Techniques d'enquête, juin 1979
Vol. 5, nº 1
Statistique Canada, nº 12-001-X au catalogue

L'inclusion de grandes unités en échantillonnage aléatoire simple[1]

M.A. Hidiroglou[2]

Résumé

Dalenius (1950) et Glasser (1962) ont énoncé des règles approximatives de partage pour la stratification d'une population en un univers à tirage complet et un univers à tirage partiel. Ils ont exprimé la valeur seuil (qui marque la frontière entre les deux types d'univers) en fonction de la moyenne, du poids de l'échantillonnage et de la variance de la population. Leurs valeurs de partage ont été calculées à partir de l'hypothèse d'un échantillon aléatoire unique de taille n tiré sans remise d'une population de taille N.

Ici, l'auteur a élaboré des règles de partage exactes et approximatives pour une situation semblable. Au lieu d'avoir la taille de l'échantillon, on dispose de la précision (coefficient de variation). Il est à noter que dans de nombreux cas d'échantillonnage le chercheur a un ensemble d'objectifs exprimés en fonction de la fiabilité et non de la taille de l'échantillon. Le résultat est particulièrement utile lorsqu'il s'agit de déterminer la limite de partage pour des échantillons tirés d'une population connue. Cette méthode est également utilisée dans le cas de l'estimation par quotient.

Mots-clés : Stratification; tirage complet; tirage partiel; valeur seuil; règles de partage.

1 Introduction

La stratification d'une population est une technique souvent utilisée dans le cadre d'un échantillonnage. Cette technique peut produire des augmentations de précision lors de l'estimation des caractéristiques de la population. Le problème traité dans la présente étude est la stratification d'une population en deux strates : la strate à tirage complet et la strate à tirage partiel. La strate à tirage complet contient certaines des plus grandes unités de la population, tandis que la strate à tirage partiel contient les unités restantes. Les unités de la strate à tirage complet sont entièrement échantillonnées, tandis qu'un échantillon aléatoire simple est tiré de la strate à tirage partiel. Ce type de stratification est particulièrement utile pour les populations dont la répartition présente une asymétrie positive marquée, qui comprend quelques grandes unités et beaucoup de petites unités. Si l'on ne reconnaît pas que ces populations fortement asymétriques doivent être stratifiées de la manière décrite ci-dessus, on risque de surestimer les

1. Présenté à la réunion annuelle de l'American Statistical Association, du 13 au 16 août 1979, Washington, D.C., États-Unis.
2. M.A. Hidiroglou, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada.

caractéristiques de la population. Ce dernier point a été étudié par Hidiroglou et Srinath (1977).

Des règles approximatives de partage pour la stratification d'une population en des univers à tirage complet et à tirage partiel ont été fournies par Dalenius (1950) et Glasser (1962). Glasser (1962) a exprimé la valeur seuil (la valeur qui délimite les sous-univers à tirage complet et à tirage partiel) en fonction de la moyenne, du poids d'échantillonnage et de la variance de la population. Ces valeurs seuil ont été calculées en supposant qu'un échantillon d'une taille prédéfinie n devait être tiré sans remise à partir d'une population de taille N. Dans le contexte actuel, des règles de partage exactes et approximatives ont été élaborées pour une situation similaire. Plutôt que de fournir la taille de l'échantillon, on indique le niveau de précision c (coefficient de variation) souhaité pour les estimations. Il convient de mentionner que dans de nombreuses situations d'échantillonnage, l'échantillonneur reçoit un ensemble d'objectifs relatifs à la fiabilité des estimations.

Singh et Singh (1975) ont adapté le résultat de Glasser pour le cas où l'échantillonnage à probabilité proportionnelle à la taille avec remise était utilisé. Le nombre d'unités n à tirer de la population est prédéfini. Ils délimitent les univers « à tirage complet » et « à tirage partiel » en fournissant une limite inférieure et supérieure pour le point limite. Ces limites sont strictement fonction des mesures de taille utilisées pour calculer les probabilités de sélection. La raison pour laquelle les limites ne comportent aucune valeur concernant la variable à l'étude est que cette dernière est exprimée en fonction des mesures de la taille par un modèle de régression. Dans la présente étude, les résultats s'étendent également au cas de l'estimateur par quotient. Pourvu qu'un modèle linéaire existe entre la variable à l'étude et la variable auxiliaire connue, un seuil peut être établi. Ce seuil sera strictement une fonction de la variable auxiliaire.

2 La procédure d'échantillonnage et la méthode d'estimation

Considérons une population finie ϕ_N composée de N unités représentées par $y_1, y_2, \ldots, y_N$. Définissons des statistiques ordonnées $y_{(N)}, y_{(N-1)}, \ldots, y_{(1)}$ où $y_{(N)} \leq y_{(N-1)} \leq \ldots \leq y_{(1)}$.

