ÉCHANTILLONNAGE ALÉATOIRE CONTRÔLÉ POUR ÉVITER LES ÉCHANTILLONS INDÉSIRABLES

Carlos A. León[1]

RÉSUMÉ

Nous proposons une méthode pour construire des plans d'échantillonnage faiblement équivalents à l'échantillonnage aléatoire simple sans remise mais qui arrivent à exclure certains échantillons jugés indésirables. Nous expliquons comment ces plans s'obtiennent comme solutions d'un problème d'optimisation sous contraintes et nous donnons quelques exemples numériques.

MOTS CLÉS : Échantillonnage contrôlé; EAS; Échantillons indésirables; Optimisation sous contraintes

ABSTRACT

We propose a method for constructing sampling plans weakly equivalent to simple random sampling without replacement that exclude certain bad samples. We show how these plans can be obtained as solutions to a constrained optimization problem and we present some numerical examples.

KEY WORDS: Controlled Sampling, SRS, Bad Samples, Optimization under Constraints

1. INTRODUCTION

1.1 Description du problème

Hedayat and Robieson (1998) puis Chang et. al. (2004) ont montré comment on peut bâtir des plans d'échantillonnage équivalents à l'échantillonnage aléatoire simple sans remise (EAS) mais en excluant une partie de l'espace échantillonal, de manière à éviter certains échantillons indésirables. Leur construction repose sur une partition de l'espace des échantillons de taille n possibles, ce qui permet d'obtenir une famille de plans d'échantillonnage pour lesquels les probabilités d'inclusion d'ordre un et deux se laissent représenter de façon simple à l'aide d'un système d'équations linéaires. En trouvant des solutions de ce système on obtient ainsi des plans faiblement équivalents à l'EAS. Cependant, bien que très intéressantes au point de vue théorique, ces solutions ne sont pas applicables en général. Elles n'évitent pas autant de mauvais échantillons que possible et les contraindre à le faire demande des conditions trop restrictives dans le contexte de la pratique des enquêtes.

Pour obtenir des plans plus flexibles, nous cherchons des solutions du système linéaire initial en introduisant un problème d'optimisation sous contraintes. Ceci nous donne notre problème d'optimisation de base, qui est ensuite étendu pour obtenir une plus grande variété de plans. L'article est organisé comme suit : nous commençons par décrire le contexte mathématique du problème et la solution proposée et nous enchaînons avec quelques résultats numériques afin d'illustrer la procédure. Finalement, nous proposons quelques applications pratiques de cette méthode.

2. ÉCHANTILLONNAGE CONTRÔLÉ

2.1 Notions préliminaires

Soit une population U composé de N unités, à partir de laquelle nous voulons tirer un échantillon de taille n. Supposons de plus que U contient un ensemble U_1 composé d'unités indésirables, le reste de la population étant dans l'ensemble U_2. Les tailles de U_1 et U_2 seront respectivement notées N_1 et N_2. Sans perte de généralité, pour simplifier l'exposé nous ferons l'hypothèse que les unités 1 à N_1 sont dans U_1. Cette partition de U induit une partition de l'ensemble des échantillons

[1] Carlos A. Leon, Statistique Canada, Tunney's Pasture, Ottawa, ON K1A 0T6, carlos.leon@statcan.gc.ca

possibles de taille n. Celle-ci est définie par les classes C_k, $k=0,...,n,$ où C_k est l'ensemble de tous les échantillons de taille n contenant exactement k unités indésirables.

Nous pouvons définir un plan d'échantillonnage de taille n sur la population U, en associant une probabilité commune α_k à chaque élément dans les classes C_k. Pour le moment n'importe quel choix est valable pourvu que la probabilité globale donne 1. Puisque U_1 et U_2 sont respectivement de taille N_1 et N_2, alors le nombre d'éléments que contient C_k est donné par le coefficient hypergéométrique $\binom{N_1}{k}\binom{N_2}{n-k}$. Les probabilités $\alpha_0, \alpha_1,..., \alpha_n$ doivent alors satisfaire l'équation

$$\sum_{k=0}^{n}\binom{N_1}{k}\binom{N_2}{n-k}\alpha_k = 1, \tag{1}$$

Tel que nous devrions nous attendre, on voit que l'EAS correspond à la solution uniforme $\alpha_k = \binom{N}{n}^{-1}$. L'équation (1) définit tous les plans de taille n possibles avec notre schéma de sélection; si nous voulons nous restreindre à ceux ayant des probabilités de sélection d'ordre un et deux identiques à celles de l'EAS, il faut imposer des conditions additionnelles. Ces plans, que nous appellerons faiblement équivalents à l'EAS, sont caractérisés dans le résultat suivant, dont la démonstration a des points communs avec celle dans Hedayat and Robieson (1998) mais est plus simple.

