ECHANTILLONNAGE EQUILIBRE PAR LA METHODE DU CUBE, VARIANCE ET ESTIMATION DE VARIANCE

J-C. DEVILLE et Y. TILLE

CREST-ENSAI

Résumé

La méthode du cube permet de sélectionner des échantillons avec des probabilités d'inclusion fixes et tel que l'estimateur de Horvitz-Thompson restitue exactement les totaux des variables de contrôle. Elle généralise la plupart des techniques utilisant de l'information auxiliaire dans un plan de sondage, comme la stratification ou le tirage à probabilités inégales. Après un exposé de la méthode, nous donnons une approximation de la variance de ces plans, et une estimation de cette approximation.

Un plan de sondage équilibré est une procédure d'échantillonnage qui sélectionne des échantillons répondants aux contraintes suivantes : l'estimateur de Horvitz-Thompson de variables de contrôle est égal aux totaux de ces variables. Bien que l'exposé soit simple, sa réalisation est difficile. La recherche d'algorithmes permettant de sélectionner des plans équilibrés n'est pourtant pas une préoccupation nouvelle. Yates (1949), Thionet (1953, pp. 203-207) et Deville, Grosbras et Roth, (1988) ont proposé des méthodes intéressantes mais insuffisante à divers titres. Royall et Herson (1973), dans le cadre d'une inférence basée sur le modèle, mettent en évidence l'intérêt des plans équilibrés mais préconisent, faute de mieux, l'usage d'un plan aléatoire simple sans remise qui équilibre.....en moyenne! La méthode du cube que nous proposons a l'avantage d'être exacte quand il existe une solution exacte au problème d'équilibrage, et donne une approximation optimale quand une solution exacte est impossible.

Elle permet enfin dans certains cas, de fournir une méthode assez simple de calcul et d'estimation de la variance des estimateurs de Horvitz-Thompson.

L'application de la macro Cube au tirage des unités primaires de l'échantillon-maître est traitée dans l'article de G. BOURDALLE, M. CHRISTINE et L. WILMS : "échantillons maître et Emploi".

1 Plans équilibrés, définition

Supposons que p caractères auxiliaires $z_1, ..., z_p$ soient disponibles c'est-à-dire que les vecteurs des valeurs $\mathbf{z}_k = (z_{k1} \ \ldots \ z_{kj} \ \ldots \ z_{kp})'$ prises sur les p caractères auxiliaires soient connus sur toutes les unités d'une population U de taille N. La connaissance des valeurs des caractères auxiliaires $\mathbf{z}_k$ permet de calculer les p totaux

$$t_{z_j} = \sum_{k \in U} z_{kj}, j = 1, ..., p,$$

ou sous forme vectorielle

$$\mathbf{t}_z = \sum_{k \in U} \mathbf{z}_k.$$

Quand l'échantillon est sélectionné, on peut calculer l'estimateur de Horvitz-Thompson des p variables auxiliaires

$$\widehat{t}_{z_j\pi} = \sum_{k \in S} \frac{z_{kj}}{\pi_k},$$

ce qui peut également s'écrire sous forme vectorielle

$$\widehat{\mathbf{t}}_{z\pi} = \sum_{k \in S} \frac{\mathbf{z}_k}{\pi_k}.$$

L'objectif est d'estimer le total t_y du caractère d'intérêt dont les valeurs ne sont connues que sur les unités sélectionnées dans l'échantillon à l'aide de l'estimateur de Horvitz-Thompson

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

L'objectif d'un tirage équilibré consiste à exploiter toute l'information auxiliaire. On utilise à la définition suivante :

Définition 1 *Un plan de sondage $p(s)$ est dit équilibré sur les caractères auxiliaires $z_1, ..., z_p$, si et seulement si il vérifie les équations d'équilibrage données par*

$$\widehat{\mathbf{t}}_{z\pi} = \mathbf{t}_z, \tag{1}$$

ce qui peut également s'écrire

$$\sum_{k \in s} \frac{z_{kj}}{\pi_k} = \sum_{k \in U} z_{kj}, \tag{2}$$

pour tout $s \subset U$ tel que $p(s) > 0$, et pour tout $j = 1, ..., p$.

Remarque
L'expression (2) équivaut à

$$Var\left(\widehat{\mathbf{t}}_{z\pi}\right) = 0.$$

Exemple 1
Un plan de sondage de taille fixe n est équilibré sur la variable $z_k = \pi_k, k \in U$. En effet,

$$\sum_{k \in S} \frac{z_k}{\pi_k} = \sum_{k \in S} 1 = n.$$

Exemple 2
On suppose que le plan de sondage est stratifié et que dans chaque strate $U_h, h = 1, ..., H$, de taille N_h, un échantillon de taille n_h est sélectionné à probabilités égales, alors le plan est équilibré sur les H variables δ_{kh} de valeurs

$$\delta_{kh} = \begin{cases} 1 & \text{si } k \in U_h \\ 0 & \text{si } k \notin U_h. \end{cases}$$

En effet,

$$\sum_{k \in S} \frac{\delta_{kh}}{\pi_k} = \sum_{k \in S} \delta_{kh} \frac{N_h}{n_h} = N_h,$$

pour $h = 1, ..., H$.

Cependant dans beaucoup de cas, il n'est pas possible de vérifier exactement les équations d'équilibrage données.

Exemple 3
Supposons que $N = 10, n = 7, \pi_k = 7/10, k \in U$, et que le seul caractère auxiliaire soit $z_k = k, k \in U$. Alors, un plan équilibré est tel que

$$\sum_{k \in S} \frac{k}{\pi_k} = \sum_{k \in U} k,$$

ce qui implique que $\sum_{k \in S} k$ doit être égal à $55 \times 7/10 = 38,5$, ce qui est impossible, car $38,5$ n'est pas entier.

L'exemple précédant montre met l'accent sur la difficulté considérable des plans équilibrés: Il existe un problème d'"arrondi" qui empêche les contraintes d'équilibrage d'être exactement satisfaites. C'est pourquoi, on cherche un plan de sondage qui vérifie exactement les contraintes si c'est possible, et approximativement si ce ne l'est pas. On verra plus loin que le problème d'arrondi devient négligeable si la taille de l'échantillon est grande relativement au nombre de contraintes.

2 Représentation d'un plan de sondage sous forme de cube

La méthode du cube (voir Deville et Tillé, 2000) est basée sur la représentation géométrique d'un plan de sondage. Les 2^N échantillons possibles (ou

