Techniques d'enquête, décembre 2002
Vol. 28, N° 2, pp. 215-224
Statistique Canada, N° 12-001-XPB au catalogue

Méthode pratique de stratification multiple par programmation linéaire

Wilson Lu et Randy R. Sitter [1]

Résumé

Sitter et Skinner (1994) présentent une méthode qui consiste à appliquer la programmation linéaire à la conception d'enquêtes avec stratification multiple, principalement dans des situations où la taille souhaitée de l'échantillon est inférieure ou à peine supérieure au nombre total de cellules de stratification. Leur méthode repose sur une idée simple, facile à comprendre et à appliquer. Cependant, en pratique, elle a le désavantage de devenir rapidement coûteuse en raison de l'importance des calculs, à mesure qu'augmente le nombre de cellules de la stratification multiple, au point de ne pouvoir être utilisée dans la plupart des situations réelles. Dans le présent article, nous étendons cette méthode de programmation linéaire et élaborons des méthodes en vue de réduire le nombre de calculs, de sorte qu'il soit possible de résoudre des problèmes de grande taille.

Mots clés : Échantillonnage PPT; répartition proportionnelle; groupement aléatoire; échantillonnage.

1. Introduction

En pratique, dans le cas de nombreuses enquêtes, il existe plusieurs variables de stratification et le concepteur de l'enquête peut donc choisir de définir des strates sous forme de cellules obtenues par recoupement des catégories de ces variables. À cet égard, consulter, par exemple, Engle, Marden et Pollock (1971), Hess, Riedel et Fitzpatrick (1976), Vihma (1981) et Skinner, Holmes et Holt (1994). Cette stratification multiple donne souvent lieu à des situations où la taille souhaitée de l'échantillon est inférieure ou à peine supérieure au nombre total de cellules de stratification (phénomène particulièrement courant lors du choix des unités primaires d'échantillonnage (upe) dans le cas de plans de sondage stratifiés à plusieurs degrés), ce qui empêche parfois d'appliquer les méthodes classiques de répartition de l'échantillon entre les strates.

Nous présentons au tableau 1 une illustration fondée sur un exemple hypothétique de Bryant, Hartley et Jessen (1960). Les collectivités (upe) sont classées en fonction de deux variables de stratification, le type et la région, comportant trois et cinq catégories, respectivement. La taille souhaitée de l'échantillon, $n = 10$, est inférieure au nombre total de cellules, c'est-à-dire 15. Cet exemple illustre aussi un problème connexe. Dans le tableau 1, les entrées sont les dénombrements prévus en cas de stratification proportionnelle, c'est-à-dire quand les tailles d'échantillon de strate sont proportionnelles aux tailles de population de strate. En général, à cause des contraintes de taille d'échantillon, les nombres prévus d'unités d'échantillonnage dans les cellules ne sont pas entiers. Si les nombres prévus sont très faibles, leur arrondissement à l'unité près ne produira pas de bons choix et enfreindra sérieusement la contrainte de répartition proportionnelle. Il est également courant que les totaux marginaux ne soient pas entiers, ce qui peut aussi causer des difficultés. Goodman et Kish (1950) ont été les premiers à essayer de résoudre ce problème selon une démarche qu'ils ont appelée sélection contrôlée, par une méthode d'échantillonnage rentrant dans la catégorie des méthode d'échantillonnage systématiques aléatoires (voir Hess, Riedel et Fitzpatrick 1961; Waterton 1983). Bryant et coll. (1960) ont présenté une méthode fort simple d'attribution aléatoire de la taille d'échantillon à chaque cellule en cas de stratification double et ont donné deux estimateurs fondés sur ce scénario d'échantillonnage. Toutefois, puisque les tailles prévues d'échantillon de cellule ne tenaient pas compte de la représentation proportionnelle de chaque cellule (autrement dit, la méthode n'est pas une méthode de sélection contrôlée convenable, puisque seules les probabilités des distributions marginales sont respectées), les propriétés de l'EQM de ces estimateurs pourraient ne pas être satisfaisantes (consulter Sitter et Skinner 1994). Selon Jessen (1970), une autre limite de la méthode de Bryant et coll. (1960) tient à ce qu'il est impossible de forcer la taille de cellules spécifiques à être nulle, alors que cela pourrait être souhaitable dans certaines situations (voir les méthodes connexes à la rubrique « échantillonnage en treillis », par exemple Jessen 1973, 1975). Il propose deux méthodes de stratification double et triple, mais elles sont assez difficiles à appliquer et, comme le font remarquer Causey, Cox et Ernst (1985), pourraient ne pas donner de solution. En s'inspirant de l'idée exposée par Rao et Nigam (1990, 1992) afin d'éviter des échantillons indésirables (voir aussi Lahiri et Mukerjee 2000), Sitter et Skinner (1994) ont proposé une méthode de programmation linéaire qui vise à tirer parti de la puissance des méthodes modernes de calcul. Cette méthode de programmation linéaire est de conception

1. Wilson Lu, étudiant de doctorat, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6; Randy R. Sitter, professeur, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6.

simple, est adaptable à diverses situations, produit toujours une solution et donne une EQM dont les propriétés sont meilleures. Sa principale limite pratique est que les calculs deviennent ardus à mesure qu'augmente le nombre de cellules de la stratification multiple, et sont rapidement infaisables. Dans le présent article, nous présentons une méthode simple qui permet d'appliquer la méthode de programmation linéaire à la résolution de problèmes de taille nettement plus grande. À la section 2, nous décrivons la méthode de programmation linéaire de Sitter et Skinner (1994) pour introduire la notation et nous discutons brièvement de ses limites numériques. À la section 3.1, nous examinons d'abord certaines stratégies simples qui permettent de rendre les calculs moins ardus à titre de justification de la proposition éventuelle. Aux sections 3.2 et 3.3, nous discutons de la méthode proposée, en supposant que les totaux marginaux sont des nombres entiers et nous donnons certains exemples où le nombre de cellules de stratification varie de 80 à 300 pour illustrer la capacité qu'a la nouvelle méthode de traiter des problèmes de grande taille. À la section 3.4, nous décrivons l'extension simple de la méthode aux totaux marginaux non entiers et nous l'illustrons en l'appliquant à un exemple réel tiré de données publiées sur la santé au travail (Vihma 1981).

