Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 65–75
Statistique Canada

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

JOHN G. SLANTA et THOMAS R. KRENZKE[1]

RÉSUMÉ

On a utilisé la méthode de Lavallée-Hidiroglou (L-H) pour le calcul des bornes de stratification dans le cadre de l'enquête annuelle du Bureau of the Census sur les dépenses en capital (Annual Capital Expenditures Survey, ou ACES), afin de stratifier une portion de l'univers de cette enquête aux fins de l'étude pilote et des enquêtes préliminaires subséquentes. Cette méthode itérative minimise la taille de l'échantillon tout en fixant le niveau souhaité de fiabilité en établissant des bornes appropriées. Toutefois, deux problèmes se sont posés lors de son application. D'abord, des bornes de départ différentes ont engendré des bornes finales différentes. Deuxièmement, la convergence vers les bornes localement optimales a été lente; autrement dit, le nombre d'itérations était élevé et la convergence n'était pas garantie. Le présent article décrit les difficultés rencontrées avec la méthode L-H et montre comment elles ont été résolues de manière à ce que cette méthode fonctionne correctement pour l'ACES. Nous décrivons notamment comment les tracés de contours ont été établis et utilisés afin d'illustrer l'importance des problèmes relevés lors de l'utilisation de la méthode L-H. Nous décrivons en outre les changements apportés à la méthode L-H afin d'en faire un moyen utile de détermination des bornes de stratification pour l'ACES.

MOTS CLÉS: Convergence; tracés de contours; enquêtes économiques.

1. INTRODUCTION

Le plan de sondage de l'enquête annuelle du Bureau of the Census sur les dépenses en capital (Annual Capital Expenditures Survey ou ACES) a pour objectifs principaux de procurer les niveaux de fiabilité souhaités à l'aide de méthodes faisables au plan opérationnel tout en respectant les contraintes budgétaires imposées. Pour répondre à ces objectifs, nous avons mis au point un plan d'échantillonnage aléatoire simple stratifié fondé sur une version modifiée de la méthode de détermination des bornes de stratification de Lavallée et Hidiroglou (L-H) (1988). Cette méthode de stratification pour les populations asymétriques donne des bornes optimales en minimisant la taille de l'échantillon total, compte tenu d'un coefficient de variation (c.v.) souhaité. Les personnes qui sont chargées d'une enquête spécialisée et qui ont accès à une variable de stratification unique peuvent tirer parti de la facilité d'exécution de cette méthode et des économies qu'elle permet de réaliser.

Nous avons examiné plusieurs articles qui portaient sur d'autres méthodes de détermination des bornes des strates fondées sur la taille. Hess, Sethi et Balakrishnan (1966) ont comparé plusieurs méthodes de stratification. La populaire méthode de Dalenius et Hodges (Cochran 1977, p. 129) a été jugée facile à mettre en oeuvre dans notre cas, mais nous l'avons rejetée au départ puisqu'elle ne tient pas compte des strates à tirage complet au moment de sa mise au point. La méthode de Sethi (1963) utilisant les distributions normales n'a pas été retenue parce que nous avons jugé qu'il serait encombrant d'identifier la distribution et que l'utilisation des distributions normales pour chacune des 80 types d'industrie examinés dans le cadre de l'enquête ne constituerait pas une solution optimale. La règle de l'égalité du produit des poids et de la portée des strates proposée par Eckman (1959) nous a semblé exiger des calculs plutôt rébarbatifs.

La méthode L-H est donc apparue la plus propice à nos besoins. Conçue précisément pour les populations asymétriques, une situation qu'on observe souvent avec des enquêtes économiques, cette méthode calcule une borne qui définit la strate à tirage complet, ainsi que les bornes optimales pour les strates à tirage partiel. En outre, elle crée parfois une strate à tirage complet supplémentaire si, sous l'effet de la répartition de Neyman, la taille de l'échantillon de la strate est plus grande ou égale à la taille de la strate.

La méthode L-H utilise un algorithme itératif qui débute avec le calcul ou la détermination arbitraire des bornes initiales des strates. Ensuite, on procède au calcul des paramètres tels que la taille, la moyenne et la variance des strates. Ces paramètres sont incorporés dans les formules de calcul des bornes, elles-mêmes dérivées d'une minimisation de la taille de l'échantillon en fonction d'une valeur souhaitée du c.v. Si les nouvelles bornes ne convergent pas, les paramètres des strates sont calculés à nouveau pour la strate nouvellement définie. Le cycle se poursuit jusqu'à ce que les bornes convergent.

Schneeberger (1979) a abordé le problème de la détermination des bornes de stratification optimales. Il a montré dans son article que lorsqu'on aborde ce problème sous l'angle d'un programme non linéaire, résolu par une méthode

[1] John G. Slanta, Manufacturing and Construction Division; Thomas R. Krenzke, Decennial Statistical Studies Division, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.

du gradient, la solution obtenue peut être une valeur minimale, maximale, tant locale que globale, ou encore un point de selle de la variance de la moyenne de l'échantillon. Detlefsen et Veum (1991) ont jugé que cette caractéristique de la méthode L-H constituait un inconvénient lorsqu'ils en ont étudié l'application pour l'enquête mensuelle sur le commerce au détail du Bureau of the Census. Ils ont observé que les bornes obtenues avec la méthode L-H étaient souvent sensiblement différentes des bornes fixées au départ, de sorte que la taille minimale de l'échantillon obtenu correspondait à un minimum local. Représentée géométriquement, la taille de l'échantillon en fonction de deux bornes de stratification prend l'aspect d'un paysage marqué d'une ou plusieurs vallées en forme de cuvettes. La méthode L-H commence dans une région et descend jusqu'au point le plus bas. S'il existe plus d'un minimum, elle ne poursuit pas sa recherche du minimum global. En conséquence, un des objectifs visés consiste à avoir des bornes initiales qui se situent au voisinage du minimum global. Le recours à des bornes de départ découlant d'une méthode comme celle de Dalenius et Hodges pourrait permettre de répondre à cette exigence.

Detlefsen et Veum (1991) ont également relevé des cas de convergence lente ou de non-convergence. Toutefois, ils ont également noté que la convergence est plus rapide lorsque le nombre de strates est réduit et lorsque les bornes de départ sont les mêmes que les bornes de sélection d'échantillon de l'enquête précédente. Afin de nous prémunir contre un nombre de boucles infinies dans le programme informatique ou contre un nombre très élevé d'itérations, nous avons décidé de faire deux choses: premièrement, nous avons mis en oeuvre un plan d'échantillonnage dans lequel la méthode L-H créerait des ensembles de seulement trois strates fondées sur la taille. Deuxièmement, nous avons décidé de mettre en oeuvre des règles d'interruption qui feraient en sorte d'interrompre le programme lorsque le taux de convergence semblait diminuer.

Dans le présent article, nous fournissons des informations de base sur l'ACES et nous décrivons brièvement notre utilisation de la méthode L-H. Nous montrons comment les tracés de contours et les graphiques tridimensionnels justifient l'utilisation de la méthode L-H pour obtenir les bornes finales. Nous expliquons comment les tracés de contours permettent de résoudre le problème de la convergence en montrant comment les contraintes peuvent être réglées de manière à être prises en compte après chaque itération. Ce procédé nous protège contre la convergence trop lente ou la non-convergence, lorsqu'on présume que les avantages attendus ne valent pas les efforts supplémentaires requis.

2. CONTEXTE DE L'ACES

L'ACES de 1992 a été conçue par le Bureau of the Census pour devenir un test opérationnel des méthodes d'échantillonnage, de traitement, de programmation, de saisie des données, de révision et d'estimation qui se prolongerait au-delà d'une étude pilote conduite en 1991, en prévision de la tenue de l'enquête définitive, en 1993. Les valeurs estimées des dépenses en capital correspondant aux activités intérieures ont été publiées par groupes de branches à partir de l'enquête de 1992. En outre, les enquêtes préliminaires de 1991 et de 1992 ont fourni de précieuses données sur les dépenses en capital qui serviront aux améliorations futures des plans d'échantillonnage.