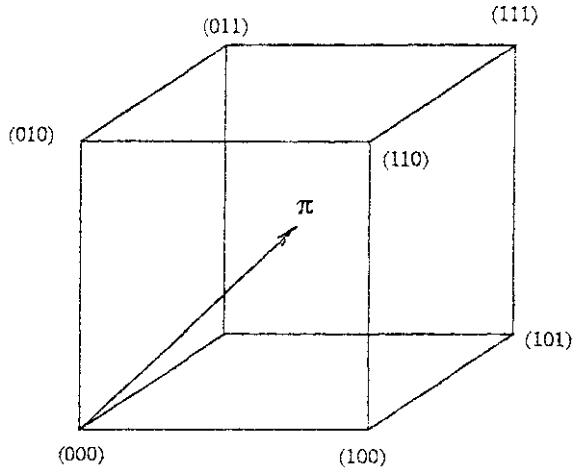

FIG. 1 – *Echantillons possibles dans une population de taille* $N = 3$

sous-ensembles) de U (si on considère que l'ensemble vide $\emptyset$ est un échantillon) peut être représenté par 2^N vecteurs de $\mathbb{R}^N$ de la manière suivante : $\mathbf{s} = (I[1 \in s], ..., I[k \in s], ..., I[N \in s])'$, où $I[k \in s]$ prend comme valeur 1 si $k \in s$ et 0 sinon. On peut alors interpréter chaque vecteur $\mathbf{s}$ comme le sommet d'un N-cube (hypercube de $\mathbb{R}^N$). Un plan de sondage de probabilités d'inclusion π_k peut alors être défini comme une loi de probabilités $p(\mathbf{s})$ sur l'ensemble des sommets du N-cube telle que

$$E(\mathbf{s}) = \sum_{s \in \mathcal{S}} p(\mathbf{s})\mathbf{s} = \boldsymbol{\pi},$$

où $\boldsymbol{\pi} = [\pi_k]$ est le vecteur de probabilités d'inclusion. Un plan de sondage est donc une combinaison linéaire convexe (à coefficients positifs) des sommets d'un N-cube.

Un algorithme d'échantillonnage peut donc être interprété comme

Un cheminement aléatoire qui permet d'atteindre un sommet du N-cube à partir d'un vecteur $\boldsymbol{\pi}$ peut donc être interprété comme un algorithme d'échantillonnage. De plus, se ce cheminement respecte des contraintes, il peut être considéré comme un plan équilibré. La figure 1 montre la représentation géométrique d'un plan dans une population de taille $N = 3$.

3 Echantillons équilibrés

La méthode du cube se décompose en deux phases : la *phase de vol* et la *phase d'atterrissage*. Dans la phase de vol, l'objectif est de transformer aléatoirement le vecteur des probabilités d'inclusion $\boldsymbol{\pi}$ en 0 ou 1 tout en vérifiant exactement

les contraintes. On montrera que l'on peut toujours arriver à un vecteur de 0 ou de 1 à l'exception de p coordonnées. On aboutit ainsi à un vecteur π^* dont $N-p$ valeurs valent 0 ou 1 et p sont comprises strictement entre 0 et 1. C'est ce qu'on appelle une p-face du cube. La phase d'atterrissage consiste à gérer le mieux possible le fait que les équations d'équilibrage (1) ne peuvent pas être exactement satisfaites et à obtenir un échantillon issu de la p-face où a conduit la phase de vol.

Les équations d'équilibrage (1) peuvent également s'écrire

$$\begin{cases} \sum_{k\in U} \mathbf{a}_k c_k = \sum_{k\in U} \mathbf{a}_k \pi_k \\ c_k \in \{0,1\}, k \in U, \end{cases} \tag{3}$$

où $\mathrm{a}_k = \mathrm{z}_k/\pi_k, k \in U$, et c_k est égal à 1 si l'unité k est dans l'échantillon et 0 sinon. Le système d'équations (3) avec les z_k donnés et les c_k inconnus définit un sous-espace affine de $\mathbb{R}^N$ de dimension $N-p$ que l'on note Q.

Si on définit $\mathbf{A}$ comme la matrice de dimension $p \times N$ donnée par $\mathbf{A} = (\mathrm{a}_1 \ldots \mathrm{a}_k \ldots \mathrm{a}_N)$, on remarque que $Q = \pi + \mathrm{Ker}\ \mathbf{A}$, où Ker $\mathbf{A}$ est le noyau de $\mathbf{A}$, autrement dit Ker $\mathbf{A}$ est le sous-espace linéaire engendré par les vecteurs de $\{\mathbf{u} \in \mathbb{R}^N | \mathbf{Au} = 0\}$.

L'idée principale pour obtenir un échantillon équilibré consiste à choisir un sommet du N-cube (un échantillon) qui reste sur le sous-espace linéaire Q ou qui est proche de Q quand ce n'est pas possible. Si C représente le N-cube dans $\mathbb{R}^N$ dont les sommets sont les échantillons de U, l'intersection entre C et Q est non-vide, car π est à l'intérieur de C et appartient à Q. L'intersection entre un N-cube et un sous-espace linéaire définit un polyèdre convexe K qui est défini formellement par $K = C \cap Q = \{[0,1]^N \cap (\pi + \mathrm{Ker}\ \mathbf{A})\}$ et est de dimension $N-p$, car il est l'intersection d'un N-cube et d'un plan de dimension $N-p$ qui a un point a l'intérieur de C.

Définition 2 *Un point extrême d'un convexe de $\mathbb{R}^N$, est un point qui ne peut pas s'écrire comme une combinaison linéaire de points de ce convexe.*

Définition 3 *Un polyèdre convexe est un convexe dont le nombre de points extrêmes est fini.*

Définition 4 *Soit D un polyèdre convexe, un sommet de D est défini comme un point qui ne peut être écrit comme une combinaison linéaire convexe à coefficients positifs des autres points de D. L'ensemble de tous les sommets des D s'écrit $Ext(D)$.*

Par exemple, on a $\#Ext(C) = 2^N$. Comme K est un polyèdre convexe, le nombre de sommets $Ext(K)$ de K est fini. De plus, chaque point intérieur de K peut s'écrire comme au moins une combinaison linéaire convexe à coefficients positifs des sommets.

Définition 5 *Un échantillon* s *est exactement équilibré si et seulement si* $\mathrm{s} \in Ext(C) \cap Q$.

On remarque qu'une condition nécessaire pour trouver un échantillon exactement équilibré est que $Ext(C) \cap Q \neq \emptyset$.

Définition 6 *Un système d'équations d'équilibrage peut être*

(i) exactement satisfait si $Ext(C) \cap Q = Ext(C \cap Q)$,

(ii) approximativement satisfait si $Ext(C) \cap Q = \emptyset$,

(iii) parfois satisfait si $Ext(C) \cap Q \neq Ext(C \cap Q)$, *et* $Ext(C) \cap Q \neq \emptyset$.

Théorème 1 *Si* $\mathbf{r} = [r_k]$ *est un point extrême de* K *alors* $\#\{k|0 < r_k < 1\} \leq p$.

Démonstration (par l'absurde) Soit $\mathbf{A}^*$ la matrice $\mathbf{A}$ restreinte aux unités non-entières de $\mathbf{r}$ autrement dit restreinte à $U^* = \#\{k|0 < r_k < 1\}$. Si $q = \#U^* > p$, alors Ker $\mathbf{A}^*$ est de dimension $q - p > 0$, et $\mathbf{r}$ n'est pas un point extrême de K. □

Les trois exemples qui suivent montrent que le problème d'arrondi peut être vu géométriquement. En effet, les équations d'équilibrage ne peuvent être exactement vérifiées quand les sommets de K ne sont pas des sommets de C, c'est-à-dire quand $q > 0$.

Exemple 4
En figure 2, un plan de sondage dans une population de taille $N = 3$ est examiné. La seule contrainte consiste à fixer la taille de l'échantillon $n = 2$, donc $p = 1$ et $z_k = \pi_k, k \in U$. De plus, les probabilités d'inclusion sont telles que $\pi_1 + \pi_2 + \pi_3 = 2$. Dans ce cas, l'équation d'équilibrage peut être vérifiée exactement.

