Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 177-183
Statistique Canada, N° 12-001-XPB au catalogue

Approche de la stratification par une méthode géométrique et par optimisation : Une comparaison de l'efficacité

Marcin Kozak et Med Ram Verma [1]

Résumé

L'article donne une comparaison des approches de la stratification par une méthode géométrique, par optimisation et par la méthode de Lavallée et Hidiroglou (LH). L'approche géométrique de stratification est une approximation, tandis que les deux autres, qui s'appuient sur des méthodes numériques, peuvent être considérées comme des méthodes de stratification optimales. L'algorithme de la stratification géométrique est très simple comparativement à ceux des deux autres approches, mais il ne tient pas compte de la construction d'une strate à tirage complet, qui est habituellement produite lorsque l'on stratifie une population positivement asymétrique. Dans le cas de la stratification par optimisation, on peut prendre en considération toute forme de la fonction d'optimisation et de ses contraintes. Une étude numérique comparative portant sur cinq populations artificielles positivement asymétriques a indiqué que, dans chaque cas étudié, l'approche par optimisation était plus efficace que la stratification géométrique. En outre, nous avons comparé les approches géométrique et par optimisation à l'algorithme LH. Cette comparaison a révélé que la méthode géométrique de stratification était moins efficace que l'algorithme LH, tandis que l'approche par optimisation était aussi efficace que cet algorithme. Néanmoins, les limites de strate déterminées par la stratification géométrique peuvent être considérées comme de bons points de départ pour l'approche par optimisation.

Mots clés : Stratification optimale; stratification géométrique; optimisation numérique; algorithme de Lavallée-Hidiroglou.

1. Introduction

Gunning et Horgan (2004) ont proposé un algorithme de stratification basé sur une progression géométrique. Par souci de simplicité, nous appellerons cette technique « approche géométrique de stratification », « stratification géométrique » ou simplement « approche géométrique ». La stratification géométrique vise à produire des valeurs égales du coefficient de variation d'une variable de stratification dans les diverses strates, en émettant l'hypothèse que la variable suit une loi uniforme dans chaque strate. Gunning et Horgan (2004) ont montré que leur algorithme est nettement plus facile à appliquer et plus efficace que la méthode classique de la fonction cumulative de la racine carrée des fréquences (Dalenius et Hodges 1959) et que l'algorithme de Lavallée et Hidiroglou (LH) (Lavallée et Hidiroglou 1988). Horgan (2006) a comparé la stratification géométrique aux méthodes de Dalenius et Hodges (1959), d'Ekman (1959), et de Lavallée et Hidiroglou (1988); de nouveau, son étude a montré que la stratification géométrique était la plus efficace parmi les méthodes comparées. Gunning, Horgan et Yancey (2004) ont appliqué cette méthode en vue de stratifier des populations comptables.

À l'instar de la méthode de la fonction cumulative de la racine carrée des fréquences, l'approche géométrique est une technique de stratification approximative, si bien que les points de stratification qu'elle fournit peuvent s'écarter considérablement des points de stratification optimaux. Par ailleurs, il existe des approches, particulièrement pour la stratification univariée, qui produisent des stratifications quasi optimales. Ces approches sont fondées sur l'utilisation d'algorithmes auto-appliqués ou de méthodes numériques d'optimisation pour produire les limites de strate (par exemple, Lavallée et Hidiroglou 1988; Lednicki et Wieczorkowski 2003; Kozak 2004). Toutefois, les méthodes de ce genre requièrent habituellement des limites initiales pour lancer le processus d'optimisation; les méthodes de stratification approximatives peuvent être utilisées pour rechercher ces points initiaux. Naturellement, les limites de strate initiales doivent être de haute qualité; sinon, l'optimisation risque de fournir un minimum local (Rivest 2002).

De nombreuses enquêtes comportent des variables d'intérêt positivement asymétriques. Le cas échéant, il est important de tenir compte de cet attribut lors de la stratification d'une population. Nombre de chercheurs ont essayé de créer des méthodes de stratification permettant de construire une strate dite « à tirage complet » (par exemple, Glasser 1962; Hidiroglou 1986) dont tous les éléments sont sélectionnés dans l'échantillon avec une probabilité égale à 1. Dans le contexte de l'échantillonnage stratifié, il s'agit du meilleur moyen de traiter les variables positivement asymétriques. Ces méthodes sont habituellement plus efficaces (de façon certaine, uniquement si une population est

1. Marcin Kozak, département de biométrie, Université agricole de Varsovie, Nowoursynowska 159, 02-776 Varsovie, Pologne. Courriel : marcin.kozak@omega.sggw.waw.pl; Med Ram Verma, Division of Agricultural Economics & Statistics, ICAR Research Complex for N.E.H. Region, Umroi Road, Umiam (Barapani) Meghalaya, Inde, Pin 793 103. Courriel : mrverma19@yahoo.co.in.

positivement asymétrique) que les méthodes de stratification ne comportant pas la construction d'une strate à tirage complet. La stratification géométrique ne comprend pas la création d'une telle strate (Gunning et Horgan 2004).