Tableau 1
Exemple tiré de Bryant et coll. (1960). Nombres prévus d'unités d'échantillonnage dans les cellules sous stratification proportionnelle (n = 10)

	Type de collectivité			
Région	Urbaine	Rurale	Métropolitaine	Total
1	1,0	0,5	0,5	2,0
2	0,2	0,3	0,5	1,0
3	0,2	0,6	1,2	2,0
4	0,6	1,8	0,6	3,0
5	1,0	0,8	0,2	2,0
Total	3,0	4,0	3,0	10,0

2. La méthode de programmation linéaire

2.1 Notions fondamentales

Nous considérons, pour introduire la méthode de programmation linéaire de Sitter et Skinner (1994), la forme la plus simple de stratification double. Supposons que N unités d'une population finie sont disposées, conformément à une classification à double entrée, en R lignes correspondant aux catégories de l'une des variables et en C colonnes correspondant aux catégories de l'autre variable. Représentons par N_{ij} le nombre d'unités de la population dans la i^{e} ligne et la j^{e} colonne (c'est-à-dire, dans la ij^{e} cellule) du tableau à double entrée et par $P_{ij} = N_{ij}/N$, la proportion de la population totale comprise dans la ij^{e} cellule. Soit $\bar{Y}$, la valeur moyenne d'une caractéristique étudiée y pour la population et $\bar{Y}_{ij}$, la valeur moyenne de y pour la ij^{e} cellule.

Nous sélectionnons l'échantillon comme suit.

i) Nous déterminons la taille d'échantillon n_{ij} aléatoirement pour chaque cellule, conformément à une méthode précisée *a priori*. Si nous représentons par s la matrice $R \times C$ $(n_{ij},\ i = 1,\ ...,\ R,\ j = 1,\ ...,\ C)$, nous attribuons par cette méthode une probabilité $p(s)$ à chaque s compris dans l'ensemble S de telles matrices possibles et nous sélectionnons une matrice unique, s, à partir de S. Pour représenter la dépendance de n_{ij} à l'égard de s, nous écrivons $n_{ij}(s)$.

ii) Puis, nous sélectionnons un échantillon aléatoire simple de $n_{ij}(s)$ unités dans la ij^{e} cellule et nous obtenons les valeurs de y.

Nous nous limitons aux plans d'échantillonnage à taille d'échantillon constante $n > 0$, autrement dit, à des matrices $s \in S_n$ telles que $\sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij}(s) = n$. Nous aimerions aussi nous limiter à la stratification proportionnelle de sorte que

$$\sum_{s \in S_n} n_{ij}(s)\ p(s) = nP_{ij} \text{ pour } i = 1,\ \ldots,\ R,\ j = 1,\ \ldots,\ C, \quad (1)$$

qui implique que la moyenne d'échantillon simple non pondérée $\bar{y}(s)$ est un estimateur non biaisé de $\bar{Y}$. Nous donnons à (1) le nom de contrainte de répartition proportionnelle prévue (RPP).

La méthode de programmation linéaire de Sitter et Skinner (1994) consiste à choisir un plan d'échantillonnage $p(s)$ qui minimise le manque prévu de « désirabilité » des échantillons par résolution du problème de programmation linéaire :

$$\min \sum_{s \in S_n} w(s)\ p(s) \quad (2)$$

subordonné à la contrainte (1), où $w(s)$ est une fonction de perte pour l'échantillon s, qu'il convient de spécifier, et où les probabilités $p(s)$ sont inconnues. Sitter et Skinner (1994) ont exploité l'observation principale de Rao et Nigam (1990, 1992) lors de travaux visant à éviter les échantillons indésirables, à savoir que, dans (2), la fonction objective est linéaire en probabilités $p(s)$ (voir aussi Lahiri et Mukerjee 2000).

Dans la fonction objective (2), la fonction de perte $w(s)$ joue un rôle important. Si $w(s)$ est bien définie, nous avons la souplesse de rechercher l'existence d'une solution optimale de (2) dans un ensemble S_n de taille économique et, par-dessus tout, de rendre l'estimation plus efficace. Sitter et Skinner (1994) proposent de choisir

$$w(s) = \sum_{i=1}^{R} \left(n_{i.}(s) - nP_{i.}\right)^2 + \sum_{j=1}^{C} \left(n_{.j}(s) - nP_{.j}\right)^2, \quad (3)$$

où $n_{i.}(s) = \sum_j n_{ij}(s), n_{.j}(s) = \sum_i n_{ij}(s), P_{i.} = \sum_j P_{ij}$ et $P_{.j} = \sum_i P_{ij}$. De toute évidence, la fonction objective (2) est effectivement $E(w(s))$ pour toute $p(s)$ déterminée par le plan de sondage et peut être interprétée comme étant

l'erreur quadratique moyenne de l'estimateur $\bar{y}$ sous un modèle d'analyse de la variance (voir Sitter et Skinner 1994). Alors, en résolvant le problème de programmation linéaire susmentionné, nous pouvons obtenir une EQM minimisée au sens de l'analyse de la variance tout en maintenant la propriété de répartition proportionnelle prévue (RPP) de $n_{ij}(s)$. Notons que, si nous obtenons un plan d'échantillonnage dont la fonction objective est nulle, toutes les contraintes de marge sont respectées. En général cette situation ne se présente que dans le cas où les totaux marginaux sont des nombres entiers.

Selon Sitter et Skinner (1994), un moyen simple de réduire la taille de S_n consiste à limiter les valeurs réelles que peut prendre n_{ij} à $\lfloor nP_{ij} \rfloor$ ou $\lfloor nP_{ij} \rfloor + 1$, où $\lfloor nP_{ij} \rfloor$ est le plus grand nombre entier inférieur ou égal à nP_{ij}. En écrivant $\tilde{n}_{ij} = n_{ij} - \lfloor nP_{ij} \rfloor$ et $r_{ij} = nP_{ij} - \lfloor nP_{ij} \rfloor$, nous pouvons alors imposer

$$E(\tilde{n}_{ij}) = r_{ij}, \tag{4}$$

où $\tilde{n}_{ij} = 0$ ou 1 et $0 \le r_{ij} < 1$. Alors, nous pouvons appliquer la méthode de programmation linéaire aux $\tilde{n}_{ij}$ et, finalement, nous pouvons utiliser $\lfloor nP_{ij} \rfloor + \tilde{n}_{ij}$ comme taille réelle des échantillons de cellule. Par conséquent, sans perte de généralité, nous supposerons que

$$n_{ij} = 0,1 \quad \text{et } 0 \le r_{ij} = nP_{ij} < 1. \tag{5}$$

2.2 Stratification plus profonde

La méthode de Sitter de Skinner (1994) s'étend facilement à un plus grand nombre de variables de stratification si l'on pose que *s* représente la matrice correspondant à r dimensions. La fonction de perte inclura alors un plus grand nombre de termes; par exemple, pour une stratification triple, l'équation (3) pourrait être remplacée par

$$w(s) = \gamma_1 \sum_{i=1}^{R_1} \left(n_{i..}(s) - nP_{i..}\right)^2 + \gamma_2 \sum_{j=1}^{R_2} \left(n_{.j.}(s) - nP_{.j.}\right)^2 + \gamma_3 \sum_{k=1}^{R_3} \left(n_{..k}(s) - nP_{..k}\right)^2$$

en notation évidente, où γ_1, γ_2 et γ_3 pourraient représenter l'importance relative de ces trois variables d'après des renseignements *a priori* (voir Sitter et Skinner 1994).