Exemple 5
Dans la figure 3, on donne l'exemple du cas où le sous-espace des contraintes ne passe par aucun sommet du cube. La seule contrainte ($p = 1$) est donnée par la variable auxiliaire $z_1 = 0, z_2 = 6 \times \pi_2$ et $z_3 = 4 \times \pi_3$. De plus, les probabilités d'inclusion doivent satisfaire l'équation $6 \times \pi_2 + 4 \times \pi_3 = 5$. Il est alors impossible de vérifier exactement l'équation d'équilibrage. L'équation d'équilibrage peut seulement être vérifiée approximativement.

Exemple 6
Dans la figure 4, on donne l'exemple du cas où le sous-espace des contraintes passe par deux sommets du cube mais un sommet de l'intersection n'est pas un sommet du cube. La seule contrainte ($p = 1$) est donnée par la variable auxiliaire $z_1 = \pi_1, z_2 = 3 \times \pi_2$ et $z_3 = \pi_3$. Les probabilités d'inclusion doivent satisfaire l'équation : $\pi_1 + 3 \times \pi_2 + \pi_3 = 4$. L'équation d'équilibrage peut seulement être parfois satisfaite.

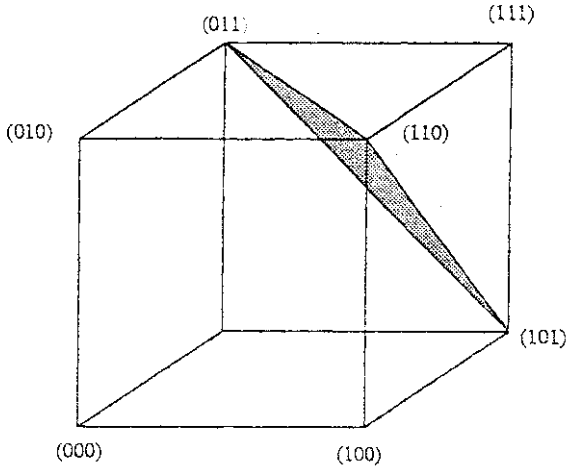

FIG. 2 – *Contrainte de taille fixe : tous les sommets de K sont des sommets du cube*

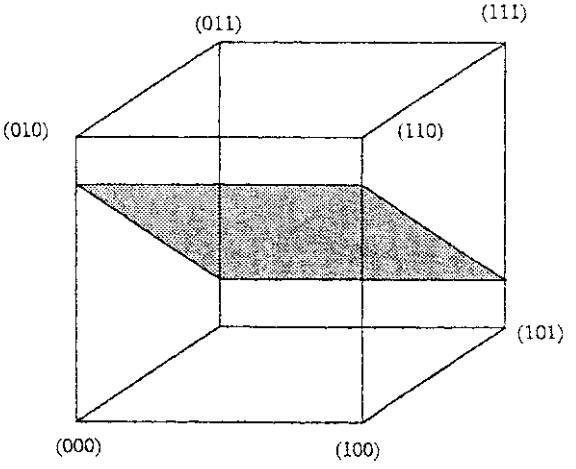

FIG. 3 – *Aucun des sommets de K n'est un sommet du cube*

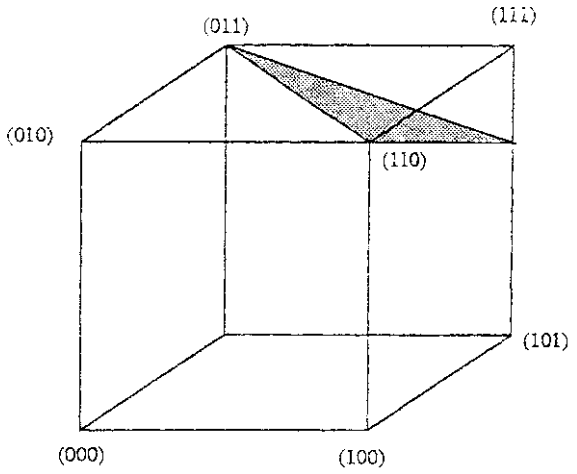

FIG. 4 – *Deux des sommets de K sont des sommets du cube, mais l'autre ne l'est pas*

4 La martingale équilibrante

Les deux phases de la procédure peuvent être résumées comme suit. A la fin de la première phase, un sommet de K est choisi au hasard de manière à ce que les probabilités d'inclusion $\pi_k, k \in U$, et les équations d'équilibrage (1) soient exactement satisfaites. La phase d'atterrissage est nécessaire seulement si le sommet de K atteint n'est pas un sommet de C, et consiste à gérer au mieux le relâchement des contraintes (1) de manière à sélectionner un échantillon, c'est-à-dire un sommet C.

Une méthode générale pour réaliser la phase de vol consiste à utiliser une martingale équilibrante définie comme suit :

Définition 7 *Un processus stochastique à temps discret $\pi(t) = [\pi_k(t)], t = 0, 1, ...$ dans $\mathbb{R}^N$ est appelé martingale équilibrante pour un vecteur de probabilités d'inclusion π et les variables auxiliaires $z_1, ..., z_p$, si*

1. $\pi(0) = \pi$,
2. $E[\pi(t)|\pi(t-1), ..., \pi(0)] = \pi(t-1), t = 1, 2, ...$
3. $\pi(t) \in K = \left\{[0,1]^N \cap (\pi + Ker\ \mathbf{A})\right\}$, *où* $\mathbf{A}$ *est une matrice de dimension* $p \times N$ *donnée par* $\mathbf{A} = (\mathbf{z}_1/\pi_1 \ldots \mathbf{z}_k/\pi_k \ldots \mathbf{z}_N/\pi_N)$.

Une martingale équilibrante est donc telle que $\pi(t-1)$ est en quelque sorte au milieu des valeurs possibles de $\pi(t)$.

Théorème 2 *Si $\pi(t)$ est une martingale équilibrante, alors*
(i) $E[\pi(t)] = E[\pi(t-1)] = ... = E[\pi(0)] = \pi$.

(ii) $\sum_{k\in U} a_k \pi_k(t) = \sum_{k\in U} a_k \pi_k = t_x, t = 0, 1, 2, ...$

(iii) Quand la martingale équilibrante atteint une face de K, alors elle reste "collée" sur cette face.

Démonstration (i) est évident. (ii) découle du fait que $\boldsymbol{\pi}(t) \in K$. Pour (iii), quand $\boldsymbol{\pi}(t-1)$ appartient à une face, $\boldsymbol{\pi}(t-1)$ est la moyenne des valeurs possibles de $\boldsymbol{\pi}(t)$ qui doit donc aussi appartenir à cette face. □

Le théorème 2 (iii) implique directement que (i) si $\pi_k(t) = 0$, alors $\pi_k(t+h) = 0, h \geq 0$, (ii) si $\pi_k(t) = 1$, alors $\pi_k(t+h) = 1, h \geq 0$, (iii) les sommets de K sont donc des états absorbants.

5 Implémentation de la phase de vol

Le problème pratique consiste à trouver une méthode qui atteint rapidement un sommet mais qui reste néanmoins fortement randomisée. La famille de solutions suivantes permet d'atteindre un sommet de K en N étapes au plus : D'abord on initialise avec $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$. Ensuite, aux instants $t = 1, ..., T$, on répète les trois étapes suivantes :

1. On génère un vecteur quelconque (aléatoire ou non) $\mathbf{u}(t) = [u_k(t)] \neq 0$ tel que *(i)* $\mathbf{u}(t)$ est dans le noyau de la matrice $\mathbf{A}$, *(ii)* $u_k(t) = 0$ si $\pi_k(t)$ soit entier.

