Techniques d'enquête, juin 1989
Vol. 15, n° 1, pp. 49–60
Statistique Canada

Répartition de l'échantillon dans les enquêtes à plusieurs variables

JAMES BETHEL[1]

RÉSUMÉ

Dans les enquêtes à objectifs multiples, on réalise souvent la répartition optimale de l'échantillon en définissant des contraintes linéaires pour la variance puis en utilisant la programmation convexe pour minimiser le coût de l'enquête. En nous servant du théorème de Kuhn-Tucker, nous définissons dans cet article une formule de répartition optimale en fonction des multiplicateurs de Lagrange. Cette formule sert ensuite à déterminer la dérivée partielle de la fonction de coût par rapport à la k-ième contrainte de variance; cette dérivée partielle est $-2\alpha_k^* g(x^*)/v_k$, où $g(x^*)$ est le coût de la répartition optimale et α_k^* et v_k sont respectivement, le k-ième multiplicateur de Lagrange normalisé et la borne supérieure pour la précision de la k-ième variable. Enfin, nous présentons un algorithme de calcul simple et en analysons les propriétés de convergence. Nous illustrons l'application de ces résultats à un plan de sondage à l'aide des données d'une enquête sur les établissements commerciaux.

MOTS CLÉS: Répartition de l'échantillon avec objectifs multiples; programmation non linéaire; échantillonage stratifié.

1. INTRODUCTION

La répartition optimale de l'échantillon dans les enquêtes à objectifs multiples est un sujet qui a été traité pour la première fois par Neyman (1934), lorsque celui-ci élaborait sa théorie sur la répartition optimale dans les enquêtes à une variable. Depuis lors, de nombreux statisticiens se sont penchés sur la question et ont proposé plusieurs méthodes qui, pour la plupart, peuvent être classées en deux grandes catégories. La première catégorie consiste à calculer la moyenne pondérée des variances de strates et à trouver la répartition optimale pour la «variance moyenne» ainsi calculée. Dalenius (1953), Yates (1960), Folks et Antle (1965), Hartley (1965) et Kish (1976) analysent des méthodes de ce genre. La seconde catégorie de méthodes consiste à imposer une contrainte à chaque variance (sous forme d'une inéquation) puis à utiliser la programmation convexe pour déterminer le mode de répartition optimale qui satisfait toutes les contraintes. Dalenius (1957), Yates (1960), Kokan (1963), Hartley (1965), Kokan et Khan (1967), Chatterjee (1969, 1972), Huddleston, Claypool et Hocking (1970), Bethel (1985) et Chromy (1987) étudient tous l'un après l'autre l'utilisation de la programmation convexe par rapport à la répartition optimale dans les enquêtes à plusieurs variables. Chaque catégorie de méthodes a ses avantages et ses inconvénients. La méthode de la «moyenne pondérée» est simple au point de vue du calcul, intéressante sur le plan intuitif et peut être appliquée suivant une hypothèse de coût fixe, mais les poids sont choisis de façon arbitraire et les propriétés d'optimalité ne sont pas clairement définies. La méthode de la «programmation convexe» apporte une solution optimale au problème posé, mais à un coût qui peut sembler exorbitant, de sorte qu'il sera nécessaire de chercher une solution optimale qui tienne compte des considérations budgétaires.

Dans cet article, nous allons définir une expression en forme analytique pour la répartition optimale en fonction des multiplicateurs de Lagrange; cette expression est soumise à des contraintes linéaires (sous forme d'inéquations). Cela va nous permettre de faire deux constatations

[1] James Bethel, Westat, 1650 Research Boulevard, Rockville, MD. 20850 USA.

qui compensent largement les inconvénients de la méthode de la «programmation convexe». La première constatation est qu'une réduction proportionnelle de la répartition optimale (dans les enquêtes à plusieurs variables) donne une répartition qui est optimale suivant des contraintes qui sont proportionnelles aux contraintes initiales. En conséquence, si la solution optimale est trop coûteuse, elle peut être ramenée directement à un niveau plus conforme au montant budgété et on peut évaluer directement les effets de cette opération sur la précision des estimations échantillonnales. La seconde constatation est que l'on obtient une expression simple pour les dérivées partielles du coût de la répartition de l'échantillon par rapport aux contraintes de variance. Ces quantités, que l'on appelle prix fictifs, illustrent la sensibilité du coût aux contraintes de variance et sont utiles pour l'évaluation du rapport coût/efficacité du plan de sondage.

