

Une généralisation de l'algorithme de Lavallée et Hidirolou pour la stratification dans les enquêtes auprès des entreprises

Louis-Paul Rivest¹

Résumé

Le présent article décrit des algorithmes de stratification permettant de tenir compte d'une divergence entre la variable de stratification et la variable étudiée lors de l'élaboration d'un plan de sondage stratifié. Nous proposons deux modèles pour caractériser la relation entre ces deux variables. L'un est un modèle de régression log-linéaire; l'autre suppose que la variable étudiée et la variable de stratification coïncident pour la plupart des unités, mais que des divergences importantes existent pour certaines unités. Puis, nous modifions l'algorithme de stratification de Lavallée et Hidirolou (1988) afin d'intégrer ces modèles dans la détermination des tailles d'échantillon et des limites de strate optimales pour un plan de sondage stratifié. Ensuite, nous donnons un exemple pour illustrer la performance du nouvel algorithme de stratification. Enfin, nous décrivons l'application numérique de cet algorithme.

Mots clés : Répartition optimum de Neyman; répartition par la méthode puissance; échantillonnage aléatoire stratifié.

1. Introduction

L'élaboration de plans d'échantillonnage stratifiés par les statisticiens ne date pas d'aujourd'hui. Dans Cochran (1977), les chapitres 5 et 5A sont consacrés à l'examen de plusieurs méthodes de répartition d'une population en strates. La création de strates est une question abordée couramment dans les publications statistiques. Les contributions récentes incluent Hedlin (2000), qui réexamine la règle de stratification d'Ekman (1959), et Dorfman et Valiant (2000), qui comparent la stratification basée sur un modèle à l'échantillonnage équilibré. La stratification basée sur un modèle fait l'objet d'une discussion dans Godfrey, Roshwalb et Wright (1984) et dans le chapitre 12 de Särndal, Swensson et Wretman (1992).

Les populations visées par les enquêtes auprès des entreprises ont une distribution asymétrique; un petit nombre d'unités représentent une part importante du total de la variable étudiée. Par conséquent, il convient d'inclure toutes les grandes unités dans l'échantillon (Dalenius 1952; Glasser 1962). Un bon plan d'échantillonnage comprend une strate à tirage complet pour les grandes entreprises, où toutes les unités sont échantillonnées, ainsi que des strates à tirage partiel pour les petites et moyennes entreprises. Habituellement, la fraction d'échantillonnage diminue parallèlement à la taille de l'unité; des poids d'échantillonnage importants sont appliqués aux petites entreprises. L'algorithme de stratification de Lavallée et Hidirolou (1988) est souvent utilisé pour déterminer les limites de strate et les tailles d'échantillon de strate dans ce contexte (consulter, par exemple, Slanta et Krenzke 1994, 1996). Cet algorithme comprend une variable de stratification, que l'on connaît pour toutes les unités de la population. Il donne les limites de strate et les tailles d'échantillon de strate qui

réduisent au minimum la taille totale de l'échantillon nécessaire pour atteindre le niveau voulu de précision. Il est basé sur une méthode itérative, élaborée par Sethi (1963) pour déterminer les limites optimales de strate. L'algorithme de Lavallée et Hidirolou ne tient pas compte des différences éventuelles entre la variable de stratification et la variable étudiée. À mesure que le temps passe, cette différence augmente et le plan d'échantillonnage produit par cet algorithme risque de ne plus satisfaire les critères de précision.

Dalenius et Gurney (1951), ainsi que Cochran (1977, chapitre 5A) considèrent la stratification dans les situations où la variable étudiée et la variable de stratification diffèrent. Nombre d'auteurs ont étudié des formules approximatives pour déterminer les limites de strate et pour évaluer le gain de précision dû à la stratification sur une variable auxiliaire. À cet égard, Serfling (1968), Singh et Sukatme (1969), Singh (1971), Singh et Parkash (1975), Anderson, Kish et Cornell (1976), Oslo (1976), Wang et Aggarwal (1984) et Yavada et Singh (1984) sont des contributions pertinentes. Hidirolou et Srinath (1993) et Hidirolou (1994) proposent des méthodes de mise à jour des limites de strate basées sur une nouvelle variable de stratification. Cependant, ces articles ne fournissent explicitement aucun algorithme de stratification tenant compte de la divergence entre la variable de stratification et la variable étudiée. Le présent article comble cette lacune grâce à la construction de généralisations de l'algorithme de Lavallée et Hidirolou (1988) qui expriment la relation entre ces deux variables sous forme de modèle statistique.

Nous passons d'abord brièvement en revue l'échantillonnage stratifié et les méthodes de répartition d'échantillon. Puis, nous proposons des modèles de la relation entre la variable de stratification et la variable étudiée. Ensuite,

1. Louis-Paul Rivest, Département de mathématiques et de statistique, Université Laval, Ste-Foy, (Québec) Canada, G1K 7P4.

nous présentons l'application de l'algorithme de Sethi quand la variable de stratification et la variable étudiée diffèrent. Enfin, nous donnons des exemples numériques.

2. Une revue de l'échantillonnage aléatoire stratifié

Dans la suite de l'article, nous utiliserons la notation générale relative à l'échantillonnage aléatoire stratifié suivante :

L = nombre de strates;

$W_h = N_h / N$ représente, pour $h = 1, \dots, L$, le poids relative de la strate h , N_h représente la taille de la strate h et $N = \sum N_h$, la taille totale de la population;

n_h représente, pour $h = 1, \dots, L$, la taille d'échantillon dans la strate h et $f_h = n_h / N_h$ représente la fraction d'échantillonnage;

\bar{Y}_h et \bar{y}_h représente les moyennes de population et d'échantillon de Y dans la strate h ;

S_{yh} représente l'écart-type de population de Y dans la strate h .

