N° 12-001-XWF au catalogue
ISSN 1712-5685

Techniques d'enquête

Juin 2014

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

Service de renseignements 1-800-635-7943

Télécopieur 1-800-565-7757

Comment accéder à ce produit

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2014

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada (http://www.statcan.gc.ca/reference/licence-fra.html).

This publication is also available in English.

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^{s}	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans *The ISI Web of knowledge* (*Web of science*), *The Survey Statistician*, *Statistical Theory and Methods Abstracts* et *SRM Database of Social Research Methodology*, *Erasmus University*. On peut en trouver les références dans *Current Index to Statistics*, et *Journal Contents in Qualitative Methods*. La revue est également citée par *SCOPUS* sur les bases de données *Elsevier Bibliographic Databases*.

COMITÉ DE DIRECTION

Président C. Julien

Anciens présidents J. Kovar (2009-2013)
D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

Membres G. Beaudoin
S. Fortier (Gestionnaire de la production)
J. Gambino
M.A. Hidiroglou
H. Mantel

COMITÉ DE RÉDACTION

Rédacteur en chef M.A. Hidiroglou, *Statistique Canada*

Ancien rédacteur en chef J. Kovar (2006-2009)
M.P. Singh (1975-2005)

Rédacteurs associés

J.-F. Beaumont, *Statistique Canada*
J. van den Brakel, *Statistics Netherlands*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
R. Chambers, *Centre for Statistical and Survey Methodology*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
D. Haziza, *Université de Montréal*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
D. Judkins, *Abt Associates*
D. Kasprzyk, *National Opinion Research Center*
J.K. Kim, *Iowa State University*
P.S. Kott, *RTI International*
P. Lahiri, *JPSM*, *University of Maryland*
P. Lavallée, *Statistique Canada*
P. Lynn, *University of Essex*
D.J. Malec, *National Center for Health Statistics*
J. Opsomer, *Colorado State University*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
F.J. Scheuren, *National Opinion Research Center*
P. do N. Silva, *Escola Nacional de Ciências Estatísticas*
P. Smith, *Office for National Statistics*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
M. Thompson, *University of Waterloo*
D. Toth, *Bureau of Labor Statistics*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *National Opinion Research Center*
C. Wu, *University of Waterloo*
W. Yung, *Statistique Canada*
A. Zaslavsky, *Harvard University*

Rédacteurs adjoints C. Bocci, K. Bosa, C. Boulet, C. Leon, H. Mantel, S. Matthews, Z. Patak, S. Rubin-Bleuer et Y. You, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée en version électronique deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préférablement en Word au rédacteur en chef, (rte@statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca/Techniquesdenquete).

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 40, numéro 1, juin 2014

Table des matières

Articles réguliers

Benmei Liu, Partha Lahiri et Graham Kalton
Modélisation hiérarchique bayésienne de proportions dans de petits domaines pondérées par les poids de sondage 1

Kelly Cristina M. Gonçalves, Fernando A. S. Moura et Helio S. Migon
Estimation bayésienne linéraire en population finie en considérant en particulier les données catégoriques 17

Qi Dong, Michael R. Elliott et Trivellore E. Raghunathan
Une méthode non paramétrique de production de populations synthétiques qui tient compte des caractéristiques des plans de sondage complexes 33

Stephen Ash
Application de la méthode des répliques des différences successives pour estimer les variances 53

Eric Graf et Yves Tillé
Estimation de variance par linéarisation pour des indices de pauvreté et d'exclusion sociale 69

Jun Shao, Eric Slud, Yang Cheng, Sheng Wang et Carma Hogue
Propriétés théoriques et empiriques d'estimateurs par la régression fondés sur un test de décision assistés par modèle 89

Natalja Menold
L'influence de la méthode d'échantillonnage et des intervieweurs sur la réalisation de l'échantillon dans l'Enquête sociale européenne 115

Communications brèves

Daniel Manrique-Vallier et Jerome P. Reiter
Imputation multiple bayésienne pour des données catégoriques à grande échelle contenant des zéros structurels 137

Techniques d'enquête, juin 2014
Vol. 40, N° 1, pp. 1-15
Statistique Canada, n° 12-001-X au catalogue

Modélisation hiérarchique bayésienne de proportions dans de petits domaines pondérées par les poids de sondage

Benmei Liu, Partha Lahiri et Graham Kalton [1]

Résumé

L'article décrit les résultats d'une étude par simulation Monte Carlo réalisée en vue de comparer l'efficacité de quatre modèles hiérarchiques bayésiens d'estimation sur petits domaines pour estimer des proportions au niveau de l'État au moyen de données provenant d'échantillons aléatoires simples stratifiés tirés d'une population finie fixe. Deux des modèles reposent sur les hypothèses fréquentes selon lesquelles, pour chaque petit domaine échantillonné, la proportion pondérée par les poids de sondage estimée suit une loi normale et sa variance d'échantillonnage est connue. L'un de ces modèles comprend un modèle de lien linéaire et l'autre, un modèle de lien logistique. Les deux autres modèles utilisent tous deux un modèle de lien logistique et reposent sur l'hypothèse que la variance d'échantillonnage est inconnue. L'un de ces deux modèles suppose que le modèle d'échantillonnage obéit à une loi normale et l'autre, qu'il obéit à une loi bêta. L'étude montre que, pour chacun des quatre modèles, la couverture sous le plan de sondage de l'intervalle de crédibilité des proportions au niveau de l'État en population finie s'écarte considérablement du niveau nominal de 95 % utilisé pour construire les intervalles.

Mots-clés : Proportions pondérées; modélisation hiérarchique bayésienne; loi bêta; intervalle de crédibilité.

1 Introduction

Les méthodes d'estimation sur petits domaines sont souvent utilisées pour estimer les proportions d'unités ayant une caractéristique donnée dans de petits domaines. Ainsi, ces méthodes sont appliquées aux données du programme des Small Area Income and Poverty Estimates (SAIPE) du Census Bureau pour estimer les taux de pauvreté au niveau de l'État, du comté et du district scolaire (Citro et Kalton, 2000; Maples et Bell 2005), aux données de la National Survey on Drug Use and Health (NSDUH) pour estimer les taux de consommation de drogues au niveau de l'État (Wright, Sathe et Spagnola 2007), et aux données de la National Assessment of Adult Literacy (NAAL) pour estimer les proportions de personnes ayant le plus faible niveau d'alphabétisation au niveau de l'État et du comté (Mohadjer, Rao, Liu, Krenzke et Van De Kerckhove 2012). Dans chaque cas, les tailles d'échantillon dans les petits domaines sont trop faibles pour produire des estimations directes d'une précision suffisante. Une grande variété de méthodes ont été élaborées pour résoudre ces problèmes d'estimation sur petits domaines. Voir Rao (2003) et Jiang et Lahiri (2006a) pour une revue de ces méthodes, et Chattopadhyay, Lahiri, Larsen et Reimnitz (1999), Farrell, MacGibbon and Tomberlin (1997) ainsi que Malec, Sedransk, Moriarity et LeClere (1997) et Malec, Davis et Cao (1999) pour des méthodes conçues spécialement pour estimer les proportions dans les petits domaines. La gamme de méthodes englobe l'approche hiérarchique bayésienne (HB) et celle du meilleur prédicteur empirique (MPE), ainsi que les modèles élaborés au niveau du domaine et au niveau de l'unité. Nous nous concentrons ici sur les modèles HB au niveau du domaine.

Lorsqu'on fait appel à un modèle HB au niveau du domaine pour produire des estimations des proportions d'unités possédant une caractéristique donnée dans de petits domaines, on suppose

1. Benmei Liu, Division of Cancer Control and Population Sciences, National Cancer Institute, 9609 Medical Center Drive Room 4E524, Bethesda, Maryland 20892; Courriel : liub2@mail.nih.gov; Partha Lahiri, JPSM, University of Maryland, 1218 Lefrak Hall, College Park, Maryland 20742; Graham Kalton, Westar, 1600 Research Boulevard, Rockville, Maryland 20850. Une grande partie de ce travail de recherche a été effectué alors que la première auteure était étudiante au cycle supérieur du *Joint Program in Survey Methodology* de l'Université du Maryland.

habituellement que, pour chaque domaine échantillonné, la proportion pondérée par les poids de sondage suit une loi d'échantillonnage normale et que sa variance d'échantillonnage est connue. Cependant, ces hypothèses posent problème lorsque la taille de l'échantillon de petit domaine est petite ou que la proportion réelle est proche de 0 ou de 1. Afin de pouvoir appliquer le théorème central limite pour approximer la normalité de la distribution d'échantillonnage d'une proportion, les échantillons doivent être raisonnablement grands, particulièrement si la proportion dans la population est très petite ou très grande (p. ex., inférieure à 0,1 ou supérieure à 0,9). En outre, quand les proportions sont très petites ou très grandes, la variance d'échantillonnage d'une proportion dans l'échantillon est extrêmement sensible à la valeur réelle de la proportion, de sorte qu'il est difficile d'établir une valeur appropriée pour la variance d'échantillonnage. Pour tenter de contourner ces problèmes, nous proposons deux modèles de rechange pour estimer les proportions dans les petits domaines et nous les comparons à deux modèles utilisés fréquemment. Les modèles sont décrits à la section 3. Les quatre modèles sont comparés au moyen d'une étude par simulation Monte Carlo dans laquelle des échantillons aléatoires simples stratifiés sont générés à partir d'une population finie fixe. L'étude par simulation est décrite à la section 4 et les résultats sont présentés à la section 5. Certaines conclusions sont formulées à la section 6. Cependant, pour commencer, nous présentons la notation pour un plan de sondage aléatoire simple stratifié à la section 2.

2 Notation

Soit N_{ih} la taille de la population dans la strate h dans le domaine i d'une population finie ($i = 1,...,m;\ h = 1,...,H_i$). Soit y_{ihk} la réponse binaire pour la caractéristique d'intérêt pour l'unité k dans la strate h dans le domaine i $(k = 1,...,N_{ih})$. Les paramètres à estimer sont les proportions dans les petits domaines $P_i = \sum_h \sum_k y_{ihk} / N_{ih}$.

Sous le plan de sondage aléatoire simple stratifié choisi pour l'étude, n_{ih} unités sont sélectionnées parmi les N_{ih} unités dans la strate (ih). L'estimateur direct classique pour P_i est :

$$p_{iw} = \frac{\sum_h^{H_i} \sum_k^{n_{ih}} w_{ih} y_{ihk}}{\sum_h^{H_i} \sum_k^{n_{ih}} w_{ih}},\ i = 1,...,m, \tag{2.1}$$

où w_{ih} désigne le poids de sondage donné par $w_{ih} = N_{ih} / n_{ih}$.

La variance de p_{iw} peut s'exprimer sous la forme

$$VAR_{st}(p_{iw}) = \frac{P_i(1-P_i)}{n_i} DEFF_i, \tag{2.2}$$

où $DEFF_i$ est l'effet de plan qui traduit l'effet du plan de sondage complexe (Kish 1965). Pour un échantillon aléatoire simple stratifié avec fractions d'échantillonnage négligeables dans toutes les strates, l'effet de plan est donné approximativement par :

$$DEFF_i = \frac{\sum_h W_{ih}^2 P_{ih}(1-P_{ih}) / n_{ih}}{P_i(1-P_i) / n_i}, \tag{2.3}$$

où $W_{ih} = N_{ih} / N_i$, $N_i = \sum_h N_{ih}$, $n_i = \sum_h n_{ih}$, et P_{ih} est la proportion dans la population de la strate h dans le domaine i.

L'effet de plan $DEFF_i$ est une fonction des P_{ih}, qui sont inconnus. Si $P_{ih}(1-P_{ih}) \approx P_i(1-P_i)$, une approximation de $DEFF_i$ peut être donnée par $deff_{iw} = n_i \sum_h W_{ih}^2 / n_{ih}$. La valeur de $deff_{iw}$ se calcule facilement, puisqu'elle ne dépend d'aucun des paramètres inconnus.

Les méthodes d'estimation sur petits domaines peuvent être utilisées pour résoudre le problème dû au fait que p_{iw} est très imprécis quand la taille de l'échantillon n_i est petite. La section 3 donne la description des modèles HB au niveau du domaine sur lesquels porte la présente étude.

3 Modèles étudiés

Un modèle général d'estimation sur petits domaines au niveau du domaine comprend deux composantes. L'une — le modèle d'échantillonnage — est un modèle pour l'erreur d'échantillonnage des estimations directes d'après l'enquête. L'autre — le modèle de lien — relie la valeur de population pour un domaine aux variables auxiliaires propres au domaine $x_i = (x_{i1}, ..., x_{ip})'$.

La section 3.1 décrit deux modèles au niveau du domaine souvent utilisés pour estimer les proportions dans les petits domaines et la section 3.2 décrit certains problèmes associés à ces modèles. La section 3.3 décrit deux autres modèles dont on peut se servir pour résoudre ces problèmes.

3.1 Deux modèles souvent utilisés

Nous étudions deux modèles souvent utilisés en vue de les comparer aux deux nouveaux modèles décrits à la section 3.4. Le premier est le modèle de Fay-Herriot (Fay et Herriot 1979), qui suppose que les variances d'échantillonnage sont connues et que les variables suivent des lois normales dans le modèle d'échantillonnage ainsi que dans le modèle de lien. Le second est le modèle normal logistique, qui ne diffère du modèle de Fay-Herriot que par le remplacement de la loi normale par une loi logit-normale dans le modèle de lien.

Modèle 1 : (modèle normal-normal de Fay-Herriot)

Modèle d'échantillonnage :

$$p_{iw} \mid P_i \overset{ind}{\sim} N(P_i,\ \psi_i) \tag{3.1}$$

Modèle de lien :

$$P_i \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x_i^{'}\beta, \sigma_v^2) \tag{3.2}$$

Modèle 2 : (modèle normal-logistique)

Modèle d'échantillonnage :

$$p_{iw} \mid P_i \overset{ind}{\sim} N(P_i,\ \psi_i) \tag{3.3}$$

Modèle de lien :

$$g(P_i) \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x_i^{'}\beta, \sigma_v^2) \tag{3.4}$$

Les deux modèles reposent sur l'hypothèse que la variance d'échantillonnage ψ_i est connue. Le modèle 1 est considéré comme un modèle apparié parce que les modèles d'échantillonnage et de lien peuvent être combinés pour produire un modèle mixte linéaire relativement simple. Cependant, on préfère souvent utiliser un modèle de lien non linéaire pour modéliser les proportions, ce qui donne des modèles d'échantillonnage et de lien non appariés, comme dans le modèle 2 (voir, par exemple, You et Rao 2002). La fonction lien $g(\cdot)$ peut être déterminée empiriquement en vérifiant l'adéquation du modèle. Les fonctions lien *log* et *logit* ont été utilisées. Le modèle de lien $\text{logit}(P_i)$ est choisi ici afin d'être certain que l'estimation de P_i soit toujours comprise dans l'intervalle de valeurs permises (0, 1).

3.2 Problèmes posés par les modèles 1 et 2

Deux problèmes importants sont associés aux modèles 1 et 2. Le premier tient au fait que ces modèles reposent tous deux sur l'hypothèse que les variances d'échantillonnage ψ_i sont connues, alors qu'en pratique elles doivent être estimées. Une approche simple consiste à utiliser l'estimation directe de la variance, mais celle-ci est très imprécise quand la valeur de P_i est très petite ou très grande et que la taille de l'échantillon n_i est petite. Une autre approche, plus complexe, consiste à élaborer une estimation approximative de P_i, disons p_{isyn}, en partant d'un modèle simple, tel qu'un modèle logistique pour p_{iw} en fonction des variables auxiliaires, puis à utiliser cette estimation dans l'estimateur de variance synthétique suivant :

$$\text{var}_{stsyn} = \frac{p_{isyn}(1 - p_{isyn})}{n_i} deff_{iw}. \tag{3.5}$$

Lorsqu'on ne dispose pas de variables auxiliaires, la proportion dans l'échantillon global peut être utilisée pour p_{isyn} dans le calcul de l'estimateur de variance synthétique.

Le deuxième problème a trait à l'hypothèse de normalité dans le modèle d'échantillonnage, qui est fondée sur une approximation en grand échantillon. Comme il est mentionné à la section 1, lorsque la taille de l'échantillon n_i est petite et que P_i est proche de 0 ou de 1, ce qui est fréquent dans le cas de l'estimation sur petits domaines, cette hypothèse pose problème.

3.3 Deux modèles de rechange

Sous les modèles 1 et 2, les variances d'échantillonnage inconnues ψ_i sont estimées, puis les estimations résultantes sont traitées comme s'il s'agissait des valeurs vraies connues. Une autre approche possible consiste à considérer que les ψ_i sont des paramètres inconnus dans le modèle HB, comme cela a été fait dans un certain nombre d'études. Par exemple, Arora et Lahiri (1997) ont appliqué un modèle HB pour modéliser les variances sous le plan de sondage pour les estimations sur échantillon. Singh, Folsom et Vaish (2005) ont proposé d'utiliser un modèle à effet de plan généralisé pour lisser la matrice de

covariance d'échantillonnage dans la modélisation pour petits domaines au moyen de données d'enquête. Récemment, You (2008) a proposé d'utiliser des effets de plan égaux au cours du temps pour modéliser les variances d'échantillonnage dans l'estimation des taux de chômage dans les petits domaines en utilisant un modèle log-linéaire transversal et chronologique. Dans le modèle 3, en tant que variante du modèle 2, les variances d'échantillonnage ψ_i sont traitées comme étant inconnues. Afin de tenir compte de la non-normalité des distributions d'échantillonnage des proportions de petit domaine pondérées par les poids de sondage, l'hypothèse d'une loi normale peut-être remplacée par celle d'une autre loi. Cette approche est appliquée dans le modèle 4 en faisant l'hypothèse d'une distribution d'échantillonnage bêta, c'est-à-dire une distribution qui possède la propriété désirable d'avoir l'intervalle de valeurs (0,1). Pour le reste, le modèle 4 est semblable au modèle 3, y compris le traitement des ψ_i, $i = 1,...,m$ comme des paramètres inconnus. Le modèle 4 a été pris en considération antérieurement par Jiang et Lahiri (2006b) à titre d'exemple de l'estimation des moyennes de domaines de population finie en utilisant l'approche du meilleur prédicteur empirique.

Modèle 3 (modèle normal-logistique avec variance d'échantillonnage inconnue) :

Modèle d'échantillonnage :

$$p_{iw} \mid P_i \overset{ind}{\sim} N(P_i,\ \psi_i) \tag{3.6}$$

Modèle de lien :

$$logit(P_i) \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x'\beta, \sigma_v^2) \tag{3.7}$$

Modèle 4 : (modèle bêta-logistique avec variance d'échantillonnage inconnue)

Modèle d'échantillonnage :

$$p_{iw} \mid P_i \overset{ind}{\sim} beta(a_i, b_i) \tag{3.8}$$

Modèle de lien :

$$logit(P_i) \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x_i'\beta, \sigma_v^2) \tag{3.9}$$

La fonction de variance approximative utilisée dans les modèles 3 et 4 est $\psi_i = [P_i(1 - P_i)/n_i]deff_{iw}$. Dans le modèle 4, les paramètres a_i et b_i sont donnés par :

$$a_i = P_i\left(\frac{n_i}{deff_{iw}} - 1\right), \text{ et } b_i = (1 - P_i)\left(\frac{n_i}{deff_{iw}} - 1\right).$$

Les estimations HB sur petits domaines peuvent être calculées au moyen des quatre modèles en utilisant l'algorithme de Metropolis-Hastings dans l'échantillonneur de Gibbs. Une description détaillée de l'algorithme, qui tire des échantillons aléatoires en se basant sur les lois conditionnelles complètes des paramètres inconnus à partir d'un ou de plusieurs jeux de valeurs initiales, figure dans Robert et Casella

(1999) et dans Chen, Shao et Ibraham (2000). You et Rao (2002) décrivent également en détail comment l'algorithme de Metropolis-Hastings fonctionne dans l'échantillonneur de Gibbs pour des modèles similaires aux modèles 1 et 2. Pour les modèles 3 et 4, l'algorithme fonctionne de la même façon que pour le modèle 2. Les distributions conditionnelles complètes sous chaque modèle sont données à l'annexe A.

4 Étude par simulation

4.1 Population étudiée et plan de sondage

La présente section décrit l'étude par simulation réalisée pour comparer l'efficacité des estimations sur petits domaines fournies par les quatre modèles HB. L'étude avait pour base le fichier de données à grande diffusion sur la natalité de 2002 qui englobait toutes les naissances survenues aux États-Unis cette année-là. Le fichier contenait des données extraites des certificats produits pour les naissances enregistrées dans chaque État et territoire (pour des renseignements détaillés, voir U.S. National Center for Health Statistics, 2009).

La population finie étudiée a été limitée aux 4 024 378 enregistrements de naissances vivantes ayant eu lieu en 2002 dans les 50 États des États-Unis et dans le district de Columbia (DC) pour lesquelles le poids à la naissance était consigné. Le paramètre d'intérêt était le taux de faible poids à la naissance au niveau de l'État P_i, $i = 1,...,51$, où le faible poids de naissance est défini comme étant moins de 2 500 grammes. La valeur de P_i variait de 5 % à 11 % selon l'État.

Dans chaque État, un plan EAS stratifié a été utilisé pour tirer des échantillons d'enregistrements de naissance. La race de la mère (blanche, noire et autre) a servi de variable de stratification. La taille de l'échantillon national a été fixée à environ 1 500 enregistrements de naissance pour chaque groupe racial. Une fraction d'échantillonnage uniforme a été appliquée à tous les États pour chaque groupe racial, sous la contrainte que deux enregistrements de naissance soient échantillonnés pour chaque groupe racial dans chaque État. La taille de l'échantillon national résultant était de $n = 4\,526$ enregistrements de naissance. Les tailles d'échantillon au niveau de l'État n_i variaient de 7 (pour les petits États tel que le Vermont) à 690 (pour la Californie), avec une taille d'échantillon médiane de 61. Cette procédure d'échantillonnage a été répétée $R = 1\,000$ fois, pour créer 1 000 jeux de données d'échantillon indépendants. Les poids d'échantillonnage sont restés les mêmes pour les diverses exécutions de la simulation.

4.2 Calcul des estimations HB

Pour simplifier, les hypothèses qui suivent ont été faites pour les modèles HB :

1. Aucune variable auxiliaire n'a été utilisée, de sorte que $x_i'\beta = \mu$.
2. Pour les modèles 1 et 2, on a posé que les variances d'échantillonnage étaient données par $\psi_i = [p_w(1 - p_w)/n_i]deff_{iw}$, où $p_w = \sum\sum\sum w_{ih}y_{ihk} / \sum\sum n_i w_{ih}$ est l'estimation nationale de la proportion de naissances vivantes avec faible poids de naissance. (Une vérification de l'utilisation de $deff_{iw}$ comme approximation de $DEFF_i$ a montré que l'approximation était raisonnable : les

deux quantités étaient proches, avec un coefficient de corrélation de Pearson de 0,96 et un ratio moyen de 1,08 entre $deff_{iw}$ et $DEFF_i$.)

3. Loi a priori uniforme pour μ, c.-à-d. $f(\mu) \propto 1$, et gamma inverse pour σ_v^2, c.-à-d. $\sigma_v^2 \sim IG(0{,}001;\ 0{,}001)$.

Pour chaque jeu de données d'échantillon, la première étape consistait à calculer les estimations directes sur échantillon au niveau de l'État. Les estimations pour chaque jeu de données d'échantillon ont ensuite été utilisées tour à tour comme entrée dans le logiciel WinBUGS (Lunn, Thomas, Best et Spiegelhalter 2000) employé pour produire les estimations HB pour les quatre modèles.

Dans un nombre important d'États où n_i était petit, les estimations directes étaient nulles dans certains jeux de données d'échantillon. Puisque WinBUGS ne peut traiter les estimations directes nulles que pour le modèle 1, pour les autres modèles, les estimations directes nulles ont été perturbées afin de les transformer en nombres positifs très petits.

Pour chaque exécution de WinBUGS, on s'est servi de trois chaînes indépendantes. Pour chaque chaîne, on a effectué un rodage de 10 000 échantillons, qui a été suivi de 10 000 autres échantillons. Le nombre d'échantillons après rodage a été réduit d'un facteur deux afin de diminuer l'autocorrélation des échantillons MCMC. Les 15 000 échantillons MCMC résultant des trois chaînes après rodage ont ensuite servi à calculer la moyenne et les centiles a posteriori pour chaque modèle HB à partir de chaque jeu de données d'échantillon. Le facteur de réduction d'échelle possible (*potential scale reduction factor*) $\hat{R}$ a été utilisé comme mesure principale de la convergence (voir Gelman et Rubin 1992). Le code WinBUGS est donné à l'annexe B.

5. Résultats des simulations

À la section 5.1, nous présentons les principaux résultats relatifs aux intervalles de crédibilité obtenus pour les proportions de naissances vivantes avec faible poids de naissance au niveau de l'État en appliquant chacun des quatre modèles. Ensuite, à la section 5.2, nous examinons les biais et les racines carrées des erreurs quadratiques moyennes de ces estimations.

5.1 Estimations des modèles et intervalles de crédibilité

Soit P_i^{HB} un estimateur HB de P_i, le pourcentage de naissances vivantes avec faible poids de naissance dans l'État i, et soit $P_{i,q}^{HB}$ le q^e centile de la distribution a posteriori de P_i. Fondé sur les résultats des 1 000 jeux de données de simulation, le tableau 5.1 donne les résultats qui suivent pour chaque modèle : la probabilité de non-couverture des intervalles de crédibilité à 95 % de P_i , c.-à-d. la probabilité que l'intervalle allant de $P_{i;0,025}^{HB}$ à $P_{i;0,975}^{HB}$ ne contienne pas P_i, et la largeur moyenne des intervalles de crédibilité $P_{i;0,975}^{HB} - P_{i;0,025}^{HB}$. Les erreurs-types de simulation Monte Carlo correspondantes sont également présentées entre parenthèses dans le tableau.

Pour examiner l'effet de la taille de l'échantillon de l'État sur les résultats des simulations, les 50 États et le district de Columbia sont répartis en trois groupes en fonction de la taille de leur échantillon, à savoir

les 15 États dont l'échantillon est de petite taille $(n_i \leq 30)$, les 24 États dont l'échantillon est de taille moyenne $(30 < n_i \leq 100)$, et les 12 États dont l'échantillon est de grande taille $(n_i > 100)$. Les résultats présentés au tableau 5.1 sont les moyennes globales sur l'ensemble des États et les moyennes pour les trois groupes distincts.

L'examen de la moitié supérieure du tableau 5.1 montre que les intervalles de crédibilité pour le modèle de Fay-Herriot (M1) sont très prudents, ne donnant quasiment aucune non-couverture. La moitié inférieure du tableau révèle que ce résultat est obtenu au prix de la plus grande largeur moyenne de l'intervalle de crédibilité parmi les quatre modèles. La largeur des intervalles de crédibilité de M1 est très stable. Une faible proportion de ces intervalles possède une borne inférieure négative.

Le faible niveau de non-couverture observé pour M1 pourrait tenir au fait que les variances d'échantillonnage ont été surestimées, peut-être parce que $deff_{iw}$ a été utilisé au lieu de $DEFF_i$. Pour examiner cette possibilité, nous avons utilisé $DEFF_i$ pour calculer la variance d'échantillonnage et nous n'avons constaté presque aucune différence de taux de non-couverture. Nous avons également exécuté le modèle en nous servant de la vraie variance définie dans (2.2) et n'avons de nouveau observé aucune différence appréciable entre les taux de non-couverture. La non-normalité de la distribution d'échantillonnage de p_{iw} pourrait également être à l'origine de ce problème.

Tableau 5.1
Pourcentage de fois que les intervalles de crédibilité à 95 % ne contiennent pas P_i, largeur moyenne des intervalles de crédibilité à 95 %, avec les erreurs-types de simulation Monte Carlo fondées sur 1 000 simulations (en pourcentage)

Taille de l'échantillon de l'État n_i	M1*	M2	M3	M4
	Pourcentage de non-couverture (erreur-type de simulation Monte Carlo)			
Échantillon global	0,40 (0,028)	8,24 (0,109)	6,52 (0,101)	4,36 (0,088)
$n_i \leq 30$ (15 États)	0,05 (0,019)	11,39 (0,239)	8,45 (0,216)	6,21 (0,190)
$30 < n_i \leq 100$ (24 États)	0,46 (0,043)	9,44 (0,167)	7,61 (0,156)	4,52 (0,132)
$n_i > 100$ (12 États)	0,70 (0,076)	1,91 (0,122)	1,94 (0,124)	1,74 (0,119)
	Largeur moyenne de l'intervalle de crédibilité à 95 % (erreur-type de simulation Monte Carlo)			
Échantillon global	9,05 (0,004)	5,52 (0,009)	6,20 (0,009)	8,45 (0,014)
$n_i \leq 30$ (15 États)	10,27 (0,009)	5,94 (0,020)	6,78 (0,021)	9,30 (0,034)
$30 < n_i \leq 100$ (24 États)	9,16 (0,005)	5,60 (0,013)	6,28 (0,013)	8,71 (0,021)
$n_i > 100$ (12 États)	7,29 (0,004)	4,84 (0,012)	5,30 (0,013)	6,88 (0,017)

*Nota : Pour le modèle 1, une faible proportion d'intervalles de crédibilité possède une borne inférieure négative.

À 8,2 %, le taux global de non-couverture des intervalles de crédibilité pour le modèle normal-logistique (M2) est sensiblement supérieur au taux nominal de 5 %. Ce modèle donne la plus petite largeur moyenne de l'intervalle. Le taux de non-couverture pour le modèle normal-logistique avec variance

inconnue (M3) s'approche davantage du taux nominal, avec une largeur globale de l'intervalle un peu plus grande que dans le cas de M2.

Le taux global de non-couverture de 4,4 % observé pour le modèle bêta-logistique (M4) est celui qui est le plus proche du taux de non-couverture nominal. Cependant, la largeur moyenne des intervalles de crédibilité est plus grande que celle obtenue pour M2 et M3, et l'erreur-type Monte Carlo de la largeur de l'intervalle est plus grande que celle observée pour les trois autres modèles. Cette instabilité pourrait découler de la complexité de la distribution conditionnelle complète pour le modèle bêta. La forte proportion des 1 000 estimations directes qui étaient nulles pour certains États dont la taille d'échantillon était petite pourrait aussi avoir causé d'importants problèmes d'ajustement de la loi bêta.

Comme prévu, pour les quatre modèles, la largeur moyenne des intervalles de crédibilité diminue lorsque la taille de l'échantillon de l'État augmente, et la variation de la largeur diminue également lorsque la taille de l'échantillon augmente. Toutefois, malgré ces diminutions, les taux de non-couverture diminuent aussi lorsque la taille de l'échantillon augmente pour les modèles 2, 3 et 4. En fait, les taux de non-couverture sont très petits pour les États dont la valeur de n_i est grande, ce qui fait penser que les intervalles de crédibilité ne reflètent pas adéquatement l'effet de la plus grande précision des estimations directes dans les États où la taille d'échantillon est grande.

5.2 Biais et REQM des estimations fondées sur les modèles

Afin d'étudier ces résultats plus en détail, nous avons examiné le biais et la racine carrée de l'erreur quadratique moyenne (REQM) des estimations P_i^{HB} pour chaque modèle. Les résultats sont présentés au tableau 5.2 dans le même format qu'au tableau 5.1. Les biais des estimations sous les modèles M1, M2 et M3 présentent une tendance similaire : les biais observés pour les petits États sont grands et positifs, et sont compensés dans une certaine mesure par les biais négatifs relativement faibles observés pour les États de moyenne et de grande taille. Dans le cas du modèle M4, les biais des estimations présentent une tendance très différente : ils sont presque nuls pour les petits États et leur valeur est grande et négative pour les États de moyenne et de grande taille. Cela indique que M4 donnerait de meilleurs résultats que les trois autres modèles en ce qui concerne le biais lorsque les échantillons des petits domaines sont de petite taille.

Tableau 5.2
Biais et racine carrée de l'erreur quadratique moyenne des estimations de P_i fondées sur les quatre modèles (en pourcentage)

Taille de l'échantillon de l'État n_i	M1		M2		M3		M4	
	Biais	REQM	Biais	REQM	Biais	REQM	Biais	REQM
Échantillon global	0,165	1,518	0,071	1,346	-0,009	1,411	-0,214	1,712
$n_i \leq 30$ (15 États)	0,621	1,651	0,572	1,630	0,466	1,652	0,009	1,922
$30 < n_i \leq 100$ (24 États)	-0,006	1,547	-0,123	1,386	-0,201	1,452	-0,319	1,775
$n_i > 100$ (12 États)	-0,063	1,294	-0,167	0,911	-0,219	1,026	-0,283	1,323

6. Discussion

Le présent article décrit les résultats d'une étude par simulation réalisée à partir d'une population finie réelle pour évaluer les intervalles de crédibilité produits par quatre modèles hiérarchiques, en se basant sur leur longueur et sur leurs propriétés de couverture sous le plan de sondage. Autant que nous sachions, ce genre d'évaluation des intervalles de crédibilité (ou de confiance) fondée sur le plan de sondage n'a encore jamais été effectuée dans le contexte de l'estimation sur petits domaines.

Dans l'étude par simulation, nous avons comparé les propriétés de couverture sous le plan de sondage des intervalles de crédibilité résultant de divers modèles hiérarchiques bayésiens élaborés pour estimer les proportions dans de petits domaines sous un plan de sondage aléatoire simple stratifié. Globalement, aucun modèle ne l'emporte clairement sur les autres, si bien que nous ne pouvons pas en recommander un plutôt que les autres.

La version hiérarchique bayésienne du modèle bien connu de Fay-Herriot semble produire des intervalles de crédibilité exagérément prudents. La non-normalité du modèle d'échantillonnage ainsi que du modèle de lien pourrait être la source de ce problème. Pour le modèle hiérarchique bêta-logistique, les intervalles de crédibilité donnent presque la couverture nominale pour les proportions de population finie et, des quatre modèles étudiés, ce modèle est celui qui possède les meilleures propriétés de biais quand la taille d'échantillon est petite. Cependant, comme l'une des distributions conditionnelles complètes pour le modèle bêta-logistique fait intervenir les proportions pondérées par les poids de sondage, la méthode MCMC pose problème chaque fois qu'une de ces proportions est nulle. Les intervalles de crédibilité pour ce modèle sont également plus larges que ceux observés pour les deux autres modèles comportant un modèle de lien logistique. La largeur de l'intervalle de crédibilité sous le modèle bêta-logistique pourrait peut-être être réduite si le modèle est modifié, par exemple en utilisant un modèle de mélange de lois à effet aléatoire en deux parties approprié qui permettrait d'éviter le problème des proportions pondérées par les poids de sondage nulles. Une étude plus approfondie à ce sujet est nécessaire. Il pourrait aussi être utile d'envisager d'autres modèles, peut-être un modèle probabiliste discret pour le niveau 1, en vue d'améliorer l'estimation des intervalles des petites proportions pour les petits domaines.

L'étude par simulation a montré que la couverture des intervalles de crédibilité bayésiens des proportions dans la population finie s'écartait considérablement du niveau nominal de 95 % pour les quatre modèles, et une constatation semblable a été faite pour la couverture sous le plan de sondage du modèle de Fay-Herriot dont l'usage est très répandu. Compte tenu de ces résultats, nous avons effectué un certain nombre d'analyses supplémentaires en vue de trouver une explication. Ces analyses comprenaient l'ajout de variables prédictives aux modèles, l'utilisation d'une loi a priori uniforme pour σ_v^2 (fondée sur des arguments formulés par Gelman 2006), l'utilisation de l'approche de la meilleure prédiction empirique pour le modèle M1, l'accroissement de la taille d'échantillon dans les États ne comptant que quelques naissances en fixant le nombre minimum à 50, et l'application de ces méthodes pour estimer la proportion de naissances dont le poids à la naissance était inférieur à la médiane nationale dans chaque État. Même si les propriétés de couverture des intervalles des proportions dans la population finie au niveau de l'État présentaient certaines différences, aucune de ces analyses n'a produit des taux de couverture proches des taux nominaux. Le seul cas où ces derniers coïncidaient avec les taux de couverture réels était celui d'un jeu de données simulé construit sous le modèle M1 pour les proportions au niveau de l'État de naissances pour lesquelles le poids de naissance était inférieur à la médiane nationale; les taux de couverture moyens

étaient de 5,1 % et de 5,2 % pour les approches du meilleur prédicteur empirique (MPE) et HB, respectivement.

L'étude par simulation a été limitée à un plan de sondage à un seul degré. En outre, pour simplifier, aucune variable auxiliaire n'a été incluse dans les modèles de lien dans les analyses principales, alors qu'en pratique, l'ajout de ce genre de variables est habituel et presque essentiel. D'autres études par simulation doivent être réalisées en vue d'examiner différents plans de sondage et différentes tailles d'échantillon, et d'intégrer certaines variables auxiliaires dans les modèles de lien. Nous espérons que notre étude encouragera d'autres chercheurs à exécuter des simulations fondées sur le plan de sondage similaires pour évaluer les méthodes d'estimation sur petits domaines. À la lumière de nos résultats limités, il convient de mettre en garde les utilisateurs d'estimations sur petits domaines quant à l'interprétation des intervalles de crédibilité des estimations.

Remerciements

Les auteurs tiennent à remercier les rédacteurs associés ainsi que deux réviseurs pour leurs suggestions constructives ayant permis d'améliorer significativement l'article original. Les travaux de recherche du deuxième auteur ont été financés par la *National Science Foundation* SES-085100.

Annexe A

A1. Distributions conditionnelles complètes pour les paramètres de chaque modèle

Soit $\vec{p} = (p_{1w}, ..., p_{mw})^t$ et $r_i = \dfrac{\psi_i}{\psi_i + \sigma_v^2}$.

Les distributions conditionnelles complètes pour le modèle de Fay-Herriot (M1) sont les suivantes :

i) $\theta_i \mid \mu, \sigma_v^2, \vec{p} \sim N((1-r_i)p_{iw} + r_i\mu, \quad \psi_i(1-r_i))$;

ii) $\mu \mid \theta_i, \sigma_v^2, p \sim N\left(\dfrac{1}{m}\sum_{i=1}^{m}\theta_i, \dfrac{\sigma_v^2}{m}\right)$;

iii) $\sigma_v^2 \mid \mu, \theta_i, \vec{p} \sim ING\left(a + \dfrac{1}{2}m, b + \dfrac{1}{2}\sum_{i=1}^{m}(\theta_i - \mu)^2\right)$.

Les distributions conditionnelles complètes pour le modèle normal-logistique (M2) sont les suivantes :

i) $\theta_i \mid \mu, \sigma_v^2, \vec{p} \propto \dfrac{1}{\theta_i(1-\theta_i)\sigma_v\sqrt{\psi_i}} \exp\left(-\dfrac{(p_{iw} - \theta_i)^2}{2\psi_i} - \dfrac{(\text{logit}(\theta_i) - \mu)^2}{2\sigma_v^2}\right)$;

ii) $\mu \mid \theta_i, \sigma_v^2, p \sim N\left(\dfrac{1}{m}\sum_{i=1}^{m}\text{logit}(\theta_i), \ \dfrac{\sigma_v^2}{m}\right)$;

iii) $\sigma_v^2 \mid \mu, \theta_i, \vec{p} \sim ING\left(a + \frac{1}{2}m, b + \frac{1}{2}\sum_{i=1}^{m}(\text{logit}(\theta_i) - \mu)^2\right)$.

Les distributions conditionnelles complètes pour le modèle normal-logistique avec variance inconnue (M3) sont les mêmes que pour le modèle M2 excepté que ψ_i est remplacé par $\theta_i(1-\theta_i)deff_{iw} / n_i$ pour la distribution de θ_i sachant les autres paramètres.

Soit $\delta_{iw} = \frac{n_i}{deff_{iw}} - 1$. Les distributions conditionnelles complètes pour le modèle bêta-logistique (M4) sont les suivantes :

i) $\theta_i \mid \mu, \sigma_v^2, \vec{p} \propto \frac{1}{\theta_i(1-\theta_i)\sigma_v} \frac{p_{iw}^{\theta_i\delta_{iw}-1}(1-p_{iw})^{(1-\theta_i)\delta_{iw}-1}}{\Gamma(\theta_i\delta_{iw})\Gamma((1-\theta_i)\delta_{iw})} \exp\left(-\frac{(\text{logit}(\theta_i)-\mu)^2}{2\sigma_v^2}\right)$;

ii) $\mu \mid \theta_i, \sigma_v^2, p \sim N\left(\frac{1}{m}\sum_{i=1}^{m}\text{logit}(\theta_i),\ \frac{\sigma_v^2}{m}\right)$;

iii) $\sigma_v^2 \mid \mu, \theta_i, \vec{p} \sim ING\left(a + \frac{1}{2}m, b + \frac{1}{2}\sum_{i=1}^{m}(\text{logit}(\theta_i) - \mu)^2\right)$.

Annexe B

Code WinBUGS pour le modèle 1 :

```
Model {
      for ( i in 1 :N)  {
         pobs[i] ~ dnorm(theta[i], D[i])
         D[i] <- 1/varhat[i]
         theta[i]<-u+v[i]
         v[i]~dnorm(0, tau)
                          }
        u~dflat()
       tau~dgamma(0.001, 0.001)
       sigma_v2<-1/tau
          }
```

Code WinBUGS pour le modèle 2 :

```
Model {
      for ( i in 1 :N)  {
         pobs[i] ~ dnorm(theta[i], D[i])
         D[i] <- 1/varhat[i]
         logit(theta[i])<-u+v[i]
         v[i]~dnorm(0, tau)
                          }
```

```
u~dflat()
tau~dgamma(0.001, 0.001)
sigma_v2<-1/tau
 }
```

Code WinBUGS pour le modèle 3 :

```
Model {
      for ( i in 1 :N)  {
         pobs[i] ~ dnorm(theta[i], E[i])
         E[i] <- SAMPn[i]/(theta[i]*(1-theta[i])*DEFF_kish[i])
         logit(theta[i])<-u+v[i]
         v[i]~dnorm(0, tau)
         D[i]<-1/E[i]
                          }
       u~dflat()
       tau~dgamma(0.001, 0.001)
       sigma_v2<-1/tau
          }
```

Code WinBUGS pour le modèle 4 :

```
Model {
      for ( i in 1 :N) {
         pobs[i] ~ dbeta(a[i], b[i])
         a[i] <- theta[i]*(theta[i]*(1-theta[i])/D[i]-1)
         b[i] <- (1-theta[i])*(theta[i]*(1-theta[i])/D[i]-1)
         logit(theta[i])<-u+v[i]
         v[i]~dnorm(0, tau)
         D[i]<-theta[i]*(1-theta[i])*DEFF_kish[i]/SAMPn[i]
                          }
       u~dflat()
       tau~dgamma(0.001, 0.001)
       sigma_v2<-1/tau
          }
```

Bibliographie

Arora, V. et Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Chattopadhyay, M., Lahiri, P., Larsen, M., et Reimnitz, J. (1999). Estimation composite de la prevalence des drogues pour des zones infraétats. *Techniques d'enquetes*, 25, 91-97.

Chen, M., Shao, Q., et Ibraham, J.G. (2000). *Monte Carlo Methods in Bayesian Computation.* New York: Springer-Verlag.

Citro, C., et Kalton, G. (Eds.) (2000). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*. Washington, DC: National Academy Press.

Farrell, P.J., MacGibbon, B., et Tomberlin, T.J. (1997). Empirical Bayes estimators of small area proportions in multistage designs. *Statistical Sinica*, 7, 1065-1083.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Gelman, A., et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

Jiang, J., et Lahiri, P. (2006a). Mixed model prediction and small area estimation. *Test*, 15, 111-999.

Jiang, J., et Lahiri, P. (2006b). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Lunn, D.J., Thomas, A., Best, N., et Spiegelhalter, D. (2000). WinBUGS – A Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.

Malec, D., Davis, W., et Cao, X. (1999). Small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine*, 18, 3189-3200.

Malec, D., Sedransk, J., Moriarity, C.L., et LeClere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association,* 92, 815-826.

Maples, J., et Bell, W.R. (2005). Evaluation of school district poverty estimates: Predictive models using IRS income tax data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1322-1329.

Mohadjer, L., Rao, J.N.K., Liu, B., Krenzke, T., et Van De Kerckhove, W. (2012). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Journal of the Indian Society of Agricultural Statistics,* 66 (1), 55-63.

Rao, J.N.K. (2003). *Small area estimation*. New York: John Wiley and Sons.

Robert, C.P., et Casella, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

Singh, A.C., Folsom, R.E., Jr. et Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. *Federal Committee on Statistical Methods Statistical Working Paper* No. 39. (http://www.fcsm.gov/05papers/Singh_etal_IIIC.pdf)

U.S. National Center for Health Statistics (2009). *National Vital Statistics System*. *Birth Data*. (http://www.cdc.gov/nchs/births.htm)

Wright, D., Sathe, N., et Spagnola, K. (2007). *State Estimates of Substance Use from the 2004-2005 National Surveys on Drug Use and Health.* (DHHS Publication No. SMA 07-4235, NSDUH Series H-31). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

You, Y. (2008). Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada. *Techniques d'enquêtes*, 34, 21-30.

You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics*, 30, 3-15.

ELECTRONIC PUBLICATIONS AVAILABLE AT | PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À

www.statcan.gc.ca

Techniques d'enquête, juin 2014
Vol. 40, N° 1, pp. 17-31
Statistique Canada, N° 12-001-X au catalogue

Estimation bayésienne linéaire en population finie en considérant en particulier les données catégoriques

Kelly Cristina M. Gonçalves, Fernando A.S. Moura et Helio S. Migon[1]

Résumé

L'estimateur bayésien linéaire en population finie est obtenu en partant d'un modèle de régression à deux degrés spécifié uniquement par les moyennes et les variances de certains paramètres du modèle associés à chaque degré de la hiérarchie. Nombre d'estimateurs fondés sur le plan de sondage usuels décrits dans la littérature peuvent être obtenus en tant que cas particuliers. Un nouvel estimateur par le ratio est également proposé pour la situation pratique où de l'information auxiliaire est disponible. L'application de la même approche bayésienne linéaire est proposée pour estimer des proportions pour des données catégoriques multiples associées aux unités de la population finie, ce qui constitue la principale contribution des présents travaux, et est illustrée au moyen d'un exemple numérique.

Mots-clés : Échangeabilité; modèle linéaire; prédiction bayésienne linéaire.

1 Introduction

Les sondages sont depuis longtemps un moyen important d'obtenir de l'information exacte auprès d'une population finie. Par exemple, les gouvernements ont besoin de statistiques descriptives sur la population pour évaluer et mettre en œuvre leurs politiques. Au cours du premier tiers du XXe siècle, la principale préoccupation des spécialistes de la statistique officielle était d'établir une norme de pratique acceptable. Neyman (1934) a créé un cadre de ce type en introduisant des méthodes de randomisation dans le processus d'échantillonnage. Il préconisait l'utilisation de la distribution de randomisation, ou distribution aléatoire, induite par le plan de sondage pour évaluer les propriétés fréquentistes de diverses procédures. Il a également introduit la notion de stratification avec répartition optimale de la taille de l'échantillon et l'utilisation de probabilités de sélection inégales. Ses travaux sont considérés comme la pierre angulaire de la théorie des sondages fondée sur le plan de sondage et ont inspiré de nombreux auteurs. Par exemple, Horvitz et Thompson (1952) ont proposé une théorie générale de l'échantillonnage avec probabilités inégales et la méthode d'estimation pondérée par les probabilités, qui est connue sous le nom d'« estimateur de Horvitz-Thompson ».

Partout dans le monde, les organismes statistiques officiels ont trouvé la théorie des sondages fondée sur le plan de sondage très intéressante. Comme l'ont fait remarquer Skinner, Holt et Smith (1989), page 2, la raison principale de cet intérêt est qu'essentiellement, cette théorie ne dépend pas d'une loi de probabilité. En effet, depuis les travaux de Neyman, tous les progrès réalisés en théorie des sondages ont été fortement influencés par l'usage descriptif du sondage. Cette approche a eu pour conséquence un manque de développements théoriques concernant l'utilisation analytique des sondages, en particulier à des fins de prédiction. Dans certaines situations précises, l'approche fondée sur le plan de sondage s'est avérée inefficace, fournissant des prédicteurs inadéquats. Ainsi, l'estimation sur petits domaines et la présence de non-réponse ne peuvent pas être traitées par l'approche fondée sur le plan de sondage sans

1. Kelly Cristina M. Gonçalves, Departamento de Estatística, Universidade Federal do Rio de Janeiro (UFRJ), RJ, Brésil. Courriel : kelly@im.ufrj.br; Fernando A. S. Moura, Departamento de Estatística, Universidade Federal do Rio de Janeiro (UFRJ), RJ, Brésil. Courriel : fmoura@im.ufrj.br; Helio S. Migon, Departamento de Estatística, Universidade Federal do Rio de Janeiro (UFRJ), RJ, Brésil. Courriel : migon@im.ufrj.br.

faire appel à des hypothèses implicites, ce qui équivaut à émettre l'hypothèse d'un modèle. Les partisans de l'approche fondée sur le plan de sondage soutiennent que l'inférence fondée sur un modèle dépend en grande partie des hypothèses du modèle, lesquelles pourraient ne pas être vérifiées. Par ailleurs, l'inférence par intervalle sur les paramètres de la population cible (habituellement les totaux et les moyennes) s'appuie sur le théorème central limite, qui n'est pas applicable dans de nombreuses situations pratiques, où la taille de l'échantillon n'est pas suffisamment grande et/ou les hypothèses d'indépendance des variables aléatoires concernées ne sont pas raisonnables.

Basu (1971) rejetait les estimations des quantités de population dépendantes de la règle d'échantillonnage, comme les probabilités d'inclusion. Selon lui, cette méthode d'estimation ne satisfaisait pas le principe de vraisemblance dont il était un partisan. Basu (1971) a créé l'exemple de l'éléphant de cirque pour montrer que l'estimateur de Horvitz-Thompson pouvait produire des estimations inappropriées et a proposé un estimateur de rechange. La question qui se pose est celle de savoir s'il est possible de concilier les deux approches. Dans le contexte du modèle de superpopulation, Zacks (2002) a montré que l'on peut retrouver certains estimateurs fondés sur le plan de sondage en utilisant une approche basée sur un modèle de régression généralisée. Little (2003) affirme qu'une spécification prudente du modèle, sensible au plan de sondage, permet de répondre aux préoccupations concernant les spécifications du modèle, et que la statistique bayésienne offre un traitement cohérent et unifié de l'inférence par sondage descriptive et analytique. Il a donné certains exemples afin d'illustrer comment l'inférence fondée sur le plan de sondage classique peut découler de la perspective bayésienne en utilisant certains modèles avec priors non informatifs.

Dans le contexte bayésien, une autre proposition intéressante en vue de concilier l'approche fondée sur le plan de sondage et celle fondée sur un modèle a été avancée par Smouse (1984). La méthode incorpore l'information a priori dans les modèles d'inférence en population finie en s'appuyant sur des méthodes bayésiennes d'estimation par les moindres carrés et nécessite seulement la spécification des premier et deuxième moments de la distribution concernée, qui décrivent les connaissances a priori au sujet des structures présentes dans la population. L'approche peut remplacer les méthodes de randomisation et se situe à mi-chemin entre deux visions extrêmes, à savoir les procédures fondées sur le plan de sondage d'une part et celles fondées sur des modèles de superpopulation d'autre part. Dans un rapport inédit, O'Hagan (1985) a présenté des estimateurs bayésiens linéaires dans certains contextes de sondage particuliers et O'Hagan (1987) a également dérivé des estimateurs bayésiens linéaires pour certains modèles de réponse randomisés. O'Hagan (1985) a pris en considération plusieurs structures de population, dont la stratification et la mise en grappes, en émettant des hypothèses appropriées au sujet des premier et deuxième moments, et a montré comment obtenir certains estimateurs fondés sur le plan fréquemment utilisés en temps que cas particuliers de cette approche plus générale. Il a également fait remarquer que ses estimations ne tenaient pas compte de l'échantillonnage non informatif. Il a cité Scott (1977) et fait le commentaire que l'échantillonnage informatif devrait être exécuté par une analyse bayésienne complète. Une référence importante concernant le traitement de l'échantillonnage informatif avec des modèles hiérarchiques figure dans Pfeffermann, Moura et Silva (2006).

La présentation de l'article est la suivante. La section 2 décrit de manière générale l'approche d'estimation bayésienne linéaire appliquée à un modèle de régression linéaire général pour la prédiction en population finie et montre comment obtenir certains estimateurs fondés sur le plan de sondage à titre de cas particuliers. À la section 3, un nouvel estimateur est proposé pour la situation pratique où de

l'information auxiliaire est disponible. À la section 4, l'approche d'estimation bayésienne linéaire est étendue aux données catégoriques multiples. Enfin, la section 5 offre certaines conclusions et des suggestions de futures recherches.

2 Estimation bayésienne linéaire en population finie

L'approche bayésienne s'est avérée fructueuse dans de nombreuses applications, particulièrement lorsque l'analyse des données a été améliorée par des jugements d'expert. Cependant, si les modèles bayésiens possèdent de nombreuses caractéristiques intéressantes, leur application requiert souvent la spécification complète d'une loi a priori, ou prior, pour un grand nombre de paramètres. Goldstein et Wooff (2007), section 1.2, soutiennent que, à mesure que le problème se complexifie, notre aptitude réelle à spécifier complètement la loi a priori et/ou le modèle d'échantillonnage en détail diminue. Ils concluent que, dans de telles situations, il est nécessaire d'élaborer des méthodes qui reposent sur une spécification faisant appel à la croyance partielle.

Hartigan (1969) a proposé une méthode d'estimation, qu'il a nommée approche d'estimation bayésienne linéaire, qui ne nécessite que la spécification des premier et deuxième moments. Les estimateurs résultants ont la propriété de minimiser la perte quadratique a posteriori parmi tous les estimateurs qui sont linéaires en les données et peuvent être considérés comme des approximations des moyennes a posteriori. L'approche d'estimation bayésienne linéaire, qui est employée pleinement dans le présent article, est décrite brièvement ci-dessous.

2.1 Approche bayésienne linéaire

Soit $\mathbf{y}_s$ le vecteur des observations et $\boldsymbol{\theta}$, le paramètre à estimer. Pour chaque valeur de $\boldsymbol{\theta}$ et chaque estimation possible $\mathbf{d}$, appartenant à l'espace paramétrique $\boldsymbol{\Theta}$, nous associons une fonction de perte quadratique $L(\boldsymbol{\theta},\mathbf{d}) = (\boldsymbol{\theta}-\mathbf{d})'(\boldsymbol{\theta}-\mathbf{d}) = tr(\boldsymbol{\theta}-\mathbf{d})(\boldsymbol{\theta}-\mathbf{d})'$. Nous souhaitons avant tout trouver la valeur de $\mathbf{d}$ qui minimise $r(\mathbf{d}) = E[L(\boldsymbol{\theta},\mathbf{d})\,|\,\mathbf{y}_s]$, la valeur prévue conditionnelle de la fonction de perte quadratique fournie par les données.

Supposons que la distribution conjointe de $\boldsymbol{\theta}$ et $\mathbf{y}_s$ est spécifiée partiellement par leurs deux premiers moments seulement :

$$\begin{pmatrix}\boldsymbol{\theta}\\ \mathbf{y}_s\end{pmatrix} \sim \left[\begin{pmatrix}\mathbf{a}\\ \mathbf{f}\end{pmatrix}, \begin{pmatrix}\mathbf{R} & \mathbf{AQ}\\ \mathbf{QA}' & \mathbf{Q}\end{pmatrix}\right], \tag{2.1}$$

où $\mathbf{a}$ et $\mathbf{f}$, respectivement, désignent les vecteurs des moyennes et $\mathbf{R}$, $\mathbf{AQ}$ et $\mathbf{Q}$, les éléments de la matrice de covariance de $\boldsymbol{\theta}$ et $\mathbf{y}_s$.

L'estimateur bayésien linéaire (EBL) de $\boldsymbol{\theta}$ est la valeur de $\mathbf{d}$ qui minimise la valeur prévue de cette fonction de perte quadratique dans la classe de toutes les estimations linéaires de la forme $\mathbf{d} = \mathbf{d}(\mathbf{y}_s) = \mathbf{h} + \mathbf{H}\mathbf{y}_s$, pour un vecteur $\mathbf{h}$ et une matrice $\mathbf{H}$. Donc, l'EBL de $\boldsymbol{\theta}$, $\hat{\mathbf{d}}$, et sa variance associée, $\hat{V}(\hat{\mathbf{d}})$, sont donnés respectivement par

$$\hat{\mathbf{d}} = \mathbf{a} + \mathbf{A}(\mathbf{y}_s - \mathbf{f}) \text{ et } \hat{V}(\hat{\mathbf{d}}) = \mathbf{R} - \mathbf{AQA}'. \tag{2.2}$$

Il convient de souligner que l'EBL dépend de la spécification des premier et deuxième moments de la distribution conjointe partiellement spécifiée en (2.1). Le problème de l'obtention de ces quantités est traité aux sections 2.3.1 et 4.1 pour certains cas particuliers.

2.2 Approche bayésienne linéaire en population finie

Considérons $U = \{u_1, \ldots, u_N\}$ une population finie comprenant N unités. Soit $\mathbf{y} = (y_1, \ldots, y_N)'$ le vecteur des valeurs d'intérêt des unités dans U. Le vecteur de réponses $\mathbf{y}$ est divisé en le vecteur des valeurs connues observées sur l'échantillon $\mathbf{y}_s$ de taille n, et en le vecteur des valeurs non observées $\mathbf{y}_{\bar{s}}$ de dimension $N - n$. Le problème général consiste à prédire une fonction du vecteur $\mathbf{y}$, telle que le total $T = \sum_{i=1}^{N} y_i = \mathbf{1}_s' \mathbf{y}_s + \mathbf{1}_{\bar{s}}' \mathbf{y}_{\bar{s}}$, où $\mathbf{1}_s$ et $\mathbf{1}_{\bar{s}}$ sont les vecteurs de 1 de dimensions n et $N - n$, respectivement. Dans l'approche fondée sur un modèle, cela se fait habituellement en se servant d'un modèle paramétrique hypothétique pour les valeurs de population y_i, puis en obtenant le meilleur prédicteur linéaire sans biais empirique (EBLUP) pour le vecteur inconnu $\mathbf{y}_{\bar{s}}$ sous ce modèle. Habituellement, l'erreur quadratique moyenne de l'EBLUP de T s'obtient par approximation d'ordre deux, de même qu'un estimateur sans biais. Voir Valliant, Dorfman et Royall (2000), chapitre 2, pour des renseignements détaillés.

L'approche bayésienne de la prédiction en population finie repose souvent sur l'hypothèse d'un modèle paramétrique, mais elle vise à trouver la loi a posteriori de T sachant $\mathbf{y}_s$. On peut obtenir les estimations ponctuelles en spécifiant une fonction de perte, quoique dans de nombreux problèmes pratiques, on considère souvent la moyenne a posteriori dont la variance associée est donnée par la variance a posteriori, c'est-à-dire :

$$E(T \mid \mathbf{y}_s) = \mathbf{1}_s' \mathbf{y}_s + \mathbf{1}_{\bar{s}}' E(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s) \text{ et } V(T \mid \mathbf{y}_s) = \mathbf{1}_{\bar{s}}' V(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s) \mathbf{1}_{\bar{s}}. \tag{2.3}$$

Il est possible d'obtenir une approximation des quantités dans (2.3) en utilisant une approche d'estimation bayésienne linéaire. Ici, nous obtiendrons en particulier les estimateurs en émettant l'hypothèse d'un modèle hiérarchique à deux degrés général en population finie, spécifié uniquement par sa moyenne et sa matrice de variance-covariance, présenté dans Bolfarine et Zacks (1992), page 76. Les cas particuliers décrivant les structures de population habituellement observées en pratique peuvent être dérivés facilement de (2.4). Le modèle général peut s'écrire :

$$\mathbf{y} \mid \boldsymbol{\beta} \sim [\mathbf{X}\boldsymbol{\beta}, \mathbf{V}] \text{ et } \boldsymbol{\beta} \sim [\mathbf{a}, \mathbf{R}], \tag{2.4}$$

où $\mathbf{X}$ est une matrice de covariables de dimensions $N \times p$, avec les lignes $\mathbf{X}_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, N$; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ est un vecteur de dimension $p \times 1$ de paramètres inconnus, et $\mathbf{y}$, sachant $\boldsymbol{\beta}$, est un vecteur aléatoire de moyenne $\mathbf{X}\boldsymbol{\beta}$ et de matrice de covariance connue $\mathbf{V}$ de dimensions $N \times N$. De manière analogue, $\mathbf{a}$ et $\mathbf{R}$ sont, respectivement, le vecteur des moyennes a priori de dimension $p \times 1$ et la matrice de covariance a priori de dimensions $p \times p$ de $\boldsymbol{\beta}$.

Puisque le vecteur de réponses $\mathbf{y}$ est divisé en $\mathbf{y}_s$ et $\mathbf{y}_{\bar{s}}$, la matrice $\mathbf{X}$, qui est supposée connue, est divisée de manière analogue en $\mathbf{X}_s$ et $\mathbf{X}_{\bar{s}}$, et $\mathbf{V}$ est divisée en $\mathbf{V}_s$, $\mathbf{V}_{\bar{s}}$, $\mathbf{V}_{s\bar{s}}$ et $\mathbf{V}_{\bar{s}s}$. L'objectif premier est de prédire $\mathbf{y}_{\bar{s}}$ sachant l'échantillon observé $\mathbf{y}_s$, puis le total T. Voici les étapes que nous avons

suivies : premièrement, nous avons utilisé une loi a priori conjointe qui n'est spécifiée que partiellement en ce qui concerne les moments comme il suit :

$$\begin{pmatrix}\mathbf{y}_{\bar{s}}\\ \mathbf{y}_s\end{pmatrix}\Bigg|\boldsymbol{\beta} \sim \left[\begin{pmatrix}\mathbf{X}_{\bar{s}}\boldsymbol{\beta}\\ \mathbf{X}_s\boldsymbol{\beta}\end{pmatrix}, \begin{pmatrix}\mathbf{V}_{\bar{s}} & \mathbf{V}_{s\bar{s}}\\ \mathbf{V}_{s\bar{s}} & \mathbf{V}_s\end{pmatrix}\right].$$

Donc, en appliquant le résultat général dans l'équation (2.2), l'EBL de $E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ et la perte quadratique prévue minimale (variance associée) sont donnés par :

$$\hat{E}(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta}) = \mathbf{X}_{\bar{s}}\boldsymbol{\beta} + \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}) \text{ et } \hat{V}(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta}) = \mathbf{V}_{\bar{s}} - \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}\mathbf{V}_{s\bar{s}}. \tag{2.5}$$

Remarque 1 : Il convient de souligner que, sous l'hypothèse de normalité, $E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ et $V(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ sont données, respectivement, par les membres de droite des équations de (2.5). L'EBL et sa variance associée donnés en (2.5) peuvent être considérés respectivement, comme des approximations de $E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ et $V(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ pour les cas de non-normalité.

