ÉCHANTILLONNAGE À DEUX REPRISES AVEC PPTSR

G.H. Choudhry et Jack E. Graham[1]

On décrit une théorie d'échantillonnage à deux reprises avec probabilités inégales et sans remise. La méthode de Fellegi (1963), où à chaque reprise la même probabilité de sélection est attribuée à une unité donnée, est utilisée pour choisir les unités de l'échantillon de renouvellement. On examine ensuite le développement mathématique de la variance des estimateurs composites de la valeur globale de la population à la deuxième reprise. Des résultats quantitatifs pour des échantillons de petite taille sont présentés et on compare l'efficacité de cette méthode à celle d'une technique différente.

1. INTRODUCTION

Dans les enquêtes à caractère répétitif, l'utilisation d'un plan d'échantillonnage avec remise partielle présente certains avantages tant du point de vue de l'efficacité que de la réduction du fardeau du répondant. Essentiellement, après chaque tirage d'un échantillon, une fraction des unités sont supprimées et remplacées par un nouvel sous-échantillon de la population. Un très grand nombre d'ouvrages statistiques examinent les méthodes et les estimateurs que l'on utilise pour l'échantillonnage à deux reprises ou plus avec probabilités égales. Mais les cas où les probabilités de sélection sont inégales présentent encore plus d'importance du point de vue de la pratique. Ainsi, prenons une population finie de N unités $\{1,2,\ldots,N\}$ qui est échantillonnée à deux reprises, soit en période 1 (l'échantillon précédent) et en période 2 (l'échantillon courant). Soient y_{1i} et y_{2i} les valeurs d'une charactéristique y de la $i^{ième}$ unité dans les échantillons 1 et 2 et Y_1 et Y_2 les valeurs globales de la population calculées à partir des échantillons correspondants. Une mesure de taille, x_i, est connue pour toutes les unités de la population.

[1] G.H. Choudhry, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada, et Jack E. Graham, Université Carleton.

Raj (1965) a étudié le plan d'échantillonnage suivant avec PPT (probabilité proportionnelle à la taille) : un premier échantillon s de taille n est constitué avec probabilités p_i proportionnelles aux valeurs de x_i et avec remise (AR). Dans une deuxième période, un échantillon aléatoire simple s_1 de m unités est tiré de s sans remise (SR) et un échantillon indépendant s_2 de taille u = n - m est sélectionné avec PPT à partir de l'ensemble de la population. Les valeurs de Y_1 et Y_2 sont alors estimées à l'aide des équations :

$$\hat{Y}_1 = \sum_s y_{1i}/(np_i) \text{ and } \hat{Y}^*_{2R} = Q^*\hat{Y}_{2u} + (1-Q^*)\hat{Y}'_2,$$

où $$\hat{Y}_{2u} = \sum_{s_2} y_{2i}/(up_i),\ \hat{Y}'_2 = \hat{Y}_1 + \sum_{s_1} (y_{2i} - y_{1i})/(mp_i),$$

et Q^* est un poids, $0 \leq Q^* \leq 1$.

La variance de $\hat{Y}^*_{2r}$ a été réduite au minimum à partir de l'hypothèse selon laquelle l'expression

$$v_{pps}(y_t) = \sum_{i=1}^{N} p_i(y_{ti}/p_i - Y_t)^2$$

est identique pour les temps t = 1 et t = 2.

Les ouvrages statistiques ont prêté une attention considérable au problème de l'échantillonnage à une reprise avec PPTSR. Une des difficultés principales est la spécification de méthodes pratiques pour obtenir certaines probabilités données à chaque tirage. Fellegi (1963) a proposé une technique où la probabilité de sélectionner l'unité i à chacun des n tirages est égale à p_i si l'on détermine n - 1 ensembles de "probabilités de travail". Ce procédé est extrêmement important pour le renouvellement d'échantillons, où il est essentiel que l'estimateur habituel de Y_2 pour les échantillons constitués avec PPT ne soit pas biaisé; cet objectif ne sera pas atteint si p_i n'est pas constant à chacun des n tirages d'un plan d'échantillonnage avec renouvellement partiel. Jusqu'à récemment, toutefois, les calculs nécessaires pour la technique de

Fellegi étaient beaucoup trop compliqués pour $n > 2$. Choudhry (1981) a conçu une méthode itérative pour mettre en oeuvre la technique de Fellegi et il a élaboré un programme informatique qui évalue les probabilités de travail pour $n \leq 5$. Même si la convergence est rapide du point de vue du nombre d'itérations, le nombre de calculs augmente par une proportion égale à N^n. Le programme indique aussi pour le calcul de la variance la probabilité composée que les unités i et j soient incluses en même temps dans l'échantillon.