Dans le présent article, les strates sont créées en prenant X pour variable de stratification. La strate h comprend toutes les unités pour lesquelles la valeur de X est comprise dans l'intervalle $(b_{h-1}, b_h]$, où $-\infty = b_0 < b_1 < \dots < b_{L-1} < b_L = \infty$ sont les limites de la strate.

Nous pouvons exprimer l'estimateur d'échantillon de \bar{Y} sous la forme $\bar{y}_{st} = \sum W_h \bar{y}_h$; sa variance est donnée par:

$$\text{Var}(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{yh}^2. \quad (2.1)$$

Dans les enquêtes auprès des entreprises, toutes les grandes entreprises sont sélectionnées; nous posons que la strate L est la strate à tirage complet, de sorte que $n_L = N_L$. Si $h < L$, n_h , on peut écrire la taille de l'échantillon dans la strate à triage partiel h sous la forme $(n - N_L) a_h$, où n est la taille totale de l'échantillon et a_h dépend de la règle de répartition. Les deux règles de répartition envisagées ici sont:

- la règle de répartition par la méthode puissance

$$a_h = \frac{(W_h \bar{Y}_h)^p}{\sum_{k=1}^{L-1} (W_k \bar{Y}_k)^p} \quad (2.2)$$

où p est un nombre positif compris dans $(0, 1]$;

- la règle de répartition optimum de Neyman

$$a_h = \frac{W_h S_{yh}}{\sum_{k=1}^{L-1} W_k S_{yk}}. \quad (2.3)$$

La résolution de (2.1) pour trouver la valeur de n donne

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 S_{yh}^2 / a_h}{\text{Var}(\bar{y}_{st}) + \sum_{h=1}^{L-1} W_h S_{yh}^2 / N}. \quad (2.4)$$

Les limites optimales de strate sont les valeurs de b_1, \dots, b_{L-1} qui minimisent n , sous une condition de précision de \bar{y}_{st} telle que $\text{Var}(\bar{y}_{st}) = \bar{Y}^2 c^2$, où c est le coefficient de variation (c.v.) cible. Pour les enquêtes auprès des entreprises, on utilise souvent la fourchette $c = 1\%$ à 10% .

3. Certains modèles tenant compte de la divergence entre la variable de stratification et la variable étudiée

Dans cette section, $\{x_i, i = 1, \dots, N\}$ représente la variable de stratification connue pour les N unités de la population. Nombre d'algorithmes de stratification, y compris celui de Lavallée et Hidiroglou, sont basés sur l'hypothèse que $\{x_i, i = 1, \dots, N\}$ représente aussi les valeurs de la variable étudiée. Nous proposons ici des modèles statistiques permettant de tenir compte d'une différence entre ces deux variables.

Dans la suite, il est pratique de considérer X et Y comme des variables aléatoires continues et de représenter par $f(x)$, $x \in R$ la densité de X . Les données $\{x_i, i = 1, \dots, N\}$ peuvent être considérées comme N réalisations indépendantes de la variable aléatoire X . Puisque la strate h comprend les unités de population pour lesquelles la valeur de X est comprise dans l'intervalle $(b_{h-1}, b_h]$ la stratification est basée sur les valeurs de $E(Y | b_h \geq X > b_{h-1})$ et $\text{Var}(Y | b_h \geq X > b_{h-1})$ c'est-à-dire la moyenne et la variance conditionnelles de Y , étant donné que l'unité est comprise dans la strate h , pour $h = 1, \dots, L-1$. Suivent trois modèles de la relation entre X et Y , ainsi que leurs moyenne et variance conditionnelles respectives pour Y .

3.1 Un modèle log-linéaire

Notre premier modèle a la forme $\log(Y) = \alpha + \beta_{\log} \log(X) + \varepsilon$, où ε est une variable aléatoire normale de moyenne nulle et de variance σ_{\log}^2 , qui est indépendante de X , et α et β_{\log} sont des paramètres qu'il faut déterminer. Si $\alpha = 0, \beta_{\log} = 1$ et $\sigma_{\log}^2 = 0$, on a $X = Y$; la variable étudiée et la variable de stratification sont identiques. En général, $Y = e^\alpha X^{\beta_{\log}} e^\varepsilon$. On peut évaluer les moments conditionnels de Y d'après les propriétés fondamentales de la loi de

distribution lognormale (voir Johnson et Kotz 1970), c'est-à-dire

$$E(e^\varepsilon) = e^{\sigma_{\log}^2/2} \text{ et } \text{Var}(e^\varepsilon) = e^{\sigma_{\log}^2} (e^{\sigma_{\log}^2} - 1).$$

Nous avons

$$E(Y|b_h \geq X > b_{h-1}) = \exp(\alpha + \sigma_{\log}^2/2) E(X^{\beta_{\log}} | b_h \geq X > b_{h-1})$$

tandis que $\text{Var}(Y|b_h \geq X > b_{h-1})$ est égale à

$$\text{Var}(E(Y|X)|b_h \geq X > b_{h-1}) + E(\text{Var}(Y|X)|b_h \geq X > b_{h-1})$$

$$= \exp(2\alpha + \sigma_{\log}^2) \{ \text{Var}(X^{\beta_{\log}} | b_h \geq X > b_{h-1})$$

$$+ (e^{\sigma_{\log}^2} - 1) E(X^{2\beta_{\log}} | b_h \geq X > b_{h-1}) \}$$

$$= \exp(2\alpha + \sigma_{\log}^2) \{ e^{\sigma_{\log}^2} E(X^{2\beta_{\log}} | b_h \geq X > b_{h-1})$$

$$- E(X^{\beta_{\log}} | b_h \geq X > b_{h-1}) \}.$$

Dans certains cas, les valeurs des paramètres β_{\log} et σ_{\log} peuvent être calculées d'après des données historiques. Des valeurs simples spéciales de ces paramètres sont $\beta_{\log} = 1$ et $\sigma_{\log}^2 = (1 - \rho^2) \text{Var}(\log(X))$. Ici, ρ représente la corrélation supposée entre $\log(X)$ et $\log(Y)$, à laquelle on peut donner des valeurs prédéterminées, comme 0,95 ou 0,99.