Choisissons un échantillon aléatoire simple de taille $n(\ell)$. Il faut mentionner que $n(\ell)$ n'est plus une taille d'échantillon fixe. Il s'agit plutôt d'une variable qui dépend du nombre d'unités à tirage complet ℓ à inclure dans l'échantillon. Supposons que le niveau de précision souhaité pour le total estimé soit obtenu par c. Le total Y peut être écrit comme suit :

$$Y = \sum_{i=1}^{\ell} y_{(i)} + \sum_{i=\ell+1}^{N} y_{(i)}. \tag{2.1}$$

Étant donné que ℓ unités sont à tirage complet et que $n(\ell) - \ell$ unités sont à tirage partiel, un estimateur du total Y serait :

$$\hat{Y} = \sum_{i=1}^{\ell} y_{(i)} + \frac{N-\ell}{n(\ell)-\ell} \sum_{i=1}^{n(\ell)-\ell} z_i \tag{2.2}$$

où $y_{(N)} \leq z_i \leq y_{(\ell+1)}$ pour $i = 1, 2, \ldots, n(\ell) - \ell$.

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

$$V(\hat{Y}) = \frac{(N-\ell)\{N-n(\ell)\}}{n(\ell)-\ell} S^2_{N-\ell} \tag{2.3}$$

où

$$S^2_{N-\ell} = \frac{1}{N-\ell-1} \sum_{i=\ell+1}^{N} (y_{(i)} - \mu_{N-\ell})^2$$

$$\mu_{N-\ell} = \frac{1}{N-\ell} \sum_{i=\ell+1}^{N} y_{(i)}.$$

Sur le plan de la fiabilité, c, $V(\hat{Y})$ peut être réexprimé en tant que $V(\hat{Y}) = c^2Y^2$ si nous substituons $V(\hat{Y}) = c^2Y^2$ dans l'équation (2.3) et que nous trouvons la solution pour $n(\ell)$:

$$n(\ell) = \ell + \frac{(N-\ell)^2 S^2_{N-\ell}}{c^2Y^2 + (N-\ell) S^2_{N-\ell}}. \tag{2.4}$$

3 Le point optimal

L'objectif consiste à trouver la valeur optimale de y qui réduit au minimum la taille de l'échantillon $n(\ell)$ pour le niveau de précision c donné. Une condition nécessaire pour le point optimal est que l'équation (2.4) comprenant $\ell = m$ ne doit pas dépasser l'équation (2.4) comprenant $\ell = m-1$ ou $\ell = m+1$. Cela signifie que la valeur optimale de $y(y^*)$ est trouvée chaque fois que $n(m-1) \geq n(m)$ et que $n(m) \leq n(m+1)$. Cette condition peut être assouplie si nous introduisons un nombre réel b dans les inégalités, à savoir,

$$n(m-1) \geq n(m) + b - 1 \tag{3.1}$$

et

$$n(m) \leq n(m+1) + b - 1,$$

où b peut être utilisé pour contrôler le nombre d'unités à inclure dans la strate à tirage complet.

La règle d'arrêt (3.1) est la règle exacte qui permet de trouver le seuil optimal pour un nombre b donné. Pour exprimer l'équation (3.1) en valeurs voisines du seuil optimal $y_{(m)}$ et $y_{(m+1)}$, nous avons besoin des deux relations suivantes.

$$(N-m)S^2_{\nu+1} = (N-m-1)S^2_{\nu} + \frac{N-m}{N-m+1}\,(y_{(m)} - \mu_{\nu})^2 \tag{3.2}$$

et

$$(N-m-2)S^2_{\nu-1} = (N-m-1)S^2_{\nu} - \frac{N-m}{N-m-1}\,(y_{(m+1)} - \mu_{\nu})^2,$$

où

$$S^2_{\nu+k} = \frac{1}{N-m+k-1} \sum_{i=m-k+1}^{N} (y_{(i)} - \mu_{\nu+k})^2$$

et

$$\mu_{\nu+k} = \frac{1}{N-m+k} \sum_{i=m-k+1}^{N} y_{(i)}$$

pour $k = -1, 0, 1$ et $\nu = N - m$.

En substituant les équations (2.4) et (3.2) dans l'équation (3.1), nous pouvons montrer que :

$$(y_{(m)} - \mu_\nu)^2 \geq \left\{\frac{[bN-n_m-(b-1)m](N-m)}{(n_m-m)(N-n_m-b+1)} + \frac{1}{N-m}\right\} S_\nu^2$$

et que (3.3)

$$(y_{(m+1)} - \mu_\nu)^2 \leq \left\{\frac{[bN-n_m-(b-1)m](N-m-2)}{(n_m-m)(N-n_m+b-1)} + \frac{1}{N-m}\right\} S_\nu^2.$$

