_1+\pi_2}. \end{cases}$$

- si $\pi_1 + \pi_2 > 1$, une des unités est sélectionnée et l'autre survit :

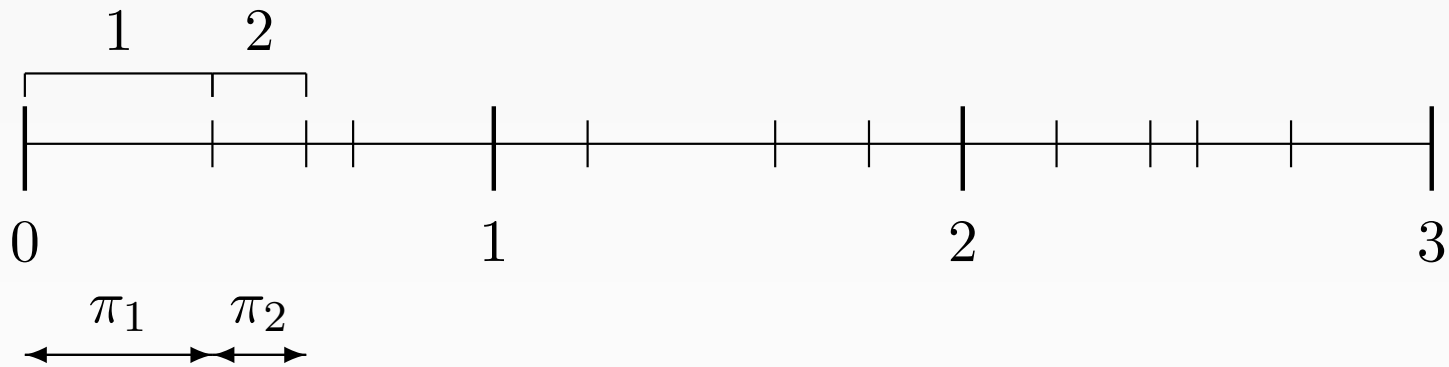
$$(\pi_1, \pi_2) = \begin{cases} (1, \pi_1 + \pi_2 - 1) & \text{avec proba } \frac{1-\pi_2}{2-\pi_1-\pi_2}, \\ (\pi_1 + \pi_2 - 1, 1) & \text{avec proba } \frac{1-\pi_1}{2-\pi_1-\pi_2}. \end{cases}$$

A l'étape t, l'unité survivante affronte l'unité suivante $t+1$ selon le même principe. A l'étape $N-1$, un échantillon de taille n a été sélectionné et les probabilités d'inclusion sont exactement respectées.

Exemple

Population U de taille $N = 11$, avec $n = 3$ et

$$\pi = (0.4 \;\; 0.2 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$



On a $(\pi_1, \pi_2) = (0.4, 0.2) = \begin{cases} (0.6, 0) & \text{avec proba } 0.4/0.6, \\ (0, 0.6) & \text{avec proba } 0.2/0.6 \end{cases}$

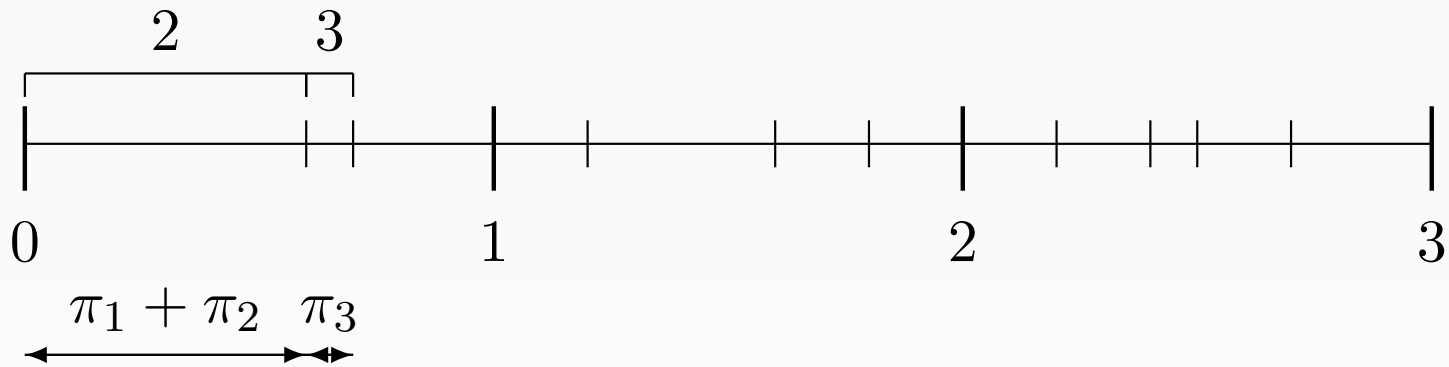
Si l'unité 2 survit, on obtient

$$\pi^{(1)} = (0 \;\; 0.6 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$

Exemple

Population U de taille $N = 11$, avec $n = 3$ et

$$\pi^{(1)} = (0 \quad 0.6 \quad 0.1 \quad 0.5 \quad 0.4 \quad 0.2 \quad 0.4 \quad 0.2 \quad 0.1 \quad 0.2 \quad 0.3)^\top .$$



On a $(\pi_2^{(1)}, \pi_3^{(1)}) = (0.6, 0.1) = \begin{cases} (0.7, 0) & \text{avec proba } 0.6/0.7, \\ (0, 0.7) & \text{avec proba } 0.1/0.7 \end{cases}$.

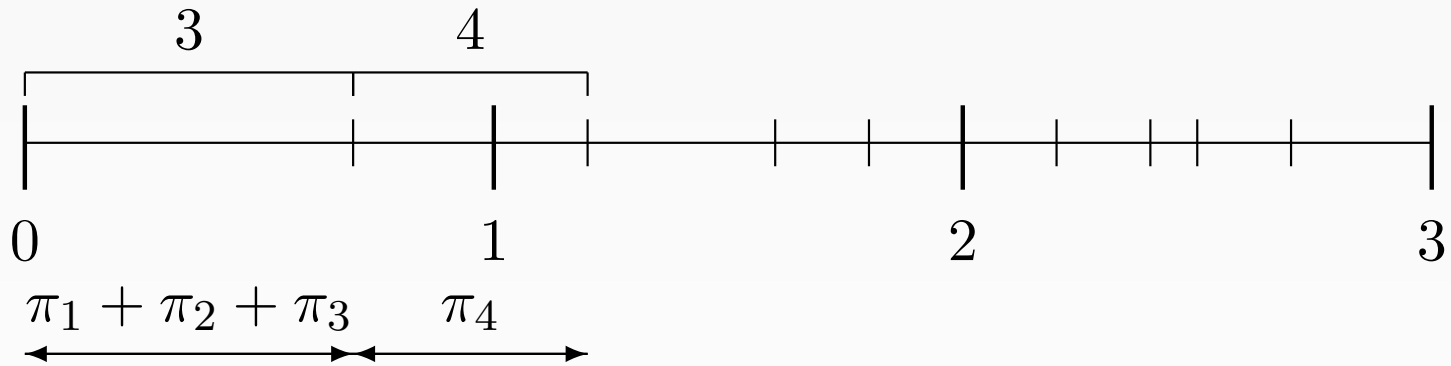
Si l'unité 3 survit, on obtient

$$\pi^{(2)} = (0 \quad 0 \quad 0.7 \quad 0.5 \quad 0.4 \quad 0.2 \quad 0.4 \quad 0.2 \quad 0.1 \quad 0.2 \quad 0.3)^\top .$$

Exemple

Population U de taille $N = 11$, avec $n = 3$ et

$$\pi^{(2)} = (0 \;\; 0 \;\; 0.7 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$



On a $(\pi_3^{(2)}, \pi_4^{(2)}) = (0.7, 0.5) = \begin{cases} (1, 0.2) & \text{avec proba } 0.5/(2 - 1.2), \\ (0.2, 1) & \text{avec proba } 0.3/(2 - 1.2) \end{cases}$.

Si l'unité 3 survit, on obtient

$$\pi^{(3)} = (0 \;\; 0 \;\; 1 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top , \ldots$$

Variance de l'estimateur de Horvitz-Thompson

On suppose que la variable π appartient au vecteur $\mathbf{x}$ de variables d'équilibrage, de sorte que l'échantillonnage est de taille fixe. La variance du π-estimateur est donnée par la formule de Yates-Grundy

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

et peut être estimée sans biais par

$$v_{YG} = -\frac{1}{2}\sum_{k\neq l\in S}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}},$$

si tous les π_{kl} sont >0, avec $\Delta_{kl}=\pi_{kl}-\pi_k\pi_l$.

Problème : les π_{kl} sont difficiles à calculer exactement.

Approximation de variance de Deville et Tillé

Deville et Tillé (2005) ont proposé une classe d'estimateurs de variance, sous les hypothèses suivantes :

1. le plan de sondage est **exactement équilibré**,
2. le plan de sondage est à **entropie maximale**, parmi les plans équilibrés sur les mêmes variables $\mathbf{x}_k$, avec les mêmes probabilités d'inclusion π.

L'entropie d'un plan de sondage $p(\cdot)$ est définie par

$$\mathcal{L}(p) = -\sum_{s \subset U} p(s) \ln p(s) \quad \text{avec} \quad 0 \ln(0) = 0.$$

C'est une mesure de désordre. Un plan à forte entropie autorise la sélection d'un grand nombre d'échantillons, et laisse donc une grande place à l'aléatoire (par opposition à des plans à faible entropie tels que le tirage systématique, par exemple).

Approximation de variance de Deville et Tillé (2)

La condition 2 (entropie maximale) n'est pas nécessairement réalisée, notamment si l'algorithme du Cube est appliqué sur un fichier trié préalablement selon une variable auxiliaire. On obtient alors un **effet de stratification**.

On peut obtenir un plan proche de l'entropie maximale en effectuant un **tri aléatoire** des unités de la population avant d'appliquer la méthode du Cube.

La condition 1 (équilibrage exact) n'est généralement pas vérifiée en raison de la phase d'atterrissage. L'approximation de variance de Deville et Tillé (2005) prend essentiellement en compte la **variance due à la phase de vol**.

Approximation de variance de Deville et Tillé (3)

Sous les deux conditions précédentes, Deville et Tillé (2005) montrent que le plan équilibré peut être vu comme un plan poissonien, conditionnel à $\hat{t}_{\mathbf{x}\pi} = t_{\mathbf{x}}$. Ils en déduisent l'approximation de variance

$$
\begin{aligned}
V_{app}(\hat{t}_{y\pi}) &= \sum_{k\in U} b_k \left(\frac{y_k}{\pi_k} - \frac{\mathbf{x}_k^\top}{\pi_k}\mathbf{B} \right)^2, \\
\mathbf{B} &= \left(\sum_{k\in U} b_k \frac{\mathbf{x}_k}{\pi_k} \frac{\mathbf{x}_k^\top}{\pi_k} \right)^{-1} \sum_{k\in U} b_k \frac{\mathbf{x}_k}{\pi_k} \frac{y_k}{\pi_k}.
\end{aligned}
$$

Plusieurs valeurs des b_k sont proposées. Le choix

$$
b_{1k} = \pi_k(1-\pi_k)\frac{N}{N-p}
$$

redonne la bonne formule de variance dans le cas du SRS.

Approximation de variance de Deville et Tillé (4)

En utilisant le principe de substitution, on obtient l'estimateur de variance

$$\hat{V}_{DT}(\hat{t}_{y\pi}) = \frac{n}{n-p}\sum_{k\in S}(1-\pi_k)\left(\frac{y_k}{\pi_k}-\frac{\mathbf{x}_k^\top}{\pi_k}\hat{\mathbf{B}}\right)^2, \quad (12)$$

$$\hat{\mathbf{B}} = \left(\sum_{k\in S}(1-\pi_k)\frac{\mathbf{x}_k}{\pi_k}\frac{\mathbf{x}_k^\top}{\pi_k}\right)^{-1}\sum_{k\in S}(1-\pi_k)\frac{\mathbf{x}_k}{\pi_k}\frac{y_k}{\pi_k}.$$

Dans le cas $\mathbf{x}_k = x_k = \pi_k$, on obtient

$$\hat{\mathbf{B}} = \widehat{R} = \frac{\sum_{k\in S}\frac{y_k}{\pi_k}(1-\pi_k)}{\sum_{k\in S}(1-\pi_k)}$$

et on retrouve l'estimateur simplifié de variance pour le tirage réjectif.

Approximation de la matrice de variance-covariance

Une autre possibilité consiste à approcher directement les π_{kl} à l'aide de simulations. En sélectionnant C échantillons indépendants, on peut approximer la matrice $\Delta = [\Delta_{kl}]$ de variance-covariance par

$$\Delta_{SIM} = \frac{1}{C} \sum_{c=1}^{C} \left[I(S_c) - \pi \right] \left[I(S_c) - \pi \right]^\top,$$

où $I(S_c) = \left[I(1 \in S_c), \ldots, 1_(N \in S_c) \right]^\top$ est le vecteur des indicatrices d'appartenance à l'échantillon.

D'où l'estimateur de variance

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

avec $\pi_{SIM,kl} = \Delta_{SIM,kl} + \pi_k \, \pi_l$.

Approximation de la matrice de variance-covariance (2)

Notons que la matrice de variance-covariance Δ est donnée par

$$\Delta = V\left[\pi(T)\right],$$

où $\pi(T)$ donne le résultat de l'échantillonnage et peut s'écrire :

$$\pi(T) = \pi + \sum_{t=1}^{T} \delta^{(t)}.$$

Les $\{\delta^{(t)}\}$ sont **non corrélés** (accroissements de la martingale équilibrante), d'où

$$\begin{aligned} V\left[\pi(T)\right] &= \sum_{t=1}^{T} V\left[\delta^{(t)}\right] = E\left\{\sum_{t=1}^{T} V\left[\delta^{(t)}|\mathcal{F}_{t-1}\right]\right\} \\ &= E\left\{\sum_{t=1}^{T} \lambda_1(t)\lambda_2(t)u(t)u(t)^\top\right\}. \end{aligned}$$

Approximation de la matrice de variance-covariance (3)

La matrice de variance-covariance est donc estimée sans biais par

$$\hat{\Delta}(s) = \sum_{t=1}^{T} \lambda_1(t)\lambda_2(t)u(t)u(t)^\top.$$

Une seconde approximation par simulations est obtenue à l'aide de C échantillons indépendants :

$$\Delta_{MD} = \frac{1}{C}\sum_{c=1}^{C} \hat{\Delta}(S_c), \tag{13}$$

d'où l'estimateur de variance

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

avec $\pi_{MD,kl} = \Delta_{MD,kl} + \pi_k\ \pi_l$.

Etude par simulations

Adapté d'une des études par simulations proposées par Deville and Tillé (2005).

On considère une population U de taille $N = 40$. Le plan de sondage considéré est de taille $n = 15$, avec $q = 4$ variables d'équilibrage $x_{k1} = \pi_k$, $x_{k2} = k$, $x_{k3} = 1/k$ et $x_{k4} = 1/k^2$.

Les probabilités d'inclusion sont générées à l'aide d'une loi uniforme, afin que les probabilités soient comprises entre 0.3 and 0.45. Les variables x_2, x_3, x_4 sont centrées et réduites.

Il s'agit d'un plan imparfaitement équilibré.

Etude par simulations

Dans la population U, cinq variables d'intérêt $y_1, \ldots, y_5$ sont générées selon le modèle de régression linéaire

$$y_{ik} = \beta_1 + \beta_2 \; x_{2k} + \beta_3 \; x_{3k} + \beta_4 \; x_{4k} + \sigma_i \; \epsilon_k, \tag{14}$$

pour $i = 1, \ldots, 5$.

On prend $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$, et les ϵ_k sont générés selon une loi normale de moyenne 0 et de variance 1.

Le coefficient σ_i est choisi pour donner un R^2 approximativement égal à 0.1 pour y_1, 0.2 pour y_2, ..., et 0.5 pour y_5.

