L'échantillonnage équilibré par la méthode du cube et la méthode réjective

Ibrahima Ousmane Ida et Louis-Paul Rivest

Université Laval

12 Octobre 2016

Plan

1. Contexte.
2. Qu'est-ce que l'échantillonnage équilibré?
3. La méthode du cube : une technique exacte.
4. Sélection par la méthode du cube : cas de l'enquête sur la pêche sportive en Gaspésie.
5. Quel choix entre la méthode du cube et la méthode réjective?
6. Conclusion.

Contexte

- La pêche sportive, une activité qui contribue à l'essor des économies locales et nationale.
- Intérêt des enquêtes sur la pêche sportive.
- Difficultés des enquêtes sur la pêche sportive.
- Préoccupation : comment intégrer des contraintes dans le processus de sélection d'échantillon?

Qu'est-ce que l'échantillonnage équilibré?

Qu'est-ce que l'échantillonnage équilibré?

Notation

- Si x est une variable auxiliaire dont les valeurs sont connues pour toutes les unités de la population, le total t_x de cette variable est :

$$t_x = \sum_{i=1}^{N} x_i$$

- Noté $\widehat{t}_{xHT}$, l'estimateur de Horvitz-Thompson du total de x est calculé :

$$\widehat{t}_{xHT} = \sum_{i \in s} \frac{x_i}{\pi_i}$$

Qu'est-ce que l'échantillonnage équilibré?

Qu'est-ce qu'un plan de sondage équilibré?

- Un plan de sondage $p(s)$ est dit équilibré sur la variable x si et seulement si il permet de sélectionner d'un échantillon tel que :

$$\widehat{t}_{xHT} = t_x$$

$$\sum_{i \in s} \frac{x_i}{\pi_i} = \sum_{i=1}^{N} x_i$$

- Objectif visé à travers l'équilibrage : tirer un échantillon dans le sous-espace des contraintes :

$$\mathcal{Q} = \left\{ s \in \mathcal{S} \,\middle|\, \sum_{i \in s} \frac{x_i}{\pi_i} = \sum_{i=1}^{N} x_i \right\}$$

Qu'est-ce que l'échantillonnage équilibré?

Quelques particularités d'échantillonnage équilibré

- L'équilibrage d'un plan peut se faire sur des variables quantitatives et aussi sur des variables qualitatives.

- C'est pourquoi on dit que l'échantillonnage équilibré est une généralisation de la stratification.

- Équilibrer un plan sur des variables x_h $(h = 1, \cdots, H)$ revient à faire de la stratification, si :

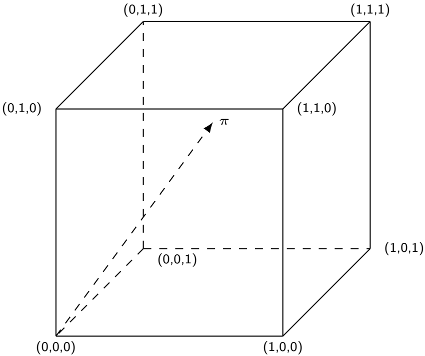
$$x_{hi} = \begin{cases} 1 & \text{si } i \in U_h \\ 0 & \text{si } i \notin U_h \end{cases} \qquad i = 1, \cdots, N$$

où U_h $(h = 1, \cdots, H)$ désignent des strates.

Méthode du cube : une technique exacte

Pourquoi l'appellation méthode du cube?

- C'est à cause de la représentation géométrique du plan de sondage. Par exemple, avec un tirage sans remise dans une population de taille $N = 3$, les échantillons possibles sont : $(0, 0, 0), \cdots, (1, 1, 1)$.



- Elle a été proposée par Deville et Tillé (2004).

Méthode du cube : une technique exacte

Sur quoi se base-t-elle?

- Si un plan de sondage $p(s)$ est équilibré sur des variables $x_1, \cdots, x_p$, alors on a :

$$\widehat{T}_{xHT} = T_x$$

$$\begin{pmatrix} \frac{x_{11}}{\pi_1} & \cdots & \frac{x_{1i}}{\pi_i} & \cdots & \frac{x_{1N}}{\pi_N} \\ \vdots & & \vdots & & \vdots \\ \frac{x_{p1}}{\pi_1} & \cdots & \frac{x_{pi}}{\pi_i} & \cdots & \frac{x_{pN}}{\pi_N} \end{pmatrix} \begin{pmatrix} Z_1 \\ \vdots \\ Z_N \end{pmatrix} = \begin{pmatrix} t_{x_1} \\ \vdots \\ t_{x_p} \end{pmatrix}$$

  où $(Z_1, \cdots, Z_N)' = s$ représente l'échantillon et Z_i des v.a. telles que $Z_i = 0$ ou 1.

