Composante du produit nº 12-001-X
Division des méthodes d’enquêtes auprès des entreprises

Article

Détermination des bornes optimales de strate au moyen de la programmation dynamique

par Mohammad G. Mostafa Khan, Niraj Nand et Nesar Ahmad

Décembre 2008

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 227-236
Statistique Canada, N° 12-001-X au catalogue

Détermination des bornes optimales de strate au moyen de la programmation dynamique

Mohammad G. Mostafa Khan, Niraj Nand et Nesar Ahmad [1]

Résumé

La stratification optimale est la méthode qui consiste à choisir les meilleures bornes qui rendent les strates intérieurement homogènes, étant donné la répartition de l'échantillon. Afin de rendre les strates intérieurement homogènes, celles-ci doivent être construites de façon que les variances de strate de la caractéristique étudiée soient aussi faibles que possible. Un moyen efficace d'y arriver, si l'on connaît la distribution de la principale variable étudiée, consiste à créer des strates en découpant l'étendue de la distribution à des points appropriés. Si la distribution des fréquences de la variable étudiée est inconnue, on peut l'approximer en se fondant sur l'expérience passée ou sur certains renseignements a priori obtenus au cours d'une étude récente. Dans le présent article, le problème de la détermination des bornes optimales de strate (BOS) est considéré comme étant le problème de la détermination des largeurs optimales de strate (LOS). Il est formulé comme un problème de programmation mathématique (PPM) consistant à minimiser la variance du paramètre de population estimé sous la répartition de Neyman en imposant que la somme des largeurs des strates soit égale à l'étendue totale de la distribution. La variable étudiée est considérée comme suivant une loi continue dont la densité de probabilité est triangulaire ou normale standard. Les PPM formulés, qui s'avèrent être des problèmes de décision à plusieurs degrés, peuvent alors être résolus en utilisant la méthode de programmation dynamique proposée par Bühler et Deutler (1975). Des exemples numériques sont présentés pour illustrer les calculs. Les résultats obtenus sont également comparés à ceux donnés par la méthode de Dalenius et Hodges (1959) dans le cas d'une distribution normale.

Mots clés : Échantillonnage aléatoire stratifié ; stratification optimale ; loi triangulaire ; loi normale standard ; problème de programmation mathématique ; problème de décision à plusieurs degrés ; méthode de programmation dynamique.

1. Introduction

L'aspect fondamental à prendre en considération dans la détermination des bornes optimales de strate (BOS) est que l'homogénéité interne des strates doit être aussi grande que possible, c'est-à-dire que les variances de strate σ_h^2 doivent être aussi faibles que possible, étant donné la répartition de l'échantillon. Quand on étudie une seule caractéristique et que l'on connaît la distribution de la variable étudiée, les BOS peuvent être déterminées en découpant l'étendue de cette distribution en des points appropriés. Ce problème de détermination des BOS a été discuté pour la première fois par Dalenius (1950), dans le cas où la variable étudiée proprement dite est utilisée comme variable de stratification. Il a présenté un ensemble d'équations minimales pouvant être résolues pour trouver les BOS. Malheureusement, en général, ces équations ne pouvaient pas être résolues, à cause de leur nature implicite. Par conséquent, plusieurs auteurs ont tenté d'obtenir les bornes approximatives de strate en utilisant des méthodes classiques. Le nombre de strates étant donné, Dalenius et Gurney (1951) ont proposé de déterminer les bornes des strates quand $W_h\sigma_h$ demeure constant, où W_h est le poids de la strate h. Mahalanobis (1952), ainsi que Hansen et Hurwitz (1953) ont proposé que l'on détermine les bornes des strates quand $W_h\mu_h$ demeure constant. Aoyama (1954) a suggéré d'appliquer une règle approximative et recommandé de produire des strates de même largeur $x_h - x_{h-1}$, où x_{h-1} et x_h sont les bornes de la strate h. Ekman (1959) a déterminé les bornes des strates sous la contrainte que $W_h(x_h - x_{h-1})$ soit constant. Dalenius et Hodges (1959) ont recommandé de construire les strates en prenant des intervalles égaux sur la cumulative de $\sqrt{f(x)}$. Sethi (1963) a proposé une méthode de détermination des bornes par calcul infinitésimal au moyen des équations

$$\frac{(x_h - \mu_h)^2 + \sigma_h^2}{\sigma_h} = \frac{(x_{h+1} - \mu_{h+1})^2 + \sigma_{h+1}^2}{\sigma_{h+1}}$$

pour une distribution continue standard ressemblant à la population étudiée.

La comparaison de certaines méthodes classiques d'approximation révèle que la méthode d'Ekman et celle de Dalenius et Hodges donnent systématiquement de bons résultats (voir Cochran 1961, Hess, Sethi et Balakrishnan 1966, Murthy 1967), mais que la seconde est plus commode et plus facile à appliquer (voir Nicoloni 2001).

Unnithan (1978) a proposé une méthode itérative s'appuyant sur la méthode de Newton modifiée de Shanno pour déterminer les bornes de strate qui donnent lieu à un

1. Mohammad G. Mostafa Khan, Department of Statistics, Iowa State University, Ames, Iowa 50011, É.-U., (M.G.M. Khan était en congé de la School of Computing, Information and Mathematical Sciences, University of the South Pacific, Suva, Fiji) ; Niraj Nand et Nesar Ahmad, School of Computing, Information and Mathematical Sciences, University of the South Pacific, Suva, Fiji.

minimum local de la variance sous la répartition de Neyman, si une solution initiale appropriée est choisie. Il est prouvé que cette méthode est plus rapide que la méthode itérative de Dalenius et Hodges. Plus tard, Unnithan et Nair (1995) ont donné une méthode pour sélectionner un point de départ approprié pour la méthode de Newton modifiée qui permet d'aboutir à un minimum global de la variance.

Lavallée et Hidiroglou (1988) ont proposé un algorithme pour construire les bornes de strate pour un échantillon stratifié et la répartition des unités échantillonnées sans certitude selon une méthode de puissance. Hidiroglou et Srinath (1993) ont présenté une forme plus générale de l'algorithme, qui, lorsqu'on affecte différentes valeurs aux paramètres opérationnels, produit une répartition de puissance, une répartition de Neyman ou une combinaison de ces deux répartitions. Sweet et Sigman (1995), ainsi que Rivest (2002) ont revu l'algorithme de Lavallée et Hidiroglou et proposé leurs versions modifiées dans lesquelles sont intégrées les différentes relations entre la stratification et les variables étudiées. Detlefsen et Veum (1991) ont étudié l'algorithme de Lavallée et Hidiroglou pour plusieurs strates et observé que la convergence de l'algorithme était lente ou inexistante. Ils ont également découvert que différents points de départ aboutissaient à des BOS différentes pour la même population.

Niemiro (1999) a proposé une méthode de recherche aléatoire dans le problème de stratification, mais l'algorithme ne garantissait pas d'aboutir à un optimum global. De surcroît, si la population était grande, il fonctionnait mal, car il nécessitait un trop grand nombre d'étapes d'itération (voir Kozak 2004).