L'unité de sondage de l'ACES était l'entreprise, qui pouvait comprendre plusieurs établissements. La population échantillonnée comprenait toutes les entreprises actives comptant cinq employés ou plus et oeuvrant dans tous les principaux secteurs industriels, exception faite du gouvernement. Ces secteurs comprenaient les mines, la construction, le secteur manufacturier, le transport, le commerce en gros et au détail, les finances, les services ainsi qu'une portion du secteur agricole comprenant les services agricoles, la foresterie, la pêche, la chasse et le trappage. Seules les entreprises nationales ont été incluses dans la base de sondage. Le personnel du service de recherche et de méthodologie de la Division de l'industrie du Bureau of the Census s'est chargé d'établir la base de sondage, de choisir l'échantillon et de produire les estimations.

La base d'échantillonnage de l'ACES a été construite à partir de la Standard Statistical Establishment List (SSEL) du Bureau of the Census en novembre 1992, en utilisant les données finales de 1991 pour les établissements à une seule unité (Single Unit ou SU) et les données de 1990 pour les établissements associés à des entreprises à unités multiples (multiunit ou MU). Ont été exclus de la base de sondage les services gouvernementaux, le service postal, les établissements internationaux, les établissements de Porto Rico, de Guam, des îles Vierges et des îles Mariannes. Les données des registres secondaires d'identification des employeurs, qui correspondent à des établissements SU figurant sur la liste SSEL, mais associés à des établissements MU, les établissements associés à la production agricole et les ménages privés ont également été exclus de la base de sondage.

Le fichier fondé sur les établissements a été fondu en un fichier fondé sur les entreprises. En outre, les codes à quatre chiffres de la classification type des industries (Standard Industrial Classification ou SIC) correspondant à chaque entreprise ont été remplacés par des codes de catégorie ACES. Les 80 catégories ACES étaient constituées soit de codes SIC à trois chiffres ou de combinaison de ces codes. La base de sondage de l'ACES comprenait environ deux millions d'entreprises.

3. APPLICATION DE LA MÉTHODE L-H À L'ACES

L'univers de l'enquête a été divisé en deux strates principales. La strate I a été définie arbitrairement comme une strate à tirage complet constituée de grandes entreprises comptant plus de 500 employés et disposant d'un actif de plus de 100 millions de dollars. Les entreprises de la strate I n'ont pas été classées dans une industrie ACES. Pour les totaux estimés de la paye utilisés dans le calcul de la taille des échantillons du niveau industriel, les entreprises de la strate I pouvaient contribuer à plus d'une industrie ACES, selon le nombre des branches ACES différentes pour lesquelles elles possédaient des feuilles de paye, tel qu'indiqué dans la liste SSEL.

La strate II contenait les entreprises comptant entre cinq et 500 employés. Les entreprises de la strate II ont été classées en une seule industrie, même si elles s'adonnaient à plus d'une activité. Chaque entreprise possédait des informations de base pour chacune des branches ACES dans lesquelles elle était active. Toutefois, la feuille de paye d'une entreprise donnée contribuait uniquement à la paye totale estimée versée dans la branche dans laquelle elle était classée. Par la suite, à l'intérieur de la strate II, pour chaque catégorie d'industrie ACES, trois strates fondées sur la taille ont été créées en fonction de la paye totale annuelle de l'entreprise à l'aide de la méthode L-H.

Ce plan de sondage présente un risque de mauvaise classification des entreprises par suite de la mesure de la taille choisie. Nous avons classé chaque entreprise de la strate II dans la branche où elle versait la paye la plus élevée. Toutefois, les entreprises indiquent elles-mêmes leurs dépenses en capital consacrées aux branches ACES dans le questionnaire ACES. Elles peuvent indiquer des dépenses dans de multiples branches. Si trop d'entreprises signalent des dépenses dans des branches autres que celle dans laquelle elles ont été classées, on perd alors toute emprise sur la fiabilité des estimations.

On s'inquiète par ailleurs d'un écart possible entre la variation de la paye et la variation des dépenses. Comme la taille de l'échantillon est directement liée à la variance, il se peut que les tailles d'échantillons soient différentes de ce qui est réellement nécessaire. En conséquence, comme la corrélation entre la paye et les dépenses n'est pas élevée, les possibilités de répondre aux exigences de fiabilité s'amenuiseront.

L'application de la méthode L-H au plan de sondage préliminaire de l'ACES de 1992 exigeait la séparation de la strate II en une strate à tirage complet et en deux strates à tirage partiel pour chaque industrie de l'ACES. Les bornes ont été dérivées pour chaque industrie en prenant la dérivée partielle de la taille de l'échantillon pour une des bornes, tout en fixant l'autre. Toutefois en pratique, nous avons permis aux deux bornes de se déplacer simultanément. Nous avons ainsi obtenu un processus itératif de minimisation de la taille de l'échantillon pour chaque industrie, compte tenu des contraintes imposées par le c.v. À l'intérieur de la strate II, pour chaque industrie de l'ACES, et en tenant pour acquise la répartition de Neyman (Detlefsen et Veum 1991), l'équation de la taille de l'échantillon qui est minimisée se présente comme suit:

$$n = n_{TA} + \frac{N\left(\sum_{j=1}^{2} W_j S_j\right)^2}{\frac{cv^2 Y^2}{N} + \sum_{j=1}^{2} W_j S_j^2}, \qquad (1)$$

où n_{TA} désigne le nombre d'entreprises dans la strate à tirage complet à l'intérieur de la strate II définie par la méthode L-H, N désigne le nombre d'entreprises de la strate II dans la branche ACES qui nous intéresse, $W_j = N_j/N$ désigne la proportion de la strate, N_j désigne le nombre d'entreprises de la strate II pour la strate j, c.v. désigne le coefficient de variation désiré pour la branche ACES qui nous intéresse, Y désigne la paye totale pour les strates I et II de la branche ACES qui nous intéresse, définie par

$$Y = \sum_{k=1}^{N_I} y_k + \sum_{j=1}^{3} \sum_{i=1}^{N_j} y_{ji},$$

N_I désigne le nombre d'entreprises de la strate I et S_j désigne l'écart-type de la paye tiré de la liste SSEL pour la strate j dans la strate II définie par

$$S_j = \sqrt{\frac{\sum_{i=1}^{N_j} (y_{ji} - \bar{Y}_j)^2}{N_j - 1}},$$

où y_{ji} désigne la valeur de la paye de l'entreprise i appartenant à la strate j pour la branche ACES qui nous intéresse et $\bar{Y}_j$ désigne la moyenne de la paye pour la strate j.

Le taux de fiabilité pour chaque industrie correspondait à une valeur attendue du c.v. de 5% sur la paye. Nous ignorons toutefois quelles seraient les valeurs de l'écart-type découlant des dépenses en capital, puisqu'il n'existait pas de données sur les dépenses en capital pour la base de sondage. Les entreprises répondant dans des branches ACES différentes de celles pour lesquelles elles avaient contribué dans le plan de sondage ont également causé des fluctuations des valeurs du c.v. Le nombre total d'entreprises sélectionnées pour l'enquête préliminaire de l'ACES de 1992 était de 11,194, soit 1,500 entreprises de la strate I et 9,694 entreprises de la strate II.