Le compromis de l'équation (3.3), si m représente le nombre optimal d'unités à inclure avec certitude, est

$$(y^*-\mu_\nu)^2 = \left\{\frac{b(N-m-1)}{n_m-m} + \frac{1}{2}\left[\frac{(b-1)(N-m)}{N-n_m-b+1} + \frac{(b-1)(N-m-2)}{N-n_m+b-1}\right]\right.$$

$$\left. + \frac{1}{2}\,\frac{b(b-1)}{(n_m-m)}\left[\frac{N-m}{N-n_m-b+1} - \frac{N-m-2}{N-n_m+b-1}\right]\right\}. \quad (3.4)$$

Il convient de mentionner que si m est le nombre optimal d'unités à inclure dans l'échantillon avec certitude, alors $y_{(m+1)} < y^* \leq y_{(m)}$. De plus, l'équation (3.4) est une solution du système d'inégalités obtenue avec l'équation (3.3). Bien que l'équation (3.4) soit une condition nécessaire à l'existence d'une valeur optimale, elle n'est pas nécessairement suffisante. Il peut exister plus d'une solution, auquel cas nous choisirions celle qui réduit au minimum $n(\ell)$ pour un nombre b donné. Comme le souligne Glasser (1962), s'il n'est pas forcément rentable d'inclure avec certitude une unité donnée à elle seule, il peut être profitable de l'inclure avec plusieurs autres unités.

Soulignons que

$$n_m = m + \frac{(N-m)^2 S_\nu^2}{c^2 Y^2 + (N-m) S_\nu^2} \quad (3.5)$$

et que

$$N-n_m = \frac{(N-m)c^2Y^2}{c^2Y^2+(N-m)S_\nu^2} \tag{3.6}$$

la substitution des équations (3.5) et (3.6) dans l'équation (3.4) donne

$$(y^*-\mu_\nu)^2 \doteq \frac{bc^2Y^2}{N-m} + (2b-1)S_\nu^2 + \frac{(N-m)(b-1)S_\nu^4}{c^2Y^2}, \tag{3.7}$$

à condition que $b-1$ soit d'un ordre de grandeur très inférieur à $NcY^2/(cY^2+NS_\nu^2)$.

Une limite supérieure pour y^* peut s'obtenir, en fonction de la variance de la population S_N^2, la taille de la population N et la moyenne μ_N, en utilisant les inégalités suivantes :

$$N(y^*-\mu_N) \leq (N-m)(y^*-\mu_\nu), \tag{3.8}$$

$$(N-m-1)S_\nu^2 \leq (N-1)S_N^2 - \frac{mN(y^*-\mu_N)^2}{N-m} \tag{3.9}$$

où il est vrai que lorsque $m>0, y^* < \mu_m, \mu_m$ est la moyenne des m plus grandes unités de la population. En substituant les inégalités des équations (3.8) et (3.9) dans l'équation (3.7), nous obtenons après quelques simplifications la règle approximative de partage

$$y^* < \mu_N + \left[\frac{bc^2Y^2}{N} + S_N^2\left\{(2b-1) + \frac{N(b-1)S_N^2}{c^2Y^2}\right\}\right]^{1/2}. \tag{3.10}$$

Cette inégalité ne dépend que de la taille de la population, du coefficient de variation c, b, μ_N et de S_N. Cette approximation n'est bonne que lorsque m est relativement petit par rapport à N. Plus les grandes unités sont extrêmes et variables, moins la limite se rapproche de la solution exacte. Bien que la programmation informatique et le temps nécessaire à l'obtention de la limite exacte soient assez minimes, il est néanmoins instructif de caractériser la limite en fonction de valeurs de population connues.

L'approximation de l'équation (3.10) révèle une caractéristique concernant l'effet de b sur le point limite. Si $b_2 > b_1$, alors le point limite associé à b_2 sera plus élevé que celui associé à b_1. Soulignons que l'inverse s'applique également. Le choix de b dépend de l'utilisateur. Dans diverses situations, le nombre d'unités dans la strate à tirage complet peut varier. Par exemple, dans les enquêtes auprès des entreprises, un facteur déterminant possible de la règle de partage pourrait être la portion de la population que les unités à tirage complet représentent en ce qui concerne la variable étudiée. Dans ce cas, l'utilisateur choisirait probablement $b \leq 1$. Un autre facteur pourrait être le fardeau de réponse. L'utilisateur introduirait très probablement un schéma de rotation qui permettrait à certaines des grandes unités d'entrer dans l'échantillon par rotation et d'en sortir. Dans ce cas, moins d'unités seraient incluses dans le tirage complet si on choisissait $b \geq 2$.

