Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 35–45
Statistique Canada

Sur la stratification de populations asymétriques

PIERRE LAVALLÉE et MICHEL A. HIDIROGLOU[1]

RÉSUMÉ

Pour un degré de précision donné, Hidiroglou (1986) a défini un algorithme permettant de diviser la population en une strate à tirage complet et en une strate à tirage partiel de manière à minimiser la taille de l'échantillon global en supposant un échantillonnage aléatoire simple sans remise dans la strate à tirage partiel. Sethi (1963) a proposé un algorithme permettant un découpage optimal de la population en un certain nombre de strates à tirage partiel. Dans cet article, il est question d'un algorithme itératif qui vise à déterminer les bornes de strates pour une population fortement asymétrique découpée en une strate à tirage complet et en un certain nombre de strates à tirage partiel. Ces bornes de strates sont calculées de manière à minimiser la taille de l'échantillon global étant donné un degré de précision relative, un échantillonnage aléatoire simple sans remise dans les strates à tirage partiel et une répartition à la puissance "p" de l'échantillon entre ces mêmes strates. L'algorithme présenté dans cet article est une combinaison des travaux d'Hidiroglou (1986) et de Sethi (1963).

MOTS CLÉS: Algorithme itératif; bornes optimales; tirage complet; tirage partiel.

1. INTRODUCTION

Dans le cas de populations fortement asymétriques comme celles que l'on retrouve dans les enquêtes entreprises, un échantillonnage efficace n'est possible qui si l'on découpe ces populations en une strate à tirage complet et en un certain nombre de strates à tirage partiel. Toutes les unités de la strate à tirage complet sont prélevées avec une probabilité égale à un (1) tandis que les unités des strates à tirage partiel sont prélevées suivant un modèle probabiliste. Glasser (1962) et Hidiroglou (1986) ont établi des règles approximatives pour le découpage d'une population en une strate à tirage complet et en une strate à tirage partiel. Glasser (1962) a déterminé l'emplacement de la coupure entre les strates en supposant qu'un échantillon de taille déterminée devait être prélevé dans les deux strates et que dans le cas de la strate à tirage partiel, il devait s'agir d'un échantillonnage aléatoire simple sans remise. Pour sa part, Hidiroglou (1986) a déterminé l'emplacement de la coupure entre les strates en supposant qu'il fallait respecter un degré de précision voulu. Ces deux approches sont duales en ce sens que Glasser cherche à minimiser la variance d'échantillonnage pour une taille d'échantillon donnée tandis que Hidiroglou cherche à minimiser la taille d'échantillon pour une variance d'échantillonnage donnée.

Dans cet article, nous proposons un algorithme qui vise à découper une population fortement asymétrique en une strate à tirage complet et en un certain nombre de strates à tirage partiel. Par cet algorithme, nous cherchons à minimiser la taille de l'échantillon global étant donné le coefficient de variation de l'estimateur et le mode de répartition de l'échantillon entre les strates à tirage partiel. Les bornes de strates sont déterminées en fonction d'une variable auxiliaire qui est en corrélation étroite avec les renseignements recueillis dans l'enquête. Si, par exemple, le chiffre des ventes annuelles est une des variables mesurées dans un recensement des détaillants, cette variable auxiliaire peut servir à déterminer les bornes de strates pour une enquête spécialisée qui

[1] Pierre Lavallée, méthodologiste, et Michel A. Hidiroglou, chef, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa (Ontario) K1A 0T6, Canada. Les auteurs tiennent à remercier France Bilocq de la Division des méthodes d'enquêtes-entreprises de Statistique Canada pour avoir programmé les exemples.

vise à recueillir des données sur les ventes mensuelles. Pour ce qui a trait aux enquêtes polyvalentes, si l'on utilise des variables auxiliaires qui ne sont pas en corrélation étroite avec la variable principale, les bornes de strates ne seront plus aussi optimales. L'algorithme que nous proposons est une version modifiée de la méthode de stratification de Sethi (1963). Les bornes de strates qui découlent de cet algorithme sont optimales et déterminent la taille d'échantillon minimum requise.