Mesures de Monte-Carlo

Biais relatif Monte Carlo (en %)

$$RB_{MC}(\hat{\theta}) = \frac{E_{MC}(\hat{\theta}) - \theta}{\hat{\theta}} \times 100.$$

Erreur quadratique moyenne de Monte Carlo :

$$EQM_{MC}(\hat{\theta}) = E_{MC}(\hat{\theta} - \theta)^2$$

Stabilité relative :

$$RS_{MC}(\hat{\theta}) = \frac{\sqrt{EQM_{MC}(\hat{\theta})}}{\theta} \times 100.$$

Les intervalles de confiance sont obtenus en utilisant une distribution t avec $n - q$ degrés de liberté.

Résultats obtenus

		Taux de couverture 5 %				
Var.	Méthode	L	U	$L+U$	% bias	Stab.
y_1	MD	4,35	5,65	10,00	0,2	33,8
	DT	4,90	5,15	10,05	-0,8	29,3
y_2	MD	3,95	6,65	10,60	0,0	44,3
	DT	6,00	5,80	11,80	-11,5	28,6
y_3	MD	3,25	7,65	10,90	-0,2	57,4
	DT	7,70	6,85	14,55	-22,8	32,2
y_4	MD	2,65	8,80	11,45	-0,4	71,7
	DT	9,25	7,95	17,20	-34,2	39,3
y_5	MD	2,15	10,90	13,05	-0,6	86,5
	DT	11,40	9,75	21,15	-45,7	48,4

Plans à deux degrés

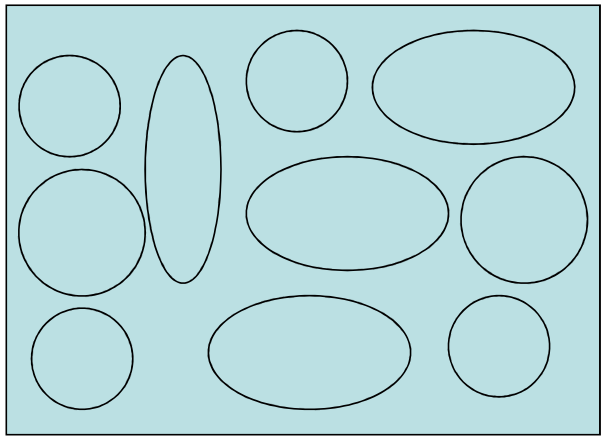
Principe de l'échantillonnage à deux degrés

- On partitionne la population U d'individus en M grosses unités appelées **Unités Primaires** (UP) ; les petites unités de U sont appelées les **Unités Secondaires** (US).
- Premier degré : on tire un échantillon d'UP.
- Second degré : dans chaque UP sélectionnée, on tire un échantillon d'US.

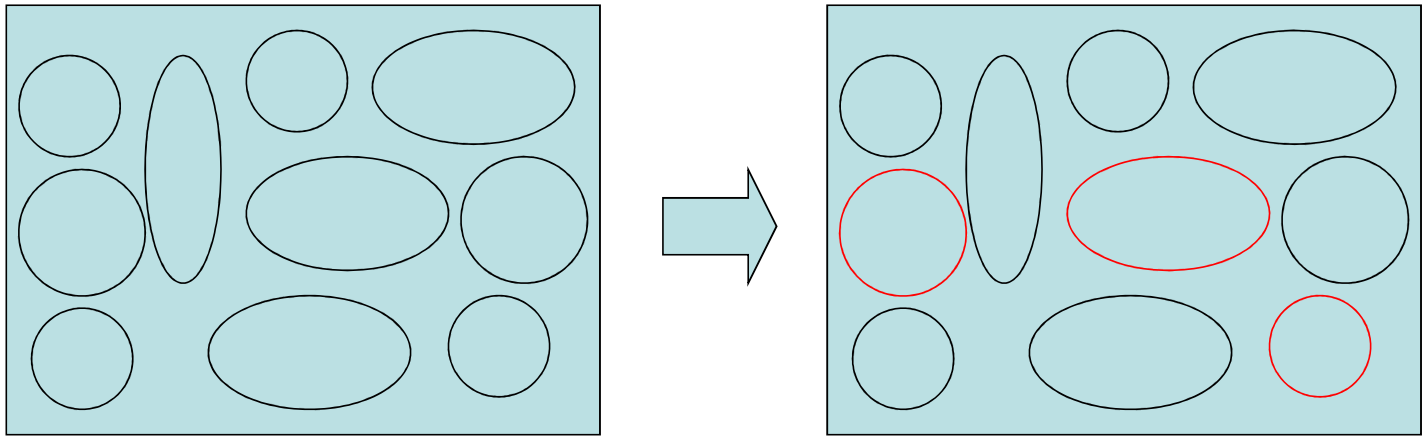
L'échantillonnage multi-degrés consiste en trois degrés de tirage ou plus. Dans le cas des enquêtes-ménage, un plan de sondage habituel consiste à

- tirer un échantillon de communes (UP),
- tirer un échantillon de quartiers dans les communes sélectionnées (US),
- tirer un échantillon de ménages dans les quartiers sélectionnés (UT).

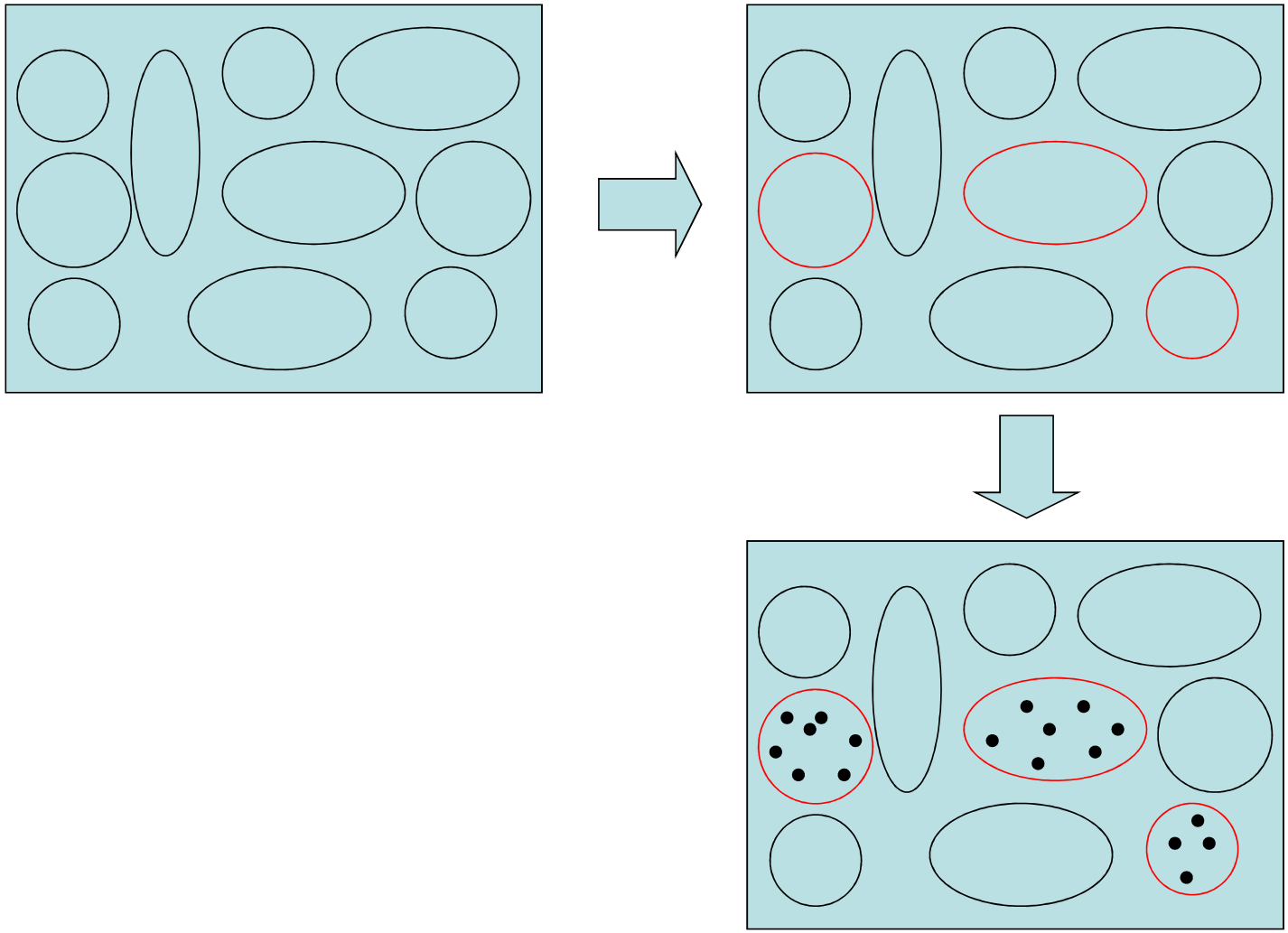
Principe du tirage multidegrés



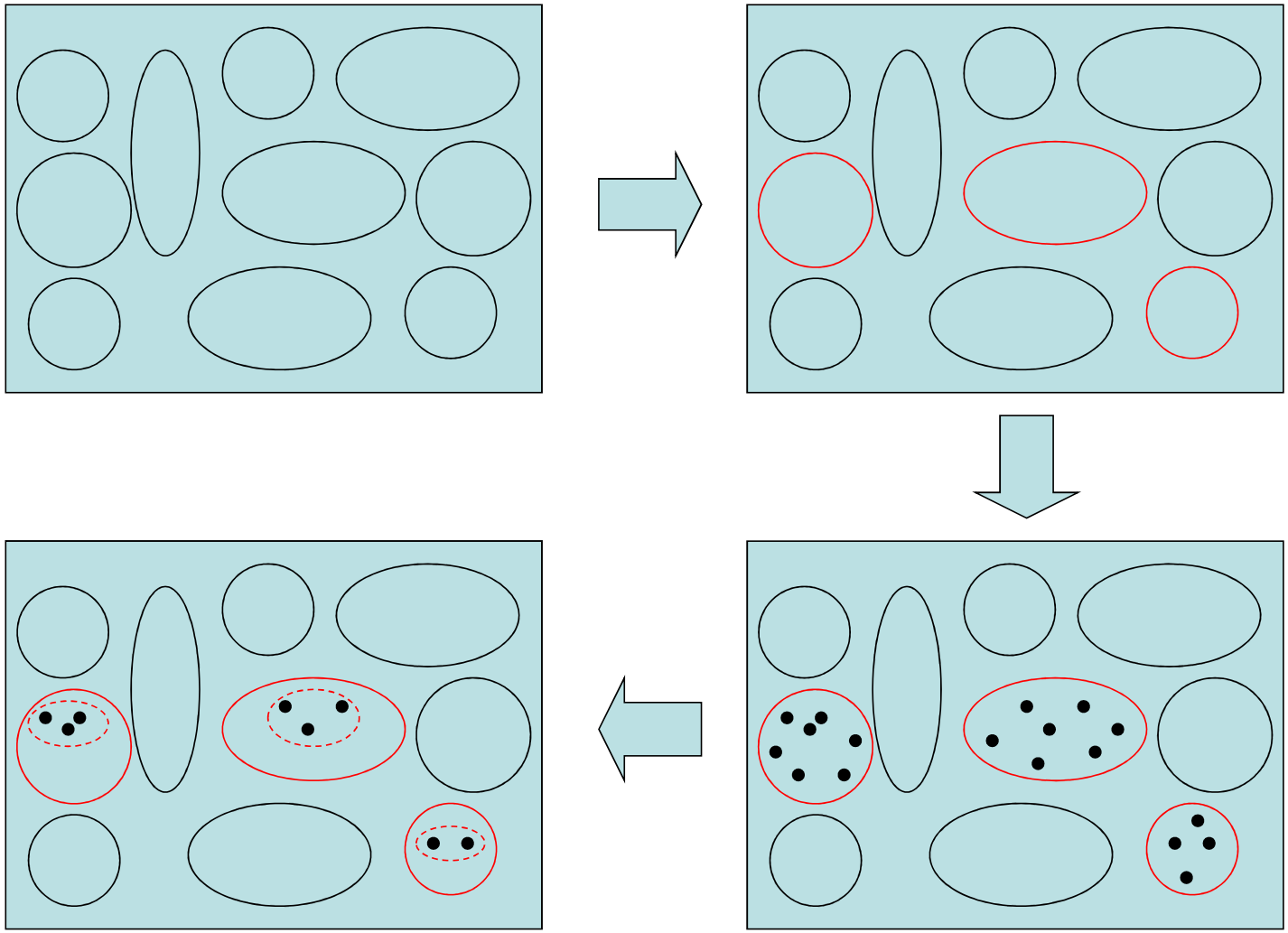
Principe du tirage multidegrés



Principe du tirage multidegrés



Principe du tirage multidegrés



Tirage sans remise des unités primaires

Dans le cas d'un sondage aléatoire simple à chaque degré, l'estimateur de Horvitz-Thompson du total est donné par

$$\hat{t}_{y\pi} = \frac{M}{m} \sum_{u_i \in S_I} N_i \, \bar{y}_i \quad \text{avec} \quad \bar{y}_i = \frac{1}{n_i} \sum_{k \in S_i} y_k.$$

Sa variance s'obtient à partir de la formule de Sen-Yates-Grundy :

$$V_p(\hat{t}_{y\pi}) = \underbrace{M^2 \left(1 - \frac{m}{M}\right) \frac{S_{yI}^2}{m}}_{V_{UP}} + \underbrace{\frac{M}{m} \sum_{u_i \in U_I} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{yi}^2}{n_i}}_{V_{US}}. \qquad (15)$$

On l'estime sans biais par

$$v_{YG}(\hat{t}_{y\pi}) = M^2 \left(1 - \frac{m}{M}\right) \frac{s_{yI}^2}{m} + \frac{M}{m} \sum_{u_i \in S_I} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{s_{yi}^2}{n_i}. \qquad (16)$$

Tirage sans remise des unités primaires (2)

Pour un plan de sondage quelconque au second degré, l'estimateur de Horvitz-Thompson vaut

$$\hat{t}_{y\pi} = \frac{M}{m} \sum_{u_i \in S_I} \hat{Y}_i$$

avec $\hat{Y}_i$ l'estimateur du sous-total Y_i.

La variance de $\hat{t}_{y\pi}$ vaut

$$V(\hat{t}_{y\pi}) = \frac{M^2}{m} \left\{ (1 - \frac{m}{M}) S_{yI}^2 + \frac{1}{M} \sum_{u_i \in U_I} V_i \right\}$$

avec $V_i = V(\hat{Y}_i)$.

Tirage sans remise des unités primaires (3)

Cette variance peut être estimée sans biais par

$$v(\hat{Y}) = \frac{M^2}{m}\left\{(1-\frac{m}{M})s_{yI}^2 + \frac{1}{M}\sum_{u_i \in S_I}\hat{V}_i\right\} \quad (17)$$

avec $\hat{V}_i$ un estimateur sans biais de V_i.

Pour un plan de sondage à d degrés, l'estimateur de variance va donc comporter un terme pour chaque degré d'échantillonnage.

Si le taux de sondage du 1er degré est faible et que les unités primaires sont de petite taille, on peut utiliser l'estimateur de variance simplifié

$$v(\hat{Y}) = \frac{M^2}{m}\left(1-\frac{m}{M}\right)s_{yI}^2. \quad (18)$$

Tirage avec remise des unités primaires

Supposons maintenant que l'échantillon S_I soit sélectionné selon un sondage aléatoire simple **avec remise**. L'estimateur de Hansen-Hurwitz vaut

$$\hat{t}_{yHH} = \frac{M}{m}\sum_{j=1}^{m}\hat{Y}_{(j)}.$$

Sa variance est égale à

$$V\left(\hat{t}_{yHH}\right) = \frac{M^2}{m}\left\{\frac{M-1}{M}S_{yI}^2 + \frac{1}{M}\sum_{u_i\in U_I}V_i\right\}.$$

Elle peut être estimée sans biais par

$$v_{WR}\left(\hat{t}_{yHH}\right) = \frac{M^2}{m}s_{yI}^2. \quad (19)$$

Tirage avec remise des unités primaires (2)

La forme simple de l'estimateur de variance vient du fait que $\hat{t}_{yHH}$ s'écrit comme une somme de variables aléatoire i.i.d. :

$$\hat{t}_{yHH} = \frac{M}{m} \sum_{j=1}^{m} \hat{Y}_{(j)}.$$

On n'a donc pas besoin d'estimateur de variance pour les tirages dans les unités primaires.