4 Application à l'estimateur par quotient

Supposons que la population ϕ_N soit constituée de N vecteurs bidimensionnels représentés par $(y_1, x_1), (y_2, x_2), \ldots, (y_N, x_N)$. Définissons des statistiques ordonnées $x_{(N)}, x_{(N-1)}, \ldots, x_{(1)}$, où $x_{(N)} \leq x_{(N-1)} \leq \ldots \leq x_{(1)}$. Supposons que les valeurs y correspondantes sont $y_{(N)}, y_{(N-1)}, \ldots, y_{(1)}$, où $y_{(i)}$ est la valeur associée à $x_{(i)}$ dans le vecteur bidimensionnel $(y_{(i)}, x_{(i)})$. Si un échantillon de taille $n(\ell)$ est sélectionné à l'aide d'un échantillonnage aléatoire simple sans remise, ℓ étant le nombre d'unités à tirage complet, un estimateur du total Y serait :

$$\hat{Y}_R = \sum_{i=1}^{\ell} y_{(i)} + \frac{\bar{y}_{n(\ell)-\ell}}{\bar{x}_{n(\ell)-\ell}} x_{N-\ell} \tag{4.1}$$

$$\bar{y}_{n(\ell)-\ell} = \frac{1}{n(\ell)-\ell} \sum_{i=1}^{n(\ell)-\ell} z_i$$

$$\bar{x}_{n(\ell)-\ell} = \frac{1}{n(\ell)-\ell} \sum_{i=1}^{n(\ell)-\ell} u_i$$

$$x_{N-\ell} = \sum_{i=\ell+1}^{N} x_{(i)},$$

$x_{(N)} \leq u_i \leq x_{(\ell+1)}$ et $y_{(N)} \leq z_i \leq y_{(\ell+1)}$ pour $i = 1, 2, \ldots, n(\ell) - \ell$. La variance du total obtenu par l'équation (4.1) est :

$$V(\hat{Y}_R) = \frac{(N-\ell)\{N-n(\ell)\}}{n(\ell)-\ell} \quad \frac{1}{N-\ell-1} \quad \sum_{i=\ell+1}^{N} (y_{(i)} - R_{N-\ell}\, x_{(i)})^2$$

où $R_{N-\ell} = Y_{N-\ell} / X_{N-\ell}$, $Y_{N-\ell} = \sum_{i=\ell+1}^{N} y(i)$, est le rapport entre le total de la population y et le total de la population x. Il faut mentionner que la valeur optimale de $y(y^*)$ sera trouvée chaque fois pour r, $n(r-1) \geq n(r)$ et $n(r) \geq n(r+1)$, où

$$n(\ell) = \ell + \frac{(N-\ell)^2 S^2_{R.N-\ell}}{c^2 Y^2 + (N-\ell) S^2_{R.N-\ell}}, \tag{4.2}$$

$$S^2_{R.N-\ell} = \frac{1}{N-\ell-1} \quad \sum_{i=\ell+1}^{N} (y_{(i)} - R_{N-\ell}\, x_{(i)})^2.$$

Il faut souligner que l'expression (4.2) fait intervenir la variable à l'étude y. Considérons la population finie ϕ_N comme un échantillon d'une superpopulation pour laquelle le modèle suivant s'applique :

$$y_i = \beta x_i + e_i$$

où

$$E(e_i | x_i) = 0,$$

$$E(e_i, e_j | x_i, x_j) = 0,$$

$$E(e_i^2 | x_i) = \alpha x_i \tag{4.3}$$

pour $\alpha \geq 0$ et $i = 1, 2, \ldots, N$. Selon ce modèle, $n(\ell)$ peut s'exprimer comme une fonction des x, c'est-à-dire

$$n(\ell) = \ell + \frac{(N-\ell)^2 \alpha\{X_{N-\ell}^2 - \sum_{i=\ell+1}^{N} x_{(i)}^2\}/(N-\ell-1)}{c^2 g X_N + (N-\ell)\alpha\{X_{N-\ell}^2 - \sum_{i=\ell+1}^{N} x_{(i)}^2\}/(N-\ell-1)} \qquad (4.4)$$

où $g = \alpha X + \beta^2 X^2$. La relation ci-dessus peut être facilement obtenue en soulignant que selon le modèle (4.3),

$$EY^2 = \alpha X + \beta^2 X^2$$

et

$$E\, s_{R\cdot N-\ell}^2 = \frac{\alpha}{N-\ell-1}\left(X_{N-\ell} - \frac{1}{X_{N-\ell}} \sum_{i=\ell+1}^{N} x_{(i)}^2\right).$$

Lorsque la valeur optimale y a été trouvée, il est possible de démontrer que $n(r-1) \geq n(r)$ est équivalent à

$$\{(N-r+1)\,\alpha\, c^2\, g\, [X_\rho^2 + (N-r)\, s_{X\cdot\rho}^2] - c^4 g^2 X_\rho\}\, x_{(r)}$$

$$\geq c^4 g^2 X_\rho^2 + \frac{N-r+1}{N-r}\,\alpha\, c^2\, g\, X_\rho^3 - (N-r+1)\,\alpha\, c^2\, g\, s_{X\cdot\rho}^2\, X_\rho \qquad (4.5)$$

où

$$X_{\rho+k} = \sum_{i=r-k+1}^{N} x_{(i)},$$

$$s_{X\cdot\rho+k}^2 = \frac{1}{N-r-1}\left[\sum_{i=r-k+1}^{N} x_{(i)}^2 - \frac{X_{\rho+k}^2}{N-r+k}\right],$$

selon $\rho = N - r$ et $k = -1, 0, 1$.

