

# Estimation de variance pour un échantillon imparfaitement équilibré

Jay Breidt (Colorado State University)  
Guillaume Chauvet (ENSAI)

Colloque sur les méthodes de sondages en l'honneur de Jean-Claude  
Deville  
24-26/06/2009

## Résumé

Une information auxiliaire disponible au stade de l'échantillonnage peut être utilisée pour sélectionner un échantillon équilibré, afin de limiter la variance.

La *méthode du Cube* permet de sélectionner des échantillons équilibrés, ou approximativement équilibrés quand l'équilibrage exact est impossible. La variance est sous-estimée si le défaut d'équilibrage n'est pas pris en compte dans le calcul de précision.

Nous proposons une méthode permettant de prendre en compte ce défaut d'équilibrage. L'estimateur de variance proposé est non biaisé, mais généralement plus instable que l'estimateur simplifié de Deville et Tillé (2005).

- 1 La méthode du Cube
- 2 Variance d'un échantillon équilibré
- 3 Etude par simulations

## Contexte

Un échantillon aléatoire  $S$  est sélectionné dans une population finie  $U$  d'individus selon un plan de sondage  $p(\cdot)$ , avec des probabilités d'inclusion  $\pi_k$ ,  $k \in U$ .

La variable d'intérêt  $y$  prend la valeur  $y_k$  sur l'individu  $k$  de la population. L'objectif est d'estimer le total  $t_y = \sum_{k \in U} y_k$ .

On suppose que l'ensemble de l'échantillon  $S$  est effectivement observé (absence de non-réponse). Un estimateur sans biais sous le plan de sondage est donné par le  $\pi$ -estimateur

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

où  $d_k = 1/\pi_k$  désigne le poids de sondage de l'individu  $k$ .

## Redressement sur information auxiliaire

Après échantillonnage, le  $\pi$ -estimateur peut être amélioré par redressement si on dispose d'une information auxiliaire.

Soit  $\mathbf{x}_k = (x_{1k}, \dots, x_{qk})'$  un vecteur de variables auxiliaires. Si on connaît (au moins) les  $\{\mathbf{x}_k\}_{k \in S}$  et le vecteur des totaux  $t_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$ , on peut utiliser l'estimateur par la régression généralisée (GREG) :

$$\hat{t}_{y,greg} = \hat{t}_{y\pi} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi})' \hat{\mathbf{B}},$$

avec

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

## Redressement sur information auxiliaire

L'estimateur GREG est **calé**, au sens où il restitue exactement les totaux connus :

$$\hat{t}_{\mathbf{x},greg} = t_{\mathbf{x}}.$$

Plus généralement, Deville et Särndal (1992) proposent la classe des estimateurs par calage.

- On minimise une fonction de distance entre anciens poids  $\{d_k\}_{k \in S}$  et nouveaux poids  $\{w_k\}_{k \in S}$ ,
- On assure de respecter parfaitement les équations de calage

$$\sum_{k \in S} w_k \mathbf{x}_k = t_{\mathbf{x}}.$$

L'estimateur GREG  $\hat{t}_{y,greg}$  est un cas particulier de l'estimateur par calage  $\hat{t}_{yw}$ , obtenu avec une fonction de distance de type Chi-deux.

## Redressement sur information auxiliaire

Si l'information auxiliaire est disponible au stade de l'échantillonnage, on peut l'utiliser pour contraindre la sélection de l'échantillon  $S$ . On dira que l'échantillon est équilibré sur les variables  $x$  si

$$\hat{t}_{x\pi} = t_x. \quad (1)$$

Par extension, un plan de sondage est dit équilibré sur les variables  $x$  si seuls les échantillons équilibrés ont une probabilité non nulle d'être sélectionnés (calage au niveau du plan de sondage).

Comme il est généralement impossible de sélectionner un échantillon exactement équilibré, le but est de définir un plan de sondage assurant que (1) soit approximativement respecté.

# La méthode du Cube

## Représentation du Cube

Deville et Tillé (2004) proposent un algorithme général pour la sélection d'échantillons équilibrés sur un nombre quelconque de variables, avec des probabilités d'inclusion  $\pi = (\pi_1, \dots, \pi_N)'$  quelconques : la méthode du Cube.

Un échantillon  $s$  est vu comme un sommet  $(s_1, \dots, s_N) \in \{0, 1\}^N$  du  $N$ -cube  $C = [0, 1]^N$ .

Les équations d'équilibrage définissent l'espace des contraintes :

$$\pi + Ker(A) \text{ où } A = (\mathbf{x}_k / \pi_k)_{k \in U}$$

L'algorithme consiste à arrondir aléatoirement des composantes du vecteur  $\pi$  par une marche aléatoire dans l'espace des contraintes.

## Etape de base : la martingale équilibrante

On initialise avec  $\pi^{(0)} = \pi$ .