2.3 Échantillonnage à plusieurs degrés

Une application importante de la stratification multiple est la sélection d'unités primaires d'échantillonnage (upe) en cas d'échantillonnage à plusieurs degrés, où il est courant de disposer de plusieurs variables de stratification.

À la section 2.1, la probabilité d'inclusion de chaque unité est $E(n_{ij}(s)/N_{ij}) = n/N$. Si nous sélectionnons les upe avec probabilités égales, nous pouvons étendre directement la méthode en remplaçant les unités par les upe et les valeurs observées de *y*, par des estimateurs non biaisés des totaux des upe. Cependant, si nous devons sélectionner les upe avec probabilités inégales, disons nz_{ijk} pour l'upe *k* dans la cellule de stratification *ij* (z_{ijk} sera habituellement égale à $M_{ijk} / \sum_{ijk} M_{ijk}$, où M_{ijk} représente une mesure de taille de l'upe *k* dans la cellule *ij*), nous pouvons facilement modifier la méthode en posant que P_{ij} est égale à $z_{ij.}/z...$, où $z_{ij.} = \sum_k z_{ijk}$ et $z... = \sum_{ijk} z_{ijk}$. Alors, si $n_{ij}(s) > 0$, nous sélectionnons un échantillon de $n_{ij}(s)$ upe dans la cellule *ij* selon une méthode donnée d'échantillonnage avec probabilité proportionnelle à z_{ijk}.

2.4 Exemple

Nous pouvons illustrer la méthode de programmation linéaire au moyen de l'exemple hypothétique de Bryant et coll. (1960) donné au tableau 1. Premièrement nous simplifions le problème comme indiqué au tableau 2 afin de satisfaire l'hypothèse exprimée en (5). Puis, nous utilisons un progiciel type de programmation linéaire pour résoudre ce problème réduit (2). Puisque nous pouvons faire correspondre exactement les totaux marginaux entiers des nombres prévus d'unités d'échantillonnage dans les cellules aux totaux marginaux des tailles d'échantillon $n_{i.}$ et $n_{.j}$, ce qui signifie que la fonction de perte *w*(*s*) peut prendre une valeur minimale nulle, la fonction objective en (2) pour cet exemple est, elle aussi, minimale lorsque sa valeur est nulle. La solution optimale de ce problème est donnée au tableau 3. Notons que cette solution a été reconvertie de façon à ce qu'elle concorde avec l'exemple original présenté au tableau 1.

Tableau 2
Exemple modifié de Bryant et coll. (1960)

	Type de collectivité			
Région	Urbaine	Rurale	Métropolitaine	Total
1	0,0	0,5	0,5	1,0
2	0,2	0,3	0,5	1,0
3	0,2	0,6	0,2	1,0
4	0,6	0,8	0,6	2,0
5	0,0	0,8	0,2	1,0
Total	1,0	3,0	2,0	6,0

Tableau 3
Solution par programmation linéaire de l'exemple tiré de Bryant et coll. (1960)

s			*p*(*s*)	*s*			*p*(*s*)	*s*			*p*(*s*)
1	1	0		1	1	0		1	1	0	
1	0	0		0	0	1		0	0	1	
0	1	1	0,2	0	1	1	0,1	0	0	2	0,2
0	2	1		1	1	1		1	2	0	
1	0	1		1	1	0		1	1	0	
1	0	1		1	0	1		1	0	1	
0	1	0		0	1	0		0	0	1	
0	1	0	0,2	0	1	1	0,1	0	1	1	0,2
0	2	1		1	1	1		1	2	0	
1	1	0		1	1	0		1	1	0	

La méthode de programmation linéaire est simple et facile à appliquer. Son principal désavantage est d'ordre computationnel. Le nombre de paramètres du problème de programmation linéaire résultant est égal au nombre d'échantillons de taille n provenant des $RC > n$ cellules, $\binom{RC}{n}$, dont la grandeur devient assez rapidement impossible à traiter. À la section suivante, nous examinons divers moyens de rendre la méthode de programmation linéaire plus efficace sur le plan des calculs tout en retenant toutes ses bonnes propriétés.

3. Méthode pratique de programmation linéaire

L'idée fondamentale de la méthode de programmation linéaire consiste à obtenir un plan d'échantillonnage optimal en ce qui concerne le manque (minimal) prévu de « désirabilité » de l'échantillon en résolvant directement un problème de programmation linéaire dont les inconnues sont les $p(s), s \in S_n$, tout en maintenant la propriété de répartition proportionnelle prévue (RPP). Le seul obstacle est que le nombre d'éléments contenus dans S_n est souvent très grand et que, même avec la puissance de calcul moderne, il devient difficile d'effectuer la programmation linéaire si le nombre d'inconnues est grand.

Pour réduire la tâche de calcul pour ce problème de programmation linéaire déterminé par la cardinalité de S_n, nous voulons obtenir un sous-ensemble de S_n, disons S_{n0}, qui soit presque aussi représentatif que S_n, mais beaucoup plus petit, et donc résoudre le problème de programmation linéaire qui suit en nous servant d'un ensemble beaucoup plus petit de $p(s), s \in S_{n0}$ en tant qu' inconnues :

$$\min \sum_{s \in S_{n0}} w(s)\, p(s). \qquad (6)$$

De cette façon, du moins faut-il l'espérer, nous pouvons résoudre les problèmes pratiques comptant un grand nombre d'inconnues sans perdre les bonnes propriétés de la méthode de programmation linéaire.

3.1 Certaines stratégies justificatives

La stratégie susmentionnée est facile à énoncer, mais son application n'est pas entièrement évidente. En fait, nous pouvons explorer plusieurs directions pour déterminer un tel sous-ensemble $S_{n0} \subset S_n$. Nous décrivons ici une méthode élémentaire liée aux fonctions de perte à laquelle ont fait allusion Sitter et Skinner (1994) et nous montrons comment elle permet d'augmenter modestement la taille des problèmes qui peuvent être résolus. Puis, nous discutons de certaines directions à prendre évidentes qui n'ont toutefois produit que peu d'amélioration. En décrivant ces efforts non fructueux, nous justifions la proposition éventuelle.

La principale souplesse de la méthode de programmation linéaire tient au choix de la fonction de perte $w(s)$. Donc, il est naturel que nous considérions d'abord la fonction de perte lorsque nous essayons d'augmenter l'efficacité computationnelle de la méthode. L'examen de la fonction objective du problème de programmation linéaire (2) nous donne à penser qu'en tant que coefficients des inconnues $p(s)$, la fonction de perte $w(s)$ n'aura pas une valeur très grande lorsque la fonction objective sera minimisée. Autrement dit, dans un plan d'échantillonnage optimal, toutes les $p(s)$ positives seront attribuées uniquement à des échantillons dont le manque de « désirabilité » est faible. Partant de cette observation, nous émettons l'hypothèse qu'un bon remplacement pour S_n pourrait être le sous-ensemble qui suit

$$S_{n0} = \left\{ \begin{array}{l} s \in S_n : w(s) = \sum_{i=1}^{R} \left(n_{i.}(s) - nP_{i.}\right)^2 \\ + \sum_{j=1}^{C} \left(n_{.j}(s) - nP_{.j}\right)^2 \le w_0 \end{array} \right\}, \qquad (7)$$

où w_0 est une constante positive prédéterminée. Dans le cas de valeurs de marge entières, on pourrait même poser $w_0 = 0$ et se limiter aux échantillons où les valeurs de marge concordent. Par exemple, au tableau 3, la solution n'attribue une probabilité positive qu'à six échantillons et, pour chacun d'eux, la fonction objective est nulle.