Nicolini (2001) a proposé une méthode, nommée *méthode des classes naturelles* (MCN), en opposition à celle, plus utilisée, de Dalenius et Hodges, mais ni l'une ni l'autre de ces méthodes ne s'est avérée plus efficace que l'autre.

Lednicki et Wieczorkowski (2003) ont présenté une méthode de stratification fondée sur la méthode du simplexe de Nelder et Mead (1965). Plus tard, Kozak (2004) a présenté l'algorithme de recherche aléatoire modifié en tant que méthode de stratification optimale. L'algorithme de Kozak était plus rapide et efficace que celui de Rivest et celui de Lednicki et Wieczorkowski, mais il n'était pas possible de garantir qu'il mène à un optimum global.

Bühler et Deutler (1975) ont formulé le problème de la détermination des BOS comme un problème d'optimisation pouvant être résolu par une méthode de programmation dynamique. Cette approche est également utilisée par Lavallée (1987, 1988) pour trouver les BOS qui diviseraient le domaine de population de deux variables de stratification en sous-ensembles distincts, de telle façon que la précision des variables d'intérêt soit maximisée.

Khan, Khan et Ahsan (2002) ont considéré le problème de la découverte des BOS comme étant équivalant au problème de détermination des largeurs optimales de strate (LOS). Ils ont formulé le problème des LOS comme un problème de programmation mathématique (PPM). En suivant l'approche de programmation dynamique de Bühler et Deutler, ils ont résolu le PPM qui donne la solution exacte si la distribution de fréquence de la variable étudiée est connue et que le nombre de strates est établi d'avance. Khan et coll. (2002) ont appliqué leur méthode pour déterminer les BOS pour une population suivant une loi uniforme et une loi triangulaire droite (*right triangular distribution*). Plus tard, Khan, Najmussehar et Ahsan (2005) ont étendu cette approche de programmation dynamique à la détermination des BOS pour une variable étudiée exponentielle.

Dans le présent article, nous discutons du problème de détermination des BOS pour des variables étudiées suivant une loi triangulaire ou normale standard. Vu que ces problèmes sont équivalents à des problèmes de détermination des LOS, nous les formulons comme des problèmes de programmation mathématique et les résolvons en suivant l'approche de programmation dynamique de Bühler et Deutler. Les PPM formulés minimisent la variance du paramètre de population estimé dans le cas d'une répartition de Neyman sous la contrainte que la somme totale des largeurs des strates est égale à l'étendue totale de la distribution de la variable étudiée. À la section 2, nous passons en revue l'approche de programmation dynamique proposée par Bühler et Deutler (1975). À la section 3, nous donnons les détails de la formulation des problèmes de détermination des LOS comme des problèmes de programmation mathématique. À la section 4, nous discutons de l'algorithme fondé sur la technique de programmation dynamique utilisée pour résoudre les PPM. À la section 5, nous illustrons les calculs de l'algorithme au moyen d'exemples numériques. Enfin, à la section 6, nous comparons les résultats obtenus par la méthode de programmation dynamique et par la méthode de la cumulative de $\sqrt{f}$ de Dalenius et Hodges (1959) en nous basant sur un exemple tiré d'une population suivant une loi normale. Cette comparaison montre que la méthode de programmation dynamique proposée est plus efficace que la méthode de la cumulative de $\sqrt{f}$.

2. Détermination des BOS par des méthodes de programmation dynamique : une revue de l'approche de Bühler et Deutler

Soit X une variable étudiée aléatoire, discrète ou continue, dont la densité de probabilité est $f(x)$, $a \le x \le b$. Pour estimer la moyenne de population μ à l'aide d'un échantillon stratifié, X est partitionnée en L strates $[a,\ x_1],\ (x_1,\ x_2],\ ...,\ (x_{L-1},\ b]$ telles que

$$a = x_0 \leq x_1 \leq x_2 \leq, ..., \leq x_{L-1} \leq x_L = b. \tag{1}$$

Supposons que, pour la strate $h\ (h = 1, 2, ..., L)$, qui contient N_h unités, on sélectionne un échantillon de taille n_h contenant les unités $y_{hj}\ (h = 1, 2, ..., L;\ j = 1, 2, ..., n_h)$. Alors, la moyenne stratifiée $\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$ est une estimation sans biais de μ dont la variance est

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

où $W_h = N_h/N,\ \bar{x}_h = 1/n_h \sum_{j=1}^{n_h} y_{hj},\ \sigma_h^2 = [1/(N_h - 1)] \times \sum_{j=1}^{N_h} (y_{hj} - \mu_h)^2$ et $\mu_h = 1/N_h \sum_{j=1}^{N_h} y_{hj}$.

Quand la fonction de fréquence $f(x)$ est connue, les valeurs de W_h et σ_h dans (2) peuvent être obtenues par

$$W_h = \int_{x_{h-1}}^{x_h} f(x)\, dx, \tag{3}$$

$$\sigma_h^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x^2 f(x)\, dx - \mu_h^2, \tag{4}$$

où

$$\mu_h = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x f(x) dx \tag{5}$$

est la moyenne et $(x_{h-1},\ x_h)$ sont les bornes de la h^{e} strate.

Alors, (2) se lit comme étant la fonction des bornes de strate et des tailles d'échantillon, autrement dit

$$V(\bar{x}_{st}) = V(\bar{x}_{st} \,|\, x_1,\ ...,\ x_{L-1},\ n_1,\ ...,\ n_L).$$

Si les n_h sont fixes, l'objectif de la stratification optimale est de déterminer les bornes de strate $(x_1,\ ...,\ x_{L-1})$ telles que $V(\bar{x}_{st})$ est minimale. En outre, si les fractions d'échantillonnage n_h/N_h sont faibles ou que l'échantillonnage est fait avec remise, nous obtenons les problèmes d'optimisation qui suivent, selon le type de répartition de la taille d'échantillon totale $(n = \sum_{h=1}^{L} n_h)$ entre les strates.

1. Répartition proportionnelle $(n_h = n \cdot W_h)$

   Minimiser $\sum_{h=1}^{L} W_h \sigma_h^2$

   sous la contrainte

   $$a = x_0 \leq x_1 \leq x_2 \leq, ..., \leq x_{L-1} \leq x_L = b \tag{6}$$

2. Répartition égale $(n_h = n/L)$

   Minimiser $\sum_{h=1}^{L} W_h^2 \sigma_h^2$

   sous la contrainte

   $$a = x_0 \leq x_1 \leq x_2 \leq,\ ...,\ \leq x_{L-1} \leq x_L = b \tag{7}$$

3. Répartition de Neyman $(n_h = n \cdot W_h \sigma_h / \sum_{h=1}^{L} W_h \sigma_h)$

   Minimiser $\sum_{h=1}^{L} W_h \sigma_h$

   sous la contrainte

   $$a = x_0 \leq x_1 \leq x_2 \leq, ..., \leq x_{L-1} \leq x_L = b. \tag{8}$$

Les problèmes (6) à (8) ont la structure suivante :

Minimiser $\sum_{h=1}^{L} \phi_h(x_{h-1},\ x_h)$,

sous la contrainte

$$a = x_0 \leq x_1 \leq x_2 \leq, ..., \leq x_{L-1} \leq x_L = b. \tag{9}$$

Bühler et Deutler (1975) ont proposé une méthode d'optimisation récurrente pour résoudre (9) en utilisant une méthode de programmation dynamique comme il suit.