Lemme 1. Un plan de sondage faiblement équivalent à l'EAS sera donné par les nombres $\alpha_0, \alpha_1,..., \alpha_n$ dès qu'ils satisfont le système linéaire

$$\sum_{k=0}^{n}\binom{N_1}{k}\binom{N_2}{n-k}\alpha_k = 1, \tag{2}$$

$$\sum_{k=1}^{n}\binom{N_1-1}{k-1}\binom{N_2}{n-k}\alpha_k = \frac{n}{N}, \tag{3}$$

$$\sum_{k=1}^{n-1}\binom{N_1-1}{k-1}\binom{N_2-1}{n-k-1}\alpha_k = \frac{n(n-1)}{N(N-1)} \tag{4}$$

Démonstration. Il n'est pas difficile de voir que ce plan comporte plusieurs symétries, ce qui fait que les probabilités d'inclusion d'ordre un appartiennent aux catégories $\{\pi_i : i \in U_1\}$ et $\{\pi_i : i \in U_2\}$, à l'intérieur desquelles elles sont constantes. De même, les probabilités d'inclusion d'ordre deux appartiennent aux catégories: $\{ \pi_{ij} : i,j \in U_1\}$, $\{ \pi_{ij} : i,j \in U_2\}$ et $\{ \pi_{ij} : (i,j) \in U_1 \times U_2\}$, où elles sont constantes. Une fois cela établi, on peut voir par un argument combinatoire classique que (3) et (4) caractérisent les probabilités $\{\pi_i : i \in U_1\}$ et $\{ \pi_{ij} : (i,j) \in U_1 \times U_2\}$ qui sont alors les mêmes que pour l'EAS. Il reste à montrer que cela caractérise les probabilités pour toutes les catégories. L'équation (2) implique que nous avons un plan de taille n. Or, comme notre plan est de taille fixe les identités suivantes doivent alors être satisfaites (voir Särndal, Swensson & Wretman (1992), résultat 2.6.2):

$$\begin{aligned} N_1\pi_1 + N_2\pi_N &= n \\ (N_1-1)\pi_{12} + N_2\pi_{1N} &= (n-1)\pi_1 \\ N_1\pi_{1N} + (N_2-1)\pi_{(N-1)N} &= (n-1)\pi_N \end{aligned} \tag{6}$$

Pour déterminer l'ensemble des valeurs dans (6), il suffit de fixer une seule des quantités π_1 , π_N de même qu' une seule parmi les quantités π_{12} , π_{1N} , $\pi_{(N-1)N}$. En remplaçant π_1 et π_{1N} par les valeurs qui apparaissent du côté droit de (3) et (4) on vérifie que toutes les autres probabilités sont conformes à celles de l'EAS ■

Il appert que si les contraintes ne se réduisent pas à des cas triviaux, les solutions de ce problème forment alors un simplexe de dimension $n-2$. On supposera donc toujours $n \geq 2$ ainsi que $N_1 \geq 2$ et $N_2 \geq 2$, même si (2)-(4) admettent une solution dans des cas dégénérés.

En réécrivant nos équations en termes des variables $\beta_k = \binom{N_1}{k}\binom{N_2}{n-k}\alpha_k$, après un peu d'algèbre notre système prend la forme

$$(7) \qquad \sum \beta_k = 1, \quad \sum k\beta_k = \frac{N_1 n}{N}, \quad \sum k(n-k)\beta_k = \frac{N_1 N_2 n(n-1)}{N(N-1)}$$

Si *B* représente le nombre d'unités indésirables dans notre échantillon, alors ceci est équivalent à spécifier des conditions sur les deux premiers moments de la distribution de *B*. Il en découle que dans cette famille de plans, le nombre moyen ainsi que la variance du nombre d'unités indésirables sont toujours fixes : $E(B)= N_1 n / N$, $V(B)=N_1N_2 n(N-n) / N^2(N-1)$.

2.2 Échantillonnage contrôlé

Pour réduire le nombre d'unités indésirables qu'il est possible d'obtenir avec le plan $\beta_0, \beta_1, ..., \beta_n$, il suffit de faire en sorte que l'on ait $\beta_k=0$ à partir d'un certain rang. Au lieu de chercher à imposer ce type de contrainte au problème initial, nous voulons sélectionner des plans qui ont la propriété désirée en optimisant une fonction objectif appropriée sous les contraintes (7). Une approche naturelle serait de chercher à minimiser une fonction telle que coût moyen, mais toute fonction quadratique de *B* a une espérance constante dans notre famille de plans. Un candidat possible est la probabilité d'excéder le coût moyen, qui équivaut à minimiser la probabilité $P[B > E(B)]$. Une autre possibilité est de minimiser l'espérance coditionnelle du coût lorsque le coût excède le coût moyen, ce qui est équivalent à minimiser l'espérance conditionnelle $E[B \mid B > E(B)]$. Nos expériences numériques montrent que la fonction suivante est mieux appropriée:

$$(8) \qquad E\left[B \mid B \le E(B)\right] = \sum_{k \le \mu} k\beta_k \Big/ \sum_{k \le \mu} \beta_k \text{, où } \mu = E(B)$$

Nous avons maintenant tous les éléments requis pour énoncer notre problème de base

Déterminer $\beta_0, \beta_1, ..., \beta_n$ qui maximisent $E [B \mid B \le E(B)]$ sous les contraintes (7).