Le mode de répartition que nous avons choisi pour illustrer la méthode est la répartition à la puissance p. L'utilisation de ce mode de répartition permet de publier des estimations de strates dont les coefficients de variation ne sont pas très différents les uns des autres. La répartition à la puissance p a été proposée par Carrol (1970), Fellegi (1981) et Bankier (1988). En pratique, elle est un compromis entre la répartition de Neyman et la nécessité d'avoir des coefficients de variation égaux pour chaque strate. Un inconvénient de la répartition de Neyman est qu'elle produit des coefficients de variation très différents d'une strate à l'autre si l'on doit avoir des estimations pour chaque strate. Par ailleurs, un mode de répartition qui produit les mêmes coefficients de variation d'une strate à l'autre peut exiger un échantillon beaucoup plus grand que celui qu'exige la répartition de Neyman. Dans le cadre de notre analyse, la répartition à la puissance p nous permettra de produire des estimations pour des strates d'entreprises de tailles variées (petites, moyennes et grandes) avec des coefficients de variation similaires.

Nous comparerons la méthode élaborée dans cet article, au point de vue des bornes et de la taille d'échantillon, à la règle de la racine de f cumulative de Dalenius — Hodges (1959) ainsi qu'à une combinaison des méthodes de stratification d'Hidiroglou (1986) et de Dalenius — Hodges (1959). L'algorithme, qui est récursif par définition, est facile à programmer et converge rapidement vers les bornes optimales. Il permet également une réduction sensible de la taille d'échantillon pour des critères de fiabilité donnés.

2. LE PROBLÈME

Considérons une population ordonnée finie de N unités:

$$y_{(1)}, y_{(2)}, \ldots, y_{(N)},$$

où $y_{(i)} \leq y_{(i+1)}$ pour $i = 1, 2, \ldots, N-1$. Cette population est découpée en L strates. Le nombre d'unités dans chaque strate est désigné par N_h, $h = 1, 2, \ldots, L$. Selon le plan d'échantillonnage, n_h unités doivent être prélevées dans chaque strate à tirage partiel de taille N_h ($h = 1, 2, \ldots, L-1$) au moyen d'un échantillonnage aléatoire simple sans remise; notons par ailleurs que $n_L = N_L$. La moyenne à estimer est

$$\bar{Y} = \sum_{h=1}^{L} \sum_{j=M_{h-1}+1}^{M_h} y_{(j)}/N, \tag{2.1}$$

où $M_h = \sum_{i=1}^{h} N_i$ pour $h = 1, 2, \ldots, L$ et $M_o = 0$.

Compte tenu de cette définition, l'estimateur de la moyenne de population $\bar{Y}$ est

$$\hat{\bar{Y}} = \left[\sum_{h=1}^{L-1} \frac{N_h}{n_h} \sum_{j=m_{h-1}+1}^{m_h} z_j + \sum_{j=M_{L-1}+1}^{N} y_{(j)} \right] /N \tag{2.2}$$

où $y_{M_{h-1}+1} \leq z_j \leq y_{M_h}$ pour $j = m_{h-1}+1, \ldots, m_h$ $(h = 1, 2, \ldots, L-1)$, $m_h = \sum_{i=1}^{h} n_i$ pour $h = 1, 2, \ldots, L$ et $m_o = 0$.

Supposons que le degré de précision voulu pour la moyenne estimée est défini par c (coefficient de variation) et que la proportion d'unités échantillonnées devant être réparties entre les $L-1$ premières strates est a_h $(h = 1, 2, \ldots, L-1)$ où $\sum_{h=1}^{L-1} a_h = 1$. Le terme "a_h" représente en fait n'importe quel mode de répartition entre les strates. Dans le cas, par exemple, de la répartition proportionnelle à N à la puissance p,

$$a_h = \frac{N_h^P}{\sum_{h=1}^{L-1} N_h^P} \quad (h = 1, 2, \ldots, L-1)$$

et dans le cas de la répartition proportionnelle à Y à la puissance p,

$$a_h = \frac{Y_h^P}{\sum_{h=1}^{L-1} Y_h^P},$$

où $0 < p < \infty$. Étant donné des hypothèses relativement simples et une valeur de p appropriée, la répartition à la puissance p a la propriété d'engendrer des coefficients de variation relativement uniformes pour les strates à tirage partiel sans accroître notablement le coefficient de variation global. Cette uniformité des coefficients de variation est souvent souhaitée par les utilisateurs des données d'enquête.

