TECHNIQUES D'ENQUÊTE

STATISTICS CANADA STATISTIQUE CANADA
LIBRARY BIBLIOTHÈQUE
C3

N° 12-001-XPB au catalogue

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

JUIN 2002

•

VOLUME 28

•

NUMÉRO 1

Statistique Canada Statistics Canada

Canada

Ca OOS

STATISTICS CANADA LIBRARY
BIBLIOTHEQUE STATISTIQUE CANADA
1010350709

TECHNIQUES D'ENQUÊTE

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

JUIN 2002 • VOLUME 28 • NUMÉRO 1

Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, 2002

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrement sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juin 2002

N° 12-001-XPB au catalogue

Périodicité: semestrielle

ISSN 0714-0045

Ottawa

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans 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.

COMITÉ DE DIRECTION

Président G.J. Brackstone

Membres D.A. Binder
G.J.C. Hole
C. Patrick
R. Platek (Ancien président)
E. Rancourt (Gestionnaire de la production)
D. Roy
M.P. Singh

COMITÉ DE RÉDACTION

Rédacteur en chef M.P. Singh, *Statistique Canada*

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*
D.A. Binder, *Statistique Canada*
J.M. Brick, *Westat, Inc.*
C. Clark, *U.S. Bureau of the Census*
J.-C. Deville, *INSEE*
J. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
M.A. Hidiroglou, *Statistique Canada*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *Joint Program in Survey Methodology*
S. Linacre, *Official National Statistics*
G. Nathan, *Hebrew University, Israel*
D. Norris, *Statistique Canada*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
L.-P. Rivest, *Université Laval*
F.J. Scheuren, *National Opinion Research Center*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Rédacteurs adjoints J.-F. Beaumont, P. Dick, H. Mantel et W. Yung, *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 deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de *Techniques d'enquête* (n° 12-001-XPB au catalogue) est de 47 $ CA par année. Le prix n'inclus pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada: États-Unis 12 $ CA (6 $ × 2 exemplaires); autres pays, 20 $ CA (10 $ × 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et staticiens du Québec.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 28, numéro 1, juin 2002

TABLE DES MATIÈRES

Dans ce numéro 1

Article Sollicité Waksberg

W. FULLER
Estimation par régression appliquée à l'échantillonnage 5

Section spéciale "En souvenir de Leslie Kish"

GRAHAM KALTON
L'influence de Leslie Kish sur la statistique d'enquête 27

LESLIE KISH
Nouveaux paradigmes (modèles) pour l'échantillonnage probabiliste 33

CHARLES H. ALEXANDER
Les échantillons successifs de Leslie Kish et l'American Community Survey 39

JEAN-MICHEL DURR et JEAN DUMAIS
La rénovation du recensement français 47

Articles Réguliers

IAN CAHILL et EDWARD J. CHEN
Étalonnage des paramètres estimés des modèles logit de choix binaire et des modèles semiparamétriques de survie 55

STEVEN T. GARREN et TED C. CHANG
Estimation améliorée des ratios dans le cas des enquêtes téléphoniques avec correction pour la non-couverture 67

YVES TILLÉ
Estimation sans biais par calage sur la répartition dans les plans simples sans remise 83

JUN SHAO et SHAIL BUTANI
Estimation de variance dans le cadre de la « Current Employment Survey » 93

MICHAEL P. COHEN
Application de l'estimation de la variance selon Rao-Shao en utilisant des poids de rééchantillonnage 103

RICHARD VALLIANT
Estimation de la variance de l'estimateur de régression généralisée 109

Techniques d'enquête, juin 2002
Vol. 28, n° 1, pp. 1-2
Statistique Canada

Dans ce numéro

Ce numéro de *Techniques d'enquête* contient le deuxième d'une série d'articles annuels sollicités et publiés en l'honneur de Joseph Waksberg. Une brève description de la série et une courte biographie de Joseph Waksberg ont été présentées dans le numéro de juin 2001. L'auteur de l'article de 2002 de la série Waksberg est Wayne Fuller. J'aimerais remercier les membres du Comité, Graham Kalton (président), Chris Skinner, David Binder et Paul Biemer d'avoir choisi un statisticien aussi prestigieux, qui a largement contribué à de nombreux aspects de la théorie et de la pratique statistiques, comme auteur du deuxième article sollicité de la série Waksberg.

Dans son article intitulé *Estimation par régression appliquées à l'échantillonnage*, Wayne Fuller présente un aperçu général des faits historiques et des faits nouveaux en ce qui a trait à l'utilisation des modèles de régression dans les enquêtes pour l'estimation, la calibration des poids et le facteur de compensation de la non-réponse. Après une brève introduction et un aperçu historique, il aborde la question de l'utilisation des modèles de régression pour l'estimation dans le cadre d'enquêtes complexes du point de vue de la variance due au plan de sondage. Il enchaîne avec une étude du point de vue de la variance due au modèle. Il aborde également l'utilisation des modèles de régression pour les données multinominales, les techniques à employer quand on dispose de variables auxiliaires pour chaque unité de la population et la régression de manière à tenir compte des effets de la non-réponse dans les enquêtes. Enfin, en exposant certains des aspects pratiques des applications, l'auteur complète cet aperçu éclairant d'un important domaine d'inférence à partir de les données d'enquête auquel Wayne Fuller a lui-même fait de nombreux apports considérables.

Ce numéro renferme également une section spéciale intitulée *L'héritage de Leslie Kish* qui comprend quatre articles, dont un rédigé par Leslie Kish lui-même qui fait part de ses dernières pensées sur la combinaison des échantillons et des enquêtes. Deux autres articles traitent de la mise en œuvre de l'idée de Leslie Kish concernant les recensements consécutifs. Ces deux articles ont aussi été présentés dans le cadre du Symposium de 2001 de Statistique Canada lors d'une séance spéciale intitulée *L'héritage de Leslie Kish.*

Le premier article de la section spéciale, par Graham Kalton, présente un aperçu inspirant de l'apport de Kish à de nombreux secteurs de la statistique. Bon nombre des problèmes sur lesquels Kish a travaillé sont placés dans un contexte historique, et leur importance pratique est soulignée.

L'article de Kish présente des problématiques sur lesquelles il travaillait au moment de son décès survenu en octobre 2000. Je remercie Graham Kalton et Jack Gambino des corrections qu'ils ont apportées à l'article, bien que celui-ci soit reproduit en grande partie tel qu'il l'était au moment du décès de Kish. Dans l'article, il prétendait que, tout comme la statistique représentait un nouveau paradigme dans la méthode scientifique et que l'échantillonnage nécessitait un nouveau paradigme en statistique, les échantillons consécutifs et les enquêtes couvrant plusieurs populations nécessitent de nouveaux paradigmes en ce qui a trait aux méthodes d'enquête. Nous ne pouvons que formuler des hypothèses sur ce que l'article final aurait été si Kish vivait toujours.

Alexander décrit l'*American Community Survey*, que doit lancer le *U.S. Census Bureau* dans les prochaines années à titre de remplacement du questionnaire détaillé du recensement décennal. Il s'agit là d'une enquête très vaste qui se fonde en grande partie sur le concept d'échantillons et de recensements consécutifs que Kish a lancé il y plus de 20 ans. Cet article traite des concepts, de la base de sondage, du plan de sondage, du cumul d'échantillons et de la pondération.

Le dernier article de la section spéciale, rédigé par Durr et Dumais, décrit le nouveau recensement consécutif lancé en France dans le but de remplacer le recensement plus traditionnel. Dans le cadre de ce recensement consécutif, chaque petite commune fera l'objet d'une enquête une fois au cours d'une période de cinq ans; les grandes communes seront divisées en cinq groupes de renouvellement dont chacun fera l'objet d'une enquête une fois tous les cinq ans. Dans l'article, on traite des objectifs, des procédures quant au plan de sondage et à l'estimation qui s'appliquent aux recensements consécutifs.

Dans leur article, Cahill et Chen élaborent une méthode d'exploitation des données provenant de plusieurs enquêtes et périodes de référence grâce à l'étalonnage des paramètres estimés des modèles logit de choix binaire et des modèles semiparamétriques de survie. Les estimations calculées d'après les données d'une enquête riche en variables explicatives sont étalonnées en fonction des renseignements fournis par une enquête dont l'horizon temporel est considérable. Cahill et Chen montrent comment appliquer la méthode à l'aide du module sur le congé de maternité du projet de microsimulation dynamique LifePaths de Statistique Canada.

Garren et Chang examinent le problème de la population ne possédant pas de téléphone dans les enquêtes par téléphone au moyen de la composition aléatoire. D'après les échantillons de microdonnées à grande diffusion, la propension qu'un ménage ait un téléphone est estimée à l'aide d'une régression linéaire généralisée, propension dont on se sert pour l'estimation. Les biais et les variances asymptotiques sont présentés à la fois pour les estimateurs stratifiés a posteriori ainsi que non stratifiés a posteriori qui tiennent compte ou non de la propension estimée. Ces quatre estimateurs sont par la suite comparés dans le cadre d'une étude de simulation.

L'article de Tillé développe un estimateur qui permet d'éviter le problème des post-strates vides qui peut arriver avec l'estimateur post-stratifié usuel. L'idée consiste à utiliser un estimateur pondéré conditionnellement et à conditionner sur les rangs dans la population d'une variable auxiliaire connue pour toutes les unités de cette population. De cette façon, les tailles des post-strates sont fixées dans l'échantillon et sont aléatoires dans la population. Ensuite, on calcule la moyenne d'estimateurs pondérés conditionnellement pour obtenir plus de stabilité. L'estimateur obtenu est calé sur la répartition, linéaire et exactement sans biais. On montre au moyen d'une étude de simulation que l'estimateur proposé est plus robuste que l'estimateur par la régression généralisée quand la relation entre la variable d'intérêt et la variable auxiliaire n'est pas linéaire. On propose finalement un estimateur approché de la variance validé au moyen de simulations.

Dans leur article, Shao et Butani traitent du problème de l'estimation des variances dans le cas des estimateurs d'enquête imputés. Ils montrent que les variances qui résultent peuvent être estimées en deux volets : il s'agit d'abord de procéder à une estimation à l'aide d'une méthode de regroupement en demis-échantillons groupé qui incorpore des ajustements de sorte que l'on tienne compte de l'imputation. Comme l'estimation du deuxième volet peut donner lieu à de nombreuses dérivations, Shao et Butani proposent que l'on procède à une correction de la méthode de regroupement en demis-échantillons groupé qui donnerait lieu à des estimations des variances à peu près sans biais.

Dans son article, Cohen décrit une méthode pour mettre en œuvre la méthode jackknife de Rao et Shao concernant l'estimation des variances de manière à tenir compte de l'imputation de valeurs au moyen des poids de rééchantillonnage. La méthode de Rao et Shao suppose le calcul, pour chaque répétition selon la méthode du jackknife, des valeurs corrigées des données imputées. La méthode peut être utilisée soit avec l'imputation de la moyenne ou soit avec l'imputation par la méthode du hot deck. Pour appliquer la méthode de Cohen, il faut ajouter des rangées supplémentaires au fichier de poids de rééchantillonnage. Pour chaque valeur imputée, on ajoute une rangée supplémentaire pour chaque répondant faisant partie de la même classe d'imputation.

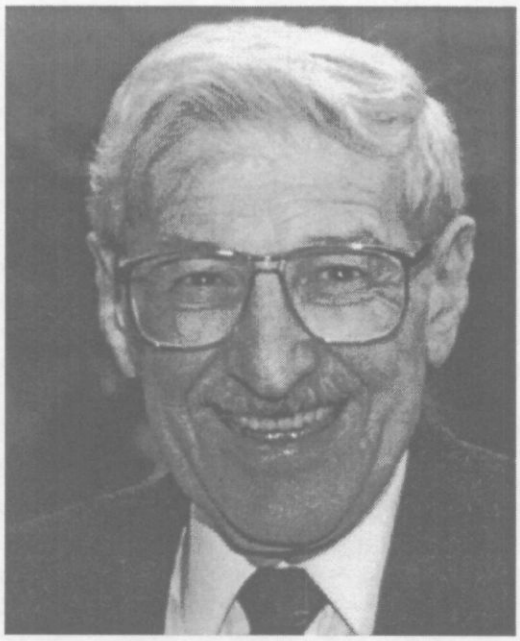
Dans le dernier article du présent numéro, Valliant étudie plusieurs estimateurs des variances pour l'estimateur de régression généralisée (GREG). Ce qui nous intéresse c'est de trouver des estimateurs de la variance qui, sous réserve de certaines conditions, ne renferment à peu près pas de biais tant du point de vue de la variance due au plan de sondage que du point de vue de la variance due au modèle même si le modèle qui sous-tend l'estimateur GREG renferme un paramètre de variance qui est faux. Soulignons qu'une caractéristique importante de ces estimateurs robustes est l'ajustement des carrés des résidus au moyen de facteurs analogues aux effets leviers utilisés en analyse par régression classique. Il semble que l'estimateur jackknife avec suppression d'une unité inclut les ajustements pour tenir compte des effets leviers et qu'il soit un bon choix du point de vue tant de la variance due au plan de sondage que de celle due au modèle. Une étude de simulation montre que ces estimateurs de la variance sont caractérisés par un biais faible et produisent des intervalles de confiance dont le taux de couverture est quasi nominal.

M.P. Singh

Techniques d'enquête, Juin 2002
Vol. 28, n° 1, pp. 3-4
Statistique Canada

Série Waksberg d'articles sollicités

Le comité de rédaction de *Techniques d'enquête* a décidé de publier une série d'articles annuels sollicités en l'honneur de Joseph Waksberg, pour souligner sa contribution importante à la méthodologie d'enquête. Chaque année nous inviterons un spécialiste renommé de la recherche en sondages à rédiger un article consacré à la rétrospective et à l'examen de la situation courante d'un domaine important de la méthodologie d'enquête. L'auteur reçoit un prix monétaire grâce à une subvention offerte par Westat en reconnaissance de la contribution de Joe Waksberg durant les nombreuses années où il a travaillé pour l'entreprise. L'*American Statistical Association* est chargée de la gestion financière et administrative de la subvention. L'auteur de l'article est choisi par un comité de quatre personnes désignées par *Techniques d'enquête* et l'*American Statistical Association*.



JOSEPH WASKBERG

ARTICLE SOLLICITÉ WAKSBERG 2002

Auteur : Wayne A. Fuller

Wayne A. Fuller est Professeur distingué émérite en statistique et économie à l'Iowa State University. Il a publié environ 100 articles dans plus de vingt revues et est auteur des livres *Introduction to Statistical Time Series* et *Measurement Error Models*. En tant que membre du Survey Group à l'Iowa State University, il a été principal responsable du développement de procédures d'estimation pour une grande enquête longitudinale appellée *U.S. National Resources Inventory*. Son intérêt pour la recherche en théorie des sondages comprend l'estimation par la régression, l'estimation pour petits domaines, l'imputation et l'échantillonnage à plusieurs phases. Il est présentement président du Comité consultatif des méthodes statistiques à Statistique Canada.

MEMBRES DU COMITÉ DE SÉLECTION DE L'ARTICLE WASKBERG (2002-2003)

David A. Binder (Président), *Statistique Canada*
J. Michael Brick, *Westat, Inc.*
David R. Bellhouse, *University of Western, Ontario*
Paul Biemer, *Research Triangle Institute, U.S.A.*

Présidents précédents:

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)

Auteurs précédents:

Gad Nathan (2001)

Nominations:

Les nominations d'individus à considérés comme auteurs ou les suggestions pour des sujets devraient être envoyés au président du comité, D.A. Binder, Statistique Canada, 3e, étage immeuble R.H. Coats, Parc Tunney, Ottawa, (Ontario), Canada, K1A 0T6, par courrier électronique: binderdav@statcan.ca ou par télécopieur (613) 951-5711. Les nominations et suggestions de sujets doivent être reçues au plus tard le 6 décembre 2002.

Techniques d'enquête, juin 2002
Vol. 28, n° 1, pp. 5-25
Statistique Canada

Estimation par régression appliquée à l'échantillonnage

WAYNE A. FULLER[1]

RÉSUMÉ

L'utilisation de la régression et des méthodes connexes est devenue courante dans le domaine de l'estimation à partir de données d'enquête. Nous passons en revue les propriétés fondamentales des estimateurs par régression, discutons de la mise en application de l'estimation par régression et étudions l'estimation de la variance des estimateurs par régression. Nous examinons aussi le rôle des modèles dans la construction des estimateurs par régression et l'utilisation de la régression en vue du rajustement pour tenir compte de la non-réponse.

MOTS-CLÉS : Information auxiliaire; calage; moindres carrés; convergence selon le plan; prédiction linéaire.

1. INTRODUCTION

Dans le domaine de l'échantillonnage, l'utilisation de renseignements sur la population étudiée est nécessaire lors de l'élaboration du plan d'échantillonnage et de l'estimation si l'on veut que les méthodes soient efficaces. Bien que ces deux activités soient intimement liées, puisque les estimateurs dépendent du plan d'échantillonnage, un traitement distinct leur est souvent réservé dans la documentation sur l'échantillonnage. Conformément à cette façon de faire, nous commençons par étudier l'estimation en traitant le plan d'échantillonnage tel que conçu. L'estimation consiste à combiner les renseignements disponibles sur la population et les données d'échantillon afin d'obtenir une bonne représentation des caractéristiques étudiées.

L'estimation par régression est l'une des méthodes importantes qui s'appuient sur l'information démographique ou sur celle provenant d'un échantillon plus grand pour construire des estimateurs efficaces. Cette information, parfois appelée *information auxiliaire*, peut être utilisée lors de l'élaboration du plan d'échantillonnage ou ne pas être disponible à cette étape. Lors d'une enquête sur une population humaine, l'information provient souvent de sources officielles, telles que le recensement national. Des sources comparables peuvent fournir l'information auxiliaire pour d'autres catégories d'enquêtes. Par exemple, dans le cas d'une enquête sur l'utilisation des terres, les données sur la superficie totale, la superficie appartenant au gouvernement fédéral et la superficie immergée en permanence peuvent être extraites des archives nationales.

Trois situations peuvent se présenter en ce qui concerne l'information auxiliaire dont on dispose. Dans la première, on connaît les valeurs du vecteur auxiliaire $\mathbf{x}$ pour chaque élément de la population au moment de la sélection de l'échantillon. Dans ces conditions, on peut utiliser la variable auxiliaire pour établir le plan d'échantillonnage.

Dans la deuxième situation, on connaît toutes les valeurs du vecteur $\mathbf{x}$, mais on ne peut associer une valeur particulière à un élément particulier tant que l'on n'a pas observé l'échantillon. Dans ces conditions, on ne peut utiliser l'information auxiliaire pour élaborer le plan d'échantillonnage, mais on dispose d'un éventail d'options d'estimation une fois que les observations sont disponibles. Par exemple, le recensement de la population peut fournir la répartition âge-sexe de la population, mais les organismes non gouvernementaux qui sélectionnent les échantillons ne disposent d'aucune liste de personnes et de leurs caractéristiques.

Dans la troisième situation, on connaît uniquement la moyenne de la population pour $\mathbf{x}$, ou la moyenne calculée d'après un grand échantillon. Dans ce cas, on ne peut utiliser l'information auxiliaire pour élaborer le plan d'échantillonnage et les méthodes d'estimation sont limitées. Par exemple, le U.S. Department of Agriculture pourrait diffuser une estimation du nombre total d'animaux d'un type particulier recensés sur les entreprises agricoles à une date donnée. Notre discussion se concentre sur cette dernière situation.

On peut aussi définir deux situations d'estimation. Dans l'une, on considère une seule variable et un paramètre, ou un très petit nombre de paramètres. L'analyste est disposé à consacrer beaucoup d'énergie à l'analyse, il possède un modèle de population bien formulé et il est prêt à appuyer la méthode d'estimation en établissant le caractère raisonnable du modèle. Dans la deuxième situation, l'analyste s'attend à exécuter un grand nombre d'analyses portant sur un grand nombre de variables. Aucun modèle particulier n'est considéré adéquat pour toutes les variables. Le cas où le spécialiste de l'échantillonnage prépare un ensemble de données qui doivent être analysées par d'autres spécialistes est un exemple typique de cette deuxième situation. Comme la personne qui prépare l'ensemble de données ne sait pas quelles sont les variables analytiques, l'accent est mis sur l'utilisation d'estimateurs que l'on peut défendre en recourant à un nombre minimal de modèles.

Les estimateurs par régression entrent dans la catégorie des estimateurs linéaires. Ces derniers présentent un avantage particulier dans le cas de l'échantillonnage, car,

[1] Wayne Fuller, Professeur distingué émérite Iowa State University, 221 Snedecor Hall, Ames, IA 50011-1210, U.S.A.

une fois calculés, les coefficients de pondération s'appliquent à toute variable analysée. Nous examinerons plusieurs propriétés des estimateurs au cours de notre discussion. En nous appuyant sur un modèle, nous acceptons l'objectif classique consistant à minimiser l'erreur quadratique moyenne pour une classe d'estimateurs. Cette classe peut être celle des estimateurs linéaires qui sont dépourvus de biais dans les conditions du modèle, mais elle peut aussi faire l'objet de restrictions supplémentaires.

Les estimateurs qui ne varient en fonction ni de l'échelle ni de la localisation peuvent être utilisés dans des conditions générales. Mickey (1959) a proposé qu'on réserve l'expression « estimateur par régression » aux estimateurs linéaires qui ne varient en fonction ni de la localisation ni de l'échelle. Bien que nous ne puissions respecter strictement cette définition, nous appuyons la distinction entre les estimateurs invariants en fonction de la localisation et de l'échelle et ceux qui ne le sont pas. Selon nous, l'invariance selon la localisation est une propriété importante dans le cas des plans d'échantillonnage où l'unité visée par l'analyse est également une unité d'échantillonnage. Dans le cas des plans d'échantillonnage en grappes ou à deux degrés pour lesquels les coefficients de pondération sont construits pour les totaux au niveau de l'unité primaire d'échantillonnage, l'invariance selon la localisation est moins importante.

Les modèles jouent un rôle considérable dans la construction des estimateurs par régression. Il est souhaitable que les propriétés des estimateurs restent bonnes même si la spécification du modèle n'est pas exacte. Par conséquent, les propriétés concernant la population finie visée par l'échantillonnage sont toutes aussi importantes que celles concernant le modèle.

Les estimateurs linéaires qui reproduisent la moyenne connue des variables auxiliaires sont dits étalonnés ou calés. Cette propriété est désirable car, par exemple, les totaux de marge des tableaux où une variable auxiliaire est une variable analysée concordent avec les totaux connus. Si la variable auxiliaire ne présente aucun intérêt analytique, le calage est moins important.

2. CONTEXTE

Jessen (1942) et Cochran (1942) comptent parmi les premiers auteurs qui ont mentionné l'estimation par régression dans le cas de l'échantillonnage. Toutefois, le recours à la régression dans des contextes comparables a fort probablement eu lieu plus tôt, Cochran (1977, page 189) parlant en effet d'une régression en vue d'estimer la superficie d'une feuille réalisée par Watson (1937). Il est intéressant de noter que Jessen a utilisé la régression essentiellement dans le cas d'une estimation composite où la régression était destinée à améliorer les estimations calculées à deux points dans le temps, étant donné des échantillons ayant certains éléments communs à chacun de ces points. Cochran (1942) a énoncé la théorie de base de la régression dans le cas de l'échantillonage d'enquête en s'appuyant fortement sur la théorie des modèles linéaires. Il a montré que le modèle linéaire ne doit pas nécessairement rester valable pour que l'estimateur par régression donne de bons résultats. Il a établi une expression $O(n^{-1})$ pour le biais et une approximation $O(n^{-2})$ pour la variance. Il a également montré que, dans le cas du modèle où la droite de régression passe par l'origine et la variance d'erreur est proportionnelle à x, l'estimateur par ratio est l'estimateur par les moindres carrés généralisés.

Durant les années 1950, l'estimation par régression a suscité un intérêt théorique qui a souvent pris la forme d'études du biais. À cet égard, consulter Mickey (1959). Brewer (1963) est l'un des premiers auteurs qui a considéré l'estimation linéaire lors de l'utilisation d'un modèle de superpopulation pour déterminer une méthode optimale. Il recherchait le plan d'expérience optimal pour l'estimateur par ratio et a examiné le conflit éventuel entre un plan optimal dans les conditions du modèle et un plan qui dépendrait moins de ce dernier. Consulter aussi Brewer (1979). Royall (1970) a soutenu que, dans le cas de l'utilisation de modèles, les propriétés conditionnelles importantes sont celles qui sont subordonnées à l'information auxiliaire contenue dans l'échantillon et que le plan d'échantillonnage devrait être choisi de façon à optimiser ces propriétés. Royall et ses collaborateurs, par exemple Royall et Cumberland (1981), ont étudié les propriétés conditionnelles des estimateurs par régression, découlant de *i.e.*, conditionnellement aux valeurs des variables auxiliaires obtenues dans l'échantillon.

Un grand nombre d'études effectuées durant les années 1970 et 1980 portaient sur la nature générale de l'estimateur par régression dans le contexte de l'échantillonnage et sur la mesure dans laquelle il est possible de rapprocher la méthode de prédiction par modèle et la perspective du plan d'échantillonnage. Fuller (1973, 1975) a énoncé les propriétés, dans le cas d'un grand échantillon, d'un vecteur de coefficients de régression calculé d'après un échantillon d'enquête. Isaki (1970) a étudié les estimateurs par régression et a publié des versions étoffées des résultats dans Isaki et Fuller (1982) et dans Fuller et Isaki (1981). Ces auteurs ont montré qu'un estimateur par régression construit dans les conditions d'un modèle est convergent par rapport au plan d'échantillonnage pour la moyenne de la population si le modèle contient certaines variables. Cassel, Särndal et Wretman (1976) ont tenu compte des éléments principaux du modèle et du plan d'échantillonnage pour construire l'estimateur et ont proposé l'expression « estimateurs par régression généralisée » pour les estimateurs de total convergents par rapport au plan d'échantillonnage de la forme

$$\hat{T}_{y,\mathrm{GREG}} = \hat{T}_{y,\mathrm{HT}} + (T_{x,N} - \hat{T}_{x,\mathrm{HT}})\hat{\boldsymbol{\beta}},$$

où $\hat{T}_{y,\mathrm{HT}}$ et $\hat{T}_{x,\mathrm{HT}}$ sont les estimateurs d'Horvitz-Thompson des totaux de y et x, respectivement, $T_{x,N}$ est le total connu de la population pour x et $\hat{\boldsymbol{\beta}}$ est un coefficient de régression

estimé. Särndal (1980), Wright (1983), ainsi que Särndal et Wright (1984) ont discuté des classes d'estimateurs par régression. L'ouvrage publié par Särndal, Swensson et Wretman (1992) contient une discussion détaillée de l'estimation par régression dont Mukhopadhyay (1993) a fait une revue.

Il a fallu attendre les années 1970 pour que l'on applique la méthode de régression aux données d'enquêtes générales à plusieurs variables et les années 1990, pour que l'application de la pondération par régression soit d'usage généralisé. Doane Agricultural Services Inc., appelé aujourd'hui Doane Marketing Research, est l'un des premiers organismes qui a utilisé les coefficients de régression pour la pondération. De 1971 à 1972, Doane a réalisé une étude d'audience en milieu agricole, à laquelle ont répondu 6 920 agriculteurs, sous la direction de M. John Wilkin. Les coefficients de pondération appliqués aux répondants ont été calculés par des méthodes de régression, en se servant de valeurs de contrôle provenant du U.S. Agricultural Census et du Department of Agriculture. Doane a offert un appui financier à la Iowa State University pour que celle-ci développe un programme de calcul de coefficients de pondération par régression. Lors de l'étude de Doane, pour s'assurer que les coefficients de pondération soient positifs, on a regroupé les observations pour lesquelles les coefficients de pondération étaient faibles et on leur a attribué un coefficient de pondération commun. Le regroupement s'est poursuivi jusqu'à ce que le coefficient commun soit positif. Plus tard, une version modifiée de la méthode de Huang et Fuller (1978) a été introduite dans les programmes informatiques pour garantir que les coefficients de pondération soient positifs. Doane applique les coefficients de pondération basés sur la régression pour ses études de marché souscrites depuis 1972.

Statistique Canada a utilisé l'estimation par régression pour la première fois en 1988 pour l'Enquête sur la population active du Canada. En 1992, l'estimation par régression a été appliquée lors du *Recensement de la population du Canada* de 1991 afin de s'assurer que la somme pondérée des valeurs des variables recueillies au moyen du questionnaire détaillé (rempli par un échantillon systématique au 1/5 de l'ensemble des ménages du Canada) soit égale aux totaux connus au niveau des ménages et de la population recueillis lors du Recensement de 1991. À cet égard, consulter Bankier, Rathwell et Majkowski (1992) et Bankier, Houle et Luc (1997). L'estimateur par régression est également la composante essentielle du Système généralisé d'estimation (SGE) développé par Statistique Canada et utilisé pour un grand nombre d'enquêtes auprès des entreprises et d'enquêtes sociales depuis sa diffusion en 1992. La méthodologie qui sous-tend le système est décrite dans Estevao, Hidiroglou et Särndal (1995). Consulter aussi Hidiroglou, Särndal et Binder (1995). L'estimation par régression est maintenant utilisée pour construire les estimateurs composites pour l'Enquête sur la population active du Canada. À cet égard, consulter Singh, Kennedy et Wu (2001), Gambino, Kennedy et Singh (2001), et Fuller et Rao (2001).

Bethlehem et Keller (1987) décrivent l'utilisation de l'estimation par régression au Bureau central de la statistique des Pays-Bas (appelé aujourd'hui Statistique Pays-Bas) dans un programme appelé LIN WEIGHT. Nieuwenbrock, Renssen et Hofman (2000) décrivent le progiciel Bascula, qui a remplacé LIN WEIGHT. Deville, Särndal et Sautory (1993) décrivent un programme informatique appelé CALMAR mis au point par l'Institut national de la statistique et des études économiques (INSEE) qui calcule les coefficients de pondération de type régression en offrant diverses options de fonctions objectives. Un programme appelé CLAN97 mis au point par Statistiques Suède est décrit dans Anderson et Nordberg (1998). Enfin, Folsom et Singh (2000) discutent d'une méthode élaborée au Research Triangle Institute.

3. LE MODÈLE LINÉAIRE CLASSIQUE

Le modèle linéaire classique est le fondement de l'estimation par régression à partir de données d'enquête, mais l'application du modèle à ces données nécessite certaines adaptations. En guise d'introduction à l'estimation par régression d'après les données obtenues sur échantillon, nous passons en revue le modèle linéaire classique. Supposons que

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + e_i,\ i = 1, 2, \ldots, n,$$

$$e_i \sim \mathrm{NI}(0, \sigma_e^2), \tag{3.1}$$

où le terme e_i est indépendant des vecteurs lignes $\mathbf{x}_j$ à k dimensions pour toutes les valeurs de i et j, et $\boldsymbol{\beta}$ est le vecteur colonne des valeurs inconnues du paramètre. Nous utiliserons aussi la représentation matricielle pour les quantités obtenues sur échantillon. Donc, pour un échantillon de n éléments,

$$\mathbf{X}' = (\mathbf{x}_1', \mathbf{x}_2', \cdots, \mathbf{x}_n') \text{ et } \mathbf{y}' = (y_1, y_2, \cdots, y_n).$$

Étant donné un échantillon de taille n et en considérant les $\mathbf{x}_i$ comme étant fixes, le meilleur estimateur (erreur quadratique moyenne minimale) de $\boldsymbol{\beta}$ est

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i\in A} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \sum_{i\in A} \mathbf{x}_i' y_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{3.2}$$

où A est l'ensemble d'indices des éléments de l'échantillon et nous supposons, comme nous le ferons jusqu'à la fin, que la matrice à inverser est non singulière. Si les e_i n'obéissent pas à la loi de distribution normale, $\hat{\boldsymbol{\beta}}$ est l'estimateur de la classe d'estimateurs linéaires non biaisés dont la variance est la plus faible. Nous pouvons écrire l'estimateur d'une combinaison linéaire de coefficients, disons $\theta_\alpha = \sum_{j=1}^{k} \alpha_j \beta_j$, sous la forme

$$\hat{\theta}_\alpha = \sum_{i \in A} w_{\alpha i}\, y_i$$

où les coefficients de pondération, $w_{\alpha i}$, minimisent le lagrangien

$$\sum_{i \in A} w_{\alpha i}^2 + \sum_{j=1}^{k} \lambda_j \left(\sum_{i \in A} w_{\alpha i}\, x_{ij} - \alpha_j \right)$$

et les λ_j sont des multiplicateurs de Lagrange. La variance de $\hat{\theta}_\alpha$ est

$$V\{\hat{\theta}_\alpha\} = V\left\{ \sum_{i \in A} w_{\alpha i} e_i \right\} = \sum_{i \in A} w_{\alpha i}^2 \sigma_e^2$$

car les coefficients de pondération sont des fonctions de $\mathbf{x}_i$ et non de y_i.

La matrice des covariances de $\hat{\boldsymbol{\beta}}$ est

$$V\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} V\left\{ \sum_{i \in A} \mathbf{b}_i' \right\} \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \right)^{-1}$$

$$= V\left\{ \sum_{i \in A} \mathbf{c}_i \right\} \tag{3.3}$$

où $\mathbf{b}_i' = \mathbf{x}_i' e_i$ et $\mathbf{c}_i = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i' e_i$. Comme e_i est indépendant de $\mathbf{x}_j$ pour toutes les valeurs de i et de j,

$$V\left\{ \sum_{i \in A} \mathbf{b}_i' \right\} = \sum_{i \in A} V\{\mathbf{b}_i'\} = \sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \sigma_e^2$$

et nous obtenons l'expression bien connue

$$V\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \sigma_e^2 .$$

Nous obtenons l'estimateur sans biais habituel de la matrice des covariances de $\hat{\boldsymbol{\beta}}$ en remplaçant σ_e^2 par l'estimateur sans biais de σ_e^2 obtenu par le carré de la moyenne des résidus, $\hat{e}_i = y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}$. Un estimateur de la matrice des covariances qui estime directement $V\{\sum_{i \in A} \mathbf{b}_i'\}$ est donné par

$$\tilde{V}_b\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \sum_{i \in A} \hat{\mathbf{b}}_i' \hat{\mathbf{b}}_i \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \right)^{-1}$$

$$= \sum_{i \in A} \hat{\mathbf{c}}_i' \hat{\mathbf{c}}_i , \tag{3.4}$$

où $\hat{\mathbf{b}}_i' = \mathbf{x}_i' \hat{e}_i$ et $\hat{\mathbf{c}}_i = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i' \hat{e}_i$. De la même façon

$$\hat{V}_b\{\hat{\theta}_\alpha\} = \sum_{i \in A} w_{\alpha i}^2 \hat{e}_i^2 \tag{3.5}$$

est une combinaison linéaire des éléments de (3.4) et un estimateur convergent de $V\{\hat{\theta}_\alpha\}$. L'estimateur (3.4) est un estimateur convergent de $V\{\hat{\boldsymbol{\beta}}\}$ si la matrice des covariances des e_i est une matrice diagonale à éléments bornés. Donc, il s'agit d'un estimateur plus robuste. Cependant, l'estimateur (3.4) est entaché d'un biais par défaut, car la variance de $\hat{e}_i$ est habituellement inférieure à la variance de e_i. Nous disposons de deux méthodes pour réduire le biais. La première consiste à rajuster le nombre de degrés de liberté en multipliant $\tilde{V}_b\{\hat{\boldsymbol{\beta}}\}$ par $(n-k)^{-1} n$, où k est la dimension de $\mathbf{x}_i$. Un autre ajustement consiste à remplacer $\hat{e}_i$ par

$$\tilde{e}_i = (1 - \psi_{ii})^{-0.5}\, \hat{e}_i ,$$

où ψ_{ii} représente le $i^{ième}$ élément diagonal de $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Consulter Horn, Horn et Duncan (1975), Royall et Cumberland (1978) et Cook et Weisberg (1982, section 2.2).

Si nous observons la valeur $\mathbf{x}_i$ pour un élément, mais que nous n'observons pas y_i, alors le meilleur prédicteur de y_i pour cet élément est $\hat{y}_i = \mathbf{x}_i \hat{\boldsymbol{\beta}}$. Pareillement, si nous connaissons la somme des $\mathbf{x}_i$ pour un ensemble de $\mathbf{x}$, alors le meilleur prédicteur de la somme des y_i est la somme des $\mathbf{x}_i \hat{\boldsymbol{\beta}}$. Donc, étant donné un ensemble de N éléments qui satisfont le modèle (3.1), un sous-ensemble d'observations $(y_i, \mathbf{x}_i)$ sur un sous-ensemble représenté par A et les valeurs connues de $\mathbf{x}_i$ pour les $N-n$ éléments restants, l'expression

$$\hat{Y}_{N-n,\,\mathrm{reg}} = \sum_{i \in \bar{A}} \hat{y}_i = \sum_{i \in \bar{A}} \mathbf{x}_i \hat{\boldsymbol{\beta}} ,$$

où $\bar{A}$ représente l'ensemble d'éléments pour lesquels y n'est pas observé, est le meilleur prédicteur de la somme des valeurs non observées de y. Consulter Goldberger (1962), Brewer (1963), Royall (1970), Harville (1976) et Graybill (1976, section 12.2). Par conséquent,

$$\hat{T}_{y,\mathrm{reg}} = \sum_{i \in A} y_i + \hat{Y}_{N-n,\,\mathrm{reg}} \tag{3.6}$$

est le meilleur prédicteur du total des N observations.

Si le premier élément du vecteur des x est toujours égal à l'unité, nous pouvons procéder à la partition du vecteur des x de sorte que $\mathbf{x}_i = (1, \mathbf{x}_{1,i})$ et écrire l'estimateur par régression de la moyenne sous la forme

$$\bar{y}_{\mathrm{reg}} = N^{-1} \hat{T}_{y,\mathrm{reg}} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}} = \bar{y}_n + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n}) \hat{\boldsymbol{\beta}}_1 , \tag{3.7}$$

où le $\hat{\boldsymbol{\beta}}$ de l'équation (3.2) est partitioné de sorte que $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1')'$ et $(\bar{y}_n, \bar{\mathbf{x}}_n)$ est le vecteur des simples moyennes d'échantillon. Nous appelons $\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}$ l'estimateur par régression des moyennes.

Étant donné le modèle (3.1), l'espérance de la moyenne de y pour la population finie de N éléments générée par le modèle est $\bar{\mathbf{x}}_N \boldsymbol{\beta}$ et $\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}$ est un estimateur sans biais de la moyenne de la population finie. Selon nous, cette étape est celle où débute l'estimation par régression de la moyenne d'une population finie dans le cas de plans d'échantillonnage plus complexes.

4. ESTIMATION FONDÉE SUR LE PLAN D'ÉCHANTILLONNAGE

À la présente section, nous traitons la population finie comme une réalisation, sous forme d'échantillon, d'une population infinie. Dans le domaine de l'échantillonnage, l'utilisation de ce genre de modèle remonte assez loin. Une série de références allant jusqu'à 1970 inclut Cochran (1939, 1942, 1946), Deming et Stephan (1941), Madow et Madow (1944), Yates (1949), Godambe (1955), Hájek (1959), Rao, Hartley et Cochran (1962), Konijn (1962), Brewer (1963), Godambe et Joshi (1965), Hanurav (1966), Ericson (1969), Isaki (1970) et Royall (1970).

Pour discuter des propriétés des estimateurs par régression dans le cas de grands échantillons, nous considérons des séries de populations finies et d'échantillons probabilistes connexes. L'ensemble d'indices des éléments de la $N^{\text{ième}}$ population finie est $U_N = \{1, \cdots, N\}$, où $N = 1, 2, \cdots$. Un vecteur ligne de caractéristiques $\mathbf{z}_{iN} = (y_{iN}, \mathbf{x}_{iN})$ est associé au $i^{\text{ième}}$ élément de la $N^{\text{ième}}$ population. Supposons que

$$\mathbf{F}_N = \left[(y_{1N}, \mathbf{x}_{1N}), (y_{2N}, \mathbf{x}_{2N}), \cdots, (y_{NN}, \mathbf{x}_{NN})\right]$$

est l'ensemble de vecteurs pour la $N^{\text{ième}}$ population finie. L'indice N est souvent omis sur les vecteurs. La moyenne de population finie est

$$\bar{\mathbf{z}}_N = (\bar{y}_N, \bar{\mathbf{x}}_N) = N^{-1} \sum_{i=1}^{N} (y_i, \mathbf{x}_i). \tag{4.1}$$

Nous représentons par A_N l'ensemble d'indices qui apparaissent dans l'échantillon sélectionné à partir de la $N^{\text{ième}}$ population finie.

Si la population finie est un échantillon tiré d'une superpopulation infinie, les propriétés probabilistes de l'échantillon sont déterminées d'après les propriétés de la superpopulation et celles de la méthode probabiliste utilisée pour sélectionner l'échantillon. Nous pouvons considérer les propriétés inconditionnelles, les propriétés conditionnelles associées à la population finie particulière ou les propriétés conditionnelles associées à une certaine partie de l'échantillon réalisé.

Les propriétés conditionnelles associées à la population finie dépendent principalement du plan d'échantillonnage et sont souvent appelées propriétés liées au plan d'échantillonnage. Donc, un estimateur $\hat{\theta}$ est dit convergent par rapport au plan d'échantillonnage pour le paramètre de population finie θ_N si, pour tout $\varepsilon > 0$,

$$\lim_{N, n \to \infty} \text{prob}\left\{ | \hat{\theta} - \theta_N | > \varepsilon \mid \mathbf{F}_N \right\} = 0,$$

où la notation signifie que nous imposons les contraintes de la population finie réalisée $\mathbf{F}_N$ et, donc, que la probabilité est déterminée d'après le plan d'échantillonnage.

Supposons que la population finie est générée par sélections indépendantes à partir d'une superpopulation pour laquelle $E\{\mathbf{z}_i' \mathbf{z}_i\}$ est définie et positive, où $\mathbf{z}_i = (y_i, \mathbf{x}_i)$. Nous définissons un vecteur de superpopulation des coefficients de régression par les moindres carrés par

$$\boldsymbol{\beta} = \left[E\{\mathbf{x}_i' \mathbf{x}_i\}\right]^{-1} E\{\mathbf{x}_i' y_i\}. \tag{4.2}$$

Étant donné un échantillon de n observations sur $\mathbf{z}_i$, nous définissons la matrice $n \times (k+1)$ des observations que nous représentons par $\mathbf{Z} = (\mathbf{y}, \mathbf{X})$, où la $i^{\text{ième}}$ ligne de $\mathbf{Z}$ est $(y_i, \mathbf{x}_i)$. Si nous supposons le modèle

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{4.3}$$

$$E\{\mathbf{u}, \mathbf{u}\mathbf{u}'\} = (\mathbf{0}, \boldsymbol{\Phi}),$$

l'estimateur par les moindres carrés généralisés de $\boldsymbol{\beta}$ est

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}' \boldsymbol{\Phi}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}' \boldsymbol{\Phi}^{-1} \mathbf{y}. \tag{4.4}$$

Le modèle (4.3) sert de motif pour les estimateurs de la forme (4.4), mais nous considérerons les estimateurs où $\boldsymbol{\Phi}$ est une matrice générale symétrique définie positive de coefficients de pondération qui n'est pas nécessairement la matrice des covariances des erreurs.

Nous nous inspirons de Fuller (1975) pour donner les propriétés liées aux grands échantillons du vecteur de coefficients estimés de régression (4.4). Consulter aussi Hidiroglou (1974), Scott et Wu (1981), et Robinson et Särndal (1983).

Supposons que la superpopulation possède huit moments et que le plan d'échantillonnage est tel que l'erreur dans l'estimateur d'Horvitz-Thompson de la moyenne est $O_p(n^{-1/2})$, où l'estimateur d'Horvitz-Thompson de la moyenne est

$$\bar{\mathbf{z}}_{\text{HT}} = (\bar{y}_{\text{HT}}, \bar{\mathbf{x}}_{\text{HT}}) = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{z}_i \tag{4.5}$$

et π_i est la probabilité de sélection pour l'élément i. Alors, l'erreur dans le vecteur des coefficients de régression est

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N \mid \mathbf{F}_N = \mathbf{Q}_{xxN}^{-1} \bar{\mathbf{b}}_{\text{HT}}' + O_p(n^{-1}), \tag{4.6}$$

où

$$\boldsymbol{\beta}_N = \mathbf{Q}_{xxN}^{-1} \mathbf{Q}_{xyN}, \tag{4.7}$$

$$(\mathbf{Q}_{xxN}, \mathbf{Q}_{xyN}) = E\left\{(\hat{\mathbf{Q}}_{xx}, \hat{\mathbf{Q}}_{xy}) \mid \mathbf{F}_N\right\}, \tag{4.8}$$

$$(\hat{\mathbf{Q}}_{xx}, \hat{\mathbf{Q}}_{xy}) = n^{-1} (\mathbf{X}' \boldsymbol{\Phi}^{-1} \mathbf{X}, \mathbf{X}' \boldsymbol{\Phi}^{-1} \mathbf{y}),$$

$$\bar{\mathbf{b}}_{\text{HT}} = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{b}_i, \tag{4.9}$$

$\mathbf{b}_i' = n^{-1} N \pi_i \zeta_i e_i$, $e_i = y_i - \mathbf{x}_i \boldsymbol{\beta}_N$, et ζ_i est la colonne i de $\mathbf{X}' \boldsymbol{\Phi}^{-1}$. L'erreur dans l'estimateur de $\boldsymbol{\beta}_N$ est approximativement égale à l'erreur dans un estimateur d'Horvitz-

Thompson de la moyenne. Dans l'expression (4.6), le $\boldsymbol{\beta}_N$ est défini comme étant une fonction des valeurs attendues des quantités d'échantillons $(\hat{\mathbf{Q}}_{xx}, \hat{\mathbf{Q}}_{xy})$. Par conséquent, $\boldsymbol{\beta}_N$ n'est pas nécessairement le coefficient de régression par les moindres carrés ordinaires pour la population finie. Le vecteur $\mathbf{b}_i$ de (4.9) est la généralisation du vecteur $\mathbf{b}_i$ de (3.3). Si la distribution limite de l'estimateur d'Horvitz-Thompson normalisé correctement est la distribution normale et qu'il existe un estimateur de la variance de cet estimateur, convergent par rapport au plan d'échantillonnage, alors il est possible de construire des tests et des intervalles de confiance pour les coefficients. Supposons que le plan d'échantillonnage soit tel que

$$\mathbf{V}_{\bar{z}\bar{z}}^{-1/2}(\bar{\mathbf{z}}_{\mathrm{HT}} - \bar{\mathbf{z}}_N) \mid \mathbf{F}_N \xrightarrow{\mathrm{L}} N(\mathbf{0}, \mathbf{I}), \tag{4.10}$$

lorsque $N, n \to \infty$, où $\mathbf{V}_{\bar{z}\bar{z}}$ est la matrice des covariances de $\bar{\mathbf{z}}_{\mathrm{HT}} - \bar{\mathbf{z}}_N$. Si $\mathbf{V}_{\bar{z}\bar{z}}$ est $O(n^{-1})$ et que l'estimateur $\hat{\mathbf{V}}_{\bar{z}\bar{z}}$ est approprié pour $\mathbf{V}_{\bar{z}\bar{z}}$, alors

$$\left[\hat{V}\{\hat{\boldsymbol{\beta}}\}\right]^{-1/2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\right) \mid \mathbf{F}_N \xrightarrow{\mathrm{L}} N(\mathbf{0}, \mathbf{I}), \tag{4.11}$$

où

$$\hat{V}\{\hat{\boldsymbol{\beta}}\} = \hat{\mathbf{Q}}_{xx}^{-1}\hat{\mathbf{V}}_{\bar{b}\bar{b}}\hat{\mathbf{Q}}_{xx}^{-1}, = \hat{V}\{\bar{\mathbf{c}}'_{\mathrm{HT}}\}, \tag{4.12}$$

$\hat{\mathbf{V}}_{\bar{b}} = \hat{V}\{\bar{\mathbf{b}}'_{\mathrm{HT}}\}$ est l'estimation de la variance de $\bar{\mathbf{b}}_{\mathrm{HT}}$ due au plan d'échantillonnage calculée avec $\hat{\mathbf{b}}'_i = n^{-1}N\pi_i\zeta'_i\hat{e}_i$ $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$, et $\hat{V}\{\bar{\mathbf{c}}'_{\mathrm{HT}}\}$ est l'estimation de la variance de $\bar{\mathbf{c}}'_{\mathrm{HT}}$ due au plan d'échantillonnage calculée avec $\hat{\mathbf{c}}'_i = \hat{\mathbf{Q}}_{xx}^{-1}\hat{\mathbf{b}}'_i$. Les propriétés limites tiennent pour des échantillons stratifiés et pour des échantillons stratifiés à deux degrés si l'on impose de légères restrictions sur la séquence des populations.

Par analogie à (3.7), nous obtenons un estimateur par régression de la moyenne de la population finie par calcul de la valeur de la fonction estimée de régression à la moyenne de population de $\mathbf{x}$ pour obtenir

$$\bar{y}_{\mathrm{reg}} = \bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}, \tag{4.13}$$

où $\hat{\boldsymbol{\beta}}$ est de la forme (4.4) avec une matrice $\boldsymbol{\Phi}$ générale. Nous pouvons écrire l'estimateur sous la forme $\mathbf{w}'\mathbf{y}$, où le vecteur des coefficients de pondération peut être construit par minimisation du lagrangien

$$\mathbf{w}'\boldsymbol{\Phi}\mathbf{w} + (\mathbf{w}'\mathbf{X} - \bar{\mathbf{x}}_N)\boldsymbol{\lambda}$$

et $\boldsymbol{\lambda}$ est le vecteur des multiplicateurs de Lagrange.

Si les vecteurs colonnes $\boldsymbol{\gamma}$ sont tels que

$$\mathbf{X}\boldsymbol{\gamma} = \boldsymbol{\Phi}\mathbf{D}_\pi^{-1}\mathbf{J} \tag{4.14}$$

pour tous les échantillons possibles, où $\mathbf{D}_\pi = \mathrm{diag}(\pi_1, \pi_2, \cdots, \pi_n)$ et $\mathbf{J}$ est un vecteur colonne de 1 à n dimensions, l'estimateur par régression $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ de (4.13), $\hat{\boldsymbol{\beta}}$ étant défini dans (4.4), est un estimateur convergent par rapport au plan d'échantillonnage de $\bar{y}_N$. Il découle de (4.11) que

$$\left[\bar{\mathbf{x}}_N\hat{V}\{\hat{\boldsymbol{\beta}}\}\bar{\mathbf{x}}'_N\right]^{-1/2}\left(\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}} - \bar{y}_N\right) \xrightarrow{\mathrm{L}} N(0, 1). \tag{4.15}$$

L'exigence (4.14) selon laquelle $\boldsymbol{\Phi}\mathbf{D}_\pi^{-1}\mathbf{J}$ doit se trouver dans l'espace colonne de $\mathbf{X}$, est une condition essentielle à la convergence par rapport au plan d'échantillonnage. Un moyen simple de satisfaire cette exigence consiste à permettre qu'une colonne de $\mathbf{X}$ soit une colonne de 1 et à utiliser un multiple de $\mathbf{D}_\pi$ pour $\boldsymbol{\Phi}$, ou à permettre qu'une colonne de $\mathbf{X}$ comprennent les éléments π_i^{-1} et fixer $\boldsymbol{\Phi} = \mathbf{I}$, ou à permettre qu'une colonne de $\mathbf{X}$ contienne les éléments π_i et fixer $\boldsymbol{\Phi} = \mathbf{D}_\pi^2$. Si $\mathbf{X}$ est composée du vecteur colonne unique contenant les éléments π_i et que $\boldsymbol{\Phi} = \mathbf{D}_\pi^2$, alors l'estimateur (4.13) se réduit à l'estimateur d'Horvitz-Thompson (4.5) pour des plans d'échantillonnage à taille fixe. Si $\mathbf{X} = \mathbf{J}$ et $\boldsymbol{\Phi} = \mathbf{D}_\pi$, l'estimateur (4.13) se réduit à l'estimateur par ratio,

$$\bar{y}_\pi = \left(\sum_{i\in A}\pi_i^{-1}\right)^{-1}\sum_{i\in A}\pi_i^{-1}y_i, \tag{4.16}$$

qui est invariant en fonction de la localisation.

Pour déterminer la nature de l'estimateur lorsque la condition (4.14) est satisfaite, supposons, sans perte de généralité, que $\mathbf{X} = (\mathbf{x}_0, \mathbf{X}_1)$, où $\mathbf{x}_0 = \boldsymbol{\Phi}\mathbf{D}_\pi^{-1}\mathbf{J}$ et $\mathbf{x}_i = (x_{0,i}, \mathbf{x}_{1,i})$. Alors

$$\bar{y}_{\mathrm{reg}} = \bar{x}_{0,N}\bar{x}_{0,\pi}^{-1}\bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{x}_{0,N}\bar{x}_{0,\pi}^{-1}\bar{\mathbf{x}}_{1,\pi})\hat{\boldsymbol{\beta}}_1, \tag{4.17}$$

où

$$\hat{\boldsymbol{\beta}}_1 = \left[(\mathbf{X}_1 - \mathbf{x}_0\hat{\boldsymbol{\mu}}_{x1})'\boldsymbol{\Phi}^{-1}(\mathbf{X}_1 - \mathbf{x}_0\hat{\boldsymbol{\mu}}_{x1})\right]^{-1} \times (\mathbf{X}_1 - \mathbf{x}_0\hat{\boldsymbol{\mu}}_{x1})'\boldsymbol{\Phi}^{-1}\mathbf{y},$$

$\hat{\boldsymbol{\mu}}_{x1} = \bar{x}_{0,\pi}^{-1}\bar{\mathbf{x}}_{1,\pi}$, et $(\bar{y}_\pi, \bar{\mathbf{x}}_\pi)$ est défini dans (4.16). Les ratios, tels que $\bar{x}_{0,\pi}^{-1}\bar{y}_\pi$, peuvent également s'écrire sous forme de ratios d'estimateurs d'Horvitz-Thompson. Si $\mathbf{J}$ se trouve dans l'espace colonne de $\mathbf{X}$, l'estimateur (4.17) est invariant en fonction de la localisation. Si $\boldsymbol{\Phi} = \mathbf{D}_\pi$, alors $\bar{x}_{0,\pi}^{-1}\bar{x}_{0,N} = 1$, et

$$\bar{y}_{\mathrm{reg}} = \bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}} = \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})\hat{\boldsymbol{\beta}}_1, \tag{4.18}$$

où

$$\hat{\boldsymbol{\beta}}_1 = \left[\sum_{i\in A}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})'\pi_i^{-1}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})\right]^{-1} \times \sum_{i\in A}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})'\pi_i^{-1}(y_i - \bar{y}_\pi). \tag{4.19}$$

En outre, quand $\boldsymbol{\Phi} = \mathbf{D}_\pi$, le $\boldsymbol{\beta}_N$ de (4.7) est le coefficient de régression pour la population

$$\boldsymbol{\beta}_N = \left[\sum_{i\in U}\mathbf{x}'_i\mathbf{x}_i\right]^{-1}\sum_{i\in U}\mathbf{x}'_i y_i. \tag{4.20}$$

Comme l'estimateur par régression de la moyenne est une combinaison linéaire de coefficients de régression, il s'agit du coefficient de régression d'une combinaison linéaire des variables x originales. Pour montrer ceci, supposons $\mathbf{x}_i = (x_{0,i}, \mathbf{x}_{1,i}) = (1, \mathbf{x}_{1,i})$, et définissons un nouveau vecteur dont la première valeur est 1 et un deuxième vecteur dont la moyenne de la population est nulle obtenue en soustrayant la moyenne de population originale $\bar{\mathbf{x}}_{1,N}$ du vecteur original $\mathbf{x}_{1,i}$. Représentons par $\mathbf{q}_i = (1, \mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})$ le vecteur transformé. Alors, le modèle de régression transformé est

$$y_i = \mathbf{q}_i \boldsymbol{\gamma} + e_i, \tag{4.21}$$

où le vecteur de coefficients pour la population finie est

$$\boldsymbol{\gamma}_N = (\bar{y}_N, \boldsymbol{\beta}'_{1,N})' = \left(\sum_{i \in U} \mathbf{q}_i' \mathbf{q}_i \right)^{-1} \sum_{i \in U} \mathbf{q}_i' y_i. \tag{4.22}$$

L'expression de l'estimateur par régression de la moyenne devient

$$\bar{y}_{\text{reg}} = \bar{\mathbf{q}}_N \hat{\boldsymbol{\gamma}} = \hat{\gamma}_0, \tag{4.23}$$

où $\hat{\boldsymbol{\gamma}}$ est obtenu d'après (4.4) en remplaçant $\mathbf{x}_i$ par $\mathbf{q}_i$. Puisqu'il s'agit d'un estimateur linéaire de la forme $\mathbf{w}'\mathbf{y}$, nous pouvons écrire

$$\bar{y}_{\text{reg}} = \sum_{i \in A} w_i y_i = \sum_{i \in A} \pi_i^{-1} g_i y_i, \tag{4.24}$$

où $w_i = \pi_i^{-1} g_i$. De surcroît, la variance estimée d'après (4.12) est

$$\hat{V}\{\bar{y}_{\text{reg}}\} = \hat{V}\{\hat{\gamma}_0\} = \hat{V}\left\{ \sum_{i \in A} \pi_i^{-1} (g_i \hat{e}_i) \right\}, \tag{4.25}$$

où il est sous-entendu que l'estimation de la variance liée au plan d'échantillonnage de (4.25) est calculée pour la variable $g_i \hat{e}_i$, $\hat{e}_i = y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}$, et que $\hat{\boldsymbol{\beta}}$ est défini dans (4.4) L'estimateur de la variance (4.25) est une généralisation directe de l'expression (3.5). Si nous transformons la variable de sorte que la moyenne de population du vecteur auxiliaire soit nulle, le premier élément du vecteur de coefficients de régression est la régression de la moyenne et le premier élément de (4.12) est un estimateur de la variance de l'estimateur par régression qui contient une composante due à l'estimation de $\boldsymbol{\beta}$. Ce fait a été mentionné dans Hidiroglou, Fuller et Hickman (1978). Consulter aussi Särndal (1982). Särndal, Swensson et Wretman (1989) ont proposé la terminologie du facteur g pour le calcul de la variance estimative d'un total estimé par régression.

Partant de (4.17), nous pouvons écrire

$$\begin{aligned} \bar{y}_{\text{reg}} &= \bar{x}_{0,N} \bar{x}_{0,\pi}^{-1} \left[\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi} \boldsymbol{\beta}_{1N} - (\bar{y}_N - \bar{\mathbf{x}}_{1,N} \boldsymbol{\beta}_{1N}) \right] \\ &\quad + O_p(n^{-1}), \\ &= \bar{e}_\pi + O_p(n^{-1}), \end{aligned}$$

où $e_i = y_i - \mathbf{x}_i \boldsymbol{\beta}$. Donc, nous pouvons estimer la variance de l'estimateur par régression au moyen de

$$\hat{V}\{\bar{e}_\pi\} = \hat{V}\left\{ \left(\sum_{i \in A} \pi_i^{-1} \right)^{-1} \sum_{i \in A} \pi_i^{-1} \hat{e}_i \right\} \tag{4.26}$$

où $\hat{e}_i = y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}$. Nous recommandons d'utiliser l'estimateur (4.25), puisqu'il est aussi facile à calculer que l'estimateur (4.26) et qu'il est applicable lorsque $\bar{\mathbf{x}}_{1,\pi} - \bar{\mathbf{x}}_{1,N}$ n'est pas $O_p(n^{-1/2})$.

Nous pouvons aussi calculer la variance de l'estimateur par régression par la méthode du jackknife ou par d'autres méthodes de répétition qui sont utilisées de plus en plus fréquemment. Consulter Frankel (1971), Kish et Frankel (1974), Woodruff et Causey (1976), Royall et Cumberland (1978), et Duchesne (2000). Yung et Rao (1996) ont montré que l'estimateur (4.25) est identique à un estimateur par linéarisation jackknife dans le cas des plans d'échantillonnage stratifiés à plusieurs degrés.

La méthode d'estimation par régression associée à (4.18) et (4.19) s'inscrit entièrement dans le cadre de la formulation d'un plan d'échantillonnage. Outre l'existence des moments, aucun modèle de population n'est utilisé, bien qu'il puisse être soutenu que le recours à la régression n'est sans doute envisagé que s'il existe une certaine corrélation linéaire entre $\mathbf{x}_{1,i}$ et y_i.

L'estimateur (4.19) est un estimateur fort naturel, car l'estimateur du coefficient de régression est un estimateur convergent par rapport au plan d'échantillonnage du coefficient de régression de la population. Il est un peu frustrant que (4.18) ne produise pas systématiquement la valeur la plus faible de la variance liée au plan d'échantillonnage pour les grands échantillons pour l'estimation de la moyenne. Si nous traitons $\hat{\boldsymbol{\beta}}_1$ dans (4.18) comme un vecteur fixe, la valeur qui minimise la variance de la combinaison linéaire des moyennes est

$$\boldsymbol{\beta}_{1,\text{dopt}} = \left[V\{\bar{\mathbf{x}}_{1,\pi} \mid \mathbf{F}_N\} \right]^{-1} C\{\bar{\mathbf{x}}_{1,\pi}, \bar{y}_\pi \mid \mathbf{F}_N\}. \tag{4.27}$$

À cet égard, consulter Cochran (1977, page 201), Fuller et Isaki (1981), Montanari (1987, 1999) et Rao (1994). S'il existe un estimateur de la variance de $\bar{\mathbf{x}}_{1,\pi}$ convergent par rapport au plan d'échantillonnage, alors le $\boldsymbol{\beta}_{1,d}$ qui minimise l'estimateur de la variance

$$\hat{V}\{\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi} \boldsymbol{\beta}_{1,d}\}, \tag{4.28}$$

représenté par $\hat{\boldsymbol{\beta}}_{1,\text{dopt}}$, est un estimateur convergent de $\boldsymbol{\beta}_{1,\text{dopt}}$. Il s'ensuit que l'estimateur

$$\bar{y}_{d,\text{reg}} = \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \hat{\boldsymbol{\beta}}_{1,\text{dopt}} \tag{4.29}$$

est caractérisé par la variance limite minimale pour les estimateurs convergents par rapport au plan d'échantillonnage de la forme $\bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \boldsymbol{\beta}_{1,d}$. En outre

$$\left[\hat{V}\{\bar{e}_\pi\} \right]^{-1/2} (\bar{y}_{d,\text{reg}} - \bar{y}_N) \xrightarrow{L} N(0,1), \tag{4.30}$$

où $\hat{V}\{\bar{e}_\pi\}$ est l'estimateur de (4.26) construit en prenant $\hat{e}_i = y_i - \bar{y}_\pi - (\mathbf{x}_{1,i} - \mathbf{x}_{1,\pi})\hat{\boldsymbol{\beta}}_{1,\text{dopt}}$.

Dans le contexte des grands échantillons, (4.29) montre comment construire un estimateur par régression dont les propriétés par rapport au plan d'échantillonnage sont optimales. En pratique, plusieurs questions restent néanmoins à résoudre. L'estimateur est obtenu sous l'hypothèse d'un grand échantillon et d'un vecteur $\mathbf{x}$ de dimension fixe. En pratique, il pourrait exister plusieurs variables auxiliaires possibles et, si un grand nombre d'entre elles sont incluses dans la régression, les termes exclus de l'approximation du grand échantillon deviennent importants. Il en est surtout ainsi pour les échantillons en grappes où le nombre d'unités primaires d'échantillonnage dans l'échantillon est faible. Le cas échéant, le nombre de degrés de liberté dans $\hat{V}\{\bar{\mathbf{x}}_{1,\pi}\}$ est faible et l'inverse peut être instable. Nous examinons ces questions plus en détail à la section 9.

L'estimateur de $\boldsymbol{\beta}_{1,\text{dopt}}$ dans (4.29) est linéaire en y pour la plupart des plans d'échantillonnage. Consulter Rao (1994). Par exemple, pour un plan stratifié avec échantillonnage aléatoire simple à l'intérieur des strates,

$$\hat{C}\{\bar{\mathbf{x}}_{1,\pi}, \bar{y}_\pi\} = \sum_{h=1}^{H} K_h \sum_{j=1}^{n_h} (\mathbf{x}_{1,hj} - \bar{\mathbf{x}}_{1,h})'(y_{hj} - \bar{y}_h), \tag{4.31}$$

où

$$K_h = W_h^2 (1 - f_h)(n_h - 1)^{-1} n_h^{-1} = N^{-2} \pi_h^{-2} (1 - f_h)(n_h - 1)^{-1} n_h,$$

$N^{-1}N_h = W_h$, N_h est la taille de la strate h, $f_h = \pi_h = N_h^{-1} n_h$, et n_h est la taille de l'échantillon dans la strate h. Il s'ensuit que les coefficients de pondération associés à l'estimateur (4.29) sont

$$w_{hi} = N^{-1}\pi_h^{-1} + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \times \left[\sum_{t=1}^{H} K_t \sum_{j=1}^{n_h} (\mathbf{x}_{1,tj} - \bar{\mathbf{x}}_{1,t})'(\mathbf{x}_{1,tj} - \bar{\mathbf{x}}_{1,t})\right]^{-1} \times K_h (\mathbf{x}_{1,hi} - \bar{\mathbf{x}}_{1,h})'. \tag{4.32}$$

Voir aussi Särndal (1996). Nous pouvons construire les coefficients de pondération de (4.32), à condition que soient satisfaites les contraintes $\sum_{hi \in A} w_{hi}^2 K_h^{-1}$

$$\sum_{i \in A_h} w_{hi} = N^{-1} N_h, \; h = 1, 2, \cdots, H,$$

et

$$\sum_{hi \in A} w_{hi} \mathbf{x}_{1,hi} = \bar{\mathbf{x}}_{1,N},$$

où A_h représente l'ensemble d'éléments dans la strate h.

L'estimateur de (4.19) quand $\boldsymbol{\Phi} = \mathbf{D}_\pi$ est une fonction des estimateurs d'Horvitz-Thompson des moments de la population. L'estimateur (4.17), tel que $\boldsymbol{\Phi}^{-1} = \text{diag}\{K_t\}$, où K_t est sur la diagonale pour les éléments de la strate t, et contenant des variables indicatrices pour les effets de strates, donne l'estimateur de la moyenne dans la classe

$$\bar{y}_{\text{reg}} = \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})\hat{\boldsymbol{\beta}}_1$$

caractérisé par l'estimation la plus faible de la variance liée au plan d'échantillonnage. Si les pentes réelles dans les strates sont identiques et que les probabilités de sélection sont proportionnelles à la racine carrée des variances à l'intérieur des strates, alors l'utilisation de $\boldsymbol{\Phi} = \mathbf{D}_\pi^2$ produit pour les petits échantillons une EQM plus faible que l'utilisation de $\boldsymbol{\Phi}^{-1} = \text{diag}\{K_t\}$, car la somme des $w_{hi}^2 \sigma_h^2$ est plus faible. Fuller et Isaki (1981) ont remarqué que l'estimateur optimal par rapport au plan d'échantillonnage est souvent bien approximé par l'estimateur construit en prenant $\boldsymbol{\Phi} = \mathbf{D}_\pi^2$.

Nous avons introduit une estimation par régression de la moyenne, mais ce sont souvent les totaux que l'on estime et que l'on utilise comme contrôles. Considérons l'estimateur par régression du total de y défini par

$$\hat{T}_{y,\text{reg}} = \hat{T}_{y,\pi} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi})\hat{\boldsymbol{\beta}}_{y\cdot x}, \tag{4.33}$$

où $\mathbf{T}_{x,N}$ est le total connu de $\mathbf{x}$ et $(\hat{T}_{y,\pi}, \hat{\mathbf{T}}_{x,\pi})$ est le vecteur des estimateurs de $(T_{y,N}, \mathbf{T}_{x,N})$ convergents par rapport au plan d'échantillonnage. Par analogie à (4.28), l'estimateur du coefficient $\boldsymbol{\beta}$ optimal est

$$\hat{\boldsymbol{\beta}}_{y\cdot x} = \left[\hat{V}\{\hat{\mathbf{T}}_{x,\pi}\}\right]^{-1} \hat{C}\{\hat{\mathbf{T}}'_{x,\pi}, \hat{T}_{y,\pi}\}, \tag{4.34}$$

où $\hat{V}\{\hat{\mathbf{T}}_{x,\pi}\}$ est un estimateur convergent par rapport au plan d'échantillonnage de la variance de $\hat{\mathbf{T}}_{x,\pi}$ et $\hat{C}\{\hat{T}'_{x,\pi}, \hat{T}_{y,\pi}\}$ est un estimateur convergent par rapport au plan d'échantillonnage de la covariance de $\hat{\mathbf{T}}_{x,\pi}$ et $\hat{T}_{y,\pi}$.

L'estimateur du total est $N\bar{y}_{\text{reg}}$ dans le cas de l'échantillonnage aléatoire simple, mais l'équivalence exacte pourrait ne pas être vérifiée dans le cas d'échantillons plus complexes, car l'estimateur de la moyenne pourrait être un estimateur par ratio. Cependant, si nous construisons l'estimateur par régression des deux totaux au moyen de (4.34), la variance pour grands échantillons du ratio des deux totaux estimés est égale à l'estimateur par régression de la moyenne. Pour montrer ceci, écrivons l'erreur dans les totaux de y et u estimés par régression sous la forme

$$\hat{T}_{y,\text{reg}} - T_{y,N} = \hat{T}_{y,\pi} - T_{y,N} + \left(\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}\right)\boldsymbol{\beta}_{y\cdot x,N} + O_p(Nn^{-1})$$

et

$$\hat{T}_{u,\text{reg}} - T_{u,N} = \hat{T}_{u,\pi} - T_{u,N} + \left(\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}\right)\boldsymbol{\beta}_{u\cdot x,N} + O_p(Nn^{-1}), \tag{4.35}$$

où nous supposons que $\hat{T}_{y,\pi} - T_{y,N}$, $\hat{\boldsymbol{\beta}}_{y\cdot x} - \boldsymbol{\beta}_{y\cdot x,N}$ et les quantités correspondantes pour u, sont $O_p(Nn^{-1/2})$ et $O_p(n^{-1/2})$, respectivement. Alors, l'erreur dans $\hat{T}_{u,\text{reg}}^{-1}\hat{T}_{y,\text{reg}}$ est

$$\begin{aligned}\hat{T}_{u,\text{reg}}^{-1}\hat{T}_{y,\text{reg}} - T_{u,N}^{-1}T_{y,N} = T_{u,N}^{-1}\Big[&\left(\hat{T}_{y,\pi} - T_{y,N}\right) \\ &- R_N\left(\hat{T}_{u,\pi} - T_{u,N}\right) \\ &+ \left(\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}\right)\left(\boldsymbol{\beta}_{y\cdot x,N} - R_N\boldsymbol{\beta}_{z\cdot x,N}\right)\Big] \\ &+ O_p(Nn^{-1}),\end{aligned} \tag{4.36}$$

où $R_N = T_{u,N}^{-1}\,T_{y,N}$. Si nous construisons l'estimateur par régression pour R_N en partant de $\hat{R} = \hat{T}_{u,\pi}^{-1}\hat{T}_{y,\pi}$, nous avons

$$\hat{R}_{\text{reg}} = \hat{R} + \left(\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}\right)\hat{\boldsymbol{\beta}}_{R\cdot x}, \tag{4.37}$$

où

$$\hat{\boldsymbol{\beta}}_{R\cdot x} = \left[\hat{V}\left\{\hat{\mathbf{T}}_{x,\pi}\right\}\right]^{-1}\hat{C}\left\{\hat{\mathbf{T}}_{x,\pi}', \hat{R}\right\}$$

et

$$\hat{C}\left\{\hat{\mathbf{T}}_{x,\pi}, \hat{R}\right\} = \hat{C}\left\{\hat{\mathbf{T}}_{x,\pi}, T_{u,N}^{-1}\left(\hat{T}_{y,\pi} - R_N\hat{T}_{u,\pi}\right)\right\}.$$

Il s'ensuit que le coefficient pour grand échantillon optimal par rapport au plan d'échantillonnage pour le ratio est $T_{u,N}^{-1}(\boldsymbol{\beta}_{y\cdot x,N} - R_N\boldsymbol{\beta}_{u\cdot x,N})$ et que le ratio des estimateurs par régression optimaux pour le plan d'échantillonnage est l'estimateur par régression du ratio optimal pour le plan d'échantillonnage dans le cas des grands échantillons.

5. MODÈLES ET ESTIMATION PAR RÉGRESSION

À la présente section, nous émettons l'hypothèse que l'analyste postule un modèle détaillé de superpopulation. Nous supposons aussi que l'échantillon est tiré avec probabilités inégales ou (et) que la structure précisée de la covariance des erreurs n'est pas un multiple de la matrice d'identité. Alors, l'estimateur optimal par rapport au plan d'échantillonnage ne concordera que dans des cas spéciaux avec le meilleur estimateur construit sous le modèle, en imposant comme contraintes les valeurs d'échantillon de **x**. Pour étudier ce conflit éventuel, écrivons le modèle de la population en notation matricielle comme suit

$$\begin{aligned}\mathbf{y}_U &= \mathbf{X}_U\boldsymbol{\beta} + \mathbf{e}_U \\ \mathbf{e}_U &\sim \left(\mathbf{0}, \boldsymbol{\Sigma}_{eeUU}\right),\end{aligned} \tag{5.1}$$

où $\mathbf{y}_U = (y_1, y_2, \cdots, y_N)'$, $\mathbf{e}_U = (e_1, e_2, \cdots, e_N)'$ et $\mathbf{X}_U = (\mathbf{x}_1', \mathbf{x}_2', \cdots, \mathbf{x}_N)'$. Nous supposons que $\boldsymbol{\Sigma}_{eeUU}$ est connue ou connue jusqu'à un multiple. Le modèle d'un échantillon contenant n observations est

$$\begin{aligned}\mathbf{y}_A &= \mathbf{X}_A\boldsymbol{\beta} + \mathbf{e}_A, \\ \mathbf{e}_A &\sim \left(\mathbf{0}, \boldsymbol{\Sigma}_{eeAA}\right),\end{aligned}$$

où $\mathbf{y}_A = (y_1, y_2, \cdots, y_n)'$, $\mathbf{e}_A = (e_1, e_2, \cdots, e_n)'$, $\mathbf{X}_A = (\mathbf{x}_1', \mathbf{x}_2', \cdots, \mathbf{x}_n')'$ et où les éléments de l'échantillon sont représentés par les indices 1, 2, ..., n, par souci de commodité. Nous adoptons l'indice U pour identifier les quantités de la population et l'indice A, les quantité de l'échantillon, mais nous omettrons souvent le second en vue de simplifier la notation. Par exemple, nous représenterons parfois la matrice $n\times n$ des covariances par $\boldsymbol{\Sigma}_{ee}$. La moyenne inconnue de la population finie est

$$\bar{y}_N = \bar{\mathbf{x}}_N\boldsymbol{\beta} + \bar{\mathbf{e}}_N. \tag{5.2}$$

Sous le modèle (5.1), le meilleur prédicteur linéaire, conditionnellement sans biais, de $\theta_N = \bar{y}_N$, compte tenu de **X**, est

$$\begin{aligned}\hat{\theta} = N^{-1}\Big[&\sum_{i\in A} y_i + (N-n)\,\bar{\mathbf{x}}_{N-n}\hat{\boldsymbol{\beta}} \\ &+ \mathbf{J}_{N-n}'\,\boldsymbol{\Gamma}_{\bar{A}A}\left(\mathbf{y}_A - \mathbf{X}_A\hat{\boldsymbol{\beta}}\right)\Big],\end{aligned} \tag{5.3}$$

où $\boldsymbol{\Gamma}_{\bar{A}A} = \boldsymbol{\Sigma}_{ee\bar{A}A}\boldsymbol{\Sigma}_{eeAA}^{-1}$, $\bar{\mathbf{x}}_{N-n} = (N-n)^{-1}(N\bar{\mathbf{x}}_N - n\bar{\mathbf{x}}_n)$, $\Sigma_{ee\bar{A}A} = E\{\mathbf{e}_{\bar{A}}\mathbf{e}_A'\}$,

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\,\boldsymbol{\Sigma}_{eeAA}^{-1}\,\mathbf{X}\right)^{-1}\mathbf{X}'\,\boldsymbol{\Sigma}_{eeAA}^{-1}\mathbf{y},$$

$\mathbf{e}_{\bar{A}} = (e_{n+1}, e_{n+2}, \cdots, e_N)$, $\mathbf{J}_{N-n}$ est un vecteur colonne de 1 à N- n dimensions, $\bar{\mathbf{x}}_n$ est la simple moyenne d'échantillon et $\bar{A}$ est l'ensemble des éléments contenus dans U qui ne sont pas contenus dans A. Consulter Royall (1976). Dans les conditions du modèle,

$$\hat{\theta} - \bar{y}_N = \mathbf{C}_{x\bar{A}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + N^{-1}\mathbf{J}_{N-n}'(\boldsymbol{\Gamma}_{\bar{A}A}\mathbf{e}_A - \mathbf{e}_{\bar{A}})$$

et

$$\begin{aligned}V\left\{\hat{\theta} - \bar{y}_N|\mathbf{X}_A\right\} = &\mathbf{C}_{x\bar{A}}V\left\{\hat{\boldsymbol{\beta}}\right\}\mathbf{C}_{x\bar{A}}' \\ &+ N^{-2}\mathbf{J}_{N-n}'\left(\boldsymbol{\Sigma}_{ee\bar{A}\bar{A}} - \boldsymbol{\Gamma}_{\bar{A}A}\boldsymbol{\Sigma}_{ee\bar{A}A}\right)\mathbf{J}_{N-n},\end{aligned} \tag{5.4}$$

où

$$\mathbf{C}_{x\bar{A}} = N^{-1}\left[(N-n)\,\bar{\mathbf{x}}_{N-n} - \mathbf{J}_{N-n}'\,\boldsymbol{\Gamma}_{\bar{A}A}\mathbf{X}_A\right].$$

La convergence de l'estimateur (5.3) par rapport au plan d'échantillonnage et les situations dans lesquelles l'estimateur basé sur un modèle se réduit à l'estimateur d'Horvitz-Thompson ont été considérées, entre autres, par Isaki (1970), Royall (1970, 1976), Scott et Smith (1974), Cassel, Särndal, et Wretman (1976, 1979, 1983), Zyskind (1976), Tallis (1978), Isaki et Fuller (1982), Wright (1983), Pfefferman (1984), Tam (1986), Brewer, Hanif et Tam (1988), Montanari (1999), et Gerow et McCulloch (2000).

L'estimateur (5.3) se réduit à $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ s'il existe un η tel que

$$\mathbf{X}_A\eta = \boldsymbol{\Sigma}_{eeAA}\mathbf{J}_n + \boldsymbol{\Sigma}_{eeA\bar{A}}\mathbf{J}_{N-n}, \tag{5.5}$$

pour tous les échantillons pour lesquels la probabilité est positive. S'il existe aussi une valeur de γ telle que

$$\mathbf{X}_A\gamma = \boldsymbol{\Sigma}_{eeAA}\mathbf{D}_\pi^{-1}\mathbf{J}_n \tag{5.6}$$

pour tous les échantillons pour lesquels la probabilité est positive, alors l'estimateur $\hat{\theta}$ de (5.3) est convergent par rapport au plan d'échantillonnage, où $\mathbf{D}_\pi$ est défini pour (4.14). Étant donné un $\mathbf{k}$ tel que

$$\mathbf{X}_A\mathbf{k} = \boldsymbol{\Sigma}_{eeAA}\left(\mathbf{D}_\pi^{-1}\mathbf{J}_n - \mathbf{J}_n\right) - \boldsymbol{\Sigma}_{eeA\bar{A}}\mathbf{J}_{N-n}, \tag{5.7}$$

alors l'estimateur $\hat{\theta}$ de (5.3) peut s'exprimer sous la forme

$$\hat{\theta} = \bar{y}_\pi + \left(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi\right)\hat{\boldsymbol{\beta}} \tag{5.8}$$

et, si le plan d'échantillonnage est tel que $\bar{y}_\pi$ est convergent par rapport au plan d'échantillonnage pour $\bar{y}_N$, l'estimateur $\hat{\theta}$ de (5.8) est convergent par rapport au plan d'échantillonnage pour $\bar{y}_N$.

Nous appelons un modèle de régression de la forme (5.1) pour lequel (5.5) et (5.6), ou (5.7), comprend un modèle complet. Si la condition (5.6) ou (5.7) n'est pas vérifiée, nous considérons que le modèle est réduit ou restreint. Nous ne pouvons pas nous attendre à ce que les conditions d'un modèle complet soient vérifiées pour chaque variable analysée dans le cas d'une enquête générale, car $\boldsymbol{\Sigma}_{ee}$ sera différente pour diverses variables y. Par conséquent, étant donné un modèle réduit, on pourrait rechercher un bon estimateur du modèle dans la classe des estimateurs convergents par rapport au plan d'échantillonnage.

Pour construire un estimateur convergent par rapport au plan d'échantillonnage de la forme $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ lorsque le modèle (5.1) est un modèle réduit, nous pouvons ajouter un vecteur satisfaisant (5.7) à la matrice $\mathbf{X}$ pour créer un modèle complet. Dans ce cas, deux situations peuvent se présenter. Dans la première, on connaît la moyenne (ou le total) de la population de la variable ajoutée. Connaissant la moyenne, nous pouvons construire l'estimateur par régression habituel et appliquer les formules habituelles d'estimation de la variance liée au plan d'échantillonnage.