Si le taux de sondage du 1er degré est faible et que les unités primaires sont de petite taille, cet estimateur de variance sera approximativement non biaisé dans le cas d'un sondage aléatoire simple **sans remise** au premier degré.

En résumé : plans à un degré

Plan de sondage	Estimateur de t_y	Estimateur de variance	
		Expression	Formule
Tirage de Poisson	$\sum_{k\in S} \frac{y_k}{\pi_k}$	$\sum_{k\in S} \left(\frac{y_k}{\pi_k}\right)^2 (1-\pi_k)$	(4)
SRS	$N\bar{y}$	$N^2(1-f)\frac{s_y^2}{n}$	(8)
SRS stratifié	$\sum_{h=1}^{H} N_h\ \bar{y}_h$	$\sum_{h=1}^{H} N_h^2(1-f_h)\frac{s_{yh}^2}{n_h}$	(10)
Tirage réjectif	$\sum_{k\in S} \frac{y_k}{\pi_k}$	$\frac{n}{n-1}\sum_{k\in S} \frac{1-\pi_k}{\pi_k^2}\left(y_k - \pi_k\ \widehat{R}\right)^2$	(11)
Tirage équilibré	$\sum_{k\in S} \frac{y_k}{\pi_k}$	$\frac{n}{n-p}\sum_{k\in S} \frac{1-\pi_k}{\pi_k^2}\left(y_k - \mathbf{x}_k^\top \hat{\mathbf{B}}\right)^2$	(12)

En résumé : plans à plusieurs degrés

Plan de sondage	Estimateur de t_y	Estimateur de variance	
		Expression	Formule
Tirage avec remise des UP	$\hat{t}_{yHH} = \frac{M}{m}\sum_{u_i \in S_I} \hat{Y}_i$	$\frac{M^2}{m} s^2_{yI}$	(19)
Tirage sans remise des UP	$\hat{t}_{y\pi} = \frac{M}{m}\sum_{u_i \in S_I} \hat{Y}_i$	$\frac{M^2}{m}(1-\frac{m}{M})s^2_{yI} + \frac{M}{m}\sum_{u_i \in S_I} \hat{V}_i$	(17)
		$\frac{M^2}{m}(1-\frac{m}{M})s^2_{yI}$ *	(18)
		$\frac{M^2}{m} s^2_{yI}$ *	(19)

*Estimateur de variance approché utilisable si $m << M$

Estimation d'un paramètre complexe

Estimateur par substitution

Contexte

Si les probabilités d'inclusion d'ordre 1 et 2 sont connues, il est possible d'estimer sans biais un total, et d'obtenir une mesure de précision de cette estimation.

En pratique, on peut s'intéresser à des paramètres plus complexes :

- Estimation d'un ratio (ex : moyenne dans un domaine, taux de chômage),
- Estimation d'un coefficient de régression, ou d'un coefficient de corrélation,
- Estimation d'un fractile ou d'un indice (ex : mesure des inégalités avec l'indice de Gini ou l'indice de Theil).

Contexte

On suppose que le paramètre à estimer est de la forme

$$\theta = f(t_{\mathbf{y}})$$

où $\mathbf{y}_k = (y_{1k}, \ldots, y_{qk})^T$ désigne un q-vecteur de variables d'intérêt, et $f : \mathbf{R}^q \to \mathbf{R}$.

Définition

*On dit que la fonction f (et par extension, le paramètre θ) est **homogène d'ordre** m si*

$$\forall \alpha > 0 \quad \forall \mathbf{x} \in \mathbf{R}^q \quad f(\alpha \mathbf{x}) = \alpha^m f(\mathbf{x})$$

Exemples

1. $\theta = t_{y1} \quad f(x) = x$
   paramètre homogène d'ordre 1

2. $\theta = \frac{t_{y1}}{t_{y2}} \quad f(x_1, x_2) = \frac{x_1}{x_2}$
   paramètre homogène d'ordre 0

3. $\theta = t_{y1} \times t_{y2} \quad f(x_1, x_2) = x_1 \times x_2$
   paramètre homogène d'ordre 2

Remarques

Les paramètres usuels sont homogènes d'ordre $m = 0$ ou 1, voire 2 dans certains cas.

Si f est homogène d'ordre m, alors

$$\begin{aligned} \theta &= f(N\ \mu_{\mathbf{y}}) \\ &= N^m f(\mu_{\mathbf{y}}) \end{aligned}$$

où la moyenne $\mu_{\mathbf{y}} = t_{\mathbf{y}}/N$ est homogène d'ordre 0 (ou sans dimension). Le paramètre θ est donc "de l'ordre de N^m".

Si f est homogène d'ordre m et deux fois différentiable, alors les différentielles f' et f'' sont homogènes d'ordre $m-1$ et $m-2$, respectivement.

Estimation du paramètre

Il semble naturel d'estimer $\theta = f(t_{\mathbf{y}})$ en remplaçant le total $t_{\mathbf{y}}$ inconnu par son π-estimateur. On obtient l'**estimateur par substitution** :

$$\hat{\theta}_\pi = f(\hat{t}_{\mathbf{y}\pi}).$$

La **technique de linéarisation** va nous fournir un développement de $\hat{\theta}_\pi - \theta$, qui permet de montrer que $\hat{\theta}_\pi$ est approximativement non biaisé pour θ, et de calculer une approximation de variance pour cet estimateur.

La technique de linéarisation permet de **prendre en compte la forme du paramètre** θ pour se ramener d'un estimateur complexe $\hat{\theta}_\pi$ à un estimateur de Horvitz-Thompson particulier $\hat{t}_{u\pi}$, dont les propriétés sont bien connues.

Technique de linéarisation

Principe

Soit $\theta = f(t_{\mathbf{y}})$, avec f une fonction homogène d'ordre m. En utilisant un développement de Taylor à l'ordre 1, on obtient :

$$\begin{aligned} \hat{\theta}_\pi - \theta &= f(\hat{t}_{\mathbf{y}\pi}) - f(t_{\mathbf{y}}) \\ &\simeq \left[f'(t_{\mathbf{y}})\right]^T \left[\hat{t}_{\mathbf{y}\pi} - t_{\mathbf{y}}\right] \\ &= \hat{t}_{u\pi} - t_u, \end{aligned} \tag{20}$$

en notant

$$\begin{aligned} u_k &= \left[f'(t_{\mathbf{y}})\right]^T [\mathbf{y}_k] \\ &= \sum_{i=1}^{q} \frac{\partial f}{\partial x_i}(t_{\mathbf{y}})\ y_{ik}. \end{aligned}$$

Sous l'approximation (20), on obtient donc

$$E_p\left[\hat{\theta}_\pi - \theta\right] \simeq 0.$$

Proposition

Soit $\theta = f(t_{\mathbf{y}})$, où f est une fonction homogène d'ordre m, deux fois différentiable et dont les dérivées secondes sont continues au point $\mu_{\mathbf{y}}$. Alors l'approximation par linéarisation de la variance de $\hat{\theta}_\pi$ est donnée par :

$$V_p\left[\hat{t}_{u\pi}\right],$$

où

$$u_k \equiv u_k(\theta) \quad = \quad \left[f'(t_{\mathbf{y}})\right]^T \left[\mathbf{y}_k\right]$$

*est appelée la **variable linéarisée** du paramètre θ.*

Pour obtenir la variance (approchée) de $\hat{\theta}_\pi$, il suffit donc :

1. de calculer la variable linéarisée du paramètre θ.
2. d'utiliser la formule de variance correspondant au plan de sondage utilisé.

Approximation de variance

Pour un plan de sondage quelconque, on obtient :

$$\begin{aligned} V_p\left[\hat{\theta}_\pi\right] &\simeq V_p\left[\hat{t}_{u\pi}\right] \\ &= \sum_{k,l\in U} \frac{u_k}{\pi_k}\frac{u_l}{\pi_l}\Delta_{kl}. \end{aligned}$$

Pour un plan de sondage de taille fixe, on a

$$\begin{aligned} V_p\left[\hat{\theta}_\pi\right] &\simeq V_p\left[\hat{t}_{u\pi}\right] \\ &= -\frac{1}{2}\sum_{k\neq l\in U}\left(\frac{u_k}{\pi_k}-\frac{u_l}{\pi_l}\right)^2\Delta_{kl}. \end{aligned}$$

Quelques règles de calcul d'une variable linéarisée

Linéarisée d'une somme : $u_k(\theta_1 + \theta_2) = u_k(\theta_1) + u_k(\theta_2)$.

Linéarisée d'un produit : $u_k(\theta_1 \times \theta_2) = \theta_1 \times u_k(\theta_2) + \theta_2 \times u_k(\theta_1)$.

Linéarisée d'une fonction : Si θ est un paramètre scalaire, et g est une fonction dérivable $\mathbf{R} \to \mathbf{R}$, alors

$$u_k\left[g(\theta)\right] \quad = \quad g'(\theta)\; u_k(\theta).$$

Corollaire de la règle précédente :

$$u_k\left[\ln(\theta)\right] \quad = \quad \frac{u_k(\theta)}{\theta}.$$

Applications

L'estimation par substitution permet de traiter les cas particuliers importants :

- de l'estimation d'un ratio,
- de l'estimation dans un domaine de la population.

Le ratio est sans doute le paramètre le plus utilisé avec le total. On s'intéresse par exemple au revenu moyen des ménages, au taux de chômage (dans la population active), ...

On parle d'**estimation sur domaine** quand on souhaite produire des estimations sur des sous-populations, définies selon un critère géographique (région), socio-démographique (classe d'âge), ...

Estimation d'un ratio

Paramètre $R = \frac{t_y}{t_x}$, homogène d'ordre $m = 0$. On l'estime par substitution par $\hat{R}_\pi = \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}}$.

Calcul direct de la variable linéarisée :

$$f(x_1, x_2) = \frac{x_1}{x_2} \qquad f'(x_1, x_2) = \left(\frac{1}{x_2}, -\frac{x_1}{(x_2)^2} \right)$$

$$u_k(R) = \frac{1}{t_x} y_k - \frac{t_y}{(t_x)^2} x_k = \frac{1}{t_x}(y_k - R x_k)$$

Utilisation des règles de calcul :

$$\begin{aligned} \ln(R) &= \ln(t_y) - \ln(t_x) \\ \frac{u_k(R)}{R} &= \frac{y_k}{t_y} - \frac{x_k}{t_x} \end{aligned}$$

Estimation sur domaine

Soit U_d un domaine de la population U, i.e. une sous-population dont la taille N_d est supposée ici inconnue.

Le total d'une variable d'intérêt y dans le domaine peut s'estimer de façon habituelle, en remarquant que

$$t_{yd} = \sum_{k \in U_d} y_k = \sum_{k \in U} z_k,$$

avec $z_k = y_k \ 1(k \in U_d)$. Un estimateur sans biais est donc donné par

$$\hat{t}_{yd} = \sum_{k \in S} \frac{z_k}{\pi_k} = \sum_{k \in S_d} \frac{y_k}{\pi_k}$$

avec $S_d = S \cap U_d$.

Estimation sur domaine

La moyenne de la variable y dans le domaine, notée μ_{yd}, peut être estimée par

$$\tilde{\mu}_{yd} = \frac{\hat{t}_{yd}}{\hat{N}_d} \quad \text{avec} \quad \hat{N}_d = \sum_{k \in S_d} \frac{1}{\pi_k}.$$

La variable linéarisée de μ_{yd} est donnée par

$$u_k = \frac{1}{N_d} (y_k - \mu_{yd}) \, 1(k \in U_d).$$

Dans le cas d'un SRS(n), on obtient :

$$V_p \left[\tilde{\mu}_{yd} \right] \simeq N^2 \frac{1-f}{n} S_u^2.$$

Estimation de variance

Estimation de variance

On a vu que pour l'estimateur par substitution $\hat{\theta}_\pi$, une **approximation de variance** s'obtient

1. en calculant la variable linéarisée du paramètre θ,
2. en utilisant la formule de variance correspondant au plan de sondage utilisé.

$$\begin{aligned} V_p\left[\hat{\theta}_\pi\right] &\simeq V_p\left[\hat{t}_{u\pi}\right] \\ &= \sum_{k,l\in U} \frac{u_k}{\pi_k}\frac{u_l}{\pi_l}\Delta_{kl}. \end{aligned}$$

Estimation de variance

Pour l'**estimation de variance**, le problème est que la variable linéarisée u_k dépend généralement de paramètres inconnus. On procède en deux étapes :

1. on calcule la variable linéarisée estimée $\hat{u}_k$, obtenue en remplaçant dans u_k les paramètres inconnus par des estimateurs,
2. on injecte cette variable linéarisée estimée dans l'estimateur de variance correspondant au plan de sondage utilisé.

Par exemple, dans le cas du ratio $R = t_y/t_x$:

$$u_k = \frac{1}{t_x}(y_k - R\ x_k) \quad \Rightarrow \quad \hat{u}_k = \frac{1}{\hat{t}_{x\pi}}(y_k - \hat{R}_\pi\ x_k)$$

et on obtient l'estimateur de variance

$$v\left[\hat{R}_\pi\right] \quad = \quad \sum_{k,l \in S} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}}.$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS. Estimation des totaux t_x et t_y.

k	x_k	y_k	
1	5	1	
2	1	3	
3	4	2	
4	8	10	
	$\bar{x} = 4.5$ $s_x^2 = 8.3$	$\bar{y} = 4$ $s_y^2 = 16.7$	

$$\hat{t}_{x\pi} = N\bar{x} = 45 \qquad v\left[\hat{t}_{x\pi}\right] = N^2 \frac{1-f}{n} s_x^2 = 125$$
$$\hat{t}_{y\pi} = N\bar{y} = 40 \qquad v\left[\hat{t}_{y\pi}\right] = N^2 \frac{1-f}{n} s_y^2 = 250$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS. Estimation du ratio t_y/t_x.

k	x_k	y_k	$\hat{u}_k = \frac{1}{\hat{t}_{x\pi}}(y_k - \hat{R}x_k)$
1	5	1	-0.08
2	1	3	0.05
3	4	2	-0.03
4	8	10	0.06
	$\bar{x} = 4.5$ $s_x^2 = 8.3$	$\bar{y} = 4$ $s_y^2 = 16.7$	$\bar{\hat{u}} = 0$ $s_{\hat{u}}^2 = 4.4\ 10^{-3}$

$$\hat{t}_{x\pi} = N\bar{x} = 45 \qquad v\left[\hat{t}_{x\pi}\right] = N^2 \frac{1-f}{n} s_x^2 = 125$$
$$\hat{t}_{y\pi} = N\bar{y} = 40 \qquad v\left[\hat{t}_{y\pi}\right] = N^2 \frac{1-f}{n} s_y^2 = 250$$
$$\hat{R} = \frac{\bar{y}}{\bar{x}} = 0.89 \qquad v\left[\hat{R}\right] = N^2 \frac{1-f}{n} s_{\hat{u}}^2 = 0.07$$

Cas du sondage aléatoire simple stratifié

Le paramètre $\theta = f(t_{\mathbf{y}})$ est estimé (approximativement sans biais) par $\hat{\theta} = f(\hat{t}_{\mathbf{y}\pi})$ avec

$$\hat{t}_{\mathbf{y}\pi} = \sum_{h=1}^{H} N_h \, \bar{\mathbf{y}}_h \quad \text{où} \quad \bar{\mathbf{y}}_h = \frac{1}{n_h} \sum_{k \in S_h} \mathbf{y}_k.$$

L'estimateur de variance par linéarisation est

$$v_{YG}(\hat{\theta}) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} s_{\hat{u}h}^2 \quad \text{avec} \quad s_{\hat{u}h}^2 = \frac{1}{n_h - 1} \sum_{k \in S_h} (\hat{u}_k - \bar{\hat{u}}_h)^2,$$

et avec $\hat{u}_k$ la variable linéarisée estimée du paramètre θ.

