C 3

TECHNIQUES D'ENQUÊTE

Nº 12-001-XPB au catalogue

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

JUIN 1997

•

VOLUME 23

•

NUMÉRO 1

Statistique Canada Statistics Canada

Canada

DATE DUE

TECHNIQUES D'ENQUÊTE

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

JUIN 1997 • VOLUME 23 • NUMÉRO 1

Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, 1997

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.

Juillet 1997

N° 12-001-XPB au catalogue

Périodicité: semestrielle

ISSN 0714-0045

Ottawa

800

10102437 63

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. 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. Binder
G.J.C. Hole
F. Mayda (Directeur de la Production)
C. Patrick
R. Platek (Ancien président)
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. Binder, *Statistique Canada*
J.-C. Deville, *INSEE*
J.D. Drew, *Statistique Canada*
W.A. Fuller, *Iowa State University*
R.M. Groves, *University of Maryland*
M.A. Hidiroglou, *Statistique Canada*
D. Holt, *Central Statistical Office, U.K.*
G. Kalton, *Westat, Inc.*
R. Lachapelle, *Statistique Canada*
S. Linacre, *Australian Bureau of Statistics*
G. Nathan, *Central Bureau of Statistics, Israel*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
I. Sande, *Bell Communications Research, U.S.A.*
F.J. Scheuren, *George Washington University*
J. Sedransk, *Case Western Reserve University*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
R. Valliant, *U.S. Bureau of Labor Statistics*
V.K. Verma, *University of Essex*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Rédacteurs adjoints J. Denis, P. Dick, H. Mantel et D. Stukel, *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 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 $ par année au Canada et de 47 $ US par année à l'extérieur du Canada. Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division des opérations et de l'intégration, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au (613) 951-7277 ou au 1 800 700-1033, par télécopieur au (613) 951-1584 ou au 1 800 889-9734 ou par Internet : 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 et la Société Statistique du Canada.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 23, numéro 1, juin 1997

TABLE DES MATIÈRES

Dans ce numéro 1

J.E. STAFFORD et D.R. BELLHOUSE
Une algèbre informatique pour la théorie des enquêtes par échantillonnage 3

S. HINKINS, H.L.OH et F. SCHEUREN
Algorithmes de plan de sondage inverses 13

P.L.D. NASCIMENTO SILVA et C.J. SKINNER
Sélection des variables pour l'estimation par régression dans le cas des populations finies 25

J.L. ELTINGE et I.S. YANSANEH
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 37

M.S. KOVAČEVIĆ et W. YUNG
Estimation de la variance des mesures de l'inégalité et de la polarisation du revenu - Étude empirique 47

K. HUMPHREYS et C.J. SKINNER
Estimation des variables instrumentales des flux bruts en présence de l'erreur de mesure 61

J. WAKSBERG, D. JUDKINS et J.T. MASSEY
Suréchantillonnage géographique dans les enquêtes démographiques aux États-Unis 69

W.C. LOSINGER
Nouvel estimateur de l'erreur-type pour les groupes aléatoires 81

K. ZEELENBERG
Dérivation simple de l'estimateur de régression par linéarisation 85

Techniques d'enquête, juin 1997
Vol. 23, N° 1, pp. 1-2
Statistique Canada

Dans ce numéro

Ce numéro de *Techniques d'enquête* contient des articles qui traitent de divers sujets. Dans le premier article, Stafford et Bellhouse présentent les blocs fonctionnels de base pour l'élaboration d'une algèbre informatique complète pour la théorie de l'échantillonnage. Ils montrent que trois techniques de base de la théorie de l'échantillonnage dépendent de l'application répétée des règles qui donnent lieu aux partitions. La méthodologie est illustrée au moyen d'applications au calcul des moments de la moyenne de l'échantillon, de l'estimateur par quotient et de l'estimateur de régression dans le cas spécial de l'échantillonnage aléatoire simple sans remise. L'application machine de la méthodologie décrite a été faite à l'aide du langage de programmation *Mathematica.*

Hinkins, Oh et Scheuren exposent une nouvelle stratégie pour l'analyse des données d'enquêtes complexes. Ils ont prélevé un sous-échantillon de telle manière que le sous-échantillon puisse être considéré comme un échantillon aléatoire simple de la population initiale, puis ils y ont appliqué la méthode standard pour les variables aléatoires indépendantes de même distribution. Ils proposent de répéter la méthode plusieurs fois pour récupérer l'information perdue durant le sous-échantillonnage de l'échantillon initial. Ils montrent comment appliquer leur méthode à l'échantillonnage stratifié, à l'échantillonnage en grappes à un ou deux degrés ainsi qu'aux plans avec deux unités primaires d'échantillonnage par strate.