3.2 Un modèle linéaire

Dans les textes traitant de l'échantillonnage, la relation entre Y et X est souvent modélisée au moyen d'un modèle linéaire hétéroscédastique,

$$Y = \beta_{\text{lin}} X + \varepsilon, \quad (3.5)$$

où la distribution conditionnelle de ε , étant donné X , a une moyenne nulle et une variance $\sigma_{\text{lin}}^2 X^\gamma$, pour un paramètre donné non négatif γ . Des calculs simples donnent $E(Y|b_h \geq X > b_{h-1}) = \beta_{\text{lin}} E(X|b_h \geq X > b_{h-1})$, tandis que $\text{Var}(Y|b_h \geq X > b_{h-1}) = \beta_{\text{lin}}^2$

$$\left\{ \begin{array}{l} \text{Var}(X|b_h \geq X > b_{h-1}) \\ + (\sigma_{\text{lin}}/\beta_{\text{lin}})^2 E(X^\gamma|b_h \geq X > b_{h-1}) \end{array} \right\}.$$

Pour une valeur arbitraire $\gamma \geq 0$, la variance conditionnelle de Y dépend de trois moments conditionnels de X . La généralisation de l'algorithme de Sethi présentée à la section 5 ne marche pas dans cette situation. Notons, cependant que, si $\gamma = 2$, les moyenne et variance conditionnelles de Y sont proportionnelles à celles du modèle log-linéaire avec

$$\beta_{\log} = 1 \text{ et } \sigma_{\log}^2 = \log(1 + (\sigma_{\text{lin}}/\beta_{\text{lin}})^2); \quad (3.6)$$

les facteurs de proportionnalité sont $\exp(\alpha + \sigma_{\log}^2/2)/\beta_{\text{lin}}$ et $\exp(2\alpha + \sigma_{\log}^2)/\beta_{\text{lin}}^2$ pour l'espérance conditionnelle et la variance conditionnelle, respectivement. Donc, les deux modèles de la relation entre la variable de stratification et la

variable étudiée, à savoir le modèle log-linéaire de la section 3.1 ou le modèle linéaire (3.5) avec le paramètre $\gamma = 2$, mènent, à la section 5, au même plan d'échantillonnage stratifié, à condition que les égalités (3.6) soient vérifiées. Plus loin, nous utiliserons le modèle log-linéaire pour représenter la relation entre X et Y . Ce modèle devrait donner de bons résultats quand la relation réelle entre Y et X est modélisée par (3.5) avec $\gamma \approx 2$. Si l'on suppose que le modèle (3.5) est vérifié pour une valeur plus faible de γ , l'algorithme de la section 5 reste applicable lorsqu'on fixe la valeur de γ à 0 ou à 1. Toutefois, nous ne nous attardons pas sur ce problème ici.

3.3 Un modèle à remplacement aléatoire

Ce modèle se fonde sur l'hypothèse que la variable de stratification est égale à la variable étudiée, c'est-à-dire $X = Y$, pour la plupart des unités. Il existe cependant une faible probabilité ε qu'une unité ait changé considérablement; le cas échéant, la valeur de Y est caractérisée par la fonction de densité $f(x)$ et est distribuée indépendamment de la valeur de X . Cette approche est celle utilisée dans Rivest (1999) pour modéliser l'occurrence des unités qui sautent d'une strate à une autre, unités pour lesquelles X n'est pas représentative de Y . Plus formellement, on peut écrire

$$Y = \begin{cases} X \text{ avec probabilité } 1 - \varepsilon \\ X_{\text{nouv}} \text{ avec probabilité } \varepsilon, \end{cases}$$

où X_{nouv} représente une variable aléatoire dont la densité $f(x)$ est distribuée indépendamment de X . Sous ce modèle, la moyenne conditionnelle de Y est donnée par

$$E(Y|b_h \geq X > b_{h-1}) =$$

$$(1 - \varepsilon) E(X|b_h \geq X > b_{h-1}) + \varepsilon E(X),$$

tandis que sa variance conditionnelle est égale à

$$\text{Var}(Y|b_h \geq X > b_{h-1})$$

$$= (1 - \varepsilon) E(X^2|b_h \geq X > b_{h-1}) + \varepsilon E(X^2)$$

$$- \{(1 - \varepsilon) E(X|b_h \geq X > b_{h-1}) + \varepsilon E(X)\}^2.$$

4. Un exemple

Avant de passer aux détails techniques de la construction des algorithmes, il serait utile de donner un exemple. Considérons la population MU284 de Särndal, Swensson et Wretman (1992) contenant des données sur 284 municipalités suédoises.