2. On calcule $\lambda_1^*(t)$ et $\lambda_2^*(t)$, les deux plus grandes valeurs de $\lambda_1(t)$ et $\lambda_2(t)$ telles que $0 \leq \boldsymbol{\pi}(t) + \lambda_1(t)\mathbf{u}(t) \leq 1$, et $0 \leq \boldsymbol{\pi}(t) - \lambda_2(t)\mathbf{u}(t) \leq 1$. On remarque que $\lambda_1(t) > 0$ et $\lambda_2(t) > 0$.

3. On sélectionne enfin

$$\boldsymbol{\pi}(t) = \begin{cases} \boldsymbol{\pi}(t-1) + \lambda_1^*(t)\mathbf{u}(t) & \text{avec une probabilité } q_1(t) \\ \boldsymbol{\pi}(t-1) - \lambda_2^*(t)\mathbf{u}(t) & \text{avec une probabilité } q_2(t), \end{cases} \tag{4}$$

où $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$ et $q_2(t) = \lambda_1^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$.

Cette étape générale est répétée jusqu'à ce qu'il ne soit plus possible de trouver un vecteur $\mathbf{u}(t) = [u_k(t)]$ (aléatoire ou non) tel que *(i)* $\mathbf{u}(t)$ soit dans le noyau de $\mathbf{A}$, *(ii)* $u_k(t) = 0$ si $\pi_k(t-1)$ est entier.

On voit bien que cette procédure définit une martingale équilibrante. En effet,

1. Le processus est initialisé par $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$.

2. De l'expression (4), on a $E[\boldsymbol{\pi}(t)|\boldsymbol{\pi}(t-1), ..., \boldsymbol{\pi}(0)] = \boldsymbol{\pi}(t-1), t = 1, 2, \ldots,$ car $E[\boldsymbol{\pi}(t)|\boldsymbol{\pi}(t-1), \mathbf{u}(t)] = \boldsymbol{\pi}(t-1), t = 1, 2, \ldots$

3. Comme $\mathbf{u}(t)$ est dans le noyau de $\mathbf{A}$, on obtient par (4), que $\boldsymbol{\pi}(t)$ reste toujours dans $K = \left\{[0,1]^N \cap (\boldsymbol{\pi} + \text{Ker } \mathbf{A})\right\}$.

A chaque étape, au moins une composante du processus est définitivement arrondie à 0 ou à 1, autrement dit $\pi(t)$ a au moins une composante entière de plus que $\pi(t-1)$. Donc $\pi(1)$ est sur une face du N-cube (sur un cube de dimension $N-1$ au plus), $\pi(2)$ est sur un cube de dimension $N-2$ au plus, et ainsi de suite.

Soit T l'instant où la phase de vol est terminée. Le fait de ne plus trouver de vecteur $\mathbf{u}(T+1)$ tel que *(i)* $\mathbf{u}(T+1)$ est dans le noyau de la matrice $\mathbf{A}$, *(ii)* $u_k(T+1)=0$ si $\pi_k(T)$ est entier, montre que la martingale équilibrante a atteint un sommet de K, et donc par le théorème 1 que $\#\{0<\pi_k(T)<1\}\leq p$.

6 Implémentation de la phase d'atterrissage

A la fin de la phase de vol, la martingale équilibrante a atteint un sommet de K mais pas nécessairement un sommet de C. Soit q le nombre de composantes non-entières de ce sommet. Si $q=0$, l'algorithme est terminé. Si $q>0$, certaines contraintes ne peuvent pas être atteintes exactement. Les contraintes doivent alors être relâchées pour atteindre un sommet du N-cube proche en un certain sens du point extrême déjà atteint. Soit T la dernière étape de la phase 1, par simplicité on note $\pi^*=[\pi_k^*]=\pi(T)$.

Définition 8 *Un échantillon* s *est dit compatible avec un vecteur* π^* *si* $\pi_k^*-I[k\in s]=0$ *pour tout* k *tel que* π_k^* *est entier. Soit* $C(\pi^*)$ *l'ensemble des échantillons compatibles avec* π^*.

Remarque
Le vecteur π^* est donc sur une q face.

Le problème consiste à chercher un plan de sondage $p(s), s\in\mathcal{S}$ qui fournit des échantillons $s\in C(\pi^*)$ qui donnent une bonne approximation pour les équations d'équilibrage, ce qui peut s'écrire:

$$\sum_{k\in s|s\in C(\pi^*)}\mathbf{a}_k\approx\sum_{k\in U}\mathbf{a}_k\pi_k^*=\sum_{k\in U}\mathbf{a}_k\pi_k. \tag{5}$$

Une bonne approximation devrait minimiser une fonction de la matrice variance-covariance de l'estimateur de Horvitz-Thompson des totaux des variables auxiliaires donnée par

$$Var\left(\widehat{\mathbf{t}}_{z\pi}\right)=\sum_{k\in U}\sum_{\ell\in U}\frac{\mathbf{z}_k\mathbf{z}_k'}{\pi_k\pi_\ell}\Delta_{k\ell}=\mathbf{A}\boldsymbol{\Delta}\mathbf{A}', \tag{6}$$

où

$$\Delta_{k\ell}=\begin{cases}\pi_{k\ell}-\pi_k\pi_\ell & \text{si } k\neq\ell\\ \pi_k(1-\pi_k) & \text{si } k=\ell.\end{cases}$$

et $\boldsymbol{\Delta}=[\Delta_{k\ell}]=Var(\mathbf{S})$. La matrice $\boldsymbol{\Delta}$ est l'opérateur variance-covariance de l'estimateur de Horvitz-Thompson. Cet opérateur peut être scindé en deux parties

$$\boldsymbol{\Delta}=\boldsymbol{\Delta}_F+E\boldsymbol{\Delta}_{L|\pi^*},$$

où $\boldsymbol{\Delta}_F$ est la partie due à la phase de vol

$$\boldsymbol{\Delta}_F = VarE\left(\mathrm{S} \mid \pi^*\right) = Var\left(\pi^*\right),$$

et $\boldsymbol{\Delta}_{L|\pi^*}$ est la partie due à la phase d'atterrissage

$$\boldsymbol{\Delta}_{L|\pi^*} = Var\left(\mathrm{S} \mid \pi^*\right) = \sum_{s \in \mathcal{C}(\pi^*)} p(s|\pi^*)(\mathrm{s} - \pi^*)(\mathrm{s} - \pi^*)',$$

et $p(s|\pi^*)$ est la probabilité de sélectionner l'échantillon s étant donné que la phase de vol s'est terminée sur π^*.

