Techniques d'enquête, décembre 1992
Vol. 18, n° 2, pp. 269–278
Statistique Canada

Une méthode de répartition de l'échantillon pour des plans d'échantillonnage à deux phases

J.B. ARMSTRONG et C.F.J. WU[1]

RÉSUMÉ

Dans la foulée de l'élaboration d'un plan de sondage pour des enquêtes-entreprises à Statistique Canada, nous exposons le problème de la répartition de l'échantillon pour un plan de sondage général à deux phases comme un problème de programmation non linéaire sous contrainte. En exploitant la structure mathématique du problème, nous proposons une solution qui consiste en des itérations entre deux sous-problèmes qui sont beaucoup moins complexes sur le plan du calcul. Lorsqu'on utilise une solution approximative comme valeur de départ, la méthode proposée donne des résultats très satisfaisants dans une étude empirique.

MOTS CLÉS: Répartition optimale; programmation convexe.

1. INTRODUCTION

Le but de cet article est de proposer une méthode de répartition de l'échantillon pour des plans de sondage à deux phases. Supposons qu'il faille répartir en L strates une population de taille N en fonction d'une variable auxiliaire z dont nous ne connaissons pas la valeur avant l'échantillonnage. En revanche, nous connaissons la valeur d'une deuxième variable auxiliaire (taille), x, qui est corrélée avec la variable étudiée, y, pour toutes les unités de la population. Dans la première phase de l'échantillonnage, la population est divisée en G strates selon x. On tire tout d'abord un échantillon aléatoire simple de la strate g ($g = 1, 2, \ldots, G$) avec comme fraction de sondage v_g, et on observe la valeur z pour chaque unité échantillonnée. Dans la seconde phase, on tire un nouvel échantillon parmi les unités qui forment l'échantillon de la strate g et pour lesquelles la valeur z appartient à la classe h ($h = 1, 2, \ldots, L$); la fraction de sondage est alors v_{gh}. On observe ensuite la valeur y pour les unités de l'échantillon de la seconde phase.

Dans le cas où il n'y a pas de stratification en fonction de la taille (c.-à-d. $G = 1$), Cochran (1977) définit la formule de répartition qui réduit au minimum la variance de l'estimation $\hat{Y} = \sum_h \sum_{i \in s2 \cap h} y_i/(v \cdot v_h)$ du total de population $Y = \sum_h N_h \cdot \bar{Y}_h$, étant donné un coût fixe, C, pour l'enquête, N_h et $\bar{Y}_h$ étant, respectivement, la taille et la moyenne de la population pour la strate h et $\sum_{i \in s2 \cap h} y_i$ désignant la somme des valeurs de y pour les unités de l'échantillon de la seconde phase, $s2$, pour lesquelles la valeur z est incluse dans la classe h. Si les estimations d'enquête servent à l'analyse, la variance du total estimé pour la classe h, $\hat{Y}_h = \sum_{i \in s2 \cap h} y_i/(v \cdot v_h)$, mérite aussi d'être considérée. Sedransk (1965), Booth et Sedransk (1969), Rao (1973) et Smith (1989) se penchent sur des problèmes de répartition où l'on minimise une fonction de variances de totaux de classe estimés, étant donné une contrainte de coût.

La méthode que nous décrivons dans cet article peut servir à résoudre le problème de la répartition dans tous les cas de stratification où la variance du total estimé pour chaque classe z est assujettie à une contrainte. L'idée de cette méthode nous est venue à la suite d'une expérimentation faite au cours d'une enquête-entreprises de Statistique Canada, où étaient échantillonnés des dossiers fiscaux d'entreprises.

[1] J.B. Armstrong, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa (Ontario) K1A 0T6 et C.F.J. Wu, Department of Statistics and Acturial Science, University of Waterloo, Waterloo (Ontario) N2L 3G1.