4. CONVERGENCE EN VOISINAGES

Il peut parfois arriver que la méthode L-H exige un grand nombre d'itérations avant que les bornes ne convergent; il arrive même que ces bornes ne convergent jamais. En règle générale, après quelques itérations seulement, une vaste proportion des améliorations possibles de la taille de l'échantillon ont déjà été obtenues. Notre but était de mettre en pratique des règles d'interruption qui permettraient d'arrêter le processus dès qu'on atteindrait une certaine superficie autour d'un minimum local. C'est ce qui nous a poussés à utiliser des tracés de contours dans notre analyse de l'effet des bornes sur la taille de l'échantillon obtenue. Cela nous a également permis d'obtenir un aperçu graphique des voisinages des minimum locaux. Nous utiliserons deux distributions pour illustrer les avantages d'une révision des tracés de contours.

4.1 Distribution symétrique

Notre premier exemple est une distribution symétrique tirée de l'article de Schneeberger. Cette distribution est symétrique au point $x = 1$, comme l'illustre la figure 1.

$$f(x) = \begin{matrix} 0 \\ 2x \\ 2(1-x) \\ 2(x-1) \\ 2(2-x) \\ 0 \end{matrix} \quad \text{pour} \quad \begin{matrix} x \leq 0 \\ 0 < x \leq 0.5 \\ 0.5 < x \leq 1 \\ 1 < x \leq 1.5 \\ 1.5 < x \leq 2 \\ 2 < x \end{matrix}$$

Schneeberger cherchait à définir les bornes de trois strates à tirage partiel en utilisant une méthode du gradient. Les résultats obtenus à l'aide de la fonction objective de $z = (\sum W_h \sigma_h)^2$ sont présentés au tableau 1.

Tableau 1

Bornes optimales pour une distribution non asymétrique

	b_1	b_2	Point optimal
(2a)	.50241	1.03985	Point minimal
(2b)	.70910	1.29090	Point de selle
(2c)	.96015	1.49759	Point minimal

Source: Schneeberger (1979).

Nous avons généré cinq ensembles de données de tailles différentes (N = 50, 100, 200, 1,000 et 5,000) en utilisant la formule $F(x) = (j - 1/2)/N$. Pour cet exemple, nous avons adapté la méthode L-H afin d'établir trois strates à tirage partiel et aucune à tirage complet et de pouvoir ainsi comparer nos résultats avec ceux de l'article de Schneeberger. Dans notre application de l'estimation des totaux, avec minimisation de la taille de l'échantillon sous réserve d'une valeur du c.v. de 5%, nous utilisions la méthode L-H pour chacune des cinq tailles de population en utilisant trois méthodes différentes de démarrage. Les résultats de cette opération sont présentés au tableau 2.

Les informations du tableau 2 appellent trois commentaires principaux. Premièrement, la convergence des algorithmes dépend de la taille de la population. La théorie sous-jacente de la méthode L-H est fondée sur des distributions continues. Nos exemples, à l'image de toute application de sondage, sont faits de données discrètes tirées de populations finies. On observe également qu'à mesure que la valeur de N augmente, les bornes obtenues se rapprochent de la valeur minimale correspondant à une population de taille infinie. La figure 2 montre la rugosité de la surface de la taille de l'échantillon lorsque la valeur de N est petite (N = 50). La représentation tridimensionnelle de la figure 2 montre bien la forme en selle de la surface obtenue. Dans ce graphique, les axes représentent les bornes inférieure et supérieure et la surface représente les tailles d'échantillons obtenues. Ce graphique montre le point de selle, les deux minimum locaux, et donne également un aperçu de l'importance des réductions de taille de l'échantillon découlant d'un déplacement des bornes. Par comparaison, la figure 3 montre que la surface obtenue lorsque la valeur de N est élevée (N = 5000) est lisse. Ces résultats portent à conclure que la rugosité de la surface de la taille de l'échantillon et, en conséquence, la taille de la population auront un effet sur le lieu de convergence des bornes.

Cet exemple nous permet en outre de rappeler encore une fois que les bornes finales dépendent des bornes initiales. Pour cet exemple, Schneeberger indique qu'avec un point de départ symétrique à $x = 1$, où $b_1 = 1 - \lambda$ et $b_2 = 1 + \lambda (0 < \lambda < 1)$ et qui définit la ligne $b_2 = 2 - b_1$, la méthode du gradient déplace le gradient sur la ligne $b_2 = 2 - b_1$ jusqu'au point d'équilibre. Lorsque nous posons les bornes de départ sur cette ligne, comme cela s'est produit lorsque nous avons commencé avec $N_1 = N_2 = N_3$, la méthode L-H converge également au point d'équilibre (voir tableau 1). Avec les bornes de départ de

Tableau 2

Bornes L-H de trois strates à tirage partiel pour une distribution symétrique

N	Méthode de démarrage	1ère itération			Itération à moins de 5% de la taille de l'échantillon				Itération finale			
		b_1	b_2	n	b_1	b_2	n	itér. n°	b_1	b_2	n	itér. n°
50	$N_1 = N_2 = N_3$	.59	1.41	10.89	.66	1.34	9.98	2	.70	1.31	9.77	4
100	$N_1 = N_2 = N_3$	.59	1.41	12.60	.66	1.34	10.91	2	.70	1.30	10.55	5
200	$N_1 = N_2 = N_3$	.59	1.41	13.42	.66	1.34	11.43	2	.71	1.29	10.99	6
1000	$N_1 = N_2 = N_3$	.59	1.41	13.85	.66	1.34	11.75	2	.71	1.29	11.37	7
5000	$N_1 = N_2 = N_3$	.59	1.41	14.12	.66	1.34	11.84	2	.71	1.29	11.45	9
50	Dalenius-Hodges	.70	1.40	10.09	.70	1.40	10.09	1	.77	1.37	9.63	4
100	Dalenius-Hodges	.70	1.40	10.90	.84	1.40	10.14	7	.93	1.47	9.65	13
200	Dalenius-Hodges	.70	1.40	11.42	.83	1.40	10.44	7	.95	1.49	9.96	17
1000	Dalenius-Hodges	.70	1.40	11.86	.86	1.42	10.67	8	.96	1.50	10.27	23
5000	Dalenius-Hodges	.70	1.40	11.95	.86	1.42	10.74	8	.96	1.50	10.34	28
50	Hors ligne	.50	1.30	10.87	.57	1.20	9.43	3	.55	1.11	9.11	6
100	Hors ligne	.50	1.30	11.95	.57	1.18	10.04	3	.53	1.07	9.65	8
200	Hors ligne	.50	1.30	12.64	.56	1.14	10.28	4	.51	1.05	9.96	12
1000	Hors ligne	.50	1.30	13.24	.56	1.14	10.59	4	.50	1.04	10.27	18
5000	Hors ligne	.50	1.30	13.37	.56	1.14	10.67	4	.50	1.04	10.34	24