Comme $\mathrm{A}\boldsymbol{\Delta}_F\mathrm{A}' = 0$, on a

$$Var\left(\widehat{\mathrm{t}}_{z\pi}\right) = EVar\left(\widehat{\mathrm{t}}_{z\pi} \mid \pi^*\right) = E\left(\mathrm{A}\boldsymbol{\Delta}_{L|\pi^*}\mathrm{A}'\right).$$

A la fin de la phase de vol, le problème consiste à trouver le plan de sondage $p(s|\pi^*)$ qui minimise une fonction de la matrice $Var(\widehat{\mathrm{t}}_{z\pi} \mid \pi^*)$. Considérons la matrice positive $\mathrm{M} = [m_{ij}]$ de dimension $q \times q$. On ne considérera que des fonctions particulières de $Var\left(\widehat{\mathrm{t}}_{z\pi} \mid \pi^*\right)$ de type M-trace :

$$\begin{aligned} \text{M-trace } Var\left(\widehat{\mathrm{t}}_{z\pi} \mid \pi^*\right) &= \text{trace}\left\{\mathrm{M} \times Var\left(\widehat{\mathrm{t}}_{z\pi} \mid \pi^*\right)\right\} \\ &= \sum_{s \in \mathcal{C}(\pi^*)} p(s|\pi^*)(\mathrm{s} - \pi^*)'\mathrm{A}'\mathrm{MA}(\mathrm{s} - \pi^*). \quad (7) \end{aligned}$$

Si $C(s) = (\mathrm{s} - \pi^*)'\mathrm{A}'\mathrm{MA}(\mathrm{s} - \pi^*)$ définit le "coût" associé à l'échantillon s, minimiser la M-trace consiste à minimiser le coût conditionnel moyen par rapport à π^*, et peut être obtenu en résolvant le programme linéaire suivant, sur tous les $s \in \mathcal{C}(\pi^*)$, de manière à trouver le meilleur plan de sondage $p(.|\pi^*)$:

$$\left|\begin{array}{l} \min_{p(.|\pi^*)} \sum_{s \in \mathcal{C}(\pi^*)} C(s)p(s|\pi^*), \\ \text{sous les contraintes} \\ \\ \sum_{s \in \mathcal{C}(\pi^*)} p(s|\pi^*) = 1, \\ \sum_{s \in \mathcal{C}(\pi^*)|s \ni k} p(s|\pi^*) = \pi_k^*, k \in U, \\ 0 \leq p(s|\pi^*) \leq 1, s \in \mathcal{C}(\pi^*). \end{array}\right. \quad (8)$$

Soit $U^* = \{k \in U | 0 < \pi_k^* < 1\}$, $q = \#U^*$, et $s^* = s \cap U^*$. Le programme linéaire (8) peut aussi s'écrire

$$\left| \begin{array}{l} \min\limits_{p^*(.)} \sum\limits_{s^* \subset U^*} C(s^*) p^*(s^*), \\ \text{sous les contraintes} \\ \\ \sum\limits_{s^* \subset U^*} p^*(s^*) = 1, \\ \sum\limits_{s^* \subset U^* | s^* \ni k} p^*(s^*) = \pi_k^*, k \in U^*, \\ 0 \leq p^*(s^*) \leq 1, s \subset U^*. \end{array} \right. \tag{9}$$

Ce programme linéaire ne dépend plus de la taille de la population mais seulement du nombre de variables d'équilibrage, car $q \leq p$. Il se restreint donc aux 2^q échantillons possibles et peut donc être généralement résolu avec une quinzaine de variables d'équilibrage.

Il est important de souligner qu'un programme linéaire aboutit à la sélection d'un plan à support minimal au sens de la définition suivante :

Définition 9 *Soit $p(.)$ un plan de sondage sur la population U de probabilités d'inclusion π_k, et $\mathcal{B} = \{s | p(s) > 0\}$. Un plan de sondage $p(.)$ est défini sur un support minimal, si et seulement si il n'existe pas d'ensemble $\mathcal{B}_0 \subset \mathcal{B}$ tel que $\mathcal{B}_0 \neq \mathcal{B}$, et*

$$\sum_{s \in \mathcal{B}_0} p_0(s) = \pi_k, k \in U, \tag{10}$$

a une solution en $p_0(s)$.

Théorème 3 *Le programme linéaire (8) a au moins une solution définie sur un support minimal.*

Démonstration
Supposons que le plan de sondage $p(.)$ ne soit pas défini sur un support minimal. Alors le système linéaire (10) a un nombre fini de solutions définies sur un support minimal noté $p_1(.), ..., p_J(.)$. Comme $p(.)$ peut s'écrire $p(s) = \sum_{j=1}^{J} \lambda_j p_j(s)$, avec des coefficients positifs λ_j, et $\sum_{j=1}^{J} \lambda_j = 1$. Le plan à support minimal avec le plus petit coût moyen $\sum_{s \in S} p_j(s) C(s)$ a un plus petit coût moyen que $p(.)$. □

Si le nombre de variables auxiliaires est trop grand pour résoudre le programme linéaire par l'algorithme du simplexe, alors, à la fin de la phase d'atterrissage, une variable auxiliaire peut directement être supprimée. Une contrainte est alors relâchée et il est possible de retourner à la phase de vol jusqu'à ce qu'il ne soit à nouveau plus possible de se déplacer dans le sous-espace des

contraintes. Les contraintes sont donc relâchées successivement. Pour cette raison, il est nécessaire de trier les variables par ordre d'importance de manière à ce que les contraintes les moins importantes soient d'abord relâchées.

Quelle que soit la variante choisie pour la phase d'atterrissage, l'écart entre l'estimateur de Horvitz-Thompson et le total peut être majorée.

Théorème 4 *Pour n'importe quelle application de la méthode du cube*

$$\left|\widehat{t}_{z_j\pi} - t_{z_j}\right| \leq p \times \max_{k\in U}\left|\frac{z_{kj}}{\pi_k}\right|.$$

La démonstration est donnée dans Deville et Tillé (2000). L'écart est donc négligeable quand le nombre de variables d'équilibrage est petit. En particulier, quand le tirage est à probabilités égales $\pi_k = n/N$, on obtient

$$\left|\widehat{t}_{z_j\pi} - t_{z_j}\right| \leq \frac{pN}{n} \times \max_{k\in U}|z_{kj}| = O\left(\frac{pN}{n}\right).$$

Il faut évidemment ajouter que cette borne considère la pire des situations, alors que la phase d'atterrissage vise justement à trouver la meilleure.

7 Choix de la fonction de coût

Un problème délicat consiste à choisir le coût $C(s)$. Le choix de $C(.)$ est une décision qui dépend des priorités du gestionnaire de l'enquête. Comme on l'a vu dans l'expression (7), le coût est défini au moyen d'une matrice M.

Un coût simple peut être construit au moyen de la somme des carrés des coefficients de variation

$$C_1(s) = \sum_j \frac{\left(\widehat{t}_{z_j\pi}(s) - t_{z_j}\right)^2}{t_{z_j}^2}, \tag{11}$$

où $\widehat{t}_{z_j\pi}(s)$ est la valeur prise par $\widehat{t}_{z_j\pi}$ sur l'échantillon s. Le coût $C_1(.)$ est une M-trace où M est une matrice diagonale avec les $1/t_{z_j}^2$ sur la diagonale.