Statistique Canada se procure des renseignements sur la population des déclarants auprès de Revenu Canada. L'organisme statistique est tenu de produire des estimations de variables financières pour des domaines définis par le croisement des provinces et des codes à quatre chiffres de la Classification type des industries (CTI4). Or, à Revenu Canada, seuls les codes à deux chiffres offrent un niveau de précision acceptable. Afin de normaliser le niveau de précision des estimations pour les domaines définis par le croisement des "CTI4" et des provinces, on a mis en application un plan d'échantillonnage à deux phases. Dans la première phase, on prélève un échantillon de déclarants à Revenu Canada dans des strates définies en fonction du code à deux chiffres de la CTI et du revenu brut de l'entreprise (taille). Avant de passer à la seconde phase de l'échantillonnage, Statistique Canada attribue un code à quatre chiffres de la CTI à chaque unité échantillonnée; ce code est réputé plus précis que ceux fournis par Revenu Canada. Dans la seconde phase de l'échantillonnage, on se sert de strates définies en fonction du code à quatre chiffres de la CTI et de la taille de l'entreprise. Les limites de classe sont les mêmes dans les deux phases. Pour une description détaillée du plan d'échantillonnage, le lecteur est prié de se référer à Choudhry, Lavallée et Hidiroglou (1989b).

L'échantillon de la première phase est formé suivant un échantillonnage de Bernoulli (appelé aussi échantillonnage de Poisson). Supposons que le déclarant i d'une case particulière formée par le croisement des provinces et des codes à deux chiffres de la CTI (CTI2) appartient à la strate de première phase g. Pour déterminer si ce déclarant appartient à l'échantillon de première phase, on génère un nombre pseudo-aléatoire compris dans l'intervalle (0,1), disons R_i, en se servant du numéro d'identification unique du déclarant. Le déclarant sera inclus dans l'échantillon de première phase si $R_i \in (0, v_g)$. Pour ce qui est du tirage de l'échantillon de la seconde phase, on se sert aussi d'un échantillonnage de Bernoulli, mais avec un ensemble de nombres pseudo-aléatoires différent. L'échantillonnage de Bernoulli permet de procéder à l'échantillonnage et au traitement avant d'avoir toute l'information voulue sur l'univers du déclarant. Cet avantage est précieux car la collecte d'informations s'échelonne normalement sur deux ans. Avec l'échantillonnage de Bernoulli, la taille des échantillons est aléatoire. Dans Choudhry, Lavallée et Hidiroglou (1989b), on calcule la variance de $\hat{Y}_{h-\mathrm{STRAT}} = \sum_g \sum_{i \in s2 \cap g \cap h} y_i / (v_g \cdot v_{gh})$ au moyen d'un échantillon aléatoire simple qui tient lieu d'échantillon de Bernoulli (voir Sunter 1986). Suivant cette méthode, on prélève à la première phase un échantillon aléatoire simple de taille fixe $n'_g = v_g \cdot N_g$ dans la strate g. Désignons par n'_{gh} le nombre d'unités de la strate g (taille) dans l'échantillon de première phase qui font partie de la strate h (CTI4). Dans la seconde phase, on prélève un échantillon aléatoire simple de taille $n_{gh} = v_{gh} \cdot n'_{gh}$ dans la strate h et la strate g, v_{gh} étant considérée fixe. La variance de $\hat{Y}_{h-\mathrm{STRAT}}$ est définie

$$V_h = \sum_g \left(\frac{1}{v_g \cdot v_{gh}} - 1 \right) \cdot A_{gh} + \sum_g \left(\frac{1}{v_g} - 1 \right) \cdot B_{gh},$$

où

$$A_{gh} = N_{gh} \cdot S^2_{gh},$$

$$B_{gh} = \left(\frac{N_g - N_{gh}}{N_g - 1} \right) \cdot \left(\frac{Y^2_{gh}}{N_{gh}} - S^2_{gh} \right),$$

et S^2_{gh} est la variance dans la strate de seconde phase gh (taille $\times$ CTI4).

Voici comment est structuré cet article. Dans la section 2, nous exposons le problème de la répartition optimale par rapport à l'échantillonnage à deux phases de dossiers fiscaux. Nous proposons une solution itérative appelée "méthode exacte". Dans la section 3, nous décrivons une version approchée de la répartition optimale qui peut servir à générer des valeurs de départ pour la méthode exacte. La section 4 contient les résultats d'une étude empirique où l'on compare diverses valeurs de départ pour la méthode exacte. Enfin, la section 5 sert de conclusion.