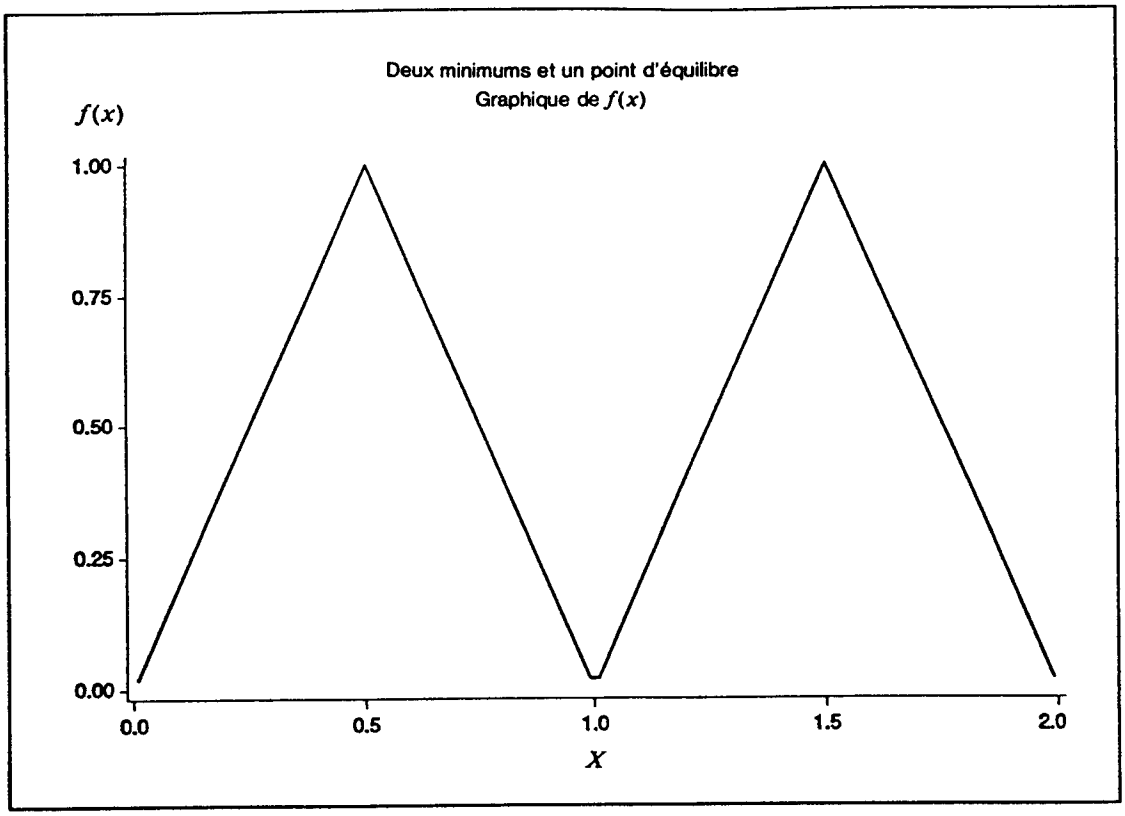

Figure 1. Graphique d'une distribution symétrique.

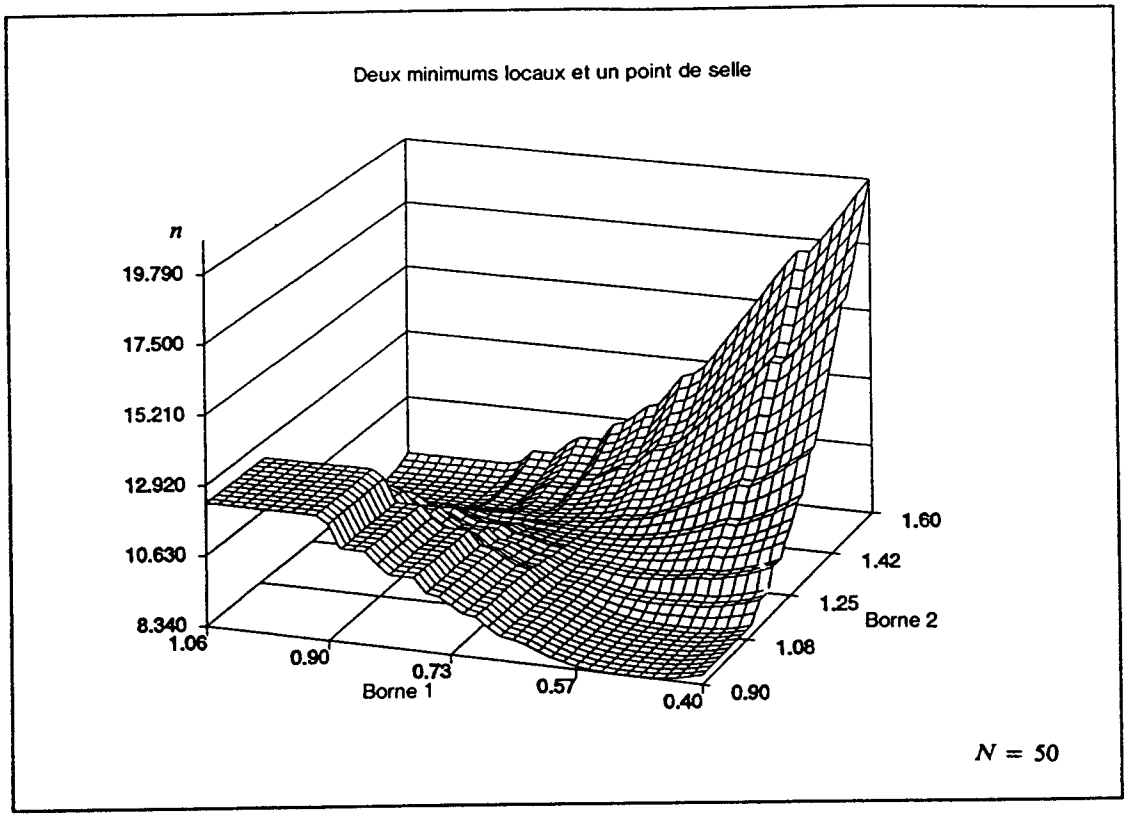

Figure 2. Surface de la taille de l'échantillon pour une distribution symétrique ($N = 50$).

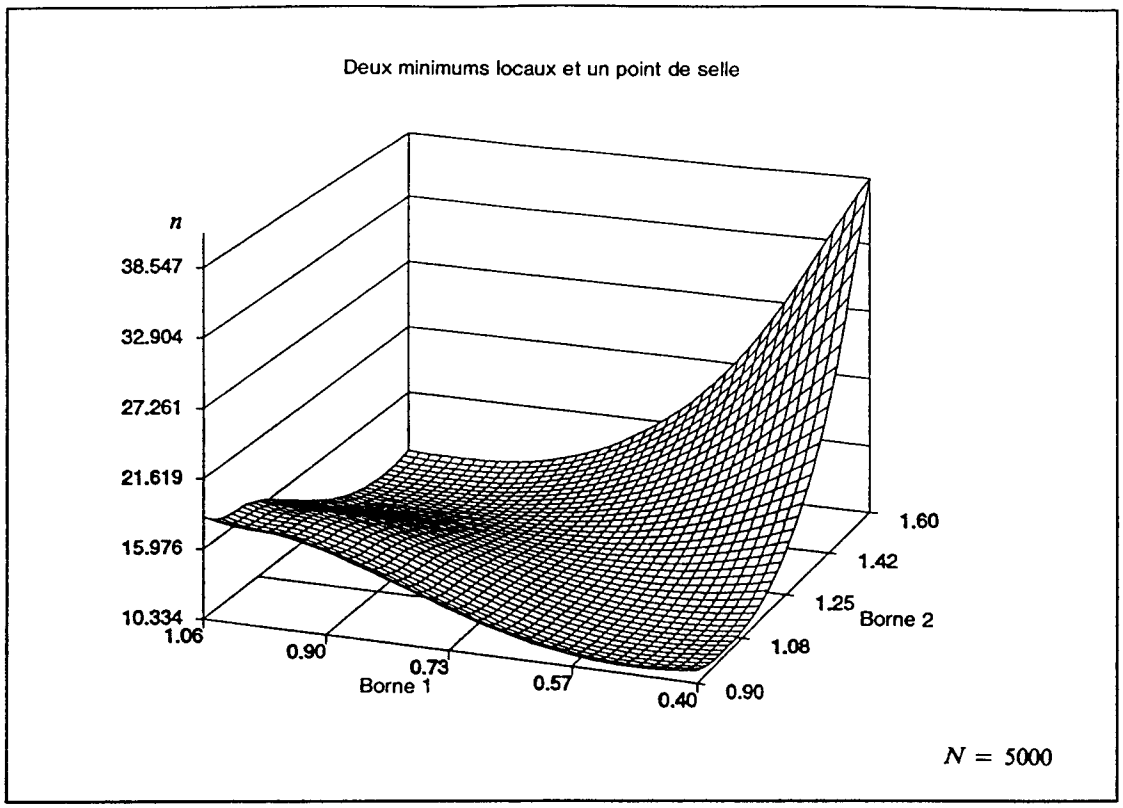

Figure 3. Surface de la taille de l'échantillon pour une distribution symétrique ($N = 5000$).