Une autre possibilité consiste à utiliser $\mathbf{M} = [m_{k\ell}] = (\mathbf{A}\mathbf{A}')^{-1}$, alors

$$C_2(s) \quad = \quad (\mathbf{s} - \pi^*)'\mathbf{A}'\left(\mathbf{A}\mathbf{A}'\right)^{-1}\mathbf{A}(\mathbf{s} - \pi^*).$$

Le choix de $C_2(.)$ a une interprétation naturelle donnée par le théorème suivant :

Théorème 5 *La distance entre un échantillon* s *et sa projection euclidienne sur le sous-espace des contraintes est donnée par*

$$C_2(s) = (\mathbf{s} - \pi^*)'\mathbf{A}'\left(\mathbf{A}\mathbf{A}'\right)^{-1}\mathbf{A}(\mathbf{s} - \pi^*). \tag{12}$$

Démonstration La projection d'un échantillon s sur le sous-espace des contraintes vaut

$$\mathbf{s} - \mathbf{A}' \left(\mathbf{A}\mathbf{A}'\right)^{-1} \mathbf{A}(\mathbf{s} - \pi).$$

La distance euclidienne entre s et sa projection est donc

$$(\mathbf{s}-\pi)'\mathbf{A}' \left(\mathbf{A}\mathbf{A}'\right)^{-1} \mathbf{A}(\mathbf{s}-\pi) = (\mathbf{s}-\pi^*+\pi^*-\pi)'\mathbf{A}' \left(\mathbf{A}\mathbf{A}'\right)^{-1} \mathbf{A}(\mathbf{s}-\pi^*+\pi^*-\pi)$$

et comme $\mathbf{A}(\pi - \pi^*) = 0$, (12) découle directement. □

Le coût $C_2(.)$ peut donc être interprété comme une simple distance dans $\mathbb{R}^N$.

8 Echantillonnage non-équilibré

La méthode du cube peut être utilisée sans information auxiliaire. Une procédure d'échantillonnage intéressante est le plan de Poisson qui consiste à réaliser des tirages indépendants pour chaque unité. Le plan de Poisson peut être obtenu au moyen de la méthode du cube en prenant $\mathbf{u}(t)$ tel que $u_t(t) = 1$, et $u_k(t) = 0$, si $k \neq t$. Ensuite $\lambda_1(t) = 1 - \pi_t$, $\lambda_2(t) = \pi_t$,

$$\pi(t+1) = \begin{cases} (\pi_1(t) \dots \ \pi_{t-1}(t) \ \ 1 \ \ \pi_{t+1} \dots \pi_N)' & \text{avec la probabilité } q_1(t) \\ (\pi_1(t) \dots \ \pi_{t-1}(t) \ \ 0 \ \ \pi_{t+1} \dots \pi_N) & \text{avec la probabilité } q_2(t), \end{cases}$$

où $q_1(t) = \pi_t$ et $q_2(t) = 1 - \pi_t$. Chaque unité est donc sélectionnée indépendamment des autres.

9 Application aux plans simples

Le plan simple sans remise de taille fixe peut également s'écrire très simplement comme un cas particulier de la méthode du cube. Supposons que $\pi = (n/N \ \dots \ n/N \ \dots \ n/N)$, et que la seule variable auxiliaire utilisée soit $z_k = n/N, k \in U$. Pour conserver la taille fixe, on peut utiliser un vecteur $\mathbf{u}(1)$ défini à partir d'un vecteur quelconque $\mathbf{v}(1)$ selon

$$u_k(1) = v_k(1) - \frac{1}{N} \sum_{k \in U} v_k(1).$$

10 Application à la stratification

La stratification peut être obtenue en prenant $z_{kh} = \delta_{kh} n_h / N_h, h = 1, ..., H$, où N_h est la taille de la strate dans la population U_h, n_h est la taille de la strate dans l'échantillon, et

$$\delta_{kh} = \begin{cases} 1 & \text{si } k \in U_h \\ 0 & \text{si } k \notin U_h. \end{cases}$$

À la première étape, la projection d'un vecteur $\mathbf{v}(t)$ sur l'espace des contraintes donne

$$u_k(1) = v_k(1) - \frac{1}{N_h} \sum_{\ell \in U_h} v_\ell(1), k \in U_h.$$

Les trois stratégies décrites dans la section 9 permettent d'obtenir un plan simple à l'intérieur de chacune des strates.

L'intérêt de la méthode du cube est que la stratification peut être généralisée à des strates qui se chevauchent. Un cas intéressant que l'on pourrait appeler "plan aléatoire par quotas marginaux" ou "stratification croisée" consiste à considérer deux critères de stratification. Par exemple dans une enquête sur des entreprises on utilise le "secteur d'activités" et la "région". Les strates de la première variable sont notées $U_{h.}, h = 1, ..., H$, et les strates de la seconde variable $U_{.i}, i = 1, ..., I$. On peut dès lors définir les $p = H + I$ variables auxiliaires d'équilibrage

$$z_{kj} = \pi_k \times \begin{cases} I[k \in U_{j.}] & j = 1, ..., H \\ I\left[k \in U_{.(j-H)}\right] & j = H+1, ..., H+I, \end{cases}$$

où $I[.]$ est une variable indicatrice qui prend la valeur 1 si la condition est vérifiée et 0 sinon. Ensuite, l'échantillon peut être directement sélectionné au moyen de la méthode du cube. La généralisation à l'utilisation de quotas multiples est évidente.

11 Application au tirage à probabilités inégales

Le problème du tirage à probabilités inégales peut être résolu par la méthode du cube. Supposons que l'objectif soit de sélectionner un échantillon de taille fixe n avec des probabilités d'inclusion $\pi_k, k \in U$, telles que $\sum_{k \in U} \pi_k = n$. Dans ce cas, la seule variable auxiliaire est $z_k = \pi_k$. On utilise un vecteur $u_k(t)$ tel que

$$\sum_{k \in U} u_k(t) = 0. \tag{13}$$

Chaque choix (aléatoire ou non) de vecteurs $\mathbf{u}(t)$ qui satisfont (13) produit une nouvelle méthode de tirage à probabilités inégales. Presque toutes les méthodes existantes peuvent être facilement représentées au moyen de la méthode du cube. La méthode du cube n'est qu'une représentation géométrique de la méthode de Scission (Deville et Tillé 1998).

Les techniques d'échantillonnage à probabilités inégales peuvent toujours être améliorées au moyen de la méthode du cube. En effet, dans toutes les méthodes d'échantillonnage à probabilités inégales de taille fixe, le plan est équilibré seulement sur une seule variable. Cependant, il y a toujours au moins deux variables disponibles $z_{k1} = \pi_k, k \in U$, et $z_{k2} = 1, k \in U$. La première implique la taille fixe de l'échantillon, la seconde implique que $\widehat{N}_\pi = \sum_{k \in S} 1/\pi_k = N$. Dans toutes les méthodes classiques d'échantillonnage à probabilités inégales, le plan est équilibré sur z_{k1} mais pas sur z_{k2}. La méthode du cube permet de prendre en compte les deux contraintes.

12 Approximation de la variance

La variance de $\widehat{t}_{y\pi}$ peut théoriquement être exprimée en utilisant les probabilités d'inclusion d'ordre deux. Malheureusement, même dans des cas très simples, le calcul s'avère impossible. Cependant, dans le cas du plan de Poisson, la variance de $\widehat{t}_{y\pi}$ peut être facilement calculée, car elle dépend seulement des probabilités d'inclusion d'ordre un.