2. MÉTHODE EXACTE

Dans cette section, nous exposons le problème de la répartition optimale et nous lui opposons une solution itérative appelée "méthode exacte". Pour exposer ce problème par rapport à l'échantillonnage à deux phases de dossiers fiscaux, il suffit de considérer une case formée par le croisement des codes à deux chiffres de la CTI (CTI2) et d'une province en particulier qui contient N unités. Le coût de l'échantillonnage d'une unité dans la première phase est K_1, quelle que soit la strate dans laquelle se trouve cette unité, tandis qu'il est de K_2 dans la seconde phase, peu importe la strate. Suivant l'échantillonnage de Bernoulli, la fonction de coût est

$$F^* = K_1 \cdot \sum_g n'_g + K_2 \cdot \sum_g \sum_h n_{gh}.$$

Puisque la taille des échantillons, n'_g et n_{gh}, est aléatoire, nous utilisons l'espérance de coût

$$F = K_1 \cdot \sum_g v_g \cdot N_g + K_2 \cdot \sum_g \sum_h v_g \cdot v_{gh} \cdot N_{gh}. \tag{1}$$

Rao (1973) et Smith (1989) se servent aussi de l'espérance mathématique de fonctions de coût aléatoires pour résoudre des problèmes de répartition dans des plans à deux phases. Pour ce qui a trait à l'échantillonnage de dossiers fiscaux, le coût total pour une province équivaut à la somme des coûts de l'échantillonnage pour toutes les cases "CTI2" dans la province. Le coefficient de variation estimé du coût de l'échantillonnage à deux phases de dossiers fiscaux pour la province de Québec, fondé sur des données de 1988, était environ 1.85%. Ce coefficient était moindre pour le coût de l'échantillonnage à l'échelle nationale.

Il est nécessaire de minimiser (1) par rapport à v_g, $g = 1, 2 \ldots, G$, et à v_{gh}, $g = 1, 2, \ldots, G$, $h = 1, 2, \ldots, H$, étant donné les contraintes

$$\sum_g \left(\frac{1}{v_g \cdot v_{gh}} - 1 \right) \cdot A_{gh} + \sum_g \left(\frac{1}{v_g} - 1 \right) \cdot B_{gh} \leq C_h^2 \cdot Y_h^2, \quad h = 1, 2, \ldots, H, \tag{2}$$

$$0 < v_g \leq 1, \quad g = 1, 2, \ldots, G,$$

$$0 < v_{gh} \leq 1, \quad g = 1, 2, \ldots, G, \quad h = 1, 2, \ldots, H,$$

où C_h désigne le coefficient de variation cible pour le domaine h (CTI4).

Lorsqu'on a tenté de résoudre directement le problème en appliquant l'algorithme de programmation quadratique successive de Schittkowski (1985) à la manière d'IMSL (1987), on a obtenu des résultats variables. L'algorithme fonctionnait bien pour des problèmes qui renfermaient un petit nombre de variables et de contraintes. Cependant, il n'a pas été possible d'obtenir des solutions satisfaisantes pour les problèmes où il y avait plus de 35 variables ou plus de 50 contraintes environ.

Le tableau 1 contient quelques chiffres de coût obtenus par l'application directe de l'algorithme de Schittkowski en ce qui concerne l'échantillonnage de dossiers fiscaux. Cette opération visait à résoudre les problèmes de répartition liés à certaines cases "CTI2" dans la province de Québec où il fallait composer avec un grand nombre de variables et de contraintes; à cette fin, nous nous sommes servis de données pour l'année d'imposition 1988. Avec l'approche directe, toutes les fractions de sondage (première phase comme seconde phase) étaient posées égales à un au départ. Le tableau 1 donne aussi le coût (moins élevé que l'autre)

Tableau 1

Résultats pour la méthode directe et la méthode exacte

CTI 2	Nombre de variables	Nombre de constraintes	Coût ($) - méthode directe	Coût ($) - méthode exacte
30	62	86	5155**	1897
35	37	51	551	512
39	38	50	1667	1450
427*	39	48	27528**	3383

* En ce qui à trait aux industries de la construction, on se sert des codes à trois chiffres pour la stratification de la première phase.
** Le programme d'IMSL s'est interrompu à cause d'une erreur interne, qui n'a pu être corrigée après consultation de la documentation.

calculé au moyen de la méthode que nous appelons "méthode exacte" et qui est décrite plus bas. Les données du tableau indiquent que l'application directe de l'algorithme de Schittkowski à la manière d'IMSL est inadéquate pour des cases "CTI2" lorsqu'il y a un grand nombre de variables et de contraintes.