En pratique, on choisit souvent 1/2 ou 1/3 comme valeur de p. Une faible valeur de p (c.-à-d., une valeur proche de 0) donne habituellement des coefficients de variation comparables d'une strate à l'autre tandis qu'une valeur plus élevée élargit l'écart entre les coefficients mais accroît du même coup la précision des estimations globales.

Soulignons également que la répartition à la puissance p définie ci-dessus équivaut au mode proposé par Bankier (1988) lorsque les coefficients de variation des strates à tirage partiel sont égaux.

La variance de $\hat{\bar{Y}}$ est

$$V(\hat{\bar{Y}}) = \frac{1}{N^2} \sum_{h=1}^{L-1} \frac{N_h}{n_h} (N_h - n_h) S_h^2, \tag{2.3}$$

où S_h^2 désigne la variance de population de chaque strate h. Si l'on tient compte du degré de précision voulu (coefficient de variation c), la variance de Y peut être redéfinie $V(\hat{\bar{Y}}) = c^2 \bar{Y}^2$. En remplaçant n_h et $V(\hat{\bar{Y}})$ dans l'équation (2.3) par $(n - N_L)\, a_h$ et $c^2 \bar{Y}^2$ respectivement et en isolant n, on obtient l'équation

$$n = N_L + \frac{\sum_{h=1}^{L-1} N_h^2 S_h^2 / a_h}{(N c \bar{Y})^2 + \sum_{h=1}^{L-1} N_h S_h^2}. \tag{2.4}$$

Le problème consiste à déterminer des bornes $b_{(1)}, b_{(2)}, \ldots, b_{(L-1)}$ (où $y_{(1)} < b_{(1)} < \ldots < b_{(L-1)} < y_{(N)}$) de manière à minimiser la taille n de l'échantillon global, étant donné le degré de fiabilité c et le mode de répartition spécifié (représenté par a_h).

3. L'ALGORITHME

La méthode utilisée dans le présent article pour calculer des bornes de strates pour un degré de précision voulu a déjà été utilisée par Dalenius (1950) pour calculer des bornes de strates pour une taille d'échantillon donnée. On suppose en premier lieu que l'échantillonnage se fait à partir d'une population dont la distribution de fréquences peut être représentée avec suffisamment de précision par une fonction continue de densité $f(y)$. Ainsi, pour un ensemble donné de limites $b_{(1)}, \ldots, b_{(L-1)}$ nous définissons les quantités suivantes:

$$W_h = \int_{b_{(h-1)}}^{b_{(h)}} f(y)\, dy, \tag{3.1}$$

$$\mu_h = \int_{b_{(h-1)}}^{b_{(h)}} y f(y)\, dy / W_h, \tag{3.2}$$

$$\sigma_h^2 = \int_{b_{(h-1)}}^{b_{(h)}} y^2 f(y)\, dy / W_h - \mu_h{}^2, \tag{3.3}$$

pour $h = 1, \ldots, L$, où $b_{(o)} = -\infty$, $b_{(L)} = +\infty$.

On peut alors reformuler l'équation (2.4) comme suit:

$$n = NW_L + \frac{N\left(\sum_{h=1}^{L-1} W_h^2 \sigma_h^2 / a_h\right)}{N c^2 \mu^2 + \sum_{h=1}^{L-1} W_h \sigma_h^2}, \tag{3.4}$$

où

$$\mu = \int_{b_{(o)}}^{b_{(L)}} y f(y)\, dy.$$

Il convient de souligner que même s'il s'agit d'une population de grande taille, le facteur de correction pour population finie (cpf) est toujours présent dans l'équation (3.4) - voir Dalenius-Gurney (1951). Par définition, la strate à tirage complet doit avoir une population finie si l'on veut que la taille d'échantillon soit finie. En outre, le fait d'ignorer le facteur de c.p.f. ne se traduirait pas par une variance nulle pour la strate à tirage complet.

