L'échantillonnage équilibré

Laurent Costa & Thomas Merly-Alpa

***Résumé* — L'objectif de cette note méthodologique est de décrire l'échantillonnage équilibré et ses contextes d'application. La méthode du cube et sa mise en oeuvre sont également abordées ainsi que brièvement l'estimation de variance.**

I. Introduction

Lors de la constitution d'un échantillon, la question du plan de sondage et de son efficacité doit se poser. Le but est d'obtenir un échantillon qui reflète au mieux possible l'hétérogénéité de la population sondée en réduisant la variance des estimateurs et en respectant les coûts. Les plans de sondage les plus connus pour réduire la variance sont les plans stratifiés et le tirage à probabilités inégales. Cependant, en pratique, il n'est pas toujours souhaitable de procéder à une stratification si n est faible ou si on ne veut pas calculer les allocations pour des problèmes d'arrondis par exemple.

L'idée du sondage équilibré repose sur l'utilisation d'informations disponibles et corrélées avec la variable d'intérêt dans l'élaboration du plan. La précision d'un plan de sondage repose sur des propriétés d'équilibrage : l'échantillon est sélectionné de façon à respecter une information connue. Par exemple :

- respect de structure âge-sexe ;
- répartition par effectif salarié.

Lorsqu'un échantillon sélectionné restitue exactement les informations disponibles conformément à ce qu'on retrouve dans la population, alors il restituera bien l'information sur la variable d'intérêt grâce à la corrélation entre les deux types d'information. C'est ce qui explique la capacité du plan de sondage équilibré à améliorer l'efficacité des estimateurs.

Malgré la difficulté d'appliquer une méthode générale de manière algorithmique qui respecte à la fois toutes les contraintes d'équilibrage ainsi qu'une sélection aléatoire de l'échantillon [1], nous verrons que la méthode du CUBE, développée par Deville et Tillé en 2004, permet de tirer des échantillons approximativement équilibrés.

II. Définition d'un tirage équilibré

Un échantillon est dit équilibré sur une ou plusieurs variables disponibles dans la base de sondage, lorsque pour chacune d'entre elles, l'estimateur Horvitz-Thompson du total coïncide exactement avec le vrai total issu de la base de sondage.

On rappelle la définition de l'estimateur sans biais d'Horvitz-Thompson du total d'une variable x noté $\hat{t}_{x\pi}$ sur un échantillon S :

$$\hat{t}_{x\pi} = \sum_{i \in S} \frac{x_i}{\pi_i}$$

où π_i est la probabilité d'inclusion de l'individu i dans l'échantillon S.

Un échantillon S d'une population U équilibré sur la variable de contrôle x respecte donc la contrainte suivante :

$$\sum_{i \in S} \frac{x_i}{\pi_i} = \sum_{i \in U} x_i \quad \text{soit} \quad \hat{\mathrm{t}}_{\mathrm{x}\pi} = \mathrm{t}_{\mathrm{x}}$$

Il s'agit donc d'une sorte de calage au niveau plan de sondage sur les variables auxiliaires. Par construction, l'estimateur du total de x est sans biais et de variance nulle.
Etudions le cas d'un modèle de travail du type :

$$y_i = \beta x_i + \varepsilon_i$$

que l'on peut réécrire en divisant par π_i puis en sommant sur chaque individu i sous la forme :

$$\hat{t}_{y\pi} = \beta \hat{t}_{x\pi} + \hat{t}_{\varepsilon\pi}$$

Comme on équilibre le plan de sondage sur la variable x corrélée à y, son total est parfaitement estimé [2], on a donc :

$$\hat{t}_{y\pi} = \beta t_x + \hat{t}_{\varepsilon\pi}$$

Et on obtient [3] :

$$V(\hat{t}_{y\pi}) = V(\hat{t}_{\varepsilon\pi})$$

Ainsi, on voit que :

- Le respect des probabilités d'inclusion permet d'obtenir une estimation sans biais $\rightarrow E(\hat{t_{y\pi}})=t_y$;
- La restriction du support du plan de sondage aux échantillons équilibrés permet d'annuler la variabilité du premier terme en x ;
- La variance n'est plus donnée que par les résidus du modèle.