Toutefois, l'exécution même de l'optimisation convexe demeure un problème. Beaucoup d'articles ont été écrits sur des méthodes permettant de résoudre les problèmes de programmation de ce genre et il existe de nombreux logiciels conçus à cette fin. Néanmoins, nous allons faire ici des considérations spéciales à ce propos et nous allons exposer une méthode de résolution simple. Cet algorithme, qui est essentiellement une méthode du gradient, est convergent, facile à programmer, et facile à utiliser puisqu'il ne requiert pas de valeur initiale. Un exemple nous permettra d'exposer cet algorithme et les autres méthodes mentionnées ci-dessus.

2. LE MODÈLE DE RÉPARTITION

Considérons le cas d'un échantillonnage aléatoire stratifié avec I strates et J variables. Supposons que la j-ième variable doit satisfaire l'équation

$$\mathrm{Var}(\bar{y}_j) \approx \sum_{i=1}^{I} W_i^2 S_{ij}^2 / n_i \leq v_j^2, \tag{1}$$

où S_{ij}^2, n_i, et W_i^2, désignent respectivement la variance de la j-ième variable de réponse, la répartition de la taille de l'échantillon et la proportion de la population qui se trouve dans la i-ième strate, et où i-ième strate, et où v_j est une constante positive arbitraire. Nous supposerons dans cet article que les facteurs de correction pour population finie sont négligeables. En pratique, les effets de cette hypothèse, dont nous discuterons plus longuement dans la section 7, devraient être limités.

Soit

$$\begin{aligned} x_i &= 1/n_i \text{ si } n_i \geq 1 \\ &= \infty \quad \text{dans le cas contraire} \end{aligned}$$

et supposons la fonction de coût

$$g(x) = \sum_{i=1}^{I} c_i / x_i, \quad c_i > 0, \, i = 1, 2, \ldots, I. \tag{2}$$

Nous pourrions inclure dans cette expression un terme constant pour les coûts fixes, mais cela ne changerait rien au processus de minimisation; par conséquent, nous omettrons ce terme afin de simplifier la notation. Définissons les constantes

$$a_{ij} = w_i^2 S_{ij}^2 / v_j^2 \tag{3}$$

que nous appellerons «unités de précision normalisées». Notons que $a_{ij} \geq 0$. En nous servant de cette notation, nous pouvons exprimer le problème de la répartition optimale de la façon suivante:

$$\begin{array}{lll} \text{Minimiser} & g(x) & \\ \text{à la condition que} & a_j' x \leq 1, \quad j = 1, 2, \ldots, J & \\ & x > 0 & \end{array} \tag{4}$$

où a_j est le j-ième vecteur colonne de la matrice $A = \{a_{ij}\}$.

Kokan (1963) analyse en profondeur ce modèle de répartition et montre comment on peut l'adapter à de nombreux problèmes courants de répartition de l'échantillon, notamment l'échantillonnage en grappes et l'échantillonnage double. Kokan et Khan (1967) poussent plus loin l'analyse dans ce contexte; Arthanari et Dodge (1981) exposent à nouveau les résultats de Kokan et Khan. Dans la même ligne de pensée, Kish (1976) décrit une catégorie de «formes linéaires» qui reviennent souvent dans la conception d'enquêtes et auxquelles s'appliqueront une bonne partie des résultats obtenus ici.

3. RÉPARTITION OPTIMALE

Le modèle de répartition optimale pour une variable unique est bien connu. Dans ce cas $J = 1$, et le minimum de $g(x)$ pourvu que $a_i' x \leq 1$ où $x > 0$, désigné par x^*, est défini

$$\begin{aligned} x_i^* &= \sqrt{c_i} \Big/ \left(\sqrt{a_{i1}} \sum_{k=1}^{I} \sqrt{c_k a_{k1}} \right) && \text{si } a_{i1} > 0,\ 1 \leq i \leq I \\ &= \infty && \text{dans le cas contraire.} \end{aligned} \tag{5}$$

Dans cette section, nous allons généraliser la formule (5) pour les cas où $J > 1$.

La fonction g définie en (2) est strictement convexe pour $x > 0$, et les contraintes indiquées en (4) sont linéaires, de sorte que les résultats de base de la programmation convexe s'appliquent ici sans problème. Kokan et Khan (1967) ont démontré qu'il existe toujours une solution optimale. Comme ci-dessus, désignons la solution optimale par x^*. Alors, d'après le théorème de Kuhn-Tucker (1951), il existe des $\lambda_j \geq 0$ de telle sorte que

$$\nabla g(x^*) + \sum_{j=1}^{J} \lambda_j a_j = 0 \tag{6}$$

(∇ désigne le gradient) et

$$\lambda_j \left(a_j' x^* - 1 \right) = 0 \tag{7}$$

pour $j = 1, 2, \ldots, J$. Si $x > 0$ satisfait $\sum_{j=1}^{J} \lambda_j a_j' x \leq \sum_{j=1}^{J} \lambda_j$, alors en combinant (6) et (7), nous avons

$$-x' \nabla g(x^*) = \sum_{j=1}^{J} \lambda_j a_j' x \leq \sum_{j=1}^{J} \lambda_j = \sum_{j=1}^{J} \lambda_j a_j' x^* = -x^{*\prime} \nabla g(x^*). \tag{8}$$