Il faut mentionner que l'utilisation des deux relations suivantes est nécessaire pour dériver (4.5) :

$$X_{\rho+1} = X_\rho + x_{(r)}$$

et

$$X_{\rho+1}^2 - \sum_{i=r}^{N} x_{(i)}^2 = X_\rho^2 - \sum_{i=r+1}^{N} x_{(i)}^2 + 2\, x_{(r)}\, X_\rho.$$

De même, en utilisant

$$X_{\rho-1} = X_\rho - x_{(r)}$$

et

$$X_{\rho-1}^2 - \sum_{i=r+2}^{N} x_{(i)}^2 = X_\rho^2 - \sum_{i=r+1}^{N} x_{(i)}^2 + 2x_{(r+1)}^2 - 2x_{(r+1)} X_\rho,$$

il est possible de démontrer que $n(r) \leq n(r+1)$ est équivalent à

$$\{- c^4 g^2 X_\rho - \frac{N-r-1}{N-r-2} (N-r-2)\, \alpha\, c^2 g X_\rho^2 - (N-r)(N-r-1)\, \alpha\, c^2 g S_{X\cdot\rho}^2$$

$$+ 2\, \frac{(N-r)(N-r-1)}{N-r-2}\, \alpha\, c^2 g X_\rho\, x_{(r+1)}\}x_{(r+1)}$$

$$\leq - c^4 g^2 X_\rho^2 - \alpha\, c^2 g X_\rho^3\, \frac{N-r-1}{N-r-2} + \frac{(N-r)(N-r-1)}{N-r-2}\, \alpha\, c^2 g X_\rho S_{X\cdot\rho}^2 . \qquad (4.6)$$

Si r représente le nombre optimal de grandes unités à inclure avec certitude, alors $x_{(r)} \geq x^* \geq x_{(r+1)}$. Cette condition est remplie chaque fois que l'équation quadratique suivante s'avère exacte.

$$\{(N-r)\, \alpha\, c^2 g S_{X\cdot\rho}^2 - \frac{N-r}{N-r-2}\, \alpha\, c^2\, g X_\rho^2 - c^4 g^2 X_\rho^2$$

$$+ \frac{(N-r)(N-r-1)}{N-r-2}\, \alpha\, c^2 g X_\rho\, x^*\}x^*$$

$$= - \frac{\alpha\, c^2 g X_\rho^3}{(N-r)(N-r-2)} + \frac{\alpha\, c^2 g X_\rho S_{X\cdot\rho}^2}{(N-r-2)}. \qquad (4.7)$$

Il faut souligner que l'expression ci-dessus dépend de la connaissance de α et β. Ainsi, il est possible de tirer grand profit de toute estimation des valeurs α et β obtenues à partir d'enquêtes antérieures ou d'autres sources.

5 Quelques illustrations utiles

L'utilisation de la procédure décrite dans la section 2 suppose que la population de laquelle l'échantillon doit être tiré est représentative de la population cible. Voici un exemple où une telle procédure peut être utilisée : toutes les valeurs associées aux unités d'un univers d'entreprise sont connues au moment t_1. Un échantillon est tiré de cet univers au moment $t_1 + k_1$, $k_1 \geq 0$, et sera utilisé comme base pour inférer les caractéristiques de l'univers de $t_1 + k_1$ à $t_1 + k_2$ où $k_2 > k_1$. Dans ce cas, l'univers au moment t_1, $\phi(t_1)$, peut être différent de l'univers au moment t_2, $\phi(t_2)$, $t_2 > t_1$. Toutefois, s'il est possible de supposer que la valeur seuil calculée au temps t_1 n'est pas trop différente de celle qui serait calculée au temps t_2, alors la subdivision de la population $\phi(t_1)$ produira toujours des gains.

Les données utilisées pour illustrer les résultats présentés à la section 3 proviennent de l'Enquête annuelle sur l'industrie des aliments et boissons de 1976. Cette enquête est essentiellement un recensement de tous les établissements servant de la nourriture et des boissons couverts par la Division du commerce et des services de Statistique Canada. Les établissements visés par cette enquête comprennent toutes les entreprises connues dont les établissements sont classés selon le code 886 (1970) de la Classification type des industries. Le code de la Classification type des industries se divise davantage en sept types d'entreprises qui comprennent les restaurants titulaires d'un permis d'alcool, les bars et les boîtes de nuit. Les données de cette enquête sont actuellement publiées selon un classement infraprovincial recoupé par type d'entreprise. L'exemple présente une situation où l'univers de l'entreprise est connu au moment t_1 (Enquête sur l'industrie des aliments et boissons de 1976) et où un échantillon doit être tiré au moment $t_1 + k_1$ (Enquête mensuelle projetée des restaurants, traiteurs et tavernes).