Considérons un problème d'optimisation ayant la structure spéciale :

Minimiser $\sum_{h=1}^{m} u_h(z_{h-1}, y_h)$,

sous les contraintes

$$\begin{aligned} & z_h = v_h(z_{h-1}, y_h), \\ & z_h \in Z_h, \\ & y_h \in S_h(z_{h-1}), \\ & z_0 = z'\ ;\ h = 1, 2, ..., m, \end{aligned} \tag{10}$$

où $m =$ nombre de degrés, $u_h =$ fonctions de rendement de degré, $v_h =$ fonctions de transformation de degré, $Z_h =$ espaces d'états, $S_h =$ espaces de décisions et $z' =$ état initial. Alors, nous pouvons utiliser une méthode de programmation dynamique en appliquant le principe d'optimalité de Bellman (Bellman 1957) pour résoudre (10).

Si $m = L,\ z_0 = a,\ z_L = b,\ Z_h = [a,\ b],\ Z_{h-1} = [a, b - y_h]$, $S_h(z_{h-1}) = [0,\ b - z_{h-1}]$ avec $z_{h-1} \in Z_{h-1},\ u_h(z_{h-1},\ y_h) = \phi_h(z_{h-1},\ y_h + z_{h-1})$ avec $y_h \in S_h(z_{h-1}),\ v_h(z_{h-1},\ y_h) = y_h + z_{h-1}$, alors (10) se transforme en le problème suivant :

Minimiser $\sum_{h=1}^{L} \phi_h(z_{h-1}, y_h + z_{h-1})$,

sous les contraintes

$$\begin{aligned} & z_h = y_h + z_{h-1}, \\ & z_h \in [a,\ b], \\ & y_h \in [0,\ b - z_{h-1}], \\ & z_0 = a,\ z_L = b\ ;\ h = 1,\ 2,\ ...,\ L. \end{aligned} \tag{11}$$

Le problème (11) et le problème (9) sont équivalents, car ils comportent les résultats suivants :

1. Si $(x_1^*, \ldots, x_{L-1}^*)$ est une solution optimale de (9), alors $y_h^* = x_h^* - x_{h-1}^*$, $z_h^* = x_h^*$ est un optimum de (11).
2. Si $y_h^*(h=1, \ldots, L)$, $z_h^*(h=1, \ldots, L-1)$ est une solution optimale de (11), alors $x_h^* = z_h^*(h=1, \ldots, L-1)$ est une solution optimale de (9).

Si $\Phi_h(z_{h-1})$ est la valeur optimale de la fonction objectif au degré h avec l'état disponible z_{h-1}, alors l'équation récurrente rétrograde pour résoudre (11) en utilisant une méthode de programmation dynamique est donnée par

$$\Phi_h(z_{h-1}) = \min[\phi_h(z_{h-1}, y_h + z_{h-1}) + \Phi_{h+1}(z_h) \mid z_h = y_h + z_{h-1}] \qquad (12)$$

sur $y_h \in S_h(z_{h-1})$ avec au départ $\Phi_{L+1} \equiv 0$.

3. Formulation du problème de détermination des LOS comme un PPM

À la présente section, nous étendons l'approche de Bühler et Deutler discutée plus haut au cas d'une variable étudiée dont la densité de probabilité $f(x)$ est continue. Le problème (11) est transformé en un problème équivalent de détermination des LOS en considérant $y_h = z_h - z_{h-1} = x_h - x_{h-1}$ comme les largeurs de strates, puis en construisant la fonction objectif et les contraintes comme des fonctions de y_h. Le PPM est traité comme un problème de décision à plusieurs degrés dans lequel, à chaque degré, la valeur de la LOS, donc de la BOS pour une strate, est déterminée en utilisant une méthode de programmation dynamique avec une équation récurrente progressive.

Soit $f(x)$ la fonction de probabilité, et x_0 et x_L, la plus petite et la plus grande valeurs de x. Si la moyenne de population est estimée sous une répartition de Neyman, le problème de la détermination des bornes des strates consiste à découper l'étendue,

$$x_L - x_0 = d, \qquad (13)$$

aux points intermédiaires $x_1 \leq x_2 \leq, \ldots, \leq x_{L-1}$, de sorte que dans (8), $\sum_{h=1}^{L} W_h \sigma_h$ soit minimale.

Considérons que $f(x)$ comprend n fonctions linéaires ou non linéaires continues par morceaux telles que :

$$f(x) = \begin{cases} g_1(x)\,; & x_0 = a_0 \leq x \leq a_1, \\ g_2(x)\,; & a_1 < x \leq a_2, \\ \vdots & \vdots \\ g_n(x)\,; & a_{n-1} < x \leq a_n = x_L. \end{cases} \qquad (14)$$

Supposons aussi que, parmi les L strates, l_i est le nombre de strates qui doivent être formées sous la densité de probabilité $g_i(x)$; $i = 1, 2, \ldots, n$ et que $\sum_{i=1}^{n} l_i = L$.

Si $f(x)$ dans (14) est intégrable, en utilisant les expressions (3), (4) et (5), nous obtenons W_h, σ_h^2 et μ_h sous forme d'une fonction des bornes x_h et x_{h-1}. Donc, dans (8), la fonction objectif peut être exprimée comme une fonction des bornes x_h et x_{h-1} uniquement. Soit

$$\phi_h(x_h, x_{h-1}) = W_h \sigma_h.$$

Notons que la fonction susmentionnée possède des valeurs différentes pour les diverses densités de probabilité dans (14).

Donc, le problème (8) peut être traité comme un problème d'optimisation en vue de trouver $x_1, x_2, \ldots, x_{L-1}$ tel qu'il est énoncé dans (9).

Soit $y_h = x_h - x_{h-1} \geq 0$ la largeur de la h^{e} $(h = 1, 2, \ldots, L)$ strate.

Partant de la définition susmentionnée de y_h, l'étendue de la distribution donnée par (13) est exprimée comme la fonction des largeurs de strate :

$$\sum_{h=1}^{L} y_h = \sum_{h=1}^{L} (x_h - x_{h-1}) = x_L - x_0 = d. \qquad (15)$$

La k^{e} borne de strate x_k ; $k = 1, 2, \ldots, L-1$ est alors exprimée par :

$$\begin{aligned} x_k &= x_0 + y_1 + y_2 + \ldots + y_k \\ &= x_{k-1} + y_k, \end{aligned}$$

qui est une fonction de la k^{e} largeur de strate et de la $(k-1)^{\text{e}}$ borne de strate.