Pour élaborer un plan d'échantillonnage stratifié sur l'estimation de la moyenne de RMT85, c'est-à-dire les recettes provenant de l'imposition municipale de 1985, nous utilisons REV84, c'est-à-dire la valeur des biens

immobiliers conformément à l'évaluation de 1984, comme variable de stratification. Nous posons $L = 5$ et fixons le c.v. cible à 5 %. Deux plans d'échantillonnage stratifiés obtenus à l'aide de l'algorithme de Lavallée et Hidiroglou sont présentés au tableau 1, pour la répartition par la méthode puissance où $p = 0,7$ et pour la répartition optimum de Neyman. Dans les deux cas, $n = 19$. Appliqués à la variable étudiée RMT85, ces deux plans d'échantillonnage donnent des estimateurs des recettes totales dont le coefficient de variation est égal à 8,3 % et à 7,3 %, respectivement. Si l'on ne tient pas compte d'une divergence entre la variable étudiée et la variable de stratification, on obtient des estimateurs plus variables que prévu.

Tableau 1

Plans d'échantillonnage stratifiés obtenus au moyen de l'algorithme de Lavallée et Hidiroglou pour la population MU284 en utilisant REV84 comme variable de stratification et un c.v. cible de 5 %

Répartition puissance avec $p = 0,7$							
	b_h	moyenne	variance	N_h	n_h	f_h	n
strate 1	1 251	874	56 250	86	1	0,01	19
strate 2	2 352	1 696	100 898	82	2	0,02	19
strate 3	4 603	3 114	351 547	65	3	0,05	19
strate 4	10 606	6 442	2 027 436	41	3	0,05	19
strate 5	59 878	19 631	275 502 518	10	10	1	19
Répartition optimum de Neyman							
	b_h	moyenne	variance	N_h	n_h	f_h	n
strate 1	1 273	878	57 260	87	2	0,02	19
strate 2	2 336	1 701	99 688	81	2	0,02	19
strate 3	4 619	3 114	351 547	65	3	0,05	19
strate 4	11 776	6 921	3 724 610	46	7	0,15	19
strate 5	59 878	28 418	426 851 844	5	5	1	19

Pour modéliser la relation entre REV84 et RMT85, nous utilisons le modèle log-linéaire de la section 3.1. La régression linéaire de $\log(\text{RMT85})$ sur $\log(\text{REV84})$ produit des valeurs extrêmes qui rendent les estimations de β_{\log} et σ_{\log} par les moindres carrés non représentatives de la relation entre les deux variables. Nous utilisons donc plutôt des estimations robustes obtenues au moyen de la fonction `lmRobMM` de `Splus`. Ces estimations sont données par $\hat{\beta}_{\log} = 1,1$ et $\hat{\sigma}_{\log} = 0,2116$. Le tableau 2 montre les plans de sondage stratifiés obtenus à l'aide de l'algorithme généralisé de Lavallée et Hidiroglou pour les deux règles de répartition. Celles-ci donnent toutes deux des estimateurs du total de RMT85 dont le c.v. est de 5,7 %, c'est-à-dire un c.v. encore supérieur à la valeur cible de 5 %. Puisque la régression log-linéaire produit des valeurs extrêmes, l'hypothèse selon laquelle les erreurs suivent la loi de distribution normale émise à la section 3.1 n'est pas vérifiée, ce qui pourrait expliquer l'impossibilité d'atteindre exactement le c.v. cible. L'augmentation de la taille de l'échantillon de

$n = 19$ à $n = 28$ vaut la peine d'être mentionnée. Pour les deux méthodes de répartition, le plan de sondage obtenu à l'aide du modèle log-linéaire produit des strates à tirage complet plus petites que celles obtenues par Lavallée et Hidiroglou.

Tableau 2

Plans d'échantillonnage stratifiés obtenus à l'aide de l'algorithme généralisé de Lavallée et Hidiroglou pour la population MU284 en utilisant REV84 comme variable de stratification, un modèle log-linéaire avec $\beta_{\log} = 1,1$ et $\sigma_{\log} = 0,216$ de la relation entre REV84 et RMT85, et un c.v. cible de 5 %

Algorithme de stratification à modèle log-linéaire avec répartition puissance où $p = 0,7$							
	b_h	moyenne	variance	N_h	n_h	f_h	n
strate 1	1 558	1 023	97 245	121	4	0,03	28
strate 2	3 031	2 219	168 204	81	5	0,06	28
strate 3	5 706	4 022	464 471	44	6	0,14	28
strate 4	11 107	7 602	2 659 061	32	7	0,22	28
strate 5	59 878	25 536	39 131 413	6	6	1	28
Algorithme de stratification à modèle log-linéaire avec répartition optimum de Neyman							
	b_h	moyenne	variance	N_h	n_h	f_h	n
strate 1	1 582	1 023	97 245	121	4	0,03	28
strate 2	3 040	2 219	168 204	81	5	0,06	28
strate 3	5 608	4 022	464 471	44	5	0,11	28
strate 4	11 476	7 709	2 952 313	33	9	0,27	28
strate 5	59 878	28 418	4,27e+08	5	5	1	28

Une autre méthode que celle de l'algorithme généralisé de Lavallée et Hidiroglou pour construire des plans d'échantillonnage stratifiés consiste à utiliser l'algorithme original en choisissant un c.v. cible plus faible. On augmente ainsi la taille de l'échantillon, ce qui réduit la variance de l'estimateur du total de la variable étudiée. La construction d'un plan d'échantillonnage pour RMT85 en prenant REV84 comme variable de stratification, à l'aide de l'algorithme standard de Lavallée et Hidiroglou avec la règle de répartition puissance ($p = 0,7$) et un c.v. cible de 3,6 % produit un plan d'échantillonnage stratifié pour lequel $n = 28$. Ce plan d'échantillonnage a la même taille d'échantillon que ceux présentés au tableau 2. Le c.v. de l'estimateur du total de RMT85 est 5,7 %, valeur identique à celle des c.v. obtenus pour les plans d'échantillonnage du tableau 2. La principale différence entre ces plans d'échantillonnage est la taille de la strate à tirage complet. Celle-ci est égale à $N_5 = 13$ pour le plan d'échantillonnage conçu au moyen de l'algorithme de Lavallée et Hidiroglou, mais est $N_5 = 5$ et $N_5 = 6$ pour les plans d'échantillonnage du tableau 2. Permettre que la variable de stratification et la variable étudiée diffèrent semble réduire l'importance relative de la strate à tirage complet dans le plan d'échantillonnage. Des études approfondies seront nécessaires pour confirmer cette hypothèse.