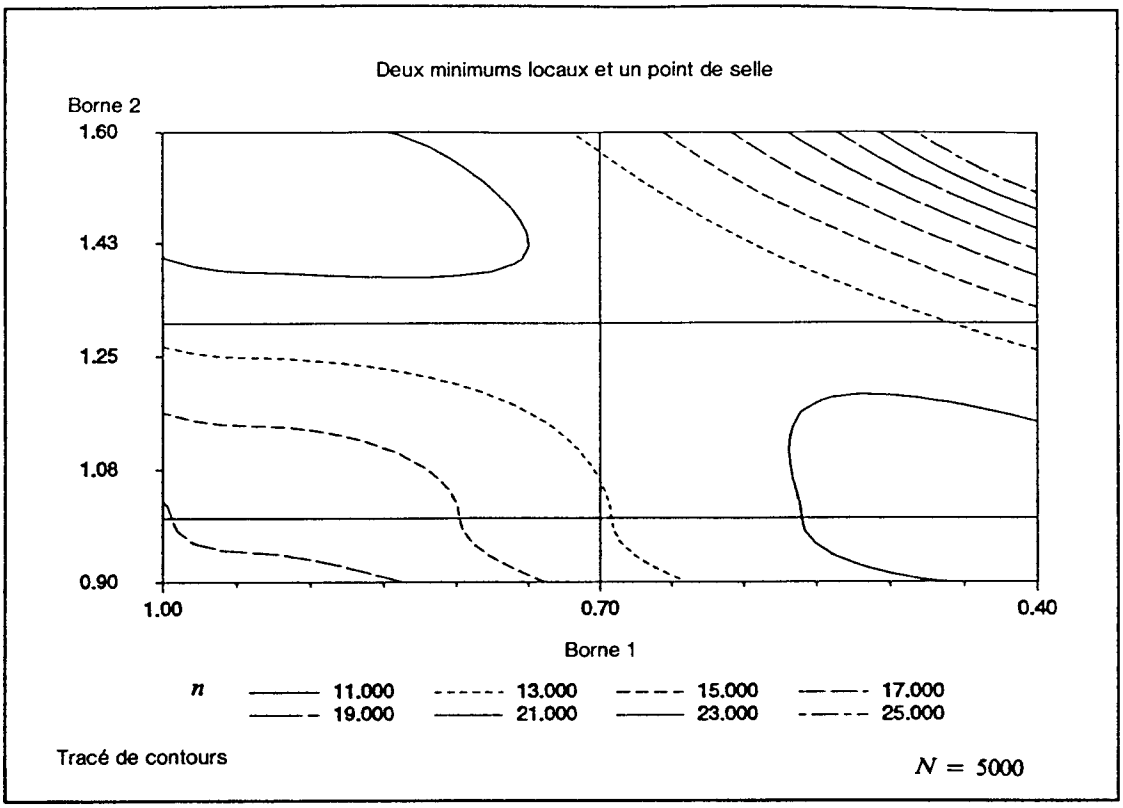

Figure 4. Tracé de contours pour une distribution symétrique ($N = 5000$).

la méthode de Dalenius-Hodges, lesquelles ne se situent pas sur la ligne lorsque $b_2 > 2 - b_1$, la méthode L-H converge vers un minimum (2c). La méthode de Dalenius-Hodges fonctionne bien dans cet exemple à cause des trois strates à tirage partiel. Avec des bornes de départ qui ne se situent pas sur la ligne dans le cas où $b_2 < 2 - b_1$ (dans ce cas-ci, $b_1 = .5$ et $b_2 = 1.3$), la méthode L-H converge vers un minimum différent (2a). Ce problème n'est pas unique à la méthode L-H puisque Schneeberger indique que les bornes obtenues par la méthode du gradient sont également dépendantes des bornes de départ.

On observe finalement dans cet exemple que les quelques premières itérations donnent des réductions relativement grandes de la taille de l'échantillon et que les itérations qui suivent donnent plusieurs réductions plus petites de cette taille. Nous présentons au tableau 2 les résultats de l'itération dans laquelle l'algorithme produit une taille de l'échantillon qui s'écarte de moins de 5% de la taille finale. Ceci signifie que l'algorithme L-H tend rapidement vers un voisinage autour de la borne optimale. Lorsque nous sommes près d'une taille optimale de l'échantillon, il semble exister une vaste gamme de bornes produisant un nombre limité de tailles d'échantillon. L'application de règles d'interruption peut effectivement réduire le temps de calcul sans nous conduire à renoncer à une véritable réduction de la taille de l'échantillon puisque les valeurs de ce paramètre sont données en nombres entiers.

Nous présentons à la figure 4 un tracé de contours de la surface illustrée à la figure 3. Ici encore, les axes représentent les bornes inférieures et supérieures des strates et la surface est définie par la taille d'échantillon qui en découle. Les lignes du tracé représentent une valeur de taille de l'échantillon. L'espace délimité correspond à une gamme de valeurs de la taille de l'échantillon. Par exemple, la ligne continue représente une taille d'échantillon de 11 et la ligne en tirets courts représente une taille d'échantillon de 13. La superficie délimitée par la ligne continue et la ligne en tirets courts contient les tailles d'échantillon comprises entre 11 et 13. Ce tracé de contours illustre une amélioration marginale de la taille de l'échantillon et montre que lorsqu'on atteint la portion inférieure de la surface, il n'est plus nécessaire de poursuivre l'opération. À partir de ce point, la plupart des améliorations de la taille de l'échantillon obtenue d'une itération à l'autre sont inférieures à l'unité. Il devient évident qu'après les quelques premières itérations, l'amélioration obtenue dans la taille de l'échantillon d'une itération à l'autre diminue rapidement. Par exemple, dans le tableau 2, lorsque $N = 5000$ et lorsqu'on a utilisé la méthode de Dalenius-Hodges pour déterminer les bornes de départ, les huit premières itérations ont donné 74% de la réduction totale de la taille de l'échantillon obtenue entre la première et la 28e et dernière itération.

4.2 Distribution asymétrique

Les données économiques sont habituellement hautement asymétriques, ce qui nous porte à accorder notre préférence à une strate à tirage complet. Le prochain exemple est tiré de la distribution de Pareto, une distribution extrêmement typique des univers économiques où il existe un grand nombre de petites entreprises et un petit nombre de grandes entreprises.

La fonction de la distribution de Pareto est définie par $F(x) = 1 - 1/(1 + x)^b$, $0 \leq x < \infty$. De cette fonction, nous avons encore une fois tiré cinq ensembles de données de tailles différentes en utilisant la formule $F(x) = (j - 1/2)/N$. Nous avons laissé les valeurs de b changer à mesure que la taille de la population évoluait.