Maintenant, si nous revenons au modèle (2.4), nous devons adapter la structure (2.1) et utiliser les résultats de (2.2) pour obtenir l'EBL de $\boldsymbol{\beta}$ et sa variance associée, $\hat{V}(\hat{\boldsymbol{\beta}})$, donnés respectivement par :

$$\hat{\boldsymbol{\beta}} = \mathbf{a} + \mathbf{R}\mathbf{X}_s'(\mathbf{X}_s\mathbf{R}\mathbf{X}_s' + \mathbf{V}_s)^{-1}(\mathbf{y}_s - \mathbf{X}_s\mathbf{a}) \text{ et } \hat{V}(\hat{\boldsymbol{\beta}}) = \mathbf{C} = \mathbf{R} - \mathbf{R}\mathbf{X}_s'(\mathbf{X}_s\mathbf{R}\mathbf{X}_s' + \mathbf{V}_s)^{-1}\mathbf{X}_s\mathbf{R}. \tag{2.6}$$

Il est facile de voir que, dans (2.6), la première équation peut être réécrite sous la forme $\hat{\boldsymbol{\beta}} = \mathbf{C}(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s + \mathbf{R}^{-1}\mathbf{a})$, où $\mathbf{C}^{-1} = \mathbf{R}^{-1} + \mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s$. Il convient de souligner que, si nous plaçons une loi a priori vague sur $\boldsymbol{\beta}$, en prenant $\mathbf{R}^{-1} \to 0$, nous obtenons l'estimateur par les moindres carrés minimal de $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s$.

Maintenant, en appliquant les propriétés bien connues des espérances et des variances conditionnelles, nous obtenons :

$$E[\mathbf{y}_{\bar{s}}|\mathbf{y}_s] = E(E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})|\mathbf{y}_s) \text{ et } V[\mathbf{y}_{\bar{s}}|\mathbf{y}_s] = E(V(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})|\mathbf{y}_s) + V(E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})|\mathbf{y}_s). \tag{2.7}$$

En remplaçant $E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ et $V(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ dans (2.7) par leur EBL respectif donné en (2.5) et puis, en remplaçant $E(\boldsymbol{\beta}|\mathbf{y}_s)$ et $V(\boldsymbol{\beta}|\mathbf{y}_s)$ par $\hat{\boldsymbol{\beta}}$ et $\hat{V}(\hat{\boldsymbol{\beta}})$ donnés en (2.6), nous obtenons l'EBL de $E[\mathbf{y}_{\bar{s}}|\mathbf{y}_s]$ et sa variance associée sous la forme :

$$\begin{aligned}\hat{E}[\mathbf{y}_{\bar{s}}|\mathbf{y}_s] &= \mathbf{X}_{\bar{s}}\hat{\boldsymbol{\beta}} + \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}) \text{ et}\\ \hat{V}[\mathbf{y}_{\bar{s}}|\mathbf{y}_s] &= \mathbf{V}_{\bar{s}} - \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}\mathbf{V}_{s\bar{s}} + (\mathbf{X}_{\bar{s}} - \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}\mathbf{X}_s)\mathbf{C}(\mathbf{X}_{\bar{s}} - \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}\mathbf{X}_s)'.\end{aligned} \tag{2.8}$$

Remarque 2 : De manière analogue à la remarque 1, sous l'hypothèse de normalité, nous avons que les membres de droite des équations (2.8) sont, respectivement, les valeurs de $E[\mathbf{y}_{\bar{s}}|\mathbf{y}_s]$ et $V[\mathbf{y}_{\bar{s}}|\mathbf{y}_s]$.

L'expression générale de l'EBL du total T et de sa variance associée est obtenue en remplaçant $E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s)$ et $V(\mathbf{y}_{\bar{s}}|\mathbf{y}_s)$ dans les équations (2.3) par leurs équivalents respectifs $\hat{E}[\mathbf{y}_{\bar{s}}|\mathbf{y}_s]$ et $\hat{V}[\mathbf{y}_{\bar{s}}|\mathbf{y}_s]$:

$$\hat{T} = \mathbf{1}_s'\mathbf{y}_s + \mathbf{1}_{\bar{s}}'\hat{E}[\mathbf{y}_{\bar{s}}|\mathbf{y}_s] \text{ et } \hat{V}(\hat{T}) = \mathbf{1}_{\bar{s}}'\hat{V}[\mathbf{y}_{\bar{s}}|\mathbf{y}_s]\mathbf{1}_{\bar{s}}. \tag{2.9}$$

Il convient de souligner que, dans de nombreuses applications de (2.9), la matrice $\mathbf{V}$ est supposée être diagonale, ce qui implique que $\mathbf{V}_{\bar{s}\bar{s}} = \mathbf{0}$ et nous avons alors :

$$\hat{T} = \mathbf{1}_s' \mathbf{y}_s + \mathbf{1}_{\bar{s}}' \mathbf{X}_{\bar{s}} \hat{\boldsymbol{\beta}} \text{ et } \hat{V}\left(\hat{T}\right) = \mathbf{1}_{\bar{s}}' \left[\mathbf{V}_{\bar{s}} + \mathbf{X}_{\bar{s}} \mathbf{C} \mathbf{X}_{\bar{s}}'\right] \mathbf{1}_{\bar{s}}. \tag{2.10}$$

En guise d'illustration, nous considérons certains exemples discutés par O'Hagan (1985) et proposons un nouvel estimateur par le ratio, qui est l'une des contributions des présents travaux. Tous les exemples peuvent être traités comme des cas particuliers du modèle (2.4).

2.3 Retour sur certains plans de sondage fréquents

2.3.1 Échantillonnage aléatoire simple sans remise : échangeabilité

O'Hagan (1985) a examiné le cas simple où la population ne présente aucune structure pertinente, ce qui peut se faire en spécifiant :

$$E(y_i) = m, V(y_i) = v \text{ et } \operatorname{Cov}\left(y_i, y_j\right) = c, i, j = 1, \ldots, N, \forall i \neq j. \tag{2.11}$$

Remarque 3 : On peut justifier la corrélation introduite dans le modèle (2.11) en invoquant l'imitation de l'échantillonnage aléatoire simple sans remise.

En appliquant le résultat général établi dans (2.10) à (2.11) avec $\boldsymbol{\beta}$ de dimension 1, $\mathbf{X} = \mathbf{1}_N$, $\mathbf{a} = m$, $\mathbf{R} = c$ et $\mathbf{V} = \sigma^2 \mathbf{I}$, où $\sigma^2 = v - c$, nous obtenons l'EBL de T et sa variance associée :

$$\hat{T}_{srs} = n\bar{y}_s + (N - n)\hat{\mu} \text{ et } \hat{V}\left(\hat{T}_{srs}\right) = (N - n)\sigma^2 + (N - n)^2 c\sigma^2 \left(\sigma^2 + nc\right)^{-1}, \tag{2.12}$$

où

$\bar{y}_s = n^{-1} \mathbf{1}_s' \mathbf{y}_s$ est la moyenne d'échantillon,

$\hat{\mu} = \omega \bar{y}_s + (1 - \omega) m$ est la valeur prévue des valeurs non observées de $\mathbf{y}$ et

$\omega = \dfrac{n\sigma^{-2}}{c^{-1} + n\sigma^{-2}}$, où $\sigma^2 = v - c$.

Il convient de souligner que $\hat{\mu}$ est une moyenne pondérée de la moyenne a priori m et de la moyenne d'échantillon $\bar{y}_s$, où ω est le ratio entre les deux quantités de population. La moyenne m peut être considérée comme le prior du chercheur pour la moyenne de population réelle $\bar{y}$. L'incertitude au sujet de y_i est divisée en deux composantes : l'incertitude au sujet du niveau global des y_i (inter-variation) et l'incertitude quant à la part de chaque y_i qui peut différer de ce niveau global (intra-variation). Une mesure utile de la variabilité des unités dans la population est donnée par

$$S^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (y_i - \bar{y})^2.$$

Il n'est pas difficile de montrer que $E\left(S^2\right) = v - c = \sigma^2$. Par conséquent, σ^2 peut être interprétée comme une estimation a priori de la variabilité à l'intérieur de la population. Nous obtenons aussi

$V(\bar{y}) = c + N^{-1}\sigma^2$. Dans de nombreuses applications, N est grand et la constante c peut donc être considérée comme l'inter-variation.

En posant que $v \to \infty$ et en maintenant σ^2 fixe, c'est-à-dire en supposant que l'on ne connaît pas les priors, les estimations dans (2.12) donnent :

$$\hat{T}_{srs} = N\,\bar{y}_s \text{ et } \hat{V}\left(\hat{T}_{srs}\right) = N^2\left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}.$$

Ces expressions sont fort semblables à l'estimation bien connue du total et à sa variance dans le contexte fondé sur le plan de sondage pour le cas de l'échantillonnage aléatoire simple. O'Hagan (1985) a discuté de certains moyens possibles d'éviter la tâche difficile d'attribuer une valeur à σ^2. Le moyen le plus naturel de le faire consiste à trouver son EBL, mais linéaire en les carrés et les termes de variance des produits croisés. Cependant, il est nécessaire de spécifier les moments d'ordre quatre des y_i. Goldstein (1979) a proposé un EBL de la variance qui n'utilise que des fonctions linéaires des données. Néanmoins, on obtient une expression compliquée de la variance associée de son EBL modifié. O'Hagan (1985) a soutenu que, si l'information a priori au sujet des composantes de la variance est faible, toute estimation a posteriori s'approche des estimations non bayésiennes classiques obtenues en utilisant uniquement les données, lorsque ce gendre d'estimations est disponible. Par conséquent, il a proposé, en guise de procédure bayésienne approximative, d'introduire ces estimations de variance classiques par substitution dans l'EBL et dans sa variance associée lorsque cela est approprié. Dans le cas qui nous occupe, nous pouvons remplacer σ^2 par $s^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y}_s)^2$, qui est sans biais sous le plan pour S^2.

2.3.2 Échantillonnage aléatoire simple stratifié sans remise

Soit y_{hi} la i^{e} unité, $i = 1,...,N_h$ appartenant à la strate h, $h = 1,..,H$. Nous supposons que la taille de strate, N_h, est connue pour toutes les strates. L'échangeabilité d'ordre deux dans chaque strate est énoncée dans O'Hagan (1985) sous la forme :

$$E(y_{hi}) = m_h, V(y_{hi}) = v_h, \text{Cov}(y_{hi}, y_{hj}) = c_h, i \neq j \text{ et } \text{Cov}(y_{hi}, y_{lj}) = d_{hl}, h \neq l.$$

Remarque 4 : Il est raisonnable de supposer que l'information obtenue au sujet d'une strate pourrait modifier les croyances au sujet des autres strates dans certaines applications spéciales. Cependant, si nous voulons imiter l'échantillonnage aléatoire simple stratifié, nous devons supposer que les observations dans les diverses strates ne sont pas corrélées, en posant que $d_{hl} = 0$.

Le modèle général (2.4) peut être appliqué à ce cas en prenant $\mathbf{X} = \text{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_H)$ et $\mathbf{V} = \text{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_H)$, avec $\mathbf{X}_h = \mathbf{1}_{N_h}$ et $\mathbf{V}_h = \sigma_h^2 \mathbf{I}_{N_h}$, où $\sigma_h^2 = v_h - c_h$, $\forall\, h = 1, \ldots, H$, $\mathbf{a} = (m_1, \ldots, m_H)'$, $\mathbf{R}$ est une matrice de dimensions $H \times H$ avec $R_{hl} = c_h$, si $h = l$ et $R_{hl} = d_{hl}$ autrement. L'EBL de T et sa variance associée sont obtenus au moyen de (2.10) et figurent dans O'Hagan (1985). Les modèles pour l'échantillonnage en grappes sont donnés dans Bolfarine et Zacks (1992), page 11. L'EBL des modèles avec grappes figurent dans O'Hagan (1985).

3 Information auxiliaire : estimateur par le ratio

Dans de nombreuses situations pratiques, il est possible d'obtenir de l'information au sujet d'une variable auxiliaire x_i (corrélée à y_i) pour toutes les unités de la population, ou du moins pour chaque unité échantillonnée, ainsi que la moyenne de population $\bar{X}$. En pratique, x_i est souvent la valeur de y_i à une période antérieure où a été effectué un recensement complet. Cette approche est utilisée dans des situations où la valeur prévue et la variance de y_i sont proportionnelles à x_i, de sorte que dans la spécification de l'EBL, nous remplaçons certaines hypothèses au sujet des y par des hypothèses au sujet des deux premiers moments du taux y_i / x_i. Autant que nous sachions, le nouvel estimateur par le ratio proposé ci-après est une nouvelle contribution à la théorie des sondages.

Le nouvel estimateur par le ratio est obtenu en tant que cas particulier du modèle (2.4) et en se servant de l'hypothèse d'échangeabilité, utilisée dans l'approche bayésienne linéaire, appliquée au taux y_i / x_i pour tout $i = 1, \ldots, N$, comme il est décrit ci-dessous :

$$E\left(\frac{y_i}{x_i}\right) = m, V\left(\frac{y_i}{x_i}\right) = v \text{ et } \operatorname{Cov}\left(\frac{y_i}{x_i}, \frac{y_j}{x_j}\right) = c, i, j = 1, \ldots, N, \forall i \neq j. \tag{3.1}$$

En appliquant le résultat général établi dans (2.10) à (3.1) avec $\mathbf{X} = (x_1, \ldots, x_N)'$, le vecteur $N \times 1$ de dimension $\mathbf{a} = m$, de variables auxiliaires $\mathbf{R} = c$ et $\mathbf{V} = \sigma^2 \operatorname{diag}(x_1, \ldots, x_N)$, où $\sigma^2 = v - c$, nous obtenons l'EBL de T et sa variance associée comme il suit :

$$\hat{T}_{ra} = n\bar{y}_s + (N - n)\hat{\mu}\bar{x}_{\bar{s}} \text{ et}$$

$$\hat{V}\left(\hat{T}_{ra}\right) = (N - n)\bar{x}_{\bar{s}}\sigma^2 + (N - n)^2 \bar{x}_{\bar{s}}^{\,2}\left(c^{-1} + \sigma^{-2} n\bar{x}_s\right)^{-1}, \text{ où}$$

$$\hat{\mu} = \omega\frac{\bar{y}_s}{\bar{x}_s} + (1 - \omega)m \text{ et } \omega = \frac{\sigma^{-2} n\bar{x}_s}{\left(c^{-1} + \sigma^{-2} n\bar{x}_s\right)},$$

où $\bar{x}_{\bar{s}} = (N\bar{X} - n\bar{x}_s)/(N - n)$ est la moyenne des x pour les unités non échantillonnées. En posant que $v \to \infty$ et $n \to \infty$, mais en maintenant σ^2 fixe, nous réobtenons l'estimateur de type ratio de l'approche fondée sur le plan de sondage : $\hat{T}_{ra} = N\bar{X}(\bar{y}_s / \bar{x}_s)$.

4 Méthode bayésienne linéaire pour données catégoriques

Il arrive souvent que l'on s'intéresse à des cas où la caractéristique observée est celle de savoir si l'unité de population possède ou non un certain attribut d'intérêt. Nous pouvons définir une variable dichotomique $y_i = 1$, si la i^{e} unité possède cet attribut, ce qui est désigné comme une réussite, et $y_i = 0$ autrement. Pour le cas binaire lorsque la taille de l'échantillon n'est pas suffisamment grande pour appliquer le théorème central limite, l'approche fondée sur le plan de sondage pourrait faire appel à la randomisation introduite par le plan de sondage pour justifier la distribution des quantités aléatoires binaires. Par exemple, Cochran (1977), sections 3.4 et 3.5, montre comment appliquer les lois hypergéométrique et binomiale pour obtenir les intervalles de confiance pour les proportions de population quand on se sert de plans d'échantillonnages aléatoires simples avec et sans remise, respectivement. Par ailleurs, des approches dépendantes d'un modèle ont également été avancées et

appliquées pour prédire les totaux ou les moyennes dans les catégories d'intérêt. Malec, Sedransk, Moriarity et LeClere (1997) ont considéré un modèle hiérarchique logistique à deux niveaux, où les grappes forment le deuxième niveau. Ils ont également comparé les estimations bayésiennes entièrement hiérarchiques aux estimations bayésiennes empiriques et aux méthodes classiques. Moura et Migon (2002) ont présenté une approche basée sur un modèle hiérarchique logistique pour la prédiction de proportions sur petits domaines, en tenant compte des effets spatiaux ainsi que des effets d'hétérogénéité non structurée possibles. Nandram et Choi (2008) ont proposé un modèle multinomial-Dirichlet dépendant du temps pour prédire les résultats d'une élection sous non-réponse ignorable et non ignorable. Ils ont également utilisé une approche bayésienne pour répartir les électeurs indécis entre les candidats.

De nouveau, ici, nous n'avons pas besoin d'utiliser des hypothèses au sujet du modèle complet ni une approche de randomisation, mais nous devons émettre certaines hypothèses au sujet des premier et deuxième moments des quantités aléatoires concernées. L'EBL pour les données binaires a été introduit brièvement par O'Hagan (1985), mais ici, nous le développons d'une manière plus générale pour le cas où nous nous intéressons à l'analyse de plus d'un attribut dans une population. L'objectif est de décrire l'estimation de la proportion de réussites avec des données catégoriques. Soit y_{ij} la variable qui indique que l'unité $i,\ i = 1,\ldots,N$ se trouve dans la catégorie $j,\ j = 1,\ldots,k$ donnée par

$$y_{ij} = \begin{cases} 1, \text{ si la } i^{\text{e}} \text{ unité possède le } j^{\text{e}} \text{ attribut;} \\ 0, \text{ autrement.} \end{cases}$$

Le but principal est d'estimer un vecteur $\mathbf{p} = (p_1,\ldots,p_k)'$, où $p_j = N^{-1}\sum_{i=1}^{N} y_{ij},\ j = 1,\ldots,k,$ est la proportion d'unités dans la catégorie $j,$ sachant $\mathbf{y}_s,$ un vecteur de dimension $nk,$ défini comme étant $\mathbf{y}_s = (y_{11}, y_{21},\ldots, y_{n1},\ldots, y_{1k}, y_{2k},\ldots, y_{nk})'.$ Comme nous avons affaire à des situations dans lesquelles il n'est possible d'associer qu'un seul attribut à chaque unité, nous avons $\sum_{j=1}^{k} p_j = 1.$ Donc, nous ne devons estimer que $k-1$ paramètres, puisqu'il s'ensuit que $\hat{p}_k = 1 - \sum_{j=1}^{k-1}\hat{p}_j$ et que l'estimation de la variance est également obtenue de manière analogue par $\hat{V}(\hat{p}_k) = \sum_{j=1}^{k-1}\hat{V}(\hat{p}_j) + \sum_{l\neq j=1}^{k-1}\hat{\text{Cov}}(\hat{p}_j, \hat{p}_l).$

En l'absence de toute autre information structurelle, nous supposons que les unités dans une catégorie donnée sont échangeables d'ordre deux, mais nous ne supposons aucune échangeabilité entre les unités de différentes catégories. Nos croyances a priori sont exprimées pour $i = 1,\ldots,N,\ j = 1,\ldots,k-1,$ comme il suit :

$$\begin{aligned} m_j &= E(y_{ij}) = P(y_{ij} = 1), v_j = V(y_{ij}) = m_j(1 - m_j) \text{ et} \\ \text{cov}(y_{ij}, y_{i'j}) &= P(y_{i'j} = 1 \mid y_{ij} = 1)P(y_{ij} = 1) - P(y_{ij} = 1)P(y_{i'j} = 1) \\ &= m_j(m_{jj} - m_j) = c_j,\ \forall i \neq i' \text{ et } \sigma_j^2 = v_j - c_j = m_j(1 - m_{jj}), \end{aligned}$$

où $m_{jj} = P(y_{i'j} = 1 \mid y_{ij} = 1),$ pour tout $i \neq i'.$

Pour $j \neq j',$ nous obtenons de manière analogue la covariance entre ces catégories sous la forme

$$\text{cov}(y_{ij}, y_{i'j'}) = \begin{cases} m_j(m_{j'j} - m_{j'}), & \text{si } i \neq i', \\ -m_j m_{j'}, & \text{si } i = i'. \end{cases}$$

Souvent, nous ne possédons pas toutes les données $\mathbf{y}_s$, mais seulement une statistique exhaustive, comme la proportion dans l'échantillon pour chaque catégorie, $\bar{\mathbf{y}}_s$. Soit $\bar{\mathbf{y}}_s$ le vecteur de dimension $k-1$- dont la j^{e} position est donnée par la moyenne d'échantillon pour la catégorie j. En utilisant le modèle général donné par (2.4), nous obtenons :

$$E(\bar{\mathbf{y}}_s) = E(E(\bar{\mathbf{y}}_s \mid \boldsymbol{\beta})) = \mathbf{a} \text{ et } \operatorname{Var}(\bar{\mathbf{y}}_s) = E(V(\bar{\mathbf{y}}_s \mid \boldsymbol{\beta})) + V(E(\bar{\mathbf{y}}_s \mid \boldsymbol{\beta})) = \mathbf{V}_s + \mathbf{R}.$$

En appliquant le modèle général donné dans (2.4), où la variable de réponse est donnée par $\bar{\mathbf{y}}_s$, le vecteur $\boldsymbol{\beta}$ est de dimension $k-1$, $\mathbf{X}_s = \mathbf{I}_s$ et $\mathbf{V} = \operatorname{diag}(\mathbf{V}_{\bar{s}}, \mathbf{V}_s)$, nous obtenons à partir de (2.10) :

$$\hat{\mathbf{p}} = \frac{n\bar{\mathbf{y}}_s + (N-n)\hat{\boldsymbol{\beta}}}{N} \text{ et } \hat{V}(\hat{\mathbf{p}}) = \frac{(N-n)^2[\mathbf{V}_{\bar{s}} + \mathbf{C}]}{N^2}, \tag{4.1}$$

où $\mathbf{C}^{-1} = \mathbf{R}^{-1} + \mathbf{V}_s$ et $\hat{\boldsymbol{\beta}} = \mathbf{C}\left(\mathbf{V}_s^{-1}\bar{\mathbf{y}}_s + \mathbf{R}^{-1}\mathbf{a}\right)$, comme il est énoncé en (2.6).

Soit $\mathbf{Q} = \mathbf{V}_s + \mathbf{R}$. L'EBL de $\mathbf{p}$ et sa variance associée donnés par (4.1) peuvent s'écrire en fonction des quantités a priori m_j, $m_{jj'}$ et $j = 1, \ldots, k-1$, en notant que $\mathbf{a} = (m_1, \ldots, m_{k-1})'$, $Q_{jj} = c_j + \sigma_j^2/n$ et $Q_{jj'} = m_j\left(m_{j'j} - m_{j'}\right) - m_j m_{j'j}/n$. Par conséquent, la matrice $\mathbf{R} = \{r_{jj'}\}, j, j' = 1, .., k-1$ avec $r_{jj} = c_j$ et $r_{jj'} = m_j\left(m_{j'j} - m_{j'}\right)$ et $\mathbf{V}_s = 1/n\,\{v_{jj'}\}$, $j, j' = 1, .., k-1$ avec $v_{jj} = \sigma_j^2$ et $v_{jj'} = -m_j m_{j'j}$. De manière analogue, nous obtenons $\mathbf{V}_{\bar{s}} = n/(N-n)\,\mathbf{V}_s$.

4.1 Obtention des priors

L'obtention des priors est le processus consistant à formuler les connaissances et les croyances d'une personne au sujet d'une ou de plusieurs quantités incertaines sous forme d'une loi de probabilité pour ces quantités. Selon Garthwaite, Kadane et O'Hagan (2005), il est commode de concevoir la tâche d'obtention des priors comme faisant intervenir un facilitateur qui aide l'expert à formuler ses connaissances spécialisées sous forme probabiliste. Dans le contexte de l'obtention d'une loi a priori pour une analyse bayésienne, ce sont les connaissances a priori de l'expert qui sont tirées au clair, mais en général, l'objectif est d'exprimer les connaissances courantes de l'expert sous forme probabiliste. Si l'expert est un statisticien ou s'il connaît très bien les concepts statistiques, l'intervention d'un facilitateur pourrait ne pas être formellement nécessaire, mais cela est rare en pratique. O'Hagan (1998) a illustré au moyen d'un exemple pratique comment obtenir les premier et deuxième moments. En particulier, il a adopté l'approche bayésienne linéaire parce qu'elle permet aux ingénieurs d'appliquer facilement une procédure d'obtention des priors.

À la présente section, nous présentons certaines contraintes concernant les quantités a priori et une solution de rechange pour faciliter le processus d'obtention des priors en vue d'obtenir l'EBL pour des données catégoriques. Comme m_j et $m_{jj'}$ sont des probabilités et que $\mathbf{R}$ et $\mathbf{V}_s$ sont les matrices de covariance dans le modèle (2.4), les contraintes qui suivent doivent être satisfaites :

1. $0 < m_j < 1$ et $0 \le m_{jj'} \le 1$, $j, j' = 1, \ldots, k-1$;
2. $\mathbf{R}$ et $\mathbf{V}_s$ sont des matrices symétriques définies positives.

Afin de vérifier si la condition (2.2) est satisfaite, on peut exécuter les étapes suivantes :

i. vérifier si $\mathbf{R}$ et $\mathbf{V}_s$ sont symétriques en vérifiant que $m_j m_{jj'} = m_{j'} m_{j'j}$;

ii. vérifier si $\mathbf{R}$ et $\mathbf{V}_s$ sont des matrices définies positives en trouvant les valeurs propres de $\mathbf{R}$ et $\mathbf{V}_s$. Si les valeurs propres sont positives, alors les matrices sont définies positives.

Il convient de mentionner que les valeurs propres sont les racines du polynôme caractéristique et que si ce polynôme est de degré $n, n \leq 4$, il est possible d'obtenir analytiquement ses racines en appliquant Bhaskara, Cardan ou Ferrari; voir Jacobson (2009), chapitre 4, pour les formules. Cependant, si $n \geq 5$, il est habituellement nécessaire d'appliquer une méthode itérative pour les obtenir. Néanmoins, pour les matrices de dimensions supérieures à 2×2, il n'est pas simple d'obtenir analytiquement ces contraintes en se basant sur les valeurs propres. La proposition qui suit présente les conditions que m_j et $m_{jj'}$, $j = 1, \ldots, k-1$, doivent satisfaire afin d'obtenir un prior convenable pour un modèle multinomial comprenant trois catégories en utilisant l'approche d'estimation bayésienne linéaire.

Proposition 1 Supposons que nous obtenons m_j, tel que $0 < m_j < 1$, $j = 1, 2$. Alors, sachant ρ_{11}, ρ_{12} et ρ_{22}, nous obtenons m_{11}, m_{12}, m_{21} et m_{22} au moyen de (4.2). Les quantités a priori m_j et $m_{jj'}$, pour $j, j' = 1, 2$, doivent satisfaire les contraintes qui suivent pour que les matrices $\mathbf{R}$ et $\mathbf{V}_s$ soient définies positives :

$$m_{11} > m_1 \ et \ m_{22} > m_2, m_{11}m_{22} - m_{11} - m_{22} + 1 > m_{12}m_{21} \text{ et}$$
$$m_{11}m_{22} - m_{11}m_2 - m_1 m_{22} > m_{12}m_{21} - 2m_2 m_{12}.$$

La vérification de la proposition 1 nécessite certaines opérations algébriques. Nous vérifions que les matrices $\mathbf{R}$ et $\mathbf{V}_s$ sont définies positives en utilisant (i) et (ii) susmentionnés. Nous faisons appel au fait que les valeurs propres d'une matrice de dimensions 2×2 sont positives si et seulement si son déterminant est positif et nous obtenons alors $m_{jj'}$, $j, j' = 1, 2$ qui satisfait cette contrainte pour les deux matrices. Pour les cas comprenant plus de trois catégories, nous devons vérifier numériquement si les matrices $\mathbf{R}$ et $\mathbf{V}_s$ sont définies positives en remplaçant m_j et $m_{jj'}$, $j = 1, \ldots, k-1$ par leur valeur numérique.

Par ailleurs, si un expert a de la difficulté à spécifier certaines de ces probabilités conditionnelles $m_{jj'}$, il pourrait être plus simple d'attribuer un prior au coefficient de corrélation. Définissons $\rho_{jj'}$ comme étant le prior du coefficient de corrélation entre deux unités différentes dans les catégories j et j', c'est-à-dire :

$$\rho_{jj'} = \operatorname{corr}\left(y_{ij}, y_{i'j'}\right) = \begin{cases} \dfrac{m_{jj} - m_j}{1 - m_j}, & j = j', \\ \dfrac{m_j\left(m_{j'j} - m_{j'}\right)}{\sqrt{m_j\left(1 - m_j\right) m_{j'}\left(1 - m_{j'}\right)}}, & j \neq j', \end{cases}$$

pour $i, i' = 1, \ldots, n$, $i \neq i'$, $j, j' = 1, \ldots, k-1$.

Par conséquent, sachant $\rho_{jj'}$, $j, j' = 1, \ldots, k-1$, nous obtenons

$$m_{jj'} = \begin{cases} m_j + \rho_{jj}\left(1 - m_j\right) & j = j', \\ \dfrac{m_j m_{j'} + \rho_{j'j}\sqrt{m_j\left(1 - m_j\right)m_{j'}\left(1 - m_{j'}\right)}}{m_{j'}}, & j \neq j'. \end{cases} \tag{4.2}$$

Il convient de mentionner que, si l'on dispose de données provenant d'une enquête antérieure, il est possible qu'un expert utilise cette information. Par exemple, m_j peut être obtenu en estimant la proportion d'unités dans la catégorie j, $j = 1, \ldots, k-1$ à partir de l'enquête antérieure. De façon analogue, $\rho_{jj'}$ peut être obtenu en utilisant les données d'une enquête antérieure. Comme l'indique la contrainte (2.1), m_j ne peut pas prendre les valeurs 0 et 1, sinon les corrélations ne seraient pas définies.

4.2 Analyse de la sensibilité aux priors

Il est utile de vérifier si l'estimateur et sa variance associée dépendent des priors attribués. Nous traitons le cas simple ne comprenant que deux catégories. Soulignons que, dans le cas où il y a plus de deux catégories, le nombre de quantités a priori qu'il faut obtenir augmente rapidement, mais que l'on peut étendre les conclusions obtenues. Par ailleurs, en l'absence d'information a priori, nous pouvons utiliser des priors non informatifs et, comme il est décrit à la section 2.2, on retrouve alors les estimateurs de l'approche fondée sur le plan de sondage.

L'EBL pour la proportion en cas de données binaires peut être obtenu en tant que cas particulier de l'estimateur (4.1),

$$\hat{p}_1 = \frac{n\bar{y}_1 + (N - n)\hat{\mu}}{N},$$

où

$\hat{\mu} = \omega\bar{y}_1 + (1 - \omega)m_1$ est la valeur prévue des valeurs non observées dans la catégorie 1,

$$\omega = \frac{n\sigma_1^{-2}}{n\sigma_1^{-2} + c_1^{-1}},$$

et $\hat{p}_2 = 1 - \hat{p}_1$. Notons que σ_1^2 et c_1 dépendent de $m_{11} = m_1 + \rho_{11}(1 - m_1)$, voir page 13. Nous analysons comment les estimations sont affectées par ρ_{11}.

1. Si $\rho_{11} \to 0$, alors $\omega \to 0$ et $\hat{\mu} \to m_1$. Donc, l'estimateur pour les valeurs non observées dépend en grande partie de la valeur du prior.
2. Si $\rho_{11} \to 1$, alors $\omega \to 1$ et $\hat{\mu} \to \bar{y}_1$. Donc, l'estimateur pour les valeurs non observées ne dépend pas de la valeur du prior.

En outre, il est facile de voir que $n/N \to 1$, $\hat{p}_1 \to \bar{y}_1$. Pour illustrer ces résultats, nous avons créé un jeu de données artificielles en fixant la proportion réelle à $\mathbf{p} = (0{,}2380; 0{,}7620)'$ et la moyenne d'échantillon à $\bar{\mathbf{y}}_s = (0{,}2614; 0{,}7386)'$. Ces valeurs ont été tirées de Moura et Migon (2002). Puis, nous avons déterminé comment les valeurs de m_1, N, $f = n/N$ et ρ_{11} affectent l'estimateur $\hat{p}_1$. La

figure 4.1 donne la représentation graphique en deux dimensions de l'erreur absolue de $\hat{p}_1$ en fonction de ρ_{11} pour certains cas particuliers. La courbe grise représente l'erreur absolue entre la proportion d'échantillon $\bar{y}_1$ et la proportion réelle p_1.

Il faut souligner que, à mesure que f ou N augmente, l'erreur absolue diminue pour toute valeur du prior. De surcroît, quand $\rho_{11} \to 0$, l'erreur absolue augmente quand m_1 diffère considérablement de la proportion réelle p_1, mais elle diminue à mesure que la taille de l'échantillon augmente. Enfin, quand $\rho_{11} \to 1$, nous observons que l'erreur absolue de $\hat{p}_1$ tend vers l'erreur absolue de la proportion d'échantillon $\bar{y}_1$. Donc, si nous avons une bonne information a priori, en ce qui concerne m_1, l'estimateur proposé donne de bons résultats pour toutes les valeurs de ρ_{11}. Cependant, si aucune information a priori n'est disponible, des priors non informatifs caractérisés par $\rho_{11} \to 1$ peuvent être utilisés et nous obtenons des résultats similaires à ceux de l'approche fondée sur le plan de sondage.

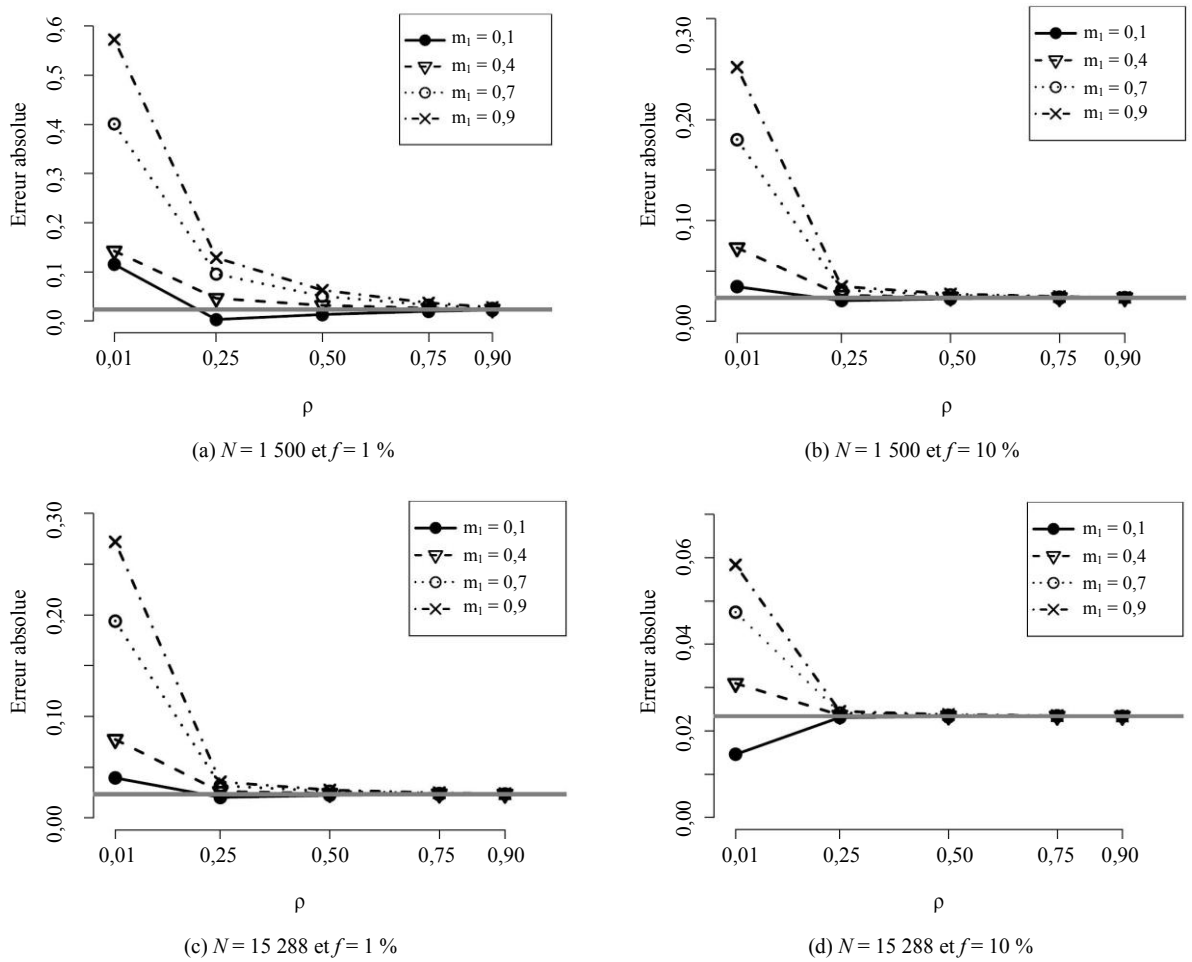

Figure 4.1 Erreur absolue pour $m_1 \in \{0{,}1; 0{,}4; 0{,}7; 0{,}9\}$, $N \in \{1\,500, 15\,288\}$ et $f \in \{1\,\%, 10\,\%\}$ fixes et $\rho_{11} \in \{0{,}01; 0{,}25; 0{,}5; 0{,}75; 0{,}9\}$ variable. La courbe grise représente l'erreur absolue de la proportion d'échantillon $\bar{y}_1$

5 Conclusion

Obtenir une loi a priori conjointe complète pour un grand nombre de dimensions représenterait une tâche énorme. La méthode bayésienne linéaire nécessite seulement l'obtention des moyennes, des variances et des covariances a priori des paramètres. Elle est particulièrement utile lorsqu'un statisticien n'est pas disponible pour procéder à l'obtention complète des priors. Un exemple d'obtention réussie des priors en utilisant cet estimateur figure dans O'Hagan (1998).

Nous avons dérivé les estimateurs fondés sur le plan de sondage bien connus en appliquant la structure de l'EBL à une approche faisant appel à un modèle de régression généralisée. Nous avons étendu l'estimateur aux données catégoriques et conclu que même si cet estimateur nécessite l'obtention des priors de nombreuses quantités, il est possible de les reparamétriser ou de travailler avec des priors non informatifs. L'exemple numérique illustre le comportement des estimations sous forme d'une fonction de la taille de l'échantillon et des spécifications des paramètres a priori. Cependant, nous sommes conscients que l'obtention des priors pour un grand nombre des paramètres n'est pas une tâche facile si l'on ne dispose pas d'information provenant d'enquêtes antérieures. Néanmoins, les exemples dont il est question dans l'article montrent que même si on ne dispose pas d'information a priori, il est possible d'obtenir les estimateurs fondés sur le plan de sondage équivalent en fixant une variance suffisamment grande pour les priors. En outre, les praticiens des sondages qui doivent obtenir des estimations pour un grand nombre de variables se rendraient également compte qu'ils ne seraient pas capables de produire des estimations d'une précision satisfaisante pour toutes les variables, quelle que soit l'approche utilisée. Enfin, nous avons montré comment l'approche de l'EBL et l'approche fondée sur le plan de sondage peuvent être conciliées.

Remerciements

Les présents travaux ont été réalisés dans le cadre de la thèse de maîtrise de Kelly C.M. Gonçalves sous la supervision de Fernando Moura et de Helio Migon, dans le programme d'études supérieures de l'UFRG. Kelly est titulaire d'une bourse décernée par le Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior (CAPES). Fernando Moura et Helio Migon ont reçu l'appui financier du Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-Brazil, BPPesq).

Les auteurs remercient le rédacteur en chef, le rédacteur associé et deux examinateurs de leurs commentaires très constructifs.

Bibliographie

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part 1 (avec discussion). Dans *Foundations of Statistical Inference*, (Éds., Godambe et Sprott), 203-242. Holt, Reinhart and Wilnston, Toronto.

Bolfarine, H., et Zacks, S. (1992). *Prediction Theory for Finite Populations*. New York : Springer-Verlag.

Cochran, W. (1977). *Sampling Techniques*. New York : John Wiley & Sons, Inc.

Garthwaite, P., Kadane, J. et O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100, 680-701.

Goldstein, M., et Wooff, D. (2007). *Bayes Linear Statistics: Theory and Methods*. Durham University, UK : Wiley series in probability and statistics.

Goldstein, M. (1979). The variance modified linear Bayes estimator. *Journal of the Royal Statistical Society*, 41, 96-100.

Hartigan, J. (1969). Linear bayesian methods. *Journal of the Royal Statistical Society*, Séries B (Méthodologique), 446-454.

Horvitz, D., et Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Jacobson, N. (2009). *Basic Algebra,* Vol 1. Dover Books on Mathematics.

Little, R.J. (2003). The Bayesian approch to sample survey inference. Dans *Analysis of Survey Data*, (Éds., R.L. Chambers et C.J. Skinner), chapitre 4, 49-52. New York : John Wiley & Sons Inc.

Malec, D., Sedransk, J., Moriarity, C.L. et LeClere, F.B. (1997). Small area inference for binary variables in National Health Interview Survey. *Journal of the American Statistical Association*, 92, 815-826.

Moura, F., et Migon, H. (2002). Bayesian spatial models for small area estimation of proportions. *Statistical Modelling*, 2, 183-201.

Nandram, B., et Choi, J. (2008). Une répartition bayésienne des électeurs indécis. *Techniques d'enquête*, 34, 1, 41-54.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

O'Hagan, A. (1985). Bayes linear estimators for finite populations. *Rapport technique 58*, Department of Statistics - University of Warwick.

O'Hagan, A. (1987). Bayes linear estimators for randomized response models. *Journal of the American Statistical Association*, 82, 580-585.

O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications. *The Statistician*, 47, 21-35.

Pfeffermann, D., Moura, F.A.S. et Silva, P.L.N. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93, 943.

Scott, A.J. (1977). Large-sample posterior distributions for finite populations. *Annals of Mathematical Statistics*, 42, 1113-1117.

Skinner, C., Holt, D. et Smith, T. (1989). *Analysis of complex surveys*. New York : John Wiley & Sons, Inc.

Smouse, E. (1984). A note on bayesian least squares inference for finite population models. *Journal of the American Statistical Association*, 79, 390-392.

Valliant, R., Dorfman, A. et Royall, R. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York : John Wiley & Sons, Inc.

Zacks, S. (2002). In the footsteps of Basu: The predictive modelling approach to sampling from finite population. *Sankhyā: The Indian Journal of Statistics*, Séries A, 64, 532-544.

ELECTRONIC PUBLICATIONS AVAILABLE AT | PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À

www.statcan.gc.ca

Techniques d'enquête, juin 2014
Vol. 40, N° 1, pp. 33-52
Statistique Canada, N° 12-001-X au catalogue

Une méthode non paramétrique de production de populations synthétiques qui tient compte des caractéristiques des plans de sondage complexes

Qi Dong, Michael R. Elliott et Trivellore E. Raghunathan[1]

Résumé

Dans la littérature n'ayant pas trait aux sondages, il est fréquent de supposer que l'échantillonnage est effectué selon un processus aléatoire simple qui produit des échantillons indépendants et identiquement distribués (IID). De nombreuses méthodes statistiques sont élaborées en grande partie dans cet univers IID. Or, l'application de ces méthodes aux données provenant de sondages complexes en omettant de tenir compte des caractéristiques du plan de sondage peut donner lieu à des inférences erronées. Donc, beaucoup de temps et d'effort ont été consacrés à l'élaboration de méthodes statistiques permettant d'analyser les données d'enquêtes complexes en tenant compte du plan de sondage. Ce problème est particulièrement important lorsqu'on génère des populations synthétiques en faisant appel à l'inférence bayésienne en population finie, comme cela se fait souvent dans un contexte de données manquantes ou de risque de divulgation, ou lorsqu'on combine des données provenant de plusieurs enquêtes. En étendant les travaux antérieurs décrits dans la littérature sur le bootstrap bayésien en population finie, nous proposons une méthode pour produire des populations synthétiques à partir d'une loi prédictive a posteriori d'une façon qui inverse les caractéristiques du plan de sondage complexe et génère des échantillons aléatoires simples dans une optique de superpopulation, en ajustant les données complexes afin qu'elles puissent être analysées comme des échantillons aléatoires simples. Nous considérons une étude par simulation sous un plan de sondage en grappes stratifié avec probabilités inégales de sélection, et nous appliquons la méthode non paramétrique proposée pour produire des populations synthétiques pour la National Health Interview Survey (NHIS) et la Medical Expenditure Panel Survey (MEPS) de 2006, qui sont des enquêtes à plan de sondage en grappes stratifié avec probabilités inégales de sélection.

Mots-clés : Populations synthétiques; loi prédictive a posteriori; bootstrap bayésien; échantillonnage inverse.

1 Introduction

Hors du contexte des techniques d'enquête, les méthodes statistiques ont habituellement été élaborées sans beaucoup se soucier du plan d'échantillonnage, souvent en supposant implicitement avoir affaire à des échantillons aléatoires simples ou, parfois, à des échantillons en grappes à un degré. En statistique d'enquête contemporaine, d'importants travaux ont pour objectif d'étendre les méthodes à l'analyse de données d'enquêtes complexes (Skinner, Holt et Smith, 1989), en tenant compte de problèmes tels que la stratification, les probabilités inégales de sélection, le biais de non-réponse ou le calage. Hinkins, Oh et Scheuren (1997) ont proposé un algorithme de plan de sondage inverse qui relie la statistique d'enquête et la statistique classique sous un autre angle. Leur idée fondamentale consiste à choisir un sous-échantillon qui possède inconditionnellement une structure d'échantillon aléatoire simple. Le sous-échantillon est souvent nettement plus petit que l'échantillon original, de sorte qu'ils proposent de répéter le processus indépendamment un grand nombre de fois et de prendre la moyenne des résultats pour augmenter la précision. Ils décrivent aussi des schémas d'échantillonnage inverse exacts ou approximatifs pour l'échantillonnage aléatoire simple stratifié, l'échantillonnage en grappes à un degré et l'échantillonnage en grappes à deux degrés. Cependant, l'application de cette nouvelle idée n'est pas très répandue en pratique,

1. Qi Dong, Netflix Inc., 100, Winchester Cir, Los Gatos (CA) 95032, courriel : qidong@umich.edu; Michael R. Elliott, Department of Biostatistics, University of Michigan, 1420, Washington Heights, Ann Arbor (MI) 48109, Survey Methodology Program, Institute for Social Research, University of Michigan, 426, Thompson St., Ann Arbor (MI) 48106, courriel : mrelliot@umich.edu; Trivellore E. Raghunathan, Department of Biostatistics, University of Michigan, 1420, Washington Heights, Ann Arbor (MI) 48109, Survey Methodology Program, Institute for Social Research, University of Michigan, 426, Thompson St., Ann Arbor (MI) 48106. courriel : teraghu@umich.edu.

peut-être parce qu'elle est très gourmande en temps de calcul et que les pertes de précision sont souvent considérables. En outre, produire des populations synthétiques à partir d'une loi prédictive a posteriori de population conditionnellement aux données d'enquêtes complexes en tenant compte du plan de sondage complexe n'est pas chose simple (Little, 1991). Néanmoins, ces dernières années, la demande de populations synthétiques s'est accrue en vue de pouvoir traiter les problèmes de troncation des pondérations ou de windsorisation (Lazzeroni et Little, 1998; Elliott et Little, 2000; Elliott, 2007; Chen, Elliott et Little, 2010), de risque de divulgation (Little, 1993; Raghunathan, Reiter et Rubin, 2003; Reiter, 2004, 2005) ou de combinaison de données provenant de plusieurs enquêtes (Raghunathan, Xie, Schenker, Parsons, Davis, Dodd et Feuer, 2007; Dong, 2012). Les populations synthétiques sont souvent générées sous une hypothèse distributionnelle (normale, binomiale, Poisson) en approximant la loi a posteriori des paramètres du modèle par la loi normale asymptotique. La moyenne et la matrice de covariance de la loi normale sont estimées après avoir tenu compte des caractéristiques du plan de sondage complexe (Raghunathan et coll., 2007).

Une grande faiblesse des méthodes fondées sur un modèle tient au fait que, si le modèle est très mal spécifié, il donnera lieu à des inférences invalides (Little, 2004). Dans un contexte multivarié, nous devons prendre en considération les liens qui existent entre les variables d'intérêt et déterminer un modèle approprié qui est ajusté aux données, ce qui peut être difficile si les données contiennent différents types de variables. Dans le présent article, nous proposons une méthode non paramétrique qui fait pendant aux méthodes fondées sur un modèle pour générer des populations synthétiques. Les travaux que nous présentons étendent le bootstrap bayésien en population finie et les modèles a posteriori de Pólya connexes de Lo (1988), Ghosh et Meeden (1983) et Cohen (1997) en vue de tenir compte des plans de sondage complexes. Puisqu'elle atteint le même objectif que la méthode d'échantillonnage inverse, elle peut être traitée comme la version bayésienne en population finie de l'échantillonnage inverse. Pour faire des inférences en utilisant ce bootstrap bayésien en population finie pondéré, nous pouvons soit nous servir directement des tirages, soit, par souci d'efficacité des calculs, utiliser les résultats établis antérieurement dans la littérature sur le risque de divulgation et l'imputation multiple, puisque ces populations produites non paramétriquement peuvent être considérées comme des imputations multiples des éléments non observés de la population.

Le plan de l'article est le suivant. À la section 2, nous discutons brièvement des populations synthétiques dans le contexte de l'inférence bayésienne en population finie. À la section 3, nous passons en revue et résumons la méthode du bootstrap bayésien et son extension en population finie, et montrons que, pour un échantillonnage avec probabilités inégales, la loi de probabilité des populations synthétiques générées sous une variante du modèle de l'urne de Pólya concorde avec la loi prédictive a posteriori d'un bootstrap bayésien en population finie. À la section 4, nous présentons la méthode proposée sous échantillonnage en grappes stratifié avec probabilités de sélection inégales. À la section 5, nous montrons que l'inférence à partir de ces populations synthétiques générées non paramétriquement peut être obtenue en utilisant les résultats tirés de la littérature sur le risque de divulgation et l'imputation multiple, où chaque population synthétique possède une variance « intra-imputation » nulle. À la section 6, nous décrivons une étude par simulation réalisée pour évaluer la performance de la méthode non paramétrique dans un contexte de rééchantillonnage. À la section 7, nous appliquons la méthode pour générer des populations synthétiques qui peuvent être utilisées pour estimer les taux de couverture par une assurance maladie en utilisant les données de la NHIS et de la MEPS de 2006, et nous comparons le résultat à celui

d'une approche de modélisation paramétrique (log-linéaire). Enfin, nous présentons nos conclusions à la section 8.

2 Production de populations synthétiques à partir de données d'enquête

Le concept fondamental de l'inférence bayésienne en population finie consiste à imputer les valeurs non échantillonnées de la population à partir de la loi prédictive a posteriori basée sur les données observées. Supposons que les valeurs de population sont $Y = (Y_1, \ldots, Y_N)$ et que les données observées, $Y_{\text{obs}} = (y_1, \ldots, y_n)$, sont obtenues dans un sondage dont les indicatrices d'échantillonnage sont $I = (I_1, \ldots, I_N)$. L'inférence bayésienne sur la population permet d'utiliser le modèle paramétrique $\Pr(Y \mid \theta)$ pour les données de population basé sur la loi prédictive a posteriori pour les éléments non observés de la population $\Pr(Y_{\text{nob}} \mid Y_{\text{obs}})$:

$$\Pr(Y_{\text{nob}} \mid Y_{\text{obs}}) = \int \Pr(Y_{\text{nob}} \mid Y_{\text{obs}}, \theta) \Pr(\theta \mid Y_{\text{obs}}) \, d\theta$$

(Ericson, 1969; Little, 1993; Rubin, 1987; Scott, 1977; Skinner et coll., 1989). Ici, nous utilisons le modèle $\Pr(Y \mid \theta)$ pour approximer la distribution de la population complète $\Pr(Y)$ et prenons la moyenne sur la distribution a posteriori basée sur les données d'échantillon $\Pr(\theta \mid Y_{\text{obs}})$. S'il existe des variables de plan de sondage connues pour la population entière, le modèle susmentionné peut être étendu naturellement en conditionnant sur ces variables.

Le fait implicite dans la dérivation susmentionnée est que l'indicatrice d'échantillonnage I ne doit pas être modélisée. Cela exige que l'échantillonnage soit ignorable (Rubin, 1987) (la distribution de I ne doit pas dépendre des données observées), et que le modèle $\Pr(Y \mid \theta)$ utilisé pour les données tienne compte des caractéristiques du plan de sondage et soit suffisamment robuste pour saisir comme il convient tous les aspects pertinents de la distribution de la variable Y d'intérêt. Notre objectif ici est d'élaborer une méthode pour générer des tirages à partir de $\Pr(Y_{\text{nob}} \mid Y_{\text{obs}})$ qui tiennent compte de toutes les caractéristiques du plan de sondage dans Y_{obs}, de manière que les tirages à partir de la loi a posteriori de $Y_{\text{nob}} \mid Y_{\text{obs}}$ puissent être traités comme un échantillon aléatoire simple dans l'analyse.

3 Bootstrap bayésien en population finie pondéré

3.1 Bootstrap bayésien en population finie (BBPF)

Supposons que les éléments (scalaires) de population $Y_i, i = 1, \ldots, N$ soient échangeables et puissent prendre $K \leq N$ valeurs possibles $(b_1, \ldots, b_K)$; donc, $Y_i \mid \theta \sim \text{MULTI}(1; \theta_1, \ldots, \theta_K)$. Supposons aussi qu'une loi a priori conjuguée de Dirichlet pour $\theta \sim \text{DIR}(\alpha_1, \ldots, \alpha_K)$ donne (Ghosh et Meeden, 1983)

$$
\begin{aligned}
P(Y_{\text{nob}} \mid y) &= P\left(b_1^{\text{nob}} = N_1 - n_1, \ldots, b_K^{\text{nob}} = N_K - n_K \mid b_1^{\text{obs}} = n_1, \ldots, b_K^{\text{obs}} = n_K\right) \\
&= \frac{\int_0^1 \ldots \int_0^1 p(Y_{\text{nob}} \mid y, \theta)\, p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K}{\int_0^1 \ldots \int_0^1 p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K} \\
&= \frac{\int_0^1 \ldots \int_0^1 p(Y_{\text{nob}} \mid \theta)\, p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K}{\int_0^1 \ldots \int_0^1 p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K} \qquad (3.1) \\
&= \frac{\int_0^1 \ldots \int_0^1 \prod_{i=1}^K \theta_i^{N_i - n_i} \prod_{i=1}^K \theta_i^{n_i} \prod_{i=1}^K \theta_i^{\alpha_i - 1} d\theta_1 \ldots d\theta_K}{\int_0^1 \ldots \int_0^1 \prod_{i=1}^K \theta_i^{n_i} \prod_{i=1}^K \theta_i^{\alpha_i - 1} d\theta_1 \ldots d\theta_K} \\
&= \frac{\left(\prod_{i=1}^K \Gamma(N_i + \alpha_i)/\Gamma(\alpha_i)\right) \Big/ \left(\Gamma(N + \alpha_0)/\Gamma(\alpha_0)\right)}{\prod_{i=1}^K \Gamma(n_i + \alpha_i) \Big/ \Gamma(n + \alpha_0)}
\end{aligned}
$$

où $\alpha_0 = \sum_{i=1}^K \alpha_i$, $\sum_{i=1}^K N_i = N$, et $n_1, \ldots, n_K$ désigne le nombre de valeurs distinctes que nous observons à partir de notre échantillon $y = (y_1, \ldots, y_n)$, $\sum_{i=1}^K n_i = n$. Si $\alpha_i \equiv 0$, alors $p(Y_{\text{nob}} \mid y)$ se réduit à

$$\left(\prod_{i=1}^K \Gamma(N_i)/\Gamma(n_i)\right) \Big/ \left(\Gamma(N)/\Gamma(n)\right).$$

Pour faciliter la mise en œuvre, Lo (1988) a proposé de faire des tirages à partir de la loi prédictive a posteriori du BBPF en utilisant une procédure fondée sur le « modèle de l'urne de Pólya». Supposons qu'une urne contient n boules possédant chacune comme étiquette un nombre réel distinct $b_i, i = 1, \ldots, K$. Nous tirons un échantillon de Pólya de taille m en sélectionnant d'abord une boule au hasard dans l'urne et en remettant la boule sélectionnée dans l'urne, puis en plaçant une boule identique dans l'urne et en répétant ce processus jusqu'à ce que m boules aient été sélectionnées. On peut montrer que la probabilité d'obtenir m_i boules de type b_i est donnée par

$$p(b_1 = m_1, \ldots, b_K = m_K) = \frac{\prod_{i=1}^k \Gamma(n_i + m_i) \Big/ \Gamma(n_i)}{\Gamma(n + m) / \Gamma(n)} \qquad (3.2)$$

où n_i est le nombre de boules de type b_i se trouvant au départ dans l'urne. La distribution des nombres de boules de type b_i est invariante sous n'importe quelle permutation des tirages. Notons que cela correspond directement à la probabilité a posteriori d'un total de $(m_1, \ldots, m_K)$ éléments de type $(b_1, \ldots, b_K)$ dans une population, sachant que $(n_1, \ldots, n_K)$ éléments ont été observés dans un échantillon (aléatoire simple) de taille $\sum_{i=1}^K n_i = n$. Donc, nous pouvons tirer un échantillon réplique de cette loi a posteriori de Pólya en procédant aux étapes suivantes :

Étape 1. Tirer un échantillon de Pólya de taille $m = N - n$, noté $\left(y_1^*, \ldots, y_{N-n}^*\right)$ à partir de l'urne $\{y_1, \ldots, y_n\}$; en vertu de (3.2), avec $m_k = N_k - n_k$ tirages de la valeur b_k^{obs} pour $k = 1, \ldots, K$, cela correspond à un tirage de $P(Y_{\text{nob}} \mid y)$ à partir de (3.1).

Étape 2. Former la population BBPF $y_1, \ldots, y_n,\ y_1^*, \ldots, y_{N-n}^*$.

3.2 BBPF avec probabilités de sélection inégales

Cohen (1997) a étendu la procédure du BBPF afin de faire un ajustement pour les probabilités de sélection inégales. Supposons que $(y_1, \ldots, y_n)$ est un échantillon tiré d'une population finie $(Y_1, \ldots, Y_N)$ avec les poids de sondage $(w_1, \ldots, w_n)$, où

$$w_i = \frac{1}{P(I_i = 1)}$$

et I est l'indicatrice d'échantillonnage. La procédure comprend deux étapes :

Étape 1. Tirer un échantillon de taille $N - n$, noté $(y_1^*, \ldots, y_{N-n}^*)$, en tirant y_k^* à partir de $(y_1, \ldots, y_n)$ de manière que y_i soit sélectionné avec la probabilité

$$\frac{w_i - 1 + l_{i,k-1} * (N - n)/n}{N - n + (k - 1) * (N - n)/n},$$

où w_i est le poids de l'unité i et $l_{i,k-1}$ est le nombre de sélections bootstrap de y_i parmi les $y_1^*, \ldots, y_{k-1}^*$. (La fonction wtpolyap du module R *polypost* peut être utilisée pour obtenir des tirages à partir d'une urne de Pólya pondérée.)

Étape 2. Former la population BBPF $y_1, \ldots, y_n,\ y_1^*, \ldots, y_{N-n}^*$.

Cohen (1997) n'a pas fourni la preuve théorique de cette procédure, mais elle peut être obtenue comme une extension simple de l'équivalence du BBPF et de l'urne de Pólya classique décrite à la section 3.1. Premièrement, nous déterminons la loi a posteriori de l'échantillon BBPF avec probabilités de sélection inégales qu'implique la procédure BBPF pondérée. La vraisemblance multinomiale fondée sur notre échantillon pondéré est donnée par

$$p(y_{\text{obs}} \mid \theta) = \prod_{i=1}^{K} \theta_i^{w_i^*},$$

où

$$w_i^* = \left(\frac{n}{N - n}\right) \sum_{j=1}^{n} I\left(y_j = b_i\right)\left(w_j - 1\right)$$

est la somme des poids de sondage moins une unité sur l'ensemble des éléments échantillonnés ayant la valeur $b_i, i = 1, \ldots, K$, normalisée pour qu'elle soit égale à n. (Soulignons que cela élimine de la vraisemblance les sujets échantillonnés dont le poids est égal à un, c'est-à-dire les éléments de l'« échantillon sélectionné avec certitude », car ils n'ont aucune chance de se trouver dans la partie non observée de la population, et donc n'apportent aucune information au sujet des éléments non observés.) En émettant l'hypothèse d'une loi a priori de Dirichlet impropre $p(\theta) = \prod_{i=1}^{k} \theta_i^{-1}$, la loi a posteriori du bootstrap bayésien en population finie pondéré est donnée par

$$
\begin{aligned}
P(Y_{\text{nob}} \mid y, w) &= P\left(b_1^{\text{nob}} = r_1, \ldots, b_K^{\text{nob}} = r_K \mid w_1^*, \ldots, w_K^*\right) \\
&= \frac{\int_0^1 \cdots \int_0^1 p(Y_{\text{nob}} \mid \theta)\, p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K}{\int_0^1 \cdots \int_0^1 p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K} \\
&= \frac{\int_0^1 \cdots \int_0^1 \prod_{i=1}^K \theta_i^{r_i} \prod_{i=1}^K \theta_i^{w_i^*} \prod_{i=1}^K \theta_i^{-1} d\theta_1 \cdots d\theta_K}{\int_0^1 \cdots \int_0^1 \prod_{i=1}^K \theta_i^{w_i^*} \prod_{i=1}^K \theta_i^{-1} d\theta_1 \cdots d\theta_K} \\
&= \prod_{i=1}^K \frac{\Gamma\left(w_i^* + r_i\right)}{\Gamma\left(w_i^*\right)} \Big/ \frac{\Gamma(N)}{\Gamma(n)}
\end{aligned}
\tag{3.3}
$$

puisque $\sum_{j=1}^n r_i = N - n$ et $\sum_{j=1}^n w_i^* = n$.