Si nous considérons $z_k = x_k$ et ajoutons (15) comme contrainte, le problème (11) peut être réécrit sous la forme d'un problème équivalent de détermination des LOS :

$$\begin{aligned} &\text{Minimiser } \sum_{h=1}^{L} \phi_h(y_h, x_{h-1}), \\ &\text{sous les contraintes} \\ &\sum_{h=1}^{L} y_h = d, \\ &y_h \geq 0;\ h = 1, 2, \ldots, L. \end{aligned} \qquad (16)$$

Au départ, x_0 est connue. Par conséquent, le premier terme, c'est-à-dire $\phi_1(y_1, x_0)$, de la fonction objectif du PPM (16) est une fonction de y_1 uniquement. Une fois que y_1 est connue, le point de stratification suivant $x_1 = x_0 + y_1$ sera connu et le deuxième terme de la fonction objectif $\phi_2(y_2, x_1)$ deviendra une fonction de y_2 uniquement.

Par conséquent, en énonçant la fonction objectif comme une fonction de y_h uniquement, le PPM (16) s'exprime par :

$$\text{Minimiser} \quad \sum_{h=1}^{L} \phi_h(y_h),$$
sous les contraintes
$$\sum_{h=1}^{L} y_h = d,$$
$$y_h \geq 0\,;\; h = 1,\, 2,\, ...,\, L. \tag{17}$$

Les sections 3.1 et 3.2 illustrent la formulation du problème de la détermination des LOS comme un PPM pour les variables étudiées de loi triangulaire et de loi normale standard, respectivement.

3.1 PPM pour la loi triangulaire

Posons que la variable étudiée x suit la loi triangulaire sur l'intervalle [a, b] avec la densité de probabilité :

$$f(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)}\,; & a \leq x \leq c \\[2ex] \dfrac{2(b-x)}{(b-a)(b-c)}\,; & c < x \leq b, \end{cases} \tag{18}$$

où a est un paramètre d'emplacement, b est un paramètre d'échelle et c est le paramètre de forme.

Elle possède deux fonctions par morceaux.

Quand $a \leq x \leq c$, en partant de (18) et en utilisant (3), (5) et (4), nous obtenons W_h, μ_h, et σ_h^2 comme il suit :

$$W_h = \frac{y_h(y_h + 2a_h)}{(b-a)(c-a)}, \tag{19}$$

$$\mu_h = \frac{\frac{2}{3}y_h^2 + 2y_h x_{h-1} - a y_h + 2a_h x_{h-1}}{y_h + 2a_h},$$

et

$$\sigma_h^2 = \frac{y_h^2\,[\,y_h^2 + 6a_h y_h + 6a_h^2\,]}{18(y_h + 2a_h)^2}, \tag{20}$$

où $y_h = x_h - x_{h-1}$, $a_h = x_{h-1} - a$ et $a \leq x_{h-1} \leq x_h \leq c.$

Donc, de (19) et (20), il découle que

$$W_h\,\sigma_h = \frac{y_h^2\sqrt{y_h^2 + 6a_h\,y_h + 6a_h^2}}{3\sqrt{2}(b-a)(c-a)}. \tag{21}$$

De même, quand $c < x \leq b$, en partant de (18) et en utilisant (3), (5) et (4), nous pouvons démontrer que

$$W_h = \frac{y_h(2b_h - y_h)}{(b-a)(b-c)}, \tag{22}$$

$$\mu_h = \frac{3b_h y_h - 3y_h x_{h-1} + 6b_h x_{h-1} - 2y_h^2}{3(2b_h - y_h)},$$

et

$$\sigma_h^2 = \frac{y_h^2(6b_h^2 - 6b_h y_h + y_h^2)}{18(2b_h - y_h)^2}, \tag{23}$$

où $y_h = x_h - x_{h-1}$, $b_h = b - x_{h-1}$ et $c < x_{h-1} \leq x_h \leq b.$

Donc, de (22) et (23), il découle que

$$W_h\sigma_h = \frac{y_h^2\sqrt{6b_h^2 - 6b_h y_h + y_h^2}}{3\sqrt{2}(b-a)(b-c)}. \tag{24}$$

Soit λ_1 et λ_2 la dernière et la première strates formées sous la première et la seconde fonctions par morceaux de (18), respectivement. Si toute strate (disons l) tombe sous les deux fonctions, λ_1 et λ_2 ne sont pas considérées comme étant deux strates distinctes, mais comme les fractions de la même l^{e} strate. Alors, en utilisant (21) et (24), le PPM (17) peut être exprimé comme le problème de détermination des LOS pour la variable étudiée dont la fonction de fréquence est triangulaire, sous la forme :

$$\text{Minimiser} \left\{ \sum_{h=1}^{\lambda_1} \frac{y_h^2\sqrt{y_h^2 + 6a_h y_h + 6a_h^2}}{3\sqrt{2}(b-a)(c-a)} + \sum_{h=\lambda_2}^{L} \frac{y_h^2\sqrt{6b_h^2 - 6b_h y_h + y_h^2}}{3\sqrt{2}(b-a)(b-c)} \right\},$$

$$\text{sous les contraintes} \quad \sum_{h=1}^{L} y_h = d,$$

$$\text{et} \quad y_h \geq 0\,;\; h = 1,\, 2,\, ...,\, L, \tag{25}$$

où $d = b - a.$

3.2 PPM pour la loi normale

La variable étudiée x suit une loi normale standard si sa fonction de densité de probabilité est donnée par

$$f(x) = \frac{1}{\sqrt{2\pi}}\ \exp\left(-\frac{x^2}{2}\right); \quad -\infty < x < \infty.$$

Comme à la section 3.1, en utilisant les définition (3), (5) et (4), nous pouvons voir que

$$W_h = \frac{erf\left(\dfrac{y_h + x_{h-1}}{\sqrt{2}}\right) - erf\left(\dfrac{x_{h-1}}{\sqrt{2}}\right)}{2}, \tag{26}$$

$$\mu_h = \frac{\sqrt{2}\left[\exp\left(-\dfrac{x_{h-1}^2}{2}\right) - \exp\left(-\dfrac{(y_h + x_{h-1})^2}{2}\right)\right]}{\sqrt{\pi}\left[erf\left(\dfrac{y_h + x_{h-1}}{\sqrt{2}}\right) - erf\left(\dfrac{x_{h-1}}{\sqrt{2}}\right)\right]}$$

et

$$\sigma_h^2 = \Bigg\{ \sqrt{2\pi} \Bigg[x_{h-1} \exp\left(-\frac{x_{h-1}^2}{2} \right) erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right)$$

$$- (y_h + x_{h-1}) \exp\left(-\frac{(y_h + x_{h-1})^2}{2} \right) erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right)$$