Cas du sondage à deux degrés

Pour un sondage aléatoire simple à chaque degré, le paramètre $\theta = f(t_{\mathbf{y}})$ est estimé (approximativement sans biais) par $\hat{\theta} = f(\hat{t}_{\mathbf{y}\pi})$ avec

$$\hat{t}_{\mathbf{y}\pi} = \frac{M}{m} \sum_{u_i \in S_I} N_i \, \bar{\mathbf{y}}_i.$$

L'estimateur de variance par linéarisation est

$$v_{YG}(\hat{\theta}) = M^2 \left(1 - \frac{m}{M}\right) \frac{s^2_{\hat{u}I}}{m} + \frac{M}{m} \sum_{u_i \in S_I} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{s^2_{\hat{u}i}}{n_i},$$

avec

$$s^2_{\hat{u}I} = \frac{1}{m-1} \sum_{u_i \in s_I} \left(\hat{U}_i - \frac{\hat{t}_{\hat{u}\pi}}{M}\right)^2, \qquad \hat{U}_i = \frac{N_i}{n_i} \sum_{k \in S_i} \hat{u}_k,$$

$$s^2_{\hat{u}i} = \frac{1}{n_i - 1} \sum_{k \in S_i} \left(\hat{u}_k - \frac{\hat{U}_i}{N_i}\right)^2.$$

Avantages :

- S'adapte à un plan de sondage (presque) quelconque
- Utilisable avec un logiciel standard d'estimation de variance (tel que POULPE) $\rightarrow$ mettre en entrée la variable linéarisée
- Calcul rapide

Inconvénients :

- Nécessite le calcul d'une variable linéarisée pour chaque statistique
- Nécessite une connaissance précise du plan de sondage
- Difficile à utiliser si la stratégie d'estimation est complexe

La linéarisation : résumé

Soit $\theta = f(t_{\mathbf{y}})$ un paramètre complexe, et $\hat{\theta} = f(\hat{t}_{\mathbf{y}\pi})$ son estimateur par substitution. On peut obtenir un estimateur de variance pour $\hat{\theta}$:

1. en calculant la variance linéarisée u_k du paramètre θ,
2. en remplaçant dans u_k les paramètres inconnus par leur estimateur par substitution, pour obtenir la variable linéarisée estimée $\hat{u}_k$,
3. en remplaçant dans l'estimateur de variance habituel $v(\hat{t}_{y\pi})$ associé au plan de sondage utilisé la variable y_k par la variable linéarisée estimée $\hat{u}_k$.

Estimateur par calage

Principe

On suppose ici que l'on dispose de p variables auxiliaires

$$\mathbf{x}_k = [x_{1k}, \ldots, x_{pk}]^T$$

dont le total sur la population est connu.

On cherche à passer de l'estimateur de Horvitz-Thompson

$$\hat{t}_{y\pi} = \sum_{k \in S} d_k \; y_k$$

à un nouvel estimateur

$$\hat{t}_{yw} = \sum_{k \in S} w_k \; y_k$$

en se **calant** sur l'information auxiliaire $\mathbf{x}_k$.

Modification des poids

On cherche de nouveaux poids w_k qui **restent proches** des poids de départ d_k, et qui **vérifient les équations de calage** $\sum_{k \in S} w_k \mathbf{x}_k = t_{\mathbf{x}}$.
On résoud pour cela le problème d'optimisation sous contraintes :

$$\min_{w_k} \sum_{k \in s} d_k G\left(\frac{w_k}{d_k}\right) \quad \text{s.c.} \quad \sum_{k \in s} w_k \mathbf{x}_k = t_{\mathbf{x}}$$

où G désigne une **fonction de distance**.

On obtient

$$w_k \quad = \quad d_k F(\lambda^T \mathbf{x}_k)$$

avec F la fonction inverse de G'.

Les fonctions de distance usuelles

1. la méthode linéaire : $G(r) = \frac{1}{2}(r-1)^2$ et $F(u) = 1 + u$,
2. la méthode raking ratio : $G(r) = r\log(r) - r + 1$ et $F(u) = \exp(u)$,
3. la méthode linéaire tronquée,
4. la méthode logit.

Estimation de variance

Estimation après calage

Après calage, on a pour toute variable y l'estimateur calé :

$$\hat{t}_{yw} = \sum_{k \in s} w_k y_k.$$

L'estimation est exacte pour les totaux de variables auxiliaires, et approximativement sans biais pour les autres variables d'intérêt.

Quelle que soit la fonction de distance utilisée, la variance de l'estimateur calé $\hat{t}_{yw}$ est **approximativement celle de l'estimateur par la régression** :

$$V_p\left[\hat{t}_{yw}\right] \simeq \sum_{k,l \in U} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \Delta_{kl}$$

où $E_k = y_k - \mathbf{b}^T \mathbf{x}_k$ donne les résidus de la régression de y sur les variables $\mathbf{x}$.

Estimation de variance

Deux estimateurs de variance peuvent être utilisés :

$$
\begin{aligned}
v_1\left[\hat{t}_{yw}\right] &= \sum_{k,l\in S} \frac{e_k}{\pi_k}\frac{e_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}} \\
v_2\left[\hat{t}_{yw}\right] &= \sum_{k,l\in S} \frac{g_k\ e_k}{\pi_k}\frac{g_l\ e_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}},
\end{aligned}
$$

où $g_k = w_k/d_k$, et $e_k = y_k - \hat{\mathbf{b}}_\pi^T \mathbf{x}_k$ donne les résidus estimés.

Le second estimateur est généralement (légèrement) préférable.

Estimation de variance

Un logiciel classique d'estimation de variance pour l'estimation de totaux $\hat{t}_{y\pi}$ peut être utilisé pour l'estimation de variance d'estimateurs calés $\hat{t}_{yw}$ de la façon suivante :

- Effectuer sur l'échantillon S la régression pondérée (par les poids d_k) de la variable y sur les variables auxiliaires $x_1, \ldots, x_p$,
- Prendre les résidus e_k de la régression et calculer les $g_k = w_k/d_k$,
- Utiliser le logiciel en remplaçant les y_k par les e_k (estimateur de variance v_1) ou par les $g_k\ e_k$ (estimateur de variance v_2).

Exemple

Echantillon de taille $n = 5$ tiré selon un SRS dans une population de taille $N = 100$. On suppose connu le total $t_x = 320$.

x_{0k}	x_{1k}	y_k	
1	1	3	
1	3	1	
1	2	8	
1	5	15	
1	4	3	

$$\hat{t}_{x\pi} = 300 \quad \hat{t}_{y\pi} = 600 \quad v(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} s_y^2 = 6.08\ 10^4$$

Exemple

Echantillon de taille $n = 5$ tiré selon un SRS dans une population de taille $N = 100$. On suppose connu le total $t_x = 320$.

x_{0k}	x_{1k}	y_k	$e_k = y_k - \hat{a} - \hat{b}\, x_{1k}$
1	1	3	0.8
1	3	1	-5
1	2	8	3.9
1	5	15	5.2
1	4	3	-4.9

$$\hat{t}_{x\pi} = 300 \quad \hat{t}_{y\pi} = 600 \quad v(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} s_y^2 = 6.08\ 10^4$$

$$\hat{a} = 0.3 \quad \hat{b} = 1.9$$

$$\hat{t}_{yw} = 638 \quad v(\hat{t}_{yw}) = N^2 \frac{1-f}{n} s_e^2 = 4.365\ 10^4$$

Applications

Sondage aléatoire simple stratifié

On suppose encore que l'échantillon est calé sur les totaux $t_{\mathbf{x}}$, et on note $e_k = y_k - \hat{\mathbf{b}}_\pi^T \mathbf{x}_k$.

Les estimateurs de variance pour l'estimateur par calage $\hat{t}_{yw}$ sont :

$$v_1(\hat{t}_{yw}) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} s_{e,h}^2 \quad \text{avec} \quad s_{e,h}^2 = \frac{1}{n_h - 1} \sum_{k \in S_h} (e_k - \bar{e}_h)^2,$$

$$v_2(\hat{t}_{yw}) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} s_{ge,h}^2 \quad \text{avec} \quad s_{ge,h}^2 = \frac{1}{n_h - 1} \sum_{k \in S_h} (g_k\, e_k - \overline{ge}_h)^2.$$

Cas du tirage réjectif

On suppose encore que l'échantillon est calé sur les totaux $t_{\mathbf{x}}$, et on note $e_k = y_k - \hat{\mathbf{b}}_\pi^T \mathbf{x}_k$.

Les estimateurs de variance pour l'estimateur par calage $\hat{t}_{yw}$ sont :

$$v_1(\hat{t}_{yw}) = \frac{n}{n-1} \sum_{k \in S} (1-\pi_k) \left(\frac{e_k}{\pi_k} - \widehat{R}_e \right)^2 \quad \text{où} \quad \widehat{R}_e = \frac{\sum_{k \in S} \frac{e_k}{\pi_k} (1-\pi_k)}{\sum_{k \in S} (1-\pi_k)},$$

$$v_2(\hat{t}_{yw}) = \frac{n}{n-1} \sum_{k \in S} (1-\pi_k) \left(\frac{g_k \ e_k}{\pi_k} - \widehat{R}_{ge} \right)^2 \quad \text{où} \quad \widehat{R}_{ge} = \frac{\sum_{k \in S} \frac{g_k \ e_k}{\pi_k} (1-\pi_k)}{\sum_{k \in S} (1-\pi_k)}.$$

Cas du sondage à deux degrés

Le premier estimateur de variance pour l'estimateur par calage $\hat{t}_{yw}$ est :

$$v_1(\hat{t}_{yw}) = M^2\left(1-\frac{m}{M}\right)\frac{s_{eI}^2}{m} + \frac{M}{m}\sum_{u_i\in S_I} N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{s_{ei}^2}{n_i},$$

avec

$$s_{eI}^2 = \frac{1}{m-1}\sum_{u_i\in s_I}\left(\hat{E}_i - \frac{\hat{t}_{e\pi}}{M}\right)^2, \qquad \hat{E}_i = \frac{N_i}{n_i}\sum_{k\in S_i} e_k,$$

$$s_{ei}^2 = \frac{1}{n_i-1}\sum_{k\in S_i}\left(e_k - \frac{\hat{E}_i}{N_i}\right)^2.$$

L'estimateur de variance $v_2(\cdot)$ se calcule de façon analogue.

Estimation d'un paramètre complexe avec calage

Principe

On s'intéresse au paramètre $\theta = f(t_{\mathbf{y}})$, que l'on estime par $\hat{\theta}_w = f(\hat{t}_{\mathbf{y}w})$: l'estimateur de Horvitz-Thompson $\hat{t}_{\mathbf{y}\pi}$ est remplacé par l'estimateur calé $\hat{t}_{\mathbf{y}w}$.

L'estimateur $\hat{\theta}_w$ est encore approximativement sans biais pour θ. On obtient un estimateur de variance :

1. en calculant la variance linéarisée u_k du paramètre θ,
2. en calculant le résidu $F_k = u_k - \mathbf{b}_u^T \mathbf{x}_k$ de la régression de u_k sur les variables de calage,
3. en remplaçant dans F_k les paramètres inconnus par leurs estimateurs $\Rightarrow f_k = \hat{u}_k - \hat{\mathbf{b}}_{u\pi}^T \mathbf{x}_k$
4. en remplaçant dans l'estimateur de variance habituel $v(\hat{t}_{y\pi})$ associé au plan de sondage utilisé la variable y_k par la variable de résidu f_k.

Application
Enquête Logement 2006

En résumé

L'Enquête Logement 2006 est une enquête auprès des ménages, qui a donné lieu à une extension régionale et à plusieurs extensions locales au niveau de la région Bretagne notamment. Un complément d'échantillon a également été sélectionné dans des bases externes.

Un plan de sondage et une technique d'estimation complexes ont été nécessaires pour la mise en commun et l'exploitation des différents sous-échantillons, la prise en compte de la non-réponse et le redressement des estimateurs.

Un outil SAS de calcul de précision basé sur les formules d'estimation de variance a été mis au point pour les partenaires de l'Enquête et les chargés d'étude de la Direction Régionale.

Présentation de l'enquête

L'Enquête Logement est une des plus grosses enquêtes réalisées par l'Insee auprès des ménages. Elle a lieu environ tous les quatre ans (dernières éditions en 2002 et 2006).

Le champ est celui des logements résidences principales en 2006, accessibles à l'aide du RP99 et de la Base de Sondage de Logements Neufs (BSLN).

Objectifs de l'enquête :

- connaître le parc de logements (ancienneté de la construction, nombre de maisons individuelles/appartements, nombre de propriétaires/locataires,...),
- décrire les conditions de vie des ménages (mobilités et causes de mobilité, confort du logement, emprunts,...).

Sélection de l'Enquête Logement

L'échantillon est sélectionné en quatre temps :

- Sélection de l'échantillon national dans l'Echantillon Maître de 99 (RP99, BSLN),
- Sélection d'une extension régionale dans l'EMEX, pour les régions concernées,
- Sélection d'extensions d'échantillon au niveau local, pour les régions concernées,
- Sélection d'échantillons complémentaires dans des bases externes.

L'EM 99

L'échantillon maître de 1999 (EM99) est une réserve de logements destinée à servir de base de sondage pour les enquêtes auprès des ménages.

Il est obtenu par un tirage stratifié (selon le degré d'urbanisation) et à plusieurs degrés. On sélectionne des communes dans le rural, des districts (pâtés de maisons) dans l'urbain, ... (Ardilly, 2006).

Les échantillons destinés aux enquêtes ménages seront ensuite tirés dans les zones sélectionnées. Dans ces zones, une liste à jour de logements est fournie par le RP99 et la BSLN.

L'EMEX

Pour les extensions régionales, il existe un échantillon-maître spécifique : l'EMEX (Bourdalle et al., 2000), constitué selon des principes voisins de l'EM :

- tirage stratifié (selon le degré d'urbanisation) et à plusieurs degrés,
- même système de rotation des logements situés dans l'EMEX,
- disjonction par rapport à l'EM.

A quoi servent les extensions ? La précision est liée à la taille d'échantillon : plus le domaine d'estimation est petit, plus la précision se détériore. La sélection d'un échantillon dans l'EMEX vise à assurer de meilleures estimations au niveau régional.

Schéma récapitulatif

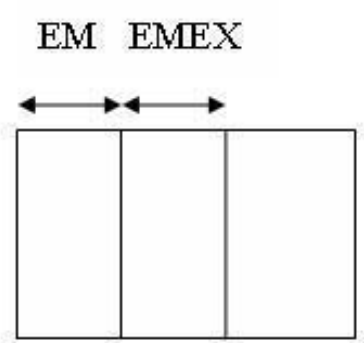


Schéma récapitulatif

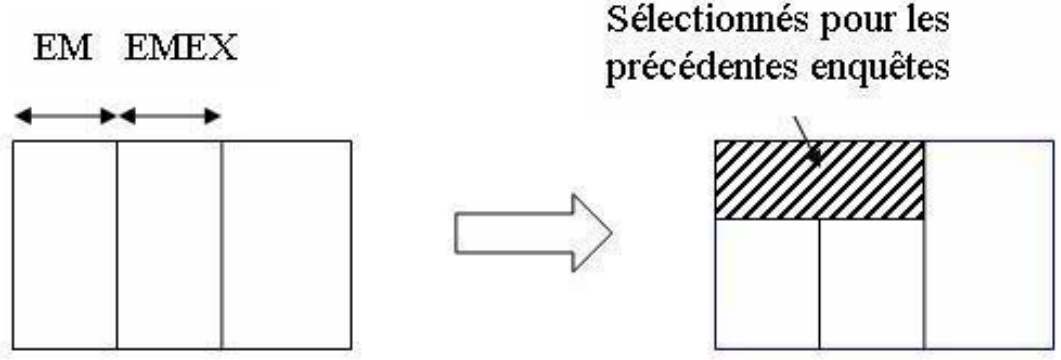
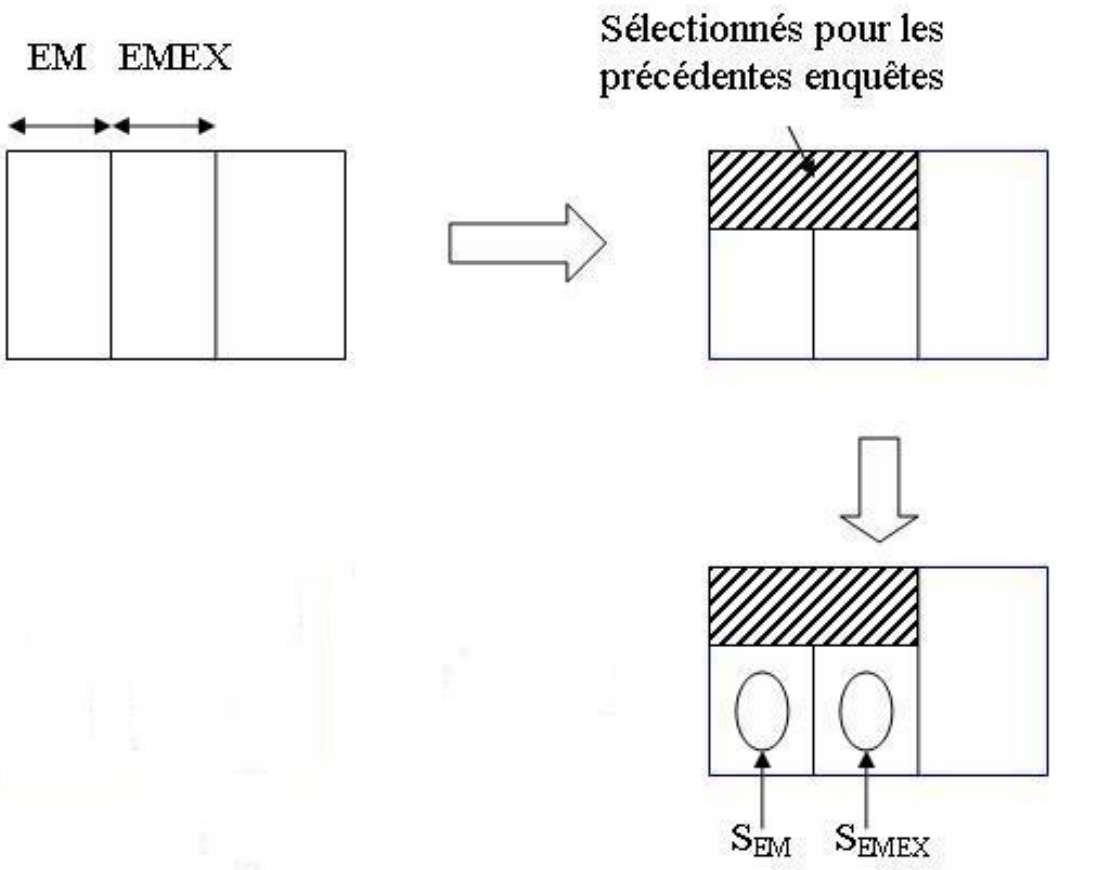