Pour décrire une méthode d'estimation applicable au cas où la moyenne de la population de la variable est inconnue, représentons par $\mathbf{q} = (q_1, q_2, ..., q_n)'$ le vecteur ajouté, où $\mathbf{q}$ est le vecteur figurant dans le deuxième membre de l'équation (5.7). Supposons que $\mathbf{H} = (\mathbf{X}, \mathbf{q})$, où $\mathbf{X}$ est la matrice des variables auxiliaires dont on connaît le vecteur des moyennes de la population, $\bar{\mathbf{x}}_N$. Nous écrivons le modèle complet pour l'échantillon sous la forme

$$\mathbf{y} = \mathbf{H}\boldsymbol{\beta}_{y\cdot h} + \mathbf{e}, \tag{5.9}$$

où $\mathbf{e} \sim (\mathbf{0}, \boldsymbol{\Sigma}_{ee})$. Le meilleur estimateur linéaire conditionnellement sans biais de $\boldsymbol{\beta}_{y\cdot h}$ est

$$\hat{\boldsymbol{\beta}}_{y\cdot h} = \left(\mathbf{H}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{y}. \tag{5.10}$$

Dans (5.9), si le coefficient de $\mathbf{q}$ n'est pas nul, il est impossible de construire un estimateur conditionnellement sans biais de $\bar{\mathbf{h}}_N\boldsymbol{\beta}_{y\cdot h}$, car la composante $\bar{q}_N$ de $\bar{\mathbf{h}}_N$ est inconnue. Cependant, comme l'estimateur $\hat{\boldsymbol{\beta}}_{y\cdot h}$ de $\boldsymbol{\beta}_{y\cdot h}$, est sans biais, il est possible de construire un estimateur conditionnellement sans biais de toute fonction linéaire de $\boldsymbol{\beta}_{y\cdot h}$. Donc, il est naturel de remplacer le terme $\bar{q}_N$ par son « meilleur estimateur disponible » et raisonnable de choisir l'estimateur par régression

$$\bar{q}_{\text{reg}} = \bar{q}_\pi + \left(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi\right)\hat{\boldsymbol{\beta}}_{q\cdot x}, \tag{5.11}$$

où $\hat{\boldsymbol{\beta}}_{q\cdot x} = (\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{q}$. Alors, l'estimateur (5.3) devient

$$\hat{\theta} = \bar{y}_\pi + \left[\left(\bar{\mathbf{x}}_N, \bar{q}_{\text{reg}}\right) - \left(\bar{\mathbf{x}}_\pi, \bar{q}_\pi\right)\right]\hat{\boldsymbol{\beta}}_{y\cdot h} \tag{5.12}$$

Nous pouvons exprimer l'estimateur (5.12) sous la forme bien connue de l'estimateur par régression,

$$\bar{y}_{\text{reg}} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}_{y\cdot x}. \tag{5.13}$$

Autrement dit, l'estimateur par régression de la moyenne de la population finie pour y fondé sur le modèle complet, lorsque l'on ne connaît pas la moyenne de q_i et qu'on l'estime au moyen de l'estimateur par régression, est l'estimateur par régression où $\boldsymbol{\beta}_{y\cdot x}$ est estimé par régression de y sur $\mathbf{x}$ par les moindres carrés généralisés en utilisant la matrice des covariances $\boldsymbol{\Sigma}_{ee}$. Consulter Park (2002). L'estimateur est conditionnellement sans biais par rapport au modèle dans le cas du modèle réduit contenant uniquement $\mathbf{x}$ si le modèle réduit est valable. Si le coefficient de population pour q_i n'est pas nul, le modèle réduit n'est pas correct. Alors, l'estimateur est conditionnellement biaisé par rapport au modèle, mais il est sans biais pour la moyenne de la population finie dans les conditions du modèle complet et d'un plan d'échantillonnage sans biais, car

$$\begin{aligned} E\left\{\bar{y}_{\text{reg}} - \bar{y}_N\right\} &= E\left\{E\left[\bar{y}_{\text{reg}} - \bar{y}_N \mid \mathbf{H}\right]\right\} \\ &= E\left\{(0, \bar{q}_{\text{reg}} - \bar{q}_N)\boldsymbol{\beta}_{y\cdot h}\right\} \doteq 0, \end{aligned} \tag{5.14}$$

où $\bar{y}_{\text{reg}}$ est défini dans (5.12) et l'approximation est due à l'espérance approximative concernant le plan d'échantillonnage dans l'estimateur par régression $\bar{q}_{\text{reg}}$.

L'estimateur (5.13) est un estimateur linéaire, où le vecteur des coefficients de pondération, $\mathbf{w}$, minimise le lagrangien

$$\mathbf{w}'\boldsymbol{\Sigma}_{ee}\mathbf{w} + \left[\mathbf{w}'\mathbf{H} - \left(\bar{\mathbf{x}}_N, \bar{q}_{\text{reg}}\right)\right]\boldsymbol{\lambda}. \tag{5.15}$$

L'estimateur est invariant en localisation si la colonne de 1 est comprise dans l'espace colonne de $\mathbf{X}$.

Puisque q est la variable dont l'omission dans le modèle complet peut produire un biais, il semble prudent de tester le coefficient de q avant d'utiliser le modèle réduit pour construire un estimateur de la moyenne de y. Nous pouvons,

pour cela, utiliser un estimateur de la variance basé sur un modèle,

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}_{y \cdot h} \mid \mathbf{H}\right\} = \left(\mathbf{H}' \hat{\boldsymbol{\Sigma}}_{ee}^{-1} \mathbf{H}\right)^{-1}$$

ou l'estimateur basé sur le plan d'échantillonnage de la variance de (4.12). Consulter Du Mouchel et Duncan (1983) et Fuller (1984).

Pour les échantillons à deux degrés, une spécification de travail pour $\boldsymbol{\Sigma}_{ee}$ pourrait être particulièrement indiquée; à cet égard, consulter Royall (1976, 1986) et Montanari (1987). Un modèle raisonnable est un modèle dans lequel il existe une corrélation commune entre les éléments d'une même unité primaire d'échantillonnage, mais une corrélation nulle entre les éléments contenus dans des unités primaires d'échantillonnage différentes. Comme la matrice des covariances associée $\boldsymbol{\Sigma}_{ee}$ est une matrice diagonale par blocs d'une forme particulière, il est assez facile de l'inverser et, donc, l'estimateur fondé sur une telle spécification $\boldsymbol{\Phi}$ de travail est assez facile à construire. L'estimateur par régression obtenu lorsque l'on utilise une spécification $\boldsymbol{\Phi}$ telle que la corrélation n'est pas nulle pour les éléments compris dans une même unité primaire d'échantillonnage est une combinaison de l'estimateur fondé sur les totaux des unités primaires d'échantillonnage et celui fondé sur les éléments. Voir Fuller et Battese (1973). Donc, l'utilisation d'une telle spécification $\boldsymbol{\Phi}$ permet d'éviter les problèmes de variance associés à l'utilisation des totaux correspondant aux unités primaires d'échantillonnage.

6. MAXIMUM DE VRAISEMBLANCE ET MÉTHODE ITÉRATIVE DU QUOTIENT (RAKING RATIO)

Le fondement théorique des estimateurs par régression décrits aux sections 3 et 4 est l'estimation du maximum de vraisemblance dans le cas du modèle linéaire dont les erreurs obéissent à la loi normale. Nous considérons maintenant le maximum de vraisemblance dans le cas de variables multinomiales. Étant donné un échantillon aléatoire simple provenant d'une multinomiale définie par les entrées dans un tableau à double entrée, le logarithme du rapport des vraisemblances est, à une constante près,

$$\sum_{i=1}^{r} \sum_{j=1}^{c} a_{ij} \log p_{ij}, \tag{6.1}$$

où a_{ij} est la fraction estimée dans la cellule ij, p_{ij} est la fraction de la population dans la cellule ij, r est le nombre de lignes et c est le nombre de colonnes. Si l'on maximise (6.1) en imposant la contrainte $\sum\sum p_{ij} = 1$, on obtient les estimateurs du maximum de vraisemblance $\hat{p}_{ij} = a_{ij}$. Si l'on connaît les fractions de marge des lignes $p_{i\cdot,N}$ et les fractions de marge des colonnes $p_{\cdot j,N}$, il est naturel de maximiser la vraisemblance subordonnée à ces contraintes en utilisant le lagrangien

$$\sum_{i=1}^{r} \sum_{j=1}^{c} a_{ij} \log p_{ij} + \sum_{i=1}^{r} \lambda_i \left(\sum_{j=1}^{c} p_{ij} - p_{i\cdot,N} \right) + \sum_{j=r+1}^{r+c} \lambda_j \left(\sum_{i=1}^{r} p_{ij} - p_{\cdot j,N} \right), \tag{6.2}$$

où λ_i, $i = 1, 2, ..., r$, tient compte des contraintes de lignes et γ_j, $j = 1, 2, ..., c$, des contraintes de colonnes. Il n'existe aucune expression explicite de la solution de (6.2) et il se pourrait qu'il n'y en ait aucune si le nombre de cellules vides est trop élevé. Une méthode qui produit des estimations s'approchant de la solution du maximum de vraisemblance est celle appelée *méthode itérative du quotient (raking ratio)* ou *ajustement proportionnel itératif*. Il s'agit d'un processus d'itération où l'on commence par faire des ajustements de ratios tenant compte des contraintes de lignes, puis des ajustements de ratios tenant compte des contraintes de colonnes, puis des ajustements de ratios tenant compte des contraintes de lignes, et ainsi de suite. La paternité de la méthode est généralement attribuée à Deming et Stephan (1940). Consulter, par exemple, Bishop, Fienberg et Holland (1975, ch. 3).

Deville et Särndal (1992) ont considéré une classe de fonctions objectives de la forme $\sum_{i \in A} G(w_i, \alpha_i)$, où $G(w, \alpha)$ est une mesure de la distance entre le coefficient de pondération initial α_i et un coefficient de pondération final w_i. La fonction objective est minimisée en imposant la contrainte

$$\sum_{i \in A} w_i \mathbf{x}_i = \bar{\mathbf{x}}_N. \tag{6.3}$$

Deville et Särndal (1992) ont utilisé le terme *calé* ou *étalonné* pour décrire les coefficients de pondération qui satisfont (6.3). Si le coefficient de pondération initial est $\alpha_i = (\sum \pi_j^{-1})^{-1} \pi_i^{-1}$ et que 1 est le premier élément de $\mathbf{x}_i$, une approximation de la solution du problème de minimisation est donnée par un estimateur par régression de la moyenne de la forme

$$\bar{y}_{\text{reg}} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}, \tag{6.4}$$

où

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i \in A} \mathbf{x}_i' \varphi_{ii}^{-1} \mathbf{x}_i\right]^{-1} \sum_{i \in A} \mathbf{x}_i' \varphi_{ii}^{-1} y_i,$$

et φ_{ii} est la dérivée seconde de $G(w, \alpha)$ par rapport à w, calculée à $(w, \alpha) = (\alpha_i, \alpha_i)$. En suivant cette méthode, Deville et Särndal (1992) ont montré que les estimateurs par le maximum de vraisemblance et par la méthode itérative du quotient ont les mêmes distributions limites que l'estimateur par régression (4.18) quand $\boldsymbol{\Phi} = \mathbf{D}_\pi$. Pour obtenir les coefficients de pondération par la méthode itérative du quotient, ils ont utilisé la fonction objective

$$\sum_{i \in A} \left[w_i \log \alpha_i^{-1} w_i + \alpha_i - w_i \right], \tag{6.5}$$

et pour obtenir les coefficients de pondération par le maximum de vraisemblance, ils ont utilisé la fonction objective

$$\sum_{i \in A} \left[w_i - \alpha_i - \alpha_i \log \alpha_i^{-1} w_i \right]. \tag{6.6}$$

Deville, Särndal et Sautory (1993) ont étudié quatre estimateurs dans la classe. Bien que les coefficients de pondération construits en se servant de diverses fonctions puissent différer considérablement, ces auteurs ont conclu que les estimations sont fort similaires, résultat qui corrobore la théorie. Singh et Mohl (1996) et Théberge (1999, 2000) discutent des estimateurs ayant la propriété de calage.

7. POPULATION DE VECTEURS AUXILIAIRES CONNUS À L'ÉTAPE DE L'ESTIMATION

Si l'on connaît le vecteur $\mathbf{x}$ pour tous les éléments de la population, le nombre d'estimateurs par régression possibles augmente considérablement. La plupart des méthodes comprennent l'ajustement d'une fonction approximative représentant la relation entre y et les variables auxiliaires. La méthode la plus courante consiste à affecter les éléments de la population à des catégories d'après les données auxiliaires et à utiliser ces catégories comme strates a posteriori. Cette méthode équivaut à approximer l'espérance de y, étant donné $\mathbf{x}$, au moyen d'une fonction en escalier. L'estimateur est formellement équivalent à l'estimateur par régression (4.19) où le vecteur $\mathbf{x}$ est un vecteur de variables indicatrices de l'appartenance à la strate a posteriori.

L'application de la méthode nécessite souvent l'élaboration des critères à utiliser pour former les strates a posteriori. Habituellement, celles-ci sont formées de sorte que chacune contienne un nombre minimal d'éléments de l'échantillon et que les coefficients de pondération pour toute strate a posteriori ne soit pas indûment grand. L'estimation par stratification a posteriori et la formation de strates a posteriori ont été étudiées, entre autres, par Fuller (1966), Holt et Smith (1979), Tremblay (1986) Kalton et Maligalig (1991), Little (1993), Eltinge et Yansaneh (1997) et Lazzeroni et Little (1998). Holt et Smith (1979) ont présenté des arguments en faveur de l'utilisation d'un estimateur conditionnel de la variance pour la stratification a posteriori.

Étant donné les vecteurs $\mathbf{x}$ de la population, nous pouvons utiliser l'échantillon pour estimer une relation fonctionnelle entre y et $\mathbf{x}$, puis prédire la valeur non observée de y. Si l'on veut que la méthode soit convergente par rapport au plan d'échantillonnage, il faut qu'une condition comparable à (4.14) soit satisfaite. Un moyen d'assurer la convergence par rapport au plan d'échantillonnage est de contraindre le modèle ajusté à satisfaire.

$$\sum_{i \in A} \pi_i^{-1} \left[y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) \right] = 0, \tag{7.1}$$

où $f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ est la valeur estimée du modèle pour la $i^{ème}$ observation.

Firth et Bennett (1998) ont fait remarquer que certains modèles non linéaires satisfont (7.1). Si le modèle initial ne satisfait pas (7.1), nous pouvons ajouter un terme estimatif de coordonnée à l'origine afin de créer un modèle complet étendu,

$$\tilde{f}_F\left(\mathbf{x}_i; \hat{\boldsymbol{\beta}}\right) = f\left(\mathbf{x}_i; \hat{\boldsymbol{\beta}}\right) + \left(\sum_{i \in A} \pi_i^{-1} \right)^{-1} \sum_{i \in A} \pi_i^{-1} \left[y_i - f\left(\mathbf{x}_i; \hat{\boldsymbol{\beta}}\right) \right].$$

Il s'agit d'une extension directe de la notion d'estimation par différence au cas non linéaire. Consulter Isaki (1970), Cassel, Särndal et Wretman (1976) et Wright (1983). Wu et Sitter (2001) ont proposé une méthode étroitement liée qui consiste à utiliser la fonction ajustée $f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ comme variable auxiliaire dans un estimateur par régression linéaire.

Outre les fonctions en escalier, plusieurs méthodes « locales » peuvent être utilisées pour approximer la relation fonctionelle entre $\mathbf{x}$ et y. Les fonctions splines et les polynômes sont des modèles linéaires qui rentrent dans la classe de la section 4. Kuo (1988), Dorfman (1993), Dorfman et Hall (1993), Chambers (1996) et Chambers, Dorfman et Wehrly (1993) ont considéré, pour les populations finies, du point de vue du modèle, des estimateurs comprenant une forme ou l'autre de lissage local pour estimer les quantités de population. Breidt et Opsomer (2000) ont montré que les estimateurs fondés sur une régression polynomiale locale sont convergents par rapport au plan d'échantillonnage. Firth et Bennett (1998) ont, eux aussi, considéré des modèles à ajustement local.

8. ESTIMATION PAR RÉGRESSION ET NON-RÉPONSE

L'estimation par régression fait souvent partie des méthodes utilisées pour corriger les données pour la non-réponse partielle. On peut se fonder, pour justifier la régression, sur un modèle tel que (3.1) ou sur le fait que la régression permet de faire la correction pour tenir compte des probabilités inégales de réponse. Voir Cassel, Särndal et Wretman (1979, 1983), Little (1982, 1986), Bethlehem (1988), Kott (1994), Fuller, Loughin et Baker (1994) et Fuller et An (1998).

Considérons un estimateur du vecteur de coefficients de régression de la population de la forme (4.4) quand $\boldsymbol{\Phi} = \mathbf{D}_\pi$ construit en se fondant sur les unités répondantes. Représentons l'estimateur par $\tilde{\boldsymbol{\beta}}$ et supposons que p_i est la probabilité conditionnelle d'observer l'unité i, étant donné que l'unité est sélectionnée dans l'échantillon. Alors, dans les conditions de régularité, l'estimateur $\tilde{\boldsymbol{\beta}}$ est un estimateur convergent de

$$\gamma_N = \left(\sum_{i \in U} \mathbf{x}_i' p_i \mathbf{x}_i \right)^{-1} \sum_{i \in U} \mathbf{x}_i' p_i y_i. \qquad (8.1)$$

Nous pouvons exprimer la moyenne de population de y sous la forme

$$\bar{y}_N = \bar{\mathbf{x}}_N \gamma_N + \bar{a}_N \qquad (8.2)$$

où $a_i = y_i - \mathbf{x}_i \gamma_N$ et $\bar{a}_N$ est la moyenne de population des a_i. L'estimateur par régression $\bar{y}_{reg} = \bar{\mathbf{x}}_N \tilde{\boldsymbol{\beta}}$ sera convergent pour $\bar{y}_N$ si la limite de la probabilité de $\bar{a}_N$ est nulle. Cette limite sera nulle si la série de populations finies est une série d'échantillons aléatoires provenant d'une population infinie dans laquelle

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + e_i, \qquad (8.3)$$

et les e_i de l'échantillon sont indépendantes de $\mathbf{x}_i$ en ayant $E\{e_i \mid \mathbf{x}_i\} = 0$.

Ou bien, une condition suffisante pour que $\bar{a}_N$ soit nulle est qu'il existe un vecteur colonne ξ tel que

$$\mathbf{x}_i \xi = p_i^{-1} \qquad (8.4)$$

pour $i = 1, 2, \cdots, N$. Donc, l'inverse de la probabilité de réponse est une fonction linéaire des variables de contrôle, l'estimateur par régression est un estimateur convergent de la moyenne de y. Un moyen de satisfaire (8.4) consiste à ce que les éléments de $\mathbf{x}_i$ soient des variables nominales qui définissent des sous-groupes et que les probabilités de réponse soient constantes dans chaque sous-groupe.

Si (8.4) est satisfaite et que la probabilité de réponse est indépendante d'une unité à l'autre, alors l'estimateur de la variance fondée sur (4.12) est un estimateur approprié de la variance de l'estimateur par régression de la moyenne. Il est particulièrement important d'utiliser un estimateur de la variance de la forme (4.12) ou (4.25), et non de la forme (4.26), car $\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi$ n'est, en général, pas $O_p(n^{-1/2})$ en cas de non-réponse. Singh et Folsom (2000) présentent un argument comparable pour l'estimateur de la variance (4.25) lorsque l'on utilise la régression pour faire la correction pour l'erreur de couverture.

Souvent, on procède à un rajustement préliminaire des probabilités de sélection pour tenir compte de la non-réponse, puis à l'estimation par régression. La méthode de rajustement de la réponse utilisée le plus fréquemment consiste à créer des cellules (strates a posteriori) et à ajuster par la méthode du ratio les coefficients de pondération calculés pour les répondants compris dans la cellule de sorte que la somme de ces coefficients soit égale au total estimé (ou connu) pour la cellule. Consulter, par exemple, Little et Rubin (1987, page 250). Les méthodes fondées sur une fonction estimée de probabilité de réponse sont examinées par Cassel, Särndal et Wretman (1983), Rosenbaum et Rubin (1983), Folsom et Witt (1994), Fuller et An (1998), et Folsom et Singh (2000). Brick, Waksberg et Keeter (1996) utilisent une probabilité estimative de contact pour faire la correction pour tenir compte de la couverture de la base de sondage.

Pour examiner les méthodes fondées sur les probabilités estimatives de réponse, supposons que l'inverse de la probabilité de réponse de l'individu i est donné par

$$p_i^{-1} = g(\mathbf{z}_i; \boldsymbol{\theta}^0), \qquad (8.5)$$

où $\mathbf{z}_i$ est un vecteur de variables qui peuvent être observées pour les répondants ainsi que par les non-répondants, $\boldsymbol{\theta}^0$ est la valeur réelle de $\boldsymbol{\theta}$ et $g(\mathbf{z}_i; \boldsymbol{\theta})$ est continue en $\boldsymbol{\theta}$ et ses dérivées première et seconde sont continues en un ensemble ouvert contenant $\boldsymbol{\theta}^0$ pour tous les $\mathbf{z}_i$. Le vecteur $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ est observé et nous supposons que la borne inférieure de p_i est un nombre positif.

Soit δ_i une variable indicatrice telle que $\delta_i = 1$ si l'on obtient une réponse et $\delta_i = 0$ si l'on n'en obtient pas. En utilisant le vecteur $(\delta_i, \mathbf{z}_i)$, nous estimons le paramètre $\boldsymbol{\theta}^0$ de la fonction de probabilité de réponse. Supposons que $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 = O_p(n^{-1/2})$, où $\hat{\boldsymbol{\theta}}$ est l'estimateur de $\boldsymbol{\theta}$. Représentons par $\boldsymbol{\beta}_N$ le vecteur des coefficients de régression pour la population finie pour la régression de y sur $\mathbf{x}$. Soit

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \pi_i^{-1} \hat{p}_i^{-1} \delta_i \right)^{-1} \sum_{i \in A} \mathbf{x}_i' y_i \pi_i^{-1} \hat{p}_i^{-1} \delta_i, \qquad (8.6)$$

où π_i sont les probabilités de sélection et $\hat{p}_i^{-1} = g(\mathbf{z}_i; \hat{\boldsymbol{\theta}})$. Dans des conditions comparables à celles utilisées à la section 4,

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = \mathbf{M}_{xx}^{-1} \sum_{i \in A} \delta_i \pi_i^{-1} p_i^{-1} \mathbf{x}_i' a_i \left[1 + p_i \mathbf{g}_{1,i} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 \right) \right] + O_p(n^{-1}),$$

où $\mathbf{g}_{1,i}$ est le vecteur ligne des dérivées premières de $g(\mathbf{z}_i; \boldsymbol{\theta})$ évaluées à $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ et $\mathbf{M}_{xx} = \sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \pi_i^{-1} p_i^{-1} \delta_i$. Si $\mathbf{g}_{1,i}$ n'est pas corrélé à a_i, alors le terme contenant $\mathbf{g}_{1,i} a_i$ est $O_p(n^{-1})$ et l'estimateur de la variance construit comme si $g(\mathbf{z}; \boldsymbol{\theta}^0)$ était connu est approprié. Les conditions sont satisfaites si $\mathbf{z}_i$ est un sous-vecteur de $\mathbf{x}_i$ et qu'il définit des cellules d'imputation (cellules d'ajustement) telles que les taux de réponse sont égaux à l'intérieur d'une cellule.

9. CONSIDÉRATIONS PRATIQUES

Si l'on veut utiliser les coefficients de pondération obtenus par régression dans le cas d'une enquête générale, aucun coefficient de pondération individuel utilisé pour estimer un total ne devrait être inférieur à l'unité. En outre, il semble raisonnable, par souci de robustesse, d'éviter les coefficients de pondération ayant une valeur très grande. Nous examinons certaines méthodes mises au point afin de réaliser ces objectifs.

Plusieurs algorithmes produisent des coefficients de pondération positifs avec une probabilité élevée. Les méthodes itératives du quotient (raking ratio) produisent des coefficients de pondération positifs pour la plupart des configurations de données. Deville, Särndal et Sautory

(1993) discutent de l'extension de la méthode itérative du quotient à des variables x générales et à des extensions visant à inclure les bornes sur les coefficients de pondération.

Tillé (1998) a proposé d'utiliser des probabilités conditionnelles approximatives, contraintes sur $\bar{\mathbf{x}}_\pi$, pour calculer un estimateur. On peut étendre son approximation de façon à produire des coefficients de pondération par régression positifs avec une forte probabilité. Représentons par $\bar{\mathbf{x}}_\pi^{(i)}$ un estimateur obtenu par suppression de l'élément i, ou de l'unité primaire d'échantillonnage i, et par modification des coefficients de pondération restants de sorte que $\bar{\mathbf{x}}_\pi^{(i)}$ soit sans biais, ou converge vers le même ordre que $\bar{\mathbf{x}}_\pi$ pour la moyenne de la population de tous les éléments, sauf i. L'estimateur $\bar{\mathbf{x}}_\pi^{(i)}$ peut être celui utilisé pour produire les écarts jackknife. Représentons par $\hat{\Sigma}_{\bar{x}\bar{x}}$ un estimateur de la matrice des covariances de $\bar{\mathbf{x}}_\pi$ et par $\hat{\Sigma}_{\bar{x}\bar{x}(i)}$, un estimateur de la matrice des covariances de $\bar{\mathbf{x}}_\pi^{(i)}$ conditionnel sur $i \in A$. Alors, dans les grands échantillons, $\bar{\mathbf{x}}_\pi$ et $\bar{\mathbf{x}}_\pi^{(i)}$ suivent approximativement la loi de distribution normale et un estimateur de la probabilité que i soit incluse dans l'échantillon, étant donné l'estimateur de la moyenne $\bar{\mathbf{x}}_\pi$, est

$$\begin{aligned}\hat{\pi}_{i|A} &= \hat{P}\left\{i \in A \mid \mathbf{F}_N, \bar{\mathbf{x}}_\pi\right\} \\ &= \pi_i \left|\hat{\Sigma}_{\bar{x}\bar{x}}\right|^{1/2} \left|\hat{\Sigma}_{\bar{x}\bar{x}(i)}\right|^{-1/2} \exp\left\{0.5\left(\mathbf{G}_{\bar{x}\bar{x}} - \mathbf{G}_{\bar{x}\bar{x}(i)}\right)\right\} \quad (9.1)\end{aligned}$$

où

$$\mathbf{G}_{\bar{x}\bar{x}} = \left(\bar{\mathbf{x}}_\pi - \bar{\mathbf{x}}_N\right)\hat{\Sigma}_{\bar{x}\bar{x}}^{-1}\left(\bar{\mathbf{x}}_\pi - \bar{\mathbf{x}}_N\right)',$$

$$\mathbf{G}_{\bar{x}\bar{x}(i)} = \left(\bar{\mathbf{x}}_\pi^{(i)} - \bar{\mathbf{x}}_N^{(i)}\right)\hat{\Sigma}_{\bar{x}\bar{x}(i)}^{-1}\left(\bar{\mathbf{x}}_\pi^{(i)} - \bar{\mathbf{x}}_N^{(i)}\right)',$$

et $\bar{\mathbf{x}}_N^{(i)} = (N-1)^{-1}(N\bar{\mathbf{x}}_N - \mathbf{x}_i)$. Dans le cas de l'échantillonnage aléatoire simple, Tillé (1998) a montré que l'estimateur

$$\bar{y}_{p\pi} = N^{-1}\sum_{i \in A} \pi_{i|A}^{-1} y_i, \quad (9.2)$$

où $\pi_{i|A}$ est la probabilité conditionnelle calculée sous l'hypothèse de normalité, est approximativement égal à l'estimateur par régression. Comme l'estimateur n'est pas calé, nous proposons une version calée obtenue par calcul de l'estimateur par régression en prenant les $\hat{\pi}_{i|A}$ comme coefficients de pondération initiaux. La différence entre l'estimateur (9.2) et l'estimateur par régression construit en prenant les poids initiaux $\hat{\pi}_{i|A}$ est $O_p(n^{-1})$. Donc, il y a beaucoup de chances que les coefficients de pondération obtenus par régression conçus de la sorte soient positifs. L'estimateur de la variance $\hat{\Sigma}_{\bar{x}\bar{x}(i)}$ est assez simple à calculer pour les échantillons stratifiés, mais demande des calculs considérables dans d'autres cas. Donc, on pourrait choisir d'approximer $\Sigma_{\bar{x}\bar{x}(i)}$.

Étant donné que les coefficients de pondération obtenus par régression sont construits par minimisation de la fonction objective, on peut ajouter des contraintes au problème afin d'imposer des bornes aux coefficients de pondération. Huang et Fuller (1978) ont proposé une méthode itérative équivalant à la construction, à chaque étape, d'une matrice $\mathbf{\Phi}$ qui réduit le coefficient de pondération sur les observations pour lesquelles l'écart du coefficient de pondération courant par rapport à la moyenne est important en valeur absolue.

Afin de discuter de méthodes supplémentaires associées aux fonctions objectives quadratiques, supposons que nous ayons une matrice des covariances de travail, que nous représentons par $\mathbf{\Phi}_{ee}$, pour le modèle (5.1) que nous utiliserons pour construire un estimateur par régression. Posons que $\boldsymbol{\alpha}$ est le vecteur colonne des coefficients de pondération initiaux et que $\mathbf{\Phi}_{ee}\boldsymbol{\alpha}$ se trouve dans l'espace colonne de $\mathbf{X}$. Alors, les coefficients de pondération qui minimisent la variance conditionnelle due au modèle sont ceux qui minimisent $\mathbf{w}'\mathbf{\Phi}_{ee}\mathbf{w}$ ou, de façon équivalente, qui minimisent

$$(\mathbf{w} - \boldsymbol{\alpha})'\mathbf{\Phi}_{ee}(\mathbf{w} - \boldsymbol{\alpha}) \quad (9.3)$$

à condition que soit satisfaite la contrainte

$$\mathbf{w}'\mathbf{X} = \bar{\mathbf{x}}_N. \quad (9.4)$$

Étant donné une fonction objective, nous pouvons appliquer des contraintes supplémentaires aux w_i de sorte que

$$L_1 \le w_i \le L_2, \quad i \in A, \quad (9.5)$$

où L_1 et L_2 sont des constantes non négatives. Minimiser (9.3), en veillant à ce que les contraintes (9.4) et (9.5) soient satisfaites, est un problème de programmation quadratique. L'utilisation de cette dernière a été proposée par Husain (1969) et utilisée par Isaki, Tsay et Fuller (2000).

Si le nombre de variables de contrôle utilisées est élevé, il peut être impossible de construire des coefficients de pondération qui satisfont les contraintes de calage et qui sont également compris entre des bornes raisonnables. Le praticien doit donc faire des compromis. La pratique la plus courante consiste à supprimer certaines variables du modèle. À cet égard, consulter Bankier, Rathwell et Maijkowski (1992) et Silva et Skinner (1997). Pour examiner une autre méthode, considérons la situation où certaines contraintes sont nécessaires, mais où d'autres peuvent être relâchées. Partitionnons la matrice des observations sur les variables auxiliaires de façon telle que $(\mathbf{X}_0, \mathbf{X}_2)$, où $\mathbf{X}_0$ représente l'ensemble de variables pour lequel des contraintes rigoureuses sont nécessaires et $\mathbf{X}_2$ celui pour lequel les contraintes peuvent être relâchées. Supposons que $\mathbf{\Phi}_{ee}\boldsymbol{\alpha}$ se situe dans l'espace colonne de $\mathbf{X}_0$. Alors, une généralisation de (9.3) et (9.4) donne la fonction

$$(\mathbf{w} - \boldsymbol{\alpha})'\mathbf{\Phi}_{ee}(\mathbf{w} - \boldsymbol{\alpha}) + (\mathbf{w}'\mathbf{X}_2 - \bar{\mathbf{x}}_{2,N})\mathbf{\Psi}(\mathbf{w}'\mathbf{X}_2 - \bar{\mathbf{x}}_{2,N})' \quad (9.6)$$

et la contrainte

$$\mathbf{w}'\mathbf{X}_0 - \mathbf{x}_{0,N} = \mathbf{0}, \quad (9.7)$$

où $\mathbf{\Phi}_{ee}$ et $\mathbf{\Psi}$ sont des matrices définies symétriques positives et $\bar{\mathbf{x}}_N = (\bar{\mathbf{x}}_{0,N}, \bar{\mathbf{x}}_{2,N})$. La valeur de $\mathbf{w}$ qui minimise (9.6) à condition que (9.7) est celle qui minimise l'erreur quadratique moyenne du prédicteur linéaire non biaisé de $\bar{\mathbf{x}}_N \boldsymbol{\beta}$ sous le modèle mixte

$$\mathbf{y} = \mathbf{X}_0 \boldsymbol{\beta}_0 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \mathbf{e},$$

où $\boldsymbol{\beta}_2 \sim (\mathbf{0}, \mathbf{\Psi}), e \sim (\mathbf{0}, \mathbf{\Phi}_{ee})$, le vecteur aléatoire $\boldsymbol{\beta}_2$ est indépendant de $\mathbf{e}$ et $\boldsymbol{\beta}_0$ est un vecteur fixe. Voir Lazzeroni et Little (1998) pour l'utilisation de modèles aléatoires pour la stratification a posteriori.

Le vecteur $\mathbf{w}'$ qui minimise (9.6) à condition que la contrainte (9.7) soit respectée est

$$\mathbf{w}' = \boldsymbol{\alpha}' + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\, \mathbf{H}_{x\psi x}^{-1} \mathbf{X}' \mathbf{\Phi}_{ee}^{-1}, \tag{9.8}$$

où

$$\mathbf{H}_{x\psi x} = \begin{pmatrix} \mathbf{X}_0' \mathbf{\Phi}_{ee}^{-1} \mathbf{X}_0 & \mathbf{X}_0' \mathbf{\Phi}_{ee}^{-1} \mathbf{X}_2 \\ \mathbf{X}_2' \mathbf{\Phi}_{ee}^{-1} \mathbf{X}_0 & \mathbf{\Psi}^{-1} + \mathbf{X}_2' \mathbf{\Phi}_{ee}^{-1} \mathbf{X}_2 \end{pmatrix}. \tag{9.9}$$

L'estimateur peut s'écrire

$$\bar{y}_{rreg} = \mathbf{w}' \mathbf{y} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\, \hat{\boldsymbol{\theta}}, \tag{9.10}$$

où $\hat{\boldsymbol{\theta}} = \mathbf{H}_{x\psi y}^{-1} \mathbf{X}' \mathbf{\Phi}_{ee}^{-1} \mathbf{y}$. Consulter Henderson (1963), Robinson (1991) et Rao (2002, ch. 6).

Husain (1969) a considéré l'application de (9.6) à un échantillon aléatoire simple provenant d'une distribution normale en posant $\mathbf{X}_0 = \mathbf{J}, \mathbf{\Phi}_{ee} = \mathbf{I}$ et $\mathbf{\Psi}^{-1} = \gamma^{-1} \hat{\mathbf{\Sigma}}_{x,22}$, où $\hat{\mathbf{\Sigma}}_{x,22}$ est l'estimateur de la matrice des covariances de $\bar{\mathbf{x}}_{2,\pi}$, et γ est une constante à déterminer. Pour ce cas, Husain a montré que la forme optimale de γ est

$$\gamma_{\text{opt}} = \left[k_2(1 - R^2)\right]^{-1} (n - k_2 - 2)\, R^2, \tag{9.11}$$

où k_2 est la dimension de $\mathbf{x}_2$ et R^2 est le carré du coefficient de corrélation multiple. Bardsley et Chambers (1984) ont considéré la fonction (9.6), la division de $\mathbf{x}_i$ en deux composantes et étudié le comportement de l'estimateur dans la perspective du modèle. La méthode associée à (9.5), (9.6) et (9.7) a été utilisée par Isaki, Tsay et Fuller (2000). Dans cette application, le vecteur $\bar{\mathbf{x}}_{0,N}$ contenait les totaux de marge d'un tableau à entrées multiples et $\bar{\mathbf{x}}_{2,N}$ contenait les totaux à l'intérieur des cellules. Rao et Singh (1997) ont étudié un estimateur étroitement apparenté dans lequel les tolérances sont données pour la différence entre les estimations finales pour les éléments de $\bar{\mathbf{x}}_{2,N}$ et pour les éléments correspondants de $\bar{\mathbf{x}}_{2,N}$.

Park (2002) a étendu les résultats d'optimalité de Husain à une matrice $\mathbf{\Psi}$ plus générale. Le vecteur $\mathbf{x}_2$ peut être transformé de sorte que $\hat{V}\{\bar{\mathbf{x}}_{2,\pi}\}$ pour le vecteur transformé soit une matrice diagonale et que $\tilde{\mathbf{X}}_2' \mathbf{\Phi}_{ee}^{-1} \tilde{\mathbf{X}}_2$ soit une matrice diagonale, où $\tilde{\mathbf{X}}_2$ représente la partie de $\mathbf{X}_2$ qui est orthogonale à $\mathbf{X}_0$ dans la matrice $\mathbf{\Phi}_{ee}$. Autrement dit,

$$\tilde{\mathbf{X}}_2 = \mathbf{X}_2 - \mathbf{X}_0 \left(\mathbf{X}_0' \mathbf{\Phi}_{ee}^{-1} \mathbf{X}_0\right)^{-1} \mathbf{X}_0' \mathbf{\Phi}_{ee}^{-1} \mathbf{X}_2.$$

Alors, la matrice diagonale $\mathbf{\Psi}$ qui minimise la variance approximative contient les éléments

$$\psi_{ii} = (m_{ii} V_{\beta\beta ii})^{-1} \beta_i^2, \tag{9.12}$$

où m_{ii} est le $i^{\text{ième}}$ élément de la matrice diagonale $\tilde{\mathbf{X}}_2' \mathbf{\Phi}_{ee}^{-1} \tilde{\mathbf{X}}_2$ et $V_{\beta\beta ii}$ est la variance de $\hat{\beta}_i$ dans l'échelle transformée. Pour appliquer la méthode, on doit estimer les paramètres de la population ou choisir des valeurs réalistes pour une matrice $\mathbf{\Psi}$ à usage général. Si l'on postule un modèle aléatoire de superpopulation pour $\boldsymbol{\beta}$, alors les β_i^2 de (9.12) sont remplacés par $E\{\beta_i^2\}$, où l'espérance est celle du modèle.

10. COMMENTAIRES

L'estimation par régression est un outil souple et puissant qui permet d'intégrer des données auxiliaires dans le processus d'estimation. Des méthodes étroitement apparentées, telles que la méthode itérative du quotient (ou raking ratio), ont des propriétés de grands échantillons équivalentes à celles des estimateurs par régression. La linéarité de ces estimateurs est de la plus haute importance, car elle permet la construction d'un ensemble de données d'usage général qui fournit de très bons estimateurs pour toute une gamme de paramètres.

Si l'on se concentre sur une variable unique y, il est possible de réaliser des gains d'efficacité en postulant un ensemble particulier de variables auxiliaires et une matrice particulière des covariances des erreurs. Étant donné la simplicité des conditions de convergence par rapport au plan d'échantillonnage, il est facile de tester la convergence de ces modèles.

REMERCIEMENTS

Cette étude a été financée en partie par le Cooperative Agreement 43-3AEU-0-80064 conclu entre la Iowa State University, le U. S. National Agricultural Statistics Service et le U. S. Bureau of the Census. Je dois énormément à Mingue Park pour son aide lors de la revue bibliographique, ses commentaires sur les résultats théoriques et leurs corrections, et l'utilisation d'éléments de sa thèse. Je remercie Michael Hidiroglou, J.N.K. Rao, Harold Mantel et Jean Opsomer de leurs commentaires constructifs sur les ébauches du manuscrit. Enfin, je remercie le rédacteur en chef adjoint de ses nombreux commentaires en vue d'améliorer la présentation.

ANNEXE

Cette annexe contient les théorèmes qui appuient les propriétés limites des estimateurs par régression décrits à la section 4.

Théorème A.1. Supposons que $\{U_N, \mathbf{F}_N, A_N, n_N : N = k+3, k+4, \ldots\}$ est une suite de populations finies et d'échantillons, où $\mathbf{F}_N$ est un échantillon provenant d'une population infinie ayant huit moments, A_N est l'échantillon de taille n_N sélectionné à partir de la $N^{\text{ième}}$ population. Supposons que $\hat{\boldsymbol{\beta}}$ est défini par (4.4) et que

$$\hat{\mathbf{Q}}_{zz} = n^{-1}\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z},$$

où $\boldsymbol{\Phi}$ est une matrice symétrique définie positive qui peut être une fonction de $\mathbf{X}$, mais non de y, $\mathbf{Z}$ est défini conformément à (4.2), et nous omettons l'indice N sur les quantités d'échantillon. Supposons que $\hat{\mathbf{Q}}_{zz}$ est définie positive et que sa probabilité est égale à un. Si $\boldsymbol{\Phi}$ est aléatoire, supposons que les lignes de $\boldsymbol{\Phi}^{-1}\mathbf{Z}$ ont borné le quatrième moment. Supposons que les probabilités de sélection satisfont

$$0 < \mathrm{K}_1 < Nn^{-1}\pi_i < K_2,$$

où π_i sont les probabilités de sélection. Supposons en outre que le plan d'échantillonnage est tel que, pour tout $\mathbf{z}$ dont le quatrième moment est borné

$$\left[(\bar{\mathbf{z}}_{\mathrm{HT}} - \bar{\mathbf{z}}_N)', (\hat{\mathbf{Q}}_{zz} - \mathbf{Q}_{zzN})\right] \mid \mathbf{F}_N = O_p(n^{-1/2}), \qquad (A.1)$$

où

$$\bar{\mathbf{z}}_{\mathrm{HT}} = (\bar{y}_{\mathrm{HT}}, \bar{\mathbf{x}}_{\mathrm{HT}}) = N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{z}_i, \qquad (A.2)$$

$\mathbf{Q}_{zzN} = E\{\hat{\mathbf{Q}}_{zz} \mid \mathbf{F}_N\}$, $\bar{\mathbf{z}}_N$ est la moyenne de la population finie de $\mathbf{z}$, $\mathbf{Q}_{zzN}$ est une matrice définie positive pour la $N^{\text{ième}}$ population et la limite de $\mathbf{Q}_{zzN}$ est définie et positive. Alors,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N \mid \mathbf{F}_N = \mathbf{Q}_{xxN}^{-1}\bar{\mathbf{b}}'_{\mathrm{HT}} + O_p(n^{-1}), \qquad (A.3)$$

où $\boldsymbol{\beta}_N = \mathbf{Q}_{xxN}^{-1}\mathbf{Q}_{xyN}$, $\bar{\mathbf{b}}_{\mathrm{HT}} = N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{b}_i$, $\mathbf{b}'_i = n^{-1}N\pi_i\zeta'_i e_i$,

$$\mathbf{Q}_{zzN} = \begin{pmatrix} Q_{yyN} & \mathbf{Q}_{yxN} \\ \mathbf{Q}_{xyN} & \mathbf{Q}_{xxN} \end{pmatrix}, \qquad (A.4)$$

$e_i = y_i - \mathbf{x}_i\boldsymbol{\beta}_N$, et ζ'_i est la colonne i de $\mathbf{X}'\boldsymbol{\Phi}^{-1}$.

Supposons que le plan d'échantillonnage est tel que

$$\mathbf{V}_{\bar{z}\bar{z}}^{-1/2}\{\bar{\mathbf{z}}_{\mathrm{HT}} - \bar{\mathbf{z}}_N \mid \mathbf{F}_N\} \xrightarrow{\mathrm{L}} N(\mathbf{0}, \mathbf{I}), \qquad (A.5)$$

lorsque $n_N \to \infty$ pour tout $\mathbf{z}$ dont le quatrième moment est fini, où $\mathbf{V}_{\bar{z}\bar{z}}$ est la matrice des covariances de $\bar{\mathbf{z}}_{\mathrm{HT}} - \bar{\mathbf{z}}_N$. Supposons que $\mathbf{V}_{\bar{z}\bar{z}}$ est $O(n^{-1})$ et que le plan d'échantillonnage admet un estimateur $\hat{\mathbf{V}}_{\bar{z}\bar{z}}$ tel que

$$n\left(\hat{\mathbf{V}}_{\bar{z}\bar{z}} - \mathbf{V}_{\bar{z}\bar{z}}\right) \mid \mathbf{F}_N = o_p(1) \qquad (A.6)$$

pour tout $\mathbf{z}$ dont le quatrième moment est borné. Alors

$$[\hat{V}\{\hat{\boldsymbol{\beta}}\}]^{-1/2}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N] \mid \mathbf{F}_N \xrightarrow{\mathrm{L}} N(\mathbf{0}, \mathbf{I}), \qquad (A.7)$$

où

$$\hat{V}\{\hat{\boldsymbol{\beta}}\} = \hat{\mathbf{Q}}_{xx}^{-1}\hat{\mathbf{V}}_{\bar{b}\bar{b}}\hat{\mathbf{Q}}_{xx}^{-1}, \qquad (A.8)$$

$\hat{\mathbf{V}}_{\bar{b}\bar{b}} = \hat{V}\{\bar{\mathbf{b}}'_{\mathrm{HT}}\}$ est l'estimation de la variance liée au plan d'échantillonnage de $\bar{\mathbf{b}}'_{\mathrm{HT}}$ calculé avec $\hat{\mathbf{b}}'_i = n^{-1}N\pi_i\zeta'_i\hat{e}_i$ et $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$.

Preuve. L'erreur dans $\hat{\boldsymbol{\beta}}$ est

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N &= \left(\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}\right)^{-1}\left[\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y} - \mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}\boldsymbol{\beta}_N\right] \\ &= \hat{\mathbf{Q}}_{xx}^{-1}(n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{e}).\end{aligned}$$

Maintenant, $\hat{\boldsymbol{\beta}}$ est un estimateur par les moindres carrés généralisés. Par conséquent,

$$\hat{\mathbf{e}}'\boldsymbol{\Phi}^{-1}\mathbf{X} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Phi}^{-1}\mathbf{X} = \mathbf{0}$$

et $\mathbf{Q}_{xyN} - \boldsymbol{\beta}'_N\mathbf{Q}_{xxN} = \mathbf{Q}_{exN} = \mathbf{0}$. Selon l'hypothèse (A.1)

$$\hat{\mathbf{Q}}'_{ex} = n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{e} = O_p(n^{-1/2}).$$

Donc,

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N &= \mathbf{Q}_{xxN}^{-1}\left(n^{-1}\sum_{i\in A}\zeta'_i e_i\right) + O_p(n^{-1}) \\ &= \mathbf{Q}_{xxN}^{-1}\left(N^{-1}\sum_{i\in A}\pi'_i\mathbf{b}'_i\right) + O_p(n^{-1}).\end{aligned}$$

Les quatrièmes moments des $\mathbf{b}_i$ sont bornés par les hypothèses. Donc, selon l'hypothèse (A.5)

$$\mathbf{V}_{\beta\beta}^{1/2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\right) \xrightarrow{\mathrm{L}} N(\mathbf{0}, \mathbf{I}),$$

où

$$\mathbf{V}_{\beta\beta} = \mathbf{Q}_{xxN}^{-1}\mathbf{V}_{\bar{b}\bar{b}}\mathbf{Q}_{xxN}^{-1}$$

et $\mathbf{V}_{\bar{b}\bar{b}} = V\{\bar{\mathbf{b}}_{\mathrm{HT}}\}$. Maintenant,

$$\begin{aligned}n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\hat{\mathbf{e}} &= n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{e} + n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}\left(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}\right) \\ &=: N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{b}'_i + N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{h}'_i,\end{aligned}$$

où

$$\mathbf{h}'_i = n^{-1}N\pi_i\zeta'_i\mathbf{x}_i\boldsymbol{\delta}_\beta$$

et $\boldsymbol{\delta}_\beta = \boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}$. Pour toute valeur fixée de $\boldsymbol{\delta}$, selon (A.6), l'estimation de la variance de $N^{-1}\sum_{i\in A}\pi_i^{-1}(\mathbf{b}'_i + \mathbf{h}'_i)$ converge vers la variance de l'estimateur de la moyenne de $\mathbf{b} + \mathbf{h}$. Par hypothèse, les éléments de $\zeta'_i\mathbf{x}_i$ ont un quatrième moment. Pour une valeur fixée de $\boldsymbol{\delta}$, la variance de $\bar{\mathbf{h}}_{\mathrm{HT}}$ est $O(n^{-1})$. Pour $\boldsymbol{\delta} = \boldsymbol{\delta}_\beta$,

$$\hat{V}\{\bar{\mathbf{h}}'_{\mathrm{HT}}\} = o_p(n^{-1}),$$

et

$$\hat{V}\{\bar{\mathbf{b}}'_{\mathrm{HT}}\} = V\{\bar{\mathbf{b}}'_{\mathrm{HT}}\} + o_p(n^{-1})$$

car $\delta_\beta = O_p(n^{-1/2})$. Le résultat (A.7) découle alors de la normalité asymptotique de $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N$.

Théorème A.2. Soit $\mathbf{y}' = (y_1, y_2, \cdots, y_n)$ et $\mathbf{X}' = (\mathbf{x}'_1, \mathbf{x}'_1, \cdots, \mathbf{x}'_n)$. Supposons que $\boldsymbol{\Phi}$ est une matrice $n \times n$ symétrique non singulière et que $\boldsymbol{\Phi}_N$ est une matrice $N \times N$ symétrique non singulière. Supposons que

$$\bar{y}_\pi, \bar{\mathbf{x}}_\pi, n^{-1}(\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}) \text{ and } n^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y}$$

sont des estimateurs convergents par rapport au plan d'échantillonnage des caractéristiques de la population finie $\bar{y}_N, \bar{\mathbf{x}}_N, \mathbf{Q}_{xxN}$ et $\mathbf{Q}_{xyN}$, respectivement, où

$$\left[\mathbf{Q}_{xxN}, \mathbf{Q}_{xyN}\right] = \left[N^{-1}\mathbf{X}'_N\boldsymbol{\Phi}_N^{-1}\mathbf{X}_N, N^{-1}\mathbf{X}_N\boldsymbol{\Phi}_N^{-1}\mathbf{y}_N\right]. \quad \text{(A.9)}$$

Soit $\boldsymbol{\beta}_N = \mathbf{Q}_{xxN}^{-1}\mathbf{Q}_{xyN}$. Supposons qu'il existe une suite de vecteurs colonnes $\{\boldsymbol{\gamma}_N\}$ tels que

$$\mathbf{X}\boldsymbol{\gamma}_N = \boldsymbol{\Phi}\mathbf{D}_\pi^{-1}\mathbf{J}_n \quad \text{(A.10)}$$

pour tous les échantillons possibles, où $\mathbf{D}_\pi = \mathrm{diag}(\pi_1, \pi_2, \cdots, \pi_n)$ et $\mathbf{J}_n$ est un vecteur colonne à n dimensions de 1. Alors, l'estimateur par régression $\bar{\mathbf{x}}_N\hat{\boldsymbol{\beta}}$ pour lequel

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y}, \quad \text{(A.11)}$$

est un estimateur convergent par rapport au plan d'échantillonnage de $\bar{y}_N$.

Preuve. Si $\hat{\boldsymbol{\beta}}$ est défini par (A.11), alors, compte tenu des propriétés des estimateurs par les moindres carrés généralisés,

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Phi}^{-1}\mathbf{X} = \mathbf{0}.$$

Si la condition (A.10) est satisfaite, alors

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{D}_\pi^{-1}\mathbf{J} = \left(\sum_{i \in A}\pi_i^{-1}\right)\left(\bar{y}_\pi - \bar{\mathbf{x}}_\pi\hat{\boldsymbol{\beta}}\right) = 0.$$

Il s'ensuit que $\bar{y}_{\mathrm{reg}}$ est convergent par rapport au plan d'échantillonnage, car

$$0 = p\lim_{N\to\infty}\left\{\left(\bar{y}_\pi - \bar{\mathbf{x}}_\pi\hat{\boldsymbol{\beta}}_n\right) \mid \mathbf{F}_N\right\}$$

$$= p\lim_{N\to\infty}\left\{\left(\bar{y}_\pi - \bar{\mathbf{x}}_\pi\boldsymbol{\beta}_N\right) \mid \mathbf{F}_N\right\}$$

$$= p\lim_{N\to\infty}\left\{\left(\bar{y}_N - \bar{\mathbf{x}}_N\boldsymbol{\beta}_N\right) \mid \mathbf{F}_N\right\}.$$

Théorème A.3. Soit une suite de populations et d'échantillons tels que définis dans le théorème A.1. Supposons que $\mathbf{z}_i$ est un vecteur de la forme $\mathbf{z}_i = (y_i, 1, \mathbf{x}_{1,i})$ et que $\mathbf{z}_{1,i} = (y_i, \mathbf{x}_{1,i})$. Supposons que $\bar{\mathbf{z}}_{1,\pi}$ est un estimateur convergent par rapport au plan d'échantillonnage de la moyenne de la population $\bar{\mathbf{z}}_{1,N}$ caractérisée par une matrice des covariances non singulières

$$V\{\bar{\mathbf{z}}_{1,\pi} \mid \mathbf{F}_N\} = O(n^{-1}) \quad \text{(A.12)}$$

et

$$n^{1/2}(\bar{\mathbf{z}}_{1,\pi} - \bar{\mathbf{z}}_{1,N}) \mid \mathbf{F}_N \xrightarrow{\mathrm{L}} N(\mathbf{0}, \boldsymbol{\Sigma}_{\bar{z}\bar{z}}), \quad \text{(A.13)}$$

où $\boldsymbol{\Sigma}_{\bar{z}\bar{z}}$ est la limite de $nV\{\bar{\mathbf{z}}_{1,\pi} \mid \mathbf{F}_N\}$. Supposons qu'il existe un estimateur de la variance de $\bar{\mathbf{z}}_{1,\pi}$, noté $\hat{V}\{\bar{\mathbf{z}}_{1,\pi}\}$, tel que

$$p\lim_{N\to\infty} n^{1+\delta}\left(\hat{V}\{\bar{\mathbf{z}}_{1,\pi}\} - V\{\bar{\mathbf{z}}_{1,\pi} \mid \mathbf{F}_N\}\right) = \mathbf{0} \quad \text{(A.14)}$$

pour $\delta > 0$. Représentons par $\hat{\boldsymbol{\beta}}_{1,\mathrm{dopt}}$ le vecteur qui minimise

$$\hat{V}\{\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi}\boldsymbol{\beta}_{1,d}\} \quad \text{(A.15)}$$

et par $\boldsymbol{\beta}_{1,\mathrm{dopt}}$ le vecteur qui minimise $V\{\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi}\boldsymbol{\beta}_{1,d}\}$. Supposons que $\bar{y}_{d,\mathrm{reg}}$ est défini par (4.29). Alors, $\bar{y}_{d,\mathrm{reg}}$ est, parmi les estimateurs convergents par rapport au plan d'échantillonnage de la forme $\bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})\boldsymbol{\beta}_{1,d}$, celui dont la variance limite est minimale. En outre,

$$\left[\hat{V}\{\bar{e}_\pi\}\right]^{-1/2}(\bar{y}_{d,\mathrm{reg}} - \bar{y}_N) \xrightarrow{\mathrm{L}} N(0, 1), \quad \text{(A.16)}$$

où $\hat{V}\{\bar{e}_\pi\}$ est l'estimateur (A.14) construit avec $\hat{e}_i = y_i - \bar{y}_\pi - (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})\hat{\boldsymbol{\beta}}_{1,\mathrm{dopt}}$.

Preuve. L'estimateur

$$\hat{\boldsymbol{\beta}}_{1,\mathrm{dopt}} = \left[\hat{V}\{\bar{\mathbf{x}}_{1,\pi}\}\right]^{-1}\hat{C}\{\bar{\mathbf{x}}_{1,\pi}, \bar{y}_\pi\}$$

minimise l'estimation de la variance de (A.15) et, en vertu de l'hypothèse (A.14), l'estimation de la variance converge vers la variance réelle. Donc, $\hat{\boldsymbol{\beta}}_{1,\mathrm{dopt}}$ est convergent par rapport au plan d'échantillonnage pour $\boldsymbol{\beta}_{1,\mathrm{dopt}}$ et $\boldsymbol{\beta}_{1,\mathrm{dopt}}$ minimise $V\{\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi}\boldsymbol{\beta}\}$. Par conséquent, aucun estimateur de la forme (4.29) n'a une distribution limite dont la variance est plus faible.

Maintenant

$$\bar{y}_{d,\mathrm{reg}} - \bar{y}_N = \bar{y}_\pi - \bar{y}_N - (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi})\hat{\boldsymbol{\beta}}_{1,\mathrm{dopt}}$$

$$= \bar{e}_\pi + o_p(n^{-1/2}),$$

où $e_i = y_i - \bar{y}_N - (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})\boldsymbol{\beta}_{1,\mathrm{dopt}}$. Par conséquent, la variance de la distribution limite de $n^{1/2}(\bar{y}_{d,\mathrm{reg}} - \bar{y}_N)$ est la variance de $n^{1/2}(\bar{e}_\pi - \bar{e}_N)$. D'après l'hypothèse (A.14), l'estimateur $\hat{V}\{\bar{\mathbf{z}}_\pi\boldsymbol{\gamma}\}$ est un estimateur convergent de la variance $V\{\bar{\mathbf{z}}_\pi\boldsymbol{\gamma}\}$ pour toute valeur fixée de $\boldsymbol{\gamma}$. Puisque $\hat{\boldsymbol{\beta}}_{1,\mathrm{dopt}} - \boldsymbol{\beta}_{1,\mathrm{dopt}} = o_p(1)$, l'estimation de la variance fondée sur $\hat{e}_i$ converge vers l'estimation de la variance fondée sur e_i et l'expression (A.16) est vérifiée.

BIBLIOGRAPHIE

ANDERSON, C., et NORDBERG, L. (1998). A user's guide to CLAN97. Statistics Sweden, Orebro, Sweden.

BANKIER, M.D., RATHWELL, S. et MAJKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Document de travail, Direction de la méthodologie, Census Operations Section, Social Survey Methods Division. Statistique Canada, Ottawa.

BANKIER, M.D., HOULE, A.M. et LUC, M. (1997). Calibration estimation in the 1991 and 1996 Canadian census. Statistique Canada (draft), 8 pages.

BARDSLEY, P., et CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

BETHLEHEM, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

BETHLEHEM, J.G., et KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

BISHOP, Y.M.M., FIENBERG, S.E. et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

BREIDT, F.J., et OPSOMER, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 1026-1053.

BREWER, K.R.W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

BREWER, K.R.W. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association*, 74, 911-915.

BREWER, K.R.W., HANIF, M. et TAM, S.M. (1988). How nearly can model-based prediction and design-based estimation be reconciled? *Journal of the American Statistical Association*, 83, 128-132.

BRICK, J.M., WAKSBERG, J. et KEETER, S. (1996). Utilisation des données sur les interruptions de service téléphonique pour ajuster la couverture. *Techniques d'enquête*, 22, 187-199.

CASSEL, C.M., SÄRNDAL, C.-E. et WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

CASSEL, C.M., SÄRNDAL, C.-E. et WRETMAN, J.H. (1979). Prediction theory for finite populations when model-based and design-based principles are combined. *Scandinavian Journal of Statistics*, 6, 97-106.

CASSEL, C.M., SÄRNDAL, C.-E. et WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. Dans *Incomplete Data in Sample Surveys*, (Éds. W.G. Madow, I. Olkin, et D. Rubin). New York: Academic Press, 3, 143-160.

CHAMBERS, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

CHAMBERS, R.L., DORFMAN, A.H. et WEHRLY, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.

COCHRAN, W.G. (1939). The use of the analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.

COCHRAN, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.

COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.

COCHRAN, W.G. (1977). *Sampling Techniques*, 3e éd., New York: John Wiley & Sons, Inc.

COOK, R.D., et WEISBERG, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

DEMING, W.E., et STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

DEMING, W.E., et STEPHAN, F.F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36, 45-49.

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

DEVILLE, J., SÄRNDAL, C.-E. et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

DORFMAN, A.H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics*, 35, 29-41.

DORFMAN, A.H., et HALL, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics*, 21, 1452-1475.

DUCHESNE, P. (2000). A note on jackknife variance estimation for the general regression estimator. *Journal of Official Statistics*, 16, 133-138.

DU MOUCHEL, W. H., et DUNCAN, G. J. (1983). Using survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.

ELTINGE, J.L., et YANSANEH, I.S. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. *Techniques d'enquête*, 23, 37-45.

ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31, 195-224.

ESTEVAO, V., HIDIROGLOU, M.A. et SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

FIRTH, D., et BENNETT, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society B*, 60, 3-21.

FOLSOM, R.E., et WITT, M.B. (1994). Testing a new attrition nonresponse adjustment method for SIPP. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 428-433.

FOLSOM, R.E., et SINGH, A.C. (2000). The generalized exponential model for a unified approach to sampling weight calibration for outlier weight treatment, nonresponse adjustment and post-stratification. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 598-603.

FRANKEL, M.R. (1971). Inference from survey samples: An empirical investigation. Institute for Social Research, University of Michigan, Ann Arbor.

FULLER, W.A. (1966). Estimation employing post strata. *Journal of the America Statistical Association*, 61, 1172-1183.

FULLER, W.A. (1973). Regression for sample surveys. Un article présenté à la réunion de International Statistical Institute. Août, 1973, Vienna, Austria.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā Series C*, 37, 117-132.

FULLER, W.A. (1984). Application de la méthode des moindres carrés et de techniques connexes aux plans de sondage complexes. *Techniques d'enquête*, 10, 107-137.

FULLER, W.A., et AN, A.B. (1998). Regression adjustments for nonresponse. *Journal of the Indian Society of Agricultural Statistics*, 51, 331-342.

FULLER, W.A., et BATTESE, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68, 626-632.

FULLER, W.A., et ISAKI, C.T. (1981). Survey design under superpopulation models. Dans *Current Topics in Survey Sampling*, (D. Krewski, J. N. K. Rao et R. Platek, Éds.), New York: Academic Press, 199-226.

FULLER, W.A., LOUGHIN, M.M. et BAKER, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20 79-89.

FULLER, W.A., et RAO, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 27, 49-56.

GAMBINO, J., KENNEDY, B. et SINGH, M.P. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada : Évaluation et application. *Techniques d'enquête*, 27, 69-79.

GEROW, K., et MCCULLOCH, C.E. (2000). Simultaneously model unbiased, design-unbiased estimation. *Biometrics* 56, 873-878.

GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17, 269-278.

GODAMBE, V.P., et JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations, 1. *Annals of Mathematical Statistics*, 36, 1707-1722.

GOLDBERGER, A.S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57, 369-375.

GRAYBILL, F.A. (1976). *Theory and application of the linear model*. Wadsworth, Belmont, CA.

HÁJEK, J. (1959). Optimum strategy and other problems in probability sampling. *Casopis Pro Pestovani Matematiky*, 84, 387-423.

HANURAV, T.V. (1966). Some aspects of unified sampling theory. *Sankhyā, Series A*, 28, 175-204.

HENDERSON, C.R. (1963). Selection index and expected genetic advance. Dans *Statistical Genetics and Plant Breeding*, 141-163. National Academy Sciences, National Research Council Publication 982, Washington, DC.

HARVILLE, D.A. (1976). Extension of the Gauss-Markov Theroem to include estimation of random effects. *Annals of Statistics*, 4, 384-395.

HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Thèse de doctorat, Iowa State University, Ames, Iowa.

HIDIROGLOU, M.A., FULLER, W.A. et HICKMAN, R.D. (1978). *Super Carp*, (6e édition, 1980) Survey Section, Statistical Laboratory, Iowa State University, Ames, Iowa.

HIDIROGLOU, M.A., SÄRNDAL, C.-E. et BINDER, D. A. (1995). Weighting and estimation in business surveys. *Business Survey Methods*, (Éds. Cox, Binder, Chinnappa, Colledge et Kott) New York: John Wiley & Sons, Inc., 477-502.

HOLT, D., et SMITH, T.M. F. (1979). Post Stratification. *Journal of the Royal Statistical Society, Serie. A*, 142, 33-46.

HORN, S.D., HORN, R.A. et DUNCAN, D.B. (1975). Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 70, 380-385.

HUANG, E.T., et FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the social statistics section*, American Statistical Association, 300-305.

HUSAIN, M. (1969). Construction of regression weights for estimation in sample surveys. Thèse de maîtrise non-publiée, Iowa State University, Ames, Iowa.

ISAKI, C.T. (1970). Survey designs utilizing prior information. Thèse de doctorat non-publiée. Iowa State University.

ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

ISAKI, C.T., TSAY, J.H. et FULLER, W.A. (2000). Estimation des facteurs de correction au recensement. *Techniques d'enquête*, 26, 37-49.

JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agriculture Experiment Station Research Bulletin*. 304

KALTON, G., et MALIGALIG, D.S. (1991). A Comparison of Methods of Weighting Adjustment for Nonresponse. *Proceedings of the 1991 Annual Research Conference*, U. S. Bureau of the Census, 409-428.

KISH, L., et FRANKEL, M.R. (1974). Inference from complex samples (avec discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.

KONIJN, H.S. (1962). Regression analysis for sample surveys. *Journal of the American Statistical Association*, 57, 590-606.

KOTT, P.S. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89, 693-696.

KUO, L. (1988). Classical and prediction approaches to estimating distribution function from survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 280-285.

LAZZERONI, L.C., et LITTLE, R.J.A. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, 14, 61-78.

LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

LITTLE, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

MADOW, W.G., et MADOW, L.H. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15, 1-24.

MICKEY, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.

MONTANARI, G.E. (1987). Post-sampling efficient Q-R prediction in large-sample surveys. *International Statistical Review*, 55, 191-202.

MONTANARI, G.E. (1999). A study on the conditional properties of finite population mean estimators. *Metron*, 57, 21-35.

MUKHOPADHYAY, P. (1993). Estimation of a finite population total under regression models: A review. *Sankhyā*, 55, 141-155.

NIEUWENBROEK, N., RENSSEN, R. et HOFMAN, L. (2000). Towards a generalized weighting system. Dans *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria, Virginia.

PARK, M. (2002). Regression estimation of the mean in Survey Sampling. Thèse de doctorat non-publiée dissertation, Iowa State University, Ames, Iowa

PFEFFERMANN, D. (1984). Note on large sample properties of balanced samples. *Journal of the Royal Statistical Society, Series B*, 46, 38-41.

RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

RAO, J.N.K. (2002). *Small Area Estimation Theory and Methods*, New York: John Wiley & Sons, Inc.

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

RAO, J.N.K., et SINGH, A.C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 57-64.

ROBINSON, G.K. (1991). The BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6, 15-32.

ROBINSON, P.M., et SÄRNDAL, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā, Series B*, 45, 240-248.

ROSENBAUM, P.R., et RUBIN, D.B. (1983). The central role of the propensity score in observational studies for casual effects. *Biometrika*, 70, 41-55.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

ROYALL, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Association*, 71, 657-664.

ROYALL, R.M. (1986). The prediction approach to robust variance estimation in two stage cluster sampling. *Journal of the American Statistical Association*, 81, 119-123.

ROYALL, R.M., et CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.

ROYALL, R.M., et CUMBERLAND, W.G. (1981). The finite population linear regression estimator and estimators of its variance, an empirical study. *Journal of the American Statistical Association*, 76, 924-930.

SÄRNDAL, C.-E. (1980). On π-weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SÄRNDAL, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.

SÄRNDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistics Association*, 91, 1289-1300.

SÄRNDAL, C.-E., SWENSON, B. et WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

SÄRNDAL, C.-E., SWENSON, B. et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SÄRNDAL, C.-E., et WRIGHT, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*. 11, 146-156.

SCOTT, A., et SMITH, T.M.F. (1974). Linear superpopulation models in survey and sampling. *Sankhyā, C*, 36, 143-146.

SCOTT, A., et WU, C.F. (1981). On the asymptotic distribution of ratio and regression estimators. *Journal of the American Statistical Association*, 76, 98-102.

SILVA, P.L.D.N., et SKINNER, C.J. (1997). Sélection des variables pour l'estimation par régression dans le cas des populations finies. *Techniques d'enquête*, 23, 25-35.

SINGH, A.C., et FOLSOM, R.E. (2000). Bias corrected estimating functions approach for variance estimation adjusted for poststratification. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 610-615.

SINGH, A.C., KENNEDY, B. et WU, S. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel. *Techniques d'enquête*, 27, 35-48.

SINGH, A.C., et MOHL, C.A. (1996). Comprendre les estimateurs de calage dans les enquêtes par échantillonnage. *Techniques d'enquête*, 22, 107-116.

TALLIS, G.M. (1978). Note on robust estimation infinite populations. *Sankhyā C*, 40, 136-138.

TAM, S.M. (1986). Characterization of best model-based predictors in survey sampling. *Biometrika*, 73, 232-235.

THÉBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

THÉBERGE, A. (2000). Calage et poids restreints. *Techniques d'enquête*, 26, 113-122.

TILLE, Y. (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. *International Statistical Review*, 66, 303-322.

TREMBLAY, V. (1986). Critères pratiques pour la définition des classes de pondération. *Techniques d'enquête*, 12, 91-103.

WATSON, D. J. (1937). The estimation of leaf area in field crops. *Journal of Agricultural Science*, 27, 474-483.

WOODRUFF, R.S., et CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

WU, C., et SITTER, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

YATES, F. (1949). *Sampling Methods for Census and Surveys*. London: Griffin.

YUNG, W., et RAO, J.N.K. (1996). Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples. *Techniques d'enquête*, 22, 23-31.

ZYSKIND, G. (1976). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Annals of Mathematical Statistics*, 38, 1092-1109.

Techniques d'enquête, juin 2002
Vol. 28, n° 1, pp. 27-32
Statistique Canada

L'influence de Leslie Kish sur la statistique d'enquête

GRAHAM KALTON[1]

RÉSUMÉ

Leslie Kish, l'un des pionniers de l'échantillonnage, est décédé le 7 octobre 2000 à l'âge de 90 ans. Le présent article donne un aperçu de l'influence qu'il a exercée sur la statistique d'enquête, principalement de ses travaux de recherche, mais aussi des efforts qu'il a déployés pour promouvoir l'application de méthodes d'échantillonnage probabiliste valides partout dans le monde. De grande portée, les travaux de recherche de Kish couvrent les méthodes d'échantillonnage, l'estimation de la variance et des effets de plans de sondage, les erreurs non dues à l'échantillonnage, l'estimation régionale, la conception d'enquête à dimensions temporelle et spatiale et les études par observation. Il a fait connaître les plans d'échantillonnage probabiliste grâce à ses activités d'expert-conseil dans de nombreux pays, à ses écrits et, tout spécialement, au programme d'été très efficace sur l'échantillonnage à l'intention des statisticiens étrangers qu'il a créé au Survey Research Center de l'Université du Michigan.

MOTS CLÉS : Plan d'échantillonnage; estimation de la variance; erreurs non dues à l'échantillonnage; échantillons successifs.

1. INTRODUCTION

Leslie Kish, l'un des pionniers de l'échantillonnage, est décédé le 7 octobre 2000 à l'âge de 90 ans. Durant sa longue et fructueuse carrière, il a exercé une influence considérable dans le domaine, grâce à ses travaux de recherche impressionnants ainsi qu'à la manière extrêmement efficace avec laquelle il a fait connaître les méthodes scientifiques d'échantillonnage probabiliste partout dans le monde, surtout dans les pays en voie de développement. Ses travaux, dont la portée est vaste, se concentraient toujours sur des questions d'importance pratique et ses innovations ont facilité l'utilisation de méthodes efficaces d'échantillonnage probabiliste dans divers domaines. Il a mis en valeur la pratique de l'échantillonnage probabiliste grâce à ses écrits magistraux (destinés surtout aux sociologues et aux démographes), à ses nombreux services de consultation et à la formation qu'il prodiguait aux statisticiens d'enquête, particulièrement ceux venant des pays en voie de développement.

Le présent article donne une idée de l'influence exercée par Kish sur la statistique d'enquête, surtout du rôle qu'il a joué dans l'avancement des méthodes d'échantillonnage d'enquête et, de façon plus générale, de ses travaux de recherche. Une brève description de sa carrière permettra de placer ses réalisations dans leur contexte. Les lecteurs qui souhaiteraient avoir plus de détails sur la vie fascinante de Kish sont invités à consulter le compte rendu de son entretien de 1994 avec Frankel et King (1996). Certains renseignements présentés ici en sont extraits.

Kish est né en 1910 à Poprad, qui faisait alors partie de l'empire Austro-hongrois, mais qui appartient aujourd'hui à la Slovaquie. En 1926, lui et sa famille ont immigré aux États-Unis. À la suite du décès de son père l'année suivante, il a occupé un poste d'assistant de laboratoire au Rockefeller Institute for Medical Research, alors qu'il fréquentait la Bay Ridge Evening High School. Il a obtenu son diplôme d'études secondaires en 1930 et s'est inscrit au cours du soir du College of the City of New York tout en continuant à travailler 54 heures par semaine au Rockefeller Institute. C'est durant son travail à l'Institut qu'il s'est passionné pour la statistique et qu'il a commencé à étudier de lui-même les traités rédigés par Fisher, Yule, Wallace et Snedecor, Tippett, Pearl et d'autres. En 1937, il a interrompu ses études pour se joindre à l'International Brigade dans sa lutte pour la cause loyaliste durant la guerre civile d'Espagne. Il est retourné aux États-Unis en 1939 où il a obtenu un baccalauréat en mathématiques avec distinction la même année. Après avoir été recruté par le U.S. Census Bureau à titre de chef de section, il a obtenu un poste de statisticien à la Division of Program Surveys du United States Department of Agriculture (USDA). En 1942, il a quitté la Division parce qu'il était appelé sous les drapeaux et y est retourné en 1945 une fois la guerre terminée. En 1947, il a déménagé avec un groupe de collègues de l'USDA dirigé par Rensis Likert afin d'établir le Survey Research Center à l'Université du Michigan, où il est resté jusqu'à sa retraite en 1981, après être devenu professeur émérite. Il a poursuivi activement ses activités professionnelles jusqu'à son décès en 2000.

2. RECHERCHE

Au début de la carrière de Kish, la science de l'échantillonnage en était à ses premiers balbutiements. Nombre d'enquêtes par sondage se fondaient sur des échantillons non probabilistes. Les méthodes d'échantillonnage

[1] Graham Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850.

probabiliste étaient à l'étude et de nombreux problèmes restaient à résoudre. Pendant qu'il travaillait au USDA, Kish a isolé trois problèmes importants dont il a poursuivi la résolution au Survey Research Center (SRC) dans le cadre de ses travaux de mise au point des méthodes d'échantillonnage.

L'un de ces problèmes consistait à déterminer comment un intervieweur pouvait sélectionner aléatoirement une personne dans un ménage échantillonné. À l'époque, on avait mis au point des méthodes d'échantillonnage probabiliste pour l'échantillonnage des ménages que l'on appliquait à la Current Population Survey. Mais cette enquête était conçue pour recueillir des données sur tous les membres des ménages échantillonnés, si bien qu'il ne fallait sélectionner personne dans les ménages. Kish a inventé une méthode de sélection objective d'un répondant et l'a présentée par écrit dans un mémoire. Son collègue Angus Campbell a insisté pour qu'il soumette son travail à une revue pour publication, ce qui a donné le fameux article qui a été sa première étude publiée (Kish 1949). La méthode, dont l'usage est maintenant très répandu, porte le nom de table de sélection de Kish.

Le deuxième problème isolé par Kish était celui du dénombrement de la non-réponse. Pour défendre son idée du dénombrement et de la déclaration de la non-réponse dans le cas de l'échantillonnage probabiliste, il a dû argumenter contre ses collègues qui craignaient que cette pratique désavantage le SRC, relativement aux organismes utilisant des méthodes non probabilistes. Kish a eu gain de cause et le SRC a adopté sa méthode, qui est maintenant pleinement reconnue comme une norme de bonne pratique.

Le troisième problème était celui de la stratification en profondeur. La stratification classique se fonde sur l'hypothèse de l'indépendance des sélections entre strates, le nombre maximal possible de strates étant égal au nombre de sélections. Surtout si le nombre de sélections est faible, comme cela est souvent le cas pour les unités primaires d'échantillonnage (UPE) dans un plan de sondage à plusieurs degrés, il est parfois souhaitable d'obtenir un échantillon mieux équilibré que ne le permet la stratification classique. En collaboration avec Roe Goodman, Kish a mis au point la technique de sélection contrôlée qui donne ce meilleur équilibre grâce à l'abandon de l'hypothèse de l'indépendance des sélections entre strates, tout en maintenant un échantillonnage probabiliste (Goodman et Kish 1950). Kish, qui cherchait toujours à trouver de bons noms, préférait qualifier la technique de « stratification multiple », expression qu'il a utilisée dans son traité sur l'échantillonnage (Kish 1965a).

Les travaux de recherche subséquents de Kish dans le domaine de la statistique d'enquête ont été de grande portée, couvrant de nombreux aspects des méthodes d'échantillonnage, des erreurs non dues à l'échantillonnage, de l'estimation régionale, de la conception d'enquêtes à composantes temporelle et spatiale et des études par observation. Ses nombreuses réalisations ont influencé considérablement l'évolution des pratiques d'échantillonnage en particulier et des études par sondage en général. Les paragraphes qui suivent, qui donnent un aperçu de ses travaux, sont classés par sujet.

Estimation de la variance. Avant les années 1970, l'analyse des données d'enquête était très limitée, compte tenu des outils analytiques existants, c'est-à-dire principalement des machines à cartes perforées, telles que des compteuses-trieuses et des totalisatrices, ainsi que des calculatrices manuelles. Donc, des statistiques comme les coefficients de pondération, particulièrement ceux qui n'étaient pas des nombres entiers, étaient difficiles à manipuler. C'est pour cette raison que Kish a étudié l'utilisation de coefficients de pondération uniformes grâce à la table de sélection de Kish, même si l'estimation non biaisée demande l'application de poids proportionnels au nombre de membres admissibles du ménage.

À cause de la complexité des calculs, avant les années 1970, on estimait rarement les erreurs d'échantillonnage d'une façon qui reflétait les plans d'échantillonnage complexes habituellement employés pour les études par sondage. Une pratique courante consistait à calculer les variances comme si l'on avait sélectionné un échantillon aléatoire simple (EAS). Kish s'est efforcé de faire valoir l'utilisation des méthodes appropriées d'estimation de la variance auprès des chercheurs du domaine social en illustrant la sous-estimation importante qui a souvent lieu lorsque l'on applique des formules d'EAS à des échantillons en grappes (Kish 1957). Au départ, il a mis au point et appliqué des méthodes de calcul simples, en insistant sur la simplicité d'un plan de sondage apparié en vertu duquel deux UPE sont échantillonnées dans chaque strate (Kish et Hess 1959a; Kish 1968). Il a inventé l'expression « effet de plan de sondage » pour désigner le rapport de la variance d'une estimation d'enquête obtenue au moyen d'un plan de sondage donné à la variance de la même estimation obtenue d'après un échantillon aléatoire simple de même taille. Il a utilisé abondamment ce concept dans son fameux traité intitulé *Survey Sampling* (Kish 1965a), qui décrit de façon encyclopédique la pratique de l'échantillonnage d'enquête et qui demeure l'un des classiques publiés par Wiley lu très fréquemment. Tout au long de sa carrière, Kish n'a cessé de s'intéresser aux effets de plan de sondage qu'il considérait comme un instrument important de conception et d'analyse des échantillons d'enquête (voir, par exemple, Kish 1982, 1995a; Kish, Frankel, Verma et Kaciroti 1995; Kish, Groves et Krotki 1976). Dans le cas d'un échantillon en grappes, un terme important de la formule de l'effet de plan de sondage est celui de la corrélation intraclasse, dont Kish a discuté dans sa thèse de doctorat (Kish 1952) et dans plusieurs autres articles (par exemple, Kish 1954, 1961a).

Lorsque l'ordinateur a été développé, Kish a rapidement saisi l'importance de cet outil pour l'estimation de la variance et, avec ses collègues du SRC, a développé un premier logiciel intitulé *Sampling Error Program Package* (Kish, Frankel et Van Eck 1972). En collaboration avec

Martin Frankel, étudiant de doctorat dont il avait la supervision, il a également étendu la gamme des statistiques pour lesquelles il était possible de calculer les erreurs d'échantillonnage dans le cas des plans de sondage complexes (Kish et Frankel 1970, 1974). Ces travaux de recherche d'une très grande importance ont mené au développement, à l'application et à l'évaluation de la méthode des répétitions équilibrées répétées, ou BRR pour balanced repeated replication, et de la méthode des répétitions jackknife répétées, ou JRR pour jackknife repeated replication, pour estimer la variance. Il a aussi proposé une définition des paramètres de population estimés d'après des statistiques analytiques fondées sur des données d'enquête dans le contexte des populations finies.

Enquêtes polyvalentes. Les articles sur l'échantillonnage traitent principalement de l'élaboration d'un plan d'échantillonnage efficace pour estimer un seul paramètre d'une population. Kish a reconnu les limites de cette approche, puisque presque toutes les enquêtes ont un caractère polyvalent. Il a rédigé plusieurs articles importants traitant des enquêtes polyvalentes et a produit des plans de sondage de compromis efficaces qui fournissent des estimations non seulement pour l'ensemble de la population, mais aussi pour divers domaines (Kish 1961b, 1969, 1976; Anderson, Kish et Cornell 1976; Kish et Anderson 1978; Kish 1980; Kish 1988). Ces dernières années, il avait étendu son étude pour s'intéresser aux enquêtes à populations multiples (par exemple, Kish 1999, 2002).

Estimation régionale. Afin de produire des estimations par domaine, Kish (1980, 1987a, 1987b) a classifié ces derniers en grands, petits et mini domaines et en éléments rares. Pour les grands domaines, on peut produire les estimations avec les données d'enquête en se servant d'estimateurs types basés sur l'échantillon, particulièrement si ce dernier est conçu pour enquêter un nombre suffisant d'unités appartenant aux domaines en question. Toutefois, la taille de l'échantillon de la plupart des enquêtes empêche de produire des estimations suffisamment précises pour les petits ou les mini domaines qui comprennent moins de, disons, un dixième de la population. Pourtant, comme Kish l'a compris rapidement, la demande d'estimations à jour pour de petits domaines, particulièrement les petites régions géographiques, allait augmenter. Sa sensibilisation à ce problème est à l'origine des travaux qu'il a menés dans deux domaines connexes.

Si la taille de l'échantillon d'une enquête est trop petite pour que l'on puisse produire des estimations régionales d'une précision suffisante, on peut recourir à des modèles statistiques pour produire des estimations indirectes. De nombreux travaux portant sur les techniques d'estimation régionale au moyen de modèles ont été réalisés ces dernières années. Durant les années 1970, Kish a contribué à l'avancement dans ce domaine à titre de directeur de thèse de trois étudiants de doctorat de l'Université du Michigan (Ericksen 1973; Kalsbeek 1973; Purcell et Kish 1979, 1980).

Parfois, il est possible de produire des estimations régionales directes ou fondées sur des données d'enquête. Les données du Recensement de la population sont une source manifeste d'estimations pour les domaines de toutes tailles et sont effectivement l'une des sources principales d'estimations régionales. Cependant, les données d'un recensement décennal deviennent périmées à mesure que l'on progresse vers la décennie suivante. Pour régler ce problème, Kish a proposé de remplacer le recensement par des échantillons à renouvellement complet, ou échantillons successifs, de sorte que l'on puisse produire des estimations plus à jour grâce à l'étalement de la collecte des données au fil du temps. Il a proposé une méthode de ce genre pour la première fois en 1979 (Kish 1979a,b) et, après cela, a rédigé de nombreux articles traitant de ce sujet (Kish 1981, 1983, 1986, 1990, 1997, 1998, 2002; Kish et Verma 1986), y compris celui sur la façon de cumuler les données d'enquête au fil du temps (Kish 1999). Dans un autre article du présent volume, Charles Alexander (2001) fait un compte rendu détaillé des travaux de Kish dans ce domaine et du rôle qu'il a joué dans l'élaboration de l'American Community Survey, enquête continue à grande échelle que le U.S. Census Bureau prévoit lancer pour remplacer le questionnaire détaillé du Recensement de 2010.