La règle de partage qui est illustrée est celle obtenue avec l'équation (3.1) lorsque b est égal à 1 et 2, respectivement. Quatre strates infraprovinciales recoupées par type d'entreprise ont été choisies pour fournir les exemples. Les entreprises comprennent, de façon respective : les bars et les boîtes de nuit à Terre-Neuve (strate 1); les bars et les boîtes de nuit dans les régions non métropolitaines du Nouveau-Brunswick (strate 2); les restaurants titulaires d'un permis d'alcool à

Halifax-Dartmouth (strate 3); et les bars et les boîtes de nuit dans les régions non métropolitaines du Québec (strate 4). Certaines des caractéristiques statistiques de ces strates sont présentées dans le tableau 5.1. Il s'agit des ventes minimales, maximales et moyennes pour chacune des strates. L'écart-type, S_N, est également fourni avec la taille de la population associée. Il faut souligner que ces statistiques supposent que les distributions de fréquences associées sont positivement asymétriques. Les figures 5.1 à 5.4 confirment cette observation.

Tableau 5.1
Caractéristiques statistiques de la strate d'intérêt

Strate	Minimum	Maximum	Moyenne	Écart-type	N
1	3 000	476 141	139 380	67 800	170
2	4 000	463 000	181 930	90 160	61
3	15 045	1 223 360	350 250	263 830	63
4	3 345	885 333	132 770	72 520	632

Pour chacune des strates en question et compte tenu du coefficient de variation souhaité, nous fournissons le nombre d'unités à inclure dans la sous-strate à tirage complet, le seuil exact et approximatif, ainsi que l'échantillon qui aurait été sélectionné si aucune sous-strate à tirage complet n'avait été formée. Ce renseignement se trouve dans le tableau 5.2. Il faut souligner que le seuil approximatif s'obtient par l'inégalité (3.10) et le seuil exact, par l'équation (3.3) à l'aide de $b = 1$.

Tableau 5.2
Renseignements concernant la procédure de tirage complet donnée par les inégalités (3.1) pour b = 1

Strate	c	Seuil exact	Seuil approximatif	m	n(m)	n(0)
1	0,1116	353 351	353 230	3	15	17
2	0,1063	339 071	357 840	4	13	16
3	0,1359	806 999	811 060	6	10	21
4	0,1209	598 192	542 450	3	12	20

Dans le tableau ci-dessus, m est égal au nombre d'unités à inclure dans l'échantillon avec certitude et n(m) est la taille de l'échantillon

global correspondant nécessaire pour atteindre la fiabilité souhaitée. Il faut souligner que n(m) < n(0) pour toutes les strates considérées, où n(0) est la taille de l'échantillon sans unités à tirage complet. Par conséquent, si des unités à tirage complet doivent être trouvées, la taille de l'échantillon global sera plus petite que celle de l'échantillon sans unités à tirage complet. Il faut mentionner que le seuil approximatif obtenu avec l'équation (3.10) est assez proche du seuil exact obtenu avec l'équation (3.3). Les résultats pour b = 2 fournis dans le tableau 5.3 mettent en évidence l'effet de b sur les points limites.

Tableau 5.3
Renseignements concernant la procédure de tirage complet obtenue avec les inégalités (3.1) pour b = 2

Strate	c	Seuil exact	Seuil approximatif	p	n(p)	n(0)
1	0,1116	476 141	450 133	1	16	17
2	0,1063	462 303	451 950	1	15	16
3	0,1359	1 223 360	1 077 048	1	19	21
4	0,1209	832 991	717 265	2	15	20

Il est nécessaire de souligner que p représente le nombre d'unités dans la strate à tirage complet et que n(p) est la taille de l'échantillon global correspondant. Encore une fois, le seuil approximatif obtenu par l'équation (3.10) est assez proche du seuil exact obtenu par l'équation (3.3) lorsque b = 2. Le seuil exact lorsque b = 2 tend à produire moins d'unités à tirage complet que le seuil exact lorsque b = 1. Nous arrivons à la même conclusion si nous utilisons le seuil approximatif.