La méthode exacte repose sur une simplification majeure du problème défini par (1) et (2); on peut réaliser cette simplification en exploitant la structure mathématique du problème. Plus précisément, nous reformulons le problème en deux grandes étapes, chacune d'elles pouvant recevoir une solution itérative. Dans la première étape, on minimise (1) par rapport à v_g, $g = 1, 2, \ldots, G$, étant donné la valeur de toutes les fractions de sondage de la seconde phase. Cette étape nécessite l'utilisation de techniques d'optimisation non linéaire. La deuxième étape consiste à minimiser (1) par rapport aux fractions de sondage de la seconde phase, étant donné les fractions de sondage de la première phase calculées dans la première étape. Cette minimisation ne nécessite pas d'itérations puisque sa solution a une forme analytique fermée. En outre, elle peut se faire pour chaque valeur de h prise individuellement, $h = 1, 2, \ldots, H$. Une fois la seconde étape terminée, on répète la première étape et on poursuit le processus itératif. On constate la convergence lorsque la variation de la fonction de coût entre deux itérations devient négligeable.

Posons $v_g^{(i)}$ et $v_{gh}^{(i)}$ comme les estimations des valeurs optimales de v_g et de v_{gh} obtenues après i itérations (chaque itération comprenant une double exécution des deux étapes décrites ci-dessus). Au début de l'itération $i + 1$, il faut transformer les variables définies par l'expression $X_g^{(i+1)} = 1/v_g^{(i+1)} - 1$. Cette opération a pour but de redéfinir le problème d'optimisation qui se pose dans la première étape de l'itération comme un problème renfermant des contraintes linéaires et une fonction-objectif convexe, c'est-à-dire comme un problème de programmation convexe. Les problèmes de ce genre sont plus faciles à résoudre.

Plus précisément, chaque itération comporte les opérations suivantes:

(i) minimisation de

$$F = \sum_g \left(N_g + \frac{K_2}{K_1} \sum_h v_{gh}^{(i-1)} \cdot N_{gh} \right) \Big/ (X_g^{(i)} + 1)$$

par rapport à $X_g^{(i)}$, $g = 1, 2, \ldots, G$, étant donné les contraintes

$$C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{X_g^{(i)} + 1}{v_{gh}^{(i-1)}} - 1 \right) \cdot A_{gh} - \sum_g X_g^{(i)} \cdot B_{gh} \geq 0, \quad h = 1, 2, \ldots, H$$

$$X_g^{(i)} \geq 0, \quad g = 1, 2, \ldots, G.$$

(ii) calcul de $v_g^{(i)} = 1/(X_g^{(i)} + 1)$, $g = 1, 2, \ldots, G$. Minimisation, pour chaque valeur de $h = 1, 2, \ldots, H$, prise individuellement, de

$$F_h = \sum_g v_g^{(i)} \cdot v_{gh}^{(i)} \cdot N_{gh}$$

par rapport à $v_{gh}^{(i)}$, $g = 1, 2, \ldots, G$, étant donné les contraintes

$$C_h^2 \cdot \hat{Y}_h^2 - \sum_g \left(\frac{1}{v_g^{(i)} \cdot v_{gh}^{(i)}} - 1 \right) \cdot A_{gh} - \sum_g \left(\frac{1}{v_g^{(i)}} - 1 \right) \cdot B_{gh} \geq 0,$$

$$0 < v_{gh}^{(i)} \leq 1, \quad g = 1, 2, \ldots, G,$$

où h est considéré comme fixe.

Nous verrons dans la section 3 qu'il n'est pas nécessaire de recourir à des méthodes numériques pour résoudre l'étape (ii). La méthode exacte ne nécessite donc que la résolution d'une série de problèmes de programmation convexe ne renfermant chacun que G variables. Il est beaucoup plus facile de résoudre un problème de programmation convexe qu'un problème de programmation non linéaire général. Dans le premier cas, une solution locale est aussi une solution globale.