Par convexité, $g(x) - g(x^*) \geq (x - x^*)' \nabla g(x^*)$ (pour tous $x > 0$, $x^* > 0$). Alors, d'après (8)

$$g(x) - g(x^*) \geq (x - x^*)' \nabla g(x^*) \geq 0.$$

Il s'ensuit que x^* est le minimum de $g(x)$ à la condition que

$$\sum_{j=1}^{J} \lambda_j a_j' x \leq \sum_{j=1}^{J} \lambda_j \text{ pour tous } x > 0.$$

Puisque la minimisation de g n'est pas influencée par des constantes multiplicatives positives, x^* minimise aussi $g(x)$ pourvu que $\sum_{j=1}^{J} \alpha_j^* a_j' x \leq 1$ et $x > 0$, où $\alpha_j^* = \lambda_j / \sum_{j=1}^{J} \lambda_j$.

Si nous voulons maintenant utiliser la formule (5) pour la répartition optimale dans les enquêtes à plusieurs variables, il suffit de l'appliquer à la somme pondérée $\sum_{j=1}^{J} \alpha_j^* a_j$:

$$x_i^* = \sqrt{c_i} \Big/ \left(\sqrt{\sum_{j=1}^{J} \alpha_j^* a_{ij}} \sum_{k=1}^{I} \sqrt{c_k \sum_{j=1}^{J} \alpha_j^* a_{ij}} \right) \quad \text{si } \sum_{j=1}^{J} \alpha_j^* a_{ij} > 0,\ 1 \leq i \leq I$$

$$= \infty \quad \text{dans le cas contraire.} \tag{9}$$

Donc, comme x^* minimise $g(x)$ à la condition que $a_j' x \leq 1, x > 0$ pour $1 \leq j \leq J$, mx^* minimisera $g(mx)$ pourvu que $a_j'(mx) \leq m$, $x > 0$ pour $1 \leq j \leq J$. Ainsi, comme nous l'avons indiqué plus haut, les contraintes de variance (ou CV) peuvent être rajustées par un facteur m (ou $\sqrt{m}$) si les coûts de l'enquête sont trop élevés.

Évidemment, la formule (9) n'est mathématiquement utile que si les α_j^* sont connus. Néanmoins, elle est utile pour calculer les prix fictifs et mettre au point un algorithme permettant de déterminer x^* et les α_j^*.

4. SENSIBILITÉ DU COÛT DE L'ENQUÊTE AUX CONTRAINTES DE VARIANCE

Dans beaucoup de problèmes d'optimisation, il est utile de savoir ce que devient la solution optimale lorsque les contraintes sont modifiées légèrement. Cela peut être particulièrement le cas dans la conception d'enquêtes, où il faut souvent trouver le juste équilibre entre les coûts, les opérations d'enquête et le niveau de précision recherché. Quoi qu'il en soit,les «prix fictifs», définis par $\partial g(x^*)/\partial v_k$, sont utiles pour discerner les faibles variations des contraintes de variance, qui pourraient réduire sensiblement le coût global de l'enquête.

En combinant les équations (2), (3), et (9), nous constatons facilement que le coût de la répartition optimale est

$$g(x^*) = \left(\sum_{i=1}^{I} \sqrt{c_i \sum_{j=1}^{J} \alpha_j^* a_{ij}} \right)^2 = \left(\sum_{i=1}^{I} \sqrt{c_i \sum_{j=1}^{J} \alpha_j^* W_i^2 S_{ij}^2 / v_j^2} \right)^2. \tag{10}$$