Figure 5.1. Distribution de fréquences de 170 unités dans la strate 1

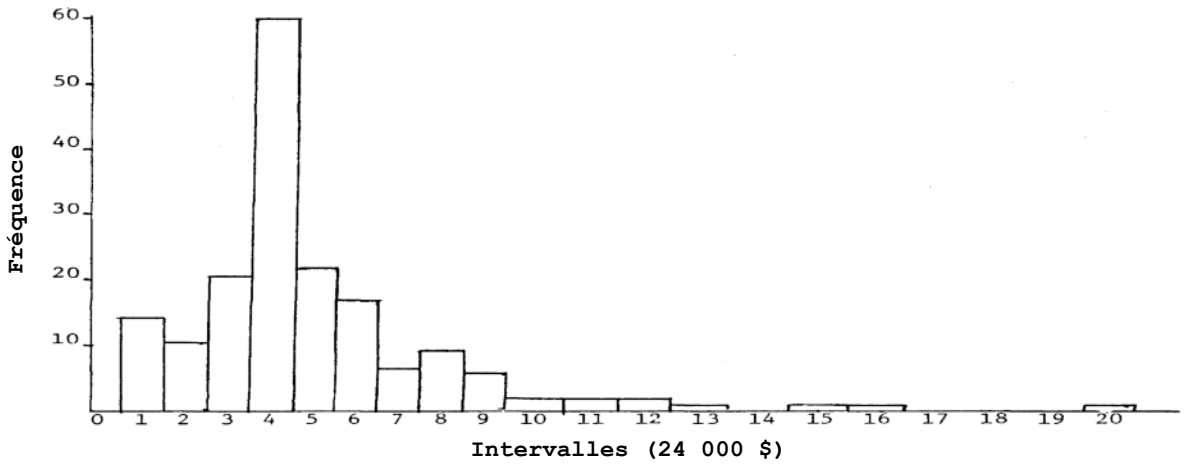


Figure 5.2. Distribution de fréquences de 61 unités dans la strate 2

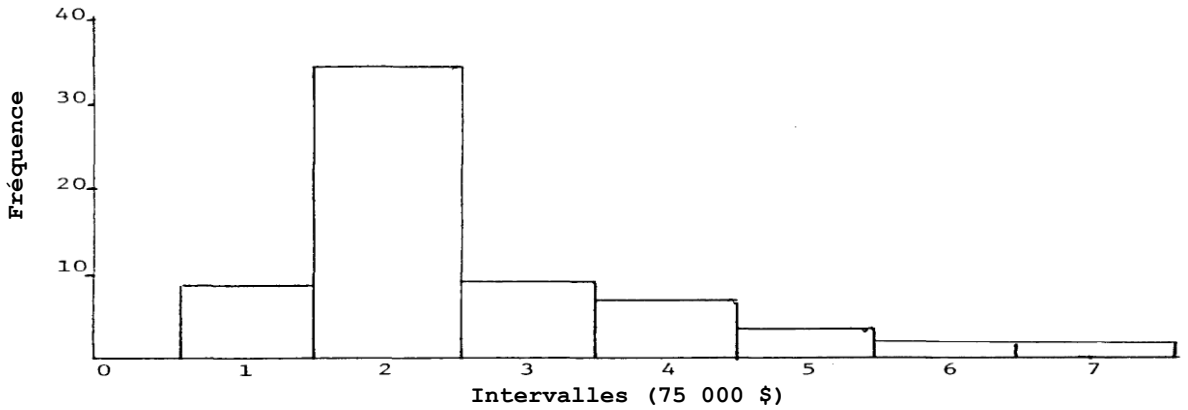


Figure 5.3 Distribution de fréquences de 63 unités dans la strate 3

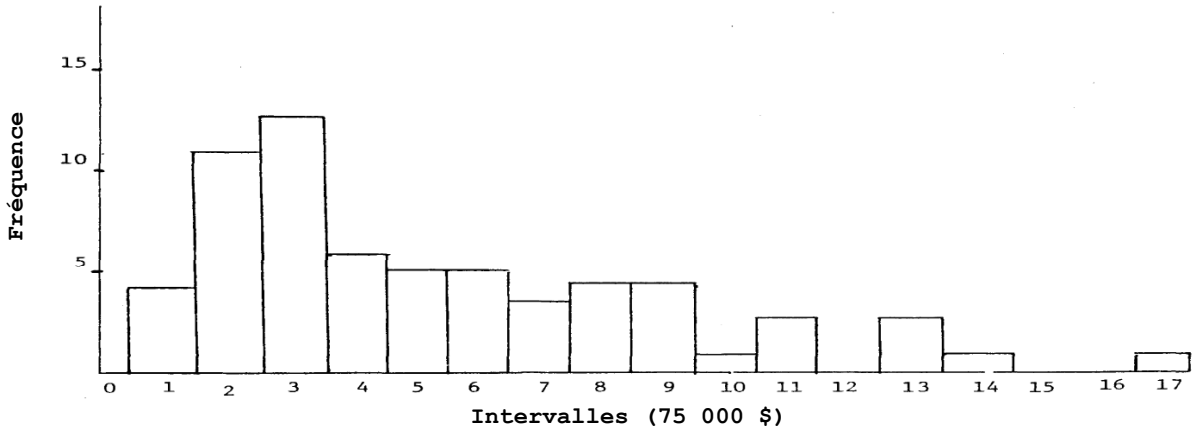
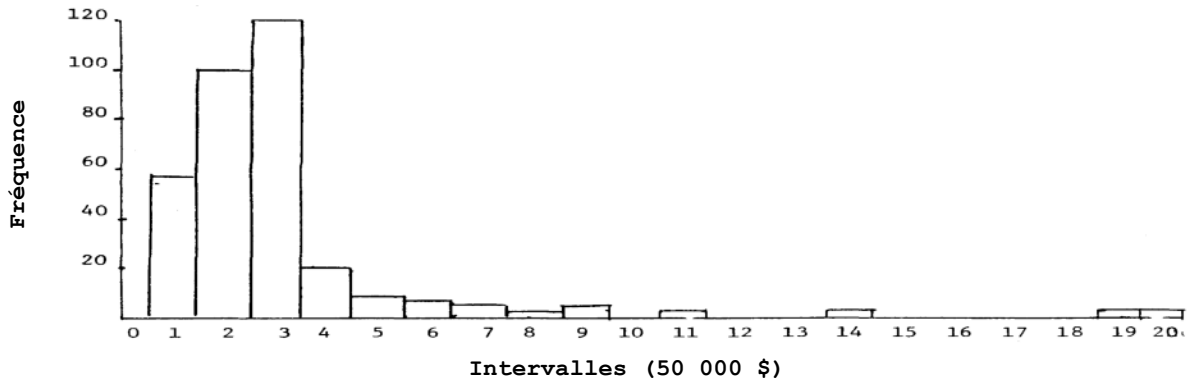


Figure 5.4 Distribution de fréquences de 632 unités dans la strate 4



6 Conclusion

Il est souhaitable de stratifier les populations fortement asymétriques sur la base de la taille des unités. L'approche suggérée dans la présente étude consiste à placer un certain nombre de grandes unités dans une strate à tirage complet et à les échantillonner avec certitude. Les unités restantes, celles rattachées à la strate à tirage partiel, sont échantillonnées à un taux approprié. Le nombre d'unités à inclure avec certitude dépend du niveau de précision souhaité c, du scalaire b, de la moyenne de la population μ_N, de la variance de la population S_N^2 et du nombre d'unités N lorsque les critères de l'équation (3.1) sont utilisés. Il faut mentionner que l'échantillonneur peut faire varier le nombre d'unités dans la strate à tirage complet en faisant varier b. La règle d'arrêt approximative (3.10) peut être utilisée pour estimer initialement le seuil exact correspondant obtenu par l'équation (3.1), à condition que les renseignements nécessaires sur la population d'intérêt soient disponibles.

Il y a plusieurs avantages à stratifier une population fortement asymétrique pour la méthode donnée. Pour obtenir un niveau de fiabilité fixe, la taille globale de l'échantillon associée à cette procédure sera invariablement inférieure à la taille de l'échantillon associée sans stratification. Cochran (1963, p. 38-39) souligne