Rao, Hartley et Cochran (1962) ont mis au point la "méthode des groupes aléatoires" pour la sélection d'un échantillon avec PPTSR. La population de N unités est divisée en n groupes de taille N_1, N_2, ..., N_n , où $\Sigma N_h = N$, et un échantillon d'une unité est prélevé indépendamment de chaque groupe avec probabilité proportionnelle aux valeurs de p_i. Ghangurde et Rao (1969) ont adapté la méthode des groupes aléatoires à l'échantillonnage à deux reprises. Pour simplifier la description, les N unités sont réparties en n groupes de taille N/n (nombre qu'on suppose entier). Dans un premier temps, une unité est tirée de chaque groupe aléatoire, tel qu'il a été mentionné plus haut, pour former un échantillon s de n unités. À un deuxième moment, un échantillon aléatoire simple s_1 de m unités appariées est sélectionné SR parmi les n unités et un échantillon indépendant s_2 de $u = n - m$ unités est recueilli à partir de l'ensemble de la population par la même méthode appliquée pour produire l'échantillon s. Ghangurde et Rao utilisent un estimateur composite $\hat{Y}'_{2G}$ de Y_2 pour réduire au minimum la variance de cette caractéristique à l'aide d'une valeur optimale du poids Q. Ensuite, la valeur optimale de $\lambda = m/n$ est déterminée. Ces auteurs font remarquer qu'il serait probablement plus efficace de tirer s_2 parmi les N - n unités de la population non incluses dans s.

Chotai (1974) a modifié la méthode de Ghangurde et Rao (G-R) pour l'échantillonnage à la deuxième reprise; les n unités de s sont choisies au hasard pour former m groupes de taille n/m (nombre qu'on suppose entier). Une unité est tirée de chacun des m groupes avec probabilité proportionnelle aux valeurs de p_i, pour constituer un échantillon s_1. Ensuite, un échantillon s_2 est prélevé à l'aide de la méthode G-R. Une fois que la variance optimale de

l'estimateur composite $\hat{Y}'_{2c}$ est calculée, Chotai détermine la valeur optimale de λ et compare l'efficacité relative de $\hat{Y}'_{2c}$ par rapport aux estimateurs optimaux de Ghangurde et Rao et de Raj. Il s'est avéré que $\hat{Y}'_{2c}$ est toujours plus efficace que $\hat{Y}^*_{2R}$ et, dans bien des cas, plus efficace que $\hat{Y}'_{2G}$. Chotai examine brièvement le cas où n/m n'est pas un nombre entier. Il convient de noter que, étant donné que λ n'est pas une fonction continue, la valeur optimale de λ devrait être calculée par des méthodes de programmation en nombres entiers. Dans la section suivante, on présente une méthode d'échantillonnage qui conduit souvent à des résultats plus efficaces que ceux obtenus au moyen des techniques d'échantillonnnage avec PPTSR proposées jusqu'à présent.

2. MÉTHODE D'ÉCHANTILLONNAGE

2.1 Sélection de l'échantillon

Dans la population de N unités (1,2,...,N), choisissons un échantillon qui contient n + u unités (u < n) tirées une à la fois sans remise selon la méthode de Fellegi, où la probabilité de sélectionner la $i^{ième}$ unité à chaque tirage est p_i, i = 1,2,...,N, $\Sigma\, p_i = 1$. Lors du premier échantillonnage, les n premières unités des n + u unités sont examinées, tandis qu'au deuxième échantillonnage, les u premières unités sont supprimées de l'échantillon et les u unités non étudiées la première fois sont incluses. Ainsi, m = n - u unités sont observées dans les deux cas. Les n unités examinées la première fois composent l'échantillon s, les unités étudiées à deux reprises constituent l'échantillon s_1 (où $s_1 \subset s$) et l'ensemble d'unités introduites seulement à la deuxième période représentent l'échantillon s_2. Soulignons que la méthode de Fellegi garantit que la probabilité de sélection d'une unité i est la même à chaque tirage et donc au cours des deux périodes. Chaudhuri (1980) a étudié une sous-catégorie d'estimateurs non biaisés de modèles linéaires non homogènes et il a démontré que le plan d'échantillonnage exposé plus haut conduit à

une méthode optimale. Ce résultat offre une raison de plus d'utiliser la méthode de Fellegi.