Le but du présent article est de comparer l'efficacité de la stratification géométrique, proposée par Gunning et Horgan (2004) à celle de deux approches de stratification par optimisation (Lavallée et Hidiroglou 1988; Lednicki et Wieczorkowski 2003; Kozak 2004) fondées sur l'utilisation de méthodes numériques d'optimisation.

2. Approches de stratification comparées

Supposons que nous souhaitions stratifier une population positivement asymétrique de N unités, U, en nous fondant sur un vecteur $\mathbf{x} = (x_1, \ldots, x_N)^T$ de dimension N connu dès le départ (c'est-à-dire avant le début de l'étude) des valeurs d'une variable de stratification X.

Dans le présent article, nous considérons deux problèmes de stratification. Le premier consiste à construire L strates sachant la taille fixe d'échantillon n. Supposons que nous recherchions un vecteur de dimension $(L+1)$ de limites de strate $\mathbf{k} = (k_0, \ldots, k_L)^T$, $(k_0 < k_1 < \ldots < k_L, k_0$ étant la valeur minimale et k_L la valeur maximale de X) qui minimise la variance d'un estimateur de la moyenne de population de X sous échantillonnage stratifié avec échantillonnage aléatoire simple sans remise dans les strates (*STSI*) et combiné à une approche avec strate à tirage complet. (Il convient de souligner que nous traitons la variable de stratification comme étant identique à la variable d'enquête correspondante.) La variance de $\bar{x}_{\text{st}}$ est donnée par

$$V(\bar{x}_{\text{st}}) = \sum_{h=1}^{L-1} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h},$$

$$\bar{x}_{\text{st}} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{x}_h, \bar{x}_h = \frac{1}{n_h} \sum_{k=1}^{n_h} x_{kh} (h = 1, \ldots, L), \qquad (1)$$

où n_h est la taille de l'échantillon provenant de la h^{e} strate, N_h est la taille de la h^{e} strate, S_h^2 est la variance de population de X restreinte à la h^{e} strate, $\bar{x}_{\text{st}}$ est l'estimateur de la moyenne de population de X sous échantillonnage *STSI*, $\bar{x}_h$ est l'estimateur de la moyenne de population de X dans la h^{e} strate sous échantillonnage aléatoire simple sans remise (*SI*) et x_{kh} est la valeur de X pour la k^{e} unité d'échantillonnage de la h^{e} strate et $h = 1, \ldots, L.$.

La répartition optimale de l'échantillon, qui s'obtient, dans le cas de notre problème, par minimisation de la variance (1) sachant la taille d'échantillon n, est donnée par la formule de l'optimum de Neyman adaptée à une approche avec strate à tirage complet (Lednicki et Wieczorkowski 2003) :

$$n_h = (n - N_L) \frac{N_h S_h}{\sum_{h=1}^{L-1} N_h S_h}, h = 1, \ldots, L-1. \qquad (2)$$

L'approche géométrique de stratification a pour objectif de rendre égales les valeurs du coefficient de variation de X dans les L strates. Elle consiste simplement à appliquer la formule qui suit basée sur une progression géométrique (Gunning et Horgan 2004)

$$k_h = ar^h, h = 0, \ldots, L, \qquad (3)$$

où $a = \min(X), k_L = \max(X)$ et $r = (k_L / k_0)^{1/L}$. La formule (3) repose sur l'hypothèse selon laquelle X suit une loi uniforme dans chaque strate.