On peut également déduire quelques propriétés :
Supposons que $x_i=\pi_i$, c'est-à-dire que l'on équilibre sur les probabilités d'inclusion.
L'équation d'équilibrage implique que

$$\hat{t}_{x\pi} = \sum_{i \in S} \frac{x_i}{\pi_i} = \sum_{i \in S} \frac{\pi_i}{\pi_i} = n(s)$$

$$et \quad t_x = \sum_{i \in U} \pi_i = E(n(S))$$

1. L'équilibrage pourrait s'avérer tellement contraint qu'il conduirait à un choix déteministe. Or la sélection doit demeurer aléatoire pour que les propriétés statistiques de biais et de variance d'échantillonnage conservent leur sens et pour respecter les probabilités d'inclusion.

2. La variance de $\hat{t}_{x\pi}$ est nulle.

3. L'expression βt_x étant constante.

Or $\hat{t}_{x\pi}$=t_x d'où

$$n(s) = E(n(S))$$

Le plan de sondage est donc de taille fixe.

Supposons que x_i=1, c'est-à-dire que l'on équilibre sur la variable constante à 1.
L'équation d'équilibrage implique que

$$\sum_{i\in S} \frac{x_i}{\pi_i} = \sum_{i\in S} \frac{1}{\pi_i} = \hat{N}_\pi = \sum_{i\in U} 1 = N$$

La taille de la population est donc parfaitement estimée.

On pourrait également raisonner en plan stratifié avec allocation proportionnelle et prouver que l'on estime exactement la taille des strates en utilisant comme variables d'équilibrage les indicatrices d'appartenance aux strates. L'échantillonnage équilibré est donc une méthode de tirage probabiliste qui assure qu'in fine, une fois l'échantillon tiré, les proportions d'individus respectant chacune des modalités, respectivement dans la population et dans l'échantillon, seront égales.

Au cours d'une enquête, on observe généralement un phénomène de non-réponse sur l'échantillon qui déstabilise l'équilibrage. Il est donc particulièrement intéressant pour un premier degré de tirage ou quand on anticipe une faible non-réponse. On peut citer par exemple [4] :

- Tirage des Unités Primaires de l'Echantillon Maître ;
- Tirage des Groupes de Rotations du Recensement.

III. La méthode du Cube

A. Principe

L'algorithme proposé par Deville et Tillé (2004) [5] a un cadre général et permet la sélection d'échantillons équilibrés sur un nombre quelconque de variables, avec un jeu de probabilités d'inclusion $\boldsymbol{\pi}$=$(\pi_1,...,\pi_N)$ quelconque. Un échantillon s est vu comme un sommet $(s_1,...,s_N) \in \{0,1\}^N$ du N-cube C=$[0,1]^N$. L'algorithme consiste en une marche aléatoire pour passer du vecteur des probabilités d'inclusion $\boldsymbol{\pi}$ au vecteur des indicatrices de sélections $\mathbf{I}$ en arrondissant aléatoirement les π_i à 0 ou 1.
On a :

$$\hat{t}_{x\pi} = \sum_{i\in U} \frac{x_i}{\pi_i} I_i = t_x = \sum_{i\in U} x_i = \sum_{i\in U} x_i \frac{\pi_i}{\pi_i}$$

$$soit \quad \sum_{i\in U} \frac{x_i}{\pi_i}(I_i - \pi_i) = 0 \quad ou \quad \mathbf{A}(\mathbf{I} - \boldsymbol{\pi}) = 0$$

avec $\mathbf{A} = (\frac{x_1}{\pi_1},...,\frac{x_N}{\pi_N})$; $\mathbf{I} = (I_1,...,I_N)^T$ le vecteur des indicatrices de sélections et $\boldsymbol{\pi} = (\pi_1,...,\pi_N)^T$ le vecteur des probabilités d'inclusion. On voit alors que I doit se situer dans l'espace des contraintes $\boldsymbol{\pi} + Ker(\mathbf{A})$ qui représente l'espace où les conditions d'équilibrage sont respectées.

4. On se reportera à la partie V. pour des exemples d'application.

5. La macro CUBE est disponible sur le site de l'Insee avec sa documentation ici : https ://www.insee.fr/fr/information/2021904

On peut dès lors représenter facilement cette méthode dans un espace de dimension 3 pour une population de 3 unités : il s'agit d'un cube. En se plaçant dans le cas d'un sondage aléatoire simple sans remise de taille 2 et en affectant les même probabilités d'inclusion à chacune des unités (π_i=2/3), on peut remarquer que l'équilibrage est toujours exact en équilibrant sur la variable constante égale à 1. On sait alors qu'il existe 3 échantillons équilibrés composés de 2 unités distinctes : il s'agit ici des sommets (0,1,1) ; (1,1,0) et (1,0,1).