Lu (2000) élabore des stratégies de programmation linéaire emboîtées pour résoudre ce problème. Pour des problèmes de taille modérée, comme des matrices 8×5 (c'est-à-dire, 40 cellules), cette méthode donne de bons résultats. Toutefois, pour des problèmes de plus grande amplitude, la taille des ensembles candidats résultants devient grande très rapidement, même dans le cas des valeurs de marge entières. Donc, pour les problèmes de grande taille, la méthode pose la même difficulté qu'auparavant, c'est-à-dire un ensemble candidat qui oblige à résoudre un problème de programmation linéaire où le nombre d'inconnues est trop élevé.

En réalité, même un ensemble d'échantillons candidats S_{n0} de la forme donnée par (7) est nettement plus grand qu'il n'est nécessaire pour que nous trouvions une solution optimale. Ce dont nous avons vraiment besoin est un sous-ensemble plus petit, mais suffisamment représentatif, où, par « petit », nous entendons suffisamment petit pour rendre *possible* la résolution du problème de programmation linéaire résultant et par « représentatif », nous entendons contenant les éléments qui promettent que ce problème de programmation linéaire soit *faisable*.

Avant de décrire la solution éventuelle que nous proposons pour ce problème, nous aimerions présenter certaines méthodes naïves d'obtention d'un tel « ensemble représentatif » qui n'ont pas donné de bons résultats. Ces méthodes ne sont pas très utiles en pratique, mais elles nous ont inspiré les idées qui sous-tendent la méthode plus élaborée que nous proposons.

1) Optimisation en deux étapes : Avant tout, dans (7), nous pourrions essayer de scinder S_{n0} en un grand nombre de sous-ensembles assez petits pour être traités individuellement par programmation linéaire, en espérant que les solutions optimales obtenues pour chacun à la première

étape de la procédure d'optimisation puissent être combinées pour former l'ensemble représentatif souhaité d'échantillons. Nous pourrions alors simplement regrouper ces solutions optimales et procéder une fois de plus à la programmation linéaire. Nous avons appliqué cette méthode à certains exemples simulés de taille, c'est-à-dire 6×6, 7×7, 8×8 et 9×9, à titre d'étude préliminaire de ses propriétés. En général, dans les deux premiers cas, la méthode a donné de très bons résultats de façon rapide; pour la matrice 8×8, son application a été lente et n'a pas toujours produit une solution optimale, et pour la matrice 9×9, elle est devenue inapplicable.

2) Rééchantillonnage à partir de S_{n0} : Nous pourrions aussi sélectionner au hasard une fraction, disons 10 %, de S_{n0} donné par (7) et espérer qu'elle soit statistiquement représentative de l'ensemble complet. Malheureusement, les résultats des simulations montrent que l'échantillon obtenu de cette façon n'est pas suffisamment « représentatif » et que, souvent, le problème de programmation linéaire résultant n'a aucune solution possible. Par exemple, par la méthode de programmation linéaire emboîtée mentionnée plus haut, nous avons pu obtenir des valeurs de marge entières concordantes pour les matrices 8×5 simulées, mais nous avons obtenu ces solutions beaucoup plus rapidement en échantillonnant de façon répétée 10 % de S_{n0} et en appliquant la méthode de Sitter et Skinner (1994) à cet ensemble jusqu'à l'obtention d'une solution possible. Par contre, lorsque nous avons considéré des matrices un peu plus grandes, l'obtention d'une solution possible a pris énormément de temps et la méthode est devenue rapidement impossible à appliquer.

Chacune de ces deux approches pose deux problèmes. Premièrement, la taille de S_{n0} devient combinatoirement énorme et même un dénombrement complet devient difficile. Le fait de devoir obtenir d'abord S_{n0}, puis de subdiviser le problème en morceaux aboutit rapidement au dépassement des limites pratiques de la méthode de programmation linéaire à cause de la taille des morceaux ou bien crée un nombre énorme de morceaux. Deuxièmement, ni l'une ni l'autre de ces stratégies ne vise à éviter les échantillons représentant un particulièrement mauvais choix en ce qui a trait au respect de la contrainte de RPP. La question qui se pose est celle de savoir s'il existe un moyen de produire un sous-ensemble d'échantillons candidats relativement « représentatif » sans choisir d'échantillons « inutiles » ou, de façon plus générale, si nous pouvons sélectionner des échantillons candidats pour lesquels la fréquence de l'apparition d'une entrée est plus ou moins reliée aux nombres prévus souhaités dans l'échantillon et, en outre, si nous pouvons le faire sans devoir d'abord dénombrer un grand S_{n0}. L'idée générale tourne autour du fait que, si nous pouvions sélectionner au hasard un sous-ensemble candidat directement à partir de S_n, sans dénombrement complet, par une méthode de sélection avec probabilités inégales qui assure simultanément que la fonction objective soit minimisée pour chaque échantillon et que la propriété de RPP soit satisfaite, nous aurions résolu le problème sans recourir du tout à la programmation linéaire. Nous nous sommes efforcés de trouver une méthode de sélection de ce genre, mais sans succès jusqu'à présent. Nous avons toutefois pu développer une procédure donnant une répartition proportionnelle prévue approximative (RPPA). Nous pouvons alors utiliser cette méthode pour générer aléatoirement un sous-ensemble candidat d'échantillons, S_{n0}, puis appliquer une méthode de programmation linéaire à ce sous-ensemble.

3.2 Une méthode d'échantillonnage avec propriété de RPPA

Ici, nous commençons par décrire l'application de la méthode au cas des valeurs de marge entières. Autrement dit, les totaux de colonne, $n_{.j} = \sum_{i=1}^{R} r_{ij}$, et les totaux de ligne, $n_{i.} = \sum_{j=1}^{C} r_{ij}$, sont des nombres entiers. Puis, nous expliquons comment nous pouvons facilement l'adapter au cas général. Le but de la méthode de programmation linéaire est de minimiser le manque prévu de « désirabilité » des échantillons tout en respectant la contrainte de RPP. Nous proposons d'accomplir ceci en deux étapes. Premièrement, nous élaborerons une méthode de sélection avec probabilités inégales qui produit des échantillons reproduisant exactement les valeurs de marge entières et ayant la propriété de RPPA. Puis, nous produirons aléatoirement un ensemble de taille modérée de telles matrices et nous appliquerons une méthode modifiée de programmation linéaire à ce sous-ensemble de toutes les matrices possibles. Nous répéterons la procédure en choisissant des ensembles de plus en plus grands. Nous décrirons la méthode d'échantillonnage, puis nous discuterons de la méthode modifiée de programmation linéaire.