Schéma récapitulatif



Extensions locales d'échantillon

Un complément d'échantillon peut être sélectionné autour de zones particulières pour lesquelles on souhaite produire des estimations fiables.

En Bretagne, c'est le cas des 6 principales aires urbaines (Brest, Lorient, Quimper, Rennes, Saint-Brieuc et Vannes).

Cet échantillon est sélectionné en excluant les logements précedemment échantillonnés dans l'EM ou l'EMEX au titre de l'Enquête Logement.

Bases externes

Enfin, pour surreprésenter des sous-populations particulières, des échantillons ont été sélectionnés dans des fichiers externes :

- Base des adresses situées dans les Zones Urbaines Sensibles (ZUS),
- Base d'allocataires de prestations.

Cette sélection n'est pas disjointe de celle des autres sous-échantillons.

Au niveau de la Bretagne, seul l'échantillon ZUS a été utilisé (fusion des autres échantillons très problématique).

Schéma récapitulatif

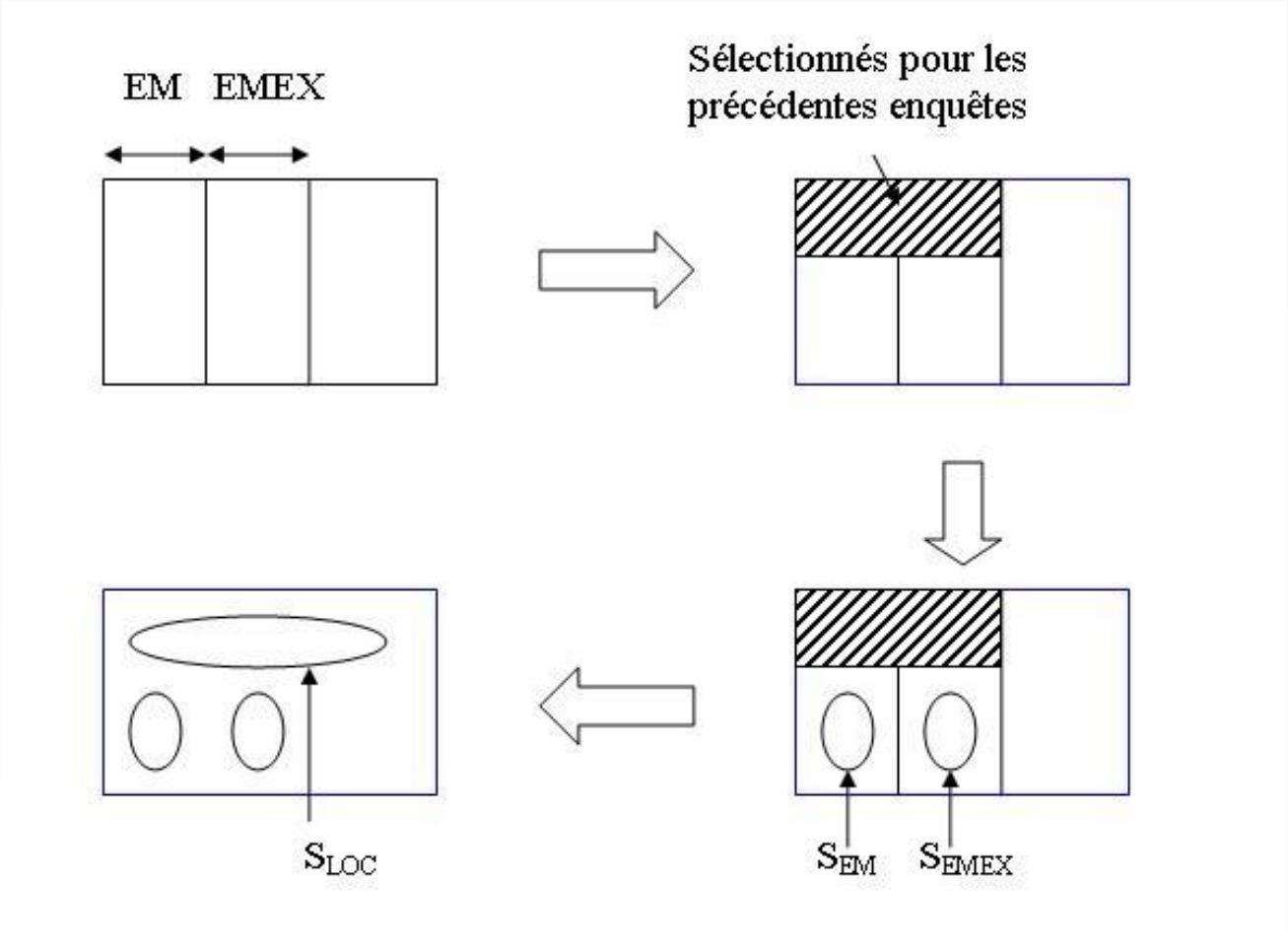


Schéma récapitulatif

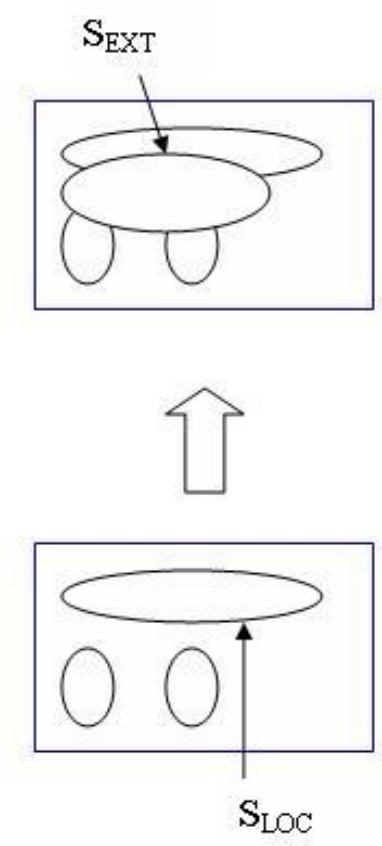
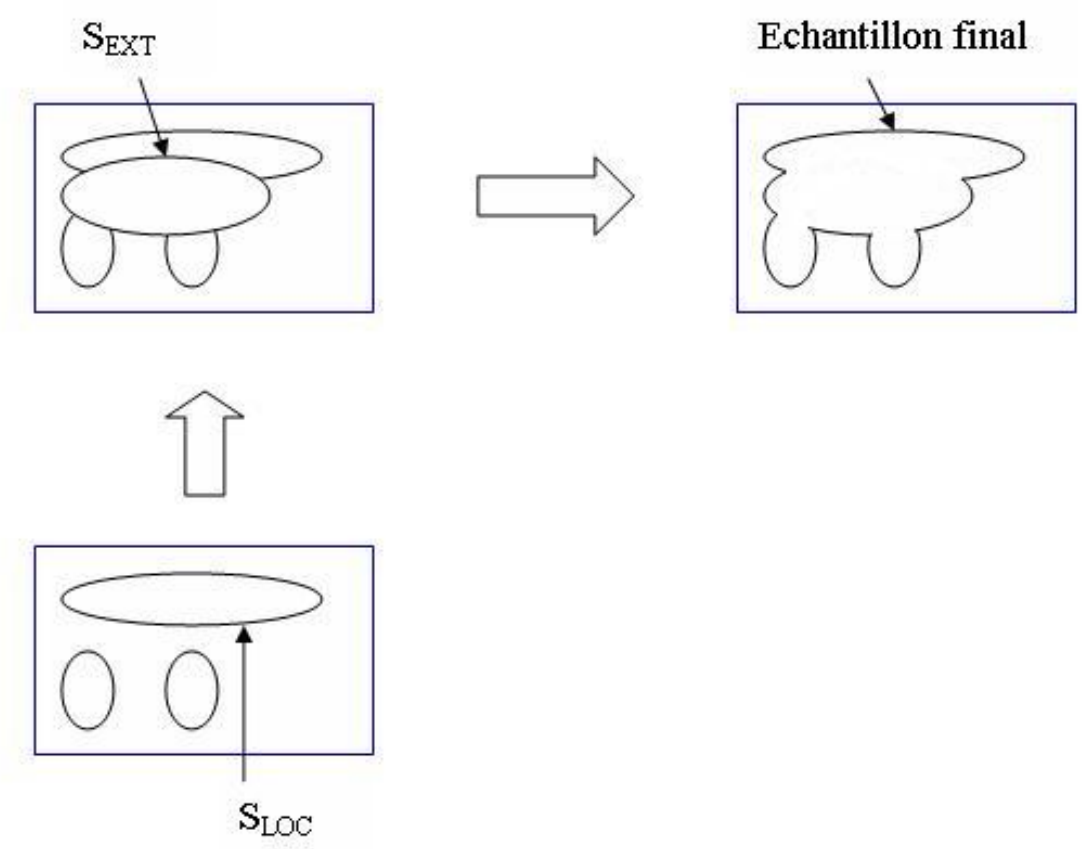


Schéma récapitulatif



Mise en commun des sous-échantillons

La difficulté consiste à produire un estimateur sans biais en gérant les trois sous-échantillons et leur intersection. Il y a essentiellement deux problèmes :

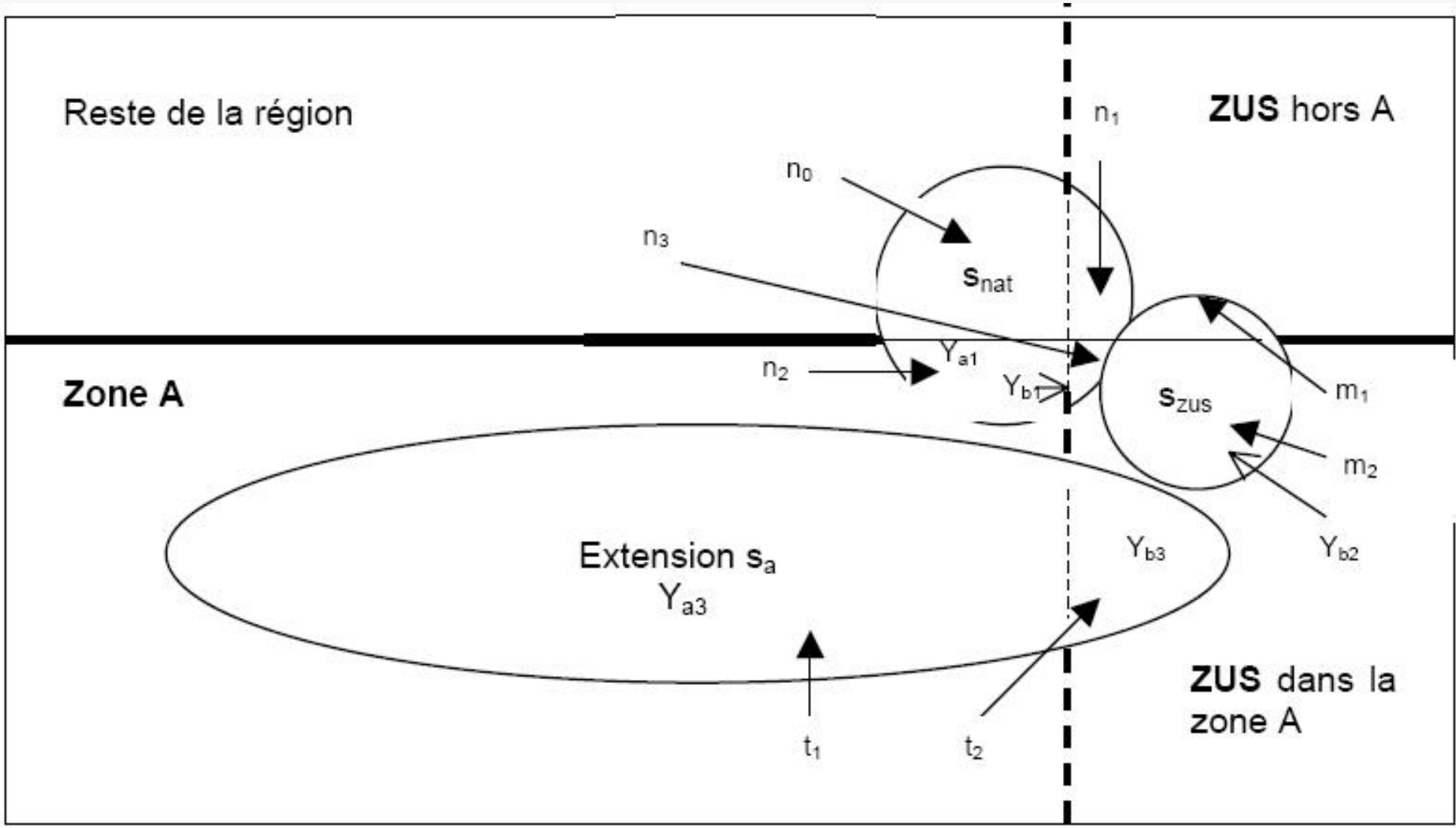
- Des sous-échantillons même disjoints peuvent représenter une même sous-population. On a donc un risque de biais dû à des doubles ou triples comptes.
- Certains logements sont sélectionnés dans deux échantillons différents.

Les sous-échantillons sont mis en commun à l'aide de la technique d'estimation composite. On note

$$\hat{t}_{yd} = \sum_{k \in S} d_k y_k$$

l'estimateur obtenu, avec S la réunion des sous-échantillons et d_k le poids du logement k.

Schéma récapitulatif (Le Guennec, 2009)



Estimation de précision

Les étapes de l'enquête

Les différentes étapes de traitement de l'enquête (Le Guennec, 2009) sont :

1. Sélection des sous-échantillons,
2. Mise en commun des sous-échantillons,
3. Redressement de la non-réponse partielle (imputation),
4. Redressement de la non-réponse totale (méthode des groupes homogènes de réponse),
5. Calage sur une information externe.

L'estimation de variance réalisée ne prend pas en compte la variance d'imputation.