Tableau 3

Bornes L-H d'une distribution asymétrique (une strate à tirage complet, deux strates à tirage partiel)

N	Méthode de démarrage	1ère itération					Itération à moins de 5% de la taille de l'échantillon					Itération finale					
		b	b_1	b_2	n_{TA}	n	b_1	b_2	n_{TA}	n	itér. n°	b	b_1	b_2	n_{TA}	n	itér. n°
50	$N_1 = N_2 = N_3$	.80	.63	2.81	17	17.2	1.66	10.20	7	9.6	5	.80	2.44	11.81	7	9.4	9
100	$N_1 = N_2 = N_3$	.90	.56	2.33	34	34.3	1.61	10.29	11	15.8	5	.90	2.58	12.44	10	15.1	12
200	$N_1 = N_2 = N_3$	.90	.56	2.36	67	67.2	2.35	17.04	15	21.8	6	.90	3.61	20.46	13	20.9	13
1000	$N_1 = N_2 = N_3$	1.00	.50	2.00	333	334.2	3.35	30.58	32	53.0	7	1.00	4.93	36.32	27	51.3	18
5000	$N_1 = N_2 = N_3$	1.05	.47	1.85	1665	1667.2	4.67	64.33	62	113.5	7	1.05	7.39	79.38	50	108.8	22
50	Dalenius-Hodges	.80	1.25	8.04	9	10.5	1.76	10.37	7	9.5	3	.80	2.44	11.81	7	9.4	6
100	Dalenius-Hodges	.90	1.39	8.98	13	16.6	1.62	10.16	11	15.8	2	.90	2.58	12.44	10	15.1	9
200	Dalenius-Hodges	.90	1.82	11.66	20	24.3	2.45	17.29	15	21.7	3	.90	3.61	20.46	13	20.9	10
1000	Dalenius-Hodges	1.00	2.37	17.28	55	65.6	3.15	29.70	33	53.5	3	1.00	4.93	36.32	27	51.3	15
5000	Dalenius-Hodges	1.05	3.09	26.27	155	175.0	4.98	66.28	60	112.3	4	1.05	7.39	79.38	50	108.8	19
50	Hidiroglou 1986	.80	.94	6.50	10	11.3	1.58	10.02	7	9.6	3	.80	2.44	11.81	7	9.4	7
100	Hidiroglou 1986	.90	.74	6.17	17	19.6	1.66	10.38	11	15.8	4	.90	2.58	12.44	10	15.1	11
200	Hidiroglou 1986	.90	1.39	9.55	24	27.2	2.50	17.58	14	21.5	4	.90	3.61	20.46	13	20.9	10
1000	Hidiroglou 1986	1.00	2.02	15.13	62	71.3	3.34	30.54	32	53.0	4	1.00	4.93	36.32	27	51.3	15
5000	Hidiroglou 1986	1.05	3.24	28.72	142	164.1	5.11	67.05	59	112.0	4	1.05	7.39	79.38	50	108.8	19

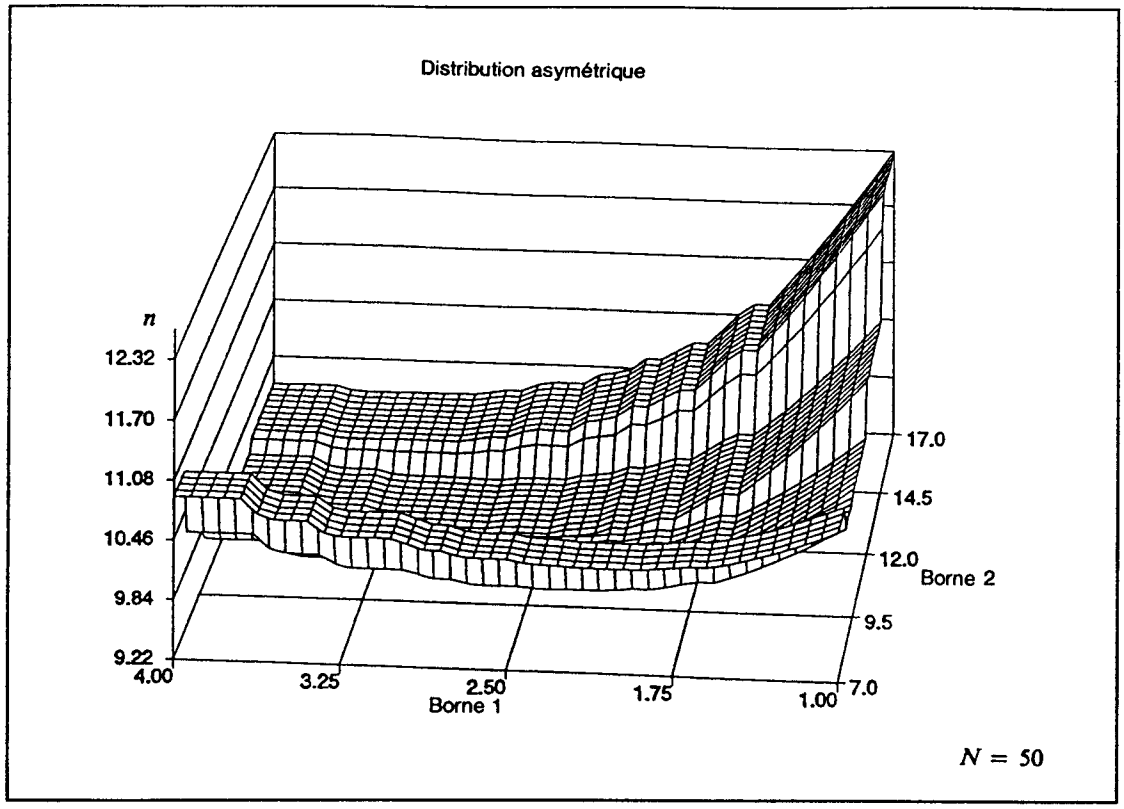

Figure 5. Surface de la taille de l'échantillon pour une distribution asymétrique ($N = 50$).

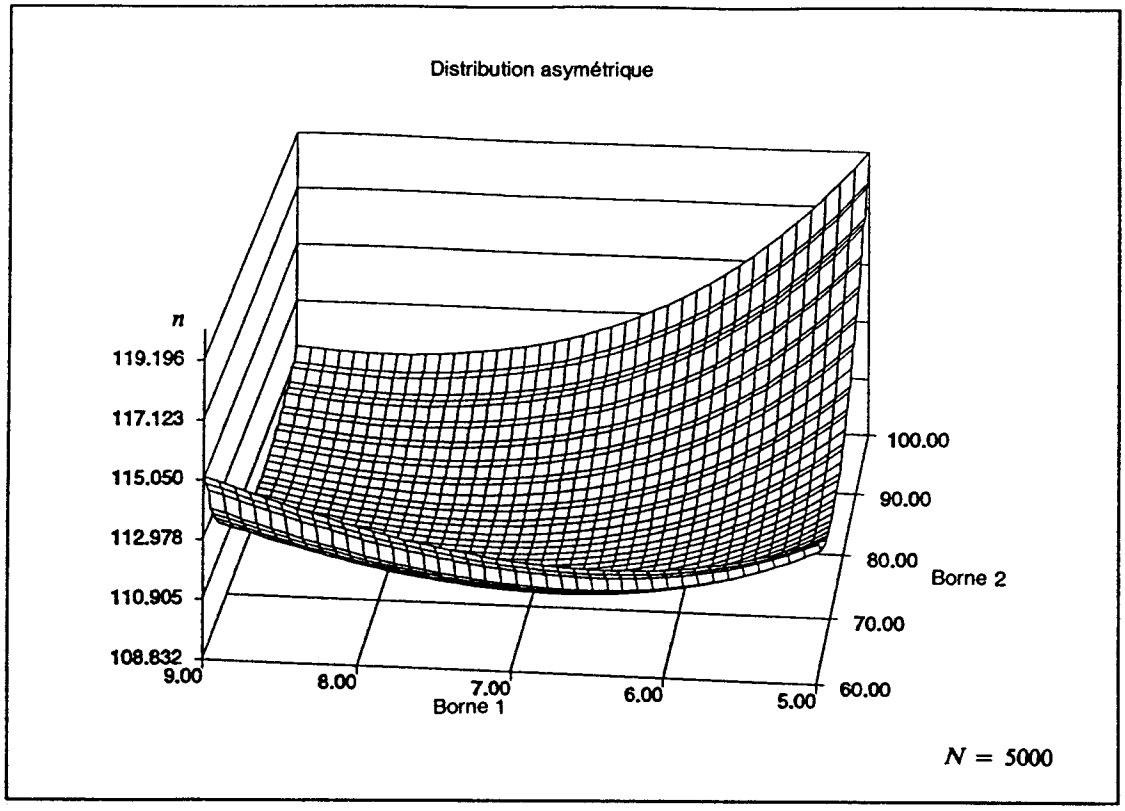

Figure 6. Surface de la taille de l'échantillon pour une distribution asymétrique ($N = 5000$).