L'approche par optimisation appliquée à ce problème de stratification particulier s'inspire de l'optimisation numérique du problème suivant : minimiser

$$f(\mathbf{k}) = V(\bar{x}_{\text{st}}), \qquad (4)$$

où $V(\bar{x}_{\text{st}})$ est la variance (1) sous la répartition optimale (2), sous les contraintes

$$N_h \geq 2 \text{ et } 2 \leq n_h \leq N_h \text{ pour } h = 1, \ldots, L-1, \qquad (5)$$

et

$$\sum_{h=1}^{L-1} n_h = n - N_L. \qquad (6)$$

Parfois, si l'on veut que le niveau de précision soit plus ou moins le même dans chaque strate, il est possible d'appliquer une méthode de « répartition avec puissance » (en anglais, *power allocation*) (Bankier 1988; Rivest 2002; Lednicki et Wieczorkowski 2003):

$$n_h = \frac{(n - N_L)(N_h \bar{x}_h)^P}{\sum_{h=1}^{L-1} (N_h \bar{x}_h)^p}, p \in (0,1]; h = 1, \ldots, L-1. \qquad (7)$$

L'approche par optimisation est plus difficile à appliquer que l'approche géométrique, en grande partie parce que cette dernière requiert un algorithme considérablement plus simple. Un choix doit être fait parmi les diverses méthodes d'optimisation disponibles. Lednicki et Wieczorkowski (2003) ont utilisé la méthode du simplexe de Nelder et Mead (1965); cependant, il est également possible d'appliquer des méthodes plus efficaces, qui nécessitent souvent l'auto-application d'algorithmes (par exemple, Kozak 2004).

Il convient de souligner que la stratification géométrique ne tient compte ni de la formule de la variance (1), ni de la répartition de l'échantillon (2), ni des contraintes (5). Or, il peut arriver que l'une des contraintes (5) ne soit pas satisfaite. Par conséquent, la stratification géométrique est une méthode de stratification approximative.

Dans la présente étude, nous avons appliqué l'algorithme proposé par Kozak (2004) pour stratifier plusieurs populations. Il s'agit d'un algorithme de recherche aléatoire adapté au problème de la stratification. Cet algorithme est simple; à chaque étape, une limite de strate est sélectionnée aléatoirement et modifiée aléatoirement. Si le nouvel ensemble de

limites de strate est meilleur que le précédent, il remplace ce dernier. L'annexe décrit en détail l'algorithme basé sur l'article publié par Kozak (2004).

Le deuxième problème examiné dans le présent article est la construction de strates qui minimisent la taille de l'échantillon provenant d'une population sachant le niveau de précision voulu de l'estimation (précision qui est donnée par la variance d'un estimateur de la moyenne ou du total de population). L'algorithme de Lavallée-Hidiroglou (LH) (Lavallée et Hidiroglou 1988) peut être considéré comme une méthode d'optimisation particulière en vue de résoudre ce problème précis de stratification; par contre, il n'est pas applicable à d'autres problèmes, par exemple celui considéré plus haut. Pour des précisions sur l'algorithme, consulter l'article de Lavallée et Hidiroglou (1988). Outre l'algorithme LH, nous avons appliqué la méthode de stratification géométrique et de recherche aléatoire pour construire les strates.

Nous avons utilisé le langage et l'environnement R (R Development Core Team 2005) pour réaliser tous les calculs de la présente étude.

3. Comparaison numérique de l'efficacité des approches de stratification sous taille d'échantillon fixe

À la présente section, nous comparons deux approches de stratification, la stratification géométrique (geom) et l'approche par optimisation (optim), appliquées à un problème de recherche des limites de strate qui minimisent la variance de l'estimateur considéré sachant une taille fixe d'échantillon. Pour réaliser la comparaison, nous avons généré cinq populations artificielles de tailles différentes (allant de 2 000 à 10 000). Les statistiques sommaires de ces populations sont présentées au tableau 1; les histogrammes des variables de stratification dans les populations sont donnés à la figure 1. Dans chaque cas, la variable de stratification était positivement asymétrique (le coefficient d'asymétrie variait de 1,40 pour la première population à 5,02 pour la cinquième). Comme cela est généralement le cas dans les populations réelles, les valeurs des variables de stratification étaient des nombres entiers. La taille d'échantillon, n_i, pour la i^{e} population était $n_i = f N_i$, où $f = 0{,}15$ est une fraction d'échantillonnage hypothétique et N_i est la taille de la i^{e} population.

Pour commencer, nous avons stratifié chaque population par la méthode de stratification géométrique en 4, 5, 6 et 7 strates. Puis, nous avons appliqué l'approche par optimisation; dans cette dernière, nous avons utilisé comme paramètres initiaux les limites de strate déterminées par la méthode de stratification géométrique.