2.2 Méthode d'estimation

Dans cette section, on propose des estimateurs composites de Y_2, la valeur globale courante, et on évalue leur variance à l'aide d'une variable nominale.

Posons que ${}_ra_i = 1$ si l'unité i, i = 1,2,...,N est sélectionnée au tirage r, r = 1,2,...,n+u et que ${}_ra_i = 0$ autrement. Comme l'espérance mathématique de ${}_ra_i$ est égale à p_i, un estimateur non biaisé de la valeur globale Y_1 de la population calculée la première fois s'écrit :

$$\hat{Y}_1 = \frac{1}{n} \sum_{r=1}^{n} \sum_{i=1}^{N} a_r a_i y_{1i} / p_i .$$

Donc, $$\hat{Y}'_2 = \hat{Y}_1 + \frac{1}{m} \sum_{r=u+1}^{n} \sum_{i=1}^{N} a_r a_i (y_{2i} - y_{1i})/p_i$$

est un estimateur non biaisé de la valeur globale Y_2 obtenue la deuxième fois. Un estimateur non biaisé de Y_2 qui est une fonction des observations recueillies à la deuxième reprise est

$$\hat{Y}_2 = \frac{1}{n} \sum_{r=u+1}^{u+n} \sum_{i=1}^{N} a_r a_i y_{2i} / p_i$$

Un estimateur composite de Y_2 correspond à la moyenne pondérée

$$\hat{Y}_{2c} = Q\hat{Y}'_2 + (1-Q)\hat{Y}_2,$$

où $0 \leq Q \leq 1$.

La variance de $\hat{Y}_{2c}$, $\mathrm{Var}(\hat{Y}_{2c}) = Q^2\mathrm{Var}(\hat{Y}'_2) + (1-Q)^2\mathrm{Var}(\hat{Y}_2) + 2Q(1-Q)\mathrm{Cov}(\hat{Y}'_2,\hat{Y}_2)$,

est le résultat des propriétés suivantes de la variable nominale ${}_r a_i$:

$$\text{Var}({}_r a_i) = p_i\,(1 - p_i),\ (i = 1,2,\ldots,\ N,\ r = 1,2,\ldots,n+u),$$

$$\text{Cov}({}_r a_i,\ {}_t a_i) = -p_i^2,\ (r \neq t),$$

$$\text{Cov}({}_r a_i,\ {}_r a_j) = -p_i p_j,\ (i \neq j),$$

$$\text{Cov}({}_r a_i,\ {}_t a_j) = E({}_r a_i \cdot {}_t a_j) - p_i p_j,\ \text{autrement},$$

où $E(.)$ correspond à l'espérance mathématique de l'expression entre parenthèses dans le plan d'échantillonnage probabiliste. Or $E({}_r a_i,\ {}_t a_j) = P({}_r a_i \cdot {}_t a_j = 1) = P({}_r a_i = 1,\ {}_t a_j = 1)$, où $P(.)$ représente une probabilité.

Soit $\Sigma_{(k-2;\ i,j)}$ la sommation pour toutes les valeurs possibles des (k-2)-uples formées de différentes unités $\{i_1, i_2, \ldots, i_{r-1}, i_{r+1}, \ldots, i_{k-2}, i_{k-1}\}$ incluses dans l'échantillon lors des k premiers tirages à partir des N-2 unités de l'ensemble $\{1,2,\ldots,i-1,i+1,\ldots,j-1,j+1,\ldots,N\}$, où la $i^{\text{ième}}$ unité est sélectionnée au tirage r et la $j^{\text{ième}}$ unité au tirage k. Cette sommation comporte $(N-2)(N-3)\ldots(N-k+1)$ termes.