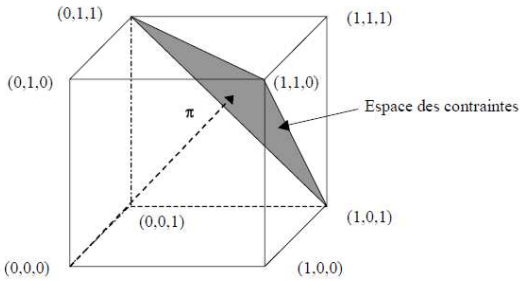

Fig. 1. Représentation graphique pour une population de 3 unités pour un sondage aléatoire simple sans remise équilibré sur la variable "1".

B. Détails de l'algorithme

On va maintenant détailler l'algorithme qui se déroule en deux phases : la phase de vol et la phase d'atterrissage.

α. La phase de vol

On initialise avec $\boldsymbol{\pi}^{(0)}$=$\boldsymbol{\pi}$. A l'étape t, on a $\boldsymbol{\pi}^{(t)}$=$\boldsymbol{\pi}^{(t-1)}$+$\boldsymbol{\delta}^{(t)}$ avec

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

- $\lambda_1(t)$,$\lambda_2(t)$>0
  → assure qu'au moins une unité est sélectionnée ou définitivement rejetée ;
- $\boldsymbol{u}(t) \in Ker(\mathbf{A})$
  → assure que les équations d'équilibrage sont exactement respectées ;
- Le choix aléatoire assure que les probabilités d'inclusion sont exactement respectées.

La phase de vol fonctionne par itérations successives où chaque pas décide du sort d'au moins un individu et de manière aléatoire une direction dans l'espace des contraintes. On va la suivre jusqu'à ce qu'elle conduise sur une face du cube. La phase de vol permet de statuer sur au moins $N-p$ individus [6] et de respecter les contraintes d'équilibrage et les probabilités d'inclusion.
À la fin de cette phase, si on est parvenu à un sommet du cube alors on obtient un échantillon parfaitement équilibré, sinon il est impossible de respecter exactement toutes les contraintes et on se retrouve "bloqué" sur une face du cube : il faudra alors déclencher la phase d'atterrissage. Celle-ci va permettre de statuer sur les individus restants en respectant exactement les probabilités d'inclusion et en respectant approximativement les contraintes d'équilibrage.

6. avec p étant le nombre de contraintes d'équilibrage.

β. La phase d'atterrissage

Il existe trois possibilités pour cette phase de l'algorithme.
La première consiste à relâcher les contraintes une par une. On introduit donc un degré de liberté à chaque étape nous permettant de poursuivre l'échantillonnage. C'est l'option la plus générale dans le sens où elle nous permet de travailler avec un nombre quelconque de variables d'équilibrage. Cependant, les premières variables relâchées peuvent être mal équilibrées.

La seconde consiste à définir un plan de sondage sur les unités restantes :

- respectant les probabilités d'inclusion de départ ;
- minimisant (en moyenne) l'écart à l'équilibre, à l'aide d'un critère du type :

$$\min \quad E||\hat{t}_{x\pi} - t_x||^2$$

Cette option permet d'obtenir un bon équilibrage global ; mais on doit définir entièrement un plan de sondage sur une population de p individus, ce qui est impossible si p est grand [7].

La troisième est identique à la seconde mais en respectant également la contrainte de taille fixe. Pour ce faire, il est nécessaire d'équilibrer sur la probabilité d'inclusion.

γ. Exemple général

On se place à nouveau dans notre cube, donc dans une population de 3 unités où l'on affecte les mêmes probabilités d'inclusion à chacune des unités (π_i=2/3), pour un exemple plus général c'est-à-dire où l'équilibrage est parfois exact [8]. On va étudier ici le cas d'un sondage aléatoire sans remise équilibré sur le numéro d'ordre [9] des individus.

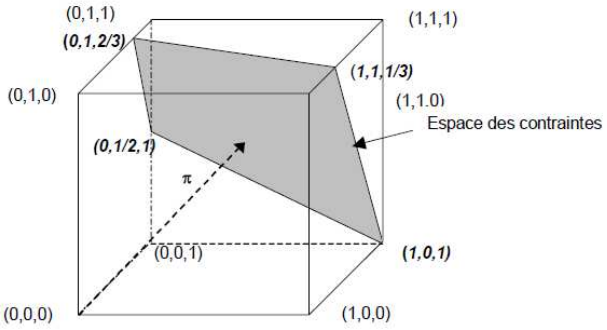

Fig. 2. Représentation graphique pour une population de 3 unités pour un sondage aléatoire sans remise équilibré sur le numéro d'ordre des unités.