Problèmes spéciaux d'échantillonnage. Au cours de sa carrière, Kish a observé à maintes reprises des problèmes particuliers d'échantillonnage qui se manifestent fréquemment et a proposé certaines solutions efficaces. Les domaines auxquels il a contribué sont :

- Échantillonnage de populations rares ou imprécises. La production d'un plan d'échantillonnage efficace pour une population rare ou imprécise (comme les personnes atteintes d'une maladie rare ou les sans-abri) est l'un des plus grands défis que doivent relever les statisticiens d'enquêtes. Kish (1965b, 1991) a rédigé des revues instructives des méthodes qui permettent de s'attaquer à ce genre de problème.
- Maximisation du chevauchement. Lorsqu'une population est échantillonnée de façon répétée au fil du temps, il faut savoir comment contrôler le chevauchement entre les échantillons des cycles successifs. Un cas particulier est celui où l'on utilise un échantillon maître d'UPE et que celui-ci doit être mis à jour lorsque les nouvelles données de recensement sont diffusées. Fréquemment, il est souhaitable de maximiser le chevauchement dans l'échantillon de l'UPE tout en mettant à jour les mesures de taille et en modifiant la stratification afin de refléter les données courantes. Kish et Scott (1971) ont proposé une méthode assez simple et efficace pour répondre à ces exigences.
- Échantillonnage d'organismes de taille inégale. Certaines enquêtes sont conçues pour produire des estimations sur des unités de niveaux différents,

comme les hôpitaux et les patients. Si le nombre de patients varie considérablement selon l'hôpital, la production de bonnes estimations au niveau de l'hôpital et au niveau des patients pose un problème. Kish (1965c) a examiné ce problème et précisé les éléments qui entrent en jeu.

Erreurs non dues à l'échantillonnage. Kish a reconnu clairement les effets néfastes que peuvent avoir les erreurs non dues à l'échantillonnage sur la qualité des estimations produites par sondage. Au tout début de sa carrière, il a étudié, en collaboration de Jack Lansing, les erreurs de réponse lors de la déclaration de la valeur du domicile faite par les répondants en comparant ces déclarations aux estimations faites par des évaluateurs professionnels (Kish et Lansing 1954). Dans son étude de la variance due à l'intervieweur, en s'inspirant de la théorie de l'échantillonnage en grappes, il a mesuré la variance due à l'intervieweur au moyen du coefficient de corrélation intraclasse et déterminé le nombre optimal d'interviews par intervieweur en se fondant sur un modèle de coût de l'échantillonnage en grappes simple (Kish 1962). Avec Irene Hess, il a réalisé une étude de la non-couverture dans le cas d'échantillons aréolaires de logements. L'étude, lancée parce que le taux de non-couverture des enquêtes du SRC était de 10 % à cette époque-là, a permis d'apporter des améliorations qui l'ont réduit à environ 3 % (Kish et Hess 1958). Toujours avec Irene Hess, il a introduit une méthode novatrice de remplacement des personnes non contactées lors d'une enquête qui consistait à les substituer par des personnes non contactées d'une enquête similaire antérieure (Kish et Hess 1959b). En ce qui concerne les scénarios d'imputation stochastique, Kish a été l'un des premiers adeptes de la répétition des imputations en vue de réduire la variance d'imputation, méthode qu'il a baptisée à l'époque imputation à répétitions répétées, ou RRIP pour repeated replication imputation procedure, et qui est connue aujourd'hui sous le nom d'imputation fractionnaire (Kalton et Kish 1984).

Études par observation. Au début de sa carrière, Kish (1959) a rédigé un article cité très fréquemment sur la conception d'études destinées à étudier les liens de causalité, particulièrement des études non randomisées. Dans ses écrits à ce sujet, il s'est appuyé sur son expérience en échantillonnage, comme celle concernant le lien entre la stratification et l'appariement (Anderson, Kish et Cornell 1980). Ses travaux ont mené à la publication de l'ouvrage intitulé *Statistical Design for Research* (Kish 1987a) dans lequel il compare l'utilisation d'enquêtes, d'expériences et d'études par observation pour étudier les effets de causalité en fonction des trois R : réalisme, randomisation et représentativité (voir aussi Kish 1975). Il a également souligné l'importance qu'il y a à évaluer le biais et la variance pour déterminer la capacité qu'ont divers plans d'études de mesurer les effets de causalité, au lieu de se concentrer uniquement sur le biais, comme le faisaient la plupart des auteurs ayant traité du sujet.

3. AUTRES CONTRIBUTIONS

Les travaux fondamentaux et de grande portée de Kish dans le domaine de la méthodologie de la statistique d'enquête sont d'une importance considérable. Toutefois, les efforts qu'il a déployés pour promouvoir l'utilisation de méthodes d'échantillonnage probabiliste valides partout dans le monde le sont peut-être plus encore.

Naturellement, les écrits de Kish, qui mettent l'accent sur l'adoption de bonnes méthodes pratiques, ont contribué à l'usage généralisé des méthodes d'échantillonnage probabiliste. Ces trois ouvrages intitulés Survey Sampling (Kish 1965a), Statistical Design for Research (Kish 1987a) et Statistical Methods for Agricultural Surveys (Kish 1989) sont tous fort précieux à cet égard, tout comme le sont ses articles magistraux à l'intention des spécialistes des sciences sociales.

Kish a toujours eu la vocation d'aider les pays en voie de développement et les pays en transition, et nombre de ses activités en témoignent. Il a été consultant en échantillonnage pour la World Fertility Survey de 1973 à 1983 et a donné des consultations dans de nombreux pays. Il s'occupait de l'organisation d'un programme de formation à l'intention des statisticiens étrangers et il a rédigé certains articles s'adressant spécialement aux statisticiens des pays en voie de développement. Ainsi, Statistical Methods for Agricultural Surveys a été écrit pour la Food and Agriculture Organisation, particulièrement à l'intention de ces pays. De 1978 à 1994, il a contribué à la rubrique Questions/Réponses du bulletin intitulé Statisticien d'enquête de l'Association internationale des statisticiens d'enquêtes. Il y donnait d'excellents conseils au sujet de nombreux problèmes pratiques d'échantillonnage qui se posent fréquemment, mais qui ne sont pas toujours bien traités dans les publications. La rubrique était jugée si utile que l'AISS a publié l'ensemble complet de questions et de réponses dans un volume spécial (Kish 1995b).

Kish était, à juste titre, très fier du programme intensif de deux mois sur l'échantillonnage à l'intention des statisticiens étrangers (Sampling Program for Foreign Statisticians) qu'il avait créé au Survey Research Center en 1961. Aujourd'hui, le SPFS a formé plus de 500 statisticiens d'enquêtes en provenance de 105 pays. Il est significatif que Kish ait choisi « Développement d'échantilllonneurs pour les pays en voie de développement » comme thème de la Morris Hansen Memorial Lecture de 1994 (Kish 1996). Pour assurer le maintien de ce programme important, lors d'une cérémonie commémorant les 90 ans de Kish, l'Université du Michigan a créé le Leslie Kish International Fellows Fund. De toutes ses réalisations, le SPFS était celle qui lui procurait le plus grand plaisir.

4. CONCLUSION

Leslie Kish est un géant dans le domaine de l'échantillonnage. Ses travaux impressionnants lui ont valu de nombreux honneurs. Ceux-ci incluent, entre autres, la fonction de président de l'Association internationale des statisticiens d'enquêtes de 1983 à 1985, la fonction de président de l'American Statistical Association en 1978 (voir Kish, 1978 pour son discours du président ayant pour thème « Chance, Statistics and Statisticians »), les titres d'Honorary Fellow de l'Institut international de statistique, d'Honorary Fellow de la Royal Statistical Society, d'Honorary Member de la Hungarian Academy of Sciences, de Fellow de l'American Association for the Advancement of Science, de Fellow de l'American Academy of Arts and Sciences, de lauréat de l'American Statistical Association's Samuel L. Wilks Award en 1997, de lauréat du Mindel Shep Award de la Population Association of America en 1998, de lauréat du Methodology Award de l'American Sociological Association en 1989, ainsi que l'octroi de grades honoris causa de l'Université de Bologne, de l'Athens University of Economics and Business et de l'Eotvos Lorand University de Budapest.

Pourtant, Kish a toujours eu les pieds sur terre et était ouvert à tous. Il discutait avec enthousiasme de nombreux sujets, y compris de sport, d'art, de littérature, de politique, de philosophie et de science. Il se souciait sans cesse d'améliorer les conditions de la population mondiale. Il s'intéressait tout spécialement aux jeunes et l'une de ses phrases préférées était « Restez jeune en étant curieux et ayez de jeunes amis ». Sans aucun doute, sa personnalité attachante a joué un rôle important dans le succès avec lequel il a fait la promotion de l'utilisation de bonnes méthodes d'échantillonnage partout dans le monde.

BIBLIOGRAPHIE

ALEXANDER, C. H. (2002). Les échantillons successifs de Leslie Kish et l'American Community Survey. *Techniques d'enquête*, 28, 37-44.

ANDERSON, D.W., KISH, L. et CORNELL, R.G. (1976). Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. *Journal of the American Statistical Association*, 71, 887-892.

ANDERSON, D.W., KISH, L. et CORNELL, R.G. (1980). On stratification, grouping, and matching. *Scandinavian Journal of Statistics*, 7, 61-66.

ERICKSEN, E. P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography*, 10, 137-160.

FELLEGI, I.P. (2000). Leslie Kish – Une vie de dévouement. *Techniques d'enquête*, 26, 133-134.

FRANKEL, M., et KING, B. (1996). A conversation with Leslie Kish. *Statistical Science*, 11, 65-87.

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

KALSBEEK, W.D. (1973). A Method for Obtaining Local Postcensal Estimates for Several Types of Variables. Thèse de doctorat, University of Michigan.

KALTON, G., et KISH, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, 13(16), 1919-1939.

KISH, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44, 380-387.

KISH, L. (1952). On the Differentiation of Ecological Units. Thèse de doctorat, University of Michigan.

KISH, L. (1954). Differentiation in metropolitan areas. *American Sociological Review*, 19, 388-398.

KISH, L. (1957). Confidence intervals for clustered samples. *American Sociological Review*, 22, 1954-1965.

KISH, L. (1959). Some statistical problems in research design. *American Sociological Review*, 24, 328-338.

KISH, L. (1961a). A measurement of homogeneity in areal units. *Bulletin of the International Statistical Institute*, 4, 201-209.

KISH, L. (1961b). Efficient allocation of a multi-purpose sample. *Econometrica*, 29, 363-385.

KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

KISH, L. (1965a). *Survey Sampling*. New York: John Wiley & Sons, Inc.

KISH, L. (1965b). Selection techniques for rare traits. *Genetics, Epidemiology, and Chronic Diseases*, Public Health Service Publication, No. 1173.

KISH, L. (1965c). Sampling organizations and groups of unequal sizes. *American Sociological Review*, 20, 564-572.

KISH, L. (1968). Standard errors for indexes from complex samples. *Journal of the American Statistical Association*, 63, 512-529.

KISH, L. (1969). Design and estimation for subclasses, comparisons, and analytical statistics. *New Developments in Survey Sampling*, (Éds. N.L. Johnson et H. Smith). New York: John Wiley & Sons, Inc.

KISH, L. (1975). Representation, randomization and control. *Quantitative Sociology*, (Éds. H.M. Blalock, A. Aganbegian, F.M. Borodkin, R. Boudon et V. Capecchi). New York: Academic Press.

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

KISH, L. (1978). Chance, statistics, and statisticians. *Journal of the American Statistical Association*, 73, 1-6.

KISH, L. (1979a). Samples and censuses. *International Statistical Review*, 47, 99-109.

KISH, L. (1979b). Rotating samples instead of censuses. *Asian and Pacific Census Forum* (East-West Center, Honolulu), 6, 1-13.

KISH, L. (1980). Design and estimation for domains. *The Statistician*, 29, 209-222.

KISH, L. (1981). Using Cumulated Rolling Samples.Washington: Library of Congress.

KISH, L. (1982). Design effects. *Encyclopedia of Statistics*, New York: John Wiley & Sons, Inc.

KISH, L. (1983). Data collection for details over space and time. *Statistical Methods and the Improvement of Data Quality*, (Éd. T. Wright). New York: Academic Press.

KISH, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 28, 1-12.

KISH, L. (1987a). *Statistical Design for Research*. New York: John Wiley & Sons.

KISH, L. (1987b). Discussion. *Small Area Statistics*, (Éd. R. Platek). New York: John Wiley & Sons, Inc.

KISH, L. (1988). Plans de sondage à usages multiples. *Techniques d'enquête*, 14, 19-33.

KISH, L. (1989). *Sampling Methods for Agricultural Surveys*. Rome: FAO.

KISH, L. (1990). Recensusement par étapes et échantillons avec renouvellement complet. *Techniques d'enquête*, 16, 66-77 et 99-100.

KISH, L. (1991). Taxonomy of elusive populations. *Journal of Official Statistics*, 7, 339-347.

KISH, L. (1995a). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.

KISH, L. (1995b). *Questions/answers from the survey statistician 1978-1994*. Libourne: International Association of Survey Statisticians.

KISH, L. (1996). Developing samplers for developing countries. *International Statistical Review*, 64, 143-162.

KISH, L. (1997). Periodic and rolling samples and censuses. *Statistics and Public Policy*, (Éd. B.D. Spencer). New York: Oxford University Press.

KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.

KISH, L. (1999). Le cumul ou la combinaison d'enquêtes démographiques. *Techniques d'enquête*, 25, 147-158.

KISH, L. (2002). Combining multi-population surveys. *Journal of Statistical Planning and Inference*, 102, 109-118.

KISH, L., et ANDERSON, D.W. (1978). Multivariate and multipurpose stratification. *Journal of the American Statistical Association*, 73, 24-34.

KISH, L., et FRANKEL, M. (1970). Balanced repeated replications for standard errors. *Journal of the American Statistical Association*, 65, 1071-1094.

KISH, L., et FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, B*, 36, 1-37.

KISH, L., FRANKEL, M.R. et VAN ECK, M. (1972). *SEPP: Sampling Error Programs Package*. Ann Arbor: Institute for Social Research.

KISH, L., FRANKEL, M.R., VERMA, V. et KACIROTI, N. (1995). Effets du plan de sondage sur les $(P_i - P_j)$ corrélés. *Techniques d'enquête*, 21, 131-139.

KISH, L., GROVES, R.M. et KROTKI, K. (1976). Sampling Errors for Fertility Surveys. Occasional Paper No. 17, World Fertility Survey.

KISH, L., et HESS, I. (1958). On noncoverage of sample dwellings. *Journal of the American Statistical Association*, 53, 509-524.

KISH, L., et HESS, I. (1959a). On variances of ratios and their differences in multi-stage samples. *Journal of the American Statistical Association*, 54, 416-446.

KISH, L., et HESS, I. (1959b). A replacement procedure for reducing the bias of nonresponse. *The American Statistician*, 13, 17-19.

KISH, L., et LANSING, J.B. (1954). Response error in estimating the value of homes. *Journal of the American Statistical Association*, 49, 520-538.

KISH, L., et SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

KISH, L., et VERMA, V. (1986). Complete censuses and samples. *Journal of Official Statistics*, 2, 381-96.

PURCELL, N.J., et KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.

PURCELL, N.J., et KISH, L. (1980). Postcensal estimates for local areas (small domains). *International Statistical Review*, 48, 3-18.

Techniques d'enquête, juin 2002
Vol. 28, n° 1, pp. 33-37
Statistique Canada

Nouveaux paradigmes (modèles) pour l'échantillonnage probabiliste

LESLIE KISH[1]

1. UN NOUVEAU PARADIGME POUR LA STATISTIQUE

À plusieurs endroits, j'aborde de nouvelles notions liées à divers aspects de l'échantillonnage, mais je ne sais trop si je devrais les appeler de nouveaux paradigmes ou de nouveaux modèles ou simplement de nouvelles méthodes. À cause de mon incertitude et de mon manque de confiance, je prie le lecteur de choisir l'expression qui lui sied le mieux. Je préfère que ce terme ne devienne pas un obstacle à notre compréhension mutuelle.

L'échantillonnage est à la fois une branche et un outil de la statistique. En 1810, après la naissance de certaines sciences comme l'astronomie, la chimie et la physique, Quetelet (Porter 1987; Stigler 1986) instituait le domaine de la statistique sous forme de paradigme. « À la fin du XVIIe siècle, les études philosophiques sur la cause et le hasard... ont commencé à se rapprocher... aux XVIIIe et XIXe siècles, on en vient de plus en plus à admettre que, contrairement à des groupes de personnes, des groupes d'événements sont susceptibles d'obéir à des lois » (Kendall 1968). Les régularités prévisibles, significatives et utiles constatées dans le comportement de groupes de populations composés d'individus non prévisibles furent nommées « statistiques » et constituèrent dès lors une grande découverte.

Quetelet et d'autres calculèrent donc des taux de natalité, des taux de décès, des taux de suicide, des taux d'homicide, des taux d'assurance, *etc.*, à l'échelle nationale (et autres) à partir d'événements individuels qui sont imprévisibles. Ces statistiques sont fondamentales à des domaines comme la démographie et la sociologie. Plus tard au XIXe siècle, Frances Galton et Karl Pearson, puis Maxwell, introduisirent les notions de statistiques en biologie et en physique respectivement, puis la théorie et les applications firent un grand pas.

La statistique et les statisticiens s'intéressent aux effets du hasard sur des données empiriques. Les mathématiques des probabilités avaient été mises au point plusieurs siècles auparavant pour les jeux de hasard ainsi que pour les calculs des erreurs d'observation en astronomie. Des données avaient aussi été compilées pour le commerce, les banques et les administrations publiques. Toutefois, un nouveau paradigme, une nouvelle vision théorique s'impose si l'on désire combiner le hasard à des données réelles. Pour cette raison, étant le produit de la maturité du développement humain, la science statistique et ses diverses branches arrivèrent sur le tard dans l'histoire et dans les universités (Kish 1985).

Le concept le plus fondamental de la statistique repose sur des populations d'individus choisis au hasard. C'est le fondement des théories de la distribution, des inférences, de la théorie de l'échantillonnage, du plan expérimental, *etc.* De plus, le paradigme de la statistique s'écarte fondamentalement de la vision déterministe de la cause et de l'effet ainsi que des relations précises observées dans les autres sciences et les mathématiques.

2. LE PARADIGME DE L'ÉCHANTILLONNAGE

Publiée presque cent ans après la naissance de la statistique, il y a maintenant plus d'un siècle, une importante monographie portant le titre *The Representative Method* est généralement acceptée comme le point de départ de l'échantillonnage moderne (Kiaer 1895). L'expression a été utilisée depuis lors dans plusieurs documents ayant fait autorité (Jensen 1926; Neyman 1934; Kruskal et Mosteller 1979a, 1979b, 1979c, 1980). Ces derniers auteurs admettent que le terme « représentatif » a été accolé à un si grand nombre de méthodes spécifiques et avec des sens tellement différents qu'il ne désigne plus aujourd'hui une méthode particulière. Toutefois, dans la forme où Kiaer l'a employé, et comme on l'utilise encore de façon générale, le terme désigne le prélèvement d'un échantillon destiné à représenter une population spécifique dans l'espace, dans le temps et sous d'autres définitions, afin de pouvoir obtenir de l'échantillon des inférences statistiques se rapportant à cette population. Pour cette raison, un échantillon représentatif national doit reposer sur un prélèvement méthodique de l'échantillon parmi tous les éléments de la population nationale, et non pas seulement au sein d'un quelconque domaine arbitraire comme une ville ou une province « typique », ou d'un sous-ensemble, qu'il soit défini ou non.

Méthode scientifiquement acceptée pour l'échantillonnage d'enquête, l'échantillonnage probabiliste assure des probabilités de prélèvement positives connues pour chacun des éléments de la population de base. La base fournit l'équivalent de listes d'unités d'échantillonnage à chaque étape du prélèvement. La base d'échantillonnage de la population complète est nécessaire pour effectuer les

[1] L'impression de cet article a été autorisée avec gentillesse par Rhea Kish, 1050 Wall St. #9A, Ann Arbor, MI 48105, courrier électronique: rheakk@umich.edu.

opérations mécaniques du prélèvement aléatoire. C'est sur quoi seront basées les inférences statistiques entre les statistiques de l'échantillon et les statistiques correspondantes de la population (paramètres) (Hansen, Hurwitz et Madow 1953a, 1953b). Cette insistance sur des inférences reposant sur des prélèvements de populations de base représente un paradigme différent des approches imprécises ou fondées sur un modèle qu'utilisent la plupart des analyses statistiques.

Il s'écoulera un demi-siècle avant que le document de Kiaer ne débouche sur une vaste acceptation de l'échantillonnage d'enquête. Outre la négligence et la résistance passive, on pouvait aussi observer une forte opposition de la part des bureaux statistiques nationaux, qui se méfiaient des méthodes d'échantillonnage destinées à remplacer les chiffres complets des recensements. D'aucuns préféraient même la « méthode de la monographie », qui présentait des chiffres complets provenant d'une province ou d'un district « typique » ou « représentatif » plutôt que d'un échantillon national prélevé de façon aléatoire (O'Muircheartaigh et Wong 1981). Outre le milieu politique, de nombreux opposants provenaient du milieu académique, dont des statisticiens. La publication du rapport de la Commission statistique des Nations Unies, sous la direction de Mahanalobis et Yates (Bureau statistique des Nations Unies 1950) renversa le courant en faveur de l'échantillonnage probabiliste. Cinq ouvrages d'influence parus entre 1949 et 1954 amorcèrent un flot d'articles sur la théorie et sur de vastes applications.

La force, l'ampleur et la durée de la résistance manifestée à l'égard des notions et de l'utilisation de l'échantillonnage probabiliste appliqué à des populations de base signifient qu'il s'agissait d'un nouveau paradigme qui commandait de la part du public et des spécialistes une nouvelle vision.

3. POPULATIONS COMPLEXES

L'échantillonnage d'enquête se distingue notamment par un prélèvement probabiliste strict d'une base de population, qui permet d'obtenir des inférences de l'échantillon pour les appliquer à une population *finie*. Cependant, les distributions complexes des éléments dans les populations représentent des problèmes encore plus importants et difficiles. Ces complexités contrastent vivement avec le modèle d'indépendance simple que présupposent, explicitement ou implicitement, presque toute la théorie statistique et toutes les statistiques mathématiques.

L'hypothèse d'observations indépendantes, ou non corrélées, de variables ou d'éléments est sous-jacente aux statistiques mathématiques et à la théorie de la distribution. Nous n'avons pas besoin ici de faire une distinction entre des variables aléatoires à distribution indépendante et identique (DII) et une « échangeabilité », et des « superpopulations ». La simplicité à la base de chacun de ces modèles est nécessaire à cause des complexités des formulations mathématiques.

Toutes les sciences ont besoin de modèles simples à leur début : par exemple, une orbite parfaitement circulaire pour les planètes, ou $d = gt^2/2$ pour les objets en chute libre dans un milieu dénué de friction. Toutefois, ces modèles ne répondent pas à la complexité du monde physique réel. De même, aucune *population*, fût-elle humaine, animale, végétale, physique, chimique ou biologique, ne comporte des éléments indépendants. Les modèles indépendants simples peuvent servir assez bien aux petits *échantillons*; et on a souvent utilisé comme exemple (précieux parce que rare) la distribution de Poisson des décès par ruades dans l'armée prussienne au cours d'une période de 43 ans (Fisher 1926).

Il y a aussi eu des tentatives de construire des populations théoriques d'éléments à DII; la plus connue est sans doute le « collectif » classique de Von Mises (1931); cependant, elles ne correspondent pas à des populations réelles. Toutefois, on a confectionné, avec beaucoup d'efforts, des tables de nombres aléatoires qui ont subi avec succès tous les tests. On les a largement utilisées dans des expériences modernes et dans des sondages. La *réplication* et la *randomisation* sont deux des notions les plus fondamentales de la statistique moderne qui s'inspirent du concept des populations.

Le simple concept d'une population composée d'éléments indépendants ne décrit pas assez bien les distributions complexes (dans l'espace, dans le temps et dans les classes) des éléments. Le groupement et la stratification sont des noms communs de complexités omniprésentes. De plus, il semble impossible de construire des modèles qui pourraient mieux décrire des populations réelles. Les distributions sont beaucoup trop complexes, outre qu'elles soient aussi différentes pour chacune des variables de l'enquête. On a examiné ces complexités et différences qu'on présente maintenant sous forme de milliers de calculs « d'effets du plan ».

L'échantillonnage d'enquête avait besoin d'un nouveau paradigme pour faire face aux complexités des populations de tous genres comportant un grand nombre de variables d'enquête et une liste croissante de statistiques d'enquête. On a donc conçu des plans de prélèvement et des formules de variance robustes susceptibles de convenir à une multitude de plans de sondage, ce qui a donné naissance à la nouvelle discipline de l'échantillonnage d'enquête. Le calcul des « effets du plan » a fait ressortir l'existence, l'ampleur et la variabilité des effets attribuables aux complexités des distributions, non seulement pour les moyennes mais aussi pour les relations à variables multiples, comme les coefficients de régression. La longue période de désaccord entre les échantillonneurs d'enquête et les économétriciens témoigne du besoin de se doter d'un nouveau paradigme.

4. REGROUPEMENT D'ÉCHANTILLONS DE POPULATION

Les échantillons de populations nationales représentent toujours des sous-populations (domaines) qui affichent des caractéristiques d'enquête parfois légèrement différentes, parfois fortement différentes. On peut distinguer ces sous-classes dans l'échantillon avec plus ou moins de peine. En premier lieu, on peut facilement séparer les échantillons de provinces lorsque leur prélèvement fait l'objet d'une opération distincte. En deuxième lieu, on peut aussi distinguer, et parfois les utiliser comme des estimations poststratifiées, des sous-classes par âge, sexe, profession et niveau de scolarité. En troisième lieu, cependant, les sous-classes définies par leurs caractéristiques sociales, psychologiques et comportementales peuvent parfois se prêter difficilement à une différenciation même si elles présentent le plus de lien avec les variables d'enquête. Par conséquent, nous admettons que les échantillons nationaux ne sont pas de simples agrégations de personnes provenant d'une population à DH, mais plutôt des combinaisons de sous-classes provenant de sous-populations affichant des caractéristiques diversifiées. La composition de populations nationales issues de domaines diversifiés mérite qu'on s'y arrête, et elle sert aussi d'exemple aux deux types de combinaison qui suivent. De plus, ces remarques s'adressent à des combinaisons non seulement d'échantillons nationaux, mais aussi de villes, d'institutions, d'établissements, *etc.*

On observe depuis quelques années deux types de plan de sondage qui exigent des efforts au-delà de ceux que commandent de simples échantillons nationaux : a) les échantillons périodiques, et b) les plans de populations multiples. Ces deux plans sont uniquement de facture récente, car ils devaient faire appel à trois types de ressource nouvelle : 1. une demande réelle appuyée par des ressources financières et politiques; 2. des ressources techniques institutionnelles suffisantes dans les bureaux statistiques nationaux; 3. de nouvelles méthodes. Pour ces deux types de plan, nous devons distinguer les besoins des *méthodes d'enquête* (définitions, variables, mesures), qu'il faut harmoniser et normaliser, et ceux liés aux *plans de sondage*, qui peuvent être librement conçus pour s'adapter à des situations nationales (et même provinciales), à condition de reposer sur un plan probabiliste (Kish 1994). Les deux types de plan ont été conçus d'abord et avant tout à des fins de comparaison, c'est-à-dire des comparaisons périodiques et des comparaisons multinationales respectivement. Toutefois, de nouvelles utilisations sont aussi apparues : des « échantillons à renouvellement complet » et des cumuls multinationaux, respectivement. Chaque type de cumul a fait face à une très forte opposition et il doit donner lieu à une nouvelle vision, à un nouveau paradigme.

On a utilisé à quelques reprises les « échantillons à renouvellement complet » dans des situations locales (Mooney 1956; Kish, Lovejoy et Rackow 1961). On les a ensuite proposés plusieurs fois à la place d'échantillons annuels nationaux et aussi comme remplacement possible de recensements décennaux (Kish 1981, 1990). Mais c'est d'abord et avant tout le Census Bureau des États-Unis (Alexander 1999; Kish 1990) qui les utilise maintenant à la place d'échantillons pour des recensements nationaux. En recommandant cette nouvelle méthode, je dois habituellement répondre à des questions au sujet des moyennes des échantillons périodiques : « Comment pouvez-vous établir la moyenne d'échantillons lorsque ces échantillons varient entre les périodes? » À mon avis, plus la variabilité est prononcée, moins grande devrait être la dépendance à une période unique, que la variation soit monotone, cyclique ou aléatoire. Je constate donc deux visions ou paradigmes contrastants. Très souvent, l'opposition disparaît après deux jours de discussions et de réflexion.

> « Par exemple, le revenu annuel est une agrégation facilement acceptée, non seulement pour les revenus stables, mais aussi pour des professions présentant des variations marquées (saisonnières ou irrégulières). Il se révélera plus facile d'accepter des moyennes d'échantillons hebdomadaires pour des statistiques annuelles que des moyennes d'échantillons cumulés sur dix ans. Néanmoins, beaucoup d'investisseurs dans les fonds communs d'actions préfèrent utiliser la moyenne mobile des gains sur dix ans ou cinq ans (malgré son caractère périmé) plutôt que les gains actualisés de l'année précédente (avec leurs variations « aléatoires » risquées. Dans la planification d'un pique-nique, la plupart des gens préféreraient aussi une température « normale » moyenne sur cinquante ans au chiffre exact de température de l'année précédente. Il y a de nombreux exemples semblables de méthodes perfectionnées d'établissement de la moyenne sur de longues périodes auquel recourt le public « naïf ». Ce public, et les décideurs, s'accommoderaient aussi rapidement des échantillons à renouvellement complet si on leur en donnait la chance ».
> (Kish 1998)

Tout comme pour les échantillons à renouvellement complet, les échantillons combinés de populations multiples ont aussi rencontré une opposition : les frontières nationales représentent différentes étapes de leur évolution, ainsi que des différences au niveau des lois, des langues, des cultures, des coutumes, des religions et des comportements. Comment peut-on alors les combiner? Cependant, nous trouvons souvent des utilisations et des sens à des moyennes continentales; par exemple, les taux de naissance et de décès en Europe, ou les taux en Amérique du Sud, en Afrique sub-saharienne ou en Afrique occidentale. C'est parfois aussi le cas de taux de naissance, de décès et de croissance à l'échelle mondiale. Cependant, n'ayant pas fait

l'objet de discussions, ces taux se combinent habituellement assez mal. Une théorie plus appropriée favoriserait cependant de meilleures combinaisons (Kish 1999). Mais il faut d'abord reconnaître le besoin en appliquant un nouveau paradigme à des combinaisons multinationales, suivi de nouvelles méthodes plus appropriées.

5. ÉCHANTILLONNAGE ESPÉRÉ

L'échantillonnage probabiliste assure, pour chaque élément de la population ($i = 1, 2, ..., N$) une probabilité de prélèvement positive connue ($P_i > 0$). Cette assurance doit se fonder sur une procédure mécanique de prélèvement au hasard plutôt que seulement sur des hypothèses, des croyances ou des modèles au sujet des distributions probabilistes. La procédure de randomisation exige une opération physique pratique qui correspond de très près (ou exactement) au modèle probabiliste (Kish 1965). La plupart des ouvrages sur l'échantillonnage d'enquête présentent une affirmation de ce genre, et j'y crois encore. Toutefois, cette définition et ses exigences souffrent de deux objections douteuses et gênantes.

La plus importante des deux concerne les situations pratiques qui se présentent fréquemment lorsque nous devons faire un choix entre l'échantillonnage probabiliste et l'échantillonnage espéré. Ces situations se produisent souvent lorsque le taux de prélèvement pratique et facile des unités de listage de $1/F$ donne non seulement la probabilité unique $1/F$ pour les éléments, mais aussi une probabilité avec la fraction variable k_i/F pour le $i^{ième}$ élément ($i = 1, 2, ..., N$) et avec $k_i > 0$. Des exemples de $k_i > 1$, habituellement un petit nombre entier, surviennent avec des listes avec répétitions, des bases de prélèvement doubles ou multiples, les résidences secondaires de ménages, des populations mobiles et nomades, ou des exploitants agricoles possédant plusieurs terrains. Des exemples de $k_i < 1$ surviennent lors du prélèvement d'un seul adulte au sein de ménages, ou de logements uniques parmi des immeubles. Dans ces exemples, il arrive souvent que k_i puisse facilement être vérifié, et il est alors plus facile, plus pratique et plus économique de recourir à des coefficients de pondération que d'essayer d'obtenir la valeur $1/F$ pour tous les éléments. Ces problèmes sont décrits dans des ouvrages et des articles.

Dans la plupart des cas, il est plus pratique et moins coûteux d'accepter les probabilités variables et de les contrebalancer par la pondération des valeurs espérées $1/k_i$ ou k_i que d'ajouter une autre étape du prélèvement. Par conséquent, pour paraphraser l'échantillonnage probabiliste : *l'échantillonnage espéré* assure pour chaque élément de la population ($i = 1, 2, ..., N$) un nombre espéré positif connu de prélèvements ($k_i/F > 0$). Dans la pratique, on utilise ces procédures pour des statistiques descriptives (du premier ordre) où les variables k_i ou $1/k_i$ ne sont ni élevées ni fréquentes. Les traitements des statistiques inférentielles – de deuxième ordre ou plus élevé – sont plus difficiles et diversifiés et sont expliqués à part dans la documentation. On remarquera que l'échantillonnage probabiliste est la situation spéciale (et souvent souhaitée) où tous les k_i sont égaux à 1.

L'autre objection à la terminologie d'échantillonnage probabiliste est plus théorique et philosophique puisqu'elle porte sur le mot « connu » dans sa définition. Ce mot semble sous-entendre une croyance. Des auteurs, à partir des classiques comme John Venn et M. G. Kendall jusqu'à des bayésiens modernes comme Dennis Lindley – et d'autres aux deux extrémités – ont clairement attribué la « probabilité » à des états de croyance et le « hasard » à des fréquences produites par des phénomènes objectifs et des opérations mécaniques. Par conséquent, notre insistance sur des opérations, comme des générateurs de nombre aléatoires, devrait sous-entendre l'expression « échantillonnage au hasard ». Toutefois, je n'en ai pas observé l'emploi et, de plus, il pourrait déboucher sur un problème philosophique : l'utilisation appropriée de bonnes tables de nombres aléatoires sous-entend une croyance dans leurs probabilités « espérées ». Je n'ai passé que très peu de temps à ces problèmes et à des discussions agréables avec seulement un très petit nombre de collègues, qui ont cependant été d'accord avec moi. Je serais heureux de participer à d'autres discussions ou de recevoir des suggestions et des corrections.

6. SUR DES SUJETS APPARENTÉS

Nous avons demandé la reconnaissance de nouveaux paradigmes pour six aspects de l'échantillonnage d'enquête, à commencer par la statistique elle-même. Enfin, nous soulignons ici le contraste entre l'échantillonnage et d'autres méthodes apparentées. Les méthodes d'enquête donnent le choix et la définition des variables, des méthodes de mesure ou d'observation et du contrôle de la qualité (Kish 1994; Groves 1989).

On a considéré l'échantillonnage d'enquête comme une méthode qui concurrence les recensements (annuels ou décennaux) et, partant, les registraires de l'état civil (Kish 1990). Dans d'autres contextes, l'échantillonnage d'enquête concurrence ou complète les expériences, les observations contrôlées et les essais cliniques. Ces contrastes doivent aussi déboucher sur des vues plus vastes et plus complètes (Kish 1987, chapitre A.1). Toutefois, ces discussions nous mèneraient bien au-delà de notre propos actuel.

BIBLIOGRAPHIE

ALEXANDER, C.H. (1999). A rolling sample survey for yearly and decennial uses. *Bulletin de l'Institut International de Statistique*, Helsinki, 52ième Session.

FISHER, R.A. (1926). *Statistical Methods for Research Workers.* London: Oliver & Boyd.

GROVES, R.M. (1989). *Survey Errors and Survey Costs.* New York: John John Wiley & Sons, Inc.

HANSEN, M.H., HURWITZ, W.N. et MADOW, W.G. (1953a). *Sample Survey Methods and Theory, I – Methods and Applications*, New York: John Wiley & Sons, Inc.

HANSEN, M.H., HURWITZ, W.N. et MADOW, W.G. (1953b). *Methods and Applications, II – Theory.* New York: John Wiley & Sons, Inc.

JENSEN, A. (1926). The representative method in practice, *Bulletin of the International Statistical Institute*, 22, pt. 1, 359-439.

KENDALL, M.G. (1968). Chance. *Dictionary of the History of Ideas*, (Éd. P.P.Wiener), New York: Chas Scribners.

KIAER, A.W. (1895). The Representative Method of Statistical Surveys, English translation, 1976, Oslo: Statistik Sentralbyro.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

KISH, L. (1981). *Using Cumulated Rolling Samples.* Washington DC: Library of Congress.

KISH, L. (1985). Chance, statistics, sampling. *Journal of Official Statistics*. 1, 35-47.

KISH, L. (1987). *Statistical Design for Research.* New York: John Wiley & Sons, Inc.

KISH, L. (1990). Recensement par étapes et échantillons avec renouvellement complet. *Techniques d'enquête*, 16, 67-86.

KISH, L. (1994). Multipopulation survey designs. *International Statistical Review*, 62, 167-186.

KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.

KISH, L. (1999). Le cumul ou la combinaison d'enquêtes démographiques. *Techniques d'enquête*, 25, 147-158.

KISH, L., LOVEJOY, W. et RACKOW, P. (1961). A multistage probability sample for continuous traffic surveys. *Proceedings of the American Statistical Association, Section on Social Statistics*, 227-230.

KRUSKAL, W.H., et MOSTELLER, F. (1979a). Representative sampling, I: Non-scientific literature. *International Statistical Review*, 47, 13-24.

KRUSKAL, W.H., et MOSTELLER, F. (1979b). Representative sampling, II: Non-scientific literature. *International Statistical Review*, 47, 111-127.

KRUSKAL, W.H., et MOSTELLER, F. (1979c). Representative sampling, III: The current statistical literature. *International Statistical Review*, 47, 245-265.

KRUSKAL, W.H., et MOSTELLER, F. (1980). Representative sampling, IV: The history of the concept in statistics. *International Statistical Review*, 48, 169-195.

MOONEY, H.W. (1956). Methodology of Two California Health Surveys, US Public Health Monograph 70, Washington DC: US Government Printing Office.

NEYMAN, J. (1934). On the 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'MUIRCHEARTAIGH, C., et Wong, S.T. (1981). The impact of sampling theory on survey sampling practice: a review. *Bulletin of the International Statistical Institute*, Buenos, 43ème Session, 1, 465-493.

PORTER, T.M. (1987). The Rise of Statistical Thinking: 1820-1900, Princeton, NJ: Princeton University Press.

STIGLER, S.M. (1986). History of Statistics, Cambridge: Harvard University Press.

UNITED NATIONS STATISTICAL OFFICE (1950). The Preparation of Sample Survey Reports, New York: UN Series C No 1; aussi revision 2 en 1964.

VON MISES, R. (1939). *Probability, Statistics, and Truth*, London: Wm. Hodge and Co.

Techniques d'enquête, juin 2002
Vol. 28, n° 1, pp. 39-46
Statistique Canada

Les échantillons successifs de Leslie Kish et l'American Community Survey

CHARLES H. ALEXANDER[1]

RÉSUMÉ

LeslieKish a longtemps prôné l'adoption de plans de sondage à « échantillons successifs », ou échantillons à renouvellement complet, avec panels mensuels non chevauchants pouvant être cumulés sur des périodes de diverses longueurs pour des domaines de taille variable. Dans ces conditions, une même enquête peut être utilisée pour atteindre plusieurs objectifs. La nouvelle enquête américaine sur les communautés (*American Community Survey (ACS)*) du Census Bureau est fondée sur un plan de sondage à échantillons successifs de ce genre. Les données qu'elle fournit servent au calcul de moyennes annuelles pour mesurer les variations au niveau de l'État et à celui de moyennes mobiles sur trois ans ou sur cinq ans pour décrire des domaines de plus en plus petits. L'article décrit l'influence exercée par Kish sur l'élaboration de l'American Community Survey, ainsi que certains problèmes méthodologiques d'ordre pratique qu'il a fallu résoudre pour mettre en œuvre le plan de sondage.

MOTS CLÉS : Échantillons successifs; moyennes pluriannuelles; cumuls asymétriques.

1. INTRODUCTION

Tel que défini plus bas, un « plan de sondage à échantillons successifs » permet de créer une enquête unique suffisamment souple pour poursuivre plusieurs objectifs. Le concept a été élaboré par Leslie Kish dans une série d'articles (y compris Kish 1979a, 1979b, 1981, 1983, 1986, 1990, 1997, 1998 et Kish et Verma 1983, 1986) dans lesquels il a énoncé les principes du cumul spatial et temporel d'information. Kish a prôné l'adoption de ce genre de plan de sondage pour atteindre des objectifs divers (Kish 1998), particulièrement dans les pays en voie de développement (Kish 1979b), mais aussi dans le contexte du Recensement des États-Unis (Kish, Lovejoy et Rackow 1981). Son utilisation personnelle des échantillons successifs, qu'il qualifiait alors d'« échantillon continu », remonte au moins à 1958 (Kish 1961); un projet antérieur (Mooney 1956) est mentionné dans Kish (1998).

L'American Community Survey (ACS), qui est destinée à remplacer l'enquête à « questionnaire détaillé » réalisée habituellement dans le cadre du recensement, s'appuiera sur une forme de plan de sondage à échantillons successifs. Le présent article décrit comment le concept des échantillons successifs est appliqué dans le cas de l'ACS, en tenant compte des objectifs et des considérations opérationnelles propres à l'enquête. Les décisions concernant le plan de sondage de l'ACS illustrent certaines questions que peut soulever l'utilisation d'échantillons successifs en général. Elles illustrent aussi les divers niveaux — conceptuel, personnel et pratique — sur lesquels s'est exercée l'influence de Leslie Kish dans le domaine du développement des enquêtes.

2. ÉCHANTILLONS SUCCESSIFS

Un plan de sondage à « échantillons successifs » consiste à sélectionner conjointement k échantillons probabilistes (panels) non chevauchants qui représentent chacun $1/F$ de la population complète. À chaque période, les membres d'un panel particulier sont interviewés. Après k périodes, tous les membres de l'échantillon auront été interviewés. Selon la précision exigée, un seul panel de $1/F$ peut suffire à la production de bonnes estimations pour la population dans son ensemble et, éventuellement, pour certains grands domaines. Pour les domaines plus petits, ou pour obtenir une plus grande précision pour les grands domaines, on peut cumuler un nombre variable de panels consécutifs, jusqu'à concurrence de k/F de la population. Un plan de sondage à échantillons successifs pour lequel $k = F$ porte le nom de « recensement continu ». Pour un échantillonnage successif mensuel, il est naturel que F soit égal à un multiple de 12 et que les cumuls soient trimestriels, semi-annuels, annuels ou pluriannuels.

Les « domaines » s'entendent des zones géographiques ainsi que des sous-groupes démographiques. Kish (1987, section 2.3) présente un cadre de référence pour le compromis entre les niveaux de détail géographique et démographique, pour un niveau donné de précision. Un élément encore plus fondamental du concept des échantillons successifs est celui du « cumul asymétrique » des données, sur des périodes de longueur variable pour des tailles de domaine diverses (Kish 1990, 1998). Ce dernier fut élargi plus tard pour donner une vue des similarités fondamentales du calcul de la moyenne spatiale et de la moyenne temporelle (Kish 1998), ainsi que du calcul de la moyenne sur divers domaines démographiques. La souplesse du plan de sondage à échantillons successifs tient

[1] Charles H. Alexander, U.S. Bureau of the Census, Suitland, Maryland, (U.S.A.) 20233.

au fait qu'il permet de réaliser divers compromis entre les niveaux de détail spatial, temporel et démographique.

Leslie Kish a laissé à ses collègues la tâche de généraliser ses idées en une « théorie de la combinaison de populations » (Kish 1999, 2001). Lors de la séance de communications offertes sur la « combinaison d'enquêtes » qu'il a organisée à l'assemblée annuelle de 1999 de l'Institut international de statistique, Leslie a expliqué aux conférenciers que leurs travaux portaient tous sur des aspects différents d'un même problème, qu'ils en soient conscients ou non. La portée de la tâche s'étend aux diverses formes de cumul de données recueillies au moyen d'échantillons successifs, ainsi qu'aux moyens de combiner les données provenant de divers pays afin de produire des statistiques pour de plus grandes entités, comme l'Union européenne. Kish (2001) soutient que ces problèmes sont fondamentalement les mêmes que celui de la combinaison d'information provenant d'expériences différentes (Cochran 1937, 1954).

3. LE QUESTIONNAIRE DÉTAILLÉ DU RECENSEMENT ET LES OPTIONS INTERCENSITAIRES

Le « questionnaire détaillé » du recensement décennal est la source principale de données infranationales sur les *caractéristiques* de la population et du logement aux États-Unis. Les estimations du *nombre* de personnes et d'unités de logement sont calculées d'après le « questionnaire abrégé » du recensement auquel répondent tous les ménages. Le questionnaire détaillé, pour lequel le taux global d'échantillonnage est de un sur six, fournit des estimations précises, détaillées (Le terme « précises » a trait aux erreurs d'échantillonnage et le terme « détaillés » signifie que les estimations sont données pour un grand nombre de domaines démographiques dans un domaine géographique.), de diverses caractéristiques démographiques et économiques pour les divers États, les grandes villes, ainsi que les grands comtés ou groupes de comtés. Il permet de produire des estimations utiles, quoique moins précises et moins détaillées, pour des régions encore plus petites, comme les petites villes et les réserves indiennes, ainsi que pour les secteurs de recensement, qui comptent, en moyenne, 4 000 habitants. Pour les unités gouvernementales les plus petites, on utilise des taux d'échantillonnage plus élevés, pouvant atteindre un sur deux pour les zones les plus petites, de sorte que l'on puisse produire des estimations utilisables pour ces domaines. Afin de compenser le taux d'échantillonnage plus élevé utilisé pour ces régions, celui appliqué aux secteurs de recensement les plus grands est de un sur huit.

Entre les recensements, les programmes statistiques du gouvernement fédéral fournissent assez peu de renseignements sur les caractéristiques de la population à un niveau infranational. Les dénombrements de base du recensement sont mis à jour grâce à un programme intercensitaire d'estimations démographiques, mais les données sur d'autres caractéristiques démographiques et économiques proviennent principalement d'enquêtes nationales. Dans le cas de la *Current Population Survey* (CPS), qui est l'enquête américaine mensuelle sur la population active, le taux d'échantillonnage est d'environ 1 sur 1 000 et le chevauchement entre les unités d'échantillonnage est considérable d'un mois à l'autre, de sorte que l'échantillon ne peut être cumulé profitablement au cours du temps comme cela est possible dans le cas d'échantillons successifs avec renouvellement complet. Un supplément à la CPS est réalisé une fois par an en mars afin de recueillir des renseignements supplémentaires pour produire des estimations du revenu et de la pauvreté au niveau de l'État, mais la précision et le niveau de détail démographique de ces estimations sont limités. Certains programmes s'appuient sur des méthodes de modélisation fondées sur des dossiers administratifs pour produire des estimations du chômage, ainsi que du revenu et de la pauvreté pour les petites régions, mais ne permettent pas de produire des estimations pour diverses autres caractéristiques.

La nécessité de recueillir plus fréquemment des renseignements pour les petits domaines (ou « communautés ») est admise depuis longtemps (Hauser 1942; Eckler 1972, page 212; Bounpane 1986). Leslie a reconnu à son ami, Philip Hauser, l'honneur d'avoir proposé un « recensement annuel par sondage » en 1941. Kish (1981) a proposé un plan de sondage à échantillons successifs comme moyen d'atteindre cet objectif et a présenté plusieurs options, y compris un plan de sondage à échantillons successifs pour la CPS. Au lieu de cette proposition, un recensement de milieu de décennie a été approuvé pour 1985, mais n'a jamais été financé. Une proposition visant à doubler la taille de l'échantillon de la CPS ne l'a pas été non plus (Tupek, Waite et Cahoon 1990).

L'intérêt du Census Bureau pour des données intercensitaires sur les caractéristiques de la population s'est ravivé à la suite d'une proposition de « programme décennal de recensement » présentée par Herriot, Bateman et McArthy (1989). Ce programme aurait permis de recueillir des données pour différents États pour des années différentes; en dernière analyse, la proposition n'a pas été acceptée. Cependant, le plaidoyer dynamique et éloquent de Roger Herriot quant à l'importance et à la valeur éventuelles des données intercensitaires de niveau infranational a sensibilisé les organismes statistiques fédéraux à la possibilité d'adopter un « nouveau paradigme » pour le cycle décennal de collecte des données. L'échantillonnage successif avec renouvellement complet proposé par Kish a été un facteur définitivement déterminant durant cette période où le Bureau a envisagé de nouvelles approches pour le Recensement de 2000 (voir Bounpane 1986).

Le Congrès ayant manifesté un regain d'intérêt pour des données intercensitaires sur les caractéristiques de la population (Melnick 1991; Sawyer 1993), l'option d'une

« mesure continue » pour remplacer le questionnaire détaillé du recensement a été envisagée durant les travaux de recherche pour le Recensement de 2000 qui ont débuté en 1992. Le plan de sondage à échantillons successifs de Kish a finalement été proposé parce qu'il offrait non seulement une certaine souplesse pour la production d'estimations, mais aussi la possibilité de recueillir efficacement des données (Alexander 1993, 1997; National Academy of Sciences 1994, 1995). De mémoire, les articles les plus influents ont été ceux de Kish (1981, 1990); ceux de Kish et Verma (1983, 1986) ont également été consultés. La « mesure continue » a été renommée plus tard « *American Community Survey (ACS)* ».

L'ACS proposée n'a pas été adoptée pour le Recensement de 2000, mais, après des essais limités réalisés de 1996 à 1998, la méthodologie de l'ACS a été mise en œuvre dans 36 comtés pour la période de 1999 à 2001, de sorte que l'on puisse comparer les résultats de cette enquête aux données du questionnaire détaillé du Recensement de 2000. Un essai à grande échelle a également été réalisé en 2000, au moyen d'un échantillon annuel représentatif des États comptant environ 700 000 adresses. Cet essai, nommé *Census 2000 Supplementary Survey*, visait à recueillir les donnés du questionnaire détaillé indépendamment du recensement, à l'aide du questionnaire de l'ACS. La réalisation de l'enquête supplémentaire se poursuit en 2001 et en 2002 dans le cadre de la transition à l'ACS.

4. PLAN DE L'AMERICAN COMMUNITY SURVEY

L'ACS, qui débutera en 2003, est financée par le Congrès et porte sur un échantillon mensuel d'environ 250 000 adresses renouvelé entièrement au début de chaque mois, ce qui donne un plan à échantillons mensuels successifs pour lequel le taux moyen d'échantillonnage est d'environ $F = 480$ ou un échantillon annuel pour lequel $F = 40$. On utilisera pour l'enquête une valeur de $k = 60$, et la période de cumul d'échantillon la plus courte pour laquelle des données seront publiées sera l'année civile. L'ACS sera réalisée par la poste, avec suivi des non-répondants par téléphone. En outre, un échantillon aléatoire d'un tiers des non-répondants persistants sera sélectionné aux fins d'un suivi sur place.

Pour les domaines ayant un taux de réponse moyen, avec une valeur mensuelle $F = 480$, l'erreur-type pour une moyenne estimée sur cinq ans d'après les données de l'ACS sera un peu plus importante que pour l'estimation correspondante calculée d'après les données du questionnaire détaillé du recensement — habituellement environ 1,33 fois plus importante. Ce résultat a été considéré « suffisamment proche » pour la plupart des applications, étant donné l'avantage quant à l'actualité des données et à la diminution prévue du taux de données manquantes, grâce à la création d'un groupe permanent d'intervieweurs. Dans les régions où le taux de réponse par la poste est inférieur à la moyenne, le sous-échantillonnage aux fins du suivi des non-répondants réduira la taille effective de l'échantillon. Cette situation est due non seulement à la réduction du nombre d'interviews, mais aussi au fait que des poids de sondage inégaux entraînent habituellement un effet de plan de sondage plus important (Kish 1965, pages 429-431). Pour tenir compte de ceci, on réalisera vraisemblablement l'ACS en appliquant un taux plus élevé de sous-échantillonnage pour la non-réponse dans les régions à faible taux de réponse, équilibré par un taux plus faible d'échantillonnage dans les régions où le taux de réponse par la poste est supérieur à la moyenne. Les détails de cette procédure restent à déterminer. On procédera aussi à un suréchantillonnage des adresses dans le cas des petites unités gouvernementales, comme cela était le cas pour l'échantillon recevant le questionnaire détaillé du recensement.

L'un des progrès importants de la dernière décennie, qui a rendu possible la réalisation de l'ACS (Kish (1981) propose comme méthode de rechange celle des « listes successives cumulatives », mais cette option rendrait onéreux le calcul d'estimations régulières pour tous les domaines les plus petits, comme les secteurs de recensement.), est la mise en place du programme créé par le Census Bureau pour tenir à jour en permanence un fichier maître d'adresses (*Master Address File/MAF*) couplé à notre base de données géographiques TIGER. La source principale de mise à jour des adresses au cours de la décennie est le *Delivery Sequence File (DSF)* tenu à jour par le Service des postes. Le Census Bureau est en train de mettre en œuvre un programme de modernisation du MAF/TIGER qui améliorera les mises à jour d'après le DSF grâce à des fichiers de données transmis par les gouvernements locaux et d'autres sources administratives et grâce au ciblage de nouvelles adresses, dans des régions plus rurales, par des intervieweurs préposés à l'ACS et à d'autres enquêtes. En fait, pour produire les échantillons mensuels, on sélectionne un échantillon annuel d'après le MAF du mois de septembre antérieur et on le divise en 12 panels mensuels. En février, on sélectionne d'après le DSF un échantillon supplémentaire de nouvelles unités que l'on répartit entre les mois restants de l'année.

Le remplacement du questionnaire détaillé du Recensement de 2010 par l'ACS est l'une des composantes du programme de remaniement du Recensement de 2010. Ce programme comprend aussi la modernisation du MAF/TIGER, ainsi qu'un programme de recherche et d'essai précoces en vue d'automatiser, de simplifier et d'améliorer les opérations du Recensement de 2010. En principe, le coût budgété de cette combinaison d'améliorations pour le cycle complet de 10 ans sera inférieur à celui de la répétition des méthodes du Recensement de 2000 en 2010. Ce plan diffère considérablement de la vision de l'ACS décrite dans National Academy of Sciences (1994, chapitre 6; 1995, chapitre 6), où je formulais l'espoir qu'en

soi, l'élimination du questionnaire détaillé, sans autres améliorations fondamentales, permette une économie suffisante pour financer l'ACS.

5. CERTAINES VARIANTES DU PLAN DE SONDAGE DE BASE ET CERTAINS PROBLÈMES

5.1 Échantillons en grappes à plusieurs degrés

L'ACS est réalisée auprès d'un échantillon systématique à un degré non mis en grappe, parce que les objectifs comprennent la production annuelle de données pour tous les petits domaines géographiques, comme les secteurs de recensement ou les groupes d'îlots. Les discussions publiées par Kish (1981, 1998) montrent clairement que la méthode des échantillons successifs s'applique aussi à l'échantillonnage par grappes et à l'échantillonnage à plusieurs degrés, ainsi qu'à diverses probabilités de sélection. Cependant, pour que l'on puisse parler d'« échantillons successifs », il faut que les unités primaires d'échantillonnage forment elles-mêmes des échantillons successifs. Un plan de sondage comportant un ensemble fixe d'unités primaires d'échantillonnage (UPE), avec échantillonnage successif dans chaque UPE, produit un « échantillon cumulatif représentatif » (Kish 1998).

Leslie insistait sur le fait que la proposition d'Herriot et coll. (1989) ne correspondait pas à ce qu'il entendait par « échantillons successifs ». Pourtant, la proposition semble correspondre à la définition si l'on considère les États comme des UPE. Selon moi, ce fait prouve qu'il est implicitement nécessaire qu'un plan de sondage à échantillons successifs produise un échantillon probabiliste représentatif utile lors de chaque période de référence, pour chaque domaine étudié, condition qui n'est pas vérifiée si les UPE sont des États. Autrement dit, les grappes, ou UPE, doivent être beaucoup plus petites que le plus petit domaine étudié. (Voir Kish 1998, page 38.)

5.2 Taux d'échantillonnage variable

Selon Kish (1998, section 4), un plan de sondage à échantillons successifs peut prévoir l'application d'un taux d'échantillonnage variable selon la strate. Ce genre de plan peut être compliqué, particulièrement si les taux d'échantillonnage varient avec le temps, car la probabilité conditionnelle de sélectionner une unité (sans remise) pour le $j^{ième}$ panel dans la $h^{ième}$ strate dépend des taux d'échantillonnage utilisés pour les panels antérieurs dans cette strate. La situation est encore plus compliquée si la strate varie au cours du temps, par exemple si les limites des unités gouvernementales changent.

Par souci de simplification, dans le cas de l'ACS, nous tirons l'échantillon en deux étapes. À la première, nous sélectionnons des « superéchantillons successifs » en appliquant pour chaque panel et chaque année un même taux d'échantillonnage, égal au taux le plus élevé nécessaire quelle que soit la strate. À la deuxième étape, l'échantillon initial est sous-échantillonné pour obtenir le taux d'échantillonnage souhaité pour chaque strate pour l'année de référence. La sélection de sous-échantillons, qui permet d'éviter le chevauchement avec la totalité des superéchantillons antérieurs, oblige uniquement à suivre le taux d'échantillonnage pour la première étape.

5.3 Mises à jour de la base de sondage

En pratique, la population varie légèrement pour chaque panel. De nouvelles adresses sont ajoutées à la base de sondage. Certaines anciennes adresses n'existent plus et peuvent être supprimées de la liste, ou y être maintenues et supprimées uniquement après que l'on ait essayé de communiquer avec les personnes vivant à ces adresses. Ceci ne présente aucun problème conceptuel fondamental et signifie qu'un « recensement continu » ne consiste pas nécessairement à communiquer avec toute unité de population ayant déjà existé, puisque certaines unités peuvent disparaître trop rapidement après leur création pour tomber dans l'échantillon.

Pour éviter l'enregistrement des divers taux d'échantillonnage conditionnels appliqués pour différentes « cohortes » d'adresses ajoutées au moment des mises à jour du fichier maître d'adresses exécutées à diverses périodes, nous avons jugé pratique d'attribuer des « échantillons rétrospectifs » artificiels en sélectionnant des adresses à partir de chaque ensemble de nouvelles adresses, non seulement pour le panel courant, mais aussi pour les anciens panels. Ces unités ne sont pas interviewées, puisque la période d'existence du panel auquel elles sont assignées est révolue, mais elles peuvent ainsi être évitées durant la sélection sans remise des futurs panels.

5.4 Que se passe-t-il après le panel k?

Une question qui, autant que je le sache, n'a pas été abordée explicitement par Leslie est la façon de tirer l'échantillon pour le panel $k + 1$. Je crois qu'il supposait que le panel $k + 1$ serait le même que le panel 1, que le panel $k + 2$ serait une répétition du panel 2, et ainsi de suite. Cette démarche donne de bons résultats pour un échantillon aléatoire simple, mais ne fonctionne pas aussi bien pour un échantillon systématique destiné à étaler l'échantillon sur une liste triée géographiquement, car, comme la base de sondage évolue avec le temps, le panel 1 ne garde pas son espacement uniforme.

Nous prévoyons sélectionner le panel $k + 1$ et les panels suivants sous forme d'échantillons systématiques frais. Chacun sera sélectionné de sorte qu'il ne chevauche aucun des $k - 1$ panels antérieurs, si bien que nous aurons systématiquement k panels consécutifs non chevauchants, sans devoir nous inquiéter du chevauchement avec les panels sélectionnés avant cela.

5.5 Date de référence des questionnaires, étant donné une période étendue d'interviews

Pour chaque panel mensuel de l'ACS, les interviews ont lieu sur une période de trois mois, deux mois étant accordés pour le retour du questionnaire dûment rempli par la poste et le suivi téléphonique, avant le lancement des visites sur place plus coûteuses le troisième mois. Donc, les données recueillies pour juin, par exemple, englobent les réponses rapides par la poste des membres du panel de juin, les réponses tardives par la poste et les interviews téléphoniques pour le panel de mai, ainsi que les cas de suivi sur place pour le panel d'avril. Cette situation oblige à se demander s'il faut que les répondants fournissent les réponses qui correspondent à la date d'envoi par la poste, c'est-à-dire la meilleure option en ce qui concerne le biais d'échantillonnage, ou celles qui correspondent au moment où les questions sont posées, c'est-à-dire la meilleure option en ce qui concerne l'erreur de réponse et d'autres erreurs non dues à l'échantillonnage, particulièrement pour les personnes qui ont changé d'adresse.

Compte tenu de ces compromis concernant la qualité des données, nous avons choisi d'utiliser comme date de référence la « date courante » et de recueillir les données sur les caractéristiques des membres du ménage au moment de l'interview. Nous avons pris cette décision notamment parce que nous estimons que les erreurs non dues à l'échantillonnage seront plus difficiles à évaluer que le biais d'échantillonnage. En outre, les biais d'échantillonnage qui entachent les estimations mensuelles auront tendance à se compenser au cours de l'année. Il s'agit de l'une des raisons de limiter l'ACS à la production d'estimations annuelles et pluriannuelles.

5.6 Utilisation d'estimations intercensitaires de population pour le contrôle des poids de sondage

Le Census Bureau a mis en place un programme d'estimations démographiques « intercensitaires » (Leslie qualifierait ces estimations de « postcensitaires », réservant le terme « intercensitaires » pour les estimations réalisées entre deux recensements achevés.) fondé sur des modèles démographiques. Ces modèles permettent de mettre à jour les données du recensement antérieur, d'après les données des dossiers de l'état civil et d'autres dossiers administratifs. Ces estimations servent de contrôles indépendants de la pondération, ou de « facteurs de stratification a posteriori », pour la plupart des enquêtes-ménages nationales (voir Kish 1965, pages 90-92). L'ajustement des poids de sondage afin qu'ils concordent avec les valeurs de contrôle permet de réduire la variance des estimations d'enquête, de rajuster les données pour tenir compte des différences de couverture selon l'âge, le sexe, la race ou l'origine hispanique, et d'augmenter la cohérence d'une enquête à l'autre. De façon analogue, dans le cas du questionnaire détaillé du recensement, on utilise les dénombrements au recensement comme contrôles de la pondération.

Habituellement, aucun contrôle de la pondération n'existait pour les domaines géographiques les plus petits, du moins pas au niveau de détail démographique existant pour les régions plus grandes. La production prévue de totaux de contrôle plus détaillés pour la pondération des données de l'ACS est décrite dans Alexander et Wetrogan (2000). Certaines améliorations résulteront de l'utilisation de meilleures sources de données administratives, mais l'ACS proprement dite fournira des données sur les variations de population qui pourront être intégrées dans les modèles démographiques. La différence entre la « règle du résident courant » appliquée pour l'ACS et la « règle du résident usuel » appliquée pour le recensement complique le problème; l'ACS comprend une question au sujet des résidents ayant occupé le logement une partie de l'année afin de permettre la correction pour cette différence de concept. Afin de faciliter cette intégration des données d'enquête dans les modèles démographiques, particulièrement en vue d'élaborer des mesures de l'erreur pour les estimations résultantes, le Census Bureau essaie de mettre au point des versions « statistiques » des modèles démographiques utilisés pour produire les estimations intercensitaires de population. Les efforts en vue d'intégrer les approches statistique et démographique s'inspirent de Purcell et Kish (1979).

6. CUMULS DIFFÉRENTS SELON L'OBJECTIF

Afin d'atteindre l'objectif principal de l'ACS, c'est-à-dire remplacer le questionnaire détaillé du recensement en tant que source de statistiques descriptives détaillées, nous prévoyons cumuler les données de l'ACS sur cinq ans pour obtenir un produit de données comparable au « fichier sommaire » traditionnellement fondé sur le questionnaire détaillé. Cette période est la plus courte pour laquelle l'erreur d'échantillonnage de l'ACS est jugée raisonnablement proche de celle relative au questionnaire détaillé du recensement. Le fichier de données sur cinq ans engloberait toutes les tailles et toutes les catégories de zones géographiques. En ce qui concerne l'affectation des fonds gouvernementaux basée sur l'évaluation des besoins courants, les études en simulation donnent à penser que le cumul des données sur trois ans serait préférable au cumul sur cinq ans, ce qui revient à sacrifier la précision pour une plus grande actualité (Alexander 1998).

Pour les régions individuelles, les données les plus importantes publiées seront les moyennes sur un an pour les régions comptant plus de 65 000 habitants et les moyennes sur trois ans pour celles comptant plus de 20 000 habitants, en plus des moyennes sur cinq ans produites pour toutes les régions. Les estimations moyennes annuelles pour les régions dont la population est inférieure à ces seuils seront offertes pour des utilisations plus « complexes » dans des modèles de séries chronologiques et pour indiquer les variations importantes des moyennes pluriannuelles, mais

nous ne leur accorderons pas autant d'importance dans nos publications ou nos sites Web.

Ces produits de données de l'ACS dont nous prévoyons la publication visent à inciter les analystes à utiliser la même période de cumul lorsqu'ils comparent des régions de tailles différentes, car ne pas le faire serait perçu comme injuste pour les régions plus petites. Ce faisant, nous avons accepté la notion de « cumul asymétrique » en ce qui concerne les niveaux de détail géographique, mais pas nécessairement à l'intérieur d'un niveau de détail géographique particulier. Par exemple, nous pourrions choisir une période d'une année pour comparer les États, mais nous recommanderions une période de cinq années pour tous les comtés repris dans un tableau comparant les grands et les petits comtés. Cette dernière recommandation nous écarte quelque peu de la démarche de Kish (1998, pages 42-43) qui permettrait d'utiliser des tableaux contenant des estimations sur un an pour les grands comtés, des moyennes sur trois ans pour ceux de taille moyenne, et des moyennes sur cinq ans pour ceux de petite taille. Il sera intéressant de voir quelle pratique les utilisateurs des données adopteront à cet égard.

7. PONDÉRATION DES ANNÉES LORS DU CUMUL PLURIANNUEL

Kish (1998) fait remarquer qu'il existe plusieurs options pour la pondération des cumuls pluriannuels. S'il existe 10 moyennes annuelles $\bar{y}_i$, alors il existe de nombreux choix de $\bar{y} = \sum w_i \bar{y}_i$, où $\sum w_i = 1$, pour les cumuls sur 10 ans.

Pour les cumuls de données de l'ACS sur cinq ans et d'autres cumuls pluriannuels, discutés à la section 6, nous prévoyons accorder un poids égal à toutes les années de référence dans les produits de données publiés types, c'est-à-dire $w_i = 0{,}2$ pour la moyenne sur cinq ans. Il s'agit là d'un sujet de désaccord avec Kish (1998) qui nous priait de considérer d'autres solutions, particulièrement des poids de la forme $w_{i+1} = Cw_i$, avec $C > 1$.

Lorsque l'on pense à des poids inégaux, l'une des questions sous-jacentes est de savoir quel problème statistique nous essayons de résoudre. Si nous prenons le cumul des données de 2003 à 2007 comme exemple, l'objectif est-il de fournir :

- une estimation « directe fondée sur le plan de sondage » de la moyenne historique pour la période de 2003 à 2007;
- une estimation « basés sur un modèle » pour la valeur de 2007;
- une estimation « directe, fondée sur le plan de sondage » de la moyenne historique pondérée pour la période allant de 2003 à 2007, en accordant plus de poids aux années récentes?

Interpréter l'estimation pour la période de 2003 à 2007 comme étant une estimation pour 2007 nécessite un modèle ou des hypothèses au sujet de la série chronologique pour la région en question. Le problème peut être considéré comme étant la combinaison d'une estimation directe pour 2007 avec une prévision pour 2007 fondée sur les années 2003, ..., 2006 à condition que la même formule soit utilisée pour toutes les régions et toutes les caractéristiques afin de maintenir l'additivité dans les tableaux et la comparabilité entre tableaux.

Précédemment, j'ai considéré la décision comme étant un choix entre les deux premiers objectifs et j'ai reculé à l'idée d'adopter la deuxième approche pour l'ACS, en dernière analyse à cause des réserves exprimées par Hansen, Madow et Tepping (1983, sections 3 et 5.5) concernant l'utilisation d'estimations fondées sur un modèle pour la production de « statistiques officielles » d'usage général. Étant donné la variété de statistiques et de zones géographiques couvertes par l'ACS, il existera inévitablement des cas où le modèle de compromis donne de fort mauvais résultats; l'utilisateur des données pourrait ne pas se rendre compte de cet échec ou en être pleinement conscient. Dans quelle mesure la moyenne de compromis peut-elle être considérée comme une estimation valide pour 2007 lorsque le modèle de compromis donne manifestement de mauvais résultats et quelle mesure de l'erreur convient-il de lui associer? Vu sous cet angle, nous avons recommandé d'utiliser les moyennes pluriannuelles non pondérées comme produits de données types polyvalents et de réserver la série chronologique d'estimations annuelles à l'utilisation dans des modèles chronologiques pour des applications spéciales et à l'interprétation des moyennes pluriannuelles lorsque l'on observe des variations durant la période de cinq ans.

Cependant, après avoir relu l'article de Kish (1998), j'interprète maintenant sa vision de la moyenne pondérée comme correspondant à la troisième formulation, c'est-à-dire un estimateur fondé sur le plan de sondage d'un paramètre de population plus à jour. Cette solution permet d'éviter les questions quant à la validité du modèle aux applications générales. Une question demeure : comment justifier et réaliser une solution consensuelle?. En outre, l'application de poids inégaux a tendance à augmenter les erreurs-types des moyennes pluriannuelles. Toutefois, Kish (1998, page 40) aura le dernier mot à ce sujet :

« D'importantes questions restent à débattre et à étudier. Peut-être à jamais, le sujet étant devenu une « branche d'activité en croissance. » [Traduction]

8. NE PAS COMBINER LA CPS ET L'ACS

Leslie a souvent déclaré qu'il était heureux de voir son idée mise en œuvre dans l'ACS, mais je pense qu'il était déçu que nous n'ayons pas essayé de remplacer le

questionnaire détaillé du recensement et la CPS par une seule enquête. Contrairement à d'autres questions sur lesquelles nous avions des discussions animées, Leslie avait adopté sur ce point une attitude de « non-intervention ». À mon avis, il considérait cette décision comme une question de compromis concernant la qualité que les organismes gouvernementaux devaient résoudre eux-mêmes. Les deux raisons principales de notre décision sont les suivantes.

Nous ne pouvons évaluer adéquatement le taux mensuel de chômage au moyen d'une enquête postale. La mesure correcte du taux de chômage nécessite des questions complexes que l'on ne pourrait poser au moyen d'un instrument envoyé par la poste, comme celles visant à s'assurer qu'une personne qui « recherche du travail » a activement entrepris des démarches pour trouver un emploi. (Voir Butani, Alexander et Esposito 1999). L'enquête supplémentaire du Recensement de 2000, qui empruntait la méthodologie de l'ACS, a produit une sur-estimation importante du taux national de chômage en 2000 (5,3 % contre 4,0 % pour la CPS). Un écart semblable a été observé pour le Recensement de 1990.

Comparativement à la CPS, les taux mensuels calculés d'après les données des enquêtes postales seraient produits avec un retard important. En outre, le fait qu'il soit impossible de réaliser toutes les interviews par la poste pour un panel donné dans le mois désigné biaise les estimations mensuelles (voir la section 5.5 plus haut). La production de moyennes mobiles trimestrielles au lieu d'estimations mensuelles, solution fréquemment suggérée par Leslie (par exemple, Kish 1999), réduirait ces problèmes, mais le rapport mensuel sur le chômage est un indicateur économique indispensable aux États-Unis.

Le remplacement du questionnaire détaillé du recensement par un autre instrument qu'un questionnaire envoyé par la poste est trop coûteux. Une enquête à échantillons successifs réalisée sur place auprès d'un échantillon suffisamment grand pour remplacer le questionnaire détaillé du recensement devrait être trois à quatre fois plus importante que la CPS. La grandeur de l'échantillon dépend de la taille de la population américaine et du nombre de secteurs de dénombrement pour lesquels des estimations doivent être produites d'après le questionnaire détaillé. Le coût par cas d'une telle enquête serait beaucoup plus élevé que pour la CPS, parce qu'il ne serait pas possible d'utiliser un échantillon en grappes ni de recourir à l'interview téléphonique pour les interviews répétées de mêmes ménages, comme cela est le cas pour la CPS. Le coût total d'une enquête de ce genre serait plusieurs fois plus élevé que le coût cumulatif de l'ACS proposée et de la CPS.

Parce qu'elle est conçue si rigoureusement pour remplacer un questionnaire détaillé, l'ACS n'illustre pas pleinement la souplesse envisagée par Leslie pour une enquête à échantillons successifs. Dans d'autres circonstances, pour une population plus petite et un besoin moins grand de l'« enquête à questionnaire détaillé » pour produire des estimations pour des domaines très petits, ou dans le cas d'exigences moins rigides quant aux délais de production et aux questions à poser, il sera peut-être possible qu'une enquête sur la population active à échantillons successifs permette de répondre à la demande de données régionales. Grâce à l'ajout d'un panel double ou d'autres composantes (Kish 1998, pages 40 et 41), on pourrait atteindre une gamme encore plus grande d'objectifs.

9. CONTRIBUTION : CONCEPTUELLE, PERSONNELLE ET PRATIQUE

La longue liste d'articles publiés par Leslie Kish au sujet des échantillons successifs démontre clairement l'intensité et la ténacité de sa campagne en faveur de ce qu'il considérait comme une idée importante. L'évolution de cette idée au fil de ses articles illustre aussi la grande attention qu'il accordait aux questions « conceptuelles » concernant les objectifs fondamentaux de qualité d'une enquête. Qu'essayons-nous de faire? Quel est le rapport entre le choix du plan d'enquête et ce que nous essayons de réaliser, et pourquoi? Cette forme d'orientation, qui est essentielle lors du lancement d'un programme d'enquête quand il faut résoudre les « grandes questions », permet de faire la distinction entre les idées qui sont abandonnées rapidement et celles qui font leur chemin.

L'appui personnel que Leslie offrait aux autres statisticiens dépassait de loin le cadre de ses articles. Alors que je n'étais nullement l'un de ses collègues les plus proches, il me prodiguait régulièrement des conseils personnels ou des encouragements lorsqu'il le jugeait nécessaire. L'expression « *still rolling* » du titre de la version anglaise de cet article était celui que je donnais aux messages électroniques que je lui envoyais pour le tenir au courant de la situation périlleuse de l'ACS aux diverses étapes du cycle budgétaire annuel, c'est-à-dire la plupart du temps. Il répondait brièvement par courriel, mais transmettait toujours les messages importants sous forme de lettres manuscrites.

Enfin, si l'on s'en tient à ses articles, il est évident que Leslie a toujours été une personne pratique, même dans ses moments de réflexion les plus théoriques, et que ses articles ne peuvent être appréciés comme il se doit si l'on ignore ce qui se passait dans le monde des enquêtes à l'époque où il les a rédigés. Lorsque je réfléchis à ses articles sur les échantillons successifs, je décèle de nombreux commentaires, au sujet de détails aussi bien que de principes généraux, qui visaient à éclairer certaines décisions que le Census Bureau devait prendre à l'époque. J'imagine que l'ensemble de ses travaux contient divers messages destinés spécifiquement à aider quelqu'un, quelque part dans le monde, qui devait prendre à l'époque une décision pratique concernant un plan de sondage.

REMERCIEMENTS

Le présent article présente les résultats de travaux de recherche et d'analyse entrepris par les employés du Census Bureau. Il a fait l'objet, de la part du Bureau, d'un examen de portée plus limitée que celui auxquelles sont soumises les publications officielles. Le présent rapport est diffusé en vue de tenir les parties intéressées au courant des travaux de recherche en cours et d'en favoriser la discussion.

BIBLIOGRAPHIE

ALEXANDER, C.H. (1993). A continuous measurement alternative for the U.S. Census. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 486-491.

ALEXANDER, C.H. (1997). Questions relatives au plan de sondage de l'« American Community Survey » et résultats préliminaires des tests. *Recueil du Symposium 97, Nouvelles Orientations pour les Enquêtes et les Recensements*, 211-217.

ALEXANDER, C.H. (1998). Recent developments in the american community survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 92-100.

ALEXANDER, C.H., et WETROGAN, S. (2000). Integrating the american community survey and the intercensal demographic estimates program. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 295-300.

BOUNPANE, P. (1986). How increased automation will improve the 1990 census. *Journal of Official Statistics*, 4, 545-553.

BUTANI, S., ALEXANDER, C. et ESPOSITO, J. (1999). Using the american community survey to enhance the current population survey: Opportunities and issues. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, Statistical Policy Working Paper 29, 3, 102-111.

COCHRAN, W.G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society (B)*, 4, 102-118.

COCHRAN, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-129.

ECKLER, A.R. (1972). *The Bureau of the Census*. New York: Praeger Publishers.

HANSEN, M.H., MADOW, W.G. et TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 384, 776-793.

HAUSER, P.M. (1942). Proposed annual census of the population. *Journal of the American Statistical Association*, 37, 81-88.

HERRIOT, R.A., BATEMAN, D.B. et MCCARTHY, W. F. (1989). The Decade Census Program - A new approach for meeting the nation's needs for sub-national data. *Proceedings of the Social Statistics Section, American Statistical Association*, 351-355.

KISH, L., LOVEJOY, W. et RACKOW, P. (1961). A multi-stage probability sample for traffic surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 227-230.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

KISH, L. (1979a). Samples and censuses. *International Statistical Review*, 47, 99-109.

KISH, L. (1979b). Rolling samples instead of censuses. *Asian and Pacific Census Forum*, G(1), August 1979, 1-2, 12-13.

KISH, L. (1981). *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau*. Washington, D.C., U.S. Government Printing Office.

KISH, L. (1983). Data collection for details over space and time. *Statistical Methods and the Improvement of Data Quality*, (Éd. T. Wright). New York: Academic, 72-84.

KISH, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 1-12.

KISH, L. (1987). *Statistical Research Design*. New York: John Wiley & Sons.

KISH, L. (1990). Recensement par étapes et échantillons avec renouvellement complet. *Techniques d'enquête*, 16, 67-86.

KISH, L. (1997). Periodic and rolling samples and censuses. Chapître 7 dans *Statistics and Public Policy*, (Éd. Bruce D. Spencer). Clarendon Press, Oxford.

KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 1, 1998, 31-46.

KISH, L. (1999). Le cumul ou la combinaison d'enquêtes démographiques. *Techniques d'enquête*, 25, 2, 147-158.

KISH, L. (2001). Combining multi-population surveys. *Journal of Statistical Planning and Inference*, à paraître dans 2001.

KISH, L., et VERMA, V (1983). Census plus samples: Combined uses and designs. *Bulletin of the International Statistical Institute*, 50(1), 66-82.

KISH, L. et VERMA, V., (1986). Complete Censuses and Samples. *Journal of Official Statistics*, 2, 381-93.

MELNICK, D. (1991). The census of 2000 A. D. and beyond. *Reviews of Major Alternatives for the Census in the Year 2000*. U.S. Government Printing Office, Washington, D.C., August 1, 1991, 60-74.

MOONEY, H.W. (1956). Methodology in two California Health Surveys. *U.S. Public Health Monograph*, 70.

NATIONAL ACADEMY OF SCIENCES (1994). *Country People in the Information Age*. Duane L. Steffey et Norman M. Bradburn, éds. National Academy Press, Washington, D.C.

NATIONAL ACADEMY OF SCIENCES (1995). *Modernizing the U.S. Census*. Barry Edmonston et Charles Schultze, éds. National Academy Press, Washington, D.C.

PURCELL, N.J., et KISH, L. (1979). Estimation for Small Domains, *Biometrics*, 35, 365-384.

SAWYER, T. C. (1993). Rethinking the Census: Reconciling the Demands for Accuracy and Precision in the 21st Century. Présenté au Research Conference on Undercounted Ethnic Populations, 7 Mai, 1993.

TUPEK, A. R., WAITE, P. J. et CAHOON, L. S. (1990). Sample Expansion Plans for the Current Population Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 72-77.

Techniques d'enquête, juin 2002
Vol. 28, n° 1, pp. 47-53
Statistique Canada

La rénovation du recensement français

J.-M. DURR et J. DUMAIS[1]

RÉSUMÉ

Il devient de plus en plus difficile de mener des recensements de façon traditionnelle. Quand on a la possibilité d'interconnecter des fichiers administratifs, on ouvre une alternative intéressante à la pratique de recensements périodiques (Laihonen 2000; Borchsenius 2000). On retrouve ce type de proposition dans un article récent de Nathan (2001). La rénovation développée à l'INSEE repose sur le concept de « recensement continu » dont l'idée remonte à Kish (1981, 1990) et Horvitz (1986). Une première approche envisageable en France peut être trouvée dans Deville et Jacod (1996). Le présent article fait le point des développements méthodologiques depuis que l'INSEE a mis en route son Programme de rénovation du recensement de la population.

MOTS CLÉS : Échantillonnage équilibré; recensement; recensement continu; calage.

1. INTRODUCTION

1.1 Les raisons de la rénovation

La France conduit depuis de nombreuses années des recensements afin de déterminer la population légale de chacune de ses circonscriptions administratives et décrire les caractéristiques socio-démographiques de ses territoires à tous les niveaux géographiques, des quartiers des communes au pays dans son ensemble. Ainsi, le recensement de 1999 s'est déroulé selon le schéma habituel : dépôt et reprise des questionnaires par des agents recenseurs, organisation, assistance technique et contrôle par l'INSEE, exécution par le Maire en tant que représentant de l'État. Toutefois, certains éléments nous ont conduits à revoir ce dispositif.