Nascimento Silva et Skinner examinent le problème de la sélection des variables pour l'estimation de régression. Ces auteurs ont mis au point une méthode axée sur la réduction de l'erreur quadratique moyenne de l'estimateur obtenu. Ils comparent de façon empirique leur approche à d'autres en utilisant les données d'un essai fait en 1988 sur les techniques brésiliennes de recensement; les méthodes proposées ont et de bonnes propriétés en ce qui a trait à l'erreur quadratique moyenne et au biais.

Eltinge et Yansaneh étudient le problème de la formation de cellules d'ajustement pour la non-réponse. Dans le contexte des paradigmes généraux des cellules basées sur l'estimation des unités et des probabilités, ils examinent une variété de diagnostics pour l'évaluation d'un ensemble de cellules d'ajustement. Les méthodes de diagnostic examinées incluent les suivantes : comparaison des estimations et des erreurs-types pour différents nombres de cellules d'ajustement; évaluation du biais à l'intérieur des cellules; évaluation de l'étendue des cellules en regard de la précision des probabilités de réponse estimées et comparaisons des estimations basées sur les cellules par rapport aux estimations non corrigées.

Kovačević et Yung ont mené une étude empirique visant à comparer des méthodes d'estimation de la variance pour les mesures de l'inégalité du revenu estimées à partir de données d'enquêtes complexes. Les méthodes d'estimation de la variance examinées sont les suivantes: méthode jackknife, méthode «bootstrap», méthode du demi-échantillon équilibré groupé, méthode itérative du demi-échantillon équilibré groupé et méthode de Taylor basée sur les équations d'estimation. Après avoir comparé les biais relatifs, la stabilité relative et les propriétés de couverture des intervalles de confiance associés, pour un certain nombre de mesures de l'inégalité du revenu, ils concluent que la méthode de Taylor est celle qui donne les meilleurs résultats, suivie de la méthode «bootstrap».

Humphreys et Skinner étudient l'utilisation de la méthode de la variable instrumentale pour l'estimation des mouvements bruts entre états discrets. Cette méthode pourrait être utile lorsque les estimations externes des taux de classification erronée ne sont pas disponibles. Ils illustrent leur méthode à l'aide de données obtenues de la U.S. Panel Study of Income Dynamics, les deux états utilisés étant «actifs» et «inactifs». Leurs résultats indiquent que, lorsqu'une erreur de mesure est présente, les estimations non corrigées peuvent comporter un biais considérable – un problème que l'on peut éviter en utilisant des variables instrumentales appropriées.

Waksberg, Judkins et Massey discutent des problèmes liés au suréchantillonnage de régions géographiques pour produire des estimations pour de petits domaines de la population lors d'enquêtes démographiques, conjointement avec la présélection des ménages. Ils présentent une évaluation empirique de la réduction de la variance, ainsi qu'une évaluation de la robustesse de l'échantillonnage dans le temps. Ils discutent du suréchantillonnage géographique simultané pour l'estimation de plusieurs petits domaines.

Dans son article, Losinger propose un estimateur modifié de l'erreur-type des groupes aléatoires pour les données obtenues de l'échantillon du recensement décennal aux États-Unis. L'estimateur habituel des groupes aléatoires comporte deux propriétés non souhaitables pour les variables binomiales: premièrement, les estimations de l'erreur-type pour les réponses «oui» et «non» ne sont pas égales; deuxièmement, si tous les répondants indiquent «oui», l'erreur-type estimée n'est pas égale à zéro. La modification proposée consiste essentiellement à appliquer un ajustement de ratio à chaque estimation par sous-groupe afin qu'il y ait concordance entre les estimations par sous-groupe de la population et la valeur totale.

Enfin, Zeelenberg présente une technique simple qui fait appel à l'utilisation des différentiels pour linéariser des estimateurs non linéaires basés sur le plan. En bout de ligne, les expressions linéarisées permettent d'obtenir des expressions simples basées sur la méthode de Taylor pour les variances des estimateurs non linéaires. L'auteur illustre la méthode proposée à l'aide de deux exemples : l'estimateur du coefficient de régression et l'estimateur de régression.