Si $\widetilde{S}$ est l'échantillon aléatoire sélectionné au moyen d'un plan de Poisson et $\widetilde{\pi}_k, k \in U$, les probabilités d'inclusion d'ordre un de ce plan de Poisson, alors

$$Var_{poiss}\left(\widehat{t}_{y\pi}\right) = Var_{poiss}\left(\sum_{k\in\widetilde{S}}\frac{y_k}{\pi_k}\right) = \sum_{k\in U}\frac{y_k^2}{\pi_k^2}\widetilde{\pi}_k\left(1-\widetilde{\pi}_k\right).$$

Le plan de Poisson est le plan qui maximise l'entropie pour des probabilités d'inclusion d'ordre un données. Si le plan équilibré est d'entropie maximale ou proche de l'entropie maximale, Deville et Tillé (2000) ont montré que la variance peut être calculée comme la variance conditionnelle d'un plan de Poisson particulier de probabilités d'inclusion $\widetilde{\pi}_k$. Autrement dit,

$$Var_{equil}\left(\widehat{t}_{y\pi}\right) = Var_{poiss}\left(\widehat{t}_{y\pi}|\widehat{\mathbf{t}}_{z\pi} = \mathbf{t}_z\right)$$

où Var_{equil} est la variance sous le plan équilibré d'entropie maximale de probabilités d'inclusion π_k et Var_{poiss} est la variance sous le plan de Poisson de probabilités d'inclusion $\widetilde{\pi}_k$. Les probabilités $\widetilde{\pi}_k$ sont inconnues et restent à déterminer.

Si on suppose que, pour un plan de Poisson, le vecteur $(\widehat{t}_{y\pi}\ \widehat{\mathbf{t}}'_{z\pi})'$ a en première approximation une distribution normale multivariée, on obtient

$$Var_{poiss}\left(\widehat{t}_{y\pi}|\widehat{\mathbf{t}}_{z\pi} = \mathbf{t}_z\right) \approx Var_{poiss}\left(\widehat{t}_{y\pi} + \left(\mathbf{t}_z - \widehat{\mathbf{t}}_{z\pi}\right)'\beta\right), \qquad (14)$$

où Var_{poiss} est la variance sous le plan de Poisson de probabilités d'inclusion $\widetilde{\pi}_k$

$$\beta = Var_{poiss}\left(\widehat{\mathbf{t}}_{z\pi}\right)^{-1} Cov_{poiss}\left(\widehat{\mathbf{t}}_{z\pi}, \widehat{t}_{y\pi}\right),$$

$$Var_{poiss}\left(\widehat{\mathbf{t}}_{z\pi}\right) = \sum_{k\in U}\frac{\mathbf{z}_k\mathbf{z}_k'}{\pi_k^2}\widetilde{\pi}_k\left(1-\widetilde{\pi}_k\right),$$

et

$$Cov_{poiss}\left(\widehat{\mathbf{t}}_{z\pi}, \widehat{t}_{y\pi}\right) = \sum_{k\in U}\frac{\mathbf{z}_k y_k}{\pi_k^2}\widetilde{\pi}_k\left(1-\widetilde{\pi}_k\right).$$

En se basant sur (14), on obtient l'approximation de variance

$$Var_{app}\left(\widehat{t}_{y\pi}\right) = \sum_{k\in U}\frac{b_k}{\pi_k^2}\left(y_k - y_k^*\right)^2, \qquad (15)$$

où

$$y_k^* = \mathbf{z}_k' \left(\sum_{\ell \in U} b_\ell \frac{\mathbf{z}_\ell \mathbf{z}_\ell}{\pi_\ell^2} \right)^{-1} \sum_{\ell \in U} b_\ell \frac{\mathbf{z}_\ell y_\ell}{\pi_\ell^2},$$

est une prévision par la régression de y_k. Les poids $b_k = \widetilde{\pi}_k(1 - \widetilde{\pi}_k)$ sont utilisés, car les probabilités d'inclusion du plan de Poisson $\widetilde{\pi}_k$ ne sont pas exactement égaux aux probabilités d'inclusion du plan équilibré. Comme les valeurs exactes des $\widetilde{\pi}_k$ ne sont pas faciles à calculer, les valeurs exactes des b_k sont inconnues également et feront l'objet d'une approximation.

On remarque que l'expression (15) peut aussi s'écrire

$$Var_{app}\left(\widehat{t}_{y\pi}\right) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} \Delta_{app k\ell},$$

où $\boldsymbol{\Delta}_{app} = \{\Delta_{\text{app}k\ell}\}$ est l'approximation de l'opérateur de variance dont l'élément général est

$$\Delta_{\text{app}k\ell} = \begin{cases} b_k - b_k \dfrac{\mathbf{z}_k'}{\pi_k} \left(\displaystyle\sum_{i \in U} b_i \frac{\mathbf{z}_i \mathbf{z}_i'}{\pi_i^2} \right)^{-1} \dfrac{\mathbf{z}_k}{\pi_k} b_k & k = \ell \\ b_k \dfrac{\mathbf{z}_k'}{\pi_k} \left(\displaystyle\sum_{k \in U} b_i \frac{\mathbf{z}_i \mathbf{z}_i'}{\pi_i^2} \right)^{-1} \dfrac{\mathbf{z}_\ell b_\ell}{\pi_\ell} & k \neq \ell. \end{cases} \tag{16}$$

Le problème principal consiste donc à trouver une bonne approximation des b_k. Quatre valeurs ont été testées dans Deville et Tillé (2000) pour les b_k. Ces valeurs sont notées respectivement $b_{k\alpha}, \alpha = 1, 2, 3, 4$, et permettent de définir quatre estimateurs de variances notés $V_\alpha, \alpha = 1, 2, 3, 4$ et quatre opérateurs de variance notés $\boldsymbol{\Delta}_\alpha, \alpha = 1, 2, 3, 4$ en remplaçant dans (15) et (16), b_k par respectivement b_{k1}, b_{k2}, b_{k3}, et b_{k4}.

1. La première approximation est obtenue en considérant que quand n est grand, $\pi_k \approx \widetilde{\pi}_k, k \in U$. On prend alors $b_{k1} = \pi_k(1 - \pi_k)$.

2. La seconde approximation est obtenue en appliquant une correction pour la perte de degrés de liberté

   $$b_{k2} = \pi_k(1 - \pi_k) \frac{N}{N - p}.$$

   Cette correction permet d'obtenir exactement le plan aléatoire simple sans remise de taille fixe avec une taille d'échantillon fixée.

3. La troisième correction découle d'un ajustement sur les éléments diagonaux de l'opérateur de variance $\boldsymbol{\Delta}$ de la vraie variance donnée en (6). En effet, ces éléments diagonaux sont toujours connus et égaux à $\pi_k(1 - \pi_k)$. Donc, on utilise

   $$b_{k3} = \pi_k(1 - \pi_k) \frac{\text{trace } \boldsymbol{\Delta}}{\text{trace } \boldsymbol{\Delta}_1},$$

   on peut ainsi définir un opérateur de variance approché $\boldsymbol{\Delta}_3$ qui a la même trace que $\boldsymbol{\Delta}$.