Tableau 1
Statistiques sommaires pour les populations artificielles étudiées

Population	Taille	Étendue	Asymétrie	Moyenne	Variance
1	4 000	3–72	1,40	16,11	45,8
2	4 000	243–28 578	2,66	2 823,95	$4{,}8 \times 10^6$
3	2 000	6–2 793	3,55	224,12	$6{,}0 \times 10^4$
4	10 000	62–74 398	4,20	3 616,41	$2{,}1 \times 10^7$
5	2 000	259–186 685	5,02	9 265,36	$1{,}1 \times 10^8$

Comme Gunning et Horgan (2004), pour comparer l'efficacité des deux approches, nous avons calculé l'efficacité relative en appliquant la formule :

$$\text{eff}_{\text{geom, optim}} = \frac{V_{\text{geom}}(\bar{x}_{\text{st}})}{V_{\text{optim}}(\bar{x}_{\text{st}})}, \qquad (8)$$

où $V_{\text{geom}}(\bar{x}_{\text{st}})$ et $V_{\text{optim}}(\bar{x}_{\text{st}})$ sont les variances (1) sous les approches géométrique et par optimisation, respectivement. En outre, nous avons calculé les coefficients de variation de l'estimateur de la moyenne de population sous les deux approches :

$$\text{cv}_{\text{geom}} = \frac{\sqrt{V_{\text{geom}}(\bar{x}_{\text{st}})}}{\bar{x}_{\text{st}}};\ \text{cv}_{\text{optim}} = \frac{\sqrt{V_{\text{optim}}(\bar{x}_{\text{st}})}}{\bar{x}_{\text{st}}}. \qquad (9)$$

Le tableau 2 contient les valeurs des efficacités relatives (8) et des coefficients de variation (9) pour chaque combinaison étudiée (population × nombre de strates).

Tableau 2
Coefficients de variation de l'estimateur de la moyenne de population sous les approches de stratification géométrique (CV_{geom}) et par optimisation (CV_{optim}), et efficacité de la stratification géométrique comparativement à l'approche par optimisation ($\text{eff}_{\text{geom,optim}}$)

Nombre de strates *L*	CV_{geom}	CV_{optim}	$\text{eff}_{\text{geom,optim}}$
	Population 1		
4	0,0086	0,0056	1,53
5	0,0070	0,0042	1,66
6	0,0057	0,0034	1,66
7	0,0051	0,0029	1,75
	Population 2		
4	0,0116	0,0084	1,37
5	0,0095	0,0065	1,47
6	0,0085	0,0051	1,66
7	0,0073	0,0042	1,72
	Population 3		
4	0,0235	0,0133	1,76
5	0,0174	0,0100	1,74
6	0,0146	0,0081	1,80
7	0,0129	0,0067	1,91
	Population 4		
4	0,0104	0,0063	1,64
5	0,0089	0,0047	1,88
6	0,0073	0,0038	1,93
7	0,0064	0,0032	2,00
	Population 5		
4	0,0235	0,0134	1,76
5	0,0185	0,0100	1,86
6	0,0161	0,0080	2,00
7	0,0134	0,0074	1,82