Par conséquent

$$\frac{\partial g(x^*)}{\partial v_k} = 2\left(\sum_{i=1}^{I}\sqrt{c_i\sum_{j=1}^{J}\alpha_j^* W_i^2 S_{ij}^2/v_j^2}\right)\sum_{i=1}^{I}\frac{-c_i\alpha_k^* W_i^2 S_{ik}^2/v_k^3}{\sqrt{c_i\sum_{j=1}^{J}\alpha_j^* W_i^2 S_{ij}^2/v_j^2}} \tag{11}$$

$$= -2\frac{\alpha_k^*}{v_k}\sqrt{g(x^*)}\sum_{i=1}^{I}\frac{a_{ik}\sqrt{c_i}}{\sqrt{\sum_{j=1}^{J}\alpha_j^* a_{ij}}}$$

$$= -2\frac{\alpha_k^*}{v_k}g(x^*)\sum_{i=1}^{I}a_{ik}\sqrt{c_i}\Big/\left(\sqrt{\sum_{j=1}^{J}\alpha_{ij}^* a_{ij}}\sum_{k=1}^{I}\sqrt{c_k\sum_{j=1}^{J}\alpha_j^* a_{kj}}\right)$$

$$= -2\frac{\alpha_k^*}{v_k}g(x^*)\ a_k' x^*.$$

L'équation (7) implique nécessairement que $\alpha_k^* = 0$ lorsque $a_k' x^* < 1$, de sorte que

$$\frac{\partial g(x^*)}{\partial v_k} = -2\frac{\alpha_k^*}{v_k}g(x^*). \tag{12}$$

Cette formule est un peu plus compliquée que l'expression utilisée habituellement pour les prix fictifs (voir, par exemple, Luenberger 1984) à cause de la relation complexe entre g et v_j.

Supposons maintenant que nous augmentons v_k de (100π) %, $0 \leq \pi \leq 1$. Nous allons représenter par $x^* + \Delta x^*$ la perturbation résultante dans x^*. Selon (12),

$$g(x^* + \Delta x^*) - g(x^*) \approx \pi v_k\frac{\partial g(x^*)}{\partial v_k} = -2\,\pi\,\alpha_k^* g(x^*). \tag{13}$$

Par conséquent, une augmentation de $(100\pi)\%$ de la k-ième contrainte de variance entraînera une réduction de $(100)(2\pi\alpha_k^*)\%$ du coût global de l'enquête.

5. CONSIDÉRATIONS RELATIVES À LA PROGRAMMATION

Dans cette section, nous examinons certains aspects techniques du calcul de x^* et présentons un algorithme simple pour déterminer x^* et les coefficients α_j^* en examinant les moyennes pondérées $\sum_{j=1}^{j}\alpha_j a_j$. Définissons δ_{ij}

$$\delta_{ij} = 1 \text{ if } i = j$$

$$= 0 \text{ if } i \neq j.$$

Pour un vecteur $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_J)'$, définissons $\tilde{x}(\alpha)$ comme suit

$$\tilde{x}_i(\alpha) = \sqrt{c_i} \Big/ \left(\sqrt{\sum_{j=1}^{J} \alpha_j a_{ij}} \sum_{k=1}^{I} \sqrt{c_k \sum_{j=1}^{J} \alpha_j a_{kj}} \right) \text{ si } \sum_{j=1}^{J} \alpha_j a_{ij} > 0,\ 1 \leq i \leq I$$

$$= \infty \qquad \text{dans le cas contraire.}$$

Notons que $\tilde{x}(\alpha^*) = x^*$. Or, l'algorithme itératif permettant de déterminer x^* est défini de la façon suivante:

1. Supposer $\alpha_j^{(1)} = \delta_{1j}$, $1 \leq j \leq J$.

2. À l'itération $n \geq 2$, trouver un indice k pour lequel

$$(a_k - a_j)' \tilde{x}(\alpha^{(n)}) \geq 0,\ 1 \leq j \leq J. \tag{14}$$

L'expression ci-dessus représente la contrainte que la solution optimale courante satisfait le plus difficilement. Si $a_k' \tilde{x}(\alpha^{(n)}) \leq 1$, on doit mettre fin à l'algorithme, sinon on doit trouver une valeur $t^{(n)} \in (0,1)$ pour laquelle

$$g(\tilde{x}(t^{(n)} \delta_k + (1 - t^{(n)})\alpha^{(n)}) \geq g(\tilde{x}(t\,\delta_k + (1 - t)\,\alpha^{(n)})) \text{ pour tous } t \in [0, 1]. \tag{15}$$

3. Supposer $\alpha_j^{(n+1)} = t^{(n)} \delta_{kj} + (1 - t^{(n)})\, \alpha_j^{(n)}$.

4. Cesser l'itération lorsque $|\alpha_j^{(n+1)} - \alpha_j^{(n)}| < \epsilon$, $1 \leq j \leq J$, où ϵ est un critère de convergence préétabli.