Ensuite, nous montrons que la distribution des échantillons obtenus à partir du modèle d'urne de Pólya sous probabilités inégales de sélection de Cohen (1997) est égale à la loi a posteriori de l'échantillon BBPF avec probabilités de sélection inégales. Sachant les données observées, la probabilité de tirer $N - n$ boules et que les premières boules r_1 aient la valeur b_1 , et ainsi de suite, et que les dernières, r_k, aient la valeur b_k est :

$$
\begin{aligned}
P(b_1 = r_1, \ldots, b_K = r_K) &= \frac{w_1^*}{n} \times \frac{w_1^* + 1}{n+1} \ldots \times \frac{w_1^* + r_1 - 1}{n + r_1 - 1} \times \ldots \times \frac{w_K^*}{n + \sum_{i=1}^{k-1} r_i} \times \ldots \times \frac{w_K^* + r_K - 1}{n + \sum_{i=1}^{k} r_i - 1} \\
&= \prod_{i=1}^K \frac{\Gamma\left(w_i^* + r_i\right)}{\Gamma\left(w_i^*\right)} \Big/ \frac{\Gamma(N)}{\Gamma(n)}
\end{aligned}
$$

où la première égalité découle du fait que la distribution des nombres de boules de type b_i est invariante sous toute permutation des tirages, comme dans le cas non pondéré, et la deuxième égalité découle de l'identité $\Gamma(x) = (x-1)\Gamma(x)$ pour $x > 0$. Donc, en notant que

$$
\frac{w_i - 1 + l_{i,k-1} * (N-n)/n}{N - n + (k-1) * (N-n)/n} = \frac{w_i^* + l_{i,k-1}}{n + (k-1)},
$$

un tirage à partir du modèle de l'urne de Pólya avec probabilités de sélection inégales donne un tirage à partir de $P(Y_{\text{nob}} \mid y, w)$ dans (3.3).

4 Méthode non paramétrique de production de populations synthétiques

À la présente section, nous étendons les méthodes du bootstrap bayésien en population finie au cas d'un plan de sondage en grappes stratifié avec probabilités de sélection inégales en vue d'élaborer une méthode non paramétrique de production de populations synthétiques qui comporte un ajustement pour tenir compte des caractéristiques du plan de sondage complexe. L'idée est de traiter la partie non observée de la population comme des données manquantes et de l'imputer en effectuant des tirages à partir des données réelles. Nous faisons l'imputation de manière que les tirages résultants à partir de la loi

a posteriori de la population reflètent les caractéristiques du plan de sondage complexe et puissent être utilisés de la façon classique pour calculer les lois a posteriori des quantités d'intérêt de la population.

4.1 Utilisation du bootstrap bayésien pour l'ajustement pour la stratification et la mise en grappe

Dans le cas d'un échantillonnage en grappes stratifié, nous devons d'abord rééchantillonner les grappes à l'intérieur des strates. Notons c le nombre total de grappes dans les données réelles, $c = \sum_{h=1}^{H} c_h$, et C le nombre de grappes dans la population, $C = \sum_{h=1}^{H} C_h$. Une approche consiste à appliquer d'abord le modèle de l'urne de Pólya avec le BBPF pour imputer les grappes non observées à l'intérieur de chaque strate, $c_1^*, \ldots, c_{C_h - c_h}^*$, qui, avec les grappes observées, fournissent les grappes dans la strate h de la population. Cependant, les données à grande diffusion disponibles ne nous permettent habituellement pas de savoir quel est le nombre de grappes dans une strate. Donc, comme alternative au tirage d'un échantillon BBPF, nous proposons le tirage d'un échantillon bootstrap bayésien classique de grappes dans chaque strate. En tenant compte de l'équivalence entre le bootstrap classique et le bootstrap bayésien, nous procédons comme Rao et Wu (1988), qui ont proposé de tirer un échantillon aléatoire simple avec remise (EASAR) de taille m_h à partir des c_h grappes et, à l'intérieur de chaque strate h, de calculer les poids de rééchantillonnage pour chaque échantillon bootstrap comme

$$w^{*(l)} = \left\{ w_{hik}^{*(l)},\ h = 1, \ldots, H,\ i = 1, \ldots, c_h,\ k = 1, \ldots, N_{hi} \right\},$$

où

$$w_{hik}^* = w_{hik} \left(\left(1 - \sqrt{\frac{m_h}{c_h - 1}} \right) + \sqrt{\frac{m_h}{c_h - 1}} \frac{c_h}{m_h} m_{hi}^* \right)$$

et m_{hi}^* désigne le nombre de fois que la grappe $i, i = 1, \ldots, c_h$ est sélectionnée. Pour être certain que tous les poids de rééchantillonnage soient non négatifs, il faut que $m_h \leq (c_h - 1)$; ici et plus loin, nous prenons $m_h = (c_h - 1)$.

Notons qu'en l'absence de grappes, nous tirons simplement un échantillon bootstrap bayésien classique à partir des données échantillonnées dans chaque strate (s'il existe une stratification) ou à partir de l'échantillon complet (en l'absence de stratification, de sorte que $H = 1$) et nous calculons les poids de rééchantillonnage comme étant $w_{hik}^* = w_{hik} m_{hi}^*$.

Nous répétons cette procédure L fois pour produire L échantillons bootstrap bayésiens (BB) notés $S_1, \ldots, S_L$. Cette étape génère L échantillons bootstrap bayésiens qui sont essentiellement L tirages à partir de la loi prédictive a posteriori des grappes non observées sachant les données réelles. Cependant, les unités formant les L échantillons bootstrap bayésiens possèdent encore des poids et ne peuvent être analysées comme s'il s'agissait d'échantillons aléatoires simples.

4.2 Utilisation du modèle de l'urne de Pólya avec le BBPF pondéré pour faire un ajustement pour la pondération

Une fois que nous avons obtenu L échantillons BB avec les poids de rééchantillonnage, la deuxième étape consiste à imputer les unités non observées en utilisant le modèle de l'urne de Pólya avec le BBPF

pondéré. En pratique, la probabilité de sélection de la k^{e} unité, y_k^*, dépend de la sélection des $k-1$ premières unités, $y_1^*, \ldots, y_{k-1}^*$. Autrement dit, pour déterminer la probabilité de sélection d'une nouvelle unité, nous devons compter le nombre de fois que chaque unité présente dans l'échantillon a été sélectionnée parmi les sélections antérieures. Si la taille de la population est très grande, il nous suffit de générer des populations synthétiques de taille $T * n$, où T est suffisamment grand pour dominer la taille d'échantillon (p. ex. 20-100). Pour accroître encore davantage l'efficacité des calculs, nous pourrions aussi tirer $F > 1$ fois une population de taille modérée, puis regrouper ces F populations pour produire une population synthétique, S_l. La taille de S_l est alors $F * T * n$.

Notons que, sous notre méthode, il ne faut connaître que les poids finaux dans les échantillons en grappes à plusieurs degrés, puisque tous les degrés d'échantillonnage avec probabilités de sélection inégales sont corrigés par l'utilisation du modèle de l'urne de Pólya avec le BBPF pondéré. Cette caractéristique de la méthode proposée est particulièrement utile, car dans de nombreux jeux de données à grande diffusion, les composantes des probabilités de sélection (p. ex. probabilités de sélection au niveau de la grappe, poids de non-réponse) ne sont pas disponibles.

5 Inférence à partir de multiples populations synthétiques non paramétriques

Supposons que nous produisons L populations synthétiques, $S_l, l = 1, \ldots, L$ en utilisant la méthode non paramétrique décrite à la section 4, et que notre cible d'inférence est $Q \equiv Q(Y)$, une fonction des données de population (p. ex. moyenne de population, corrélation, estimateur du maximum de vraisemblance de population d'un paramètre de régression). Nous pouvons calculer Q_l comme étant l'estimation de Q obtenue en regroupant les F populations synthétiques utilisées pour imputer les unités non observées de S_l; puisqu'il s'agit de tirages directs à partir de la loi prédictive a posteriori de la population, nous pouvons calculer les moyennes, les quantiles et les intervalles de crédibilité a posteriori d'après les estimations empiriques correspondantes à partir des tirages, si L est suffisamment grand.

Cependant, dans de nombreuses situations, le temps de calcul nécessaire pour imputer la population peut être très important, même s'il ne faut pas synthétiser la population complète. D'où, une autre approche de l'inférence consiste à utiliser la loi t comme approximation de la distribution prédictive a posteriori d'une statistique de population scalaire Q :

$$Q \mid S_1, \ldots, S_L \overset{\cdot}{\sim} t_{L-1}\left(\bar{Q}_L, \left(1 + L^{-1}\right) V_L\right)$$

où

$$\bar{Q}_L = \frac{\sum_{l=1}^{L} Q_l}{L} = \frac{\sum_{l=1}^{L} \sum_{f=1}^{F} Q_{lf}}{LF} \quad \text{et} \quad V_L = \frac{1}{L} \sum_{l=1}^{L} \left(Q_l - \bar{Q}_L\right)^2 .$$

Le résultat découle directement de la section 4.1 dans Raghunathan et coll. 2003, et est fondé sur les règles de combinaison classiques pour l'imputation multiple de Rubin (1987), en traitant les unités non observées de S_l comme des données manquantes et les unités échantillonnées, comme des données observées. La variance « intra-imputation » moyenne est nulle, puisque la population est entièrement synthétisée; d'où,

la variance a posteriori de Q est entièrement une fonction de la variance inter-imputations, et le nombre de degrés de liberté est simplement donné par le nombre d'échantillons BBPF. (Lorsque la population est très grande, il nous suffit de synthétiser un tirage assez grand pour que la variance « intra-imputation » moyenne soit négligeable comparativement à la variance inter-imputations V_L.) Le résultat suppose que $E(Q_{lf}) = Q$, ce que notre estimateur BBPF pondéré garantit, et que la taille de l'échantillon est suffisamment grande pour permettre d'appliquer la théorie asymptotique bayésienne.

6 Études par simulation

À la présente section, nous décrivons deux études par simulation réalisées pour évaluer les propriétés de rééchantillonnage des estimateurs de population construits en utilisant la méthode non paramétrique qui produit des populations synthétiques en effectuant un ajustement pour tenir compte des caractéristiques du plan de sondage complexe. La première simulation porte sur un plan de sondage à un degré avec probabilités de sélection inégales dans lequel nous faisons varier le nombre de tirages BBPF pondérés pour chaque population synthétique, ainsi que le nombre de populations synthétiques pour évaluer l'effet sur l'inférence. La deuxième simulation a pour objectif de comparer les propriétés inférentielles à partir des données observées et à partir de la loi a posteriori obtenue d'après la population synthétique sous un plan de sondage stratifié à plusieurs degrés avec probabilités de sélection inégales, cette fois-ci en fixant la taille de l'échantillon a posteriori, tout en considérant la moyenne de population ainsi que les paramètres de régression de population comme les cibles des inférences.

6.1 Plan de sondage à un degré avec probabilités de sélection inégales

Nous avons généré les données pour la variable de résultat Y dans une population de N sujets provenant d'une loi Gamma modérément asymétrique, conditionnellement à la covariable X qui suit une loi uniforme :

$$X_i \sim \text{UNI}(0,05; 0,65), i = 1, \ldots, N$$

$$Y_i \mid X_i = x_i \sim \text{GAMMA}(10 * x_i, 1)$$

Nous supposons que X est entièrement observé pour la population, et que la probabilité de sélection π est proportionnelle à X, de sorte que $\pi_i \doteq nx_i \big/ \sum_i x_i$ dans un plan de sondage sans remise à condition que $n << N$. La quantité à estimer est la moyenne de population $\bar{Y} = N^{-1} \sum_{i=1}^{N} y_i = 3{,}564$. Notons que $\text{corr}(Y_i, X_i) = 0{,}6794$, de sorte que les moyennes d'échantillon non pondérées présentent un biais positif, et que l'utilisation des poids de sondage $w_i = 1/\pi_i$ est nécessaire pour obtenir des estimations sans biais de $\bar{Y}$. Nous générons une population de taille $N = 1\,000$ à partir de laquelle nous tirons $n = 100$ échantillons; nous estimons ensuite le biais, la variance empirique et la variance estimée, la longueur de l'intervalle de confiance à 95 % et la couverture au niveau de confiance nominal de 95 % au moyen de 200 échantillons indépendants tirés de la population. Nous faisons varier le nombre total de populations simulées L qui prend les valeurs de 5, 20, 100 et 1 000, ainsi que le nombre F de tirages BBPF de taille $N - n$ (de manière que $K = 9$) qui prend les valeurs de 1, 20 et 100, dans un plan factoriel complet. Nous obtenons la variance, la longueur de l'intervalle et la couverture de l'intervalle au moyen de l'approximation normale; pour $L = 100$ et 1 000, nous obtenons également la variance, la longueur de

l'intervalle et la couverture de l'intervalle en utilisant les tirages directs à partir de la loi prédictive a posteriori, puisque nous disposons d'un nombre suffisant de tirages à partir de cette loi pour produire ces estimations.

Le tableau 6.1 donne les résultats de l'étude par simulation. Dans tous les cas, l'estimation ponctuelle $\bar{Q}_L$ de la moyenne de population est approximativement sans biais, ce qui témoigne de la capacité du BBPF pondéré à « défaire » les poids de sondage pour produire la population synthétique. Sous l'approximation normale, l'augmentation du nombre de populations synthétiques est associée à de plus petites variances et des intervalles plus étroits, comme il fallait s'y attendre sous un plus grand nombre de degrés de liberté, quoique la différence entre les résultats obtenus pour 20 et 100 populations soit minime, juste quand la loi t_{20} commence à s'approcher d'une loi normale standard. Enfin, l'utilisation d'un seul tirage BBPF de taille $N-n$ semble donner lieu à une surestimation de la variance et à un surdénombrement, surtout pour les petites valeurs de L. Les valeurs de L et de F égales ou supérieures à 20 semblent donner des résultats raisonnables. L'utilisation des tirages directs pour $L=100$ et 1 000 produit des estimations de variance et d'intervalle de crédibilité qui sont fort semblables à celles données par l'approximation normale, les longueurs d'intervalle étant toutefois légèrement plus courtes et les couvertures un peu moins conservatrices.

Tableau 6.1
Biais, variance empirique, moyenne de la variance estimée, longueur de l'intervalle et couverture de l'intervalle de confiance au niveau nominal de 95 % d'une moyenne de population en fonction du nombre de populations synthétiques (L) et du nombre de tirages par bootstrap bayésien en population finie pondéré qui constituent la population synthétique (F). Longueur et couverture de l'intervalle obtenues par approximation par la loi t et empiriquement par simulation directe. Plan de sondage à un degré avec probabilités de sélection inégales. Résultats pour 200 simulations.

L	**5**			**20**			**100**			**1 000**		
F	**1**	**20**	**100**	**1**	**20**	**100**	**1**	**20**	**100**	**1**	**20**	**100**
Biais	-0,020	0,009	-0,026	0,021	-0,030	0,010	-0,031	0,024	-0,028	-0,045	-0,070	0,079
Variance emp.	0,126	0,099	0,106	0,088	0,092	0,120	0,093	0,079	0,085	0,084	0,093	0,078
Variance est. : t	0,172	0,119	0,105	0,156	0,098	0,099	0,109	0,097	0,095	0,147	0,104	0,094
Longueur de l'intervalle : t	2,20	1,78	1,71	1,63	1,30	1,32	1,52	1,21	1,20	1,50	1,26	1,20
Couverture IC à 95 % : t	97	95	96	99	94	92	98	96	95	98	96	98
Variance est. : Empirique	0,138	0,095	0,084	0,148	0,093	0,094	0,108	0,096	0,094	0,084	0,093	0,078
Longueur de l'intervalle : Empirique	s.o.	s.o.	s.o.	s.o.	s.o.	s.o.	1,50	1,19	1,18	1,49	1,25	1,19
Couverture IC à 95 % : Empirique	s.o.	s.o.	s.o.	s.o.	s.o.	s.o.	96	93	94	98	96	97

6.2 Plan de sondage stratifié à plusieurs degrés avec probabilités de sélection inégales

Nous avons généré une population comprenant des strates et des grappes dans chaque strate à partir de la loi normale bivariée suivante :

$$\begin{pmatrix} X_{1ijk} \\ X_{2ijk} \end{pmatrix} \sim N\left(\begin{pmatrix} 500 + 4,5*i + u_{ij} \\ 500 + 4,5*i + u_{ij} \end{pmatrix}, \begin{pmatrix} 100 & 50 \\ 50 & 100 \end{pmatrix}\right),$$

où

$i = 1/150$ désigne l'effet de strate;

$u_{ij} \sim N(0,10)$ désigne l'effet de grappe aléatoire;

$a_i \sim \text{uniforme}(2,52)$ désigne le nombre de grappes dans la strate i;

$b_{ij} \sim \text{uniforme}(10,20)$ désigne le nombre d'unités dans la grappe j de la strate i.

La population utilisée pour l'étude par simulation compte 61 324 sujets. Nous tirons un échantillon en grappes stratifié avec probabilités de sélection inégales. Plus précisément, nous sélectionnons deux grappes dans chaque strate avec probabilités proportionnelles à la taille de grappe (PPT) données par $b_{i\bullet} = \sum_{j=1}^{a_i} b_{ij}$. Dans chaque grappe sélectionnée, nous sélectionnons environ un cinquième $(1/5)$ de la population. Donc, la probabilité que l'unité ij soit sélectionnée est donnée par

$$\pi_{ij} = \frac{2b_{i\bullet}}{\sum_{j=1}^{a_i} b_{ij}} \times \frac{\lfloor b_{ij}/5 \rfloor}{b_{ij}}$$

pour tout élément j dans la grappe i avec les poids correspondants

$$w_{ij} = \frac{b_{ij} \sum_{j=1}^{a_i} b_{ij}}{2b_{i\bullet} \lfloor b_{ij}/5 \rfloor}.$$

Puisque les nombres de grappes et d'unités sont aléatoires, la taille de l'échantillon complexe diffère légèrement d'une réplique à l'autre, la moyenne étant approximativement de 770.

Comme l'échantillon et la population sont de grande taille, nous nous concentrons sur l'inférence en utilisant les approximations t. Nous générons $L = 100$ populations synthétiques en utilisant F échantillons BBPF pondérés de taille $K = 100n$. Les quantités à estimer sont la moyenne marginale de population pour x_1

$$\bar{X}_1 = N^{-1} \sum_{i=1}^{N} X_{1i}$$

et celle pour x_2, obtenue de manière similaire, ainsi que les coefficients de régression de x_1 sur x_2, donnés par

$$B_0 = \bar{X}_1 - B_1 \bar{X}_2, B_1 = \frac{\sum_{i=1}^{N} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^{N} (X_{2i} - \bar{X}_2)^2}.$$

Nous avons tiré 200 échantillons indépendants de la population et utilisé les données d'échantillon pour calculer directement les moyennes et les coefficients de régression linéaire pour l'échantillon pondéré, ainsi que les estimations correspondantes des variances et des intervalles de confiance au niveau nominal de 95 % en utilisant des approximations par développement en série de Taylor, et les avons comparées aux estimations équivalentes obtenues en utilisant les données synthétiques non paramétriques. Les résultats sont présentés au tableau 6.2. (Puisque les moyennes marginales ont la même valeur de superpopulation, nous combinons les résultats dans le tableau 6.2.) La figure 6.1 donne le diagramme de dispersion des paires de moyennes, d'ordonnées à l'origine et de pentes estimées d'après les échantillons réels et les

populations synthétiques correspondantes, ainsi qu'une droite à 45 degrés. Les distributions d'échantillonnage des estimations d'après les échantillons réels et les populations synthétiques sont très proches. Les estimations ponctuelles et les erreurs-types des moyennes ainsi que des paramètres de régression sont très proches. Les taux de couverture des intervalles de confiance à 95 % sont très proches pour les trois statistiques, et sont proches des valeurs nominales.

Tableau 6.2
Statistiques descriptives et analytiques estimées d'après les données réelles et les populations synthétiques dans une évaluation par simulation de la méthode non paramétrique. Plan de sondage à deux degrés stratifié avec probabilités de sélection inégales. Résultats d'après 200 simulations.

Type	Données réelles				Populations synthétiques			
	Estimation	e.-t.	É.-T.	Couverture (%)	Estimation	e.-t.	É.-T.	Couverture (%)
Moyenne $\bar{X}$	836,701	0,461	0,491	93	836,793	0,476	0,493	94
Ordonnée à l'origine B_0	1,013	1,768	1,848	94	1,014	1,775	1,846	92
Pente B_1	0,999	0,002	0,002	92	0,999	0,002	0,002	92

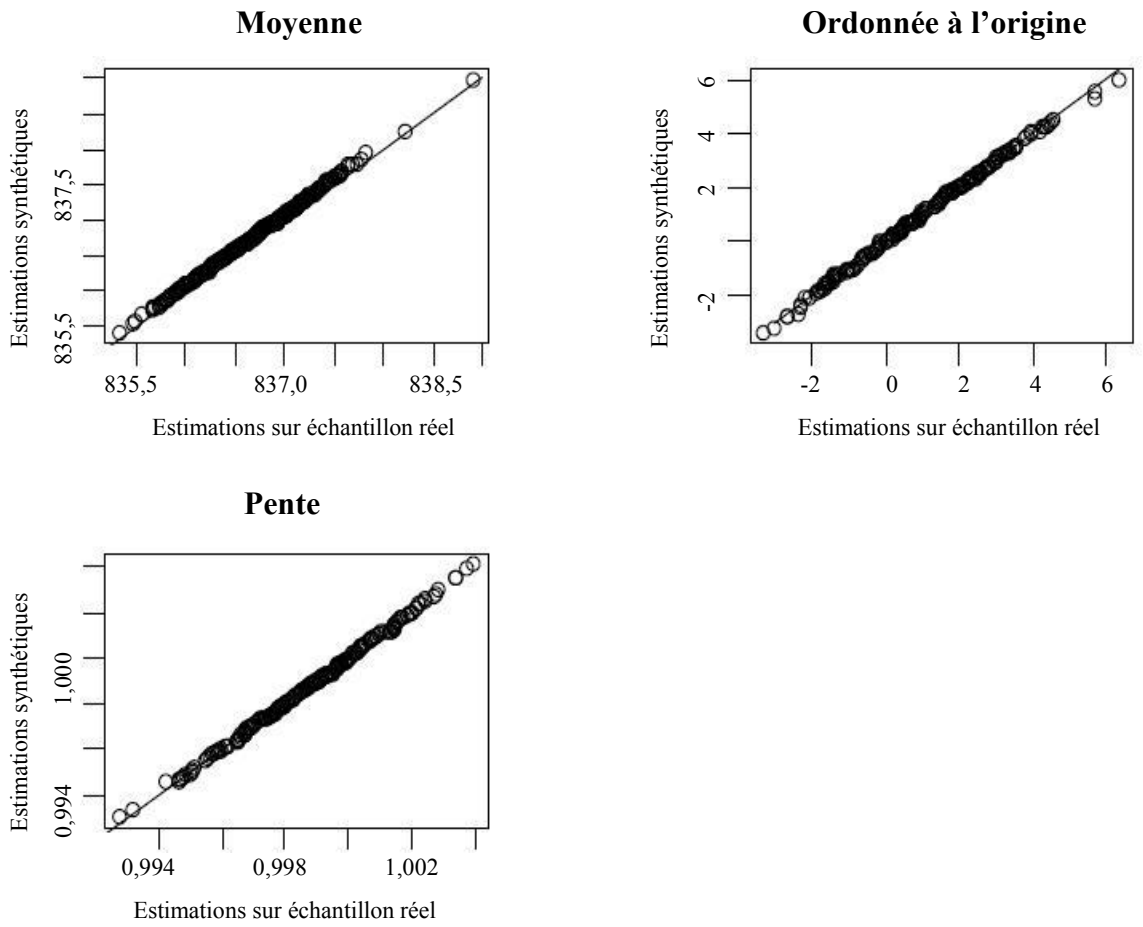

Figure 6.1 Diagramme de dispersion des statistiques descriptives et analytiques pour les populations réelles et synthétiques

7 Application

À la présente section, nous utilisons des données tirées de la National Health Interview Survey (NHIS) de 2006 et de la Medical Expenditure Panel Survey (MEPS) de 2006 pour évaluer la performance de la méthode non paramétrique sous un plan de sondage en grappes stratifié. La National Health Interview Survey (NHIS) est une enquête sur la santé de portée nationale, réalisée par interview en personne selon un plan stratifié à plusieurs degrés avec suréchantillonnage des populations noires, hispaniques et âgées. Pour des raisons de confidentialité, la stratification et les variables au niveau de l'unité d'échantillonnage (UPE) réelles ne sont pas communiquées dans les fichiers de données à grande diffusion; elles sont remplacées par des pseudo-strates et UPE (deux par strate). La MEPS est réalisée auprès d'un sous-échantillon de l'échantillon de la NHIS de l'année précédente, selon le même plan stratifié à plusieurs degrés.

Tant dans la NHIS que dans la MEPS, on demande aux participants à l'enquête s'ils sont couverts par une assurance maladie et, dans l'affirmative, quel régime d'assurance maladie ils utilisent (privé par opposition à public tel que Medicare ou Medicaid). Nous estimons les taux globaux de couverture par une assurance maladie, ainsi que les taux de couverture dans des sous-populations définies en fonction de variables démographiques telles que le sexe, la race, le niveau de revenu ou des combinaisons de ces variables; en particulier, nous estimons la couverture par une assurance maladie des hommes, des Blancs non hispaniques et des Blancs non hispaniques dont le revenu du ménage est compris entre 25 000 $ et 35 000 $ par année. Nous supprimons les cas pour lesquels les valeurs manquent pour certaines questions et nous axons notre simulation sur les cas complets. Nous obtenons ainsi 20 147 et 20 893 cas pour les données de la NHIS et de la MEPS, respectivement.

7.1 Estimation de la couverture par une assurance maladie d'après la NHIS et la MEPS

Dans la présente étude par simulation, nous utilisons la méthode non paramétrique pour apporter un ajustement pour tenir compte de l'échantillonnage en grappes stratifié utilisé dans la NHIS et la MEPS de 2006, et pour produire des populations synthétiques qui peuvent être analysées comme des échantillons aléatoires simples. Nous considérons également une approche fondée sur un modèle pour produire des populations synthétiques en utilisant un modèle log-linéaire pour la situation de couverture par une assurance maladie en fonction de six variables démographiques indépendantes : sexe, race, région de recensement, niveau de scolarité, âge (catégorique) et revenu du ménage (catégorique). Ensuite, nous évaluons la méthode en comparant les estimations du taux de couverture par une assurance maladie pour l'ensemble de la population et pour les sous-domaines choisis pour les populations synthétiques obtenues par la méthode non paramétrique et par celle du modèle log-linéaire à celles obtenues au moyen des données réelles.

7.1.1 Production de populations synthétiques non paramétriques

En utilisant la méthode non paramétrique élaborée à la section 3, nous produisons 200 populations synthétiques pour chaque enquête. Plus précisément, nous générons $B = 200$ échantillons BB et, pour chacun de ces échantillons, nous générons $F = 10$ échantillons BBPF de taille $5n\,(K = 5)$. Donc, chaque

population synthétique est 50 fois plus grande que l'échantillon réel (1 007 350 pour la NHIS, 1 044 650 pour la MEPS). Chaque population synthétique est analysée comme un échantillon aléatoire simple et les estimations sont combinées comme il est décrit à la section 5.

7.1.2 Production de populations synthétiques au moyen de modèles log-linéaires

Dans la situation fréquente où les données d'enquête d'intérêt prennent la forme d'un tableau de contingence multidimensionnel, un modèle log-linéaire pourrait être considéré comme une approche paramétrique pour générer des tirages à partir d'une loi prédictive a posteriori. Pour simplifier l'exposé, supposons que Y est la variable d'intérêt comprenant m niveaux, et que Z est une variable de plan comprenant n niveaux (p. ex. sexe ou race) dont la loi de probabilité marginale est connue pour la population. Supposons que $\pi_{ij}, i = 1,\ldots,m, j = 1,\ldots,n,$ représente la proportion dans la ij^{e} cellule, $\sum_{i=1}^{m}\sum_{j=1}^{n}\pi_{ij} = 1.$ Un modèle log-linéaire entièrement saturé est donné par (Agresti, 2002) :

$$\log\left(\pi_{ij}\right) = \lambda_0 + \lambda_i^Z + \lambda_j^Y + \lambda_{ij}^{ZY},\ i = 1,\ldots,m, j = 1,\ldots,n,$$

où $\log\left(\pi_{ij}\right)$ est le logarithme de la probabilité qu'une observation se trouve dans la cellule ij du tableau de contingence, λ_i^Z est l'effet principal pour Z, λ_j^Y est l'effet principal pour Y et λ_{ij}^{ZY} est l'effet d'interaction pour Z et Y. Ce modèle comprend tous les effets unidimensionnels et bidimensionnels possibles, et est donc saturé, car il contient le même nombre d'effets que de cellules dans le tableau de contingence. Pour éviter de surajuster les données dans l'exemple, nous pouvons considérer des modèles non saturés dont sont exclus certains termes d'interaction, voire tous, en choisissant le modèle en nous basant sur des tests de rapport de vraisemblance, ou sur le critère AIC ou BIC.

Les populations synthétiques peuvent être générées à partir de la distribution prédictive a posteriori issue du modèle. Toutefois, si les données sont recueillies selon un plan de sondage complexe, nous ne connaissons aucun logiciel statistique standard capable de produire à la fois l'estimation ponctuelle et l'estimation de covariance des coefficients de régression. Nous avons donc choisi d'utiliser une méthode de rééchantillonnage jackknife pour tenir compte de la stratification, de la mise en grappe et de la pondération. Plus précisément, les populations synthétiques paramétriques peuvent être générées selon les étapes suivantes :

1. Estimer les coefficients et la matrice de covariance :

Sous le modèle choisi (supposé être le modèle saturé bidimensionnel ici, simplement pour l'illustration), estimer les coefficients $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)'$, $i = 1,\ldots,m-1, j = 1,\ldots,n-1$ et la matrice de covariance des estimations $\hat{\lambda} = \left(\hat{\lambda}_0, \hat{\lambda}_i^Z, \hat{\lambda}_j^Y, \hat{\lambda}_{ij}^{ZY}\right)'$ après avoir tenu compte des caractéristiques du plan complexe en utilisant la méthode des répliques équilibrées jackknife (REJ) :

- Pour chaque réplique, retirer une grappe et augmenter les poids de sondage des unités des autres grappes à l'intérieur de la même strate d'un facteur $c_h/(c_h - 1)$ (poids de rééchantillonnage), où c_h désigne le nombre de grappes dans la strate h. En supposant que nous avons un total de

$\sum_{h=1}^{H} c_h = C$ grappes, nous avons alors C répliques. Pour chaque réplique, nous ajustons le modèle log-linéaire et obtenons les estimations du maximum de vraisemblance (EMV) des coefficients $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)'$, $i = 1, \ldots, m-1$, $j = 1, \ldots, n-1$.

- Pour chaque réplique, utiliser les poids de rééchantillonnage pour ajuster le modèle log-linéaire. Plus précisément, utiliser les poids de rééchantillonnage pour calculer la taille de chaque cellule du tableau de contingence, qui est utilisé pour ajuster le modèle log-linéaire. Nous notons l'EMV pour la r^{e} réplique comme un vecteur colonne, $\hat{\lambda}_r, r = 1, \ldots, c_h$ pour la strate h. Soulignons que $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)'$, $i = 1, \ldots, m-1$, $j = 1, \ldots, n-1$ est un vecteur colonne de dimension mn par 1. Nous le notons $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)' = (\lambda_0, \lambda_1, \ldots, \lambda_{mn})'$. De même, $\hat{\lambda}_r, r = 1, \ldots, c_h, h = 1, \ldots, H$ sont aussi des vecteurs colonnes de dimensions mn par 1 que nous notons $\left(\hat{\lambda}_0^{(r)}, \hat{\lambda}_1^{(r)}, \ldots, \hat{\lambda}_{mn}^{(r)}\right)'$.

L'EMV des coefficients $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)'$, $i = 1, \ldots, m-1, j = 1, \ldots, n-1$ peut être obtenu comme $\hat{\lambda}_{\text{EMV}} = \sum_{h=1}^{H} \sum_{r=1}^{c_h} \hat{\lambda}_r \big/ C$. Pour la matrice de covariance de dimensions mn par mn, l'estimation par rééchantillonnage jackknife du pq^{e} $(p, q = 1, \ldots, mn)$ élément est la covariance entre les p^{e} et q^{e} coefficients, qui est donnée par :

$$\sum_{h=1}^{H} \frac{c_h - 1}{c_h} \sum_{r=1}^{c_h} \left(\hat{\lambda}_p^{(r)} - \bar{\hat{\lambda}}_p\right)\left(\hat{\lambda}_q^{(r)} - \bar{\hat{\lambda}}_q\right),$$

où $\bar{\hat{\lambda}}_p = \sum_{h=1}^{H} \sum_{r=1}^{c_h} \hat{\lambda}_p^{(r)} \big/ C$ et $\bar{\hat{\lambda}}_q = \sum_{h=1}^{H} \sum_{r=1}^{c_h} \hat{\lambda}_q^{(r)} \big/ C$. Cela nous donne l'estimation de variance correcte de $\hat{\lambda}_{\text{EMV}}$.

2. Obtenir une approximation de la loi a posteriori des coefficients :

Soit T la décomposition de Cholesky telle que $TT^t = \text{cov}\left(\hat{\lambda}_{\text{EMV}}\right)$. Générer un vecteur z de variables aléatoires normales standardisées et définir $\Lambda_* = \hat{\lambda}_{\text{EMV}} + Tz$.

3. Imputer les valeurs non observées de la population :

Supposons que l'on procède à L tirages, $\Lambda_1, \ldots, \Lambda_L$, à partir de la loi a posteriori approximative de λ. Pour chaque

$$l = 1, \ldots, L, \Lambda_l = \left(\Lambda_0^{(l)}, \Lambda_i^{X(l)}, \Lambda_j^{Y(l)}, \Lambda_{ij}^{XY(l)}\right)', \; i = 1, \ldots, m-1, j = 1, \ldots, n-1,$$

nous pouvons générer un tableau synthétique en utilisant le modèle supposé :

$$\log\left(\pi_{ij}^{(l)}\right) = \Lambda_0^{(l)} + \Lambda_i^{X(l)} + \Lambda_j^{Y(l)} + \Lambda_{ij}^{XY(l)}, i = 1, \ldots, m-1, j = 1, \ldots, n-1.$$

Une fois que les proportions sont déterminées pour chaque cellule, nous pouvons générer un tableau synthétique de n'importe quelle taille.

Les résultats qui suivent sont fondés sur un tableau de contingence à sept dimensions (voir le tableau 7.1 pour les catégories particulières de covariables). Les mesures du BIC indiquent qu'un modèle contenant toutes les interactions bidimensionnelles mais ne contenant aucune interaction tridimensionnelle est celui qui donne l'ajustement le plus parcimonieux.

Tableau 7.1
Variables et catégories de réponse de la NHIS et de la MEPS de 2006 utilisées dans le modèle log-linéaire

Variables d'intérêt	Catégories de réponse
Âge	1 : [18; 24]; 2 : [25; 34]; 3 : [35; 44]; 4 : [45; 54]; 5 : [55; 64]; 6 : >= 65
Région de recensement	1 : Nord-Est; 2 : Mid-Ouest; 3 : Sud; 4 : Ouest
Scolarité	1 : Études secondaires partielles; 2 : Diplôme d'études secondaires; 3 : Études collégiales partielles; 4 : Diplôme d'études collégiales
Sexe	1 : Masculin; 2 : Féminin
Couverture par une assurance maladie	1 : N'importe quel régime privé; 2 : Régime public; 3 : Non assuré
Revenu	1 : (0; 10 000); 2 : [10 000; 15 000); 3 : [15 000; 20 000); 4 : [20 000; 25 000); 5 : [25 000; 35 000); 6 : [35 000; 75 000); 7 : >= 75 000
Race	1 : Hispanique; 2 : Blanche non hispanique; 3 : Noire non hispanique; 4 : Tous les autres groupes non hispaniques confondus

7.2 Résultats

Les résultats sont résumés au tableau 7.2. Pour la population totale et les sous-populations les plus grandes, nous voyons que les estimations ponctuelles (moyenne a posteriori) des taux de couverture par une assurance médicale sont les mêmes sous les approches non paramétrique et log-linéaire, et qu'elles sont presque identiques à celles obtenues au moyen des données réelles après avoir tenu compte des caractéristiques du plan de sondage complexe. Les deux méthodes donnent des populations synthétiques dont les variances (a posteriori) sont légèrement plus élevées que dans le cas des données réelles, ce qui reflète la perte d'information dans la synthèse. Dans le cas de la NHIS, la perte pour l'estimateur non paramétrique est égale, en moyenne, à un peu plus de 20 % et est légèrement supérieure à celle observée pour le modèle log-linéaire, pour lequel la perte est, en moyenne, de l'ordre de 10 %. Dans le cas de la MEPS, les estimateurs affichent tous deux une perte d'environ 10 % par rapport aux données réelles. Cependant, pour les sous-populations plus petites (Blancs non hispaniques gagnant de 25 000 $ à 35 000 $ par année), le modèle log-linéaire produit des résultats biaisés, dus au fait que le modèle log-linéaire ne contient pas toutes les interactions possibles. La méthode non paramétrique produit des estimations presque identiques à celles obtenues au moyen des données réelles après avoir tenu compte des caractéristiques du plan de sondage complexe. Le modèle log-linéaire donne également lieu à une sous-estimation importante, de l'ordre de 30 % à 40 %, de la variance de la couverture par une assurance médicale pour ces sous-populations, par opposition à une surestimation de l'ordre de 10 % à 40 % dans le cas de l'approche non paramétrique.

Tableau 7.2
Estimations d'après les données réelles et d'après les populations synthétiques (modèles non paramétrique et log-linéaire) pour la NHIS et la MEPS de 2006

		Données réelles (plan complexe)		Populations synthétiques			
				Non paramétrique		Modèle log-linéaire	
Domaine	**Type**	**NHIS**	**MEPS**	**NHIS**	**MEPS**	**NHIS**	**MEPS**
Population complète				Proportion			
	Régime privé	0,746	0,735	0,746	0,736	0,746	0,734
	Régime public	0,075	0,133	0,075	0,132	0,076	0,133
	Non assuré	0,179	0,132	0,179	0,132	0,178	0,132
				Variance			
	Régime privé	2,46E-05	2,78E-05	3,15E-05	3,31E-05	2,66E-05	2,86E-05
	Régime public	6,29E-06	1,44E-05	8,06E-06	1,59E-05	7,99E-06	1,77E-05
	Non assuré	1,84E-05	1,41E-05	2,29E-05	1,71E-05	1,81E-05	1,56E-05
Hommes				Proportion			
	Régime privé	0,740	0,735	0,740	0,736	0,740	0,735
	Régime public	0,060	0,101	0,060	0,100	0,060	0,102
	Non assuré	0,200	0,164	0,200	0,164	0,200	0,164
				Variance			
	Régime privé	3,32E-05	3,87E-05	3,93E-05	4,31E-05	3,70E-05	3,52E-05
	Régime public	6,82E-06	1,53E-05	8,81E-06	1,63E-05	7,91E-06	1,91E-05
	Non assuré	2,94E-05	2,64E-05	3,29E-05	2,79E-05	3,19E-05	2,56E-05
Race blanche non hispanique				Proportion			
	Régime privé	0,805	0,788	0,804	0,788	0,804	0,788
	Régime public	0,062	0,116	0,062	0,116	0,062	0,117
	Non assuré	0,134	0,096	0,134	0,096	0,134	0,096
				Variance			
	Régime privé	2,99E-05	3,35E-05	3,79E-05	4,12E-05	3,07E-05	3,98E-05
	Régime public	8,20E-06	1,81E-05	1,04E-05	2,00E-05	1,10E-05	2,45E-05
	Non assuré	2,02E-05	1,51E-05	2,35E-05	1,80E-05	1,82E-05	1,82E-05
Race blanche non hispanique et revenu [25 000 $; 35 000 $)				Proportion			
	Régime privé	0,827	0,813	0,827	0,814	0,840	0,838
	Régime public	0,039	0,079	0,039	0,079	0,037	0,067
	Non assuré	0,134	0,108	0,134	0,107	0,122	0,096
				Variance			
	Régime privé	1,00E-04	1,39E-04	1,48E-04	1,63E-04	6,80E-05	8,59E-05
	Régime public	2,82E-05	6,31E-05	3,86E-05	7,28E-05	1,79E-05	4,25E-05
	Non assuré	7,24E-05	8,92E-05	9,55E-05	1,11E-04	4,38E-05	5,79E-05

8 Discussion

Dans le présent article, nous proposons et évaluons une méthode non paramétrique pour produire des populations synthétiques. Cette méthode permet de tenir compte des caractéristiques du plan de sondage complexe sans utiliser de modèles hypothétiques pour les données observées, de sorte qu'elle est robuste aux erreurs de spécifications du modèle. En outre, contrairement aux méthodes fondées sur un modèle qui nécessitent l'élaboration de modèles d'imputation distincts pour les diverses variables d'intérêt, la méthode non paramétrique n'utilise que les variables de plan de sondage pour générer les populations synthétiques et n'est donc pas particulière à une variable.

Nous avons considéré les propriétés de rééchantillonnage de nos estimateurs synthétiques non paramétriques sous une loi Gamma univariée et sous une loi normale bivariée, en estimant les moyennes, les pentes et les ordonnées à l'origine. Les estimations ponctuelles étaient sans biais, les intervalles avaient une couverture correspondant approximativement au niveau nominal et les pertes d'efficacité comparativement aux données réelles étaient négligeables. Nous avons également considéré des

conditions « réelles » en générant une loi prédictive pour les données de la NHIS et de la MEPS de 2006 et en estimant les taux de couverture par une assurance médicale et les variances associées par la méthode non paramétrique ainsi que par une approche de modélisation log-linéaire entièrement paramétrique. Lorsque les modèles sont bien ajustés aux données, la méthode fondée sur un modèle est plus efficace que la méthode non paramétrique. Cependant, lorsque le modèle hypothétique n'est pas bien ajusté aux données, comme cela est le cas pour certains petits domaines, la méthode fondée sur un modèle peut produire des inférences non valides. Dans ces situations, la méthode non paramétrique est robuste à l'erreur de spécification du modèle.

Outre la robustesse à l'erreur de spécification du modèle, un autre avantage de la méthode non paramétrique tient au fait qu'elle n'utilise que les variables de plan de sondage, comme la strate, la grappe et le poids, pour imputer la partie non observée de la population. Contrairement aux méthodes fondées sur un modèle, elle ne requiert donc pas la modélisation de relations compliquées entre les variables d'intérêt, laquelle devient impossible si des valeurs manquent pour certains items dans les données réelles. La méthode non paramétrique préserve ces valeurs d'item manquantes dans les populations synthétiques produites. Cette propriété pourrait combler une lacune dans le domaine de l'imputation multiple en ce sens que les méthodes existantes consistent habituellement à imputer les valeurs manquantes dans les données comme si ces dernières avaient été obtenues par échantillonnage aléatoire simple, sans tenir compte des caractéristiques du plan de sondage complexe. Un avantage apparenté est que, même si les populations synthétiques sont produites non paramétriquement en se servant des variables de plan, il n'est pas nécessaire qu'elles contiennent elles-mêmes ces variables, puisqu'elles peuvent être analysées comme des échantillons aléatoires simples. Cela permet donc d'éliminer le risque de divulgation associé à la diffusion des variables du plan de sondage (De Waal et Willenborg, 1997; Mitra et Reiter, 2006; Reiter et Mitra, 2009).

Un quatrième avantage pratique de la méthode non paramétrique est qu'elle est plus facile à mettre en œuvre dans les progiciels statistiques existants, parce qu'elle est axée sur les variables du plan de sondage; de ce fait, il n'est pas nécessaire d'élaborer des stratégies particulières pour les divers types de variables et de structures de données.

Comme l'application du bootstrap bayésien en population finie (BBPF) pondéré ne requiert pas que l'on connaisse le nombre de grappes dans la population ni les probabilités conditionnelles de sélection à chaque degré de sélection dans le cas d'un échantillonnage à plusieurs degrés, nous utilisons un bootstrap bayésien approximatif pour tenir compte de la stratification et de la mise en grappes. Selon nous, cette approche est avantageuse à de nombreux égards, puisqu'habituellement, les ensembles de données à grande diffusion ne contiennent pas la ventilation des poids pour chaque degré d'échantillonnage. Toutefois, l'inconvénient est que, afin de s'assurer que les poids de rééchantillonnage soient positifs, le bootstrap bayésien produit moins de grappes dans les strates qu'il n'y en a dans les données réelles. Quand les probabilités de sélection sont connues pour tous les degrés d'échantillonnage, il semble probable que le BBPF pondéré puisse être mis en œuvre à chaque degré, en imputant la population de grappes non observées et la population d'éléments dans chaque grappe en deux étapes, à l'exemple de Meeden (1999), tout comme le BBPF à un degré s'inspire de Ghosh et Meeden (1983). Il s'agit d'un domaine dans lequel la recherche doit se poursuivre.

Remerciements

La présente étude a été financée par la subvention R01CA129101 du NCI. Les auteurs remercient le rédacteur, le rédacteur associé et deux examinateurs anonymes de leurs commentaires. Nous sommes tout particulièrement redevables à l'examinateur qui nous a aidés à mieux comprendre et expliquer les liens entre le bootstrap bayésien en population finie et la loi a posteriori de Pólya exposés à la section 3.

Bibliographie

Agresti, A. (2002). *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Chen, Q., Elliott, M.R. et Little, R.J.A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology*, 36, 1, 25-37.

Cohen, M.P. (1997). The Bayesian bootstrap and multiple imputation for unequal probability sample designs. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 635-638.

de Waal, A.G., et Willenborg, L.C.R.J. (1997). Statistical disclosure control and sampling weights. *Journal of Official Statistics*, 13, 417-434.

Dong, Q. (2012). Combining Information from Multiple Complex Surveys. Unpublished Thesis.

Elliott, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, 33, 1, 27-40.

Elliott, M.R., et Little, R.J.A. (2000). Model-based approaches to weight trimming. *Journal of Official Statistics*, 16, 191-210.

Ericson, W.A. (1969). Subjective Bayesian modeling in sampling finite populations. *Journal of the Royal Statistical Society*, B31, 195-234.

Ghosh, M., et Meeden, G. (1983). Estimation of the variance in finite population sampling. *Sankhyā: The Indian Journal of Statistics*, B45, 362-375.

Hinkins, S., Oh, H.L. et Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 1, 13-24.

Lazzeroni, L.C., et Little, R.J.A. (1998). Random effects models for smoothing poststratification weights. *Journal of Official Statistics*, 14, 61-78.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.

Lo, A.Y. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16, 1684-1695.

Meeden, G. (1999). A noninformative Bayesian approach for two-stage cluster sampling. *Sankhyā: The Indian Journal of Statistics*, B61, 133-144.

Mitra, R., et Reiter J.P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. Privacy in statistical databases: Lecture Notes in Computer Science, 4302, 177-188.

Raghunathan, T.E., Reiter, J.P. et Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T.E., Xie, D.W., Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W. et Feuer, D.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening, *Journal of the American Statistical Association*,102, 474-486

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30, 2, 235-242.

Reiter, J.P. (2005). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society*, A168, 185-205.

Reiter, J.P., et Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, 1, 1, Article 6.

Rubin, D.B (1987). *Multiple Imputation for Non-Response in Surveys*, New York: John Wiley & Sons, Inc.

Scott, A.J. (1977). Large sample posterior distributions in finite populations. *The Annals of Mathematical Statistics*, 42, 1113-1117.

Skinner, C., Holt, D. et Smith, T. (1989). *Analysis of Complex Surveys*, New York: John Wiley & Sons, Inc.

Techniques d'enquête, juin 2014
Vol. 40, N° 1, pp. 53-67
Statistique Canada, N° 12-001-X au catalogue

Application de la méthode des répliques des différences successives pour estimer les variances

Stephen Ash[1]

Résumé

Fay et Train (1995) présentent une méthode qu'ils nomment *successive difference replication*, c.-à-d. répliques des différences successives, qui peut être utilisée pour estimer la variance d'un total estimé au moyen d'un échantillon aléatoire systématique tiré d'une liste ordonnée. L'estimateur prend la forme générale d'un estimateur de variance par rééchantillonnage, où les facteurs de rééchantillonnage sont construits de manière à imiter l'estimateur par différences successives. Cet estimateur est une modification de celui présenté dans Wolter (1985). Le présent article étend la méthodologie en expliquant l'effet de l'attribution des lignes de matrice sur l'estimateur de variance, en montrant comment un jeu réduit de répliques mène à un estimateur raisonnable et en établissant les conditions pour que la méthode des répliques des différences successives soit équivalente à l'estimateur par différences successives.

Mots-clés : Différences successives; répliques des différences successives; échantillonnage aléatoire systématique.

1 Introduction

Fay et Train (1995) présentent une méthode qu'ils nomment *successive difference replication* (SDR), c.-à-d. répliques des différences successives, qui peut être utilisée pour estimer la variance d'un total estimé au moyen d'un échantillon aléatoire systématique tiré d'une liste ordonnée. L'estimateur prend la forme générale de l'estimateur de variance par rééchantillonnage, où les facteurs de rééchantillonnage sont construits de manière à imiter l'estimateur par différences successives (SD).

L'article décrit l'établissement et l'utilisation de nouveaux concepts en vue de mieux comprendre la méthodologie proposée au départ par Fay et Train (1995), ci-après appelés F et T. Ces nouveaux concepts aident à expliquer l'effet de l'attribution des lignes de matrice sur l'estimateur de variance, à montrer comment un jeu réduit de répliques mène à un estimateur raisonnable, et à établir les conditions pour que la méthode des répliques des différences successives soit équivalente à l'estimateur par différences successives. Nous espérons qu'en étant mieux comprise, la SDR semblera moins mystérieuse et sera donc plus facile à utiliser par toute personne qui souhaite estimer les variances dans le cas d'un échantillonnage aléatoire systématique.

L'article débute par un examen de l'estimateur SD et de la façon dont il convient à l'estimation de la variance des échantillons aléatoires systématiques. La partie principale est consacrée à l'introduction de deux théorèmes qui fournissent les conditions pour que l'estimateur SDR soit équivalent à l'estimateur SD. L'article se termine par la présentation d'exemples empiriques en vue d'examiner l'effet de différents schémas d'affectation des lignes de matrice et de montrer qu'il est approprié d'utiliser un ensemble réduit de répliques.

1. Stephen Ash, U.S. Census Bureau, 4600 Silver Hill Road, Washington DC 20233. Courriel : stephen.eliot.ash@census.gov.

Dans la suite de l'exposé, l'abréviation *sys* sera utilisée pour désigner l'échantillonnage aléatoire systématique à partir d'une liste ordonnée. Nous utilisons l'abréviation *sys* parce que l'on peut montrer que l'échantillonnage systématique à partir d'une liste non ordonnée ou d'une liste ordonnée aléatoirement est équivalent à l'échantillonnage aléatoire simple (Madow et Madow 1944). Pour les besoins de notre discussion, nous nous concentrons uniquement sur la sélection équiprobabiliste et sur les méthodes de sélection d'un échantillon dans une seule dimension. Le lecteur trouvera d'excellents résumés de l'échantillonnage *sys* et de l'estimation des variances sous *sys* dans Iachan (1982), Wolter (1985, chapitre 7), Murthy et Rao (1988), et Bellhouse (1988).

1.1 Revue de la méthode des différences successives

Wolter (1984; estimateur 2) donne un estimateur par différences successives de la variance d'une moyenne estimée $(\bar{y})$ sous un plan de sondage *sys* de la forme

$$\hat{v}_{\mathrm{SD1}}(\hat{\bar{y}}) = (1-f)\frac{1}{2n(n-1)}\sum_{k=2}^{n}(y_k - y_{k-1})^2,$$

où y_k est la variable d'intérêt, k indice les unités de l'échantillon ordonné, et $f = n/N$ est la fraction d'échantillonnage. La statistique d'intérêt est Y ou le total de y_k sur l'univers d'intérêt, et $\hat{Y}$ est un estimateur de Y. Soit N et n la taille de l'univers et de l'échantillon, respectivement. La moyenne de y_k et son estimateur sont définis comme étant $\bar{y} = Y/N$ et $\hat{\bar{y}}$, respectivement. Nous définissons aussi l'estimateur du total Y comme étant $\hat{Y} = \sum_{k=1}^{n}\breve{y}_k$, où la variable d'intérêt pondérée par des poids égaux est $\breve{y}_k = (N/n)\,y_k$; pour des poids de sondage inégaux w_k, elle est définie comme étant $\breve{y}_k = w_k y_k$. L'estimateur $\hat{v}_{\mathrm{SD1}}(\hat{\bar{y}})$ a été décrit par Yates (1953; pages 229 à 231) et recommandé par Wolter (1984). Murthy et Rao (1988, équation 32) donnent un aperçu des raisons pour lesquelles l'estimateur fonctionne. La version abrégée est que, puisque sous échantillonnage *sys* une seule unité est sélectionnée dans chaque strate implicite, la solution de l'estimateur SD consiste à fusionner les strates implicites adjacentes. Avec deux unités, nous pouvons estimer la variance d'une strate implicite. Après fusion des strates implicites, la moyenne est calculée sur toutes les paires possibles, puis multipliée par n, le nombre de strates implicites, pour donner la variance de toutes les strates implicites.

F et T donnent un estimateur de variance SD d'un total sous échantillonnage *sys* de la forme

$$\hat{v}_{\mathrm{SD1}}(\hat{Y}) = (1-f)\frac{n}{2(n-1)}\sum_{k=2}^{n}(\breve{y}_k - \breve{y}_{k-1})^2.$$

Wolter (1985, équation 7.7.4) définit le même estimateur où $w_k = (np_k)^{-1}$ et p_k est la probabilité de sélection avec remise de l'unité k. F et T définissent un deuxième estimateur SD

$$\hat{v}_{\mathrm{SD2}}(\hat{Y}) = \frac{1}{2}(1-f)\left[\sum_{k=2}^{n}(\breve{y}_k - \breve{y}_{k-1})^2 + (\breve{y}_n - \breve{y}_1)^2\right],$$

qui est « circulaire » en ce sens qu'il inclut une différence au carré supplémentaire qui relie les première et dernière unités de la liste triée.

Nous exprimons l'estimateur SD2 de manière plus générale sous une forme quadratique $\breve{\mathbf{y}}'\mathbf{C}\breve{\mathbf{y}}$, où $\breve{\mathbf{y}}' = [\breve{y}_1 \breve{y}_2 \ldots \breve{y}_n]$ est défini comme le vecteur d'observations pondérées de dimension $n \times 1$ et $\mathbf{C}$ est une matrice carrée dont tous les éléments de la diagonale principale valent 2, tous les éléments de la diagonale supérieure et de la diagonale inférieure valent -1, et l'élément inférieur gauche et l'élément supérieur droit valent -1. Ici, les diagonales supérieures sont définies comme étant les diagonales adjacentes à la diagonale principale, excepté dans le cas d'une matrice de dimensions 2×2.

2 Répliques des différences successives

2.1 Définition de la méthode des répliques des différences successives

F et T présentent une méthode qu'ils nomment *successive difference replication* (SDR), c.-à-d. répliques des différences successives, qui permet d'estimer la variance sous échantillonnage *sys* en imitant $\hat{v}_{\mathrm{SD2}}(\hat{Y})$, ce qui signifie que l'estimateur SDR est équivalent ou quasi équivalent à $\hat{v}_{\mathrm{SD2}}(\hat{Y})$. Nous montrons comment la méthode SDR peut être appliquée pour produire les facteurs et les poids de rééchantillonnage pour un estimateur de variance par rééchantillonnage général qui est équivalent à l'estimateur SD2. Avant de définir l'estimateur SDR dans le premier théorème, nous établissons certains termes et fournissons un lemme qui est utilisé dans le théorème.

Un schéma d'attribution de lignes, ou plus simplement schéma AL, correspond à l'attribution de deux lignes d'une matrice à chaque unité de l'échantillon. Nous désignons habituellement la paire de lignes par (a_i, b_i) pour l'unité i. Une boucle connectée est un schéma AL qui ne répète aucune des lignes, c.-à-d. $a_i \neq a_j$ et $b_i \neq b_j$ pour tous i et j dans la boucle connectée, et qui est circulaire, c.-à-d. $b_i = a_{i+1}$ pour tout $i < n$ et $b_n = a_1$. Un exemple de boucle connectée pour trois observations est (1,2), (2,3), (3,1).

Une matrice de décalage $\mathbf{S}$ peut être utilisée pour déplacer les lignes ou les colonnes d'une matrice. Nous expliquons le processus de déplacement des lignes, qui est similaire au processus de déplacement des colonnes. Une matrice de décalage est une matrice carrée dont tous les éléments valent 0, à l'exception d'une valeur 1 unique dans chaque colonne. Si nous voulons déplacer la ligne p jusqu'à la ligne q, nous plaçons une valeur 1 dans la q^{e} ligne de la p^{e} colonne et des 0 ailleurs. Nous insistons sur le fait que l'ordre est important lorsqu'on applique une matrice de décalage à une autre matrice. L'application de $\mathbf{S}$ à une autre matrice carrée $\mathbf{A}$ sous la forme $\mathbf{AS}$ déplace les colonnes de $\mathbf{A}$, mais sous la forme $\mathbf{SA}$, elle déplace les lignes de $\mathbf{A}$.

Lemme : Soit $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_c$ les matrices de décalage, alors $\mathrm{bloc}(\mathbf{S}_1'\mathbf{S}_1, \mathbf{S}_2'\mathbf{S}_2, \ldots, \mathbf{S}_C'\mathbf{S}_C) = \mathbf{I}$.

Preuve. Nous commençons par définir une matrice diagonale par blocs générale $\mathbf{A}$ qui est formée par les matrices carrées $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_C$ comme

$$\mathbf{A} = \mathrm{bloc}(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_C) = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{A}_C \end{bmatrix}.$$

On peut montrer que, si $\mathbf{A}$ et $\mathbf{B}$ sont toutes deux des matrices diagonales par blocs et que les matrices carrées $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_C$ ont les mêmes dimensions que $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_C$, respectivement, alors $\mathbf{AB} = \text{bloc}(\mathbf{A}_1\mathbf{B}_1, \mathbf{A}_2\mathbf{B}_2, \ldots, \mathbf{A}_C\mathbf{B}_C)$. Pour une matrice de décalage donnée, nous savons aussi que $\mathbf{S}'\mathbf{S} = \mathbf{I}$, puisque le décalage d'une ligne vers le bas d'une matrice de décalage est $\mathbf{I}$. Le lemme découle des deux éléments qui précèdent.

Nous définissons aussi une matrice de décalage d'une ligne comme étant une matrice de décalage qui décale toutes les lignes d'une autre matrice d'une ligne vers le bas et transfère la dernière ligne à la première ligne, ou qui décale toutes les lignes d'une autre matrice d'une ligne vers le haut et transfère la première ligne à la dernière ligne. Si $\mathbf{S}_D$ est une matrice de décalage d'une ligne qui déplace les lignes vers le bas, tous les éléments de la diagonale supérieure et l'élément inférieur gauche de la matrice ont une valeur de 1, par exemple $\mathbf{S}_1$. De même, si $\mathbf{S}_U$ est une matrice de décalage d'une ligne qui déplace les lignes vers le haut, tous les éléments de la diagonal inférieure et l'élément supérieur droit de la matrice ont une valeur de 1, par exemple la matrice $\mathbf{S}_2$ subséquemment définie. Notons la propriété que $\mathbf{S}_D = \mathbf{S}'_U$ et $\mathbf{S}_U = \mathbf{S}'_D$; donc, $\mathbf{S}_U + \mathbf{S}'_U = \mathbf{S}_D + \mathbf{S}'_D$. Nous présentons maintenant le théorème principal de l'article qui établit les conditions sous lesquelles l'estimateur SDR est équivalent à l'estimateur SD2.

Théorème 1 : Soit n la taille d'un échantillon *sys* donné et $\breve{\mathbf{y}}' = [\breve{y}_1 \breve{y}_2 \ldots \breve{y}_n]$, le vecteur d'observations pondérées de dimension $n \times 1$, où l'ordre des observations reflète l'ordre de tirage de l'échantillon *sys*.

(a) Choisir une matrice de Hadamard d'ordre k $(\mathbf{HH}' = k\,\mathbf{I})$, où $n \leq k$.

(b) Choisir un schéma d'attribution de lignes (AL) qui assigne deux lignes (a_i, b_i) à chaque unité i de l'échantillon. Poser que le schéma AL définit C boucles connectées c contenant chacune m_c unités.

(c) Choisir les $m = n$ lignes de $\mathbf{H}$ correspondant au schéma AL pour créer la matrice $\mathbf{M}$ de dimensions $m \times k$. L'ordre des lignes de $\mathbf{M}$ doit correspondre à la première ligne du schéma AL. Par exemple, la première ligne de $\mathbf{M}$ doit être la ligne $a_{i=1}$ de $\mathbf{H}$, la deuxième ligne doit être la ligne $a_{i=2}$ de $\mathbf{H}$, *etc*. Ensuite, définir la matrice de décalage de dimensions $m \times m$ comme étant $\mathbf{S} = \text{bloc}(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_C)$, où les matrices de décalage d'une ligne $\mathbf{S}_c$ de dimensions $m_c \times m_c$ sont définies en vue d'identifier la position de la deuxième ligne b_i du schéma AL dans $\mathbf{M}$. En général, chaque matrice de décalage $\mathbf{S}_c$ sera une matrice de décalage vers le haut, une matrice de décalage vers le bas ou une matrice de décalage de dimensions 2×2 (voir la matrice $\mathbf{S}_4$ subséquemment définie).