- Le sous-espace des contraintes peut s'écrire :

$$\mathcal{Q} = \{ s \in \mathcal{S} \,|\, A * s = T_x \}$$

- A est la matrice des contraintes d'équilibrage.

Méthode du cube : une technique exacte

Comment fonctionne-t-elle?

- Pour tirer un échantillon s de $\mathcal{S}$ $\left(\{0;1\}^N\right)$, elle opère dans l'hypercube $\mathcal{C}$ $\left([0,1]^N\right)$ et transforme le vecteur π des probabilités de sélection :

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

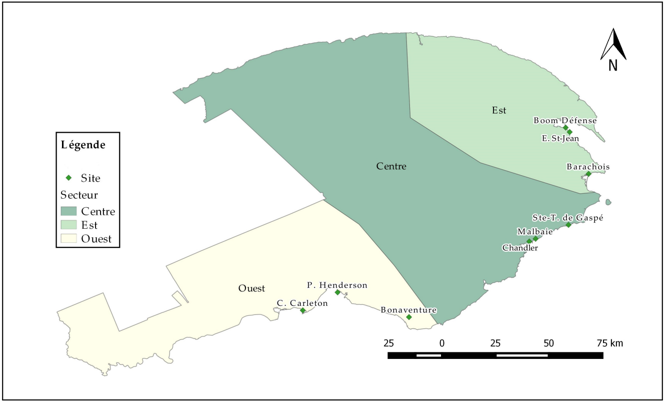
où $\delta(t)$ sont des vecteurs propres de la matrice A des contraintes d'équilibrage.

- Elle admet deux phases :
  1. Phase de vol pour chercher un échantillon exactement équilibré.
  2. Phase d'atterrissage pour trouver un échantillon approximativement équilibré. L'atterrissage se fait par progammation linéaire ou par suppression de variables.

Sélection par la méthode du cube : cas de l'enquête sur la pêche sportive en Gaspésie.

Sélection par la méthode du cube : cas de l'enquête sur la pêche sportive en Gaspésie

Zone de l'enquête



Sélection par la méthode du cube : cas de l'enquête sur la pêche sportive en Gaspésie

Plan d'échantillonnage

- Il s'agit d'un plan stratifié à trois degrés :
  1. Stratification : $h = 1, \ldots, H$ représentent les jours et sont des strates;
  2. Degré 1 : $i = 1, \ldots, I$ sont les secteurs;
  3. Degré 2 : $j = 1, \ldots, J$ sont les périodes;
  4. Stratification : $k = 1, \ldots, K$ représentent les sous-périodes et sont aussi des strates;
  5. Degré 3 : $\ell = 1, \ldots, L$ sont les sites.

- Probabilités de sélection aux différents degrés :

$$\pi_{hi}^{(1)} = n_1 \frac{x_{i\bullet}}{x_{\bullet\bullet}}, \quad \pi_{j|i}^{(2)} = \frac{n_{2|1}}{J}, \quad \pi_{\ell|ijk}^{(3)} = n_{3|1,2} \frac{x_{i\ell}}{x_{i\bullet}}.$$

Sélection par la méthode du cube : cas de l'enquête sur la pêche sportive en Gaspésie

Contraintes d'équilibrage

- Les contraintes opérationnelles veulent que :
  1. les visites de chaque secteur soient équitablement réparties entre les périodes;
  2. les visites de chaque site soient équitablement réparties entre les sous-périodes;
  3. etc.
- Au total 77 contraintes ont été imposées au plan.
- Spécification des contraintes concernant les visites des secteurs pendant les différentes périodes :

$$\sum_{h,i,j} Z_{hi} Z_{j|hi} I_{hij}^{(\beta)} = H n_1 \frac{n_{2|1}}{J} = 22$$

Sélection par la méthode du cube : cas de l'enquête sur la pêche sportive en Gaspésie

Résultats après simulation de 10 000 échantillons

Table: Moyennes des visites des sites et espérances des visites

Secteur	Site	$x_{i\ell}$	$E(n_{i\ell})$	$\overline{n}_{i\ell}$	$s_{\overline{n}_{i\ell}}$
East $(i=1)$	Boom Défense $(\ell=1)$	2	20,308	20,286	0,850
	E. St-Jean $(\ell=2)$	1	10,154	10,153	0,621
	Barachois $(\ell=3)$	2	20,308	20,296	0,881
Centre $(i=2)$	Ste-T. de Gaspé $(\ell=4)$	1	10,154	10,176	0,865
	Malbaie $(\ell=5)$	1	10,154	10,155	0,880
	Chandler$(\ell=6)$	1	10,154	10,162	0,881
West $(i=3)$	Bonaventure $(\ell=7)$	2	20,308	20,311	1,004
	P. Henderson $(\ell=8)$	1	10,154	10,153	0,681
	C. Carleton $(\ell=9)$	2	20,308	20,309	1,016

Taille d'échantillon : $n=132$.