Posons $F^{(i)}$ comme la valeur de la fonction de coût, (1), calculée au moyen de $v_g^{(i)}$ et de $v_{gh}^{(i)}$. Les valeurs $F^{(i)}$ forment une suite monotonement décroissante et, de ce fait, tendent vers une limite. C'est la valeur de départ qui détermine si cette limite et les fractions de sondage correspondantes conduisent au minimum global, cela à cause de la géométrie des contraintes définies en (2). Dans la pratique, il faudrait tester plusieurs valeurs de départ en vue de trouver la meilleure solution. Une de ces valeurs nous est fournie par la "méthode approximative", qui ne nécessite pas d'itérations. Cette méthode est décrite dans la section suivante.

3. MÉTHODE APPROXIMATIVE

Dans cette section, nous présentons une méthode de répartition qui se rapproche de la répartition optimale. Elle a été proposée la première fois par Choudhry, Lavallée et Hidiroglou (1989a). Après avoir supposé que toutes les fractions de sondage de la seconde phase sont égales à un, on détermine une formule approchée de répartition optimale pour l'échantillon de la première phase. On répartit ensuite l'échantillon de la seconde phase, compte tenu des fractions de sondage de la première phase. Comme le coût de l'échantillonnage d'une unité dans l'une et l'autre phase ne dépend pas de la strate dans laquelle se trouve l'unité, minimiser le coût revient à minimiser la taille de l'échantillon à chaque étape de la méthode.

La première étape de la méthode approximative consiste à trouver une solution approchée au problème de la répartition optimale pour un plan d'échantillonnage à une phase. Pour cela, on doit calculer le minimum - pour chaque valeur de h prise individuellement - de

$$F^{(h)} = \sum_g v_{g|h} \cdot N_g \tag{3}$$

par rapport à $v_{g|h}$, $g = 1, 2, \ldots, G$. Le symbole $v_{g|h}$ indique que la détermination de la fraction de sondage pour la strate g est assujettie à une seule contrainte de précision, à savoir celle relative au domaine h (CTI4), h étant fixe. En particulier, la minimisation doit s'effectuer eu égard aux contraintes

$$\sum_g \left(\frac{1}{v_{g|h}} - 1\right) \cdot (A_{gh} + B_{gh}) \leq C_h^2 \cdot Y_h^2, \tag{4}$$

$$0 < v_{g|h} \leq 1, \quad g = 1, 2, \ldots, G. \tag{5}$$

On peut montrer que le minimum de (3) est obtenu lorsque l'égalité dans l'équation (4) est satisfaite, de sorte que le problème défini par les équations (3), (4) et (5) équivaut à calculer le point critique du lagrangien

$$L = \sum_g v_{g|h} N_g + \lambda \cdot \left[C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{v_{g|h}} - 1\right) \cdot (A_{gh} + B_{gh})\right].$$

En posant les dérivées par rapport à $v_{g|h}$ égales à zéro, on obtient

$$v_{g|h} = ((A_{gh} + B_{gh})/N_g)^{1/2} \cdot (-\lambda)^{1/2}, \quad g = 1, 2, \ldots, G. \tag{6}$$

En posant $\partial L/\partial\lambda = 0$, on obtient

$$(-\lambda)^{1/2} = \sum_g ((A_{gh} + B_{gh}) \cdot N_g)^{1/2} \Big/ \left(C_h^2 \cdot Y_h^2 + \sum_g (A_{gh} + B_{gh})\right). \tag{7}$$

Après avoir substitué (7) dans (6), on obtient la fraction de sondage optimale pour la strate g, étant donné une seule contrainte de précision, celle relative au domaine h (CTI4),

$$v_{g|h}^* = ((A_{gh} + B_{gh})/N_g)^{1/2} \cdot$$

$$\sum_g ((A_{gh} + B_{gh}) \cdot N_g)^{1/2} \Big/ \left(C_h^2 \cdot Y_h^2 + \sum_g (A_{gh} + B_{gh})\right). \tag{8}$$

Si une ou plusieurs des fractions de sondage définies par l'expression (8) sont supérieures à un, on peut poser ces fractions égales à un et procéder à une répartition avec un nombre moindre de strates. On reconnaît dans cette technique la méthode de "surrépartition" dont fait état Cochran (1977). Il est nécessaire de calculer (8) pour $h = 1, 2, \ldots, H$. La fraction de sondage approximative de la première phase pour la strate g, v_g^*, est posée égale à la valeur la plus élevée de l'ensemble $\{v_{g|h}^*, h = 1, 2, \ldots, H\}$ pour $g = 1, 2, \ldots, G$; de cette manière, nous sommes sûrs que la contrainte de précision relative à chaque domaine "CTI4" sera satisfaite.