Comme dans l'analyse de Fellegi (1963), définissons $\{p_i(\ell);\ i=1,2,\ldots,N\}$ comme étant l'ensemble des "probabilités de travail" pour la sélection d'une unité au tirage ℓ, $\ell = 1,2,\ldots,n+u$. Aux tirages k et r, où $k > r$:

$$E({}_r a_i \cdot {}_k a_j) = \Sigma_{(k-2;i,j)}\ p_{i_1}(1)\ \frac{p_{i_2}(2)}{1-p_{i_1}(2)} \cdots \frac{p_{i_{r-1}}(r-1)}{1 - \sum_{\ell=1}^{r-2} p_{i_\ell}(r-1)}$$

$$\times\ \frac{p_i(r)}{1 - \sum_{\ell=1}^{r-1} p_{i_\ell}(r)} \times \frac{p_{i_{r+1}}(r+1)}{1 - \sum_{\ell=1}^{r-1} p_{i_\ell}(r+1) - p_i(r+1)} \times$$

$$\ldots \times \frac{p_j(k)}{1 - \sum_{\ell=1}^{r-1} p_{i_\ell}(k) - p_i(k) - \sum_{\ell=r+1}^{k-1} p_{i_\ell}(k)}$$

Or

$$\mathrm{Var}\,(\hat{Y}'_2) = \frac{1}{n^2}\,\mathrm{Var}\Big[\sum_{r=1}^{n}\sum_{i=1}^{N} {}_r a_i y_{1i}/p_i\Big] + \frac{1}{m^2}\,\mathrm{Var}\Big[\sum_{r=u+1}^{n}\sum_{i=1}^{N} {}_r a_i (y_{2i}-y_{1i})/p_i\Big]$$

$$+ \frac{2}{mn}\,\mathrm{Cov}\Big[\sum_{r=1}^{n}\sum_{i=1}^{N} {}_r a_i y_{1i}/p_i,\ \sum_{r=u+1}^{n}\sum_{i=1}^{N} {}_r a_i (y_{2i} - y_{1i})/p_i\Big]$$

À l'aide des propriétés des variables nominales ${}_r a_i$ présentées plus haut, on peut démontrer que :

$$\frac{1}{n^2}\,\mathrm{Var}\Big[\sum_{r=1}^{n}\sum_{i=1}^{N} {}_r a_i y_{1i}/p_i\Big] = \frac{1}{n}\sum_{i=1}^{N} p_i z_{1i}^2 + \frac{1}{n^2}\sum_{i\neq j}\sum P(i,j\varepsilon s) z_{1i} z_{1j} - Y_1^2 \ ,$$

$$\frac{1}{m^2}\,\mathrm{Var}\Big[\sum_{r=u+1}^{n}\sum_{i=1}^{N} {}_r a_i (y_{2i} - y_{1i})/p_i\Big] = \frac{1}{m}\sum_{i=1}^{N} p_i (z_{2i} - z_{1i})^2$$

$$+ \frac{1}{m^2}\sum_{i\neq j}\sum P(i,j\ \varepsilon s_1)(z_{2i} - z_{1i})(z_{2j} - z_{1j}) - (Y_2 - Y_1)^2 \ ,$$

où $z_{tj} = y_{tj}/p_j$, $t=1,2$ et $n-u=m$.

En outre,

$$\frac{1}{mn}\,\mathrm{Cov}\Big[\sum_{r=1}^{n}\sum_{i=1}^{N} {}_r a_i y_{1i}/p_i,\ \sum_{u+1}^{n}\sum_{1}^{N} {}_r a_i (y_{2i} - y_{1i})/p_i\Big]$$

$$= \frac{1}{n}\sum_{1}^{N} p_i z_{1i}(z_{2i} - z_{1i}) + \frac{1}{mn}\sum_{i\neq j}\sum P(i\varepsilon s,\ j\varepsilon s_1) z_{1i}(z_{2j} - z_{1j}) - Y_1(Y_2 - Y_1).$$

Si l'on substitue les trois expressions présentées ci-dessus dans l'équation de $\mathrm{Var}(\hat{Y}_2')$, on obtient la formule :

$$\mathrm{Var}(\hat{Y}_2') = \sum_1^N p_i \left[\frac{z_{2i}^2}{n} + (z_{2i}-z_{1i})^2 \left(\frac{1}{m} - \frac{1}{n}\right)\right] + \sum_{i\neq j}\sum \left[\frac{P(i,j\epsilon s)}{n^2} z_{1i}\, z_{1j} \right.$$

$$+ \frac{P(i,j\ \epsilon s_1)}{m^2}(z_{2i}-z_{1i})(z_{2j} - z_{1j})$$

$$\left. + \frac{2P(i\epsilon s, j\epsilon s_1)}{nm} z_{1i}\ (z_{2j}-z_{1j})\right] - Y_2^2 \ . \qquad (1)$$