Le terme "a_h" dans l'équation (2.3) peut aussi être exprimé à l'aide des quantités (3.1, (3.2) et (3.3). Dans le cas de la répartition proportionnelle à N à la puissance p, nous avons:

$$a_h = \frac{W_h^p}{\sum_{h=1}^{L-1} W_h^p}, \tag{3.5}$$

pour $h = 1, .., L-1$.

Pour ce qui a trait à la répartition proportionnelle à Y à la puissance p, nous avons l'équation suivante:

$$a_h = \frac{(W_h \mu_h)^p}{\sum_{h=1}^{L-1} (W_h \mu_h)^p}, \tag{3.6}$$

où $0 < p < \infty$.

Dans cet article, nous considérerons surtout le répartition proportionnelle à Y mais les calculs peuvent être effectués aussi bien pour la répartition proportionnelle à N et, de fait, pour n'importe quel mode de répartition représenté par le terme a_h où $\sum_{h=1}^{L-1} a_h = 1$. Si nous remplaçons a_h dans l'équation (3.4) par l'expression définie en (3.6), nous obtenons

$$n = N W_L + \frac{N \left[\sum_{h=1}^{L-1} (W_h \sigma_h)^2 (W_h \mu_h)^{-p} \right] \left[\sum_{h=1}^{L-1} (W_h \mu_h)^p \right]}{N c^2 \mu^2 + \sum_{h=1}^{L-1} W_h \sigma_h^2}. \tag{3.7}$$

Afin de déterminer les bornes optimales $b_{(1)}, \ldots, b_{(L-1)}$ de manière que la taille n de l'échantillon soit minimale, on calcule les dérivées partielles de l'équation (3.7) par rapport à $b_{(1)}, \ldots, b_{(L-1)}$, respectivement et on les pose égales à zéro. On obtient alors les équations suivantes:

Pour $h = 1, \ldots, L-2$,

$$[F T_h - F T_{h+1}]\ b_{(h)}^2 +$$
$$[F K_h - 2\mu_h F T_h - F K_{h+1} + 2\mu_{h+1} F T_{h+1} + 2\mu_h AB - 2\mu_{h+1} AB]\ b_{(h)} +$$
$$[F T_h \mu_h^2 + F T_h \sigma_h^2 - F T_{h+1} \mu_{h+1}^2 - F T_{h+1} \sigma_{h+1}^2 - AB\mu_h^2 + AB\mu_{h+1}^2] = 0, \tag{3.8}$$

et pour $h = L-1$,

$$[F T_{L-1} - AB]\ b_{(L-1)}^2 +$$
$$[F K_{L-1} - 2\mu_{L-1} F T_{L-1} + 2\mu_{L-1} AB]\ b_{(L-1)} +$$
$$[F T_{L-1} \mu_{L-1}^2 + F T_{L-1} \sigma_{L-1}^2 - AB\mu_{L-1}^2 - F^2] = 0, \tag{3.9}$$

où

$$A = \sum_{h=1}^{L-1} (W_h \mu_h)^p,$$

$$B = \sum_{h=1}^{L-1} (W_h \sigma_h)^2 (W_h \mu_h)^{-p},$$

$$F = N c^2 \mu^2 + \sum_{h=1}^{L-1} W_h \sigma_h^2,$$

$$K_h = B p (W_h \mu_h)^{p-1} - A p (W_h \sigma_h)^2 (W_h \mu_h)^{-p-1},$$

$$T_h = A W_h (W_h \mu_h)^{-p}.$$

Si nous identifions le coefficient de $b_{(h)}^2$ par α_h, le coefficient de $b_{(h)}$ par β_h et l'ensemble des autres termes par γ_h, nous pouvons représenter les équations (3.8) et (3.9) comme des équations quadratiques de la forme $\alpha_h b_{(h)}^2 + \beta_h b_{(h)} + \gamma_h = 0$. Or, comme le souligne Sethi (1963), les termes α_h, β_h et γ_h sont eux-mêmes des fonctions de $b_{(1)}, \ldots, b_{(L-1)}$ par les intégrales (3.1), (3.2) et (3.3). En adoptant l'approche de Sethi (1963), nous pouvons résoudre facilement les équations (3.8) et (3.9) à l'aide de la méthode itérative suivante:

ÉTAPE 1 : Prendre des bornes arbitraires $b'_{(1)} < \ldots < b'_{(L-1)}$.