Voici l'idée de base de l'élaboration d'une telle méthode d'échantillonnage : pour un tableau à double entrée (en supposant que les tailles prévues d'échantillon de cellule ont été ajustées de façon qu'elles soient comprises entre 0 et 1 comme cela a été fait pour passer du tableau 1 au tableau 2), nous commençons par tirer une série de cellules de population afin de produire $a_{11}, a_{12}, ..., a_{1C}$ à la première ligne par une méthode d'échantillonnage avec probabilités inégales sans remise fondée sur les dénombrements prévus de cette ligne, où $a_{ij} = 1$ si la ij^e est sélectionnée et = 0 autrement. Puis, nous tirons $a_{i1}, a_{i2}, ..., a_{iC}$ pour $i > 1$ en veillant à ce que toutes les $\sum_{k=1}^{i} a_{kj}$ soient inférieures ou égales aux totaux marginaux de colonne correspondants $n_{.j}$. Voici les détails de cette méthode d'échantillonnage.

Étape 1: Permuter aléatoirement les lignes et poser $i = 1$. Étant donné la première ligne de probabilités d'inclusion $r_{11}, r_{12}, ..., r_{1C}$, tirer un échantillon de $n_{1.}$ cellules à partir de C dans la strate de première ligne par une méthode d'échantillonnage avec probabilités inégales sans remise; écrire la première ligne d'échantillons en fonction de variables indicatrices $a_{11}, a_{12}, \ldots, a_{1C}$ telles que définies plus haut; poser $A_j = a_{1j}$ pour $j = 1, \ldots, C$.

Étape 2 : Poser $i = i + 1$.

Étape 2.1 : Pour $j = 1, \ldots, C$, faire ce qui suit :

a) poser $R_j = \sum_{k=1}^{i} r_{kj}$,
b) si $R_j - A_j \leq 0$ poser $a_{ij} = 0$,
c) si $R_j - A_j \geq 1$ poser $a_{ij} = 1$.

Étape 2.2 : Poser $J = \{j : 0 < R_j - A_j < 1\}$ et $rtot = \sum_{j=1}^{C} r_{ij} - \#\{j : a_{ij} = 1\}$. Si $rtot > 0$, alors $r'_{ij} = r_{ij} \times rtot / \sum_{j \in J} r_{ij}$, pour $j \in J$. S'il existe un $j_0 \in J$ tel que $r'_{ij0} > 1$, alors poser $a_{ij0} = 1$ et retourner à l'étape 2.1. Sinon, passer à l'étape 3.

Étape 3 : Tirer un échantillon de *rtot* cellules à partir de J par une méthode d'échantillonnage avec probabilités inégales sans remise et r'_{ij} pour obtenir a_{ij} pour $j \in J$. Poser $A_j = \sum_{k=1}^{i} a_{kj}$ pour $j = 1, \ldots, C$.

Étape 4 : Si $i = R$, alors s'arrêter; sinon, retourner à l'étape 2.

Notons qu'à l'étape 2, la façon de recalculer la i^e ligne de probabilités d'inclusion n'est pas unique. Cependant, les règles générales à suivre pour ce recalcul sont :

a) $0 \leq r'_{ij} \leq 1$ et, si $A_j = n_{.j}$, autrement dit si suffisamment d'unités sont sélectionnées à partir de la j^e colonne, r'_{ij} doit être fixé à 0; si $A_j = n_{.j} - (R - i + 1)$, autrement dit s'il n'est pas possible de sélectionner un nombre suffisant d'unités à partir de cette colonne à moins que toutes les unités restantes soient sélectionnées, r'_{ij} doit être fixé à 1;
b) garder $\sum_{j=1}^{C} r'_{ij} = \sum_{j=1}^{C} r_{ij} = n_{i.}$.

La méthode s'étend facilement au cas des totaux marginaux non entiers. Nous reportons toutefois la discussion détaillée de cette extension plus loin.

Nous pouvons maintenant utiliser la méthode susmentionnée pour produire un ensemble candidat, S_{n0}, et lui appliquer la méthode de programmation linéaire. Pour comprendre pourquoi nous choisissons de modifier la méthode de programmation linéaire, il faut se rendre compte que, pour le cas des valeurs de marge entières, chaque $s \in S_{n0}$ atteint déjà le minimum dans (2), de sorte qu'une application directe de la programmation linéaire revient à déterminer s'il existe ou non une solution possible. Donc, si nous générons un S_{n0} de taille égale à, disons, 500, puis 1 000, *etc.*, et que le progiciel de programmation linéaire continue de ne trouver aucune autre solution possible, nous ne savons vraiment pas si nous nous rapprochons ou non de la solution. Au lieu de cela, nous choisissons de renverser le problème d'optimisation et de résoudre un problème dual.

$$\min_{p(s)} \sum_{i,j} \left| \sum_{s \in S_{n0}} n_{ij}(s)\, p(s) - r_{ij} \right|. \qquad (8)$$

Nous savons que $w(s) = 0$ pour tout $s \in S_{n0}$ et nous cherchons une solution qui donne un nombre minimal de valeurs nulle dans (8). Essentiellement, nous avons permuté les rôles de la fonction objective et de la contrainte de RPP du problème original. La difficulté tient au fait qu'il est moins aisé d'utiliser la programmation linéaire pour résoudre (8). Voici un moyen de le faire. Nous commençons par préciser les contraintes

$$\sum_{s \in S_{n0}} n_{ij}(s)\, p(s) - r_{ij} + d_{ij} - e_{ij} = 0 \quad \text{pour } i = 1, \ldots, R \text{ et } j = 1, \ldots, C, \qquad (9)$$

où

$$d_{ij} \geq 0, e_{ij} \geq 0, d_{ij} e_{ij} = 0. \qquad (10)$$

Puis, nous notons que

$$\left| \sum_{s \in S_{n0}} n_{ij}(s)\, p(s) - r_{ij} \right| = \begin{cases} d_{ij} \text{ si } \sum_{s \in S_{n0}} n_{ij}(s)\, p(s) - r_{ij} < 0 \\ e_{ij} \text{ si } \sum_{s \in S_{n0}} n_{ij}(s)\, p(s) - r_{ij} \geq 0 \end{cases} = d_{ij} + e_{ij}. \qquad (11)$$

Donc, nous pouvons remplacer (8) par

$$\min_{p(s), d_{ij}, e_{ij}} \sum_{i,j} \left(d_{ij} + e_{ij} \right), \qquad (12)$$

à condition que

$$\sum_{s \in S_{n0}} n_{ij}(s)\, p(s) - r_{ij} + d_{ij} - e_{ij} = 0, d_{ij}, e_{ij}, p(s) \geq 0, d_{ij} e_{ij} = 0. \qquad (13)$$