En outre,

$$\mathrm{Var}(\hat{Y}_2) = \frac{1}{n}\sum_i p_i z_{2i}^2 + \frac{1}{n^2} \sum_{i\neq j}\sum P(i,j\epsilon s^*) z_{2i} z_{2j} - Y_2^2, \qquad (2)$$

où s* est l'ensemble des n unités observées à la dexième reprise et

$$\mathrm{Cov}(\hat{Y}_2', \hat{Y}_2) = \sum_{i\neq j}\sum \left[\frac{P(i\epsilon s,\ j\epsilon s^*)}{n^2} z_{1i} z_{2j} + \frac{P(i\epsilon s_1,\ j\epsilon s^*)}{nm} (z_{2i} - z_{1i}) z_{2j}\right]$$

$$+ \frac{1}{n}\sum_i p_i z_{2i}\ (z_{2i} - \frac{u}{n} z_{1i}) - Y_2^2 \qquad (3)$$

Si l'on combine les expressions (1), (2) et (3), on obtient $\mathrm{Var}(\hat{Y}_{2c})$.

La valeur optimale du poids Q qui réduit au minimum $\mathrm{Var}(\hat{Y}_{2c})$ est

$$Q_{opt} = [\mathrm{Var}(\hat{Y}_2) - \mathrm{Cov}(\hat{Y}_2', \hat{Y}_2)] / [(\mathrm{Var}(\hat{Y}_2') + \mathrm{Var}(\hat{Y}_2) - 2\ \mathrm{Cov}(\hat{Y}_2',\ \hat{Y}_2)] .$$

La variance minimum ainsi obtenue est

$$\mathrm{Var}(\hat{Y}_{2c})=[\mathrm{Var}(\hat{Y}_2').\mathrm{Var}(\hat{Y}_2)-(\mathrm{Cov}(\hat{Y}_2',\hat{Y}_2))^2]/[\mathrm{Var}(\hat{Y}_2')+\mathrm{Var}(\hat{Y}_2)-2\ \mathrm{Cov}(\hat{Y}_2',\hat{Y}_2)]\,.$$

Un autre estimateur composite $\hat{Y}^*_{2c}$ de Y_2 est

$$\hat{Y}^*_{2c} = Q^*\hat{Y}_2' + (1-Q^*)\hat{Y}_{2u},$$

où

$$\hat{Y}_{2u} = \sum_{r=n+1}^{n+u} \sum_{i=1}^{N} r^{a_i} y_{2i}/(up_i).$$

Pour calculer la variance de $\hat{Y}^*_{2c}$, on utilise la formule (1) et les expressions suivantes :

$$\mathrm{Var}(\hat{Y}_{2u})= \frac{1}{u} \sum_i p_i z_{2i}^2 + \frac{1}{u^2} \sum_{i\neq j} \sum P(i,j \in s_2) z_{2i} z_{2j} - Y_2^2 \ ,$$

$$\mathrm{Cov}(\hat{Y}_2', \hat{Y}_{2u}) = \frac{1}{nu} \sum_{i\neq j} \sum P(i \in s, j \in s_2)\ z_{1i} z_{2j}$$

$$+ \frac{1}{mu} \sum_{i\neq j} \sum P(i \in s_1, j \in s_2)(z_{2i} - z_{1i})\ z_{2j} - Y_2^2 \ .$$

2.3 Cas spécial

Pour vérifier les résultats des calculs précédents, prenons le cas de l'échantillonnage aléatoire simple sans remise.

Ainsi, $\hat{Y}_2' = N(\bar{y}_1 + (\bar{y}_{2m} - \bar{y}_{1m}))$, où $\bar{y}_1$ et $\bar{y}_{1m}$ sont les moyennes de l'échantillon calculées respectivement pour toutes les unités échantillonnées et pour toutes les unités appariées à l'échantillon précédent et $\bar{y}_{2m}$ est la moyenne de l'échantillon calculée pour toutes les unités appariées de l'échantillon courant. *Une évaluation directe conduit au résultat :*

$$\mathrm{Var}(\hat{Y}_2') = N^2 \left[\left(\frac{1}{m} - \frac{1}{n}\right) (S_1^2 - 2S_{12}) + \left(\frac{1}{m} - \frac{1}{N}\right) S_2^2 \right]$$

où, par exemple :

$$S_{12} = \sum_{i=1}^{N} (y_{1i} - \bar{Y}_1)(y_{2i} - \bar{Y}_2)/(N-1).$$

Cette expression concorde avec l'équation (1) dans le cas où $p_i = 1/N$ et $P(i,j\epsilon s) = n(n-1)/N(N-1)$ $(i \neq j)$.