ÉTAPE 2 : Calculer les proportions W'_h, les moyennes μ'_h et les variances $\sigma_h^{2'}$ (à l'aide des équations (3.1), (3.2) et (3.3), respectivement) fondées sur ces bornes, $h = 1, \ldots, L-1$.

ÉTAPE 3 : Remplacer la série de bornes initiale par $b''_{(1)}, \ldots, b''_{(L-1)}$ où

$$b''_{(h)} = \frac{-\alpha'_h + \sqrt{\beta_h'^2 - 4\alpha'_h \gamma'_h}}{2\alpha'_h}, \quad h = 1, \ldots, L-1. \tag{3.10}$$

ÉTAPE 4 : Reprendre les étapes 2 et 3 jusqu'à ce qu'on obtienne deux séries consécutives identiques ou très semblables, c'est-à-dire

$$\max_{h=1}^{L-1} | b''_{(h)} - b'_{(h)} | < \epsilon \text{ pour une valeur } \epsilon > 0. \tag{3.11}$$

Soulignons qu'il est possible de démontrer que le signe précédant la racine carrée ($\sqrt{\ }$) est positif parce que $b'_{(h)}$ se situe entre μ'_h et μ'_{h+1}.

L'inconvénient de cet algorithme est que son utilisation exige que l'on dispose de renseignements sur la fonction de densité approximative $f_{(y)}$. Comme la population considérée est finie, nous pouvons éliminer cet inconvénient en remplaçant les quantités (3.1), (3.2) et (3.3) par les expressions correspondantes fondées sur une population finie. Ainsi, conformément au procédé appliqué dans Cochran (1977), nous pouvons remplacer les paramètres de population infinie donnés par les expressions (3.1), (3.2) et (3.3) par les paramètres correspondants pour population finie, c'est-à-dire:

$$W_h = \frac{N_h}{N}, \tag{3.12}$$

$$\bar{Y}_h = \frac{1}{N_h} \sum_{\substack{j=b_{(h-1)}+1}}^{b_{(h)}} y_{(j)}, \tag{3.13}$$

$$S_h^2 = \frac{1}{N_{h-1}} \sum_{j=b_{(h-1)}+1}^{b_{(h)}} y_{(j)}^2 - N_h \bar{Y}_h^2, \tag{3.14}$$

pour $h = 1, \ldots, L$.

Ces formules peuvent alors nous permettre de déterminer sans difficulté à l'aide de la méthode itérative suivante les bournes de strates $b_{(1)}, \ldots, b_{(L-1)}$ de manière que la taille n de l'échantillon global minimum pour un degré de fiabilité (c) et un mode de répartition données:

ÉTAPE 0 : Classer les éléments de la population $y_1, \ldots, y_N$ par ordre croissant et poser $b_{(0)} = y_{(1)}$ et $b_{(L)} = y_{(N)}$.

ÉTAPE 1 : Choisir des bornes arbitraires de telle sorte que $b_{(0)} < b'_{(1)} < \ldots < b'_{(L-1)} < b_{(L)}$.

ÉTAPE 2 : Calculer la proportion W'_h, la moyenne $\bar{Y}'_h$ et la variance $S_h^{2'}$ (à l'aide des équations (3.12), (3.13) et (3.14) respectivement) fondées sur ces bornes, $h = 1, \ldots, L$-1.

ÉTAPE 3 : Remplacer la série de bornes initiale par $b''_{(1)}, \ldots, b''_{(L-1)}$ où

$$b''_{(h)} = \frac{-\alpha'_h + \sqrt{\beta_h'^2 - 4\,\alpha'_h\,\gamma'_h}}{2\,\alpha'_h}, \; h = 1, \ldots, L-1.$$

ÉTAPE 4: Reprendre les étapes 2 et 3 jusqu'à ce qu'on obtienne deux séries consécutives identiques ou très semblables, c'est-à-dire

$$\max_{h=1}^{L-1} | b''_{(h)} - b'_{(h)} | < \epsilon \text{ pour une valeur } \epsilon < 0.$$

Dans la section qui suit, nous comparons l'application de cet algorithme à celle d'autres méthodes.