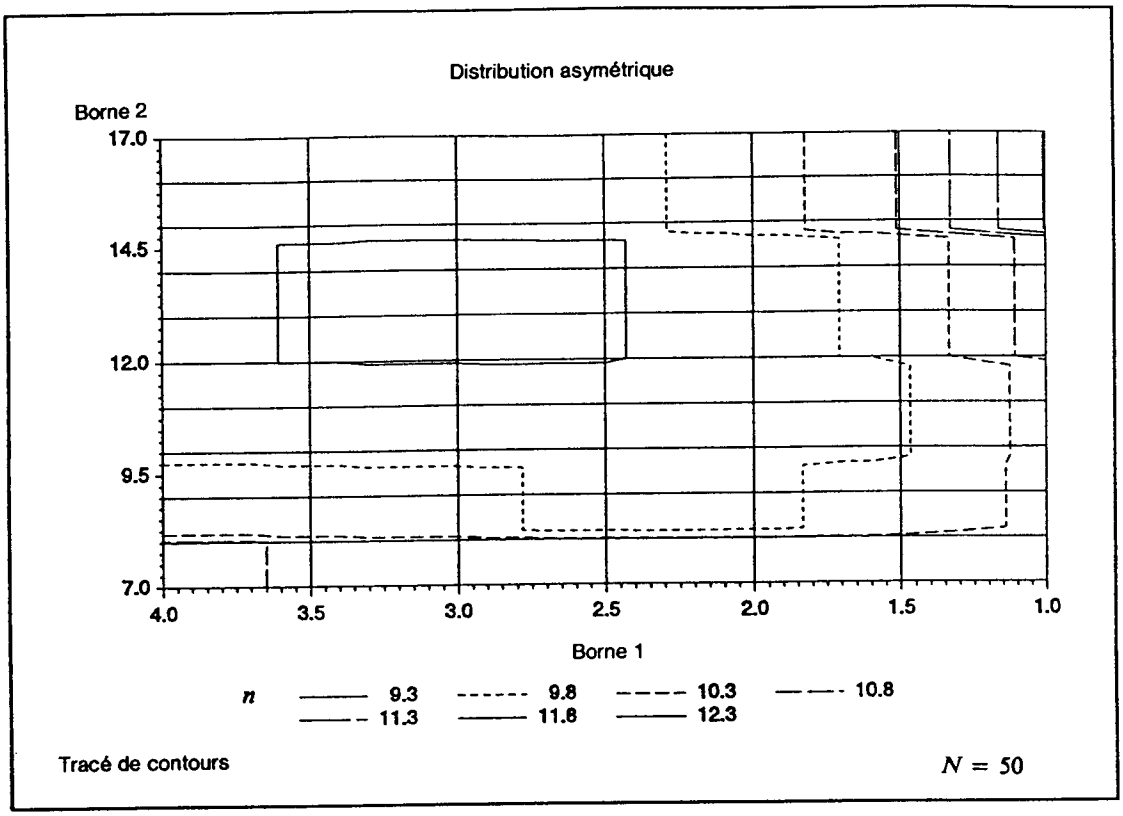

Figure 7. Tracé de contours d'une distribution asymétrique ($N = 50$).

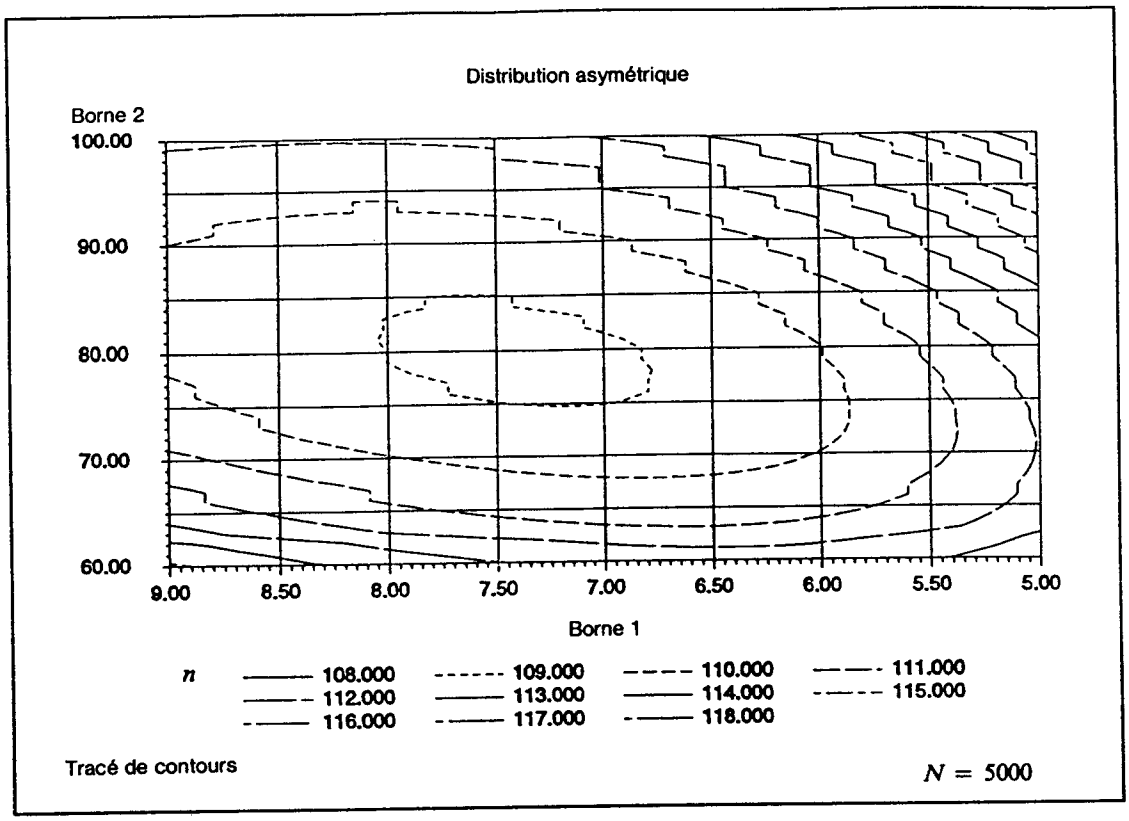

Figure 8. Tracé de contours d'une distribution asymétrique ($N = 5000$).

Nous avons procédé ainsi pour maintenir l'extrémité supérieure de la distribution discrète finie proportionnellement comparable, en gros, à la population entière pour chaque taille de population. À cette fin, nous avons choisi le paramètre b de manière qu'une proportion d'environ 90% de la somme totale soit prise en compte par les 20% supérieurs de l'ensemble des unités de sondage possibles. Puisque les ensembles de données contiennent un nombre fini de valeurs discrètes, la dérivation des variances des strates différentes lorsque les valeurs de b étaient inférieures à 2 n'a pas posé de problème.

Nous présentons au tableau 3 les résultats de l'application de la méthode L-H pour des tailles de population et des points de départ différents. Le premier groupe utilise des valeurs de départ qui donnent des populations de strates égales ($N_1 = N_2 = N_3$). Le deuxième groupe utilise la méthode de Dalenius-Hodges pour obtenir l'ensemble des bornes initiales. Le troisième groupe obtient les bornes initiales en utilisant une méthode de détermination de la borne de la strate à tirage complet proposée par Hidiroglou (1986) et utilise la méthode de Dalenius-Hodges pour les autres bornes. Ici encore, on peut observer que la surface représentant la taille de l'échantillon est beaucoup plus rugueuse pour les petites populations (voir figure 5). Par exemple, lorsque $N = 50$ et que b_1 est fixe, on n'obtient qu'une seule taille de l'échantillon lorsque b_2 varie entre 11.8 et 14.7. Ceci s'explique par l'absence de valeurs comprises à l'intérieur de cette gamme dans la population. À mesure que la taille de la population augmente, les valeurs des données se rapprochent et la surface de l'échantillon s'adoucit (voir figure 6).