Définir l'estimateur du total r pour chaque réplique comme $\hat{Y}_r = \sum_{i=1}^{n} f_{i,r}\breve{y}_i$, où la matrice des facteurs de rééchantillonnage est $\mathbf{F} = \mathbf{1}_m\mathbf{1}'_k + \left(2^{-3/2}\mathbf{I}_m - 2^{-3/2}\mathbf{S}\right)\mathbf{M}$ et les valeurs individuelles dans la matrice sont définies pour chaque unité i (lignes de $\mathbf{F}$) de la réplique r (colonnes de $\mathbf{F}$) comme étant $f_{i,r} = 1 + 2^{-3/2}h_{a_i,r} - 2^{-3/2}h_{b_i,r}$. $\mathbf{I}_m$ est une matrice identité de dimensions $m \times m$ et $\mathbf{1}_m$ est un vecteur de dimension $m \times 1$ de 1. Alors, l'estimateur de variance SDR $\hat{v}_{\text{SDR}}\left(\hat{Y}\right) = (1 - f)\,4/k \sum_{r=1}^{m}\left(\hat{Y}_r - \hat{Y}\right)^2$ est équivalent à la somme des C différents estimateurs SD2.

Preuve. L'estimateur SDR peut s'écrire en notation matricielle sous la forme

$$(1-f)\frac{4}{k}\left(\breve{\mathbf{y}}'\left(\mathbf{1}_m\mathbf{1}_k' + \left(2^{-3/2}\mathbf{I}_m - 2^{-3/2}\mathbf{S}\right)\mathbf{M}\right) - \breve{\mathbf{y}}'\mathbf{1}_m\mathbf{1}_k'\right)\left(\breve{\mathbf{y}}'\left(\mathbf{1}_m\mathbf{1}_k' + \left(2^{-3/2}\mathbf{I}_m - 2^{-3/2}\mathbf{S}\right)\mathbf{M}\right) - \breve{\mathbf{y}}'\mathbf{1}_m\mathbf{1}_k'\right)'$$

$$= (1-f)\frac{4}{k}\left(2^{-3/2}\right)^2 \breve{\mathbf{y}}'(\mathbf{I}_m - \mathbf{S})\mathbf{M}\mathbf{M}'(\mathbf{I}_m - \mathbf{S})'\breve{\mathbf{y}}$$

Comme $\{\text{lignes de } \mathbf{M}\} \subseteq \{\text{lignes de } \mathbf{H}\}$, on peut montrer que $\mathbf{M}\mathbf{M}' = k\mathbf{I}$. Partant de ce résultat, la variance devient

$$(1-f)\frac{1}{2k}\breve{\mathbf{y}}'(\mathbf{I}_m - \mathbf{S})(k\mathbf{I}_m)(\mathbf{I}_m - \mathbf{S})'\breve{\mathbf{y}} = \frac{1}{2}(1-f)\breve{\mathbf{y}}'(\mathbf{I}_m - \mathbf{S})(\mathbf{I}_m - \mathbf{S})'\breve{\mathbf{y}}$$

$$= \frac{1}{2}(1-f)\breve{\mathbf{y}}'(2\mathbf{I}_m - \mathbf{S} - \mathbf{S}')\breve{\mathbf{y}}$$

La dernière ligne découle du lemme et a une valeur constante pour tout choix de $\mathbf{H}$. En notant la structure diagonale par blocs de $\mathbf{S}$, nous pouvons écrire l'estimateur sous la forme

$$\frac{1}{2}(1-f)\sum_{c=1}^{C}\breve{\mathbf{y}}_c'(2\mathbf{I}_m - \mathbf{S}_c - \mathbf{S}_c')\breve{\mathbf{y}}_c,$$

où $\breve{\mathbf{y}}_c$ correspond au vecteur des observations pondérées dans la boucle connectée c, qui est un résultat de la partition du vecteur d'observations pondérées pour donner $\breve{\mathbf{y}}' = [\breve{\mathbf{y}}_{c=1}\breve{\mathbf{y}}_{c=2}\ldots\breve{\mathbf{y}}_{c=C}]$. Le choix du schéma AL ne modifie pas le résultat, puisque nous savons que $2\mathbf{I}_m - \mathbf{S}_c - \mathbf{S}_c'$ est constant pour une matrice de décalage d'une ligne vers le haut ou vers le bas $\mathbf{S}_c$.

Note 1 : Le théorème 1 définit l'estimateur SDR en fonctions des facteurs de rééchantillonnage, mais nous pouvons aussi l'exprimer en fonction des poids de rééchantillonnage sous la forme

$$(1-f)\frac{4}{k}\mathbf{y}'(\mathbf{W} - \mathbf{1}_m\mathbf{1}_k')(\mathbf{W} - \mathbf{1}_m\mathbf{1}_k')'\mathbf{y}.$$

Ici, $\mathbf{W}$ est la matrice de dimensions $m \times k$ des poids de rééchantillonnage définie comme étant $\mathbf{W} = \mathbf{w} * \mathbf{F}$, où $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ est le vecteur de poids de sondage pour les n unités de l'échantillon et l'opérateur $*$ multiplie les éléments du vecteur $\mathbf{w}$ par chacune des colonnes de $\mathbf{F}$, c.-à-d. que, si $W_{i,r}$ et w_i sont des entrées de $\mathbf{W}$ et $\mathbf{w}$, respectivement, les entrées de $\mathbf{W}$ sont définies comme étant $W_{i,r} = w_i \times f_{i,r}$.

Note 2 : Huang et Bell (2009) définissent similairement l'estimateur SDR sous une forme quadratique et l'utilisent pour établir certaines propriétés générales de l'estimateur quand y_k est i.i.d.(μ, σ^2). Nous souhaitons interpréter la façon dont l'estimateur SDR fonctionne et la qualité de son fonctionnement. Définir la forme quadratique avec des matrices de décalage et des boucles connectées permet de mieux comprendre les attributions de lignes et l'efficacité de l'estimateur.

Pour un échantillon de grande taille, il n'est habituellement pas pratique d'utiliser une matrice $\mathbf{H}$ où $n < k$. Le deuxième théorème offre un moyen d'utiliser $\mathbf{H}$ en prenant $k < n$ pour produire une plus grande matrice de Hadamard $\tilde{\mathbf{H}}$ où $k \geq n$ qui résultera en un estimateur SDR équivalent à l'estimateur SD2. Le deuxième théorème étoffe et clarifie aussi les instructions données par F et T pour le cas où $n > k$. Dans leurs instructions, F et T utilisent le mot cycle pour désigner chaque tranche de $m_d \leq k$ unités de l'échantillon. Le théorème 2 n'impose pas de contraintes sur le schéma AL, mais suit à part cela les conditions établies par F et T.

Théorème 2 : Soit n la taille d'un échantillon sys donné.

(a) Choisir une matrice de Hadamard $\mathbf{H}_A$ d'ordre k_A, où $n > k_A$.

(b) Choisir un schéma AL qui assigne les lignes de $\mathbf{H}_A$ à l'échantillon. En gardant l'ordre original, répartir les n unités de l'échantillon en D cycles. Chaque cycle d comprend $m_d \leq k_A$ unités. Dans chaque cycle, le schéma AL définit une ou plusieurs boucles connectées.

(c) Choisir une matrice de Hadamard semi-normale $\mathbf{H}_B$ d'ordre k_B et l'utiliser pour définir une plus grande matrice de Hadamard $\tilde{\mathbf{H}}$ d'ordre $\tilde{k}$ générée à partir de la matrice $\mathbf{H}_A$ originale. Cela peut se faire en appliquant une construction de Welsch à $\mathbf{H}_A$, c.-à-d. $\tilde{\mathbf{H}} = \mathbf{H}_B \otimes \mathbf{H}_A$.

(d) Choisir les $m = \sum_{d=1}^{D} m_d$ lignes de $\tilde{\mathbf{H}}$ qui correspondent au schéma AL pour créer la matrice $\tilde{\mathbf{M}}$ de dimensions $m \times \tilde{k}$. L'ordre des lignes de $\tilde{\mathbf{M}}$ doit correspondre à la première ligne du schéma AL. Ensuite, définir la matrice de décalage de dimensions $m \times m$ comme étant $\mathbf{S} = \text{bloc}(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_D)$ où les matrices $\mathbf{S}_d$ de dimensions $m_d \times m_d$ identifient la position de la deuxième ligne b_i du schéma AL dans $\tilde{\mathbf{M}}$.

Dans ces conditions, l'estimateur SDR est défini comme

$$\hat{v}_{\text{SDR}}\left(\hat{Y}\right) = (1 - f)\frac{4}{\tilde{k}}\sum_{r=1}^{\tilde{k}}\left(\hat{Y}_r - \hat{Y}\right)^2$$

et est équivalent à la somme d'au moins D estimateurs SD2.

Preuve. Le résultat découle de l'application du théorème 1. La valeur particulière de D découle du fait que chacun des D cycles peut posséder une ou plusieurs boucles connectées, de manière à avoir un total d'au moins D boucles connectées.

Exemple 1 : Soit $n = 14$ et choisissons la matrice de Hadamard non normale $\mathbf{H}_A = \mathbf{H}_{4b}$ d'ordre $k_A = 4$. Le nombre de cycles est $D = 4$ et le schéma AL dans chaque cycle est donné dans la deuxième colonne du tableau 2.1 pour chaque unité. Définissons $\tilde{\mathbf{H}}$ d'ordre $\tilde{k} = 16$ en utilisant une construction de Welsh de la matrice de Hadamard normale originale comme il suit

$$\mathbf{H}_{16} = \mathbf{H}_{4a} \otimes \mathbf{H}_{4b} = \begin{bmatrix} \mathbf{H}_{4b} & \mathbf{H}_{4b} & \mathbf{H}_{4b} & \mathbf{H}_{4b} \\ \mathbf{H}_{4b} & -\mathbf{H}_{4b} & \mathbf{H}_{4b} & -\mathbf{H}_{4b} \\ \mathbf{H}_{4b} & \mathbf{H}_{4b} & -\mathbf{H}_{4b} & -\mathbf{H}_{4b} \\ \mathbf{H}_{4b} & -\mathbf{H}_{4b} & -\mathbf{H}_{4b} & \mathbf{H}_{4b} \end{bmatrix}$$

où

$$\mathbf{H}_{4a} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \text{ et } \mathbf{H}_{4b} = \begin{bmatrix} 1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 \end{bmatrix}.$$

En utilisant $\mathbf{H}_{16}$, nous pouvons calculer les facteurs de rééchantillonnage pour 16 répliques comme au tableau 2.1. En notation matricielle, $\tilde{\mathbf{M}}$ englobe toutes les lignes de $\tilde{\mathbf{H}} = \mathbf{H}_{16}$ sauf les lignes 13 et 16. Les lignes de $\tilde{\mathbf{M}}$ sont ordonnées par a_i, la première ligne assignée dans le schéma AL. La matrice de décalage est définie comme $\mathbf{S} = \text{bloc}(\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \mathbf{S}_4)$, où les matrices de décalage correspondant à chaque cycle sont

$$\mathbf{S}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \mathbf{S}_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{S}_3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{S}_4 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Tableau 2.1
Matrice des facteurs de rééchantillonnage $(f_{i,r})$ pour l'exemple 1

Unité #	AL $\mathbf{H}_A = \mathbf{H}_{4b}$	AL $\tilde{\mathbf{H}} = \mathbf{H}_{16}$	Cycle	Réplique 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	(1,2)	(1,2)		1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0
2	(2,3)	(2,3)	1	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7
3	(3,4)	(3,4)		1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3
4	(4,1)	(4,1)		1,0	1,7	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7	0,3	1,0
5	(1,3)	(5,7)		1,0	1,0	1,7	1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7	1,7	1,0	1,0	0,3	0,3
6	(3,1)	(7,5)	2	1,0	1,0	0,3	0,3	1,0	1,0	1,7	1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7	1,7
7	(2,4)	(6,8)		0,3	0,3	1,0	1,0	1,7	1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7	1,7	1,0	1,0
8	(4,2)	(8,6)		1,7	1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7	1,7	1,0	1,0	0,3	0,3	1,0	1,0
9	(1,4)	(9,12)		1,0	0,3	1,7	1,0	1,0	0,3	1,7	1,0	1,0	1,7	0,3	1,0	1,0	1,7	0,3	1,0
10	(4,3)	(12,11)	3	1,0	1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0	0,3	1,0	0,3	1,0	0,3	1,0	0,3
11	(3,2)	(11,10)		1,7	1,0	1,0	0,3	1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7
12	(2,1)	(10,9)		0,3	1,0	0,3	1,0	0,3	1,0	0,3	1,0	1,7	1,0	1,7	1,0	1,7	1,0	1,7	1,0
13	(2,3)	(14,15)	4	0,3	1,0	1,0	1,7	1,7	1,0	1,0	0,3	1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7
14	(3,2)	(15,14)		1,7	1,0	1,0	0,3	0,3	1,0	1,0	1,7	0,3	1,0	1,0	1,7	1,7	1,0	1,0	0,3

Étant donné les facteurs de rééchantillonnage du tableau 2.1, l'estimateur SDR est équivalent à la somme de cinq estimateurs SD2 différents, un pour chaque boucle connectée du schéma AL, c.-à-d.

$$(1-f)\frac{4}{\tilde{k}}\sum_{r=1}^{\tilde{k}}\left(\hat{Y}_r-\hat{Y}\right)^2=\frac{1}{2}(1-f)\begin{bmatrix}\sum_{i=2}^{4}(y_i-y_{i-1})^2+(y_4-y_1)^2+2(y_6-y_5)^2\\+2(y_8-y_7)^2+\sum_{i=10}^{12}(y_i-y_{i-1})^2+(y_{12}-y_9)^2\\+2(y_{13}-y_{13})^2\end{bmatrix}. \qquad (2.1)$$

Il convient de souligner quelques éléments concernant l'exemple 1. Premièrement, le nombre de répliques nécessaires est supérieur à la taille de l'échantillon. Cela se produit lorsque m_d n'est pas constant dans tous les cycles. Le quatrième cycle ne comprend que deux unités d'échantillon, mais nous avons dû utiliser quatre répliques de chaque $\mathbf{H}_{4b}$ parce qu'au moins un des cycles utilisait quatre lignes.

Pour rendre l'exemple plus intéressant, nous avons choisi une matrice de Hadamard non normale $\mathbf{H}_{4b}$ pour $\mathbf{H}_A$. Cette matrice de Hadamard non normale a été construite en partant de la matrice de Hadamard normale $\mathbf{H}_{4a}$ et en inversant la procédure décrite par Hedayat et Wallis (1978) pour trouver une matrice de Hadamard normale. Ici nous avons simplement changé le signe de tous les éléments de la deuxième ligne, puis nous avons changé le signe de tous les éléments de la deuxième colonne.

Si nous avions utilisé la matrice de Hadamard normale $\mathbf{H}_{4a}$ pour $\mathbf{H}_A$ ainsi que $\mathbf{H}_B$, les facteurs de rééchantillonnage pour les répliques 1, 5, 9 et 13 auraient tous été égaux à 1,0. Nous disons qu'une réplique est « morte » quand chaque élément reçoit une valeur de 1,0 et que l'estimation basée sur la réplique est donc égale à l'estimation originale. Dans l'estimateur SDR, les répliques mortes sont tout à fait valables et dues simplement à la façon dont les facteurs de rééchantillonnage sont répartis par la matrice de Hadamard. En cas de réplique morte, de nombreuses valeurs 1,0 se trouvent dans celle-ci, et la composition des autres répliques est plus mélangée, avec des valeurs de 1,7 et de 0,3. Cependant, toutes les répliques, même les répliques mortes, sont nécessaires pour l'estimation.

La valeur réelle du théorème 2 tient au fait qu'il permet de comprendre la prescription originale de F et T pour l'estimateur SDR quand $n > k$. Dans F et T, le schéma AL est appliqué de manière répétée aux $m = k - 1$ lignes de $\mathbf{H}_A$ (en sautant la première ligne de $\mathbf{H}_A$), où $\mathbf{H}_A$ est choisie comme une matrice de Hadamard normale. Les répliques sont ensuite formées en utilisant les k_A colonnes de $\mathbf{H}_A$. Si nous appliquons le cadre plus vaste du théorème 2, nous dirions qu'ils ont utilisé implicitement une matrice normale $\mathbf{H}_B$, qui donne $\tilde{\mathbf{H}} = \mathbf{H}_B \otimes \mathbf{H}_A$ et n'inclut que les k_A premières répliques dans l'estimateur de variance. Puisqu'un sous-ensemble des répliques nécessaires pour que l'estimateur SDR soit équivalent à l'estimateur SD2 est utilisé, nous disons que l'estimateur résultant est une approximation de l'estimateur SD2.

Exemple 1 (suite) : Si nous utilisons seulement les quatre premières répliques du tableau 2.1, l'estimateur SDR sera équivalent à (2.1) plus le terme de reste R qui est défini comme

$$R = \begin{bmatrix} (y_1 - y_2)(y_8 - y_7) + (y_1 - y_2)(y_{11} - y_{12}) + (y_1 - y_2)(-y_{13} + y_{14}) \\ +(y_8 - y_7)(y_{11} - y_{12}) + (y_8 - y_7)(y_{14} - y_{13}) + (y_{11} - y_{12})(y_{14} - y_{13}) \\ +(y_4 - y_3)(y_8 - y_7) + (y_4 - y_3)(y_{10} - y_9) + (y_8 - y_7)(y_{10} - y_9) \\ +(y_1 - y_4)(y_5 - y_6) + (y_1 - y_4)(y_9 - y_{12}) + (y_5 - y_6)(y_9 - y_{12}) \\ +(y_2 - y_3)(y_5 - y_6) + (y_2 - y_3)(y_{10} - y_{11}) + (y_2 - y_3)(y_{13} - y_{14}) \\ +(y_5 - y_6)(y_{10} - y_{11}) + (y_5 - y_6)(y_{13} - y_{14}) + (y_{10} - y_{11})(y_{13} - y_{14}) \end{bmatrix}$$

Notons que R comprend le même nombre de termes positifs et négatifs, qui ne s'annulent pas exactement, mais qui font que la valeur de R est habituellement proche de zéro. De même, utiliser les répliques 1 à $q \times k_A$, où $q = 1, 2, \ldots, k_B$, donne un reste R comprenant un nombre égal de termes positifs et de termes négatifs. Ce n'est qu'en utilisant toutes les répliques de $\tilde{\mathbf{H}}$ que le terme de reste R est nul.

Exemple 2 : La taille de l'échantillon mensuel de la Current Population Survey (CPS) est de $n = 72\,000$ ménages par mois (U.S. Census Bureau 2006). La CPS est réalisée selon un plan de sondage à deux degrés comprenant la sélection d'un échantillon de premier degré formé d'unités primaires d'échantillonnage (UPE), qui sont habituellement des comtés ou des groupes de comtés, puis le tirage de l'échantillon de deuxième degré de ménages à partir de l'échantillon d'UPE. Certaines UPE, généralement les régions métropolitaines, sont sélectionnées avec certitude, c.-à-d. que leur probabilité de sélection au premier degré est 1,0. Dans le cas des UPE sélectionnées avec certitude, l'échantillon sys peut être traité comme le plan de sondage de premier degré dans l'estimation de la variance, c.-à-d. que la méthode SDR est appliquée pour produire les répliques. Dans le cas des UPE sélectionnées sans certitude, la méthode des répliques équilibrées répétées (BRR pour *Balanced Repeated Replication*) [McCarthy 1966] est appliquée pour produire les répliques. Environ 75 % de l'échantillon ou 54 000 unités sont comprises dans les UPE autoreprésentatives, auxquelles est appliquée la méthode SDR.

L'application de la méthode SDR à la CPS comprend l'utilisation d'une matrice de Hadamard d'ordre $k = 160$ dont sont exclues deux lignes, c.-à-d. que $m = 158$. Les poids de rééchantillonnage sont produits pour 160 répliques. Même s'il peut sembler qu'il s'agit d'une conclusion logique du présent article, nous ne suggérons pas que l'on utilise pour la CPS une matrice de Hadamard d'ordre $k = 54\,000$ ni que l'on produise 54 000 jeux de poids de rééchantillonnage. Cela donnerait en effet un nombre irraisonnable de répliques. Nous sommes plutôt d'avis que le sous-ensemble de 160 répliques utilisé pour la CPS est grand et fournit par conséquent une approximation raisonnable de l'estimateur SD2. Plus loin, dans les exemples empiriques, nous examinons l'effet de l'utilisation d'un jeu réduit de répliques.

2.2 Attribution de lignes quand $n > k$

Jusqu'ici, nous avons supposé qu'un schéma AL était donné et nous n'avons pas discuté de la façon de générer ce schéma pour un échantillon particulier, où $n > k$. À la présente section, nous examinons deux schémas AL et formulons certains commentaires au sujet de l'attribution de lignes en général. Le premier schéma AL est similaire à celui décrit par Sukasih et Jang (2003) et est destiné à être utilisé quand $k < n$ et avec le théorème 2.

AL1 : Ce schéma AL attribue une paire de lignes a_i et b_i à chaque tranche de m_d unités de l'échantillon, que nous appelons cycle d, où $m_d \leq k$. Après $m_d - 1$ cycles, le schéma AL est répété jusqu'à ce qu'une paire de lignes ait été attribuée à chacune des unités de l'échantillon.

Étape 1 : Trier l'échantillon dans l'ordre dans lequel il était trié avant la sélection de l'échantillon.

Étape 2 : Initialiser le numéro du cycle par $d = 1$ et le nombre de boucles connectées par $c = 1$.

Étape 3 : Commencer l'AL au début d'un cycle ou d'une boucle connectée en prenant $a_1 = c$.

Étape 4 : Répéter le schéma AL suivant : $b_i = \mathrm{mod}(a_i + d, k)$ et $a_i = b_i$ jusqu'à ce que chacune des m_d lignes du cycle ait été utilisée ou que l'AL devienne une boucle connectée. Ici, la fonction modulo ou $\mathrm{mod}(a, b)$ est définie comme étant le reste de la division de a par b. Si les m_d lignes du cycle ont toutes été utilisées, commencer un nouveau cycle : poser que $d = d + 1$ et retourner à l'étape 3. Sinon (fin d'une boucle connectée, mais non la fin d'un cycle), commencer une nouvelle boucle connectée : poser que $c = c + 1$ et retourner à l'étape 3.

Étape 5 : À la fin de $d = m_d - 1$ cycles, recommencer au premier cycle – retourner à l'étape 2.

Le schéma AL1 possède les caractéristiques suivantes :

- Chacun des cycles $d = 1, 2, \ldots, m_d - 1$ de l'AL attribue m_d paires de lignes. Cela crée un total de $m_d(m_d - 1)$ paires de lignes.
- Le schéma d'AL se répète après $m_d - 1$ cycles. F et T suggèrent de redémarrer l'AL après 10 cycles. Nous recommandons d'utiliser chacun des $m_d - 1$ cycles avant de redémarrer l'AL.
- Les valeurs de a_i et b_i sont toujours espacées de c unités.
- Au milieu de la séquence, le schéma se répète en ordre inverse. Si m est un nombre pair, les cycles avant et après le $(m_d + 1)/2^{\mathrm{e}}$ cycle se répètent en ordre inverse.

Le schéma AL1 diffère de du schéma AL de Sukasih et Jang (2003), en ce sens que nous ne suggérons pas de sauter la ligne 1 ni de répéter le schéma AL après 10 cycles et nous n'exigeons pas que $k - 1$ soit un nombre premier. Premièrement, une ligne dont tous les éléments valent 1 peut paraître étrange, mais cela ne pose pas de problème. Comme dans le cas d'une colonne dont tous les éléments valent 1 dans $\mathbf{M}$, ce qui donne une réplique morte, une ligne ne contenant que des 1 n'aura d'effet que sur la distribution des facteurs de rééchantillonnage. Une unité i à laquelle a été attribuée la ligne 1 (soit $a_i = 1$ ou $b_i = 1$) possédera un plus grand nombre de facteurs de rééchantillonnage valant 1,0 qu'autrement. Cela n'est pas incorrect; il s'agit simplement de la façon dont les facteurs de rééchantillonnage sont distribués par $\mathbf{H}_A$. La deuxième différence est que nous suggérons de répéter l'attribution après m cycles, c'est-à-dire au moment où le schéma se répète, plutôt qu'après un nombre fixé de 10 cycles. Enfin, nous n'exigeons pas que $k - 1$ soit un nombre premier, mais notons que si $m_d = k - 1$ et que $k - 1$ est un nombre premier, il est garanti que chaque cycle ne possédera qu'une seule boucle connectée.

Nous fournissons un deuxième schéma AL plus facile à mettre en œuvre, appelé AL2, que nous comparons au schéma AL1 dans les exemples empiriques.

AL2 : Pas de mélange des attributions de lignes. Répéter la même AL simple toutes les m_d unités, c.-à-d. $(1,2),(2,3),\ldots,(m_d,1)$.

3 Exemples empiriques

Les exemples empiriques servent à examiner les questions suivantes :

Q1. Dans quelle mesure l'estimateur SDR donne-t-il de bons résultats quand on se sert d'un sous-ensemble de toutes les répliques nécessaires pour que l'estimateur SDR soit équivalent à l'estimateur SD?

Q2. Quel schéma d'attribution de lignes est le meilleur, AL1 ou AL2?

Q3. Devrions-nous utiliser un plus grand nombre ou un moins grand nombre de boucles connectées?

Pour répondre à ces questions, nous avons appliqué l'estimateur de variance SDR à plusieurs populations. Pour chaque population, nous avons sélectionné un échantillon sys de taille $n = 64$. Le tableau 3.1 décrit les trois estimateurs SDR étudiés.

Tableau 3.1
Estimateurs SDR pour les exemples empiriques

Estimateur	k_A	$\mathbf{H}_A$	k_B	$\mathbf{H}_B$
1	4	$\mathbf{H}_{4a}$	16	$\mathbf{H}_{4a} \otimes \mathbf{H}_{4a}$
2	16	$\mathbf{H}_{4a} \otimes \mathbf{H}_{4a}$	4	$\mathbf{H}_{4a}$
3	64	$\mathbf{H}_{4a} \otimes \mathbf{H}_{4a} \otimes \mathbf{H}_{4a}$	1	1

Sous cette construction, les estimateurs SDR comprenaient $k_B = 1, 4$ ou 16 cycles, mais tous utilisaient la même matrice $\tilde{\mathbf{H}} = \mathbf{H}_{4a} \otimes \mathbf{H}_{4a} \otimes \mathbf{H}_{4a}$, qui est la matrice de Hadamard normale d'ordre $\tilde{k} = 64$. Pour les trois estimateurs du tableau 3.1, nous avons également fait varier le schéma d'attribution de lignes (AL1 ou AL2), ainsi que le nombre de répliques utilisées par chaque estimateur, soit 16, 32, 48 ou 64. Tant avec AL1 qu'avec AL2, il n'existe qu'une seule boucle connectée par cycle, de sorte que le nombre de boucles connectées que possédaient les estimateurs 1, 2 et 3 était $k_B = 16, 4$ et 1, respectivement. En annexe, les résultats pour les estimateurs SDR sont résumés au tableau A1, tandis que le tableau A2 donne les résultats pour les estimateurs de variance SD1, SD2 et $eassr$ appliqués aux fins de comparaison.

Jeux de données utilisés. Les populations « A » sont empruntées à l'exemple empirique de Wolter (1984). Pour les populations A1 à A7, nous avons généré 400 populations finies de taille $N = 64\,000$. Pour chaque population, il existait $b = 100$ échantillons possibles de taille $n = 64$. Les échantillons sont désignés par l'indice $i = 1, 2, \ldots, b = 100$ et, dans chaque échantillon, les unités sont désignées par l'indice $j = 1, 2, \ldots, n = 64$. Le tableau 3.2 résume comment la variable d'intérêt μ_{ij} est générée pour chacune des populations « A ».

Tableau 3.2
Description des populations artificielles de Wolter

Population	Description	n	b	μ_{ij}	e_{ij}
A1	Aléatoire	20	50	0	e_{ij} iid $N(0,100)$
A2	Tendance linéaire	20	50	$i + (j-1)k$	e_{ij} iid $N(0,100)$
A3	Effets de stratification	20	50	j	e_{ij} iid $N(0,100)$
A4	Effets de stratification	20	50	$j + 10$	$e_{ij} = \begin{cases} \varepsilon_{ij}, & \text{si } \varepsilon_{ij} \geq -(j+10) \\ -(j+10), & \text{autrement.} \end{cases}$ ε_{ij} iid $N(0,100), \rho = 0,8$
A5	Autocorrélée	20	50	0	$e_{ij} = \rho e_{i-1,j} + \varepsilon_{ij}$ $e_{i1} \sim N\left(0,100/\left(1-\rho^2\right)\right)$ ε_{ij} iid $N(0,100), \rho = 0,8$
A6	Autocorrélée	20	50	0	comme A5 avec $\rho = 0,4$
A7	Périodique	20	50	$20\sin\{2\pi/50[i+(j-1)k]\}$	e_{ij} iid $N(0,100)$

Mesures d'évaluation. Nous avons évalué les divers estimateurs de variance au moyen des trois mesures utilisées par Wolter, à savoir le biais relatif prévu (ERB pour *expected relative bias*), l'erreur quadratique moyenne relative (RMSE pour *relative mean squared error*) et le ratio de couverture. La première mesure, ERB, que nous avons utilisée pour examiner l'exactitude des estimateurs, est définie pour un estimateur donné θ comme $\text{ERB}(\hat{v}_\theta) = E_m\left(E_p(\hat{v}_\theta - v)\right)/E_m(v)$. Dans notre notation, E_p et E_m désignent les espérances sous le plan et sous le modèle, respectivement. Pour examiner la variance des estimateurs, nous avons également mesuré la RMSE, qui est définie comme étant $\text{RMSE}(\hat{v}_\theta) = E_m\left(E_p(\hat{v}_\theta - v)^2\right)/E_m(v)$. Nous avons calculé le ratio de couverture sous forme du pourcentage de fois que le vrai total de population était compris dans l'intervalle de confiance produit en utilisant l'estimation, c.-à-d. $\left(\hat{Y} - z_\alpha\sqrt{\hat{v}_\alpha}, \hat{Y} + z_\alpha\sqrt{\hat{v}_\alpha}\right)$. Ici, z_α est la valeur tirée d'une distribution normale qui a été choisie pour produire les intervalles de confiance à 95 %.

Résultats. En ce qui concerne Q1, les colonnes 4 à 7 du tableau A1 montrent que l'augmentation du nombre de répliques n'a qu'un effet minime sur le biais. Ce n'est que pour la population à tendance linéaire (A2) que l'estimateur SDR avec quatre boucles connectées présente une tendance cohérente de réduction du biais à mesure qu'augmente le nombre de répliques. Les autres combinaisons de population et d'estimateur ne révèlent aucune tendance décroissante ni croissante significatives lorsque le nombre de répliques augmente. Cette constatation représente un résultat positif, parce qu'elle indique que la réduction du jeu de répliques n'accroît pas le biais. Comme prévu, les RMSE dans les colonnes 8 à 11 du tableau A1 augmentent à mesure que le nombre de répliques diminue, mais curieusement, l'accroissement est relativement faible. De même, les intervalles de confiance présentés dans les colonnes 12 à 15 s'améliorent parallèlement à l'augmentation du nombre de répliques, sauf dans le cas des populations A2 et A7.

En ce qui concerne la question Q2, la comparaison des schémas AL1 et AL2 indique que l'estimateur SDR avec quatre boucles connectées produit habituellement de plus faibles biais (colonnes 4 à 7 du tableau A1) et variances (colonnes 8 à 11 du tableau A1) avec AL1 qu'avec AL2. Dans le cas de 16 boucles connectées, les biais et les variances sont similaires pour AL1 et AL2. Ces résultats laissent entendre que le biais et la variance sont tous deux améliorés, mais que l'effet est réduit à mesure que la taille des boucles connectées diminue.

En ce qui concerne Q3, les biais présentés dans les colonnes 4 à 7 diminuent lorsque le nombre de boucles connectées augmente. Fait exception la population périodique (A7). Lorsque les RMSE des estimateurs SD1 et SD2 ne sont pas similaires, comme dans le cas de la population à tendance linéaire (A2), l'augmentation du nombre de boucles connectées réduit également la RMSE. Ce résultat n'est pas étonnant. L'estimateur comprenant une seule grande boucle connectée est équivalent à l'estimateur SD2, de sorte qu'il peut présenter des biais et RMSE plus importants en raison du terme $(\hat{y}_1 - \hat{y}_{64})^2$. Dans l'autre sens, un plus grand nombre de boucles connectées réduit effectivement l'effet du terme $(\hat{y}_1 - \hat{y}_{64})^2$, de sorte que l'estimateur agit davantage comme l'estimateur SD1, dont le biais et la variance sont généralement plus faibles que ceux de l'estimateur SD2.

4 Conclusion

Le présent article décrit les conditions pour que l'estimateur SDR soit équivalent à l'estimateur SD2, et montre de quelle façon ils sont équivalents quand la taille de l'échantillon est plus petite ou plus grande que la matrice de Hadamard choisie. Lorsqu'une matrice de Hadamard $\mathbf{H}_A$ plus petite est utilisée et que les répliques sont tirées uniquement de $\mathbf{H}_A$, l'article montre comment le jeu réduit de répliques produit une approximation raisonnable de l'estimateur SD2. Les exemples empiriques indiquent qu'utiliser un jeu réduit de répliques est raisonnable, puisque la réduction du nombre de répliques n'accroît pas le biais des estimations. En outre, nous voyons que l'utilisation d'un grand nombre de boucles connectées réduit l'effet du carré de la différence entre la première et la dernière unité dans l'échantillon. Puisque le biais et la RMSE de l'estimateur SD1 sont généralement plus grands que ceux de l'estimateur SD2, les estimateurs SDR utilisant un plus grand nombre plutôt qu'un plus petit nombre de boucles connectées donneront des biais et RMSE plus faibles que les estimateurs SDR.

Remerciements

L'auteur remercie David Hornick et Brian Dumbacher de leur révision de la première ébauche du manuscrit, ainsi que les examinateurs et le rédacteur de leurs commentaires qui lui ont permis d'améliorer et de clarifier l'article.

Annexe

Tableau A1
Résultats des simulations de l'estimateur SDR

Population	k_A	AL	Biais relatif prévu selon le nombre de répliques				Erreur quadratique moyenne relative				Ratio de couverture			
			16	32	48	64	16	32	48	64	16	32	48	64
A1	4	1	0,010	0,009	0,009	0,009	0,176	0,091	0,066	0,054	93	94	94	94
		2	0,010	0,010	0,010	0,009	0,176	0,095	0,064	0,048	92	94	94	95
	16	1	0,009	0,008	0,010	0,009	0,141	0,080	0,059	0,048	93	94	94	95
		2	0,009	0,010	0,010	0,009	0,194	0,096	0,065	0,049	92	94	94	95
	64	1 ou 2	0,009	0,009	0,010	0,009	0,194	0,096	0,064	0,049	92	94	94	94
A2	4	1	-0,696	-0,840	-0,888	-0,907	0,485	0,706	0,789	0,823	62	45	38	35
		2	-0,538	-0,768	-0,845	-0,883	0,290	0,590	0,714	0,780	77	54	45	39
	16	1	0,113	-0,270	-0,500	-0,615	0,013	0,073	0,250	0,378	100	97	80	100
		2	1,302	0,152	-0,231	-0,423	1,695	0,023	0,054	0,179	100	100	99	100
	64	1 ou 2	1,302	1,379	1,404	1,417	1,695	1,901	1,972	2,008	100	100	100	100
A3	4	1	0,049	0,031	0,025	0,021	0,195	0,095	0,068	0,054	93	94	94	95
		2	0,070	0,040	0,030	0,025	0,222	0,103	0,067	0,050	93	94	94	95
	16	1	0,155	0,105	0,075	0,060	0,207	0,106	0,070	0,055	95	95	95	95
		2	0,314	0,163	0,112	0,086	0,374	0,144	0,085	0,061	96	95	95	95
	64	1 ou 2	0,314	0,324	0,327	0,327	0,374	0,245	0,199	0,176	96	97	97	97
A4	4	1	0,040	0,023	0,017	0,014	0,192	0,104	0,077	0,063	93	94	94	94
		2	0,060	0,030	0,021	0,017	0,217	0,110	0,075	0,058	93	94	94	95
	16	1	0,144	0,095	0,066	0,052	0,208	0,109	0,077	0,063	95	95	95	95
		2	0,291	0,146	0,098	0,075	0,357	0,144	0,090	0,067	96	95	95	95
	64	1 ou 2	0,291	0,299	0,303	0,305	0,357	0,232	0,191	0,170	96	97	97	97
A5	4	1	0,063	0,063	0,063	0,065	0,192	0,106	0,076	0,063	94	94	95	95
		2	0,068	0,066	0,066	0,065	0,217	0,111	0,075	0,057	93	94	95	95
	16	1	0,063	0,063	0,063	0,065	0,161	0,093	0,068	0,057	94	95	95	95
		2	0,065	0,067	0,066	0,066	0,214	0,111	0,075	0,056	93	94	95	95
	64	1 ou 2	0,065	0,066	0,066	0,065	0,214	0,110	0,074	0,056	93	94	95	95
A6	4	1	0,093	0,092	0,093	0,094	0,211	0,117	0,088	0,072	94	95	95	95
		2	0,092	0,096	0,095	0,094	0,229	0,120	0,086	0,067	94	95	95	95
	16	1	0,099	0,095	0,094	0,094	0,185	0,107	0,080	0,067	94	95	95	95
		2	0,093	0,094	0,094	0,093	0,226	0,117	0,085	0,067	94	95	95	95
	64	1 ou 2	0,093	0,096	0,095	0,095	0,226	0,118	0,084	0,066	94	95	95	95
A7	4	1	0,105	0,069	0,112	0,253	0,219	0,106	0,091	0,143	94	95	95	97
		2	0,004	0,004	0,073	0,310	0,187	0,098	0,079	0,175	92	94	95	97
	16	1	0,177	0,168	0,462	0,847	0,229	0,137	0,351	0,828	95	96	98	99
		2	0,002	0,003	0,027	1,248	0,187	0,097	0,065	1,689	92	94	95	100
	64	1 ou 2	0,002	0,003	0,030	0,115	0,187	0,097	0,065	0,062	92	94	95	96

Tableau A2
Résultats des simulations des méthodes comparatives

Population	Biais relatif prévu selon le nombre de répliques			Erreur quadratique moyenne relative			Ratio de couverture		
	SD1	SD2	EASSR	SD1	SD2	EASSR	SD1	SD2	EASSR
A1	0,009	0,009	-0,001	0,049	0,049	0,032	94	94	97
A2	-0,960	1,417	25,317	0,921	2,008	640,916	23	100	100
A3	0,015	0,327	3,462	0,049	0,176	12,203	94	97	100
A4	0,006	0,305	3,284	0,057	0,170	11,109	94	97	100
A5	0,064	0,065	0,055	0,056	0,056	0,039	95	95	97
A6	0,093	0,095	0,084	0,065	0,066	0,046	95	95	98
A7	0,112	0,115	20,641	0,063	0,062	427,141	96	96	100

Bibliographie

Bellhouse, D.R. (1988). Systematic sampling. Extrait de *Handbook of Statistics*, 6, 125-145.

Fay, R.E., et Train, G.F. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. *Proceedings of the Section on Government Statistics*, American Statistical Association, 154-159.

Hedayat, A., et Wallis, W.D. (1978). Hadamard matrices and their applications. *The Annuals of Statistics*, 6, 1184-1238.

Huang, E.T., et Bell, W.R. (2009). A simulation study of the distribution of Fay's successive difference replication variance estimator. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 5294-5308.

Iachan, R. (1982). Systematic sampling: A critical review. *International Statistical Review*, 50, 293-303.

Madow, W.G., et Madow, L.H. (1944). On the theory of systematic sampling. *Annuals of Mathematical Statistics*, 15, 1-14.

McCarthy, P.J. (1966). Pseudo-replication: Half-samples. *Review of the International Statistical Institute*, 37, 239-264.

Murthy. M.N., et Rao, T.J. (1988). Systematic sampling with illustrative examples. Extrait de *Handbook of Statistics*, 6, 147-185.

Sukasih, A.S., et Jang, D. (2003). Monte Carlo study on the successive difference replication method for non-linear statistics. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3608-3612.

Wolter, K.M. (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, 781-790.

Wolter, K.M. (1985). *Introduction to Variance Estimation*, Springer-Verlag.

Yates, F. (1953). *Sampling Methods for Censuses and Surveys*, 2nd Edition, Hafner Publishing Company, New York, NY.

U.S. Census Bureau (2006). Technical Paper 66, "Design and Methodology: Current Population Survey," Octobre 2006.

ELECTRONIC PUBLICATIONS AVAILABLE AT | PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À

www.statcan.gc.ca

Techniques d'enquête, juin 2014
Vol. 40, N° 1, pp. 69-88
Statistique Canada, N° 12-001-X au catalogue

Estimation de variance par linéarisation pour des indices de pauvreté et d'exclusion sociale

Eric Graf et Yves Tillé[1]

Résumé

Nous avons mis en œuvre la technique de linéarisation généralisée reposant sur le concept de fonction d'influence tout comme l'a fait Osier (2009) pour estimer la variance de statistiques complexes telles que les indices de Laeken. Des simulations réalisées avec le langage R montrent que, pour les cas où l'on a recours à une estimation par noyau gaussien de la fonction de densité des revenus considérés, on obtient un fort biais pour la valeur estimée de la variance. On propose deux autres méthodes pour estimer la densité qui diminuent fortement le biais constaté. L'une de ces méthodes a déjà été esquissée par Deville (2000). Les résultats publiés ici permettront une amélioration substantielle de la qualité des informations sur la précision de certains indices de Laeken diffusées et comparées internationalement.

Mots-clés : Fonction d'influence; enquête SILC; statistiques non linéaires; indices de pauvreté et d'inégalité.

1 Introduction

Deville (2000) a proposé d'estimer la précision des statistiques non linéaires dans des plans de sondage au moyen de la technique de linéarisation généralisée. Celle-ci repose sur le concept de fonction d'influence proposée par Hampel (1974) dans le domaine de la statistique robuste. Osier (2009) a appliqué ces théories pour estimer la variance de statistiques complexes telles que les indices de Laeken (Eurostat 2005) dans l'enquête européenne sur le revenu et les conditions de vie (EU-SILC). Goga, Deville et Ruiz-Gazen (2009) étendent la théorie de Deville (2000) à des enquêtes basées sur deux échantillons. Verma et Betti (2011) fournissent une liste exhaustive des indicateurs de pauvreté conventionnels et des variables linéarisées associées, ils comparent également les performances de la technique de linéarisation à la méthode du Jackknife (*Jackknife repeated replication*). Dans cet article, nous nous restreignons à des indicateurs de pauvreté publiés dans le cadre de l'enquête SILC, en focalisant sur la façon d'estimer la fonction de densité des revenus en différents points de leur distribution.

Dans la section 2, on rappelle les fondements théoriques nécessaires ainsi que les expressions des indices de pauvreté et d'inégalité étudiés, de même que celle de leur linéarisée. Certaines linéarisées dépendent de la fonction de densité de la variable d'intérêt qui est généralement estimée par noyau gaussien. Deux méthodes alternatives au noyau gaussien sont présentées dans la section 3. Les simulations réalisées en langage R sont décrites et commentées dans la section 4. On y montre que la méthode par noyau gaussien peut engendrer un fort biais pour la valeur estimée de la variance des indices pour les cas où l'on a recours à une estimation de la fonction de densité des revenus considérés. On voit aussi que les deux autres méthodes pour estimer la densité proposées dans la section 3 diminuent le biais constaté, ce qui fait aussi l'objet des conclusions dans la dernière partie de ce texte.

1. Eric Graf et Yves Tillé, Institut de Statistique, Faculté des sciences économiques, Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel, Suisse. Courriel : eric.graf@unine.ch et yves.tille@unine.ch.

2 Rappels sur les indices de pauvreté traités ainsi que leur linéarisée

Soit une population finie U constituée de N unités identifiables $u_1, ..., u_k, ..., u_N$. Pour simplifier l'écriture on désigne par la suite l'unité u_k par son indice k. En pratique la population U est une base de sondage avec un taux de couverture acceptable de la population d'intérêt pour laquelle on désire faire des inférences. On associe à chaque unité k une valeur y_k d'une caractéristique d'intérêt (ici un revenu). Sans nuire à la généralité et pour alléger les notations, on suppose que les y_k sont tous distincts et triés par ordre de grandeur, donc $y_k = y_{[k]}$. Dans les données issues d'enquêtes par échantillonnage, il arrive fréquemment qu'il y ait des doublons, c'est-à-dire plusieurs unités ayant la même valeur y, que ce soit dû à des arrondis ou à des questions-fourchettes. Dans ces cas et pour cette étude, il suffit d'ajouter un montant assez petit (c'est-à-dire négligeable) de la devise tiré au hasard selon une loi uniforme pour que les données soient triables sans équivoque.

Soit un échantillon aléatoire S de taille n obtenu par un plan d'échantillonnage $p(s) = \mathrm{P}(S = s)$, pour tout $s \subset U$. Soit également $\pi_k = \mathrm{P}(k \in s) > 0$ la probabilité d'inclusion dans l'échantillon de l'unité k de U. Soit aussi $d_k = 1/\pi_k$ le poids d'échantillonnage et $w_k = w_k(s)$ un poids d'estimation qui peut être égal à d_k mais qui peut aussi être plus raffiné. Par exemple, w_k peut avoir été obtenu suite à un calage (Deville et Särndal 1992) et refléter ainsi également une correction de non-réponse.

Les estimateurs des indices de pauvreté et d'inégalité sont des statistiques non linéaires qui ne peuvent pas s'exprimer comme des fonctions régulières (c'est-à-dire continûment différentiables jusqu'à l'ordre deux) de totaux. En effet, il s'agit de statistiques de rangs pour l'indice de Gini et de quantiles pour les autres. Comme le relève Osier (2009), leur variance ne peut donc pas être estimée par une linéarisation de Taylor mais nécessite le recours à la méthode de linéarisation généralisée (Deville 2000; Demnati et Rao 2004; Osier 2009). Une alternative pour estimer la variance serait d'utiliser des techniques de rééchantillonnage du genre bootstrap, mais dans le cadre des données de l'enquête SILC, une préférence a été donnée à la technique de linéarisation, du moins pour un certain nombre de pays participants. En effet, les méthodes de rééchantillonnage nécessitent souvent davantage de ressources humaines et machine. De plus, Eurostat collaborant avec une trentaine de pays ayant des plans de sondage différents suivis d'éventuels corrections pour la non-réponse et de calages sur des sources externes, il a semblé plus adéquat d'opter pour une solution analytique pour estimer la variance. Par ailleurs, certains pays pouvaient utiliser le logiciel SAS déjà existant POULPE (Ardilly et Osier 2007) pour produire les estimations nécessaires. Ce fut le cas pour les premiers tests avec les données SILC suisses. On applique ici un mode opératoire qui, comme le relèvent Antal, Langel et Tillé (2011), concilie l'approche introduite par Deville (2000) à celle de Demnati et Rao (2004). Les deux approches utilisent la notion de *fonction d'influence* qui fut développée initialement dans le domaine des statistiques robustes (Hampel 1974). Antal et coll. (2011) mentionnent aussi que l'on peut retrouver les mêmes linéarisées en appliquant la méthode proposée par Graf (2011, 2013) qui construit une variable linéarisée basée sur un développement en série de Taylor par rapport aux indicatrices d'inclusion dans l'échantillon. Citons aussi le travail de Kovačević et Binder (1997) où une approche de linéarisation par les équations estimantes est développée.

Selon Deville (2000), on cerne l'influence de l'unité k sur un paramètre d'intérêt θ au niveau de la population par une variation infinitésimale de l'importance attribuée à cette unité. On exprime ledit paramètre comme une fonctionnelle $\theta = T(M)$, où M est une mesure allouant une masse unité,

$M(k) = M_k = 1$, uniquement aux points du continuum correspondant aux unités $k \in U$. La spécialisation de la mesure générale M en une mesure discrète fait passer la fonctionnelle T, définie *a priori* sur un continuum, en une fonctionnelle discrète, tout comme le total Y est défini par la somme des y_k sur notre population finie. La *fonction d'influence* de T, ou la *variable linéarisée*, est définie par

$$I[T(M)]_k = z_k = \lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t}, \text{pour tout } k \in U,$$

où δ_k est la mesure de Dirac pour l'unité $k\,(\delta_k(i) = 1 \text{ si } i = k \text{ et } 0 \text{ sinon})$. En pratique on ne dispose que des données connues sur un échantillon S et Deville (2000) obtient une linéarisée $\hat{z}_k$ ou *fonction d'influence empirique*, en : 1) évaluant la limite ci-dessus par calcul différentiel, 2) remplaçant dans l'évaluation les quantités inconnues par les quantités correspondantes estimées à partir de l'échantillon. Il justifie ce procédé en montrant que :

$$T(\hat{M}) - T(M) \approx \left(\sum_{k \in S} w_k z_k - \sum_{k \in U} z_k\right).$$

Le résultat central est que, sous des conditions asymptotiques décrites dans Deville (2000), qui sont en principe satisfaites lorsque l'échantillon est « assez grand », la variance du total estimé de la variable $\hat{z}_k$ est une approximation de la variance de la statistique (complexe) $\hat{\theta}$:

$$\text{var}\left[\sum_{k \in s} \hat{z}_k w_k\right] \approx \text{var}(\hat{\theta}).$$

Le point de départ de l'approche de Deville est donc le paramètre de population et non l'estimateur qu'on se propose d'utiliser pour l'évaluation à partir de l'échantillon. Dans les cas où l'estimateur utilisé découle naturellement de l'expression du paramètre de population (comme par exemple le total Y approché par l'estimateur de Horvitz-Thompson), le procédé ne présente pas d'ambiguïté. Mais des imprécisions surviennent si l'on estime le même total Y en ayant recours à l'estimateur par le quotient grâce à une variable auxiliaire x. En effet, dans ce cas, l'approche de Deville ne précisant pas la forme de l'estimateur du total à utiliser fournira plutôt une fonction d'influence constante égale à 1, au lieu de faire intervenir le quotient inconnu d'intérêt.

Une alternative qui évite ces problèmes est celle de Demnati-Rao, lorsque rattachée au préalable au cadre de Deville comme cela est fait dans Antal et coll. (2011). Ces auteurs présentent l'approche de Demnati-Rao comme résultant du cadre de Deville lorsque la mesure M employée n'est pas la mesure discrète définie sur U présentée précédemment, mais plutôt la mesure suivante définie sur S, l'échantillon :

$$\hat{M}(k) = w_k, k \in S$$

où w_k est un poids. En définissant la mesure sur S, on se donne dans les faits comme point de départ l'estimateur et non le paramètre; c'est le paramètre qui se retrouve exprimé au départ sous la forme d'une fonctionnelle et non le paramètre de population à estimer. Autrement dit, on se donne comme fonctionnelle celle qui correspond à l'estimateur pour lequel on souhaite avoir une estimation de la variance par linéarisation généralisée. On obtient ensuite la linéarisée en fonction de cette fonctionnelle de la façon suivante :

$$I[T(\hat{M})]_k = \hat{z}_k = \lim_{t \to 0} \frac{T(\hat{M} + t\delta_k) - T(\hat{M})}{t}, \text{pour tout } k \in S.$$

Antal et coll. (2011) observent que, dans la mesure où la fonctionnelle apparaissant dans cette limite s'exprime comme une fonction explicite des variables que sont les poids assignés par la mesure $\hat{M}$ aux observations, cette linéarisée est en fait une fonction des dérivées partielles par rapport aux poids :

$$I[T(\hat{M})]_k = \frac{\partial T(\hat{M})}{\partial w_k}.$$

Antal et coll. (2011) relèvent que, les linérarisées que nous rappelons dans la suite, peuvent être obtenues par les deux approches. En effet, l'évaluation de la limite à-la-Demnati-Rao ne mène pas nécessairement à l'estimation de la variance suggérée par Deville (2000). L'approche pratique utilisée dans le cadre de cet article pourrait donc être nommée comme étant celle de Deville-Demnati-Rao en reconnaissance au cadre théorique fourni par Deville (2000) et l'algorithmique pratique de ce cadre qu'apportent Demnati et Rao (2004).

Par cette méthode, on peut estimer la variance de $\hat{\theta}$ quel que soit le plan d'échantillonnage, et donc obtenir un intervalle de confiance, en substituant la variable linéarisée dans la formule de variance pour un total correspondant au plan choisi. Sous un plan aléatoire simple sans remise, l'estimateur de la variance d'un indice d'inégalité $\hat{\theta}$ est donné par

$$\widehat{\mathrm{var}}_{\mathrm{lin}}[\hat{\theta}] = \frac{N(N-n)}{n}\frac{1}{n-1}\sum_{k\in S}(\hat{z}_k - \bar{z})^2, \tag{2.1}$$

avec

$$\bar{z} = n^{-1}\sum_{k\in S}\hat{z}_k.$$

Dans la suite, dans le cadre de la mesure du revenu de la population, nous rappelons les définitions empiriques des indices d'inégalité considérés ainsi que l'expression de leurs linéarisées telles que nous les avons mises en œuvre.

2.1 L'indice de Gini

L'indice de Gini, G, est compris entre 0 (en cas d'égalité totale, tous gagnent le même montant) et 1 (en cas d'inégalité totale, c'est-à-dire un individu gagne tout et les autres rien). L'indice G s'exprime en fonction des revenus cumulés d'une certaine proportion d'individus les plus pauvres. Si $\mathcal{Y}$ est la variable aléatoire représentant les revenus, $f(y)$ sa fonction de densité et $F(y)$ sa fonction de répartition, alors la *courbe de Lorenz* (Lorenz 1905) est définie par

$$L(\alpha) = \frac{\int_0^{F^{-1}(\alpha)} yf(y)\,dy}{\int_0^{\infty} yf(y)\,dy} = \frac{1}{\mathrm{E}(\mathcal{Y})}\int_0^{\alpha} F^{-1}(u)\,du.$$

L'indice de Gini représente deux fois la surface comprise entre la courbe de Lorenz et la ligne (diagonale $f_{eg}(x) = x$) de l'égalité parfaite (Figure 2.1). Il est donc défini par :

$$G = 2\int_0^1 [\alpha - L(\alpha)]\,d\alpha.$$

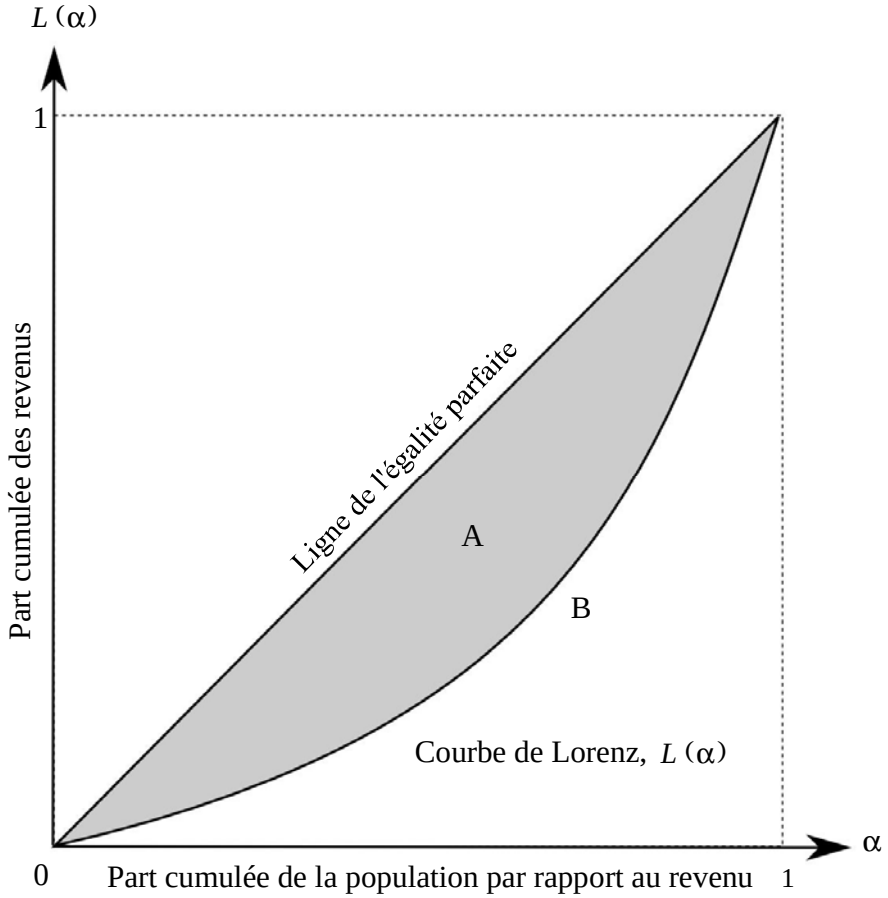

Figure 2.1 Indice de Gini G et courbe de Lorenz $L(\alpha)$. $G = 2A$, $A + B = 1/2$

Dans le cas d'une population finie, les y_k ne sont pas aléatoires et l'indice de Gini est défini sur la population U par :

$$G = \frac{2\sum_{k\in U} k y_k}{N\sum_{k\in U} y_k} - \frac{N+1}{N},$$

où les y_k ont été préalablement triés par leur rang. Calculé sur un échantillon, on l'estime par :

$$\begin{aligned}\hat{G} &= \frac{2}{\hat{N}\hat{Y}}\sum_{k\in S} w_k \hat{N}_k y_k - \left(1 + \frac{1}{\hat{N}\hat{Y}}\sum_{k\in S} w_k^2 y_k\right) \\ &= \frac{\sum_{k\in S}\sum_{\ell\in S} w_k w_\ell |y_k - y_\ell|}{2\hat{N}\hat{Y}},\end{aligned}$$

où $\hat{N}_k = \sum_{\ell\in S} w_\ell \mathbf{1}_{[y_\ell \le y_k]}$ est la somme cumulée des poids w_k, $\hat{Y} = \sum_{k\in S} w_k y_k$ étant le revenu total estimé pour la population et $\hat{N} = \sum_{k\in S} w_k$ la taille estimée de cette population. L'expression peut être simplifiée si tous les poids sont égaux et valent tous N/n :

$$\hat{G} = \frac{2\sum_{k\in S} k y_k}{n\sum_{k\in S} y_k} - \frac{n+1}{n}.$$

Notons que la définition peut varier d'un facteur $n/(n-1)$ selon les auteurs (Osier 2009; Eurostat 2004b), mais cette subtilité est négligeable dès que la taille de l'échantillon est assez grande.

Langel et Tillé (2012) ont fait une synthèse des différentes approches permettant d'obtenir la même variable linéarisée de l'indice de Gini estimée sur l'échantillon :

$$\hat{z}_k^{\text{GINI}} = \frac{1}{\hat{N}\hat{Y}}\left[2\hat{N}_k(y_k - \hat{\bar{Y}}_k) + \hat{Y} - \hat{N}y_k - \hat{G}\left(\hat{Y} + y_k\hat{N}\right)\right],$$

où $\hat{\bar{Y}}_k = \sum_{\ell=1}^{k} w_\ell y_\ell / \hat{N}_k$, les y_ℓ étant triés et distincts.

2.2 Le Quintile Share Ratio (QSR ou S_{80}/S_{20})

On trouve un bon tour d'horizon sur cet indice dans Langel et Tillé (2012). Soient q_{80} et q_{20} les 80e et 20e percentiles de la fonction de répartition $\mathrm{F}(y)$. Le QSR est le ratio de la somme des revenus des 20 % les plus riches sur les 20 % les plus pauvres. Dans le cas continu, on peut le définir ainsi :

$$\text{QSR} = \frac{\mathrm{E}(\mathcal{Y} \mid \mathcal{Y} > q_{80})}{\mathrm{E}(\mathcal{Y} \mid \mathcal{Y} < q_{20})} = \frac{1 - L(0{,}8)}{L(0{,}2)},$$

où $\mathcal{Y}$ serait une variable aléatoire représentant les revenus. Dans le cas de populations finies, le QSR peut être exprimé et estimé au niveau de l'échantillon en fonction des sommes partielles

$$\widehat{\text{QSR}} = \frac{\hat{Y} - \hat{Y}_{0,8}}{\hat{Y}_{0,2}},$$

où, suite aux résultats obtenus par Langel et Tillé (2011), nous utiliserons la définition suivante de la somme partielle, ce qui diffère très légèrement de la définition officielle d'Eurostat (2004a),

$$\hat{Y}_\alpha = \sum_{k\in S} w_k y_k H\left(\frac{\alpha\hat{N} - \hat{N}_{k-1}}{w_k}\right), \tag{2.2}$$

avec

$$H(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \le x < 1 \\ 1 & \text{si } x \ge 1. \end{cases}$$

Pour obtenir la linéarisée du QSR, il faut d'abord calculer la linéarisée de la somme partielle (2.2) qui est donnée par :

$$I(Y_\alpha)_k = y_k H(\alpha N - k + 1) + \left[\alpha - \mathbf{1}_{[y_k < Q_\alpha]}\right] Q_\alpha,$$

où $Q_\alpha = y_i$, avec $\hat{N}_{i-1} < \alpha\hat{N} \le \hat{N}_i$, correspond à la première définition du quantile d'une population finie dans l'article de Hyndman et Fan (1996). Osier (2009) obtient une linéarisée qui dépend de la densité de la variable $\mathcal{Y}$. Langel et Tillé (2011) ont cependant montré qu'une simplification permet d'éluder le

problème de l'estimation de cette densité pour le QSR et qu'il n'est donc pas nécessaire de faire une approximation par noyau de la densité des revenus comme le propose Osier (2009).