4. EXEMPLES

Afin d'illustrer l'algorithme décrit dans la section 3, nous allons nous servir de données tirées des enquêtes annuelles sur le commerce de détail et le commerce de gros réalisées par Statistique Canada. Ces enquêtes permettent d'évaluer le chiffre des ventes des entreprises spécialisées dans le commerce de détail ou le commerce de gros. Nous utilisons trois populations pour illustrer l'algorithme. Il s'agit des grossistes en produits divers du Québec (population 1), des grossistes en produits alimentaires divers du Manitoba (population 2) et des magasins de vente au détail d'appareils ménagers et d'appareils électroniques du Québec (population 3). Le choix des populations s'est fait de manière à obtenir un éventail des tailles de population: grande, moyenne et petite. Les coefficients d'asymétrie pour ces populations sont 24.2 (population 1), 6.5 (population 2) et 13.6 (population 3).

Nous allons comparer les résultats de l'algorithme proposé à ceux de deux autres méthodes. La première de ces méthodes consiste simplement à stratifier la population au moyen de la règle de la racine de f cumulative de Dalenius — Hodges (1959). La seconde méthode consiste à déterminer l'emplacement de la coupure entre la strate à tirage complet et la strate à tirage partiel au

moyen de la formule d'approximation d'Hidiroglou (1986), puis à appliquer la règle de la racine de f cumulative pour découper la population qui n'est pas incluse dans la strate à tirage complet en un certain nombre de strates à tirage partiel. Nous désignerons respectivement ces méthodes comme la règle $f^{1/2}$ CUM et la méthode combinée; quant à l'algorithme proposé, nous le désignerons comme la méthode optimale. Il n'est pas réaliste d'utiliser uniquement la méthode de Dalenius-Hodges (1959) car, dans la pratique, celle-ci ne serait appliquée qu'une fois que la strate à tirage complet aurait été définie à l'aide d'une règle arbitraire donnée. Néanmoins, nous illustrons ici l'application de cette méthode pour mettre en garde le lecteur contre une utilisation inconsidérée de cette méthode pour des populations fortement asymétriques.

Hidiroglou (1986) détermine l'emplacement de la coupure entre les strates au moyen de la formule itérative suivante:

$$b_{TA}^{''} = \mu_{[N-t']} = \left\{ \frac{N-t'-1}{(N-t')^2} N^2 c^2 \bar{Y}^2 + S^2_{[N-t']} \right\}^{1/2}, \tag{4.1}$$

où

$$\mu_{[N-t']} = \frac{1}{N-t'} \sum_{i=1}^{N-t'} y_{(i)} \tag{4.2}$$

Tableau 1

Effet de la variation du coefficient de variation et de la puissance p sur la taille d'échantillon pour trois méthodes de stratification
(population 1 — taille = 1221)

			Méthode de stratification								
			Règle $f^{1/2}$ Cum			Composée			Optimale		
c	p	Strates	N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$
0.05	0.25	1	1196	177*		1017	16		891	11	
		2	20	20	3,715,320	152	14	465,180	290	13	302,912
		3	5	5	14,786,280	52	52	1,131,961	40	40	1,835,930
		Total		202			82			64	
0.05	0.50	1	1196	178*		1017	16		863	10	
		2	20	20	3,715,320	152	13	465,180	318	14	289,422
		3	5	5	17,786,280	52	52	1,131,961	40	40	1,832,038
		Total		203			81			64	
0.01	1.00	1	1196	616*		751	37		687	36	
		2	20	20	3,715,320	215	34	196,840	374	78	162,068
		3	5	5	14,786,280	255	255	383,033	160	160	564,076
		Total		641			326			274	
0.05	1.00	1	1196	180*	3,715,320	1017	16		858	8	
		2	20	20	14,786,280	152	11	465,180	323	16	271,920
		3	5	5		52	52	1,131,961	40	40	1,867,254
		Total		205			79			64	
0.10	1.00	1	1196	56*		1073	7		1007	7	
		2	20	20	3,715,320	109	4	592,900	191	9	442,357
		3	5	5	14,786,280	39	39	1,953,113	23	23	4,032,950
		Total		81			50			39	

*Une répartition disproportionnée est nécessaire pour satisfaire au coefficient de variation.