$$- x_{h-1} \exp\left(-\frac{x_{h-1}^2}{2} \right) erf\left(\frac{x_{h-1}}{\sqrt{2}} \right)$$

$$+ (y_h + x_{h-1}) \exp\left(-\frac{(y_h + x_{h-1})^2}{2} \right) erf\left(\frac{x_{h-1}}{\sqrt{2}} \right) \Bigg]$$

$$+ \pi \left[erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right) - erf\left(\frac{x_{h-1}}{\sqrt{2}} \right) \right]^2$$

$$- 2 \left[\exp\left(-\frac{x_{h-1}^2}{2} \right) - \exp\left(-\frac{(y_h + x_{h-1})^2}{2} \right) \right]^2 \Bigg\}$$

$$\div \pi \left[erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right) - erf\left(\frac{x_{h-1}}{\sqrt{2}} \right) \right]^2, \tag{27}$$

où $y_h = x_h - x_{h-1}$, $erf(x_h) - erf(x_{h-1}) = (2/\sqrt{\pi}) \times \int_{x_{h-1}}^{x_h} \exp(-u^2)\, du$ et $h = 1, 2, ..., L$.

Par conséquent, en utilisant les valeurs données par (26) et (27), nous pouvons exprimer le PPM (17) sous la forme

$$\text{Minimiser} \sum_{h=1}^{L} \text{Sqrt} \Bigg\{ \frac{1}{2\sqrt{2\pi}} \Bigg[x_{h-1} \exp\left(-\frac{x_{h-1}^2}{2} \right) erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right)$$

$$- (y_h + x_{h-1}) \exp\left(-\frac{(y_h + x_{h-1})^2}{2} \right) erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right)$$

$$- x_{h-1} \exp\left(-\frac{x_{h-1}^2}{2} \right) erf\left(\frac{x_{h-1}}{\sqrt{2}} \right)$$

$$+ (y_h + x_{h-1}) \exp\left(-\frac{(y_h + x_{h-1})^2}{2} \right) erf\left(\frac{x_{h-1}}{\sqrt{2}} \right) \Bigg]$$

$$+ \frac{1}{4} \left[erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right) - erf\left(\frac{x_{h-1}}{\sqrt{2}} \right) \right]^2$$

$$- \frac{1}{2\pi} \left[\exp\left(-\frac{x_{h-1}^2}{2} \right) - \exp\left(-\frac{(y_h + x_{h-1})^2}{2} \right) \right]^2 \Bigg\}$$

sous les contraintes

$$\sum_{h=1}^{L} y_h = d$$

$$y_h \geq 0\,;\ h = 1, 2, ..., L. \tag{28}$$

4. L'algorithme fondé sur la méthode de programmation dynamique

Le PPM (17) est un problème de décision à plusieurs degrés dans lequel la fonction objectif et les contraintes peuvent être séparées de y_h, ce qui nous permet d'utiliser une méthode de programmation dynamique telle que celle illustrée par Bühler et Deutler (1975) pour le problème (11).

Considérons le sous-problème suivant de (17) pour les $k(< L)$ premières strates :

$$\text{Minimiser} \sum_{h=1}^{k} \phi_h(y_h),$$

sous les contraintes

$$\sum_{h=1}^{k} y_h = d_k,$$

$$y_h \geq 0\,;\ h = 1, 2, ..., k, \tag{29}$$

où $d_k < d$ est la largeur totale disponible pour la division en k strates ou la valeur d'état au degré k. Notons que $d_k = d$ pour $k = L$.

Les fonctions de transformation sont données par

$$\begin{aligned}
d_k &= y_1 + y_2 + ... + y_k, \\
d_{k-1} &= y_1 + y_2 + ... + y_{k-1} = d_k - y_k, \\
d_{k-2} &= y_1 + y_2 + ... + y_{k-2} = d_{k-1} - y_{k-1}, \\
&\vdots \\
d_2 &= y_1 + y_2 = d_3 - y_3, \\
d_1 &= y_1 = d_2 - y_2.
\end{aligned}$$

Soit $\Phi_k(d_k)$ la valeur minimale de la fonction objectif de (29), c'est-à-dire

$$\Phi_k(d_k) = \min \left[\sum_{h=1}^{k} \phi_h(y_h) \,\middle|\, \sum_{h=1}^{k} y_h = d_k, \quad \text{et} \quad y_h \geq 0; h = 1, 2, ..., k \right].$$

En utilisant la définition susmentionnée de $\Phi_k(d_k)$, le PPM (17) équivaut à calculer $\Phi_L(d)$ par récurrence en trouvant $\Phi_k(d_k)$ pour $k = 1, 2, ..., L$ et $0 \leq d_k \leq d$.

Nous pouvons écrire :

$$\Phi_k(d_k) = \min \left[\phi_k(y_k) + \sum_{h=1}^{k-1} \phi_h(y_h) \,\middle|\, \sum_{h=1}^{k-1} y_h = d_k - y_k, \quad \text{et} \quad y_h \geq 0;\ h = 1, 2, ..., k-1. \right]$$

Pour une valeur fixe de y_k ; $0 \leq y_k \leq d_k$,

$$\Phi_k(d_k) = \phi_k(y_k) + \min\left[\sum_{h=1}^{k-1}\phi_h(y_h)\,\middle|\,\sum_{h=1}^{k-1}y_h = d_k - y_k,\right.$$
$$\left.\text{et } y_h \geq 0;\, h = 1, 2, \ldots, k-1\right].$$

En appliquant le principe d'optimalité de Bellman, nous écrivons une récurrence progressive au lieu d'une récurrence rétrograde comme l'ont proposé Bühler et Deutler dans (12) pour appliquer la méthode de programmation dynamique, soit :

$$\Phi_k(d_k) = \min_{0 \leq y_k \leq d_k}\left[\phi_k(y_k) + \Phi_{k-1}(d_k - y_k)\right],\ k \geq 2. \quad (30)$$

Pour le premier degré, c'est-à-dire pour $k = 1$:

$$\Phi_1(d_1) = \phi_1(d_1) \Rightarrow\ y_1^* = d_1, \quad (31)$$

où $y_1^* = d_1$ est la largeur optimale de la première strate. Les relations (30) et (31) sont résolues par récurrence pour chaque $k = 1, 2, \ldots, L$ et $0 \leq d_k \leq d$, et nous obtenons $\Phi_L(d)$. Partant de $\Phi_L(d)$, la largeur optimale de la L^{e} strate, nous obtenons y_L^*. À partir de $\Phi_{L-1}(d - y_L^*)$, la largeur optimale de la $(L-1)^{\text{e}}$ strate, nous obtenons y_{L-1}^*, et ainsi de suite jusqu'à ce que nous obtenions y_1^*.