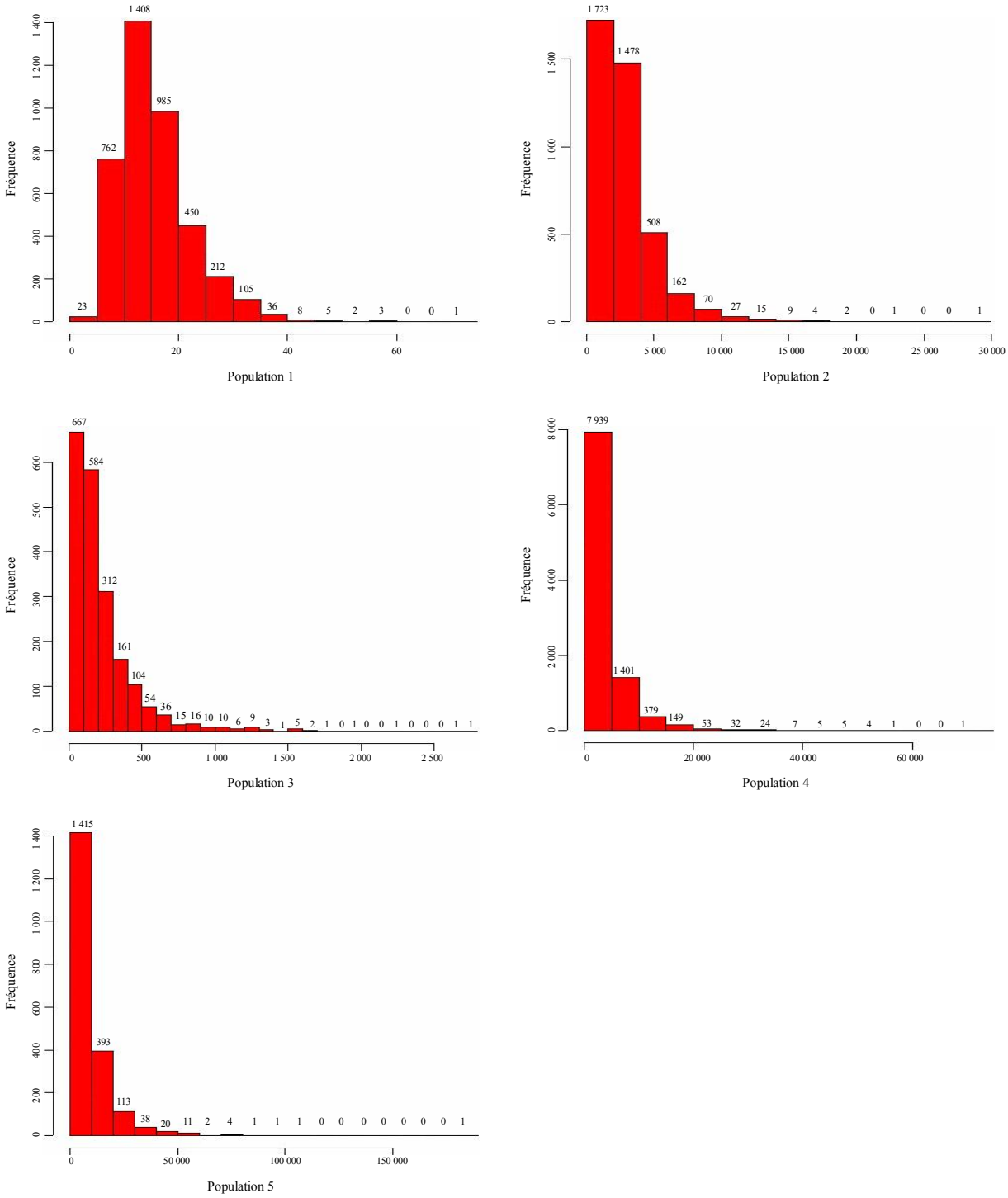

Figure 1. Histogrammes de la variable de stratification dans les populations artificielles étudiées.

Dans chaque cas, l'approche par optimisation a été plus efficace que la stratification géométrique. L'efficacité n'était inférieure à 1,5 que pour deux combinaisons; pour les autres, elle variait entre 1,5 et 2. En général, le gain d'efficacité est d'autant plus important que le nombre de strates construites est grand.

4. Comparaison numérique de l'efficacité des approches de stratification sous niveau de précision fixe de l'estimation

Gunning et Horgan (2004), ainsi que Horgan (2006) ont comparé la stratification géométrique à l'algorithme de Lavallée et Hidiroglou (Lavallée et Hidiroglou 1988) et constaté que la première était généralement plus efficace. À

la présente section, nous comparons les trois approches de stratification, à savoir la stratification géométrique, l'algorithme LH et l'approche par optimisation selon la méthode de recherche aléatoire. Nous avons utilisé pour la présente étude les cinq mêmes populations qu'à la section précédente (voir tableau 1 et figure 1).

Les efficacités relatives des approches comparées ont été évaluées au moyen de la formule

$$\text{eff}_{i,j} = \frac{n_i(\text{cv})}{n_j(\text{cv})}, \qquad (10)$$

où i et j sont les indices des approches de stratification (i, j = geom, optim, LH), et $n_i(\text{cv})$ et $n_j(\text{cv})$ sont les tailles d'échantillon minimales requises pour obtenir un niveau souhaité de précision (cv) sous les i^{e} et j^{e} approches, respectivement.

En suivant ces trois approches, nous avons stratifié chaque population en $L = 4, \ldots, 7$ strates; le niveau fixé de précision était de 0,01 dans chaque cas. Les tailles minimales d'échantillon requises pour ce niveau de précision et les efficacités relatives (10) sont données au tableau 3.