De plus, dans l'échantillonnage aléatoire simple, $\hat{Y}_2 = N\bar{y}_2$ (où $\bar{y}_2$ est la moyenne de l'échantillon calculée à partir de toutes les n unités de l'échantillon constitué à la deuxième reprise) et la variance de $\hat{Y}_2$ est

$$\mathrm{Var}(\hat{Y}_2) = N(N-n)S_2^2 .$$

Une évaluation de Var $(\hat{Y}_2)$ au moyen de l'expression (2) produit le même résultat. Enfin, on peut obtenir l'expression suivante soit par une évaluation directe ou à partir de l'équation (3) :

$$\mathrm{Cov}(\hat{Y}_2', \hat{Y}_2) = -NS_2^2$$

De façon semblable, on peut calculer Var $(\hat{Y}_{2c}^*)$.

3. EXEMPLES QUANTITATIFS

On compare ici l'efficacité des estimateurs composites $\hat{Y}_{2c}$ et $\hat{Y}_{2c}^*$, auxquels correspond une valeur optimale, Q et Q* respectivement, qui réduit au minimum la variance de l'estimation, et celle de l'estimateur $\hat{Y}_2$ calculé à partir des renseignements courants recueillis auprès d'un échantillon sélectionné avec

PPT. Comme on ne dispose pas d'une forme fermée de $\mathrm{Var}(\hat{Y}_{2c})$ et de $\mathrm{Var}(\hat{Y}^*_{2c})$ pour permettre des comparaisons analytiques, on a utilisé de petites populations pour obtenir des valeurs de variables et mesurer les contrastes entre les estimateurs. (Les populations étudiées devaient être de petite taille, comme celles qu'on trouve souvent dans les échantillons stratifiés, puisque l'effet de l'échantillonnage avec ou sans remise se voit seulement quand les fractions de sondage sont non négligeables.) Quatre plans d'échantillonnage avec renouvellement ont été appliqués à chaque population : (n,m) = (2,1), (3,2), (3,1) et (4,3), où n est le nombre d'unités de l'échantillon à chaque reprise et m est le nombre d'unités incluses dans l'échantillon à la première et à la deuxième reprise. Deux populations proviennent de l'ouvrage de Murthy (1967), qui a divisé une seule population de 34 villages en deux populations de 16 et de 17 villages (une unité extrême a été exclue). La caractéristique x qui sert à mesurer la taille est le nombre d'acres cultivés en 1961; y_1 et y_2 sont le nombre d'acres de blé en 1963 et en 1964 respectivement. Une troisième population est un ensemble de 14 fermes dans la province de la Saskatchewan où x est le nombre d'acres des fermes en 1980 et y_1 et y_2 sont le nombre d'acres ensemencés en 1980 et en 1981 respectivement. On analyse aussi deux autres ensembles de données recueillies parmi une population de 15 unités et une autre de 16 unités.

Le tableau 1 montre l'efficacité relative de $\hat{Y}_{2c}$ et $\hat{Y}^*_{2c}$ par rapport à $\hat{Y}_2$ pour chacune des cinq populations et les quatre plans d'échantillonnage. Un paramètre très important dans chaque comparaison est la correlation ρ_z entre $z_{1i} = y_{1i}/p_i$ et $z_{2i} = y_{2i}/p_i$:

$$\rho_z = \frac{\sum_{i=1}^{N} p_i z_{1i} z_{2i} - Y_1 Y_2}{\sqrt{\sum_{i=1}^{N} p_i z_{1i}^2 - Y_1^2}\ \sqrt{\sum_{i=1}^{N} p_i z_{2i}^2 - Y_2^2}}\ .$$

Les valeurs de ρ_z calculées pour les populations étudiées varient entre

0.940 et 0.213. Le tableau 1 indique également les valeurs optimales de Q et Q*. On prévoit effectuer une étude ultérieure qui portera sur l'efficacité de $\hat{Y}_{2c}$ et $\hat{Y}^*_{2c}$ lorsque les valeurs choisies pour Q et Q* ne sont pas optimales.