Soulignons que, selon la fonction par morceaux donnée dans (14) sous laquelle la strate est formée, $\phi_k(y_k)$ dans (30) prendra une valeur différente pour chaque y_k, comme il suit :

$$y_k = x_k - x_{k-1} \leq a_i - a_0, \text{ pour un } i\ (i = 1, 2, \ldots, n)$$

et

$$x_k \in [a_{i-1}, a_i], \qquad \text{pour un } i\ (i = 1, 2, \ldots, n).$$

5. Exemples numériques

À la présente section, nous présentons le détail des calculs de l'algorithme exposé à la section 4 pour les PPM (25) et (28).

5.1 Loi triangulaire

Supposons que $a = x_0 = 0,\ c = 1$ et $b = x_L = 2.$ Cela implique que $d = x_L - x_0 = 2$ et le PPM (25) s'exprime sous la forme :

$$\text{Minimiser } \left\{\sum_{h=1}^{\lambda_1}\frac{y_h^2\sqrt{y_h^2 + 6a_h y_h + 6a_h^2}}{6\sqrt{2}} + \sum_{h=\lambda_2}^{L}\frac{y_h^2\sqrt{6b_h^2 - 6b_h y_h + y_h^2}}{6\sqrt{2}}\right\},$$
$$\text{sous les contraintes } \sum_{h=1}^{L} y_h = 2,$$
$$\text{et } y_h \geq 0\,;\ h = 1, 2, \ldots, L, \quad (32)$$

où $a_h = x_{h-1}$ et $b_h = 2 - x_{h-1}$.

En utilisant (30) et (31), nous pouvons énoncer les équations récurrentes pour la résolution du PPM (32) de la façon suivante :

Pour le premier degré $(k = 1)$

$$\Phi_1(d_1) = \frac{d_1^3}{6\sqrt{2}} \qquad \text{à} \qquad y_1 = d_1, \quad (33)$$

et pour les autres degrés $(k \geq 2)$

$$\Phi_k(d_k) = \begin{cases} \min\left[\dfrac{y_k^2\sqrt{y_k^2 + 6a_k y_k + 6a_k^2}}{6\sqrt{2}} + \Phi_{k-1}(d_k - y_k)\right] & \text{si } 0 \leq d_k \leq 1, \\ \min\left[\dfrac{y_k^2\sqrt{6b_k^2 - 6b_k y_k + y_k^2}}{6\sqrt{2}} + \Phi_{k-1}(d_k - y_k)\right] & \text{si } 1 < d_k \leq 2, \end{cases} \quad (34)$$

où le minimum est calculé sur $0 \leq y_k \leq d_k,\ a_k = x_{k-1} = d_k - y_k$ et $b_k = 2 - x_{k-1} = 2 - d_k + y_k$.

En remplaçant a_k et b_k par ces valeurs dans (34), nous obtenons

$$\Phi_k(d_k) = \begin{cases} \min\left[\dfrac{y_k^2\sqrt{y_k^2 + 6(d_k - y_k)\,d_k}}{6\sqrt{2}} + \Phi_{k-1}(d_k - y_k)\right] & \text{si } 0 \leq d_k \leq 1, \\ \min\left[\dfrac{y_k^2\sqrt{y_k^2 + 6(2 - d_k + y_k)(2 - d_k)}}{6\sqrt{2}} + \Phi_{k-1}(d_k - y_k)\right] & \text{si } 1 < d_k \leq 2, \end{cases} \quad (35)$$

où le minimum est calculé sur $0 \leq y_k \leq d_k$.

Puis, en résolvant les équations récurrentes (33) et (35) à l'aide d'un programme informatique développé pour l'algorithme donné à la section 4, nous obtenons les LOS. Le tableau 1 donne les résultats des largeurs optimales de strate y_h^* et, donc, les bornes optimales de strate x_h^*, ainsi que les valeurs de la fonction objectif $\sum_{h=1}^{L}\phi_h(y_h)$ pour $L = 2, 3, 4, 5$ et 6.

Tableau 1
Largeurs et bornes optimales de strate pour la loi triangulaire

Nombre de strates L	Largeurs optimales de strate (LOS) (y_h^*)	Bornes optimales de strate (BOS) $(x_h^* = x_{h-1}^* + y_h^*)$	Valeurs optimales de la fonction objectif $\sum_{h=1}^{L} \phi_h(y_h) = \sum_{h=1}^{L} W_h \sigma_h$
2	$y_1^* = 1{,}000000$ $y_2^* = 1{,}000000$	$x_1^* = 1{,}000000$	0,2357022604
3	$y_1^* = 0{,}838081$ $y_2^* = 0{,}411608$ $y_3^* = 0{,}750311$	$x_1^* = 0{,}838081$ $x_2^* = 1{,}249689$	0,1655523797
4	$y_1^* = 0{,}645751$ $y_2^* = 0{,}354249$ $y_3^* = 0{,}354249$ $y_4^* = 0{,}645751$	$x_1^* = 0{,}645751$ $x_2^* = 1{,}000000$ $x_3^* = 1{,}354249$	0,1226262641
5	$y_1^* = 0{,}582819$ $y_2^* = 0{,}319725$ $y_3^* = 0{,}252176$ $y_4^* = 0{,}299439$ $y_5^* = 0{,}545841$	$x_1^* = 0{,}582819$ $x_2^* = 0{,}902544$ $x_3^* = 1{,}154720$ $x_4^* = 1{,}454159$	0,0998893913
6	$y_1^* = 0{,}497369$ $y_2^* = 0{,}272849$ $y_3^* = 0{,}229782$ $y_4^* = 0{,}229782$ $y_5^* = 0{,}272849$ $y_6^* = 0{,}497369$	$x_1^* = 0{,}497369$ $x_2^* = 0{,}770218$ $x_3^* = 1{,}000000$ $x_4^* = 1{,}229782$ $x_5^* = 1{,}502631$	0,0829362498

5.2 Loi normale

Posons que x suit la loi normale standard dans l'intervalle $(x_0,\ x_L)$. Pour les besoins de l'exemple, nous supposons que $x_0 = -4$ et $x_L = 4$. Alors, $d = 8$, ce qui donne le PPM (28) sous la forme :

$$
\begin{aligned}
\text{Minimize } & \sum_{h=1}^{L} \left\{ \text{Sqrt} \left\{ \frac{1}{2\sqrt{2\pi}} \times \right.\right. \\
& \left[x_{h-1} \exp\left(-\frac{x_{h-1}^2}{2}\right) erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}}\right) \right. \\
& -(y_h + x_{h-1}) \exp\left(-\frac{(y_h + x_{h-1})^2}{2}\right) erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}}\right) \\
& -x_{h-1} \exp\left(-\frac{x_{h-1}^2}{2}\right) erf\left(\frac{x_{h-1}}{\sqrt{2}}\right) \\
& \left. +(y_h + x_{h-1}) \exp\left(-\frac{(y_h + x_{h-1})^2}{2}\right) erf\left(\frac{x_{h-1}}{\sqrt{2}}\right) \right] \\
& +\frac{1}{4}\left[erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}}\right) - erf\left(\frac{x_{h-1}}{\sqrt{2}}\right) \right]^2 \\
& \left.\left. -\frac{1}{2\pi}\left[\exp\left(-\frac{x_{h-1}^2}{2}\right) - \exp\left(-\frac{(y_h + x_{h-1})^2}{2}\right) \right]^2 \right\} \right\},
\end{aligned}
$$

$$
\text{sous les contraintes } \sum_{h=1}^{L} y_h = 8,
$$

$$
\text{et } y_h \geq 0\,;\ h = 1,\ 2,\ \ldots,\ L. \tag{36}
$$

Nous avons

$$
\begin{aligned}
x_{k-1} &= x_0 + y_1 + y_2 + \ldots + y_{k-1} \\
&= -4 + y_1 + y_2 + \ldots + y_{k-1} \\
&= d_{k-1} - 4 \\
&= d_k - y_k - 4.
\end{aligned}
$$

En introduisant cette valeur de x_{k-1} par substitution dans (36) et en utilisant (30) et (31), nous obtenons les équations récurrentes pour la résolution du PPM (36) comme il suit.