Calcul de variance (1)

On note

$$\hat{t}_{yw} = \sum_{k \in S} w_k y_k$$

l'estimateur obtenu avec les poids calés w_k. On a :

$$V\left(\hat{t}_{yw}\right) \simeq V\left(\hat{t}_{ed}\right)$$

avec $e_k = y_k - \hat{\mathbf{b}}_\pi^\top \mathbf{x}_k$ le résidu de régression de y sur les variables de calage.

Cette variance se décompose en

$$V\left(\hat{t}_{ed}\right) = V_p\left(\hat{t}_{ed}\right) + V_{nr}\left(\hat{t}_{ed}\right),$$

où la variance due à l'échantillonnage $V_p(\cdot)$ et la variance due à la non-réponse $V_{nr}(\cdot)$ sont estimées séparément.

Calcul de variance (2)

L'estimateur $\hat{t}_{e,d}$ peut se décomposer sur les trois sous-échantillons tirés nationalement (N), localement (L) ou dans les ZUS (Z) :

$$\hat{t}_{ed} = \hat{t}^N_{\tilde{e}d} + \hat{t}^L_{\tilde{e}d} + \hat{t}^Z_{\tilde{e}d}$$

avec $\tilde{e}_k = f(e_k)$ la variable synthétique associée à la technique d'estimation composite.

En utilisant l'indépendance (réelle ou approchée) des trois sous-échantillons :

$$V_p\left(\hat{t}_{ed}\right) \simeq V_p\left(\hat{t}^N_{\tilde{e}d}\right) + V_p\left(\hat{t}^L_{\tilde{e}d}\right) + V_p\left(\hat{t}^Z_{\tilde{e}d}\right).$$

Ces trois termes sont ensuite estimés séparément.

Un exemple de calcul de précision

On souhaite estimer, sur l'ensemble des résidences principales de l'aire urbaine de Rennes :

- La structure des logements selon le nombre de chambres,
- La surface moyenne par habitant.

On est dans le cas d'une estimation sur un **domaine** D, c'est à dire sur une sous-population de U. Cette estimation ne pose pas de problème particulier, en remarquant que

$$t_{yD} = \sum_{k \in D} y_k = \sum_{k \in U} y_k \ 1_{k \in D}$$

où $1_{k \in D}$ vaut 1 si le logement k est dans le domaine, et 0 sinon.

On va donc estimer :

- des effectifs (nombre de logements ne comptant aucune chambre, comptant 1 chambre, ...),
- un ratio (surface totale rapportée au nombre d'habitants).

Dans le premier cas, on estime un total (variable indicatrice pour un effectif).

Dans le second cas, on estime un ratio de totaux : l'estimation de variance se fait par linéarisation (Deville, 1999).

Résultats obtenus

Paramètre	Estim.	Var.	CV (%)	BI (95%)	BS (95%)	DEFF	DCAL	NR (%)
Surf. moy.	$38,27$	$0,23$	$1,25$	$37,34$	$39,21$	$0,48$	$0,42$	$21,79$
% Log.								
0 cha.	$0,08$	$4,3\ 10^{-5}$	$7,75$	$0,07$	$0,10$	$0,46$	$0,99$	$17,34$
1 cha.	$0,18$	$1,1\ 10^{-4}$	$5,69$	$0,16$	$0,20$	$0,59$	$0,90$	$21,80$
2 cha.	$0,26$	$2,1\ 10^{-4}$	$5,68$	$0,23$	$0,29$	$0,91$	$0,96$	$19,15$
3 cha.	$0,26$	$3,0\ 10^{-4}$	$6,64$	$0,23$	$0,29$	$1,26$	$0,98$	$18,06$
4 cha.	$0,18$	$1,7\ 10^{-4}$	$7,43$	$0,15$	$0,20$	$0,97$	$0,80$	$18,82$
5 cha.	$0,04$	$8,5\ 10^{-5}$	$23,6$	$0,02$	$0,06$	$1,86$	$0,97$	$19,96$
6 cha.	$3\ 10^{-3}$	$1,7\ 10^{-6}$	$48,5$	10^{-4}	$5,2\ 10^{-3}$	$0,52$	$1,01$	$17,36$
+ 6 cha.	$4\ 10^{-4}$	$3,3\ 10^{-7}$	152	-10^{-3}	$1,5\ 10^{-3}$	$0,72$	$1,01$	$17,21$

Méthodes de rééchantillonnage

Introduction

Contexte

Cette section (y compris les notations) s'appuie largement sur Shao et Tu (1994) "The Jackknife and the Bootstrap".

On souhaite étudier les propriétés d'une population (finie ou infinie), en utilisant les données relevées sur un échantillon i.i.d. $X_1, \ldots, X_n$. Cet échantillon est généré selon une distribution inconnue F, d'espérance m et de variance σ^2 supposées finies.

On notera également :

$$T_n \equiv T_n(X_1, \ldots, X_n)$$

un estimateur quelconque.

Exemples

1) La moyenne simple :

$$T_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

estimateur naturel de l'espérance m de la loi F.

2) Une fonction de moyennes :

$$T_n = f\left(\bar{X}_n\right),$$

où on impose généralement des conditions de régularité sur la fonction $f(\cdot)$ (continuité, différentiabilité, ...).
Cas particuliers : ratio, coefficient de corrélation ou de régression, ...

Exemples (2)

3) La distribution empirique :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x)$$

qui estime la loi inconnue F, avec $1(\cdot)$ variable indicatrice.

4) Un quantile empirique :

$$T_n = F_n^{-1}(t),$$

avec $t \in [0, 1]$ et

$$F_n^{-1}(t) = \text{Inf}\{x; F_n(x) \geq t\}.$$

Quelques rappels

Moyenne

Nous considérons tout d'abord le cas de données $X_1, \ldots, X_n$ unidimensionnelles, et de l'estimation de l'espérance m de la loi F.

L'estimateur de m est donné par

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{de variance} \quad V\left(\bar{X}_n\right) = \frac{\sigma^2}{n}.$$

Cette variance peut être estimée sans biais par

$$v\left(\bar{X}_n\right) = \frac{s_X^2}{n} \qquad \text{avec} \qquad s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

Moyenne : intervalle de confiance

La loi limite de l'estimateur $\bar{X}_n$ est donnée par le théorème central-limite :

$$\frac{\sqrt{n}(\bar{X}_n - m)}{s_X} \quad \rightarrow_{\mathcal{L}} \quad \mathcal{N}(0,1),$$

On obtient l'intervalle de confiance de niveau $1-2\alpha$ pour le paramètre m :

$$\left[\bar{X}_n \pm u_\alpha \ \frac{s_X}{\sqrt{n}}\right]$$

avec u_α le fractile d'ordre α d'une loi $\mathcal{N}(0,1)$.

$$\text{Exemples :} \quad u_{0.025} = 1.96 \quad u_{0.05} = 1.64$$

Fonction de moyennes

Nous considérons maintenant le cas de p-vecteurs $X_1, \ldots, X_n$ générés selon une loi commune F de dimension p, et d'espérance m.

On souhaite estimer le paramètre $\theta = f(m)$, où la fonction $f(\cdot)$ est supposée continûment différentiable.

Ce paramètre peut être estimé par substitution (ou plug-in) par

$$\hat{\theta} = f\left(\bar{X}_n\right) = T_n.$$

Principe :

$$\begin{aligned} \hat{\theta} - \theta &= f\left(\bar{X}_n\right) - f(m) \\ &\simeq \left[f'(m)\right]^T \left(\bar{X}_n - m\right) \end{aligned}$$

Fonction de moyennes : estimation de variance

On en déduit :

$$V\left(\hat{\theta}\right) \simeq V\left(\left[f'(m)\right]^T \bar{X}_n\right).$$

Cette variance peut être estimée asymptotiquement sans biais par

$$v\left(\hat{\theta}\right) = \frac{s_U^2}{n},$$

avec

$$s_U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(U_i - \bar{U}_n)^2$$

et $U_i = \left[f'(\bar{X}_n)\right]^T X_i$ la variable linéarisée.
$\Rightarrow$ estimateur de variance par linéarisation.

Fonction de moyennes : intervalle de confiance

La loi limite de l'estimateur $\hat{\theta}$ est donnée par :

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{s_U} \quad \rightarrow_{\mathcal{L}} \quad \mathcal{N}(0,1),$$

On obtient l'intervalle de confiance de niveau $1 - 2\alpha$ pour le paramètre θ :

$$\left[\hat{\theta} \pm u_\alpha \frac{s_U}{\sqrt{n}}\right]$$

avec u_α le fractile d'ordre α d'une loi $\mathcal{N}(0,1)$.

Méthodes de rééchantillonnage

Le Jackknife

Le Jackknife

Le Jackknife a été introduit à l'origine par Quenouille (1949a,b) pour estimer le biais d'une statistique, puis a été proposé pour l'estimation de variance par Tukey (1958).

Principe du Jackknife : on recalcule la statistique estimée en supprimant chaque unité tour à tour. La variabilité des statistiques Jackknifées est utilisée afin d'estimer la variance.

Cette technique est encore appelée le delete-1 Jackknife.

Cas d'une fonction de moyennes

Nous considérons le cas de p-vecteurs $X_1, \ldots, X_n$ générés selon une loi commune F de dimension p, et d'espérance m.

On souhaite estimer le paramètre $\theta = f(m)$, où la fonction $f(\cdot)$ est supposée continûment différentiable.

Ce paramètre est estimé par substitution par

$$\hat{\theta} \quad = \quad f\left(\bar{X}_n\right) = T_n.$$

L'estimateur basé sur l'échantillon privé de l'individu j est donné par

$$\hat{\theta}_{-j} = f(\bar{X}_{n,-j}) \quad \text{où} \quad \bar{X}_{n,-j} = \frac{1}{n-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} X_i.$$

Cas d'une fonction de moyennes (2)

L'estimateur Jackknife du biais de $\hat{\theta}$ est donné par

$$b_{JACK}\left[\hat{\theta}\right] = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{-i} - \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_{-j}\right).$$

L'estimateur Jackknife de la variance de $\hat{\theta}$ est donné par

$$v_{JACK}\left[\hat{\theta}\right] = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{-i} - \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_{-j}\right)^2.$$

Pour une fonction $f(\cdot)$ continûment différentiable au point m, le Jackknife donne une estimation de variance consistante pour $\hat{\theta}$.

Cas d'une moyenne

L'estimateur Jackknife du biais de $\bar{X}_n$ est donné par

$$b_{JACK}\left[\bar{X}_n\right] = \frac{n-1}{n}\sum_{j=1}^{n}\left(\bar{X}_{n,-j} - \bar{X}_n\right).$$

On montre que cet estimateur est égal à 0
$\Rightarrow$ cohérent avec le caractère non biaisé de $\bar{X}_n$.

L'estimateur Jackknife de la variance de $\bar{X}_n$ est donné par

$$\begin{aligned} v_{JACK}\left[\bar{X}_n\right] &= \frac{n-1}{n}\sum_{j=1}^{n}\left(\bar{X}_{n,-j} - \bar{X}_n\right)^2 \\ &= \frac{s_X^2}{n}. \end{aligned}$$

$\Rightarrow$ restitue l'estimateur sans biais de la variance

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i				
1	1				
2	3				
3	2				
4	10				

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i				
1	1				
2	3				
3	2				
4	10				
	$\bar{X}_n = 4$				

$$v\left[\bar{X}_n\right] = \frac{s_X^2}{n} = 4.17$$

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i	S_{-1}			
1	1	-			
2	3	3			
3	2	2			
4	10	10			
	$\bar{X}_n = 4$				

$$v\left[\bar{X}_n\right] = \frac{s_X^2}{n} = 4.17$$

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i	S_{-1}			
1	1	-			
2	3	3			
3	2	2			
4	10	10			
	$\bar{X}_n = 4$	$\bar{X}_{n,-1} = 5$			

$$v\left[\bar{X}_n\right] = \frac{s_X^2}{n} = 4.17$$

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i	S_{-1}	S_{-2}	S_{-3}	S_{-4}
1	1	-	1	1	1
2	3	3	-	3	3
3	2	2	2	-	2
4	10	10	10	10	-
	$\bar{X}_n = 4$	$\bar{X}_{n,-1} = 5$			

$$v\left[\bar{X}_n\right] = \frac{s_X^2}{n} = 4.17$$

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i	S_{-1}	S_{-2}	S_{-3}	S_{-4}
1	1	-	1	1	1
2	3	3	-	3	3
3	2	2	2	-	2
4	10	10	10	10	-
	$\bar{X}_n = 4$	$\bar{X}_{n,-1} = 5$	$\bar{X}_{n,-2} = 4.33$	$\bar{X}_{n,-3} = 4.67$	$\bar{X}_{n,-4} = 2$

$$v\left[\bar{X}_n\right] = \frac{s_X^2}{n} = 4.17$$

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i	S_{-1}	S_{-2}	S_{-3}	S_{-4}
1	1	-	1	1	1
2	3	3	-	3	3
3	2	2	2	-	2
4	10	10	10	10	-
	$\bar{X}_n = 4$	$\bar{X}_{n,-1} = 5$	$\bar{X}_{n,-2} = 4.33$	$\bar{X}_{n,-3} = 4.67$	$\bar{X}_{n,-4} = 2$

$v\left[\bar{X}_n\right] = \frac{s_X^2}{n} = 4.17$	$v_{JACK}\left[\bar{X}_n\right] = \frac{n-1}{n}\sum_{i=1}^{n}\left(\bar{X}_{n,-i} - \bar{X}_n\right)^2 = 4.17$

Linéarisation Jackknife

L'estimateur Jackknife de la variance peut se réécrire

$$v_{JACK}\left[\hat{\theta}\right] = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\tilde{\theta}_i - \frac{1}{n} \sum_{j=1}^{n} \tilde{\theta}_j \right)^2 = \frac{s^2_{\tilde{\theta}}}{n},$$

où

$$\tilde{\theta}_i = (n-1)(\hat{\theta} - \hat{\theta}_{-i})$$

donne une approximation numérique de la variable linéarisée (Davison et Hinkley, 1997,p .50).

Le Jackknife peut donc être vu comme une méthode numérique permettant d'éviter le calcul des variables linéarisées.

Shao et Tu (1995) montrent que l'estimation de variance est également consistante sous des conditions plus faibles de différentiabilité.

Exemple

On génère une population artificielle de taille $N = 1\,000$ contenant deux variables X et Y. La variable X est générée selon une loi gamma de paramètres 2 et 5. La variable Y est générée de façon à ce que $\rho(X, Y) \simeq 0.7$.

On prélève un échantillon de taille $n = 100$ dans U, par sondage aléatoire simple avec remise. L'objectif est d'estimer le ratio

$$R = \frac{m_Y}{m_X}.$$

On compare pour le paramètre R la linéarisée de Taylor et la linéarisée Jackknife.

Exemple

i				
12				
14				
16				
71				
82				
122				
126				
128				

Exemple

i	X_i	Y_i		
12	3.28	7.37		
14	14.75	18.02		
16	16.66	13.52		
71	15.95	11.86		
82	7.31	21.80		
122	11.53	21.05		
126	12.25	4.28		
128	24.04	31.34		

Exemple

i	X_i	Y_i	Linéarisée U_i	
12	3.28	7.37	0.395	
14	14.75	18.02	0.373	
16	16.66	13.52	-0.212	
71	15.95	11.86	-0.303	
82	7.31	21.80	1.380	
122	11.53	21.05	0.939	
126	12.25	4.28	-0.682	
128	24.04	31.34	0.793	

Exemple

i	X_i	Y_i	Linéarisée U_i	Lin. Jackknife $\tilde{\theta}_i$
12	3.28	7.37	0.395	0.392
14	14.75	18.02	0.373	0.375
16	16.66	13.52	-0.212	-0.213
71	15.95	11.86	-0.303	-0.304
82	7.31	21.80	1.380	1.376
122	11.53	21.05	0.939	0.940
126	12.25	4.28	-0.682	-0.683
128	24.04	31.34	0.793	0.803

Intervalle de confiance

La consistance de l'estimateur de variance Jackknife implique que :

$$\frac{\hat{\theta} - \theta}{\sqrt{v_{JACK}\left[\hat{\theta}\right]}} \rightarrow_{\mathcal{L}} \mathcal{N}(0,1),$$

On peut donc utiliser cet estimateur de variance pour produire un intervalle de confiance de niveau $1-2\alpha$ pour le paramètre θ :

$$\left[\hat{\theta} \pm u_\alpha \sqrt{v_{JACK}\left[\hat{\theta}\right]}\right].$$

Avantage : variable linéarisée remplacée par une approximation numérique.