Les conclusions suivantes se dégagent de ces données empiriques: 1) La valeur optimale de Q est généralement plus grande quand ρ_z est grand et que la valeur optimale de Q diminue, que l'estimateur soit $\hat{Y}_{2c}$ ou $\hat{Y}^*_{2c}$. 2) La valeur optimale de Q pour Y^*_{2c} dépasse toujours celle de Q pour $\hat{Y}_{2c}$. 3) Lorsque ρ_z diminue, l'efficacité de $\hat{Y}_{2c}$ par rapport à $\hat{Y}_2$ diminue aussi (tel que prévu) et se rapproche de l'unité comme limite inférieure quand la valeur de Q est optimale. Par contre, ce genre de phénomène ne s'observe pas clairement dans le cas de $\hat{Y}^*_{2c}$ puisque $\text{Var}(\hat{Y}^*_{2c})$ n'est pas une fonction monotone de ρ_z. Pour de petites valeurs de ρ_z, on enregistre de faibles gains et de faibles pertes d'efficacité par rapport à $\hat{Y}_2$. 4) $\hat{Y}^*_{2c}$ est plus efficace que $\hat{Y}_{2c}$ quand la valeur de ρ_z est élevée, alors que $\hat{Y}_{2c}$ est plus efficace que $\hat{Y}^*_{2c}$ quand la valeur de ρ_z est basse. 5) Si $\lambda = m/n$ est faible, comme dans le plan (3,1), on peut réaliser d'importants gains d'efficacité si l'on utilise $\hat{Y}^*_{2c}$ au lieu de $\hat{Y}_2$ quand la valeur de ρ_z est élevée. Quand ρ_z est faible, le plan (4,3) conduit aux meilleurs gains d'efficacité si l'on se sert de $\hat{Y}_{2c}$; les résultats des trois autres plans d'échantillonnage sont à peu près les mêmes.

Il convient de noter que même s'il existe une bonne corrélation entre les valeurs de y_{1i} et y_{2i}, l'estimateur composite $\hat{Y}^*_{2c}$ peut toujours causer des pertes d'efficacité en comparaison de l'estimateur $\hat{Y}_2$ calculée à partir de données recueillies à une seule reprise au moyen de l'échantillonnage avec PPT. Le facteur critique est la corrélation ρ_z entre les valeurs de z_{1i} et z_{2i}. La meilleure solution est d'utiliser $\hat{Y}_{2c}$ avec la valeur optimale de Q puisque $\hat{Y}_{2c} = \hat{Y}_2$ quand $Q = 0$.

Le tableau 2 indique l'efficacité relative de $\hat{Y}_{2c}$ et $\hat{Y}^*_{2c}$ dans un plan d'échantillonnage avec PPTSR par rapport à l'estimateur $\hat{Y}^*_{2R}$ utilisé par Raj (1965) dans son plan d'échantillonnage avec PPTAR décrit plus haut. Pour que les comparaisons soient plus valables, on n'a pas supposé que $V_{ppt}(y_t)$ était égal aux deux reprises t = 1 et t = 2; on s'est servi des valeurs optimales de Q* pour chaque couple (n,m). Dans tous les cas, tel que prévu, les estimateurs $\hat{Y}_{2c}$ et $\hat{Y}^*_{2c}$ dans le plan d'échantillonnage SR s'avèrent plus efficaces que $\hat{Y}^*_{2R}$ dans le plan de Raj. Pour une valeur donnée de ρ_z, plus n est grand, plus le gain d'efficacité entraîné par l'échantillonnage SR est élevé. Enfin, on constate que Raj a réalisé des gains d'efficacité par rapport aux échantillons sans appariement seulement quand $\rho_z > 0.5$, alors que qu'on réalise toujours des gains d'efficacité si l'on utilise $\hat{Y}_{2c}$ dans l'échantillonnage SR, quelle que soit la valeur de ρ_z.

REMERCIEMENTS

Nous remercions le Professeur J.N.K. Rao pour ses suggestions et ses observations au cours de cette étude, ainsi que l'arbitre pour ses remarques pertinentes.

BIBLIOGRAPHIE

[1] CHAUDHURI, A. (1980). "On Optimal and Related Strategies for Sampling on Two Occasions with Varying Probabilités", prétirage, Indian Statistical Institute.