Tableau 3
Tailles d'échantillon minimales requises pour obtenir une valeur égale à 0,01 pour le coefficient de variation de l'estimateur de la moyenne de population, sous la stratification géométrique (n_{geom}), l'approche par optimisation (n_{optim}) et l'algorithme LH (n_{LH}); et efficacité de la stratification géométrique relativement à l'approche par optimisation ($\text{eff}_{\text{geom, optim}}$), de la stratification géométrique relativement à l'algorithme LH ($\text{eff}_{\text{geom, LH}}$) et de l'algorithme LH relativement à l'approche par optimisation ($\text{eff}_{\text{LH, optim}}$)

Nombre de strates L	n_{geom}	n_{optim}	n_{LH}	$\text{eff}_{\text{geom, optim}}$	$\text{eff}_{\text{geom, LH}}$	$\text{eff}_{\text{LH, optim}}$
Population 1						
4	805	496	496	1,63	1,63	1,00
5	613	344	344	1,78	1,78	1,00
6	460	252	252	1,83	1,83	1,00
7	357	192	192	1,86	1,86	1,00
Population 2						
4	483	248	259	1,94	1,86	1,04
5	329	154	163	2,14	2,02	1,06
6	224	113	117	1,98	1,92	1,03
7	180	83	83	2,17	2,17	1,00
Population 3						
4	782	410	411	1,91	1,90	1,00
5	601	303	304	1,98	1,98	1,00
6	495	242	241	2,04	2,05	1,00
7	422	195	195	2,11	2,16	1,00
Population 4						
4	839	409	409	2,05	2,05	1,00
5	650	301	301	2,15	2,15	1,00
6	552	240	242	2,30	2,28	1,01
7	–[1]	200	200	–	–	1,00
Population 5						
4	1 768	894	894	1,98	1,98	1,00
5	1 274	628	628	2,03	2,03	1,00
6	949	459	459	2,07	2,07	1,00
7	758	355	355	2,13	2,13	1,00

[1] L'obtention des limites de strate a posé des problèmes numériques (les tailles d'échantillon provenant de certaines strates étaient supérieures aux tailles de ces strates).

Il découle des résultats que l'approche par optimisation est plus efficace que la stratification géométrique; cette observation a été faite pour chaque population et chaque nombre de strates. L'efficacité relative était systématiquement supérieure à 1,6. En outre, une conclusion intéressante se dégage de la comparaison de l'efficacité des stratifications géométrique et LH. Comme nous l'avons déjà mentionné, Gunning et Horgan (2004), ainsi que Horgan (2006) ont constaté que la stratification géométrique était plus efficace que l'algorithme LH. Par contre, dans notre étude, l'algorithme LH était systématiquement plus efficace que la stratification géométrique, constatation que nous avons également faite pour d'autres populations de taille et d'asymétrie différentes que nous avons générées (les résultats ne sont pas présentés ici). Néanmoins, nous n'affirmons pas que l'algorithme LH est systématiquement plus efficace que la stratification géométrique. Il peut arriver que cette dernière donne de meilleurs résultats, comme Gunning et Horgan (2004) et Horgan (2006) l'ont observé lors de leurs études.

De la comparaison de l'algorithme LH à l'approche par optimisation, il découle que les deux méthodes donnent des points de stratification qui produisent des tailles d'échantillon semblables. Dans certains cas, la stratification LH est un peu meilleure et dans d'autres, un peu moins bonne, que l'approche par optimisation. Néanmoins, ces différences ne nous permettent pas de déclarer que l'une de ces deux approches est plus efficace que l'autre. En fait, elles ont toutes deux le même objectif (dans ce problème de stratification particulier) et diffèrent simplement en ce qui concerne l'algorithme utilisé pour atteindre cet objectif. Brièvement, d'après nos résultats, nous concluons qu'en général, la stratification LH et l'approche par optimisation sont plus efficaces que la stratification géométrique.

5. Conclusion

La méthode de stratification fondée sur une progression géométrique proposée par Gunning et Horgan (2004) possède un avantage significatif; plus précisément, son algorithme est très simple à appliquer comparativement à la méthode de la fonction cumulative de la racine carrée des fréquences de Dalenius et Hodges (1959) et à d'autres méthodes de stratification. Toutefois, il s'agit d'une méthode approximative, si bien que les limites de strate qu'elle produit peuvent mener à des estimations de précision médiocre (ou nécessiter l'utilisation d'un échantillon de grande taille pour obtenir le niveau requis de précision). En outre, il est probable que les strates construites ne satisferont pas toutes aux contraintes (5); autrement dit, il se peut que certaines strates soient vides (de sorte qu'elles ne contiendront aucune unité de population) ou (et) que la taille

des échantillons provenant de certaines strates soit inférieure à deux ou supérieure à la taille de population de la strate.