Étant donné les fractions de sondage de la première phase, il est facile de calculer les fractions de sondage optimales de la seconde phase. Supposons que, pour la case h formée par le croisement des codes à deux chiffres de la CTI (CTI2) et des provinces, les strates g visées par l'opération de répartition correspondent à un ensemble de nombres entiers, Γ. Nous posons

les fractions de sondage de la seconde phase égales à un pour les strates g qui ne sont pas visées par l'opération. Normalement, on devrait avoir l'ensemble $\Gamma = \{1, 2, \ldots, G\}$, mais à cause du phénomène de "surrépartition" qui survient durant la répartition de l'échantillon de la seconde phase, par exemple, Γ peut ne pas contenir tous les entiers de 1 à G. Répartir l'échantillon de la seconde phase équivaut à déterminer le minimum de

$$F_h = \sum_{g\in\Gamma} v_{gh} \cdot v_g^* \cdot N_{gh} \tag{9}$$

par rapport à v_{gh}, $g\in\Gamma$, étant donné les contraintes

$$\sum_{g\in\Gamma} \left(\frac{1}{v_{gh}} - 1\right) \cdot \frac{A_{gh}}{v_g^*} \leq M_h, \tag{10}$$

$$0 < v_{gh} \leq 1, \quad g\in\Gamma, \tag{11}$$

où

$$M_h = C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{v_g^*} - 1\right) \cdot (A_{gh} + B_{gh}).$$

Notons que le nombre prévu d'unités appartenant au domaine h (CTI4) dans l'échantillon de la seconde phase pour la strate g, $v_g^* \cdot N_{gh}$, figure dans l'équation (9). On peut montrer facilement que (9) prend une valeur minimum lorsque l'égalité dans l'équation (10) est satisfaite. Par conséquent, la minimisation équivaut à déterminer le point critique du lagrangien

$$L_h = \sum_{g\in\Gamma} v_{gh} \cdot v_g^* \cdot N_{gh} + \lambda \cdot \left(M_h - \sum_{g\in\Gamma} \cdot \left(\frac{1}{v_{gh}} - 1\right) \cdot \frac{A_{gh}}{v_g^*}\right),$$

par rapport à v_{gh}, $g\in\Gamma$, et à λ, étant donné les contraintes

$$0 < v_{gh} \leq 1, \quad g\in\Gamma.$$

En posant les dérivées premières de L_h égales à zéro et en simplifiant, on obtient

$$v_{gh} = (-\lambda \cdot A_{gh}/N_{gh})^{1/2} \cdot (1/v_g^*), \quad g\in\Gamma, \tag{12}$$

$$(-\lambda)^{1/2} = \sum_g (N_{gh} \cdot A_{gh})^{1/2}/D_{\Gamma h}, \tag{13}$$

où

$$D_{\Gamma h} = C_h^2 \cdot Y_h^2 \sum_{g\in\Gamma} \left(\frac{1}{v_g^*}\right) \cdot A_{gh} - \sum_g \left(\frac{1}{v_g^*} - 1\right) \cdot (A_{gh} + B_{gh}).$$

Notons qu'une répartition ne sera possible que si $D_{\Gamma h}$ est positif. En substituant (13) dans (12), on obtient

$$v_{gh}^{*} = (A_{gh}/N_{gh})^{1/2} \cdot (1/v_{g}^{*}) \cdot \sum_{g \in \Gamma} (N_{gh} \cdot A_{gh})^{1/2} / D_{\Gamma h}. \tag{14}$$

Si v_{gh}^{*} est supérieure à un pour certaines strates de seconde phase gh, on peut bien sûr utiliser la méthode de "surrépartition" mentionnée plus haut. Notons que l'équation (14) produit aussi la solution voulue pour l'étape (ii) de chaque itération de la méthode exacte.