La fonction d'influence dépend de celles des sommes partielles :

$$I\,(\mathrm{QSR})_k = z_k^{\mathrm{QSR}} = \frac{y_k - I\,(Y_{0,8})}{Y_{0,2}} - \frac{(Y - Y_{0,8})\,I\,(Y_{0,2})}{Y_{0,2}^2}.$$

En faisant les substitutions nécessaires, on trouve que la linéarisée estimée sur la base de l'échantillon vaut

$$\begin{aligned}\hat{z}_k^{\mathrm{QSR}} = & \frac{y_k - \left\{ y_k H\left(\dfrac{0{,}8\hat{N} - \hat{N}_{k-1}}{w_k}\right) + \hat{Q}_{0,8}\left[0{,}8 - \mathbf{1}_{\left[y_k < \hat{Q}_{0,8}\right]}\right]\right\}}{\hat{Y}_{0,2}} \\ & - \frac{\left(\hat{Y} - \hat{Y}_{0,8}\right)\left\{ y_k H\left(\dfrac{0{,}2\hat{N} - \hat{N}_{k-1}}{w_k}\right) + \hat{Q}_{0,2}\left[0{,}2 - \mathbf{1}_{\left[y_k < \hat{Q}_{0,2}\right]}\right]\right\}}{\hat{Y}_{0,2}^2}.\end{aligned} \tag{2.3}$$

2.3 Linéarisée d'un quantile

Avant de traiter les indices de pauvreté, il convient de donner quelques détails sur la linéarisée d'un quantile d'ordre α. Celle-ci s'estime par :

$$\hat{z}_k^{Q_\alpha} = -\frac{1}{f\left(\hat{Q}_\alpha\right)}\frac{1}{\hat{N}}\left[\mathbf{1}_{\left[y_k \le \hat{Q}_\alpha\right]} - \alpha\right],$$

où le quantile pondéré peut être défini de manière similaire à la somme partielle (2.2) et $f(\cdot)$ est la fonction de densité des revenus qui est discutée en détails à la section 3. Notons qu'Eurostat (2004a) préconise la deuxième définition de Hyndman et Fan (1996). On pourrait discuter de la définition d'Eurostat et utiliser une autre définition du quantile, par exemple $Q_\alpha = y_{k-1} + (y_k - y_{k-1})[\alpha N - (k-1)]$ où $\alpha N < k \le \alpha N + 1$, ce qui correspond à la quatrième définition selon Hyndman et Fan (1996). On estime alors le quantile sur l'échantillon par

$$\hat{Q}_\alpha = y_{k-1} + (y_k - y_{k-1})\left(\frac{\alpha\hat{N} - \hat{N}_{k-1}}{w_k}\right).$$

La linéarisée du quantile dépend de la valeur de la fonction de densité des revenus en ce quantile. Or, la vraie densité des revenus n'est pas connue et doit donc aussi être estimée à partir de l'échantillon. Deville (2000) et Osier (2009) proposent de l'estimer par noyau gaussien. On revient plus en détails sur le problème de l'estimation de f dans la section 3.

En plus du problème de l'estimation de la fonction de densité des revenus, Croux (1998) montre que la fonction d'influence empirique de la médiane n'est pas un estimateur convergent de la fonction d'influence (théorique) correspondante. Pour une variable positive (cas des revenus), la fonction d'influence empirique de la médiane (cas que Croux traite dans son article) converge vers une distribution exponentielle dont l'espérance est la fonction d'influence. Elle résiste mal aux valeurs extrêmes s'il y en a une trop grande proportion. On dira qu'elle manque de robustesse, dans le sens que la valeur de l'estimateur sur la base de l'échantillon peut s'éloigner fortement de la vraie valeur sur la population en

raison d'observations extrêmes (c'est-à-dire très grandes par rapport aux autres) présentes dans l'échantillon (voir Hampel (1974) pour les idées de base sur la robustesse en population infinie, et Beaumont, Haziza et Ruiz-Gazen (2013) pour des réflexions récentes à ce sujet dans le cadre d'échantillonnage en population finie).

2.4 La médiane et le seuil de risque de pauvreté (ARPT)

Soit $\hat{m} = \hat{Q}_{0,5}$ la médiane estimée sur l'échantillon, le seuil de pauvreté, noté ARPT (*At Risk of Poverty Threshold*), est défini comme étant égal à 60 % de la médiane :

$$\text{ARPT} = 0{,}6F^{-1}(0{,}5)$$
$$\widehat{\text{ARPT}} = 0{,}6\hat{Q}_{0,5} = 0{,}6\hat{m}.$$

Il s'agit d'une mesure absolue qui dépend de l'échelle. La linéarisée de l'ARPT est proportionnelle à celle de la médiane :

$$\hat{z}_k^{\text{ARPT}} = I(\text{ARPT})_k = 0{,}6I(\text{MED})_k = -\frac{0{,}6}{f(\hat{m})}\frac{1}{\hat{N}}\left[\mathbf{1}_{[y_k \le \hat{m}]} - 0{,}5\right].$$

2.5 Le taux de risque de pauvreté (ARPR)

Le taux de risque de pauvreté (*At Risk of Poverty Rate*), $\text{ARPR} \in [0,1]$, correspond à la proportion de la population au-dessous du seuil de risque de pauvreté, $\text{ARPR} = F(\text{ARPT})$. Il est indépendant de l'échelle tout comme l'indice de Gini, le QSR et le RMPG (voir section 2.7). La définition officielle d'Eurostat (2004a) de son estimation à partir de l'échantillon est

$$\widehat{\text{ARPR}} = \frac{\sum_{y_k < \widehat{\text{ARPT}}} w_k}{\hat{N}}.$$

La linéarisée de l'ARPR est donnée par Osier (2009) :

$$\begin{aligned}\hat{z}_k^{\text{ARPR}} &= \frac{1}{\hat{N}}\left(\mathbf{1}_{[y_k \le \widehat{\text{ARPT}}]} - \widehat{\text{ARPR}}\right) - \frac{f(\widehat{\text{ARPT}})}{f(\hat{m})}\frac{0{,}6}{\hat{N}}\left(\mathbf{1}_{[y_k \le \hat{m}]} - 0{,}5\right)\\ &= \frac{1}{\hat{N}}\left(\mathbf{1}_{[y_k \le \widehat{\text{ARPT}}]} - \widehat{\text{ARPR}}\right) + f(\widehat{\text{ARPT}})\,\hat{z}_k^{\text{ARPT}}.\end{aligned}$$

Ici, la fonction de densité des revenus doit être estimée en deux points : en la médiane et en l'ARPT.

2.6 La médiane des pauvres

Il s'agit du revenu médian des personnes en-dessous du seuil de risque de pauvreté : $m_p = F^{-1}(1/2\,F(\text{ARPT}))$. On l'estime selon la même procédure que tout autre quantile dont la définition exacte peut varier. La linéarisée de m_p (Osier 2009) dépend de celle de l'ARPR :

$$\hat{z}_k^{m_p} = \frac{1}{f(\hat{m}_p)}\frac{\hat{z}_k^{\text{ARPR}}}{2} - \frac{1}{\hat{N}}\left(\mathbf{1}_{[y_k \le \hat{m}_p]} - F(\hat{m}_p)\right).$$

L'estimation de la densité des revenus intervient donc trois fois : en la médiane et en l'ARPT dans $\hat{z}_k^{\text{ARPR}}$ et en la médiane des pauvres m_p.

2.7 Le Relative Median Poverty Gap (RMPG)

Il s'agit de la différence relative entre le seuil de risque de pauvreté et la médiane des pauvres. On définit $\text{RMPG} = 0$ si tous les « pauvres » gagnent un montant égal au seuil et $\text{RMPG} = 1$ si les pauvres ne gagnent rien du tout. C'est une mesure de « combien pauvre sont les pauvres » :

$$\text{RMPG} = \frac{\text{ARPT} - m_p}{\text{ARPT}}.$$

Son estimation à partir de l'échantillon est donc déjà décrite. L'influence de chaque observation sur le RMPG est donnée par Osier (2009) :

$$\hat{z}_k^{\text{RMPG}} = \frac{\hat{m}_p \hat{z}_k^{\text{ARPT}} - \widehat{\text{ARPT}}\, \hat{z}_k^{m_p}}{\widehat{\text{ARPT}}^2}.$$

L'estimation de la densité de la distribution des revenus intervient à quatre reprises : une fois dans le calcul de $\hat{z}_k^{\text{ARPT}}$ et trois fois dans celui de $\hat{z}_k^{m_p}$.

3 L'estimation de la fonction de densité des revenus

Dans une approche basée sur le plan (*design based*) en population finie, l'inférence se fait par rapport au plan de sondage $\text{P}(S)$ utilisé pour sélectionner l'échantillon S dans la population U de taille finie N. Dans cette approche, seules les indicatrices d'inclusion dans l'échantillon sont aléatoires, toutes les autres grandeurs sont fixes. La fonction de répartition des revenus au niveau de la population est alors une fonction en escaliers : $F_y(x) = \sum_{k \in U} \mathbf{1}_{y_k \le x} \big/ N$ et sa dérivée, la fonction de densité, n'existe pas à cause des discontinuités. Si l'on ne veut pas se placer dans une approche basée sur le modèle (*model based*) avec un modèle de super population pour justifier le terme de fonction de densité des revenus, il faut artificiellement lisser la fonction de répartition pour qu'elle devienne dérivable. C'est donc par abus de langage que nous nous autorisons ici à parler de fonction de densité. Avec cette volonté de lissage, Deville (2000) et Osier (2009) proposent d'estimer la fonction de densité des revenus par noyau gaussien :

$$\begin{aligned} K(u) &= \frac{1}{h\sqrt{2\pi}} e^{-u^2/2}, \qquad u = \frac{x - y_k}{h} \\ \hat{f}_1(x) &= \frac{1}{\hat{N}} \sum_{k \in S} w_k K\left(\frac{x - y_k}{h}\right) \\ &= \frac{1}{h\sqrt{2\pi}} \frac{1}{\hat{N}} \sum_{k \in S} w_k \exp\left[-\frac{(x - y_k)^2}{2h^2}\right] \end{aligned} \tag{3.1}$$

où h est la largeur de la bande qu'Osier estime par $\hat{h} = \hat{\sigma}\hat{N}^{-0,2}$ et $\hat{\sigma}$ est l'écart-type estimé de la distribution empirique des revenus :

$$\hat{\sigma} = \sqrt{\frac{\sum_{k\in S} w_k y_k^2}{\hat{N}} - \left(\frac{\sum_{k\in S} w_k y_k}{\hat{N}}\right)^2} = \sqrt{\frac{\sum_{k\in S} w_k y_k^2}{\hat{N}} - \bar{y}_w^2}.$$

Notons que cette estimation de σ n'est pas robuste étant très sensible aux valeurs extrêmes de y. Des données de revenus ont très souvent une queue de distribution étendue vers la droite avec des valeurs très élevées possibles, on parle de valeurs aberrantes représentatives (*representative outliers*) au sens de Chambers (1986) et Hulliger (1999). Comme le montrent nos simulations dans la section 4, cela peut fortement biaiser nos estimations de variance. Verma et Betti (2011) procèdent également par noyau rappelant que, selon Silverman (1986), le choix du noyau n'est pas crucial pour assurer la convergence de $\hat{f}(y)$ vers $f(y)$, alors que celui de la largeur de bande l'est. Ils utilisent une valeur conseillée par Silverman dans le cas de distributions avec un coefficient d'asymétrie positif, $h = 0{,}79(\hat{Q}_{75} - \hat{Q}_{25})\hat{N}^{-0{,}2}$. Dans leurs conclusions, ils relèvent que la méthode par linéarisation peut être problématique en raison d'irrégularités de la fonction de densité empirique. On ajoutera que ces problèmes sont d'autant plus préoccupants qu'il est fréquent, dans les données issues d'enquêtes, d'avoir des agglomérats d'observations à certaines valeurs (dues à des arrondis ou des questions-fourchettes), ce qui peut compliquer l'estimation de la densité. La suite de l'article décrit des solutions que nous avançons pour réduire le biais de la variance estimée.

3.1 Passer par le logarithme

Une solution qui, comme on le verra plus loin, donne de très bons résultats est de simplement passer par le logarithme pour estimer la densité en x. Si l'on pose $v = \log(x + a)$, où x est le revenu et a un nombre réel positif par exemple égal à $(|\min_k(y_k)| + 1)$ dans le cas où l'on aurait des revenus négatifs ou nuls (en négligeant le fait que a serait estimé), on a que

$$F_v(v) = \mathrm{P}(\mathcal{V} \le v) = \mathrm{P}(\log(\mathcal{Y} + a) \le v) = \mathrm{P}(\mathcal{Y} \le e^v - a) = F_y(e^v - a),$$

où $\mathcal{V}$ et $\mathcal{Y}$ seraient de variables aléatoires. Donc,

$$f_v(v) = \frac{dF_v(v)}{dv} = \frac{dF_y(e^v - a)}{dv} = f_y(e^v - a)e^v.$$

Autrement dit $f_v(v) = f_y(x)(x + a)$, ce qui nous donne l'estimateur suivant de la densité en x :

$$\hat{f}_2(x) = \frac{\hat{f}_v(v)}{x + a} = \frac{\hat{f}_y(\log(x + a))}{x + a}. \tag{3.2}$$

L'estimation de la densité en x de $\mathcal{Y}$ peut donc s'évaluer en estimant celle du logarithme de la variable, divisée par la valeur de la variable au point qui nous intéresse. La propriété reste valable en population finie. Le fait de passer par le logarithme a l'avantage de diminuer l'effet levier exercé par les grandes valeurs des revenus dans le calcul de l'approximation de la densité par noyau. Les simulations montrent que cette méthode très simple réduit fortement le biais.

3.2 Plus proches voisins avec largeur de bande minimale

Deville (2000) esquisse une autre manière du type « plus proches voisins » (voir Silverman 1986) d'estimer la densité en utilisant le noyau

$$K_D(u) = \begin{cases} \dfrac{1}{b-a} & \text{si } a \le u < b \\ 0 & \text{sinon,} \end{cases},$$

avec $u = y_k$ et où le choix de a et b, vérifiant $x \in [a, b]$, reste à déterminer et pourrait dépendre de x. La distance $(b - a)$ représente la largeur de bande h. L'estimation de la densité vaudrait alors

$$\begin{aligned} \hat{f}_D(x, a, b) &= \frac{1}{\hat{N}} \sum_{k \in S} K_D(y_k) \\ &= \frac{1}{\hat{N}} \sum_{k \in S} w_k \frac{1}{b-a} \mathbf{1}_{y_k \in [a,b[} \\ &= \frac{\hat{F}_y(b) - \hat{F}_y(a)}{b-a}, x \in [a, b[\end{aligned} \tag{3.3}$$

avec $\hat{F}_y(x) = \sum_{k \in S} w_k \mathbf{1}_{y_k \le x} \big/ \hat{N}$.

Notons que l'estimation de la densité (3.3) n'est pas une fonction continue et qu'elle ne serait pas très adaptée pour estimer des valeurs de densité à l'extrémité des queues de la distribution. Puisque nos travaux ne reposent pas trop sur les queues de la distribution, nous considérons cette approche comme une option.

Notre deuxième proposition d'estimation de la densité en x s'inspire de l'idée ci-dessus. Elle est du type « plus proches voisins », mais impose aussi une largeur de bande minimale : notre méthode impose d'utiliser au minimum les p plus proches observations du point x tout en imposant une largeur de bande minimale $h(p) \ge h_{\text{opt}}$ où

$$h_{\text{opt}} = \frac{0{,}9 \min(\hat{\sigma}, \hat{Q}_{75} - \hat{Q}_{25})}{1{,}34 \sqrt[5]{\hat{N}}}$$

est la règle empirique (*rule of thumb*) de Silverman (1986) pour déterminer la largeur de la bande. Cette valeur est aussi utilisée par défaut par la fonction R `density` pour la largeur de la bande si rien n'est spécifié. Cette solution est plus robuste que (3.1) et évite les problèmes que l'on rencontre lorsque plusieurs valeurs y_k sont très proches les unes des autres, ce qui arrive fréquemment parce que les personnes interrogées ont tendance à arrondir leur revenu.

Les valeurs y_k, $k = 1, ..., n$, étant supposées ordonnées par leur rang, la largeur $h(p)$ de la fenêtre autour de x est initialement déterminée par les p plus proches observations, avec $p \ll n$. Pour les simulations exposées dans la section suivante, après différents essais, le p initial a été fixé à 30. On impute comme densité en x la densité estimée au point observé y_j le plus proche inférieur ou égal à x, c'est-à-dire $j = \max(k \mid y_k \le x)$, $k = 1, ..., n$. La largeur de la bande en x dépendra en fait des p_j plus proches observations autour de y_j, avec $p_j \ge p$. On la désignera par $h(p_j)$ dans la suite pour le

rappeler. La densité n'est donc estimée qu'en des points observés sans qu'un lissage ou une interpolation soient menés entre les $\hat{f}(y_j)$. L'algorithme pour estimer $\hat{f}(y_j)$ est le suivant (voir aussi Figure 3.1) :

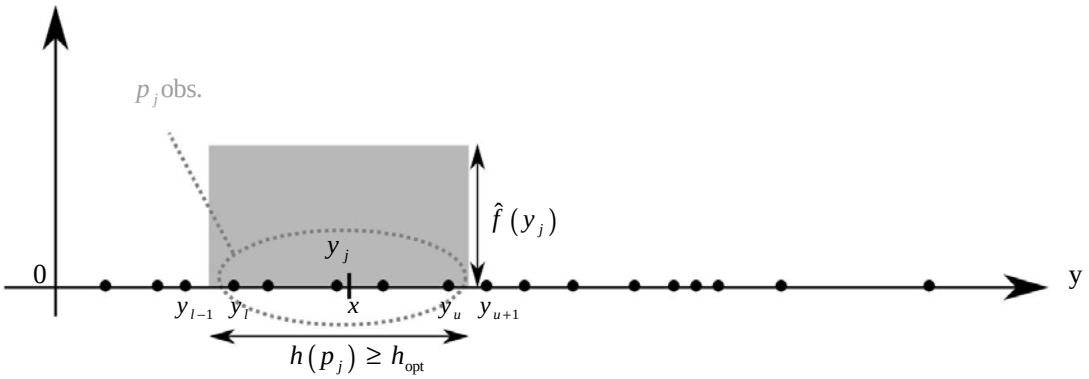

Figure 3.1 Largeur de la fenêtre $h(p_j)$

1. La largeur initiale de la fenêtre autour du point y_j, avec $p_j = p$, est définie par :

$$h(p_j) = \frac{y_u + y_{u+1}}{2} - \frac{y_\ell + y_{\ell-1}}{2}; \quad u = \begin{cases} j + p_j/2 - 1 & \text{si } p_j \text{ est pair} \\ j + \lfloor p_j/2 \rfloor & \text{si } p_j \text{ est impair} \end{cases}$$
$$\ell = j - \lfloor p_j/2 \rfloor.$$

2. Si la largeur de fenêtre $h(p_j)$ ainsi obtenue est inférieure à h_{opt}, on incrémente les deux bornes :
   borne supérieure : $u \to u + 1$, tant que $u < n$,
   borne inférieure : $l \to l - 1$, tant que $l > 1$,
   ce qui implique $p_j \to p_j + 2$, sauf si $u = n$ ou $l = 1$, on n'a alors plus le même nombre de points à gauche et à droite de y_j.

3. Répéter 2 jusqu'à ce que $h(p_j) \geq h_{\text{opt}}$.

4. La densité estimée en x est alors donnée par

$$\hat{f}(x) = \hat{f}(y_j) = \begin{cases} \dfrac{p_j}{nh(p_j)} & \text{sans pondération,} \\ \dfrac{\sum_{p_j \text{ plus proches de } y_j} w_j^{\text{std}}}{nh(p_j)} & \text{avec pondération,} \end{cases}$$

   avec les poids standardisés $w_k^{\text{std}} = w_k / \bar{w}$, $k = 1, \ldots, n$.

Le nombre d'observations p_j prises en compte pour le calcul peut varier et dépend de la courbure locale de la fonction de répartition empirique. La condition $h(p_j) \geq h_{opt}$ garantit une largeur de fenêtre minimale aux endroits où beaucoup d'observations seraient concentrées sur un petit intervalle. On rend la procédure encore plus solide en combinant cette approche avec la précédente, c'est-à-dire en estimant la densité du logarithme de la variable divisée par sa valeur (non logarithmisée) :

$$\hat{f}_3(x) = \frac{\hat{f}(\log(x+a))}{x+a}. \tag{3.4}$$

3.3 Robustesse de la linéarisée

Comme mentionné plus haut, dans le cas de la médiane ou pour les autres quantiles, Croux (1998) relève que la fonction d'influence empirique ou linéarisée estimée à partir de l'échantillon n'est pas aussi robuste qu'il n'y paraît, même si l'on connaît la fonction de densité. Nous avons vérifié cela pour les données SILC utilisées dans les simulations modélisées avec une loi Bêta Généralisée de seconde espèce (GB2) grâce à la fonction `profml.gb2` de R (Graf et Nedyalkova 2011). Sur de petits échantillons $(n \leq 100)$, le biais potentiel de la linéarisée engendré par un trop grand nombre de valeurs extrêmes peut aussi biaiser l'estimation de la variance calculée à partir de cette dernière. Pour de plus grands échantillons $(n \geq 1\,000)$, un biais relatif maximal dans la variance estimée à l'aide de la linéarisée empirique *vs.* théorique peut atteindre jusqu'à 5 %. Il est cependant en-dessous du pourcent en valeur absolue dans les trois quarts des cas.

4 Résultats

Des simulations ont été menées sur trois jeux de données réelles afin de comparer et évaluer les différentes méthodes d'estimation de la fonction de densité : $\hat{f}_1(x)$ voir (3.1), $\hat{f}_2(x)$ voir (3.2) et $\hat{f}_3(x)$ voir (3.4). Ces dernières sont nécessaires à l'estimation de la variance de certains indices de pauvreté et d'inégalité.

1. Le premier jeu de données correspond aux revenus équivalents des ménages de l'enquête SILC suisse menée par l'Office Fédéral Suisse de la Statistique en 2009. On y dispose de 17 534 individus avec un revenu équivalent non nul.
2. Le deuxième jeu de données est aussi issu de l'enquête SILC 2009 mais se restreint aux personnes salariées. Il s'agit des salaires du registre de la Centrale de Compensation qui a été couplé avec les répondants de l'enquête. Nous n'avons donc pas de problème de non-réponse et disposons de 7 922 individus avec un salaire non nul.
3. Le troisième fichier de test, nommé *Ilocos*, est celui à disposition dans le package `ineq` de R (Zeileis 2012). Il contient 632 observations. Il s'agit du revenu des ménages d'une des 16 régions des Philippines appelée Ilocos. Les données sont issues de deux enquêtes de l'Institut National de Statistiques des Philippines en 1997 et 1998.

Les trois jeux de données ont un coefficient d'asymétrie positif, ce qui est typique des distributions de revenus. Chaque jeu de données étant considéré comme une population, on a sélectionné dans un premier temps 10 000 échantillons aléatoires simples sans remise de différentes tailles. Les valeurs des différents indices sont calculées pour chacun des échantillons, ce qui nous procure une estimation Monte Carlo de leur variance, $\mathrm{var}_{\mathrm{sim}}(\hat{\theta})$, pour un indice de pauvreté ou d'inégalité θ. L'estimateur par linéarisation de la variance est noté $\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta})$ et est calculé en utilisant la linéarisée $\hat{z}^{\hat{\theta}}$ estimée pour chaque échantillon :

$$\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta}) = \frac{N(N-n)}{n}\mathrm{var}\left(\hat{z}_S^{\hat{\theta}}\right),$$

où n est la taille d'échantillon utilisée pour les simulations et

$$\mathrm{var}\left(\hat{z}_S^{\hat{\theta}}\right) = \frac{1}{n-1}\sum_{k\in S}\left(\hat{z}_{S,k}^{\hat{\theta}} - \bar{z}_S^{\hat{\theta}}\right)$$

avec $\bar{z}_S^{\hat{\theta}} = n^{-1}\sum_S \hat{z}_{S,k}^{\hat{\theta}}$, voir (2.1).

La qualité de l'estimateur de variance par linéarisation est évaluée en comparant la valeur Monte Carlo espérée de la variance estimée par linéarisation, notée $E_{\mathrm{sim}}\left[\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta})\right]$, à la « vraie » variance Monte Carlo $\mathrm{var}_{\mathrm{sim}}(\hat{\theta})$ en terme de biais relatif :

$$\mathrm{RB}\left[\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta})\right] = \frac{E_{\mathrm{sim}}\left[\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta})\right] - \mathrm{var}_{\mathrm{sim}}(\hat{\theta})}{\mathrm{var}_{\mathrm{sim}}(\hat{\theta})}. \tag{4.1}$$

Pour le deuxième jeu de données (SILC 2009, revenu des personnes salariées) nous avons, dans un deuxième temps, également sélectionné 10 000 échantillons aléatoires sans remise selon un plan stratifié, puis calé les poids de sondage de manière à respecter huit marges sociodémographiques de totaux connus sur la population de 7 922 individus. Les cinq strates utilisées correspondent à des tranches d'âge des personnes salariées (voir tableau 4.1).

Les huit cellules de calage ont été obtenues par les croisements des trois variables dichotomiques suivantes (variables auxiliaires du calage) :

1. MARIÉ indiquant si la personne est mariée ou pas,
2. CHEF indiquant si la personne a une fonction de chef ou pas dans son emploi,
3. HOMME renseignant sur le sexe de l'individu.

Les totaux sur la population de 7 922 individus pour ces cellules de calage sont indiqués dans le tableau 4.2.

Tableau 4.1
Définition des strates utilisées dans les simulations sur les données SILC 2009 et trois tailles d'échantillons (revenu des personnes salariées, $N = 7\,922$)

strate h	description	N_h	%	n_h		
1	personnes de moins de 25 ans	1 187	15,0	75	112	150
2	personnes de 26 à 35 ans	1 359	17,2	86	129	171
3	personnes de 36 à 45 ans	2 137	27,0	135	202	270
4	personnes de 46 à 55 ans	1 864	23,5	117	177	235
5	personnes de plus de 55 ans	1 375	17,4	87	130	174
	TOTAL	7 922	100,0	500	750	1 000

Tableau 4.2
Marges de calage utilisées dans les simulations sur les données SILC 2009 (revenu des personnes salariées, $N = 7\,922$)

marge	MARIÉ	CHEF	HOMME	Tot. pop.	%
1	0	0	0	1 487	18,8
2	0	0	1	1 208	15,2
3	0	1	0	323	4,1
4	0	1	1	457	5,8
5	1	0	0	1 759	22,2
6	1	0	1	1 278	16,1
7	1	1	0	328	4,1
8	1	1	1	1 082	13,7
			TOTAL	7 922	100,0

Pour chaque échantillon stratifié, un calage sur marge (méthode linéaire) a été effectué de manière à ce que les sommes des poids après calage respectent les huit marges citées ci-dessus. Les estimations ponctuelles des indicateurs ainsi que leur linéarisée ont été calculées pour chaque échantillon à l'aide des poids calés.

Pour estimer la variance, on a appliqué la méthode développée par Deville (2000). Celle-ci consiste à linéariser aussi par rapport au calage en calculant les résidus $e^{\hat{\theta}}$ de la régression (pondérée par les poids de sondage) des linérarisées des indicateurs sur les variables auxiliaires du calage. La variance du total des résidus ainsi calculés sous plan aléatoire stratifié sans remise est alors un estimateur de la variance de l'indicateur estimé, ce qui est la grandeur qui nous intéresse :

$$\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta}) = \sum_{h=1}^{H} \frac{N_h}{n_h}(N_h - n_h)\, s^2_{e_h^{\hat{\theta}}} \tag{4.2}$$

avec

$$s^2_{e_h^{\hat{\theta}}} = \frac{1}{n_h - 1} \sum_{k \in S_h} \left(e_k^{\hat{\theta}} - \bar{e}^{\hat{\theta}}\right)^2$$

La qualité de l'estimateur de variance par linéarisation est évaluée de manière analogue au cas du sondage aléatoire simple, voir (4.1).

Les tableaux 4.3, 4.4 et 4.5 présentent les biais relatifs de la variance obtenus pour les trois jeux de données utilisés et décrits plus haut dans le cas du sondage aléatoire simple. Le tableau 4.6 présente les biais relatifs de la variance obtenus dans le cas du sondage aléatoire stratifié avec poids calés. Dans les parties supérieures des tableaux, on trouve les valeurs pour l'indice de Gini et le QSR. Ces derniers ne nécessitent pas d'estimer la fonction de densité des revenus. L'estimation de leur variance fonctionne bien. On note un problème de sous-estimation de la variance pour l'indice de Gini dans le cas stratifié avec calage (tableau 4.6).

Dans le cas du premier jeu de données considéré, le tableau 4.3 ne révèle pas de très grandes différences si ce n'est que l'estimation de la densité de revenus par $\hat{f}_3(x)$ mène à des résultats plus conservateurs. En effet, le biais relatif reste du même ordre de grandeur mais positif, alors qu'il est négatif pour les deux autres manières d'estimer la densité. Pour le deuxième jeu de donnée, le tableau 4.4 montre qu'il est essentiel de passer par le logarithme ou d'utiliser la méthode par plus proches voisins avec

largeur de bande minimale. Avec cette dernière méthode, tous les biais relatifs tombent en-dessous de 10 % dès que les tailles d'échantillons sont suffisamment grandes (voir la dernière colonne du tableau). Les simulations sur les mêmes données avec un plan de sondage stratifié et un calage renforcent et confirment ces résultats (voir tableau 4.6). Pour le troisième jeu de données, le tableau 4.5 indique les mêmes tendances, bien que les résultats soient plus instables en raison des petites tailles d'échantillon et de population. Cette constatation n'est pas surprenante, le nombre minimal de voisins à prendre en compte ayant été fixé à 30. En l'occurrence, pour le jeu de données Ilocos, des simulations avec un p plus petit fixé à 10 ne changent rien au final, car la condition $h(p_j) \geq h_{\text{opt}}$ le remonte systématiquement au-dessus de 30.

De plus, de manière générale, on constate que plus il faut avoir recours à l'estimation de la densité par noyau gaussien - $\hat{f}_1(x)$ - plus l'erreur augmente. En effet, les biais relatifs de la variance pour la médiane des pauvres et le RMPG sont presque systématiquement plus grands en valeur absolue que ceux pour les autres indicateurs. Pour le RMPG, l'erreur peut se compenser (cas du tableau 4.3) s'il y a suffisamment d'observations car l'estimation de la densité intervient tant au numérateur qu'au dénominateur.

Tableau 4.3
Biais relatifs (4.1) de la variance obtenus avec 10 000 échantillons aléatoires simples sans remise sur les données SILC 2009 (revenu équivalent des ménages, N = 17 534)

	Tailles d'échantillons (taux de sondage)								
Indice	**n = 500 (2,9 %)**			**n = 750 (4,3 %)**			**n = 1 000 (5,7 %)**		
GINI	-0,02			-0,02			-0,02		
QSR	0,01			0,00			0,00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0,08	-0,06	0,04	-0,09	-0,07	0,03	-0,09	-0,07	0,04
ARPR	-0,05	-0,01	-0,00	-0,09	-0,06	-0,05	-0,08	-0,05	-0,03
RMPG	-0,09	-0,07	*0,15*	*-0,10*	-0,07	*0,12*	-0,09	-0,06	*0,14*
MEDP	*-0,16*	*-0,12*	0,09	*-0,19*	*-0,13*	0,05	*-0,18*	*-0,11*	0,07
MED	-0,08	-0,06	0,05	-0,08	-0,06	0,04	-0,08	-0,06	0,04

Tableau 4.4
Biais relatifs (4.1) de la variance obtenus avec 10 000 échantillons aléatoires simple sans remise sur les données SILC 2009 (revenu des personnes salariées, N = 7 922)

	Tailles d'échantillons (taux de sondage)								
Indice	**n = 500 (6,3 %)**			**n = 750 (9,5 %)**			**n = 1 000 (12,6 %)**		
GINI	-0,03			-0,03			-0,02		
QSR	-0,00			0,00			0,00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	0,07	0,05	*0,13*	0,06	0,04	*0,10*	0,06	0,03	0,08
ARPR	-0,05	-0,04	-0,02	-0,05	-0,04	-0,01	-0,06	-0,05	-0,02
RMPG	*0,61*	*0,12*	*0,15*	*0,60*	*0,11*	0,08	*0,59*	0,09	0,05
MEDP	*0,73*	*0,17*	*0,18*	*0,72*	*0,16*	*0,10*	*0,72*	*0,15*	0,07
MED	0,07	0,04	*0,13*	0,06	0,04	*0,10*	0,05	0,03	0,07

Tableau 4.5
Biais relatifs (4.1) de la variance obtenus avec 10 000 échantillons aléatoires simple sans remise sur les données Ilocos (revenu des ménages, $N = 632$)

Indice	Tailles d'échantillons (taux de sondage) $n = 50$ (7,9 %)			$n = 63$ (10,0 %)		
GINI	*-0,16*			*-0,13*		
QSR	0,00			0,00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0,05	-0,06	-0,01	-0,03	-0,03	-0,01
ARPR	*-0,31*	-0,01	*-0,12*	*-0,33*	-0,03	*-0,18*
RMPG	*1,55*	*0,83*	*0,26*	*1,54*	*0,16*	*0,39*
MEDP	*1,02*	*0,28*	*-0,26*	*1,05*	0,07	*-0,11*
MED	0,04	0,03	0,08	0,07	0,07	0,09

Tableau 4.6
Biais relatifs (4.1) de la variance obtenus avec 10 000 échantillons aléatoires stratifiés sans remise, avec poids calés sur huit marges sociodémographiques, sur les données SILC 2009 (revenu des personnes salariées, $N = 7\ 922$)

Indice	Tailles d'échantillons (taux de sondage) $n = 500$ (6,3 %)			$n = 750$ (9,5 %)			$n = 1\ 000$ (12,6 %)		
GINI	*-0,21*			*-0,20*			*-0,20*		
QSR	-0,06			-0,06			-0,07		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0,07	-0,09	-0,01	-0,08	*-0,10*	-0,04	-0,09	*-0,11*	-0,06
ARPR	*-0,10*	*-0,10*	-0,08	-0,07	-0,06	-0,05	-0,06	-0,06	-0,05
RMPG	*0,63*	*0,13*	*0,13*	*0,61*	*0,11*	0,08	*0,59*	*0,10*	0,04
MEDP	*0,71*	*0,16*	*0,15*	*0,68*	*0,13*	0,09	*0,66*	*0,12*	0,04
MED	-0,07	-0,09	-0,01	-0,08	*-0,10*	-0,04	-0,08	*-0,11*	-0,06

En résumé, on voit que l'on peut tout autant surestimer la variance $\left(\mathrm{RB}\left[\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta})\right] > 0\right)$ que la sous-estimer $\left(\mathrm{RB}\left[\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta})\right] < 0\right)$ selon l'indice et le jeu de données considérés. Le fait de passer au logarithme $\left(\hat{f}_2(x)\right)$ améliore très significativement la situation. La méthode par « plus proches voisins » $\left(\hat{f}_3(x)\right)$ fait disparaître tout problème si on a suffisamment de données (cas des tableaux 4.3, 4.4 et 4.6). Cette dernière a un peu plus de difficulté si les échantillons sont petits (cas du tableau 4.5). Les variations illogiques et les biais qui subsistent dans les tableaux peuvent aussi être dus au manque de robustesse des linéarisées sur certains échantillons comme mentionné à la section 3.3.

5 Conclusions

Dans plusieurs pays, des enquêtes par échantillonnage de niveau national publient les valeurs extrapolées des indices de Laeken (Eurostat 2005) car elles constituent des indicateurs importants permettant d'aiguiller les preneurs de décisions aux niveaux politique et social. Il est dès lors essentiel de pouvoir quantifier la précision de ces mesures, ce qui soulève la question de la pertinence de l'estimation

de la précision dont on dispose. Ce travail montre qu'une amélioration substantielle peut être faite dans l'estimation de la précision d'indices de pauvreté et d'inégalité qui nécessitent le recours à une estimation (locale) de la densité des revenus ou de la variable monétaire considérée.

Les simulations menées montrent que la méthode d'estimation de la densité par noyau gaussien actuellement mise en œuvre dans la plupart des cas n'est pas à recommander sans au moins passer par le logarithme comme proposé dans la section 3.1, faute de quoi le biais de la variance estimée peut devenir très important. La méthode par « plus proches voisins » (section 3.2) qui impose aussi une largeur de bande minimale peut donner des résultats encore meilleurs, surtout si on a des agglomérats d'observations en certaines valeurs dans les données étudiées. Elle nécessite cependant de choisir un nombre minimal p de voisins en fonction des données sur lesquelles on travaille. Si l'on a peu d'observations à disposition, il est préférable d'en rester à la variante du logarithme. Dans tous les cas, nous espérons que ce travail contribuera à une prise de conscience du fait qu'il faut être méticuleux lors de l'implémentation du calcul des linéarisées pour tous les indicateurs faisant intervenir des quantiles.

6 Remerciements

Ce travail a été réalisé dans le cadre d'une convention de collaboration entre l'Institut de Statistique de l'Université de Neuchâtel et l'Office Fédéral de la Statistique (OFS). Nous tenons également à remercier en particulier la section *Revenus, consommation et conditions de vie* de l'OFS pour nous avoir mis à disposition les données de la partie suisse de l'enquête européenne sur le revenu et les conditions de vie. Merci aussi à Matti Langel et Anne Massiani pour leur appui au cours de nos investigations.

Bibliographie

Antal, E., Langel, M. et Tilllé, Y. (2011). Variance estimation of inequality indices in complex sampling designs. *Proceedings 58th World Statistical Congress*, Dublin.

Ardilly, P., et Osier, G. (2007). Cross-sectional variance estimation for the French "Labour Force Survey". *Survey Research Methods*, 1, 75-83.

Beaumont, J.-F., Haziza, D. et Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 555-569.

Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

Croux, C. (1998). Limit behaviour of the empirical influence function of the median. *Statistics & Probability Letters*, 37, 331-340.

Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 1, 17-27.

Deville, J.-C. (2000). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques de résidus. *Techniques d'enquêtes*, 25, 2, 219-230.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Eurostat (2004a). Common cross-sectional eu indicators based on eu-silc; the gender pay gap. Documents de travail et d'études de l'Office for Official Publications of the European Communities, Luxembourg. EU-SILC 131-rev/04.

Eurostat (2004b). Theoretical study of the gini index. Documents de travail et d'études de l'Office for Official Publications of the European Communities, Luxembourg. EU-SILC 131-A/04.

Eurostat (2005). The continuity of indicators during the transition between ECHP and EU-SILC. Documents de travail et d'études de l'Office for Official Publications of the European Communities, Luxembourg.

Goga, C., Deville, J.-C. et Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample survey data. *Biometrika*, 96, 691-709.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. Dans *Compositional Data Analysis, Theory and Applications*, (Éds., V. Pawlosky-Glahn et A. Buccianti), Oxford : Wiley, chapitre 9, 114-127.

Graf, M. (2013). A simplified approach to linerarization variance for surveys. *University of Neuchâtel*, document de travail.

Graf, M., et Nedyalkova, D. (2011). *GB2: Generalized Beta Distribution of the Second Kind: properties, likelihood, estimation.* R package version 1.0.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383-393.

Hulliger, B. (1999). Simple and robust estimators for sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 54-63.

Hyndman, R.J., et Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician*, 50, 361-365.

Kovačević, M.S., et Binder, D.A. (1997). Variance estimation for measures of income inequality and polarization - The estimating equations approach. *Journal of Official Statistics*, 13, 41-58.

Langel, M., et Tillé, Y. (2011). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference*, 141, 2976-2985.

Langel, M., et Tillé, Y. (2012). Variance estimation of the Gini index: Revisiting a result several times published. Dans *Press in Journal of the Royal Statistical Society - Series A.*

Lorenz, M.O. (1905). Methods of measuring the concentration of wealth. *American Statistical Association*, 9, 209-219.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3, 167-195.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, Londres.

Verma, V., et Betti, G. (2011). Taylor linearization sampling errors and design effects for poverty measures and other complex statistics. *Journal of Applied Statistics*, 38, 1549-1576.

Zeileis, A. (2012). *ineq: Measuring Inequality, Concentration, and Poverty*. R package version 0.2-10.

Techniques d'enquête, juin 2014
Vol. 40, N° 1, pp. 89-114
Statistique Canada, N° 12-001-X au catalogue

Propriétés théoriques et empiriques d'estimateurs par la régression fondés sur un test de décision assistés par modèle

Jun Shao, Eric Slud, Yang Cheng, Sheng Wang et Carma Hogue[1]

Résumé

En 2009, deux enquêtes importantes réalisées par la division des administrations publiques du U.S. Census Bureau ont été remaniées afin de réduire la taille de l'échantillon, d'économiser des ressources et d'améliorer la précision des estimations (Cheng, Corcoran, Barth et Hogue 2009). Sous le nouveau plan de sondage, chaque strate habituelle, définie par l'État et le type d'administration publique, qui contient un nombre suffisant d'unités (administrations publiques) est divisée en deux sous-strates en fonction de la masse salariale totale de chaque unité afin de tirer un plus petit échantillon de la sous-strate des unités de petite taille. L'approche assistée par modèle est adoptée pour estimer les totaux de population. Des estimateurs par la régression utilisant des variables auxiliaires sont obtenus soit pour chaque sous-strate ainsi créée soit pour la strate originale en regroupant des deux sous-strates. Cheng, Slud et Hogue (2010) ont proposé une méthode fondée sur un test de décision qui consiste à appliquer un test d'hypothèse pour décider quel estimateur par la régression sera utilisé pour chaque strate originale. La convergence et la normalité asymptotique de ces estimateurs assistés par modèle sont établies ici sous un cadre asymptotique fondé sur le plan de sondage ou assisté par modèle. Nos résultats asymptotiques suggèrent aussi deux types d'estimateurs de variance convergents, l'un obtenu par substitution des quantités inconnues dans les variances asymptotiques et l'autre en appliquant la méthode du bootstrap. La performance de tous les estimateurs des totaux et des estimateurs de leur variance est examinée au moyen d'études empiriques. L'Annual Survey of Public Employment and Payroll (ASPEP) des États-Unis est utilisé pour motiver et illustrer notre étude.

Mots-clés : Normalité asymptotique; bootstrap; estimateur fondé sur un test de décision; probabilité proportionnelle à la taille; stratification; estimation de la variance.

1 Introduction

L'Annual Survey of Public Employment and Payroll (ASPEP) des États-Unis fournit des estimations courantes de l'emploi et de la rémunération à temps plein et à temps partiel dans les administrations publiques d'État et locales par fonction (par exemple, enseignement primaire et secondaire, enseignement supérieur, services de police, services de protection contre l'incendie, administration financière, services judiciaires et juridiques, *etc.*). Cette enquête a pour champ d'observation les administrations publiques d'État et locales (89 526 selon le Census of Governments de 2007), qui englobent les comtés, les villes, les cantons, les administrations appelées « districts spéciaux » et les districts scolaires. L'ASPEP, qui est la seule source de données sur l'emploi dans le secteur public par fonction administrative et catégorie d'emploi, fournit des données sur le nombre et la rémunération des employés à temps plein et à temps partiel, ainsi que le nombre d'heures travaillées par les employés à temps partiel. Habituellement, la collecte des données débute en mars et se poursuit pendant environ sept mois, en prenant la période de paye incluant le 12 mars comme période de référence.

Soit U la population finie de N unités subdivisée en H strates, $U_1, \ldots, U_H$, où U_h contient N_h unités et $N_1 + \cdots + N_H = N$. Le plan de sondage habituel de l'ASPEP est un plan avec probabilité

1. Jun Shao, Statistics Department University of Wisconsin, Madison WI , Courriel: shao@stat.wisc.edu; Eric Slud, Center for Statistical Research and Methodology, US Census Bureau, Washington DC and Mathematics Department, University of Maryland, College Park, MD, Courriel: eric.v.slud@census.gov; Yang Cheng, Demographic Statistical Methods Division, US Census Bureau, Washington DC, Courriel: yang.cheng@census.gov; Sheng Wang, Mathematica Policy Research, Princeton NJ, Courriel : swang@mathematica-mpr.com; et Carma Hogue, Governments Division, US Census Bureau, Washington DC, Courriel: carma.ray.hogue@census.gov.

proportionnelle à la taille (PPT), où les strates sont construites en se basant sur l'État et le type d'administration publique, à savoir le comté, le sous-comté (grande ou petite ville), le district spécial ou le district scolaire. La taille de chaque unité (administration publique) est mesurée par la masse salariale totale, et l'échantillonnage est effectué indépendamment dans les diverses strates. En 2009, on a élaboré un plan d'échantillonnage modifié, qui comprend la division de certaines strates U_h en deux sous-strates, U_{h1} et U_{h2} contenant N_{h1} et N_{h2} unités, respectivement, où U_{h1} contient les unités de petite taille (Cheng et coll. 2009). L'idée était d'économiser des ressources et de réduire le fardeau de réponse en sélectionnant dans U_{h1} un échantillon plus petit sous le plan modifié que sous le plan habituel. Soit S_{hj} un échantillon PPT de taille n_{hj} provenant de U_{hj}, $j = 1, 2$, $n_{h1} + n_{h2} = n_h$. Notons que n_{h1} peut encore être plus grand que n_{h2}, parce que N_{h1} est habituellement beaucoup plus grand que N_{h2}.

Pour l'unité $i \in U$, soit y_i une variable étudiée clé (p. ex., l'emploi à temps plein, la rémunération à temps plein, l'emploi à temps partiel, la rémunération à temps partiel, les heures travaillées à temps partiel), x_i une variable auxiliaire, disons la même variable que y_i provenant du recensement le plus récent, et soit z_i la covariable utilisée comme variable de taille dans l'échantillonnage PPT. Les valeurs des covariables x_i et z_i sont observées pour tout $i \in U$, tandis que y_i est observée uniquement pour chaque unité *i* échantillonnée.

L'estimateur de Horvitz-Thompson du total inconnu $Y = \sum_{i \in U} y_i$ est

$$\hat{Y}_{\mathrm{HT}} = \sum_h \sum_j \sum_{i \in S_{hj}} y_i / \pi_i \,, \tag{1.1}$$

où π_i est la probabilité d'inclusion d'ordre un de l'unité i dans S_{hj}, une fonction connue des z_i. Pour utiliser la variable auxiliaire x_i et accroître la précision de l'estimation de Y, l'approche assistée par modèle (Särndal, Swensson et Wretman 1992) a été adoptée. L'application de la régression dans chaque échantillon S_{hj} conduit à l'estimateur par la régression de Y de la forme

$$\hat{Y}_{\mathrm{reg},2} = \sum_h \sum_j \left[\frac{N_{hj} \hat{Y}_{hj}}{\hat{N}_{hj}} + \hat{\beta}_{hj} \left(X_{hj} - \frac{N_{hj} \hat{X}_{hj}}{\hat{N}_{hj}} \right) \right], \tag{1.2}$$

où $X_{hj} = \sum_{i \in U_{hj}} x_i$, $\hat{Y}_{hj} = \sum_{i \in S_{hj}} y_i / \pi_i$, $\hat{X}_{hj} = \sum_{i \in S_{hj}} x_i / \pi_i$, $\hat{N}_{hj} = \sum_{i \in S_{hj}} 1/\pi_i$, et

$$\hat{\beta}_{hj} = \frac{\sum_{i \in S_{hj}} \left(x_i - \hat{X}_{hj} / \hat{N}_{hj} \right) y_i / \pi_i}{\sum_{i \in S_{hj}} \left(x_i - \hat{X}_{hj} / \hat{N}_{hj} \right)^2 / \pi_i} .$$

Autrement, la combinaison des deux sous-strates S_{h1} et S_{h2} donne l'estimateur par la régression suivant. (Un examinateur fait remarquer correctement que $\hat{Y}_{\mathrm{reg},1}$ dans (1.3) n'est pas l'estimateur groupé que l'on utiliserait si les droites de régression dans la strate h étaient combinées mais que les deux sous-strates ne l'étaient pas; cependant, il *est* l'estimateur naturel lorsque non seulement les droites de régression, mais aussi les sous-strates sont combinées.)

$$\hat{Y}_{\text{reg},1} = \sum_h \left[\frac{N_h \hat{Y}_h}{\hat{N}_h} + \hat{\beta}_h \left(X_h - \frac{N_h \hat{X}_h}{\hat{N}_h} \right) \right], \tag{1.3}$$

où $\hat{Y}_h = \sum_j \hat{Y}_{hj}$, $\hat{X}_h = \sum_j \hat{X}_{hj}$, $\hat{N}_h = \sum_j \hat{N}_{hj}$, et

$$\hat{\beta}_h = \frac{\sum_j \sum_{i \in S_{hj}} \left(x_i - \hat{X}_h / \hat{N}_h \right) y_i / \pi_i}{\sum_j \sum_{i \in S_{hj}} \left(x_i - \hat{X}_h / \hat{N}_h \right)^2 / \pi_i}.$$

Puisque $\hat{Y}_{\text{reg},1}$ ainsi que $\hat{Y}_{\text{reg},2}$ sont des estimateurs assistés par modèle, ils sont convergents sous échantillonnage répété, que le modèle de régression soit ou non vérifié. Si les droites de régression par les moindres carrés dans les deux sous-strates U_{hj} sont les mêmes, $\hat{Y}_{\text{reg},1}$ peut être plus efficace que $\hat{Y}_{\text{reg},2}$. Par ailleurs, si les droites de régression sont différentes, $\hat{Y}_{\text{reg},2}$ peut être plus efficace que $\hat{Y}_{\text{reg},1}$.

Cheng et coll. (2010) ont proposé une méthode fondée sur un test de décision qui consiste à appliquer un test d'hypothèse pour décider s'il faut combiner S_{h1} et S_{h2}. À l'intérieur de la strate h, on teste l'hypothèse d'égalité des pentes des droites de régression dans U_{h1} et U_{h2}. Soit

$$\hat{\alpha}_{hj} = \frac{\hat{Y}_{hj} - \hat{\beta}_{hj} \hat{X}_{hj}}{\hat{N}_{hj}}, \hat{\sigma}^2_{xe,hj} = \frac{n_{hj}}{\hat{N}^2_{hj}} \sum_{i \in S_{hj}} \left(x_i - \frac{\hat{X}_{hj}}{\hat{N}_{hj}} \right)^2 \frac{\left(y_i - \hat{\alpha}_{hj} - \hat{\beta}_{hj} x_i \right)^2}{\pi_i^2},$$

$$\hat{\sigma}^2_{xhj} = \sum_{i \in S_{hj}} \frac{\left(x_i - \hat{X}_{hj} / \hat{N}_{hj} \right)^2}{\pi_i \hat{N}_{hj}}, t_h = \sqrt{n_h - 4} \left(\hat{\beta}_{h1} - \hat{\beta}_{h2} \right) \Big/ \sqrt{n_h \sum_{j=1}^{2} \frac{\hat{\sigma}^2_{xe,hj}}{n_{hj} \hat{\sigma}^4_{xhj}}}.$$

Si $|t_h| > t_{1-\tau/2, n_h - 4}$, où $t_{1-\tau/2,\nu}$ est le $(1 - \tau/2)^{\text{e}}$ quantile de la distribution t avec ν degrés de liberté, alors nous rejetons l'hypothèse d'une pente commune et nous utilisons $\hat{\beta}_{hj}$ (et fixons $\zeta_h = 1$). Ici, τ est un seuil de signification nominal fixé par défaut à 0,05, mais nous considérerons d'autres choix de la valeur de τ à la section consacrée aux simulations. La définition de la statistique de test faisant intervenir $n_h - 4$ degrés de liberté est un choix légèrement artificiel conçu afin de rendre les probabilités de rejet d'un échantillon modéré plus proches de la valeur nominale, mais la théorie asymptotique en grand échantillon justifiant ce test est donnée à la partie (c) du théorème 1. Si $|t_h| \le t_{1-\tau/2, n_h - 4}$, alors nous acceptons l'hypothèse d'une pente commune, nous combinons les sous-strates S_{h1} et S_{h2}, et nous utilisons $\hat{\beta}_h$ (en fixant $\zeta_h = 0$). Les tests sont effectués de manière indépendante dans les diverses strates $h = 1, \ldots, H$. L'estimateur de Y fondé sur le test de décision est alors

$$\hat{Y}_{\text{dec}} = \sum_h \sum_j \zeta_h \left[\frac{N_{hj} \hat{Y}_{hj}}{\hat{N}_{hj}} + \hat{\beta}_{hj} \left(X_{hj} - \frac{N_{hj} \hat{X}_{hj}}{\hat{N}_{hj}} \right) \right] + \sum_h (1 - \zeta_h) \left[\frac{N_h \hat{Y}_h}{\hat{N}_h} + \hat{\beta}_h \left(X_h - \frac{N_h \hat{X}_h}{\hat{N}_h} \right) \right]. \tag{1.4}$$

Puisque les deux droites de régression ayant une pente commune peuvent avoir des ordonnées à l'origine différentes, on pourrait tester une hypothèse supplémentaire concernant les ordonnées à l'origine pour décider s'il faut combiner les deux sous-strates. Cependant, des points de population (x_i, y_i) se

trouvant sur deux droites de régression de sous-strate parallèles, mais non identiques seraient discontinus autour du seuil entre les deux sous-strates U_{h1} et U_{h2}, ce qui ne semble ne se produire que rarement dans les situations pratiques. Par exemple, dans l'ASPEP, Cheng et coll. (2010) ont étudié les pentes et les ordonnées à l'origine de sous-strates dans les ensembles de donnés des recensements des administrations publiques de 2002 et de 2007, et ont constaté que l'hypothèse d'une ordonnée à l'origine commune ne pouvait jamais être rejetée lorsque l'hypothèse d'une pente commune ne pouvait pas l'être. Donc, l'estimateur fondé sur un test de décision donné dans (1.4) dépend uniquement du test de l'hypothèse d'égalité des pentes des droites de régression des sous-strates.

Les estimateurs à deux degrés étudiés ici sont des cas particuliers de procédures nommées antérieurement estimateurs après un test préliminaire (*preliminary test estimators*). Il existe une littérature abondante traitant de l'utilisation de ce genre de procédures dans les enquêtes, y compris une bibliographie de Bancroft et Han (1977), un livre publié par Saleh (2006) et un traitement proposé par Fuller (2009, section 6.7). Une idée de Saleh (2006) consiste à estimer les coefficients par une combinaison convexe des coefficients estimés à partir des strates distinctes en faisant dépendre les proportions d'une statistique de test. Les estimateurs lissés de ce genre pourraient être plus efficaces que nos procédures fondées sur un test de décision. Si les ordonnées à l'origine et les pentes propres aux strates étaient considérées comme aléatoires, on pourrait aussi essayer d'appliquer à l'estimation une approche bayésienne empirique fondée sur un modèle.

Les estimateurs fondés sur un test de décision (1.4) sont nouveaux, parce qu'ils sont assistés par modèle et convergents sous le plan dans le contexte des sondages, et utilisent explicitement les tailles de population de sous-strate connues. Dans un esprit à peu près semblable, Rao et Ramachandran (1974) avaient effectué antérieurement une comparaison exacte des estimateurs par le ratio distincts et combinés sous un modèle de ratio similaire au modèle de régression considéré dans le présent article.

L'objectif de l'article est d'illustrer certaines propriétés asymptotiques et empiriques des estimateurs de Y décrits plus haut et des estimateurs de leur variance. La convergence et la normalité asymptotique de $\hat{Y}_{\mathrm{reg},1}$, $\hat{Y}_{\mathrm{reg},2}$, et $\hat{Y}_{\mathrm{dec}}$ sont établies à la section 2, dans le contexte de la théorie asymptotique fondée sur le plan de sondage ou assistée par modèle. Bien que les résultats asymptotiques d'ordre un favorisent $\hat{Y}_{\mathrm{reg},2}$, $\hat{Y}_{\mathrm{reg},1}$ pourrait être meilleur quand certaines tailles d'échantillon de sous-strate n_{h2} sont modérées, un effet asymptotique d'ordre deux. L'avantage de l'estimateur fondé sur un test de décision $\hat{Y}_{\mathrm{dec}}$ tient à l'adaptation en vue d'être proche de $\hat{Y}_{\mathrm{reg},1}$ ou de $\hat{Y}_{\mathrm{reg},2}$ selon celui qui est le meilleur. Comme l'indique la discussion du paragraphe (III) de la section 4.4, les simulations montrent que l'avantage de cette adaptabilité est de réduire l'EQM d'une quantité allant jusqu'à quelques pour cent sous des conditions de paramétrisation raisonnables, et de plus grandes quantités sous des conditions plus étranges.

L'estimation de la variance de l'estimateur fondé sur un test de décision est traitée à la section 3. Même si la théorie asymptotique exposée à la section 2 laisse entendre que des estimateurs convergents de variance sont obtenus par substitution des quantités inconnues dans les formules de variance asymptotique, nous étudions aussi les estimateurs bootstrap de la variance proposés dans Cheng et coll. (2010), qui ont généralement de meilleures propriétés en échantillon fini que les estimateurs par substitution. Les résultats empiriques sont présentés à la section 4, les interprétations et les conclusions étant formulées à la sous-section 4.4. Toutes les preuves techniques sont données en annexe.

2 Convergence et normalité asymptotique

Afin d'examiner les propriétés asymptotiques, nous considérons la population U comme l'une d'une série de populations $\{U^{(m)}, m = 1, 2, \ldots\}$, où le nombre d'unités dans $U^{(m)}$ tend vers l'infini quand $m \to \infty$. Nous ne traitons ici que le cas de strates desquelles est tiré un grand échantillon n_h ; autrement dit, nous supposons que, pour chaque strate h, la taille de l'échantillon n_h dépend de m et tend vers l'infini quand $m \to \infty$, mais nous omettons l'indice m pour simplifier la notation. Tous les processus limites sont considérés pour $m \to \infty$. À l'instar d'auteurs tels que Isaki et Fuller (1982) et Deville et Särndal (1992), nous donnons à ces conditions le nom de cadre asymptotique de *superpopulation*. Sous le cadre fondé sur le plan de sondage considéré à la section 2.1, les vecteurs d'attributs dans les populations sous-jacentes ne doivent pas être considérés comme des vecteurs aléatoires. Cependant, sous le cadre assisté par modèle considéré à la section 2.2, des modèles de régression hypothétiques sont associés aux vecteurs d'attributs.

Puisque chaque estimateur est une somme d'estimateurs indépendants construits dans chaque strate, pour simplifier, nous présentons les résultats asymptotiques pour le cas où $H = 1$. Les résultats et les conclusions s'appliquent directement au cas d'une valeur fixe de H et peuvent aussi être étendus à la situation où H tend vers l'infini. (Il est habituel que les grandes enquêtes contiennent de nombreuses strates, quoique dans l'ASPEP, le nombre de strates définies selon le type d'administration publique qui ont été subdivisées en sous-strates était un peu inférieur à 100.) Puisque nous considérons seulement le cas $H = 1$, nous omettons l'indice h désignant la strate à la présente section, p. ex., $n_{hj} = n_j$, $n_h = n$, $N_{hj} = N_j$ et $N_h = N$. En outre, pour $j = 1, 2$, les estimateurs $\hat{\beta}_j$ et $\hat{\beta}$ sont définis par les formules présentées après les équations (1.2) et (1.3) avec l'indice inférieur h supprimé, considérées conjointement avec

$$\hat{\mu}_{xj} = \hat{X}_j \big/ \hat{N}_j, \quad \hat{\alpha}_j = \hat{Y}_j \big/ \hat{N}_j - \hat{\beta}_j \hat{\mu}_{xj}, \quad \hat{\sigma}^2_{xj} = \hat{N}_j^{-1} \sum_{i \in S_j} \pi_i^{-1} \left(x_i - \hat{\mu}_{xj} \right)^2$$

$$\hat{\sigma}^2_{xe,j} = n_j \sum_{i \in S_j} \left(x_i - \hat{\mu}_{xj} \right)^2 \left(y_i - \hat{\alpha}_j - \hat{\beta}_j x_i \right)^2 \Big/ \left(\pi_i^2 \hat{N}_j^2 \right).$$

De surcroît, pour simplifier, nous n'examinons les résultats asymptotiques que sous échantillonnage avec remise. Les résultats peuvent être appliqués au cas de l'échantillonnage sans remise si la fraction d'échantillonnage n/N est négligeable.

2.1 Cadre asymptotique fondé sur le plan de sondage

Premièrement, nous établissons la normalité asymptotique de $\hat{Y}_{\text{reg},1}$ et $\hat{Y}_{\text{reg},2}$ sous échantillonnage répété, c'est-à-dire quand y_i et x_i sont fixes pour $i \in U$, et S_j est un échantillon PPT aléatoire.

Théorème 1 Supposons que S_1 et S_2 sont des échantillons PPT indépendants tirés avec remise de U_1 et U_2, respectivement, où l'unité $i \in U_j$ possède la probabilité $p_{ij} = z_i \big/ \sum_{i \in U_j} z_i > 0$ d'être sélectionnée, et le poids d'échantillonnage $\pi_i^{-1} = 1 \big/ \left(n_j p_{ij} \right)$ pour $j = 1, 2$, et que les quatre conditions qui suivent sont vérifiées, à mesure que l'indice séquentiel de population m tend vers ∞.

(C1) Il existe des constantes φ_j et ω_j telles que $\sqrt{n/n_j} \to \varphi_j$ et $N_j/N \to \omega_j$.

(C2) Pour $j = 1, 2,$ il existe des constantes μ_{yj}, μ_{xj} et β_j telles que

$$\bar{Y}_j = Y_j/N_j = \sum_{i \in U_j} y_i/N_j \to \mu_{yj}, \bar{X}_j = X_j/N_j = \sum_{i \in U_j} x_i/N_j \to \mu_{xj}$$

existent, de même que les limites $N_j^{-1}\sum_{i \in U_j}(x_i - \mu_{xj})^2 \to \sigma_{xj}^2 > 0,$ et en outre,

$$\left(\sqrt{n_j}/N_j\right)\sum_{i \in U_j} x_i\left(y_i - Y_j/N_j - \beta_j\left(x_i - X_j/N_j\right)\right) \to 0 \text{ quand } n, N \to \infty.$$

(C3) Les limites $D_{N_j} = \sum_{i \in U_j} p_{ij} b_{ij} b_{ij}^T / N_j^2 \to D_j$ existent, où pour $i \in U_j$,

$$b_{ij} = \left[1/p_{ij} - N_j, x_i/p_{ij} - X_j, y_i/p_{ij} - Y_j\right]^T,$$

v^T désigne la transposée vectorielle, et D_j est définie positive. La limite $\sigma_{xe,j}^2 = \lim N_j^{-2}\sum_{i \in U_j}(x_i - \mu_{xj})^2 (y_i - \alpha_j - \beta_j x_i)^2 / p_{ij}$ existe aussi, pour $\alpha_j = \mu_{yj} - \beta_j \mu_{xj}$.