Delete-d jackknife

Le delete-1 Jackknife implique un total de n suppressions, ce qui peut être prohibitif si la taille d'échantillon est grande. Les suppressions peuvent être également réalisées par blocs de d unités à la fois (Shao et Wu, 1985). On parle alors de delete-d Jackknife.

Cette méthode a également été étudiée par Shao et Wu (1989) afin de produire une estimation consistante de variance pour des paramètres non lisses tels que les fractiles. En particulier, les nombres d et $n - d$ d'individus supprimés et d'individus conservés doivent tendre vers l'infini avec n.

Le Bootstrap

Le Bootstrap

Le Bootstrap a été introduit par Efron (1979) dans le cadre d'une population infinie. L'idée de base consiste à reproduire le mécanisme d'échantillonnage d'origine.

Le Bootstrap permet d'obtenir une approximation de la distribution d'un estimateur. Il permet d'obtenir une estimation de variance consistante, y compris pour des paramètres non lisses tels que les fractiles.

L'adaptation au cas d'une population finie est assez problématique, et fait l'objet d'une littérature abondante, voir en particulier Shao et Tu (1995), Davison et Hinkley (1997), Davison et Sardy (2007).

Principe

L'idée de base du Bootstrap est un principe de substitution ou plug-in. Soit $X_1, \ldots, X_n$ un échantillon iid de distribution inconnue $F(\cdot)$, et de fonction de répartition empirique

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x)$$

1. Plug-in : un paramètre $\theta(F)$ est estimé par $\theta(F_n)$.
2. Si $\theta(F_n)$ n'est pas calculable :
   1. On tire avec remise B fois dans $X_1, \ldots, X_n$. On obtient

$$(X_1^{*b}, \ldots, X_n^{*b}) \quad \text{et} \quad F_n^{*b}(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i^{*b} \leq x).$$

   2. On remplace $\theta(F_n)$ par $\frac{1}{B} \sum_{b=1}^{B} \theta(F_n^{*b})$.

Fonction de moyennes : estimation du biais

Nous considérons le cas de p-vecteurs $X_1, \ldots, X_n$ générés selon une loi commune F de dimension p, et d'espérance m. On estime $\theta = f(m)$, où $f(\cdot)$ est supposée continûment différentiable.

L'estimateur basé sur le rééchantillon S^* est donné par $\hat{\theta}^* = f(\bar{X}_n^*)$.

Estimateur Bootstrap du biais de $\hat{\theta}$:

$$b_{BOOT}\left[\hat{\theta}\right] = E^*\left[f(\bar{X}_n^*) - f(\bar{X}_n)\right].$$

Approximation de Monte-Carlo :

$$b_{BOOT}^B\left[\hat{\theta}\right] = \frac{1}{B}\sum_{b=1}^{B}\left[f(\bar{X}_n^{*b}) - f(\bar{X}_n)\right].$$

Fonction de moyennes : estimation de la variance

Estimateur Bootstrap de la variance de $\hat{\theta}$:

$$v_{BOOT}\left[\hat{\theta}\right] = E^*\left[f(\bar{X}_n^*) - E^*[f(\bar{X}_n^*)]\right]^2$$

Approximation de Monte-Carlo :

$$v_{BOOT}^B\left[\hat{\theta}\right] = \frac{1}{B-1}\sum_{b=1}^{B}\left[f(\bar{X}_n^{*b}) - \frac{1}{B}\sum_{c=1}^{B} f(\bar{X}_n^{*c})\right]^2 .$$

Si la fonction $f(\cdot)$ est continument différentiable au point μ_y, le Bootstrap donne une estimation de variance consistante pour $\hat{\theta}$.

Principe :

- $\hat{\theta}^* - \hat{\theta} = f(\bar{X}_n^*) - f(\bar{X}_n) \simeq \left[f'(\bar{X}_n)\right]^T \left[\bar{X}_n^* - \bar{X}_n\right]$,
- consistance de l'estimateur de variance par linéarisation.

Moyenne : estimation du biais

Le biais de $\bar{X}_n$ est donné par

$$\begin{aligned} B\left[\bar{X}_n\right] &= E\left[\bar{X}_n - m\right] \\ &= E\left[\int x dF_n(x) - \int x dF(x)\right] \\ &= 0. \end{aligned}$$

L'estimateur Bootstrap du biais de $\bar{X}_n$ est donné par

$$\begin{aligned} b_{BOOT}\left[\bar{X}_n\right] &= E^*\left[\bar{X}_n^* - \bar{X}_n\right] \\ &= \bar{X}_n - \bar{X}_n = 0. \end{aligned}$$

avec $E^*(\cdot)$ l'espérance sous le mécanisme de rééchantillonnage.
$\Rightarrow$ cohérent avec le caractère non biaisé de $\bar{X}_n$.

Moyenne : estimation de la variance

La variance de $\bar{X}_n$ est donnée par

$$\begin{aligned} V\left[\bar{X}_n\right] &= E\left[\bar{X}_n - m\right]^2 \\ &= \sigma^2/n. \end{aligned}$$

L'estimateur Bootstrap de la variance de $\bar{X}_n$ est donné par

$$\begin{aligned} v_{BOOT}\left[\bar{X}_n\right] &= E^*\left[\bar{X}_n^* - \bar{X}_n\right]^2 \\ &= \left(\frac{n-1}{n}\right)\frac{s_y^2}{n}. \end{aligned}$$

Rééchantillonner avec une taille $m = n - 1$ permet de supprimer le biais (Efron, 1982).

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i						
1	1						
2	3						
3	2						
4	10						

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i						
1	1						
2	3						
3	2						
4	10						
	$\bar{X} = 4$						

$$v\left[\bar{X}\right] = \frac{s_X^2}{n} = 4.17$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i	W_1					
1	1	0					
2	3	3					
3	2	1					
4	10	0					
	$\bar{X} = 4$						

$$v\left[\bar{X}\right] = \frac{s_X^2}{n} = 4.17$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i	W_1					
1	1	0					
2	3	3					
3	2	1					
4	10	0					
	$\bar{X} = 4$	$\bar{X}_1^* = 2.75$					

$$v\left[\bar{X}\right] = \frac{s_X^2}{n} = 4.17$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i	W_1	W_2	W_3	W_4	W_5	W_6
1	1	0	0	1	0	1	1
2	3	3	0	1	0	1	3
3	2	1	1	0	2	1	0
4	10	0	3	2	2	1	0
	$\bar{X} = 4$	$\bar{X}_1^* = 2.75$					

$$v\left[\bar{X}\right] = \frac{s_X^2}{n} = 4.17$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i	W_1	W_2	W_3	W_4	W_5	W_6
1	1	0	0	1	0	1	1
2	3	3	0	1	0	1	3
3	2	1	1	0	2	1	0
4	10	0	3	2	2	1	0
	$\bar{X} = 4$	$\bar{X}_1^* = 2.75$	$\bar{X}_2^* = 8$	$\bar{X}_3^* = 6$	$\bar{X}_4^* = 6$	$\bar{X}_5^* = 4$	$\bar{X}_6^* = 2.5$

$$v\left[\bar{X}\right] = \frac{s_X^2}{n} = 4.17$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i	W_1	W_2	W_3	W_4	W_5	W_6
1	1	0	0	1	0	1	1
2	3	3	0	1	0	1	3
3	2	1	1	0	2	1	0
4	10	0	3	2	2	1	0
	$\bar{X} = 4$	$\bar{X}_1^* = 2.75$	$\bar{X}_2^* = 8$	$\bar{X}_3^* = 6$	$\bar{X}_4^* = 6$	$\bar{X}_5^* = 4$	$\bar{X}_6^* = 2.5$

$v\left[\bar{X}\right] = \frac{s_X^2}{n} = 4.17$	$v_{BOOT}^B\left[\bar{X}\right] = 3.87$

Intervalle de confiance

Un des intérêts du Bootstrap est que l'on estime non seulement la variance de $\hat{\theta}$, mais aussi sa distribution. Les statistiques Bootstrappées peuvent être utilisées pour produire un intervalle de confiance.

La méthode des percentiles est une solution simple pour obtenir un intervalle de confiance :

- On tire B rééchantillons $S_1^*, \ldots, S_B^* \Rightarrow \hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.
- On les trie par ordre croissant $\Rightarrow \hat{\theta}_{(1)}^*, \ldots, \hat{\theta}_{(B)}^*$.
- On supprime les $\alpha\%$ les plus faibles et les $\alpha\%$ les plus grandes pour obtenir l'intervalle de confiance de niveau $(1-2\alpha)$:

$$\left[\hat{\theta}_{(L)}^*, \hat{\theta}_{(U)}^*\right].$$

Rééchantillonnage pour données d'enquête

Le Jackknife

Sondage aléatoire simple sans remise

Comme la linéarisation, le Jackknife est une technique permettant de prendre en compte la forme de la statistique dans le calcul de variance, mais pas le plan de sondage.

Pour obtenir une estimation consistante de variance avec un SAS sans remise, l'estimateur de variance doit être remplacé par

$$\begin{aligned} v_{JACK}\left[\hat{\theta}\right] &= \frac{1-f}{n} s^2_{\tilde{\theta}} \\ &= (1-f)\left[\frac{n-1}{n}\sum_{k\in S}\left(\hat{\theta}_{-k} - \frac{1}{n}\sum_{l\in S}\hat{\theta}_{-l}\right)^2\right]. \end{aligned}$$

On peut le voir comme l'estimateur de variance par linéarisation, où la variable linéarisée est remplacée par son approximation Jackknife.

Tirage stratifié

L'estimation de variance Jackknife se généralise facilement au cas d'un sondage aléatoire simple stratifié. Si le paramètre d'intérêt est de la forme $\theta = \sum_{h=1}^{H} \theta_h$, l'estimateur de variance Jackknife est donné par

$$\begin{aligned} v_{JACK}\left[\hat{\theta}\right] &= \sum_{h=1}^{H} v_{JACK}\left[\hat{\theta}_h\right] \\ &= \sum_{h=1}^{H} \frac{1-f_h}{n_h} s^2_{\tilde{\theta}_h} \\ &= \sum_{h=1}^{H} (1-f_h) \frac{n_h-1}{n_h} \sum_{k \in S_h} \left(\hat{\theta}_{h,-k} - \frac{1}{n_h} \sum_{l \in S_h} \hat{\theta}_{h,-l}\right)^2, \end{aligned}$$

où l'estimateur $\hat{\theta}_{h,-k}$ est recalculé en supprimant l'unité k de l'échantillon S_h, et en multipliant les autres unités de S_h par un facteur $n_h/(n_h-1)$.

Tirage à probabilités inégales

Dans le cas d'un tirage à probabilités inégales, Campbell (1980) propose d'utiliser l'estimateur de variance de Horvitz-Thompson ou de Yates-Grundy, en injectant l'approximation Jackknife de la variable linéarisée (voir aussi Berger et Skinner, 2005).

Berger (2007) a proposé d'utiliser l'estimateur de variance de Hajek dans le cas d'un tirage à forte entropie, i.e. d'utiliser l'estimateur de variance

$$v\left[\hat{t}_{y\pi}\right] = \sum_{k\in S}(1-\pi_k)\left(\frac{y_k}{\pi_k} - \widehat{R}\right)^2 \text{ où } \widehat{R} = \frac{\sum_{k\in S}\frac{y_k}{\pi_k}(1-\pi_k)}{\sum_{k\in S}(1-\pi_k)},$$

en remplaçant la variable y_k par une linéarisée de type Jackknife $\tilde{\theta}_k$.

Tirage multi-degrés

Le Jackknife peut être étendu au cas d'un tirage multidegrés où les unités primaires sont sélectionnées **avec remise** : on supprime successivement chaque unité primaire.

L'estimateur Jackknife de variance s'obtient en remplaçant dans l'estimateur de variance

$$v\left[\hat{t}_{y\pi}\right] = \frac{m}{m-1}\sum_{u_i \in S_I}\left[\frac{\hat{t}_{y_i\pi}}{\pi_{Ii}} - \frac{\hat{t}_{y\pi}}{m}\right]^2$$

la variable d'intérêt y par la linéarisée Jackknife.

Krewski et Rao (1981) montrent la consistance de cet estimateur de variance, voir également Rao et Wu (1985), Kovar et al. (1988).

Extensions

Le delete-1 Jackknife implique un total de n suppressions, ce qui peut être prohibitif si la taille d'échantillon est grande. Les suppressions peuvent être également réalisées par blocs de d unités à la fois (Shao et Wu, 1985). On parle alors de delete-d Jackknife.

Cette méthode a également été proposée par Shao et Wu (1989) afin de produire une estimation consistante de variance pour le Jackknife pour des paramètres non lisses tels que les quantiles.

Le Jackknife : résumé

L'estimation de variance Jackknife est fortement comparable à l'estimation de variance par linéarisation (avec ses avantages et ses inconvénients).

Le Jackknife permet de prendre en compte la forme du paramètre dans l'estimation de variance, mais la connaissance du plan de sondage reste nécessaire.

Le Jackknife permet d'éviter le calcul explicite de la variable linéarisée, mais sa mise en oeuvre peut être complexe si le nombre de suppressions est grand.

Le Bootstrap

Sondage aléatoire simple sans remise

L'extension au cas d'un sondage simple sans remise est problématique. Deux approches ont principalement été proposées dans la littérature :

- Approche 1 : Reproduire le mécanisme d'échantillonnage d'origine, dans l'esprit du Bootstrap d'Efron.
- Approche 2 : Reproduire un estimateur de variance (approximativement) sans biais dans le cas de l'estimation d'une moyenne (ou d'un total).

Approche 1 : la méthode de Gross

Gross (1980) suggère d'utiliser l'échantillon S afin de recréer une pseudo-population U^*, dans laquelle on rééchantillonne selon un SAS sans remise :

1. Obtenir U^* en dupliquant N/n fois chaque unité k de S,
2. Tirer dans U^* un rééchantillon S^* selon un SAS sans remise de taille n, pour obtenir un estimateur $\hat{\theta}^*$,
3. Répéter l'étape 2 indépendamment B fois, et estimer $V[\theta]$ par

$$v_{BOOT}^{B} = \frac{1}{B-1}\sum_{b=1}^{B}\left[f(\bar{y}_b^*) - \frac{1}{B}\sum_{c=1}^{B} f(\bar{y}_c^*)\right]^2 .$$

Procédure connue sous le nom de Bootstrap sans remise (BWO), ou Bootstrap populationnel.

Approche 1 : la méthode de Gross

Dans le cas où la variable y est unidimensionnelle, avec $\theta = \mu_y$ et $\hat{\theta} = \bar{y}$, on a :

$$\begin{aligned} v_{BOOT}\left[\hat{\theta}\right] &= \frac{N(n-1)}{n(N-1)}\frac{1-f}{n}s_y^2 \\ &= \frac{N(n-1)}{n(N-1)}v\left[\bar{y}\right]. \end{aligned}$$

L'estimateur de variance Bootstrap est donc biaisé. Le biais est négligeable si la taille d'échantillon est grande, mais peut être appréciable si n est borné (cas d'un sondage aléatoire simple stratifié avec une stratification très fine).

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 5$ est sélectionné selon un SAS sans remise.

Valeurs de la variable y échantillonnées :

$$S \equiv \{1, 3, 2, 10, 8\}.$$

Pseudo-population obtenue :

$$U^* \equiv \{1, 1, 3, 3, 2, 2, 10, 10, 8, 8\}.$$

On rééchantillonne dans la population U^* selon un SAS sans remise de taille 5. Une variable de poids Bootstrap donne le nombre de fois où l'unité est sélectionnée dans le rééchantillon.