A l'étape  $t$ ,  $\pi^{(t)} = \pi^{(t-1)} + \delta^{(t)}$ , avec

$$\delta^{(t)} = \begin{cases} +\lambda_1(t) u(t) & \text{avec la probabilité } \lambda_2(t)/(\lambda_1(t) + \lambda_2(t)) \\ -\lambda_2(t) u(t) & \text{avec la probabilité } \lambda_1(t)/(\lambda_1(t) + \lambda_2(t)) \end{cases},$$

où

- $\lambda_1(t), \lambda_2(t) > 0$   
→ choisis afin qu'au moins une unité soit sélectionnée ou définitivement rejetée.
- $u(t) \in \text{Ker}(A)$  est un vecteur (non aléatoire)  
→ assure que les équations d'équilibrage sont exactement respectées.
- le choix aléatoire assure que les probabilités d'inclusion sont exactement respectées.

## Fin de l'échantillonnage

L'algorithme précédent s'arrête quand il n'est plus possible de trouver un vecteur  $u(t)$  respectant les contraintes précédentes : c'est la fin de la **phase de vol**.

La **phase d'atterrissage** permet de terminer l'échantillonnage pour les unités restantes (au plus  $q$ ). Plusieurs options sont possibles (Deville et Tillé, 2004), on suppose ici que les contraintes d'équilibrage sont relâchées successivement.

Le vecteur  $\pi(T)$  obtenu à la dernière étape de l'algorithme donne le résultat de l'échantillonnage.

# Applications de la méthode du Cube

- Sélection des Unités Primaires de l'Echantillon Maître (Bourdalle et al., 2000),
- Sélection des échantillons du Nouveau Recensement (Godinot, 2004),
- Imputation aléatoire équilibrée afin de limiter la variance d'imputation (Deville 2006, Chauvet, Deville et Haziza 2009).

# Variance d'un échantillon équilibré

## Estimateur de Yates-Grundy

On suppose que la variable  $\pi$  appartient au vecteur  $\mathbf{x}$  de variables d'équilibrage (échantillonnage de taille fixe). La variance du  $\pi$ -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  $\pi_{kl}$  sont  $> 0$ , avec  $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ .

Problème :  $\pi_{kl}$  difficiles à calculer exactement.

## Estimation de la matrice de covariance

Une première estimation de la matrice  $\Delta = [\Delta_{kl}]$  de variance-covariance peut être obtenue par simulations directes. On sélectionne  $C$  échantillons indépendants, et on estime  $\Delta$  par

$$\Delta_{sim} = \frac{1}{C} \sum_{c=1}^C (I(S_c) - \pi)(I(S_c) - \pi)',$$

où  $I(S_c) = (I(1 \in S_c), \dots, I(N \in S_c))'$  est le vecteur des indicatrices d'appartenance à l'échantillon.

Une autre possibilité consiste à s'appuyer sur l'algorithme d'échantillonnage utilisé.

## Estimation de la matrice de covariance

Notons que la matrice de variance-covariance  $\Delta$  est donnée par

$$\Delta = V(\pi(T)),$$

où  $\pi(T)$  donne le résultat de l'algorithme d'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

$$\begin{aligned} V(\pi(T)) &= \sum_{t=1}^T V(\delta^{(t)}) \\ &= E[\sum_{t=1}^T V(\delta^{(t)} | \mathcal{F}_{t-1})] \\ &= E[\sum_{t=1}^T \lambda_1(t) \lambda_2(t) u(t) u'(t)]. \end{aligned}$$

## Estimation de la matrice de covariance

La matrice de variance-covariance peut donc être estimée sans biais par

$$\hat{\Delta}(s) = \sum_{t=1}^T \lambda_1(t)\lambda_2(t)u(t)u'(t), \quad (2)$$

et une seconde approximation par simulations est obtenue à l'aide de  $C$  échantillons indépendants :

$$\Delta_{app} = \frac{1}{C} \sum_{c=1}^C \hat{\Delta}(S_c). \quad (3)$$

## Comparaison entre les deux méthodes

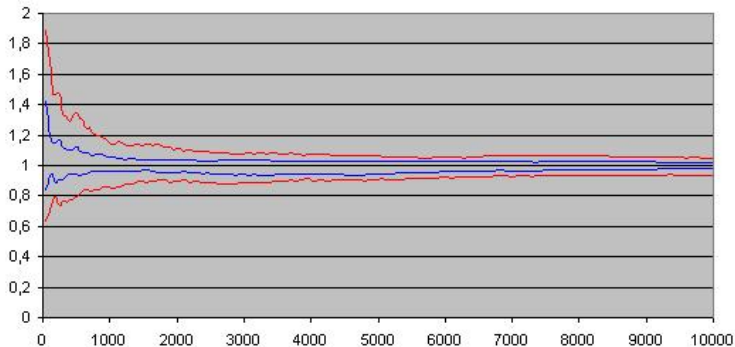
La méthode proposée permet-elle une convergence plus rapide de la matrice de variance-covariance estimée ?