(C4) Les éléments de $\Lambda_j = \sum_{i \in U_j} p_{ij} c_{ij} c_{ij}^T / N_j^4$ forment une séquence bornée, où pour $i \in U_j$, $c_{ij} = \left[(1/p_{ij} - N_j)^2, (x_i/p_{ij} - X_j)^2, (y_i/p_{ij} - Y_j)^2\right]^T$.

Alors, quand $m \to \infty,$ les conclusions qui suivent sont vérifiées.

(a) Pour $j = 1, 2,$ $\hat{\mu}_{xj} \to_P \mu_{xj}, \hat{\mu}_{yj} \to_P \mu_{yj}, \hat{\beta}_j \to_P \beta_j, \hat{\alpha}_j \to_P \alpha_j,$ et $\hat{\sigma}_{xj}^2 \to_P \sigma_{xj}^2,$ où $\to_P$ désigne la convergence en probabilité.

(b) L'estimateur pour la strate combinée $\hat{\beta}$ possède l'expression exacte

$$\hat{\beta} = \frac{\sum_{j=1}^2 \hat{\beta}_j \hat{\sigma}_{xj}^2 \hat{N}_j + \left(\hat{X}_2 - \hat{X}_1\right)\left(\hat{Y}_2 - \hat{Y}_1\right)\hat{N}_1\hat{N}_2/\left(\hat{N}_1 + \hat{N}_2\right)}{\sum_{j=1}^2 \hat{\sigma}_{xj}^2 \hat{N}_j + \left(\hat{X}_2 - \hat{X}_1\right)^2 \hat{N}_1\hat{N}_2/\left(\hat{N}_1 + \hat{N}_2\right)} \tag{2.1}$$

et la limite en probabilité

$$\beta = \frac{\sum_{j=1}^2 \beta_j \sigma_{xj}^2 \omega_j + (\mu_{x2} - \mu_{x1})(\mu_{y2} - \mu_{y1})\omega_1\omega_2}{\sum_{j=1}^2 \sigma_{xj}^2 \omega_j + (\mu_{x2} - \mu_{x1})^2 \omega_1\omega_2}.$$

(c) $\sqrt{n_j}\left(\hat{\beta}_j - \beta_j\right) \to_d N\left(0, \sigma_{xe,j}^2/\sigma_{x,j}^4\right),$ où $\to_d$ désigne la convergence en loi, et $\hat{\sigma}_{xe,j}^2 \to_P \sigma_{xe,j}^2$.

(d) Pour $k = 1, 2$,

$$\sqrt{n}\left(\hat{Y}_{\mathrm{reg},k} - Y\right)/N \to_d N\left(0, \sigma_k^2\right) \tag{2.2}$$

où $\sigma_k^2 = \sum_{j=1}^{2} a_{kj}^T D_j a_{kj}$ et

$$a_{1j} = \omega_j \varphi_j \left[-\left(\mu_y - \beta\mu_x\right), -\beta, 1\right]^T, \quad a_{2j} = \omega_j \varphi_j \left[-\left(\mu_{yj} - \beta_j \mu_{xj}\right), -\beta_j, 1\right]^T,$$

$\mu_x = \omega_1\mu_{x1} + \omega_2\mu_{x2}$, $\mu_y = \omega_1\mu_{y1} + \omega_2\mu_{y2}$, et D_j est donnée dans la condition (C3).

Les conditions (C1) à (C4) du théorème 1 fournissent une formulation générale du cadre de superpopulation pour l'inférence statistique sous le plan de sondage en grand échantillon, dans laquelle les coefficients de régression selon l'enquête estiment des paramètres descriptifs bien définis de la population servant de base de sondage. Les résultats des parties (a) à (b) montrent que les limites en probabilité β_j, α_j de $\hat{\beta}_j, \hat{\alpha}_j$ possèdent l'interprétation classique de pentes et d'ordonnées à l'origine de droites des moindres carrés de superpopulation. (Ces paramètres de pente et d'ordonnée à l'origine conservent aussi leur interprétation sous un modèle habituelle sous le modèle (2.7) présenté à la section 2.2.) La théorie asymptotique pour $\hat{\beta}_j$ dans la conclusion (c) nous permet de déduire le comportement en grand échantillon de $\hat{Y}_{\mathrm{dec}}$ à partir de celui fourni dans (d) pour $\hat{Y}_{\mathrm{reg},k}$.

Sous les conditions supplémentaires

$$\beta_1 = \beta_2, \alpha_1 = \alpha_2, \tag{2.3}$$

il découle clairement de la partie (b) du théorème 1 que $\beta_j = \beta$, et $\sigma_1^2 = \sigma_2^2$ dans (2.2), de sorte que $\hat{Y}_{\mathrm{reg},1}$, $\hat{Y}_{\mathrm{reg},2}$ et $\hat{Y}_{\mathrm{dec}}$ sont tous les trois asymptotiquement les mêmes jusqu'à des restes d'ordre plus faible que $N/\sqrt{n}$, comme nous allons le montrer maintenant. En outre, si $\beta_1 \neq \beta_2$, alors $\hat{Y}_{\mathrm{reg},2} - \hat{Y}_{\mathrm{dec}}$ continue d'être $o_P\left(N/\sqrt{n}\right)$, et le test d'égalité des pentes aboutit au rejet, c.-à-d. $P\left(\hat{Y}_{\mathrm{dec}} = \hat{Y}_{\mathrm{reg},2}\right) \to 1$, et par conséquent $\hat{Y}_{\mathrm{dec}}$ suit la même loi asymptotique que $\hat{Y}_{\mathrm{reg},2}$, qui est plus efficace que $\hat{Y}_{\mathrm{reg},1}$ selon le résultat de la section 2.2.

Théorème 2 Supposons que l'on formule les mêmes hypothèses (C1) à (C4) que pour le théorème 1.

(a) Quand la condition (2.3) est vérifiée, alors quand $m \to \infty$

$$\sqrt{n}\left(\hat{\beta}_2 - \hat{\beta}_1\right) \to_d N\left(0, \sigma_d^2\right), \qquad \sigma_d^2 = \sum_{j=1}^{2} \frac{\sigma_{xe,j}^2}{\varphi_j^2 \sigma_{xj}^4}, \tag{2.4}$$

et les estimateurs $\hat{Y}_{\mathrm{reg},1}, \hat{Y}_{\mathrm{reg},2}$ et $\hat{Y}_{\mathrm{dec}}$ suivent tous une loi asymptotiquement normale et sont équivalents au sens où

$$\frac{n}{N^2}\left[\left(\hat{Y}_{\mathrm{reg},1} - \hat{Y}_{\mathrm{reg},2}\right)^2 + \left(\hat{Y}_{\mathrm{reg},2} - \hat{Y}_{\mathrm{dec}}\right)^2\right] \to_P 0. \tag{2.5}$$

(b) Quand $\beta_1 \neq \beta_2$, $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$ et $\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)/N \to_d N\left(0, \sigma_2^2\right)$.

Une étude plus perfectionnée du comportement asymptotique des estimateurs $\hat{Y}_{\text{dec}}$ peut être entreprise dans l'esprit de Saleh (2006), comme dans le cas des versions contiguës ou de Pitman pour les modèles statistiques hors du contexte des sondages, en supposant que $\sqrt{n}(\beta_1 - \beta_2) \to r$ pour une constante r. Sous cette hypothèse, on peut montrer que $\hat{Y}_{\text{reg},1} - \hat{Y}_{\text{reg},2} = o_P\left(N/\sqrt{n}\right)$ et, par conséquent, que les trois estimateurs centrés et réduits $\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)$, $\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)$ et $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)$ suivent tous la même loi normale asymptotique de moyenne 0. En outre,

$$P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to \Phi\left(-z_{\tau/2} + r/\sigma_d\right) + \Phi\left(-z_{\tau/2} - r/\sigma_d\right), \tag{2.6}$$

où σ_d^2 est donné dans (2.4), et $z_{\tau/2}$ et Φ sont, repectivement, le point de pourcentage et la fonction de répartition de la loi normale centrée réduite. Donc, $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right)$ possède une limite différente de 1. En particulier, dans (2.6), la limite est égale à τ quand $\beta_1 = \beta_2$ (c.-à-d. quand $r = 0$).

2.2 Cadre asymptotique assisté par modèle

À la présente section, nous examinons le comportement des estimateurs $\hat{Y}_{\text{reg},k}, \hat{Y}_{\text{dec}}$ sous le modèle probabiliste hypothétique selon lequel les triplets (x_i, y_i, z_i) dans la population finie, $i \in U_j$, sont indépendants et identiquement distribués (iid), où les variables de taille $z_i > 0$ sont utilisées pour définir les probabilités de sélection PPT avec remise $p_{ij} = z_i \Big/ \sum_{i' \in U_j} z_{i'}$, et où x_i et y_i suivent le modèle

$$y_i = \alpha_j + \beta_j x_i + \varepsilon_i,\ i \in U_j, \tag{2.7}$$

avec α_j et β_j représentant les paramètres ordonnée à l'origine et pente inconnus pour la régression dans la strate U_j. Nous supposons que les erreurs $\varepsilon_i, i \in U_j$, sont iid de moyenne 0 et de variance finie σ_ε^2, et qu'elles sont indépendantes de (x_i, z_i), et que les variables x_i pour $i \in U_j$ ont une variance finie. En outre, pour permettre l'échantillonnage PPT, nous supposons que $\max_{i \in U_j} n_j p_{ij} < 1$ avec la probabilité s'approchant de 1 quand m est grand, c.-à-d. quand n_j, N_j sont grands.

À la présente section, les propriétés asymptotiques des estimateurs $\hat{Y}_{\text{reg},k}, \hat{Y}_{\text{dec}}$ sont considérées en regard du modèle de régression et de l'échantillonnage répété. En vertu du théorème 1, les estimateurs assistés par modèle $\hat{Y}_{\text{reg},1}$ et $\hat{Y}_{\text{reg},2}$ sont encore convergents et asymptotiquement normaux pour les triplets (x_i, y_i, z_i) iid à l'intérieur des strates, puisque les conditions (C1) à (C4) sont satisfaites sous les hypothèses de moments sur $z_i, 1/z_i$, même si le modèle (2.7) est incorrect. Cependant, les estimateurs $\hat{Y}_{\text{reg},k}$ sont efficaces quand le modèle (2.7) est correct.

Théorème 3 Supposons que l'on a le modèle (2.7) ainsi que la condition (C1), avec $E\left(x_i^4\right) < \infty, E\left(\varepsilon_i^4\right) < \infty, E\left(z_i\right) < \infty$, et $E\left(\left(1 + x_i^4\right)/z_i^3\right) < \infty$. Alors, toutes les conclusions du théorème 1 et du théorème 2 sont encore vérifiées. En particulier, quand $\beta_1 \neq \beta_2, \sigma_1^2$, la variance

asymptotique de $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)/N$ est plus grande que σ_2^2, la variance asymptotique de $\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)/N$. En outre,

$$\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)/N \to_d N\left(0, (1-\pi)\sigma_1^2 + \pi\sigma_2^2\right), \tag{2.8}$$

où π est la limite de $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right)$.

Notons que, dans (2.8), π est égal à 1 quand $\beta_1 \neq \beta_2$ et égal à τ quand $\beta_1 = \beta_2$.

Selon le théorème 3, sous le modèle (2.7), les trois estimateurs définis dans (1.2) à (1.4) ont tous la même efficacité asymptotique quand $\alpha_1 = \alpha_2$ et $\beta_1 = \beta_2$ (condition (2.3)). De surcroît, $\hat{Y}_{\text{reg},1}$ est asymptotiquement pire que $\hat{Y}_{\text{reg},2}$ quand $\beta_1 \neq \beta_2$. Donc, pourquoi n'utiliserions-nous pas systématiquement $\hat{Y}_{\text{reg},2}$?

Les assertions du théorème 3 sont des résultats asymptotiques d'ordre un. Un résultat asymptotique d'ordre deux, plus affiné, sous les conditions du théorème 3 et la condition (2.3) quand les tailles z_i sont toutes égales est que, jusqu'à un terme d'ordre $n_1^{-2} + n_2^{-2}$,

$$\text{eqm}\left(\frac{\hat{Y}_{\text{reg},1}}{N}\right) - \frac{\sigma_\varepsilon^2}{n} \leq \left[\text{eqm}\left(\frac{\hat{Y}_{\text{reg},2}}{N}\right) - \frac{\sigma_\varepsilon^2}{n}\right]\left[1 - \frac{n_1 n_2 \left(\bar{X}_1 - \bar{X}_2\right)^2}{nD_n}\right], \tag{2.9}$$

où l'eqm est l'erreur quadratique moyenne conditionnellement aux x_i, $\bar{X}_j = N_j^{-1}\sum_{i \in U_j} x_i$, et

$$D_n = \sum_{j=1}^{2}\sum_{i \in U_j}\left(x_i - \bar{X}_j\right)^2 + \frac{n_1 n_2 \left(\bar{X}_1 - \bar{X}_2\right)^2}{n}.$$

Le résultat (2.9) indique que, lorsque les poids sont égaux et que $\beta_1 = \beta_2$ et $\alpha_1 = \alpha_2$, la performance en échantillon fini de $\hat{Y}_{\text{reg},1}$ pourrait être meilleure que celle de $\hat{Y}_{\text{reg},2}$ pour des valeurs modérées de n_1 et n_2. Voir les résultats des simulations à la section 4. La preuve de (2.9) est un cas particulier d'un résultat plus général donné dans Slud (2012) et est donc omise.

Dans les applications, nous ne savons pas si $\beta_1 = \beta_2$. Donc, l'estimateur fondé sur un test de décision $\hat{Y}_{\text{dec}}$ est une procédure adaptative pour sélectionner un bon estimateur. Compte tenu de (2.8), la performance de $\hat{Y}_{\text{dec}}$ est proche (un peu moins bonne) de celle de $\hat{Y}_{\text{reg},2}$ quand $\beta_1 \neq \beta_2$, et est proche (un peu moins bonne) de celle de $\hat{Y}_{\text{reg},1}$ quand $\alpha_1 = \alpha_2$ et $\beta_1 = \beta_2$. Ces constatations sont également corroborées par les résultats des simulations à la section 4.

3 Estimation de la variance

Il est d'usage de communiquer une estimation de la variance ou de l'erreur-type pour chaque estimation d'après des données d'enquête. L'estimation de la variance est également essentielle pour l'inférence statistique lorsqu'on établit un intervalle de confiance pour un paramètre d'intérêt inconnu.

Les résultats asymptotiques de la section 2 suggèrent un estimateur de variance pour $\hat{Y}_{\text{reg},k}$ obtenu en substituant dans (2.2) des estimateurs pour les quantités inconnues dans σ_k^2. Puisque la variance totale est une somme de H variances intrastrate, sans perte de généralité, nous considérons une strate $(H = 1)$. Pour $j = 1, 2,$ soit

$$\hat{D}_{n_j} = \sum_{i \in S_j} \frac{\hat{b}_{ij}\hat{b}_{ij}^T}{(n_j - 1)\hat{N}_j}, \quad \hat{b}_{ij} = \left[1/p_{ij} - \hat{N}_j, x_i/p_{ij} - \hat{X}_j, y_i/p_{ij} - \hat{Y}_j\right]^T, \quad i \in S_j,$$

$$\hat{a}_{1j} = \frac{\hat{N}_j n^{1/2}}{\hat{N} n_j^{1/2}}\left[-\left(\bar{y}_j - \hat{\beta}_j \bar{x}_j\right), -\hat{\beta}_j, 1\right]^T, \quad \hat{a}_{2j} = \frac{\hat{N}_j n^{1/2}}{\hat{N} n_j^{1/2}}\left[-\left(\bar{y} - \hat{\beta}\bar{x}\right), -\hat{\beta}, 1\right]^T,$$

$$\bar{y}_j = \hat{Y}_j / \hat{N}_j, \quad \bar{x}_j = \hat{X}_j / \hat{N}_j, \quad \bar{y} = \sum_{j=1}^{2} \hat{Y}_j / \left(\hat{N}_1 + \hat{N}_2\right), \quad \bar{x} = \sum_{j=1}^{2} \hat{X}_j / \left(\hat{N}_1 + \hat{N}_2\right).$$

Alors, sous les conditions du théorème 1,

$$\hat{\sigma}_k^2 = \sum_{j=1}^{2} \hat{a}_{kj}^T \hat{D}_{n_j} \hat{a}_{kj} \to_P \sigma_k^2, \quad k = 1, 2.$$

C'est-à-dire que $\hat{\sigma}_k^2$ est convergent pour σ_k^2. Les résultats des théorèmes 2 et 3 montrent aussi que $\hat{\sigma}_2^2$ est un estimateur de variance convergent pour l'estimateur fondé sur un test de décision $\hat{Y}_{\text{dec}}$, parce que nous avons soit $\sigma_1^2 = \sigma_2^2$ soit $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$.

Cependant, ces estimateurs de variance obtenus par substitution peuvent ne pas donner d'aussi bons résultats lorsque la valeur de n_1 ou de n_2 est modérée (voir la section 4). Une autre méthode est celle du bootstrap proposée par Cheng et coll. (2010). Soit $\hat{\theta}$ l'estimateur pris en considération. L'estimateur bootstrap de sa variance peut être obtenu comme il suit.

1. Tirer un échantillon bootstrap S_j^* de taille n_j par échantillonnage aléatoire simple avec remise à partir de S_j, où S_1^* et S_2^* sont obtenus de manière indépendante. S'il existe k_j unités autoreprésentatives (AR) dans S_j, comme il est discuté à la section 4.1 qui suit, on tire alors des échantillons de tailles $n_j - k_j$ avec remise, avec $j = 1, 2$.

2. Utiliser les poids de sondage et les données observées provenant de l'ensemble de données originales pour former un ensemble de données bootstrap $S_1^* \cup S_2^*$. À partir de cet ensemble de données, calculer l'analogue bootstrap $\hat{\theta}^*$ de $\hat{\theta}$.

3. Répéter indépendamment les étapes qui précèdent B fois pour obtenir $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$. La variance d'échantillon de $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$ est l'estimateur bootstrap de la variance de $\hat{\theta}$.

Sous les conditions des théorèmes 1 et 2, les estimateurs bootstrap de la variance de $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$ et $\hat{Y}_{\text{dec}}$ sont des estimateurs convergents. La preuve pour le bootstrap est similaire aux preuves des théorèmes et est donc omise.

4 Résultats des simulations pour $H = 1$

La théorie en grand échantillon présentée plus haut ne convient pas pour indiquer si les résultats asymptotiques décrivent adéquatement le comportement des estimateurs $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$ et $\hat{Y}_{\text{dec}}$, et des estimateurs de leur variance dans des échantillons de taille modérée, ou si $\hat{Y}_{\text{reg},1}$ et $\hat{Y}_{\text{dec}}$ fournissent jamais des améliorations utiles de l'erreur quadratique moyenne dans des échantillons de taille moyenne. Nous présentons certains résultats de simulations pour étudier ces questions, ainsi que les problèmes de petit échantillon qui se posent lorsqu'on applique ces méthodes dans le contexte de l'ASPEP.

Dans les simulations, les valeurs dans la population servant de base de sondage U sont soit générées sous un modèle soit tirées des recensements des administrations publiques de 2002 et 2007 en utilisant les poids de sondage de l'ASPEP de 2007. Le premier jeu de simulations (présenté dans les tableaux 4.1 à 4.6) résume le comportement moyen sur de nombreuses populations servant de bases de sondage générées par un modèle. Dans le deuxième jeu de simulations portant sur des données artificielles, résumé au tableau 4.8, la population servant de base de sondage demeure fixe tout au long de la simulation. Toutes les populations servant de bases de sondage sont constituées d'une seule strate $(H = 1)$ décomposée en deux sous-strates $(j = 1, 2)$ selon que la valeur d'une variable de taille se situe en-dessous ou au-dessus d'un quantile particulier, habituellement le quantile 0,8. Dans toutes les simulations décrites à la présente section, l'échantillonnage des populations servant de bases de sondage est effectué selon un plan PPT avec remise.

4.1 Considérations concernant les petits échantillons

Avant de décrire les simulations, nous discutons de certaines caractéristiques particulières de l'échantillonnage PPT avec remise (PPTAR) qui, lorsqu'il est appliqué dans des conditions où les échantillons sont petits et les variables de taille ne sont pas équilibrées, requiert une approche de calcul spéciale. Des résultats numériquement irréguliers peuvent être obtenus lorsque les petits échantillons sélectionnés sont utilisés par strate, puis soumis au bootstrap pour estimer les variances.

Les poids $\pi_i^{-1} = 1/(n_j p_{ij})$ dans l'échantillonnage PPTAR ne sont tous supérieurs à 1 que si les probabilités de tirage simples $p_{ij} = z_i \Big/ \sum_{i' \in U_j} z_{i'}$ sont toutes inférieures à $1/n_j$. Pour éviter les résultats anormaux en petit échantillon et pour que les plans PPTAR imitant les plans PPT sans remise demeurent pertinents, toute unité $i \in U_j$ avec $n_j p_{ij} \geq 1$ est rendue *autoreprésentative* (AR), c.-à-d. qu'elle est échantillonnée avec certitude mais une seule fois, et si ces unités sont au nombre de k_j, alors les probabilités $\{p_{ij} : i \in U_j, n_j p_{ij} < 1\}$ sont renormalisées pour tirer un échantillon PPTAR de taille $n_j - k_j$. S'il reste des probabilités renormalisées $\geq 1/(n_j - k_j)$, leurs unités deviennent aussi autoreprésentatives et une nouvelle renormalisation est effectuée. Ce processus est répété aussi souvent qu'il est nécessaire. Donc, les petits échantillons dont les distributions des variables de taille sont très inégales pourraient ne pas être compatibles avec l'échantillonnage PPTAR, situation qui se présente dans certains cas de données réelles de l'ASPEP examinés plus loin.

Nous aurions pu faire un autre choix, mais nous nous conformons à la pratique de l'ASPEP consistant à inclure toutes les unités autoreprésentatives dans l'ajustement des estimateurs par régression pondérés

par les poids de sondage $\hat{\beta}_2$ et $\hat{\beta}$. Cependant, sous ce choix, l'échantillonnage PPTAR suivi par le rééchantillonnage bootstrap des petits échantillons peut donner lieu à un comportement très imprévisible, qui doit être reconnu quand on résume le comportement des estimateurs bootstrap de la variance. Le problème tient au fait que, quand un petit nombre m d'unités non autoreprésentatives sont échantillonnées selon un plan PPTAR, en plus d'un ensemble d'unités autoreprésentatives, puis sont traitées par la méthode du bootstrap, la probabilité que l'échantillon bootstrap contienne seulement une unité non autoreprésentative unique peut être étonnamment grande, ce qui donne lieu à une très forte variabilité du bootstrap. Ce phénomène a été observé dans les simulations présentées plus loin, pour une sous-strate de grande taille contenant 20 éléments ou moins et des variables de taille ayant une distribution très asymétrique, dans les cas de variables x_i lognormales ou de l'ASPEP.

4.2 Données artificielles générées par un modèle

Toutes les populations artificielles servant de bases de sondage ont été générées au moyen de N = 2 000 triplets (x_i, y_i, z_i) iid satisfaisant l'équation (2.7), pour U_1 constituée des N_1 = 1 600 d'entre-eux pour lesquels la valeur de x_i était inférieure à leur 80e percentile empirique $c = \left(x_{(1,600)} + x_{(1,601)}\right)/2$, et U_2 constituée des 400 autres. Dans la plupart des cas, les variables z_i ont été générées comme $N(30 + x_i, 100)$ variables conditionnées pour qu'elles soient positives (ce qui a nécessité à l'occasion une resimulation dans les modèles lognormaux de x_i ci-dessous) et étaient conditionnellement indépendantes de y_i sachant x_i. (Cependant, dans certains cas, des échantillons non pondérés ont été tirés en prenant les z_i identiquement égales.) Des échantillons PPT avec remise stratifiés de tailles $(n_1, n_2) = (100, 50), (100, 20)$, ou $(50, 20)$ ont été tirés dans des exécutions de simulation successives, en utilisant les variables de taille z_i, à partir de la même base de sondage.

Les modèles générant (x_i, y_i) sont indexés comme il suit. Dans les modèles dont le préfixe est **M1**, les variables indépendantes x_i suivent une loi Gamma $(4; 0,1)$ dont le quantile 0,8 est égal à 55,2, tandis que dans les modèles **M2**, les variables x_i suivent une loi lognormale (1;6,25) dont le quantile 0,8 est égal à 22,3. Les populations **M1**, et les modèles **M2** avec le suffixe **E** ont une variance conditionnelle de 100 pour y_i sachant x_i, tandis que les modèles **M2** sans le suffixe **E** ont une variance conditionnelle de $20x_i$. Les moyennes conditionnelles $E(y_i | x_i)$ sont toutes linéaires, égales à $20 + 1,5x_i$ dans les modèles indicés **H0** et égales à $20 + x_i + 0,5(x_i - c)I_{[j=2]}$ dans la sous-strate U_j dans les modèles **H1**. Les ordonnées à l'origine des modèles de régression sont choisies de manière que les droites se coupent à $x = c$, que les pentes soient égales ou non (voir la discussion à la section 1). Le tableau 4.1 donne la moyenne et l'écart-type (É.T.) pour les totaux Y générés à partir des attributs de la population servant de base de sondage $\{y_i\}_{i=1}^{2\,000}$ sous les divers modèles. Les variables x_i ainsi que les totaux Y ont une distribution à queue plus longue sous les modèles lognormaux.

Tableau 4.1
Moyennes et écarts-types des totaux Y sous les modèles de simulation

	Gamma		Lognormaux			
Modèle	**M1.H0**	**M1.H1**	**M2.H0**	**M2.H0E**	**M2.H1**	**M2.H1E**
E(Y)	160 000	123 177	225 603	225 603	173 485	173 485
É.T.(Y)	1 414,2	653,5	94 380	94 368	62 362	62 344

Modèles de population simulés

M1.H0 : $x_i \sim \text{Gamma}(4;0,1)$ (paramètre de forme 4, paramètre d'échelle 10), $y_i \sim N(20+1,5x_i;100)$ (variance 100), tout $i \in U$.

M1.H1 : $x_i \sim \text{Gamma}(4;0,1)$, $y_i \sim N\left(20+x_i+0,5(x_i-c)I_{[x_i \geq c]};100\right)$, tout i.

M2.H0 : $\log(x_i) \sim N(1;6,25)$, $y_i \sim N(20+1,5x_i;20x_i)$, tout i.

M2.H0E : $\log(x_i) \sim N(1;6,25)$, $y_i \sim N(20+1,5x_i;100)$, tout i.

M2.H1 : $\log(x_i) \sim N(1;6,25)$, $y_i \sim N\left(20+x_i+0,5(x_i-c)I_{[x_i \geq c]};20x_i\right)$, tout i.

M2.H1E : $\log(x_i) \sim N(1;6,25)$, $y_i \sim N\left(20+x_i+0,5(x_i-c)I_{[x_i \geq c]};100\right)$, tout i.

Les résultats des simulations et les résultats bootstrap présentés dans les tableaux 4.2 à 4.5 ont été générés suivant le plan de sondage et de présentation des résultats qui suit. Pour chaque type de population, 60 populations servant de bases de sondage distinctes ont été générées, et 50 expériences d'échantillonnage indépendantes ont été exécutées avec chacune de ces populations. Dans les cas où les résultats de l'échantillonnage pondéré et non pondéré ont été comparés, ces échantillons ont été tirés indépendamment l'un de l'autre à partir du même ensemble de 60 populations servant de bases de sondage. Donc, on disposait de 3 000 répliques indépendantes pour le calcul de la moyenne Monte Carlo des résultats statistiques, pour trois tailles d'échantillon stratifiées différentes, et 400 itérations bootstrap ont été effectuées pour chaque échantillon généré de cette façon.

Tableau 4.2
É.T. empiriques et estimés et couverture de l'IC, d'après les simulations du modèle M1

		M1.H0			M1.H1		
Tailles	**Stat.**	$\hat{Y}_{\text{reg},1}$	$\hat{Y}_{\text{reg},2}$	$\hat{Y}_{\text{dec}}$	$\hat{Y}_{\text{reg},1}$	$\hat{Y}_{\text{reg},2}$	$\hat{Y}_{\text{dec}}$
100,50	É.T.$_{MC}$	1 785,5	1 794,3	1 788,0	1 817,6	1 773,5	1 774,4
	É.T.$_S$	1 757,1	1 751,5	1 755,6	1 794,6	1 735,2	1 735,8
	É.T.$_B$	1 752,4	1 762,0	1 758,4	1 788,1	1 742,9	1 747,0
	PC$_S$	94,47	94,37	94,50	93,93	93,73	93,77
	PC$_B$	94,60	94,53	94,67	93,93	94,03	94,07
100,20	É.T.$_{MC}$	1 930,0	1 944,8	1 934,0	2 008,4	1 944,4	1 960,4
	É.T.$_S$	1 888,3	1 876,6	1 884,1	1 944,4	1 861,0	1 866,5
	É.T.$_B$	1 878,8	1 901,4	1 895,8	1 936,1	1 885,6	1 897,9
	PC$_S$	94,20	93,83	94,13	93,53	93,20	93,07
	PC$_B$	93,80	94,00	93,97	93,60	93,83	93,97
50,20	É.T.$_{MC}$	2 583,5	2 610,7	2 593,5	2 591,3	2 522,8	2 535,4
	É.T.$_S$	2 509,2	2 490,8	2 505,1	2 562,2	2 465,0	2 474,5
	É.T.$_B$	2 498,5	2 538,0	2 522,9	2 550,3	2 508,5	2 525,6
	PC$_S$	93,70	93,13	93,57	93,97	93,63	93,43
	PC$_B$	93,63	93,73	93,87	93,83	93,77	94,10

Tableau 4.3
É.T. empiriques et estimés et couverture de l'IC, d'après les simulations du modèle M2

		M2.H0			M2.H1		
Tailles	**Stat.**	$\hat{Y}_{\text{reg},1}$	$\hat{Y}_{\text{reg},2}$	$\hat{Y}_{\text{dec}}$	$\hat{Y}_{\text{reg},1}$	$\hat{Y}_{\text{reg},2}$	$\hat{Y}_{\text{dec}}$
100,50	É.T.$_{MC}$	3 400,1	3 475,4	3 406,8	3 481,9	3 483,8	3 482,2
	É.T.$_{S}$	3 420,6	3 400,0	3 417,0	3 537,8	3 405,0	3 463,7
	É.T.$_{B}$	3 590,0	3 715,2	3 623,4	3 852,0	3 921,9	3 898,4
	PC$_{S}$	95,10	93,43	94,83	95,03	93,40	94,13
	PC$_{B}$	95,67	95,77	95,77	95,63	95,77	95,70
100,20	É.T.$_{MC}$	5 655,2	6 184,0	5 698,6	5 853,0	6 181,1	5 955,6
	É.T.$_{S}$	5 644,9	5 575,7	5 640,9	5 798,3	5 587,3	5 697,3
	É.T.$_{B}$	5 565,1	6 687,3	5 857,8	5 907,8	6 838,0	6 466,6
	PC$_{S}$	93,83	88,47	93,40	92,77	88,30	90,70
	PC$_{B}$	92,33	93,67	93,37	92,63	94,33	94,17
50,20	É.T.$_{MC}$	5 773,2	6 319,2	5 833,9	5 934,2	6 230,6	6 009,8
	É.T.$_{S}$	5 800,2	5 677,2	5 785,8	6 012,6	5 755,4	5 919,2
	É.T.$_{B}$	5 728,5	6 825,2	6 086,0	6 102,2	6 978,1	6 522,1
	PC$_{S}$	94,60	88,67	93,97	94,07	89,37	92,27
	PC$_{B}$	93,40	94,23	94,27	93,47	95,03	94,80

Tableau 4.4
É.T. pour $\hat{Y}_{\text{HT}}$ *vs* $\hat{Y}_{\text{dec}}$, et couverture des intervalles de confiance percentiles bootstrap pour $\hat{Y}_{\text{dec}}$, pour $\tau = 0,05$ *vs* $0,20$, pour les modèles M1 et M2, H0 et H1

		$\hat{Y}_{\text{dec}}, \tau = 0,05$		$\hat{Y}_{\text{HT}}$	$\hat{Y}_{\text{dec}}, \tau = 0,20$	
Modèle	**Échantillons**	**É.T$_{MC}$**	**PC$_{PB}$**	**É.T.$_{\text{HT}}$**	**É.T$_{MC}$**	**PC$_{PB}$**
M1.H0	100,50	1 788,0	94,23	2 774,0	1 745,5	94,60
	100,20	1 934,0	93,50	3 032,6	1 915,9	94,10
	50,20	2 593,5	93,17	3 000,7	2 500,1	94,43
M1.H1	100,50	1 774,4	93,70	2 387,3	1 737,3	94,43
	100,20	1 960,4	93,27	2 678,9	1 948,0	93,23
	50,20	2 535,4	93,90	3 035,0	2 509,8	94,23
M2.H0	100,50	3 406,8	95,20	4 160,0	3 398,8	94,83
	100,20	5 698,6	91,13	6 720,2	5 705,7	92,57
	50,20	5 833,9	92,60	7 080,0	5 979,8	92,17
M2.H1	100,50	3 482,2	95,13	4 393,6	3 423,9	94,03
	100,20	5 955,6	92,07	7 413,1	5 917,3	92,40
	50,20	6 009,8	92,33	7 840,4	6 105,6	92,17

Tableau 4.5
Comparaisons des estimations de l'É.T. et de la couverture de l'IC pour H0 et H1 pour trois modèles lognormaux, pondérés (W) et non pondérés (U) dans M2, et pondérés (E) dans M2.E. Les couvertures en % des IC sont données pour les É.T. ainsi que les intervalles percentiles bootstrap

Modèle	Taille	Stat	É.T.	$\widehat{\text{É.T.}}_S$	$\widehat{\text{É.T.}}_B$	PC_S	PC_B	PC_{PB}
H0.W	100,50	$\hat{Y}_{reg,1}$	3 400,1	3 420,6	3 590,0	95,10	95,67	94,93
		$\hat{Y}_{reg,2}$	3 475,4	3 400,0	3 715,2	93,43	95,17	95,33
		$\hat{Y}_{dec}$	3 406,8	3 417,0	3 623,4	94,83	95,77	95,20
H0.U		$\hat{Y}_{reg,1}$	5 481,6	3 674,8	5 571,9	81,43	93,50	92,07
		$\hat{Y}_{reg,2}$	5 782,8	3 646,6	6 076,3	80,13	93,67	91,90
		$\hat{Y}_{dec}$	5 525,5	3 669,0	5 726,8	81,07	93,83	92,20
H0.E		$\hat{Y}_{reg,1}$	1 888,8	1 930,1	1 904,7	94,73	94,53	94,23
		$\hat{Y}_{reg,2}$	1 888,6	1 911,1	1 893,2	94,43	94,30	94,20
		$\hat{Y}_{dec}$	1 892,9	1 926,5	1 905,0	94,67	94,57	94,20
H0.W	50,20	$\hat{Y}_{reg,1}$	5 773,2	5 800,2	5 728,5	94,60	93,40	92,00
		$\hat{Y}_{reg,2}$	6 319,2	5 677,2	6 825,2	88,67	94,23	92,60
		$\hat{Y}_{dec}$	5 833,9	5 785,8	6 086,0	93,97	94,27	92,60
H0.U		$\hat{Y}_{reg,1}$	10 000,3	5 136,5	9 905,6	71,10	90,73	89,80
		$\hat{Y}_{reg,2}$	11 192,8	5 085,0	12 806,8	68,70	92,90	89,37
		$\hat{Y}_{dec}$	10 134,1	5 120,7	11 245,9	70,73	92,37	90,27
H0.E		$\hat{Y}_{reg,1}$	2 811,4	2 831,6	2 769,5	94,13	94,00	93,93
		$\hat{Y}_{reg,2}$	2 811,9	2 753,8	2 741,1	93,47	93,77	93,30
		$\hat{Y}_{dec}$	2 817,4	2 821,8	2 777,0	93,83	93,90	93,77
H1.W	100,50	$\hat{Y}_{reg,1}$	3 481,9	3 537,8	3 852,0	95,03	95,63	95,27
		$\hat{Y}_{reg,2}$	3 483,8	3 405,0	3 921,9	93,40	95,77	95,10
		$\hat{Y}_{dec}$	3 482,2	3 463,7	3 898,4	94,13	95,70	95,13
H1.U		$\hat{Y}_{reg,1}$	5 631,4	3 774,8	5 614,6	80,90	92,33	91,07
		$\hat{Y}_{reg,2}$	5 838,3	3 699,6	6 010,5	79,13	92,73	91,37
		$\hat{Y}_{dec}$	5 727,0	3 732,8	5 870,5	80,40	92,93	91,63
H1.E		$\hat{Y}_{reg,1}$	2 005,5	2 094,2	2 019,1	95,60	94,97	94,60
		$\hat{Y}_{reg,2}$	1 909,9	1 908,2	1 892,5	94,83	94,77	94,17
		$\hat{Y}_{dec}$	1 931,9	1 941,7	1 934,6	94,97	95,20	94,83
H1.W	50,20	$\hat{Y}_{reg,1}$	5 934,2	6 012,6	6 102,2	94,07	93,47	91,97
		$\hat{Y}_{reg,2}$	6 230,6	5 755,4	6 978,1	89,37	95,03	92,23
		$\hat{Y}_{dec}$	6 009,8	5 919,2	6 522,1	92,27	94,80	92,33
H1.U		$\hat{Y}_{reg,1}$	9 315,8	5 350,9	10 040,0	74,17	93,10	90,57
		$\hat{Y}_{reg,2}$	10 583,8	5 229,6	12 476,8	71,23	94,57	90,87
		$\hat{Y}_{dec}$	9 989,6	5 295,4	11 479,5	72,53	94,33	91,47
H1.E		$\hat{Y}_{reg,1}$	3 096,1	3 137,7	2 795,6	94,63	93,43	93,37
		$\hat{Y}_{reg,2}$	2 880,6	2 766,8	2 745,7	93,10	93,40	93,47
		$\hat{Y}_{dec}$	2 977,3	2 929,2	2 882,0	93,77	93,77	93,77

Nous avons calculé les quantités qui suivent pour chaque combinaison de modèles, pondérations et tailles d'échantillon : les biais en pourcentage de $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$, $\hat{Y}_{\text{dec}}$ (avec $\tau = 0{,}05$ dans tous les tableaux, sauf le tableau 4.4 où $\tau = 0{,}05$ ou 0,20) en tant qu'estimateurs de Y; les écarts-types (É.T.) Monte Carlo, É.-T.$_{MC}$, de ces trois estimateurs; les É.T. estimés des estimateurs, en utilisant les estimateurs de l'É.T. par substitution $(\widehat{\text{É.T.}}_S)$ et bootstrap $(\widehat{\text{É.T.}}_B)$, respectivement, décrits à la section 3; la probabilité de couverture, PC_u, des intervalles de confiance à 95 % nominaux pour $Y : \hat{Y} \pm 1{,}960 \cdot \widehat{\text{É.T.}}_u$, où $\hat{Y}$ est l'un des trois estimateurs de Y, et $u = S$ ou B; et les intervalles de confiance bootstrap percentiles (et leur pourcentage de couverture PC_{BP}) obtenus d'après les quantiles 0,025 et 0,975 empiriques des (400) valeurs bootstrap de chacun des trois estimateurs $\hat{Y}$ de Y. En outre, nous avons calculé les biais empiriques des estimations de Horvitz-Thompson $\hat{Y}_{\text{HT}}$ dans (1.1) et leurs écarts-types empiriques É.T.$_{\text{HT}}$. (De ces quantités calculées, seuls les biais ne sont pas présentés, puisqu'ils étaient tous nettement inférieurs à 0,5 % sauf pour le modèle **M2.H1.U**, et même dans ce cas, la valeur la plus importante du biais était de l'ordre de 1 %.) Deux autres statistiques, calculées et présentées au tableau 4.6 pour chacun des estimateurs $\hat{Y}$ de Y, sont les erreurs-types sur l'ensemble des populations servant de bases de sondage générées aléatoirement des estimations Monte Carlo et bootstrap intrapopulation des É.T. des estimateurs $\hat{Y}$.

Tableau 4.6
Erreurs-types sur l'ensemble des populations des É.T. empiriques et bootstrap estimés pour les estimateurs $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$, et $\hat{Y}_{\text{dec}}$, pour certains modèles et pondérations

		$\hat{Y}_{\text{reg},1}$		$\hat{Y}_{\text{reg},2}$		$\hat{Y}_{\text{dec}}$	
Modèle	**Tailles**	**É.T.**	$\widehat{\text{É.T.}}_B$	**É.T.**	$\widehat{\text{É.T.}}_B$	**É.T.**	$\widehat{\text{É.T.}}_B$
M1.H0	100,50	198	35	196	35	197	35
	50,20	210	52	208	51	210	51
M1.H1	100,50	204	39	183	40	184	41
	50,20	319	57	298	62	302	62
M2.H0	100,50	404	345	450	383	405	351
	50,20	825	518	1,075	916	889	631
M2.H0.E	100,50	187	49	185	45	184	47
	50,20	294	85	293	71	298	82
M2.H1	100,50	409	409	410	421	408	414
	50,20	767	624	946	929	841	730
M2.H1.E	100,50	208	59	196	46	204	50
	50,20	258	141	261	82	239	102
M2.H1.U	100,50	1 676	1 351	1 773	1 539	1 726	1 467
	50,20	2 397	2 543	3 425	3 454	3 102	3 159

4.3 Données réelles du recensement des administrations publiques

Nos simulations fondées sur l'échantillonnage répété à partir de bases de sondage contenant des données réelles s'appuient sur un ensemble de données nationales au niveau des États rassemblées par Yang Cheng. La base de sondage de l'ASPEP réalisée auprès des administrations publiques pour l'année de référence 2007, qui était aussi une année de recensement, est la même que celle du fichier du

recensement des administrations publiques (*Census of Governments*) de 2007. Notre ensemble de données contient les valeurs des variables de l'ASPEP de 2002 et de 2007 (nombre d'employés, rémunération et heures travaillées à temps plein et à temps partiel) tirées des recensements de ces années, ainsi que les poids de sondage de 2007 et les variables indicatrices de présence dans l'échantillon pour l'ASPEP. Un poids égal à 1 signifie que l'administration publique en question était autoreprésentative, au sens où elle a été choisie avec certitude en vue d'être incluse dans l'ASPEP. La variable de taille z_i pour l'échantillonnage PPT dans l'ASPEP est égale à la somme des masses salariales à temps plein et à temps partiel provenant du recensement le plus récent, de sorte que nous nous limitons à l'examen des 53 402 administrations publiques figurant dans le fichier pour lesquelles la valeur de cette variable était positive. Le tableau 4.7 donne les administrations publiques de type sous-comté et district spécial (les seules qui sont subdivisées en sous-strates de petites et de grandes unités) dans neuf États, ainsi que les nombres d'unités autoreprésentatives et les nombres d'unités échantillonnées en 2007. Comme il est mentionné à la sous-section 4.1, le nombre final d'unités autoreprésentatives (AR) pour l'échantillonnage PPT avec remise peut dépasser le nombre d'unités sélectionnées initialement en vue d'être incluses avec certitude, et les nombres plus élevés, qui correspondent à la taille de l'échantillon effectivement sélectionné en 2007, sont indiqués dans les colonnes AR du tableau 4.7. L'inspection de ce tableau montre que plusieurs combinaisons État-type d'administration publique ont une population nulle dans une sous-strate ou ne contiennent qu'un nombre trop faible d'unités non autoreprésentatives pour être utile dans la simulation d'échantillons répétés. À titre de règle empirique, nous prenons 15 comme nombre minimal d'unités non autoreprésentatives et nous recommandons que les paires de sous-strates contenant un nombre plus faible d'unités non autoreprésentatives dans la strate des grandes unités soient fusionnées sans recourir à la stratégie fondée sur un test de décision étudiée dans le présent article.

Tableau 4.7
Population de recensement, tailles d'échantillon de l'ASPEP et nombre d'administrations publiques de types sous-comté et district spécial autoreprésentatives par sous-strate en 2007, pour neuf États choisis

	Sous-comté					District spécial				
	Petites unités		Grandes unités			Petites unités		Grandes unités		
	Pop.	Éch.	Pop.	Éch.	AR	Pop.	Éch.	Pop.	Éch.	AR
AL	378	15	55	45	26	0	0	400	102	64
CA	0	0	475	104	86	1 595	39	107	107	107
CO	0	0	265	34	18	627	16	65	55	33
FL	317	16	81	54	36	0	0	330	48	24
GA	461	17	49	36	20	0	0	293	70	32
MO	980	25	101	101	101	799	27	106	66	42
NY	1 473	25	69	69	69	606	16	33	23	4
PA	2 409	55	123	81	31	921	21	37	37	37
WI	1 702	36	129	71	44	281	16	61	40	20

Pour neuf combinaisons d'administration publique par type comprenant 15 unités non autoreprésentatives ou plus et au moins 17 unités non autoreprésentatives non échantillonnées de la strate des grandes unités (sauf pour les États AL, CO, et GA pour lesquels il existait respectivement 9, 10 et 11 unités non autoreprésentatives non échantillonnées), le tableau 4.8 donne les résultats pour les

estimateurs fondés sur un test de décision et les estimations de la variance dans les paires de sous-strates. Dans chacune des combinaisons État-type d'administration publique, 3 000 échantillons PPTAR stratifiés ayant les tailles indiquées ont été tirés de la base de sondage de l'ASPEP et du recensement des administrations publiques décrites plus haut, avec x_i et y_i désignant, respectivement, la masse salariale des employés à temps plein de l'administration publique concernée telle qu'enregistrée aux recensements des administrations publiques de 2002 et de 2007, et z_i désignant la masse salariale totale (temps plein plus temps partiel) en 2002. Pour chaque échantillon simulé, on a calculé les estimateurs $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}, \hat{Y}_{\text{dec}}$ et estimé les variances empiriques. La variance de $\hat{Y}_{\text{dec}}$ a également été estimée par les méthodes de la formule de substitution et du bootstrap comme dans les simulations basées sur des données artificielles. (Mais il convient de souligner que, comme il a été décrit plus haut, dans chaque échantillon de sous-strate, les échantillons bootstrap ont été tirés uniquement parmi les unités non autoreprésentatives.) Les résultats sont présentés au tableau 4.8. Les efficacités relatives des estimateurs par la régression stratifiés combinés et distincts peuvent être évaluées d'après le ratio correspondant des É.T. donné dans la colonne 5 du tableau. Les autres É.T. présentés sont les estimateurs empiriques, par substitution et bootstrap de l'écart-type de $\hat{Y}_{\text{dec}}$.

Tableau 4.8
Sommaire des simulations par échantillonnage répété à partir de la base de sondage de l'ASPEP de 2007. La masse salariale totale des employés à temps plein (Y) est exprimée en multiples de 100 millions de dollars, et les estimations de l'É.T. données dans les colonnes 6 à 8 sont exprimées en unités de 1 million de dollars $\hat{Y}_{\text{dec}}$. É.T.$_1$/É.T.$_2$ dans la colonne 5 est le ratio de l'É.T. empirique de $\hat{Y}_{\text{reg},1}$ à celui de $\hat{Y}_{\text{reg},2}$.

État	Strate	Y	Taille	É.T.$_1$/É.T.$_2$	$\widehat{\text{É.T.}}$	$\widehat{\text{É.T.}}_S$	$\widehat{\text{É.T.}}_B$
AL	Sous-comté	1,2	25,46	2,14	4,90	3,67	5,71
CA	Distr. spécial	4,3	30,90	0,98	29,4	21,2	26,8
CO	Distr. spécial	0,6	25,55	1,14	3,77	2,58	3,00
FL	Sous-comté	4,3	25,54	1,16	11,9	9,4	12,2
GA	Sous-comté	1,5	25,38	1,15	4,38	3,26	4,88
MO	Distr. spécial	0,6	40,70	2,13	2,99	2,20	2,99
NY	Sous-comté	23,6	35,52	1,53	13,6	12,0	14,1
PA	Sous-comté	3,0	40,70	1,12	7,28	5,79	7,60
WI	Sous-comté	1,4	40,70	2,06	5,00	4,45	5,17

4.4 Discussion des résultats des simulations

L'exposé qui suit est un résumé et une interprétation des résultats des tableaux, ainsi que d'autres résultats non présentés.

I) Bon nombre des simulations au moyen de données artificielles servent à confirmer les résultats théoriques en grand échantillon des théorèmes. Nous avons déjà mentionné que, dans les tableaux 4.2 et 4.3, les biais des trois estimateurs de Y $\left(\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}, \hat{Y}_{\text{dec}}\right)$ sont généralement faibles. Dans le tableau 4.2, qui se rapporte aux modèles avec variables indépendantes et poids reliés à la loi Gamma dans les modèles **M1**, les estimateurs de variance par substitution et par bootstrap de chaque estimateur de Y sont assez précis et proches l'un de l'autre, et les intervalles de confiance ont tous une couverture proche de la couverture nominale. Sous **M1**.**H0** ainsi que **M1**.**H1**, pour les plus petites tailles d'échantillon n_2, on

note une tendance des estimateurs $\widehat{\text{É.T.}}_S$ et $\widehat{\text{É.T.}}_B$ à sous-estimer légèrement les écarts-types réels ou empiriques, mais $\widehat{\text{É.T.}}_B$ semble suivre l'écart-type de plus près que $\widehat{\text{É.T.}}_S$ pour $\hat{Y}_{\text{reg},2}$ et $\hat{Y}_{\text{dec}}$.

II) La distribution des valeurs de la variable x_i lognormale dans les modèles **M2** est beaucoup plus dispersée et asymétrique que dans les modèles **M1**, mais les résultats des simulations appuient néanmoins la théorie asymptotique quand $n_2 = 50$, quoique pas si $n_2 = 20$. Les intervalles de confiance de Y fondés sur l'estimateur par substitution en ce qui concerne $\hat{Y}_{\text{reg},2}$ ont une probabilité de couverture beaucoup trop faible lorsque l'on utilise l'estimateur de variance par substitution. Dans le tableau 4.3, pour chaque type d'estimateur de Y, l'estimateur de variance par substitution présente une tendance prononcée à sous-estimer la variance (empirique) réelle et l'estimateur par le bootstrap, à la surestimer.

Le tableau 4.5 clarifie le fait que le comportement extrême des estimateurs de variance sous les modèles **M2** résulte partiellement de ce que les distributions des variables indépendantes et de y_i sont dispersées et asymétriques, et partiellement de ce que la variable de taille utilisée dans les pondérations PPT présente aussi ces propriétés. Les cas désignés par le suffixe **W** dans ce tableau sont les mêmes que dans le tableau 4.3. Les cas portant le suffixe **E** ont les mêmes variables (x_i, z_i) que dans le tableau 4.3, mais les variances conditionnelles de y_i sachant x_i ont la valeur constante de 100; grâce à ce changement, le comportement irrégulier des estimateurs de l'écart-type disparaît. Cependant, lorsque les variances de y_i conditionnelles sont les mêmes que dans le modèle de base **M2**, mais que l'échantillonnage PPTAR est *non* pondéré, c.-à-d. lorsque toutes les variables z_i sont remplacées par la valeur 1, les estimateurs empiriques et bootstrap de l'écart-type sont très proches et très grands, tandis que l'estimateur de variance par substitution est trop faible, et ce d'un facteur spectaculairement grand variant de $1/2$ à $3/4$. Ce phénomène étrange s'observe de la même façon pour les trois estimateurs de Y. (Cependant, une variante de l'échantillonnage non pondéré dans le modèle **M1** ne modifie pas matériellement les résultats par rapport à ceux présentés au tableau 4.2.)

III) Un objectif des simulations était de savoir s'il existe jamais un avantage, en ce qui concerne l'erreur quadratique moyenne (EQM), à utiliser $\hat{Y}_{\text{reg},1}$ plutôt que $\hat{Y}_{\text{reg},2}$, faute de quoi il y aurait fort peu de raisons d'utiliser $\hat{Y}_{\text{dec}}$. En effet, les théorèmes en grand échantillon disent que le terme principal de variance en grand échantillon est toujours optimal pour $\hat{Y}_{\text{reg},2}$ (parce qu'il est le même que pour $\hat{Y}_{\text{reg},1}$ sous l'hypothèse nulle ou parce qu'il est strictement meilleur sous le modèle (2.7) avec des pentes distinctes). Toutefois, nous avons indiqué après le théorème 3, dans la borne (2.9), que $\hat{Y}_{\text{reg},1}$ peut avoir une EQM d'ordre deux plus petite que $\hat{Y}_{\text{reg},2}$, et les colonnes **H0** des tableaux 4.2 et 4.3 révèlent un avantage faible mais consistant de $\hat{Y}_{\text{reg},1}$ par rapport à $\hat{Y}_{\text{reg},2}$ en ce qui concerne l'écart-type, avantage qui est plus prononcé pour **M2**. Cet avantage disparaît sous la version fixe **M1**.**H1**, mais curieusement, pas sous **M2**.**H1**. L'avantage léger, mais réel, de $\hat{Y}_{\text{reg},1}$ en ce qui concerne l'EQM conditionnelle quand les pentes dans les sous-strates sont très proches de l'égalité est discuté plus en détail par Slud (2012).

Les estimateurs $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}, \hat{Y}_{\text{dec}}$ considérés ici sont du type régression et il pourrait être intéressant de comparer le comportement de leur EQM dans les populations simulées à celles de l'estimateur plus simple de Horvitz-Thompson $\hat{Y}_{\text{HT}}$ dans (1.1). Tous ces estimateurs sont presque sans biais, de sorte que les EQM sont essentiellement les mêmes que les variances, et une comparaison des troisième et cinquième colonnes du tableau 4.4 montre que les variances de $\hat{Y}_{\text{HT}}$ sont considérablement plus grandes que celles de $\hat{Y}_{\text{dec}}$. La

différence est moins prononcée pour les échantillons de plus grande taille, mais même dans ce cas, elle est de 30 % à 55 %. L'avantage de $\hat{Y}_{\text{dec}}$ reste encore très prononcé dans le modèle **M2**, où les variances sous le modèle et l'asymétrie de la distribution sont plus importantes, mais moins que dans le modèle **M1**.

IV) La définition de $\hat{Y}_{\text{dec}}$ contient le seuil de signification nominal arbitraire τ, qui dans tous les tableaux sauf le tableau 4.4 a été fixé à 0,05. Comme le laisse entendre la théorie en grand échantillon, les propriétés de l'estimateur fondé sur un test de décision sont comprises entre celles de $\hat{Y}_{\text{reg},1}$ et de $\hat{Y}_{\text{reg},2}$, et de plus grandes valeurs de τ rendent $\hat{Y}_{\text{dec}}$ plus souvent égal à $\hat{Y}_{\text{reg},1}$. Comme le montre la comparaison des colonnes 6 et 7 du tableau 4.4, le choix $\tau = 0{,}20$ semble aboutir, dans les modèles simulés, à des écarts-types de $\hat{Y}_{\text{dec}}$ très légèrement plus faibles sous le modèle **M1**, tandis que sous le modèle **M2**, l'écart-type est plutôt plus grand pour les petites tailles d'échantillon. La conclusion est faible, parce que les différences sont relativement petites comparativement aux différences d'écart-type observées d'une population servant de base de sondage à l'autre. Nous préférons laisser une plus petite valeur de τ dicter le groupement fréquent de sous-strates, sauf quand il existe des différences prononcées de pente estimée entre les sous-strates. Cette constatation selon laquelle de plus grands seuils de signification τ n'améliorent pas les propriétés de $\hat{Y}_{\text{dec}}$ diffère de celle de Saleh (2006) voulant que de plus grands seuils de signification soient très avantageux dans d'autres contextes de tests préliminaires.

V) Le tableau 4.6 renseigne sur la variabilité des estimateurs de l'écart-type des estimateurs de Y selon la population servant de base de sondage. Les estimateurs bootstrap de la variance semblent moins susceptibles de varier d'une population servant de base de sondage à l'autre, parce que la moyenne réalisée par le bootstrap les stabilise. Dans ce tableau, la principale constatation semble être que la variabilité entre les populations servant de bases de sondage est modérée, sauf sous le modèle **M2** non pondéré, où elle est remarquablement grande. Ce résultat semble expliquer l'inflation extrême des variances sous **M2.U** observées dans le tableau 4.5.

VI) Dans de nombreuses applications bootstrap avec statistique suivant approximativement une loi normale, la mauvaise couverture des intervalles de confiance fondés sur la théorie normale due à la non-normalité de la statistique obtenue par bootstrap peut être atténuée en utilisant les intervalles bootstrap percentiles (BP) (Shao et Tu 1995, section 4.1). Dans les présentes simulations, le tableau 4.4 (colonnes 4 et 6) donne les pourcentages de couverture des intervalles BP pour $\hat{Y}_{\text{dec}}$ dans les conditions où les tableaux 4.2 et 4.3 donnent les couvertures des IC sous la théorie de la loi normale basées sur l'écart-type estimé par bootstrap. Quelle qu'en soit la raison, les tableaux montrent que, sous la théorie de la loi normale, PC_B a systématiquement tendance à être légèrement inférieur à la valeur nominale mais néanmoins légèrement supérieur à la couverture des intervalles BP, PC_{BP}. Donc, nos simulations indiquent que, dans ces conditions, la préférence va à l'intervalle plus simple $\hat{Y}_{\text{dec}} \pm 1{,}96 \cdot \widehat{\text{É.T.}}_B$.

VII) Il reste à tirer les leçons des simulations portant sur des données réelles du recensement des administrations publiques présentées à la section 4.3. Le premier commentaire qui s'impose est que l'étalement et l'asymétrie de la distribution des variables indépendantes x_i correspondant à la masse salariale des employés à temps plein et de la variable de taille z_i correspondant à la masse salariale totale sont très importants, et ressemblent davantage à ceux observés pour les modèles lognormaux **M2** que pour les modèles Gamma **M1**. Le tableau 4.8 indique (dans la colonne 5) un avantage constant de $\hat{Y}_{\text{reg},2}$ par rapport à $\hat{Y}_{\text{reg},1}$ en ce qui concerne l'EQM, sauf dans le cas CA-district spécial, bien que la différence

soit faible dans le cas CO-district spécial et dans les cas FL, GA et PA-sous-comté. Il convient de souligner que, dans presque tous ces exemples, l'estimateur bootstrap de l'écart-type pour $\hat{Y}_{\text{dec}}$ est plus précis que l'estimateur par la formule de substitution, malgré les nombres assez faibles d'unités non autoreprésentatives échantillonnées et non échantillonnées et (dans plusieurs cas, comme le montre le tableau 4.7) des nombres relativement élevés d'unités autoreprésentatives. Les estimations de l'écart-type par substitution sont systématiquement trop petites, tandis que les estimations bootstrap sont habituellement légèrement élevées (c.-à-d. qu'en général $\widehat{\text{É.T.}}_S < \widehat{\text{É.T.}} < \widehat{\text{É.T.}}_B$). L'erreur relative de $\widehat{\text{É.T.}}_B$ par rapport à $\widehat{\text{É.T.}}$ ne dépasse pas environ 5 % dans ces exemples, sauf dans les cas (AL, CO, GA) où les unités non autoreprésentatives non échantillonnées sont particulièrement peu nombreuses dans la sous-strate de grandes unités.

Les sous-strates de grandes unités dans l'ASPEP ont habituellement une petite population totale dans la base de sondage et contiennent souvent un nombre relativement grand d'unités autoreprésentatives. Bien que nos simulations aient montré que cela n'invalide pas complètement les inférences faites au moyen de $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}$ ou $\hat{Y}_{\text{dec}}$, ces statistiques ont des distributions assez différentes de celles prévues par la théorie en grand échantillon, et de futures subdivisions des sous-strates permettraient peut-être d'obtenir des sous-strates de grandes unités un peu plus importantes en vue d'obtenir des inférences statistiques se comportant de la manière attendue.

Plus généralement, les résultats des simulations indiquent que l'estimateur fondé sur un test de décision avec l'estimateur des intervalles défini d'après les variances bootstrap se comporte bien et peut être recommandé, sauf pour des populations très dispersées et asymétriques ou des populations pour lesquelles les tailles d'échantillon de grandes unités sont plus petites que 20 à 25.

Remerciements

Le présent article décrit les travaux de recherche et analyses des auteurs et est diffusé en vue d'informer les parties intéressées et de favoriser la discussion. Les conclusions n'engagent que les auteurs et n'ont pas été approuvées par le Census Bureau. Nous tenons à remercier trois examinateurs et un rédacteur associé de leurs commentaires et suggestions utiles qui nous ont permis d'améliorer l'article. Les travaux de recherche de Jun Shao ont été financés partiellement par la bourse NSF Grant DMS-1007454.

Annexe

Preuve du théorème 1. Sous échantillonnage PPT, $\pi_i = n_j p_{ij}$ pour l'unité $i \in U_j$, et à chaque tirage avec remise, l'indice échantillonné $i_t \in U_j, t = 1, \ldots, n_j$ possède $P(i_t = i) = p_{ij}$ pour chaque $i \in U_j$. En calculant les moyennes et les variances (sous échantillonnage répété) de $\hat{N}_j$, $\hat{X}_j$, $\hat{Y}_j$, $N_j^{-1} \sum_{i \in S_j} x_i y_i / \pi_i$ et $N_j^{-1} \sum_{i \in S_j} x_i^2 / \pi_i$, nous constatons que les variances sont d'ordre n_j^{-1} au moyen

des limites données dans (C2) et (C3) et des bornes données dans (C4). Les assertions de la partie (a) s'ensuivent directement.