4. ÉTUDE EMPIRIQUE

La méthode approximative est utile à deux points de vue. Premièrement, elle constitue une bonne valeur de départ pour la méthode exacte et deuxièmement, elle est probablement plus facile à appliquer dans la pratique. Dans cette section, nous présentons les résultats d'une étude comparative empirique dans laquelle nous avons utilisé des données de la province de Québec pour l'année d'imposition 1988. Nous communiquons les résultats obtenus au moyen de la méthode exacte, pour diverses valeurs de départ, comme au moyen de la méthode approximative. Étant donné que les quantités N_{gh}, Y_h et S_{gh}^2, nécessaires aux deux méthodes, étaient inconnues, nous les avons estimées à l'aide des données.

Nous avons procédé à une stratification selon la taille, comme dans la vraie enquête: quatre strates à tirage partiel et une à tirage complet. Des schèmes de répartition ont été déterminés pour 64 cases "CTI2" (toutes tirées des données du Québec sauf quelques-unes à effectif réduit). Le nombre de fractions de sondage calculé au cours de cette opération variait de 8 à 92, avec une médiane de 24. Le nombre de contraintes, lui, variait de 9 à 115, avec une médiane de 31. Une vingtaine de cases "CTI2" renfermaient plus de 35 variables et dix-huit de ces cases renfermaient aussi plus de 50 contraintes. En tout, nous avions 1,850 strates de seconde phase, qui comptaient environ 230,000 unités de population.

Le coût de l'échantillonnage à la première phase a été établi à \$1.40 l'unité; ce coût comprenait le microfilmage ou la reproduction par photocopie des déclarations d'impôt à Revenu Canada, l'envoi de l'information à Statistique Canada et la détermination des codes à quatre chiffres de la Classification type des industries. Quant au coût de l'échantillonnage de la seconde phase, qui correspondait essentiellement au coût de la transcription de valeurs pour des variables financières, il a été établi à \$7.00. Ces coûts reflètent assez bien ceux enregistrés normalement dans la vraie enquête.

Pour calculer les schèmes de répartition à l'aide de la méthode exacte, nous nous sommes servis de trois valeurs de départ: I – la solution de la méthode approximative; II – l'ensemble des fractions de sondage de la première phase posées égales à un, plus les fractions de la seconde phase conditionnellement optimales correspondantes; et III – un ensemble aléatoire de fractions de sondage de première phase possibles, avec les fractions de seconde phase conditionnellement optimales correspondantes. En outre, nous avons éprouvé la méthode exacte avec une perturbation de chacune de ces valeurs de départ. En ce qui concerne la valeur de départ I, la valeur perturbée de la fraction de sondage de première phase pour la strate g était $v_g^{(0)} = 0.1 + 0.9 \cdot v_g^{*}$, où v_g^{*} est la solution de la méthode approximative. Les fractions de sondage de la seconde phase ont pris au départ des valeurs optimales, étant donné la valeur perturbée des fractions de la première phase. La valeur de départ III a été perturbée de façon analogue. En ce qui a trait à la valeur de départ II, la valeur perturbée de la fraction de sondage était $v_{gh}^{(0)} = 0.1 + 0.9 \cdot v_{gh}^{**}$, où v_{gh}^{**} est optimale, à condition qu'il y ait recensement à la première phase de l'échantillonnage. Pour chaque valeur de départ, nous avons retenu le meilleur résultat observé, peu importe qu'il ait été obtenu à l'aide de la valeur authentique ou de la valeur perturbée correspondante. La convergence était constatée dès que la variation relative de la

Tableau 2

Résultats pour la méthode exacte et la méthode approximative

Méthode	Exacte - Valeur de départ			Approximative
	I	II	III	
Coût total ($)	122779	139347	200998	130228
Nombre de cases pour lesquelles la méthode a produit le meilleur résultat*	48	17	1	

* Pour deux cases, les valeurs de départ I et II ont produit le même résultat et celui-ci est supérieur au résultat obtenu au moyen de la valeur de départ III. C'est pourquoi la somme des chiffres de cette ligne est 66 plutôt que 64.

fonction de coût entre deux itérations était inférieure à 10^{-4}. La version d'IMSL de l'algorithme de programmation quadratique successive de Schittkowski était utilisée pour résoudre les problèmes de programmation non linéaire.