Tout d'abord, l'intervalle intercensitaire a tendance à s'allonger. En effet, la périodicité des recensements n'est pas inscrite dans la loi, et la date de chaque recensement est fixée par décret. On est passé de recensements quinquennaux avant la guerre, à des écarts entre les recensements de 7, puis 8 ans. Le dernier recensement, prévu initialement en 1997, a été repoussé à 1999, soit 9 ans après le précédent, pour des raisons budgétaires. De plus, le public ne comprend pas toujours la nécessité d'une opération aussi lourde dans un contexte de fichiers administratifs toujours plus nombreux, même s'il redoute par ailleurs leur utilisation croisée. Enfin, le mouvement de décentralisation que connaît la France depuis plus de vingt ans a généré de nombreux besoins de données statistiques afin d'éclairer les politiques locales. Le recensement, source d'information locale par excellence, doit donc s'adapter et fournir des données plus fraîches et toujours finement localisées.

C'est pourquoi un programme de rénovation du recensement de la population a été engagé à l'Insee dès la fin des années 90. La France ne disposant pas de registre de population et le contexte national ne s'y prêtant pas, il a donc été décidé d'envisager une voie intermédiaire combinant la réalisation d'enquêtes annuelles par sondage et l'utilisation de fichiers administratifs non nominatifs que l'Insee est habilité à utiliser à des fins exclusivement statistiques. Pour les communes dont la population est inférieure à un seuil, fixé pour l'instant à 10 000 habitants, les enquêtes seront exhaustives et auront lieu chaque année par roulement au cours d'une période de cinq ans. Pour les autres communes, une enquête par sondage sera effectuée chaque année, la totalité du territoire de ces communes étant prise en compte au terme de la même période de cinq ans. Pour mener à bien cette rénovation, un cadre juridique nouveau s'est avéré nécessaire. Le Conseil d'État, consulté sur le projet, a préconisé, dans son avis du 2 juillet 1998, que le gouvernement soumette au Parlement un projet de loi.

Outre la nécessité de renforcer l'assise légale du recensement, il a considéré que le changement important des modalités d'élaboration des chiffres de population, alors que plus de 200 textes législatifs ou réglementaires s'y réfèrent, nécessitait de passer par la voie législative. Dans le cadre ainsi défini, le projet de loi vise essentiellement à définir les principes et à fixer les règles de base applicables à l'organisation du recensement. L'opération est placée sous la responsabilité et le contrôle de l'État : l'Insee organise le cadre de la collecte (concepts, protocoles), réalise le tirage des échantillons, veille à la qualité des informations collectées, exploite les données et les diffuse. Les communes, en tant que collectivités locales, préparent et réalisent les enquêtes de recensement. En compensation, l'État leur verse une dotation financière. Ces dispositions clarifient le rôle de chacun des partenaires et les responsabilités qui leur incombent.

[1] Jean-Michel Durr, Programme de rénovation du recensement de la population, INSEE, Direction générale, 18 boul. Adolphe Pinard, 75675 Paris CEDEX 14, France; Jean Dumais, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa K1A 0T6, Canada; Cet article a été préparé pendant que l'auteur était en contrat au Programme de rénovation du recensement de la population, INSEE.

1.2 Objectifs de qualité

Les **objectifs de qualité** suivants sont assignés au Programme :

1.2.1 Qualité des données :

Actualité : L'objectif est de pouvoir diffuser avant la fin de chaque année A la population légale de toutes les circonscriptions administratives au 1er janvier A-2, une description statistique de toutes les zones du territoire (communes et groupements de communes, quartiers des grandes villes, pays, *etc.*), relative au 1er janvier A-2 et une description statistique au 1er janvier de l'année A pour la France et ses grands territoires (régions...). Par comparaison avec un recensement général, le recensement rénové fournira des résultats analogues sur la population et les logements pour un gain en termes de fraîcheur moyenne des données de l'ordre de 3 à 4 ans.

Pertinence : Les données produites doivent être pertinentes pour une utilisation locale. En particulier, les données dont l'étude ne se justifie qu'à des niveaux géographiques très supérieurs à la commune seront écartées au profit de données plus utiles à l'action locale. Le choix des données à collecter est effectué dans le cadre du Conseil national de l'information statistique (CNIS), qui comprend des représentants des différentes catégories de producteurs et d'utilisateurs des données statistiques publiques. Un groupe de travail du CNIS a proposé des évolutions tout en préservant une nécessaire continuité avec les recensements précédents et en limitant la charge de réponse.

Précision : Le recensement a pour objet de fournir des données significatives pour tous les niveaux géographiques du pays. Les données produites doivent avoir une précision suffisante, y compris aux niveaux infra-communaux, pour les répartitions les plus utiles à ces niveaux. Ceci concerne, en particulier, les structures par sexe et âge, par type d'activité, par catégorie socioprofessionnelle et par catégorie de logement. La précision des données devra pouvoir être estimée et indiquée aux utilisateurs.

Compréhensibilité : Les données produites devront être compréhensibles aisément et d'utilisation comparable à celle des données produites par un recensement général, afin de ne pas perturber les utilisateurs.

1.2.2 Qualité du processus :

Charge de réponse : Afin de limiter la charge de réponse pour la population, les informations collectées devront se limiter au strict nécessaire. En particulier, les informations disponibles aux mêmes échelons géographiques dans d'autres sources ne seront pas collectées, sauf si l'obtention de ces informations dans le cadre du recensement permet des croisements utiles avec les autres variables. Le questionnaire individuel se limitera à un recto-verso, à l'instar des recensements précédents.

Questionnaire : la méthode de collecte étant basée sur le dépôt-retrait, les questionnaires doivent être accessibles à l'ensemble de la population. Afin de s'assurer de la compréhension des questions, des tests qualitatifs, par la méthode des groupes de discussion, ont été menés. Par ailleurs, un test de collecte a été réalisé au premier semestre 2001 auprès de 4 000 logements.

Confidentialité : Les données mobilisées lors du recensement sont protégées par la loi. Les informations individuelles collectées ne doivent être accessibles qu'aux seules personnes autorisées. Ces informations sont destinées à l'Insee et ne peuvent être utilisées qu'à des fins exclusivement statistiques. Seules les informations indispensables à la préparation et à la réalisation des enquêtes de recensement sont partagées, pour ce qui les concerne, avec les communes ou leurs groupements.

Qualité de couverture : la couverture des recensements généraux n'était pas systématiquement évaluée. A l'occasion du recensement de 1990, une enquête post-censitaire avait conclu à l'existence d'un sous-dénombrement de l'ordre de 1,8 % et d'un sur-dénombrement de l'ordre de 0,9 %, soit une précision globale d'environ 0,9 %. Les sous-dénombrements les plus importants étaient observés dans les grandes agglomérations. Le fait de procéder annuellement par sondage dans les communes de plus de 10 000 habitants, et donc de diminuer le nombre de personnes à recenser, devrait permettre de concentrer les efforts pour obtenir les réponses des enquêtés. Une évaluation régulière de la couverture du recensement rénové sera également effectuée, par comparaison avec les sources administratives et par enquêtes ponctuelles.

Robustesse technique et organisationnelle : Compte tenu de l'ampleur des volumes traités et de l'enjeu attaché au recensement, le Programme doit s'appuyer sur des innovations techniques maîtrisées. En corollaire, le démarrage de l'opération devra privilégier la robustesse du dispositif. L'introduction d'innovations techniques ou fonctionnelles pourra se faire tout au long de la vie du recensement dans le cadre de la maintenance évolutive ou de projets spécifiques. Ces projets pourront ainsi s'appuyer sur les opérations annuelles pour tester leurs apports avant généralisation. Cependant, la notion de cycle quinquennal sera privilégiée pour l'introduction de modifications sensibles, telles que des évolutions des questionnaires. L'organisation doit s'appuyer sur un partenariat équilibré entre l'Insee et les communes. L'organisation envisagée devra être réalisable par l'Insee compte tenu de ses moyens et de son programme de travail, moyennant des aménagements de l'organisation de la production de l'Institut. De même, l'organisation doit être supportable par les communes ou les établissements de coopération intercommunale. Le rythme annuel en grande commune et la possibilité pour les petites et moyennes communes de déléguer la collecte sur leur territoire à une structure intercommunale sont de nature à favoriser la professionnalisation des acteurs de la collecte.

Pour l'Insee, l'intégration des opérations de recensement dans le programme de travail annuel des directions régionales, de même que le volume de l'opération, sept fois

moins important que celui du recensement général, permettra un contrôle plus approfondi de l'opération. En effet, au lieu de 60 millions d'individus à recenser dans 36 700 communes avec 110 000 agents recenseurs une année donnée, il n'y aura chaque année que 18 000 agents recenseurs qui rendront visite à près de 9 millions d'habitants, et ce dans 8 000 communes environ.

Afin de définir précisément le cadre d'organisation entre l'Insee et les communes, un décret prévoira les responsabilités respectives, les moyens à mettre en place par les communes ainsi que les processus de validation des différentes étapes.

Maîtrise des coûts : la réalisation de la collecte sur un cycle de 5 ans permettra d'étaler les charges financières affectées à l'opération. En ce qui concerne les communes de plus de 10 000 habitants, la charge du recensement rénové sera plus faible que celle d'un recensement général de population en régime courant. En revanche, la charge devrait, pour les communes de moins de 10 000 habitants, être identique à celle d'un recensement général, mais tous les 5 ans alors que la périodicité d'un recensement général était de l'ordre de 8 ans. Le coût du recensement rénové en régime permanent sera équivalent au septième du coût d'un recensement général, ce qui permet de réaliser la réforme sans surcoût budgétaire. Cependant, les premières années de collecte pourront supporter un budget quelque peu supérieur à ce montant afin de permettre le rodage du processus.

2. STRATÉGIE D'ÉCHANTILLONNAGE

La commune constitue le point d'ancrage de la rénovation : les « petites et moyennes communes » (celles de moins de 10 000 habitants) seront sondées au taux (moyen) d'un cinquième par an et tous leurs logements seront visités; toutes les « grandes communes » seront visitées chaque année, mais seulement une fraction de leurs logements sera enquêtée.

2.1 Les petites et moyennes communes

Considérons d'abord le domaine des « petites et moyennes communes ». Dans chaque région, cinq groupes de rotation de communes seront formés. Ces groupes de rotation seront créés à partir des renseignements du recensement de la population (RP99) par tirage d'échantillons équilibrés (Deville, Tillé (1999, 2000)) sur la distribution âge-sexe des communes; cette approche devrait permettre de minimiser les variations interannuelles dues au seul sondage.

Les figures 1 et 2 illustrent comment les 5 groupes de rotation sont équilibrés. Ces deux figures donnent les « diagrammes à moustaches » de deux variables mesurées sur les 2 811 petites et moyennes communes de Rhône-Alpes au Recensement général de la population de 1990. Pour chaque groupe de rotation, on voit et les quartiles et l'étendue de la distribution; il est intéressant de noter la superposition des diagrammes. La variable « Nombre de femmes âgées de 20 à 39 ans » a été utilisée pour la composition des groupes; le nombre de résidences principales, ni aucune des variables associées au ménage ou au logement, n'intervient pas dans l'établissement de l'équilibre.

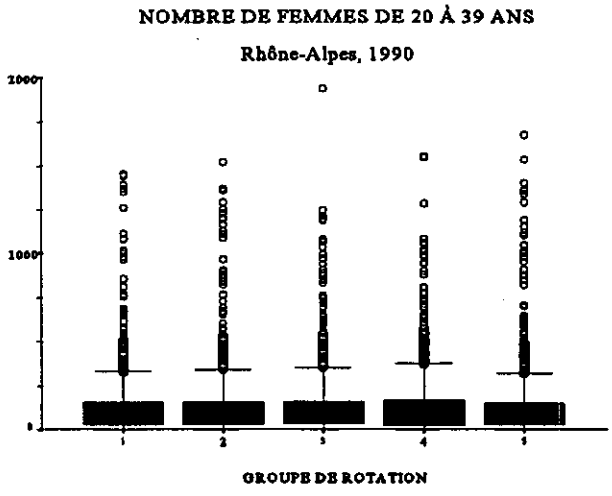


Figure 1.

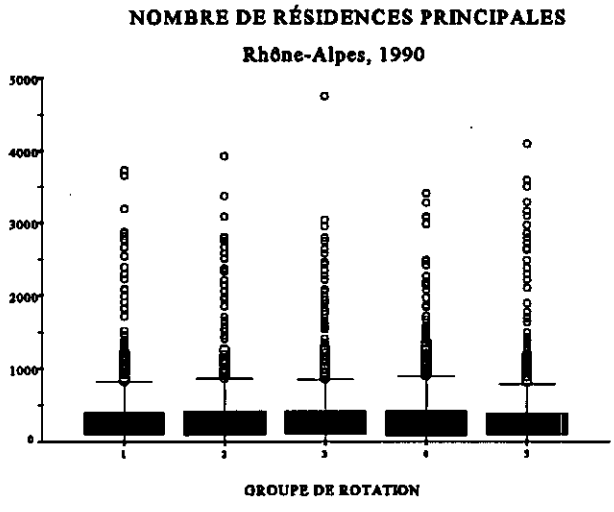


Figure 2.

Chaque année, on fera le recensement (exhaustif) de la population et des logements de toutes les communes d'un des groupes de rotation. Ainsi, chacune des « petites et moyennes communes » sera recensée une fois tous les 5 ans, et toutes les « petites et moyennes communes » à raison d'un cinquième par année.

2.2 Les grandes communes

Le sondage en « grande commune » utilisera le « répertoire d'immeubles localisés » (RIL). Ce répertoire est une

liste d'édifices (résidentiels, institutionnels ou commerciaux) repérés individuellement de façon à créer une cartographie numérisée. Le RIL sera d'abord alimenté par les résultats du RP99 permettant ainsi de décrire statistiquement chaque immeuble résidentiel (Au RP99, un « immeuble » est l'ensemble des logements desservis par la même cage d'escalier; un même immeuble physique peut devenir plusieurs « immeuble du R.P. »).

Le RIL sera mis à jour en continu à partir de permis de construire, de permis de démolir, de fichiers d'abonnés (eau, gaz, électricité, etc.), de renseignements fournis par les administrations locales et par l'observation directe sur le terrain. Ainsi, le RIL peut servir à la constitution d'une base de sondage « immeubles » en « grande commune ».

Dans chaque IRIS2000 (« îlots regroupés selon des indicateurs statistiques » zone homogène d'environ 2 000 habitants) de chaque « grande commune », on créera 5 groupes de rotation d'adresses sur le modèle du sondage des « petites et moyennes communes ». Trois strates supplémentaires seront prévues dans chaque commune : une pour les immeubles d'activité (usines, entrepôts,...), une seconde pour les logements collectifs (établissements, collectivités, communautés, internats,...), et une dernière pour les adresses neuves.

On visitera chaque année un cinquième des immeubles d'activité pour s'assurer qu'ils sont toujours vides de logements (logement de gardien, ou espace converti à l'habitation); les logements éventuellement trouvés dans de tels immeubles seraient considérés autoreprésentatifs parce qu'exceptionnels. L'ensemble des logements collectifs sera couvert chaque année; un cinquième d'entre eux seront visités alors que l'effectif des quatre autres cinquièmes sera éventuellement mis à jour par enquête téléphonique. Finalement, les immeubles d'habitation neufs seront insérés dans les groupes de rotation.

Comme décrit plus haut, les groupes de rotation d'adresses seront visités à tour de rôle au cours d'une période de 5 ans. Un sous-échantillon d'adresses correspondant à 40 % des logements du groupe sera alors tiré. Dans chaque adresse sélectionnée, l'ensemble des logements sera enquêté.

En résumé, l'échantillon annuel comptera environ 8 millions de bulletins individuels, 6 millions des « petites et moyennes communes » et 2 millions des « grandes communes ».

3. ESTIMATIONS GLOBALES ET DÉTAILLEES

En régime courant, trois séries d'estimations seront produites et diffusées chaque année : une série d'estimations des populations légales, une série d'estimations détaillées – dont on tirera les populations légales – et une série d'estimations globales servant au calage des précédentes.

3.1 Estimations globales

A ce jour, les plans de diffusion prévoient la publication, au 31 décembre *A*, des résultats nationaux et régionaux pour l'enquête tenue en début d'année *A*; ces estimations forment les estimations globales pour l'année *A*. Il est aussi prévu de publier à la même date les résultats pour chacune des « petites et moyennes communes » visitées durant la campagne de collecte de l'année *A*.

3.2 Estimations détaillées

Les fichiers administratifs fourniront des informations complémentaires à un niveau de détail assez fin. Il sera alors possible de mesurer la distorsion entre ce qui a été observé et ce qui est inscrit au fichier pour des objets similaires (immeubles, îlots,...). Cette distorsion sur des agrégats bien déterminés peut se traduire en facteur de correction à appliquer aux données administratives de sorte que la somme corrigée de celles-ci corresponde bien aux estimations censitaires.

A ce jour, il est prévu d'exploiter les fichiers administratifs à un niveau d'agrégation géographique (immeuble, îlot, district d'agent recenseur,...) qui renseigne sur les individus (âge, sexe d'après les fichiers de l'assurance maladie) ou leurs logements (fichiers de taxe d'habitation, ci-dessous TH).

Les résultats détaillés relatifs à l'année *A-2* seront mis à disposition au 31 décembre *A* (Il est prévu que l'acquisition et le traitement des fichiers administratifs prendront environ 2 ans); ces résultats détaillés seront l'amalgame d'observations faites par sondage (dans les grandes communes) ou recensement (dans les petites et moyennes communes) et de données synthétiques.

Les données synthétiques seront obtenues à partir de la relation entre données observées et administratives sur un même point en un même instant. Par exemple, pour une commune *C* du Groupe II recensée en *A-3*, dont le recensement est établi à $R_{C,II}^{A-3}$, on obtient une imputation de son recensement pour l'année cible *A-2* en faisant :

$$\tilde{R}_{C,II}^{A-2} = R_{C,II}^{A-3} \times \frac{\mathrm{Adm}_{II}^{A-2}}{\mathrm{Adm}_{II}^{A-3}} = R_{C,II}^{A-3} \times \frac{\sum_{c \in II} \mathrm{Adm}_c^{A-2}}{\sum_{c \in II} \mathrm{Adm}_c^{A-3}} ,$$

où Adm_c^a est la valeur des sources administratives pour la commune *c* et l'année *a*.

En régime permanent, pour une commune enquêtée en *A-5* et *A* (voir la figure ci-dessous), on aura mesuré des variables sur les personnes (âge, sexe, activité, profession,...) et sur les logements (taille du ménage, nombre de pièces, mode d'occupation, confort...) aux deux moments.

	A-6		A-5		A-4		A-3		A-2		A-1		A	
Gr I		Adm		Adm	R	Adm		Adm	?	Adm		Adm		Adm
Gr II		Adm		Adm		Adm	R_{II}^{A-3}	Adm	$\tilde{R}_{C,II}^{A-2}$	Adm		Adm		Adm
Gr III		Adm		Adm		Adm		Adm	R_{III}^{A-2}	Adm		Adm		Adm
Gr IV	R	Adm		Adm		Adm		Adm	?	Adm	R	Adm		Adm
Gr V		Adm	R	Adm		Adm		Adm	?	Adm		Adm	R	
Total	5R	ΣAdm	5R	ΣAdm	5R	ΣAdm	5R	ΣAdm	5R	ΣAdm	5R	ΣAdm	5R	ΣAdm

De plus, pour les groupes IV et V, l'estimation synthétique pour *A-2* pourrait profiter des informations recueillies durant la campagne de l'année *A-1* (respectivement *A*); en effet, il serait possible de calculer des facteurs d'ajustement par rapport au plus récent recensement et rétropoler sur la période intercensitaire. Par exemple, pour une commune *D* du groupe IV, on peut faire :

$$\Theta_1 = R_{D,IV}^{A-6} \times \frac{\sum_{c \in IV} \mathrm{Adm}_c^{A-2}}{\sum_{c \in IV} \mathrm{Adm}_c^{A-6}} \text{ et } \Theta_2 = R_{D,IV}^{A-1} \times \frac{\sum_{c \in IV} \mathrm{Adm}_c^{A-2}}{\sum_{c \in IV} \mathrm{Adm}_c^{A-1}}.$$

Il est à peu près certain que ces deux séries, extrapolations et rétropolations, ne coïncideront pas. Toutefois, il est souhaitable de publier une et une seule série d'estimations pour toute zone pour tout moment. Il apparaît naturel de produire une série « composite » dont les extrémités soient ancrées aux valeurs du recensement. La combinaison linéaire suivante peut jouer ce rôle en donnant plus d'important à la collecte la plus récente :

$$\tilde{R}_{D,IV}^{A-2} = 0{,}2 \times \Theta_1 + 0{,}8 \times \Theta_2.$$

De même, en donnant les définitions appropriées à Θ_1 et Θ_2, on ferait pour une commune *E* du groupe *V* :

$$\tilde{R}_{E,V}^{A-2} = 0{,}4 \times \Theta_1 + 0{,}6 \times \Theta_2.$$

Ces facteurs d'ajustement Θ devront être calculés pour des strates relativement fines de la population, des classes d'âge-sexe par exemple, de façon à préserver la plus grande souplesse démographique et géographique à l'ajustement des recensements. La qualité des fichiers administratifs et les disparités locales dicteront le niveau auquel l'ajustement peut être réalisé convenablement (des départements, des régions métropolitaines,...). On peut tenir un raisonnement analogue en grande commune si on remplace une « petite et moyenne commune » par une « adresse ».

Finalement, quand toutes les communes de tous les groupes auront été imputées, il est improbable que l'estimation du total d'une variable d'intérêt obtenu du fichier imputé (estimations détaillées) ne corresponde plus au total estimé à partir des seules observations (estimations globales publiées deux ans auparavant). Il est donc convenu que les estimations détaillées soient calées aux estimations globales. Le niveau de calage dépendra encore une fois des tendances locales et de la qualité des estimations globales.

3.3 Estimations de populations légales

La série des estimations des populations légales sont la troisième série tirée du recensement. Il s'agit des chiffres de population auxquels se réfèrent les textes de loi pour le financement des communes, le découpage électoral, la composition du conseil municipal, ...

La « population totale » légale d'une commune comprend :

- les individus dont la résidence principale est située dans la commune,
- ceux qui résident dans un établissement ou un logement collectif situé sur le territoire de la commune,
- ceux qui résident dans un établissement ou un logement collectif situé dans une autre commune mais qui ont gardé un logement dans leur commune d'origine,
- les personnes qui ont une résidence dans la commune et qui vivent dans un logement collectif d'une autre commune pour leur travail ou dans une autre commune pour leurs études,
- et les populations administrativement rattachées à la commune (forains, mariniers, *etc.*).

On voit donc que ces populations ne peuvent être estimées qu'une fois l'ensemble du territoire couvert, c'est-à-dire qu'au moment des estimations détaillées.

3.4 Estimation de la variance d'échantillonnage

Il est prévu d'accompagner les estimations, globales et détaillées, d'une mesure de leur qualité statistique. Les travaux portant sur cet aspect ont débuté à l'automne 2001; l'option privilégiée pour l'instant est le recours à des tables de référence, comme on le fait pour l'enquête canadienne sur la population active, par exemple. Les variances d'échantillonnage seront vraisemblablement obtenues par ré-échantillonnage sur la base de sondage.

3.5 Imprécision due à la synthèse

On a montré, à la section 3.2, comment la production des estimations par synthèse utilise l'information amassée : d'abord une extrapolation pour un « vieux » recensement, pour deux groupes de rotation (disons I et II); ensuite l'utilisation directe des résultats du recensement pour un troisième groupe de rotation (disons III); enfin, la combinaison des extrapolations et rétropolations pour caler les deux derniers groupes (disons IV et V).

Cette synthèse peut être formalisée sous l'angle d'un modèle de non-réponse (Särndal 1990) : la campagne annuelle s'apparente alors à un sondage à 100 % qui subirait 80 % de non-réponse, laquelle est palliée par le recours à l'imputation par le ratio. Si l'échantillon complet est noté s, les répondants sont notés r et les non-répondants sont notés s-r, on peut écrire

$$y_{\cdot k} = \begin{cases} y_k & \text{si } k \in r \\ \hat{\beta}\, x_k & \text{si } k \in s-r \end{cases} \quad \text{avec} \quad \hat{\beta} = \frac{\bar{y}_r}{\bar{x}_r}.$$

C'est-à-dire que le modèle d'imputation est

$$\xi : \begin{cases} y_k = \beta x_k + \varepsilon_k. \\ E(\varepsilon_k) = 0 \\ V(\varepsilon_k) = \sigma^2 x_k \end{cases}$$

où les erreurs ε_k sont non corrélées.

Avec un tel modèle d'imputation, sous sondage aléatoire simple,

$$\hat{Y}_{\cdot} = \frac{N}{n} \sum y_{\cdot k} = \frac{N}{n} \left\{ \sum_r y_k + \sum_{s-r} \hat{\beta} x_k \right\} = \ldots$$
$$= N \frac{\bar{y}_r}{\bar{x}_r} \bar{x}_s$$

L'incertitude autour de l'estimation avec imputation dépend des aléas de sondage et de la qualité du modèle d'imputation ξ :

$(\hat{Y}_{\cdot} - Y)$	=	$(\hat{Y} - Y)$	+	$(\hat{Y}_{\cdot} - \hat{Y})$
incertitude totale	=	incertitude du sondage	+	incertitude du modèle

Si on suppose que l'imputation se fait sans biais :

$$E_\xi E_s E_r (\hat{Y}_{\cdot} - Y) = 0$$

on a,

$$V_{\text{totale}} = E_\xi E_s E_r (\hat{Y}_{\cdot} - Y)^2 = \ldots$$
$$= E_\xi E_s E_r (\hat{Y} - Y)^2 + E_\xi E_s E_r (\hat{Y}_{\cdot} - \hat{Y})^2$$
$$= E_\xi V_s + E_s E_r V_\xi$$
$$V_{\text{totale}} = V_{\text{échantillon}} + V_{\text{imputation}}$$

en supposant que le sondage et la réponse sont indépendants du mécanisme d'imputation.

L'utilisation de données imputées comme si elles avaient été observées dans le calcul de l'estimation de V_s même à une sous-estimation de $V_{\text{échantillon}}$. En espérance,

$$E_\xi (\hat{V}_s - \hat{V}_{\cdot s}) = V_{dif}.$$

Pour les estimateurs de ces variances, Särndal montre que l'on obtient

$$\hat{V}_{\text{sondage}} = N^2 \left(\frac{1}{n} \frac{1}{N}\right) \left\{ S_{\cdot}^2 + C_0 \hat{\sigma}^2 \right\}$$

avec C_0 près de $\left(1 - \frac{m}{n}\right) \bar{x}_{s-r}$ et $\hat{\sigma}^2$ près de $\frac{\sum_r e_k^2}{\sum_r x_k}$ et

$$\hat{V}_{\text{imputation}} = N_2 \left(\frac{1}{m} - \frac{1}{n} \right) A \bar{x}_s \hat{\sigma}^2,$$

avec $A = \bar{x}_{s-r} / \bar{x}_r$, qu'on peut comprendre comme un effet de sélection des répondants. On remarque que si $x_k = 1$, alors, on obtient un sondage à deux phases de taille m parmi n et n parmi N. De plus, si $s = r$, $V_{\text{totale}} = V_{\text{sondage}}$.

Dans le modèle de Särndal, les x (données administratives) et y (données censitaires) sont contemporains; à tout le moins, on aura observé certains des y. En reprenant la structure développée à la section précédente, on aurait :

Année A-2		
y_k	x_k	m répondants (Groupe III)
$y_{\cdot k}$	x_k	n-m imputations (autres groupes)

Dans l'application du RRP, tout n'est pas synchrone :

... A-4		A-3		A-2		A-1	A
Y_I^{A-4}	X_I^{A-4}		X_I^{A-3}		X_I^{A-2}		
	X_{II}^{A-4}	Y_{II}^{A-3}	X_{II}^{A-3}	$Y_{\cdot II}^{A-2}$	X_{II}^{A-2}		
	X_{III}^{A-4}		X_{III}^{A-3}	Y_{III}^{A-2}	X_{III}^{A-2}		
	X_{IV}^{A-4}		X_{IV}^{A-3}		X_{IV}^{A-2}	... ...	
	X_V^{A-4}		X_V^{A-3}		X_V^{A-2}		... ...

En effet, $Y_{II}^{A-3}, X_{II}^{A-3}, Y_{\cdot II}^{A-2}$, et X_{II}^{A-2} ne sont pas toutes mesurées ou observées la même année. En fait, quand on ne regarde que le seul groupe de rotation III par exemple, on a un échantillon de taille n en *A-2* et un échantillon identique en *A-3* entièrement non-répondant. En conséquence, certains paramètres de l'estimation de V_{totale} ne sont plus calculables.

En revanche, en regardant le problème pour un temps donné, on a bien un échantillon de taille n répondants et $4n$ non-répondants. On pourrait approcher l'incertitude du

processus d'imputation asynchrone (celui du recensement rénové) par celle du processus d'imputation synchrone (proche du modèle de Särndal).

Cette approche a été testée sur les petites et moyennes communes de Rhône-Alpes, pour lesquelles les groupes de rotation, la Taxe d'Habitation (TH90) et le Recensement général de la population de 1990 (RGP90) sont disponibles (Kauffmann 2000). La méthode donne de bons résultats pour les variables bien corrélées avec la TH; les résultats indiquent aussi qu'une source de données administratives proches des variables décrivant les personnes sera nécessaire afin de maintenir les erreurs du modèle à un niveau tolérable.

4. TRAVAUX EN COURS

Les travaux de méthodologie autour de la rénovation du recensement sont loin d'être complétés. Au nombre des chantiers ouverts, notons

- l'établissement des règles de passage de seuil, les problèmes d'oscillation autour du seuil des 10 000 habitants et le calcul des populations légales;
- la sensibilité des bornes de strate en grande commune et leur robustesse dans le temps;
- la mise à jour et la maintenance des bases de sondage et des échantillons, en particulier, les ajustements éventuels suite aux passages de seuil et l'incorporation de nouveaux objets dans les groupes de rotation;
- l'imputation massive et la synthèse, tant les modèles que leur précision;
- l'estimation de la précision des estimateurs; et
- la collecte auprès des populations mobiles.

BIBLIOGRAPHIE

BORCHSENIUS, L. (2000). From a Conventional to a Register-based Census of Population. Les Recensements après 2001, Séminaire Eurostat-INSEE, Paris.

DEVILLE, J.C., et JACOD, M. (1996). Replacing the Traditional French Census by a Large Scale Continuous Population Survey. *Annual Research Conference Proceedings*, USBC, Washington.

DEVILLE, J.C., et TILLÉ, Y. (1999). *Balanced Sampling by Means of the Cube Method.* CREST-ENSAI, document interne, soumis pour publication.

DEVILLE, J.C., et TILLÉ, Y. (2000). Échantillonnage équilibré par la méthode du cube et estimation de variance. *Journées de Méthodologie*, décembre 2000, INSEE, Paris.

HORVITZ, D.G. (1986). Statement to the Subcommittee on Census and Population. Committee on Post Office and Civil Service, House of Representatives, Research Triangle Park, North Carolina.

KAUFFMANN, B. (2000). *Estimations annuelles dans la rénovation du recensement de la population.* Note de travail interne, Département de la démographie, INSEE.

KISH, L. (1981). Population Counts from Cumulated Samples. Congressional Research Service. *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau*, Prepared for the Subcommittee on Census and Population, Committee on Post Office and Civil Service, House of Representatives, Washington.

KISH, L. (1990). Recensement par étapes et échantillons avec renouvellement complet. *Techniques d'enquête*, 16, 1, 63-71, Statistique Canada, Ottawa.

LAIHONEN, A. (2000). 2001 Round Population Censuses in Europe. *Les Recensements après 2001*, Séminaire Eurostat-INSEE, Paris.

NATHAN, G. (2001). Models for combining longitudinal data from administrative sources and panel surveys. Présentation invitée, ISI, Séoul, aôut 2001.

SÄRNDAL, C.-E.(1990). Méthodes pour estimer la précision des estimations d'enquête losqu'il y a eu imputation. *Recueil : Symposium 90, Mesure et amélioration de la qualité des données*, Statistique Canada, 337-350.

Techniques d'enquête, juin 2002
Vol. 28, n° 1, pp. 55-66
Statistique Canada

Étalonnage des paramètres estimés des modèles logit de choix binaire et des modèles semiparamétriques de survie

IAN CAHILL et EDWARD J. CHEN[1]

RÉSUMÉ

Nous élaborons une méthode d'exploitation des données provenant de plusieurs enquêtes et périodes de référence grâce à l'étalonnage des paramètres estimés des modèles logit de choix binaire et des modèles semiparamétriques de survie. L'objectif est de mettre à profit la source assez riche de covariables socioéconomiques que constituent l'Enquête sur la dynamique du travail et du revenu (EDTR) de Statistique Canada et l'horizon temporel de l'Enquête sur la population active (EPA) conjugué au suivi des individus lors de chaque interview durant les six mois où ils font partie de l'échantillon de l'enquête. Nous démontrons comment la méthode peut être appliquée à l'aide du module sur le congé de maternité du projet de microsimulation dynamique LifePaths de Statistique Canada. Le choix consistant à prendre un congé de maternité plutôt que d'arrêter de travailler est spécifié sous forme de modèle logit binaire, tandis que la durée du congé est spécifiée sous forme de modèle semiparamétrique de survie à hasards proportionnels comprenant des covariables ainsi qu'une fonction de hasard de base qui peut varier chaque mois. Nous estimons d'abord les deux modèles par la méthode du maximum de vraisemblance au moyen des données regroupées de l'EDTR sur les congés de maternité à compter de la période de 1993 à 1996, puis nous les étalonnons en fonction des estimations annuelles calculées d'après les données de l'EPA recueillies pour la période de 1976 à 1992. Dans le cas du modèle logit, nous ajustons le prédicteur linéaire d'après une estimation du logarithme de la cote exprimant la chance (log-odds) basée sur les données de EPA. Pour le modèle de survie, nous utilisons un estimateur de Kaplan-Meier de la fonction de hasard calculé d'après les données de l'EPA pour rajuster la valeur prévue de la fonction de hasard dans le modèle semiparamétrique.

MOTS CLÉS : Microsimulation; étalonnage; modèles semiparamétriques de survie; logit binaire.

1. INTRODUCTION

Les chercheurs fondent souvent leurs modèles économétriques sur les données d'une enquête réalisée sur une courte période. Le cas échéant, il peut être souhaitable d'intégrer des renseignements provenant d'une source complémentaire couvrant une période plus longue, même si des mesures ne sont disponibles que pour la variable dépendante. Pour une classe générale de modèles non linéaires, nous mettons au point une méthode simple d'étalonnage des estimations des paramètres calculées d'après les données d'une enquête riche en variables explicatives en fonction des renseignements fournis par une enquête dont l'horizon temporel est considérable. L'un des objectifs principaux est de faire concorder les prévisions du modèle avec l'information provenant de la source secondaire de données. Nous décrivons d'abord l'application de la méthode à un modèle logit simple de choix binaire, puis à un modèle semiparamétrique de survie. Puisqu'il peut être considéré comme une série de choix binaires, tout en étant interprété comme un modèle temporel continu incomplètement observé, le modèle de survie représente une généralisation naturelle de la première application.

Nous procédons, pour illustrer la méthode, à une étude du congé de maternité. L'Enquête sur la dynamique du travail et du revenu (EDTR) de Statistique Canada fournit des données sur la fréquence à laquelle le congé de maternité est choisi de préférence au retrait du marché du travail et sur la durée du congé de maternité, ainsi que sur un riche ensemble de variables explicatives. Par conséquent, nous utilisons les données de cette enquête pour estimer les paramètres de base, y compris ceux qui déterminent les effets des variables explicatives sur la fréquence (modèle logit) et sur la probabilité de retourner au travail (modèle de survie). L'enquête sur la population active (EPA) réalisée par Statistique Canada fournit des approximations raisonnables de la fréquence et de la durée du congé de maternité remontant jusqu'à 1976. Par conséquent, les paramètres estimés d'après les données de l'EDTR sont étalonnés en fonction des estimations calculées d'après les données de l'EPA de la fréquence et de la probabilité de retour au travail pour la période allant de 1976 à 1992, période antérieure à celle pour laquelle existent des données de l'EDTR.

Les travaux ont été réalisés durant la mise au point du module sur le congé de maternité du modèle de microsimulation LifePaths de Statistique Canada. L'objectif du projet LifePaths est de construire un modèle de microsimulation dynamique englobant autant de détails que possible sur les processus socioéconomiques au Canada, ainsi que les profils chronologiques d'évolution de ces processus. Le modèle LifePaths a été utilisé dans le cadre

[1] Ian Cahill, Partenariat et évaluation continue, DRHC, 140 Promenade du Portage, Phase IV, 3e étage, Bureau 3D475, Gatineau (Québec) K1A 0J9; Edward J. Chen, Division des méthodes des enquêtes auprès des ménages, Statistique Canada, Immeuble R.-H.- Coats, 16e étage, Parc Tunney, Ottawa (Ontario) K1A 0T6.

d'une gamme variée d'analyses de politiques et de travaux de recherche stratégique. À titre d'exemple, mentionnons l'analyse de la politique canadienne de prêts aux étudiants (étude réalisée pour le compte de Développement des ressources humaines Canada et du gouvernement de l'Ontario), l'étude du retour aux études (Appleby, Boothby, Rouleau et Rowe, 1999), l'étude de l'utilisation du temps (Wolfson et Rowe 1996; Wolfson 1997; Wolfson et Rowe 1998a), l'étude du régime de transfert d'impôt et des pensions (Wolfson, Rowe, Gribble et Lin 1998; Wolfson et Rowe 1998b) et l'étude des carrières au sein de la population active (Rowe et Lin 1999). En outre, la nécessité de produire des données pour LifePaths a donné lieu à de nouveaux travaux de recherche portant, par exemple, sur les carrières dans l'enseignement (Chen et Oderkirk 1998; Rowe et Chen 1997; Plager et Chen 1999) et sur la corrélation des gains (Chen et Rowe 1999).

Le modèle LifePaths est conçu pour intégrer les renseignements socioéconomiques provenant de toutes les sources pertinentes disponibles à Statistique Canada. Aussi sa construction a-t-elle motivé l'étude de méthodes permettant d'exploiter des sources multiples de données. L'intégration d'un modèle estimé dans LifePaths est un outil puissant si l'on veut faire, d'après le modèle, des inférences que l'on peut comparer à des renseignements provenant d'autres sources. Par exemple, Rowe et Lin (1999) ont obtenu les données sur la durée d'occupation d'un emploi par simulation au moyen d'un modèle estimé d'après des données longitudinales couvrant une courte période, puis par comparaison des résultats aux données d'une enquête transversale. Nous décrivons ici l'un des volets des efforts déployés en vue de créer un outil qui fournira le maximum d'information pouvant être extraite des sources de données de Statistique Canada.

La présentation de l'article vise à illustrer la façon dont les problèmes techniques sont souvent identifiés durant l'élaboration du modèle LifePaths et comment leurs solutions sont intégrées dans le processus de développement du modèle. Par conséquent, nous fournissons une quantité assez importante de renseignements généraux sur certaines questions connexes. À la section 2, nous décrivons dans les grandes lignes le contexte du problème d'étalonnage et à la section 3, nous présentons la théorie qui sous-tend notre solution, en mentionnant certaines extensions possibles qui pourraient faire l'objet de futurs travaux. À la section 4, nous décrivons les modèles auxquels elle sera appliquée, y compris certains détails concernant l'explication de leur paramètres pour la période de référence. Puis, à la section 5, nous décrivons l'application de la méthode d'étalonnage à ces modèles. À la section 6, nous présentons nos résultats empiriques et nous en discutons, puis, à la section 7, nous présentons certaines conclusions générales.

2. CONTEXTE DU PROBLÈME

Pour placer le problème dans son contexte, nous commençons par donner un aperçu de la structure du modèle LifePaths, une brève description des sources de données utilisées et un exposé de la façon dont s'est posé le problème de l'étalonnage.

2.1 Structure du modèle LifePaths

Le modèle LifePaths permet de simuler le cycle de vie d'individus sous forme d'une série d'événements qui modifient l'ensemble de « variables d'état » décrivant les caractéristiques démographiques, sociales et économiques individuelles. Pour tout individu, des périodes d'attente sont associées à chaque événement possible, avec la possibilité d'être infinies. Les périodes d'attente peuvent aussi dépendre conditionnellement des valeurs des variables d'état. Le type d'événement pour lequel la période d'attente est la plus courte est celui qui se produit (c'est-à-dire que le modèle active les fonctions associées à cet événement). La modification de toute variable d'état au moment où a lieu un événement peut mener à la production de nouvelles périodes d'attente pour d'autres événements.

Pour initialiser un cas, le modèle LifePaths attribue de façon aléatoire le sexe, la province de résidence, l'âge au moment de l'immigration et l'année de naissance de l'individu « dominant ». L'année de naissance peut varier de 1892 à 2051. Les hypothèses concernant la mortalité et l'immigration sont conçues de façon à reproduire les structures provinciales âge-sexe. Lorsqu'un individu dominant se marie, établit une union de fait ou a un enfant, le modèle crée un individu non dominant ayant les caractéristiques appropriées qui est relié à l'individu dominant et fait alors partie du cas. Une fois créés, les individus non dominants subissent les mêmes événements éventuels que les individus dominants. Cependant, comme leur raison d'être est de compléter le profil de l'acteur dominant, les données qui les concernent sont généralement filtrées de tous les rapports tabulaires.

À l'heure actuelle, LifePaths comprend des modèles de fécondité, de mortalité, de nuptialité (y compris les unions de fait), de cheminement dans la formation, de carrière au sein de la population active, de congé de maternité, de nombre d'heures de travail, de revenu, d'impôt et de transferts. Le modèle des carrières au sein de la population active décrit les transitions entre les états d'« employé salarié », de « travailleur autonome » et de « non-emploi ». Il comprend aussi un modèle du départ à la retraite et du travail d'etudiant. Le modèle de cheminement dans la formation au secondaire et postsecondaire au niveau provincial est bien au point et fortement développé.

2.2 Sources des données

Les paramètres de base du modèle du congé de maternité ont été calculés d'après les données de l'EDTR couvrant les congés de maternité à compter de la période allant de 1993 à 1996. L'utilisation des données recueillies pour 1997 nous a permis de suivre la plupart des congés de maternité jusqu'à leur achèvement plutôt que d'utiliser des données fortement censurées. L'EDTR est une enquête-ménage conçue pour permettre l'analyse longitudinale et transversale des situations financière et professionnelle des individus. L'enquête, qui a été lancée en 1993, suit les mêmes personnes pendant six ans, un nouveau groupe de rotation étant créé tous les trois ans. Chaque groupe de rotation compte 15 000 ménages regroupant 30 000 adultes. L'enquête nous fournit les données sur le mois de naissance de l'enfant, les données mensuelles sur la situation d'activité et sur un riche ensemble de variables explicatives, y compris la durée de l'emploi, un indicateur du travail autonome, le rang de naissance de l'enfant, l'existence d'un conjoint employé, la province de résidence, le niveau de scolarité et l'âge. Nous pouvons aussi déterminer si une mère qui a quitté un emploi dans les quatre mois avant un accouchement a repris le même emploi dans les 16 mois qui suivent. Ces spécifications, que nous utilisons comme définition pratique du congé de maternité, deviennent notre unité d'analyse avec une légère extension afin d'inclure le un pourcent de cas où une mère a repris un emploi différent après un état d'activité correspondant à une absence au cours du mois précédent. En utilisant cette unité d'analyse, nous obtenons un échantillon de 835 naissances. Comme nous le montrons à la section 6, cet échantillon est de taille suffisante pour dégager certains facteurs explicatifs importants. Plus précisément, nous constatons que l'effet de plusieurs facteurs est significatif au niveau de confiance de 95 %. Cet échantillon contient environ 730 mères distinctes, qui représentent plus de 87 % de l'échantillon de naissances. Autrement dit, il existera une certaine corrélation entre les observations, parce que certaines mères ont pris deux congés de maternité ou plus durant la période de référence, mais nous estimons que cette corrélation n'est pas suffisamment forte pour justifier l'utilisation d'outils statistiques spéciaux.

L'EPA est une enquête-ménage mensuelle qui se concentre sur la situation d'activité, mais qui fournit aussi des données sur plusieurs caractéristiques démographiques. Normalement, les données de l'enquête servent exclusivement à l'analyse transversale. Cependant, pour le projet LifePaths, nous avons créé un fichier couvrant la période de 1976 à 1995 afin de suivre les individus à mesure qu'ils passent par les six groupes de rotation mensuels de l'enquête, et d'obtenir ainsi une fenêtre sur six mois de l'activité de chaque individu sur le marché du travail. Comme le nombre d'enfants et leur âge est enregistré chaque mois, il est possible d'observer l'arrivée d'un nouvel enfant. Puisque nous utilisons toutes les enquêtes réalisées durant la période de référence, l'échantillon est de très grande taille et nous observons environ 26 000 naissances.

Dans la fenêtre de six mois de l'EPA, nous notons la situation d'activité d'une nouvelle mère au moment où la présence de l'enfant est déclarée pour la première fois. Cette variable est la clé de l'estimation de la probabilité que la mère choisisse un congé de maternité plutôt qu'un retrait du marché du travail. Nous commençons par considérer P(E), c'est-à-dire la proportion de mères de cette catégorie qui sont employées. Si la mère est « employée, au travail », nous supposons qu'elle s'est absentée brièvement de son travail, c'est-à-dire moins d'un mois. Si elle est « employée, absente du travail », il se peut qu'elle ait choisi de prendre un congé de maternité, puis de reprendre son emploi après ce congé. Cependant, il n'en n'est pas toujours ainsi. Une nouvelle mère que nous considérons comme employée et absente (*EA*) peut plus tard quitter son emploi et passer à l'état de non-employée (*NE*). Pour tenir compte de cette situation, considérons les mères ayant un enfant de moins d'un an observées dans une fenêtre. Nous calculons la proportion $P(EA \rightarrow NE)$ de transitions de l'état d'« employée, absente du travail » à l'état de « non-employée ». Nous estimons aussi la proportion $P(NE \rightarrow AE)$ de mères qui retournent à leur ancien emploi (*AE*) après avoir quitté le marché du travail. Nous calculons cette estimation d'après les données sur les mères ayant un jeune enfant qui font la transition d'un état de non-employée à un emploi pour lequel la date de début est antérieure au mois précédent. Notre estimation de la probabilité de choisir le congé de maternité devient maintenant $P(E) - P(EA \rightarrow NE) + P(NE \rightarrow AE)$.

Il est également possible d'observer des mères ayant un enfant de moins d'un an qui passent de l'état d'« employée, absente du travail » pour des raisons personnelles ou familiales à l'état d'« employée, au travail ». Nous utilisons cette transition comme approximation du retour au travail après un congé de maternité. Puisque la durée de l'absence est déclarée le mois précédent, cet élément de donnée est essentiel à l'étalonnage du modèle de survie.

La discussion qui précède illustre la faiblesse des données de l'EPA comparativement à celles de l'EDTR dans le cadre d'une étude du congé de maternité. L'EPA fournit non seulement un moins grand nombre de variables explicatives que l'EDTR, mais elle nous oblige aussi à accepter des approximations pour les variables dépendantes. Néanmoins, nous avons besoin de la profondeur chronologique des données de l'EPA. Cette relation entre les ensembles de données constitue le contexte du problème d'étalonnage décrit à la section suivante.

L'EDTR et l'EPA ont toutes deux un plan d'échantillonnage complexe comprenant une stratification détaillée et des méthodes complexes de calcul des coefficients de pondération des observations. Nous nous servons systématiquement des coefficients de pondération aussi bien pour le calcul des estimations que pour celui des fréquences. Nous

utilisons des méthodes assez simples dont nous discutons aux sections 4 et 5.

2.3 Le problème de l'étalonnage

Notre problème d'étalonnage a pour contexte un modèle représentant les femmes qui font le choix entre le retrait du marché du travail ou le départ en congé de maternité et qui, si elles choisissent ce dernier, décident de la durée de ce congé. La première décision est représentée par un modèle logit binaire et la deuxième, par un modèle semiparamétrique de survie. Ces modèles incluent l'un et l'autre un vecteur de variables explicatives et de paramètres connexes. Dans le modèle de simulation LifePaths, les décisions sont prises dans le cadre de l'événement représentant le choix ayant trait au congé de maternité, choix qui a toujours lieu au milieu de la grossesse. Les données de l'EDTR conviennent assez bien à l'estimation des paramètres de base de ces deux modèles. Cependant, puisqu'un des objectifs principaux du projet LifePaths est d'encapsuler les profils d'évolution chronologique des processus socioéconomiques, il a fallu étalonner les estimations des paramètres fondées sur les données de l'ETDR en fonction des estimations annuelles de la moyenne des variables dépendantes calculées d'après les données de l'EPA.

Pour traiter ce problème, nous supposons, pour deux raisons, que les caractéristiques observées de la population sont stables. En premier lieu LifePaths est en cours d'élaboration et l'exercice d'étalonnage dont nous présentons les résultats a été réalisé à un stade où d'autres éléments du modèle destinés à prédire ces caractéristiques faisaient l'objet d'une révision approfondie. Dans la section 3.3, nous effleurons la question des conséquences de l'évolution des caractéristiques de la population. En deuxième lieu, nous supposons que l'évolution systématique des résultats observés entre les périodes de référence est due principalement à la variation de certains facteurs non inclus dans les caractéristiques mesurées des individus. Dans le cas de notre application, nous avons constaté une tendance à choisir le congé de maternité plutôt que le retrait du marché du travail qui semble due à une évolution sociale plutôt qu'à des modifications de la composition de la population. Nous avons également observé un changement dans la distribution des durées des congés de maternité qui semblent résulter des modifications apportées au programme de l'assurance-chômage (a.-c.) mis en œuvre lors de l'adoption du projet de loi C-21 en 1990. C'est à cette époque qu'ont été introduites les prestations parentales prolongeant de 15 à 25 semaines la période durant laquelle nombre de mères ont droit aux prestations. Beaucoup de mères retournent au travail à peu près au moment où elles ont épuisé les prestations d'assurance-chômage.

3. MÉTHODE D'ÉTALONNAGE

À la présente section, nous présentons la méthode sous forme abstraite, afin de préciser les hypothèses, d'élaborer la notation et de montrer la similarité entre l'application au choix binaire et à l'analyse de survie.

3.1 Application au choix binaire

Le modèle de base de la méthode d'étalonnage a trait au choix binaire. Puisque notre objectif principal n'est pas d'observer les changements dans la population, nous simplifions l'analyse en supposant que les variables explicative ou les caractéristiques individuelles durant la période τ sont représentées par une série de vecteurs aléatoires indépendants et identiquement distribués X^τ. Nous reconnaissons qu'il s'agit d'une hypothèse assez forte. Néanmoins, pour des raisons dont nous discutons à la section 2.3, nous l'utilisons pour nos travaux empiriques. Nous montrons à la section 3.3 qu'il est assez simple d'étendre la théorie afin d'intégrer les tendances observées pour les variables indépendantes.

Considérons un prédicteur linéaire donné par

$$\eta^\tau(x) = \beta' x + \gamma^\tau \tag{3.1}$$

où β est un vecteur de coefficients constant au cours du temps, x est un résultat possible de X^τ et γ^τ représente un paramètre particulier à la période τ. À noter que x ne contient aucun « terme constant ». Posons que Y^τ est une variable aléatoire, à distribution conjointe avec X^τ, qui prend la valeur 1 si un événement a lieu et 0 s'il n'a pas lieu. Supposons que la probabilité de l'événement, conditionnelle aux caractéristiques x, est donnée par

$$E(Y^\tau \mid X^\tau = x) = \pi^\tau(x) = F(\eta^\tau(x)) \tag{3.2}$$

où nous posons que F est une fonction de distribution continue. Cette fonction sera alors bornée par les valeurs zéro et un, et possédera une fonction inverse g, telle que

$$\eta^\tau(x) = g(\pi^\tau(x)). \tag{3.3}$$

Dans le contexte des modèles linéaires généralisés, g est appelée fonction de lien. Nous commençons par calculer les estimations du maximum de vraisemblance des paramètres de base $\hat{\beta}$ et $\hat{\gamma}^{\tau_0}$ à l'aide des données recueillies pour la période de référence τ_0 (dans notre cas, il s'agit de la période pour laquelle les données de l'EDTR sont disponibles). Naturellement, ces données doivent inclure les variables qui correspondent aux résultats de X^τ ainsi que de Y^τ. Il reste à estimer γ^τ pour chaque période τ. Les équations (3.1) et (3.3) impliquent que

$$E\{\eta^\tau(X^\tau) - \eta^{\tau_0}(X^{\tau_0})\} = \gamma^\tau - \gamma^{\tau_0} = E\{g(\pi^\tau(X^\tau))\} - E\{g(\pi^{\tau_0}(X^{\tau_0}))\}. \tag{3.4}$$

Puisque nous ne disposons, sur les résultats de Y^τ, que des données provenant de l'EPA pour chaque période, nous estimons les termes γ^τ comme suit

$$\hat{\gamma}^\tau = \hat{\gamma}^{\tau_0} + g(\hat{\pi}^\tau) - g(\hat{\pi}^{\tau_0}) \tag{3.5}$$

où $\hat{\pi}^{\tau}$ est une estimation de $E(Y^{\tau})$. Si nous utilisons les données de l'EPA, cette estimation représente la fréquence pondérée de l'événement à la période τ (en prenant chaque coefficient de pondération du mois pour lequel on observe un enfant pour la première fois). Pour justifier cette méthode, nous utilisons l'équation (3.4) et supposons une approximation

$$E\{g(\pi^{\tau}(X^{\tau}))\} - E\{g(\pi^{\tau_0}(X^{\tau_0}))\} \cong g(E\{\pi^{\tau}(X^{\tau})\}) - g(E\{\pi^{\tau_0}(X^{\tau_0})\}) \quad (3.6)$$

En raison de l'inégalité de Jensen, les résultats seront inexacts dans les régions où g est convexe ou concave. Néanmoins, si, localement, g s'approche d'une fonction linéaire dans les régions où $\pi^{\tau}(X^{\tau})$ et $\pi^{\tau_0}(X^{\tau_0})$ sont concentrées, alors (3.6) pourrait être relativement exacte. Le fait que g possède un point d'inflexion à 0,5 pourrait faciliter l'approximation lorsque les probabilités sont dispersées autour de cette valeur.

Heureusement, nous pouvons tester l'adéquation de l'estimateur par simulation du modèle estimé dans LifePaths et comparaison des fréquences prévues de l'événement aux fréquences pondérées correspondantes calculées d'après les données. Les résultats indiquent que le modèle est suffisamment adéquat pour notre application.

3.2 Application à l'analyse de survie

Nous montrerons à la section 5.2 que la méthode décrite plus haut peut être étendue au modèle semiparamétrique de survie grâce à l'ajout d'un indice t représentant la durée dans l'état courant, si bien que (3.5), devient

$$\hat{\gamma}^{\tau}(t) = \hat{\gamma}^{\tau_0}(t) + g(\hat{\pi}^{\tau}(t)) - g(\hat{\pi}^{\tau_0}(t)) \quad (3.7)$$

où $\hat{\pi}^{\tau}(t)$ représente la fonction de hasard empirirque.

3.3 Tendances qui caractérisent les variables indépendantes

Nous pouvons améliorer la méthode d'étalonnage en tenant compte de l'évolution des caractéristiques observées. Comme nous l'avons mentionné à la section 2.3, cette correction serait envisagée si d'autres éléments du modèle LifePaths étaient à un stade de développement plus poussé. Pour cela, nous relâchons l'hypothèse selon laquelle les vecteurs aléatoires X^{τ} sont identiquement distribués. L'équation (3.4) devient alors

$$\begin{aligned} E\{\eta^{\tau}(X^{\tau}) - \eta^{\tau_0}(X^{\tau_0})\} &= \gamma^{\tau} - \gamma^{\tau_0} + \beta'\{E(X^{\tau}) - E(X^{\tau_0})\} \\ &= E\{g(\pi^{\tau}(X^{\tau}))\} \\ &\quad - E\{g(\pi^{\tau_0}(X^{\tau_0}))\} \end{aligned} \quad (3.8)$$

D'après cette équation, nous pourrions estimer γ^{τ} par

$$\hat{\gamma}^{\tau} = \hat{\gamma}^{\tau_0} + g(\hat{\pi}^{\tau}) - g(\hat{\pi}^{\tau_0}) - \hat{\beta}'(\bar{x}^{\tau} - \bar{x}^{\tau_0}) \quad (3.9)$$

où $\bar{x}^{\tau}$ est le vecteur des valeurs moyennes des caractéristiques à la période τ. Naturellement, il pourrait être impossible d'obtenir toutes les valeurs moyennes à partir de la même source de données. La méthode s'étendrait au modèle de survie de la même façon que (3.7) pour donner

$$\hat{\gamma}^{\tau}(t) = \hat{\gamma}^{\tau_0}(t) + g(\hat{\pi}^{\tau}(t)) - g(\hat{\pi}^{\tau_0}(t)) - \hat{\beta}'(\bar{x}^{\tau}(t) - (\bar{x}^{\tau_0}(t)). \quad (3.10)$$

4. MODÈLES ET ESTIMATION DES PARAMÈTRES DE BASE

Comme nous l'expliquons à la section 3.1, les paramètres de base $\hat{\beta}$ et $\hat{\gamma}^{\tau_0}$ sont estimés par la méthode du maximum de vraisemblance au moyen de données recueillies pour la période τ_0. Nous utilisons les données de l'EDTR sur les congés de maternité recueillies à compter de la période allant de 1993 à 1996 (notre période de référence τ_0). Nous n'essayons pas d'estimer la variation annuelle du terme constant γ durant cette période.

4.1 Modèle logit binaire

Nous adoptons le modèle logit pour représenter le choix fait par une mère entre le congé de maternité et le retrait du marché du travail. À partir d'ici, nous adoptons une notation économétrique plus conventionnelle et utilisons l'indice i pour représenter une variable aléatoire ou un résultat associé à un individu i. Nous supposons que la variable aléatoire Y_i^{τ} prend la valeur 0 ou 1, $Y_i^{\tau} = 1$ indiquant que la nouvelle mère i à laquelle est associé le vecteur de caractéristiques x_i à la période τ choisit de prendre un congé de maternité, à condition qu'elle ait été employée, et que

$$\pi_i^{\tau} = P(Y_i^{\tau} = 1) = F(\eta_i^{\tau}) = \frac{\exp(\eta_i^{\tau})}{1 + \exp(\eta_i^{\tau})} \quad (4.1)$$

où $\eta_i^{\tau} = \beta' x_i + \gamma^{\tau}$ est le prédicteur linéaire de l'équation (3.1) et F est la fonction de distribution logistique. Nous estimons les paramètres de base $\hat{\beta}$ et $\hat{\gamma}^{\tau_0}$ au moyen de N observations provenant de l'EDTR en maximisant le logarithme du rapport de vraisemblance $\ln L(\beta, \gamma^{\tau_0})$ où

$$\begin{aligned} L(\beta, \gamma^{\tau}) &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) \\ &= \prod_{y_i = 0} [1 - F(\eta_i^{\tau})] \prod_{y_i = 1} F(\eta_i^{\tau}) \\ &= \prod_i [F(\eta_i^{\tau})]^{y_i} [1 - F(\eta_i^{\tau})]^{1 - y_i} \end{aligned} \quad (4.2)$$

et

$$\ln L(\beta, \gamma^{\tau}) = \sum_i \{y_i \ln F(\eta_i^{\tau}) + (1 - y_i) \ln [1 - F(\eta_i^{\tau})]\}. \quad (4.3)$$

Les coefficients de pondération longitudinaux de l'EDTR calculés pour l'année de la naissance de l'enfant sont mis à l'échelle de sorte que leur somme soit égale à l'effectif de l'échantillon, puis utilisés pour pondérer les termes du logarithme du rapport de vraisemblance et de ses dérivées.

Les équations pondérées de caractérisation (équations de score) sont

$$\frac{\partial L(\beta,\gamma^\tau)}{\partial \beta} = \sum_i w_i x_i y_i - \sum_i w_i x_i F(\eta_i^\tau) = 0$$

$$\frac{\partial L(\beta,\gamma^\tau)}{\partial \gamma^\tau} = \sum_i w_i y_i - \sum_i w_i F(\eta_i^\tau) = 0. \qquad (4.4)$$

Nous avons obtenu la solution, qui maximise le logarithme du rapport de vraisemblance, par itération de Newton-Raphson. Le modèle logit, qui est utilisé fréquemment par les statisticiens et les économétriciens, a fait l'objet de nombreux articles. Par exemple, consulter Chambless et Boyle (1985), Roberts, Rao et Kumar (1987) et Morel (1989).

4.2 Modèle semiparamétrique de survie : Forme élémentaire

Dans le cas où la mère choisit de prendre un congé de maternité, nous utilisons un modèle de survie pour décrire la durée du congé. La densité de probabilité de la distribution a une forme complexe, comme l'illustre les graphiques de la section 6.4. On observe un pic pour les durées inférieures à un mois et le mode semble correspondre au montant maximal des prestations spéciales d'assurance-chômage dont ont bénéficié les mères après 1990 (15 semaines de prestations de maternité, plus 10 semaines de prestations parentales, plus une période d'attente de deux semaines). Pour commencer, nous avons estimé divers modèles entièrement paramétriques, y compris un modèle log-logistique de survie combiné à un modèle logit pour prédire les durées inférieures à un mois, mais nous n'avons pu obtenir un ajustement convenable. Pour résoudre ce problème, nous nous inspirons de Prentice et Gloeckler (1978), de Han et Hausman (1986) et de Meyer (1990) et nous estimons de façon non paramétrique les effets du temps sur la fonction de hasard de retour au travail. Cette dernière est exprimée sous forme d'une équation à hasards proportionnels :

$$\lambda_i^\tau(t) = \lambda_0^\tau(t) \exp\{\beta' x_i(t)\} \qquad (4.5)$$

où $\lambda_0^\tau(t)$ est la fonction de hasard de base, inconnue pour la durée t du congé et la période τ, $x_i(t)$ est un vecteur des variables explicatives pour la mère i, et β est un vecteur de coefficient. Les données nous indiquent lequel des intervalles [0,1), [1,2), [2,3), ... contient la durée du congé (dans notre cas l'unité est le mois) et le modèle peut être interprété comme un modèle de hasard en temps continu incomplètement observé sans restriction quant à la forme de la fonction de hasard de base. Si T_i^τ est la durée du congé de la mère i à la période τ, alors, pour $t = 1, 2, 3, \ldots$, la probabilité que le congé dure jusqu'à la période t, étant donné qu'il a duré jusqu'à la période $t - 1$, peut être écrite sous la forme

$$P(T_i^\tau > t \mid T_i^\tau \geq t-1) = \exp\left[-\int_{t-1}^{t} \lambda_i^\tau(u)\,du\right]$$

$$= \exp\left[-\exp\{\beta' x_i(t)\} \int_{t-1}^{t} \lambda_0^\tau(u)\,du\right] \qquad (4.6)$$

si nous supposons que $x_i(t)$ est constante sur l'intervalle entre $t - 1$ et t. Afin d'appliquer la théorie de la section 3, nous pouvons réécrire l'équation (4.6) sous la forme

$$1 - \pi_i^\tau(t) = P(T_i^\tau \geq t \mid T_i^\tau \geq t-1)$$

$$= \exp[-\exp\{\beta' x_i(t) + \gamma^\tau(t)\}] \qquad (4.7)$$

$$= \exp[-\exp\{\eta_i^\tau(t)\}]$$

où

$$\gamma^\tau(t) = \ln\left[\int_{t-1}^{t} \lambda_0^\tau(u)\,du\right]. \qquad (4.8)$$

Nous pouvons censurer toute observation courante à partir d'une valeur élevée T de la durée. De nouveau, nous pouvons estimer les paramètres de base $\hat{\beta}$ et $\hat{\gamma}^{\tau_0}$ au moyen de N observations provenant de l'EDTR par maximisation du logarithme du rapport de vraisemblance $\ln L(\gamma^{\tau_0}, \beta)$. Comme nous nous référons systématiquement à des données recueillies pour la période de référence, dans la suite de la section 4, nous abandonnons les indices τ_0.

La fonction de vraisemblance est donnée par

$$L(\gamma,\beta) = \prod_{i=1}^{N} [[1 - \exp\{-\exp(\eta_i(k_i))\}]^{\delta_i} \prod_{t=1}^{k_i} \exp\{-\exp(\eta_i(t))\}] \qquad (4.9)$$

où $\gamma = [\gamma(1), \gamma(2), \ldots, \gamma(T)]'$, C_i est une période de censuration, $\delta_i = 1$ si $T_i \leq C_i$ et 0 autrement, $k_i = \min(\mathrm{int}(T_i), C_i)$. Par conséquent, le logarithme du rapport de vraisemblance est donné par

$$\ln L(\gamma,\beta) = \sum_{i=1}^{N} [\delta_i \ln[1 - \exp\{-\exp(\eta_i(k_i))\}] - \sum_{t=1}^{k_i} \exp(\eta_i(t))]. \qquad (4.10)$$

Les coefficients de pondération applicables aux mois durant lesquels un enfant est observé pour la première fois sont mis à l'échelle de sorte que leur somme corresponde à l'effectif de l'échantillon, puis utilisés pour pondérer les termes de la fonction de vraisemblance et ses dérivées. La fonction pondérée de vraisemblance est maximisée au moyen du quasi-algorithme de Newton proposé par Broyden, Fletcher, Goldfarb et Shanno (BFGS), selon une méthode d'application fondée sur Dennis et Schnabel (1983).

4.3 Modèle semiparamétrique de survie : Avec décision concernant l'intervalle arrêt du travail-naissance

Dans notre application, la situation est compliquée du fait que nous voulons modéliser la durée entre l'arrêt du travail et la naissance (intervalle arrêt de travail-naissance), ainsi que la fonction de hasard du retour au travail après un congé de maternité. Le modèle de l'intervalle arrêt du travail-naissance est estimé séparément, d'après des données de l'EDTR. L'examen de l'intervalle moyen pour chaque année de données de l'EPA indique que la durée de cet intervalle est restée assez stable au fil du temps, si bien que nous n'étalonnons pas le modèle. Néanmoins, il est nécessaire d'apporter une modification au modèle semi-paramétrique de survie afin d'intégrer le modèle distinct d'intervalle arrêt du travail-naissance. Nous pouvons pour cela supposer que la décision concernant l'intervalle arrêt du travail-naissance, qui tient éventuellement compte de l'état de santé, agit comme une contrainte sur la durée totale souhaitée du congé. Autrement dit, le modèle susmentionné s'appliquerait à la durée totale souhaitée, que l'on ne peut observer, et pourrait être appelée T^*.

Dans les cas où la durée souhaitée est plus courte que l'intervalle arrêt du travail-naissance, la mère pourrait retourner au travail aussitôt que possible après la naissance. Autrement dit, dans les cas où nous observons un intervalle arrêt de travail-naissance important (supérieur à un mois) et que la mère retourne au travail peu de temps après la naissance (dans le mois qui suit), tout ce que nous savons au sujet de la durée souhaitée est que

$$T^* \leq T$$

où T représente la durée totale du congé. Ceci équivaut à la situation appelée « censuration à gauche » par Cox et Oaks (1984, page 178), où l'observation ne débute pas immédiatement et où certains individus sont déjà éliminés avant que l'observation ne commence.

Pour ce genre d'observation, nous obtenons la contribution à la fonction de vraisemblance et à son logarithme par

$$\begin{aligned} L_i = 1 - \prod_{t=1}^{k_i} P(T^* \geq t \mid T^* \geq t-1) = 1 \\ - \prod_{t=1}^{k_i} \exp[-\exp(\eta_i(t))] \end{aligned} \tag{4.11}$$

et

$$\ln(L_i) = \ln\{1 - \exp[-\sum_{t=1}^{k_i} \exp(\eta_i(t))]\}. \tag{4.12}$$

Malheureusement, l'expression du logarithme du rapport de vraisemblance ne se simplifie pas comme l'expression correspondante pour les observations « censurées à droite ». Malgré cela, les expériences de Monte Carlo indiquent que l'estimation ne pose pas de difficulté même si les ensembles de données sont fortement censurés.

Nous appliquons les coefficients de pondération longitudinaux de l'EDTR pour l'année de la naissance de l'enfant de la même façon que dans le cas de la forme élémentaire du modèle de survie.

5. ÉTALONNAGE DES MODÈLES

Pour entamer la procédure d'étalonnage, nous devons inverser la fonction de distribution F donnée par l'équation (3.2) afin d'obtenir la fonction de lien g. Puis, nous appliquons l'équation (3.5) dans le cas du modèle logit et l'équation (3.7) dans le cas du modèle de survie.

5.1 Application du modèle logit binaire

Pour étalonner le modèle logit, nous devons d'abord inverser la fonction de distribution logistique dans l'équation (4.1) afin d'obtenir

$$\eta_i^\tau = g(\pi_i^\tau) = \ln\left(\frac{\pi_i^\tau}{1-\pi_i^\tau}\right) \tag{5.1}$$

où g est la fonction logit bien connue. Nous pouvons alors appliquer les équations (3.5) et (5.1) pour obtenir

$$\hat{\gamma}^\tau = \hat{\gamma}^{\tau_0} + g(\hat{\pi}^\tau) - g(\hat{\pi}^{\tau_0}) = \hat{\gamma}^{\tau_0} + \ln\left(\frac{\hat{\pi}^\tau/(1-\hat{\pi}^\tau)}{\hat{\pi}^{\tau_0}/(1-\hat{\pi}^{\tau_0})}\right) \tag{5.2}$$

où, pour $\tau < \tau_0$, chaque $\hat{\pi}^\tau$ représente la fréquence à laquelle est choisi le congé de maternité calculée d'après les données de l'EPA pour les congés de maternité commençant dans l'année τ, et $\hat{\pi}^{\tau_0}$ représente la fréquence calculée d'après les données de l'EDTR.

5.2 Extension au modèle de survie

Partant de l'équation (4.7) nous obtenons

$$\pi_i^\tau(t) = 1 - \exp[-\exp\{\eta_i^\tau(t)\}] = F\{\eta_i^\tau(t)\} \tag{5.3}$$

où

$$\eta_i^\tau(t) = \beta' x_i(t) + \gamma^\tau(t). \tag{5.4}$$

Ici, F est une distribution de valeurs extrêmes qu'il est facile d'inverser pour obtenir

$$\eta_i^\tau(t) = \ln[-\ln(1-\pi_i^\tau(t))] = g(\pi_i^\tau(t)). \tag{5.5}$$

Pour l'étalonnage, nous pouvons utiliser l'équation (3.7) ainsi que les fréquences observées pour l'année τ représentée par le hasard empirique, ou ratio survenue/exposition, donné par

$$\hat{\pi}^\tau(t) = d^\tau(t) \big/ r^\tau(t) \tag{5.6}$$

où, pour les congés commençant à l'instant τ, $d^\tau(t)$ représente le nombre de mères dont le statut a changé dans l'intervalle $(t-1, t]$ et $r^\tau(t)$ représente le nombre de mères observées pour la durée t, y compris celles éliminées par

censuration à l'instant t (la censuration ne peut avoir lieu qu'à la fin des intervalles). Le nombre de mères a été calculé d'après des dénombrements sur échantillon en appliquant le coefficient de pondération de l'EPA pour le mois durant lequel une nouvelle mère retourne au travail. La fonction de hasard empirique et l'estimateur correspondant de la fonction de survie découlant de la loi du produit des probabilités ont été étudiés par Kaplan et Meier (1958). L'utilisation de la fonction de hasard empirique dans l'équation (3.7) simultanément à l'équation (5.5) donne

$$\hat{\gamma}^{\tau}(t) = \hat{\gamma}^{\tau_0}(t) + \ln\left(\frac{\ln[1 - \hat{\pi}^{\tau}(t)]}{\ln[1 - \hat{\pi}^{\tau_0}(t)]}\right). \qquad (5.7)$$

6. RÉSULTATS EMPIRIQUES

Nous présentons les résultats de l'estimation pour la période de référence et les résultats de la simulation basée sur les estimations étalonnées des paramètres. Nous comparons les résultats de la simulation aux fréquences annuelles, calculées d'après les données d'enquête, du choix du congé de maternité dans le cas du modèle logit et aux distributions des fréquences annuelles, calculées d'après les données d'enquête, des durées du congé de maternité dans le cas du modèle de survie.

6.1 Estimations dans le cas du modèle logit binaire

Les estimations obtenues par application du modèle logit aux données de l'EDTR sont présentées au tableau 1. Y sont omises les variables indicatrices formant les catégories de référence pour les variables utilisées dans le modèle, c'est-à-dire « Ontario » pour la province de résidence et « Certaines études postsecondaires » pour le niveau le plus élevé de scolarité. Nous avons testé les variables de revenu personnel et de revenu familial et constaté que leur effet n'était pas significatif, si bien que nous ne les avons pas incluses dans la régression.

Les estimations peuvent être entachées d'un certain biais, particulièrement celles de l'erreur-type, parce qu'on n'a tenu compte du plan d'échantillonnage complexe de l'EDTR qu'au moyen des coefficients de pondération appliqués au logarithme du rapport de vraisemblance.

L'effet positif significatif de la durée de l'emploi semble raisonnable pour plusieurs raisons. Une durée importante d'occupation d'un emploi pourrait indiquer que la femme concernée a accumulé un capital humain propre à l'entreprise et a atteint un certain niveau d'ancienneté. Elle pourrait aussi indiquer un attachement puissant au marché du travail en général. Par ailleurs, le congé que l'entreprise accordera à la femme avec la garantie qu'elle retrouve son emploi par la suite sera d'autant plus long que l'employée aura de l'ancienneté. En outre, les garanties d'emploi offertes par les gouvernements provinciaux dépendent aussi de l'ancienneté professionnelle. Enfin, une grande ancienneté professionnelle signifie que la femme répondra aux exigences d'admissibilité aux prestations d'assurance-chômage (20 semaines d'emploi assuré). Nous avons testé dans le modèle une variable nominale indiquant que les critères d'admissibilité aux prestations d'assurance-chômage étaient satisfaits et constaté que son effet était à peine significatif au niveau de 5 %. Toutefois, nous n'avons pas inclus la variable dans le modèle, parce qu'à ce stade, nous ne sommes pas capables de modéliser les variations du programme d'assurance-chômage sous l'influence de covariables, à cause de l'incertitude de l'interprétation et de la forte corrélation avec la durée de l'emploi. Dans le cas de l'EPA, les travailleurs autonomes ne sont considérés comme ayant quitté le marché du travail que lorsqu'ils mettent fin à leur entreprise. Puisque prendre un congé signifie simplement ne pas mettre fin à l'entreprise, on devrait s'attendre à un effet positif significatif de la variable de travail autonome. Le fait d'avoir travaillé pour son propre compte avant la naissance augmente de 333 % la cote exprimant la chance de prendre un congé de maternité, ce qui représente l'effet le plus important observé pour une variable explicative.

Tableau 1
Estimation des paramètres du modèle logit binaire

Paramètre	Estimation du coefficient	Contribution au rapport de cotes*	Erreur-type du coefficient	Valeur *p*
Constante	-6,432	0,002	2,995	0,0318
T.-N.	-0,829	0,436	0,741	0,2636
Î.-P.-É.	0,931	2,537	1,612	0,5633
N.-É.	-0,456	0,634	0,541	0,3992
N.-B.	0,207	1,230	0,675	0,7596
Qc	-0,361	0,697	0,247	0,1437
Man.	-0,490	0,613	0,503	0,3306
Sask.	-0,163	0,850	0,458	0,7218
Alb.	-0,200	0,819	0,325	0,5379
C.-B.	-0,120	0,887	0,300	0,6899
Durée de l'emploi (mois)/10	0,094	1,099	0,026	0,0003
Travail autonome?	1,203	3,330	0,418	0,0040
Âge (années)	0,479	1,614	0,199	0,0160
(Âge^2)/10	-0,071	0,931	0,033	0,0296
< Diplôme d'études secondaires	-0,702	0,496	0,357	0,0490
Diplôme d'études secondaires	-0,148	0,862	0,276	0,5913
Diplôme universitaire	-0,292	0,747	0,229	0,2027
Premier enfant?	-0,525	0,592	0,192	0,0063

Log du rapport de vraisemblance = -381,553
Nombre d'observations = 835
Les coefficients de pondération longitudinaux de l'EDTR pour l'année de la naissance, mis à l'échelle de sorte que leur somme soit égale à l'effectif de l'échantillon, sont appliqués aux observations

* Il s'agit de l'exponentielle du coefficient. Elle peut être interprétée comme étant la variation proportionnelle du rapport de cotes due à une variation unitaire de la variable indépendante correspondante.

L'effet de la variable de premier enfant semble également raisonnable. Pour une femme qui donne naissance à son premier enfant, la cote exprimant la possibilité de prendre un congé de maternité ne représente que 59 % de celle observée pour une femme qui a déjà des enfants, toutes les autres caractéristiques étant par ailleurs égales; autrement dit, les femmes qui deviennent mère pour la première fois sont plus susceptibles de cesser de travailler que celles qui ont déjà eu des enfants. Ces résultats pourraient tenir au fait que notre échantillon comprend les mères qui travaillaient dans les quatre mois avant et après la naissance. Les femmes qui ont plus d'un enfant ont tendance à espacer les naissances d'au moins quelques années. Si elles sont employées juste avant une deuxième naissance ou avant les naissances subséquentes, elles auront déjà prouvé qu'elles sont retournées au travail après une absence dont la durée doit avoir été inférieure à l'intervalle entre les naissances. Cette observation permet d'exclure une tendance commune au retrait du marché du travail – par exemple, rester à la maison jusqu'à ce que les enfants aillent à l'école.

L'influence de l'âge est plus difficile à interpréter, puisque l'effet de celui-ci sur le logarithme du rapport de cotes n'est pas linéaire. Si nous traçons le graphique du terme $-0{,}479 * \hat{age} - 0{,}0071 * \hat{age}^2$, nous constatons que, à mesure que l'âge augmente, le logarithme de la cote exprimant la chance de prendre un congé de maternité commence par augmenter, mais que le taux de croissance diminue jusqu'à un plafond où le maximum du logarithme de la cote est atteint, à l'âge de 34 ans. Puisque le nombre de femmes qui donnent naissance à des enfants diminuent considérablement après cet âge, la diminution subséquente du logarithme pourrait ne pas être significative. Nous pourrions conjecturer que, chez les jeunes mères, le fait d'être, relativement parlant, plus âgée indique un attachement plus prononcé au marché du travail et, donc, une plus forte tendance à prendre un congé de maternité, alors que chez les mères plus âgées, qui ont passé l'étape de la première entrée sur le marché du travail, cet effet est moins prononcé. Cependant, les résultats ne sont sans doute pas suffisamment précis pour tirer des conclusions catégoriques à ce sujet.

6.2 Résultats des simulations pour le modèle logit binaire étalonné

L'exercice d'étalonnage consiste à ajuster le terme constant du modèle de la façon décrite par (5.2) pour chaque année, de 1975 à 1992. Nous n'ajustons pas le terme constant après 1992, en partie parce que les données de l'EPA ne révèlent aucune tendance marquée après 1992. Ensuite, nous intégrons le modèle dans LifePaths et exécutons une simulation. Pour chaque année, de 1976 à 1995, la figure 1 montre la fréquence à laquelle est choisi le congé de maternité dans la simulation LifePaths et la fréquence correspondante estimée d'après les données de l'EPA. Pour la période allant de 1993 à 1995, nous présentons aussi les estimations calculées d'après les données de l'EDTR.

La simulation reflète la variation au fil du temps qui se dégage des données de l'EPA pour la période allant de 1976 à 1992. La simulation LifePaths ne comporte aucune correction d'étalonnage après 1992, de sorte que les paramètres de base estimés d'après les données regroupées de l'EDTR pour 1993 à 1996 soient efficaces. La fréquence simulée est légèrement inférieure à la fréquence observée dans le cadre de l'EDTR durant cette période. Les deux sources possible d'erreurs sont le manque de souplesse de la spécification du modèle de choix binaire et les différences entre les estimations des variables explicatives basées sur les données de l'EDTR et celles fournies par le modèle LifePaths.

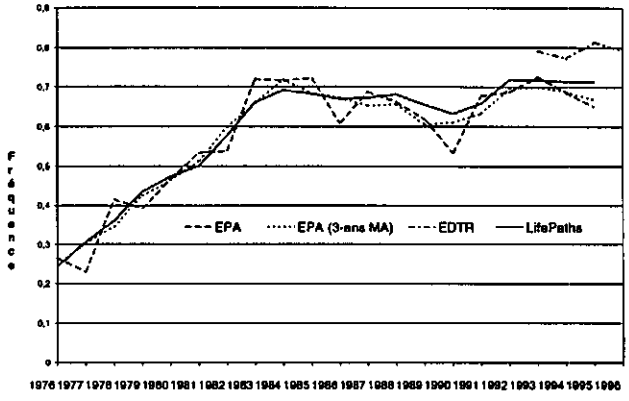


Figure 1. Fréquence du choix du congé de maternité, 1976 à 1996

6.3 Estimations pour le modèle de survie

Les estimations obtenues par application du modèle semiparamétrique de survie aux données de l'EDTR sont présentées au tableau 2. Comme dans le cas de l'estimation au moyen du modèle logit binaire, sont omises les variables indicatrices servant de catégorie de référence, c'est-à-dire « Ontario » pour la province de résidence et « Certaines études postsecondaires » pour le plus haut niveau de scolarité. Puisque la variable dépendante est la probabilité de retourner au travail, un coefficient de covariable positif indique une influence qui a tendance à raccourcir la durée du congé de maternité.

Les estimations des termes constants dans le prédicteur linéaire dépendant de la durée donné par (4.7) sont représentées dans le tableau 2 par GAMMAi, $i = 1, 2, \ldots, 15$. Ces estimations correspondent à l'effet de la fonction de hasard de base à laquelle est intégré l'effet de la durée.

De nouveau, nous avons testé les variables de revenu personnel et de revenu familial et constaté que leur effet n'était pas significatif. Ces deux résultats, ainsi que l'importance de la variable de travail autonome en tant que prédicteur d'un retour précoce au travail concordent avec les résultats Marshall (1999). Selon ce dernier, les variables

de niveau de scolarité n'ont aucun effet significatif sur le fait qu'une mère retourne au travail dans le mois. Toutefois, nous constatons que le fait de posséder un diplôme universitaire a un effet négatif significatif sur la fonction de hasard (effet positif sur la durée). La durée de l'emploi a un effet négatif significatif sur la fonction de hasard, ce qui reflète peut-être le lien entre cette variable et le droit à l'assurance-chômage ainsi que la sécurité de l'emploi.