Pour l'assertion (b), nous avons, en vertu de la définition de $\hat{\beta}$, que

$$\hat{\beta} = \frac{\sum_{j=1}^{2}\sum_{i\in S_j}\left(x_i - \hat{\mu}_{x,j} + \hat{\mu}_{x,j} - \left(\hat{X}_1 + \hat{X}_2\right)/\left(\hat{N}_1 + \hat{N}_2\right)\right) y_i/\pi_i}{\sum_{j=1}^{2}\sum_{i\in S_j}\left(x_i - \hat{\mu}_{x,j} + \hat{\mu}_{x,j} - \left(\hat{X}_1 + \hat{X}_2\right)/\left(\hat{N}_1 + \hat{N}_2\right)\right)^2/\pi_i}$$

$$= \frac{N^{-1}\left(\sum_{j=1}^{2}\hat{\beta}_j\hat{\sigma}_{xj}^2\hat{N}_j + \left(\hat{N}_1\hat{N}_2/\left(\hat{N}_1 + \hat{N}_2\right)\right)(\hat{\mu}_{x1} - \hat{\mu}_{x2})\left(\hat{\mu}_{y1} - \hat{\mu}_{y2}\right)\right)}{N^{-1}\left(\sum_{j=1}^{2}\hat{\sigma}_{xj}^2\hat{N}_j + \left(\hat{N}_1\hat{N}_2/\left(\hat{N}_1 + \hat{N}_2\right)\right)(\hat{\mu}_{x1} - \hat{\mu}_{x2})^2\right)},$$

d'où l'égalité (2.1) dans (b) découle immédiatement par substitution des limites de la partie (a) ainsi que des limites $N_j/N \to \omega_j$.

Soit Σ_N la matrice diagonale par blocs avec deux blocs diagonaux D_{N_1} et D_{N_2}, et pour $j = 1, 2$, soit

$$\begin{aligned}
\Omega_{1j} &= \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\left(\frac{1}{p_{ij}} - N_j\right), \qquad \Omega_{2j} = \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\left(\frac{x_i}{p_{ij}} - X_j\right),\\
\Omega_{3j} &= \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\left(\frac{y_i}{p_{ij}} - Y_j\right), \qquad \Omega_{4j} = \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\frac{x_i - \mu_{x,j}}{p_{ij}}\left(y_i - \alpha_j - \beta_j x_i\right).
\end{aligned} \tag{A.1}$$

Puisque S_1 et S_2 sont indépendants, $\{\Omega_{k1}\}_{k=1}^4$ est indépendant de $\{\Omega_{k2}\}_{k=1}^4$. Notons que, ici et tout au long de la présente preuve, les sommes sur $i \in S_j$ utilisées pour définir $\hat{X}_j, \hat{Y}_j, \Omega_{kj}$, et les estimateurs de variance doivent être interprétés comme étant des sommes *avec multiplicité* compte tenu du plan d'échantillonnage PPT avec remise. La condition (C4) permet d'appliquer le théorème central limite de Liapounov pour montrer que

$$\Sigma_N^{-1/2}\left[\Omega_{11}, \Omega_{21}, \Omega_{31}, \Omega_{12}, \Omega_{22}, \Omega_{32}\right]^T \to_d N(0, I_6), \ \Omega_{4j} \to_d N\left(0, \sigma_{xe,j}^2\right), \tag{A.2}$$

où I_6 est la matrice identité de dimensions 6×6, et $\sigma_{xe,j}^2$ est donné dans l'énoncé de (d). Les limites qui définissent les variances asymptotiques dans (A.2) existent conformément à (C3).

Preuve de (c). Il est facile de vérifier d'après la définition que

$$\begin{pmatrix}\hat{\beta}_j - \beta_j\\ \hat{\alpha}_j - \alpha_j\end{pmatrix} = \frac{1}{\hat{N}_j\hat{\sigma}_{xj}^2}\sum_{i\in S_j}\begin{pmatrix}x_i - \hat{\mu}_{xj}\\ \hat{\sigma}_{xj}^2 - (x_i - \hat{\mu}_{xj})\hat{\mu}_{xj}\end{pmatrix}\frac{y_i - \alpha_j - \beta_j x_i}{\pi_i}.$$

Puisqu'il a été établi dans (a) que $\hat{\sigma}_{xj}^2 \to_P \sigma_{xj}^2$ et $\hat{N}_j/N_j \to_P 1$, il s'ensuit que la distribution limite de $\sqrt{n_j}\left(\hat{\beta}_j - \beta_j\right)$ est la même que celle de

$$\sqrt{n}_j \left(N_j \sigma_{xj}^2\right)^{-1} \sum_{i \in S_j} \left(x_i - \mu_{xj}\right)\left(y_i - \alpha_j - \beta_j x_i\right) / \pi_i ,$$

qui est clairement la même que celle de $\sigma_{xj}^{-2}\Omega_{4j}$ dans (A.1). La première assertion de (c) découle immédiatement de (A.2). La convergence de $\hat{\sigma}_{xe,j}^2$ s'ensuit en notant en vertu de (a) que

$$\hat{\sigma}_{xe,j}^2 - N_j^{-2} \sum_{i \in S_j} \frac{\left(x_i - \mu_{xj}\right)^2}{\pi_i p_{ij}} \left(y_i - \alpha_j - \beta_j x_i\right)^2 \to_P 0. \tag{A.3}$$

Le deuxième terme du premier membre de (A.3) contient une variance d'échantillonnage PPT avec remise calculée de manière qu'elle soit bornée par $1/n_j$ conformément à (C4), dont l'espérance en vertu de (C3) converge vers $\sigma_{xe,j}^2$.

Preuve de (d). De (1.2) et (a), il découle que $\left(\hat{Y}_{\text{reg},2} - Y\right)/N \to_P 0,$ qui peut aussi être considéré comme la représentation

$$\begin{aligned}
\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)/N &= \frac{\sqrt{n}}{N} \sum_{j=1}^{2} \left[\frac{N_j \hat{Y}_j}{\hat{N}_j} - Y_j + \hat{\beta}_j \left(X_j - \frac{N_j \hat{X}_j}{\hat{N}_j}\right)\right] \\
&= \frac{\sqrt{n} N_1^2}{\sqrt{n_1} N \hat{N}_1} \left[\left(-\bar{Y}_1 + \hat{\beta}_1 \bar{X}_1\right)\Omega_{11} - \hat{\beta}_1 \Omega_{21} + \Omega_{31}\right] \\
&\quad + \frac{\sqrt{n} N_2^2}{\sqrt{n_1} N \hat{N}_2} \left[\left(-\bar{Y}_2 + \hat{\beta}_2 \bar{X}_2\right)\Omega_{12} - \hat{\beta}_2 \Omega_{22} + \Omega_{32}\right] \\
&= d_{n1}^T \bar{\Omega}_1 + d_{n2}^T \bar{\Omega}_2 ,
\end{aligned}$$

où la deuxième égalité découle des définitions notationnelles de Ω_{kj} de même que $\pi_i = n_j p_{ij}$, $\hat{Y}_j = \sum_{i \in S_j} y_i / \pi_i$, $\hat{X}_j = \sum_{i \in S_j} x_i / \pi_i$, et la troisième de

$$d_{nj} = \frac{\sqrt{n} N_j^2}{\sqrt{n_j} N \hat{N}_j} \left[-\bar{Y}_j + \hat{\beta}_j \bar{X}_j, -\hat{\beta}_j, 1\right]^T, \quad \bar{\Omega}_1 = \left[\Omega_{11}, \Omega_{21}, \Omega_{31}\right]^T, \quad \bar{\Omega}_2 = \left[\Omega_{21}, \Omega_{22}, \Omega_{32}\right]^T.$$

En vertu de (A.2), $\bar{\Omega}_1 = O_p(1)$ et $\bar{\Omega}_2 = O_p(1)$. En vertu de la condition (C2), $d_{nj}^T = a_{2j}^T + o_p(1)$. Par conséquent, en vertu de (A.2), de la condition (C3) et de la méthode delta,

$$\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)/N = a_{21}^T \bar{\Omega}_1 + a_{22}^T \bar{\Omega}_2 + o_p(1) \to_d N\left(0, \sigma_2^2\right),$$

où la variance asymptotique $\sigma_2^2 = \sum_{j=1}^{2} a_{2j}^T D_j a_{2j}$ est systématiquement estimée par

$$\frac{n}{N^2} \sum_{j=1}^{2} \sum_{i \in S_j} \frac{1}{\pi_i^2} \left(y_i - \hat{\beta}_j x_i - \left(\hat{Y}_j - \hat{\beta}_j \hat{X}_j\right)/\hat{N}_j\right)^2 ,$$

qui est en accord avec la formule (9) de Cheng et coll. (2010). La preuve que $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)/N \to_d N\left(0, \sigma_1^2\right)$ est similaire.

Preuve du théorème 2. En vertu de la conclusion (c) du théorème 1,

$$\sqrt{n}\left(\hat{\beta}_2 - \hat{\beta}_1 - \beta_2 + \beta_1\right) \to_d N\left(0, \sum_{j=1}^{2} \sigma_{xe,j}^2 \Big/ \left(\varphi_j^2 \sigma_{xj}^4\right)\right). \tag{A.4}$$

La conclusion (2.4) dans la partie (a) de ce théorème s'ensuit directement.

Dans la preuve du théorème 1, nous avons montré que

$$\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)/N = a_{21}^T \bar{\Omega}_1 + a_{22}^T \bar{\Omega}_2 + o_p(1), \tag{A.5}$$

où les vecteurs constants a_{kj} (et μ_x, μ_y) ont été définis dans la partie (d) du théorème 1. De même,

$$\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)/N = a_{11}^T \bar{\Omega}_1 + a_{12}^T \bar{\Omega}_2 + o_p(1). \tag{A.6}$$

Quand (2.3) est vérifiée, $\beta_j = \beta$ (en vertu de la partie (b) du théorème 1) et $\mu_y - \beta\mu_x = \sum_{j=1}^{2} \omega_j \left(\mu_{yj} - \beta_j \mu_{xj}\right) = \mu_{y2} - \beta\mu_{x2}$, de sorte que $a_{1j} = a_{2j}$ pour $j = 1, 2$. Il découle directement de (A.5) et (A.6) que $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - \hat{Y}_{\text{reg},2}\right)/N \to_P 0$, et donc que les estimateurs $\hat{Y}_{\text{reg},k}$ suivent la même loi asymptotique, qui est normale comme nous l'avons montré à la partie (d) du théorème 1. Enfin, la définition de $\hat{Y}_{\text{dec}}$ implique que $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},1} \text{ ou } \hat{Y}_{\text{reg},2}\right) = 1$, et (A.5) et (A.6) impliquent que

$$\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)/N = a_{21}^T \bar{\Omega}_1 + a_{22}^T \bar{\Omega}_2 + o_p(1), \tag{A.7}$$

ce qui achève la preuve de (2.5) dans (a).

Preuve de (b). Si $\beta_1 \neq \beta_2$, alors (A.4) implique que $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$, c.-à-d. que le test t pour l'égalité de $\hat{\beta}_j$ donne lieu au rejet avec certitude à la limite. Alors (A.7) continue d'être vérifiée, et la loi asymptotique de $\hat{Y}_{\text{dec}}$ demeure la même que celle de $\hat{Y}_{\text{reg},2}$.

Preuve du théorème 3. Dans ce théorème, les hypothèses (C2) à (C4) sont remplacées par les hypothèses selon lesquelles les triplets iid (y_i, x_i, z_i) satisfont les conditions de moments et le modèle (2.7). Les assertions dans (C2) à (C4) restent alors vérifiées lorsque la probabilité tend vers 1 quand n, N sont grands, ce qui est établi à l'aide de la loi (forte) des grands nombres.

Outre les conclusions des théorèmes 1 et 2, il reste à montrer que $\hat{Y}_{\text{reg},2}$ possède une plus petite variance asymptotique que $\hat{Y}_{\text{reg},1}$. Soit $\vartheta = (\vartheta_1, \vartheta_2)$ et

$$F_j(\vartheta) = [-\vartheta_1, -\vartheta_2, 1]\, D_j\, [-\vartheta_1, -\vartheta_2, 1]^T.$$

Selon la définition de σ_1^2 et σ_2^2 dans (2.2), il suffit de montrer que $F_j(\vartheta)$ prend sa valeur minimale à $\vartheta = (\alpha_j, \beta_j)$. Nous allons maintenant prouver cela pour $j = 1$. La preuve pour $j = 2$ est similaire. Soit $m_{ii'}$ l'élément (i, i') de D_1. Puisque D_1 est symétrique et définie positive sous la condition (C3), $m_{12} = m_{21}$ et il existe un $\theta^* = (\theta_1^*, \theta_2^*)$ unique tel que $F_1(\theta^*) = \min_\vartheta F_1(\vartheta)$ et $\partial F_1(\vartheta)/\partial\vartheta^T|_{\vartheta=\theta^*} = 0$. Cela implique que θ^* est la solution des deux équations suivantes :

$$m_{11}\vartheta_1 + m_{12}\vartheta_2 = m_{13}, \quad m_{12}\vartheta_1 + m_{22}\vartheta_2 = m_{23} \tag{A.8}$$

Par conséquent, il suffit de montrer que $\theta^* = (\alpha_1, \beta_1)$. Puisque D_1 est définie positive, le système d'équations (A.8) possède une solution unique. Étant donné la définition de D_1,

$$\begin{aligned} m_{11}\alpha_1 + m_{12}\beta_1 &= \lim_{N_1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)^2 p_{i1}\alpha_1 + \sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)\left(\frac{x_i}{p_{i1}} - X_1\right)p_{i1}\beta_1\right] \\ &= \lim_{N_1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)(\alpha_1 - N_1\alpha_1 p_{i1} + \beta_1 x_i - \beta_1 p_{i1} X_1)\right], \end{aligned}$$

et

$$\begin{aligned} m_{13} &= \lim_{N_1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)\left(\frac{y_i}{p_{i1}} - Y_1\right)p_{i1}\right] \\ &= \lim_{N_1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)(\alpha_1 + \beta_1 x_i + \varepsilon_i - N_1\alpha_1 p_{i1} - \beta_1 p_{i1} X_1)\right] \\ &= \lim_{N1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)(\alpha_1 - N_1\alpha_1 p_{i1} + \beta_1 x_i - \beta_1 p_{i1} X_1)\right], \end{aligned}$$

où la dernière égalité découle de l'hypothèse que ε_i est indépendant de x_i et z_i, et est de moyenne 0 et de variance finie, et chacune des séquences z_i, $1/z_i$, et x_i/z_i est iid avec une espérance finie. Par conséquent, $m_{11}\alpha_1 + m_{12}\beta_1 = m_{13}$. On prouve de même que $m_{12}\alpha_1 + m_{22}\beta_2 = m_{23}$. Par conséquent, (α_1, β_1) est la solution unique du système d'équations (A.8), c.-à-d. que $F_1(\vartheta)$ prend sa valeur minimale à $\vartheta = (\alpha_1, \beta_1)$. D'où, $\sigma_2^2 < \sigma_1^2$. Cela termine la preuve du théorème 3.

Bibliographie

Bancroft, T., et Han, C.-P. (1977). Inference based on conditional specifications: A note and a bibliography. *International Statistical Review*, 45, 117-127.

Cheng, Y., Corcoran, C., Barth, J. et Hogue, C. (2009). An estimation procedure for the new public employment survey design. Washington, DC: American Statistical Association. *Survey Research Methods Section*, American Statistical Association, 3032-3046.

Cheng, Y., Slud, E. et Hogue, C. (2010). Variance estimation for decision-based estimators with application to the annual survey of public employment and payroll. *Government Statistics Section of the American Statistical Association*. Vancouver: American Statistical Association.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Fuller, W.A. (2009). *Sampling Statistics*. New York: John Wiley & Sons, Inc.

Isaki, C., et Fuller, W. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Rao, J.N.K., et Ramachandran, V. (1974). Comparison of the separate and combined ratio estimators. *Sankhyā*, C, 36, 151-156.

Saleh, A.K. Md. (2006). *Theory of Preliminary Test and Stein-type Estimation, with Applications*. Hoboken: Wiley-Interscience.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shao, J., et Tu, D. (1995) *The Jackknife and Bootstrap*. New York: Springer.

Slud, E.V. (2012). Moderate-sample behavior of adaptively pooled stratified regression estimators. U.S. Census Bureau preprint.

Techniques d'enquête, juin 2014
Vol. 40, N° 1, p. 115-136
Statistique Canada, N° 12-001-X au catalogue

L'influence de la méthode d'échantillonnage et des intervieweurs sur la réalisation de l'échantillon dans l'Enquête sociale européenne

Natalja Menold[1]

Résumé

Le présent article traite de l'effet de différentes méthodes d'échantillonnage sur la qualité de l'échantillon réalisé. On s'attendait à ce que cet effet dépende de la mesure dans laquelle les intervieweurs ont la liberté d'interviewer des personnes avec lesquelles il leur est facile de prendre contact ou dont il leur est facile d'obtenir la coopération (donc d'effectuer des substitutions). L'analyse a été menée dans un contexte transculturel en utilisant des données provenant des quatre premières vagues de l'Enquête sociale européenne (ESS, pour *European Social Survey*). Les substitutions sont mesurées par les écarts par rapport au ratio hommes-femmes de 50/50 dans des sous-échantillons constitués de couples hétérosexuels. Des écarts importants ont été observés dans de nombreux pays qui participent à l'ESS. Ces écarts se sont également avérés les plus faibles lorsque des registres officiels de résidents avaient servi de base de sondage pour le tirage des échantillons (échantillonnage à partir de registres de personnes) dans le cas où l'un des conjoints était plus difficile à joindre que l'autre. Cette portée des substitutions ne variait pas d'une vague à l'autre de l'ESS et était faiblement corrélée au mode de rémunération et aux procédures de contrôle des intervieweurs. Les résultats permettent de conclure que les échantillons tirés de registres de personnes sont de plus haute qualité.

Mots-clés : Méthodes d'échantillonnage; substitutions par les intervieweurs; erreurs non dues à l'observation.

1 Introduction

Les biais dans les statistiques d'enquête sont décrits par les modèles d'erreur d'enquête totale (Groves, Fowler, Couper, Lepkowski, Singer et Tourangeau 2004; Smith 2007). L'erreur d'enquête totale résulte de deux types d'erreurs, les erreurs d'observation et les erreurs non dues à l'observation. Le présent article porte sur l'influence qu'exercent les facteurs associés aux erreurs non dues à l'observation sur la comparabilité transculturelle, c'est-à-dire sur l'échec des statistiques d'enquête à représenter adéquatement la population cible. Ces types d'erreurs – également appelés erreurs de représentation – résultent de différences entre l'échantillon net obtenu (nombre de participants à l'enquête, Haeder et Lynn 2007) et la population cible.

Des travaux de recherche antérieurs menés dans des contextes transculturels ont révélé des différences de taux de réponse très prononcées entre les pays (Billiet, Phillipsen, Fitzgerald et Stoop 2007; Couper et de Leeuw 2003; de Heer 1999; de Heer et Israis 1992; de Leeuw et de Heer 2002; Hox et de Leeuw 2002; Johnson, O'Rourke, Burris et Owens 2002; Stoop 2005; Symons, Matsuo, Beullens et Billiet 2008), des différences entre les procédures utilisées sur le terrain (Billiet et coll. 2007; Kohler 2007; Kreuter et Kohler 2009; Smith 2007; Stoop 2005; Symons et coll. 2008) et des différences entre les méthodes d'échantillonnage (Lynn, Haeder, Gabler et Laaksonen 2007). Ces dernières font référence aux procédures de construction des bases de sondage et de sélection des éléments de l'échantillon. Toutes les différences énumérées peuvent nuire à la comparabilité transculturelle. En particulier, cette dernière pourrait être réduite en raison de méthodes d'échantillonnage différentes pour obtenir un échantillon probabiliste,

1. Natalja Menold, GESIS - Leibniz Institute for the Social Sciences, Survey Design et Methodology, P.O.Box 12 21 55, D-68072 Mannheim. Courriel : natalja.menold@gesis.org.

puisque la standardisation des méthodes d'échantillonnage est limitée par la disponibilité locale des bases de sondage, par leur qualité et leur convivialité, et par le budget d'enquête (Lynn et coll. 2007).

Lynn et coll. (2007) ont abordé la question des effets de plan et des tailles d'échantillon nécessaires pour que les échantillons nets soient comparables dans le contexte de l'ESS. Ils se sont concentrés dans leur étude sur la sélection de l'échantillon avant l'étape du travail sur le terrain. Toutefois, la comparabilité des échantillons pourrait également être influencée par les intervieweurs durant leur travail sur le terrain. La mesure dans laquelle les intervieweurs sont libres de remplacer les personnes échantillonnées par des personnes qui ne le sont pas (personnes avec lesquelles il est facile de prendre contact et qui sont coopératives) diffère selon la méthode d'échantillonnage (Hoffmeyer-Zlotnik 2006; Kohler 2007; Sodeur 2007). Il y a substitution sur le terrain lorsqu'une unité non répondante est remplacée par une unité de substitution (réserve) durant l'étape du travail sur le terrain du processus d'enquête (Vehovar 1999, page 335). Les substitutions dont parle Vehovar sont des substitutions légitimes qui sont permises par le protocole. En revanche, le présent article traite du problème des substitutions illégitimes (appelées simplement « substitutions » dans la suite de l'exposé) effectuées sans permission. Selon l'AAPOR (2003), les substitutions délibérées faites par les intervieweurs constituent une sorte de falsification.

Le but du présent article est de déterminer si l'effet des intervieweurs, que l'on suppose être associé aux substitutions, varie selon la méthode d'échantillonnage utilisée pour obtenir les échantillons probabilistes dans un contexte transculturel. En outre, on déterminera si le mode de rémunération, les procédures de contrôle, l'organisme de collecte des données (institution qui procède à la collecte des données sur le terrain) ou des facteurs temporels sont associés à cet effet d'intervieweur. Les résultats pourraient aider les responsables de l'enquête à décider des méthodes d'échantillonnage qu'il convient d'utiliser – un facteur de coût très important dans les enquêtes – et quand il convient de choisir des méthodes favorisant la motivation de l'intervieweur à ne pas effectuer de substitutions. Les résultats ont aussi de l'importance en ce qui concerne la recherche sur le comportement de l'intervieweur, puisqu'ils donnent un indice des erreurs associées à l'influence de l'intervieweur sur la comparabilité transculturelle.

La section suivante (section 2) expose le contexte théorique et empirique de l'étude présentée dans l'article. Les hypothèses sont décrites à la section 3. La section 4 renseigne sur la procédure et la méthode utilisées pour analyser l'influence de l'intervieweur. Les résultats sont présentés à la section 5. Enfin, la section 6 contient une discussion des résultats et les conclusions.

2 Contexte théorique et empirique

Des substitutions peuvent avoir lieu au cours des tâches que les intervieweurs effectuent avant l'interview. Les intervieweurs 1) créent une base de sondage, par exemple en dressant des listes d'adresses pour les enquêtes; puis, 2) ils cherchent à obtenir la coopération des unités sélectionnées (adresses, logements, ménages) et ils sélectionnent aussi les personnes pour l'interview à partir de ces unités. Enfin, les intervieweurs 3) cherchent à obtenir la coopération des personnes échantillonnées (Groves et coll. 2004). Quand les méthodes d'échantillonnage diffèrent, les tâches qu'effectuent les intervieweurs diffèrent comme il est décrit plus bas (figure 2.1).

La première méthode examinée est celle de l'échantillonnage à partir de registres de personnes (désigné par ERP). Les registres de population officiels au niveau de la personne sont utilisés comme bases de sondage pour l'ERP. La sélection des personnes est effectuée avant l'étape du travail sur le terrain, de sorte que les tâches des intervieweurs se réduisent simplement à obtenir la coopération des personnes échantillonnées (figure 2.1). Dans le cas de l'ERP, les intervieweurs peuvent influer sur la non-réponse (p. ex., Couper et Groves 1992; de Leeuw et Hox 1996; Durrant, Groves, Staetsky et Steele 2010), mais théoriquement parlant, ils n'ont aucune influence sur la base de sondage ni sur la sélection des éléments de l'échantillon. Ce niveau d'effet de l'intervieweur sur la non-représentation dans le cas de l'ERP est illustré par la flèche dans la figure 2.1.

Cependant, comme le montre l'exemple donné par Groves et coll. (2004), les éléments sélectionnés (personnes dans le cas de l'ERP) peuvent différer en ce qui concerne la probabilité d'être contacté par un intervieweur (facilité de prise de contact) et la probabilité d'obtenir la participation à l'enquête après la prise de contact (coopération). Par exemple, il est plus difficile de prendre contact avec les personnes qui habitent les régions urbaines ou celles qui sont jeunes, célibataires, sans enfants, mieux instruites et socialement actives (Stoop 2004). En revanche, les personnes âgées, les femmes, les personnes moins instruites et les personnes isolées socialement refusent plus souvent de coopérer (Dohrenwend et Dohrenwend 1968; Stoop 2004; Williams, Irvine, McGinnis, McMurdo et Crombie 2007). Si des difficultés surviennent lors des tentatives de prise de contact et d'obtention de la coopération des personnes cibles, des substitutions peuvent avoir lieu. Par exemple, Koch (1995) a donné le nombre de substitutions dans une enquête dans laquelle avait été utilisé l'ERP.

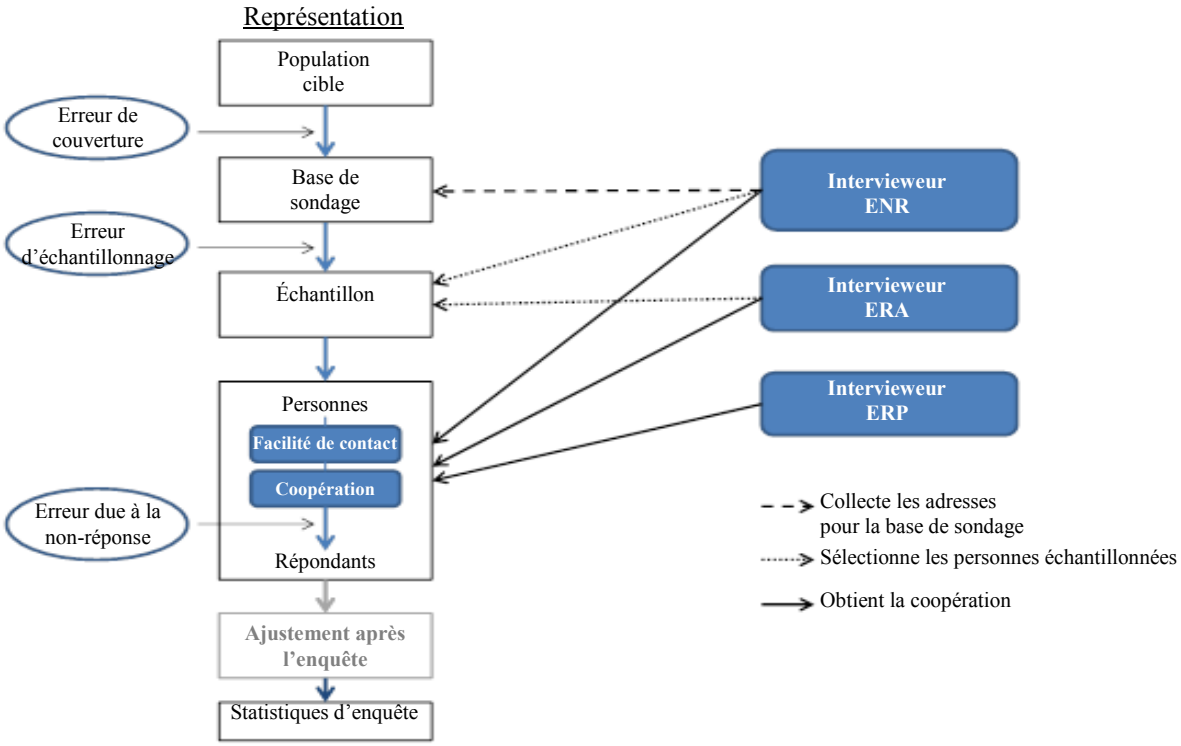

Figure 2.1 Tâches de l'intervieweur sous différentes méthodes d'échantillonnage reliées aux erreurs de couverture et de non-réponse. Le chemin de la représentation est adapté de Groves et coll., 2004, page 48.

La méthode d'échantillonnage suivante est celle de l'échantillonnage à partir de registres d'adresses/de ménages (ERA). Dans le cas de l'ERA, des listes de ménages ou d'adresses sont employées comme bases de sondage. Les ménages ou les adresses sont sélectionnés par les bureaux d'enquêtes avant l'étape du travail sur le terrain. Dans ce cas, les intervieweurs effectuent les tâches deux et trois (voir plus haut) : ils prennent contact avec les unités sélectionnées et sélectionnent les personnes pour l'interview si plus d'une personne admissible vit dans une unité sélectionnée. Les intervieweurs peuvent délibérément s'écarter des règles de sélection aléatoire et avoir ainsi un effet négatif sur l'échantillon sélectionné (figure 2.1). Puisque les intervieweurs ont plus de liberté pour sélectionner les personnes échantillonnées dans le cas de l'ERA que dans celui de l'ERP, on suppose que l'effet d'intervieweur associé aux substitutions est plus important pour l'ERA que pour l'ERP (figure 2.2). En outre, dans le cas de l'ERA, le résultat de la sélection de l'échantillon n'est pas connu d'avance et est, par conséquent, plus difficile à contrôler que dans le cas de l'ERP.

L'échantillonnage non effectué à partir de registres (ENR), lorsque ni des listes de personnes ni des listes d'adresses ou de ménages ne sont disponibles comme bases de sondage, représente la troisième méthode d'échantillonnage. Cette dernière englobe l'échantillonnage par marche aléatoire (p. ex., Arber 2002; ESS Sampling Plans), ainsi que l'établissement de listes d'adresses et l'échantillonnage (LAE). Dans le cas de l'ENR, les intervieweurs produisent eux-mêmes une base de sondage en dressant la liste des adresses dans une région géographique sélectionnée aléatoirement. Les intervieweurs doivent suivre rigoureusement les instructions concernant les procédures de collecte des adresses. Ils effectuent cette tâche en plus de la sélection des personnes à une adresse, comme il est décrit pour l'ERA, et de prendre contact avec les personnes sélectionnées et d'obtenir leur coopération, comme il est décrit pour l'ERP ainsi que l'ERA (figure 2.1). Dans le cas de l'ENR, les intervieweurs peuvent influer non seulement sur la sélection de l'échantillon, mais aussi sur la création de la base de sondage. Un intervieweur peut s'écarter des instructions et ne choisir que les adresses où il s'attend à pouvoir prendre contact avec la personne cible et obtenir sa coopération. Des substitutions sont particulièrement susceptibles d'avoir lieu lorsqu'on utilise une procédure (marche aléatoire) dans laquelle l'intervieweur réalise les interviews à des adresses qu'il choisit dans une région en se conformant à des instructions concernant la collecte des adresses et le chemin à suivre à travers la région. Dans un autre type d'ENR, plus restreint, la liste des adresses dans une région géographique est établie par l'intervieweur, mais la sélection réelle est effectuée par une équipe de coordination (établissement de la liste d'adresses et échantillonnage, LAE). Les adresses sélectionnées sont ensuite assignées à un intervieweur différent qui procède aux interviews. Le degré de liberté de l'intervieweur semble être le même dans le cas de la LAE que dans celui de l'ERA. Cependant, les instructions pour dresser la liste des adresses ou collecter les adresses peuvent être ambiguës dans les deux types d'ENR (Schnell, Hill et Esser 2011). Par conséquent, les intervieweurs ont plus de liberté pour effectuer des substitutions dans le contexte de l'ERA que dans celui de l'ERA (figure 2.2).

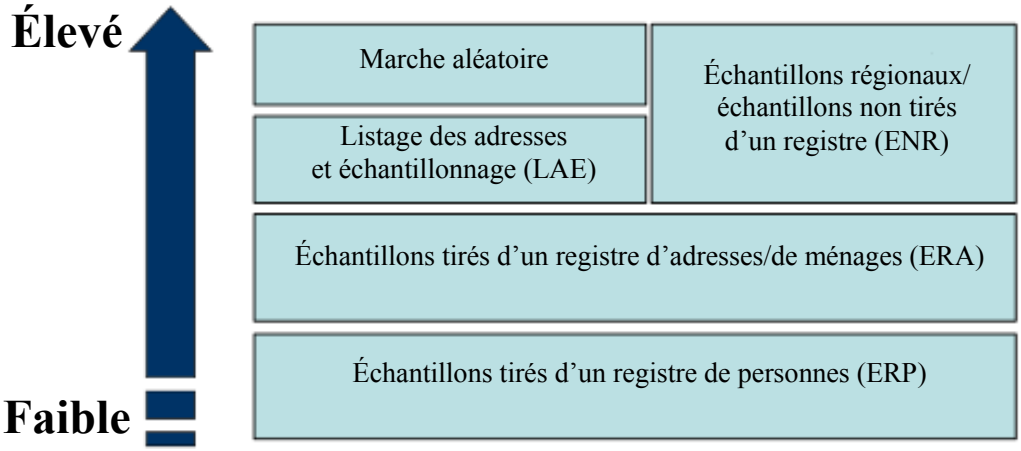

Figure 2.2 Degré de liberté des intervieweurs concernant les substitutions dans diverses méthodes d'échantillonnage

Les écarts dans les échantillons nets obtenus, qui sont associés aux écarts par rapport aux règles de sélection aléatoire de l'échantillon (p. ex., substitutions), peuvent être analysés empiriquement au moyen d'une méthode élaborée par Sodeur (1997). Cette méthode consiste à définir une sous-population caractérisée par un paramètre fixe et connu, puis à observer les statistiques représentant ce paramètre dans un sous-échantillon défini de façon correspondante. L'erreur non due à l'observation est d'autant plus forte que la statistique observée s'écarte du paramètre de population. Dans le présent article, nous considérons le ratio hommes-femmes des couples hétérosexuels que l'on sait être un paramètre de population valant 50/50. Dans les limites des fluctuations aléatoires, tout échantillon tiré de la population de couples hétérosexuels devrait produire une proportion d'hommes d'environ 50 %. Des écarts importants par rapport à ce niveau de 50 % indiquent des écarts par rapport aux normes de sélection de l'échantillon, par exemple sous forme de substitutions (voir la section 4.2 pour des précisions).

En utilisant cette méthode, Sodeur (2007) et Hoffmeyer-Zlotnik (2006) ont constaté que les écarts par rapport au ratio hommes-femmes vrai de 50/50 variaient entre les diverses vagues de l'Enquête sociale générale allemande (ALLBUS), qui comprend aussi différentes méthodes d'échantillonnage. Ces auteurs ont constaté que les hommes avec lesquels il est difficile d'entrer en contact sont interviewés moins fréquemment que les femmes (puisque les hommes sont les soutiens de famille des ménages avec enfants). Outre la facilité de prise de contact, les différences de coopération entre les conjoints peuvent jouer un rôle (Hoffmeyer-Zlotnik 2006). Si les conjoints sont à la retraite, la facilité de prendre contact avec l'un et l'autre est la même, mais ils diffèrent sur le plan de la coopération. L'homme à la retraite, maintenant à la maison, se sent responsable de fournir à l'intervieweur l'information sur le ménage (puisqu'il en est le « chef »). La femme peut refuser de participer puisque le mari aime coopérer. Un intervieweur qui prend contact avec ce genre de ménage pourrait interviewer les hommes plutôt que les femmes afin d'éviter les refus (Hoffmeyer-Zlotnik 2006).

Kohler (2007) a observé des écarts plus importants par rapport au ratio hommes-femmes de 50/50 choisi comme paramètre dans les échantillons de type ENR que dans ceux obtenus par d'autres méthodes d'échantillonnage dans six enquêtes transculturelles (Eurobarometer 62.1, European Quality of Life Survey EQLS'03, ESS 2002, ESS 2004, European Value Study 1999, International Social Survey

Program, ISSP 2002). Malheureusement, l'effet de la méthode d'échantillonnage observé par Kohler (2007) était particulier à l'enquête. Les échantillons les plus mal conçus – échantillons régionaux avec ENR – étaient surtout utilisés dans une enquête (EQLS). Les différences observées par Kohler entre une méthode de marche aléatoire et d'autres méthodes d'échantillonnage pourraient être attribuables à des différences entre l'EQLS et les autres enquêtes. D'autres chercheurs (Hoffmeyer-Zlotnik 2006; Souder 1997) ont étudié l'effet des méthodes d'échantillonnage sur l'effet d'intervieweur associé aux substitutions en ne considérant qu'une seule enquête nationale allemande; les résultats de cette étude ne sont donc pas applicables au contexte transculturel. Par conséquent, il est important de se pencher sur la question de la relation entre les méthodes d'échantillonnage et l'effet d'intervieweur associé aux substitutions dans les enquêtes transculturelles. Il importe aussi de prendre en considération d'autres facteurs explicatifs susceptibles d'affecter les substitutions. Les substitutions faites par les intervieweurs peuvent dépendre non seulement des méthodes d'échantillonnage, mais aussi des procédures mises en œuvre sur le terrain qui ont un effet sur la motivation qu'a l'intervieweur de produire des données d'enquête exactes. Par conséquent, les substitutions peuvent varier en fonction de l'organisme qui collecte les données (Hoffmeyer-Zlotnik 2006; Sodeur 1997; 2007) ou des contrôles utilisés durant une enquête (Kohler 2007). Les contrôles impliquent qu'un élément de l'échantillon est recontacté pour confirmer le résultat produit par un intervieweur. En plus des contrôles, les modes de rémunération des intervieweurs peuvent influer sur leur rendement. Si les intervieweurs sont rémunérés par interview achevée, ils assument les risques de coûts élevés associés aux longues distances entre les adresses sélectionnées, aux nombreuses tentatives de prise de contact ou à la longueur de l'interview (Sodeur 2007). Par conséquent, un changement d'organisme de collecte des données, de procédures de contrôle et de mode de rémunération doit être envisagé lorsqu'on analyse l'effet d'intervieweur associé aux substitutions. Outre ces facteurs, il est intéressant de savoir comment ces effets d'intervieweur varient au cours du temps. Par exemple, la prise en considération du temps dans un contexte transculturel permet de déterminer plus facilement si cet effet d'intervieweur est propre à un pays. Un effet d'intervieweur propre à un pays doit être stable dans le pays en question au cours de différentes vagues de l'enquête, même si la méthode d'échantillonnage a changé.

3 Hypothèses de recherche

Si l'effet de l'intervieweur en termes de substitutions est opérationnalisé en utilisant la méthode de Sodeur, on s'attend à pouvoir l'observer dans les statistiques d'enquête sous forme d'écarts par rapport au ratio hommes-femmes de 50/50 dans des sous-échantillons de répondants représentatifs des couples hétérosexuels. Cet effet d'intervieweur devrait différer en fonction du degré variable de facilité de prendre contact ou de coopération manifesté par les conjoints. Ceux-ci diffèrent sur le plan de la facilité de prendre contact dans les ménages constitués d'un couple ayant de jeunes enfants dans lequel l'homme est le soutien de famille (Hoffmeyer-Zlotnik 2006; Sodeur 2007; Stoop 2004). Si les intervieweurs recourent à des substitutions, la proportion d'hommes devrait être significativement plus faible que la valeur vraie (50 %) dans ces ménages, puisqu'il est plus difficile de prendre contact avec les hommes qu'avec les femmes. Cette situation change si l'on considère des couples dont les conjoints sont retraités. Ici, comme l'a exposé antérieurement Hoffmeyer-Zlotnik (2006), la facilité de prendre contact est de même niveau chez les deux conjoints, mais ceux-ci peuvent différer sur le plan de la coopération. Dans les sous-

échantillons de couples retraités, on s'attendrait à ce que la proportion d'hommes soit significativement supérieure à 50 % en cas d'effet d'intervieweur dû aux substitutions. Les hypothèses décrivant cet effet dans différents types de ménages sont les suivantes :

Hypothèse 1 (H1) : Les écarts par rapport au ratio hommes-femmes vrai (50/50) varient selon le type de ménage. Dans les ménages formés de couples avec (jeunes) enfants, la proportion d'hommes interviewés est inférieure à 50 %, tandis que dans les ménages formés de conjoints retraités, cette proportion est supérieure à 50 %.

Comme il est illustré à la section 2, les différentes méthodes d'échantillonnage peuvent être associées à différents degrés de liberté des intervieweurs en ce qui concerne les substitutions (figure 2.2). Par conséquent, on s'attend à observer les différences qui suivent entre les méthodes d'échantillonnage :

Hypothèse 2 (H2) : Les écarts par rapport au ratio hommes-femmes vrai (50/50) varient selon la méthode d'échantillonnage utilisée dans une enquête. L'ERP donne les écarts les plus faibles et l'ENR, les écarts les plus importants.

Si les écarts par rapport au paramètre de population sont dus au fait que les intervieweurs s'écartent des normes prescrites, ils devraient varier en fonction de la méthode d'échantillonnage utilisée ou du type de ménage, qui à son tour est associé à des niveaux variables de facilité de prise de contact ou de coopération de la part des conjoints. Les écarts devraient être stables au cours du temps si les méthodes d'échantillonnage demeurent constantes. Cependant, ils peuvent être corrélés au mode de rémunération des intervieweurs ainsi qu'aux procédures de contrôle, ou aux organismes de collecte des données, qui, en principe, devraient différer en ce qui a trait aux pratiques influant sur la motivation de l'intervieweur à travailler.

Hypothèse 3 (H3) : Mis à part les changements de méthode d'échantillonnage dans un pays, les écarts par rapport au ratio hommes-femmes de 50/50 sont indépendants de l'influence d'autres changements au fil du temps. Donc, ils ne varient pas d'une vague à l'autre d'une enquête. Cependant, le mode de rémunération des intervieweurs, les procédures de contrôle et les changements d'organisme de collecte des données devraient être corrélés aux écarts par rapport au ratio hommes-femmes de 50/50.

4 Méthodes

4.1 Données

Afin d'isoler tout effet dû à la méthode d'échantillonnage des autres effets propres à l'enquête, on peut utiliser des données provenant d'une enquête réalisée dans plusieurs pays qui ont appliqué des méthodes d'échantillonnage différentes. Les données provenant de nombreuses vagues de l'enquête devraient être disponibles afin de pouvoir tenir compte de l'effet temporel. Par conséquent, nous avons utilisé les données provenant des vagues 1 à 4 de l'ESS (European Social Survey Round 1-4 Data 2011). L'ESS a

été réalisée dans 20 à plus de 30 pays, qui diffèrent en ce qui a trait aux méthodes d'échantillonnage. En outre, le programme de l'ESS impose des normes rigoureuses aux organismes d'enquête, comme un échantillonnage aléatoire strict et des procédures de prise de contact étendues, ou des procédures de contrôle sur le terrain (Koch, Blom, Stoop et Kappelhof 2009; Philippens et Billiet 2004). L'efficacité des normes utilisées pour l'ESS a été illustrée par Kohler (2007), qui a montré que les écarts par rapport au ratio hommes-femmes de 50/50 étaient moins nombreux pour la vague 1 de l'ESS que pour d'autres enquêtes. De plus, les méthodes de collecte des données de l'ESS ont été améliorées régulièrement (Koch et coll., 2009). De surcroît, l'ESS fournit de la documentation détaillée sur les procédures d'échantillonnage, ainsi que sur la collecte des données (voir les ESS Documentation Reports), qui permettent d'opérationnaliser les variables d'intérêt.

4.2 Méthode d'évaluation de l'effet de l'intervieweur

La méthode mise au point par Sodeur (1997) a été choisie pour l'analyse. Cette méthode permet d'évaluer la qualité de l'échantillon net dans le cas des échantillons probabilistes. La qualité de la sélection aléatoire de l'échantillon a été examinée fréquemment au moyen d'autres statistiques disponibles dans un pays (critères externes). Toutefois, ces statistiques externes sont souvent inconnues, ce qui a poussé Sodeur à proposer l'utilisation de critères internes – c'est-à-dire l'utilisation d'information provenant de l'échantillon net seulement. Sodeur (1997) décrit la méthode comme comprenant les étapes suivantes : 1) tirer un sous-échantillon à partir de l'échantillon complet afin de se concentrer sur les répondants qui sont représentatifs des couples hétérosexuels : les conjoints doivent vivre ensemble dans un ménage et doivent appartenir tous deux à la population cible de l'enquête; 2) définir les unités qui doivent être retirées du sous-échantillon : célibataires, conjoints ne vivant pas ensemble dans un ménage et ménages comptant d'autres personnes apparentées qui appartiennent à la population cible. Alors, l'étape trois consiste à 3) définir une statistique d'enquête – p. ex., le pourcentage d'hommes – comme étant la variable dépendante qu'il faut comparer au paramètre de population.

Une analyse pour déterminer les causes des écarts par rapport au paramètre de population – par exemple, le comportement de l'intervieweur – requiert des spécifications supplémentaires aux étapes 1 et 2 pour s'assurer que le comportement de l'intervieweur varie (conceptuellement) en fonction de la facilité de prendre contact avec les personnes cibles ou de la coopération de ces dernières. Des spécifications de ce genre ont été formulées dans le présent article sous forme de définitions de différents types de ménages (voir l'hypothèse H1), dont la sélection est décrite à la section 4.3.

Le ratio hommes-femmes vrai de 50/50 chez les couples hétérosexuels n'est relié à aucun autre ratio hommes-femmes, tel que celui observé pour la population totale de résidents d'un pays. Par conséquent, comme le soutient Kohler (2007), ce ratio hommes-femmes ne peut être affecté par aucune sorte d'erreur de mesure et ne dépend pas de la taille du ménage, puisque l'analyse est limitée à deux personnes dans le ménage et que ces personnes appartiennent toutes deux à la population cible.

La méthode de Sodeur offre des avantages par rapport à d'autres méthodes, car aucune information ou données externes supplémentaires ne sont nécessaires. Cependant, elle requiert que les caractéristiques définies pour sélectionner les sous-échantillons soient connues non seulement pour les répondants, mais aussi pour leurs conjoints (p. ex., le sexe du conjoint). En outre, il ne doit pas exister de différence systématique de comportement de refus entre les hommes et les femmes (différences à l'égard du refus),

ce qui peut se manifester même si les intervieweurs travaillent honnêtement. Dans la pratique, on a constaté que les femmes sont plus hésitantes à participer que les hommes (Pickery et Loosveldt 2002; Schnauber et Daschmann 2008; Stoop 2004; Williams et coll. 2007). Cela semble être également le cas dans l'ESS, dans laquelle il s'est avéré que les femmes refusaient plus souvent de participer que les hommes. L'analyse des données des vagues 1 à 4 de l'ESS provenant des formulaires de prise de contact effectuée par l'auteur montre que 30,3 % d'hommes et 37,9 % de femmes ont refusé de participer à l'ESS1 (dans certains pays, aucune donnée concernant la variable de sexe n'a été fournie; par conséquent, la proportion de données manquantes était de 32,4 %). Dans l'ESS2, 30,8 % d'hommes et 37,9 % de femmes ont refusé de coopérer (31,3 % de données manquantes); dans l'ESS3, 33,8 % d'hommes et 39,0 % de femmes ont refusé de coopérer (27,2 % de données manquantes) et dans l'ESS4, 38,4 % d'hommes et 45,8 % de femmes ont refusé de coopérer (avec une proportion de données manquantes réduite à 15,8 %). Par conséquent, le fait que les hommes étaient présents dans un sous-échantillon de données de l'ESS moins de 50 % du temps peut être expliqué plausiblement par les substitutions, tandis qu'une fréquence des hommes supérieure à 50 % peut être expliquée par des différences à l'égard du refus. Cependant, si le pourcentage d'hommes variait selon la méthode d'échantillonnage – comme le prévoit l'hypothèse H2 –, il serait difficile d'expliquer un tel résultat uniquement par les différences à l'égard du refus, qui semblent être une caractéristique relativement stable.

4.3 Procédure

La section qui suit décrit les procédures utilisées pour tester les hypothèses H1 à H3. Le tirage des sous-échantillons à partir de l'échantillon complet de l'ESS est décrit pour commencer. Les écarts d par rapport au ratio hommes-femmes vrai de 50/50 dans un sous-échantillon représentent la variable dépendante dans toutes les analyses subséquentes. Les valeurs de d sont comparées entre différents types de ménages pour tester l'hypothèse H1. Ensuite, l'opérationnalisation de la variable « méthode d'échantillonnage » (pour tester l'hypothèse H2) est décrite. Enfin, l'hypothèse H3 est reliée aux variables de temps, de changement d'organisme de collecte des données, de mode de rémunération et procédures de contrôle des intervieweurs, dont l'opérationnalisation est décrite à la dernière section. Les hypothèses H2 et H3 ont été testées en faisant appel à l'analyse de covariance multivariée (MANCOVA) suivie d'analyses de covariance (ANCOVA) dans lesquelles la méthode d'échantillonnage a été utilisée comme variable indépendante et la vague de l'ESS, le changement d'organisme de collecte des données, la prime de rémunération et les contrôles de l'intervieweur ont servi de covariables.

Tirage des sous-échantillons

La population cible de l'ESS correspond dans chaque pays aux personnes de 15 ans et plus qui résident dans des logements privés, indépendamment de leur nationalité et leur citoyenneté, leur langue ou leur statut légal (p. ex., ESS-1 2002 Documentation Report, page 2). Les répondants ($n = 88\,375$) qui vivent avec un conjoint de sexe opposé âgé de 15 ans ou plus ont été sélectionnés à partir de l'échantillon total des vagues 1 à 4 de l'ESS ($n = 184\,988$). Le tirage de cet échantillon a réduit la base de données d'analyse à la moitié environ de l'échantillon total. Cependant, cette sélection était nécessaire pour s'assurer que le pourcentage prévu d'hommes soit de 50 %.

Trois types de ménages ont été distingués dans le sous-échantillon sélectionné : les couples avec enfants de 0 à 6 ans (< 7ans; $n = 18\,791$), les couples avec enfants de 7 à 14 ans ($n = 53\,651$) et les couples dont les deux conjoints étaient à l'âge de la retraite (retraités, $n = 15\,933$). Afin de déterminer l'âge de la retraite, on s'est servi de l'âge légal de la retraite en vigueur dans chaque pays (voir l'annexe). Les deux premiers groupes avec enfants ont été formés parce que l'on supposait que les différences de facilité de prise de contact entre les conjoints étaient particulièrement prononcées dans ces ménages. Pour le troisième groupe, on a supposé que les différences de facilité de prise de contact selon le sexe étaient relativement modestes, mais que les hommes et les femmes différaient en ce qui concerne la coopération.

Le fait que les hommes sont les soutiens de famille dans les deux sous-échantillons contenant des ménages avec enfants est corroboré par l'analyse effectuée par l'auteur en utilisant les données de l'ESS. L'examen des activités des répondants au cours des sept jours qui ont précédé l'enquête dans les ménages avec enfants de moins de sept ans a montré que 58 % d'hommes et 42 % de femmes avaient un emploi rémunéré. Pour ce qui est du conjoint du répondant, 64 % d'hommes et 36 % de femmes avaient un emploi rémunéré. Des résultats comparables ont été obtenus pour les répondants dans les ménages avec enfants de 7 à 14 ans (pour les répondants, 54 % d'hommes et 46 % de femmes avaient un emploi rémunéré et pour les conjoints des répondants, 60,5 % d'hommes et 39,5 % de femmes étaient dans cette situation). Dans les ménages dont les conjoints étaient retraités, 80,6 % des répondants étaient retraités, 11,5 % faisaient des travaux ménagers et 1,3 % étaient atteints d'une maladie ou d'une incapacité de longue durée. En ce qui concerne le conjoint du répondant, 84,4 % étaient retraités, 17 % faisaient des travaux ménagers et 2,1 % étaient atteints d'une maladie ou d'une incapacité de longue durée.

Catégorisation des méthodes d'échantillonnage

Les méthodes d'échantillonnage ont été classées en se basant sur les rapports documentaires produits pour chaque vague de l'ESS (European Social Survey (2011): ESS 1-4 Documentation Reports). Le tableau 4.1 résume les principales caractéristiques des méthodes d'échantillonnage utilisées dans l'ESS. Le tableau 4.2 montre quelles méthodes d'échantillonnage ont été utilisées dans chaque pays à chaque vague. Pour des renseignements plus détaillés sur les procédures de sélection de l'ESS, consulter les rapports documentaires ou Lynn et coll. (2007).

Dans le cas de l'ERA, il est important d'examiner le traitement des immeubles à logements multiples correspondant à une adresse, puisque ce sont également les intervieweurs qui gèrent la situation. La documentation de l'enquête ne décrivait ce traitement que pour quelques pays (Irlande, Israël, Pays-Bas et Royaume-Uni). En Irlande, par exemple, les intervieweurs dressaient la liste des ménages et sélectionnaient l'un d'entre eux en utilisant la grille de Kish (Kish 1965).

En Autriche, une méthode d'ENR a été appliquée à 50 % seulement de l'échantillon, tandis que les 50 % restants ont été sélectionnés par une méthode d'ERA. Puisque l'ENR peut entraîner un plus grand nombre de substitutions que l'ERA seulement, on peut s'attendre à ce que les résultats en Autriche soient plus semblables aux résultats obtenus dans les pays ayant utilisé un ENR que dans ceux ayant utilisé un ERA. Par conséquent, l'auteur a classé l'Autriche dans la catégorie de l'ENR.

Tableau 4.1
Méthodes d'échantillonnage dans les pays participant à l'EES (vagues 1 à 4)

	Échantillonnage à partir d'un registre de personnes	Échantillonnage à partir d'un registre d'adresses/de ménages	Échantillonnage non effectué à partir d'un registre
Base de sondage	Listes fiables de résidents	Listes fiables d'adresses/de ménages	Régions (pas de listes de résidents, d'adresses ni de ménages)
1er degré : Sélection des UPE			
Définition d'une unité	Grappes régionales, régions, municipalités	Circonscriptions électorales, régions de codes postaux	Grappes régionales, régions, municipalités
Processus de sélection	Échantillonnage aléatoire systématique	Échantillonnage aléatoire systématique	Échantillonnage aléatoire systématique
Résultat	Communauté, municipalité	Circonscriptions électorales, circonscriptions de code postal	Régions géographiques, municipalités
2e degré : Sélection des ménages	Sans objet		
Définition d'une unité		Un ménage, une adresse	Un ménage/logement
Processus de sélection		Échantillonnage aléatoire simple ou systématique	Marche aléatoire/LAE échantillonnage aléatoire simple
Résultat		Adresses des ménages	Ménage/adresse/logement
3e degré : Sélection des personnes			
Définition d'une unité	Personne cible	Personne cible	Personne cible
Processus de sélection	Échantillonnage aléatoire simple ou systématique	Sélection aléatoire par l'intervieweur par la méthode de la grille de Kish ou du dernier anniversaire	Sélection aléatoire par l'intervieweur par la méthode de la grille de Kish ou du dernier anniversaire
Résultat	Nom et adresse des personnes échantillonnées	Personnes échantillonnées	Personnes échantillonnées

Tableau 4.2
Classification des pays participant à l'ESS en fonction des méthodes d'échantillonnage

Vague de l'ESS	Échantillonnage à partir d'un registre de personnes (ERP)	Échantillonnage à partir d'un registre d'adresses/de ménages (ERA)	Échantillonnage non effectué à partir d'un registre (ENR)
ESS 1	BE, DE, HU, NO, PL, SI, DK, FI, SE	Adresse : IE, IT, NL, GB, CH Ménage : CZ, LU, ES	FR, GR, PT, AT
ESS 2	BE, DE, HU, NO, PL, SI, DK, FI, SE, ES, EE, IS, SK	Adresse : IE, NL, GB, CH Ménage : LU, TR	FR, GR, PT, AT, CZ, UA
ESS 3	BE, DE, NO, PL, SI, DK, FI, SE, ES, EE, SK	Adresse : IE, NL, GB, CH, LV Ménage : CY, BG, HU	FR, PT, AT, UA, RU, RO
ESS 4	BE, DE, HU, NO, PL, SI, DK, FI, SE, ES, EE	Adresse : IE, NL, GB, CH, IL, LV Ménage : CZ, CY, LT, GR, KRO, TR, BG	FR, PT, SK, UA, RU, RO

Note La Roumanie n'est pas incluse dans le fichier de données intégré de l'ESS; aucune information sur la méthode d'échantillonnage n'a été fournie par l'Italie dans les rapports documentaires ESS2 à ESS4. Les pays sont étiquetés conformément à la norme ISO 3166-1, voir l'annexe.

Le type de méthode d'ENR utilisé par un pays a rarement été décrit dans la documentation. Dans le cas de l'ESS1, il est évident que la Grèce est la seule à avoir utilisé la méthode LAE. L'utilisation d'une méthode LAE est décrite pour la République tchèque et la Slovaquie pour les dernières vagues. Pour l'ESS4, l'Ukraine, la Russie et le Portugal décrivent une procédure comparable à la méthode LAE. Cependant, dans ces pays, les intervieweurs (et non les bureaux) ont sélectionné un nombre fixe d'unités à partir des listes produites par d'autres intervieweurs.

Variables explicatives

L'information se rapportant à une vague particulière de l'ESS a été utilisée comme variable pour tenir compte de l'effet temporel. Les rapports documentaires ont fourni des renseignements sur les autres variables explicatives, le changement d'organisme de collecte des données, ainsi que le mode de rémunération et les procédures de contrôle des intervieweurs. Les pays qui ont changé d'organisme de collecte des données entre les vagues sont présentés en annexe. Pour ce qui est du mode de rémunération, il s'est avéré que la principale méthode employée pour l'ESS était la rémunération par interview achevée. Un taux de rémunération horaire n'a été utilisé que dans quelques pays qui ont également utilisé un ERP (pour les ESS1 et 2 en Norvège et en Suède, ainsi que les ESS3 et 4 en Norvège et en Finlande). Par conséquent, les modes de rémunération variaient peu et il n'a pas été possible d'effectuer une analyse des données correspondantes. Toutefois, le paiement de primes variait selon le pays et selon la vague de l'enquête. Par conséquent, cette information a été utilisée pour produire une variable de contrôle dichotomique (paiement d'une prime : oui/non).

Deux variables ont été utilisées pour décrire les procédures de contrôle : le nombre d'éléments de l'échantillon admissibles sélectionnés pour les contrôles divisé par le nombre d'éléments de l'échantillon admissibles (ratio d'éléments sélectionnés), ainsi que le nombre de résultats confirmés divisé par le nombre d'éléments de l'échantillon sélectionnés pour les contrôles (ratio de résultats confirmés). La première variable décrit le nombre de contrôles dans un pays, tandis que la seconde décrit l'efficacité de ces contrôles. Le « ratio d'éléments sélectionnés » varie de 10 % pour l'ERP, à 13 % pour l'ENR et à 16 % pour l'ERA. Le « ratio de résultats confirmés » est un peu plus élevé pour l'ENR ($M = 75{,}21$, $E.T. = 24{,}81$) que pour les deux autres méthodes d'échantillonnage (ERP : $M = 61{,}89$, $E.T. = 31{,}95$; ERA : $M = 66{,}49$; $E.T. = 32{,}56$).

5 Résultats

5.1 Différences entre les types de ménages

Pour commencer, nous présenterons les résultats de la vérification de l'hypothèse H1. Selon cette hypothèse, les écarts par rapport au ratio hommes-femmes de 50/50 varient en fonction du type de ménage. La figure 5.1 montre les différences (d) entre le pourcentage réel d'hommes et la valeur vraie prévue de 50 % dans trois sous-échantillons. Un intervalle de confiance (IC) à 95 % a été utilisé pour tenir compte des fluctuations aléatoires. Comme la proportion prévue d'hommes est $p = 0{,}5$, la variance vaut,

en moyenne, $0,25/n$, où n est le nombre de cas dans le sous-échantillon dans un pays. L'IC à 95 % a été calculé comme il suit (voir Kohler 2007, page 59) :

$$IC = 0,5 \pm 1,96 \times \sqrt{0,25/2}.$$

La figure 5.1 montre que pour les deux sous-échantillons contenant des ménages avec enfants, les valeurs significatives de d sont négatives dans la majorité des cas, ce qui signifie que la proportion d'hommes dans ces sous-échantillons est inférieure à 50 % (tel que prévu par H1). La plupart de ces valeurs de d étaient de l'ordre de 10 % ou plus. Des valeurs de d positives significatives (inattendues) plus faibles (de l'ordre de 5 %) sont observées pour trois pays qui ont utilisé l'ERP (Belgique et Norvège pour l'ESS1, Finlande pour l'ESS2). Cependant, ces différences n'ont pas pu être discernées dans d'autres vagues de l'enquête.

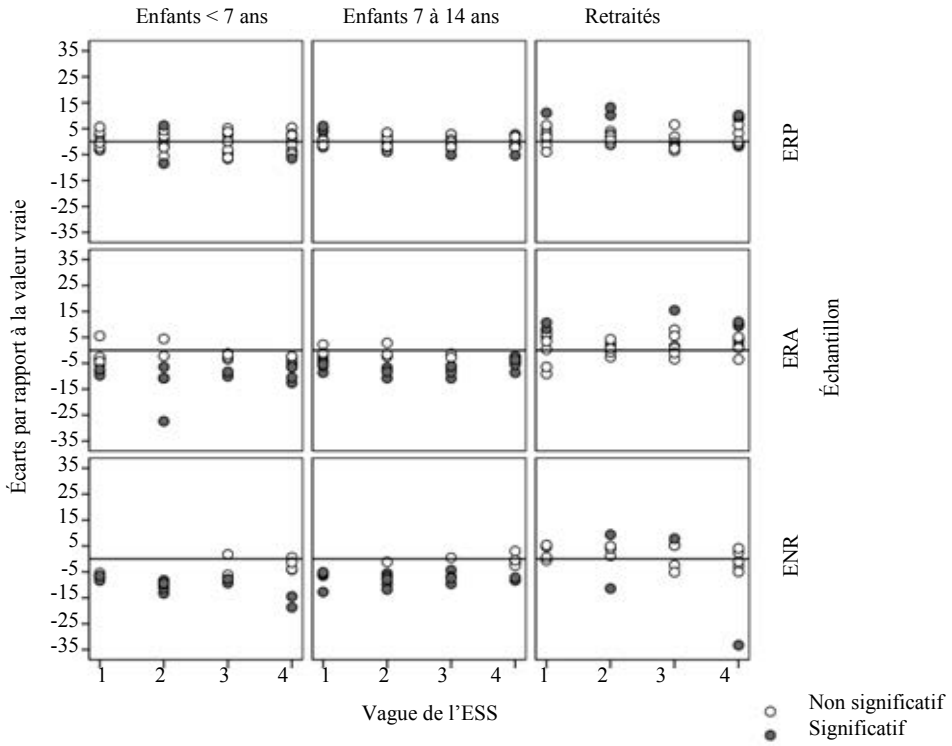

Figure 5.1 Écarts des pourcentages d'hommes par rapport à la valeur vraie de 50 % (d) pour différents types de ménages des vagues ESS1 à ESS4

L'examen des résultats pour les sous-échantillons de ménages dont les conjoints sont à l'âge de la retraite (retraités) permet de constater des valeurs de d significativement élevées (de l'ordre de 10 % ou plus) ayant le signe prévu (positif, ce qui signifie que les pourcentages d'hommes sont supérieurs à 50 %) pour certains pays dans le cas de toutes les méthodes d'échantillonnage (dans l'ESS1, pour la Norvège, la République tchèque et les Pays-Bas; dans l'ESS2, pour la Norvège, la Pologne et la France; dans l'ESS3, pour Chypre et la Russie; dans l'ESS4, pour l'Allemagne, la Hongrie, Chypre et le Royaume-Uni). Curieusement, la proportion d'hommes est nettement inférieure à 50 % en Slovaquie dans l'ESS4 (de l'ordre de 33 %) et pour le Portugal dans l'ESS2 (de l'ordre de 11 %). Ce résultat peut être expliqué par

des modèles particuliers de répartition des rôles entre les conjoints. Ici, la femme semble représenter le ménage, même si l'homme est à la maison.