Les résultats de l'étude empirique figurent dans le tableau 2. On y trouve le coût total de l'échantillonnage calculé selon quatre solutions, de même que le nombre de cases "CTI2" pour lesquelles chacune des valeurs de départ utilisées pour la méthode exacte a produit de meilleurs résultats que d'autres valeurs de départ. Nous n'avons pas indiqué les coûts du calcul puisqu'ils n'étaient pas suffisamment élevés pour avoir une influence quelconque.

Les résultats montrent que la méthode approximative offre les meilleures valeurs de départ pour la méthode exacte. Bien que la valeur de départ II ait donné de meilleurs résultats que la valeur de départ I pour 17 cases "CTI2", le coût total associé à la valeur II est plus élevé que celui établi à l'aide de la méthode approximative. La méthode exacte laisse à désirer lorsque les valeurs de départ correspondent à un ensemble aléatoire de fractions de sondage de première phase possibles.

Bien que le coût total établi au moyen de la valeur de départ I de la méthode exacte n'est que 5.7% moins élevé que le coût établi à l'aide de la méthode approximative, il convient de souligner que la méthode exacte (avec valeur de départ I) sera toujours supérieure à la méthode approximative. Celle-là a donné de meilleurs résultats que celle-ci pour 42 cases.

5. CONCLUSION

Dans les deux premières sections de cet article, nous avons exposé le problème de la répartition d'un échantillon pour les plans de sondage à deux phases comme un problème d'optimisation sous contrainte. Si le nombre de variables et de contraintes est peu élevé, on peut obtenir une solution par l'application directe de méthodes numériques. En revanche, l'approche directe laisse à désirer lorsque les variables et les contraintes sont en grand nombre.

En exploitant la structure mathématique du problème, on peut reformuler celui-ci en deux sous-problèmes: premièrement, un problème de programmation convexe avec contraintes linéaires qui renferme beaucoup moins de variables et deuxièmement, un problème que l'on peut résoudre sans recourir à des méthodes numériques. L'algorithme que nous avons proposé dans la section 2 consiste en des itérations entre les deux sous-problèmes. Plus simple sur le plan du calcul, il est plus efficace que l'approche directe pour résoudre des problèmes qui renferment un grand nombre de variables et de contraintes. Dans la section 3, nous avons présenté une méthode approximative de répartition de l'échantillon qui ne nécessite pas l'emploi de méthodes numériques. L'étude empirique de la section 4 montre que la méthode approximative offre une valeur de départ particulièrement bonne pour l'algorithme proposé dans la section 2.

REMERCIEMENTS

Les auteurs tiennent à exprimer leur reconnaissance à Pierre Lavallée, qui, le premier, a établi les expressions relatives aux fractions de sondage de seconde phase conditionnellement optimales utilisées dans la méthode approximative. Ils remercient également l'arbitre et le rédacteur associé qui leur ont formulé des commentaires utiles. C.F.J. Wu est bénéficiaire d'une subvention du Conseil de recherches en sciences naturelles et en génie du Canada.

BIBLIOGRAPHIE

ARMSTRONG, J.B., BLOCK, C., et SRINATH, K.P. (1991). Two-phase sampling of tax records for business surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 228-233.

BOOTH, G., et SEDRANSK, J. (1969). Planning some two-factor comparative surveys. *Journal of the American Statistical Association*, 64, 560-573.

CHOUDHRY, G.H., LAVALLÉE, P., et HIDIROGLOU, M. (1989a). Two-phase sample design for tax data. Document non publié, Division des méthodes d'enquêtes-entreprises, Statistique Canada.

CHOUDHRY, G.H., LAVALLÉE,P., et HIDIROGLOU, M. (1989b). Two-phase sample design for tax data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 646-651.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3ième éd.). New York: John Wiley.

IMSL (1987). Math/Library FORTRAN Subroutines for Mathematical Applications. Houston: IMSL Inc.

RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

SCHITTKOWSKI, K. (1985). NLPQL: A FORTRAN subroutine solving constrained nonlinear programming problems. *Annals of Operations Research*, 5, 485-500.

SEDRANSK, J. (1965). A double sampling scheme for analytical surveys. *Journal of the American Statistical Association*, 60, 985-1004.

SMITH, P.J. (1989). Is two-phase sampling really better for estimating age composition? *Journal of the American Statistical Association*, 84, 916-921.

SUNTER, A.B. (1986). Implicit longitudinal sampling from administrative files: a useful technique. *Journal of Official Statistics*, 2, 161-168.