Nous avons également appliqué à la variable REV84 l'algorithme de stratification choisi pour le modèle à remplacement aléatoire de la section 3.3 (avec répartition optimum de Neyman). Si l'on suppose qu'il existe des variations pour 2 % des unités ($\epsilon = 0,02$), l'algorithme généralisé de Lavallée et Hidiroglou produit un plan d'échantillonnage stratifié avec $n = 37$ unités d'échantillonnage; l'estimateur résultant du total de RMT85 a un c.v. de 5,5 %. Une propriété intéressante de ce plan d'échantillonnage stratifié est que la fraction d'échantillonnage la plus faible est $\min_h f_h = 9,3\%$ cette valeur est nettement plus grande que celle de $\min_h f_h$ pour les plans d'échantillonnage des tableaux 1 et 2. Malgré l'existence de valeurs extrêmes, le modèle à remplacement aléatoire ne décrit pas aussi bien que le modèle log-linéaire les divergences entre REV84 et RMT85. C'est pourquoi une plus grande taille d'échantillon, 37 au lieu de 28, est nécessaire pour obtenir un estimateur dont la variance est comparable à celle obtenue pour la stratification basée sur un modèle log-linéaire.

5. Une méthode de construction d'algorithmes de stratification

Le but d'un algorithme de stratification est de déterminer les limites de strate et les tailles d'échantillon optimales pour l'échantillonnage de Y en se servant des valeurs connues $\{x_i; i = 1, \dots, N\}$ de la variable X pour toutes les unités de la population. Un modèle, comme ceux décrits à la section 3, caractérise la relation entre X et Y . Dans cette section, nous étendons l'algorithme de stratification de Lavallée et Hidiroglou (1988) à des situations où X et Y diffèrent. Nous nous servons du modèle log-linéaire de la section 3.1 pour tenir compte des différences entre Y et X . Les modifications nécessaires pour traiter le modèle à remplacement aléatoire sont faciles à appliquer (voir Rivest 1999).

5.1 Une généralisation de la méthode de stratification de Sethi (1963)

Il est pratique de considérer une population infinie analogue à l'équation (2.4) pour n . Puisque la variable aléatoire X a une densité $f(x)$, les deux premiers moments conditionnels de Y , étant donné que $b_{h-1} < X \leq b_h$, peuvent s'écrire en fonction de

$$W_h = \int_{b_{h-1}}^{b_h} f(x) dx, \varphi_h = \int_{b_{h-1}}^{b_h} \alpha^\beta f(x) dx,$$

$$\text{et } \psi_h = \int_{b_{h-1}}^{b_h} x^{2\beta} f(x) dx,$$

où β est la pente du modèle log-linéaire donné à la section 3.1 (à la présente section, β et σ représentent les paramètres du modèle log-linéaire de la section 3.1, mais, puisqu'il n'y a aucun risque de confusion, nous n'utilisons plus l'indice

log). Aux fins de la stratification, il est utile de réécrire (2.4) en fonction des moyenne et variance conditionnelles de Y ,

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 \text{Var}(Y | b_h \geq X > b_{h-1}) / a_{h,X}}{\bar{Y}^2 c^2 + \sum_{h=1}^{L-1} W_h \text{Var}(Y | b_h \geq X > b_{h-1}) / N}, \quad (5.7)$$

où $a_{h,X}$ représente la règle de répartition écrite en fonction de la variable connue X . Par exemple, dans le cas de la répartition par la méthode puissance,

$$a_{h,X} = \frac{\{W_h E(Y | b_h \geq X > b_{h-1})\}^p}{\sum_{k=1}^{L-1} \{W_k E(Y | b_k \geq X > b_{k-1})\}^p},$$

pour $h = 1, \dots, L-1$. Étant donné un modèle de la relation entre Y et X , $\text{Var}(Y | b_h \geq X > b_{h-1})$ et $E(Y | b_h \geq X > b_{h-1})$ peuvent s'écrire en fonction de W_h, φ_h , et ψ_h . Donc, nous pouvons évaluer les dérivées partielles de n par rapport à b_h , pour $h < L-1$, en appliquant la règle de dérivation d'une fonction composée, ou règle d'enchaînement,

$$\begin{aligned} \frac{\partial n}{\partial b_h} &= \frac{\partial n}{\partial W_h} \frac{\partial W_h}{\partial b_h} + \frac{\partial n}{\partial \varphi_h} \frac{\partial \varphi_h}{\partial b_h} + \frac{\partial n}{\partial \psi_h} \frac{\partial \psi_h}{\partial b_h} \\ &+ \frac{\partial n}{\partial W_{h+1}} \frac{\partial W_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \varphi_{h+1}} \frac{\partial \varphi_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \psi_{h+1}} \frac{\partial \psi_{h+1}}{\partial b_h}. \end{aligned}$$