Afin de vérifier la convergence de l'algorithme, il convient tout d'abord de noter que $\tilde{x}(\alpha)$ minimise $g(x)$ à la condition que $\sum_{j=1}^{J} \alpha_j a_j' x \leq 1$. Par conséquent, comme $\sum_{j=1}^{J} \alpha_j a_j' x^* \leq \sum_{j=1}^{J} \alpha_j = 1$,

$$0 \leq g(\tilde{x}(\alpha^{(n)})) \leq g(x^*) \tag{16}$$

pour tous n. Par ailleurs, d'après (15), $g(\tilde{x}(\alpha^{(n)}))$ est non décroissante, ce $g(\tilde{x}(\alpha^{(n)}))$ qui implique qu'elle est convergente. Afin de montrer que $\tilde{x}(\alpha^{(n)}) \to x^*$, définissons tout d'abord

$$h_{k\alpha}(t) = \sum_{i=1}^{I} \sqrt{c_i \sum_{j=1}^{J} (t\delta_{kj} + (1-t)\alpha_j) a_{ij}} = \sqrt{g(\tilde{x}(t\delta_k + (1 - t)\alpha))}. \tag{17}$$

Puisque $h_{k\alpha}(t)$ est concave (c.-à-d., $-h_{k\alpha}(t)$ est convexe),

$$h_{k\alpha}(t) - h_{k\alpha}(0) = t\, h'(0) + O(t^2) \tag{18}$$

$$= -t \sum_{i=1}^{I} \frac{\sum_{j=1}^{J} (\delta_{kj} - \alpha_j)\, a_{kj}\sqrt{c_k}}{2\sqrt{c_i \sum_{j=1}^{J} \alpha_j\, a_{ij}}} + O(t^2)$$

$$= (t/2)\sqrt{g(\tilde{x}(\alpha))}\,(a_k'\tilde{x}(\alpha) - 1) + O(t^2).$$

Si nous acceptons que t tende vers zéro, alors il existe une valeur $t \in (0,1)$ pour laquelle

$$\sqrt{g(\tilde{x}(t\delta_k + (1-t)\alpha))} = h_{k\alpha}(t) > h_{k\alpha}(0) = \sqrt{g(\tilde{x}(\alpha))}$$

si et seulement si $a_k'\tilde{x}(\alpha) > 1$. Nous pouvons donc conclure, d'après (15), que les contraintes sont satisfaites à la convergence; cette conclusion et l'expression (16) impliquent que $\lim_{n\to\infty} \tilde{x}(\alpha^{(n)}) = x^*$.

Dans l'exécution de l'algorithme, l'étape 2 nécessite la recherche d'une valeur $t^{(n)}$. Définissons $h_{k\alpha}(t)$ comme dans l'équation (17). Il est clair d'après l'analyse précédente que $a_k'\tilde{x}(t\delta_k + (1-t)\alpha^{(n)}) = 1$ lorsque $h(t)$ (et par conséquent g) est à un maximum. De plus, comme $h_{k\alpha}(t)$ est strictement concave, $h_{k\alpha}'(t)$ est non croissance en t et il y a donc un seul point où $h_{k\alpha}'(t) = 0$. On peut donc exécuter une recherche dichotomique pour le point où est maximisée $h_{k\alpha}(t)$ en vérifiant simplement si $a_k'\tilde{x}(t\delta_k + (1-t)\alpha^{(n)}) = 1$, ce qui est un moyen rapide d'obtenir une valeur approchée de $t^{(n)}$.

Nous avons vu plus haut que l'algorithme pose a_1 comme valeur initiale. Cela est tout à fait arbitraire puisque n'importe lequel des a_j, $1 \le j \le J$, conviendrait. En pratique, la contrainte pour laquelle la répartition optimale (c.-à-d. formule (5)) entraîne le coût le plus élevé est généralement un choix judicieux pour la valeur initiale.

Il convient de souligner que la seconde étape de l'algorithme nécessitera IJ calculs pour la formule (14) ainsi qu'une recherche dichotomique en 10 étapes par exemple, chacune d'elles comportant $3I + J + 1$ calculs pour la formule (15), tandis que la troisième étape nécessitera J calculs. Ainsi,chaque itération de l'algorithme est $O(IJ)$. D'après (18), à la n-ième itération $O(IJ)$.

$$h_{k\alpha}'(0) \approx \frac{1}{2} h_{k\alpha}(0)\ (a_k'\tilde{x}(\alpha^{(n)}) - 1)$$

de sorte que $a_k'\tilde{x}(\alpha^{(n)})$ est approximativement proportionnelle à $h_{k\alpha}'(0)$ (jusqu'à une constante additive) Au point de vue heuristique $h_{k\alpha}'(0)$ est la «pente» de h dans la direction de a_k, ce qui donne à penser que l'algorithme est essentiellement une méthode du gradient (ou une méthode de la plus grande ascension en l'occurrence). Cela suppose, en retour, un taux de convergence linéaire (voir, par exemple, Forsyth 1968).