Tableau 2

Effet de la variation du coefficient de variation et de la puissance p sur la taille d'échantillon pour trois méthodes de stratification
(population 2 — taille = 44)

			Méthode de stratification								
			Règle $f^{1/2}$ Cum			Composée			Optimale		
c	p	Strates	N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$
0.05	0.25	1	42	38		32	1		29	1	
		2	1	1*	137,939,900	6	1	4,708,409	11	1	3,029,455
		3	1	1	459,739,000	6	6	10,622,301	4	4	17,461,464
		Total		40			8			6	
0.05	0.50	1	42	38		32	1		28	1	
		2	1	1*	137,939,900	6	1	4,708,409	12	1	2,582,819
		3	1	1	459,739,000	6	6	10,622,301	4	4	17,640,325
		Total		40			8			6	
0.01	1.00	1	42	42		25	1		25	1	
		2	1	1	137,939,900	5	1	1,059,550	10	4	1,153,322
		3	1	1	459,739,000	14	14	3,742,377	9	9	5,969,271
		Total		44			16			14	
0.05	1.00	1	42	38		32	1		26	1	
		2	1	1*	137,939,900	6	1	4,708,409	14	2	1,779,500
		3	1	1	459,739,000	6	6	10,622,301	4	4	17,349,902
		Total		40			8			7	
0.10	1.00	1	42	30		34	1		28	1	
		2	1	1*	137,939,900	6	1	4,848,218	13	1	2,413,800
		3	1	1	459,739,000	4	4	16,749,625	3	3	30,091,449
		Total		32			6			5	

*Une répartition disproportionnée est nécessaire pour satisfaire au coefficient de variation.

et

$$S^2_{[N-t']} = \frac{1}{N-t'-1} \sum_{i=1}^{N-t'} (y_{(i)} - \mu_{[N-t']})^2.$$

Le nombre d'unités à tirage complet obtenues pour chaque étape du processus itératif est t'. Le point de départ de cette approximation est

$$b'_{TA} = \mu_{[N]} + \{N c^2 \bar{Y}^2 + S^2_{[N]}\}^{1/2} . \tag{4.3}$$

Le processus itératif cesse lorsque l'inéquation suivante est satisfaite:

$$0 \leq 1 - n(t'')/n(t') < 0.10 \tag{4.4}$$

Tableau 3

Effet de l'accroissement du nombre de strates sur la taille d'échantillon pour deux méthodes de stratification
$p = 1$, $c = 0.05$

Population 1 ($N = 1221$) Méthode de stratification	Strates	Nombre de strates 3 N_h	3 n_h	3 $b_{(h)}$	4 N_h	4 n_h	4 $b_{(h)}$	5 N_h	5 n_h	5 $b_{(h)}$
Composée	1	1017	16		897	6		823	3	
	2	152	11	465,180	194	5	311,117	194	2	245,090
	3	52	52	1,131,961	78	4	641,252	101	2	465,180
	4				52	52	1,131,961	51	2	751,297
	5							52	52	1,131,961
	Total		79			67			61	
Optimale	1	858	8		704	3		655	2	
	2	323	16	271,920	373	7	173,981	358	4	149,327
	3	40	40	1,867,254	112	6	604,869	163	5	453,114
	4				32	32	2,676,449	29	4	1,522,329
	5							16	16	5,810,487
	Total		64			48			31	
Population 3 ($N = 161$)										
Composée	1	106	6		84	2		71	1	
	2	39	6	265,480	38	2	185,320	35	1	155,260
	3	16	16	553,255	23	2	335,620	22	1	265,480
	4				16	16	553,255	17	1	385,720
	5							16	16	553,255
	Total		28			22			20	
Optimale	1	86	4		55	1		34	1	
	2	65	9	199,415	61	3	125,572	51	1	83,594
	3	10	10	680,942	39	5	312,769	42	2	192,215
	4				6	6	826,942	29	3	382,236
	5							5	5	906,894
	Total		23			15			12	

où

$$n(t') = t' + \frac{(N-t')^2 S^2_{[N-t']}}{(N c \bar{Y})^2 + (N-t') S^2_{[N-t']}} . \tag{4.5}$$