[2] CHOTAI, J. (1974). "A Note on the Rao-Hartley-Cochran Method for PPS Sampling Over Two Occasions", Sankhya 36, C, 173-180.

[3] CHOUDHRY, G.H. (1981). "Construction of Working Probabilities and Joint Selection Probabilities for Fellegi's PPS Sampling Scheme", Techniques d'enquête 7, no. 1, 93-108.

[4] FELLEGI, I.P. (1963). "Sampling With and Without Replacement: Rotating and Non-rotating Samples", J. Amer. Stat. Ass. 58, 183-201.

[5] GHANGURDE, P.D. et RAO, J.N.K. (1969). "Some Results on Sampling Over Two Occasions", Sankhya 31, A, 463-472.

[6] MURTHY, M.N. (1967). Sampling Theory and Methods. Calcutta : Statistical Publishing Society.

[7] RAJ, D. (1965). "On Sampling Over Two Occasions With Probability Proportional to Size", Ann. Math. Statist. 36, 327-330.

[8] RAO, J.N.K., HARTLEY, H.O. et COCHRAN, W.G. (1962). "On a Simple Procedure of Unequal Probability Sampling Without Replacement", J.R. Statist. Soc. B. 24, no. 2, 482-491.

TABLEAU 1

Efficacité des estimateurs composites des plans d'échantillonnage SR par rapport à celle de l'estimateur du plan d'échantillonnage avec PPTSR

Population	N	(n,m)	ρ_z	Q_{opt}	$\frac{Var(\hat{Y}_2)}{Var(\hat{Y}_{2c})}$	Q^*_{opt}	$\frac{Var(\hat{Y}_2)}{Var(\hat{Y}^*_{2c})}$
Murthy	17	(2,1)	0.940	0.443	1.337	0.640	1.476
Groupe 1		(3,2)		0.456	1.230	0.738	1.367
		(3,1)		0.419	1.524	0.545	1.656
		(4,3)		0.462	1.188	0.793	1.309
Murthy	16	(2,1)	0.867	0.377	1.205	0.615	1.364
Groupe 2		(3,2)		0.402	1.151	0.727	1.296
		(3,1)		0.336	1.282	0.499	1.459
		(4,3)		0.431	1.191	0.786	1.255
Nombre	14	(2,1)	0.546	0.181	1.083	0.466	0.927
d'acres		(3,2)		0.215	1.070	0.646	0.927
		(3,1)		0.137	1.092	0.295	0.933
		(4,3)		0.279	1.117	0.736	0.931
Ensemble de	15	(2,1)	0.392	0.113	1.019	0.506	1.013
données 1		(3,2)		0.140	1.017	0.670	1.013
		(3,1)		0.082	1.020	0.340	1.013
		(4,3)		0.330	1.170	0.752	1.012
Ensemble de	16	(2,1)	0.213	0.061	1.007	0.451	0.898
données 2		(3,2)		0.078	1.007	0.636	0.896
		(3,1)		0.042	1.007	0.280	0.908
		(4,3)		0.285	1.142	0.730	0.901

TABLEAU 2

Efficacité des estimateurs composites des plans d'échantillonnage SR par rapport à celle de l'estimateur composite de Raj

Population	N	(n,m)	ρ_z	$Var(\hat{Y}^*_{2R})/Var(\hat{Y}_{2c})$	$Var(\hat{Y}^*_{2R})/Var(\hat{Y}^*_{2c})$
Murthy	17	(2,1)	0.940	1.038	1.146
Groupe 1		(3,2)		1.127	1.252
		(3,1)		1.215	1.320
		(4,3)		1.248	1.375
Murthy	16	(2,1)	0.867	1.001	1.133
Groupe 2		(3,2)		1.106	1.246
		(3,1)		1.130	1.287
		(4,3)		1.309	1.380
Nombre	14	(2,1)	0.546	1.244	1.065
d'acres		(3,2)		1.330	1.151
		(3,1)		1.351	1.154
		(4,3)		1.507	1.257
Ensemble de	15	(2,1)	0.392	1.095	1.089
données 1		(3,2)		1.197	1.192
		(3,1)		1.200	1.192
		(4,3)		1.522	1.317
Ensemble de	16	(2,1)	0.213	1.197	1.068
données 2		(3,2)		1.293	1.152
		(3,1)		1.280	1.154
		(4,3)		1.589	1.254