3.3 Certains exemples pour des valeurs de marge entières

Nous allons maintenant utiliser deux exemples pour illustrer la méthode d'échantillonnage. Nous décrivons le premier, c'est-à-dire une matrice 10×8, en détail pour illustrer la procédure complète. Nous donnons le deuxième, pour une matrice de plus grande taille (20×15), pour illustrer la taille des problèmes qu'elle permet de résoudre (il s'agit presque de la taille limite des problèmes que la méthode proposée permet de résoudre en pratique). Toute procédure d'échantillonnage avec probabilités inégales sans remise peut y être intégrée. Dans l'exemple 1 qui suit, nous choisissons d'utiliser la méthode par groupement aléatoire de Rao, Hartley et Cochran (1962), puisqu'elle est simple et que nous ne devons vraiment reproduire qu'approximativement les probabilités de sélection, ce qu'elle permet de faire. Cependant, la méthode Rao-Hartley-Cochran ne donne de bons résultats que pour des problèmes, au maximum, de taille moyenne. Dans le cas des exemples 2 et 3, nous devrions utiliser une méthode qui reproduit exactement les probabilités de sélection. Il en existe plusieurs, mais nous choisissons celle élaborée par Lu (2000).

Exemple 1. Matrice 10×8 avec valeurs de marge entières : Le tableau 4 donne un problème de stratification double avec le nombre prévu d'unités d'échantillonnage dans les cellules et la taille d'échantillon.

Tableau 4
Nombres prévus d'unités d'échantillonnage dans les cellules sous stratification proportionnelle (n = 40)

N° de ligne	N° de colonne								Total marginal de ligne
	1	2	3	4	5	6	7	8	
1	0,41	0,55	0,58	0,80	0,23	0,61	0,70	0,12	4
2	0,52	0,15	0,07	0,90	0,28	0,10	0,37	0,61	3
3	0,72	0,15	0,65	0,73	0,39	0,34	0,85	0,17	4
4	0,70	0,55	0,46	0,10	0,41	0,05	0,24	0,49	3
5	0,07	0,63	0,45	0,81	0,52	0,02	0,70	0,80	4
6	0,61	0,33	0,79	0,21	0,02	0,61	0,67	0,76	4
7	0,88	0,48	0,73	0,69	0,44	0,64	0,86	0,28	5
8	0,22	0,14	0,85	0,37	0,69	0,45	0,49	0,79	4
9	0,85	0,44	0,80	0,76	0,31	0,71	0,60	0,53	5
10	0,02	0,58	0,62	0,63	0,71	0,47	0,52	0,45	4
Total marginal de colonne	5	4	6	6	4	4	6	5	40

Voici les étapes élémentaires de notre plan d'échantillonnage.

Étape 1. Obtenir un sous-ensemble d'échantillons candidates S_{n0} par la méthode d'échantillonnage proposée avec propriété de RPPA pour tirer, disons, 500 échantillons (obtenus en trois minutes). La proportion de l'échantillon que représente chaque cellule, qui est donnée au tableau 5, peut être comparée à celle du tableau 4 pour déterminer dans quelle mesure elle s'approche de celle satisfaisant la propriété de RPP.

Étape 2. Résoudre le problème de programmation linéaire donné par (12) et (13) pour obtenir

$$\min_{p(s),\, s\in S_{n0}} \sum_{i,j} \Big| \sum_{s} n_{ij}(s)\, p(s) - nP_{ij} \Big|. \qquad (14)$$

Si la valeur objective de (14) est supérieure à zéro, répéter l'étape 1 avec un ensemble S_{n0} plus grand. Si la valeur objective de (14) est nulle, arrêter, car une solution optimale a été obtenue.

Tableau 5
Nombres d'unités d'échantillonnage dans les cellules sous stratification proportionnelle (n = 40)

N° de ligne	N° de colonne								Total marginal de ligne
	1	2	3	4	5	6	7	8	
1	0,408	0,554	0,582	0,776	0,250	0,594	0,734	0,102	4
2	0,554	0,150	0,062	0,916	0,280	0,122	0,366	0,550	3
3	0,690	0,144	0,638	0,720	0,402	0,360	0,838	0,208	4
4	0,692	0,542	0,452	0,120	0,416	0,044	0,260	0,474	3
5	0,060	0,602	0,446	0,814	0,568	0,016	0,708	0,786	4
6	0,558	0,348	0,780	0,216	0,012	0,634	0,682	0,770	4
7	0,866	0,480	0,734	0,676	0,470	0,664	0,842	0,268	5
8	0,254	0,158	0,848	0,400	0,654	0,412	0,490	0,784	4
9	0,870	0,418	0,830	0,772	0,292	0,692	0,624	0,502	5
10	0,026	0,564	0,636	0,658	0,714	0,416	0,500	0,486	4
Total marginal de colonne	5	4	6	6	4	4	6	5	40

Dans cet exemple, un sous-ensemble S_{n0} candidat de 500 échantillons a suffi pour obtenir une valeur objective nulle.

Exemple 2. Matrice 20×15 avec valeurs de marge entières : Dans cet exemple, nous considérons la matrice 20×15 avec valeurs de marge entières donnée au tableau 6.

Les étapes réelles de calcul sont les suivantes :

Première itération :

Étape 1. Tirer 500 échantillons pour former S_{n0}.

Étape 2. La valeur objective de (14) est 0,1659.

Deuxième itération :

Étape 1. Tirer 500 échantillons à ajouter à S_{n0}.

Étape 2. La valeur objective de (14) est nulle. Le plan d'échantillonnage final est obtenu.

Cette procédure a duré de 30 à 60 secondes en se servant d'un programme Fortran sur poste de travail Sun Ultra 10.

3.4 Extension aux valeurs de marge non entières

La méthode s'étend facilement aux valeurs de marge non entières. Il suffit simplement de remplacer $n_{i.}$ dans tout l'algorithme par $n_{i.}^{*}$ qui prend la valeur $\lfloor r_{i.} \rfloor + 1$ avec une probabilité $\alpha = r_{i.} - \lfloor r_{i.} \rfloor$ et la valeur $\lfloor r_{i.} \rfloor$ avec une probabilité $1-\alpha$. La seule difficulté supplémentaire est que $E[w(s)]$ ne peut devenir nulle. Donc, nous n'avons pas de borne inférieure de référence évidente pour savoir si nous approchons ou non de la solution optimale. Cependant, la stratégie de randomisation susmentionnée garantit que, pour tout échantillon RPPA obtenu, nous ayons

$$\big|n_{i.}(s) - r_{i.}\big| < 1 \text{ et } \big|n_{.j}(s) - r_{.j}\big| < 1 \quad \text{pour } i = 1, \ldots, R, \; j = 1, \ldots, C. \qquad (15)$$

Ceci, conjugué à la propriété de RPP, $E[n_{ij}(s)] = \sum_s n_{ij}(s)\, p(s) = r_{ij}$, implique que l'espérance de la fonction de manque de désirabilité $w(s)$ définie en (3) est constante

$$E[w(s)] = \sum_i \big(r_{i.} - \lfloor r_{i.} \rfloor\big)\big(1 + \lfloor r_{i.} \rfloor - r_{i.}\big) + \sum_j \big(r_{.j} - \lfloor r_{.j} \rfloor\big)\big(1 + \lfloor r_{.j} \rfloor - r_{.j}\big). \qquad (16)$$

Nous en donnons la preuve à l'annexe 1. Donc, si l'expression (14) devient nulle lorsqu'on applique la stratégie susmentionnée, la solution résultante donnera la valeur minimale de $E[w(s)]$ comme en (16).