Les tableaux 1 et 2 donnent les résultats pertinents pour une grande population (population 1) et une petite population (population 2) étant donné divers coefficients de variation et diverses valeurs de p (pour la répartition à la puissance p). Le tableau 3 donne les résultats obtenus pour la grande population (population 1) et une population de taille moyenne (population 3) lorsqu'on modifie le nombre de strates. Pour les trois tableaux, la répartition de l'échantillon entre les strates à tirage partiel s'est faite suivant le mode de répartition proportionnelle à Y à la puissance p.

Les tableaux 1 et 2 nous permettent de tirer les conclusions suivantes. La règle de la racine de f cumulative est très inefficace dans les présentes circonstances. La méthode composée réduit

de façon substantielle le taille de l'échantillon. L'application de la méthode optimum entraîne une réduction additionnelle de l'ordre de 20% pour les trois strates figurant dans ces deux tableaux. Pour un coefficient de variation donné, la variation de la puissance "p" influe peu sur la taille de l'échantillon. Comme prévu, la taille d'échantillon augmente lorsque le coefficient de variation (c) diminue (pour une même valeur de p). La méthode optimale définit moins d'unités à tirage complet (strate 3) que la méthode composée ou, en d'autres termes, la borne de la strate à tirage complet est plus élevée pour la première méthode que pour la seconde. La règle de la racine de f cumulative n'est plus aussi efficace lorsqu'il s'agit de déterminer la borne de la strate à tirage complet. Il est facile de voir que la borne calculée à l'aide de cette règle est considérablement plus élevée que celles calculées à l'aide des autres méthodes.

Dans le tableau 3, nous comparons les résultats obtenus par la méthode composée et la méthode optimale pour deux populations lorsque le nombre de strates varie, étant donné un coefficient de variation et une valeur de p (pour la répartition proportionnelle à Y à la puissance p). Des conclusions semblables à celles tirées des deux premiers tableaux se dégagent du tableau 3. L'accroissement du nombre de strates a pour effet de réduire le nombre d'unités échantillonnées, peu importe la méthode utilisée. Toutefois, la réduction est plus prononcée dans le cas de la méthode optimale.

5. CONCLUSION

Le découpage optimal d'une population asymétrique en une strate à tirage complet et en un certain nombre de strates à tirage partiel a permis de réduire sensiblement la taille de l'échantillon global pour un degré de précision relative donné. La méthode proposée peut être adaptée à n'importe quel mode de répartition et à n'importe quel nombre de strates. On peut également faire abstraction de la condition relative au tirage complet.

L'algorithme, qui est récursif par définition, converge rapidement. Il peut être appliqué facilement par ordinateur par l'intermédiaire du SAS, du FORTRAN ou de n'importe quel autre langage évolué.

BIBLIOGRAPHIE

BANKIER, M.D. (1988). Power allocations, determining sample sizes for sub-national areas. À paraître dans *The American Statistician*.

CARROL, J. (1970). Allocation of a sample between States. Note de service non publiée de l'Australian Bureau of Census and Statistics.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3^e éd.). New York: John Wiley & Sons.

DALENIUS, T. (1950). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift*, 33, 203-213.

DALENIUS, T., et GURNEY, M. (1951). The problem of optimum stratification.II, *Skandinavisk Aktuarietidskrift*, 34, 133-148.

DALENIUS, T., et HODGES, J.L.Jr. (1959). Minimum variance stratification, *Skandinavisk Aktuarietidskrift*, 54, 88-101.

FELLEGI, I.P. (1981). Should the census counts be adjusted for allocation purposes? — Equity considerations. Dans *Current Topics in Survey Sampling* (éds. D. Krewski, R. Platek et J.N.K. Rao), New York: Academic Press, 47-76.

GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Revue Internationale de Statistique*, 30, 28-32.

HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design, *The American Statistican*, 40, 27-31.

SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.