Pour le premier degré $(k = 1)$:

$$
\begin{aligned}
\Phi_1(d_1) = & \left\{ \text{Sqrt} \left\{ \frac{1}{2\sqrt{2\pi}} \left[-\exp\left(-\frac{1}{2}\right) erf\left(\frac{(d_1 - 4)}{\sqrt{2}}\right) \right.\right.\right. \\
& - (d_1 - 4) \exp\left(-\frac{(d_1 - 4)^2}{2}\right) erf\left(\frac{(d_1 - 4)}{\sqrt{2}}\right) \\
& + \exp\left(-\frac{1}{2}\right) erf\left(-\frac{1}{\sqrt{2}}\right) \\
& \left. + (d_1 - 4) \exp\left(-\frac{(d_1 - 4)^2}{2}\right) erf\left(-\frac{1}{\sqrt{2}}\right) \right] \\
& + \frac{1}{4}\left[erf\left(\frac{(d_1 - 4)}{\sqrt{2}}\right) - erf\left(-\frac{1}{\sqrt{2}}\right) \right]^2 \\
& \left.\left. - \frac{1}{2\pi}\left[\exp\left(-\frac{1}{2}\right) - \exp\left(-\frac{(d_1 - 4)^2}{2}\right) \right]^2 \right\} \right\}
\end{aligned} \tag{37}
$$

à $y_1 = d_1$,
et pour les degrés $k \geq 2$:

$$
\begin{aligned}
\Phi_k(d_k) = \min_{0 \leq y_k \leq d_k} & \left\{ \text{Sqrt} \left\{ \frac{1}{2\sqrt{2\pi}} \times \right.\right. \\
& \left[(d_k - y_k - 4) \exp\left(-\frac{(d_k - y_k - 4)^2}{2}\right) erf\left(\frac{d_k - 4}{\sqrt{2}}\right) \right. \\
& - (d_k - 4) \exp\left(-\frac{(d_k - 4)^2}{2}\right) erf\left(\frac{d_k - 4}{\sqrt{2}}\right) \\
& - (d_k - y_k - 4) \exp\left(-\frac{(d_k - y_k - 4)^2}{2}\right) erf\left(\frac{(d_k - y_k - 4)}{\sqrt{2}}\right)
\end{aligned}
$$

$$+ (d_k - 4)\exp\left(-\frac{(d_k-4)^2}{2}\right)erf\left(\frac{(d_k-y_k-4)}{\sqrt{2}}\right)\Bigg]$$

$$+\frac{1}{4}\left[erf\left(\frac{d_k-4}{\sqrt{2}}\right)-erf\left(\frac{(d_k-y_k-4)}{\sqrt{2}}\right)\right]^2$$

$$-\frac{1}{2\pi}\left[\exp\left(-\frac{(d_k-y_k-4)^2}{2}\right)-\exp\left(-\frac{(d_k-4)^2}{2}\right)\right]^2\Bigg\}$$

$$+\ \Phi_{k-1}(d_k - y_k)\Bigg\}. \tag{38}$$

En résolvant les équations récurrentes (37) et (38), nous obtenons les largeurs optimales de strate y_h^* et, donc, les bornes optimales de strate x_h^*. Le tableau 2 donne ces résultats, ainsi que les valeurs de la fonction objectif $\sum_{h=1}^{L}\phi_h(y_h)$ pour $L = 2$, 3, 4, 5 et 6.

6. Discussion

À la présente section, nous effectuons une étude numérique de l'efficacité de la méthode de programmation dynamique comparativement à la méthode de la cumulative de $\sqrt{f}$ de Dalenius et Hodges, qui est celle utilisée le plus fréquemment et la mieux connue. Pour cela, nous avons produit des données de taille $N = 10\,000$ pour une population de densité de probabilité normale standard $f(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$, que nous avons groupées en 19 classes égales. Au tableau 3, nous donnons les fréquences de classe dans la colonne 2 et leurs racines cumulatives dans la colonne 3.

Dans cet exemple, la plus petite et la plus grande valeurs de x sont $x_0 = -3{,}98$ et $x_L = 3{,}62$, respectivement. Par conséquent, l'étendue de la distribution est $d = 7{,}60$.

Pour déterminer les BOS pour cette distribution, nous utilisons la méthode de la cumulative de $\sqrt{f}$ ainsi que la méthode de programmation dynamique. Pour chaque $L = 2, 3, 4, 5$ et 6, nous calculons la variance $\sum_{h=1}^{L} W_h\sigma_h$, que nous utilisons pour évaluer l'efficacité des deux méthodes de stratification. Les résultats de cette comparaison sont donnés au tableau 4. L'examen de la dernière colonne du tableau montre que les BOS obtenues par la méthode de programmation dynamique sont plus efficaces pour tout $L = 1, 2, ..., 6$. En outre, l'efficacité de la méthode de la cumulative de $\sqrt{f}$ dépend du choix initial du nombre de classes, mais il n'existe aucun fondement théorique donnant le meilleur nombre de classes (voir Hedlin 2000).