Exemple 3. Exemple réel de matrice 27×3 avec valeurs de marge non entières : Nous illustrons la méthode au moyen d'un exemple réel tiré d'une étude de l'hygiène du milieu portant sur la santé au travail dans diverses branches d'activité en Finlande (Vihma 1981). La population choisie pour l'étude comprenait 1 430 petits lieux de travail industriel (5 à 49 employés), représentant en tout 22 893 employés, dans la province d'Uusimaa, c'est-à-dire la province la plus au sud et la plus industrialisée de la Finlande.

Tableau 6
Nombres prévus d'unités d'échantillonnage dans les cellules sous stratification proportionnelle (n = 151)

0,73	0,58	0,08	0,59	0,69	0,84	0,04	0,17	0,27	0,8	0,02	0,84	0,79	0,03	0,53	7
0,43	0,39	0,35	0,57	0,35	0,38	0,47	0,53	0,39	0,96	0,52	0,27	0,68	0,40	0,31	7
0,73	0,25	0,15	0,73	0,48	0,32	0,91	0,49	0,03	0,61	0,14	0,61	0,73	0,25	0,87	7
0,13	0,28	0,35	0,60	0,26	0,38	0,37	0,39	0,71	0,01	0,93	0,72	0,30	0,66	0,91	7
0,32	0,06	0,86	0,47	0,80	0,93	0,96	0,30	0,65	0,72	0,67	0,54	0,51	0,77	0,44	9
0,12	0,78	0,81	0,34	0,28	0,02	0,89	0,41	0,94	0,82	0,37	0,81	0,85	0,51	0,05	8
0,48	0,51	0,50	0,62	0,35	0,11	0,85	0,78	0,29	0,39	0,69	0,07	0,67	0,78	0,91	8
0,86	0,41	0,11	0,17	0,75	0,89	0,48	0,48	0,91	0,20	0,53	0,67	0,34	0,19	0,01	7
0,81	0,00	0,13	0,93	0,36	0,12	0,19	0,86	0,33	0,04	0,79	0,69	0,56	0,37	0,82	7
0,82	0,22	0,54	0,82	0,61	0,46	0,74	0,33	0,24	0,53	0,41	0,18	0,30	0,03	0,77	7
0,95	0,60	0,35	0,33	0,95	0,43	0,06	0,63	0,71	0,02	0,55	0,23	0,87	0,21	0,11	7
0,96	0,65	0,96	0,83	0,41	0,58	0,49	0,27	0,74	0,88	0,93	0,46	0,6	0,13	0,11	9
0,83	0,54	0,05	0,96	0,79	0,70	0,33	0,81	0,86	0,45	0,45	0,84	0,29	0,30	0,80	9
0,75	0,65	0,63	0,04	0,32	0,36	0,38	0,80	0,50	0,23	0,37	0,23	0,85	0,69	0,20	7
0,79	0,31	0,55	0,26	0,04	0,05	0,91	0,11	0,43	0,79	0,14	0,64	0,44	0,48	0,06	6
0,23	0,92	0,81	0,42	0,49	0,10	0,74	0,56	0,24	0,47	0,34	0,57	0,60	0,56	0,95	8
0,13	0,77	0,65	0,66	0,05	0,23	0,58	0,74	0,19	0,94	0,26	0,75	0,16	0,71	0,18	7
0,31	0,01	0,60	0,38	0,01	0,55	0,70	0,72	0,20	0,87	0,55	0,82	0,77	0,44	0,07	7
0,63	0,67	0,21	0,02	0,16	0,68	0,14	0,17	0,95	0,78	0,58	0,55	0,94	0,96	0,56	8
0,99	0,40	0,31	0,26	0,85	0,87	0,77	0,75	0,42	0,49	0,76	0,51	0,75	0,53	0,34	9
12	9	9	10	9	9	11	10	10	11	10	11	12	9	9	151

Les unités primaires d'échantillonnage étaient les lieux de travail et un échantillon de $n = 100$ de ces unités était souhaité. Cette taille d'échantillon était la seule acceptable étant donné le coût de l'enquête éventuelle. Les lieux de travail ont été stratifiés en fonction de deux variables, à savoir le type de branche d'activité (27 catégories) et le nombre d'employés (3 catégories). Les nombres prévus d'unités dans les cellules sous stratification proportionnelle sont donnés au tableau 7. Le plan réel d'échantillonnage utilisé pour l'étude était fondé sur la méthode de Bryant et coll. (1960) après avoir regroupé certaines strates, car il s'agissait de la seule méthode disponible au moment de la réalisation de l'étude.

Nous avons appliqué notre méthode à ce problème. La valeur minimale possible de $E[w(s)]$ en utilisant la méthode que nous proposons est 5,0418. Les étapes réelles de calcul sont les suivantes :

Première itération :

Étape 1. Tirer 500 échantillons pour former S_{n0}, en générant au hasard les $n_{i.}^{*}$, indépendamment pour chaque échantillon.

Étape 2. La valeur objective de (14) est 0,45088.

Deuxième itération :

Étape 1. Tirer 500 échantillons à ajouter à S_{n0}.

Étape 2. La valeur objective de (14) est nulle. Le plan d'échantillonnage final est obtenu et produit la valeur minimale $E[w(s)]$ 5,0418.

Cette procédure a duré environ 30 secondes en se servant d'un programme Fortran sur poste de travail Sun Ultra 10.