D'après notre expérience (voir Bethel 1985), l'algorithme converge rapidement pour la plupart des problèmes de complexité moyenne. Par exemple, des problèmes de répartition d'échantillon où il y a de 20 à 30 strates et de 5 à 10 contraintes ont été résolus en 3 à 5 secondes à l'aide de l'algorithme (sur un ordinateur Compaq 386/20 avec un co-processeur arithmétique 30387) alors qu'ils l'ont été en 6 à 8 secondes à l'aide d'une technique séquentielle de minimisation sans contrainte (sequential unconstrained minimization technique - SUMT), qui comprend un algorithme avec peine fondé sur la méthode du gradient. Le temps d'exécution varie

considérablement selon la complexité du problème, le nombre de contraintes agissantes et, bien sûr, les caractéristiques du matériel. Les divers systèmes utilisés dans les circonstances (20 à 30 strates et 5 à 10 contraintes) ont été le Macintosh SE (temps d'exécution: de 30 secondes à 2 ou 3 minutes), le Leading Edge Model D (de 1 à 5 minutes), le Zilog System 8000 (de 5 à 60 secondes) et le Compaq précité (de 5 à 10 secondes). Néanmoins, les temps d'exécution sont généralement négligeables par rapport au temps nécessaire pour créer des fichiers et exécuter d'autres tâches préliminaires. L'algorithme SUMT, par exemple, peut prendre plusieurs heures pour trouver une valeur initiale acceptable. Un avantage notable de l'algorithme décrit ci-dessus en quatre étapes est qu'il ne nécessite pas de valeurs initiales externes. En outre, il est relativement facile à programmer, n'exigeant que 40 ou 50 lignes de programmation.

Chromy (1987) propose un algorithme encore plus simple. Nous pouvons adapter cet algorithme à notre notation et à notre approche générale de la façon suivante: Posons $\alpha_j^{(1)} \equiv 1/J$,et, pour $n \geq 2$, Soit

$$\alpha_j^{(n)} = \alpha_j^{(n-1)}(a_j'\tilde{x}(\alpha^{(n-1)}))^2 / \sum_{j=1}^{J} \alpha_j^{(n-1)}(a_j'\tilde{x}(\alpha^{(n-1)}))^2 \quad 1 \leq j \leq J. \tag{19}$$

Comme l'algorithme décrit précédemment en quatre étapes, l'expression (19) ne nécessite pas de valeurs initiales externes; en outre, cet algorithme est encore plus facile à programmer et, selon plusieurs comparaisons, semble converger beaucoup plus rapidement. Malheureusement, il semblerait que cette convergence ne puisse se vérifier de façon formelle mais un nombre considérable d'applications (voir Chromy 1987, pour une analyse plus détaillée) permettent de croire que cet algorithme possède des propriétés de convergence intéressantes.

6. EXEMPLE

Les tableaux 1 à 3 présentent un exemple tiré d'une enquête sur les établissements commerciaux. (Seules les strates relatives aux établissements d'enseignement sont présentées ici.) Ces tableaux contiennent quatre des principales variables d'intérêt: surface utile, âge du bâtiment, nombre d'employés à plein temps et proportion (en pourcentage) dans laquelle le bâtiment est chauffé au mazout. Le tableau 1 renferme les données relatives à la variance de strate. Les unités de précision normalisées sont calculées en l'occurrence par la formule

$$a_{ij} = \frac{W_i^2 S_{ij}^2}{\bar{Y}_j^2 v_j^2}$$

où $v_j = .06$ pour toutes les variables (de sorte que la demi-largeur d'un intervalle de confiance à 90 % équivaille à environ 10 % de la moyenne). Le tableau 1 donne également la taille d'échantillon requise pour une répartition de Neyman pour chacune des variables. On suppose que les coûts d'enquête sont les mêmes d'une strate à l'autre.

Le tableau 2 donne la solution pour un premier passage en machine; cette solution nécessite un échantillon de 241 unités. Dans le même tableau on trouve les coefficients de Lagrange normalisés et les niveaux de précision atteints; ces chiffres indiquent clairement que la surface utile et l'âge du bâtiment sont les variables prédominantes tandis que les autres variables ne sont pas «agissantes». Dans ce cas, on a utilisé la valeur initiale $\alpha^{(1)} = (1,0,0,0)$; comme les troisième et quatrième contraintes étaient toujours satisfaites, il n'y a eu qu'une itération avec une recherche dichotomique en 9 étapes pour $t^{(1)}$. (Les estimations successives pour la valeur t optimale ont été 1/2, 1/4, 3/8, 5/16, 11/32, 21/64, 43/128, 85/256 et 171/512.) Le tableau 2

Tableau 1.