Tableau 2
Estimations des paramètres du modèle de survie

Paramètre	Estimation	Erreur-type	Valeur *p*
Durée de l'emploi (mois) /10	-0,030	0,010	0,0024
T.-N.	0,195	0,426	0,6470
I.P.E.	0,307	0,490	0,5313
N.-E.	0,173	0,253	0,4940
N.-B.	0,109	0,293	0,7091
QUÉ	0,111	0,117	0,3411
MAN	-0,402	0,253	0,1116
SASK	-0,303	0,213	0,1539
ALB	0,270	0,154	0,0798
C.B.	-0,440	0,148	0,0030
Travail autonome?	1,665	0,157	0,0000
Âge	-0,253	0,041	0,0000
Âge** 2 / 10	0,043	0,007	0,0000
Premier enfant?	-0,301	0,090	0,0009
< Diplôme d'études secondaires	0,508	0,206	0,0135
Diplôme d'études secondaires	-0,124	0,125	0,3212
Diplôme universitaire	-0,374	0,108	0,0006
Conjoint employé?	0,109	0,151	0,4703
Gamma1	2,570	0,609	0,0000
Gamma2	-1,136	0,816	0,1636
Gamma3	-0,466	0,719	0,5176
Gamma4	0,78	0,64	0,2232
Gamme5	1,425	0,627	0,0231
Gamma6	2,755	0,613	0,0000
Gamma7	3,64	0,612	0,0000
Gamma8	3,413	0,62	0,0000
Gamma9	3,465	0,63	0,0000
Gamma10	3,387	0,649	0,0000
Gamma11	4,579	0,655	0,0000
Gamma12	4,285	0,785	0,0000
Gamma13	3,645	1,110	0,0010
Gamma14	3,746	1,281	0,0034
Gamma15	6,215	2,415	0,0101
Log du rapport de varisemblance = -1165,06			
Nombres d'observations = 3 411			

Les coefficients de pondération longitudinaux de l'EDTR pour l'année de la naissance, mis à l'échelle de sorte que leur somme soit égale à l'effectif de l'échantillon, sont appliqués aux observations.

6.4 Résultats de la simulation pour le modèle étalonné de survie

Dans le cas du modèle semiparamétrique de survie, l'étalonnage consiste à rajuster tous les termes GAMMA*i*, *i* = 1, 2, ..., 15 de la section précédente conformément à (5.8) pour chacune des années de la période allant de 1975 à 1992. Nous intégrons le modèle dans LifePaths et exécutons la simulation.

Nous présentons la distribution des fréquences des durées simulées des congés de maternité et la comparons à la distribution correspondante des fréquences observées d'après les données de l'EPA. Aux fins de la présentation des résultats, nous avons calculé la moyenne des fréquences sur trois ans. Un aspect important de la courbe de distribution des fréquences est la variation abrupte apparemment liée à l'introduction des prestations parentales au moment de la mise en application du projet de loi C-21 à la fin de 1990. Comme les mères ayant présenté une demande de prestations pour maternité à l'époque de la mise en application de la loi avaient droit aux prestations parentales, les demandes enregistrées au début de 1990 représentent un mélange de régimes de prestations. Par conséquent, nous n'avons inclus l'année 1990 dans aucune des moyennes sur trois ans. Aux figures 2 et 3, nous utilisons des périodes de trois ans disjointes couvrant la période de 1976 à 1984. Pour équilibrer les périodes étudiées avant et après 1990 en nous servant des données disponibles, nous utilisons dans les figures 4 et 5 les périodes chevauchantes allant de 1985 à 1987, de 1987 à 1989, de 1991 à 1993 et de 1993 à 1995.

La distribution des durées des congés obtenue d'après les données de l'EDTR recueillies pour la période de 1993 à 1996 est présentée à la figure 6. Cette distribution peut être comparée aux données simulées présentées à la figure 4 pour la période allant de 1993 à 1995, puisqu'aucun étalonnage n'a été appliqué après 1992.

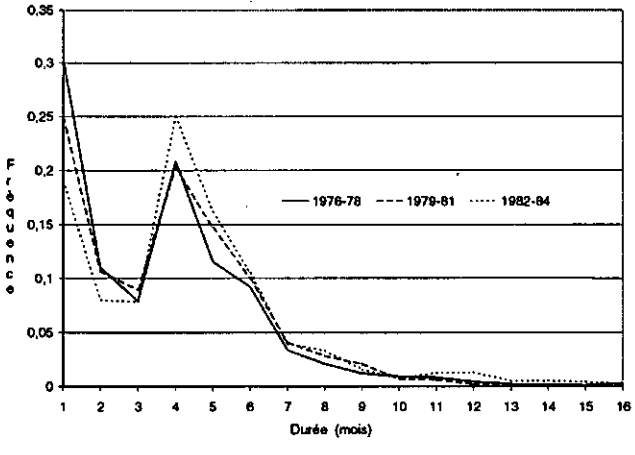


Figure 2. LifePaths : Distribution des durées des congés pour 1976 à 1984

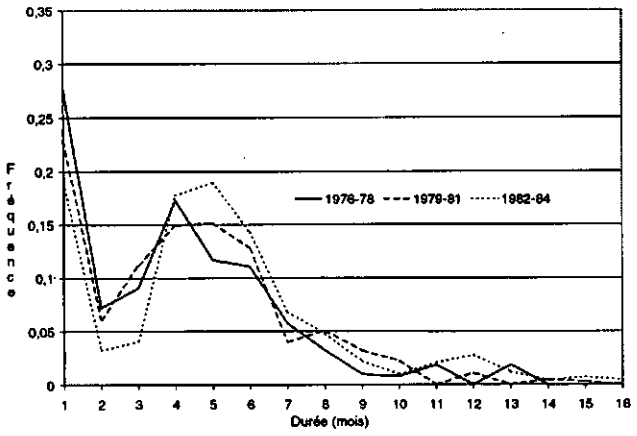


Figure 3. Données de l'EPA : Distribution des durées des congés pour 1976 à 1984

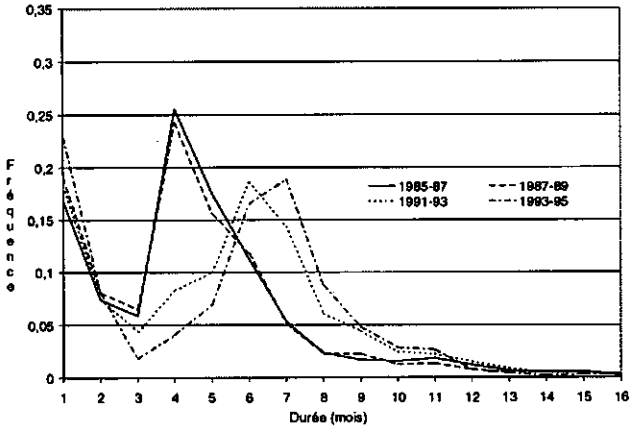


Figure 4. LifePaths : Distribution des durées des congés pour 1985 à 1989 et 1991 à 1995

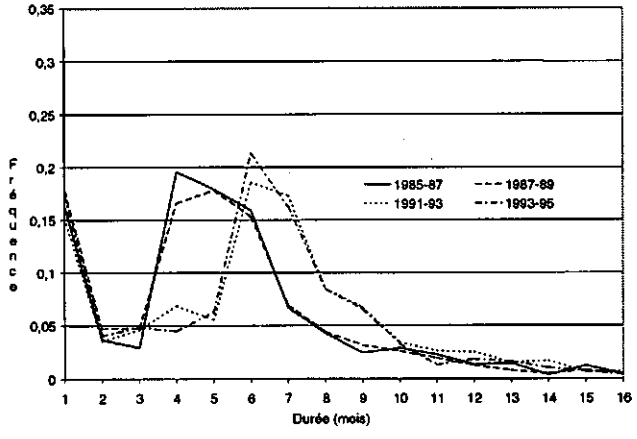


Figure 5. Données de l'EPA : Distribution des durées des congés pour 1985 à 1989 et 1991 à 1995

À la figure 7, nous présentons la durée moyenne des congés de maternité débutant chaque année de la période observée. Nous comparons la moyenne des durées simulées à celles des durées calculées d'après les données d'enquête.

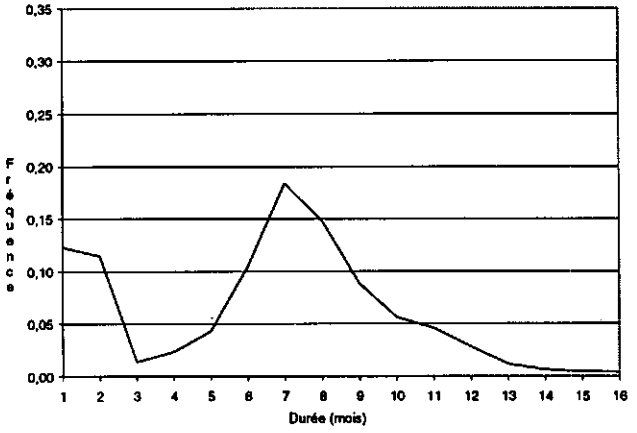


Figure 6. Données de l'EDTR : Distribution des durées des congés pour 1993 à 1996

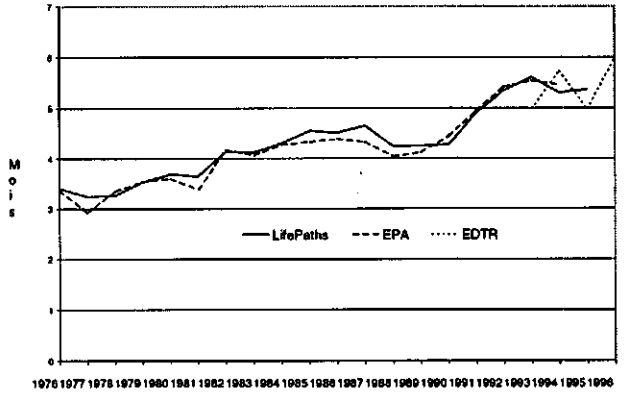


Figure 7. Durée moyenne du congé de maternité, 1976 à 1996

6.5 Évaluation du rendement de l'étalonnage

La méthode d'étalonnage semble être très efficace dans le cas du modèle logit binaire. La tendance qui se dégage des données de l'EPA est bien reflétée par la simulation LifePaths. Dans le cas du modèle de survie, l'aspect clé des données de l'EPA est le déplacement abrupt du mode de la distribution des fréquences après 1990, déplacement qui semble être dû à l'introduction des prestations parentales. Les données simulées reflètent ce décalage. En outre, la durée moyenne du congé de maternité établie d'après la simulation concorde très étroitement avec les données de l'EPA.

Une divergence notable entre la simulation et les données de l'EPA est la hauteur du mode dans l'intervalle (3,4] mois dans la distribution des fréquences des durées établies d'après la simulation LifePaths pour la période de 1982 à 1989. Cette divergence pourrait être due à l'effet de tendances caractérisant les valeurs des variables explicatives, variables que nous avons considérées comme étant stables. D'autres travaux seront nécessaires pour établir ce fait. Nous avons discuté d'une extension possible du modèle à la section 3.3.

7. CONCLUSIONS

La technique que nous avons mise au point semble permettre d'étalonner les paramètres du modèle logit et du modèle de survie de façon à ce que les caractéristiques essentielles des données de l'EPA soient reflétées par les prévisions du modèle LifePaths. L'élément clé de l'étalonnage du modèle logit est l'ajustement du paramètre correspondant aux « termes constants » dans le prédicteur linéaire qui est intégré dans la fonction de distribution logistique afin de prédire l'espérance conditionnelle de la variable dépendante. À la section 3.1, nous élaborons la technique dans un cadre général englobant d'autres modèles de choix binaires. Cette technique pourrait notamment être étendue au modèle probit bien connu où un prédicteur linéaire est intégré dans la fonction de distribution normale type. L'étalonnage du modèle semiparamétrique de survie repose sur l'ajustement de tous les paramètres représentant la fonction de hasard de base. Nos résultats illustrent comment il est possible de faire évoluer en fonction du temps la forme entière de la distribution des durées prévues par le modèle selon un profil révélé par des données supplémentaires.

REMERCIEMENTS

Les auteurs remercient Steve Gribble et les membres du Groupe de la modélisation socioéconomique de Statistique Canada pour leurs commentaires utiles durant le développement du module sur le congé de maternité, Geoff Rowe et Huan Nguyen pour l'utilisation de leur programme informatique afin de suivre les individus durant leur participation à l'échantillon de l'EPA, Katherine Marshall pour ses conseils sur l'utilisation des données de l'EDTR et pour le partage de ses programmes informatiques, Adrienne ten Cate pour les discussions fructueuses qu'ils ont eues avec elle et un évaluateur anonyme pour plusieurs améliorations proposées. Ces travaux ont été réalisés à l'époque où les deux auteurs travaillaient pour le Groupe de la modélisation socioéconomique, Statistique Canada, immeuble R.-H.-Coats, 24e étage, Parc Tunney

BIBLIOGRAPHIE

APPLEBY, J., BOOTHBY, D., ROULEAU, M. et ROWE, G. (1999). Level and Distribution of Individual Returns to Post-Secondary Education: Simulation Results from the LifePaths Model. Présenté à la réunion de 1999 de Canadian Economics Association.

CHAMBLESS, L.E., et BOYLE, K.E. (1985). Maximum Likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics, A: Theory and Methods*, 14, 177-192.

CHEN, E.J., et ODERKIRK, J. (1997). Varied Pathways: The Undergraduate Experience in Ontario, Feature article. *Education Quarterly Review*, Statistique Canada, 4, 3, 47-62.

CHEN, E.J., et ROWE, G. (1999). Trend Correlation of Labour Market Earnings in Canada: 1982 to 1995. *Statistical Society of Canada 1999 Proceedings of the Survey Methods Section*, 173-179.

COX, D.R., et OAKS, D. (1984). *Analysis of Survival Data*. New York: Chapman and Hall.

DENNIS, J.E. Jr, et SCHNABEL, R.B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall.

HAN, A., et HAUSMAN, J. A. (1986). Semiparametric Estimation of Duration and Competing Risk Models. M.I.T. Document de travail, numéro 450.

KAPLAN, E.L., et MEIER, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-81.

MARSHALL, K. (1999). Employment after childbirth. *Perspectives on labour and income*. Statistique Canada, 18-25.

MEYER, B.D. (1990). Unemployment Insurance and Unemployment Spells. *Econometrica*, 58, 757-782.

MOREL, J.G. (1989). Régression logistique selon des plans de sondage complexes. *Techniques d'enquête* 15, 213-233.

PLAGER, L., et CHEN, E.J. (1999). Student Debt from 1990-91 to 1995-96: An Analysis of Canada Student Loans Data. MAJOR RELEASES, *THE DAILY* and *Education Quarterly Review*, Statistique Canada, 5, 4, 10-35.

PRENTICE, R., et GLOECKLER, L. (1978). Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data, *Biometrics*, 34, 57-67.

ROBERTS, G.A., RAO, J.N.K. et KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

ROWE, G., et CHEN, E.J. (1998). Un modèle à entrées et sorties de la progression scolaire au secondaire dans les provinces Canadiennes. *Recueil : Symposium 98, Analyse longitudinale pour les enquêtes complexes*, Statistique Canada, 183-192.

ROWE, G., et LIN, X. (1999). Modélisation des carrières au sein de la population active pour le modèle de simulation des *Lifepaths*, *Recueil : Symposium 99, Combiner des données de sources différentes*, Statistique Canada, 61-69.

WOLFSON, M.C. (1997). Sketching LifePaths: A New Framework for Socio-Economic Statistics. *Simulating Social Phenomena*, (Éds. Conte, R. Gegselmann et P. Terna), Lecture Notes in Economics and Mathematical Systems, 456, Springer.

WOLFSON, M.C., et ROWE, G. (1996). Perspectives on Working Time Over the Life Cycle, Canadian Employment Research Forum Conference on Changes to Working Time, Ottawa.

WOLFSON, M.C., et ROWE, G. (1998a). LifePaths - Toward an Integrated Microanalytic Framework for Socio-Economic Statistics. 26th General Conference of the International Association for Research in Income and Wealth, Cambridge, U.K.

WOLFSON, M.C., et ROWE, G. (1998b). Public Pension Reforms - Analyses Based on the LifePaths Generational Accounting Framework, 26th General Conference of the International Association for Research in Income and Wealth, Cambridge, U.K.

WOLFSON, M.C., ROWE, G., GRIBBLE, S. et LIN, X. (1998). Historical Generational Accounting with Heterogeneous Populations. *Government Finances and Generational Equity* (Éd. M. Corak). Statistique Canada, 107-127.

Techniques d'enquête, juin 2002
Vol. 28, n° 1, pp. 67-82
Statistique Canada

Estimation améliorée des ratios dans le cas des enquêtes téléphoniques avec correction pour la non-couverture

STEVEN T. GARREN et TED C. CHANG[1]

RÉSUMÉ

Comme certains membres d'une population peuvent ne pas avoir le téléphone, les enquêtes téléphoniques par la méthode de composition aléatoire à l'intérieur des strates donnent parfois lieu à des estimateurs asymptotiquement biaisés des ratios. Nous examinons l'effet associé à l'impossibilité d'échantillonner la population non abonnée au téléphone. Nous tenons compte de la propension d'un ménage à avoir le téléphone pour proposer un estimateur stratifié a posteriori pondéré en fonction de l'abonnement au téléphone qui semble donner de meilleurs résultats que l'estimateur stratifié a posteriori type si l'on prend pour critère l'erreur quadratique moyenne. Nous estimons les propensions à posséder le téléphone d'après les échantillons de microdonnées à grande diffusion (PUMS pour *Public Use Microdata Samples*) provenant du Recensement des États-Unis. Nous considérons des estimateurs non stratifiés a posteriori lorsque la taille de l'échantillon est faible. Nous calculons, pour chaque échantillon, l'erreur quadratique moyenne asymptotique, ainsi que son estimation fondée sur un échantillon. Enfin, nous examinons des exemples réels en nous servant des PUMS. Les autres formes de non-réponse ne sont pas examinées dans le cadre de cette étude.

MOTS CLÉS : Asymptotique; échantillons de microdonnées à grande diffusion du Recensement; stratification a posteriori; enquête téléphonique.

1. INTRODUCTION

Considérons les enquêtes où l'on échantillonne la population d'abonnés au téléphone. Les problèmes importants que posent les enquêtes téléphoniques sont la non-réponse (c'est-à-dire le refus de participer à l'enquête) et la non-couverture (c'est-à-dire le non-abonnement au service téléphonique). La non-réponse peut être la cause d'un biais plus important que la non-couverture, puisque la propension à ne pas répondre est habituellement nettement plus élevée que la propension à ne pas être abonné au téléphone. Cependant, ici, nous n'examinons que brièvement le cas de la non-réponse, car l'étude se concentre sur la non-couverture.

1.1 Revue de la littérature

Khurshid et Sahai (1995) ont produit une bibliographie de grande envergure des articles traitant des enquêtes téléphoniques. Steeh, Groves, Comment et Hansmire (1983, pages 189 à 197) donnent des exemples de taux de non-réponse. Little (1986) et Rubin (1987) examinent des méthodes de correction de la non-réponse, par pondération ou par imputation. Rao (1997) donne une vue d'ensemble des enquêtes par sondage, y compris une discussion des méthodes de rééchantillonnage, en particulier le jackknife pour l'estimation de variance. La discussion inclue des techniques pour l'estimation de la variance due à l'imputation.

En ce qui concerne la non-couverture, Brick, Waksberg et Keeter (1994) constatent qu'aux États-Unis, en tout temps, 94 % des ménages sont abonnés au téléphone. Ils notent aussi que les ménages dont le service téléphonique est interrompu sont habituellement indigents. Keeter (1995) mentionne que, durant une enquête réalisée de 1992 à 1993, plus de la moitié des ménages n'ayant pas été raccordés de façon continue au service téléphonique durant cette année-là étaient des abonnés irréguliers, c'est-à-dire des ménages qui avaient été abonnés et non-abonnés au téléphone à différentes périodes durant l'année en question. Il observe aussi que la plupart des caractéristiques socioéconomiques (sauf la propriété du logement) des ménages « abonnés irréguliers » étaient comparables à celles des ménages perpétuellement sans le téléphone. Ces similarités entre les populations d'abonnés irréguliers et de ménages non-abonnés au téléphone donnent à penser que l'on peut faire des inférences valides au sujet de l'ensemble de la population (abonnés, non-abonnés et abonnés irréguliers) d'après les enquêtes téléphoniques. Thornberry et Massey (1988) examinent la non-couverture pour divers groupes sociodémographiques de 1963 à 1986 et constatent que le revenu est le déterminant le plus important de la probabilité qu'un ménage ait le téléphone.

1.2 Notre approche

À partir de plusieurs caractéristiques, telles que la propriété du logement et la langue parlée par les membres du ménage, nous estimons ici la propension d'un ménage à être abonné au service téléphonique en nous servant de la partie correspondant à la Virginie des échantillons de microdonnées à grande diffusion (PUMS pour *Public Use*

[1] Steven T. Garren, Department of Mathematics and Statistics, MSC 7803, James Madison University, Harrisonburg, Virginia, 22807, U.S.A. Travaux de recherche financés en partie par la subvention MH53259-01A2 du NIMH; Ted C. Chang, Division of Statistics, 108 Halsey Hall, University of Virginia, Charlottesville, Virginia, 22903, U.S.A. Travaux de recherche financés en partie par la subvention N-00014-92-J-1009 de l'ONR.

Microdata Samples) du Recensement de 1990, échantillon qui représente 5 % de la population. Une variable du fichier indique si les ménages sont ou non abonnés au téléphone. Nous estimons la propension (probabilité) à être abonné au service téléphonique par régression linéaire généralisée avec fonction de lien bilogarithmique, car le lien logit produit un ajustement médiocre du modèle. Nous recommandons d'utiliser notre modèle ajusté de régression, ainsi que les paramètres estimés, pour estimer ces probabilités de façon générale, lorsqu'un échantillon aléatoire est sélectionné à partir de la population d'abonnés au téléphone de Virginie.

Parce qu'il s'agit d'énormes ensembles de données, les PUMS jouent un autre rôle utile dans la présente étude. Nous les utilisons pour comparer le biais et la variance des estimateurs en examinant l'ensemble de la population d'abonnés au téléphone et en tirant des échantillons répétés de cette population. Les PUMS contiennent des données nominales correspondant à 75 variables du ménages et 75 variables personnelles pour tous les membres des ménages sélectionnés.

Dans les exemples donnés à la section 6, nous estimons le taux de titulaires d'un diplôme d'études secondaires, le nombre moyen d'automobiles par ménage et le revenu moyen du ménage au moyen d'estimateurs stratifiés et non stratifiés a posteriori pour des échantillons de 500 unités provenant des PUMS. Pour le taux de titulaires d'un diplôme d'études secondaires, les variables de stratification a posteriori sont le sexe, l'âge et la race du chef du ménage. Pour le nombre moyen d'automobiles par ménage, les variables de stratification a posteriori sont le revenu du ménage uniquement. Nous analysons les estimateurs du revenu moyen du ménage deux fois, d'abord en utilisant comme variable de stratification a posteriori uniquement la race du chef du ménage, puis en utilisant comme variables de stratification a posteriori le sexe, l'âge et la race du chef du ménage. Chacune des ces variables de stratification a posteriori est répartie en deux catégories, sauf le revenu, qui est réparti en trois catégories.

Le fait que les estimateurs ne tiennent pas compte de la propension à être abonné au téléphone a pour grand inconvénient que ces estimateurs ne sont pas asymptotiquement non-biaisés lorsque la taille de l'échantillon devient importante. L'un des principaux objectifs du présent article est de montrer que le biais est réduit considérablement lorsque les estimateurs tiennent compte de la propension à être abonné au téléphone, telle qu'estimée d'après les PUMS.

Puisque nous considérons les estimateurs *stratifiés a posteriori* ainsi que *non stratifiés a posteriori*, de même que *l'usage* ou le *non-usage* de la propension à être abonné et à ne pas être abonné au service téléphonique, nous examinons ici quatre estimateurs d'une moyenne de population qui sont la moyenne d'échantillon, l'estimateur stratifié a posteriori habituel, un estimateur pondéré en fonction de l'abonnement au téléphone et, enfin, l'estimateur stratifié a posteriori pondéré en fonction de l'abonnement au téléphone que nous proposons. L'erreur quadratique moyenne (EQM) de l'estimateur pondéré en fonction de l'abonnement au téléphone et celle de l'estimateur stratifié a posteriori pondéré en fonction de l'abonnement au téléphone tendent vers zéro lorsque la taille de l'échantillon devient importante, contrairement à l'EQM des deux autres estimateurs.

Nous adoptons un modèle à deux phases pour nos quatre estimateurs. La première phase comprend la sélection de ménages à partir de la population entière pour faire partie de la population d'abonnés au téléphone. Nous traitons la propension d'un ménage à être abonné au service téléphonique comme étant la probabilité que le ménage soit sélectionné pour faire partie de la population ayant le téléphone et nous supposons que cette probabilité est positive (quoique, éventuellement, faible) pour chaque ménage. La deuxième phase est la sélection d'un échantillon aléatoire simple stratifié (peut-être selon des variables géographiques) de la population ayant le téléphone. Dans les exemples donnés à la section 6, nous considérons la stratification a posteriori selon des caractéristiques telles que la race et l'âge du chef du ménage. Puisque la taille de nos échantillons est faible, nous ne stratifions pas la population de la Virginie selon des variables géographiques, bien que notre formule permette à la fois la stratification et la stratification a posteriori.

Idéalement, on procéderait à la stratification a posteriori en se servant des mêmes covariables que celles utilisées pour estimer la propension à être abonné au téléphone à la première phase du modèle. Dans ce cas, les trois estimateurs fondés sur la propension à être abonné au service téléphonique et (ou) la stratification a posteriori seront presque identiques. Cependant, la taille d'échantillon de chaque catégorie stratifiée a posteriori ne doit pas être faible, de sorte que des limites pratiques limitent le nombre de catégories qu'il convient d'utiliser pour la stratification a posteriori. Néanmoins, de nombreuses catégories peuvent être utilisées pour déterminer la propension à être abonné au service téléphonique d'après les PUMS, parce que l'on utilise l'entièreté de la population.

Même s'il est possible de procéder à une stratification a posteriori au moyen d'un grand nombre de covariables, les formules habituelles de la variance en cas de stratification a posteriori exigent que l'on tire un échantillon aléatoire stratifié à partir de la population entière. Or, dans notre situation, nous sélectionnons un échantillon aléatoire stratifié à partir de la population ayant le téléphone, si bien que les formules habituelles de la variance ne sont pas applicables. Les techniques mises au point par Politz et Simmons (consulter Cochran 1977, pages 374 à 377) demandent qu'on utilise la population entière comme base de sondage plutôt que simplement la population ayant le téléphone et ne s'appliquent donc pas à notre scénario puisqu'il permet la non-couverture.

Nous calculons la variance asymptotique de chacun des quatre estimateurs d'un ratio de population et nous

déterminons des estimations raisonnables de ces variances. Puisqu'une moyenne de population est un cas spécial de ratio et qu'un total de population est un multiple de ratio, les résultats concernant les estimateurs des moyennes ou des totaux découlent des résultats concernant les estimateurs des ratios.

2. NOTATION

Considérons N ménages dans une population U. Pour chaque ménage compris dans U, représentons deux variables étudiées par y_{1k} et y_{2k}, pour $k \in U$. À tous point dans le temps, le fait que le $k^{ième}$ ménage possède ou ne possède pas le téléphone est traité comme étant aléatoire, tandis que y_{ik} est traité comme étant fixe.

En posant que

$$\alpha_i = N^{-1} \sum_{k \in U} y_{ik},$$

pour $i = 1, 2$, l'objectif est d'estimer α_1, α_2 et le ratio

$$\mu = \alpha_1 / \alpha_2.$$

En toute généralité, nous nous concentrons sur l'estimation de α_1 et μ.

Un cas spécial important de l'estimation d'un ratio μ se pose lorsque l'on souhaite estimer la moyenne d'une variable z_k pour $k \in D$ pour une sous-population $D \subset U$ que l'on ne peut pas échantillonner directement. Les sous-populations définies selon la race en sont des exemples. Posons que x_k est égal à 1 si $k \in D$ et égal à 0 autrement. Posons aussi que $y_{1k} = z_k x_k$ et $y_{2k} = x_k$. Alors, μ représente la moyenne de population de z_k sur la sous-population D.

Supposons qu'il existe H strates et utilisons la notation h comme indice de la strate. Supposons qu'il existe G groupes et choisissons la notation g comme indice des groupes qui sont utilisés pour construire les strates a posteriori. Nous connaissons les strates avant l'échantillonnage, mais nous n'observons les groupes qu'après la sélection de l'échantillon final. Par conséquent, U_{gh} représente tous les ménages dans le groupe g et la strate h; N_{gh} représente la taille de U_{gh} et N_h représente la taille de U_h. Les autres variables sont définies de la même façon en fonction de g et de h.

Représentons par U_T la population de ménages dans U qui sont abonnés au téléphone à l'heure actuelle et représentons par N_T la taille de U_T. La probabilité, ou propension, que le $k^{ième}$ ménage dans U soit également dans U_T est représentée par p_k et nous supposons que $p_k > 0$ pour toute valeur de k. Un échantillon aléatoire simple de taille n_h est tiré à partir de U_{Th} pour $h = 1, ..., H$. Représentons par s_h l'échantillon final dans la strate h. La taille de l'échantillon final, s, est représentée par n. Pour l'analyse asymptotique, ici, nous supposons que $n/N \to 0$ lorsque $n \to \infty$ dans le même esprit que Särndal, Swensson et Wretman (1992, pages 166 à 169).

3. LES ESTIMATEURS

Le plan d'échantillonnage est traité comme un plan à deux phases avec échantillonnage de Poisson à la première phase et échantillonnage aléatoire simple stratifié à la deuxième phase. Chaque individu entre dans la population d'abonnés au téléphone avec la probabilité p_k, pour $k \in U$, puis entre dans l'échantillon final conformément à la sélection d'un échantillon aléatoire simple de taille n_h, $h = 1, ..., H$. Nous supposons que les p_k sont connues ou qu'elles peuvent être estimées exactement, comme nous le montrons à la section 5. La validation des estimateurs de μ examinés à la présente section figure à l'annexe.

3.1 Estimateur stratifié a posteriori et estimateur par ratio

Les estimations stratifiées a posteriori de α_1 et α_2 sont

$$\hat{\alpha}_{ps(i)} = N^{-1} \sum_{h=1}^{H} \sum_{g=1}^{G} N_{gh} n_{gh}^{-1} \sum_{k \in s_{gh}} y_{ik},$$

pour $i = 1, 2$, et l'estimation stratifiée a posteriori de μ est $\hat{\mu}_{ps} = \hat{\alpha}_{ps(1)} / \hat{\alpha}_{ps(2)}$. On sait qu'une estimation valide de la variance, subordonnée à U_T, (consulter Särndal et coll. 1992, pages 270-271) est

$$\widehat{\text{var}}\,(\hat{\mu}_{ps} | U_T)$$

$$= (N \hat{\alpha}_{ps(2)})^{-2} \sum_{h=1}^{H} \sum_{g=1}^{G} N_{gh}^2 \left[\frac{1 - (n_h / N_{Th})}{n_{gh}(n_{gh} - 1)} \right]$$

$$\sum_{k \in s_{gh}} \left[y_{1k} - \hat{\mu}_{ps} y_{2k} - n_{gh}^{-1} \sum_{j \in s_{gh}} (y_{1j} - \hat{\mu}_{ps} y_{2j}) \right]^2. \quad (3.1)$$

Bien que l'on ne puisse estimer le biais d'après l'échantillon final, il est bien connu que le biais théorique de $\hat{\mu}_{ps}$ est

$$\text{biais } \hat{\mu}_{ps} = \frac{\sum_{h=1}^{H} \sum_{g=1}^{G} N_{gh} \left[\sum_{j \in U_{gh}} p_j \right]^{-1} \sum_{k \in U_{gh}} p_k y_{1k}}{\sum_{h=1}^{H} \sum_{g=1}^{G} N_{gh} \left[\sum_{j \in U_{gh}} p_j \right]^{-1} \sum_{k \in U_{gh}} p_k y_{2k}}$$

$$- \frac{\sum_{k \in U} y_{1k}}{\sum_{k \in U} y_{2k}} + O(n^{-1}) \quad (3.2)$$

lorsque $n \to \infty$. Notons, en examinant (3.2), qu'en général, l'EQM de $\hat{\mu}_{ps}$ ne tend pas vers zéro à mesure que la taille n de l'échantillon augmente.

Pour déterminer la variance et le biais de $\hat{\alpha}_{ps(1)}$, posons $y_{2k} = 1$ pour toutes les valeurs de k, de sorte que $\hat{\mu}_{ps}$ et μ

deviennent $\hat{\alpha}_{ps(1)}$ et α_1, respectivement. Nous pouvons alors appliquer les équations (3.1) et (3.2) de sorte que

$$\widehat{\text{var}}\ \hat{\alpha}_{ps(1)} = N^{-2}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh}^2\left[\frac{1-(n_h/N_{Th})}{n_{gh}(n_{gh}-1)}\right]$$

$$\sum_{k\in s_{gh}}\left[y_{1k} - n_{gh}^{-1}\sum_{j\in s_{gh}} y_{1j}\right]^2, \qquad (3.3)$$

et

biais $\hat{\alpha}_{ps(1)}$

$$= N^{-1}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh}\left(\sum_{j\in U_{gh}} p_j\right)^{-1}\sum_{k\in U_{gh}} p_k y_{1k} - \alpha_1 + O(n^{-1})$$

$$= O(1)$$

lorsque $n \to \infty$. Cochran (1977, pages 134-135) a produit un facteur de correction, d'ordre n^{-2}, pour (3.3). Toutefois, ce facteur de correction ne s'applique pas à (3.1) puisque le terme d'erreur due à l'estimation d'après le ratio est $O(n^{-2})$.

Comme d'habitude, l'estimateur par ratio, représenté par $\bar{y}_1/\bar{y}_2$, est défini comme étant le ratio de la moyenne d'échantillon de y_1 à la moyenne d'échantillon y_2, c'est-à-dire

$$\bar{y}_1/\bar{y}_2 = \sum_{k\in s} y_{1k}\Big/\sum_{j\in s} y_{2j}.$$

L'estimateur stratifié a posteriori et l'estimateur par ratio sont identiques si $G = H = 1$. Puisque nous n'utiliserons qu'une seule strate (c'est-à-dire $H = 1$) à la section 6, il n'est pas nécessaire que nous fassions référence à une théorie distincte pour l'estimateur par ratio.

3.2 Estimateur pondéré en fonction de l'abonnement au téléphone

Puisque l'estimateur stratifié a posteriori, $\hat{\mu}_{ps}$, est biaisé, nous proposons deux autres estimateurs. L'un est l'estimateur pondéré en fonction de l'abonnement au téléphone, qui tient compte de la probabilité qu'un individu ait le téléphone. À la présente section, nous supposons que les p_k sont connues pour toutes les valeurs de $k \in s$ ou que l'on peut les estimer exactement. Nous décrivons l'estimation des p_k d'après les PUMS à la section 5.

Pour obtenir une estimation grossière de α_i pour $i = 1, 2$, utilisons

$$\tilde{\alpha}_{w(i)} = N^{-1}\sum_{h=1}^{H} N_{Th} n_h^{-1}\sum_{k\in s_h} p_k^{-1} y_{ik}. \qquad (3.4)$$

Puis, estimons μ au moyen de

$$\hat{\mu}_w = \tilde{\alpha}_{w(1)}/\tilde{\alpha}_{w(2)}, \qquad (3.5)$$

qui est asymptotiquement non biaisé pour μ, puisque $\tilde{\alpha}_{w(i)}$ est un estimateur non biaisé de $\alpha_{w(i)}$, pour $i = 1, 2$. Nous montrons qu'une estimation valide de la variance de $\hat{\mu}_w$ est donnée par

$$\widehat{\text{var}}\ \hat{\mu}_w = [N\tilde{\alpha}_{w(2)}]^{-2}\sum_{h=1}^{H}\frac{N_{Th}^2\,[1-(n_h/N_{Th})]}{n_h(n_h-1)}$$

$$\sum_{k\in s_h}\left[\frac{y_{1k}-\hat{\mu}_w y_{2k}}{p_k} - n_h^{-1}\sum_{j\in s_h}\frac{y_{1j}-\hat{\mu}_w y_{2j}}{p_j}\right]^2. \qquad (3.6)$$

Puisque l'estimateur $\hat{\mu}_w$ est asymptotiquement non biaisé, alors une estimation valide de l'EQM de $\hat{\mu}_w$ est identique à l'estimation de la variance.

Poser $y_{2j} = 1$ dans (3.4) et (3.5) permet qu'une estimation valide de $\alpha_{w(1)}$ soit

$$\tilde{\alpha}_{w(1)} = \left[\sum_{h=1}^{H} N_{Th} n_h^{-1}\sum_{k\in s_h} p_k^{-1}\right]^{-1}\sum_{h=1}^{H} N_{Th} n_h^{-1}\sum_{k\in s_h} p_k^{-1} y_{1k}.$$

On peut estimer la variance de $\hat{\alpha}_{w(1)}$ en posant $y_{2j} = 1$ dans (3.6).

3.3 Estimateur stratifié a posteriori pondéré en fonction de l'abonnement au téléphone

L'autre estimateur que nous proposons, qui combine la stratification a posteriori et l'estimateur pondéré en fonction de l'abonnement au téléphone, est peut-être le meilleur des quatre si la taille des échantillons est suffisante pour justifier la stratification a posteriori. Toutefois, pour utiliser ce nouvel estimateur, il faut que tous les N_{gh} soient suffisamment grands pour que, lorsque la probabilité est élevée, les n_{gh} ne soient pas trop petits. Pour estimer α_i, nous utilisons

$$\tilde{\alpha}_{psw(i)} = N^{-1}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh}\left[\sum_{j\in s_{gh}} p_j^{-1}\right]^{-1}\sum_{k\in s_{gh}} p_k^{-1}\, y_{ik},$$

pour $i = 1, 2$. Puis, nous estimons μ par $\hat{\mu}_{psw} = \hat{\alpha}_{psw(1)}/\hat{\alpha}_{psw(2)}$. L'estimation de la variance de $\hat{\mu}_{psw}$ est donnée par

$$\widehat{\text{var}}\ \hat{\mu}_{psw} = [N\tilde{\alpha}_{psw(2)}]^{-2}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh}^2\left(\sum_{j\in s_{gh}} p_j^{-1}\right)^{-2}$$

$$\left(\frac{n_{gh}}{n_{gh}-1}\right)\left(1-\frac{n_h}{N_{Th}}\right)\sum_{k\in s_{gh}} p_k^{-2}$$

$$\left[y_{1k} - \hat{\mu}_{psw} y_{2k} - \left(\sum_{j\in s_{gh}} p_j^{-1}\right)^{-1}\sum_{m\in s_{gh}} p_m^{-1}(y_{1m} - \hat{\mu}_{psw} y_{2m})\right]^2. \qquad (3.7)$$

Si la valeur de tout terme n_{gh} est faible, nous pourrions alors préférer l'estimateur

$$\widehat{\widehat{\text{var}}}\,\hat{\mu}_{\text{psw}} = [N\tilde{\alpha}_{\text{psw}(2)}]^{-2}\sum_{h=1}^{H} N_h^2\left(\sum_{j\in s_h} p_j^{-1}\right)^{-2}$$

$$\left(\frac{n_h}{n_h-1}\right)\left(1-\frac{n_h}{N_{Th}}\right)\sum_{g=1}^{G}\sum_{k\in s_{gh}} p_k^{-2}$$

$$\left[y_{1k}-\hat{\mu}_{\text{psw}}y_{2k}-\left(\sum_{j\in s_{gh}} p_j^{-1}\right)^{-1}\sum_{m\in s_{gh}} p_m^{-1}(y_{1m}-\hat{\mu}_{\text{psw}}y_{2m})\right]^2. \quad (3.8)$$

Notons que, si l'on connaissait la valeur des N_{Tgh}, situation fort peu probable, un estimateur mieux connu et plus intuitif de $\hat{\mu}_{\text{psw}}$ serait

$$\widetilde{\text{var}}\,\hat{\mu}_{\text{psw}} = [N\tilde{\alpha}_{\text{psw}(2)}]^{-2}\sum_{h=1}^{H}\sum_{g=1}^{G}\frac{N_{Tgh}^2}{n_{gh}(n_{gh}-1)}$$

$$\left(1-\frac{n_{gh}}{N_{Tgh}}\right)\sum_{k\in s_{gh}} p_k^{-2}$$

$$\left[y_{1k}-\hat{\mu}_{\text{psw}}y_{2k}-\left(\sum_{j\in s_{gh}} p_j^{-1}\right)^{-1}\sum_{m\in s_{gh}} p_m^{-1}(y_{1m}-\hat{\mu}_{\text{psw}}y_{2m})\right]^2. \quad (3.9)$$

Puisque ordinairement, on ne connaît pas les N_{Tgh}, (3.9) n'est généralement pas un estimateur pratique. Cependant, (3.9) facilite la justification de (3.7) et (3.8) qui sont, elles, des équations assez pratiques.

Puisque l'estimateur $\hat{\mu}_{\text{psw}}$, est asymptotiquement non biaisé, alors une estimation valide de l'EQM est identique à l'estimation de la variance. En outre, poser que $y_{2j} = 1$ dans (3.7) et (3.8) permet d'estimer la variance de $\hat{\alpha}_{\text{psw}(1)}$.

Si $G = 1$, l'estimateur $\hat{\mu}_{\text{psw}}$ ne se réduit pas à $\hat{\mu}_w$, comme on pourrait naïvement s'y attendre. Lorsque $G = 1$ l'estimateur de choix est $\hat{\mu}_w$, puisqu'il est fondé sur un seul ratio, tandis que $\hat{\mu}_{\text{psw}}$ est fondé sur un ratio de ratios. L'utilisation de l'estimateur $\hat{\mu}_{\text{psw}}$ nécessite un échantillon de plus grande taille pour chaque catégorie strate-groupe, tandis que $\hat{\mu}_w$ nécessite uniquement un échantillon global de grande taille. Néanmoins, lorsque $H = G = 1$, les estimateurs $\hat{\mu}_w$ et $\hat{\mu}_{\text{psw}}$ sont identiques; les estimateurs de la variance fondés sur $\hat{\mu}_w$ sont préférables à ceux fondés sur $\hat{\mu}_{\text{psw}}$, car les estimations de la variance de $\hat{\mu}_{\text{psw}}$ sont basées sur un ratio de ratios.

4. ERREUR QUADRATIQUE MOYENNE ASYMPTOTIQUE

Nous allons maintenant donner l'erreur quadratique moyenne asymptotique des estimateurs définis à la section 3. Les preuves, qui découlent de la linéarisation par série de Taylor, sont exposées à l'annexe, ainsi que les conditions mineures de régularité requises.

4.1 Estimateur stratifié a posteriori

Pour trouver la variance asymptotique théorique de l'estimateur stratifié a posteriori de μ, nous commençons par définir

$$\alpha_i^* = \text{plim}_{n\to\infty}\,\hat{\alpha}_{\text{ps}(i)} = N^{-1}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh}$$

$$\left(\sum_{j\in U_{gh}} p_j\right)^{-1}\sum_{k\in U_{gh}} p_k y_{ik}, \quad (4.1)$$

pour $i = 1, 2$, ainsi que

$$\mu^* = \alpha_1^*/\alpha_2^*. \quad (4.2)$$

Notons qu'en général, $\alpha_i^* \neq \alpha_i$ et $\mu^* \neq \mu$. La variance asymptotique théorique de $\hat{\mu}_{\text{ps}}$ est

$$\text{var}\,\hat{\mu}_{\text{ps}} = (N\alpha_2^*)^{-2}\sum_{h=1}^{H}\sum_{g=1}^{G}$$

$$\frac{N_{gh}^2\left[\left(\sum_{j\in U_h} p_j\right)-n_h\right]}{n_h\left(\sum_{j\in U_{gh}} p_j\right)\left[\left(\sum_{j\in U_{gh}} p_j\right)-1\right]}\sum_{k\in U_{gh}} p_k$$

$$\left[y_{1k}-\mu^* y_{2k}-\frac{\sum_{j\in U_{gh}} p_j(y_{1j}-\mu^* y_{2j})}{\sum_{j\in U_{gh}} p_j}\right]^2 + O\left(n^{-2}+N^{-1}\right) \quad (4.3)$$

lorsque $n \to \infty$. Nous avons montré dans (3.2) que le biais asymptotique de $\hat{\mu}_{\text{ps}}$ est $O(1)$ lorsque $n \to \infty$. Par conséquent, l'EQM asymptotique de $\hat{\mu}_{\text{ps}}$ est également $O(1)$ si $n \to \infty$.

4.2 Estimateur pondéré en fonction de l'abonnement au téléphone

La variance asymptotique théorique de l'estimateur pondéré en fonction de l'abonnement au téléphone de μ est

$$\text{var}\,\hat{\mu}_w = (N\alpha_2)^{-2}\sum_{h=1}^{H}\frac{\left[\left(\sum_{j\in U_h} p_j\right)-n_h\right]\left(\sum_{j\in U_h} p_j\right)}{n_h\left[\left(\sum_{j\in U_h} p_j\right)-1\right]}$$

$$\sum_{k\in U_h} p_k\left[\frac{y_{1k}-\mu\, y_{2k}}{p_k}-\frac{\sum_{j\in U_h}(y_{1j}-\mu\, y_{2j})}{\sum_{j\in U_h} p_j}\right]^2 + O\left(n^{-2}+N^{-1}\right). \quad (4.4)$$

Puisque $\hat{\mu}_w$ est asymptotiquement non biaisé, alors son EQM est identique au deuxième membre de l'équation (4.4).

4.3 Estimateur stratifié a posteriori pondéré en fonction de l'abonnement au téléphone

La variance asymptotique théorique de l'estimateur stratifié a posteriori pondéré en fonction de l'abonnement au téléphone de μ est

$$\operatorname{var} \hat{\mu}_{psw} = (N\alpha_2)^{-2} \sum_{h=1}^{H} \sum_{g=1}^{G} \frac{\left(\sum_{j \in U_{gh}} p_j\right)\left[\left(\sum_{j \in U_h} p_j\right) - n_h\right]}{n_h\left[\left(\sum_{j \in U_{gh}} p_j\right) - 1\right]} \sum_{k \in U_{gh}} p_k^{-1}$$

$$\left[y_{1k} - \mu y_{2k} - N_{gh}^{-1} \sum_{j \in U_{gh}} (y_{1j} - \mu y_{2j})\right]^2$$

$$+ O\left(n^{-2} + N^{-1}\right). \qquad (4.5)$$

Puisque $\hat{\mu}_{psw}$ est asymptotiquement non biaisé, son EQM est identique au deuxième membre de l'équation (4.5).

5. ESTIMATION DES p_k AU MOYEN DES ÉCHANTILLONS DE MICRODONNÉES À GRANDE DIFFUSION

Le United States Bureau of the Census a produit les échantillons de microdonnées à grande diffusion (PUMS), c'est-à-dire des échantillons au 1/100 et au 1/20 de la population de chacun des 50 États et de Washington, D.C., pour l'année de référence 1990. Pour chaque personne sélectionnée dans l'échantillon, le fichier contient les données sur 75 variables du ménage et 75 variables personnelles, chaque ménage ayant un chef de ménage clairement défini. Nous utilisons les PUMS pour deux raisons. À la présente section, nous nous en servons pour estimer les p_k tandis qu'à la section 6, nous les utilisons pour l'exécution de simulations en vue de construire des exemples permettant de comparer les estimateurs.

Dans le présent article, nous utilisons l'échantillon au 1/20 pour la Virginie. Puisque 1/20, c'est-à-dire 5 %, représente un nombre énorme de ménages, nous traitons cet échantillon comme s'il s'agissait de la population totale de la Virginie. Comme nous nous intéressons aux enquêtes téléphoniques, nous sélectionnerons les ménages à partir de cet échantillon au 1/20. Nous pouvons faire des inférences concernant les variables personnelles, telles que le taux de titulaires d'un diplôme d'études secondaires et les variables du ménage, telles que le nombre d'automobiles ou le revenu d'un ménage. Les fichiers des PUMS contiennent, pour chaque ménage, des données indiquant si celui-ci est abonné ou non au téléphone. Nous avons éliminé de notre étude tous les ménages pour lesquels le fichier contenait, pour la situation d'abonnement au téléphone, la mention « sans objet ». Ces ménages correspondaient soit à des logements vacants soit à des logements collectifs (institutionnels ou non). Pour 1990, le nombre de ménages retenus est de 110 744, dont 104 606 possédaient le téléphone, donc, la proportion de ménages retenus ayant le téléphone est de 94,5%

Pour estimer p_k, qui est la probabilité, ou la propension, que le $k^{\text{ième}}$ ménage ait le téléphone, nous utilisons la régression linéaire généralisée avec un lien bilogarithmique sur l'échantillon au 1/20 pour la Virginie, ainsi que les coefficients de pondération attribués aux ménages dans le fichier PUMS. McCullagh et Nelder (1991, pages 107 à 110) recommandent d'utiliser un lien bilogarithmique lorsque la valeur des probabilité s'approche de l'unité et nous avons constaté que ce lien produit un bon ajustement du modèle. Nous avons également observé que la fonction de lien logit donne un mauvais ajustement.

Nous utilisons les coefficients de pondération des ménages du PUMS pour estimer les p_k, mais nous ne les utilisons nulle part ailleurs. En particulier, à la section 6, lors de la création d'échantillons de Monte Carlo de la population PUMS, il s'agit d'échantillons aléatoires simples de la population d'abonnés au téléphone.

Exemples d'estimation des p_k

Nous utilisons, pour estimer les p_k, six covariables, à savoir le nombre de personnes dans le ménage, le mode d'occupation du logement (propriétaire ou locataire), la date à laquelle le chef de ménage est entré dans le logement, le revenu du ménage, la langue parlée par le ménage et la race du chef de ménage. Nous avons choisi ces six covariables ainsi que les catégories pour chacune d'elles, en nous basant sur une analyse approfondie du PUMS de 1990 par les techniques de régression linéaire généralisée de SAS. L'analyse a montré que toutes ces covariables étaient hautement significatives. Pour obtenir les estimations des p_k, nous additionnons les estimations appropriées des covariables figurant dans le tableau 1. La covariable pour le nombre de personnes doit être multipliée par le nombre de personnes que compte le ménage; cependant, si le nombre de personnes est supérieur à 5, nous le fixons à 5 aux fins des calculs. Par exemple, si le ménage compte 3 Américains d'ascendance asiatique parlant l'anglais, possédant deux automobiles et vivant dans une maison achetée en 1987 et que le revenu du ménage est égal à 75 000$, alors le tableau 1 indique que l'estimation de p_k est la solution de

$$\log(-\log p_k) = 3 \times 0{,}2747 - 0{,}5552 + 0{,}5920$$

$$+ 0{,}1896 + 1{,}0004 + 0{,}6156 + 0{,}0000.$$

Notons que, dans le tableau 1, pour les covariables de date *d'entrée dans le logement*, *de nombre d'automobiles* et *de revenu*, les valeurs correspondant aux catégories diminuent, comme prévu, de façon monotonique, sauf si le revenu est négatif.

Une correction recommandée si l'on recourt à la composition aléatoire consiste à demander à chaque répondant combien de lignes téléphoniques possède le ménage et à multiplier le nombre indiqué par l'estimation de p_k établie d'après le tableau 1 pour obtenir une nouvelle estimation de p_k. Par conséquent, p_k est maintenant un coefficient de pondération plutôt qu'une probabilité. Pour les simulations dont nous discutons à la section 6, cette correction n'est pas nécessaire, puisque les ménages sont tous aussi susceptibles les uns que les autres d'être sélectionnés par échantillonnage aléatoire simple du PUMS, quel que soit le nombre de lignes téléphoniques.

Donc, on peut se servir du tableau 1 pour estimer les p_k lorsque l'on réalise des enquêtes téléphoniques. Si l'on utilise un modèle de régression linéaire généralisé calculé d'après le PUMS produit pour une date de référence antérieure pour analyser les données d'une enquête ultérieure, on peut procéder à un rééchelonnement afin de tenir compte des modifications de la distribution des revenus des ménages au cours du temps. Le tableau 1 donne aussi les coefficients d'un modèle calculé en se basant sur le PUMS de 1980. Nous donnons à la section 6 un exemple d'utilisation d'un modèle fondé sur un PUMS de 1980 pour calculer les p_k pour un échantillon tiré de la population du PUMS de 1990. Nous notons que le PUMS de 1980 ne contenait pas la covariables (date d'entrée dans le logement) et qu'on obtenait un modèle mieux ajusté lorsque les catégories de langue « Espagnol » et « Autre » étaient combinées. En outre, le revenu médian du ménage a presque doublé de 1980 à 1990, si bien que le nombre de catégories de revenu utilisées en 1980 était plus faible qu'en 1990.

Bien que l'utilisation du tableau 1 soit pratique lorsque l'on procède à un échantillonnage à partir d'un PUMS et que l'on exécute des simulations, les données sur les covariables énumérées dans ce tableau pourraient ne pas être disponible dans le cas d'enquêtes réelles où l'on recourt à la composition aléatoire. Il est possible de reproduire le tableau 1 en utilisant d'autres covariables ou d'estimer les p_k selon la méthode de rechange qui suit.

Tableau 1

Valeurs des covariables pour l'estimation des p_k d'après le PUMS au 1/20 pour la Virginie. Les erreurs-types figurent entre parenthèses. Si le nombre de personnes est supérieur à 5, le convertir à la valeur 5. La covariable « mode d'occupation du logement » ne figurait pas dans le PUMS de 1980. La catégorie « 40 000$ à 49 999$ » de 1980 englobe effectivement « 40 000$ et plus ». La catégorie « autre » pour la covariable « langue » de 1980 comprend l'espagnol.

Covariable	Catégorie	Valeur de 1990		Valeur de 1980	
Nombre de personnes		0,2747	-0,0022	0,1929	-0,002
Mode d'occupation du logement	Propriétaire	-0,5552	(0,0079)	-0,7845	(0,0057)
	Locataire		(0,0000)		(0,0000)
Date d'entrée dans le logement	1989 ou 1990	0,9742	(0,0121)	NA	
	1985 à 1988	0,5920	(0,0119)	NA	
	1980 à 1984	0,3489	(0,0138)	NA	
	1970 à 1979	0,2185	(0,0136)	NA	
	1969 ou avant	0,0000	(0,0000)	NA	
Nombre d'automobiles	0	1,2927	(0,0152)	0,8633	(0,0118)
	1	0,6842	(0,0143)	0,3981	(0,0109)
	2	0,1896	(0,0145)	0,0399	(0,0112)
	3 ou plus	0,0000	(0,0000)	0,0000	(0,0000)
Revenu	Moins de 0$	3,5325	(0,1294)	2,3639	(0,0830)
	0$ to 9 999$	3,7929	(0,0539)	2,5238	(0,0260)
	10 000$ to 19 999$	3,4878	(0,0538)	1,9763	(0,0258)
	20 000$ to 29 999$	3,0299	(0,0539)	1,0220	(0,0269)
	30 000$ to 39 999$	2,4297	(0,0543)	0,3889	(0,0317)
	40 000$ to 49 999$	1,8899	(0,0556)	0,0000	(0,0000)
	50 000$ to 59 999$	1,5992	(0,0578)	NA	
	60 000$ to 69 999$	1,2144	(0,0631)	NA	
	70 000$ to 79 999$	1,0004	(0,0704)	NA	
	80 000$ et plus	0,0000	(0,0000)	NA	
Langue	Anglais	0,6156	(0,0164)	0,4232	(0,0153)
	Espagnol	0,4889	(0,0216)	NA	
	Autre	0,0000	(0,0000)	0,000	(0,0000)
Race	Noire	-0,4233	(0,0064)	-0,3837	(0,0058)
	Autre	0,0000	(0,0000)	0,0000	(0,0000)
Ordonnée à l'origine		-7,6707	(0,0588)	-4,9024	(0,0322)

Méthode de rechange pour l'estimation des p_k

On pourrait poser aux participants à une enquête téléphonique par composition aléatoire les deux questions suivantes : « 1) Au cours des 12 derniers mois, combien de lignes téléphoniques votre ménage possédait-il? 2) Au cours des 12 derniers mois, pendant combien de mois chaque ligne téléphonique a-t-elle été en service? » Maintenant, représentons par p_k la somme des réponses à la question (2). Par exemple, dans un ménage possédant deux lignes téléphoniques, où l'une de ces lignes a été en service pendant les 12 mois et l'autre pendant 5 mois seulement, l'estimation de p_k serait de 12+5=17. De nouveau, ici, p_k représente un coefficient de pondération plutôt qu'une probabilité. Poser cette deuxième question au répondant est une démarche comparable à celle recommandée par Brick et coll. (1994) qui ont également proposé de pondérer les données pour tenir compte de la probabilité qu'un ménage soit abonné au téléphone.

6. INFÉRENCES D'APRÈS LES VARIABLES DU MÉNAGE ET LES VARIABLES PERSONNELLES

Nous allons comparer les quatre estimateurs proposés de μ lorsque nous faisons des inférences concernant le taux de titulaires d'un diplôme d'études secondaires parmi les personnes de 21 ans et plus, le nombre moyen d'automobiles par ménage et le revenu moyen des ménages dans l'État de Virginie. Nous avons réalisé 100 000 simulations d'échantillons aléatoires simples comptant 500 ménages ayant le téléphone à partir du PUMS au 1/20 pour la Virginie pour 1990 en n'utilisant qu'une seule strate (c'est-à-dire H=1).

À la section 6.1, nous utilisons deux ensembles de valeurs de p_k. L'un se fonde sur l'ajustement d'un modèle de régression GLIM au PUMS de 1990 et l'autre sur l'ajustement d'un modèle GLIM au PUMS de 1980 en se servant, pour rajuster les catégories de revenu, du ratio du revenu médian du ménage de 1990 (32 800$) au revenu médian du ménage de 1980 (17 510$). L'utilisation de la valeur de p_k pour 1980 en vue d'estimer un paramètre pour 1990 montre à quel point notre méthode donne de bons résultats lorsque l'on utilise les coefficients du modèle GLIM pour les ensembles de données futures, à condition de faire à une correction pour tenir compte de l'inflation. Aux sections 6.2 et 6.3, nous utilisons uniquement les valeurs de p_k pour 1990.

Lorsque la taille des échantillons est suffisamment grande, on devrait recourir à la stratification a posteriori. Nous pouvons comparer entre-eux les estimateurs non stratifiés a posteriori, d'une part, et les estimateurs stratifiés a posteriori, d'autre part. La comparaison de $\bar{y}_1/\bar{y}_2$ à $\hat{\mu}_w$ est appropriée, ainsi que celle de $\hat{\mu}_{ps}$ à $\hat{\mu}_{psw}$. Ces comparaisons montrent les améliorations dues à l'utilisation de p_k dans les estimateurs.

6.1 Estimation du taux de titulaires d'un diplôme d'études secondaires

Si l'on utilise l'entièreté de la population du PUMS au 1/20 pour la Virginie en 1990, le taux moyen de titulaires d'un diplôme d'études secondaires pour l'ensemble des Virginiens de 21 ans et plus est μ=0,75118. Pour estimer le taux de titulaires d'un diplôme d'études secondaires au moyen d'un échantillon aléatoire simple et de $\hat{\mu}_{ps}$ ou $\hat{\mu}_{psw}$, nous procédons à une stratification a posteriori selon le sexe (masculin, féminin), l'âge (moins de 45 ans, 45 ans et plus) et la race (noire, autre) du chef du ménage. Nous estimons les valeurs de p_k d'après le tableau 1. Les valeurs des biais et des écarts-types dont nous discutons plus loin sont présentées au tableau 2 pour les p_k de 1990.

Tableau 2
Biais et écarts-types des estimations du taux de titulaires d'un diplôme d'études secondaires

	Estimateur			
	Non stratifié a posteriori		Stratifié a posteriori	
	$\bar{y}_1/\bar{y}_2$	$\hat{\mu}_w$	$\hat{\mu}_{ps}$	$\hat{\mu}_{psw}$
Biais agrégé	0,01471	0,00722	0,01461	0,00874
Biais téléphone	0,01472	0,00720	0,01463	0,00850
Biais deuxième phase	0,00000	0,00002	-0,00002	0,00024
Biais théorique	0,00777	0	0,00663	0
Écart-type simulé	0,01683	0,01737	0,01605	0,01643
Écart-type estimé	0,01680	0,01734	0,01601	0,01635*
Écart-type théorique	0,01700	0,01752	0,01617	0,01658
Racine de l'erreur quadratique moyenne	0,02236	0,01881	0,02171	0,01861

Le taux réel de titulaires d'un diplôme d'études secondaires est 0,75118. La stratification a posteriori est fondée sur le sexe, l'âge et la race. La taille des échantillons sélectionnés est de 500 et le nombre de simulations exécutées est de 100 000.

* Cette valeur est basée sur l'équation (3.7), tandis que la valeur obtenue en se basant sur (3.8) est 0,01610.

Pour estimer le *biais agrégé* de chacun des quatre estimateurs de μ, nous calculons la moyenne sur 100 000 simulations de la différence entre l'estimation pour un échantillon de taille 500 et μ. L'estimation du *biais agrégé* produit par $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$ et $\hat{\mu}_{psw}$ sont 0,01471, 0,00722, 0,01461 et 0,00874, respectivement, si l'on utilise les valeurs de p_k pour 1990. Donc, l'utilisation des p_k réduit de 51 % le biais de l'estimateur non stratifié a posteriori et de 40 % celui de l'estimateur stratifié a posteriori.

L'utilisation des valeurs de p_k pour 1980 produit des résultats comparables. Les estimations du *biais abrégé* produit par $\hat{\mu}_w$ et $\hat{\mu}_{psw}$ sont 0,00578 et 0,00856, respectivement, si l'on se sert des valeurs de p_k pour 1980. Cependant, ces résultats ne sont pas résumés dans les tableaux.

Le *biais de téléphone* présenté dans le tableau 2 est le biais obtenu lorsque l'on échantillonne la population totale

d'abonnés au téléphone, U_T, pour calculer $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$ et $\hat{\mu}_{psw}$. Ce biais est dû au fait que l'on échantillonne U_T plutôt que U. Dans la suite de l'exemple, nous adoptons la convention de mentionner entre parenthèses les estimations fondées sur les p_k de 1980 si ces estimations diffèrent de celles calculées en se fondant sur les p_k de 1990. Les valeurs du *biais de téléphone* sont 0,01472, 0,00720 (0,00577), 0,01463 et 0,00850 (0,00838), et s'approchent donc de celles du *biais agrégé*.

Le *biais de deuxième phase*, qui correspond à la différence entre le *biais agrégé* et le *biais de téléphone*, est dû au fait que l'estimateur est l'approximation d'un ratio. Les estimations de ce *biais de deuxième phase*, tenant compte de l'erreur d'arrondissement, pour $\bar{y}_1/\bar{y}_2$ $\hat{\mu}_w$, $\hat{\mu}_{ps}$ et $\hat{\mu}_{psw}$ sont 0,00000, 0,00002 (0,00001), - 0,00002 et 0,00024 (0,00018), respectivement. Donc, dans notre exemple, le *biais de deuxième phase* est négligeable comparativement au *biais de téléphone*.

Les estimations du *biais théorique*, basées sur (3.2) de $\bar{y}_1/\bar{y}_2$ et $\hat{\mu}_{ps}$ sont 0,00777 (0,00905) et 0,00663 (0,00678), respectivement. Ces valeurs diffèrent de celles du *biais agrégé*, puisque (3.2) se base sur toutes les populations d'abonnés au téléphone possibles, tandis que le *biais agrégé* est subordonné à la réalisation en question de la population d'abonnés au téléphone. Le biais théorique est calculé en se basant sur le modèle voulant que chaque ménage soit caractérisé par une probabilité p_k de posséder le téléphone et sa valeur dépend donc du modèle utilisé pour l'ajustement au p_k. Puisque $\hat{\mu}_w$ et $\hat{\mu}_{psw}$ sont asymptotiquement non biaisés, leur *biais théorique* est défini comme étant nul.

Les valeurs de *l'écart type simulé* des 100 000 estimations simulées de μ pour $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$ et $\hat{\mu}_{psw}$ sont 0,01683, 0,01737 (0,01734), 0,01605 et 0,01643 (0,01634). Ces quatre nombres sont assez proches des valeurs de *l'écart type estimé*, qui sont égaux à la racine carrée de la variance estimative moyenne de l'estimateur de μ, basée sur(3.1), (3.6) et (3.7). Précisément, ces valeurs de *l'écart type estimé* sont 0,01680, 0,01734 (0,01732), 0,01601 et 0,01635 (0,01628), respectivement. L'écart type estimé de rechange, fondé sur (3.8), de $\hat{\mu}_{psw}$ est 0,01610 (0,01606), valeur qui, de nouveau, s'approche de la valeur 0,01635 (0,01628). Les estimations de *l'écart type théorique* sont 0,01700 (0,01697), 0,01752 (0,01749), 0,01617 (0,01621) et 0,01658 (0,01653), si l'on se fonde sur le PUMS au 1/20 entier pour la Virginie pour 1990 et sur les équations (4.3), (4.4) et (4.5). Ces valeurs de l'*écart type théorique* s'approchent aussi des autres écarts types calculés.

L'utilisation des p_k réduit de 51 % (61 %) le *biais agrégé* de l'estimateur non stratifié a posteriori et de 40 % (41 %) celui de l'estimateur stratifié a posteriori. Toutefois, l'écart-type augmente légèrement. Les estimations de la racine carrée de l'erreur quadratique moyenne sont calculées d'après le *biais agrégé* et *l'écart-type simulé* des estimateurs $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$ et $\hat{\mu}_{psw}$ sont 0,02236 (0,02236), 0,01881 (0,01828), 0,02171 (0,02171) et 0,01861 (0,01844), respectivement. Donc, l'utilisation des p_k réduit de 29 % (33 %) l'EQM de l'estimateur non stratifié a posteriori et de 27 % (28 %) celle de l'estimateur stratifié a posteriori. Notons qu'en ce qui concerne l'EQM, l'écart entre $\bar{y}_1/\bar{y}_2$ et $\hat{\mu}_{ps}$, ainsi qu'entre $\hat{\mu}_w$ et $\hat{\mu}_{psw}$, est faible. Par conséquent, la stratification a posteriori produit peu d'amélioration.

6.2 Estimation du nombre moyen d'automobiles par ménage

Calculé d'après le PUMS au 1/20 entier pour la Virginie pour 1990, le nombre moyen d'automobiles par ménage est égal à 1,80397. La stratification a posteriori a été réalisée selon le revenu du ménage, en utilisant les catégories {moins de 20 000$, au moins 20 000$ mais moins de 35 000$ et au moins 35 000$}. Nous estimons de nouveau les p_k, mais celle fois-ci, nous excluons la covariable « nombre d'automobiles » du modèle GLIM ajusté au PUMS de 1990, puisque le nombre moyen d'automobiles est la variable que nous estimons.

Comme le montre le tableau 3, les estimations du *biais agrégé* sur 100 000 simulations d'échantillons aléatoires simples de taille égale à 500 sont 0,04872, 0,01629, 0,02226 et 0,01471, et celles du *biais de téléphone* sont 0,04872, 0,01623, 0,02220 et 0,01458 pour les estimateurs $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$ et $\hat{\mu}_{psw}$, respectivement. Par conséquent, les valeurs du *biais de deuxième phase* sont assez faibles. L'utilisation des p_k réduit de 67 % le biais de l'estimateur non stratifié a posteriori et de 34 % celui produit par l'estimateur stratifié a posteriori. Il se pourrait que cette dernière valeur de la proportion de biais éliminée soit plus faible que la première parce que le revenu est un prédicteur puissant de l'abonnement ou non d'un ménage au service téléphonique (consulter Groves 1989, pages 116 à 119; Thornberry et Massey 1988), et que les groupes résultant de la stratification a posteriori pour la détermination de $\hat{\mu}_{ps}$ et $\hat{\mu}_{psw}$ sont fondés sur le revenu.

Les écarts-types des simulations sont 0,04694, 0,04764, 0,04162 et 0,04172, respectivement. Les valeurs de la racine de l'erreur quadratique moyenne pour les quatre estimateurs sont égales à environ 0,06765, 0,05035, 0,04720 et 0,04424, respectivement, si bien que l'utilisation des p_k réduit de 45 % l'EQM de l'estimateur non stratifié a posteriori et de 12 % celle de l'estimateur stratifié a posteriori.

Nous avons également réalisé des simulations, dont les résultats ne sont pas résumés dans les tableaux, où nous avons retenu la variable « nombre d'automobiles » lors de l'ajustement du modèle GLIM au PUMS de 1990. Les estimations du *biais agrégé* pour les estimateurs $\hat{\mu}_w$ et $\hat{\mu}_{psw}$ sont 0,00116 et 0,00006, respectivement, c'est-à-dire des valeurs nettement plus faibles que 0,01629 et 0,01471, qui sont les valeurs respectives du *biais relatif* obtenues en éliminant la variable « nombre d'automobiles » lors de l'ajustement du modèle GLIM. En outre, nous estimons

qu'il est nécessaire d'éliminer la variable étudiée (ici le nombre d'automobiles) du modèle GLIM ajusté au PUMS pour que l'analyse soit appropriée.

Tableau 3
Biais et écarts-types des estimations du nombre moyen d'automobiles par ménage

	Estimateur			
	Non stratifié a posteriori		Stratifié a posteriori	
	$\bar{y}_1/\bar{y}_2$	$\hat{\mu}_w$	$\hat{\mu}_{ps}$	$\hat{\mu}_{psw}$
Biais agrégé	0,04872	0,01629	0,02226	0,01471
Biais téléphone	0,04872	0,01623	0,02220	0,01458
Biais deuxième phase	0,00000	0,00006	0,00006	0,00013
Biais théorique	0,03388	0	0,00859	0
Écart-type simulé	0,04694	0,04764	0,04162	0,04172
Écart-type estimé	0,04682	0,04753	0,04148	0,04158*
Écart-type théorique	0,04715	0,04791	0,04152	0,04161
Racine de l'erreur quadratique moyenne	0,06765	0,05035	0,04720	0,04424

Le nombre moyen réel d'automobiles par ménage est égal à 1,80397. La stratification a posteriori est fondée sur le revenu. La taille des échantillons sélectionnés est de 500 et le nombre de simulations, de 100 000.

* Cette valeur du nombre moyen d'automobiles est fondée sur (3.7), tandis que la valeur basée sur (3.8) est égale à 0,04142.

6.3 Estimation du revenu moyen des ménages

Calculé pour le PUMS au 1/20 entier pour la Virginie en 1990, le revenu moyen des ménages est égal à 40 187$. De nouveau, nous estimons les p_k mais cette fois-ci, c'est la covariable « revenu » que nous excluons du modèle GLIM ajusté au PUMS pour 1990, puisque c'est le revenu moyen des ménages que nous estimons.

Au tableau 4, pour estimer la valeur du revenu des ménages d'après un échantillon aléatoire simple de taille égale à 500 et d'après $\hat{\mu}_{ps}$ ou $\hat{\mu}_{psw}$, nous n'effectuons une stratification a posteriori que selon la race (noire, autre) du chef du ménage. Les estimations du *biais agrégé* basées sur 100 000 simulations sont 1 412$, 640$, 1 192$ et 633$, respectivement, et celles du *biais de téléphone* sont 1 414$, 640$, 1 193$ et 630$ pour les estimateurs $\bar{y}_1/\bar{y}_2$, $\hat{\mu}_w$, $\hat{\mu}_{ps}$ et $\hat{\mu}_{psw}$, respectivement. Donc, les valeurs du *biais de deuxième phase* sont faibles comparativement à celles du *biais de téléphone*. Dans l'ensemble, l'utilisation des p_k réduit de 55 % le biais produit par l'estimateur non stratifié a posteriori et de 47 % celui produit par l'estimateur stratifié a posteriori.

Les valeurs de l'écart-type des simulations sont 1 534$, 1 518$, 1 502$ et 1 488$, respectivement. Donc, les valeurs de la racine de l'erreur quadratique moyenne pour les quatre estimateurs sont environ égales à 2 085$, 1 647$, 1 918$ et 1 617 $, respectivement, si bien que l'utilisation des p_k réduit de 38 % l'EQM de l'estimateur non stratifié a posteriori et de 29 % celle de l'estimateur stratifié a posteriori. Si l'on prend l'EQM comme critère, l'amélioration due à la stratification a posteriori est moins importante.

Au tableau 5, nous estimons de nouveau le revenu du ménage, mais celle fois-ci, nous procédons à une stratification a posteriori selon le sexe (masculin, féminin), l'âge (moins de 45 ans, au moins 45 ans) et la race (noire, autre) du chef du ménage. Notons que cette nouvelle stratification a posteriori n'affecte pas les estimateurs non stratifiés a posteriori. Les estimations du *biais agrégé* fondées sur 100 000 simulations sont 1 173$ et 757$, et celles du *biais de téléphone* sont égales à 1 177$ et 747$ pour les estimateurs stratifiés a posteriori $\hat{\mu}_{ps}$ et $\hat{\mu}_{psw}$, respectivement. De nouveau, les valeurs du *biais de deuxième phase* sont faibles comparativement à celles du *biais de téléphone*. L'utilisation des p_k réduit de 35 % le biais dû simplement à la stratification a posteriori.

Pour l'estimateur stratifié a posteriori, le biais théorique est égal à 463$. Les valeurs de l'écart-type des simulations sont 1 445$ et 1 435$ pour les estimateurs $\hat{\mu}_{ps}$ et $\hat{\mu}_{psw}$, respectivement. Les racines carrées de l'erreur quadratique moyenne sont égales à 1 861$ et 1 622$ pour les estimateurs $\hat{\mu}_{ps}$ et $\hat{\mu}_{psw}$, respectivement. Donc, l'utilisation des p_k réduit de 24 % la valeur de l'EQM pour l'estimateur stratifié a posteriori.

La valeur de l'EQM de $\hat{\mu}_{psw}$ est à peu près la même au tableau 4 et au tableau 5. Cependant, l'EQM de $\hat{\mu}_{ps}$ diminue légèrement lorsque l'on passe du tableau 4 au tableau 5.

Tableau 4
Biais et écart-types des estimations du revenu des ménages, stratification a posteriori selon la race

	Estimateur			
	Non stratifié a posteriori		Stratifié a posteriori	
	$\bar{y}_1/\bar{y}_2$	$\hat{\mu}_w$	$\hat{\mu}_{ps}$	$\hat{\mu}_{psw}$
Biais agrégé	1 412$	640$	1 192$	633$
Biais téléphone	1 414$	640$	1 193$	630$
Biais deuxième phase	-2$	0$	-2$	3$
Biais théorique	789$	0$	586$	0$
Écart-type simulé	1 534$	1 518$	1 502$	1 488$
Écart-type estimé	1 537$	1 521$	1 506$	1 491$*
Écart-type théorique	1 535$	1 518$	1 503$	1 488$
Racine de l'erreur quadratique moyenne	2 085$	1 647$	1 918$	1 617$

La valeur réelle du revenu moyen des ménages est 40 187$. Notons que $\bar{y}_1/\bar{y}_2$ et $\hat{\mu}_w$ ne dépendent pas de la stratification a posteriori, si bien que les résultats obtenus ici sont les mêmes que ceux obtenus au tableau 5. La taille des échantillons sélectionnés est égale à 500 et le nombre de simulations est égal à 100 000.

* Cette valeur est fondée sur (3.7), tandis que la valeur fondée sur (3.8) est 1 490$.

Tableau 5
Biais et écarts-types des estimations du revenu des ménages, stratification a posteriori selon le sexe, l'âge et la race

	Estimateur			
	Non stratifié a posteriori		Stratifié a posteriori	
	$\bar{y}_1/\bar{y}_2$	$\hat{\mu}_w$	$\hat{\mu}_{ps}$	$\hat{\mu}_{psw}$
Biais agrégé	1 412$	640$	1 173$	757$
Biais téléphone	1 414$	640$	1 177$	747$
Biais deuxième phase	-2$	0$	-4$	10$
Biais théorique	789$	0$	463$	0$
Écart-type simulé	1 534$	1 518$	1 445$	1 435$
Écart-type estimé	1 537$	1 521$	1 448$	1 438$*
Écart-type théorique	1 535$	1 518$	1 440$	1 430$
Racine de l'erreur quadratique moyenne	2 085$	1 647$	1 861$	.1 622$

La valeur réelle du revenu moyen des ménages est 40 187$. Notons que $\bar{y}_1/\bar{y}_2$ et $\hat{\mu}_w$ ne dépendent pas de la stratification a posteriori, si bien que les résultats obtenus ici sont les mêmes que ceux obtenus au tableau 4. La taille des échantillons sélectionnés est égale à 500 et le nombre de simulations est égal à 100 000.

* Cette valeur est fondée sur (3.7), tandis que la valeur fondée sur (3.8) est 1 421$.

7. DISCUSSION

Nous proposons ici d'utiliser de grandes bases de données à grande diffusion (c'est-à-dire les PUMS) pour élaborer un modèle de la propension p_k d'un ménage à avoir le téléphone. Nous avons utilisé, pour la Virginie en 1990, un modèle GLIM avec une fonction de lien bilogarithmique et les variables indépendantes « nombre de personnes », « mode d'occupation du logement », « date d'entrée dans le logement », « nombre d'automobiles », « revenu du ménage », « langue » et « race ».

Nous avons proposé d'utiliser les coefficients de pondération en fonction de l'abonnement au téléphone p_k pour réduire le biais de non-couverture qui entache les estimateurs en cas d'enquêtes téléphoniques. Nous pouvons nous attendre à ce que ce biais se manifeste lorsque la variable étudiée est liée à la possession d'un téléphone. Les exemples que nous avons choisis sont tous des variables de ce type et, par conséquent, les améliorations observées lorsque l'on utilise des coefficients de pondération en fonction de l'abonnement au téléphone sont plus importantes que l'on ne s'y attendrait pour des variables pour lesquelles l'association à la possession d'un téléphone est faible.

Les coefficients de pondération peuvent être combinés à la stratification a posteriori. Nous avons constaté que l'utilisation de ce genre de coefficient de pondération réduit fortement le biais qui entache l'estimateur non stratifié a posteriori ainsi que l'estimateur stratifié a posteriori.

Il faut, pour procéder à la stratification a posteriori, que la taille de l'échantillon soit suffisamment grande pour que la probabilité d'être vide soit négligeable pour toute strate a posteriori. Comme nos expériences étaient basées sur des échantillons de taille égale à 500, le nombre de strates a posteriori était assez limité. Il est certain que, si l'on disposait d'un échantillon suffisamment grand pour pouvoir stratifier a posteriori en fonction des mêmes variables indépendantes que celles utilisées pour élaborer les p_k, l'utilisation de coefficients de pondération en fonction de l'abonnement au service téléphonique n'offrirait qu'une amélioration négligeable comparativement à la stratification a posteriori. Cependant, nombre de sondages d'opinion téléphoniques de portée nationale sont réalisés auprès d'un échantillon d'environ 1 000 personnes et nous sommes convaincus que, pour des échantillons de cette taille, l'application de coefficients de pondération en fonction de l'abonnement au téléphone produirait une réelle amélioration.

Nous avons également présenté les résultats obtenus en appliquant aux données de 1990 les coefficients de pondération en fonction de l'abonnement au téléphone calculés d'après le PUMS de 1980 en corrigeant pour l'inflation les catégories définies pour la variable de revenu du ménage. Les résultats sont comparables à ceux obtenus en utilisant les coefficients de pondération calculés d'après le PUMS de 1990. Par conséquent, bien que les données des PUMS ne soient produites que tous les dix ans et qu'elles puissent remonter jusqu'à douze ans dans le passé, l'utilisation de modèles de la propension à être abonné au téléphone élaborés d'après les anciens ensembles de données des PUMS permet de réduire considérablement le biais dû à l'échantillonnage téléphonique, à condition que les catégories soient corrigées comme il convient pour l'inflation.

Enfin, les PUMS sont ventilés par État et par grande région métropolitaine, de sorte que l'on peut créer des modèles pondérés en fonction de l'abonnement au téléphone distincts pour les grandes unités géographiques, mesure qui semble appropriée en cas de grande enquête.

REMERCIEMENTS

Les auteurs sont très reconnaissants à un évaluateur anonyme de ses nombreuses suggestions constructives.

ANNEXE : ÉTABLISSEMENT DES ÉQUATIONS

Avant de montrer comment ont été obtenues les équations présentées aux sections 3 et 4, nous devons supposer que les séries $\{\alpha_{i1}, \alpha_{i2}, ...\}$, pour $i = 1, 2$ obéissent à certaines conditions de régularité. En outre, nous devons prouver certains lemmes. Ensuite, nous pourrons produire les équations des estimateurs $\hat{\mu}_{ps}$, $\hat{\mu}_w$ et $\hat{\mu}_{psw}$. Chaque fois que nous introduisons la variable d'erreur ξ_k dans ce qui suit, alors $\xi_k = O_p(1)$ et $E(\xi_k)^2 = O(1)$ quand $k \to \infty$. Afin de simplifier la notation

(mais en commettant un léger abus), nous permettrons que la série représentée par $\{\xi_1, \xi_2, \ldots\}$ diffère d'une équation à l'autre.

Condition A : Chaque α_{ik} représente une moyenne d'observations sur échantillon telle que $E\alpha_{ik} - \alpha_i = O(k^{-1})$, $E|\alpha_{ik} - \alpha_i|^3 = O(k^{-3/2})$ et $\alpha_{ik} - \alpha_i = O_p(k^{-1/2})$ lorsque $k \to \infty$ pour $i = 1, 2$. Posons que $\mu_k = \alpha_{1k}/\alpha_{2k}$ pour $k = 1, 2, \ldots$.

LEMME A.1 La condition A implique $E\mu_k - \mu = O(k^{-1})$ lorsque $k \to \infty$.

PREUVE : Définissons la fonction $f(\gamma_1, \gamma_2) = \gamma_1/\gamma_2$. En vertu d'un développement linéaire en série de Taylor,

$$\mu_k - \mu = \alpha_{1k}/\alpha_{2k} - \alpha_1/\alpha_2$$

$$= (\alpha_{1k} - \alpha_1)\frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_1} + (\alpha_{2k} - \alpha_2)\frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_2} + k^{-1}\xi_k$$

$$= (\alpha_{1k} - \alpha_1)(\alpha_2)^{-1} - (\alpha_{2k} - \alpha_2)\mu(\alpha_2)^{-2} + k^{-1}\xi_k.$$

Le résultat découle de la condition A.

Condition B : La série $\{\alpha_{i1}, \alpha_{i2}, \ldots\}$ pour $i = 1, 2$ satisfait

$$k^{1/2}\left[\begin{pmatrix}\alpha_{1k}\\ \alpha_{2k}\end{pmatrix} - \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}\right] \underline{\underline{d}}$$

$$N\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right) + \begin{pmatrix}k^{-1/2}\xi_{1k}\\ k^{-1/2}\xi_{2k}\end{pmatrix}$$

pour certaines constantes σ_1^2, σ_2^2 et ρ.