Observons que

$$\frac{\partial W_h}{\partial b_h} = -\frac{\partial W_{h+1}}{\partial b_h} = f(b_h)$$

$$\frac{\partial \varphi_h}{\partial b_h} = -\frac{\partial \varphi_{h+1}}{\partial b_h} = b_h^\beta f(b_h)$$

$$\frac{\partial \psi_h}{\partial b_h} = -\frac{\partial \psi_{h+1}}{\partial b_h} = b_h^{2\beta} f(b_h)$$

Ceci nous mène au résultat suivant, pour $h < L-1$,

$$\frac{\partial n}{\partial b_h} = f(b_h)$$

$$\left\{ \left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}} \right) + \left(\frac{\partial n}{\partial \varphi_h} - \frac{\partial n}{\partial \varphi_{h+1}} \right) b_h^\beta + \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}} \right) b_h^{2\beta} \right\}.$$

Pareillement,

$$\frac{\partial n}{\partial b_{L-1}} = f(b_{L-1}) \left\{ -N + \frac{\partial n}{\partial W_{L-1}} + \frac{\partial n}{\partial \varphi_{L-1}} b_{L-1}^\beta + \frac{\partial n}{\partial \psi_{L-1}} b_{L-1}^{2\beta} \right\}.$$

Nous utilisons l'algorithme de Sethi (1963) pour résoudre $\partial n / \partial b_h = 0$. Il se fonde sur l'hypothèse que les dérivées partielles sont proportionnelles à des fonctions quadratiques en b_h^β . La valeur mise à jour de b_h^β est donnée par la racine

de la fonction quadratique correspondante ayant la valeur la plus grande. Si $h < L - 1$, ceci nous donne

$$b_h^{\beta \text{ nouv}} = \frac{-\left(\frac{\partial n}{\partial \varphi_h} - \frac{\partial n}{\partial \varphi_{h+1}}\right) / \left\{2\left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right)\right\} + \left\{\left(\frac{\partial n}{\partial \varphi_h} - \frac{\partial n}{\partial \varphi_{h+1}}\right)^2 - 4\left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right)\left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}}\right)\right\}^{1/2}}{\left\{2\left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right)\right\}}$$

tandis que si $h = L - 1$, nous avons

$$b_{L-1}^{\beta \text{ nouv}} = \frac{-\frac{\partial n}{\partial \varphi_{L-1}} + \left\{\left(\frac{\partial n}{\partial \varphi_{L-1}}\right)^2 - 4\frac{\partial n}{\partial \psi_{L-1}}\left(\frac{\partial n}{\partial W_{L-1}} - N\right)\right\}^{1/2}}{\left(2\frac{\partial n}{\partial \psi_{L-1}}\right)}$$

Les dérivées partielles de n par rapport à W_h , φ_h , et ψ_h dépendent des moments d'ordre 0, 1 et 2 de x^β dans la strate h . Nous calculons ces moments d'après les N valeurs de x dans la population. Par exemple,

$$\varphi_h = \frac{1}{N} \sum_{i: b_{h-1} < x_i \leq b_h} x_i^\beta.$$

Nous donnons plus loin des applications de cette méthode générale.

Lors de l'utilisation de l'algorithme de Sethi, on pose habituellement que $L \geq 3$. Notons toutefois que l'algorithme marche aussi si $L = 2$. Dans ce cas, l'algorithme recherche la limite entre une strate à tirage complet et une strate à tirage partiel. Les évaluations successives de $b_{L-1}^{\beta \text{ nouv}}$ présentées plus haut produisent une limite optimale. Quand on suppose que la variable de stratification et la variable étudiée coïncident, c'est-à-dire $X = Y$, cette limite est presque identique à celle obtenue au moyen de l'algorithme présenté dans Hidiroglou (1986).

5.2 Un algorithme pour la répartition par la méthode puissance

Pour le modèle log-linéaire de la section 3.1, l'espérance conditionnelle est $E(Y|b_h \geq X > b_{h-1}) = C\varphi_h / W_h$ tandis que la variance conditionnelle est

$$\text{Var}(Y|b_h \geq X > b_{h-1}) = C^2 \{e^{\sigma^2} \psi_h / W_h - (\varphi_h / W_h)^2\},$$

où $C = \exp(\alpha + \sigma^2 / 2)$. Aux termes de la règle de répartition par la méthode puissance, $a_{h,X} = \varphi_h^p / \sum_{h=1}^{L-1} \varphi_k^p$ et la formule (5.7) pour n devient

$$n = NW_L + \frac{\sum_{h=1}^{L-1} \varphi_h^p \sum_{h=1}^{L-1} (e^{\sigma^2} W_h \psi_h - \varphi_h^2) / \varphi_h^p}{\left(\sum x_i^\beta / N\right)^2 c^2 + \sum_{h=1}^{L-1} (e^{\sigma^2} \psi_h - \varphi_h^2 / W_h) / N}$$

Les dérivées partielles nécessaires pour appliquer l'algorithme de stratification se calculent facilement; pour $h \leq L - 1$,

$$\frac{\partial n}{\partial W_h} = \frac{Ae^{\sigma^2} \psi_h / \varphi_h^p}{F} - \frac{AB(\varphi_h / W_h)^2 / N}{F^2}$$

$$\frac{\partial n}{\partial \varphi_h} = \frac{A\{-pe(\sigma^2 W_h \psi_h - \varphi_h^2) / \varphi_h^{p+1} - 2 / \varphi_h^{p-1}\} + p\varphi_h^{p-1} B}{F} + 2\frac{AB\varphi_h / (nW_h)}{F^2}$$

$$\frac{\partial n}{\partial \psi_h} = e^{\sigma^2} \frac{AW_h / \varphi_h^p}{F} - e^{\sigma^2} \frac{AB / N}{F^2},$$

où

$$A = \sum_{h=1}^{L-1} \varphi_h^p, B = \sum_{h=1}^{L-1} (e^{\sigma^2} W_h \psi_h - \varphi_h^2) / \varphi_h^p,$$

et

$$F = \left(\sum x_i^\beta / N\right)^2 c^2 + \sum_{h=1}^{L-1} (e^{\sigma^2} \psi_h - \varphi_h^2 / W_h) / N.$$