Exemple de répartition de l'échantillon: enquête sur les établissements d'enseignement.

	Écart type de strate				
	Poids	Surface utile	Âge du bâtiment	Nombre d'employés	Proportion du bâtiment (%) chauffé au mazout
Strate					
1	.5158	22,319.11	43.71	25.72	48.15
2	.2632	24,056.21	16.68	27.09	36.79
3	.1184	54,201.75	24.70	17.11	48.04
4	.0711	155,514.21	16.01	59.46	38.07
5	.0184	125,239.21	14.74	51.27	48.80
6	.0132	355,392.69	20.90	212.13	57.74
Moyenne:		54,641.85	43.03	45.23	67.58
v_k:		.06	.06	.06	.06

	Unités de précision normalisées			
	Surface	Âge du bâtiment	Nombre d'employés	Proportion du bâtiment (%) chauffée au mazout
Strate				
1	12.33	76.24	23.90	37.52
2	3.73	2.89	6.93	5.70
3	3.83	1.28	.56	1.96
4	11.36	.19	2.44	.45
5	7.37	.01	.12	.05
6	2.03	.01	1.06	.04
Taille d'échantillon requise:	222	149	127	121

indique également les prix fictifs à 10 %: une hausse de 10 % de la première (ou de la seconde) contrainte entraînera une réduction de la taille de l'échantillon d'environ 32 (ou 16) unités. Comme les troisième et quatrième contraintes ne sont pas «agissantes» dans la solution, une modification des CV correspondantes n'aura aucun effet sur la répartition ou les coûts de l'échantillonnage.

Le tableau 3 donne une solution pour un second passage en machine moyennant que la taille de l'échantillon ne dépasse pas 200. Les solutions optimales sont donc multipliées par 241/200 (de sorte que chaque élément de la répartition optimale soit réduit par un facteur de 200/241) et les CV sont multipliées par $\sqrt{241/200}$. Les nouveaux prix fictifs à 10 % sont -27 et -13 pour les première et seconde contraintes respectivement, reflétant ainsi la diminution du coût global de l'enquête. Il convient de noter que les CV du tableau 3 sont environ 10 % plus élevées que celles du tableau 1, de sorte quela réduction de la taille de l'échantillon annoncée par les prix fictifs du tableau 2 (soit 48) se compare assez bien à la réduction réelle, qui est de 41 unités. (Les prix fictifs seront toujours quelque peu optimistes à cause de l'approximation linéaire).

Tableau 2.

Exemple de répartition de l'échantillon: solution optimale pour un premier passage en machine.

Strate	$\sum \alpha_j^* a_{ji}$	x_i^*	Répartition optimale
1	33.6749	.0111	90
2	3.4495	.0347	29
3	2.9783	.0373	27
4	7.6294	.0233	43
5	4.9119	.0291	34
6	1.3554	.0553	18
Total:			241

	Surface utile	Âge du bâtiment	Nombre d'employés	Proportion du bâtiment chauffée au mazout
Multiplicateur de Lagrange (normalisé):	.6660	.3340	.0000	.0000
Niveau de précision atteint:	.0600	.0600	.0481	.0502
Prix fictifs à 10 %:	-32	-16	0	0

Tableau 3.

Exemple de répartition de l'échantillon: solution optimale pour un échantillon dont la taille ne doit pas excéder 200.

Strate	$\sum \alpha_j^* a_{ji}$	x_i^*	Répartition optimale
1	33.6749	.0134	75
2	3.4495	.0418	24
3	2.9783	.0449	22
4	7.6294	.0281	36
5	4.9119	.0351	29
6	1.3554	.0666	15
Total:			201

	Surface utile	Âge du bâtiment	Nombre d'employés	Proportion du bâtiment chauffée au mazout
Multiplicateur de Lagrange (normalisé):	.6660	.3340	.0000	.0000
Niveau de précision atteint:	.0657	.0658	.0528	.0551
Prix fictifs à 10 %:	-27	-13	0	0

7. ANALYSE

Cet article nous a permis d'exposer de façon formelle la répartition optimale de l'échantillon dans les enquêtes à objectifs multiples avec des contraintes de variance linéaires et de définir des expressions pour les dérivées partielles de la fonction de coût par rapport aux contraintes de précision. Ces expressions, notamment, produisent des approximations utiles pour la planification d'enquêtes, permettant ainsi d'exécuter de nombreux travaux préliminaires sans les calculs rigoureux faits par ordinateur.