Le tracé de contours correspondant à $N = 50$ (figure 7) présente des formes irrégulières délimitées par des lignes droites. Par contre, le tracé de contours correspondant à $N = 5000$ (figure 8) présente des formes elliptiques concentriques presque lisses. Une forme comparable et concentrique des tracés de contours semblerait être une qualité souhaitable puisque cela signifierait que le minimum global est le seul minimum local.

Le tracé de contours correspondant à $N = 50$ illustre le cas où la méthode L-H n'a pas donné une convergence vers des bornes optimales. Puisque, dans cet exemple, nous avons laissé le programme L-H tourner jusqu'à la convergence, il est permis de se demander pourquoi la méthode L-H n'a pas abouti à une convergence aux bornes optimales. Ce problème s'explique plus facilement à l'examen de la figure 5. Nous constatons en effet que lorsque la taille de la population est petite, la surface de la taille de l'échantillon n'est pas aussi lisse que dans la figure 6. Nous distinguons dans la figure 5 des crêtes importantes causées par les vastes lacunes dans les données discrètes asymétriques ($x_{43} = 9.71, x_{44} = 11.81, x_{45} = 14.79, x_{46} = 19.29$). Ceci signifie que pour une valeur donnée b_1, toute valeur de b_2 comprise entre 11.81 et 14.79 donnerait la même taille d'échantillon. Lorsque nous avons utilisé le programme L-H avec des bornes de départ différentes autres que les trois énumérées au tableau 3, nous avons obtenu les bornes finales du tableau 3 ainsi que d'autres bornes avec leurs valeurs correspondantes de la taille de l'échantillon. Il semble que la méthode L-H converge vers une région inférieure sur l'une des crêtes principales, à condition que cette région se trouve dans le voisinage des bornes optimales. La taille d'échantillon minimale est de 9.22 et la méthode L-H du tableau 3 donne une taille d'échantillon de 9.36. La valeur entière la plus petite de la taille de l'échantillon pour chaque résultat qui répond à la contrainte ou qui la dépasse est 10. Ici encore, nous constatons que la méthode L-H fonctionne exceptionnellement bien même avec des distributions discrètes dont les populations sont petites, puisque les bornes convergent à l'intérieur du voisinage contenant la solution optimale.

Il convient en outre de signaler que les bornes peuvent prendre une large gamme de valeurs tout en maintenant la même valeur entière de la taille de l'échantillon. À mesure que la taille du voisinage s'élargit, la gamme des bornes augmente également. Signalons finalement que même si la gamme des valeurs b_1 pour un voisinage donné est plus petite que la gamme des valeurs b_2, il existe beaucoup plus d'unités de sondage dans la gamme des b_1 que dans celle des b_2 à cause de l'asymétrie de la distribution.

5. CONCLUSION

Les graphiques que nous avons présentés montrent qu'une vaste gamme de bornes donne une gamme étroite de valeurs de la taille de l'échantillon lorsqu'elle se trouve au voisinage d'une valeur optimale (portion inférieure en forme de cuvette des graphiques). Une amélioration additionnelle de la taille de l'échantillon (p. ex., un léger gain marginal) risque de ne pas justifier les efforts supplémentaires nécessaires pour l'obtenir. Il n'est même pas sûr qu'un tel gain marginal contribuera à améliorer la taille de l'échantillon puisque cette dernière est en réalité une valeur entière et que le gain marginal en question pourrait n'être qu'une petite fraction. La méthode L-H s'est avérée très efficace pour le calcul des bornes dans un voisinage souhaité autour d'une valeur optimale, et elle s'est également avérée relativement rapide.

En mesurant le taux de convergence à l'aide de la taille de l'échantillon au lieu des bornes, nous avons été en mesure de mieux déterminer le moment où on atteignait un voisinage souhaitable autour d'une valeur optimale. Les bornes varient en effet grandement dans un tel voisinage tandis que la taille de l'échantillon (le paramètre qui nous intéresse surtout) ne varie que légèrement. Lorsque l'amélioration obtenue dans la taille de l'échantillon d'une itération à l'autre était devenue minimale ou nulle, nous arrêtions immédiatement le programme en présumant que nous avions atteint le voisinage désiré. Les règles d'interruption suivantes sont recommandées. On interrompt le programme lorsque:

1) la différence entre la nouvelle borne supérieure et la borne supérieure de l'itération précédente est inférieure à un. Nous utilisons le nombre entier de un, dans notre cas, puisque les valeurs de la paye n'existent qu'en nombres entiers et qu'un changement des bornes d'une valeur inférieure à un n'influera sur aucune entreprise;
2) la différence entre la nouvelle borne inférieure et celle de l'itération précédente est inférieure à un;
3) la différence entre la nouvelle taille de l'échantillon et celle de l'itération antérieure est inférieure à une petite valeur arbitraire. Nous recommandons un nombre inférieur à un puisque les valeurs de la taille de l'échantillon sont habituellement arrondies et que toute amélioration fractionnaire de la taille de l'échantillon est négligeable. Il convient d'être prudent lors du choix de cette valeur puisqu'il est possible que le taux de réduction de la taille de l'échantillon augmente d'une itération à l'autre par suite d'un changement de la pente de la surface;
4) le programme atteint la 30^{e} itération. Il s'agit évidemment là d'une valeur arbitraire qui peut dépendre du nombre de fois (industries) où il convient d'utiliser la méthode L-H.

Il convient en outre de mentionner que les populations petites peuvent provoquer une convergence des bornes à un point sous-optimal, comme nous l'avons montré dans les exemples. Les graphiques de la taille de l'échantillon montrent une surface rugueuse pour les petites populations et une surface lisse pour les grandes. C'est cette surface rugueuse, due à la nature discrète de la petite population, qui contribue en partie à déterminer le lieu de convergence de la méthode L-H.

Soulignons finalement, à titre de conclusion, que la méthode de Dalenius-Hodges présume que toutes les strates obtenues seront échantillonnées. La méthode L-H est conçue pour établir une sous-strate analytique à tirage complet. Par conséquent, la strate supérieure élaborée par la méthode de Dalenius-Hodges au moment de la création des bornes initiales pour les branches ACES, sera déséquilibrée puisqu'elle ne sera pas échantillonnée. Dans une telle situation, nous avons observé des améliorations dans la taille de l'échantillon en passant de la méthode de Dalenius-Hodges à la première itération de la méthode L-H. Il peut cependant arriver que les bornes de départ conduisent à un minimum local qui ne représente pas la meilleure solution.

REMERCIEMENTS

Les auteurs tiennent à remercier M. Michel Hidiroglou pour ses commentaires utiles. Nous remercions également Carol (Veum) Caldwell, Easley Hoy, les critiques de la revue Techniques d'enquête et les gestionnaires du service de la recherche et de la méthodologie de la Manufacturing and Construction Division pour leurs commentaires utiles formulés au cours de l'examen du présent article.

BIBLIOGRAPHIE

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

DETLEFSEN, R., et VEUM, C. (1991). Design issues for the retail trade sample surveys of the U.S. Bureau of the Census. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 214-219.

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

HESS, I., SETHI, V.K., et BALAKRISHNAN, T.R. (1966). Stratification: A practical investigation. *Journal of the American Statistical Association*, 61, 74-90.

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

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

SCHNEEBERGER, H. (1979). Saddle-points of the variance of the sample mean in stratified sampling. *Sankhȳa*, Series C, 41, 92-96.

SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *American Journal of Statistics*, 5, 20-23.