LEMME A.2 Si les conditions A et B sont remplies,

$$\text{EQM}\,\mu_k = (\alpha_2)^{-2}\,\text{var}(\alpha_{1k} - \mu\,\alpha_{2k}) + O(k^{-2}),$$

et

$$\text{var}\,\mu_k = (\alpha_2)^{-2}\,\text{var}(\alpha_{1k} - \mu\,\alpha_{2k}) + O(k^{-2}),$$

lorsque $k \to \infty$.

PREUVE : En vertu d'un développement linéaire en série de Taylor,

$$\begin{aligned}
\mu_k - \mu &= \alpha_{1k}/\alpha_{2k} - \alpha_1/\alpha_2\\
&= (\alpha_{1k} - \alpha_1)\frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_1} + (\alpha_{2k} - \alpha_2)\frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_2}\\
&\quad + \frac{1}{2}\left[(\alpha_{1k} - \alpha_1)^2\frac{\partial^2 f(\alpha_1, \alpha_2)}{\partial(\alpha_1)^2} + (\alpha_{2k} - \alpha_2)^2\right.\\
&\quad \left.\frac{\partial^2 f(\alpha_1, \alpha_2)}{\partial(\alpha_2)^2} + 2\,(\alpha_{1k} - \alpha_1)(\alpha_{2k} - \alpha_2)\frac{\partial^2 f(\alpha_1, \alpha_2)}{\partial\alpha_1\,\partial\alpha_2}\right]\\
&\quad + k^{-3/2}\xi_k\\
&= (\alpha_{1k} - \alpha_1)(\alpha_2)^{-1} - (\alpha_{2k} - \alpha_2)\mu(\alpha_2)^{-1}\\
&\quad + (\alpha_{2k} - \alpha_2)^2\mu(\alpha_2)^{-2} - (\alpha_{1k} - \alpha_1)(\alpha_{2k} - \alpha_2)(\alpha_2)^{-2}\\
&\quad + k^{-3/2}\xi_k\\
&= \alpha_2^{-1}(\alpha_{1k} - \mu\,\alpha_{2k})\left[1 - \alpha_2^{-1}(\alpha_{2k} - \alpha_2)\right] + k^{-3/2}\xi_k.
\end{aligned}$$

Par conséquent,

$$(\mu_k - \mu)^2 = \alpha_2^{-2}(\alpha_{1k} - \mu\,\alpha_{2k})^2\left[1 - 2\alpha_2^{-1}(\alpha_{2k} - \alpha_2)\right] + k^{-2}\xi_k,$$

ce qui implique

$$\text{EQM}\,\mu_k = \alpha_2^{-2}\,\text{var}(\alpha_{1k} - \mu\,\alpha_{2k}) - 2\alpha_2^{-3}$$
$$\text{cov}\,\{(\alpha_{1k} - \mu\,\alpha_{2k})^2, (\alpha_{2k} - \alpha_2)\} + k^{-2}\xi_k. \qquad \text{(A.1)}$$

Maintenant, nous allons montrer que le terme de covariance de (A.1) est asymptotiquement négligeable. Puisque

$$k^{1/2}(\alpha_{1k} - \mu\,\alpha_{2k}) \underline{\underline{d}}\; N(0, \sigma_3^2) + k^{-1/2}\xi_k$$

pour certaines constantes σ_3^2, alors

$$k(\alpha_{1k} - \mu\,\alpha_{2k})^2 \underline{\underline{d}}\; \sigma_3^2\chi_1^2 + k^{-1/2}\xi_k,$$

où χ_1^2 représente une variable chi carré aléatoire à un degré de liberté. En outre,

$$k^{1/2}(\alpha_{2k} - \alpha_2) \underline{\underline{d}}\; N(0, \sigma_2^2) + k^{-1/2}\xi_k.$$

Si les signes des α_{ik} sont inversés pour $i = 1, 2$, alors $k(\alpha_{1k} - \mu\,\alpha_{2k})^2$ ne change pas, mais $k^{1/2}(\alpha_{2k} - \alpha_2)$ est inversé. Par conséquent, par symétrie,

$$\text{cov}\{k(\alpha_{1k} - \mu\,\alpha_{2k})^2,\ k^{1/2}(\alpha_{2k} - \alpha_2)\} = O(k^{-1/2})$$

lorsque $k \to \infty$. Donc,

$$\text{cov}\{(\alpha_{1k} - \mu\,\alpha_{2k})^2, (\alpha_{2k} - \alpha_2)\} = O(k^2) \qquad \text{(A.2)}$$

lorsque $k \to \infty$. La première partie du lemme découle de la combinaison de (A.1) et (A.2). Puisque le lemme A.1 implique

$$\text{biais } u_k = O(k^{-1})$$

lorsque $k \to \infty$, alors la deuxième partie de ce lemme en découle.

Condition C : Si nous définissons $\alpha_{Ti} = \text{plim}_{n\to\infty}\hat{\alpha}_i$ étant donné U_T, l'estimateur $\hat{\alpha}_i$, de α_i satisfait les conditions suivantes pour $i = 1, 2$:

$$E\,(\hat{\alpha}_i \mid U_T) - \alpha_{Ti} = O(n^{-1});$$

$$\text{étant donné } U_T, \hat{\alpha}_i - \alpha_{Ti} = O_p(n^{-1/2});$$

et

$$E(\{|\hat{\alpha}_i - \alpha_{Ti}|^3\} \mid U_T) = O(n^{-3/2})$$

lorsque $k \to \infty$.

Condition D : Étant donné U_T,

$$n^{1/2}\left[\begin{pmatrix}\hat{\alpha}_1\\ \hat{\alpha}_2\end{pmatrix} - \begin{pmatrix}\alpha_{T1}\\ \alpha_{T2}\end{pmatrix}\right] \stackrel{d}{=}$$

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right) + \begin{pmatrix}n^{-1/2}\,\xi_n\\ n^{-1/2}\,\xi_n\end{pmatrix}$$

pour une certaine matrice définie positive Σ, où $\alpha_{Ti} = \text{plim}_{n\to\infty}\hat{\alpha}_i$ étant donné U_T. En outre,

$$E(\{|\hat{\alpha}_i - \alpha_{Ti}|^3\} \mid U_T) = O(n^{-3/2})$$

lorsque $n \to \infty$, pour $i = 1, 2$.

THÉORÈME A.1 Si les conditions C et D sont remplies, nous avons

$$\text{var}\,\hat{\mu} = \alpha_2^{-2}\,E\,\text{var}\,(\hat{\alpha}_1 - \mu_T\hat{\alpha}_2 \mid U_T) + O(n^{-2} + N^{-1})$$

lorsque $n \to \infty$, où $\mu_T = \alpha_{T1}/\alpha_{T2}$.

PREUVE : Pour commencer, nous déterminons $E\,\text{var}\,(\hat{\mu} \mid U_T)$. Sous la condition D, nous appliquons le lemme A.2 pour obtenir

$$\text{var}\,(\hat{\mu} \mid U_T) = \alpha_{T2}^{-2}\,\text{var}\,(\hat{\alpha}_1 - \mu_{U_T}\hat{\alpha}_2 \mid U_T) + n^{-2}\xi_n. \qquad \text{(A.3)}$$

Puisque

$$\alpha_{T2}^{-2} = (\alpha_2)^{-2} + N^{-1/2}\xi_n$$

et que

$$\text{var}\,(\hat{\alpha}_1 - \mu_T\hat{\alpha}_2 \mid U_T) = n^{-1}\xi_n,$$

alors (A.3) implique

$$E\,\text{var}(\hat{\mu} \mid U_T) =$$

$$\alpha_2^{-2}E\,\text{var}(\hat{\alpha}_1 - \mu_T\hat{\alpha}_2 \mid U_T) + O(n^{-2} + n^{-1}N^{-1/2}) \qquad \text{(A.4)}$$

lorsque $n \to \infty$. Maintenant, nous déterminons $\text{var}\,E(\hat{\mu} \mid U_T)$. La condition C et le lemme A.1 implique

$$E(\hat{\mu} \mid U_T) = \mu_T + n^{-1}\xi_n = \mu + (n^{-1} + N^{-1/2})\xi_n.$$

Donc,

$$\text{var}\,E(\hat{\mu} \mid U_T) = O(n^{-2} + N^{-1}) \qquad \text{(A.5)}$$

lorsque $n \to \infty$. Le résultat découle de la combinaison de (A.4) avec(A.5).

A.1 Estimateur stratifié a posteriori

Ici, nous établissons les équations liées à l'estimateur stratifié a posteriori, $\hat{\mu}_{ps}$, où $\hat{\alpha}_{ps(1)}$ et $\hat{\alpha}_{ps(2)}$ satisfont les conditions C et D. Notons que

$$E(\hat{\alpha}_{ps(i)} \mid U_T) = N^{-1}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh}N_{Tgh}^{-1}\sum_{k\in U_{Tgh}} y_{ik}$$

pour $i = 1, 2$, et définissons $\mu_{T,ps} = E(\hat{\alpha}_{ps(1)} \mid U_T)/E(\hat{\alpha}_{ps(2)} \mid U_T)$. Rappelons les définitions de α_i^* et μ^* dans (4.1) et (4.2).

Obtention de (4.3), la variance asymptotique de $\hat{\mu}_{ps}$:

Puisque

$$\text{var}\,(\hat{\alpha}_{ps(1)} - \mu_{T,ps}\,\hat{\alpha}_{ps(2)} \mid U_T, n_{gh})$$

$$= N^{-2}\sum_{h=1}^{H}\sum_{g=1}^{G}\frac{N_{gh}^2\,[1 - n_{gh}/N_{Tgh}]}{n_{gh}(N_{Tgh} - 1)}\sum_{k\in U_{Tgh}}$$

$$\left[y_{1k} - \mu_{T,ps}\,y_{2k} - N_{Tgh}^{-1}\sum_{j\in U_{Tgh}}(y_{1j} - \mu_{T,ps}y_{2j})\right]^2. \qquad \text{(A.6)}$$

Alors

$$E\,\text{var}\,(\hat{\alpha}_{ps(1)} - \mu_{T,ps}\,\hat{\alpha}_{ps(2)} \mid U_T, n_{gh})$$

$$= N^{-2}\sum_{h=1}^{H}\sum_{g=1}^{G}\frac{N_{gh}^2\left[\left(\sum_{j\in U_h} p_j\right) - n_h\right]}{n_h\left(\sum_{j\in U_{gh}} p_j\right)\left[\left(\sum_{j\in U_{gh}} p_j\right) - 1\right]}$$

$$\sum_{k\in U_{gh}} p_k\left[y_{1k} - \mu^* y_{2k} - \frac{\sum_{j\in U_{gh}} p_j(y_{1j} - \mu^* y_{2j})}{\sum_{j\in U_{gh}} p_j}\right]^2$$

$$+ O(n^{-2} + n^{-1}N^{-1/2}) \qquad \text{(A.7)}$$

lorsque $n \to \infty$. En outre, puisque

$$E(\hat{\alpha}_{ps(1)} - \mu_{T,ps}\hat{\alpha}_{ps(2)} \mid U_T, n_{gh})$$

$$= N^{-1}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh} N_{Tgh}^{-1} \sum_{k \in U_{Tgh}} (y_{1k} - \mu_{T,ps} y_{2k}), \qquad (A.8)$$

alors

$$E[\operatorname{var} E(\hat{\alpha}_{ps(1)} - \mu_{T,ps}\hat{\alpha}_{ps(2)} \mid U_T, n_{gh}) \mid U_T] = 0. \qquad (A.9)$$

Puisque le théorème A.1 et (A.7) impliquent

$$\operatorname{var}\hat{\mu}_{ps} = (\alpha_2^*)^{-2} E \operatorname{var}(\hat{\alpha}_{ps(1)} - \mu_{T,ps}\hat{\alpha}_{ps(2)} \mid U_T, n_{gh}) + O(n^{-2} + N^{-1}) \qquad (A.10)$$

lorsque $n \to \infty$, alors (A.9) implique (4.3).

Obtention de (3.1), la variance estimative de $\hat{\mu}_{ps}$:

Étant donné (A.6), nous avons l'estimateur

$$\widehat{\operatorname{var}}\left(n_{gh}^{-1}\sum_{k \in s_{gh}} \{y_{1k} - \mu_{T,ps}\, y_{2k}\} \mid U_T, n_{gh}\right)$$

$$= \frac{1 - n_{gh}/N_{Tgh}}{n_{gh}(n_{gh} - 1)} \sum_{k \in s_{gh}}$$

$$\left[y_{1k} - \mu_{T,ps} y_{2k} - n_{gh}^{-1}\sum_{j \in s_{gh}} (y_{1j} - \mu_{T,ps} y_{2j})\right]^2.$$

Le résultat découle de l'utilisation de (A.10).

Obtention de (3.2), biais estimatif de $\hat{\mu}_{ps}$:

Le lemme A.1 implique

$$\hat{\mu}_{ps} - \mu^* = O(n^{-1})$$

lorsque $n \to \infty$. Puisque

$$E\,\hat{\alpha}_{ps(i)} = \alpha_i^* + O(N^{-1})$$

lorsque $N \to \infty$ pour $i = 1, 2$, le résultat s'ensuit.

A.2 Estimateur pondéré en fonction de l'abonnement au téléphone

Ici nous obtenons les équations liées à l'estimateur pondéré en fonction de l'abonnement au téléphone, $\hat{\mu}_w$, sous les conditions C et D, où $\tilde{\alpha}_{w(1)}$ et $\tilde{\alpha}_{w(2)}$ satisfont les conditions C et D. Notons que

$$E(\tilde{\alpha}_{w(i)} \mid U_T) = N^{-1}\sum_{k \in U_T} y_{ik}/p_k$$

pour $i = 1, 2$, et définissons $\mu_{T,w} = E(\tilde{\alpha}_{w(1)} \mid U_T)/E(\tilde{\alpha}_{w(2)} \mid U_T)$.

Obtention de (4.4), la variance asymptotique de $\hat{\mu}_{ps}$:

Puisque

$$\operatorname{var}\left(\tilde{\alpha}_{w(1)} - \mu_{T,w}\,\tilde{\alpha}_{w(2)} \mid U_T\right)$$

$$= N^{-2}\sum_{h=1}^{H} \frac{(N_{Th} - n_h) N_{Th}}{n_h (N_{Th} - 1)} \sum_{k \in U_{Th}}$$

$$\left[\frac{y_{1k} - \mu_{T,w} y_{2k}}{p_k} - N_{Th}^{-1}\sum_{j \in U_{Th}} \frac{y_{1j} - \mu_{T,w} y_{2j}}{p_j}\right]^2,$$

alors

$$E \operatorname{var}\left(\tilde{\alpha}_{w(1)} - \mu_{T,w}\,\tilde{\alpha}_{w(2)} \mid U_T\right)$$

$$= N^{-2}\sum_{h=1}^{H} \frac{\left[\left(\sum_{j \in U_h} p_j\right) - n_h\right]\left(\sum_{j \in U_h} p_j\right)}{n_h\left[\left(\sum_{j \in U_h} p_j\right) - 1\right]} \qquad (A.11)$$

$$\sum_{j \in U_h} p_k \left[\frac{y_{1k} - \mu y_{2k}}{p_k} - \frac{\sum_{j \in U_h} (y_{1j} - \mu y_{2j})}{\sum_{j \in U_h} p_j}\right]^2$$

$$+ O(n^{-1} N^{-1/2}) \qquad (A.12)$$

lorsque $n \to \infty$. Le résultat découle de l'application du théorème A.1 à (A.11).

Obtention de (3.6), la variance estimative de $\hat{\mu}_{ps}$:

Étant donné le théorème A.1, une estimation valide de $\operatorname{var}\hat{\mu}_w$ estime aussi

$$(\alpha_2)^{-2} E \operatorname{var}\left(\hat{\alpha}_{w(1)} - \mu_{T,w}\hat{\alpha}_{w(2)} \mid U_T\right),$$

qui est équivalent à

$$(\alpha_2)^{-2} E \operatorname{var}\left(N^{-1}\sum_{h=1}^{H} \frac{N_{Th}}{n_h}\sum_{k \in s_h} \frac{y_{1k} - \mu_{T,w} y_{2k}}{p_k} \mid U_T\right).$$

Le résultats s'ensuit.

A.3 Estimateur stratifié a posteriori pondéré en fonction de l'abonnement au téléphone

Ici, nous obtenons les équations liées à l'estimateur stratifié a posteriori, $\hat{\mu}_{psw}$, sous les conditions C et D, où $\hat{\alpha}_{psw(1)}$ et $\hat{\alpha}_{psw(2)}$ satisfont les conditions C et D. Le lemme A.1 implique

$$E(\hat{\alpha}_{psw(i)} \mid U_T) = N^{-1}\sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh} \frac{\sum_{k \in U_{Tgh}} p_k^{-1} y_{ik}}{\sum_{k \in U_{Tgh}} p_k^{-1}} + n^{-1}\xi_n$$

pour $i = 1, 2$, et nous définissons $\mu_{T,\text{psw}} = E(\hat{\alpha}_{\text{psw}(1)} | U_T)/E(\hat{\alpha}_{\text{psw}(2)} | U_T)$.

Obtention de (4.5), la variance asymptotique de $\hat{\mu}_{\text{psw}}$:
Il découle du lemme A.2 que

$$\text{var}\left(\frac{\sum_{k \in s_{gh}} p_k^{-1} \{y_{1k} - \mu_{T,\text{psw}} y_{2k}\}}{\sum_{k \in s_{gh}} p_k^{-1}} | U_T, n_{gh} \right)$$

$$= \frac{1 - n_{gh}/N_{Tgh}}{n_{gh}(N_{Tgh} - 1)} \left(N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} p_j^{-1} \right)^{-2}$$

$$\sum_{k \in U_{Tgh}} p_k^{-2} \left[y_{1k} - \mu_{\text{psw},U_T} y_{2k} - \left(N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} p_j^{-1} \right)^{-1} \right.$$

$$\left. \left(N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} \left\{ \frac{y_{1j} - \mu_{T,\text{psw}} y_{2j}}{p_j} \right\} \right) \right]^2 + n^{-2} \xi_n.$$

Par conséquent,

$$E\left[\text{var}\left(\frac{\sum_{k \in s_{gh}} p_k^{-1} \{y_{1k} - \mu_{T,\text{psw}} y_{2k}\}}{\sum_{k \in s_{gh}} p_k^{-1}} | U_T, n_{gh} \right) | U_T \right.$$

$$= \frac{(N_{Th} - n_h)}{n_h N_{Tgh}(N_{Tgh} - 1)} \left(N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} p_j^{-1} \right)^{-2}$$

$$\sum_{k \in U_{Tgh}} p_k^{-2} \left[y_{1k} - \mu_{T,\text{psw}} y_{2k} - \left(N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} p_j^{-1} \right)^{-1} \right.$$

$$\left. \left(N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} \left\{ \frac{y_{1j} - \mu_{T,\text{psw}} y_{2j}}{p_j} \right\} \right) \right]^2 + n^{-2} \xi_n. \tag{A.13}$$

Puisque

$$E\, N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} p_j^{-1} = \left(\sum_{k \in U_{gh}} p_k \right)^{-1} N_{gh} + O(N^{-1})$$

lorsque $N \to \infty$, alors (A.13) implique l'espérance inconditionnelle

$$E\, \text{var}\left(\frac{\sum_{k \in s_{gh}} p_k^{-1} \{y_{1k} - \mu_{T,\text{psw}} y_{2k}\}}{\sum_{k \in s_{gh}} p_k^{-1}} | U_T, n_{gh} \right)$$

$$= \frac{N_{gh}^{-2} \left(\sum_{j \in U_{gh}} p_j \right) \left[\left(\sum_{j \in U_h} p_j \right) - n_h \right]}{n_h \left[\left(\sum_{j \in U_{Tgh}} p_j \right) - 1 \right]}$$

$$\sum_{k \in U_{gh}} p_k^{-1} \left[y_{1k} - \mu y_{2k} - N_{gh}^{-1} \sum_{j \in U_{gh}} (y_{1j} - \mu y_{2j}) \right]^2$$

$$+ O(n^{-2} + N^{-1}) \tag{A.14}$$

lorsque $n \to \infty$. En vertu du théorème A.1,

$$\text{var}\, \hat{\mu}_{\text{psw}}$$

$$= \alpha_2^{-2} E\, \text{var}(\hat{\alpha}_{\text{psw}(1)} - \mu_{T,\text{psw}} \hat{\alpha}_{\text{psw}(2)} | U_T, n_{gh}, \forall g, h)$$

$$+ \alpha_2^{-2} E\, \text{var}\left[E(\hat{\alpha}_{\text{psw}(1)} - \mu_{T,\text{psw}} \hat{\alpha}_{\text{psw}(2)} | \right.$$

$$\left. U_T, n_{gh}, \forall g, h) | n_{gh}, \forall g, h \right] + O(n^{-2} + N^{-1}) \tag{A.15}$$

lorsque $n \to \infty$. Le lemme A.1 implique

$$\text{var}\left[E(\hat{\alpha}_{\text{psw}(1)} - \mu_{T,\text{psw}} \hat{\alpha}_{\text{psw}(2)} | \right.$$

$$\left. U_T, n_{gh}, \forall g, h) | n_{gh}, \forall g, h \right] = n^{-2} \xi_n. \tag{A.16}$$

Puisque (4.1) implique :

$$\text{var}(\hat{\alpha}_{\text{psw}(1)} - \mu_{T,\text{psw}} \hat{\alpha}_{\text{psw}(2)} | U_T, n_{gh})$$

$$= N^{-2} \sum_{h=1}^{H} N_h^2$$

$$\text{var}\left(N_h^{-1} \sum_{g=1}^{G} N_{gh} \frac{\sum_{k \in s_{gh}} p_k^{-1} \{y_{1k} - \mu_{T,\text{psw}} y_{2k}\}}{\sum_{k \in s_{gh}} p_k^{-1}} | U_{T,n_{gh}} \right),$$

alors (A.14), (A.15) et (A.16) implique le résultat.

Obtention de (3.7), la variance estimée de $\hat{\mu}_{\text{psw}}$:
Observons que

$$\frac{N_{Tgh} \sum_{j \in U_h} p_j}{n_{gh} n_h} \approx \frac{N_{Tgh} N_{Th}}{n_{gh} n_h} \approx \left[\frac{N_{gh}}{\sum_{k \in s_{gh}} p_k^{-1}} \right]^2.$$

Notons (4.5), le résultat s'ensuit.

Obtention de (3.8), autre estimation de la variance de $\hat{\mu}_{psw}$:

Puisque

$$\frac{N_{gh}}{\sum_{k \in s_{gh}} p_k^{-1}} \approx \frac{N_{Tgh}}{n_{gh}} \approx \frac{N_{Th}}{n_h} \approx \frac{N_h}{\sum_{k \in s_h} p_k^{-1}},$$

le résultat découle de (3.7).

BIBLIOGRAPHHIE

BRICK, J.M., WAKSBERG, J. et KEETER, S. (1994). Evaluating the use of data on interruptions in telephone service for nontelephone households. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 19-28.

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

GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.

KEETER, S. (1995). Estimating telephone noncoverage bias with a telephone survey. *Public Opinion Quarterly*, 59, 196-217.

KHURSHID, A., et SAHAI, H. (1995). A bibliography on telephone survey methodology. *Journal of Official Statistics*, 11, 325-367.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

McCULLAGH, P., et NELDER, J.A. (1991). *Generalized Linear Models*. New York: Chapman and Hall.

RAO, J.N.K. (1997). Developments in sample survey theory: An appraisal. *Canadian Journal of Statistics*, 25, 1-21.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

SÄRNDAL, C.-E., SWENSSON, B. et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

STEEH, C.G., GROVES, R.M., COMMENT, R. et HANSMIRE, E. (1983). Report on the survey research center's surveys of consumer attitudes. *Incomplete Data in Sample Surveys*, (Éd. W.G. Madow, H. Nisselson et I. Olkin), Academic Press, New York, 1.

THORNBERRY, JR., O.T., et MASSEY, J.T. (1988). Trends in United States telephone coverage across time and subgroups. *Telephone Surveys*, (Éds. R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, et J. Waksberg). New York: John Wiley & Sons, Inc., 25-49.

Techniques d'enquête, juin 2002
Vol. 28, n° 1, pp. 83-91
Statistique Canada

Estimation sans biais par calage sur la répartition dans les plans simples sans remise

YVES TILLÉ[1]

RÉSUMÉ

L'estimateur post-stratifié a parfois des post-strates vides. Pour pallier ce problème, on construit un estimateur post-stratifié dont les tailles des post-strates sont fixées dans l'échantillon. Les tailles des post-strates sont alors aléatoires dans la population. Ensuite, on construit un estimateur lissé en effectuant une moyenne mobile des estimateurs post-stratifiés. Cette technique permet de construire une théorie exacte du calage sur la répartition. L'estimateur obtenu est non seulement calé sur la répartition, il est linéaire, et exactement sans biais. On compare ensuite l'estimateur calé à l'estimateur par la régression. On propose enfin un estimateur approché de la variance validé au moyen de simulations.

MOTS CLÉS : Estimation sans biais; calage sur une fonction de répartition; probabilités d'inclusion conditionnelles; pondération.

1. INTRODUCTION

Lors d'une enquête par sondage, il arrive que l'on connaisse les valeurs d'un caractère auxiliaire pour toutes les unités de la population. Cette information peut être disponible quand les unités sont sélectionnées dans une base de données qui contient d'autres variables d'intérêt. On est alors tenté de caler les résultats d'une enquête sur cette information auxiliaire. Soit, on ne retient de cette variable auxiliaire que certaines fonctions (moments, effectifs) en vue d'utiliser une méthode de calage (voir par exemple Deville et Särndal 1992 ou Estevao, Hidiroglou et Särndal 1995), soit on peut découper cette variable en classes en vue d'utiliser un estimateur post-stratifié.

Si l'on opte pour l'estimateur post-stratifié, le découpage en strates s'avère délicat. Théoriquement, les strates doivent être définies avant la sélection de l'échantillon. Où faut-il placer les bornes des post-strates ? De quelles tailles doivent être les post-strates ? Cette dernière question est la plus embarrassante, car le problème principal de la post-stratification est la possibilité d'obtenir des post-strates vides. Les tailles des post-strates doivent donc être suffisamment grandes pour que la probabilité d'obtenir une post-strate vide soit négligeable. Ces problèmes ne se limitent pas aux estimateurs post-stratifiés, En effet, les estimateurs par la régression ou calés peuvent également ne pas exister pour certains échantillons.

Notre objectif est de définir une nouvelle méthode permettant d'utiliser l'information auxiliaire dans la population. Cette méthode est basée sur la définition de post-strates pour lesquelles le nombre d'unités est fixé dans l'échantillon, et non dans la population. On peut ainsi importer dans l'estimateur une information auxiliaire complexe issue de la connaissance de toutes les valeurs prises par la variable auxiliaire, tout en évitant à la fois le problème de la définition des bornes des post-strates et le problème des post-strates vides.

Cet article est organisé comme suit : En section 2, la notation est définie, en section 3, on donne le principe du conditionnement par les rangs, qui permet de définir des estimateurs sans biais en section 4. Ensuite, en section 5, on définit l'estimateur lissé, et un cas particulier est examiné en détail en section 6. En section 7, on donne une application à l'estimation d'une fonction de répartition. En section 8, on compare ce nouvel estimateur à l'estimateur par la régression et l'estimateur du plan simple sans remise. Le calcul de variance est abordé en section 9. Suite à l'impossibilité de donner une solution exacte, on donne une approximation en section 10, qui est testée par des simulations en section 11. Enfin, des conclusions générales sont données en section 12.

2. NOTATION

On suppose que l'on a une population composée de N unités d'observation dont les étiquettes des observations sont notées $\{1, ..., k, ..., N\}$. Dans cette population, on s'intéresse à un caractère d'intérêt $Y_k, k \in U$. L'objectif consiste à estimer le total $Y = \sum_{k \in U} Y_k$. On sélectionne un échantillon aléatoire S de taille fixe n au moyen d'un plan aléatoire simple sans remise. On note I_k la variable aléatoire indicatrice qui prend la valeur 1 si l'unité k est dans l'échantillon et 0 sinon. Les probabilités d'inclusion d'ordre un sont donc définies par $\Pr(k \in S) = \pi_k = n/N, k \in U$, et les probabilités d'inclusion d'ordre deux par $\Pr(k, l \in S) = \pi_{kl} = n(n-1)/(N(N-1)), k \neq l \in U$.

On s'intéressera à la classe des estimateurs linéaires de Y qui s'écrit

[1] Yves Tillé, Groupe de Statistique, Université de Neuchâtel, Espace de l'Europe 4, Case postale 827, 2002 Neuchâtel, Suisse.
Courrier électronique : yves.tille@unine.ch

$$\hat{Y}_w = \sum_{k \in S} w_k Y_k,$$

où les poids w_k peuvent dépendre de l'échantillon S et donc être aléatoires.

Une des possibilités consiste à prendre $w_k = 1/\pi_k = n/N$, ce qui donne l'estimateur de Horvitz-Thompson,

$$\hat{Y}_{\mathrm{HT}} = \sum_{k \in S} \frac{Y_k}{\pi_k} = \frac{N}{n} \sum_{k \in S} Y_k,$$

qui est sans biais.

Nous allons cependant nous intéresser à la classe plus générale des estimateurs pondérés conditionnellement (Tillé 1998, 1999a) où les unités sont pondérées par des inverses de probabilités d'inclusion conditionnelles. Si Z est une statistique quelconque, alors l'estimateur pondéré conditionnellement

$$\hat{Y}_Z = \sum_{k \in S} \frac{Y_k}{E(I_k|Z)} \tag{1}$$

est strictement sans biais si et seulement si $E(I_k \mid Z) > 0$, pour tout $k \in U$. En effet,

$$E(\hat{Y}|Z) = \sum_{k \in U} \frac{E(I_k|Z)Y_k}{E(I_k|Z)} = Y.$$

Comme l'estimateur est conditionnellement sans biais, il est également non-conditionnellement sans biais. L'estimateur (1) généralise, selon le choix de la statistique Z utilisée, l'estimateur stratifié, mais aussi (à une approximation près) l'estimateur par la régression (voir Tillé 1998).

3. CONDITIONNEMENT SUR DES RANGS

Supposons maintenant que les N valeurs $X_1, \ldots, X_k, \ldots, X_N$ d'un caractère auxiliaire x soient connues sur les N unités de la population. Dans un premier temps, on suppose que tous les X_k prennent des valeurs distinctes, cette hypothèse sera ensuite levée en section 5. Le rang R_k de l'unité k est

$$R_k = \#\{l \in U \mid X_l \le X_k\}.$$

On note par ailleurs r_j, $j = 1, \ldots, n$, les rangs de la population ordonnés des n unités sélectionnées dans l'échantillon donc $r_1 < r_2 < \ldots < r_{n-1} < r_n$. Les r_j sont des variables aléatoires ayant une distribution hypergéométrique négative (voir Tillé 1999b).

La statistique utilisée pour définir les probabilités d'inclusion conditionnelles est un sous-ensemble de $\{r_1, \ldots, r_j, \ldots, r_n\}$. On définit d'abord

- un entier q tel que $2 \le q \le n$, définissant la période,
- un entier b tel que $2 \le b$, définissant la bordure,
- un entier l tel que $b \le l \le b+q-1$, définissant le décalage.

Les quantité q, b, et l servent à définir un sous-ensemble d'indices :

$$E_l = \{r_l, r_{l+q}, r_{l+2q}, \ldots, r_{l+hq}, \ldots, r_{l+Hq}\},$$

$$\text{pour } l = b, \ldots, b+q-1.$$

Par exemple, si $n = 18$, $q = 4$, $b = 3$, alors

$$E_3 = \{r_3, r_7, r_{11}, r_{15}\},$$

$$E_4 = \{r_4, r_8, r_{12}, r_{16}\},$$

$$E_5 = \{r_5, r_9, r_{13}\},$$

$$E_6 = \{r_6, r_{10}, r_{14}\}.$$

La probabilité d'inclusion conditionnelle est calculée par rapport à l'un des E_l.

La valeur de H est définie de manière à ce que $l + Hq \le n-b+1$ et donc H est le plus grand entier tel que $H \le (n-b-l+1)/q$. Il est donc clair que H dépend de l.

Ensuite, on peut calculer les probabilités d'inclusion :

$$E(I_k|E_l) = \begin{cases} 1 & \text{si } k \in E_l \\ \dfrac{q-1}{r_{l+hq} - r_{l+(h-1)q} - 1} & \text{si } r_{l+(h-1)q} < k < r_{l+hq},\ h = 1, \ldots, H \\ \dfrac{l-1}{r_l - 1} & \text{si } k < r_l \\ \dfrac{n-(l+Hq)}{N - r_{l+Hq}} & \text{si } k > r_{l+Hq}. \end{cases}$$

Ces probabilités d'inclusion sont donc assez contrastées. Cependant elles sont toutes positives, y compris sur les bords. Il est important de prendre une bordure $b \ge 2$ afin que la première et la dernière post-strate ne soit pas vide.

4. UNE CLASSE D'ESTIMATEURS SANS BIAIS

Comme $E(I_k|E_l) > 0$, on peut construire un estimateur qui est sans biais et même conditionnellement sans biais par rapport à E_l. En notant $y_1, \ldots, y_j, \ldots, y_n$ les n valeurs prises par les unités dans l'échantillon ordonnées selon les R_k, on obtient

$$\hat{Y}_l = \sum_{k \in S} \frac{Y_k}{E(I_k|E_l)}$$

$$= \frac{r_l - 1}{l-1} \sum_{j=1}^{l-1} y_j + y_l$$

$$+ \sum_{h=1}^{H} \left(\frac{r_{l+hq} - r_{l+(h-1)q} - 1}{q-1} \sum_{j=1}^{q-1} y_{l+(h-1)q+j} + y_{l+hq} \right)$$

$$+ \frac{N - r_{l+Hq}}{n-(l+Hq)} \sum_{j=l+Hq+1}^{n} y_j$$

$$= N_{0|l}\hat{\bar{y}}_{0|l} + y_l + \sum_{h=1}^{H} \left(N_{h|l}\hat{\bar{y}}_{h|l} + y_{l+hq} \right) + N_{H+1|l}\hat{\bar{y}}_{H+1|l}$$

où

$$N_{0|l} = r_l - 1,$$

$$N_{h|l} = r_{l+hq} - r_{l+(h-1)q} - 1, h = 1, \ldots, H,$$

$$N_{H+1|l} = N - r_{l+Hq},$$

$$\hat{\bar{y}}_{0|l} = \frac{1}{l-1} \sum_{j=l}^{l-1} y_j,$$

$$\hat{\bar{y}}_{h|l} = \frac{1}{q-1} \sum_{j=1}^{q-1} y_{l+(h-1)q+j}, h = 1, \ldots, H,$$

et

$$\hat{\bar{y}}_{H+1|l} = \frac{1}{n-(l+Hq)} \sum_{j=l+Hq+1}^{n} y_j.$$

Cet estimateur est en réalité un estimateur post-stratifié où les tailles des post-strates sont fixées dans l'échantillon. Comme $E(I_k|E_l) > 0$, $\hat{Y}_l$ est strictement sans biais non-conditionnellement et conditionnellement à E_l, ce qui n'est évidemment pas le cas de l'estimateur post-stratifié classique, car ce dernier a une probabilité non-nulle d'avoir une post-strate vide. En fixant la taille des post-strates dans l'échantillon, la constitution de post-strates vides devient impossible. La taille correspondante de la post-strate dans la population est une variable aléatoire $N_{h|l}$.

L'estimateur $\hat{Y}_l$ possède une autre propriété intéressante. En utilisant la définition des $E(I_k|E_l)$, on montre assez facilement que

$$\sum_{k \in S} \frac{1}{E(I_k|E_l)} = N.$$

L'estimateur est donc calé sur la taille de la population. Cette propriété, que possède également l'estimateur de Horvitz-Thompson dans les plans simples, n'est donc pas perdue. Les unités dont les rangs sont dans E_l sont appelées unités pivots, et sont affectées d'un poids égal à 1, ce qui rend les poids très inégaux. On peut donc reprocher à $\hat{Y}_l$ d'utiliser des poids fortement dispersés. Ce problème peut être résolu en réalisant un lissage des estimateurs.

5. LISSAGE DES ESTIMATEURS

Pour résoudre le problème de la dispersion des poids, on réalise une moyenne mobile d'estimateurs de la manière suivante :

$$\hat{Y}_c = \frac{1}{q} \sum_{l=b}^{b+q-1} \hat{Y}_l.$$

$\hat{Y}_c$ garde toutes les propriétés des $\hat{Y}_l$. Il est donc sans biais, calé sur N et linéaire, il peut donc s'écrire sous la forme

$$\hat{Y}_c = \sum_{j=1}^{n} w_j y_j,$$

où $w_j =$

$$\begin{cases} \frac{1}{q} \sum_{l=b}^{b+q-1} \frac{r_l - 1}{l-1}, & j < b, \\ \frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_{j+l-b} - r_{m^-(j+l-b-q)} - 1}{j+l-b-m^-(j+l-b-q)-1} + 1 \right), & b \le j < b+q-1, \\ \frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_{j+l-b} - r_{j+l-b-q} - 1}{q-1} + 1 \right), & b+q-1 \le j \le n-b+2-q, \\ \frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_{m^+(j+l-b)} - r_{j+l-b-q} - 1}{m^+(j+l-b)-j+l-b-q-1} + 1 \right), & n-b+2-q < j \le n-b+1, \\ \frac{1}{q} \sum_{l=b}^{b+q-1} \frac{N+1-r_{n+1-l}-1}{n+1-(n+1-l)-1} = & \\ \qquad \frac{1}{q} \sum_{l=b}^{b+q-1} \frac{N - r_{n+1-l}}{l-1}, & n-b+1 < j, \end{cases}$$

$$m^-(x) = \begin{cases} 0 & \text{si } x < b \\ x & \text{sinon}, \end{cases} \qquad (2)$$

$$m^+(x) = \begin{cases} n+1 & \text{si } x > n-b+1 \\ x & \text{sinon}, \end{cases}$$

$r_0 = 0$ et $r_{n+1} = N + 1$.

Sous l'apparente complexité due au traitement particulier des bords, le système de pondération est relativement simple. Dans le cas où l'on ne se trouve pas trop près des bords, il vaut alors

$$w_j = \frac{1}{q}\left(\sum_{l=b}^{b+q-1} \frac{r_{j+l-b} - r_{j+l-b-q} - 1}{q-1} + 1 \right)$$

$$= \frac{1}{q\,(q-1)} \sum_{\alpha=0}^{q-1} (r_{j+\alpha} - r_{j+\alpha-q}).$$

Si toutes les valeurs de la variable auxiliaire ne sont pas distinctes, on peut affecter les rangs unités ayant des valeurs communes au hasard. Par exemple, si on a $X_1 = 2, X_2 = 5, X_3 = 5, X_4 = 7, X_5 = 8$, on choisit avec une probabilité 1/2, soit les rangs $R_1 = 1, R_2 = 2, R_3 = 3, R_4 = 4, R_5 = 5$, soit $R_1 = 1, R_2 = 3, R_3 = 2, R_4 = 4, R_5 = 5$. On calcule alors l'estimateur lissé pour chaque permutation, et on fait leur moyenne. Cette méthode a l'avantage de conserver un estimateur sans biais. En effet, pour chacune des permutations possibles, l'estimateur est sans biais. En pratique, il est n'est pas nécessaire de calculer les estimateurs pour toutes les permutations. On peut calculer l'estimateur d'une permutation, et ensuite, on égalise simplement les poids des unités ayant les mêmes valeurs pour la variable x.

6. CAS OÙ $q = 2$, $b = 2$

Quand $q = 2$, et $b = 2$, on obtient après quelques calculs

$$\hat{Y}_c = \frac{1}{2}\left\{ \sum_{j=3}^{n-2} y_j (r_{j+1} - r_{j-1}) + \frac{r_3 + 2r_2 - 3}{2} y_1 + \frac{r_3 + 1}{2} y_2 + \frac{r_{n+1} - r_{n-2} + 1}{2} y_{n-1} + \frac{3r_{n+1} - 2r_{n-1} - r_{n-2} - 3}{2} y_n \right\}$$

$$= \frac{1}{2}\left\{ \sum_{j=1}^{n} y_j (r_{j+1} - r_{j-1}) + y_1 \frac{r_3 - 3}{2} + y_2 \frac{2r_1 + 1 - r_3}{2} + y_{n-1} \frac{r_{n+1} + r_{n-2} + 1 - 2r_n}{2} + y_n \frac{r_{n+1} - r_{n-2} - 3}{2} \right\},$$

où $r_0 = 0$ et $r_{n+1} = N + 1$. On retrouve un estimateur proposé par Ren (2000, page 140) et obtenu avec un argument de calage. Seule la gestion des bords est légèrement différente.

Example 1 : Soit une population de taille $N = 20$. Supposons que les valeurs de la variable d'intérêt se trouvent dans la Table 1. On suppose, en outre que l'échantillon de taille $n = 7$ est composé des unités de rangs $\{3, 7, 8, 11, 12, 15, 17\}$. Si on prend $q = 2$, $l = 2$, $b = 2$ on obtient $E_2 = \{r_2, r_4, r_6\} = \{7, 11, 15\}$. Ensuite on peut calculer $E(I_k \mid E_2 = \{7, 11, 15\})$. Les probabilités d'inclusion conditionnelles sont les suivantes :

$$E(I_3 \mid E_2 = \{7, 11, 15\}) = 1/6,$$
$$E(I_7 \mid E_2 = \{7, 11, 15\}) = 1,$$
$$E(I_8 \mid E_2 = \{7, 11, 15\}) = 1/3,$$
$$E(I_{11} \mid E_2 = \{7, 11, 15\}) = 1,$$
$$E(I_{12} \mid E_2 = \{7, 11, 15\}) = 1/3,$$
$$E(I_{15} \mid E_2 = \{7, 11, 15\}) = 1,$$
$$E(I_{17} \mid E_2 = \{7, 11, 15\}) = 1/5.$$

Tableau 1
Exemple d'une population de taille $N = 20$

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_k	9	71	72	35	91	14	3	36	64	38	81	52	78	62	86	16	20	59	84	55
R_k	2	14	15	6	20	3	1	7	13	8	17	9	16	12	19	4	5	11	18	10

L'estimateur

$$\hat{Y}_0 = \sum \frac{y_k}{E(I_k \mid E_2 = \{7, 11, 15\})}$$

est donc sans biais et conditionnellement sans biais. De plus, il est linéaire et

$$\sum_{k \in S} \frac{1}{E(I_k \mid E_2 = \{7, 11, 15\})} = N.$$

D'autre part, si on prend $q = 2$, $l = 3$, $b = 2$, on obtient $E_3 = \{r_3, r_5\} = \{8, 12\}$. Alors, avec le même exemple, on calcule $E(I_k \mid E_3 = \{8, 12\})$, et on obtient

$$E(I_3 \mid E_3 = \{8, 12\}) = 2/7,$$
$$E(I_7 \mid E_3 = \{8, 12\}) = 2/7,$$
$$E(I_8 \mid E_3 = \{8, 12\}) = 1,$$
$$E(I_{11} \mid E_3 = \{8, 12\}) = 1/3,$$
$$E(I_{12} \mid E_3 = \{8, 12\}) = 1,$$
$$E(I_{15} \mid E_3 = \{8, 12\}) = 2/8 = 1/4,$$
$$E(I_{17} \mid E_3 = \{8, 12\}) = 2/8 = 1/4.$$

L'estimateur

$$\hat{Y}_1 = \sum \frac{y_k}{E(I_k \mid E_3 = \{8, 12\})}$$

est également sans biais et linéaire.

Enfin, on calcule la moyenne des deux estimateurs :

$$\hat{Y}_c = \frac{\hat{Y}_0 + \hat{Y}_1}{2}.$$

Les poids s'obtiennent simplement en faisant la moyenne des poids des estimateurs $\hat{Y}_0$ et $\hat{Y}_1$, et valent

$$\begin{aligned} w_3 &= (6 + 7/2)/2 = 19/4, \\ w_7 &= (1 + 7/2)/2 = 9/4, \\ w_8 &= (3 + 1)/2 = 2, \\ w_{11} &= (1 + 3)/2 = 2, \\ w_{12} &= (3 + 1)/2 = 2, \\ w_{15} &= (1 + 4)/2 = 5/2, \\ w_{17} &= (5 + 4)/2 = 9/2. \end{aligned}$$

L'estimateur $\hat{Y}_c$ est linéaire et strictement sans biais.

7. APPLICATION À L'ESTIMATION DE LA RÉPARTITION

Il existe de multiples méthodes permettant d'utiliser de manière appropriée de l'information auxiliaire pour estimer une fonction de répartition. On peut trouver une exposé de ces techniques dans Ren (2000) et dans Wu et Sitter (2001). La méthode que nous proposons permet également d'estimer la répartition. La répartition dans la population est définie par

$$F_1(y) = \frac{1}{N} \sum_{k \in U} I\{y_k \le y\},$$

et peut être estimée par

$$\hat{F}_1(y) = \frac{\sum_{k \in S} w_k I\{y_k \le y\}}{\sum_{k \in S} w_k},$$

où $I\{y \le y_k\}$ est la fonction indicatrice, et les w_k sont les poids associés aux unités k qui valent $1/\pi_k = N/n$ pour l'estimateur de Horvitz-Thompson, et qui sont donnés en (2) pour l'estimateur calé.

Notons que les deux fonctions sont discrètes, mais que les sauts sont beaucoup moins nombreux sur S que sur U. Pour atténuer les différences des répartitions entre l'échantillon et la population, nous avons lissé les fonctions de répartition, en utilisant comme Deville (1995) une interpolation linéaire des centres des contremarches, ce qui consiste à définir $F_2(y)$ en reliant les points

$$\frac{1}{2}\left\{F_1(y_k) - F_1(y_k - \varepsilon)\right\},$$

pour $k \in U$, où ε est un réel strictement positif arbitrairement petit. Ensuite, on définit $\hat{F}_2(y)$ en reliant les points

$$\frac{1}{2}\left\{\hat{F}_1(y_k) - \hat{F}_1(y_k - \varepsilon)\right\},$$

pour l'échantillon.

Example 2 : Une population de taille $N = 1\ 000$ a été générée au moyen de variables logarithmico-normales indépendantes et équidistribuées. Un échantillon de taille $n = 16$ a ensuite été sélectionné et on a fixé $h = 5$. La figure 1 donne $F_2(x)$ dans la population.

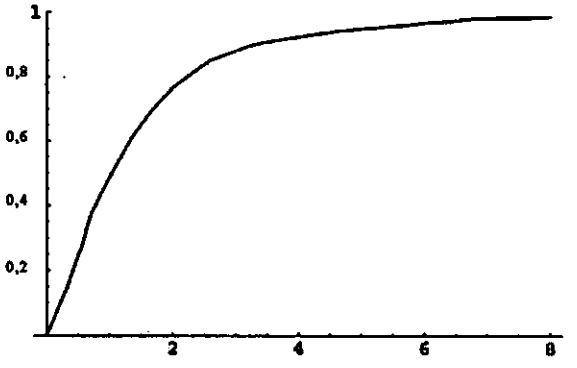


Figure 1 . Fonction de répartition dans la population

Ensuite, la figure 2 montre $F_2(x)$ et la répartition estimée par l'estimateur de Horvitz-Thompson. Enfin, la figure 3 montre $F_2(x)$ et la répartition estimée par l'estimateur calé.

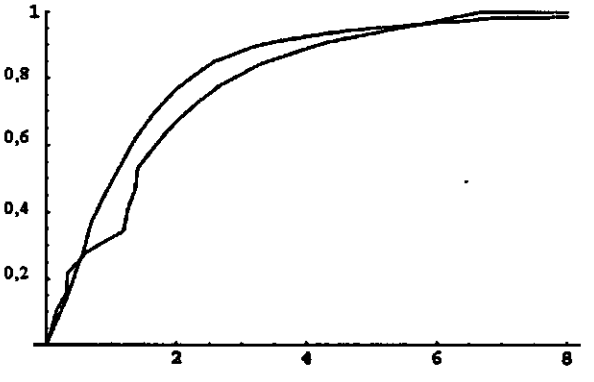


Figure 2. Fonction de répartition dans la population et estimateur de Horvitz-Thompson de la répartition

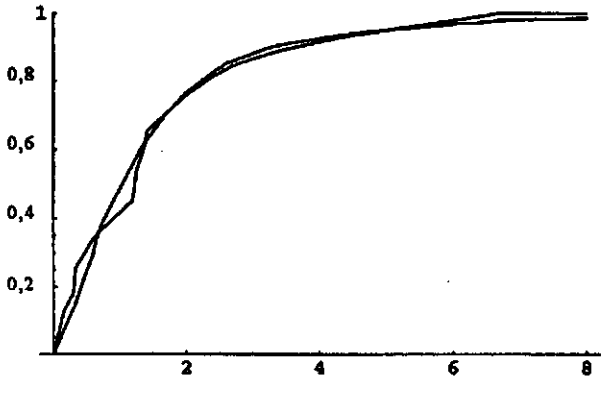


Figure 3. Fonction de répartition dans la population et estimateur calé de la répartition

8. COMPARAISON À L'ESTIMATEUR PAR LA RÉGRESSION

Afin de comparer les qualités de l'estimateur proposé, un ensemble de simulations a été réalisé pour comparer l'estimateur calé sur la répartition avec l'estimateur de Horvitz-Thompson et l'estimateur par la régression. Trois populations de taille 1 000 on été générées selon les modèles suivants.

- *Modèle A Dépendance linéaire* : La population est générée selon le modèle $X_k \sim N(0, 1)$ et $Y_k = X_k + 1{,}33333 \times \varepsilon_k$ où $\varepsilon_k \sim N(0, 1)$. Le coefficient de corrélation obtenu dans la population est $\rho = 0{,}616154$.

- Modèle B Dépendance non linéaire 1 : La population est générée selon le modèle $X_k \sim N(0, 1)$ et $Y_k = (0.2 + X_k)^2 + 1{,}33333 \times \varepsilon_k$ où $\varepsilon_k \sim N(0, 1)$. Le coefficient de corrélation obtenu dans la population est $\rho = 0{,}28975$.

- Modèle C Dépendance non linéaire 2 : La population est générée selon le modèle $X_k \sim N(0, 1)$ et $Y_k = (0{,}4 + X_k)^2 + 1{,}33333 \times \varepsilon_k$ où $\varepsilon_k \sim N(0, 1)$. Le coefficient de corrélation obtenu dans la population est $\rho = 0.476158$.

Dans chaque population 100 000 échantillons de taille 100 ont été sélectionnés. Pour chaque échantillon trois systèmes de pondération ont été calculés.

1. les poids associés au plan simple $w_k = N/n$,
2. les poids de l'estimateur calé sur la répartition donné en (2) en prenant la fenêtre $q = 10$ et la bordure $b = 6$,
3. les poids de l'estimateur par la régression donnés par

$$w_k = \frac{N}{n} + \left(X - \hat{X}_{\mathrm{HT}}\right) \frac{\left(X_k - \hat{\bar{X}}\right)}{\sum_{k \in S} \left(X_k - \hat{\bar{X}}\right)^2},$$

où X est le total des X_k sur la population, $\hat{X}_{\mathrm{HT}}$ est l'estimateur de Horvitz-Thompson de X, et $\hat{\bar{X}} = \hat{X}_{\mathrm{HT}}/N$.

Au moyen de ces poids, l'estimateur de la moyenne et des neuf déciles ont été calculés pour chaque échantillon. Ensuite, on estime la variance de ces estimateurs au moyen des simulations.

Les résultats sont donnés dans les tableaux 2, 3 et 4. Les variances ont été ramenées à 1 pour le plan simple. Pour le modèle linéaire, l'estimateur par la régression est légèrement préférable. Cependant, pour le cas non-linéaire, le gain de précision de l'estimateur calé sur la répartition est très important sur la moyenne comme sur les quantiles. L'estimateur que nous proposons se comporte donc de manière robuste en cas de relation non-linéaire entre la variable auxiliaire et la variable d'intérêt.

Tableau 2
Modèle A : variance des estimateurs
(référence : Horvitz-Thompson=1)

Paramètre	Calage répartition	Estim. Régression
moyenne	0,674422	0,632608
1er décile	0,905273	0,893876
2ème décile	0,815403	0,802082
3ème décile	0,842681	0,815071
4ème décile	0,806749	0,768283
5ème décile	0,783731	0,740765
6ème décile	0,818051	0,782549
7ème décile	0,794411	0,773794
8ème décile	0,857114	0,844874
9ème décile	0,884424	0,884032

Tableau 3
Modèle B : variance des estimateurs
(référence : Horvitz-Thompson=1)

Paramètre	Calage répartition	Estim. Régression
moyenne	0,429689	0,953025
1er décile	0,913598	0,958656
2ème décile	0,919394	1,009270
3ème décile	0,829860	0,987950
4ème décile	0,792094	0,989114
5ème décile	0,703908	0,992023
6ème décile	0,622705	1,009830
7ème décile	0,550028	0,981249
8ème décile	0,443828	1,010340
9ème décile	0,549615	1,029120

Tableau 4
Modèle C : variance des estimateurs
(référence : Horvitz-Thompson=1)

Paramètre	Calage répartition	Estim. Régression
moyenne	0,30768	0,808114
1er décile	0,95560	0,983582
2ème décile	0,85920	0,970913
3ème décile	0,73854	0,930401
4ème décile	0,65728	0,950651
5ème décile	0,60500	0,956807
6ème décile	0,52139	0,930514
7ème décile	0,45709	0,907537
8ème décile	0,40752	0,903593
9ème décile	0,39820	0,860050

9. VARIANCE ET ESTIMATION DE VARIANCE

Pour calculer la variance de $\hat{Y}_c$, on commence par calculer la variance de $\hat{Y}_l$. Comme $\hat{Y}_l$ est sans biais conditionnellement à E_l, on a

$$V(\hat{Y}_l) = E\,V(\hat{Y}_l \mid E_l).$$

Comme dans chacune des post-strates, conditionnellement à E_l le plan est simple sans remise de taille fixe, on a

$$V(\hat{Y}_l \mid E_l) = \sum_{h=0}^{H+1} N_{h|l}^2\, V(\hat{\bar{y}}_{h|l})$$

$$= \sum_{h=0}^{H+1} N_{h|l}^2 \frac{N_{h|l} - n_{h|l}}{N_{n|l}} \frac{S_{h|l}^2}{n_{h|l}}, \qquad (3)$$

où

$$n_{0|l} = l - 1,$$

$$n_{h|l} = q - 1, h = 1, \ldots, H,$$

$$n_{H+1|l} = n - (l + Hq),$$

$$\bar{Y}_{0|l} = \frac{1}{N_{0|l}} \sum_{k=1}^{r_{l-1}} Y_{(k)},$$

$$\bar{Y}_{h|l} = \frac{1}{N_{h|l}} \sum_{k=r_{l+(h-1)q}+1}^{r_{l+hq-1}} Y_{(k)}, h = 1, \ldots, H,$$

$$\bar{Y}_{H+1|l} = \frac{1}{N_{H|l}} \sum_{k=N-r_{l+Hq}+1}^{N} Y_{(k)},$$

$$S_{0|l}^2 = \frac{1}{N_{0|l} - 1} \sum_{k=1}^{r_{l-1}} (Y_{(k)} - \bar{Y}_{0|l})^2,$$

$$S_{h|l}^2 = \frac{1}{N_{h|l} - 1} \sum_{k=r_l+(h-1)q+1}^{r_l+h_q-1} (Y_{(k)} - \bar{Y}_{h|l})^2, h = 1, \ldots, H,$$

et

$$S_{H+1|l}^2 = \frac{1}{N_{H+1|l} - 1} \sum_{k=N-r_{l+Hq}+1}^{N} (Y_{(k)} - \bar{Y}_{H+1|0})^2,$$

où les $Y_{(k)}$ représentent les valeurs de Y_k triées selon l'ordre croissant des X_k.

On remarque qu'il est très difficile de calculer la variance non-conditionnelle de $\hat{Y}_l$, c'est-à-dire l'espérance de $V(\hat{Y}_l|E_l)$. En effet, $N_{h|l}$ et $S_{h|l}^2$ sont aléatoires. Cependant, on peut estimer simplement $V(\hat{Y}_l|E_l)$ et obtenir un estimateur sans biais de la variance conditionnelle (et donc de la variance) en estimant simplement (3), par

$$\hat{V}(\hat{Y}_l \mid E_l) = \sum_{h=0}^{H+1} N_{h|l}^2 \frac{N_{h|l} - n_{h|l}}{N_{h|l}\, n_{h|l}} s_{h|l}^2, \qquad (4)$$

où

$$s_{0|l}^2 = \frac{1}{n_{0|l} - 1} \sum_{j=1}^{l-1} (y_j - \hat{\bar{y}}_{0|l})^2,$$

$$s_{h|l}^2 = \frac{1}{n_{h|l} - 1} \sum_{j=1}^{q-1} (y_{l+(h-1)q+j} - \hat{\bar{y}}_{h|l})^2, h = 1, \ldots, H,$$

et

$$s_{H+1|l}^2 = \frac{1}{n_{h|l} - 1} \sum_{j=l+Hq+1}^{n} (y_j - \hat{\bar{y}}_{H+1|l})^2.$$

L'estimateur $\hat{V}(\hat{Y}_l|E_l)$ est non seulement sans biais pour $V(\hat{Y}_l|E_l)$ mais aussi pour $V(\hat{Y}_l)$.

10. APPROXIMATIONS POUR LE CALCUL DE LA VARIANCE

Malheureusement, le calcul de la variance de $\hat{Y}_c$ devient plus complexe à cause des covariances. En effet, on a

$$V(\hat{Y}_c) = \frac{1}{q^2} \sum_{l=b}^{b+q-1} \sum_{i=b}^{b+q-1} \mathrm{Cov}(\hat{Y}_l, \hat{Y}_i).$$

Quand $l = i$, le problème consiste à estimer $V(\hat{Y}_i)$, ce qui se fait sans difficulté. Quand $l \neq i$, il faut calculer

$$\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i) = E\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i|E_l)$$

$$+ \mathrm{Cov}(E(\hat{Y}_l|E_l), E(\hat{Y}_i|E_l)).$$

Comme $E(\hat{Y}_l|E_l) = Y$, obtient

$$\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i) = E\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i|E_l)$$

$$= E\,E(\hat{Y}_l \hat{Y}_i|E_l) - Y^2.$$

Le calcul $E(\hat{Y}_l, \hat{Y}_i|E_l)$ semble malheureusement inextricable, il faut donc chercher une approximation.

Une première voie consiste à chercher un majorant de la variance, comme

$$\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i) \leq \sqrt{V(\hat{Y}_l)\,V(\hat{Y}_i)},$$

on a un majorant donné par

$$V(\hat{Y}_c) \leq \frac{1}{q^2} \sum_{l=b}^{b+q-1} \sum_{i=b}^{b+q-1} \sqrt{V(\hat{Y}_l)\,V(\hat{Y}_i)}$$

$$= \frac{1}{q^2} \left(\sum_{l=b}^{b+q-1} \sqrt{V(\hat{Y}_l)} \right)^2,$$

ce qui peut être estimé par

$$\hat{V}_1(\hat{Y}_c) = \frac{1}{q^2} \left(\sum_{l=b}^{b+q-1} \sqrt{\hat{V}(\hat{Y}_l \mid E_l)} \right)^2,$$

ce qui revient à estimer l'écart-type des moyennes par la moyenne des écart-types.

Enfin, une seconde voie peut-être donnée par une technique de résidus. En effet, de manière générale, quand on redresse un estimateur au moyen d'une technique de calage, on estime la variance au moyen d'une technique de résidus (voir à ce sujet Deville et Särndal 1992, et Deville 1999). Dans le cas du calcul de la variance de $\hat{Y}_l$, on peut utiliser une technique de résidus pour obtenir la variance exacte. En effet, considérons la variable

$$v_k(l) = \begin{cases} \left(\dfrac{N^2(N-n)}{Nn(n-1)}\right)^{-1/2} \left(\dfrac{N_{h|l}^2(N_{h|l}-n_{h|l})}{N_{h|l}n_{h|l}(N_{h|l}-1)}\right)^{1/2} (Y_k - \bar{Y}_{h|l}) \\ \qquad \text{si } k = r_{l+(h-1)q+1}, \ldots, r_{l+hq-1} \\ 0 \qquad \text{si } k = r_{l+(h-1)q} \text{ ou } k = r_{l+hq} \end{cases}$$

qui peut apparaître comme un résidu associé à l'estimateur $\hat{Y}_l$. La variable $v_k(l)$ injectée dans l'expression classique du plan simple sans remise de taille fixe est exactement égale à la variance conditionnelle $\hat{Y}_l$ donnée en (3). En effet,

$$N^2 \frac{N-n}{nN} \frac{1}{N-1} \sum_{k \in U} \left(v_k - \frac{\sum_{k \in U} v_k}{N} \right)^2 = V(\hat{Y}_l | E_l).$$

Cette variable dépend cependant des $\bar{Y}_{h|l}$ qui sont inconnus. On peut estimer $v_k(l)$ par

$$\hat{v}_j(l) = \begin{cases} \left(\dfrac{N^2(N-n)}{Nn(n-1)}\right)^{-1/2} \left(\dfrac{N_{h|l}^2(N_{h|l}-n_{h|l})}{N_{h|l}n_{h|l}(n_{h|l}-1)}\right)^{1/2} (y_j - \hat{\bar{y}}_{h|l}) \\ \qquad \text{si } j = l+(h-1)q+1, \ldots, l+hq-1 \\ 0 \qquad \text{si } j = l+(h-1)q \text{ ou } j = l+hq \end{cases}$$

Si on injecte $\hat{v}_k(l)$ dans l'estimateur de la variance du plan simple sans remise, on obtient un estimateur sans biais de la variance conditionnelle, et donc de la variance.

$$N^2 \frac{N-n}{nN} \frac{1}{n-1} \sum_{j=1}^{n} \left(\hat{v}_j - \frac{\sum_{j=1}^{n} \hat{v}_j}{n} \right)^2 = \hat{V}(\hat{Y}_l | E_l).$$

Deville (1999) montre que la variance d'une somme d'estimateurs peut se calculer en faisant la somme des résidus associés à ces estimateurs, les résidus étant calculés par linéarisation. Pour estimer la variance de $\hat{Y}_c$, on pourrait donc simplement prendre la moyenne des résidus $\hat{v}_k(l)$, ce qui s'écrit

$$\hat{v}_k = \frac{1}{q} \sum_{l=b}^{b+q-1} \hat{v}_k(l).$$

On pourrait ainsi estimer la variance par

$$\hat{V}_2(\hat{Y}_c) = \frac{N^2(N-n)}{nN} \frac{1}{n-1} \sum_{k \in S} \left(\hat{v}_k - \frac{\sum_{k \in S} \hat{v}_k}{n} \right)^2.$$

Ces deux estimateurs de variances sont à évaluer par des simulations.

11. SIMULATIONS POUR LES ESTIMATEURS DE VARIANCE

Les simulations présentées dans le tableau (5) sont basées sur des populations de taille $N = 100$, qui sont générées au moyen de variables aléatoires indépendantes normales. Pour chaque cas étudié, on donne la valeur de q, et le coefficient de corrélation entre la variable d'intérêt Y_k et le rang R_k de la variable auxiliaire X_k. La borne b est définie en prenant la partie entière de $q/2+1$. L'objectif étant de valider l'estimateur de variance, on tire 3 000 échantillons de taille $n = 20$ pour chaque simulation, et on compare la variance estimée par les simulations de l'estimateur calé $V_{si}(\hat{Y}_c)$ aux espérances sous les simulations des deux estimateurs de variance notées $E_{si}(\hat{V}_\alpha(\hat{Y}_c))$, $\alpha = 1, 2$. Les deux dernières colonnes des tableaux présentent le biais relatif défini par

$$\mathrm{BR}_{si}\hat{V}_\alpha(\hat{Y}_c) = \frac{E_{si}\hat{V}_\alpha(\hat{Y}_c) - V_{si}(\hat{Y}_c)}{V_{si}(\hat{Y}_c)}, \ \alpha = 1, 2.$$

Les simulations montrent que les deux estimateurs proposés surestiment la variance. La surestimation semble diminuer quand q croît. L'estimateur $\hat{V}_2(\hat{Y}_c)$ a manifestement le plus petit biais. On préconisera donc l'utilisation de $\hat{V}_2(\hat{Y}_c)$.

Tableau 5
Résultats des simulations

	Corrélation: 0,802				
q	$V_{si}(\hat{Y}_c)$	$E_{si}\hat{V}_1(\hat{Y}_c)$	$E_{si}\hat{V}_2(\hat{Y}_c)$	$BR_{si}\hat{V}_1(\hat{Y}_c)$	$BR_{si}\hat{V}_2(\hat{Y}_c)$
4	0,045	0,065	0,054	0,444	0,200
5	0,045	0,066	0,057	0,467	0,267
6	0,056	0,076	0,070	0,357	0,250
7	0,058	0,079	0,059	0,362	0,017
8	0,063	0,088	0,087	0,397	0,381
	Corrélation: 0,481				
q	$V_{si}(\hat{Y}_c)$	$E_{si}\hat{V}_1(\hat{Y}_c)$	$E_{si}\hat{V}_2(\hat{Y}_c)$	$BR_{si}\hat{V}_1(\hat{Y}_c)$	$BR_{si}\hat{V}_2(\hat{Y}_c)$
4	0,048	0,066	0,059	0,375	0,229
5	0,045	0,060	0,054	0,333	0,200
6	0,044	0,056	0,051	0,273	0,159
7	0,044	0,054	0,051	0,227	0,159
8	0,045	0,052	0,048	0,156	0,067
	Corrélation: 0,111				
q	$V_{si}(\hat{Y}_c)$	$E_{si}\hat{V}_1(\hat{Y}_c)$	$E_{si}\hat{V}_2(\hat{Y}_c)$	$BR_{si}\hat{V}_1(\hat{Y}_c)$	$BR_{si}\hat{V}_2(\hat{Y}_c)$
4	0,281	0,471	0,363	0,676	0,292
5	0,297	0,420	0,356	0,414	0,199
6	0,279	0,363	0,316	0,301	0,133
7	0,267	0,342	0,324	0,281	0,213
8	0,282	0,327	0,281	0,160	-0,004

12. CONCLUSIONS

L'estimateur que nous proposons est un des rares estimateurs qui soit à la fois sans biais, linéaire, qui utilise de l'information auxiliaire, et qui est calé sur la taille de la population. Il est paramétrable au moyen de la largeur de la fenêtre q. Ce nouvel estimateur est robuste par rapport à l'estimateur par la régression. Il peut prendre en compte une information auxiliaire ayant une relation non-linéaire avec la variable d'intérêt. La plupart des simulations semblent montrer que la largeur de la fenêtre n'a pas beaucoup d'impacts sur la précision d'un estimateur de moyenne. Cependant, il apparaît aussi qu'une petite largeur de fenêtre n'est pas pénalisante, même si il n'y a pas de dépendance entre la variable auxiliaire et la variable d'intérêt. Plus q est petit, plus le calage sera serré, et l'estimateur de la variance sera alors fortement pénalisé, car un degré de liberté est perdu dans chaque post-strate. Le choix de q doit donc tenir compte de ce problème.

Il existe beaucoup d'autres méthodes permettant d'utiliser l'information donnée par une fonction de répartition (voir Ren 2000) pour améliorer un estimateur. Les résultats que nous avons présentés se réduisent aux plans simples, mais nous pensons qu'ils sont importants, au même titre que la post-stratification est importante comme cas particulier des techniques de calage. En effet, la post-stratification est un des rares exemples où l'on peut montrer avec exactitude que le calage correspond à une approche conditionnelle. De plus, notre approche peut être vue comme un calage sur une fonction de répartition fournissant un estimateur sans biais. Une bonne technique générale de calage sur la répartition devrait donc retrouver dans les plans simples la méthode que nous avons présentée.

REMERCIEMENTS

Nous remercions Jean-Claude Deville et Anne-Catherine Favre, deux arbitres et un éditeur associé pour leurs commentaires constructifs qui ont permis d'améliorer considérablement cet article.

BIBLIOGRAPHIE

DEVILLE, J.-C. (1995). *Estimation de la variance du coefficient de Gini mesuré par sondage*. INSEE Méthode, document de travail, Méthodologie F9510.

DEVILLE, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : techniques de résidus et de linéarisation. *Techniques d'enquête*, 25, 219-230.

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

ESTEVAO, V., HIDIROGLOU, M.A. et SÄRNDAL, C.-E. (1995). Methodological principle for a generalized estimation system in Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

REN, R. (2000). *Estimation par calage sur la répartition*. Thèse de Doctorat en préparation, Paris, Université Paris Dauphine.

TILLÉ, Y. (1998). Estimation in surveys using conditional inclusion probabilities : simple random sampling. *International Statistical Review*, 66, 303-322.

TILLÉ, Y. (1999a). Sur la détermination a posteriori des bornes des post-strates. Dans *Les sondages*, (Éds. G. Brossier, et A.-M. Dussaix). Dunod, 202-208.

TILLÉ, Y. (1999b). Estimation dans des enquêtes par sondage avec des probabilités d'inclusion conditionnelles : enquêtes à plan d'échantillonnage complexe. *Techniques d'enquêtes*, 25, 57-66.

WU, C., et SITTER, R.R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *Canadian Journal of Statistics*, 29, 289-307.

Techniques d'enquête, juin 2002
Vol. 28, n° 1, pp. 93-101
Statistique Canada

Estimation de variance dans le cadre de la « Current Employment Survey »

JUN SHAO et SHAIL BUTANI[1]

RÉSUMÉ

Comme dans la plupart des autres enquêtes, il y a souvent un phénomène de non-réponse dans la « Current Employment Survey » effectuée tous les mois aux États-Unis par le Bureau of Labor Statistics (BLS). Dans un mois quelconque, les estimateurs seront plus efficients si on procède à une imputation à l'aide des données déclarées des mois antérieurs au lieu de ne pas tenir compte des répondants et de modifier la pondération de l'enquête en conséquence. Il faut cependant dire que l'imputation influe aussi sur l'estimation de variance. Si on traite les valeurs imputées comme des valeurs déclarées et qu'on applique une méthode type d'estimation de variance, les estimateurs de variance seront entachés d'un biais négatif. Dans le présent article, nous proposons un certain nombre de ces estimateurs par regroupement en demi-échantillons équilibrés et par réimputation afin de tenir compte de l'imputation. Nous présentons certains résultats de simulation pour le rendement d'échantillonnel pour population finie des estimateurs par données d'imputation et leurs estimateurs de variance.

MOTS CLÉS : Demi-échantillons équilibrés; taux d'échantillonnage non négligeables; imputation de rapport; échantillonnage stratifié.

1. INTRODUCTION

La « Current Employment Survey » (CES), communément appelée enquête sur la rémunération, est menée tous les mois par le Bureau of Labor Statistics (BLS) aux États-Unis. De mois en mois, on recueille par ce moyen des données auprès des établissements, et ce, par divers systèmes automatisés : interviews téléphoniques assistées par ordinateur, entrée de données sur clavier téléphonique, télécopie, transmission électronique de données, poste, *etc.* Les principales variables de l'enquête sont l'emploi, les travailleurs de la production (sans le personnel de surveillance), leurs heures de travail et leurs gains dans les établissements on agricoles. Chaque année, on tire des chiffres de dénombrement des travailleurs des dossiers administratifs du régime d'assurance-chômage.

La non-réponse est un phénomène fréquent dans la CES. Dans un mois quelconque, on obtient des estimateurs d'enquête plus efficients par une imputation à l'aide des données déclarées des mois antérieurs que par une prise en compte des seules données de déclaration du mois en cours et une modification de la pondération de l'enquête. Cette constatation vaut particulièrement pour la CES, puisque le taux de non-réponse y est de 60 % à 80 % et qu'environ 60 % des non-répondants d'un mois peuvent devenir disponibles un ou plusieurs mois après, d'où la possibilité d'utiliser ces données dans un mois à venir comme « données déclarées des mois antérieurs » (données chronologiques).

On sait bien toutefois que, si on traite des valeurs d'imputation comme des valeurs de déclaration et applique une méthode type d'estimation de variance, les estimateurs de variance seront entachés d'un biais (souvent négatif). Il est possible d'établir des estimateurs de variance valides par certaines hypothèses posées au sujet des plans d'échantillonnage, des méthodes d'imputation et des mécanismes de réponse (et parfois aussi des modèles qui produisent les données).

Dans le présent article, nous étudions l'estimation de variance aux fins de la CES. Après avoir décrit le plan d'échantillonnage et la méthode d'imputation de cette enquête à la section 2, nous dégageons à la section 3 des estimateurs de variance valides (c'est-à-dire asymptotiquement non biaisés et convergents) pour les estimateurs d'enquête par données imputées. Pour simplifier le calcul des estimateurs de variance, nous proposons des approximations à la section 4 et en étudions les propriétés par simulation à la section 5. Nous tirons un certain nombre de conclusions à la section 6. Bien que notre étude ait la CES pour objet, ses résultats en sont, à notre avis, généralisables à toute enquête ayant un plan d'échantillonnage et une méthode d'imputation semblables.

2. PLAN D'ÉCHANTILLONNAGE ET IMPUTATION

La CES applique le plan suivant d'échantillonnage probabiliste stratifié. Soit P une population finie constituée d'un ensemble d'établissements $\{1, ..., N\}$ et stratifiée selon l'industrie et la taille des établissements. Dans la strate $h^{\text{ième}}$, nous prélevons un échantillon de taille $n_h \geq 2$ sans remise sur N_h unités de population par application indépendante d'un plan d'échantillonnage probabiliste

[1] Jun Shao, Département de statistique, Université du Wisconsin, Madison WI 53706; Shail Butani, Statistical Methods Division, Bureau of Labor Statistics, Washington, D.C. 20212.

interstrates. Les taux d'échantillonnage n_h / N_h ne sont pas nécessairement négligeables. Pour un certain nombre de strates se caractérisant par une forte taille des établissements, $n_h = N_h$. Soit S l'échantillon en question. Pour $i \in S$ et le mois $t = 0, 1, ..., T$, nous observons (s'il n'y a pas non-réponse) le nombre de travailleurs ($y_{t,i}^E$), le nombre de travailleurs hors personnel de surveillance ($y_{t,i}^W$), le nombre d'heures travaillées ($y_{t,i}^H$) et la rémunération hebdomadaire ($y_{t,i}^P$). Soit $y_{t,i}$ tout $y_{t,i}^E, y_{t,i}^W, y_{t,i}^H$ ou $y_{t,i}^P$. Dans la CES, les grands paramètres d'intérêt sont les totaux de population $Y_t = \sum_{i \in p} y_{t,i}$, $t = 1, ..., T$. Comme on peut tirer chaque année ces totaux des dossiers administratifs, nous supposons sans perte de généralité que Y_0 est connu. S'il n'y a pas non-réponse, nous estimons Y_t par un estimateur de rapport

$$\hat{Y}_t = Y_0 \sum_{i \in S} w_i y_{t,i} \Big/ \sum_{i \in S} w_i y_{0,i}, \quad t = 1, ..., T, \qquad (1)$$

où w_i est la valeur de pondération d'enquête de la $i^{\text{ième}}$ unité échantillonnée dans la $h^{\text{ième}}$ strate.

Dans notre recherche, nous appliquons, à partir du mois 1, la méthode d'imputation proposée par Butani, Harter et Wolter (1997) (voir la description qui suit) à l'égard des non-répondants. L'imputation se fait à l'intérieur d'une cellule d'imputation, c'est-à-dire dans une strate ou une union de strates. Les valeurs imputées des mois $1, ..., t-1$ sont reprises en imputation pour les non-répondants du mois t, sauf si les non-répondants des mois $1, ..., t-1$ deviennent répondants avant le mois t.

1. Nombre de travailleurs. Si $y_{t,i}^E$ est non-répondant, il y a imputation par

$$\tilde{y}_{t,i}^E = \hat{\alpha}_t \tilde{y}_{t-1,i}^E,$$

où $\tilde{y}_{t-1,i}^E = y_{t-1,i}^E$ (valeur de déclaration) si $y_{t-1,i}^E$ est disponible au mois t, $\tilde{y}_{t-1,i}^E$ étant alors une valeur d'imputation,

$$\hat{\alpha}_t = \frac{\sum_{j \in R_t} w_j y_{t,j}^E}{\sum_{j \in R_t} w_j y_{t-1,j}^E},$$

et où R_t est l'ensemble des unités déclarantes des mois t et $t-1$.

2. Nombre de travailleurs hors personnel de surveillance. Si $y_{t,i}^W$ est non-répondant, il y a imputation par

$$\tilde{y}_{t,i}^W = \tilde{y}_{t-1,i}^W \tilde{y}_{t,i}^E / \tilde{y}_{t-1,i}^E,$$

où $\tilde{y}_{t-1,i}^W$ est défini comme $\tilde{y}_{t-1,i}^E$.

3. Nombre d'heures travaillées. Si $y_{t,i}^H$ est non-répondant, il y a imputation par

$$\tilde{y}_{t,i}^H = \hat{\gamma}_t \tilde{y}_{t-1,i}^H \tilde{y}_{t,i}^W / \tilde{y}_{t-1,i}^W,$$

où $\tilde{y}_{t-1,i}^H$ est défini comme $\tilde{y}_{t-1,i}^E$ et

$$\hat{\gamma}_t = \frac{\sum_{j \in R_t} w_j y_{t,j}^H \Big/ \sum_{j \in R_t} w_j y_{t,j}^W}{\sum_{j \in R_t} w_j y_{t-1,j}^H \Big/ \sum_{j \in R_t} w_j y_{t-1,j}^W}.$$

4. Rémunération hebdomadaire. Si $y_{t,i}^P$ est non-répondant, il y a imputation par

$$\tilde{y}_{t,i}^P = \hat{\beta}_t \tilde{y}_{t-1,i}^P \tilde{y}_{t,i}^H / \tilde{y}_{t-1,i}^H,$$

où $\tilde{y}_{t-1,i}^P$ est défini comme $\tilde{y}_{t-1,i}^E$ et

$$\hat{\beta}_t = \frac{\sum_{j \in R_t} w_j y_{t,j}^P \Big/ \sum_{j \in R_t} w_j y_{t,j}^H}{\sum_{j \in R_t} w_j y_{t-1,j}^P \Big/ \sum_{j \in R_t} w_j y_{t-1,j}^H}.$$

Après imputation à l'égard des non-répondants, on établit les totaux mensuels estimatifs par la formule (1) en assimilant les valeurs d'imputation à des données de déclaration.

Posons que la population P se divise en K cellules d'imputation disjointes $P_1, ..., P_K$ et que, pour chaque k,

$$y_{t,i} = \alpha_{t,k}\, y_{t-1,i} + \sqrt{y_{t-1,i}}\, e_{t,i},$$

$$E_m(y_{t,i}) = \mu_{t,k}, \qquad E_m(e_{t,i}) = 0, \qquad i \in P_k, t = 1, 2, ...,$$

$$V_m(y_{t,i}) = \nu_{t,k}, \qquad V_m(e_{t,i}) = \sigma_k^2, \qquad (2)$$

où $y_{t,i}$ désigne tout $y_{t,i}^E, y_{t,i}^W, y_{t,i}^H$ ou $y_{t,i}^P$, E_m et V_m sont respectivement l'espérance et la variance (en valeur marginale) du modèle, $\alpha_{t,k}$ et σ_k^2 sont des paramètres inconnus, les $e_{t,i}$ sont indépendants et identiquement distribués et les deux opérations $\{y_{t,i}\}$ et $\{e_{t,i}\}$ sont indépendantes. Nous posons aussi que, dans chaque P_k, l'indicateur de réponse $a_{h,i}$ (=1 si $y_{t,i}$ est répondant et = 0 dans les autres cas) et $y_{t,i}$ sont indépendants, étant donné $y_{t-s,i}, a_{t-s,i}$, $s = 1, 2, ..., t$. Dans ce que l'on appelle le mécanisme de réponse sans confusion (Lee, Rancourt et Särndal 1994), $a_{t,i}$ et $y_{t,i}$ sont dépendants, mais par $y_{t-s,i}, a_{t-s,i}$, $s = 1, 2, ..., t$. C'est un cas plus général que celui de l'hypothèse d'indépendance de $(y_{1,i}, ..., y_{t,i})$ et de $(a_{1,i}, ..., a_{t,i})$. Posons enfin que les indicateurs de réponse d'unités différentes sont indépendants. Dans ces hypothèses, les estimateurs $\hat{Y}_t$ par les données d'imputation

(comme nous l'avons décrit à la section précédente) sont asymptotiquement sans biais à l'égard de l'espérance conjointe du modèle (2) et de l'échantillonnage dans la population finie.

Dans la CES, les cellules d'imputation sont des unions de strates et, ainsi,

$$\sum_{i \in S \cap P_k} w_i = M_k, \quad k = 1, ..., K,$$

où M_k est le nombre d'unités de population dans la $k^{\text{ième}}$ cellule d'imputation P_k. Les $\hat{Y}_t$ sont donc conditionnellement sans biais à l'égard de l'espérance du modèle (étant donné S), c'est-à-dire :

$$E_m(\hat{Y}_t - Y_t) = 0.$$

3. ESTIMATION DE VARIANCE

Soit E_s et V_s l'espérance et la variance respectives d'échantillonnage et V, la variance globale. Ainsi,

$$V(\hat{Y}_t - Y_t) = E_s[V_m(\hat{Y}_t - Y_t)] + V_s[E_m(\hat{Y}_t - Y_t)]$$

$$= E_s[V_m(\hat{Y}_t - Y_t)], \qquad (3)$$

puisque $E_m(\hat{Y}_t - Y_t) = 0$. Nous démontrons en outre en annexe que

$$V_m(\hat{Y}_t - Y_t) = V_m(\hat{Y}_t) - V_m(Y_t). \qquad (4)$$

À noter que (4) est évident en cas de non-réponse.

À cause de (3), l'estimation de $V(\hat{Y}_t - Y_t)$ est la même que celle de $V_m(\hat{Y}_t - Y_t)$. De plus, nous pouvons, à cause de (4), calculer d'abord les estimateurs v_{t1} et v_{t2} de $V_m(\hat{Y}_t)$ et $V_m(Y_t)$ respectivement, puis prendre la différence $v_{t1} - v_{t2}$ comme notre estimateur de variance pour $\hat{Y}_t$. Comme $V_m(\hat{Y}_t)$ est une variance conditionnelle étant donné S, nous n'avons pas à tenir compte des taux d'échantillonnage n_h/N_h dans l'estimation de $V_m(\hat{Y}_t)$.