Exemple

k	y_k						
1	1						
2	3						
3	2						
4	10						
5	8						

Exemple

k	y_k						
1	1						
2	3						
3	2						
4	10						
5	8						
	$\bar{y} = 4.8$						

$$v\left[\bar{y}\right] = (1-f)\frac{s_y^2}{n} = 1.57$$

Exemple

k	y_k	W_1					
1	1	2					
2	3	0					
3	2	1					
4	10	1					
5	8	1					
	$\bar{y} = 4.8$						

$$v\left[\bar{y}\right] = (1-f)\frac{s_y^2}{n} = 1.57$$

Exemple

k	y_k	W_1					
1	1	2					
2	3	0					
3	2	1					
4	10	1					
5	8	1					
	$\bar{y} = 4.8$	$\bar{y}_1^* = 4.4$					

$$v\left[\bar{y}\right] = (1-f)\frac{s_y^2}{n} = 1.57$$

Exemple

k	y_k	W_1	W_2	W_3	W_4	W_5	W_6
1	1	2	2	0	0	1	1
2	3	0	2	2	2	1	0
3	2	1	1	2	0	2	1
4	10	1	0	1	2	0	1
5	8	1	0	0	1	1	2
	$\bar{y} = 4.8$	$\bar{y}_1^* = 4.4$					

$$v\left[\bar{y}\right] = (1-f)\frac{s_y^2}{n} = 1.57$$

Exemple

k	y_k	W_1	W_2	W_3	W_4	W_5	W_6
1	1	2	2	0	0	1	1
2	3	0	2	2	2	1	0
3	2	1	1	2	0	2	1
4	10	1	0	1	2	0	1
5	8	1	0	0	1	1	2
	$\bar{y} = 4.8$	$\bar{y}_1^* = 4.4$	$\bar{y}_2^* = 2$	$\bar{y}_3^* = 4$	$\bar{y}_4^* = 6.8$	$\bar{y}_5^* = 3.2$	$\bar{y}_6^* = 5.8$

$$v\left[\bar{y}\right] = (1-f)\frac{s_y^2}{n} = 1.57$$

Exemple

k	y_k	W_1	W_2	W_3	W_4	W_5	W_6
1	1	2	2	0	0	1	1
2	3	0	2	2	2	1	0
3	2	1	1	2	0	2	1
4	10	1	0	1	2	0	1
5	8	1	0	0	1	1	2
	$\bar{y} = 4.8$	$\bar{y}_1^* = 4.4$	$\bar{y}_2^* = 2$	$\bar{y}_3^* = 4$	$\bar{y}_4^* = 6.8$	$\bar{y}_5^* = 3.2$	$\bar{y}_6^* = 5.8$

$v\left[\bar{y}\right] = (1-f)\frac{s_y^2}{n} = 1.57$	$B = 50 \Rightarrow v_{BOOT}^B\left[\bar{y}\right] = 1.35$

Extensions de la méthode de Gross

L'algorithme de Bootstrap suppose que le nombre de duplications N/n est entier. Dans le cas contraire, une adaptation est nécessaire.

Bickel et Freedman (1985) et Chao et Lo (1984) proposent une randomisation entre deux pseudo-populations. Booth et al. (1994) proposent d'arrondir N/n pour la duplication, et de compléter la pseudo-population U^* par sondage aléatoire simple dans S.

Sitter (1992) propose une randomisation sur le nombre de duplications et la taille de rééchantillon afin de retrouver (en moyenne) le bon estimateur de variance pour $\bar{y}$.

Approche 2 : le Bootstrap avec remise

Mac Carthy et Snowden (1985) proposent de rééchantillonner avec remise dans S, pour obtenir un rééchantillon S^* de taille m. Procédure connue sous le nom de Bootstrap avec remise (BWR).

Mac Carthy et Snowden suggèrent le choix

$$m = \frac{n-1}{1-f},$$

qui conduit à une estimation sans biais de variance pour $\bar{y}$. Si ce choix est impossible, une randomisation sur la taille de rééchantillon est nécessaire.

Approche 2 : le Rescaling Bootstrap

Rao et Wu (1985) proposent également de rééchantillonner avec remise dans S, pour obtenir un rééchantillon S^* de taille m. Dans chaque rééchantillon S^*, les poids Bootstrap sont ajustés afin de restituer l'estimateur sans biais de variance pour $\bar{y}$, voir également Rao et al. (1992).

Rao et Wu argumentent de la consistance de l'estimateur de variance, dans le cas où le paramètre est une fonction lisse de moyennes. Ils suggèrent une valeur optimale pour la taille de rééchantillon m, mais cette valeur peut être non entière.

Approche 2 : le Mirror-Match Bootstrap

Cette méthode proposée par Sitter (1992) combine les deux approches. Un rééchantillon S^* est obtenu en sous-échantillonnant k fois selon un SAS de taille n^*, et en agglomérant les sous-échantillons obtenus.

Le choix $k = \frac{n(1-f)}{n^*(1-f^*)}$ où $f^* = n^*/n$ permet de retrouver l'estimateur de variance sans biais de $\bar{y}$. Sitter suggère le choix $f^* = f$, et de randomiser k et n^* si nécessaire.

Là encore, Sitter argumente de la consistance de l'estimateur de variance, dans le cas où le paramètre est une fonction lisse de moyennes.

Quelle approche choisir ?

La littérature est assez contradictoire à ce sujet. Certains jeux de simulation (Rao et Wu, 1984, Sitter, 1993) plaident en faveur des méthodes ad-hoc (approche 2). D'autres études (Presnell et Booth, 1994, Davison et Hinkley, 1997, Davison et Sardy, 2007) suggèrent le contraire.

La méthode proposée par Gross est intuitivement plus proche du principe de Bootstrap proposé par Efron. Toutes les méthodes de Bootstrap proposées présentent des difficultés pratiques. La généralisation à un plan de sondage complexe est laborieuse, en dehors de quelques plans de sondage particuliers.

Tirage stratifié

Les procédures proposées se généralisent facilement au cas d'un sondage aléatoire simple stratifié : la même procédure est appliquée dans chaque strate.

Si la stratification est fine, un biais même faible d'un estimateur de variance dans chaque strate peut conduire à un biais global important. Pour cette raison, Rao et Wu (1984) et Sitter (1993) recommandent l'utilisation des méthodes du Rescaled Bootstrap et du Mirror-Match Bootstrap.

Les résultats obtenus par Davison et Hinkley (1997) montrent au contraire qu'une approche de type Gross donne de bons résultats, même avec un nombre important de strates.

Tirage à probabilités inégales

Le principe de la méthode de Gross peut se généraliser à un tirage à probabilités inégales, en utilisant le principe d'estimation de Horvitz-Thompson : une unité k de l'échantillon représente $1/\pi_k$ unités de la population.

La pseudo-population U^* est obtenue $1/\pi_k$ fois chaque individu k de l'échantillon. Le plan de sondage d'origine est appliqué dans U^*. Une simulation (Chauvet, 2007) montre un bon comportement de cette méthode pour un tirage à forte entropie, mais :

- problème de la duplication ($1/\pi_k$ rarement entier)
- validation théorique de la méthode ?
- échec pour un tirage à faible entropie (ex : tirage systématique).

Voir également Antal et Tillé (2009), pour une approche ad-hoc.

Exemple

Population U de taille $N = 10$, dans laquelle un échantillon S de taille $n = 2$ est sélectionné selon un tirage systématique ordonné à probabilités égales (tri de la population selon une variable auxiliaire x).

k	1	2	3	4	5	6	7	8	9	10

Exemple

Population U de taille $N = 10$, dans laquelle un échantillon S de taille $n = 2$ est sélectionné selon un tirage systématique ordonné à probabilités égales (tri de la population selon une variable auxiliaire x).

k	1	2	3	4	5	6	7	8	9	10
x_k	1	3	4	9	12	15	20	35	36	39

Exemple

Population U de taille $N = 10$, dans laquelle un échantillon S de taille $n = 2$ est sélectionné selon un tirage systématique ordonné à probabilités égales (tri de la population selon une variable auxiliaire x).

k	1	2	3	4	5	6	7	8	9	10
x_k	1	3	4	9	12	15	20	35	36	39
y_k	2	1	7	5	9	11	8	14	5	20

Exemple

Population U de taille $N = 10$, dans laquelle un échantillon S de taille $n = 2$ est sélectionné selon un tirage systématique ordonné à probabilités égales (tri de la population selon une variable auxiliaire x).

k	1	**2**	3	4	5	6	7	8	9	10
x_k	1	**3**	4	9	12	15	20	35	36	39
y_k	2	**1**	7	5	9	11	8	14	5	20

On tire un individu au hasard parmi les 5 premiers.

Exemple

Population U de taille $N = 10$, dans laquelle un échantillon S de taille $n = 2$ est sélectionné selon un tirage systématique ordonné à probabilités égales (tri de la population selon une variable auxiliaire x).

k	1	**2**	3	4	5	6	**7**	8	9	10
x_k	1	**3**	4	9	12	15	**20**	35	36	39
y_k	2	**1**	7	5	9	11	**8**	14	5	20

On tire un individu au hasard parmi les 5 premiers. Puis on fait un bond de taille 5.

Exemple

Si on applique la méthode de Gross, on obtient la pseudo-population U^* :

k	2	7	2	7	2	7	2	7	2	7

Exemple

Si on applique la méthode de Gross, on obtient la pseudo-population U^* :

k	2	7	2	7	2	7	2	7	2	7
x_k	3	20	3	20	3	20	3	20	3	20

Exemple

Si on applique la méthode de Gross, on obtient la pseudo-population U^* :

k	2	7	2	7	2	7	2	7	2	7
x_k	3	20	3	20	3	20	3	20	3	20
y_k	1	8	1	8	1	8	1	8	1	8

On la trie selon la variable x avant de réaliser le tirage systématique de taille 2 :

k	2	2	2	2	2	7	7	7	7	7
x_k	3	3	3	3	3	20	20	20	20	20
y_k	1	1	1	1	1	8	8	8	8	8

$\Rightarrow$ on échantillonne toujours les mêmes valeurs.

Tirage multi-degrés

Le Bootstrap peut être facilement appliqué au cas d'un tirage multidegrés où les unités primaires sont sélectionnées **avec remise** : on rééchantillonne parmi les unités primaires uniquement.

Dans le cas où les unités primaires sont sélectionnées sans remise, on obtient une estimation de variance approximativement sans biais si la fraction du sondage du 1er degré est faible. Le Bootstrap permet en fait de capter le 1er terme de variance donné par l'approche renversée (Haziza, 2009).

Cas d'un tirage multidegrés général encore peu traité (Funaoka et al., 2006).

Sur le rééchantillonnage en Statistique classique

- **Davison, A.C., and Hinkley, D.V. (1997). *Bootstrap Methods and their application*. Cambridge University Press.**
- Efron, B. (1979). *Bootstrap methods : another look at the Jackknife*. Annals of Statistics, 7, p. 1-26.
- Hampel, F.R., and Ronchetti, E.M., and Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics : The Approach Based on Influence Functions*. New York : Wiley.
- Quenouille, M. (1949). *Approximation tests of correlation in time series*. JRSS B, 11, p. 18-84.
- Shao, J., and Tu, D. (1995). *The jackknife and Bootstrap*. New-York, Springer.
- Shao, J., and Wu, C.F.J. (1989). *A general theory for jackknife variance estimation*. Annals of Statistics, 17, p. 1176-1197.
- Tukey, J. (1958). *Bias and confidence in not quite large samples*. Annals of Mathematical Statistics, 29, p. 614.

Sur le Bootstrap en Sondage

- Bickel, P.J., and Freedman, D.A. (1981). *Asymptotic normality and the bootstrap in stratified sampling*. Annals of Statistics, 12, 470-482.
- Booth, J.G., and Butler, R.W., and Hall, P. (1994). *Bootstrap Methods for Finite Populations*. JASA, 89, p. 1282-1289.
- Chao, H., and Lo, K.Y. (1985). *A Bootstrap Method for Finite Populations*. Sankhya, 47, p. 399-405.
- Chauvet, G. (2007). *Méthodes de Bootstrap en population finie*. PhD Thesis, université de Rennes 2.
- Kovar J.G., and Rao J.N.K., and Wu C.F.J (1988). *Bootstrap and other methods to measure errors in survey estimates*. Canadian Journal of Statistics, 16, p. 25-45.
- Mc Carthy, P.J., and Snowden, C.B. (1985). *The Bootstrap and finite population sampling*. Public Health Service Publication 1369.
- Gross, S.T. (1980). *Median estimation in sample surveys*. Proceedings of the Survey Research Methods Section, American Statistical Association, p. 181-184.
- Rao, J.N.K., and Wu, C.F.J. (1988). *Resampling inference with complex survey data*. JASA, 83, p. 231-241.
- Rao, J.N.K., and Wu, C.F.J., and Yue, K. (1992). *Some recent work on resampling methods for complex surveys*. Survey Methodology, 18, p. 209-217.
- Sitter, R.R. (1992). *A resampling procedure for complex survey data*. JASA, 87, p. 755-765.

Sur le rééchantillonnage en Sondage

- Berger, Y.G. (2007). *A Jackknife Variance Estimator for Unistage Stratified Samples with Unequal Probabilities*, Biometrika, 94, p. 953-964.
- Berger Y. G., and Skinner C. J. (2005). *A jackknife variance estimator for unequal probability sampling*. JRSS B, 67, p. 79-89.
- Campbell C. (1980). *A different view of finite population estimation.* Proceedings of the Survey Research Methods Section, American Statistal Association, p. 319-324.
- **Davison, A.C., and Sardy, S. (2006). *Méthodes de rééchantillonnage pour l'estimation de variance en sondages*. Journal de la SFDS, 147, p. 3-32.**
- Krewski D., and Rao, J.N.K. (1981). *Inference from stratified samples : properties of the linearization, jackknife and balanced repeated replication methods*. Annals of Statistics, 9, p. 1010-1019.
- Mc Carthy, P.J. (1969). *Pseudo-replication : Half-samples*. International Statistical Review, 37, p. 239-264.
- Rao, J.N.K., and Shao, J. (1996). *On balanced half-sample variance estimation in stratified sampling*. JASA.
- Rao, J.N.K., and Wu, C.F.J. (1985). *Inference from stratified samples : second-order analysis of three methods for nonlinear statistics*. JASA, 80, p. 620-630.

Sur l'estimation de variance en Sondage

- Berger, Y.G. (1996), *Asymptotic Variance for Sequential Sampling without Replacement with Unequal Probabilities*, Survey Methodology, 22.
- Berger, Y.G. (1998), *Variance Estimation Using List Sequential Scheme for Unequal Probability Sampling*, Journal of Official Statistics, 14.
- Caron, N. (1998). *Le logiciel POULPE : aspects méthodologiques*, Actes des Journées des Méthodologie Statistique, Insee.
- Deville, J-C. (1999), *Variance estimation for complex statistics and estimators : linearization and residual techniques*, Survey Methodology, 25.
- Haziza, D., and Beaumont, J-F. (2005). *Estimation de la variance dans le cas d'échantillonnage à deux phases*. Actes du colloque francophone sur les Sondages, Québec.
- Kovacevic, M. S., and Binder, D.A. (1997). *Variance estimation for measures of income inequality and polarization*. Journal of Official Statistics, 13, p. 41-58.
- Matei, A, Tillé, Y. (2005), *Evaluation of variance approximations and estimators in unequal probability sampling with maximum entropy*, Journal of Official Statistics, 21.
- Petit, J-N. (1998), *Le logiciel POULPE : modélisation informatique*, Actes des Journées des Méthodologie Statistique, Insee.
- Shao, J., Steel, P. (1999). *Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions*. JASA 94, p. 254-265.

Sur les Sondages

- **Ardilly, P. (2006). *Les Techniques de Sondage*. Technip.**
- Cochran, W.G. (1977). *Sampling Techniques*. Wiley, New-York.
- Deville, J-C., and Tillé, Y. (1998). *Unequal probability sampling without replacement through a splitting method*. Biometrika, 85, p. 89-101.
- Deville, J-C., and Tillé, Y. (2004). *Efficient balanced sampling : the cube method*. Biometrika, 128, p. 569-591.
- Goga, C., Deville, J-C., and Ruiz-Gazen, A. (2009). *Use of functionals in linearization and composite estimation with application to two-sample survey data*. Biometrika, 96, P. 691-710.
- Hájek, J. (1981). *Sampling from a finite population*, New-York, Marcel Dekker.
- **Haziza, D. (2009). *Imputation and inference in the presence of missing data*. Handbook of Statistics, Volume 29, Sample Surveys : Theory Methods and Inference, Editors : C.R. Rao and D. Pfeffermann.**
- Madow, L.H, and Madow, W.G. (1944). *On the theory of systematic sampling, II*. Annals of Mathematical Statistics, 20, p. 333-354.