Tableau 7
Enquête sur la santé au travail, Vihma (1981)
Nombres prévus d'unités d'échantillonnage dans les cellules sous stratification proportionnelle (n = 100)

Type de branche d'activité	Nombre d'employés			
	5-9	10-19	20-49	$r_{i.}$
Produits alimentaires	2,38	3,56	3,78	9,72
Aliments	0,35	0,14	0,56	1,05
Boissons	0,14	0,07	0,21	0,42
Textiles	1,33	1,26	1,46	4,05
Habillement	3,15	3,71	2,09	8,95
Cuir	0,56	0,14	0,07	0,77
Chaussures	0,07	0,07	0,21	0,35
Produits du bois	2,37	1,89	0,91	5,17
Mobilier	1,33	0,84	0,91	3,08
Produits du papier	0,42	0,49	0,42	1,33
Imprimerie	7,20	6,01	4,20	17,41
Produits chimiques industriels	0,56	0,35	0,28	1,19
Produits chimiques	1,82	1,54	1,53	4,89
Pétrole	0,14	0,07	0,00	0,21
Produits divers du charbon et du pétrole	0,07	0,07	0,14	0,28
Produits en caoutchouc	0,14	0,21	0,07	0,42
Produits en plastique	1,40	1,05	1,19	3,64
Produits en verre	0,42	0,21	0,21	0,84
Produits minéraux non métalliques	1,12	0,98	0,84	2,94
Fer et acier	0,14	0,07	0,35	0,56
Métaux non ferreux	0,35	0,14	0,28	0,77
Produits métalliques fabriqués	4,96	4,06	2,59	11,61
Machines	2,80	1,96	3,21	7,97
Matériel électrique	1,89	1,60	1,33	4,82
Matériel de transport	0,84	0,84	0,84	2,52
Matériel scientifique	0,56	0,42	0,49	1,47
Industries de la fabrication	1,68	0,91	0,98	3,57
$n_{.j}$	38,19	32,66	29,15	100,00

4. Conclusion

Nous proposons une méthode de stratification à deux dimensions qui étend l'applicabilité de la méthode de programmation linéaire de Sitter et Skinner (1994) à des problèmes de taille nettement plus grande. La méthode se concentre sur la façon de construire un petit ensemble d'échantillons candidats « représentatifs » en utilisant une méthode d'échantillonnage avec probabilités inégales qui génère des échantillons candidats satisfaisant presque les contraintes de RPPA du problème de programmation linéaire, puis sur l'application de la méthode de programmation linéaire à cet ensemble plus petit d'échantillons.

Il convient de souligner que la méthode de programmation linéaire s'étend facilement aux plans d'échantillonnage stratifiés à plusieurs degrés. Puisqu'il n'existe aucune différence fondamentale entre la méthode de programmation linéaire originale et l'extension proposée ici cette généralisation tient aussi pour la méthode proposée. Dans le même esprit, on peut aussi envisager une discussion des questions ayant trait à l'estimation de la variance des estimateurs résultant de Sitter et Skinner (1994).

Notons toutefois que, lorsqu'on impose des contraintes en fixant autour des nP_{ij} des bornes telles qu'ils soient des entiers, le problème ressemble à un problème d'arrondissement contrôlé (voir Kelly, Golden et Assad 1993, et les références qu'ils citent), mais nous n'explorons pas cette question ici.

Remerciements

Ces travaux ont été financés par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada.

Annexe 1

Preuve de (16) : $n_{i.}(s) - [r_{i.}] \sim$ Bernoulli $(r_{i.} - \lfloor r_{i.} \rfloor)$ et a la variance $(r_{i.} - \lfloor r_{i.} \rfloor)(1 + \lfloor r_{i.} \rfloor - r_{i.})$. Ceci implique que

$$\begin{aligned}\sum_s (n_{i.}(s) - r_{i.})^2 \, p(s) &= E(n_{i.}(s) - r_{i.})^2 \, V(n_{i.}(s)) \\ &= V(n_{i.}(s) - \lfloor r_{i.} \rfloor) \\ &= (r_{i.} - \lfloor r_{i.} \rfloor)(1 + \lfloor r_{i.} \rfloor - r_{i.}),\end{aligned}$$

et, suivant un argument similaire, que

$$\sum_s (n_{.j}(s) - r_{.j})^2 \, p(s) = (r_{.j} - \lfloor r_{.j} \rfloor)(1 + \lfloor r_{.j} \rfloor - r_{.j}).$$

Par conséquent, avec $w(s)$ définie en (3),

$$\begin{aligned}&E[w(s)] \\ &= \sum_s w(s)\, p(s) = \sum_s \left\{ \sum_i (n_{i.}(s) - r_{i.})^2 + \sum_j (n_{.j}(s) - r_{.j})^2 \right\} p(s) \\ &= \sum_i \sum_s (n_{i.}(s) - r_{i.})^2 \, p(s) + \sum_j \sum_s (n_{.j}(s) - r_{.j})^2 \, p(s) \\ &= \sum_i (r_{i.} - \lfloor r_{i.} \rfloor)(1 + \lfloor r_{i.} \rfloor - r_{i.}) + \sum_j (r_{.j} - \lfloor r_{.j} \rfloor)(1 + \lfloor r_{.j} \rfloor - r_{.j}).\end{aligned}$$

Bibliographie

Bryant, E.C., Hartley, H.O. et Jessen, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*, 55, 105-124.

Causey, B.D., Cox, L.H. et Ernst, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

Engle, M., Marsden, G. et Pollock, S.W. (1971). Child work and social class. *Psychiatry*, 34, 140-150.

Goodman, R., et Kish, L. (1950). Controlled selection-a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

Hess, I., Riedel, D.C. et Fitzpatrick, T.B. (1976). *Probability Sampling of Hospitals and Patients*. University of Michigan, Ann Arbor, deuxième édition.

Jessen, R.J. (1970). Probability sampling with marginal constraints. *Journal of the American Statistical Association*, 65, 776-795.

Jessen, R.J. (1973). Some properties of probability lattice sampling. *Journal of the American Statistical Association*, 68, 20-28.

Jessen, R.J. (1975). Square and cubic lattice sampling. *Biometrics*, 31, 449-471.

Kelly, J.K., Golden, B.L. et Assad, A.A. (1993). The controlled rounding problem: complexity and computational experience. *European Journal of Operational Research*, 65, 207-217.

Lahiri, P., et Mukerjee, R. (2000). On a simplification of the linear programming approach to controlled sampling. *Statistical Sinica*, 10, 1171-1178.

Lu, W. (2000). Multi-way stratification by linear programming made practical. Thèse de maîtrise, Simon Fraser University.

Rao, J.N.K., Hartley, H.O. et Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, Serie B, 24, 482-491.

Rao, J.N.K., et Nigam, A.K. (1990). Optimal controlled sampling design. *Biometrika*, 77, 807-814.

Rao, J.N.K., et Nigam, A.K. (1992). 'Optimal' controlled sampling: a unified approach. *International Statistical Review*, 60, 89-98.

Sitter, R.R., et Skinner, C.J. (1994). Stratification multidimensionnelle par programmation linéaire. *Techniques d'enquête*, 20, 69-78.

Skinner, C.J., Holmes, D.J. et Holt, D. (1994). Multiple frame sampling for multiple stratification. *International Statistical Review*, 62, 333-347.

Vihma, T. (1981). Health hazards and stress factors in small industry-Prevalence study in the province of Uusimaa with special reference to the type of industry and the occupational title as classifications for the description of occupational health problems. *Scandinavian Journal of Work, Environment and Health*, 7, Suppl. 3, 1-149.

Waterton, J.J. (1983). An exercise in controlled selection. *Applied Statistics*, 32, 150-164.