L'équilibrage n'est exact ici que sur le sommet (1,0,1) qui représente la seule intersection d'un sommet du cube avec l'espace des contraintes.
À la fin de la phase de vol, l'algorithme aura sélectionné ce sommet [10] ou aura abouti sur un des trois autres points d'intersection du cube et de l'espace des contraintes : il faudra alors déclencher la phase d'atterrissage. Cette phase conduira en fonction du critère d'optimalité choisi à un des 5 échantillons (approximativement équilibrés donc) se situant sur les sommets des arêtes correspondant aux points d'intersections [11].
Le critère d'optimalité choisi pour la phase d'atterrissage permet de retenir plus probablement les échantillons les plus "proches" de l'équilibre. Cependant, pour préserver la nature aléatoire du tirage, la méthode ne peut pas garantir d'obtenir l'unique échantillon parfaitement équilibré.

IV. Estimation de variance

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

- le plan de sondage est exactement équilibré ;
- le plan de sondage est à entropie maximale [12] parmi les plans équilibrés sur les mêmes variables $\mathbf{x}_i$ avec les mêmes probabilités d'inclusion $\boldsymbol{\pi}$.

Ainsi, sous les deux conditions, le plan équilibré peut être vu comme un plan poissonnien conditionnel à $\hat{t}_{x\pi}$=t_x. L'approximation de variance résultante est donnée par :

$$V_{app}(\hat{t}_{y\pi}) \quad = \quad \sum_{i\in U} b_i(\frac{y_i}{\pi_i} - \frac{\mathbf{x}_i^T}{\pi_i}\mathbf{B})^2$$

où B est le vecteur de coefficient de la régression [13] de $\frac{y_i}{\pi_i}$ sur les variables d'équilibrage $\frac{\mathbf{x}_i}{\pi_i}$ et les b_i sont solutions d'un système non linéaire dont une première approximation est donnée dans l'article de Deville et Tillé par b_i=π_i(1-π_i).
En utilisant le principe d'expansion, on obtient alors l'estimateur de variance de Deville et Tillé.

Cependant, les deux conditions ci-dessus ne sont généralement pas vérifiées, en raison, pour la première, de la phase d'atterrissage (mais raisonnable si le nombre de variables d'équilibrage p est faible devant N) et, pour la seconde, de la difficulté à rendre un plan aussi aléatoire que possible (par exemple si on trie le fichier selon une variable auxiliaire on a un effet de stratification qui joue sur l'entropie).

V. Applications

A. L'équilibrage dans l'Échantillon Maître

L'Échantillon Maître (EM) est un échantillon de zones, utilisé comme réserve de logements pour les enquêtes auprès des ménages. L'Échantillon Maître de 1999 (EM99) a été utilisé pour les enquêtes réalisées entre 1999 et 2009. Chacune des zones était confiée à un enquêteur "stable dans le temps et localisé à proximité". On parle de Zones d'Action Enquêteur (ZAE). Le passage depuis 2004 à des Enquêtes de Recensement a nécessité de modifier le système de tirage de l'EM.

7. si p=19 par exemple alors on a 500 000 échantillons possibles environ.
8. Dans l'exemple précédent, on le rappelle, l'équilibrage était toujours exact et l'algorithme se concluait à la phase de vol pour déterminer un échantillon.
9. Variable correspondant à la valeur de la ligne où se situe l'individu dans le fichier.
10. Et dans ce cas, il s'arrêtera ici en ayant respecté toutes les contraintes d'équilibrage exactement.
11. Ici, il s'agit donc des sommets (0,1,0) ; (0,1,1) ; (0,0,1) ; (1,1,0) et (1,1,1).
12. L'entropie d'un plan de sondage p est définie par

$$L(p) = -\sum_{s\in U} p(s) \quad ln(p(s))$$

C'est une mesure de désordre : plus elle est forte plus le plan autorise la sélection d'un grand nombre d'échantillons (et laisse donc une grande place à l'aléatoire).
13. Si on considère E_i=y_i-$\mathbf{x}_i^T\mathbf{B}$ dans la formule de la variance on voit alors que E_i représente les résidus de la régression.