5.3 Un algorithme pour la répartition optimum de Neyman

Dans le cas de la répartition optimum de Neyman, la règle de répartition (2.3) écrite en fonction de W_h , φ_h et ψ_h est

$$a_{h,X} = \frac{\{e^{\sigma^2} \psi_h W_h - \varphi_h^2\}^{1/2}}{\sum_{h=1}^{L-1} \{e^{\sigma^2} \psi_h W_h - \varphi_h^2\}^{1/2}}$$

et la formule pour n est

$$n = NW_L + \frac{\left\{\sum_{h=1}^{L-1} (e^{\sigma^2} \psi_h W_h - \varphi_h^2)^{1/2}\right\}^2}{\left(\sum x_i^\beta / N\right)^2 c^2 + \sum_{h=1}^{L-1} (e^{\sigma^2} \psi_h - \varphi_h^2 / W_h) / N}$$

Les dérivées partielles nécessaires pour appliquer l'algorithme itératif de Sethi (1963) sont

$$\frac{\partial n}{\partial W_h} = \frac{Ae^{\sigma^2} \psi_h / (e^{\sigma^2} \psi_h W_h - \varphi_h^2)^{1/2}}{F} - \frac{A^2 (\varphi_h / W_h)^2 / N}{F^2}$$

$$\frac{\partial n}{\partial \varphi_h} = \frac{-2A\varphi_h / \{e^{\sigma^2} W_h \psi_h - \varphi_h^2\}^{1/2}}{F} + \frac{2A^2 \varphi_h / (W_h N)}{F^2}$$

$$\frac{\partial n}{\partial \psi_h} = \frac{e^{\sigma^2} A W_h / \{e^{\sigma^2} W_h \psi_h - \varphi_h^2\}^{1/2}}{F} - e^{\sigma^2} \frac{A^2 / N}{F^2},$$

où

$$A = \sum_{h=1}^{L-1} (e^{\sigma^2} \psi_h W_h - \varphi_h^2)^{1/2},$$

et

$$F = \left(\sum x_i^\beta / N \right)^2 c^2 + \sum_{h=1}^{L-1} (e^{\sigma^2} \psi_h - \varphi_h^2 / W_h) / N.$$

6. Considération numériques

Slanta et Krenzke (1994, 1996) ont éprouvé des difficultés d'ordre numérique lors de l'utilisation de l'algorithme de Lavallée et Hidioglou avec la répartition optimum de Neyman : la convergence était lente et l'algorithme ne convergait pas toujours vers la valeur minimale réelle de n . En effet, Schneeberger (1979) et Slanta et Krenzke (1994) ont montré que, pour une population bimodale particulière, le problème présente un col; autrement dit, les dérivées partielles sont toutes nulles aux limites b_h qui ne donnent pas de valeur minimale réelle de n .

Lors de l'utilisation des algorithmes présentés ici, nous avons aussi éprouvé les difficultés numériques évoquées dans Slanta et Krenzke (1994). Ceux construits sous la répartition puissance sont généralement plus stables que ceux construits sous la répartition optimum de Neyman; les difficultés numériques sont fréquentes lorsque le nombre L de strates est grand. En outre, à mesure que la distribution de Y s'écarte de celle de X , c'est-à-dire à mesure que σ^2 augmente, la non-convergence de l'algorithme et l'impossibilité d'atteindre la valeur minimale globale de n deviennent plus fréquentes. Dans ces situations, les valeurs de départ de l'algorithme de stratification jouent un rôle de la plus grande importance. Par exemple, dans le tableau 2, le plan d'échantillonnage tenant compte des divergences entre Y et X obtenu dans le cas de la répartition optimum de Neyman dépend fortement des valeurs de départ. L'algorithme présenté au tableau 2 a comme valeurs de départ les limites présentées au tableau 2 pour la répartition puissance. En choisissant pour valeurs de départ de l'algorithme les limites obtenues au tableau 1 pour l'algorithme de Lavallée et

Hidioglou avec répartition optimum de Neyman, nous obtenons un plan d'échantillonnage différent pour lequel $n = 29$.

Une bonne stratégie de calcul consiste à exécuter l'algorithme de stratification pour plusieurs plans d'échantillonnage intermédiaires afin d'obtenir un plan d'échantillonnage final, en utilisant les limites de strate obtenues lors d'une étape comme valeurs de départ de l'algorithme à l'étape suivante. L'application de l'algorithme log-linéaire se fait toujours en deux étapes. On commence par exécuter l'algorithme de Lavallée et Hidioglou en fixant $\sigma = 0$, puis on utilise ces limites comme valeurs de départ pour l'exécution de l'algorithme en donnant une valeur non nulle à σ . On utilise aussi comme valeurs de départ pour la répartition optimum de Neyman les limites correspondantes calculées sous répartition puissance pour une valeur de p d'environ 0,7.

7. Conclusion

Le présent article propose des généralisations de l'algorithme de stratification de Lavallée et Hidioglou qui tiennent compte d'une différence entre la variable de stratification et la variable étudiée. Deux modèles statistiques sont introduits à cette fin. La nouvelle classe d'algorithmes s'appuie sur la règle de dérivation d'une fonction composée, ou règle d'enchaînement, pour calculer les dérivées partielles et sur la méthode de Sethi (1963) pour déterminer les limites optimales de strate.