En résumé, les écarts significatifs par rapport à la valeur vraie dans différents types de ménages étaient principalement en harmonie avec les attentes de l'hypothèse H1.

5.2 Différences entre les méthodes d'échantillonnage

L'effet de la méthode d'échantillonnage (tel que prévu par H2) a été testé par la méthode d'analyse de covariance multivariée (MANCOVA). Les valeurs de d pour les trois types de ménages (les trois sous-échantillons sélectionnés) ont été considérées comme étant les valeurs de trois variables dépendantes qui ont été analysées simultanément par la méthode MANCOVA. Puisque cette dernière est basée sur une analyse des moyennes, nous avons considéré les valeurs absolues de d. Autrement, il aurait été impossible de tenir compte des différences dont le signe était inattendu, qui seraient également associées à l'effet de la méthode d'échantillonnage. Puisque la plupart des différences étaient négatives dans les sous-échantillons composés de ménages avec enfants, les valeurs absolues de d représentent une proportion d'hommes inférieure à 50 %. Dans le cas des sous-échantillons composés de conjoints à l'âge de la retraite, il faut tenir compte du fait que la proportion d'hommes n'était pas seulement supérieure à 50 %, mais qu'elle était aussi plus faible que 50 % au Portugal (ESS2) et en Slovaquie (ESS4). En outre, les différences significatives ainsi que non significatives sont prises en considération pour pouvoir comparer les pays pour lesquels les valeurs de d sont faibles et ceux pour lesquels elles sont élevées.

La MANCOVA a mis en évidence un effet multivarié significatif élevé du facteur « méthode d'échantillonnage » (Lambda de Wilks (LW) $F_{(6,174)} = 6,87, p < 0,001,$ ampleur de l'effet $\eta^2 = 0,21$). En revanche, aucun résultat significatif n'a été observé pour les variables explicatives ($p > 0,10,$ $\eta^2 \max = 0,04$). Afin de prendre en considération les valeurs de d dans différents types de ménages, on a procédé à des analyses de covariance univariées (ANCOVA). L'homogénéité de la variance – en tant que présupposé pour une ANCOVA – est donnée selon le test de Levene pour le sous-échantillon de retraités, et aussi selon le test $F_{\max}$ pour les sous-échantillons de ménages avec enfants. Les ANCOVA ont révélé des différences significatives entre les moyennes des valeurs de d pour les diverses méthodes d'échantillonnage dans les deux sous-échantillons composés de ménages avec enfants (tableau 5.1). Les variances expliquées dans les ANCOVA pour ces sous-échantillons sont assez élevées (voir R^2 dans le tableau 5.1). En moyenne, la valeur de d la plus faible est observée pour l'ERP, tandis que la valeur la plus élevée est observée pour l'ENR (tableau 5.1 et figure 5.2). Toutefois, des comparaisons post-*hoc* simples en utilisant les sous-échantillons avec enfants ne révèlent des différences significatives qu'entre l'ERP et les deux autres méthodes d'échantillonnage (tableau 5.2). En outre, aucune différence remarquable entre les valeurs de d n'a été constatée entre les pays utilisant la méthode LAE et ceux utilisant l'échantillonnage par marche aléatoire.

Dans l'ensemble, les résultats montrent que l'hypothèse H2 est appuyée partiellement si l'on prend en considération les ménages avec enfants.

Tableau 5.1
Statistique descriptive $(M(E.T.))$ et résultats des ANCOVA pour la comparaison de d entre les trois types de ménages

	Types de ménages			
	Enfants < 7 ans	**Enfants 7 à 14 ans**	**Retraités**	n **(pays)**
Méthode d'échantillonnage (traitement)				
ERP	3,28(2,07)	2,21(1,37)	3,34 (3,35)	43
ERA	6,61(4,98)	4,87 (2,74)	4,94(3,83)	31
ENR	7,85 (4,4)	5,92 (3,55)	5,78(6,87)	21
$F(ddl_1 = 2, ddl_2 = 88)$	14,52***	20,9***	1,93	
Période : vague de l'ESS				
1	4,49(2,67)	4,08(2,94)	4,75(3,22)	22
2	6,92(5,73)	4,33(3,3)	3,63(3,71)	24
3	4,78(3,04)	4,02(3,18)	3,74(3,44)	23
4	5,23(4,41)	3,24(2,22)	5,39(6,66)	26
$F(ddl_1 = 1, ddl_2 = 88)$	0,00	1,18	0,02	
Paiement d'une prime				
Non	5,83(4,37)	4,41(3,10)	4,10(3,73)	54
Oui	4,78(3,99)	3,23(2,52)	4,81(5,49)	41
$F(ddl_1 = 1, ddl_2 = 88)$	0,57	$3{,}21^{+}$	0,49	
Ratio d'éléments sélectionnés				
$F(ddl_1 = 1, ddl_2 = 88)$	0,11	0,51	1,09	
Ratio de résultats confirmés				
$F(ddl_1 = 1, ddl_2 = 88)$	$3{,}11^{+}$	0,11	0,00	
R^2	0,22	0,31	0,01	

Notes $***p < 0,001$, $^{+}p < 0,10$.

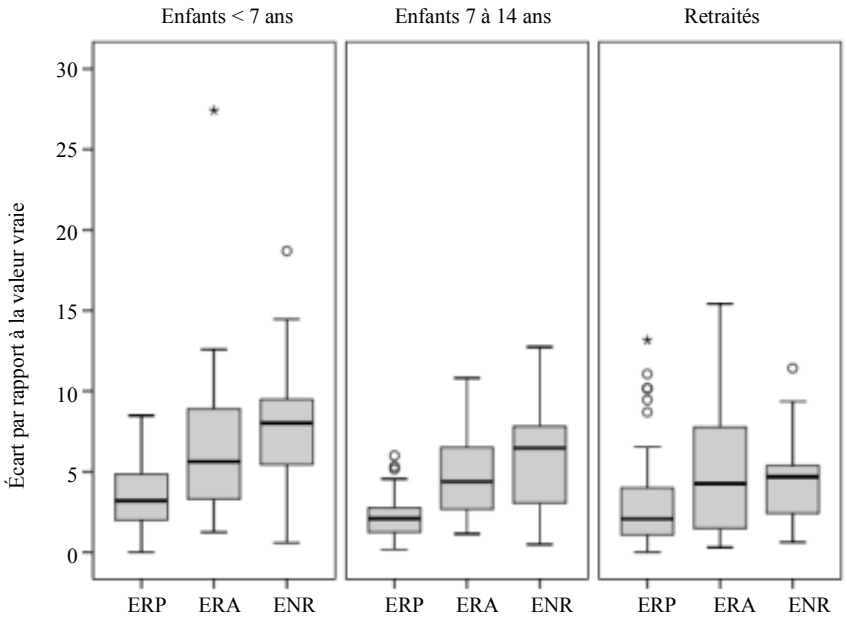

Figure 5.2 Boîtes à moustaches pour les valeurs absolues de d présentées pour différentes méthodes d'échantillonnage dans les trois types de ménages

Tableau 5.2
Différences des moyennes de $(d\,(E.T.))$ entre les méthodes d'échantillonnage dans les sous-échantillons avec enfants

	Enfants < 7 ans	Enfants 7 à 14 ans
Différences entre		
ERP et ERA	-3,34 (0,89)**	-2,66 (0,58)**
ERP et ENR	-4,58 (1,0)**	-3,71 (0,65)**
ERA et ENR	-1,24 (1,07)	-1,05 (0,7)

Note ** $p < 0,01$. Tests post-*hoc* pour comparaison simple avec correction de Bonferroni.

5.3 Effet des variables explicatives

L'effet des variables explicatives a été analysé pour tester l'hypothèse H3, qui prévoit que les écarts par rapport au ratio hommes-femmes de 50/50 seront stables au cours du temps et corrélés au mode de rémunération, aux contrôles de l'intervieweur et au changement d'organisme de collecte des données.

Certains pays qui participent à l'ESS ont procédé à un changement de méthode d'échantillonnage et/ou d'organisme de collecte des données entre les vagues (voir annexe). Les résultats ont montré que ni les effets multivariés $(\mathrm{LW}\,F_{(3,85)} = 0,81, p > 0,10)$ ni les effets univariés ne sont significatifs pour le changement d'organisme de collecte des données. Donc, le tableau 5.1 présente les résultats de l'ANCOVA sans cette variable. Si la variable « changement d'organisme de collecte des données » est incluse dans les analyses, l'effet de la variable « ratio de résultats confirmés » n'est plus significatif, mais cela n'a pas d'incidence sur les effets de toute autre variable. Ce résultat montre qu'un changement d'organisme de collecte des données peut être corrélé aux procédures de contrôle. Les différences entre les valeurs de d d'une vague à l'autre de l'ESS ne sont pas significatives non plus, que ce soit dans l'analyse multivariée $(\mathrm{LW}\,F_{(3,86)} = 0,51, p > 0,10)$ ou dans les analyses univariées (pour ces dernières, voir le tableau 5.1).

Le tableau 5.1 montre que, dans les sous-échantillons avec enfants, les moyennes de la valeur de d sont plus faibles si une prime est versée que si elle ne l'est pas. Cependant, cette différence n'est significative qu'à un seuil de signification de 10 % $(p < 0,10)$ et seulement pour les ménages avec enfants plus âgés. En conséquence, ce résultat montre que le mode de rémunération peut jouer un rôle et réduire l'écart par rapport à la valeur vraie lorsque la rémunération est plus élevée.

En ce qui concerne les procédures de contrôle, le nombre de contrôles (« ratio d'éléments sélectionnés ») n'est pas relié à la valeur de d (tableau 5.1). Le taux de succès des contrôles (« ratio de résultats confirmés ») est quant à lui relié à la valeur de d dans le sous-échantillon avec enfants de moins de sept ans. Cette relation est négative $(B = -0,06; E.T. = 0,04)$, ce qui signifie que plus les taux de contrôle confirmés sont faibles, plus les valeurs de d sont élevées. Cependant, cette relation n'est, elle aussi, significative qu'à un seuil de signification de 10 %.

En ce qui concerne l'hypothèse H3, on a montré que l'effet des méthodes d'échantillonnage est indépendant de l'effet temporel. Les résultats corroborent les prévisions de H3 au sujet du mode de rémunération et des contrôles de l'intervieweur. Cependant, les résultats pour ces variables indiquent que ces effets sont faibles et qu'ils ne peuvent être constatés que pour certains types de ménages.

6 Résumé et conclusion

Les résultats de la présente étude indiquent que des écarts significatifs par rapport au paramètre de population (50 % d'hommes) ont été observés dans de nombreux pays participant à l'ESS et que ces écarts étaient associés à la facilité de prendre contact avec les conjoints ou à la coopération de ces derniers dans les couples hétérosexuels (appui pour l'hypothèse H1). La grandeur de ces écarts variait en fonction de la méthode d'échantillonnage lorsque les conjoints différaient également en ce qui concerne la facilité de prendre contact avec eux (dans les sous-échantillons avec enfants). Donc, l'hypothèse H2 est partiellement corroborée. Dans les sous-échantillons avec enfants, l'ERP était la méthode d'échantillonnage associée aux données de la plus haute qualité puisque les écarts qu'elle produisait par rapport au paramètre de population étaient les plus faibles. Cependant, les résultats pour les sous-échantillons composés de conjoints retraités montrent que des écarts très prononcés sont également possibles sous ERP.

Les résultats pour les sous-échantillons avec enfants appuient l'explication voulant que le comportement des intervieweurs concernant les substitutions joue un rôle, puisque les écarts prévus par rapport au paramètre de population variaient en fonction du degré de liberté dont disposait l'intervieweur pour influer sur la réalisation de l'échantillon. Des résultats comparables ont été signalés par Sodeur (1997) et Kohler (2007). Il est moins plausible d'expliquer le fait que les hommes sont interviewés moins de 50 % du temps par une différence d'attitude concernant le refus de participer puisque, lorsqu'il en est ainsi, on s'attend à ce que la proportion d'hommes soit supérieure à 50 %. La différence d'attitude concernant le refus de participer ne devrait pas non plus varier selon la méthode d'échantillonnage. Pour les retraités, l'interview d'hommes plus de 50 % du temps a été observée dans plusieurs pays, mais dans des vagues isolées de l'enquête seulement. Cette faible stabilité des écarts par rapport au ratio hommes-femmes de 50 % peut aussi être associée à l'effet de l'intervieweur plutôt qu'à la différence d'attitude concernant le refus de participer, puisque cette dernière serait relativement stable dans un pays au cours de la période de référence de l'analyse. Cependant, puisque la présente étude n'a pas été réalisée selon un plan d'expérience, il est important de poursuivre l'étude des effets de la différence d'attitude à l'égard du refus de participer et des substitutions afin de pouvoir mieux les différencier et d'examiner les relations causales.

Même si les écarts par rapport au paramètre de population de 50 % variaient dans certains pays d'une vague à l'autre, dans l'ensemble, leur grandeur ne variait pas de manière significative au fil du temps, malgré l'amélioration des procédures de collecte de données de l'ESS (voir Koch et coll. 2009). De surcroît, les écarts par rapport au paramètre de population ne dépendaient pas de l'organisme chargé de la collecte des données ni du pays.

Les résultats donnent aussi à penser que le mode de rémunération et les procédures de contrôle de l'intervieweur peuvent réduire les cas de substitution. Cependant, il convient de souligner que la possibilité de prendre en compte le mode de rémunération et les procédures de contrôle a été limitée en raison de la faible variation dans les données ou du peu d'information disponible dans la documentation de l'enquête.

Il faut aussi tenir compte du fait que les résultats présentés ici sont fondés sur des sous-échantillons particuliers et ne peuvent pas être généralisés à l'échantillon complet de l'ESS. Cependant, l'absence de biais dans les sous-échantillons ne garantit pas l'absence de biais dans l'échantillon complet (Kohler 2007,

page 55). En outre, une analyse portant sur des groupes particuliers peut souvent présenter un intérêt (p. ex., savoir quelles sont les opinions des parents avec enfants ou des personnes salariées).

Les résultats de la présente étude donnent à penser que l'ERP est associé à un échantillon de meilleure qualité, de sorte que le biais de non-représentation dans les enquêtes transculturelles est plus faible que si l'on utilise les autres méthodes d'échantillonnage. La présente étude fait ressortir ce fait plus clairement que les études antérieures. Les analyses portant sur les données ALLBUS réalisées par Sodeur (2007) et par Hoffmeyer-Zlotnik (2006) comparaient plusieurs vagues d'une seule enquête dans un seul pays (Allemagne); dans l'analyse de Kohler (2007), l'effet de la méthode d'échantillonnage était confondu avec l'effet de l'enquête (voir la section 1). Ces situations ont été évitées dans l'analyse présentée ici.

Pour conclure, des écarts significatifs par rapport au paramètre de population, qui semblent être associés à des substitutions effectuées par les intervieweurs, ont été observés dans de nombreux pays participant à l'ESS. Afin de réduire cet effet d'intervieweur, il est préférable d'utiliser des méthodes d'échantillonnage, telles que l'ERP, qui réduisent le degré de liberté dont disposent les intervieweurs pour sélectionner les répondants et influer sur la qualité de l'échantillon. En outre, les procédures d'enquête qui accroissent la motivation des intervieweurs à produire des données d'enquête exactes sont très pertinentes et doivent être prises en considération dans de futures études, ainsi que dans les pratiques d'enquête.

Annexe

Codes des pays participant à l'ESS, changement de méthode d'échantillonnage et d'organisme chargé de la collecte des données, et âge légal de la retraite des hommes et des femmes dans chaque pays

Code de pays : ISO 3166-1	Pays	Changement de méthode d'échantillonnage (entre les vagues)	Changement d'organisme de collecte des données : entre les vagues	Âge légal de la retraite	
				Hommes	Femmes
BE	Belgique		1-2; 2-3	65	65
BG	Bulgarie		3-4	63	60
DE	Allemagne			65	65
DK	Danemark			65	65
EE	Estonie		2-3; 3-4	63	60
ES	Espagne	ERA-ENR (1-2)	2-3	65	65
FI	Finlande			65	65
HU	Hongrie	ERP – ERA (2-3) et inverse (3-4)	2-3	62	62
NO	Norvège			67	67
PL	Pologne			65	60
SE	Suède			65	65
SI	Slovénie			63	60
SK	Slovaquie	ERP – ENR (3-4)	2-3; 3-4	62	59
CH	Suisse			65	64
CZ	République tchèque	ERA-ENR (1-2)	1-2	65	62
CY	Chypre		3-4	65	65
GB	Royaume-Uni		1-2; 3-4	65	60
GR	Grèce	ENR-ERA (2-4)		65	60
IE	Irlande		3-4	65	65
IL	Israël			67	64
IT	Italie			65	60
LU	Luxembourg			65	65
NL	Pays-Bas			65	65
TR	Turquie			47	44
AT	Autriche			65	60
FR	France			60	60
PT	Portugal			65	65
RU	Fédération russe			60	55
UA	Ukraine			60	55

Notes Sources des données sur l'âge légal de la retraite :

1) http://www.oecd-ilibrary.org/finance-and-investment/pensions-at-a-glance-2011_pension_glance-2011-en

2) http://ec.europa.eu/employment_social/missoc/db/public/compareTables.do

3) Israel: http://www.btl.gov.il/

Bibliographie

AAPOR (2003). *Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection and Repair of Its Effects*. (http://www.aapor.org/pdfs/falsification.pdf; 14 mai 2009).

Arber, S. (2002). Design samples. Dans *Researching Social Life*, (Éd., N. Gilbert), Thousand Oaks: Sage, 58-84.

Billiet, J., Phillipsen, M., Fitzgerald, R. et Stoop, I. (2007). Estimation of nonresponse bias in the European Social Survey: Using information from reluctant respondents. *Journal of Official Statistics*, 23, 135-162.

Couper, M.P., et Groves, R.M. (1992). Le rôle de l'intervieweur dans la participation aux enquêtes. *Techniques d'enquête*, 18, 2, 279-294.

Couper, M.P., et De Leeuw, E.D. (2003). Nonresponse in cross-cultural and crossnational surveys. Dans *Cross-Cultural Survey Methods*, (Éds., J.A. Harkness, F.J.R. van de Vijver et P.Ph. Mohler), New York: John Wiley & Sons, Inc., 157-177.

De Heer, W. (1999). International response trends: Results of an international survey. *Journal of Official Statistics*, 15, 129-142.

De Heer, W., et Israis, A.Z. (1992). Response Trends in Europe. Article presenté à l'American Statistical Association, août 1992.

De Leeuw, E., et De Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. Dans *Survey Nonresponse*, (Éds., R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little), New York: John Wiley & Sons, Inc., 41-54.

De Leeuw, E.D., et Hox, J. (1996). The effect of the interviewer on the decision to cooperate in a survey of the elderly. Dans *International Perspectives on Nonresponse*, (Éd., S. Laaksonen). Helsinki: Statistics Finland, 46-52.

Dohrenwend, B.S., et Dohrenwend, B.P. (1968). Sources of refusals in surveys. *The Public Opinion Quarterly*, 32(1), 74-83.

Durrant, G.B., Groves, R.M., Staetsky, L. et Steele, F. (2010). Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly*, 74, 1-36.

ESS1-ESS4 data from Contact forms. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services. Extrait le 8 juin 2013 de: http://ess.nsd.uib.no/ess/round1/download.html.

European Social Survey Round 1-4 Data (2011). Data file edition ESS1-4e01.0_F1. Norwegian Social Science Data Services, Norway – Data Archive and distributor of ESS data.

European Social Survey (2011). ESS-4 2008 Documentation Report. Edition 4.0. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.

European Social Survey (2011). ESS-3 2006 Documentation Report. Edition 3.3. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.

European Social Survey (2011). ESS-2 2004 Documentation Report. Edition 3.3. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.

European Social Survey (2011). ESS-1 2002 Documentation Report. Edition 6.2. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. et Tourangeau, R. (2004). *Survey Methodology*. New Jersey: Wiley.

Haeder, S., et Lynn, P. (2007). How representative can a multi-nation survey be? In *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, (Éds., R. Jowell, C. Roberts, R. Fitzgerald and E. Gillian), London *et al.*: Sage, 33-52.

Hoffmeyer-Zlotnik, J.H.P. (2006). Stichprobenziehung in der Umfragepraxis. Die unterschiedlichen Ergebnisse von Zufallsstichproben in face-to-face Umfragen. Dans *Stichprobenqualität in Bevölkerungsumfragen*, (Éds., F. Faulbaum and Ch. Wolf). Informationszentrum Sozialwissenschaften: Bonn, 19-36.

Hox, J.J., et De Leeuw, E.D. (2002). The Influence of interviewers' attitude and behavior on household survey nonresponse: An international comparison. Dans *Survey Nonresponse*, (Éds., R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little), New York: John Wiley & Sons, Inc., 103-118.

Johnson, T.P., O'Rourke, D., Burris, J. et Owens, L. (2002). Culture and survey nonresponse. Dans *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little), New York: John Wiley & Sons, Inc., 55-69.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Koch, A. (1995). Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994. *ZUMA-Nachrichten*, 36, 89-105.

Koch, A., Blom, A.G., Stoop, I. et Kappelhof, J. (2009). Data collection quality assurance in cross-national surveys at the example of the ESS. *Methoden Daten Analysen*, 3, 219-247.

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, 1, 2, 55-67.

Kreuter, F., et Kohler, U. (2009). Analyzing contact sequences in call record data. Potential and limitations of sequence indicators for nonresponse adjustments in the European Social Survey. *Journal of Official Statistics*, 25, 203-226.

Lynn, P., Haeder, S., Gabler, S. et Laaksonen, S. (2007). Methods for achiving equivalence of samples in cross-national surveys: The European Social Survey Experience. *Journal of Official Statistics*, 1, 107-124.

Pickery, J., et Loosveldt, G. (2002). A multilevel multinomial analysis of interviewer effects on various components of unit non response. *Quality and Quantity*, 36, 427-437.

Philippens, M., et Billiet, J. (2004). Monitoring and evaluating nonresponse issues and fieldwork efforts in the European Social Survey. Article présenté lors de la *European Conference on Quality and Methodology in Official Statistics*. Mainz, Allemagne.

Schnauber, A., et Daschmann, G. (2008). States oder traits? Was beeinflußt die Teilnahmebereitschaft an telefonischen Interviews? *Zeitschrift Für Empirische Sozialforschung*, 2, 97-123.

Schnell, R., Hill, P.B. et Esser, E. (2011). *Methoden der Empirischen Sozialforschung.* München: R. Oldenbourg Verlag.

Smith, T.W. (2007). Survey nonresponse procedures in cross-national perspective: The 2005 ISSP non-response survey. *Survey Research Methods*, 1, 45-54.

Sodeur, W. (1997). Interne kriterien zur Beurteilung von Wahrscheinlichkeitsauswahlen. *ZA-Information*, 41, 58-82.

Sodeur, W. (2007). Entscheidungsspielräume von Interviewern bei der Wahrscheinlichskeitsauswahl. *Methoden Daten Analysen*, 1, 2, 107-130.

Stoop, I.A.L. (2004). Surveying nonrespondents. *Field Methods*, 16, 23-54.

Stoop, I.A.L. (2005). *The Hunt for the Last Respondent*. The Hague: Social and Cultural Planning Office.

Symons, K., Matsuo, H., Beullens, K. et Billiet, J. (2008). *Response Based Quality Assessment in the ESS – Round 3: An Update for 19 countries*. London: Centre for Comparative Social Surveys, City University.

Vehovar, V. (1999). Field substitution and unit nonresponse. *Journal of Oficial Statistics*, 2, 335-350.

Williams, B., Irvine, L., McGinnis, A.R., McMurdo, M.E.T. et Crombie, I.K. (2007). When "no" might not quite mean "no"; the importance of informed and meaningful non-consent: results from a survey of individuals refusing participation in a health-related research project. *BMC Health Services Research*, 7, 59.

Techniques d'enquête, juin 2014
Vol. 40, N° 1, p. 137-147
Statistique Canada, N° 12-001-X au catalogue

Imputation multiple bayésienne pour des données catégoriques à grande échelle contenant des zéros structurels

Daniel Manrique-Vallier et Jerome P. Reiter[1]

Résumé

Nous proposons une approche d'imputation multiple des réponses manquant aléatoirement dans les enquêtes à grande échelle qui ne portent que sur des variables catégoriques présentant des zéros structurels. Notre approche consiste à utiliser des mélanges de lois multinomiales comme outils d'imputation et à tenir compte des zéros structurels en concevant les données observées comme un échantillon tronqué issu d'une population hypothétique ne contenant pas de zéros structurels. Cette approche possède plusieurs caractéristiques intéressantes : les imputations sont générées à partir de modèles bayésiens conjoints cohérents qui tiennent compte automatiquement des dépendances complexes et s'adaptent facilement à de grands nombres de variables. Nous décrivons un algorithme d'échantillonnage de Gibbs pour mettre en œuvre l'approche et illustrons son potentiel au moyen d'une étude par échantillonnage répété en utilisant des microdonnées de recensement à grande diffusion provenant de l'État de New York, aux États-Unis.

Mots-clés : classes latentes; log-linéaire; valeur manquante; mélange; multinomiale; non-réponse.

1 Introduction

De nombreux organismes réalisent des enquêtes qui comprennent un grand nombre de variables exclusivement catégoriques. Inévitablement, ces enquêtes souffrent de non-réponse partielle qui, non prise en compte, peut réduire la précision ou augmenter le biais (Little et Rubin 2002). L'une des approches pour traiter la non-réponse partielle est l'imputation multiple (Rubin 1987), suivant laquelle l'organisme procède à un échantillonnage répété à partir de lois de probabilité prédictives pour remplacer les réponses manquantes. Cela crée $M > 1$ ensembles de données complets qui peuvent être analysés ou diffusés aux membres du public. Si les modèles d'imputation satisfont certaines conditions (Rubin 1987, chapitre 4), les analystes des M ensembles de données complets peuvent faire des inférences valides en utilisant des méthodes et des logiciels statistiques pour données complètes. Pour une revue de l'imputation multiple, voir Rubin (1996), Barnard et Meng (1999), Reiter et Raghunathan (2007), et Harel et Zhou (2007).

L'imputation multiple peut généralement être mise en œuvre selon deux stratégies. La première consiste à postuler un modèle conjoint pour toutes les variables et à estimer ce modèle en utilisant des techniques bayésiennes, qui incluent habituellement une augmentation des données et un échantillonnage Monte Carlo par chaîne de Markov (MCMC). Les modèles conjoints habituels comprennent les modèles normaux multivariés pour les données continues et les modèles log-linéaires pour les données catégoriques (Schafer 1997). La deuxième stratégie consiste à adopter des approches basées sur des équations chaînées (Van Buuren et Oudshoorn 1999; Raghunathan, Lepkowski, van Hoewyk et Solenberger 2001; White, Royston et Wood 2011). L'analyste estime une série de modèles conditionnels univariés et impute les valeurs manquantes séquentiellement en se servant de ces modèles. Les modèles conditionnels types comprennent les régressions normales pour variables dépendantes continues et les régressions logistiques ou multinomiales logistiques pour les variables dépendantes catégoriques.

1. Daniel Manrique-Vallier est professeur adjoint au Département de Statistique, Indiana University, Bloomington, IN 47408. Courriel : dmanriqu@indiana.edu; Jerome P. Reiter est professeur *Mrs. Alexander Hehmeyer of Statistical Science*, Duke University, Durham, NC 27708-0251. Courriel : jerry@stat.duke.edu.

Comme l'ont mentionné Vermunt, Ginkel, der Ark et Sijtsma (2008) ainsi que Si et Reiter (2013), les stratégies axées sur des équations chaînées conviennent mal pour les ensembles de données catégoriques présentant des dépendances complexes. Pour toute régression logistique (multinomiale) conditionnelle, le nombre de modèles possibles est énorme si l'on tient compte des effets d'interaction potentiels. Spécifier minutieusement chaque modèle conditionnel demande beaucoup de temps sans que l'on soit certain d'obtenir un ensemble de modèles cohérents sur le plan théorique. En effet, de nombreux praticiens des équations chaînées utilisent pour cette raison des paramètres par défaut qui n'incluent que les effets principaux dans les modèles conditionnels. En excluant les interactions, les analystes risquent de produire des ensembles de données complets qui fournissent des estimations biaisées. Il convient de souligner que les mêmes difficultés de sélection des modèles frappent les approches fondées sur des modèles log-linéaires.

Pour éviter ces problèmes, Si et Reiter (2013) proposent une approche conjointe, entièrement bayésienne, de modélisation de l'imputation multiple basée sur des modèles à classes latentes pour les données catégoriques de haute dimensionnalité. L'idée est de modéliser le tableau de contingence implicite des variables catégoriques comme un mélange de lois multinomiales indépendantes, en estimant le mélange de lois non paramétriquement au moyen de lois *a priori* issues du processus de Dirichlet. Les mélanges de lois multinomiales peuvent décrire des dépendances arbitrairement complexes et les calculs sont commodes et rapides, de sorte qu'ils constituent des outils d'imputation multiple d'usage général efficace. Par exemple, Si et Reiter (2013) ont appliqué leurs modèles pour imputer les valeurs manquantes pour 80 variables catégoriques de la *Trends in International Mathematics and Science Study*.

Dans leur approche, Si et Reiter (2013) ne traitent pas la complication importante et répandue dans les données d'enquête due au fait que certaines combinaisons de variables pourraient être impossibles *a priori*. On donne à ces cas le nom de zéros structurels (Bishop, Fienberg et Holland 1975). Par exemple, aux États-Unis, il est impossible que des enfants de moins de 15 ans se marient. Des zéros structurels peuvent aussi résulter des enchaînements de questions dans les enquêtes. Les algorithmes d'imputation de Si et Reiter (2013), s'ils sont appliqués directement, permettent d'obtenir une probabilité non nulle pour les zéros structurels, ce qui à son tour biaise les estimations des probabilités pour les combinaisons possibles.

Dans le présent article, nous présentons le problème de modélisation conjointe entièrement bayésienne pour l'imputation multiple de grands ensembles de données catégoriques contenant des zéros structurels. Notre approche combine le modèle d'imputation à classes latentes de Si et Reiter (2013) et la méthode de traitement des zéros structurels élaborée par Manrique-Vallier et Reiter (à paraître en 2014). Au moyen de simulations, nous montrons que l'approche produit des ensembles de données multi-imputés qui ne violent pas les contraintes des zéros structurels et peuvent avoir des propriétés d'échantillonnage répété bien calé.

2 Modèle d'imputation bayésien à classes latentes avec zéros structurels

Supposons que nous avons un échantillon de n individus mesurés sur J variables catégoriques. Chaque individu est associé à un vecteur de réponses $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$, dont les composantes prennent des valeurs provenant d'un ensemble de L_j niveaux. Pour simplifier, nous étiquetons ces

niveaux en utilisant des nombres consécutifs, $x_{ij} \in \{1,\ldots,L_j\}$, de sorte que $\mathbf{x}_i \in \mathcal{C} = \prod_{j=1}^{J}\{1,\ldots,L_j\}$. Notons que $\mathcal{C}$ inclut toutes les combinaisons des J variables, y compris les zéros structurels, et que chaque combinaison $\mathbf{x}$ peut être considérée comme une cellule dans le tableau de contingence formé par $\mathcal{C}$. Soit $\mathbf{x}_i = \left(\mathbf{x}_i^{\text{obs}}, \mathbf{x}_i^{\text{manq}}\right)$, où $\mathbf{x}_i^{\text{obs}}$ inclut les variables dont les valeurs sont observées et $\mathbf{x}_i^{\text{manq}}$ inclut les variables dont les valeurs manquent. Enfin, soit $S = \{s_1,\ldots,s_C\}$, où $s_c \in \mathcal{C}$ et $c = 1,\ldots,C < |S|$, l'ensemble de cellules contenant un zéro structurel, c.-à-d. $\Pr(\mathbf{x}_i \in S) = 0$.

2.1 Modèles à classes latentes

Pour commencer, nous décrivons le modèle à classes latentes bayésien sans nous préoccuper des zéros structurels et sans aucune donnée manquante, c.-à-d. $\mathbf{x}_i = \mathbf{x}_i^{\text{obs}}$. Ce modèle est un mélange fini de produits de lois multinomiales,

$$p(\mathbf{x}\,|\,\boldsymbol{\lambda},\boldsymbol{\pi}) = f^{MCL}(\mathbf{x}\,|\,\boldsymbol{\lambda},\boldsymbol{\pi}) = \sum_{k=1}^{K}\pi_k \prod_{j=1}^{J}\lambda_{jk}\left[x_j\right], \tag{2.1}$$

où $\boldsymbol{\lambda} = \left(\lambda_{jk}[l]\right)$, avec tous les $\lambda_{jk}[l] > 0$ et $\sum_{l=1}^{L_j}\lambda_{jk}[l] = 1$. Ici, $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_K)$ avec $\sum_{k=1}^{K}\pi_k = 1$. Ce modèle correspond au processus générateur,

$$x_{ij}\,|\,z_i \overset{\text{indép}}{\sim} \text{Discrète}_{1:L_j}\left(\lambda_{jz_i}[1],\ldots,\lambda_{jz_i}\left[L_j\right]\right) \text{ pour tout } i \text{ et } j \tag{2.2}$$

$$z_i\,|\,\boldsymbol{\pi} \overset{\text{iid}}{\sim} \text{Discrète}_{1:K}(\pi_1,\ldots,\pi_K) \text{ pour tout } i. \tag{2.3}$$

En ce qui concerne la notation, soit $(\mathcal{X},\mathcal{Z})$ un échantillon de n variables obtenu au moyen de ce processus, avec $\mathcal{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)$ et $\mathcal{Z} = (z_1,\ldots,z_n)$. Pour K suffisamment grand, (2.1) peut représenter arbitrairement les lois conjointes de $\mathbf{x}$ (Suppes et Zanotti 1981; Dunson et Xing 2009). Et, en utilisant la représentation d'indépendance conditionnelle dans (2.2) et (2.3), le modèle peut être estimé et simulé efficacement même si J est grand.

Pour les lois *a priori* sur $\boldsymbol{\pi}$, nous suivons Si et Reiter (2013) et Manrique-Vallier et Reiter (à paraître en 2014). Nous avons

$$\lambda_{jk}[\cdot] \overset{\text{indép}}{\sim} \text{Dirichlet}\left(\mathbf{1}_{L_j}\right) \tag{2.4}$$

$$\pi_k = V_k \prod_{h<k}(1 - V_h) \tag{2.5}$$

$$V_k \overset{\text{iid}}{\sim} \text{Beta}(1,\alpha) \text{ pour } k = 1,\ldots,K-1; V_K = 1 \tag{2.6}$$

$$\alpha \sim \text{Gamma}(0,25;0,25) \tag{2.7}$$

Dans (2.4), les lois *a priori* sont équivalentes à des lois uniformes sur le support des $J \times K$ probabilités multinomiales conditionnelles et représentent donc de vagues connaissances *a priori*. La loi *a priori* pour

$\boldsymbol{\pi}$ dans (2.5) à (2.7) est un exemple de loi *a priori stick-breaking* de dimension finie (Sethuraman 1994; Ishwaran et James 2001). Comme il est discuté dans Dunson et Xing (2009) et Si et Reiter (2013), elle attribue habituellement $\mathcal{Z}$ à moins de K classes, ce qui réduit les calculs et évite le surajustement. Pour une discussion et une justification plus approfondies de ce modèle en tant qu'outil d'imputation, voir Si et Reiter (2013).

2.2 Modèles tronqués à classes latentes

Le modèle à classes latentes donné dans (2.1) ne spécifie pas naturellement les cellules contenant des zéros structurels *a priori*, parce qu'il repose sur la supposition d'une probabilité positive dans chaque cellule. Donc, pour représenter les tableaux contenant des zéros structurels, nous devons tronquer le modèle de sorte que

$$f^{\mathrm{MCLT}}(\mathbf{x} \mid \boldsymbol{\lambda}, \boldsymbol{\pi}, S) \propto 1\{\mathbf{x} \notin S\} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}\left[x_j\right]. \tag{2.8}$$

Comme le montrent Manrique-Vallier et Reiter (à paraître en 2014), l'obtention d'échantillons à partir de la loi *a posteriori* de paramètres $(\boldsymbol{\lambda}, \boldsymbol{\pi})$, conditionnellement à un échantillon $\mathcal{X}^1 = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, peut être facilitée considérablement par l'adoption d'une stratégie d'augmentation analogue à celles décrites dans Basu et Ebrahimi (2001) et dans O'Malley et Zaslavsky (2008). Nous considérons que $\mathcal{X}^1$ représente la part de variables qui n'ont pas été incluses dans l'ensemble S provenant d'un échantillon plus grand, $\mathcal{X}$, généré directement à partir de (2.1). Soit n_0, $\mathcal{X}^0$, et $\mathcal{Z}^0$ la taille (inconnue) de l'échantillon, les vecteurs de réponses et les étiquettes des classes latentes pour la partie de $\mathcal{X}$ qui n'est pas comprise dans S. En utilisant une loi *a priori* empruntée à Meng et Zaslavsky (2002), Manrique-Vallier et Reiter (à paraître en 2014) montrent que, si $p(N) \propto 1/N$, où $N = n_0 + n$, la loi *a posteriori* de $(\boldsymbol{\lambda}, \boldsymbol{\pi})$ sous le modèle tronqué (2.8) peut être obtenue en effectuant sur $\left(n_0, \mathcal{X}^0, \mathcal{Z}^0, \mathcal{Z}^1\right)$ l'intégration de la loi *a posteriori* sous le modèle d'échantillon augmenté.

En procédant ainsi, Manrique-Vallier et Reiter (à paraître en 2014) élaborent un algorithme efficace sur le plan des calculs pour traiter de grands ensembles de zéros structurels lorsqu'ils peuvent être exprimés comme l'union d'ensembles définis par des *conditions marginales*. Il s'agit d'ensembles définis en fixant certains niveaux pour un sous-ensemble de variables catégoriques, par exemple, l'ensemble de toutes les cellules, de façon que $\{\mathbf{x} \in \mathcal{C} : x_3 = 1, x_6 = 3\}$. Manrique-Vallier et Reiter (à paraître en 2014) introduisent pour exprimer les conditions marginales une notation vectorielle que nous utilisons ici également. Soit $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_J)$ où, pour $j = 1, \ldots, J$, nous posons que $\mu_j = x_j$ quand x_j est fixé à un certain niveau et $\mu_j = *$ autrement, où $*$ est une notation spéciale représentant un paramètre substituable. En utilisant cette notation et en supposant que $J = 8$, les conditions qui définissent l'ensemble susmentionné servant d'exemple ($x_3 = 1$ et $x_6 = 3$) correspondent au vecteur $(*, *, 1, *, *, 3, *, *)$. Pour éviter d'encombrer la notation, nous utilisons les vecteurs $\boldsymbol{\mu}$ pour représenter les conditions marginales ainsi que les cellules définies par ces conditions marginales, le contexte permettant de déterminer s'il s'agit des premières ou des secondes.

2.3 Estimation et imputation multiple

Discutons maintenant de la façon d'estimer le modèle décrit à la section 2.2, puis de le convertir en un outil d'imputation multiple, lorsque certaines réponses manquent au hasard. La stratégie de base consiste à utiliser un échantillonneur de Gibbs. Étant donné un ensemble de données complet $\left(\mathbf{x}^{\text{obs}}, \mathbf{x}^{\text{manq}}\right)$, nous effectuons une sélection des paramètres en utilisant l'algorithme de Manrique-Vallier et Reiter (à paraître en 2014). Étant donné une sélection des paramètres, nous effectuons la sélection de $\mathbf{x}^{\text{manq}}$ tel que décrit plus bas.

Formellement, l'algorithme procède comme il suit. Supposons que l'ensemble de zéros structurels peut être défini comme l'union de C conditions marginales disjointes, $S = \cup_{c=1}^{C} \boldsymbol{\mu}_c$, et que nous utilisons les lois *a priori* pour α, $\boldsymbol{\lambda}$ et $\boldsymbol{\pi}$ définies à la section 2.1. Sachant $\mathbf{x}_i = \left(\mathbf{x}_i^{\text{obs}}, \mathbf{x}_i^{\text{manq}}\right)$ pour $i = 1, \ldots, n$, l'algorithme de Manrique-Vallier et Reiter (à paraître en 2014) échantillonne les paramètres comme il suit.

1. Pour $i = 1, \ldots, n$, échantillonner $z_i^1 \sim \text{Discrète}_{1:K}(p_1, \ldots, p_k)$, avec $p_k \propto \pi_k \prod_{j=1}^{J} \lambda_{jk}\left[x_{ij}^1\right]$.
2. Pour $j = 1, \ldots, J$ et $k = 1, \ldots, K$, échantillonner $\lambda_{jk[\cdot]} \sim \text{Dirichlet}\left(\xi_{jk1}, \ldots, \xi_{jkL_j}\right)$, avec $\xi_{jkl} = 1 + \sum_{i=1}^{n} 1\left\{x_{ij}^1 = l, z_i^1 = k\right\} + \sum_{i=1}^{n_0} 1\left\{x_{ij}^0 = l, z_i^0 = k\right\}$.
3. Pour $k = 1, \ldots, K-1$ échantillonner $V_k \sim \text{Beta}\left(1 + \nu_k, a + \sum_{h=k+1}^{K} \nu_k\right)$ où $\nu_k = \sum_{i=1}^{n} 1\left\{z_i^1 = k\right\} + \sum_{i=1}^{n_0} 1\left\{z_i^0 = k\right\}$. Poser que $V_K = 1$ et faire $\pi_k = V_k \prod_{h<k}(1 - V_h)$ pour tout $k = 1, \ldots, K$.
4. Pour $c = 1, \ldots, C$, calculer $\omega_c = \Pr(\mathbf{x} \in \boldsymbol{\mu}_c \mid \boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{\mu_{cj} \neq *} \lambda_{jk}\left[\mu_{cj}\right]$.
5. Échantillonner $(n_1, \ldots, n_C) \sim NM(n, \omega_1, \ldots, \omega_C)$, où NM est la loi multinomiale négative, et poser que $n_0 = \sum_{c=1}^{C} n_c$.
6. Poser que $\kappa \leftarrow 1$. Répéter ce qui suit pour chaque $c = 1, \ldots, C$.
   a) Calculer le vecteur normalisé $(p_1, \ldots, p_K)$, où $p_k \propto \pi_k \prod_{j:\mu_{cj} \neq *} \lambda_{jk}\left[\mu_{cj}\right]$.
   b) Répéter les trois étapes suivantes n_c fois :
      i. échantillonner $z_\kappa^0 \sim \text{Discrète}(p_1, \ldots, p_k)$,
      ii. pour $j = 1, \ldots, J$, échantillonner

$$x_{\kappa j}^0 \sim \begin{cases} \text{Discrète}_{1:L_j}\left(\lambda_{jz_\kappa^0}[1], \ldots, \lambda_{jz_\kappa^0}\left[L_j\right]\right) & \text{si } \mu_{cj} = * \\ \delta_{\mu_{jc}} & \text{si } \mu_{cj} \neq * \end{cases}$$

      où $\delta_{\mu_{cj}}$ est une distribution de masse ponctuelle à μ_{cj},

      iii. poser que $\kappa \leftarrow \kappa + 1$.
7. Échantillonner $\alpha \sim \text{Gamma}(a - 1 + K, b - \log \pi_K)$.

Après avoir échantillonné les paramètres, nous devons effectuer un tirage de $\mathbf{x}^{\text{manq}}$. Pour $i = 1, \ldots, n$, soit $\mathbf{m}_i = (m_{i1}, \ldots, m_{iJ})$ un vecteur tel que $m_{ij} = 1$ si la composante j dans $\mathbf{x}_i$ est manquante et $m_{ij} = 0$ autrement. En supposant que les données manquent au hasard, nous ne devons échantillonner que les composantes de chaque $\mathbf{x}_i$ pour lesquelles $m_{ij} = 1$, conditionnellement aux composantes pour lesquelles $m_{ij} = 0$. Donc, nous ajoutons une huitième étape à l'algorithme.

8. Pour $i = 1, \ldots, n$, échantillonner $\mathbf{x}_i^{\text{manq}}$ à partir de sa distribution conditionnelle complète,

$$p\left(\mathbf{x}_i^{\text{manq}} \middle| \ldots\right) \propto 1\{\mathbf{x}_i \notin S\} \prod_{j:m_{ij}=1} \lambda_{jz_i}\left[x_{ij}\right]. \tag{2.9}$$

En l'absence de zéros structurels, les x_{ij} qu'il faut imputer sont conditionnellement indépendants sachant z_i, ce qui transforme la tâche d'imputation en un exercice d'échantillonnage multinomial ordinaire (Si et Reiter 2013). Cependant, les zéros structurels que contient S induisent une dépendance entre les composantes. Donc, nous ne pouvons pas simplement échantillonner les composantes indépendamment les unes des autres. Une approche naïve consiste à utiliser un scénario d'acceptation-rejet en effectuant un échantillonnage répété à partir de la loi proposée $p\left(\mathbf{x}^{\text{manq}*}\right) = \prod_{j:m_{ij}=1} \lambda_{jz_i}\left[x_{ij}\right]$ jusqu'à l'obtention d'une variable telle que $\mathbf{x}^{\text{manq}*} \notin S$. Cependant, si la région de rejet est grande ou possède une probabilité élevée, cette approche peut être très inefficace.

Nous proposons plutôt de former des étapes d'échantillonnage de Gibbs supplémentaires, en calculant les lois conditionnelles de toutes les composantes manquantes afin de pouvoir les échantillonner individuellement. Soit $\text{Rep}(\mathbf{x}_i, j, l)$ le vecteur qui résulte du remplacement de la composante j dans $\mathbf{x}_i$ par une valeur arbitraire $l \in \{1, 2, \ldots, L_j\}$. La loi conditionnelle complète de la composante manquante j de $\mathbf{x}_i$ (quand $m_{ij} = 1$) est $p\left(x_{ij} \middle| \ldots\right) \propto 1\{\text{Rep}(\mathbf{x}_i, j, x_{ij}) \notin S\} \lambda_{jz_i}\left[x_{ij}\right]$. Donc, nous remplaçons l'étape 8 dans l'algorithme par

8'. Pour chaque $(i, j) \in \{(i, j) : m_{ij} = 1\}$, échantillonner $x_{ij} \sim \text{Discrète}_{1:L_j}\left(p_1, \ldots, p_{L_j}\right)$, où $p_l \propto \lambda_{jz_i}[l] 1\{\text{Rep}(\mathbf{x}_i, j, l) \notin S\}$.

La définition de p_l implique de tronquer le support de la loi conditionnelle complète de x_{ij}, c'est-à-dire $\{1, \ldots, L_j\}$, de manière à ne garder que les valeurs qui évitent $\mathbf{x}_i \in S$, sachant les valeurs courantes de $\{x_{ij'} : \text{tout } j' \neq j\}$.

Pour obtenir M ensembles de données complets à utiliser pour l'imputation multiple, les analystes sélectionnent M des $\mathbf{x}^{\text{manq}}$ échantillonnés après convergence de l'échantillonneur de Gibbs. Ces ensembles de données doivent être suffisamment espacés pour être approximativement indépendants (sachant $\mathbf{x}^{\text{obs}}$). Cela requiert de réduire les échantillons MCMC de manière que les autocorrélations entre les paramètres soient proches de zéro.

3 Étude par simulation

Afin d'illustrer empiriquement la performance de cet outil d'imputation, nous avons effectué une expérience d'échantillonnage répété en utilisant un extrait des 5 % de l'échantillon de microdonnées à grande diffusion provenant des données du Recensement des États-Unis de 2000 pour l'État de New York (Ruggles, Alexander, Genadek, Goeken, Schroeder et Sobek 2010). Les données comprennent $H =$ 953 076 individus et dix variables catégoriques : propriété du logement (3 niveaux), situation de prêt hypothécaire (4 niveaux), âge (9 niveaux), sexe (2 niveaux), état matrimonial (6 niveaux), identification unique de la race (5 niveaux), niveau d'études (11 niveaux), situation d'emploi (4 niveaux), situation d'incapacité d'emploi (3 niveaux) et situation d'ancien combattant (3 niveaux). Ces variables définissent un tableau de contingence contenant 2 566 080 cellules, dont 2 317 030 correspondent à des zéros structurels.

Nous traitons les H enregistrements comme une population à partir de laquelle nous tirons 500 échantillons indépendants de taille $n = 1\,000$. Pour chaque échantillon, nous imposons des données manquantes en remplaçant aléatoirement par un blanc 30 % des valeurs au niveau de la réponse enregistrées pour chaque variable. Puis nous estimons le modèle tronqué à classes latentes de la section 2.3 en utilisant 10 000 itérations MCMC et en écartant les 5 000 premières à titre de rodage. À partir de chaque chaîne restante, nous créons $M = 50$ ensembles de données complets par un échantillonnage systématique de chaque tranche de 100 itérations. Dans chacune des 500 simulations, nous utilisons un nombre maximal de classes latentes $K = 50$. Habituellement, le nombre effectif de composantes, c.-à-d. celles comprenant au moins un individu, est compris entre 10 et 15 (selon le sous-échantillon particulier) et ne dépasse pas 26.

Comme quantités à estimer, nous utilisons toutes les probabilités à trois dimensions dont les valeurs sont supérieures à 0,1 dans la population (les $H = 953\,076$ individus). Cela équivaut à 279 quantités à estimer. Dans chaque échantillon, nous estimons les intervalles de confiance à 95 % pour chacune des 279 probabilités en utilisant les règles de combinaison pour l'imputation multiple de Rubin (1987). Nous calculons aussi les intervalles correspondants au moyen des données avant l'introduction de valeurs manquantes, qui sont dénommées données complètes.

La figure 3.1 donne le pourcentage des 500 intervalles de confiance à 95 % qui couvrent leur valeur de population. Dans la plupart des cas, les taux de couverture simulés pour l'imputation multiple sont compris dans l'intervalle d'erreur Monte Carlo par rapport au niveau nominal. Quelques intervalles de confiance fondés sur l'imputation multiple ont un faible taux de couverture; en particulier, trois d'entre eux ont un taux inférieur à 85 % tandis que leurs analogues avec données complètes sont plus proches du niveau nominal. Cependant, comme le montre la figure 3.2, les grandeurs absolues des biais des estimations ponctuelles de ces quantités ont tendance à être modestes. Ces résultats encourageants corroborent les résultats présentés par Si et Reiter (2013), dont les simulations comprenaient jusqu'à 50 variables (sans aucun zéro structurel).

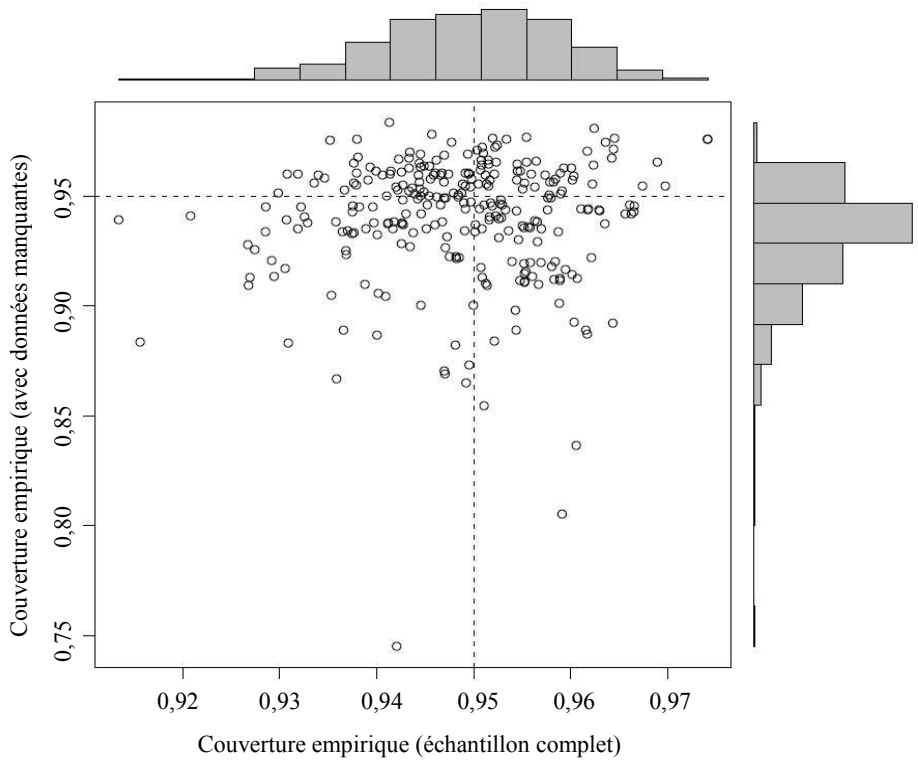

Figure 3.1 Comparaison des taux de couverture empiriques (sur 500 essais) des intervalles de confiance pour les estimations des probabilités marginales à trois dimensions calculées à partir des échantillons complets et à partir des ensembles de données traités par imputation multiple. Les droites en trait interrompu indiquent le niveau de couverture nominal. Un bruit aléatoire Unif (-0,004, 0,004) est ajouté par souci de clarté.

Pour chaque quantité à estimer, nous calculons aussi la fraction d'information manquante (FMI, Rubin 1987, page 77) moyenne estimée sur les 500 essais. Ces données sont présentées à la figure 3.3. La plupart des FMI moyennes sont proches du taux de réponses manquantes de 30 % que nous avons imposé sur chaque variable dans le plan de simulation. Cependant, bon nombre de FMI moyennes sont significativement inférieures à 30 %, y compris quatre qui sont exactement égales à zéro. Les quantités à estimer ayant une FMI moyenne significativement inférieure à 0,30 correspondent aux entrées des tableaux des probabilités marginales à trois dimensions où les zéros structurels restreignent sévèrement les imputations possibles. En fait, les zéros structurels réduisent la perte d'information due aux données manquantes. Par exemple, les quatre quantités à estimer avec une FMI moyenne $= 0$ correspondent à des combinaisons de variables pour lesquelles les restrictions ne laissent qu'un seul schéma d'imputation possible. Donc, aucune information n'est perdue même si les valeurs des données manquent effectivement. En incorporant les zéros structurels, nous imputons automatiquement les cas de ce type de manière appropriée et nous pouvons tirer parti de l'information fournie par les contraintes liées aux zéros structurels.

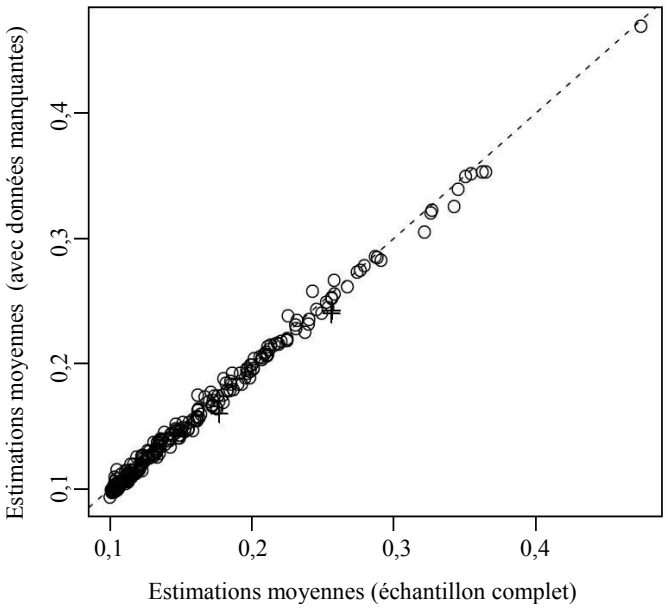

Figure 3.2 Estimations moyennes (sur 500 essais) des probabilités marginales à trois dimensions à partir des ensembles de données avec imputation multiple en fonction de celles calculées à partir des échantillons complets. Les points marqués d'une croix sont les estimations pour lesquelles la couverture empirique des intervalles de confiance à 95 % fondés sur l'imputation multiple est inférieure à 85 %.

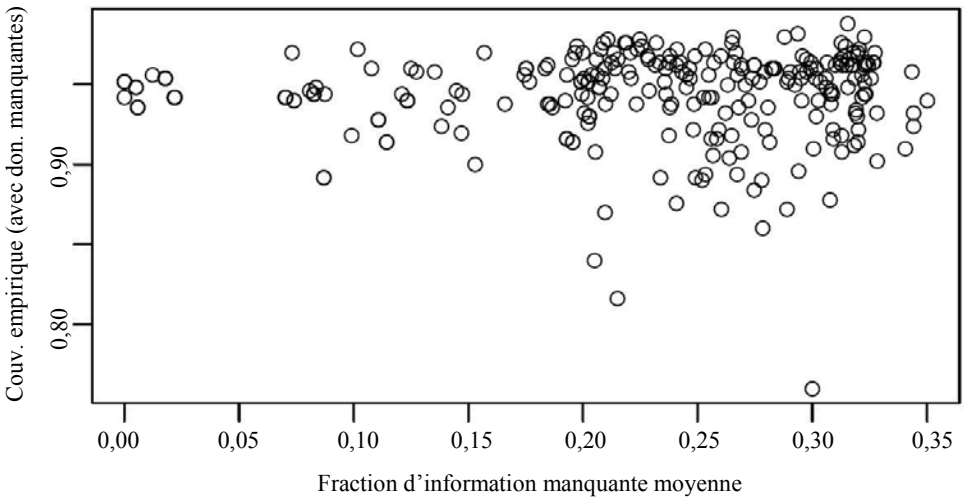

Figure 3.3 Taux de couverture empiriques (sur 500 essais) des intervalles de confiance pour 279 estimations des probabilités marginales à trois dimensions calculées à partir des ensembles de données traités par imputation multiple en fonction de la fraction d'information manquante moyenne estimée (sur les 500 essais) correspondante.

4 Conclusion

Les contraintes liées aux zéros structurels, comme les combinaisons impossibles et les enchaînements de questions, sont une caractéristique importante de nombreuses enquêtes. Elles jouent aussi un rôle clé dans l'imputation. Ne pas tenir compte des zéros structurels lors de l'estimation des modèles peut donner lieu à des biais importants lorsqu'on estime des quantités qui dépendent des probabilités conjointes ou conditionnelles. Cela se traduit par la génération de valeurs imputées qui ne reflètent pas exactement la structure de dépendance des données et qui peuvent subséquemment mener à des inférences biaisées en présence d'imputation multiple. En outre, les zéros structurels jouent souvent le rôle de règles de cohérence. Ne pas appliquer ces règles dans l'imputation pourrait donner des ensembles de données complets contenant des réponses incohérentes — comme des enfants veufs ou des locataires payant des impôts fonciers — que de nombreux organismes hésiteraient à diffuser et que de nombreux utilisateurs trouveraient difficiles à analyser. Fondée sur des modèles bayésiens tronqués à classes latentes, l'approche proposée ici offre aux spécialistes des sondages un moyen d'éviter ce genre de problèmes en utilisant des imputations multiples à partir de modèles cohérents sur le plan théorique et commodes sur le plan des calculs qui peuvent saisir des dépendances complexes et, simultanément, réduire les difficultés et les conjectures de spécification des modèles qui caractérisent souvent les approches classiques d'imputation multiple appliquées aux données catégoriques. Le code informatique en C++ et en R pour la mise en œuvre des algorithmes décrits dans le présent article peut être obtenu directement auprès des auteurs.

Remerciements

Cette étude a été financée par une subvention de la Fondation nationale des sciences (SES 11-31897).

Bibliographie

Barnard, J. et Meng, X. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8, 17-36.

Basu, S. et Ebrahimi, N. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, 88, 269-279.

Bishop, Y., Fienberg, S. et Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press, réimprimé en 2007, New York: Springer-Verlag.

Dunson, D. et Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104, 1042-1051.

Harel, O. et Zhou, X.H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in Medicine*, 26, 3057-3077.

Ishwaran, H. et James, L.F. (2001). Gibbs sampling for stick-breaking priors. *Journal of the American Statistical Association*, 96, 161-173.

Little, R.J.A. et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

Manrique-Vallier, D. et Reiter, J.P. (à paraître en 2014). Bayesian estimation of discrete multivariate truncated latent structure models. *Journal of Computational and Graphical Statistics*.

Meng, X.L. et Zaslavsky, A.M. (2002). Single observation unbiased priors. *The Annals of Statistics*, 30, 1345-1375.

O'Malley, A.J. et Zaslavsky, A.M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, 103, 1405-1418.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. et Solenberger, P. (2001). Une technique multidimensionelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Techniques d'enquête*, 27, 1, 91-103.

Reiter, J.P. et Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Ruggles, S., Alexander, T., Genadek, K., Goeken, R., Schroeder, M.B. et Sobek, M. (2010). Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]. University of Minnesota, Minneapolis. http://usa.ipums.org.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.

Si, Y. et Reiter, J.P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, à paraître.

Suppes, P. et Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese*, 48, 191-199.

Van Buuren, S. et Oudshoorn, C. (1999). Flexible multivariate imputation by MICE. Rapport technique, Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054.

Vermunt, J.K., Ginkel, J.R.V., der Ark, L.A.V. et Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38, 369-397.

White, I.R., Royston, P. et Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30, 377-399.