Nous considérons d'abord l'estimation de $V_m(\hat{Y}_t)$. Si nous pouvons calculer une formule approximative de $V_m(\hat{Y}_t)$, nous pourrons directement estimer cette valeur par substitution. Il reste que la forme explicite de $\hat{Y}_t$ est fort complexe lorsque t n'est pas petit, d'où la grande difficulté de calculer $V_m(\hat{Y}_t)$. Ainsi, dans la CES, nous adoptons une méthode de regroupement en demi-échantillons où la correction ou (réimputation) de Rao et Shao (1992) permet de tenir compte de l'imputation. Il s'agit plus précisément de constituer deux groupes aléatoires à partir des unités échantillonnées dans chaque strate. On crée des demi-échantillons R à l'aide d'une matrice d'Hadamard; $H+1 \le R \le H+4$ est le nombre de strates. Pour le $r^{\text{ième}}$ demi-échantillon et la $i^{\text{ième}}$ unité échantillonnée, nous définissons :

$$w_i^{(r)} = \begin{cases} (1+0{,}5)\, w_i & \text{si l'unité est dans le } r^{\text{ième}} \text{ demi-échantillon} \\ (1-0{,}5)\, w_i & \text{si l'unité n'est pas dans le } r^{\text{ième}} \text{ demi-échantillon,} \end{cases}$$

où w_i est la valeur initiale de pondération d'enquête. Posons que $\hat{Y}_t^{(r)}$ est le même que $\hat{Y}_t$ sauf remplacement des valeurs de pondération w_i par les valeurs de pondération $w_i^{(r)}$, ce qui comprend les valeurs utilisées dans l'imputation (en d'autres termes, $\hat{\alpha}_t$, $\hat{\gamma}_t$ et $\hat{\beta}_t$ sont recalculés pour chaque r, ce qui équivaut à la correction de Rao et Shao). Un estimateur de variance par regroupement en demi-échantillons de $V_m(\hat{Y}_t)$ sera

$$v_{t1} = \frac{4}{R} \sum_{r=1}^{R} \left(\hat{Y}_t^{(r)} - \frac{1}{R} \sum_{t=1}^{R} \hat{Y}_t^{(t)} \right)^2. \qquad (5)$$

Il convient de noter que, si nous prenons 0,5 au lieu de 1 dans l'élaboration de la pondération $w_i^{(r)}$, c'est en application de la méthode de Fay (Dippo, Fay et Morganstein 1984; Judkins 1990; Rao et Shao 1999). Asymptotiquement, v_{t1} est sans biais et convergent pour $V_m(\hat{Y}_t)$ (Shao, Chen et Chen 1998; Rao et Shao 1999; Shao et Chen 1999).

Considérons maintenant l'estimation de $V_m(Y_t)$. Dans le modèle (2),

$$V_m(Y_t) = \sum_k M_k v_{t,k},$$

ce qui est de l'ordre $O(N)$, où N est la taille de la population P. D'ordinaire, $V_m(\hat{Y}_t)$ est de l'ordre $O(N^2/n)$, où $n = \sum_h n_h$ est la taille d'échantillon. $V_m(Y_t)/V_m(\hat{Y}_t)$ est donc de l'ordre $O(n/N)$, et il est inutile d'estimer $V_m(Y_t)$ si n/N est négligeable (bien que certains taux d'échantillonnage n_h/N_h ne le soient pas).

Dans la CES cependant, n/N est d'environ 15 %, valeur non négligeable. Ainsi, l'estimation de $V_m(Y_t)$ est nécessaire. Un estimateur asymptotiquement sans biais et convergent de $V_m(Y_t)$ sera

$$v_{t2} = \sum_k M_k s_{k,t}^2, \qquad (6)$$

où $s_{k,t}^2$ est la variance habituelle d'échantillonnage selon les répondants $y_{t,i}$ dans la $k^{\text{ième}}$ cellule d'imputation.

4. ESTIMATEURS APPROXIMATIFS DE VARIANCE

On peut voir à la section 3 qu'un juste estimateur de variance de $\hat{Y}_t$ est $v_{t1} - v_{t2}$, où v_{t1} et v_{t2} sont respectivement donnés par (5) et (6). Bien qu'on puisse facilement étendre la détermination de v_{t1} au cas où $\hat{Y}_t$ est remplacé

par un estimateur non linéaire comme $\hat{Y}_t^P / \hat{Y}_t^H$ (rémunération hebdomadaire rapportée aux heures), il faut, pour une extension semblable de v_{t2}, procéder par développement de Taylor pour chaque estimateur non linéaire. Ainsi, dans la CES, il est souhaitable de calculer un estimateur approximatif de variance qui ne soit pas parfaitement juste et n'exige pas de calcul de v_{t2}.

À noter que, si n/N est négligeable, nous pouvons simplement nous servir de v_{t1} comme estimateur de $V(\hat{Y}_t - Y_t)$. Dans la CES toutefois, l'utilisation de v_{t1} est source de surestimation, puisque n/N n'est pas d'une valeur négligeable (voir aussi les résultats de simulation à la section 5). Comme cette surestimation est causée par le taux d'échantillonnage, une solution possible est d'intégrer les taux d'échantillonnage à la méthode de regroupement en demi-échantillons. Là où il n'y a pas de non-réponse, l'intégration peut se faire par la formule (2), $w_i^{(r)}$ étant alors remplacé par

$$\tilde{w}_i^{(r)} = \begin{cases} (1 + 0{,}5\sqrt{1 - n_h/N_h})\, w_i & \text{si l'unité est dans le } r^{\text{ième}} \text{ demi-échantillon} \\ (1 - 0{,}5\sqrt{1 - n_h/N_h})\, w_i & \text{si l'unité n'est pas dans le } r^{\text{ième}} \text{ demi-échantillon,} \end{cases} \quad (7)$$

où i se trouve dans la strate h.

Soit $\tilde{v}_{t1}$ l'estimateur de variance obtenu à l'aide de (5), mais après substitution de $\tilde{w}_i^{(r)}$ à $w_i^{(r)}$. Si nous utilisons $\tilde{v}_{t1}$ comme estimateur de $V(\hat{Y}_t - Y_t)$, il y a toutefois biais négatif, mais on doit préférer cet estimateur au simple estimateur qui assimile des valeurs d'imputation à des valeurs d'observation (voir les résultats de simulation à la section 5).

Si v_{t1} surestime la juste variance $V(\hat{Y}_t - Y_t)$ et que $\tilde{v}_{t1}$ la sous-estime, un moyen terme est possible où on remplacera le taux d'échantillonnage n_h/N_h dans (7) par le « taux d'échantillonnage estimé » $r_{h,t}/N_h$, où $r_{h,t}$ est le nombre de répondants dans la strate h pour le mois t. Soit $\hat{v}_{t1}$ l'estimateur de variance obtenu par (5) et (7), mais avec remplacement de n_h/N_h par $r_{h,t}/N_h$. Dans ce cas,

$$\tilde{v}_{t1} \le \hat{v}_{t1} \le v_{t1}.$$

Les trois estimateurs de variance sont asymptotiquement sans biais et approximativement égaux si n/N est négligeable. Pour un n/N non négligeable toutefois, ils sont asymptotiquement entachés d'un biais.

Pour dégager l'ordre de grandeur des biais de $\tilde{v}_{t1}$, $\hat{v}_{t1}$ et v_{t1}, nous considérons le cas le plus simple où il n'y a pas de strates et où $t = 1$. Soit $y_i = y_{1,i}^E$, $x_i = y_{0,i}^E$ et

$$\hat{Y} = \sum a_i y_i + \sum (1 - a_i)\hat{R} x_i,$$

où $a_i = 1$ si y_i est un répondant et $a_i = 0$ dans les autres cas, $\hat{R} = \sum a_i y_i / \sum a_i x_i$ et toutes les sommations sont sur $i \in S$. Soit $\hat{U} = (\sum x_i/n)/(\sum a_i x_i/r)$, où r est le nombre de répondants y. Le juste estimateur de variance pour $\hat{Y}$ est alors $v_1 - v_2$ avec

$$v_1 = \frac{N^2 \hat{U}^2 s_d^2}{r} + \frac{2N^2 \hat{U}\hat{R} s_{dx} + N^2 \hat{R}^2 s_x^2}{n}$$

et

$$v_2 = N\hat{U} s_d^2 + 2N\hat{U}\hat{R} s_{dx} + N\hat{R}^2 s_x^2,$$

où $s_d^2 = (r-1)^{-1} \sum a_i (y_i - \hat{R} x_i)^2$, $s_{dx} = (r-1)^{-1} \sum a_i x_i (y_i - \hat{R} x_i)$, et s_x^2 est la variance d'échantillon fondée sur les x_i. De plus,

$$\tilde{v}_1 = \left(1 - \frac{n}{N}\right)\left(\frac{N^2 \hat{U}^2 s_d^2}{r} + \frac{2N^2 \hat{U}\hat{R} s_{dx} + N^2 \hat{R}^2 s_x^2}{n}\right)$$

$$= v_1 - \frac{nN\hat{U}^2 s_d^2}{r} - 2N\hat{U}\hat{R} s_{dx} - N\hat{R}^2 s_x^2$$

et

$$\hat{v}_1 = \left(1 - \frac{r}{N}\right)\left(\frac{N^2 \hat{U}^2 s_d^2}{r} + \frac{2N^2 \hat{U}\hat{R} s_{dx} + N^2 \hat{R}^2 s_x^2}{n}\right)$$

$$= v_1 - N\hat{U}^2 s_d^2 - \frac{2rN\hat{U}\hat{R} s_{dx} + rN\hat{R}^2 s_x^2}{n}.$$

Comme $v_1 - v_2$ est asymptotiquement sans biais, le biais de $v_{t1} = v_1$ est du même ordre que v_2 et toujours non négatif. Le biais de $\tilde{v}_{t1} = \tilde{v}_1$ est du même ordre que

$$N\hat{U} s_d^2 \left(1 - \frac{n\hat{U}}{r}\right) = -N\hat{U} s_d^2 \frac{\sum (1 - a_i) x_i}{\sum a_i x_i}$$

et toujours non positif. Le biais de $\hat{v}_{t1} = \hat{v}$ est du même ordre que

$$N\hat{U}(1 - \hat{U}) s_d^2 + \left(1 - \frac{r}{n}\right)\left(2N\hat{U}\hat{R} s_{dx} + N\hat{R}^2 s_x^2\right). \quad (8)$$

Le biais en (8) est non négatif si $s_{dx} \ge 0$ et $\hat{U} \approx 1$ (ce qui se vérifie lorsque a_i est indépendant de x_i).

5. CERTAINS RÉSULTATS DE SIMULATION

Pour étudier plus avant les biais des estimateurs de variance v_{t1}, $\tilde{v}_{t1}$ et $\hat{v}_{t1}$, nous avons effectué une étude de simulation à l'aide d'un ensemble de données de la CES (des années 1980) comprenant 149 044 unités comme

population P. Chaque unité $i \in P$ a un vecteur $\boldsymbol{y}_i = (y_{t,i}^E, y_{t,i}^W, y_{t,i}^H, y_{t,i}^P,\ t = 0, 1, ..., 7)$ et un vecteur $\boldsymbol{r}_i$ formé des indicateurs de réponse des éléments de $\boldsymbol{y}_i$, bien que toutes les valeurs de $\boldsymbol{y}_i$ soient disponibles (dans les dossiers administratifs). Dans la simulation, nous obtenons l'échantilon S par échantillonnage aléatoire simple stratifié $\{\boldsymbol{y}_i\}$ pour une taille de 23 092 unités de P selon la répartition indiquée au tableau 1. Dans cette simulation, les indicateurs de réponse de $\{\boldsymbol{y}_i\}$ sont issus d'un autre échantillonnage aléatoire simple stratifié (indépendant) $\{\boldsymbol{r}_i\}$ d'unités de P. Ainsi, les non-répondants sont des éléments aléatoires qui se répartissent selon les valeurs des $\boldsymbol{r}_i$ dans l'ensemble de données P, mais en toute indépendance des $\boldsymbol{y}_i$.

Après obtention des non-répondants et des données d'échantillon, il y a eu imputation à l'égard des premiers comme le décrit la section 2. Nous avons calculé les totaux mensuels $\hat{Y}_t$ et les variations mensuelles $\hat{Y}_t - \hat{Y}_{t-1}$ estimés en fonction des données d'imputation. Le calcul de leurs estimateurs de variance $\tilde{v}_{t1}, \hat{v}_{t1}, v_{t1}$ et $v_{t1} - v_{t2}$ est celui que décrivent les sections 3 et 4. À des fins de comparaison, nous avons enfin calculé le simple estimateur de variance v_{t0} par assimilation des valeurs d'imputation aux données d'observation.

Les tableaux 2 à 5 présentent respectivement pour 4 variables les valeurs obtenues en 1 000 simulations des biais relatifs (BR) et des variances (Var) des totaux $\hat{Y}_t$ et des variations $\hat{Y}_t - \hat{Y}_{t-1}$ estimés, du BR et du coefficient de variation (CV) des estimateurs de variance pour $\hat{Y}_t$ et $\hat{Y}_t - \hat{Y}_{t-1}$, de la probabilité de « couverture » (PC) des intervalles de confiance approximatifs à 95 % (sous la forme :

$$\text{estimation} \pm 1{,}96 \sqrt{\text{variance estimative}})$$

et de la largeur moyenne (LM) des intervalles. Les erreurs-types estimées de simulation sont de 2 % pour le BR, le CV et la LM et de 0,5 % pour la PC.

Aux tableaux 2 à 5, les biais relatifs des estimateurs des totaux et des variations mensuels sont négligeables pour toutes les variables. Voici en résumé les résultats de simulation pour les estimateurs de variance (BR et CV).

1. Comme on pouvait le prévoir, l'estimateur simple de variance v_{t0} est entaché d'un important biais relatif en valeur négative.

2. L'estimateur de variance asymptotiquement non biaisé $v_{t1} - v_{t2}$ est d'un bon rendement général. Son biais relatif est toujours de moins de 10 % en valeur absolue. Souvent, il n'atteint pas les 5 %.

3. L'estimateur de variance v_{t1} est entaché dans tous les cas d'un grand biais relatif en valeur positive, ce qui indique que le terme v_{t2} n'est pas négligeable dans la CES où le taux d'échantillonnage global n/N est d'environ 15 %.

4. L'estimateur de variance $\tilde{v}_{t1}$, qui est le même que v_{t1} mais après intégration des taux d'échantillonnage n_h/N_h (section 4), présente en général un biais relatif en valeur négative. Ce biais négatif peut être important, plus particulièrement dans l'estimation de variance pour les variations mensuelles.

5. L'estimateur de variance $\hat{v}_{t1}$, qui est le même que $\tilde{v}_{t1}$ mais après remplacement des taux d'échantillonnage n_h/N_h par $r_{h,t}/N_h$, est d'un bon rendement dans l'étude de simulation, bien que n'étant pas asymptotiquement sans biais (section 4). Son biais relatif est important dans quelques cas, entre autres dans les estimations de variance pour le total de la rémunération hebdomadaire dans les mois 1 et 4, pour le total des heures dans le mois 1 et pour la variation d'emploi dans le mois 7. Dans bien des cas toutefois, $\hat{v}_{t1}$ est encore d'un meilleur rendement que l'estimateur asymptotiquement sans biais $v_{t1} - v_{t2}$.

Voici en résumé les résultats de simulation des intervalles de confiance pour ce qui est de la PC et de la LM.

1. La PC de l'intervalle de confiance pour l'estimateur simple de variance v_{t0} se situe bien en deçà dans la plupart des cas du niveau nominal de 95 %.

2. La PC de l'intervalle de confiance pour l'estimateur de variance asymptotiquement valide $v_{t1} - v_{t2}$ est comprise entre 90 % et 93 % dans la plupart des cas. Tel est souvent le cas avec un estimateur de variance asymptotiquement valide : le biais relatif est petit, mais la PC de l'intervalle de confiance correspondant est inférieure au niveau nominal. La raison en est peut-être que la convergence de distribution (normalité asymptotique qui est à la base même des intervalles de confiance asymptotiques) exige plus en taille d'échantillon que la convergence du deuxième moment (estimation de variance).

3. En ce qui concerne la PC, l'intervalle de confiance pour v_{t1} est le meilleur, peut-être parce que la surestimation de variance compense la sous-couverture de l'estimation d'intervalle. La largeur moyenne de l'intervalle pour v_{t1} peut être bien supérieure à celle d'autres intervalles, notamment pour la variable de la rémunération hebdomadaire.

4. La PC de l'intervalle de confiance pour $\hat{v}_{t1}$, qui n'est pas asymptotiquement valide, est semblable à celle de l'intervalle pour $v_{t1} - v_{t2}$.

Tableau 1
Taille d'échantillon par strate

CTI	Taille	Taille de strate	Taille d'échantillon	Taux d'échantillonnage	CTI	Taille	Taille de strate	Taille d'échantillon	Taux d'échantillonnage
10, 12-14	1	567	14	0,02439	50-51	1	3631	66	0,01812
	2	433	303	0,70000		2	3678	183	0,04987
	3	526	526	1,00000		3	4300	403	0,09375
	4	210	210	1,00000		4	1831	289	0,15789
	5	165	165	1,00000		5	833	320	0,38461
15-17	1	5055	129	0,02549	52-59	1	7084	149	0,02103
	2	4476	570	0,12731		2	5701	440	0,07724
	3	5281	1154	0,21854		3	8363	1037	0,12403
	4	2111	836	0,39583		4	4511	763	0,16915
	5	1005	1005	1,00000		5	4087	1002	0,24528
24-25, 32-29	1	3103	73	0,02349	60-62, 67	1	1384	17	0,01230
	2	3905	331	0,08475		2	971	38	0,03906
	3	6381	891	0,13966		3	1529	115	0,07500
	4	4273	1036	0,24242		4	981	67	0,06818
	5	4143	2127	0,51351		5	728	73	0,10000
20-23, 26-31	1	1754	40	0,02276	63-64	1	1364	15	0,01119
	2	1953	128	0,06564		2	652	20	0,03125
	3	3591	524	0,14599		3	754	87	0,11538
	4	3108	596	0,19167		4	435	48	0,11110
	5	3448	1041	0,30189		5	344	57	0,16667
40-49	1	1648	31	0,01902	7, 70-99	1	9641	230	0,02385
	2	1463	101	0,06918		2	6701	643	0,09602
	3	1988	221	0,11111		3	7833	1275	0,16275
	4	1171	211	0,18033		4	4839	1317	0,27215
	5	759	108	0,14286		5	4352	2067	0,47500

Tableau 2
Résultats de simulation pour la variable de l'emploi

	Estimation du total			Estimation de variance pour le total estimatif																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Mois	Total*	BR	Var*	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC	LM
1	6,7E6	0,0	5,5E7	-37,0	47,6	85,3	7,7	-4,1	67,5	92,3	9,2	4,9	69,8	93,1	9,6	19,5	76,1	95,1	10,3	7,4	67,4	92,8	9,7
2	6,8E6	0,0	8,8E7	-34,3	28,8	86,9	9,6	-7,3	40,4	92,6	11,4	0,9	42,9	93,6	12,4	15,3	47,6	94,7	12,7	4,4	49,1	92,3	12,1
3	6,9E6	0,0	1,4E8	-26,1	30,4	88,2	12,9	-4,1	42,3	91,8	14,7	1,4	44,2	92,9	15,1	18,8	49,9	94,8	16,3	3,6	50,5	90,8	15,2
4	6,9E6	0,0	2,1E8	-22,5	32,9	89,3	16,1	-2,4	44,0	92,1	18,1	3,8	46,3	92,7	18,7	22,3	53,1	94,7	20,3	2,7	51,3	91,4	18,6
5	6,9E6	0,0	2,7E8	-21,9	35,0	88,3	18,4	-7,7	45,2	90,9	20,0	-1,1	47,9	92,0	20,7	16,2	55,6	94,4	22,4	-4,7	54,2	90,9	20,3
6	6,9E6	0,0	2,0E8	-8,8	40,5	91,7	17,1	-5,2	41,7	91,9	17,4	0,0	43,6	93,1	17,9	19,7	51,8	95,5	19,6	-3,1	52,5	90,5	17,6
7	6,9E6	0,0	1,5E8	-12,4	34,8	91,8	14,5	-8,6	36,1	92,5	14,8	-2,0	38,3	93,6	15,3	16,8	45,0	96,2	16,7	-6,6	42,4	92,7	15,0

	Estimation de la variation			Estimation de variance pour la variation estimative																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Mois	Variation*	BR	Var*	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC	LM
2	8,0E4	-0,1	6,1E7	-43,0	25,4	84,9	7,5	-11,3	41,4	92,3	9,3	-4,5	43,9	93,7	9,7	9,4	48,7	95,6	10,3	8,6	51,7	93,5	10,3
3	9,7E4	-1,8	7,4E7	-35,0	31,7	85,0	8,7	-8,5	46,0	90,5	10,4	-3,2	47,7	91,0	10,7	11,7	53,1	93,4	11,5	-3,1	48,8	90,9	10,7
4	1,8E4	2,9	1,1E8	-31,8	42,3	87,4	11,0	-0,9	60,6	93,1	13,2	4,9	63,2	93,6	13,6	25,0	73,5	95,9	14,8	-2,5	47,7	89,9	13,1
5	4,4E4	3,4	1,1E8	-41,9	34,5	83,1	10,1	-10,8	57,3	91,4	12,5	-4,9	60,4	92,3	12,9	13,2	69,4	94,6	14,1	0,8	94,1	93,1	13,3
6	-1,1E4	9,3	1,1E8	-41,0	29,9	84,1	10,2	-12,6	42,0	91,1	12,4	-6,4	44,2	93,0	12,8	9,4	50,2	94,6	13,9	-4,1	53,9	93,0	13,0
7	1,6E3	3,2	1,2E8	-43,8	38,4	82,9	10,4	-15,9	57,5	89,6	12,7	-11,3	60,1	90,5	13,1	5,6	69,9	92,6	14,2	-0,2	75,5	90,0	13,8

Total : total de population.
Variation : différence de population entre le mois en cours et le mois précédent.
Var : variance du total ou de la variation estimés.
BR : biais relatif = 100 (biais/juste valeur) %.
CV : coefficient de variation = 100 (erreur-type/juste valeur) %.
PC : Probabilité de couverture de l'intervalle de confiance asymptotique avec variance estimée (en %).
LM : (Largeur moyenne de l'intervalle de confiance asymptotique) / 10^7.
* : Notation scientifique (Exemple 6 700 000 est 6,7E6).

Tableau 3
Résultats de simulation pour les travailleurs hors personnel de surveillance

	Estimation du total			Estimation de variance pour le total estimatif																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Mois	Total*	BR	Var*	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC	LM
1	5,4E6	-0,1	4,6E7	-33,3	49,7	80,9	7,0	-4,4	66,1	88,1	8,4	4,6	68,6	89,9	8,8	19,8	75,5	92,3	9,4	3,4	65,3	89,9	8,7
2	5,5E6	-0,1	7,6E7	-30,6	31,4	84,0	9,2	-7,4	41,1	89,4	10,6	0,9	43,7	91,0	11,1	15,8	48,7	93,8	11,9	4,2	50,8	88,3	11,3
3	5,6E6	-0,1	1,2E8	-23,6	31,2	85,6	12,1	-4,8	41,0	89,5	13,5	0,7	42,9	90,0	13,9	18,4	48,7	93,1	15,1	3,9	50,9	90,8	14,1
4	5,6E6	-0,1	1,9E8	-19,0	34,5	88,4	15,7	-2,4	43,8	91,7	17,2	3,8	46,3	91,9	17,8	22,5	53,2	94,1	19,3	1,8	71,7	90,5	17,6
5	5,7E6	-0,1	2,4E8	-18,9	36,8	87,8	17,6	-7,1	45,3	89,7	18,9	-0,4	48,2	90,7	19,6	17,2	56,0	93,0	21,2	-4,1	54,7	90,4	19,2
6	5,7E6	0,0	1,8E8	-7,6	41,7	91,8	16,3	-4,7	42,8	92,4	16,6	0,6	44,8	92,7	17,0	20,6	53,1	95,4	18,6	-3,3	53,1	90,5	16,7
7	5,7E6	0,0	1,4E8	-10,9	36,1	91,9	14,1	-7,7	37,2	92,2	14,4	1,0	39,4	93,6	15,0	18,3	46,3	95,9	16,3	-8,5	42,5	92,6	14,3

	Estimation de la variation			Estimation de variance pour la variation estimative																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Mois	Variation*	BR	Var*	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC	LM
2	7,7E4	-0,8	5,1E7	-40,8	27,0	84,5	7,0	-12,9	41,2	91,5	8,4	-6,0	43,7	92,4	8,8	8,2	48,8	94,4	9,4	9,9	54,8	93,0	9,5
3	9,1E4	-1,4	6,2E7	-31,2	32,1	86,4	8,3	-8,7	42,8	91,2	9,5	-3,2	44,5	91,7	9,8	12,3	49,9	94,1	10,6	-3,1	47,7	91,3	9,8
4	1,6E4	19,6	9,1E7	-27,2	44,0	87,1	10,3	-1,1	59,4	92,8	12,0	4,7	62,1	94,1	12,3	24,9	73,0	95,8	13,5	-5,3	91,8	91,4	11,7
5	4,4E4	-0,4	9,5E7	-37,5	38,4	83,4	9,7	-10,0	58,6	90,8	11,7	-3,9	61,8	91,3	12,1	14,5	71,4	93,4	13,2	-2,1	79,4	92,3	12,2
6	-1,0E4	-19,3	9,0E7	-37,0	32,4	83,4	9,5	-11,1	43,1	89,6	11,3	-4,7	45,5	90,4	11,7	11,7	51,8	92,4	12,7	-3,3	54,7	90,9	11,8
7	7,9E2	48,7	1,0E8	-39,3	42,6	83,7	9,9	-14,5	59,7	89,2	11,7	-9,8	62,4	90,2	12,0	7,6	72,6	92,6	13,1	-1,3	76,8	90,6	12,6

Total : total de population.
Variation : différence de population entre le mois en cours et le mois précédent.
Var : variance du total ou de la variation estimés.
BR : biais relatif = 100 (biais/juste valeur) %.
CV : coefficient de variation = 100 (erreur-type/juste valeur) %.
PC : Probabilité de couverture de l'intervalle de confiance asymptotique avec variance estimée (en %).
LM : (Largeur moyenne de l'intervalle de confiance asymptotique) / 10^7.
* : Notation scientifique (Exemple 6 700 000 est 6,7E6).

Tableau 4
Résultats de simulation pour les heures travaillées

	Estimation du total			Estimation de variance pour la variation estimative n																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Mois	Total*	RB	Var*	RB	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM
1	1,9E8	-0,1	5,8E10	-31,5	28,0	79,0	8,0	2,3	44,4	88,3	9,7	12,3	46,5	90,5	10,2	33,4	53,4	93,6	11,1	8,0	48,7	90,9	10,0
2	2,0E8	-0,1	1,2E11	-30,2	32,8	84,7	11,6	-7,7	40,4	90,6	13,3	0,1	42,8	91,7	13,9	19,7	49,4	94,3	15,2	3,8	49,1	90,1	14,1
3	2,0E8	-0,1	1,8E11	-23,3	30,0	86,3	14,9	-6,3	36,7	90,3	16,4	-1,0	38,1	91,2	16,9	19,6	43,8	94,6	18,6	1,4	45,2	90,7	17,1
4	2,0E8	0,0	3,2E11	-20,2	35,6	90,2	20,2	-0,5	47,1	93,4	22,6	5,6	49,7	93,3	23,3	27,9	59,8	95,3	25,6	-0,4	79,7	91,2	22,6
5	2,1E8	0,0	4,4E11	-21,2	40,5	88,9	23,6	-7,9	52,3	90,7	25,5	-1,6	55,1	92,0	26,3	18,0	64,4	94,2	28,8	-5,1	64,2	90,9	25,8
6	2,1E8	0,0	3,4E11	-10,4	46,3	92,1	22,1	-5,9	48,9	92,2	22,6	-1,0	50,7	93,0	23,2	20,8	59,9	94,7	25,6	-3,3	65,7	90,3	22,9
7	2,1E8	0,0	2,3E11	-7,0	40,8	93,0	18,5	-2,2	42,8	93,2	19,0	4,2	44,7	94,1	19,6	27,2	53,2	95,8	21,6	-7,7	49,0	90,9	18,4

	Estimation de la variation			Estimation de variance pour la variation estimative																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Mois	Variation*	RB	Var*	RB	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM
2	5,0E6	0,1	8,8E10	-38,8	25,9	89,0	9,3	-9,7	35,1	92,4	11,3	-2,2	37,2	93,7	11,7	16,3	43,0	96,1	12,8	6,7	43,8	93,6	12,3
3	3,8E6	-1,0	1,1E11	-36,5	25,2	88,4	10,6	-12,6	34,5	91,9	12,4	-6,7	36,0	92,4	12,8	10,4	41,2	93,9	13,9	0,4	41,3	93,2	13,3
4	1,0E6	11,0	2,1E11	-31,2	45,6	87,3	15,2	-5,0	59,3	90,9	17,9	0,6	62,4	91,6	18,4	21,6	75,2	93,9	20,2	-3,4	98,8	91,5	18,0
5	2,1E6	-0,5	2,2E11	-41,6	39,9	85,6	14,3	-14,3	63,9	91,1	17,4	-8,4	66,6	90,1	18,0	10,5	76,0	94,9	19,7	1,5	95,1	93,2	18,9
6	-7,7E5	-7,8	1,9E11	-40,1	35,1	82,5	13,5	-12,7	47,5	89,5	16,3	-6,5	50,3	90,7	16,9	12,7	60,1	94,1	18,5	-9,5	55,1	91,1	16,6
7	2,5E5	-7,2	2,1E11	-39,0	48,4	82,9	14,3	-15,1	60,3	89,5	16,9	-10,6	62,4	90,3	17,3	8,0	72,3	94,0	19,0	-3,9	82,0	91,8	18,0

Total : total de population.
Variation : différence de population entre le mois en cours et le mois précédent.
Var : variance du total ou de la variation estimés.
BR : biais relatif = 100 (biais/juste valeur) %.
CV : coefficient de variation = 100 (erreur-type/juste valeur) %.
PC : Probabilité de couverture de l'intervalle de confiance asymptotique avec variance estimée (en %).
LM : (Largeur moyenne de l'intervalle de confiance asymptotique) / 10^{10}.
* : Notation scientifique (Exemple 6 700 000 est 6,7E6).

Tableau 5

Résultats de simulation pour la rémunération hebdomadaire

	Estimation du Total			Estimation de variance pour le total estimatif																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Mois	Total*	RB	Var*	RB	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM
1	2,0E9	-0,1	9,5E12	-30,7	30,4	81,8	10,3	1,7	41,0	90,0	12,4	17,2	44,3	92,4	13,3	39,8	54,4	94,4	14,6	4,3	48,9	91,0	12,6
2	2,1E9	-0,1	1,7e[13]	-27,2	27,8	84,3	14,1	-3,4	38,7	89,2	16,2	7,9	41,2	91,2	17,1	31,1	48,1	93,5	18,9	3,3	51,5	91,6	16,8
3	2,1E9	-0,1	2,2e[13]	-14,3	34,7	85,6	17,4	1,1	42,2	88,1	18,9	8,0	43,9	89,5	19,5	34,9	51,4	93,5	21,8	2,6	50,4	91,4	19,0
4	2,2E9	-0,1	3,7e[13]	-12,3	40,3	90,1	22,8	6,4	50,6	92,8	25,1	13,8	53,0	94,1	26,0	41,2	63,0	96,1	28,9	-0,9	84,5	92,8	24,2
5	2,2E9	-0,1	5,0e[13]	-16,0	41,6	89,0	25,9	-1,5	51,8	91,4	28,1	5,9	54,8	92,0	29,1	29,3	64,6	94,3	32,2	-5,4	56,0	92,4	27,5
6	2,2E9	-0,1	4,5e[13]	-9,4	44,1	92,0	25,5	-3,8	46,9	92,6	26,3	1,8	48,7	92,8	27,1	27,8	57,8	95,0	30,3	-0,4	54,1	94,2	26,8
7	2,2E9	-0,1	3,5e[13]	-7,3	43,1	92,1	22,8	-0,7	48,3	92,8	23,6	6,8	50,0	93,9	24,5	31,9	57,0	96,4	27,2	-0,0	54,3	95,3	23,7

	Estimation de la variation			Estimation de variance pour la variation estimative																			
				v_{t0}				$\tilde{v}_{t1}$				$\hat{v}_{t1}$				v_{t1}				$v_{t1} - v_{t2}$			
Mois	Variation*	RB	Var*	RB	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM
2	6,4E7	-0,1	1,5E13	-37,6	25,7	85,4	12,2	-8,2	38,4	93,0	14,8	0,2	40,4	94,1	15,5	21,6	47,7	95,8	17,1	5,5	49,2	92,6	15,9
3	3,5E7	-1,6	1,3E13	-31,7	27,9	87,7	11,9	-5,2	42,3	92,2	14,0	2,2	43,8	92,8	14,6	22,3	48,9	94,3	15,9	3,5	43,2	93,5	14,7
4	2,1E7	6,6	2,4E13	-29,5	47,1	86,7	16,5	0,4	63,2	91,9	19,6	6,7	66,2	92,6	20,2	30,7	78,7	95,2	22,4	-4,3	96,9	90,6	19,2
5	2,1E7	-0,4	2,4E13	-40,5	34,1	83,5	15,1	-9,2	55,7	90,5	18,7	-2,4	58,9	92,0	19,4	19,9	69,2	94,9	21,5	3,6	90,0	92,5	19,9
6	1,4E7	2,0	2,3E13	-40,8	31,1	84,4	14,8	-13,5	46,0	91,4	17,8	-6,7	48,9	92,1	18,5	16,8	60,1	94,5	20,7	-4,4	53,0	91,5	18,8
7	1,1E7	-0,1	2,7E13	-40,5	42,0	83,1	16,0	-13,9	56,5	89,2	19,3	-8,7	58,7	90,6	19,9	13,0	68,8	92,8	22,1	-3,7	69,5	90,8	20,4

Total : total de population.
Variation : différence de population entre le mois en cours et le mois précédent.
Var : variance du total ou de la variation estimés.
BR : biais relatif = 100 (biais/juste valeur) %.
CV : coefficient de variation = 100 (erreur-type/juste valeur) %.
PC : Probabilité de couverture de l'intervalle de confiance asymptotique avec variance estimée (en %).
LM : (Largeur moyenne de l'intervalle de confiance asymptotique) / 10^{12}.
* : Notation scientifique (Exemple 6 700 000 est 6,7E6).

6. CONCLUSION ET EXAMEN

Pour les estimateurs avec données d'imputation de la Current Employment Survey (CES), nous proposons un estimateur $v_{t1} - v_{t2}$ asymptotiquement sans biais et convergent (section 3). S'il est facile de calculer v_{t1} par la méthode de regroupement en demi-échantillons équilibrés, l'établissement de v_{t2} comporte des calculs distincts pour les estimateurs non linéaires. Ainsi, nous considérons plusieurs approximations, v_{t1}, $\tilde{v}_{t1}$ et $\hat{v}_{t1}$ (section 4), et les comparons à $v_{t1} - v_{t2}$ dans une étude de simulation où un ensemble de données de la CES nous sert de population. Les résultats indiquent que v_{t1} et $\tilde{v}_{t1}$ ont d'importants biais relatifs imputables au caractère non négligeable du taux d'échantillonnage global (15 %). L'estimateur $\hat{v}_{t1}$, qui est le même que v_{t1}, mais après intégration d'un taux d'échantillonnage estimé (par le taux de réponse) dans l'application de la méthode de regroupement en demi-échantillons équilibrés, est d'un assez bon rendement. Ainsi, nous recommandons de remplacer $v_{t1} - v_{t2}$ par $\hat{v}_{t1}$ si le calcul de v_{t2} est trop complexe. Comme le recours au « taux d'échantillonnage observé » $r_{h,t}/N_h$ ne tient pas compte de ce que des données des mois antérieurs soient disponibles sur les non-répondants, il est possible d'améliorer $\hat{v}_{t1}$ en faisant intervenir un taux d'échantillonnage estimé plus fidèle, ce qui peut être, par exemple, la « fraction de données manquantes » de Rubin (1987).

Notre étude vise la CES, mais les résultats obtenus sont applicables à toute enquête ayant un plan d'échantillonnage et une méthode d'imputation semblables. Ajoutons qu'une extension au cas où le modèle (2) comprend $y_{t,i}$, $y_{t-1,i}$, ..., $y_{t-s,i}$ avec un nombre entier $s \geq 2$ est chose simple, bien que le calcul de v_{t2} (pour un estimateur de variance asymptotiquement valide) soit plus compliqué.

REMERCIEMENTS

Les auteurs remercient un rédacteur adjoint et deux examinateurs de leurs observations et de leurs suggestions utiles. La recherche de Jun Shao est partiellement financé par la bourse NSF 9803112, 01-02223 de la DMS et de la bourse NSA 904-99-1-0032 de la MDA.

ANNEXE : DÉMONSTRATION DE (4)

Il suffit de démontrer que

$$\mathrm{Cov}_m(\hat{Y}_t, Y_t) = V_m(Y_t). \tag{9}$$

Nous présentons le cas d'une cellule d'imputation unique et d'une égalité $y_{t,i} = y_{t,i}^E$ (emploi). Le cas général peut faire l'objet d'un traitement semblable.

Nous procédons par induction mathématique. Si $t = 1$,

$$\hat{Y}_t = \hat{\alpha}_1 Y_0.$$

Par l'hypothèse (2),

$$\begin{aligned}\mathrm{Cov}_m(\hat{Y}_t, Y_t) &= \alpha_1^2 V_m(Y_0) + \sigma^2 E_m(Y_0)\\ &= N(\alpha_1^2 v_0 + \sigma^2 \mu_0)\\ &= V_m(Y_1).\end{aligned}$$

Posons maintenant que (9) se vérifie au moment $t - 1$. Soit E_t, V_t et Cov_t l'espérance, la variance et la covariance en conditionnalité par $y_{j-1,i}, R_j, j=1, ..., t$. Dans ce cas,

$$E_t(\hat{Y}_t) = \alpha_t \hat{Y}_{t-1}$$

et

$$\begin{aligned}\mathrm{Cov}_t(\hat{Y}_t, Y_t) &= \mathrm{Cov}_t(\hat{\alpha}_t \hat{Y}_{t-1}, Y_t)\\ &= \hat{Y}_{t-1}\, \mathrm{Cov}_t(\hat{\alpha}_t, Y_t)\\ &= \sigma^2 \hat{Y}_{t-1},\end{aligned}$$

où la dernière égalité procède de l'hypothèse (2). Par l'hypothèse d'induction,

$$\mathrm{Cov}_m(\hat{Y}_t, Y_{t-1}) = V_m(Y_{t-1}).$$

Alors,

$$\begin{aligned}\mathrm{Cov}_m(\hat{Y}_t, Y_t) &= \mathrm{Cov}_m\left[E_t(\hat{Y}_t), E_t(Y_t)\right] + E_m\left[\mathrm{Cov}_t(\hat{Y}_t, Y_t)\right]\\ &= \alpha_t^2 \mathrm{Cov}_m(\hat{Y}_{t-1}, Y_{t-1}) + \sigma^2 E_m(\hat{Y}_{t-1})\\ &= \sigma_t^2 V_m(Y_{t-1}) + \sigma^2 E_m(Y_{t-1})\\ &= V_m(Y_t).\end{aligned}$$

BIBLIOGRAPHIE

BUTANI, S., HARTER, R. et WOLTER, K. (1997). Estimation procedures for the Bureau of Labor Statistics Current Employment Statistics Program. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 523-528.

DIPPO, C.S., FAY, R.E. et MORGANSTEIN, D.H. (1984). Computing variances from complex samples with replicate weights. Dans *Proceedings of the Section on Survey Research Methodology*, American Statistical Association. 489-494.

JUDKINS, D.R. (1990). Fay's method of variance estimation. *Journal of the Official Statistical*, 6, 223-239.

LEE, H., RANCOURT, E and SÄRNDAL C.-E. (1994). Experiment with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

RAO, J.N.K. et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation, *Biometrika*, 79, 811-822.

RAO, J.N.K., et SHAO, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

SHAO, J., et CHEN, Y. (1999). Approximate balanced half samples and related replication methods for imputed survey data. *Sankhya, B*, Special Issue on Sample Surveys, 187-201.

SHAO, J., CHEN, Y. et CHEN, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

Techniques d'enquête, juin 2002
Vol. 28, n° 1, pp. 103-108
Statistique Canada

Application de l'estimation de la variance selon Rao-Shao en utilisant des poids de rééchantillonnage

MICHAEL P. COHEN[1]

RÉSUMÉ

Pour estimer la variance en tenant compte de l'imputation pour la non-réponse partielle, Rao et Shao (1992) ont lancé une méthode fondée sur un rééchantillonnage corrigé. Shao, Chen et Chen (1998) ont apporté plusieurs perfectionnements à la méthode (particulièrement l'extension du rééchantillonnage par le jackknife de Rao et Shao au rééchantillonnage équilibré une méthode BRR). Dans le présent article, nous examinons comment il est possible d'appliquer ces méthodes en utilisant des poids de rééchantillonnage.

MOTS CLÉS : Rééchantillonnage équilibré (méthode BRR); rééchantillonnage par le Jackknife; imputation; non-réponse partielle; méthode hot deck pondérée.

1. INTRODUCTION

L'utilisation de poids de rééchantillonnage facilite l'estimation de la variance par les méthodes de rééchantillonnage (Dippo, Fay et Morganstein 1984). Au cours de la dernière décennie, plusieurs méthodes de ce genre ont été mises au point (Rao et Shao 1992; Shao, Chen et Chen 1998) afin de tenir compte de la variation due à l'imputation pour la non-réponse partielle lors de l'estimation de la variance. Toutefois, la façon d'appliquer ces méthodes de rééchantillonnage corrigé en se servant de poids de rééchantillonnage n'est pas entièrement évidente. Le présent article examine les moyens de le faire. Il se concentre sur la façon de préparer l'ensemble de données pour que les logiciels courants d'estimation de la variance à l'aide de poids de rééchantillonnage puissent être utilisés sans aucune modification. Néanmoins, à l'avant-dernière section, nous formulons certains commentaires quant à l'utilité éventuelle de la modification du logiciel.

2. MÉTHODES DE RÉÉCHANTILLONNAGE ET POIDS DE RÉÉCHANTILLONNAGE

Wolter (1985) donne une introduction complète à l'estimation de la variance dans le cas des enquêtes par sondage. Les chapitres 3 et 4 couvrent les deux méthodes par rééchantillonnage dont il est question dans le présent article, à savoir le rééchantillonnage par le jackknife et le rééchantillonnage équilibré (balanced repeated replication). Pour un traitement plus récent et plus poussé de la question, il est conseillé au lecteur de consulter Shao et Tu (1995, chapitre 6). L'estimation de la variance des données d'enquête par rééchantillonnage continue de faire l'objet de recherche. Parmi les travaux encore plus récents, mentionnons Brick et Morganstein (1996, 1997), Kott (2001), Rao et Shao (1996, 1999), Rust et Rao (1996), Shao (1996) et Valliant (1996).

Les deux méthodes de rééchantillonnage se fondent sur la création de sous-ensembles de l'échantillon appelés *répétitions*. Le schéma selon lequel sont formées les répétitions est l'élément qui différencie les deux méthodes. Dans le cas du rééchantillonnage équilibré, encore appelé méthode du demi-échantillon équilibré répété, les répétitions correspondent à environ la moitié des unités de l'échantillon original; par conséquent, elles sont également appelées *demi-échantillons*. Dans le cas du rééchantillonnage par le jackknife (tel qu'appliqué aux données d'enquête), les répétitions correspondent habituellement à l'échantillon original, dont on a supprimé une unité primaire d'échantillonnage (UPE) ou un petit nombre d'UPE appartenant à la même strate. Dans le cas des deux méthodes, les répétitions sont considérées comme des échantillons à part entière. Par conséquent, si θ représente l'estimation d'une certains quantité θ fondée sur l'échantillon original, nous pouvons former une estimation $\hat{\theta}^{(r)}$ de θ fondée sur la répétition r. Si nous procédons à R répétitions, nous estimons la variance d'échantillonnage de $\hat{\theta}$, $\mathrm{var}(\hat{\theta})$, par

$$\widehat{\mathrm{var}}(\hat{\theta}) = C_{M,R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2 \qquad (2.1)$$

où la constante $C_{M,R}$ dépend uniquement de la méthode de rééchantillonnage M et du nombre de répétitions R.

Pour former l'estimation $\hat{\theta}$ de θ, nous utilisons les poids d'échantillonnage. Par exemple, pour estimer un total de population pour une variable particulière y, nous calculons la somme pondérée des valeurs de y. Donc, si y_u et w_u sont les valeurs de y et du poids d'échantillonnage pour l'unité d'échantillonnage u, alors $\hat{\theta} = \sum_u w_u y_u$ où la somme est calculée sur l'ensemble des unités échantillonnées. En plus du poids d'échantillonnage w_u figurant dans

[1] Michael P. Cohen, Senior Mathematical Statistical, U.S.Bureau of Transportation Statistics, 400 Seventh Street SW, Washington, D.C. 20590 U.S.A.

l'enregistrement pour l'unité u, nous pouvons ajouter les poids de rééchantillonnage $w_u^{(r)}, r = 1$ à R, dans l'enregistrement qui figure dans le fichier et calculer $\hat{\theta}^{(r)}$ de la même façon que $\hat{\theta}$, mais en remplaçant $w_u^{(r)}$ par w_u pour chaque unité d'échantillonnage u. Donc, dans l'exemple où $\hat{\theta}$ représente le total de la population pour $y, \hat{\theta}^{(r)} = \sum_u w_u^{(r)} y_u$. Si l'unité u ne fait pas partie de la répétition r, alors $w_u^{(r)} = 0$. Une partie ou l'ensemble des poids de rééchantillonnage applicables aux diverses unités *incluses* dans la répétition seront plus grands que le poids d'échantillonnage, si bien que les unités comprises dans la répétition continueront de représenter l'ensemble de la population.

L'utilisation des poids de rééchantillonnage fournis dans le fichier pour calculer les estimations de la variance d'échantillonnage présente les avantages qui suivent :

- Toute statistique, aussi compliquée soit-elle, pouvant être calculée pour l'ensemble de l'échantillon peut aussi facilement l'être pour chaque répétition. La variance d'échantillonnage est alors estimée selon (2.1).
- Les corrections pour tenir compte de la non-réponse totale et de la stratification a posteriori peuvent (et devraient) être réalisées individuellement pour chaque répétition et intégrées dans les poids de rééchantillonnage. Cette correction est habituellement confiée à un spécialiste chevronné de l'échantillonnage et les poids de rééchantillonnage corrigé sont intégrés dans le fichier de sorte que l'analyste des données puisse les utiliser sans effort supplémentaire.
- Les corrections apportées aux poids de rééchantillonnage qui figurent dans le fichier peuvent se fonder sur des données auxiliaires auxquelles l'analyste n'a pas nécessairement accès, parfois pour des raisons de confidentialité, ou qu'il pourrait avoir de la difficulté à obtenir ou à utiliser, même si leur consultation n'est pas restreinte.
- Des logiciels d'usage général sont capables de traiter les poids de rééchantillonnage. Deux logiciels qui mettent l'accent sur les méthodes de rééchantillonnage applicables aux données d'enquête sont WesVar de Westat Inc. et VPLX du U.S. Census Bureau. Pour des renseignements sur les logiciels d'analyse des données d'enquête, consulter la page Web

  //www.fas.harvard.edu/~stats/survey-soft/survey-soft.html.

Dans la présente section, nous n'avons pas tenu compte des difficultés que pose l'intégration de la composante de la variance due à l'imputation visant à compenser la non-réponse partielle dans les estimations de la variance. Nous examinons ces difficultés à la section suivante.

3. MÉTHODES PAR RÉÉCHANTILLONNAGE CORRIGÉ

Les travaux de Rao et Shao (1992) et de Shao, Chen et Chen (1998) servent de fondement au présent article. Shao et Chen (1999) et Shao et Steel (1999) traitent également de l'estimation de la variance des données d'enquête par rééchantillonnage en cas d'imputation.

Nous commençons par décrire la notation, en nous inspirant en grande partie de celle utilisée par Shao, Chen et Chen (1998). La population est divisée en L strates de sorte que N_h grappes soient comprises dans la $h^{\text{ième}}$ strate. À la première étape d'échantillonnage dans la strate h, nous tirons $n_h \geq 2$ grappes, la $i^{\text{ième}}$ grappe étant sélectionnée avec la probabilité $p_{hi}, i = 1, ..., N_h; h = 1, ..., L$. Les grappes sont sélectionnées sans remise et de façon indépendante dans les diverses strates. Nous supposons que les fractions d'échantillonnage n_h / N_h sont suffisamment faibles pour éviter d'apporter une correction pour les populations finies. D'autres étapes d'échantillonnage peuvent avoir lieu dans chaque grappe, de façon indépendante de grappe en grappe. La grappe i de la strate h contient, en dernière analyse, N_{hi} unités de population. Pour chaque unité de population (h, i, j), il existe une variable étudiée y_{hij}. Posons que S représente l'ensemble des unités d'échantillonnage et que $\{\tilde{y}_{hij}, (h, i, j) \in S\}$ est l'ensemble des données qui a fait l'objet d'une imputation : les $\tilde{y}_{hij}$ sont égales aux y_{hij} lorsque la variable est observée et égales à la valeur imputée autrement. Les unités d'échantillonnage sont réparties en *classes d'imputation* notées au moyen de l'indice k et A_k représente l'ensemble des répondants pour la variable y dans la classe d'imputation k. Nous supposons que l'ensemble de données contient des idenficateurs (« signaux ») qui permettent de repérer les non-répondants.

Dans le cas des méthodes par rééchantillonnage corrigé, $\tilde{y}_{hij}$ dans la classe d'imputation k est corrigée de sorte que

$$\tilde{y}_{hij}^{(r)} = \begin{cases} \tilde{y}_{hij} + E_{A_k}^{(r)}(\tilde{y}_{hij}) - E_{A_k}(\tilde{y}_{hij}) \\ \quad \text{si la valeur de } y_{hij} \text{ est imputée} \\ y_{hij} \\ \quad \text{si la valeur de } y_{hij} \text{ est observée,} \end{cases} \tag{3.1}$$

où E_{A_k} est l'espérance en ce qui concerne la méthode originale d'imputation dans la classe d'imputation k et $E_{A_k}^{(r)}$ est l'espérance en ce qui concerne la méthode d'imputation fondée uniquement sur les données de la $r^{\text{ième}}$ répétition dans la classe d'imputation k. Cette formule est donnée explicitement par Shao, Chen et Chen (1998, page 822) pour le rééchantillonnage équilibré et pour toute une gamme de méthodes d'imputation. Elle s'applique aussi au développement présenté par Rao et Shao (1992) pour le rééchantillonnage par le jackknife et l'imputation par la méthode hot deck pondérée.

Nous adopterons la notation selon laquelle $(h^\circ i^\circ j^\circ)$ représente une unité qui n'a pas répondu à la question y et $(h'i'j')$ représente une unité qui a répondu à la question y. Nous supposons que

$$E_{A_k}(\tilde{y}_{h^\circ i^\circ j^\circ}) = \sum_{(h'i'j')\in A_k} a_{h'i'j';h^\circ i^\circ j^\circ}\, y_{h'i'j'}$$

et

$$E^{(r)}_{A_k}(\tilde{y}_{h^\circ i^\circ j^\circ}) = \sum_{(h'i'j')\in A_k} a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ}\, y_{h'i'j'}$$

où les $a_{h'i'j';h^\circ i^\circ j^\circ}$ et $a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ}$ sont des constantes qui ne dépendent pas des valeurs de $y_{h'i'j'}$ et $a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ} = 0$ lorsque $(h'i'j')$ ne figure pas dans la répétition r. Les $a_{h'i'j';h^\circ i^\circ j^\circ}$ et $a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ}$ peuvent dépendre des données auxiliaires disponibles pour toutes les unités échantillonnées. Pour la méthode d'imputation hot deck pondérée de Rao et Shao (1992) et pour toutes les méthodes d'imputation envisagées par Shao, Chen et Chen (1998), l'espérance mathématique est de cette forme.

3.1 Exemple : Imputation par quotient

Cette méthode d'imputation s'applique aux situations où il existe des données auxiliaires $\{x_{hij}\}$ pour toutes les unités d'échantillonnage. L'imputation par quotient consiste à imputer une valeur pour une réponse manquante $y_{h^\circ i^\circ j^\circ}$ conformément à

$$x_{h^\circ i^\circ j^\circ} \sum_{(h'i'j')\in A_k} w_{h'i'j'}\, y_{h'i'j'} \Big/ \sum_{(h'i'j')\in A_k} w_{h'i'j'}\, x_{h'i'j'}.$$

Par conséquent

$$a_{h'i'j';h^\circ i^\circ j^\circ} = x_{h^\circ i^\circ j^\circ}\, w_{h'i'j'} \Big/ \sum_{(h''i''j'')\in A_k} w_{h''i''j''}\, x_{h''i''j''}$$

et

$$a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ} = x_{h^\circ i^\circ j^\circ}\, w^{(r)}_{h'i'j'} \Big/ \sum_{(h''i''j'')\in A_k} w^{(r)}_{h''i''j''}\, x_{h''i''j''}.$$

Notons que les $a_{h'i'j';h^\circ i^\circ j^\circ}$ et $a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ}$ dépendent des $\{x_{hij}\}$.

3.2 Exemple : Imputation hot deck pondérée

Cette méthode d'imputation consiste à remplacer une réponse manquante par une valeur sélectionnée au hasard parmi les réponses fournies pour la question par les autres répondants avec probabilité proportionnelle au poids des répondants dans la classe d'imputation. Consulter la section 5 pour une discussion approfondie de cette méthode. Shao, Chen et Chen (1998, page 822) montrent que

$$E_{A_k}(\tilde{y}_{h^\circ i^\circ j^\circ}) = \sum_{(h'i'j')\in A_k} w_{h'i'j'}\, y_{h'i'j'} \Big/ \sum_{(h'i'j')\in A_k} w_{h'i'j'}$$

et

$$E^{(r)}_{A_k}(\tilde{y}_{h^\circ i^\circ j^\circ}) = \sum_{(h'i'j')\in A_k} w^{(r)}_{h'i'j'}\, y_{h'i'j'} \Big/ \sum_{(h'i'j')\in A_k} w^{(r)}_{h'i'j'}.$$

Donc,

$$a_{h'i'j';h^\circ i^\circ j^\circ} = w_{h'i'j'} \Big/ \sum_{(h''i''j'')\in A_k} w_{h''i''j''}$$

et

$$a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ} = w^{(r)}_{h'i'j'} \Big/ \sum_{(h''i''j'')\in A_k} w^{(r)}_{h''i''j''}.$$

4. FICHIER DE DONNÉES POUR L'ESTIMATION DE LA VARIANCE

Par souci de simplicité, nous supposons que chaque enregistrement contient un identificateur précisant à quelle classe d'imputation l'unité appartient. Souvent, la classe d'imputation est déterminée par plusieurs variables de l'enregistrement. Un enregistrement ressemblera à :

$$ID \quad IC \quad w_{hij}\, w^{(1)}_{hij} \cdots w^{(R)}_{hij} \quad \tilde{y}_{hij} \quad IF_y \quad \tilde{z}_{hij} \quad IF_z$$

où ID est l'identificateur de l'unité, IC est l'identificateur de la classe d'imputation, w_{hij} est le poids d'échantillonnage (pour l'échantillon complet), $w^{(1)}_{hij} \cdots w^{(R)}_{hij}$ sont les poids de rééchantillonnage, $\tilde{y}_{hij}$ est la valeur (éventuellement imputée) de la variable y étudiée, IF_y est le « signal » d'imputation qui indique si la valeur de $\tilde{y}_{hij}$ est imputée, $\tilde{z}_{hij}$ est la valeur (éventuellement imputée) d'une autre variable z et IF_z est le « signal » d'imputation qui indique si la valeur de $\tilde{z}_{hij}$ est imputée. Naturellement, il pourrait exister d'autres variables dans le fichier, comme une variable auxiliaire x_{hij} dont la valeur est disponible pour toutes les unités d'échantillonnage.

Nous proposons d'ajouter des enregistrements, appelés enregistrements supplémentaires, pour faciliter l'estimation de la variance. Pour chaque unité non répondante $(h^\circ i^\circ j^\circ)$ et chaque unité répondante $(h'i'j')$ à la question y dans la classe d'imputation k, nous créons l'enregistrement

$$ID \quad IC \quad 0 \quad \tilde{w}^{(1)}_{h^\circ i^\circ j^\circ;h'i'j'} \cdots \tilde{w}^{(R)}_{h^\circ i^\circ j^\circ;h'i'j'} \quad y_{h'i'j'} \quad IF_y \quad 0 \quad IF_z$$

où $IC = k$, ID est l'identificateur de l'unité $(h^\circ i^\circ j^\circ)$ qui n'a pas répondu à la question y et

$$\tilde{w}^{(r)}_{h^\circ i^\circ j^\circ;h'i'j'} = (a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ} - a_{h'i'j';h^\circ i^\circ j^\circ})\, w^{(r)}_{h^\circ i^\circ j^\circ}, \qquad r = 1, \ldots, R. \quad (4.1)$$

Notons que le poids pour l'échantillon complet est 0 dans les enregistrements supplémentaires, si bien que ces derniers n'influent pas sur les estimations calculées pour l'ensemble de l'échantillon. Cependant, les estimations calculées d'après les répétitions concordent avec celles définies par (3.1). Notons aussi que les poids $\tilde{w}^{(r)}_{h^\circ i^\circ j^\circ;h'i'j'}$ peuvent être négatifs.

Tableau 1
Illustration numérique : Partie du fichier de données utilisé pour l'estimation de la variance

ID	IC	w_{hij}	$w_{hij}^{(1)}$	...	$w_{hij}^{(R)}$	$\tilde{y}_{hij}$	IF_y	$\tilde{z}_{hij}$	IF_z
001	1	10,1	20,2000	...	0,0000	5,4	1	1,2	1
002	1	20,3	40,6000	...	0,0000	5,1	0	1,3	0
003	1	18,4	36,8000	...	0,0000	5,2	0	1,3	0
004	1	11,1	0,0000	...	22,2000	5,1	1	1,2	0
005	1	16,3	0,0000	...	32,6000	5,1	1	1,4	0
006	1	15,4	0,0000	...	30,8000	5,4	0	1,4	0
001	1	0,0	3,0162	...	0,0000	5,1	2	0,0	3
001	1	0,0	2,7339	...	0,0000	5,2	2	0,0	3
001	1	0,0	-5,7501	...	0,0000	5,4	2	0,0	3
004	1	0,0	0,0000	...	-8,3301	5,1	2	0,0	3
004	1	0,0	0,0000	...	-7,5505	5,2	2	0,0	3
004	1	0,0	0,0000	...	15,8806	5,4	2	0,0	3
005	1	0,0	0,0000	...	-12,2325	5,1	2	0,0	3
005	1	0,0	0,0000	...	-11,0876	5,2	2	0,0	3
005	1	0,0	0,0000	...	23,3201	5,4	2	0,0	3
001	1	0,0	5,5645	...	0,0000	0,0	3	1,3	2
001	1	0,0	5,0436	...	0,0000	0,0	3	1,3	2
001	1	0,0	-2,7512	...	0,0000	0,0	3	1,2	2
001	1	0,0	-4,0400	...	0,0000	0,0	3	1,4	2
001	1	0,0	-3,8169	...	0,0000	0,0	3	1,4	2

Dans l'illustration numérique du tableau 1, les neuf enregistrements (lignes du tableau) pour lesquels $IF_y = 2$ sont les enregistrements supplémentaires pour la question y. Les six premiers enregistrements sont les enregistrements originaux pour les six unités d'échantillonnage qui représentent la classe d'imputation $IC = 1$. (Les enregistrements figurant à la fin pour lesquels $IF_z = 2$, qui sont les enregistrements supplémentaires pour la question z, seront examinés à la fin du paragraphe. Dans ces enregistrements, le signal d'imputation pour y, IF_y, a été fixé à 3 pour indiquer qu'ils correspondent à des enregistrements supplémentaires pour une autre question que y.) Il existe trois répondants ($IF_y = 0$) et trois non-répondants ($IF_y = 1$) à la question y. Nous supposons que la méthode d'imputation est la méthode hot deck pondérée. Nous présentons uniquement les premier et dernier poids de rééchantillonnage ($w_{hij}^{(1)}$ et $w_{hij}^{(R)}$), mais ils concordent avec les poids de rééchantillonnage utilisés pour la méthode par rééchantillonnage équilibré d'estimation de la variance. Nous avons $\sum w_{hij}\ \tilde{y}_{hij} = 476{,}650$, $\sum w_{hij}^{(1)}\ \tilde{y}_{hij} = 506{,}048$ et $\sum w_{hij}^{(R)}\ \tilde{y}_{hij} = 455{,}696$, où les sommes sont calculées sur l'ensemble des enregistrements. Le lecteur peut vérifier que ces résultats concordent avec $\sum w_{hij}\ \tilde{y}_{hij} = 476{,}650$, $\sum w_{hij}^{(1)}\ \tilde{y}_{hij}^{(1)} = 506{,}048$ et $\sum w_{hij}^{(R)}\ \tilde{y}_{hij}^{(R)} = 455{,}696$ obtenus au moyen de (3.1), où les sommes sont calculées sur les six premiers enregistrements uniquement.

Considérons maintenant la question z. Les enregistrements supplémentaires pour cette question prennent la forme

$$ID\ IC\ 0\ \tilde{w}_{h^\circ i^\circ j^\circ;h'i'j'}^{(1)} \cdots \tilde{w}_{h^\circ i^\circ j^\circ;h'i'j'}^{(R)}\ 0\ IF_y\ z_{h'i'j'}\ IF_z$$

où les poids $\tilde{w}_{h^\circ i^\circ j^\circ;h'i'j'}^{(1)} \cdots \tilde{w}_{h^\circ i^\circ j^\circ;h'i'j'}^{(R)}$ sont calculés selon (4.1), mais en utilisant la méthode d'imputation et le profil de réponse correspondant à la question z. La méthode d'imputation pour z ne doit pas nécessairement être la même que celle utilisée pour y, mais doit avoir la forme discutée à la section 3. Dans le tableau 1, les enregistrements supplémentaires pour la question z sont ceux pour lesquels $IF_z = 2$. Nous avons alors $\sum w_{hij}\ \tilde{z}_{hij} = 120{,}30$, $\sum w_{hij}^{(1)}\ \tilde{z}_{hij} = 124{,}349$ et $\sum w_{hij}^{(R)}\ \tilde{z}_{hij} = 115{,}400$, où les sommes sont calculées sur l'ensemble des enregistrements. Ces résultats concordent avec les sommes obtenues selon l'équation (3.1).

Manifestement, le plus gros inconvénient de cette méthode tient au grand nombre d'enregistrements supplémentaires qu'il faut ajouter au fichier. Cet inconvénient est moins prononcé lorsque les classes d'imputation sont petites. (Cependant, la taille des classes d'imputation dépend de nombreux facteurs.) Par contre, les avantages sont les suivants :

- Les estimations répétées et les estimations de la variance fondées sur les rééchantillonnages corrigés peuvent être calculées au moyen de n'importe quel logiciel conçu pour estimer la variance en se fondant sur les poids de rééchantillonnage.
- S'il existe une autre variable, disons y', présentant le même profil de non-réponse que y et pour laquelle on utilise exactement la même méthode d'imputation que pour y (autrement dit, les mêmes valeurs de a et $a^{(r)}$), le calcul des estimations répétées pour y' peut être réalisé sans ajouter de nouveaux enregistrements.
- Des estimations peuvent être calculées sur des sous-domaines, même s'ils recoupent les limites des classes d'imputation.
- Si l'on suppose que la méthode d'imputation est la méthode hot deck pondérée, on estime la variance d'une variable dérivée, disons log y où $y>0$, en ajoutant simplement la variable dérivée à chaque enregistrement et en se fondant sur cette variable pour calculer les estimations répétées. (Nous ferons d'autres commentaires sur la méthode hot deck pondérée à la section suivante.)

L'analyste des données peut choisir de supprimer les enregistrements supplémentaires d'une copie du fichier de données et utiliser le fichier réduit pour repérer les valeurs aberrantes, formuler des hypothèses, *etc.*, puis réintégrer les enregistrements supplémentaires dans le fichier au moment d'estimer les variances.

Soulignons que Rao et Shao (1992) n'ont proposé et évalué leur méthode d'estimation jackknife de la variance que pour les estimations des totaux (ou des moyennes). Il convient donc de se montrer prudent lors de l'utilisation de la méthode pour des statistiques plus complexes. Pareillement, Shao, Chen et Chen (1998) ont proposé leur méthode d'estimation de la variance par rééchantillonnage équilibré pour des fonctions des totaux et pour des quantiles, si bien qu'il ne faut pas l'appliquer à d'autres statistiques.

5. MÉTHODE HOT DECK PONDÉRÉE

Le recours à l'imputation par la méthode hot deck pondérée (par exemple Cox 1980) présente plusieurs avantages, si bien que nous y consacrons une section distincte. Rao et Shao (1992) se concentrent sur cette méthode d'imputation qui est également discutée dans Shao, Chen et Chen (1998). En vertu de cette méthode, pour remplacer une réponse manquante, on impute une valeur sélectionnée au hasard à partir des réponses à la question étudiée qui figurent dans la classe d'imputation. La probabilité de sélection est proportionnelle à $w_{h'i'j'}$, c'est-à-dire le poids appliqué au répondant. Les répondants pour lesquels la probabilité d'être sélectionné est positive sont appelés des *donneurs potentiels*; le non-répondant qui est visé par l'imputation est le receveur. S'il existe dans le fichier plus d'une question pour lesquelles on imputera une valeur par la méthode hot deck pondérée, la situation est simplifiée si l'on utilise comme donneurs potentiels des répondants ayant fourni une réponse complète (unités qui ont répondu à toutes les questions) et que l'on n'utilise qu'un seul donneur pour imputer les valeurs pour toutes les questions d'un receveur particulier nécessitant une imputation par la méthode hot deck pondérée. (Le donneur est sélectionné pour chaque unité d'échantillonnage pour laquelle il y a non-réponse partielle pour *toute* question.)

Si la probabilité de répondre à une question est la même pour chaque unité d'une classe d'imputation, la méthode hot deck pondérée produit des estimations des moyennes, des totaux et des quantiles d'échantillon conformes au plan de sondage. De surcroît, les valeurs imputées sont « plausibles » en ce sens qu'elles ont l'air de données réelles.

Une caractéristique intéressante de la méthode hot deck pondérée est son équivariance en cas de transformations bijectives. Pour expliquer l'équivariance, considérons la variable dérivée d, où $d = g(y)$ et g est une fonction bijective. Alors, par la méthode hot deck pondérée, nous imputons la valeur de y pour l'unité $(h^\circ i^\circ j^\circ)$ qui n'a pas répondu à la question au moyen de $\tilde{y}_{h^\circ i^\circ j^\circ}$ et nous utilisons $g(\tilde{y}_{h^\circ i^\circ j^\circ})$ pour d. Ceci revient à utiliser la méthode hot deck pondérée pour imputer la valeur de d au moyen de $\tilde{d}_{h^\circ i^\circ j^\circ}$ et à utiliser $g^{-1}(\tilde{d}_{h^\circ i^\circ j^\circ})$ pour y. Peu d'autres méthodes présentent cette caractéristique de l'imputation hot deck. Par exemple, en cas d'*imputation par la moyenne* (où la valeur imputée est la moyenne des valeurs des réponses comprises dans la classe d'imputation), g doit être linéaire pour que la propriété d'équivariance soit vérifiée. La pertinence de cette propriété par rapport à l'estimation de la variance par rééchantillonnage corrigé tient au fait que, en cas d'imputation par la méthode hot deck, l'analyste des données peut ajouter $d = g(y)$ au fichier et estimer les variances pour d ainsi que pour y.

Supposons que l'on applique la méthode hot deck pondérée à plusieurs variables d'un fichier et que l'on ne se serve que de répondants ayant fourni des réponses complètes comme donneurs potentiels. Dans ce cas, même si le profil de non-réponse diffère selon la variable à laquelle il faut imputer des valeurs, l'application du rééchantillonnage corrigé au moyen des poids de rééchantillonnage décrite à la section précédente peut se faire en utilisant le même ensemble de poids de rééchantillonnage supplémentaires

$$\tilde{w}^{(r)}_{h^\circ i^\circ j^\circ;h'i'j'} = (a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ} - a_{h'i'j';h^\circ i^\circ j^\circ})\, w^{(r)}_{h^\circ i^\circ j^\circ}$$

pour toutes les variables.

6. AUTRES MÉTHODES

À la présente section, nous considérons d'autres méthodes, y compris une qui nécessite la modification du logiciel.

6.1 Première méthode de rechange

Un moyen de réduire le nombre d'enregistrements consiste à inclure des enregistrements supplémentaires de la forme

$$ID' \quad IC \quad 0 \quad \ddot{w}^{(1)}_{h'i'j'} \cdots \ddot{w}^{(R)}_{h'i'j'} \quad y_{h'i'j'} \quad IF_y \quad 0 \quad IF_z$$

où ID' est l'identificateur de l'unité donneuse potentielle $(h'i'j')$ qui a répondu à la question y, B_k est l'ensemble des unités non répondantes à la question y dans la classe d'imputation k et

$$\ddot{w}^{(r)}_{h'i'j'} = \sum_{(h^\circ i^\circ j^\circ)\in B_k} (a^{(r)}_{h'i'j';h^\circ i^\circ j^\circ} - a_{h'i'j';h^\circ i^\circ j^\circ})\, w^{(r)}_{h^\circ i^\circ j^\circ},$$

$$r = 1, \ldots, R.$$

Dans ces conditions, pour une question donnée, il n'existe qu'un seul enregistrement supplémentaire par donneur potentiel. Le principal inconvénient est que, à cause de la sommation, il est impossible de calculer les estimations pour des sous-domaines qui recoupent plusieurs classes d'imputation.

6.2 Deuxième méthode de rechange

L'application la plus évidente consisterait peut-être à ajouter les $\tilde{y}^{(r)}_{hij}$ à l'enregistrement (hij) et à modifier le

logiciel afin d'utiliser les $\tilde{y}_{hij}^{(r)}$ plutôt que $\tilde{y}_{hij}$ lors du calcul des estimations répétées. Les principaux inconvénients tiennent au fait que 1) il faudrait procéder à une reprogrammation complexe du logiciel, 2) s'il faut imputer les valeurs de plusieurs variables, le nombre de champs de données nécessaires augmente considérablement et 3) la méthode selon laquelle un analyste des données estimerait la variance d'une variable dérivée, disons d, n'est pas certaine, à moins que les $\tilde{d}_{hij}^{(r)}$ figurent d'avance dans le fichier. Les caractéristiques avantageuses de cette mise en oeuvre sont que 1) aucun enregistrement supplémentaire n'est nécessaire et 2) l'estimation de la variance pour des sous-domaines ne nécessite pas de travaux supplémentaires.

7. CONCLUSIONS

Les méthodes d'estimation de la variance par rééchantillonnage corrigé de Rao et Shao (1992) et de Shao, Chen et Chen (1998) offrent un moyen de calculer des estimations de la variance qui tienne compte de l'imputation pour compenser la non-réponse partielle. Une future étape importante consistera à trouver des moyens de faciliter les calculs. Le présent article décrit les applications fondées sur l'utilisation de poids de rééchantillonnage.

REMERCIEMENTS

Une question posée par Robert E. Fay lors d'une présentation à la Washington Statistical Society a fourni l'idée à l'origine du présent article. L'auteur remercie également les deux évaluateurs, le rédacteur en chef adjoint et le rédacteur en chef de leurs commentaires utiles. L'auteur travaillait pour le National Center for Education Statistics au moment de la rédaction de la version initiale de l'article. L'article expose les opinions personnelles de l'auteur et ne sous-entend pas que ces opinions bénéficient de l'appui officiel du U.S. Department of Education ou du U.S. Department of Transportation ni ne permet d'inférer un tel soutien.

BIBLIOGRAPHIE

BRICK, J.M., et MORGANSTEIN, D. (1996). WesVarPC: Software for computing variance estimates from complex designs. *Proceedings of the 1996 Annual Research Conference*, U.S. Bureau of the Census, 861-866.

BRICK, J.M., et MORGANSTEIN, D. (1997). Computing sampling errors from clustered unequally weighted data using replication: WesVarPC. *Bulletin of the International Statistical Institute, Proceedings*, 1, 479-482.

COX, B.G. (1980). The weighted sequential hot deck imputation procedure. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 721-726.

DIPPO, C.S., FAY, R.E. et MORGANSTEIN, D.H. (1984). Computing variances from complex samples with replicate weights. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 489-494.

KOTT, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

RAO, J.N.K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

RAO, J.N.K., et SHAO, J. (1996). On balanced half sample variance estimation in stratified sampling. *Journal of the American Statistical Society*, 91, 343-348.

RAO, J.N.K., et SHAO, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

RUST, K., et RAO, J.N.K. (1996). Variance estimation for complex estimators in sample surveys. *Statistics in Medical Research*, 5, 381-397.

SHAO, J. (1996). Resampling methods in sample surveys (avec discussion). *Statistics*, 27, 203-254.

SHAO, J., et CHEN, Y. (1999). Approximate balanced half samples and related replication methods for imputed survey data. *Sankhyā*, B, Special Issue on Sample Surveys, 187-201.

SHAO, J., CHEN, Y. et CHEN, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Society*, 93, 819-831.

SHAO, J., et STEEL, P. (1999). Variance estimation for imputed survey data with non-negligible sampling fractions. *Journal of the American Statistical Society*, 94, 254-265.

SHAO, J., et TU, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

VALLIANT, R. (1996). Limitations of balanced half-sampling. *Journal of Official Statistics*, 12, 225-240.

WOLTER, K.M. (1985). *Introduction to Variance Estimation* New York: Springer-Verlag.

Techniques d'enquête, juin 2002
Vol. 28, n° 1, pp. 109-122
Statistique Canada

Estimation de la variance de l'estimateur de régression généralisée

RICHARD VALLIANT[1]

RÉSUMÉ

On trouve dans la littérature sur l'échantillonnage diverses propositions d'estimateurs de la variance de l'estimateur de régression généralisée (GREG) d'une moyenne, principalement en vue d'estimer la variance due au plan de sondage. Il est facile de concevoir des estimateurs de la variance qui, dans certaines conditions, sont approximativement non biaisés en ce qui concerne le plan de sondage et le modèle. Nous étudions ici plusieurs estimateurs bivalents dans le cas de l'échantillonnage à un degré. Il s'agit d'estimateurs robustes de la variance due au modèle, même si le modèle qui motive l'estimateur de régression généralisée (GREG) comprend un paramètre de variance incorrect.

L'une des caractéristiques principales des estimateurs robustes est le rajustement des carrés des résidus au moyen de facteurs analogues aux effets leviers utilisés en analyse par régression classique. Nous montrons aussi que l'estimateur jackknife avec suppression d'une unité inclut les ajustements pour tenir compte des effets de leviers et est un bon choix du point de vue tant de la variance due au plan de sondage que de celle due au modèle. Dans un ensemble de simulations, ces estimateurs de la variance sont caractérisés par un biais faible et produisent des intervalles de confiance dont le taux de couverture est quasi nominal pour plusieurs méthodes d'échantillonnage, tailles d'échantillon et populations dans le cas de l'échantillonnage à un degré.

Nous présentons aussi les résultats des simulations pour une population à distribution asymétrique où tous les estimateurs de la variance donnent de mauvais résultats. Les échantillons qui ne représentent pas adéquatement les unités dont la valeur est grande produisent des estimations de la moyenne trop faibles, des estimations de la variance trop faibles et des intervalles de confiance dont la couverture est nettement inférieure au taux nominal. Ces faiblesses doivent être évitées à l'étape de l'élaboration du plan de sondage grâce à la sélection d'échantillons qui couvrent bien les unités extrêmes. Cependant, ceci n'est pas faisable dans le cas de populations pour lesquelles les renseignements sur le plan de sondage sont insuffisants.

MOTS CLÉS : Couverture de l'intervalle de confiance; matrice chapeau, jackknife; effet levier; modèle non biaisé; asymétrie.

1. INTRODUCTION

L'estimation robuste de la variance est un élément important dont il faut tenir compte en cas d'échantillonnage d'une population finie où on utilise l'approche prédictive. Valliant, Dorfman et Royall (2000) résument une grande partie des études publiées dans le domaine de la modélisation. Selon l'approche prédictive, on formule un modèle de travail que l'on utilise ensuite pour construire un estimateur ponctuel d'une moyenne ou d'un total. On crée des estimateurs de la variance qui sont robustes en ce sens qu'ils sont approximativement sans biais par rapport au modèle et convergents pour la variance due au modèle, même si la spécification de la variance dans le modèle de travail est incorrecte. Dans le présent article, nous étendons cette approche à l'estimateur de régression généralisée (GREG) afin de construire des estimateurs de la variance qui sont approximativement non biaisés en ce qui concerne le modèle, mais qui sont aussi approximativement non biaisés en ce qui concerne le plan de sondage dans le cas de l'échantillonnage à un degré. Nous comparons plusieurs solutions, y compris l'estimateur jackknife et certaines variantes de ce dernier. Nous utilisons une classe particulière de modèles linéaires ainsi qu'un échantillonnage de Bernoulli ou de Poisson comme fondement des estimateurs de la variance. Cependant, en pratique, certains de ces estimateurs peuvent être appliqués avec de bons résultats à des plans d'échantillonnage à un degré où les tirages ne sont pas indépendants.

À chaque unité de la population sont associés une variable cible Y_i et un vecteur p de variables auxiliaires $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$ où $i = 1, ..., N$. Le vecteur des totaux de population des variables auxiliaires est $\mathbf{T}_x = (T_{x1}, ..., T_{xp})'$ où $T_{xk} = \sum_{i=1}^{N} x_{ki}, k = 1, ..., p$. L'estimateur de régression généralisée, défini plus bas, est motivé par un modèle linéaire dans lequel les Y sont des variables aléatoires indépendantes telles que

$$E_M(Y_i) = \mathbf{x}_i' \beta$$

$$\text{var}_M(Y_i) = v_i. \qquad (1.1)$$

Dans la plupart des situations, (1.1) est un modèle « de travail » vraisemblablement incorrect dans une certaine mesure.

Supposons que l'on sélectionne un échantillon probabiliste s et que la probabilité de sélection de l'unité d'échantillonnage i soit $P(\delta_i = 1) = \pi_i$ où δ_i est un indicateur 0-1 de la présence ou non de l'unité dans l'échantillon. Nous supposons que le mécanisme de sélection de l'échantillon est ignorable. Approximativement parlant, un mécanisme ignorable signifie que la distribution conjointe des Y et des indicateurs d'échantillonnage, étant donné les $\mathbf{x}$, peut être décomposée en un produit de la distribution de Y étant

[1] Richard Valliant, Westat, 1650 Research Boulevard, Rockville, MD 20850.

donné $\mathbf{x}$ par la distribution des indicateurs étant donné $\mathbf{x}$ (voir Sugden et Smith 1984 pour une définition formelle). Dans ce cas, l'inférence fondée sur un modèle peut se faire à l'aide du modèle sans tenir compte du mécanisme de sélection.

Le vecteur des n cibles pour les unités d'échantillonnage est $\mathbf{Y}_s = (Y_1, ..., Y_n)'$ et la matrice $n \times p$ des variables auxiliaires pour les unités d'échantillonnage est $\mathbf{X}_s = (\mathbf{x}_1, ..., \mathbf{x}_n)'$. Définissons la matrice diagonale des probabilités de sélection comme étant $\mathbf{\Pi}_s = \text{diag}(\pi_i)$, $i \in s$, et la matrice diagonale des variances dues au modèle comme étant $\mathbf{V}_s = \text{diag}(v_i)$. L'estimateur GREG du total, $T = \sum_{i=1}^{N} Y_i$, est alors défini comme étant l'estimateur d'Horvitz-Thompson ou l'estimateur π, $\hat{T}_\pi = \sum_s Y_i/\pi_i$, plus un ajustement :

$$\hat{T}_G = \hat{T}_\pi + \hat{\mathbf{B}}'(\mathbf{T}_x - \hat{\mathbf{T}}_x) \tag{1.2}$$

où $\hat{\mathbf{B}} = \mathbf{A}_{\pi s}^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{\Pi}_s^{-1}\mathbf{Y}_s$ avec $\mathbf{A}_{\pi s} = \mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{\Pi}_s^{-1}\mathbf{X}_s$, et $\hat{\mathbf{T}}_x = \sum_s \mathbf{x}_i/\pi_i$. Nous pouvons aussi écrire l'estimateur GREG sous la forme

$$\hat{T}_G = \mathbf{g}_s'\mathbf{\Pi}_s^{-1}\mathbf{Y}_s \tag{1.3}$$

où $\mathbf{g}_s = \mathbf{V}_s^{-1}\mathbf{X}_s\mathbf{A}_{\pi s}^{-1}(\mathbf{T}_x - \hat{\mathbf{T}}_x) + \mathbf{1}_s$ et $\mathbf{1}_s$ est un vecteur de n valeurs 1. L'expression (1.3) sera utile pour les calculs subséquents.

Une variante de l'estimateur GREG, appelée estimateur « esthétique », a été proposée par Särndal et Wright (1984) et amplifiée par Brewer (1995, 1999). Un estimateur esthétique peut aussi être interprété comme étant fondé sur le plan de sondage et fondé sur un modèle. Les estimateurs de la variance présentés ici pourraient également être adaptés de façon à couvrir l'estimation esthétique.

À condition que l'on connaisse N, l'estimateur GREG de la moyenne est simplement $\hat{\bar{Y}}_G = \hat{T}_G/N$. Nous nous concentrerons ici sur l'analyse de $\hat{\bar{Y}}_G$. (Dans certaines situations, particulièrement celles où l'on recourt à l'échantillonnage à plusieurs degrés, on ne connaît pas la taille de la population et il faut utiliser une estimation, $\hat{N}$, dans le dénominateur de $\hat{\bar{Y}}_G$. L'analyse qui suit concernant la moyenne ne s'applique pas dans ce cas.) Nous pouvons utiliser dans l'estimateur GREG des variables auxiliaires quantitatives ou qualitatives (ou les deux). Si nous utilisons une variable qualitative, comme le sexe (masculin ou féminin), alors deux colonnes ou plus de la matrice $\mathbf{X}_s$ seront linéairement dépendantes, auquel cas nous utiliserons dans (1.2) et (1.3) une inverse généralisée, représentée par $\mathbf{A}_{\pi s}^{-}$. Notons que, même si $\mathbf{A}_{\pi s}^{-}$ n'est pas unique, l'estimateur GREG ne varie pas en fonction du choix de l'inverse généralisée. La preuve est similaire au théorème 7.4.1 dans Valliant et coll. (2000).