Dans chaque grande commune (+ de 10 000 habitants), on a stratifié selon le type d'adresse que l'on a réparti en 5 groupes de rotation. Pour les petites communes, on a stratifié par région et on a répartiti les communes en 5 groupes de rotation par tirage équilibré selon la méthode du Cube.
Le nouvel Échantillon Maître Octopusse a été présenté ainsi :

Au niveau des Grandes Communes :
- 1 ZAE = 1 GC ;
- tirage d'un échantillon de ZAE-GC (méthode du Cube) ;
- pour une enquête réalisée au cours de l'année t+1, tirage d'un échantillon de logements parmi ceux enquêtés l'année t.

→ Tirage à 2 degrés.

Au niveau des petites communes :
- 1 ZAE = regroupement de PC, contenant au moins 300 résidences principales de chaque groupe de rotation ;
- tirage d'un échantillon de ZAE-PC (méthode du Cube) ;
- pour une enquête réalisée au cours de l'année t+1, tirage d'un échantillon de logements dans les communes recensées l'année t.

→ Tirage à 2 degrés.

Le tirage des ZAE (en PC ou en GC) a été équilibré grâce au CUBE sur le nombre de résidences principales des ZAE par groupe de rotation, le revenu fiscal ventilé par groupe de rotation, le nombre de résidences principales par types d'espaces au RP99 [14].

B. Tirage de l'enquête Care-I

L'enquête CARE-Institutions de la DREES auprès des personnes âgées vivant en institution vise à compléter l'enquête CARE (Capacités, Aides et REssources) auprès des personnes âgées vivant en ménage ordinaire et poursuit les mêmes objectifs : suivre l'évolution de la dépendance, estimer le reste à charge lié à la dépendance et mesurer l'implication de l'entourage auprès de la personne âgée.
Le tirage des institutions s'effectue en 2 phases. Il s'agit d'abord de tirer 30 départements puis 1 000 établissements dans ces derniers. Les départements sont sélectionnés en fonction de leur nombre de résidents en établissement pour personnes âgées. En effet, tous les départements ne sont pas équivalents : ils ont plus ou moins d'établissements ou de résidents. Pour tenir compte de ces différences, on réalise un tirage à probabilités inégales. Le tirage des départements a été effectué au sein de 3 groupes de départements homogènes dont la classification fut établie suite à une CAH prenant en compte des variables de type d'établissement et de tranches de capacité. Ces trois classes forment les strates de tirage.
En deuxième phase, on souhaite que la répartition des seniors de l'échantillon soit identique à la répartition de l'ensemble des résidents en établissement (sur le champ retenu). Le tirage est effectué au sein de l'ensemble des départements sélectionnés précédemment. Pour permettre un tirage représentatif des établissements, un tirage équilibré a été effectué grâce à la macro SAS CUBE selon les variables de catégorie d'établissement et de statut juridique de celui-ci en prenant en compte la probabilité d'inclusion de l'établissement dans l'échantillon, tout en respectant également la contrainte d'échantillonnage de taille fixe (même nombre d'établissements par département dans l'échantillon).

14. Dans la région Île-de-France, d'autres variables (sur la structure démographique et le type de logements notamment) viennent compléter cet équilibrage.

REFERENCES

[1] Deville, J.-C. & Tillé, Y. (2004). Efficient Balanced Sampling : The Cube Method. *Biometrika*, Vol 91, No 4, pp 893-912.
[2] Deville, J.-C. & Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, Vol 128, pp 569-591.
[3] Tillé, Y. (2011). Dix années d'échantillonnage équilibré par la méthode du cube : une évaluation. *Techniques d'enquête*, Vol. 37, No 2, pp. 233-246.
[4] Rousseau, S. & Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. *Rapport technique*, Insee, Paris.
[5] Ardilly, P. (2006). *Les techniques de sondages*. Éditions Technip, Paris.
[6] Chauvet, G. (2012). Estimation de variance pour le nouvel Échantillon-Maître, *Journées de Méthodologie Statistique*, Paris.
[7] Christine, M. & Faivre, S. (2009). OCTOPUSSE : un système d'Échantillon- Maître pour le tirage des échantillons dans la dernière EAR, *Journées de Méthodologie Statistique*, Paris.

Département des méthodes statistiques
Version n° 1, diffusée le 21 juin 2017.