Dans notre étude, l'approche par optimisation (au moyen des algorithmes LH et de recherche aléatoire) s'est avérée plus efficace que la stratification géométrique pour chaque population étudiée et chaque nombre de strates construites. Néanmoins, les limites de strate données par la stratification géométrique peuvent être considérées comme de bons paramètres initiaux pour l'approche par optimisation; par contre, elles ne devraient pas être regardées comme des limites de strate optimales ou efficaces. De surcroît, nos résultats montrent de façon concluante que la stratification géométrique est moins efficace que celle présentée par Lavallée et Hidiroglou (1988), résultat opposé à celui obtenu par Gunning et Horgan (2004) et par Horgan (2006). L'étude de ce problème doit se poursuivre sur des populations asymétriques réelles; les recherches portant sur des populations artificielles indiquent sans équivoque que l'algorithme LH et l'approche par optimisation sont plus efficaces que la stratification géométrique.

À première vue, on pourrait s'étonner du fait que le gain d'efficacité réalisé en appliquant les approches LH et par optimisation comparativement à la stratification géométrique s'accroît lorsque le nombre de strates augmente. Toutefois, l'explication est simple. Le but de la stratification géométrique est d'égaliser les coefficients de variation de la variable de stratification dans les strates. Par conséquent, il diffère de celui de la stratification consistant à optimiser l'efficacité de l'estimation ou à minimiser la taille d'échantillon. Qui plus est, il n'est pas certain que, sous la stratification optimale, la distribution de la variable de stratification/d'enquête soit uniforme dans les strates. Les deux ensembles de limites de strate (c'est-à-dire ceux fournis par les approches géométrique et par optimisation) ne sont pas nécessairement les mêmes; en fait, il est probable qu'ils diffèrent.

Notons que nous appliqué l'algorithme de recherche aléatoire dans l'approche de stratification par optimisation. Or, l'algorithme de Lavallée et Hidiroglou (1988) est également un représentant des approches par optimisation. Quand le but de la stratification est de minimiser la taille d'échantillon requise pour obtenir un niveau souhaité de précision, il est probable que les deux approches produisent des résultats semblables, comme cela a été le cas lors de notre étude. Néanmoins, l'algorithme de recherche aléatoire peut être appliqué à n'importe quel problème de stratification (c'est-à-dire à toute fonction d'optimisation et ses contraintes), contrairement à l'algorithme LH, qui n'est applicable qu'à la minimisation de la taille d'échantillon pour un niveau de précision donné. Il convient de souligner que l'algorithme de recherche aléatoire fournit, comme méthode d'optimisation globale, des résultats aléatoires.

Notre but n'était pas, toutefois, de promouvoir l'un ou l'autre de ces deux algorithmes en montrant qu'ils sont plus efficaces que la stratification géométrique. Qui plus est, nous avons appliqué la méthode du simplexe de Nelder et Mead (1965) pour stratifier les populations (résultats non présentés ici); les résultats obtenus par cette méthode étaient fort semblables à ceux produits par les algorithmes LH et de recherche aléatoire. Chacune de ces méthodes présente certains inconvénients. Par exemple, des difficultés numériques peuvent survenir lors de l'utilisation de l'algorithme LH (Slanta et Krenzke 1996), tandis que la méthode de recherche aléatoire fournit des résultats aléatoires (Kozak 2004), la méthode de Nelder et Mead (1965) peut être inefficace si le nombre de strates et la taille de la population sont grands (Kozak 2004) et, en fait, l'obtention de points de stratification optimaux n'a été prouvée pour aucune de ces méthodes. Par conséquent, il reste encore à construire un algorithme de stratification produisant des résultats optimaux quelle que soit la situation (par exemple en ce qui concerne la taille de la population ou l'asymétrie de la variable), ainsi que des résultats non aléatoires. Notre objectif principal était de prouver que la stratification géométrique n'est pas optimale, mais que les points de stratification qu'elle produit peuvent être utiles comme paramètres initiaux dans d'autres approches de stratification.

Remerciements

Les auteurs remercient vivement les examinateurs et le rédacteur adjoint de *Techniques d'enquête* de leurs commentaires précieux, qui leur ont permis d'améliorer la première version du présent article.