4. Enfin, la quatrième approximation découle du fait que les éléments diagonaux de $\boldsymbol{\Delta}_{app}$ sont donnés en (16). Or les éléments diagonaux de $\boldsymbol{\Delta}$ ne dépendent que des probabilités d'inclusion d'ordre un et valent $\pi_k(1-\pi_k)$. Les b_{k4} sont construits de sorte que $\boldsymbol{\Delta}$ et $\boldsymbol{\Delta}_{app}$ aient la même diagonale, autrement dit que

$$\pi_k(1-\pi_k) = b_k - b_k \frac{\mathbf{z}_k'}{\pi_k} \left(\sum_{k \in U} b_k \frac{\mathbf{z}_k \mathbf{z}_k'}{\pi_k^2} \right)^{-1} \frac{\mathbf{z}_k}{\pi_k} b_k, k \in U. \qquad (17)$$

Cette approximation requiert un calcul itératif trivial d'après (17) dont malheureusement la convergence n'est pas toujours assurée. Quand on traite du cas de taille fixe, par exemple, une condition nécessaire et suffisante est que

$$\max \frac{\pi_k(1-\pi_k)}{\text{trace}\boldsymbol{\Delta}} < \frac{1}{2}$$

(voir Deville et Tillé, 2000). Cependant, cette quatrième approximation est la seule qui fournit l'expression de variance exacte pour les plans stratifiées avec des sondages aléatoires simples dans les strates.

Un ensemble de simulations réalisées par Deville et Tillé (2000) montre que b_{k4} donne la meilleure approximation. Malheureusement la solution du système d'équation (17) n'existe pas toujours. Dans ce cas, on peut utiliser b_{k3}.

13 Estimation de la variance

La variance peut être estimée en utilisant le même principe. L'estimateur général est donné par

$$\widehat{Var}\left(\hat{t}_{y\pi}\right) = \sum_{k \in S} \frac{c_k}{\pi_k^2} \left(y_k - \hat{y}_k^*\right)^2, \qquad (18)$$

où

$$\hat{y}_k^* = \mathbf{z}_k' \left(\sum_{\ell \in S} c_\ell \frac{\mathbf{z}_\ell \mathbf{z}_\ell'}{\pi_\ell^2} \right)^{-1} \sum_{\ell \in S} c_\ell \frac{\mathbf{z}_\ell y_\ell}{\pi_\ell^2},$$

est l'estimation du prédicteur par la régression de of y_k. A nouveau, quatre estimateurs découlent de quatre définitions des c_k. On remarque que (18) peut également s'écrire

$$\sum_{k \in S} \sum_{\ell \in S} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} D_{k\ell},$$

où

$$D_{k\ell} = \begin{cases} c_k - c_k \dfrac{\mathbf{z}_k'}{\pi_k} \left(\displaystyle\sum_{i \in S} c_i \frac{\mathbf{z}_i \mathbf{z}_i'}{\pi_i^2} \right)^{-1} \dfrac{\mathbf{z}_k}{\pi_k} c_k & k = \ell \\ c_k \dfrac{\mathbf{z}_k'}{\pi_k} \left(\displaystyle\sum_{k \in S} c_i \frac{\mathbf{z}_i \mathbf{z}_i'}{\pi_i^2} \right)^{-1} \dfrac{\mathbf{z}_\ell c_\ell}{\pi_\ell} & k \neq \ell. \end{cases}$$

Les quatre définitions de c_k sont notées c_{k1}, c_{k2}, c_{k3}, et c_{k4}, et permettent de définir quatre approximations de la variance en remplaçant dans l'expression (18) c_k par respectivement c_{k1}, c_{k2}, c_{k3}, et c_{k4}.

1. Le premier estimateur est obtenu en prenant $c_{k1} = (1 - \pi_k)$.

2. Le deuxième est obtenu en appliquant une correction pour la perte de degrés de liberté

$$c_{k2} = (1 - \pi_k)\frac{n}{n-p}.$$

   Cette correction donne un estimateur sans biais pour le plan aléatoire simple sans remise de taille fixe.

3. Le troisième estimateur découle du fait que les éléments diagonaux de la vraie matrice des $\Delta_{k\ell}/\pi_{k\ell}$ sont toujours connus et valent $1 - \pi_k$. Alors, on peut écrire

$$c_{k3} = (1 - \pi_k)\frac{\sum_{k\in S}(1-\pi_k)}{\sum_{k\in S} D_{kk}}.$$

4. Finalement, le quatrième estimateur découle du fait que les éléments diagonaux D_{kk} sont connus. Les c_{k4} sont construits de telle manière que

$$1 - \pi_k = D_{kk}, k \in U, \tag{19}$$

   ou, en d'autres mots,

$$1 - \pi_k = c_k - c_k \frac{\mathbf{z}_k'}{\pi_k}\left(\sum_{i\in S} c_i \frac{\mathbf{z}_i\mathbf{z}_i'}{\pi_i^2}\right)^{-1}\frac{\mathbf{z}_k}{\pi_k}c_k, k \in U.$$

   Cette quatrième approximation est la seule qui fournit l'estimateur exactement sans biais pour n'importe quel plan stratifié.

Ces quatre estimateurs découlent directement des quatre approximations de variance. On préfère de nouveau les coefficient c_{k4}.

Références

ARDILLY, P. (1991). Echantillonnage représentatif optimum à probabilités inégales. *Annales d'Économie et de Statistique* **23** 91-113.

ARDILLY, P. (1994). *Les Techniques de sondage*. Paris, Technip.

BERGER, Y. (1998). *Comportements asymptotiques des plans de sondage à probabilités inégales pour un modèle de population fixe*. Ph.D., Université Libre de Bruxelles.

BOUSABAA, A, LIEBER, J. et SIROLLI, R. (1999). *La Macro Cube*, Rapport interne, ENSAI, Rennes.

BREWER, K.W.R., and HANIF, M., (1983). *Sampling with unequal probabilities.* New York, Sringer Verlag.

CHEN, X.-H., DEMPSTER, A.P., et LIU, S.L. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* **81** 457-469.

DEVILLE, J.-C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. *Proceedings of the Workshop Auxiliary Information in Surveys.* Örebro (Suède).

DEVILLE, J.-C. (2000). Variance estimation for complex statistics: residual technique and linearisation. *Survey Methodology*, **25**, 193-204.

DEVILLE, J.-C., GROSBRAS, J.-M., et ROTH N. (1988). Efficient sampling algorithms and balanced sample. *COMPSTAT, Proceeding in computational statistics.* Physica Verlag, pp. 255-266.

DEVILLE J.-C., et SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376-382.

DEVILLE, J.-C., et TILLÉ, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* **85** 89-101.

DEVILLE, J.-C., et TILLÉ, Y. (2000). Selection of several unequal probability samples from the same population. *Journal of Statistical Planning and Inference* to appear.

FAN C.T., MULLER, M.E. et REZUCHA I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computer. *Journal of the American Statistical Association* **57** 387-402.

HÁJEK, J. (1981). *Sampling from finite population.* New York, Marcel Dekker.

HANSEN, M.H., et HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* **14** 333-362.

HEDAYAT, A.S., et MAJUMDAR, DIBYEN (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference* **44** 237-247.

MADOW, W.G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics* **20** 333-354.

NEYMAN, J. (1934). On the two different aspects of representative method : the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97** 558-606.

ROYALL, R. et HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association* **68** 880-889.

SUNTER, A. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics* **26** 261-268.

SUNTER, A. (1986). Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review* **54** 33-50.

THIONET, P. (1953). *La théorie des sondages*. Etudes théorique 5, INSEE, Paris, Imprimerie Nationale.

TILLÉ, Y. (1996). A moving stratification algorithm. *Survey Methodology* **22** 85-94.

TSCHUPROW, A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observation. *Metron* **3** 461-493.

WYNN, H.P. (1977). Convex sets of finite population plans. *Annals of Statistics* **5** 414-418.

YATES, F., (1949). Sampling methods for censuses and surveys, London, C. Griffen.