Ce problème peut être généralisé au cas où les probabilités d'ordre un sont arbitraires. De plus, il s'avère que la contrainte au niveau des moments d'ordre deux limite notre capacité à réduire le support de la distribution. Ces considérations nous amènent à formuler un problème un peu plus général

Déterminer une distribution $\beta_0, \beta_1, ..., \beta_n$ qui maximise $E[B \mid B \le E(B)]$sous les contraintes

$$(9) \qquad \sum k\beta_k = N_1\pi_1 \qquad \sum k(n-k)\beta_k \le N_1 N_2 \pi_1 \pi_N \left[1 - \frac{\lambda(N-n)}{n(N-1)}\right]$$

Où le paramètre λ contrôle la dispersion de *B*. Si $\lambda=1$ on a la valeur initiale, alors que $\lambda=0$ correspond à faire en sorte que la variance de *B* soit nulle (lorsque la borne est atteinte).

3. EXEMPLES NUMÉRIQUES

3.1 Implémentation et exemples

La majeure partie du travail consistait à transformer les contraintes initiales et déterminer des fonctions objectif appropriées. Cette recherche s'est faite en s'aidant de simulations numériques. La méthode a été implémentée à l'aide de la procédure OPTmodel qui fait partie de la version 9.1 du logiciel SAS®. Pour ce faire, nous avons écrit une macro SAS qui est disponible sur demande.

EXEMPLE 1.

Ici nous prenons une population de taille N= 2397 avec N_1=300 unités indésirables et un échantillon de taille $n=100$. Nous posons λ=1 dans notre algorithme, ce qui correspond à spécifier que les probabilités d'inclusion d'ordre deux doivent être exactement celles qu'on a pour l'EAS. La deuxième colonne du tableau donne la solution optimale et la troisième donne les probabilités d'obtenir k unités indésirables sous le plan aléatoire simple. La solution optimale donne des valeurs β_k =0 pour toutes les valeurs de k qui n'apparaissent pas dans le tableau.

Tableau 1

k	β_k optimaux	β_k EAS
...	...	...
9	0	0.07385
10	0	0.09744
11	0	0.11514
12	0.6589	0.12285
13	0.2804	0.11915
14	0.0594	0.10567
15	0	0.08612
16	0	0.06478
17	0	0.04513
...	...	...
100	0.0013	<1E-10

On voit que la solution optimale réussit à faire en sorte que l'on ait β_k =0 pour la majeure partie de la distribution qui se trouve en dehors d'un voisinage immédiat de la moyenne. Cependant elle n'arrive pas à éliminer tout ce qui se trouve à droite de la moyenne et il subsiste une masse assez importante en k=100, ce qui fait que le plan optimal n'est pas vraiment acceptable.

EXEMPLE 2.

Avec les mêmes valeurs que dans l'exemple précédent nous prenons λ=0.05 dans notre algorithme, ce qui fait que les probabilités d'inclusion d'ordre deux peuvent maintenant être légèrement différentes de celles qu'on a pour l'EAS. Ici aussi la solution optimale donne β_k =0 pour toutes les valeurs de k qui n'apparaissent pas dans le tableau.

Tableau 2

k	β_k optimaux	β_k EAS
...	...	...
9	0	0.07385
10	0	0.09744
11	0	0.11514
12	0.6021	0.12285
13	0.2805	0.11915
14	0.1174	0.10567
15	0	0.08612
16	0	0.06478
17	0	0.04513
...	...	...

Maintenant la solution optimale réalise $\beta_k = 0$ pour toute la distribution qui se trouve en dehors d'un voisinage immédiat de la moyenne. Le prix à payer est que les probabilités d'inclusion d'ordre deux sont légèrement différentes de celles prescrites initialement.

Tableau 3

	π_1	π_N	π_{12}	π_{1N}	$\pi_{(N-1)N}$
Plan optimal	0.0417	0.0417	0.001612	0.001739	0.001721
EAS	0.0417	0.0417	0.001723	0.001723	0.001723

3. CONCLUSION

La méthode proposée permet de contrôler efficacement l'incidence d'échantillons contenant trop d'unités indésirables et, contrairement à d'autres approches telle la méthode du cube, est relativement simple à implémenter. Plusieurs applications sont envisageables. Par exemple, dans des contextes où les coûts d'interviewer certaines unités sont très grands comparés au coût pour les autres unités, mais qu'au vu des caractéristiques d'intérêt la population (la strate) est homogène. Une autre application du même genre pourrait être dans le cas où les unités ont des propensités à répondre particulièrement faibles ou pour lesquelles l'information auxiliaire disponible sur la base de sondage est de faible qualité.

.

RÉFÉRENCES

Chang, H. J., Wang, C. L. and Huang, K. C. (2004). «Simple random sample equivalent survey designs reducing undesirable units from a finitepopulation». *Statistical papers* **45**, 287-295.

Hedayat, A. S. and Robieson, W. Z. (1998). «Exclusion of an undesirable sample from the support of a simple random sample». The American Statistician **52**(1), 41-43.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992), Model Assisted Survey Sampling, Springer, NY