L'estimateur GREG est non biaisé en ce qui concerne le modèle dans les conditions de (1.1) et est approximativement non biaisé en ce qui concerne le plan de sondage dans le cas de grands échantillons probabilistes. Notons que l'absence de biais dû au modèle nécessite uniquement que $E_M(Y_i) = \mathbf{x}_i'\beta$; si les paramètres de variance sont mal précisés dans (1.1), l'estimateur GREG demeurera non biaisé en ce qui concerne le modèle. Par contre, si $E_M(Y_i)$ est spécifiée incorrectement, il comportera un biais de modèle et l'erreur quadratique moyenne du modèle contiendra éventuellement un terme de biais carré important. L'erreur d'estimation de l'estimateur GREG $\hat{\bar{Y}}_G$ est définie comme étant

$$\hat{\bar{Y}}_G - \bar{Y} = N^{-1}(\mathbf{a}_s'\mathbf{Y}_s - \mathbf{1}_r'\mathbf{Y}_r)$$

où $\bar{Y} = T/N$, $\mathbf{a}_s = \mathbf{\Pi}_s^{-1}\mathbf{g}_s - \mathbf{1}_s$, $\mathbf{Y}_r$ est le vecteur $(N-n)$ des variables cibles pour les unités non échantillonnées, et $\mathbf{1}_r$ est un vecteur de $N-n$ valeurs 1. Supposons maintenant que le modèle réel de Y_i est

$$E_M(Y_i) = \mathbf{x}_i'\beta$$
$$\text{var}_M(Y_i) = \psi_i, \tag{1.4}$$

autrement dit que la spécification de la variance est différente de celle donnée dans (1.1), mais que $E_M(Y_i)$ est la même. Si nous utilisons l'erreur d'estimation, la variance de l'erreur de $\hat{\bar{Y}}_G$ est alors

$$\text{var}_M(\hat{\bar{Y}}_G - \bar{Y}) = N^{-2}(\mathbf{a}_s'\mathbf{\Psi}_s\mathbf{a}_s + \mathbf{1}_r'\mathbf{\Psi}_r\mathbf{1}_r)$$

où la matrice des covariances $n \times n$ de $\mathbf{Y}_s$ est $\mathbf{\Psi}_s = \text{diag}(\mathbf{\Psi}_i)$ et $\mathbf{\Psi}_r$ est la matrice des covariances $(N-n) \times (N-n)$ de $\mathbf{Y}_r$. Si les tailles d'échantillon et de population sont toutes les deux importantes et que la fraction d'échantillonnage, $f = n/N$, est négligeable, la variance de l'erreur est approximativement

$$\text{var}_M(\hat{\bar{Y}} - \bar{Y}) \cong N^{-2}\sum_{i \in s} a_i^2\psi_i. \tag{1.5}$$

Notons que cette variance dépend des paramètres de la variance réelle, Ψ_i, et des paramètres de la variance du modèle de travail, v_i, car v_i fait partie de a_i. Puisque $\mathbf{a}_s$ est approximativement identique à $\Pi_s^{-1}\mathbf{g}_s$ lorsque les probabilités de sélection sont faibles, la variance de l'erreur est, dans ce cas, aussi approximativement égale à

$$\text{var}_M(\hat{\bar{Y}}_G - \bar{Y}) \cong N^{-2}\sum_{i \in s}\frac{g_i^2}{\pi_i^2}\Psi_i. \tag{1.6}$$

En ce qui concerne l'estimation de la variance due au modèle, nous prendrons comme cible l'une ou l'autre des formes asymptotiques (1.5) et (1.6). Cependant, si la fraction d'échantillonnage est grande, le terme $\mathbf{1}_r'\mathbf{\Psi}_r\mathbf{1}_r/N^2$ peut représenter une part importante de la variance de l'erreur et (1.5) ou (1.6) pourrait constituer une mauvaise approximation.

Nous considérerons la variance due au plan de sondage dans le cas des plans d'échantillonnage à un degré de Bernoulli et de Poisson. Dans le cas de l'échantillonnage de Poisson, les indicateurs δ_i de la présence ou non d'une unité dans l'échantillon sont indépendants et $P(\delta_i = 1) = 1 - P(\delta_i = 0) = \pi_i$ (voir Särndal, Swensson et Wretman 1992,

section 3.5, pour une description plus détaillée). L'échantillonnage de Bernoulli est un cas particulier de l'échantillonnage de Poisson où la probabilité d'inclusion est la même pour toutes les unités. En vertu de ces plans, la variance approximative due au plan de sondage de $\hat{\bar{Y}}_G$ est

$$\mathrm{var}_\pi(\hat{\bar{Y}}_G) \approx N^{-2}\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}E_i^2 \qquad (1.7)$$

où $E_i = Y_i - \mathbf{x}_i'\mathbf{B}$ et $\mathbf{B} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$ est l'estimateur des paramètres de régression évalué pour la population finie complète. Särndal (1996) recommande d'utiliser l'estimateur GREG conjugué à des plans d'échantillonnage pour lesquels (1.7) est valide étant donné que la variance (1.7) est simple et que l'utilisation de l'estimation par régression peut souvent plus que compenser l'effet des tailles aléatoires d'échantillon qui sont une conséquence de ce genre de plan d'échantillonnage.

Les plans d'échantillonnage de Bernoulli et de Poisson, ainsi que les modèles linéaires (1.1) et (1.4) servent principalement de motif à l'utilisation des estimateurs de la variance présentés aux sections 2 et 3. Comme l'ont fait remarquer Yung et Rao (1996, page 24), il est courant d'utiliser des estimateurs de la variance qui sont appropriés pour un plan de sondage à tirages indépendants ou pour un plan de sondage avec remise, même si l'échantillon a été sélectionné sans remise. Pareillement, les estimateurs de la variance motivés par un modèle linéaire sont souvent appliqués à des cas où l'on s'attend à des divergences par rapport au modèle. Cette démarche pratique, qui sous-tend la logique suivie dans le présent article, est illustrée grâce à l'étude de simulation présentée à la section 4.

2. ESTIMATEURS DE LA VARIANCE

Dans le contexte de l'estimation de la variance, notre objectif général consistera à trouver des estimateurs convergents et approximativement non biaisés autant par rapport à un modèle que par rapport à un plan de sondage. Kott (1990) envisage aussi ce problème. Notons que l'objectif ici n'est pas d'estimer la variance combinée (ou anticipée) modèle-plan de sondage,

$$E_M E_\pi\left\{\left[\left(\hat{\bar{Y}}_G - \bar{Y}\right) - E_M E_\pi\left(\hat{\bar{Y}}_G - \bar{Y}\right)\right]^2\right\}.$$

Nous cherchons plutôt des estimateurs utiles à la fois pour $\mathrm{var}_M(\hat{\bar{Y}} - \bar{Y})$ et $\mathrm{var}_\pi(\hat{\bar{Y}})$. Les arguments avancés ici sont en grande partie des arguments heuristiques utilisés comme justification des formes adoptées pour les estimateurs de la variance. En outre, nous devons appliquer des conditions formelles, telles que celles énoncées dans Royall et Cumberland (1978) ou Yung et Rao (2000) afin d'assurer la convergence et l'absence approximative de biais par rapport au modèle et au plan de sondage.

En premier lieu, considérons l'estimation de la variance approximative due au modèle donnée par (1.5). Dans le développement qui suit, nous supposons que, à mesure que les valeurs de N et n deviennent grandes,

i) $N \max_i(\pi_i) = O(n)$ et

ii) $\mathbf{A}_{\pi s}/N$ converge vers une matrice de constantes $\mathbf{A}_o$.

Le résidu associé à l'unité d'échantillonnage i est $r_i = Y_i - \hat{Y}_i$ où $\hat{Y}_i = \mathbf{x}_i'\hat{\mathbf{B}}$. Le vecteur des valeurs prévues pour les unités d'échantillonnage peut s'écrire sous la forme

$$\hat{\mathbf{Y}}_s = \mathbf{H}_s\mathbf{Y}_s \qquad (2.1)$$

où $\mathbf{H}_s = \mathbf{X}_s\mathbf{A}_{\pi s}^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{\Pi}_s^{-1}$. La valeur prévue pour une unité individuelle est $\hat{Y}_i = \sum_{j\in s} h_{ij}Y_j$ où $h_{ij} = \mathbf{x}_i'\mathbf{A}_{\pi s}^{-1}\mathbf{x}_j/(v_j\pi_j)$ est le $(ij)^{\text{ième}}$ élément de $\mathbf{H}_s$. La matrice $\mathbf{H}_s$ est l'analogue de la matrice chapeau habituelle (Belsley, Kuh et Welsch 1980) de l'analyse par régression classique. Les éléments diagonaux de la matrice chapeau portent le nom d'effet levier et représentent une mesure de l'influence qu'exerce une unité sur sa propre valeur prévue. Notons que (2.1) prend en compte les inverses des probabilités de sélection, alors que ces derniers ne joueraient aucun rôle dans une analyse basée purement sur un modèle.

Le lemme qui suit, qui est une variante de certains résultats du Lemme 5.3.1 de Valliant et coll. (2000), donne certaines propriétés des effets leviers et de la matrice chapeau.

Lemme 1. Supposons que (i) et (ii) sont vérifiées. Pour $\mathbf{H}_s = \mathbf{X}_s\mathbf{A}_{\pi s}^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{\Pi}_s^{-1}$, les propriétés qui suivent sont vérifiées pour tous les $i \in s$:

a) $h_{ij} = O(n^{-1})$;

b) $\mathbf{H}_s$ est idempotente;

c) $0 \le h_{ii} \le 1$.

Preuve : Puisque $h_{ij} = \mathbf{x}_i'\mathbf{A}_{\pi s}^{-1}\mathbf{x}_j/(v_j\pi_j)$, les conditions (i) et (ii) impliquent que $h_{ij} = O(n^{-1})$. La partie (b) découle de la multiplication directe, en utilisant la définition de $\mathbf{H}_s$. Pour prouver (c), notons que $h_{ii} \ge 0$, puisqu'il s'agit d'une forme quadratique. La partie (b) implique que $h_{ii} = h_{ii}^2 + \sum_{j\ne i} h_{ij}h_{ji}$ qui est vérifiée uniquement si $h_{ii} \le 1$.

Ensuite, nous écrivons le résidus sous la forme $r_i = Y_i(1-h_{ii}) - \sum_{j\in s(i)} h_{ij}Y_j$ où $s(i)$ est l'échantillon dont est exclue l'unité i. Puisque $E_M(r_i) = 0$, nous avons $E_M(r_i^2) = \mathrm{var}_M(r_i)$ et

$$E_M(r_i^2) = \Psi_i(1-h_{ii})^2 + \sum_{j\in s(i)} h_{ij}^2\Psi_j \qquad (2.2)$$

dans les conditions du modèle (1.4). En utilisant le lemme 1(a), nous avons $h_{ii} = o(1)$, $h_{ij}^2 = o(1)$ et, conséquemment, $E_M(r_i^2) \approx \Psi_i$. Donc, dans le cas de grands échantillons, r_i^2 est un estimateur approximativement non biaisé de la variance due au modèle correcte, même si la spécification de la variance dans le modèle (1.1) est

incorrecte. Par conséquent, r_i^2 est un estimateur robuste de la variance due au modèle pour l'unité i indépendamment de la forme de Ψ_i. Un estimateur simple, robuste, de la variance approximative due au modèle (1.5) est alors

$$v_{R1}(\hat{\hat{Y}}_G) = N^{-2}\sum_s a_i^2 r_i^2 \tag{2.3}$$

qui est un genre d'estimateur « sandwich » (voir, par exemple, White 1982). (Notons qu'un argument formel pour montrer que v_{R1} est robuste nécessiterait des conditions telles que $n^{-1}E_M(v_{R1})$ et $n^{-1}N^{-2}\sum_s a_i^2\Psi_i$ convergent vers la même quantité.) Un autre estimateur de la variance, similaire à v_{R1} si $\mathbf{a}_s \cong \Pi_s^{-1}\mathbf{g}_s$, est

$$v_{R2}(\hat{\hat{Y}}_G) = N^{-2}\sum_s \frac{g_i^2}{\pi_i^2} r_i^2. \tag{2.4}$$

Un estimateur de la variance approximative due au plan de sondage figurant dans (1.7) est

$$v_\pi(\hat{\hat{Y}}_G) = N^{-2}\sum_s \frac{1-\pi_i}{\pi_i^2} r_i^2. \tag{2.5}$$

Une autre solution, qui, selon Särndal et coll. (1989), a de meilleures propriétés conditionnelles, est

$$v_{\mathrm{SSW}}(\hat{\hat{Y}}_G) = N^{-2}\sum_s \frac{1-\pi_i}{\pi_i^2} g_i^2 r_i^2. \tag{2.6}$$

Un autre estimateur, comparable, utilisé dans le logiciel SUPERCARP (Hidiroglou, Fuller et Hickman 1980) et obtenu par des méthodes de développement en série de Taylor, est

$$v_T(\hat{\hat{Y}}_G) = N^{-2}\frac{n}{n-1}\sum_s\left(\frac{g_i r_i}{\pi_i} - \frac{1}{n}\sum_s \frac{g_i r_i}{\pi_i}\right)^2. \tag{2.7}$$

Comme nous le montrons à l'annexe, le deuxième terme entre parenthèses dans (2.7) converge en probabilité vers zéro dans les conditions du modèle (1.1). Donc, $v_T \approx v_{R2}$ dans le cas de grands échantillons.

Si la probabilité de sélection de chaque unité est faible, v_{SSW} sera similaire à v_{R1}, v_{R2} et v_T. Les trois estimateurs seront approximativement sans biais par rapport au modèle (1.4) et approximativement sans biais par rapport au plan de sondage dans les conditions de l'échantillonnage de Bernoulli et de Poisson. Par ailleurs, v_π est approximativement sans biais par rapport au plan de sondage, mais ne tient pas compte des coefficients g_i et est biaisé en ce qui concerne tant le modèle (1.1) que le modèle (1.4).

À titre d'exemple simple, considérons l'échantillonnage de Bernoulli où $\pi_i = n/N$ et le modèle de travail $E_M(Y_i) = x_i\beta$, $\mathrm{var}_M(Y_i) = \sigma^2 x_i$. Alors, l'estimateur GREG est l'estimateur par ratio $\hat{\hat{Y}}_G = \bar{Y}_s\bar{x}/\bar{x}_s$ où $\bar{x}$ est une moyenne de population finie. Dans le cas d'une spécification plus générale, la variance approximative par rapport au modèle sous la spécification plus générale, $\mathrm{var}_M(Y_i) = \psi_i$, est $(\bar{\Psi}_s/n)(\bar{x}/\bar{x}_s)^2$ où $\bar{\Psi}_s = \sum_s \Psi_i/n$. La variance approximative par rapport au plan de sondage est $(1-f)/(nN)\sum_{i=1}^N (Y_i - x_i\bar{Y}/\bar{x})^2$, où $\bar{Y}$ est une moyenne de population finie. L'estimateur $v_{R2} = n^{-2}(\bar{x}/\bar{x}_s)^2\sum_s(Y_i - x_i\bar{Y}_s/\bar{x}_s)^2$ est approximativement sans biais pour la variance due au modèle et, parce que $\bar{x}/\bar{x}_s \xrightarrow{P} 1$ dans le cas de grands échantillons de Bernoulli, v_{R2} est également approximativement sans biais pour la variance due au plan de sondage, à condition que f soit faible. En revanche, $v_\pi = n^{-2}(1-f)\sum_s(Y_i - x_i\bar{Y}_s/\bar{x}_s)^2$ est approximativement sans biais par rapport au plan de sondage, mais n'est sans biais par rapport au modèle que pour des échantillons équilibrés où $\bar{x} = \bar{x}_s$. Royall et Cumberland (1981) ont observé des résultats comparables pour l'estimateur par ratio dans le cas de l'échantillonnage aléatoire simple sans remise.

3. AUTRES ESTIMATEURS DE LA VARIANCE BASÉS SUR LES CARRÉS CORRIGÉS DES RÉSIDUS

Le premier estimateur de rechange de la variance que nous considérons est le jackknife. La version particulière que nous étudions ici est définie comme étant

$$v_J = \frac{n-1}{n}\sum_{i=1}^{n}\left[\hat{\hat{Y}}_{G(i)} - \hat{\hat{Y}}_{G(.)}\right]^2 \tag{3.1}$$

où $\hat{\hat{Y}}_{G(i)}$ a la même forme que l'estimateur basé sur l'échantillon complet après omission de l'unité d'échantillonnage i. Si la probabilité de sélection a la forme $\pi_i = np_i$, alors l'équation (3.1) peut être réécrite. Si nous adoptons comme convention de notation d'indiquer au moyen de l'indice (i) que l'unité d'échantillonnage i a été omise, nous avons

$$\hat{\hat{Y}}_{G(i)} = \hat{T}_{G(i)}/N,\ \hat{\hat{Y}}_{G(\cdot)} = \sum_{i\in s}\hat{\hat{Y}}_{G(i)}/n,\ \hat{T}_{G(i)}$$

$$= \hat{T}_{\pi(i)} + \hat{\mathbf{B}}_{(i)}'(\mathbf{T}_x - \hat{\mathbf{T}}_{x(i)}),$$

$$\hat{T}_{\pi(i)} = n\sum_{i\in s(i)} Y_j/[\pi_j(n-1)],\ \hat{T}_{x(i)}$$

$$= n\sum_{j\in s(i)} x_j/[\pi_j(n-1)], \text{ et}$$

$$\hat{B}_{(i)} = A_{\pi s(i)}^{-1}X_{s(i)}'V_{s(i)}^{-1}\Pi_{s(i)}^{-1}Y_{s(i)} \text{ où}$$

$$A_{\pi s(i)} = X_{s(i)}'V_{s(i)}^{-1}\Pi_{s(i)}^{-1}X_{s(i)}$$

Une autre version, plus prudente, mais asymptotiquement équivalente, du jackknife consiste à remplacer $\hat{\hat{Y}}_{G(.)}$ par l'estimateur basé sur l'échantillon complet $\hat{\hat{Y}}_G$. Les propriétés basées sur le plan de sondage de l'estimateur

jackknife défini par (3.1) sont habituellement étudiées sur des échantillons sélectionnés avec remise (voir, par exemple, Krewski et Rao 1981, Rao et Wu 1985, Yung et Rao 1996), mais appliquées en pratique à des plans d'échantillonnage sans remise. Notons que, pour l'estimateur linéaire $\hat{\bar{Y}}_\pi = N^{-1}\sum_{i\in s} Y_i/\pi_i$ dans le cas de l'échantillonnage avec probabilité proportionnelle à la taille sans remise, ni le jackknife, v_J, ni l'approximation de v_J donnée plus loin dans cette section ne se réduit aux estimateurs habituels de la variance d'Horvitz-Thompson ou de Yates-Grundy.

Moyennant un certain effort, nous pouvons écrire le jackknife sous une forme qui englobe les résidus et les effets leviers. La forme réécrite permettra de discerner clairement la relation entre le jackknife et les estimateurs de la variance présentés à la section 2. En premier lieu, notons les égalités suivantes qu'il est facile de vérifier :

$$\hat{T}_{\pi(i)} = \frac{n}{n-1}\left(\hat{T}_{\pi(i)} - \frac{Y_i}{\pi_i}\right), \hat{T}_{x(i)} = \frac{n}{n-1}\left(\hat{\mathbf{T}}_{x(i)} - \frac{\mathbf{x}_i}{\pi_i}\right) \quad (3.2)$$

$$\mathbf{X}'_{s(i)}\mathbf{V}^{-1}_{s(i)}\mathbf{\Pi}^{-1}_{s(i)}\mathbf{Y}_{s(i)} = \mathbf{X}_s\mathbf{V}^{-1}_s\mathbf{\Pi}^{-1}_s\mathbf{Y}_s - \mathbf{x}_i Y_i/v_i\pi_i,$$

$$\mathbf{A}_{\pi s(i)} = \mathbf{A}_{\pi s} - \mathbf{x}_i\mathbf{x}_i'/v_i\pi_i \quad (3.3)$$

Si nous utilisons une formule standard de l'inverse de la somme des deux matrices, l'estimateur de la pente, lorsque l'on omet l'unité d'échantillonnage i, est donné par

$$\mathbf{B}_{(i)} = \hat{\mathbf{B}} + n^{-1}\sum_s \frac{\mathbf{A}^{-1}_{\pi s}\mathbf{x}_i}{1-h_{ii}}\frac{r_i}{v_i\pi_i}.$$

Les détails de cette formule et des calculs qui suivent sont exposés brièvement à l'annexe. Après un nombre considérable d'opérations algébriques, nous obtenons

$$\hat{T}_{G(i)} - \hat{T}_{G(\cdot)} = -\frac{n}{n-1}(D_i - \bar{D}_s) + \frac{n}{n-1}F_i$$

où

$$D_i = \frac{g_i r_i}{\pi_i(1-h_{ii})}$$

et F_i est définie à l'annexe. Le jackknife donné par (3.1) est alors égal à

$$v_J(\hat{\bar{Y}}_G) = N^{-2}\frac{n}{n-1}\times$$

$$\left[\sum_s (D_i - \bar{D}_s)^2 + \sum_s F_i^2 - 2\sum_s F_i(D_i - \bar{D}_s)\right]. \quad (3.4)$$

L'expression (3.4) est une égalité exacte qui pourrait être utilisée comme formule de calcul de l'estimateur jackknife. Ceci éviterait de devoir éliminer mécaniquement une unité, calculer $\hat{\bar{Y}}_{G(i)}$, et ainsi de suite, pour la totalité de l'échantillon.

Dans le cas de grands échantillons, le premier terme entre crochets dans (3.4) domine, tandis que les deuxième et troisième sont presque nuls dans certaines conditions raisonnables. Donc, dans le cas de grands échantillons, on peut approximer le jackknife au moyen de $v_J(\hat{\bar{Y}}_G) \cong N^{-2}\sum_s (D_i - \bar{D}_s)^2$ ou, de façon équivalente,

$$v_J(\hat{\bar{Y}}_G) \cong \frac{1}{N^2}\times$$

$$\sum_s\left[\frac{g_i r_i}{\pi_i(1-h_{ii})}\right]^2 - \frac{1}{N^2 n}\left[\sum_s \frac{g_i r_i}{\pi_i(1-h_{ii})}\right]^2. \quad (3.5)$$

Comme nous le montrons à l'annexe, le deuxième terme de (3.5) converge en probabilité vers zéro dans les conditions du modèle (1.1). Par conséquent, une autre approximation du jackknife prend la forme

$$v_J(\hat{\bar{Y}}_G) \cong \frac{1}{N^2}\sum_s\left[\frac{g_i^2 r_i^2}{\pi_i^2(1-h_{ii})}\right]^2. \quad (3.6)$$

Comme le montrent les expressions (3.5) et (3.6), les coefficients g_i^2 nécessaires pour l'estimation de la variance due au modèle sont intégrés implicitement dans l'estimateur jackknife. Le deuxième membre de (3.6) est lui-même un estimateur de rechange que nous représentons par $v_J^*(\hat{\bar{Y}}_G)$.

Yung et Rao (1996) établissent aussi une approximation du jackknife pour l'estimateur GREG dans le cas de l'échantillonnage à plusieurs degrés. Pour l'échantillonnage à un degré, leur approximation est égale à v_T, défini en (2.7), qui revient au même que (3.5) si les effets leviers sont nuls. Duchesne (2000) a également présenté une formule pour le jackknife, nommée $\hat{V}_{JK2}$, qui tient compte des effets leviers de l'échantillon. L'avantage de l'expression (3.4) tient au fait qu'elle indique clairement quelles parties de l'estimateur sont négligeables dans le cas de grands échantillons. Duchesne a également présenté un estimateur, nommé $\hat{V}^*_{JK2}$, qui est essentiellement le même que v_{R2} et est une approximation du jackknife.

Les expressions (3.5) et (3.6) montrent explicitement comment les effets leviers influent sur la grandeur de la variance jackknife. Les effets leviers pondérés, h_{ii}, qui ne sont pas quasiment nuls gonfleront la valeur de v_J. Selon la configuration des x, cet effet pourrait être important dans le cas de certains échantillons.

Puisque h_{ii} tend vers zéro lorsque la taille de l'échantillon augmente, v_J, v_{R2}, v_{SSW} et v_T ont les mêmes propriétés asymptotiques. En particulier, l'estimateur jackknife est approximativement sans biais en ce qui concerne le modèle ou le plan de sondage et est robuste en cas de spécification erronée des termes de variance dans le modèle (1.1). Cependant, dans (3.6), le facteur $(1-h_{ii})$ est inférieur ou égal à 1, de sorte que la valeur de l'estimateur jackknife sera supérieur à celle des autres estimateurs de la variance. Habituellement, dans ces conditions, le taux de

couverture des intervalles de confiance fondés sur l'estimateur jackknife est plus élevé que celui des intervalles de confiance fondés sur v_{R2}, v_{SSW}, ou v_T.

Notons aussi que, si l'échantillonnage se fait sans remise et que certaines probabilités de sélection de premier ou de deuxième ordre sont faibles, les estimateurs v_{R2}, v_D, v_J et v_J^* donneront lieu à une surestimation de la variance due au plan de sondage ou de la variance due au modèle. Pour tenir compte des probabilités de sélection non négligeables, nous pouvons procéder à certains ajustements simples. Une version ajustée de $v_J^*(\hat{\hat{Y}}_G)$, modelée sur v_{SSW}, est donnée par

$$v_{JP}^*(\hat{\hat{Y}}_G) = \frac{1}{N^2}\sum_s \frac{(1-\pi_i)g_i^2 r_i^2}{\pi_i^2(1-h_{ii})^2}.$$

Cette expression est similaire à l'estimateur $\hat{V}_{JK3}^*$ de Duchesne (2000), quoique celui-ci ne tienne pas compte des effets leviers. L'expression (3.6) fait aussi penser à un autre estimateur qui est étroitement lié à un estimateur de la variance de l'erreur du meilleur prédicteur linéaire non biaisé de la moyenne dans les conditions du modèle (1.1) (voir Valliant, Dorfman, et Royall 2000, chapitre 5). Cet estimateur est un peu moins conservateur que (3.6), mais comporte encore des corrections tenant compte des effets leviers :

$$v_D(\hat{\hat{Y}}_G) = \frac{1}{N^2}\sum_s \frac{g_i^2 r_i^2}{\pi_i^2(1-h_{ii})}.$$

Comme $h_{ii} = o(1)$, v_D est également approximativement dépourvu de biais par rapport au modèle et au plan de sondage. Une variante de cet estimateur qui pourrait donner de meilleurs résultats lorsque la valeur de certaines probabilités de sélection est grande et

$$v_{DP}(\hat{\hat{Y}}_G) = \frac{1}{N^2}\sum_s \frac{(1-\pi_i)g_i^2 r_i^2}{\pi_i^2(1-h_{ii})}.$$

4. RÉSULTATS DES SIMULATIONS

Pour déterminer la performance des estimateurs de la variance, nous avons réalisé plusieurs études de simulation en nous servant de trois populations différentes. La première est la population d'hôpitaux énumérée dans Valliant, Dorfman et Royall (2000, Annexe B). La deuxième est la population active décrite dans Valliant (1993). La troisième est une modification de la population active susmentionnée. Pour chacune de ces trois populations, l'échantillonnage a lieu sans remise, tel que décrit plus bas. Ces plans d'échantillonnage permettront de vérifier la notion selon laquelle les estimateurs de la variance motivés, en partie, par des plans d'échantillonnage avec remise demeurent utiles lorsqu'on les applique à des plans d'échantillonnage sans remise.

La population d'hôpitaux compte $N = 393$ unités et une valeur auxiliaire unique x, qui est le nombre de lits dans chaque hôpital. La variable Y représente le nombre de patients ayant reçu leur congé durant une période de référence particulière. Pour cette population, l'estimateur GREG est fondé sur le modèle $E_M(Y) = \beta_1 x^{1/2} + \beta_2 x$, $\mathrm{var}_M(Y) = \sigma^2 x$. Nous avons tiré des échantillons de taille égale à 50 et à 100 selon un plan d'échantillonnage aléatoire simple sans remise (eassr) et selon un plan d'échantillonnage avec probabilité proportionnelle à la taille (ppt) sans remise où la taille est égale à la racine carrée de x. Pour chaque combinaison de méthode d'échantillonnage et de taille d'échantillon, nous avons tiré 3 000 échantillons. Nous avons calculé les estimateurs $\hat{\hat{Y}}_G$, v_π, v_{R1}, v_{R2}, v_{SSW}, v_D, v_{DP}, v_J^*, v_{JP}^* et v_J pour chaque échantillon. Aux fins de comparaison, nous avons également inclus l'estimateur π, $\hat{\hat{Y}}_\pi = \hat{T}_\pi/N$. Nous avons inclus l'estimateur de la variance v_T, mais les résultats ne sont pas présentés ici parce qu'ils différaient fort peu de ceux obtenus pour v_{R2}.

La population active contient 10 841 personnes. Les variables auxiliaires que nous avons utilisées sont l'âge, le sexe et le nombre d'heures travaillées par semaine. La variable Y correspond à la rémunération hebdomadaire totale. Nous avons défini quatre groupes d'âge, à savoir 19 ans et moins, 20 à 24 ans, 25 à 34 ans et 35 ans et plus. Le modèle défini pour l'estimateur GREG inclut une ordonnée à l'origine, les effets principaux pour l'âge et le sexe et la variable quantitative correspondant au nombre d'heures travaillées. Nous avons posé que la variance du modèle était constante. Nous avons tiré des échantillons de taille égale à 50, 100 et 250, selon deux méthodes, à savoir l'eassr et l'échantillonnage sans remise avec probabilité proportionnelle au nombre d'heures travaillées. (Cette population présente certains regroupements par grappe, mais nous n'en n'avons pas tenu compte dans les simulations.)

La troisième population est une version modifiée de la population active conçue de façon à ajouter certaines observations aberrantes ou une certaine asymétrie dans les valeurs observées de la variable de rémunération hebdomadaire. Nous représentons cette nouvelle version par la notation « PA(mod.) » aux fins de référence. Dans le cas de la population active originale, nous avons plafonné la rémunération hebdomadaire à 999$. Pour chaque salaire plafonné de la sorte, nous avons généré une nouvelle rémunération égale à 1 000$ ainsi qu'une variable aléatoire lognormale dont la distribution est caractérisée par des paramètres d'échelle et de forme égaux à 6,9 et 1, respectivement. Nous avons produit les rémunérations recodées pour 4,4 % de la population. Avant le recodage, la rémunération moyenne annualisée était de 19 359$ et la valeur maximale, de 51 948$; après le recodage, la moyenne était de 23 103$ et la valeur maximale, de 608 116$. Donc, PA(mod.) est caractérisée par une courbe des revenus plus asymétrique que celle que l'on observerait pour la population réelle.

La distribution résultante de PA(mod.) est présentée à la figure 1 où la rémunération hebdomadaire est représentée graphiquement en fonction du nombre d'heures travaillées pour les sous-groupes définis selon l'âge. Dans chaque panneau, les points noirs correspondent aux hommes et les cercles non remplis, aux femmes. Dans chaque panneau, une ligne de référence horizontale est tracée à 999$. Bien qu'un nombre important de points se superposent, les caractéristiques générales sont évidentes. Le niveau de rémunération et la dispersion augmentent parallèlement à l'âge, le nombre d'heures travaillées par semaine est lié, quoiqu'assez faiblement, à la rémunération, et celle-ci présente l'asymétrie la plus forte pour le groupe des 25 à 34 ans et celui des 35 ans et plus. Le fait que la rémunération des hommes est généralement plus élevée que celle des femmes est moins évident.

Le tableau 1 montre le biais relatif empirique en pourcentage, défini comme étant la moyenne de $(\hat{T} - T)/T$ sur les échantillons pour l'estimateur π et pour l'estimateur de régression généralisée, pour les diverses populations et les diverses tailles d'échantillon. Nous présentons aussi les valeurs de la racine de l'erreur quadratique moyenne (reqm), définie comme étant la racine carrée de la moyenne de $(\hat{T} - T)^2$ sur l'ensemble des échantillons. Dans le cas de la population d'hôpitaux, le biais de l'un et l'autre estimateurs est négligeable, quelle que soit la taille de l'échantillon. Pour cette population, l'estimateur GREG est nettement plus efficace que l'estimateur π, parce qu'il existe une forte relation entre Y et x. Dans le cas des deux populations actives, l'estimateur π et l'estimateur GREG sont l'un et l'autre presque dépourvus de biais, mais l'estimateur GREG est un peu plus efficace, si l'on s'en tient à la reqm, pour toutes les tailles d'échantillon et toutes les méthodes d'échantillonnage.

Le tableau 2 donne le biais relatif (biais rel.) de neuf estimateurs de la variance, défini comme étant $100(\bar{v} - \text{eqm})/\text{eqm}$, où $\bar{v}$ est la moyenne de l'estimateur de la variance sur les 3 000 échantillons et eqm est l'erreur quadratique moyenne empirique de l'estimateur GREG. Les lignes du tableau sont ordonnées selon la taille du biais relatif dans PA(mod.) pour des eassr de taille égale à 50, mais le classement serait comparable pour les autres populations, tailles d'échantillon et méthodes d'échantillonnage. Dans le cas de la population d'hôpitaux, la fraction d'échantillonnage est importante, particulièrement si n=100. Comme il faut s'y attendre, étant donné cette situation, les estimateurs qui n'incluent aucune forme de correction pour tenir compte de la population finie (*fpc*), c'est-à-dire v_{R2}, v_D, v_J^* et v_J, surestiment fortement la variance, que l'on recoure à l'échantillonnage aléatoire simple sans

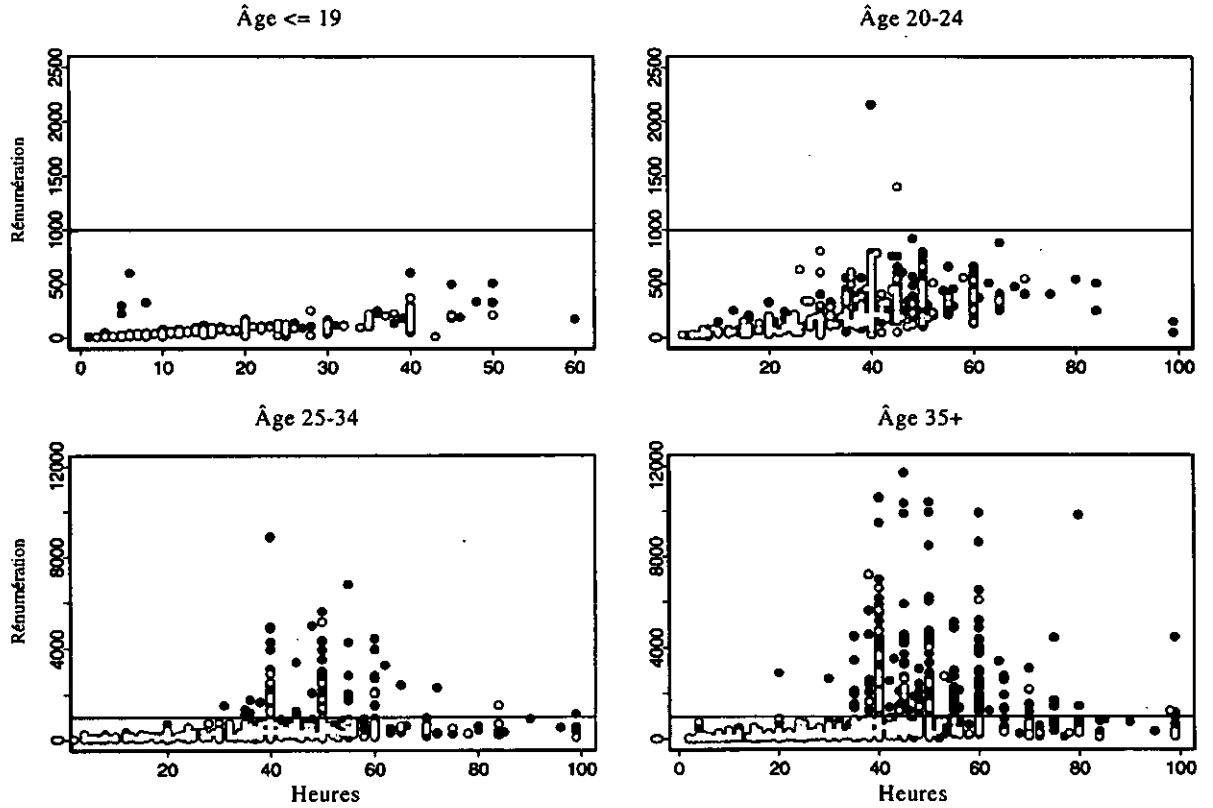


Figure 1. Diagramme de dispersion de la rémunération hebdomadaire en fonction du nombre d'heures travaillées par semaine pour quatre groupes d'âge, pour la population PA(mod.). Les cercles vides correspondent aux femmes. Les cercles pleins correspondent aux hommes. Une ligne horizontale est tracée à 999$, valeur maximale dans le cas de la population active originale.

remise (eassr) ou à l'échantillonnage avec probabilité proportionnelle à la taille (ppt). Comme il ne contient aucun terme reflétant la variance due au modèle de la somme sur les unités non comprises dans l'échantillon, v_{R1} sous-estime grandement l'eqm lorsque la fraction d'échantillonnage est importante.

Tableau 1

Biais relatif et racine carrée de l'erreur quadratique moyenne (reqm) de l'estimateur π et de l'estimateur de régression généralisée pour diverses études de simulation comptant chacune 3 000 échantillons

	Hôpitaux		Population active			PA(mod.)		
	n=50	n=100	n=50	n=100	n=250	n=50	n=100	n=250
Échantillonnage aléatoire simple								
$\hat{Y}_\pi$								
Bias rel. (%)	0,2	-0,1	-0,6	0	0	-0,1	0	-0,3
reqm	76,6	50,7	34,2	24,1	15,5	88,6	61,2	38,8
$\hat{Y}_G$								
Bias rel. (%)	0,2	0,2	0,1	0,1	0,2	0,4	0,2	-0,1
reqm	32,6	21,1	28,3	19,9	12,4	86,0	57,4	36,0
Échantillonnage avec probabilité proportionnelle à la taille								
$\hat{Y}_\pi$								
Bias rel. (%)	-0,1	0,1	-0,5	0	0	0	-0,1	-0,1
reqm	37,6	24,4	28,2	20,3	12,6	80,6	54,6	34,1
$\hat{Y}_G$								
Bias rel. (%)	0,1	0,1	-0,10	0,10	0	-0,6	-0,7	-0,4
reqm	27,2	16,9	28,2	19,3	12,0	81,8	55,1	33,5

Tableau 2

Biais relatif de neuf estimateurs de la variance pour l'estimateur de régression généralisée lors de diverses études de simulation comptant chacune 3 000 échantillons

	Hôpitaux		Population active			PA(mod.)		
	n=50	n=100	n=50	n=100	n=250	n=50	n=100	n=250
Échantillonnage aléatoire								
v_π	-8,6	-4,2	-18,1	-12,3	-7,5	-16,3	-2,8	-2,6
v_{R1}	-18,9	-27	-11,3	-9,9	-8	-9,6	-0,7	-3,3
v_{SSW}	-7,6	-3	-10,9	-9,1	-5,9	-9,3	0,1	-1,1
v_{R2}	5,9	30,1	-10,5	-8,2	-3,7	-8,8	1	1,3
v_{DP}	-1,4	0,2	0,1	-3,8	-3,8	0,6	5,1	0,8
v_D	13	34,3	0,6	-2,9	-1,6	1	6,1	3,2
v_J	18,4	37,4	13,9	2,2	0,3	11,2	10,5	4,8
v_{JP}^*	5,4	3,5	14	2,1	-1,7	12,4	10,5	2,7
v_J^*	20,8	38,8	14,5	3,1	0,7	12,9	11,5	5,2
Échantillonnage avec probabilité proportionnelle à la taille								
v_π	-5,9	-0,9	-22,1	-12,1	-6,8	-16,5	-10,6	-0,3
v_{R1}	-19,7	-32,4	-11,9	-7,7	-7,1	-9,1	-8,2	-2,7
v_{SSW}	-4	0	-11,6	-7	-4,9	-8,7	-7,3	-0,1
v_{R2}	16	52,6	-11,2	-6	-2,5	-8,3	-6,3	2,6
v_{DP}	0,1	2	0,8	-0,3	-1,6	0,9	-2,5	2,1
v_D	20,8	55,6	1,3	0,7	0,8	1,4	-1,5	4,8
v_J	23,6	57,2	22,6	11,8	5,3	14,6	4,7	7,3
v_{JP}^*	4,4	4	19,7	9,3	3,1	14,8	3,9	4,9
v_J^*	26,1	58,8	20,3	10,3	5,5	15,4	5	7,7

Dans le cas de la population active et de la population active modifiée PA(mod.), l'augmentation de la taille de l'échantillon fait diminuer le biais. Les estimateurs v_π, v_{R1}, v_{R2} et v_{SSW} sont entachés d'un biais négatif qui a tendance à devenir moins important à mesure que la taille de l'échantillon augmente. L'estimateur jackknife v_J et ses variantes, v_J^*, v_{JP}^*, surestiment la variance, particulièrement pour n=50. Pour chaque taille d'échantillon, les estimateurs v_D et v_{DP} sont presque dépourvus de biais dans de plus grandes proportions que la plupart des autres estimateurs.

Tableau 3

Taux de couverture des intervalles de confiance à 95 % pour des simulations au moyen de la population d'hôpitaux et de neuf estimateurs de la variance. En tout, 3 000 échantillons aléatoires simples ont été tirés sans remise pour des tailles d'échantillon de 50 et 100. L représente le pourcentage d'échantillons pour lesquels $(\hat{Y}_G - \bar{Y})/v^{1/2} < -1{,}96$; M représente le pourcentage d'échantillons pour lesquels $|\hat{Y}_G - \bar{Y}|/v^{1/2} \leq 1{,}96$; U représente le pourcentage d'échantillons pour lesquels $(\hat{Y}_G - \bar{Y})/v^{1/2} < 1{,}96$.

	n=50			n=100		
	L	M	U	L	M	U
Échantillonnage aléatoire simple						
v_π	3,1	92,1	4,8	2,6	93,6	3,9
v_{R1}	4,2	91,0	4,7	4,8	89,8	5,5
v_{SSW}	3,3	92,5	4,2	2,8	94,0	3,1
v_{R2}	2,8	93,9	3,3	1,4	97,0	1,6
v_{DP}	3,1	93,0	3,9	2,7	94,3	2,9
v_D	2,4	94,6	3,0	1,2	97,3	1,5
v_J	2,2	95,0	2,8	1,2	97,3	1,5
v_{JP}^*	2,9	93,6	3,5	2,6	94,6	2,9
v_J^*	2,2	95,1	2,8	1,2	97,4	1,4
Échantillonnage avec probabilité proportionnelle à la taille						
v_π	2,9	93,9	3,2	2,6	94,6	2,8
v_{R1}	4,1	92,0	3,9	5,0	89,3	5,7
v_{SSW}	2,9	94,2	2,9	2,6	94,8	2,6
v_{R2}	2,1	95,8	2,1	0,9	98,3	0,8
v_{DP}	2,7	94,5	2,8	2,5	95,0	2,5
v_D	1,9	96,2	1,9	0,9	98,3	0,8
v_J	1,8	96,3	1,9	0,9	98,4	0,7
v_{JP}^*	2,6	94,8	2,6	2,4	95,4	2,2
v_J^*	1,7	96,5	1,8	0,8	98,4	0,7

Le tableau 3 donne la couverture empirique des intervalles de confiance à 95 % sur les 3 000 échantillons pour chaque ensemble, dans le cas de la population d'hôpitaux. Les trois estimateurs de la variance contenant des ajustements pour les effets leviers, mais non la correction pour la population finie *fpc*, c'est-à-dire v_D, v_J^* et v_J^*, produisent une valeur plus grande et, donc, un taux de couverture plus élevé que v_π, v_{R2} et v_{SSW}. Stukel, Hidiroglou et Särndal (1996) ont, eux aussi constaté que l'estimateur jackknife a tendance à produire des estimations plus grandes de la variance de l'estimateur GREG que les autres estimateurs de la variance. Cette propriété est avantageuse dans le cas de la plus petite taille d'échantillon, n = 50. Si n = 100 et que la fraction d'échantillonnage est importante, les

estimateurs incluant une correction pour la population finie *fpc*, c'est-à-dire v_π, v_{SSW}, v_{DP} et v_J^*, ont un taux de couverture plus proche du taux nominal de 95 %, tandis que v_{R2}, v_D, v_J^* et v_J couvrent environ 97 % ou 98 % des échantillons. L'estimateur v_{JP}^*, qui est une approximation du jackknife, mais qui n'inclut pas de *fpc*, est un bon choix indépendamment de la taille d'échantillon ou du plan d'échantillonnage.

Les tableaux 4 et 5 montrent les taux de couverture pour la population active et la population active modifiée (PA(mod.)). Les estimateurs $v_{DP}, v_D, v_J, v_{JP}^*$ et v_J^* sont nettement meilleurs pour la population active que pour la population active modifiée pour $n = 50$, aussi bien dans le cas de l'eassr que de l'échantillonnage ppt. En revanche, pour $n = 250$, les taux de couverture sont les mêmes pour tous les estimateurs. L'estimateur fondé purement sur le plan de sondage, v_π, donne de mauvais résultats lorsque la taille d'échantillon est petite pour l'un et l'autre plans d'échantillonnage. Comme pour la population d'hôpitaux, v_{JP}^* produit une couverture quasiment égale au taux nominal pour chaque taille d'échantillon dans le cas de la population active.

Tableau 4

Taux de couverture des intervalles de confiance à 95 % pour des simulations au moyen de la population active et de la population active modifiée PA(mod.) et neuf estimateurs de la variance. En tout, 3 000 échantillons aléatoires simples ont été tirés sans remise pour des tailles d'échantillon de 50 et 100. *L* représente le pourcentage d'échantillons pour lesquels $(\hat{\bar{Y}}_G - \bar{Y})/v^{1/2} < -1{,}96$; *M* représente le pourcentage d'échantillons pour lesquels $|\hat{\bar{Y}}_G - \bar{Y}|/v^{1/2} \leq 1{,}96$; *U* représente le pourcentage d'échantillons pour lesquels $(\hat{\bar{Y}}_G - \bar{Y})/v^{1/2} > 1{,}96$.

	n=50			*n*=100			*n*=250		
	L	*M*	*U*	*L*	*M*	*U*	*L*	*M*	*U*
Population active									
v_π	5,3	91,4	3,2	4,3	92,8	2,9	2,8	94,1	3,1
v_{R1}	4,9	92,4	2,7	4,3	93,0	2,7	2,8	93,9	3,3
v_{SSW}	4,9	92,5	2,6	4,3	93,1	2,7	2,8	94,1	3,1
v_{R2}	4,9	92,5	2,6	4,2	93,2	2,6	2,5	94,6	2,9
v_{DP}	4,2	93,6	2,2	3,9	93,7	2,4	2,6	94,5	2,9
v_D	4,2	93,6	2,2	3,9	93,9	2,2	2,4	94,9	2,7
v_J	3	95,1	1,9	3,4	94,7	1,9	2,4	95,0	2,7
v_{JP}^*	3	95,1	1,9	3,3	94,7	1,9	2,5	94,8	2,7
v_J^*	3	95,1	1,9	3,3	94,8	1,9	2,4	95,0	2,7
PA(mod.)									
v_π	21,0	78,0	0,9	14,1	85,5	0,4	9,9	89,7	0,4
v_{R1}	20,9	78,7	0,3	14,1	85,7	0,2	10,2	89,5	0,3
v_{SSW}	20,9	78,8	0,3	14,0	85,8	0,2	9,9	89,9	0,3
v_{R2}	20,8	78,8	0,3	13,8	86,0	0,2	9,7	90,1	0,3
v_{DP}	19,7	80,0	0,2	13,4	86,5	0,1	9,7	90,1	0,3
v_D	19,7	80,0	0,2	13,2	86,7	0,1	9,6	90,1	0,3
v_J	18,4	81,4	0,2	12,7	87,2	0,1	9,4	90,3	0,3
v_{JP}^*	18,4	81,4	0,2	12,7	87,2	0,1	9,5	90,2	0,3
v_J^*	18,3	81,5	0,2	12,6	87,3	0,1	9,3	90,4	0,3

Tableau 5

Taux de couverture des intervalles de confiance à 95 % pour des simulations au moyen de la population active et de la population active modifiée PA(mod.) et neuf estimateurs de la variance. En tout, 3 000 échantillons ont été sélectionnés avec probabilité proportionnelle à la taille sans remise pour des tailles d'échantillon de 50, 100 et 250. *L* représente le pourcentage d'échantillons pour lesquels $(\hat{\bar{Y}}_G - \bar{Y})/v^{1/2} < -1{,}96$; *M* représente le pourcentage d'échantillons pour lesquels $|\hat{\bar{Y}}_G - \bar{Y}|/v^{1/2} \leq 1{,}96$; *U* représente le pourcentage d'échantillons pour lesquels $(\hat{\bar{Y}}_G - \bar{Y})/v^{1/2} > 1{,}96$.

	n=50			*n*=100			*n*=250		
	L	*M*	*U*	*L*	*M*	*U*	*L*	*M*	*U*
Population active									
v_π	5,7	90,2	4,1	3,7	92,9	3,4	3,1	94,3	2,6
v_{R1}	5,3	92,1	2,6	3,3	93,8	2,9	3,5	94,0	2,5
v_{SSW}	5,2	92,2	2,6	3,2	94,0	2,9	3,3	94,4	2,2
v_{R2}	5,2	92,3	2,6	3,1	94,1	2,8	3,0	94,8	2,2
v_{DP}	4,3	93,6	2,0	2,9	94,7	2,4	3,0	94,9	2,1
v_D	4,3	93,7	2,0	2,9	94,7	2,4	2,8	95,1	2,1
v_J	3,3	95,5	1,2	2,4	95,8	1,7	2,6	95,5	1,9
v_{JP}^*	3,3	95,4	1,3	2,6	95,5	1,9	2,7	95,3	1,9
v_J^*	3,3	95,4	1,3	2,6	95,6	1,8	2,6	95,6	1,8
PA(mod).									
v_π	19,6	79,7	0,7	15,0	84,4	0,7	9,9	89,8	0,4
v_{R1}	20,2	79,6	0,2	15,9	83,8	0,3	10,3	89,4	0,3
v_{SSW}	20,1	79,7	0,2	15,8	84,0	0,3	10,0	89,8	0,2
v_{R2}	20,1	79,7	0,2	15,6	84,1	0,2	9,8	90,0	0,2
v_{DP}	18,7	81,1	0,2	14,8	85,0	0,1	9,7	90,0	0,2
v_D	18,7	81,1	0,2	14,7	85,2	0,1	9,4	90,4	0,2
v_J	16,6	83,2	0,1	13,6	86,4	0,0	9,1	90,7	0,2
v_{JP}^*	16,6	83,3	0,1	13,9	86,1	0,0	9,4	90,4	0,2
v_J^*	16,5	83,4	0,1	13,8	86,2	0,0	9,1	90,7	0,2

Les résultats les plus frappants des tableaux 4 et 5 sont ceux présentés pour PA(mod.) pour laquelle tous les estimateurs de la variance produisent des intervalles de confiance dont la couverture est médiocre. Celle-ci varie de 78,0 % pour la combinaison (v_π, $n = 50$, eassr) à 90,7 % pour (v_J et v_J^*, $n = 250$, ppt). Virtuellement tous les cas de non-couverture sont dus au fait que $(\hat{\bar{Y}}_G - \bar{Y})/v^{1/2} < -1{,}96$, où v représente n'importe quel estimateur de la variance. Nous observons des taux de couverture médiocres même si l'estimateur π et l'estimateur GREG sont sans biais sur tous les échantillons (voir le tableau 1) et si, dans le cas de v_J, v_{JP}^* et v_J^*, l'estimateur produit une surestimation de la variance (voir tableau 2).

Des erreurs d'estimation négative, $\hat{\bar{Y}}_G - \bar{Y}$, ont lieu dans le cas d'échantillons qui ne comptent qu'un nombre assez faible de personnes ayant une rémunération hebdomadaire élevée. La figure 2 donne le tracé de la statistique *t* fondée sur $\sqrt{v_{JP}^*}$, *i.e.*, $(\hat{\bar{Y}}_G - \bar{Y})/\sqrt{v_{JP}^*}$, en fonction du nombre de personnes dans l'échantillon dont la rémunération hebdomadaire est égale ou supérieure à 1 000$ dans des ensembles de 1 000 échantillons pour (eassr, $n = 50$, 100, 250). Les erreurs d'estimation négatives dans le cas d'échantillons comptant un petit nombre de personnes ayant un revenu élevé produisent une statistique *t* négative et des intervalles de confiance qui manquent la moyenne de

population du côté de la borne inférieure. La gravité du problème diminue à mesure qu'augmente la taille de l'échantillon, mais la convergence vers le taux de couverture nominal est lente et se produit « de façon ascendante ». Quel que soit l'estimateur de la variance utilisé, la couverture est inférieure à 95 % à moins que l'échantillon ne soit assez grand.

Nous évaluons aussi la performance des estimateurs de la variance en fonction des caractéristiques d'échantillon. Par souci de concision, nous présentons uniquement les résultats concernant le biais des estimateurs de la variance. Dans le cas de la population d'hôpitaux, nous avons trié les échantillons en fonction de $D_x = \mathbf{1}'(\hat{\mathbf{T}}_x - \mathbf{T}_x)$, qui est la somme des différences entre les estimateurs π des totaux de $x^{1/2}$ et x et les totaux de population correspondants. Puis, nous avons formé 20 groupes contenant chacun 150 échantillons. Dans chaque groupe, nous avons calculé le biais de $\hat{\hat{Y}}_G$, ainsi que la racine de l'erreur quadratique moyenne (reqm) et la racine carrée de la moyenne de chaque estimateur de la variance. Les résultats sont présentés graphiquement à la figure 3 pour le plan eassr lorsque $n = 50$ et 100 et pour le plan ppt lorsque $n = 50$ et 100. Nous avons tracé le graphique pour un sous-ensemble d'estimateurs de la variance. Dans chaque panneau, l'axe horizontal donne les valeurs de D_x. Puisque v_J, v_J^*, v_D et v_{R2} sont similaires pour la plupart de l'étendue de D_x, nous n'avons tracé que les points obtenus pour l'estimateur jackknife v_J. En outre, v_{DP} et v_{JP}^* étant proches, seul le dernier est représenté. L'estimateur GREG est entaché d'un biais conditionnel qui influe sur la reqm si les échantillons ne sont pas équilibrés. Les propriétés conditionnelles médiocres de v_π sont surtout évidentes dans le cas de l'échantillonnage aléatoire simple pour lequel le biais de v_π en tant qu'estimation de l'eqm passe d'une valeur négative à une valeur positive sur l'étendue de D_x. Pour les autres estimations de la variance, le biais conditionnel est comparable au biais inconditionnel du tableau 2. Théoriquement, v_{JP}^* et v_{SSW} sont tous deux approximativement dépourvus de biais par rapport au plan de sondage et au modèle et suivent bien la reqm.

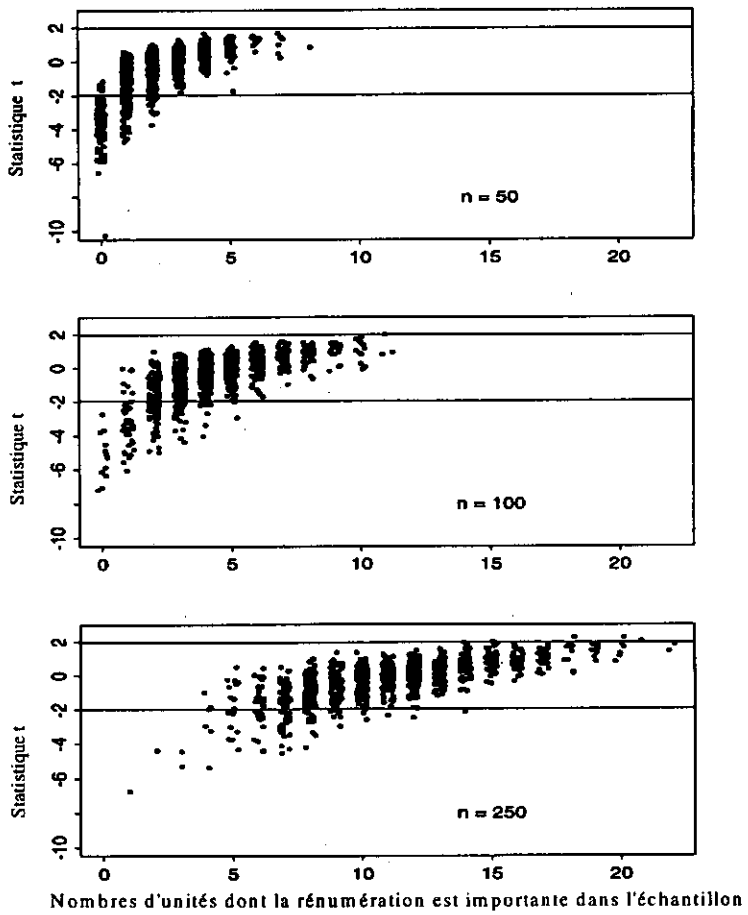


Figure 2. Tracé de la statistique t en fonction du nombre de personnes dans l'échantillon dont la rémunération hebdomadaire est supérieure à 1 000$ dans les ensembles de 1 000 échantillons aléatoires simples de taille $n = 50$, 100, 250 provenant de la population PA(mod.). Des droites horizontales des référence sont tracées à ±1,96. Les points sont déplacés légèrement par rapport à leur position normale afin de réduire au minimum les superpositions.

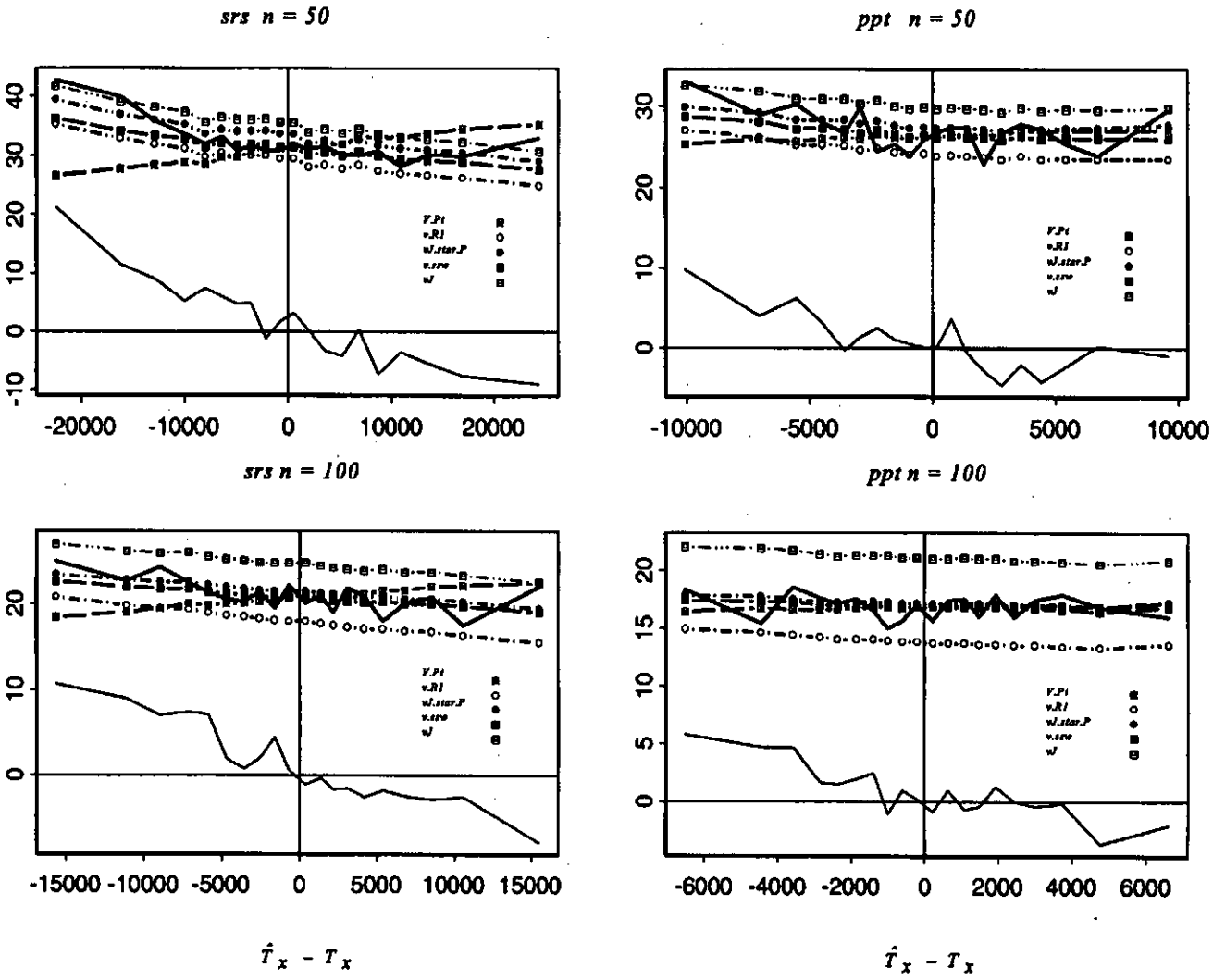


Figure 3. Tracé des valeurs du biais conditionnel, de la reqm et de la moyenne des estimations de l'erreur-type de l'estimateur GREG pour les échantillons provenant de la population des hôpitaux. Des droites horizontale et verticale de référence sont tracées au point zéro. Dans chaque panneau, la courbe la plus basse représente le biais de l'estimateur GREG. La droite en trait plein épais représente la racine de l'erreur quadratique moyenne conditionnelle.

La figure 4 présente un graphique similaire pour les échantillons tirés de la population active. Les ensembles d'estimations (v_D, v_{DP}) et (v_J, v_J^*, v_{JP}) sont fort semblables et seule la première de chaque ensemble est incluse dans les graphiques. Seuls les échantillons eassr et ppt de taille n = 50 et 250 sont inclus. De nouveau, l'axe horizontal est D_x, qui représente la somme des différences entre les estimations π et les valeurs de population des totaux pour les groupes âge-sexe et le nombre d'heures travaillées par semaine. Le biais conditionnel de v_π est évident dans le cas des échantillons pour lesquels les valeurs de D_x sont les plus faibles, mais le problème diminue lorsque la taille d'échantillon augmente, aussi bien dans le cas de l'eassr que de l'échantillonnage ppt. L'estimateur jackknife v_J est, en moyenne, celui qui produit l'estimation de la variance la plus importante sur toute l'étendue de D_x. Les différences entre les estimations de la variance et leur biais sont moindres pour les échantillons de grande taille. Les estimateurs v_D, v_{SSW} et v_J suivent tous raisonnablement bien la reqm, sauf dans les cas où D_x est la plus négative, la valeur étant alors assez faible pour tous les estimateurs.

5. CONCLUSION

Divers estimateurs de la variance de l'estimateur de régression généralisée sont proposés dans la littérature sur l'échantillonnage, principalement dans le but d'estimer la variance due au plan de sondage. Il est facile de construire des estimateurs approximativement non biaisés de la variance due au plan de sondage et, dans le cas de certains modèles, de la variance due au modèle. En outre, les estimateurs bivalents étudiés ici sont des estimateurs robustes de la variance due au modèle, même si le modèle qui sert de fondement à l'estimateur GREG contient un paramètre de variance incorrect.

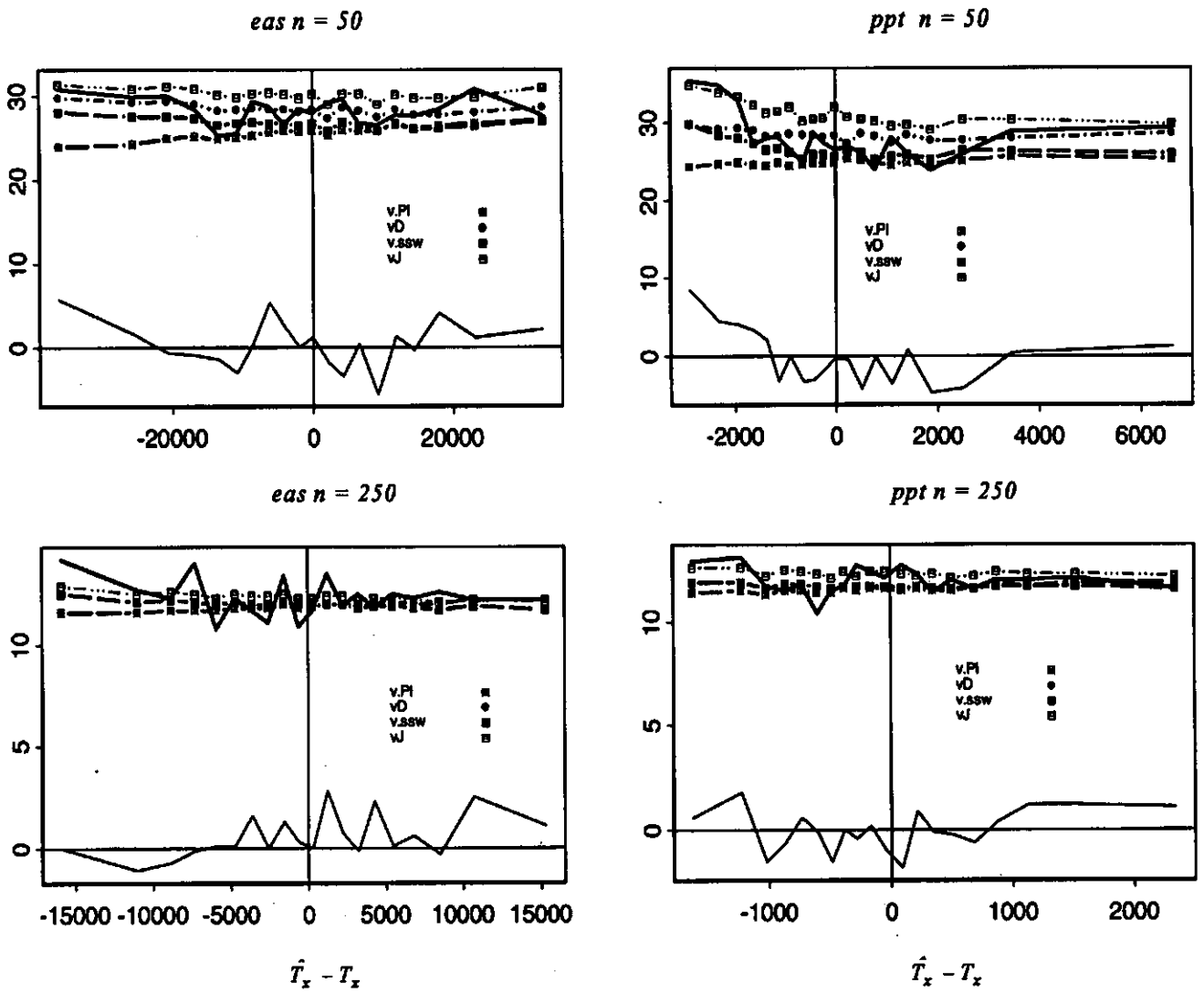


Figure 4. Tracé des valeurs du biais conditionnel, de la reqm et de la moyenne des estimations de l'erreurs-type de l'estimateur GREG pour les échantillons provenant de la population active. Des droites horizontale et verticale de référence sont tracés au point zéro. Dans chaque panneau, la courbe la plus basse représente le biais de l'estimateur GREG. La droite en trait plein épais représente la racine de l'erreur quadratique moyenne conditionnelle.

Une caractéristique importante des meilleurs de ces estimateurs est l'ajustement des carrés des résidus au moyen de facteurs analogues aux effets leviers utilisés en analyse par régression classique. Il est bien connu, en économétrie, qu'il est souhaitable de corriger pour les effets leviers les estimateurs par régression de la variance, en vue de combattre l'hétéroscédasticité. Cette correction a été proposée par MacKinnon et White (1985) et réexaminée récemment par Long et Ervin (2000). L'un des meilleurs choix est une approximation de l'estimateur jackknife, représentée ici par v_{JP}^{*}, qui inclut une forme de correction pour tenir compte de la population finie.

Les estimateurs robustes étudiés ici sont assez utiles pour les variables dont la distribution présente un « comportement » raisonnable. Si les échantillons sont de petite taille ou de taille moyenne, ils produisent des estimateurs corrigés de la variance souvent caractérisée par une meilleure couverture de l'intervalle de confiance. Cependant, ils n'offrent aucune défense si les variables sont très asymétriques et que les observations dont la valeur est forte ne sont pas bien représentées dans l'échantillon. Qu'il s'agisse d'un problème d'asymétrie ou de valeurs aberrantes, l'effet est évident. Un échantillon ne contenant pas un nombre suffisant d'unités pour lesquelles les valeurs sont élevées produira une estimation trop faible de la moyenne, souvent assortie d'un estimateur de la variance produisant une estimation trop faible. Comme l'illustrent les simulations décrites à la section 4, pour les échantillons de ce genre, même le meilleur des estimateurs proposés de la variance ne produira pas d'intervalle de confiance dont la couverture correspond au taux nominal. Les méthodes de transformation de Chen et Chen (1996) paraissent prometteuses, mais leur application devrait être testée pour les estimateurs GREG plus complexes étudiés ici.

La résolution la plus efficace du problème d'asymétrie ne consiste pas à mieux utiliser les données

échantillonnales. Il semble plutôt que le plan d'échantillonnage proprement dit doit être conçu de façon à assurer une bonne représentation des grandes unités. Toutefois, dans de nombreux cas, comme celui d'une enquête auprès des ménages afin de mesurer le revenu ou les biens d'équipement, cette solution peut être difficile, voir impossible, si l'on ne dispose pas de données auxiliaires étroitement liées à la variable étudiée. Le cas échéant, il pourrait être utile de mieux utiliser les données échantillonnales au moyen de modèles pour variables asymétriques (voir, par exemple Karlberg 2000).

REMERCIEMENTS

L'auteur remercie Alan Dorfman dont les idées sont à l'origine de la présente étude, ainsi que le rédacteur en chef adjoint et les deux évaluateurs pour leur examen minutieux.

ANNEXE : Détails des calculs de l'estimateur Jackknife

L'utilisation de (3.2), (3.3) et du résultat de la matrice type dans le lemme 5.4.1 de Valliant et coll. (2000) nous donne

$$\mathbf{A}_{\pi s(i)}^{-1} = \left[\mathbf{A}_{\pi s}^{-1} + \frac{\mathbf{A}_{\pi s}^{-1}\mathbf{x}_i\mathbf{x}_i'\,\mathbf{A}_{\pi s}^{-1}/v_i\pi_i}{1-h_{ii}}\right].$$

De ceci et de la définition de $\hat{\mathbf{B}}_{(i)}$, il découle que l'estimateur de la pente, si l'on omet l'unité i, est $\hat{\mathbf{B}}_{(i)} = \hat{\mathbf{B}} + n^{-1}\sum_s \mathbf{Q}_i$ où

$$\mathbf{Q}_i = \frac{\mathbf{A}_{\pi s}\mathbf{x}_i}{1-h_{ii}}\frac{r_i}{v_i\pi_i}.$$

L'estimateur GREG, après suppression de l'unité i, est

$$\hat{T}_{G(i)} = \frac{n}{n-1}\left(\hat{T}_\pi - \frac{Y_i}{\pi_i}\right) + \left(\hat{\mathbf{B}}-\mathbf{Q}_i\right)'\left[\mathbf{T}_x - \frac{n}{n-1}\left(\hat{\mathbf{T}}_x - \frac{\mathbf{x}_i}{\pi_i}\right)\right].$$

Après certains réarrangements, nous pouvons le réécrire sous la forme

$$\hat{T}_{G(i)} = \frac{n}{n-1}\hat{T}_G - \frac{n}{n-1}\left[\frac{g_i r_i}{\pi_i(1-h_{ii})}\right] + \frac{n}{n-1}G_i + \frac{1}{n-1}K_i$$

où

$$G_i = \frac{h_{ii}Y_i - \hat{Y}_i}{\pi_i(1-h_{ii})}$$

et

$$K_i = (\hat{\mathbf{B}} - \mathbf{Q}_i)'\left(\frac{n\mathbf{x}_i}{\pi_i} - \hat{\mathbf{T}}_x\right).$$

Il s'ensuit que $\hat{T}_{G(i)} - \hat{T}_{G(\cdot)} = -n(n-1)^{-1}(D_i - \bar{D}_s) + n(n-1)^{-1}F_i$ où $F_i = (G_i - \bar{G}_s) + n^{-1}(K_i - \bar{K}_s)$ et $\bar{G}_s$ et $\bar{K}_s$ sont des moyennes d'échantillon dont les définitions sont évidentes. La substitution dans la formule du jackknife (3.1) donne

$$v_J\left(\hat{\bar{Y}}_G\right) = N^{-2}\frac{n}{n-1}\times$$

$$\left[\sum_s (D_i - \bar{D}_s)^2 + \sum_s F_i^2 - 2\sum_s F_i(D_i - \bar{D}_s)\right]. \quad \text{(A.1)}$$

La formule (A.1) est exacte, mais, grâce à certaines approximations supplémentaires, nous pouvons obtenir les tailles relatives des termes. Partant des valeurs de G_i et K_i données plus haut et du fait que h_{ii} et les éléments de $\mathbf{Q}_i$ sont $o(1)$, nous avons

$$\begin{aligned} G_i + n^{-1}K_i &= \frac{h_{ii}Y_i - \hat{Y}_i}{\pi_i(1-h_{ii})} + \frac{1}{n}(\hat{\mathbf{B}} - \mathbf{Q}_i)'\left(\frac{n\mathbf{x}_i}{\pi_i} - \hat{\mathbf{T}}_x\right) \\ &\approx -\frac{\hat{Y}_i}{\pi_i} + \hat{\mathbf{B}}'\frac{\mathbf{x}_i}{\pi_i} - \frac{1}{n}\hat{\mathbf{B}}'\hat{\mathbf{T}}_x \\ &= -\frac{1}{n}\hat{\mathbf{B}}'\hat{\mathbf{T}}_x \end{aligned}$$

où $\approx$ signifie « aymptomatiquement équivalent à ». Il s'ensuit que $F_i \approx 0$ et que $v_J(\hat{\bar{Y}}_G) \approx \sum_s (D_i - \bar{D}_s)^2$, c'est-à-dire que (3.5) est vérifié.

Ensuite, nous pouvons montrer que le deuxième terme de (3.5) converge en probabilité vers zéro. Le vecteur des résidus peut être exprimé sous la forme $\mathbf{r}_s = (\mathbf{I} - \mathbf{H}_s)\mathbf{Y}_s$, et le deuxième terme de (3.5) est égal à $N^{-2}n^{-1}\mathbf{g}_s'\,\Pi_s^{-1}\mathbf{U}^{-1}\mathbf{r}_s\mathbf{r}_s'\mathbf{U}^{-1}\Pi_s^{-1}\mathbf{g}_s$ où $\mathbf{U} = \mathrm{diag}(1-h_{ii})$, $i \in s$. Donc, le deuxième terme de (3.5) est le carré de $B = N^{-1}n^{-1/2}\mathbf{g}_s'\,\Pi_s^{-1}\mathbf{U}^{-1}\mathbf{r}_s$ dont l'espérance est nulle dans les conditions de tout modèle pour lequel $E_M(r_i) = 0$. La variance de B due au modèle est

$$\begin{aligned} &N^{-2}n^{-1}\,\mathrm{var}_M(\mathbf{g}_s'\,\Pi_s^{-1}\mathbf{U}^{-1}\mathbf{r}_s) = \\ &N^{-2}n^{-1}\mathbf{g}_s'\,\Pi_s^{-1}\mathbf{U}^{-1}(\mathbf{I}-\mathbf{H}_s)\times \\ &\mathbf{V}_s(\mathbf{I}-\mathbf{H}_s)'\,\mathbf{U}^{-1}\Pi_s^{-1}\mathbf{g}_s \end{aligned} \quad \text{(A.2)}$$

dont l'ordre de grandeur est n^{-2} étant donné les hypothèses que nous avons formulées. Par conséquent, le deuxième terme de (3.5) est le carré d'un terme dont la moyenne est nulle et dont la variance due au modèle tend vers zéro à mesure qu'augmente la taille de l'échantillon. Alors, le deuxième terme de (3.5) converge vers zéro conformément à l'inégalité de Chebyshev. Ceci justifie (3.6).

BIBLIOGRAPHIE

BELSLEY, D.A., KUH, E. et WELSCH, R.E. (1980). *Regression Diagnostics*. New York: John Wiley & Sons, Inc.

BREWER, K.R.W. (1995). Combining design-based and model-based inference. Chapter 30 dans *Business Survey Methods*, (Éds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. College, et P.S. Kott). New York: John Wiley & Sons, Inc., 589-606.

BREWER, K.R.W. (1999). Le calage esthétique dans le cas de l'échantillonnage avec probabilités inégales. *Techniques d'enquête*, 25, 231-239.

CHEN, G., et CHEN, J. (1996) Une méthode de transformation applicable à l'échantillonnage de populations finies calée par une méthode de vraisemblance empirique. *Techniques d'enquête*, 22, 139-147.

DUCHESNE, P. (2000). A note on jackknife variance estimation for the general regression estimator. *Journal of Official Statistics*, 16, 133-138.

HIDIROGLOU, M.A., FULLER, W.A. et HICKMAN, R.D. (1980). SUPERCARP. Department of Statistics. Ames, Iowa: Iowa State University.

KARLBERG, F. (2000). Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16, 229-243.

KOTT, P.S. (1990). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference*, 24, 287-296.

KREWSKI et RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

LONG, J.S., et ERVIN, L.H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54, 217-224.

MACKINNON, J.G., et WHITE, H. (1985). Some heteroskedastic consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29, 53-57.

RAO, J.N.K., et WU, C.J.F. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

ROYALL, R.M., et CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.

ROYALL, R.M., et CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.

SÄRNDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.

SÄRNDAL, C.-E., SWENSSON, B. et WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator. *Biometrika*, 76, 527-537.

SÄRNDAL, C.-E., SWENSSON, B. et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SÄRNDAL, C.-E., et WRIGHT, R. (1984). Cosmetic form of estimators in survey sampling. *Scandanavian Journal of Statistics*, 11, 146-156.

STUKEL, D., HIDIROGLOU, M.A. et SÄRNDAL, C.-E. (1996). Estimation de la variance des estimateurs de calage : comparaison des méthodes du jackknife et de la linéarisation de Taylor. *Techniques d'enquête*, 22, 117-126.

SUGDEN, R.A., et SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.

VALLIANT, R., DORFMAN, A.H. et ROYALL, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1-25.

YUNG, W., et RAO, J.N.K. (1996). Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples. *Techniques d'enquête*, 22, 23-31.

YUNG, W., et RAO, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95, 903-915.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 18, No. 1, 2002

Evaluating Socio-economic (SES) Bias in Survey Nonresponse
John Goyder, Keith Warriner, and Susan Miller 1

Are Nonrespondents to Health Surveys Less Healthy than Respondents
G. Cohen and J.C. Duffy 13

Accounting for Biases in Election Surveys; The Case of the 1998 Quebec Election
Claire Durand, Andre Blais, and Sebastien Vachon 25

Small Area Estimation via Generalized Linear Models
Alastair Noble, Stephen Haslett, and Greg Arnold 45

Generalized Fisher Price Indexes and the Use of Scanner Data in the Consumer Price Index (CPI)
Jan de Haan 61

Research and Development in Official Statistics and Scientific Co-operation with Universities: An Empirical Investigation
Risto Lehtonen, Erkki Pahkinen, and Carl-Erik Särndal 87

Book and Software Reviews 111

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 30, No. 1, March/mars 2002, 1-176

Richard A. LOCKHART
Editor's Report/Rapport du rédacteur en chef 1

Yong YOU and J.N.K. RAO
Small area estimation using unmatched sampling and linking models 3

Debbie J. DUPUIS and Stephan MORGENTHALER
Robust weighted likelihood estimators with an application to bivariate extreme value problems 17

Patrick BÉLISLE, Lawrence JOSEPH, David B. WOLFSON and Xiaojie ZHOU
Bayesian estimation of cognitive decline in patients with Alzheimer's disease 37

Joseph G. IBRAHIM, Ming-Hui CHEN and Stuart R. LIPSITZ
Bayesian methods for generalized linear models with covariates missing at random 55

Mario TROTTINI and Fulvio SPEZZAFERRI
A generalized predictive criterion for model selection 79

Pamela OHMAN-STRICKLAND and George CASELLA
Approximate and estimated saddlepoint approximations 97

Yong B. LIM, Jerome SACKS, W.J. STUDDEN and William J. WELCH
Design and analysis of computer experiments when the output is highly correlated over the input space 109

Boxin TANG, Fengshi MA, Debra INGRAM and Hong WANG
Bounds on the maximum number of clear two-factor interactions for 2^{m-p} designs of resolution III and IV 127

Case study in data analysis: The genetic analysis of inflammatory bowel disease 137

Lucia MIREA, Shelley B. BULL, Mark S. SILVERBERG and Katherine A. SIMINOVITCH
Introduction and Analysis 1: The genetic analysis of a complex disease 138

Jiahua CHEN, John D. KALBFLEISCH and Sandra ROMERO-HIDALGO
Analysis 2: Genetic data analysis of affected sib pairs 145

Gerarda A. DARLINGTON and Andrew D. PATERSON
Analysis 3: Genetic analysis of chromosome 6 in inflammatory bowel disease 152

Nicole M. ROSLIN, J.C. LOREDO-OSTI, Celia M.T. GREENWOOD and Kenneth MORGAN
Analysis 4: Genetic analysis of the role of the HLA region in inflammatory bowel disease 158

Christopher A. FIELD and Bruce SMITH
Discussion of the evaluation of a candidate genetic locus in a genome scan of complex disease 167

Acknowledgement of referees' service/Remerciements aux membres des jurys 175

Forthcoming Papers/Articles à paraître 176

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points ci-dessous. Les articles acceptés doivent être soumis sous forme de fichiers de traitement de texte, préférablement WordPerfect. Les autres logiciels sont acceptables, mais une version sur papier sera alors exigée pour le traitement des formules et des figures.

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
1.4 Les remerciements doivent paraître à la fin du texte.
1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0; l, 1).
3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.