Tableau 2
Largeurs et bornes optimales de strate pour la loi normale standard

Nombre de strates L	Largeurs optimales de strate (LOS) (y_h^*)	Bornes optimales de strate (BOS) ($x_h^* = x_{h-1}^* + y_h^*$)	Valeurs optimales de la fonction objectif $\sum_{h=1}^{L}\phi_h(y_h) = \sum_{h=1}^{L} W_h\sigma_h$
2	$y_1^* = 4{,}000000$ $y_2^* = 4{,}000000$	$x_1^* = 0{,}000000$	0,6021710931
3	$y_1^* = 3{,}450300$ $y_2^* = 1{,}099400$ $y_3^* = 3{,}450300$	$x_1^* = -0{,}549700$ $x_2^* = 0{,}549700$	0,4265717619
4	$y_1^* = 3{,}124570$ $y_2^* = 0{,}875430$ $y_3^* = 0{,}875430$ $y_4^* = 3{,}124570$	$x_1^* = -0{,}875430$ $x_2^* = 0{,}000000$ $x_3^* = 0{,}875430$	0,3297899642
5	$y_1^* = 2{,}896360$ $y_2^* = 0{,}767900$ $y_3^* = 0{,}671480$ $y_4^* = 0{,}767900$ $y_5^* = 2{,}896360$	$x_1^* = -1{,}103640$ $x_2^* = -0{,}335740$ $x_3^* = 0{,}335740$ $x_4^* = 1{,}103640$	0,2686646379
6	$y_1^* = 2{,}722440$ $y_2^* = 0{,}702200$ $y_3^* = 0{,}575360$ $y_4^* = 0{,}575360$ $y_5^* = 0{,}702200$ $y_6^* = 2{,}722440$	$x_1^* = -1{,}277560$ $x_2^* = -0{,}575360$ $x_3^* = 0{,}000000$ $x_4^* = 0{,}575360$ $x_5^* = 1{,}277560$ $x_6^* = 4{,}000000$	0,2265979522

Tableau 3
Distribution des fréquences de x et cum $\sqrt{f(x)}$

Classe	Fréquence $f(x)$	Cum $\sqrt{f(x)}$
(-3,98)-(-3,58)	2	1,4
(-3,58)-(-3,18)	6	3,8
(-3,18)-(-2,78)	23	8,6
(-2,78)-(-2,38)	59	16,3
(-2,38)-(-1,98)	155	28,7
(-1,98)-(-1,58)	296	45,9
(-1,58)-(-1,18)	630	71,0
(-1,18)-(-0,783)	1 015	102,9
(-0,783)-(-0,383)	1 361	139,8
(-0,383)-0,017	1 551	179,2
0,017-0,417	1 495	217,9
0,417-0,817	1 315	254,2
0,817-1,22	1 003	285,9
1,22-1,62	613	310,7
1,62-2,02	285	327,6
2,02-2,42	128	338,9
2,42-2,82	38	345,1
2,82-3,22	18	349,3
3,22-3,62	7	351,9

Enfin, les autres méthodes décrites dans la littérature, telles que celles de Aoyama (1954), Ekman (1959) et Sethi (1963), sont avant tout des méthodes classiques appliquées en vue d'obtenir des bornes de strate approximatives. De nombreux auteurs, dont Unithan (1978), Lavallée et Hidiroglou (1988), Sweet et Sigman (1995), et Rivest (2002), ont proposé des méthodes itératives. Ces dernières requièrent des solutions initiales approximatives. En outre, il n'existe aucune garantie qu'une méthode itérative donnera le minimum global en l'absence d'une solution initiale approximative appropriée et la fonction de variance possède plus d'un minimum local. L'avantage de la méthode de programmation dynamique tient au fait qu'elle donne le minimum global de la fonction objectif et qu'elle ne nécessite aucune solution initiale approximative.

Tableau 4
Efficacité relative de la méthode de programmation dynamique

L	Méthode cum $\sqrt{f}$		Méthode de programmation dynamique		Efficacité relative
	BOS	$\sum_{h=1}^{L} W_h \sigma_h$	BOS	$\sum_{h=1}^{L} W_h \sigma_h$	
2	-0,017	0,60131	-0,00034	0,60126	100,00832
3	-0,783	0,43177	-0,55015	0,42576	101,41159
	0,417		0,54884		
4	-0,783	0,33067	-0,87593	0,32905	100,49233
	-0,017		-0,00081		
	0,817		0,87395		
5	-1,18	0,27066	-1,10418	0,26799	100,99631
	-3,83		-0,33656		
	0,417		0,33452		
	1,22		1,10147		
6	-1,18	0,24242	-1,27813	0,22598	107,27498
	-0,783		-0,57619		
	-0,017		-0,00115		
	0,417		0,57369		
	1,22		1,27462		

Remerciements

Les auteurs remercient le rédacteur associé et les examinateurs de leurs suggestions et commentaires précieux en vue d'améliorer le manuscrit

Bibliographie

Aoyama, H. (1954). A study of stratified random sampling. *Annals of the Institute of Statistical Mathematics*, 6, 1-36.

Bellman, R.E. (1957). *Dynamic Programming*. Princetown University Press, New Jersey.

Bühler, W., et Deutler, T. (1975). Optimal stratification and grouping by dynamic programming. *Metrika*, 22, 161-175.

Cochran, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 38, 2, 345-358.

Dalenius, T. (1950). The problem of optimum stratification-II. *Skand. Aktuartidskr*, 33, 203-213.

Dalenius, T., et Gurney, M. (1951). The problem of optimum stratification. *Skand. Aktuartidskr*, 34, 133-148.

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

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

Ekman, G. (1959). Approximate expression for conditional mean and variance over small intervals of a continuous distribution. *Annals of the Institute of Statistical Mathematics*, 30, 1131-1134.

Hansen, M.H., et Hurwitz, W.N. (1953). On the theory of sampling from finite population. *The Annals of Mathematical Statistics*, 14, 333-362.

Hedlin, D. (2000). A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*, 16, 15-29.

Hess, I., Sethi, V.K. et Balakrishnan, T.R. (1966). Stratification: A practical investigation. *Journal of the Americal Statistical Association*, 61, 71-90.

Hidiroglou, M.A., et Srinath, K.P. (1993). Problems associated with designing subannual bussiness surveys. *Journal of Bussiness and Economic Statistics*, 11, 397-405.

Khan, E.A., Khan, M.G.M. et Ahsan, M.J. (2002). Optimum stratification: A mathematical programming approach. *Culcutta Statistical Association Bulletin*, 52 (spécial), 205-208.

Khan, M.G.M., Najmussehar, et Ahsan, M.J. (2005). Optimum stratification for exponential study variable under Neyman allocation. *Journal of Indian Society of Agricultural Statistics*, 59(2), 146-150.

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

Lavallée, P. (1987). Some contributions to optimal stratification. Thèse de maîtrise, Carleton University, Ottawa, Canada.

Lavallée, P. (1988). Two-way optimal stratification using dynamic programming. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Virginia, 646-651.

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.

Mahalanobis, P.C. (1952). Some aspects of the design of sample surveys. *Sankhyā*, 12, 1-7.

Murthy, M.N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

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

Nicolini, G. (2001). A method to define strata boundaries. *Departmental Working Papers 2001-01*, Department of Economics, University of Milan, Italie (www.economia.unimi.it/pubb/wp83.pdf).

Niemiro, W. (1999). Konstrukcja optymalnej stratyfikacja metoda poszukiwan losowych. (Optimal Stratification using Random Search Method). *Wiadomosci Statystyczne*, 10, 1-9.

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.

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

Sweet, E.M., et Sigman, R.S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 491-496.

Unnithan, V.K.G. (1978). The minimum variance boundry points of stratification. *Sankhyā*, 40, C, 60-72.

Unnithan, V.K.G., et Nair, N.U. (1995). Minimum variance stratification. *Communications in Statistics*, 24(1), 275-284.