que pour les distributions de fréquences qui ne sont pas raisonnablement proches de la normalité, il est risqué d'utiliser l'approximation normale comme base pour établir les intervalles de confiance. En séparant certaines des observations les plus importantes des distributions fortement asymétriques, les intervalles de confiance sont essentiellement basés sur des populations qui sont moins asymétriques. Ce dernier point devrait inciter l'échantillonneur à avoir plus confiance en l'utilisation de l'approximation normale. Enfin, ce type de stratification permet d'éviter la surestimation des caractéristiques de la population lorsque des distributions très asymétriques sont échantillonnées.

Glasser (1962) a souligné que la définition que nous devons donner aux grandes unités dépend de la méthode d'échantillonnage du reste de la population et de la méthode d'estimation. Dans le cas de l'estimation par le quotient, l'équation quadratique (4.7) doit être résolue séquentiellement pour permettre d'obtenir le seuil x^*.

Bibliographie

[1] Cochran, W.G. (1963), *Sampling Techniques*, Wi1ey, 2nd ed., New York.

[2] Dalenius, T. (1950), "The Problem of Optimum Stratification in a Special Type of Design", *Skandinavisk Aktuariediskrift*, p. 61-70.

[3] Glasser, G.J. (1962), "On the Complete Coverage of Large Units in a Statistical Study", *Review of the International Statistical Institute*, Vol. 30, p. 28-32.

[4] Hidiroglou, M.A. et Srinath, K.P. (1977), "Some Estimators of Population Totals from Simple Random Samples Containing Large Units". *Proceedings of the American Statistical Society*, Chicago, Social Statistics Section, p. 903-908.

[5] Singh, D. et Singh, R. (1975), "On the Inclusion of Large Units in the Sample", *Journal of the Indian Society of Agricultural Statistics*, p. 24-31.