Quel choix entre la méthode du cube et la méthode réjective?

Quel choix entre la méthode du cube et la méthode réjective?

Plan d'échantillonnage et méthode réjective

- Il s'agit d'un plan stratifié à deux degrés :
  1. Stratification : $h = 1, \ldots, H$ représentent les jours et sont des strates;
  2. Degré 1 : $i = 1, \ldots, I$ sont les secteurs;
  3. Degré 2 : $j = 1, \ldots, J$ sont les sites.

- Au sens de Fuller (2009), ce plan est équilibré si et seulement si il permet de tirer d'un échantillon qui satisfait la condition :

$$\mathcal{Q}_{p,n} = (\overline{n} - E(n_e))' \left[Var\left(\overline{n}\right)\right]^{-1} (\overline{n} - E(n_e)) < \gamma^2$$

  où $\overline{n}$ est le vecteur des visites des sites, $E(n_e)$ vecteur des espérances des visites, γ^2 tolérance de sélection.

- $\mathcal{Q}_{p,n}$ suit une χ^2_p.

Résultats après simulation de 100 000 échantillons

Table: Comparaison des moyennes des visites des sites aux espérances des visites.

		Méthode du Cube				Méthode de Fuller (taux de rejet = 95%)			
Site	$E(n_{ij})$	$\overline{n}_{ij}$	$SE_{\overline{n}_{ij}}$	z	$2Pr > \|z\|$	$\overline{n}_{ij}$	$SE_{\overline{n}_{ij}}$	z	$2Pr > \|z\|$
$i = 1, j = 1$	3	3	0	na	na	3,0105	0,0020	5,2003	0
$i = 1, j = 2$	2	2	0	na	na	1,9543	0,0020	-22,5290	0
$i = 1, j = 3$	3	3	0	na	na	3,0100	0,0020	4,9411	0
$i = 2, j = 4$	2	1,9993	0,0004	-1,6687	0,0952	1,9540	0,0018	-25,2836	0
$i = 2, j = 5$	2	2,0007	0,0006	1,0999	0,2714	1,9563	0,0018	-23,9179	0
$i = 2, j = 6$	2	2,0003	0,0006	0,5978	0,5500	1,9527	0,0018	-25,9863	0
$i = 3, j = 7$	3	2,9975	0,0011	-2,2948	0,0217	3,0350	0,0018	18,9196	0
$i = 3, j = 8$	3	3,0020	0,0012	1,6343	0,1022	3,0365	0,0019	19,6968	0
$i = 3, j = 9$	4	4,0007	0,0010	0,6602	0,5091	4,0913	0,0019	47,0904	0

Taille d'échantillon : $n = 24$.

Conclusion

- La méthode du cube donne des estimations sans biais et conserve les probabilités de sélection.
- La méthode réjective modifie les probabilités de sélection et conduit souvent à des estimations biaisées.
- Dans le contexte d'enquête sur la pêche sportive, la sélection par la méthode du cube semble être une excellente alternative d'échantillonnage.

Références

- Audrey-Anne Vallée, Bastien Ferland-Raymond, Louis-Paul Rivest, and Yves Tillé. Incorporating spatial and operational constraints in the sampling designs for forest inventories. Environmetrics, 26(8):557-570, 2015.
- Caren Hasler and Yves Tillé. Fast balanced sampling for highly stratified population. Computational Statistics and Data Analysis, 74:81-94, 2014.
- Guillaume Chauvet, David Haziza, and Éric Lesage. Examining some aspects of balanced sampling in surveys. Statistica Sinica, 2015.
- Ibrahima Ousmane Ida. L'échantillonnage équilibré par la méthode du cube et la méthode réjective. Mémoire de maîtrise, Université du Québec, 2016. URL http://www.theses.ulaval.ca/2016/32818/32818.pdf.
- Jean-Claude Deville and Yves Tillé. Efficient balanced sampling: the cube method. Biometrika, 91(4):893-912, 2004.
- Wayne A Fuller. Some design properties of a rejective sampling procedure. Biometrika, 96 (4):933-944, 2009.
- Yves Tillé. Sampling algorithms. Springer, New York, 2011.

Merci !