Annexe

L'algorithme qui suit a été proposé par Kozak (2004) et nous nous sommes bornés à adapter certains de ses détails au problème général de la stratification. Dans l'algorithme, nous ne faisons pas référence au problème particulier de la stratification (autrement dit, nous ne définissons pas la fonction d'optimisation et ses contraintes), puisqu'il fonctionne pour les deux problèmes présentés dans l'article, ainsi que pour d'autres problèmes de stratification. Au besoin, nous faisons référence à la « fonction d'optimisation » (qui peut être la variance d'un estimateur étudié ou la taille d'un échantillon provenant d'une population) et aux « contraintes » (qui, selon la fonction d'optimisation, peuvent être les contraintes (5) et (6), ou les contraintes (5) combinées à la contrainte sur le niveau de précision de l'estimation); d'autres formes de la fonction d'optimisation et de ses contraintes peuvent sans aucun doute être prises en considération.

Définissons un vecteur comme il suit. Il prend des valeurs dans l'intervalle $(1, N)$, N étant la taille de population. À condition qu'une population soit triée en fonction des valeurs d'une variable de stratification X, deux éléments a_{h-1} et a_h du vecteur **a** définissent la strate h de telle façon que cette strate comprenne les éléments d'indice I (qui donne l'ordre d'un élément dans la population triée) tel que $a_{h-1} < I \le a_h$, $h = 1, \ldots, L, a_0 = 0, a_L = N$. L'algorithme est le suivant.

1. Trier la population en fonction des valeurs de la variable de stratification.
2. Choisir un vecteur initial **a**, c'est-à-dire le vecteur de limites de strate initiales. Des nombres entiers aléatoires qui satisfont les contraintes peuvent être utilisés, mais la pratique révèle que de meilleurs résultats peuvent être obtenus en utilisant les limites de strate approximatives déterminées par une méthode de stratification approximative. Calculer la valeur de la fonction d'optimisation. Vérifier les contraintes; si elles ne sont pas satisfaites, les points initiaux doivent être modifiés.
3. Pour $r = 0, 1, \ldots, R$ répéter l'étape suivante :
   a. Générer le point $\mathbf{a}'$ en tirant une limite de strate a_i puis en la modifiant comme il suit

$$a_i' = a_i + j,$$
$$a_k' = a_k \quad \text{for } k = 1, \ldots, L-1, k \ne i, \qquad (11)$$

   où j est le nombre entier aléatoire, $j \in \langle -p; -1 \rangle \cup \langle 1; p \rangle$, p étant un nombre entier donné choisi d'après la taille de population (la valeur de p est d'autant plus élevée que la population est grande); habituellement, p devrait être compris entre 3 et 5.
   b. Calculer la valeur de la fonction d'optimisation.
   c. Si les contraintes sont satisfaites et que la valeur de la fonction d'optimisation sous le vecteur $\mathbf{a}'$ est plus petite que celle obtenue sous le vecteur **a**, accepter le nouveau vecteur, c'est-à-dire $\mathbf{a}_{r+1} = \mathbf{a}'$ (où $\mathbf{a}_{r+1}$ est le vecteur de limites de strate dans une itération suivante); sinon, ne pas accepter le vecteur, c'est-à-dire $\mathbf{a}_{r+1} = \mathbf{a}$.
4. Finir l'algorithme si la règle d'arrêt est satisfaite, c'est-à-dire si $r = R$, où R est le nombre donné d'étapes ou que, lors des m (par exemple, 50) dernières étapes, la valeur de la fonction d'optimisation n'a pas varié. Enfin, calculer le vecteur **k** (le vecteur de limites de strate finales) en fonction des valeurs du vecteur **a**.

Bibliographie

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

Dalenius, T., et Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

Ekman, G. (1959). An approximation useful in univariate stratification. *Annals of Mathematical Statistics*, 30, 219-229.

Glasser, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*, 30, 28-32.

Gunning, P., et Horgan, J.M. (2004). Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques. *Techniques d'enquête*, 30, 177-185.

Gunning, P., Horgan, J.M. et Yancey, W. (2004). Geometric stratification of accounting data. *J. de Contaduria y Administration*, 214, septiembre-diciembre.

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

Horgan, J.M. (2006). Stratification of skewed populations: A review. *Revue Internationale de Statistique*, 74(1): 67-76.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5), 797-806.

Lavallée, P., et Hidiroglou, M. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 35-45.

Lednicki, B., et Wieczorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, 6, 287-306.

Nelder, J.A., et Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7, 308-313.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; URL http://www.R-project.org.

Rivest, L.-P. (2002). Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, 28, 207-214 (http://www.mat.ulaval.ca/pages/lpr/).

Slanta, J., et Krenzke, T. (1996). Utilisation de la méthode de Lavallée et Hidiroglou pour le calcul des limites de stratification aux fins de l'enquête annuelle sur les dépenses en capital du Bureau of the Census. *Techniques d'enquête*, 22, 65-75.