Les multiplicateurs de Lagrange normalisés α_j^*, occupent une place importante dans cet article. Nous avons remarqué, en particulier, que lorsque la j-ième contrainte de variance n'est pas «agissante» dans la solution du problème de répartition, le j-ième multiplicateur de Lagrange $\alpha_j^* = 0$.

La méthode d'optimisation analysée dans cet article produit une solution continue qu'il faut arrondir d'une manière quelconque si l'on veut obtenir des tailles de strates en nombres entiers. De toute évidence, cette opération d'arrondissement créera un certain écart par rapport à l'optimalité. Toutefois, on s'entend normalement pour dire que la fonction objectif considérée ici est plutôt insensible à de faibles entorses à l'optimalité (voir Cochran 1977), de sorte que les solutions en nombres entiers sont probablement rentables. De fait, les erreurs d'arrondissement seraient vraisemblablement négligeables par rapport aux erreurs d'échantillonnage qui touchent les estimations de moyennes et de variances dont on dispose normalement pour élaborer un plan de sondage optimal.

Enfin, rappelons-nous que cet article ne tenait aucunement compte des facteurs de correction pour population finie. Nous pourrions facilement les intégrer au modèle de répartition en arrangeant les équations (1) et (3), mais cela rendrait l'équation (13) quelque peu imprécise. Néanmoins, il convient de se rappeler que même lorsque le facteur de correction pour population finie n'est pas négligeable pour quelques-unes des strates, l'effet global est, habituellement négligeable. Quoi qu'il en soit, il est toujours possible de calculer le terme perturbateur $\sum_{i=1}^{I} W_i^2 S_{ij}^2 / N_i$, pour une évaluation de la situation et de l'ajouter, si nécessaire, à v_k dans la formule (13) pour obtenir des résultats exacts.

REMERCIEMENTS

L'auteur a apprécié les commentaires des arbitres et les échanges fructueux qu'il a eus avec Rick Williams du Research Triangle Institute et Patrick McCarthy d'Applied Management Sciences; ces personnes ont contribué à améliorer sensiblement le contenu de cet article.

BIBLIOGRAPHIE

ARTHANARI, T.S., et DODGE, Y., (1981). *Mathematical Programming in Statistics*. New York: John Wiley and Sons.

BETHEL, J.W. (1985). An optimum allocation algorithm for multivariate surveys. *Proceedings of the Survey Research Section, American Statistical Association*, 209-212.

CHATTERJEE, S. (1968). Multivariate stratified surveys. *Journal of the American Statistical Association*, 63, 530-534.

CHATTERJEE, S. (1972). A study of optimum allocation in multivariate stratified surveys. *Skandinavisk Actuarietidskrift*, 55, 73-80.

CHROMY, J.R. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Section, American Statistical Association*, 194-199.

DALENIUS, T. (1953). The multivariate sampling problem. *Skandinavisk Actuarietidskrift*, 36, 92-102.

DALENIUS, T. (1957). *Sampling in Sweden*. Stockholm: Almquist and Wicksall.

FOLKS, J.L., et ANTLE, C.E. (1965). Optimum allocation of sampling units when there are R responses of interest. *Journal of the American Statistical Association*, 60, 225-233.

FORSYTH, G.E. (1968). On the asymptotic directions of the s-dimensional optimum gradient method. *Numerische Mathematik*, 11, 57-76.

HARTLEY, H.O. (1965). Multiple purpose optimum allocation in stratified sampling. *Proceedings of the Social Statistics Section, American Statistical Association*, 258-261.

HUDDLESTON, H.F., CLAYPOOL, P.L., et HOCKING, R.R. (1970). Optimum sample allocation to strata using convex programming. *Applied Statistics*, 19, 273-278.

KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society A.*, 139, 80-95.

KOKAN, A.R. (1963). Optimum allocation in multivariate surveys. *Journal of the Royal Statistical Society A.*, 126, 557-565.

KOKAN, A.R., et KHAN, S. (1967). Optimum allocation in multivariate surveys: an analytical solution. *Journal of the Royal Statistical Society B.*, 29, 115-125.

KUHN, H.W., et TUCKER, A.W. (1951). Nonlinear programming. *Proceedings 2nd Berkeley Symposium Mathematical Statistics and Probability*.

LUENBERGER, D.G. (1984). *Linear and Nonlinear Programming*. Reading, Massachusets: Addison-Wesley.

NEYMAN, J. (1934). On the two different aspects of the representative method: the method of representative sampling and the method of purposive sampling. *Journal of the Royal Statistical Society*, 558-625.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*. London: Charles Griffin and Company.