On compare les deux méthodes dans le cas d'une petite population de taille  $N = 10$ , en équilibrant sur la taille d'échantillon ( $x_{k1} = \pi_k$ ) et la taille de la population ( $x_{k2} = 1$ ). Le plan de sondage associé à la méthode du Cube et la matrice  $\Delta$  peuvent être ici calculés exactement.

Dans la décomposition  $\Delta_{app} \mathbf{z} = \lambda \Delta \mathbf{z}$ , les valeurs propres  $\lambda_{max}, \lambda_{min} > 0$  représentent la situation où l'approximation de variance est la pire (Deville et Tillé, 2005).

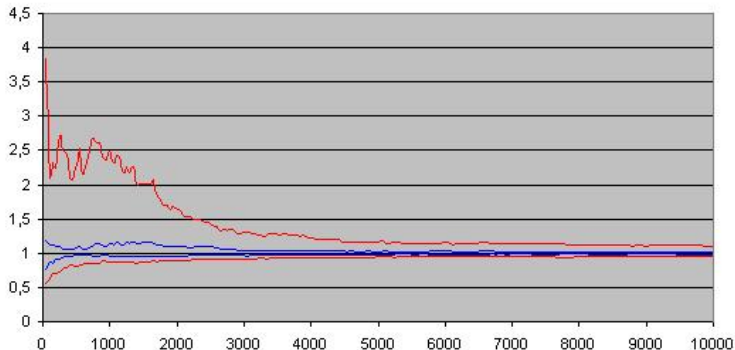
## Résultats obtenus

Convergence de la matrice de variance-covariance estimée par simulations pour un échantillon de taille  $n=3$



## Résultats obtenus

Convergence de la matrice de variance-covariance estimée par simulations pour un échantillon de taille  $n=5$



## Estimateur de variance proposé

Un nouvel estimateur de variance est donné par

$$v_{MD} = -\frac{1}{2} \sum_{k \neq l \in S} \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \frac{\Delta_{app,kl}}{\pi_{app,kl}}, \quad (4)$$

en remplaçant dans l'estimateur de Yates-Grundy

- $\Delta_{kl}$  par  $\Delta_{app,kl}$ ,
- $\pi_{kl}$  par  $\pi_{app,kl} = \Delta_{app,kl} + \pi_k \pi_l$ .

## Approximation de Deville et Tillé

Pour un plan de sondage exactement équilibré et à entropie maximale, Deville et Tillé (2005) proposent l'approximation de variance

$$V_{app}(\hat{t}_{y\pi}) = \frac{N}{N-q} \sum_{k \in U} \pi_k (1 - \pi_k) \left( \frac{y_k}{\pi_k} - \frac{y_k^*}{\pi_k} \right)^2, \quad (5)$$

où  $y_k^*$  donne une prédiction de  $y_k$  obtenue avec les  $q$  variables d'équilibrage  $\mathbf{x}_k$ .

On obtient par substitution l'estimateur de variance

$$v_{DT} = \frac{n}{n-q} \sum_{k \in S} (1 - \pi_k) \left( \frac{y_k}{\pi_k} - \frac{\tilde{y}_k^*}{\pi_k} \right)^2. \quad (6)$$

# Etude par simulations

Nous adaptons une des études par simulations proposées par Deville and Tillé (2005).

On considère une population  $U$  de taille  $N = 40$ . Le plan considéré est  $n = 15$ ,  $q = 4$ ,  $x_{k1} = \pi_k$ ,  $x_{k2} = k$ ,  $x_{k3} = 1/k$  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  $x_2, \dots, x_4$  sont centrées et réduites.

Il s'agit d'un plan imparfaitement équilibré.

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

$$y_{ik} = \beta_1 + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \sigma_i \epsilon_k, \quad (7)$$

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

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

Le coefficient  $\sigma_i$  est choisi pour donner un  $R^2$  approximativement égal à 0.1 pour  $y_1$ , 0.2 pour  $y_2$ , ...

## 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

Var.	Méthode	Taux de couverture			% bias	Stab.
		5 %				
		<i>L</i>	<i>U</i>	<i>L + U</i>		
<i>y</i> <sub>1</sub>	MD	4,35	5,65	10,00	0,2	33,8
	DT	4,90	5,15	10,05	-0,8	29,3
<i>y</i> <sub>2</sub>	MD	3,95	6,65	10,60	0,0	44,3
	DT	6,00	5,80	11,80	-11,5	28,6
<i>y</i> <sub>3</sub>	MD	3,25	7,65	10,90	-0,2	57,4
	DT	7,70	6,85	14,55	-22,8	32,2
<i>y</i> <sub>4</sub>	MD	2,65	8,80	11,45	-0,4	71,7
	DT	9,25	7,95	17,20	-34,2	39,3
<i>y</i> <sub>5</sub>	MD	2,15	10,90	13,05	-0,6	86,5
	DT	11,40	9,75	21,15	-45,7	48,4

## Conclusion et perspectives

Travail préliminaire

Modifier l'estimateur de variance pour limiter l'instabilité ?

Comparer avec d'autres estimateurs de variance prenant en compte les deux composantes de la variance.