L'algorithme de stratification à modèle log-linéaire proposé dans l'article a été utilisé avec de bons résultats dans plusieurs enquêtes conçues par le Service de consultation statistique de l'Université Laval. Pour estimer la production annuelle totale de sirop d'érable, le nombre d'érables producteurs de sève par producteur représente une variable de taille pratique. Nous avons utilisé des données historiques pour estimer les paramètres du modèle log-linéaire reliant les érables producteurs de sève et le volume produit. Un autre exemple est l'estimation du déficit total au titre de l'entretien des bâtiments hospitaliers au Québec. La valeur de chaque immeuble a été choisie comme variable de stratification connue. Des experts ont estimé le déficit relatif à l'entretien des bâtiments comme étant de l'ordre de 20 % à 40 %. La résolution de $4\sigma_{\log} = \log(40\%) - \log(20\%)$ donne $\sigma_{\log} = \log(2)/4 = 0.17$ comme valeur paramétrique possible pour le modèle log-linéaire de la section 3.1. Dans ces deux exemples, le fait de tenir compte des divergences entre la variable de stratification et la variable étudiée augmente la taille de l'échantillon n d'un pourcentage acceptable et produit des estimateurs dont les c.v. estimés sont proches des c.v. cibles.

Deux fonctions SAS IML appliquant l'algorithme décrit dans le présent article, l'une pour la répartition par la

méthode puissance et l'autre pour la répartition optimum de Neyman, sont affichées sur le site web de l'auteur à <http://www.mat.ulaval.ca/pages/lpr/>. Elles permettent à l'utilisateur de spécifier les valeurs de départ des limites de strate et peuvent être utilisées pour mettre en œuvre les stratégies de calcul présentées à la section 6.

Remerciements

L'auteur remercie Nathalie Vandal et Gaétan Daigle qui ont programmé les fonctions SAS IML pour les algorithmes de stratification utilisés dans le présent article. Il remercie aussi le rédacteur en chef et l'examineur de leurs commentaires constructifs.

Bibliographie

- Anderson, D.W., Kish, L. et Cornell, R.G. (1976). Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. *Journal of the American Statistical Association*, 71, 887-892.
- Cochran, W.G. (1977). *Sampling Techniques*. Troisième édition. New York : John Wiley & Sons, Inc.
- Dalenius, T. (1952). The problem of optimum stratification in a special type of design. *Skandinavisk Aktuarietidskrift*, 35, 61-70.
- Dalenius, T., et Gurney, M. (1951). The problem of optimum stratification II. *Skandinavisk Aktuarietidskrift*, 34, 133-148.
- Dorfman, A.H., et Valliant, R. (2000). Stratification by size revisited. *Journal of Official Statistics*, 16, 139-154.
- Eckman, G. (1959). An approximation useful in univariate stratification. *Annals of Mathematical Statistics*, 30, 219-229.
- Glasser, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*, 30, 28-32.
- Godfrey, J., Roshwalb, A. et Wright, R.L. (1984). Model-based stratification in inventory cost estimation. *Journal of Business and Economic Statistics*, 2, 1-9.
- Hedlin, D. (2000). A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*, 16, 15-29.
- Hidiroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.
- Hidiroglou, M. (1994). Sampling and Estimation for Establishment Surveys: Stumbling Blocks and Progress. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 153-162.
- Hidiroglou, M.A., et Srinath, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*, 11, 397-405.
- Johnson, N.L., et Kotz, S. (1970). *Continuous Univariate Distribution-I*. New York: John Wiley & Sons, Inc.
- Lavallée, P., et Hidiroglou, M. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 35-45.
- Oslo, I.T. (1976). A comparison of approximately optimal stratification given proportional allocation with other methods of stratification and allocation. *Metrika*, 23, 15-25.
- Rivest, L.-P. (1999). Stratum jumpers: Can we avoid them? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 64-72.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.
- Schneeberger, H. (1979). Saddle points of the variance of the sample mean in stratified sampling. *Sankhyā: The Indian Journal of Statistics, Series C*, 41, 92-96.
- Serfling, R.J. (1968). Approximate optimal stratification. *Journal of the American Statistical Association*, 63, 1298-1309.
- Sethi, V.K. (1963). A note on the optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.
- Singh, R.J. (1971). Approximately optimal stratification of the auxiliary variable. *Journal of the American Statistical Association*, 66, 829-834.
- Singh, R., et Parkash, D. (1975). Optimal stratification for equal allocation. *Annals of the Institute of Statistical Mathematics*, 27, 273-280.
- Singh, R., et Sukatme, B.V. (1969). Optimum stratification. *Annals of the Institute of Statistical Mathematics*, 21, 515-528.
- Slanta, J., et Krenzke, T. (1994). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's annual Capital Expenditure Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 693-698.
- Slanta, J., et Krenzke, T. (1996). Utilisation de la méthode de Lavallée et Hidiroglou pour le calcul des limites de stratification aux fins de l'enquête annuelle sur les dépenses en capital du Bureau of the Census. *Techniques d'enquête*, 22, 65-75.
- Wang, M.C., et Aggarwal, V. (1984). Stratification under a particular Pareto distribution. *Communications in Statistics, Part A—Theory and Methods*, 13, 711-735.
- Yavada, S., et Singh, R. (1984). Optimum stratification for allocation proportional to strata totals for simple random sampling scheme. *Communications in Statistics, Part A—Theory and Methods*, 13, 2793-2806.