Le rédacteur en chef

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 3-11
Statistique Canada

Une algèbre informatique pour la théorie des enquêtes par échantillonnage

J.E. STAFFORD et D.R. BELLHOUSE[1]

RÉSUMÉ

Les auteurs présentent un système de procédures qui peut servir à automatiser les calculs algébriques complexes que l'on retrouve souvent en théorie des enquêtes par échantillonnage. Ils montrent que trois techniques de base en théorie de l'échantillonnage dépendent de l'application répétée de règles donnant lieu à des partitions: le calcul des valeurs espérées dans un plan d'échantillonnage quelconque à un degré, la détermination d'estimateurs non biaisés ou convergents dans le cadre de ces plans et le développement en séries de Taylor. La méthode est appliquée ici à des calculs de moments de la moyenne de l'échantillon, de l'estimateur de quotients et de l'estimateur de la régression dans le cas spécial d'un sondage aléatoire simple sans remise. L'innovation présentée ici est que les calculs peuvent désormais être exécutés instantanément par ordinateur sans erreur et sans recours à des formules existantes possiblement longues et complexes. Un autre avantage immédiat est que les calculs peuvent être exécutés là où il n'existe actuellement aucune formule. Le code machine élaboré en vue de la mise en oeuvre de cette méthode est accessible par FTP anonyme au *fisher.stats.uwo.ca.*.

MOTS CLÉS: Statistiques *k*; partitions; moments de produits; estimateurs de quotient et de régression; calculs symboliques; estimation de la variance.

1. INTRODUCTION

En théorie classique de l'échantillonnage, deux problèmes généraux nous préoccupent. Il s'agit de la détermination d'un estimateur non biaisé d'un paramètre θ et du calcul des moments de $\hat{\theta}$, l'estimateur de θ.

La méthode de traitement de base des espérances et de l'estimation non biaisée consiste à effectuer, sur l'échantillon et la population, des sommes imbriquées respectivement par l'entremise des probabilités d'inclusion, c'est-à-dire les probabilités simples ou conjointes selon le cas. Une somme imbriquée est une somme couvrant l'étendue d'un ou plusieurs indices de sorte que chaque terme de la somme dépend d'indices de valeur différente. Un estimateur non biaisé d'une somme imbriquée quelconque de population est la somme imbriquée de l'échantillon associé, la quantité sous la sommation étant divisée par la probabilité d'inclusion appropriée. De même, l'espérance d'une somme imbriquée quelconque d'un échantillon est la somme imbriquée de la population associée, la quantité sous la sommation étant multipliée par la probabilité d'inclusion appropriée.

En théorie de l'échantillonnage, comme dans plusieurs autres domaines de la statistique, de nombreux calculs algébriques dépendent d'une partition quelconque. Pour ce qui est de l'échantillonnage en particulier, Wishart (1952) a montré que les calculs de moments de base dans le cadre d'un sondage aléatoire simple sans remise dépendent largement de partitions. Nous utiliserons ici des partitions pour exprimer la somme de produits de moyennes ou de totaux sous forme de combinaisons linéaires de sommes imbriquées et inversement.

Dans les résultats que nous présentons ici, nous considérons la situation dans laquelle θ et $\hat{\theta}$ peuvent être exprimés sous forme de fonctions continues de moyennes ou de totaux (population ou échantillon selon le cas). Il existe deux possibilités: la fonction continue en question peut être exprimée comme la somme de produits de moyennes ou de totaux, ou bien la fonction continue ne peut pas être exprimée de cette façon. Lorsque la deuxième possibilité s'applique, la fonction $\hat{\theta}$ est d'abord linéarisée par un développement en série de Taylor et θ est exprimé sous forme de racine d'une équation d'estimation. Nous utilisons des partitions de nombres entiers afin d'obtenir les termes de la linéarisation de Taylor d'une fonction ou pour la racine d'une fonction. Le résultat final est que θ et $\hat{\theta}$ peuvent être exprimés, exactement ou approximativement, comme la somme de produits de moyennes ou de totaux qui, à leur tour, peuvent être exprimés sous forme de combinaisons linéaires de sommes imbriquées et inversement. L'estimation de θ ou le calcul des moments de $\hat{\theta}$ représente dès lors une procédure en trois étapes: a) exprimer une équation d'estimation pour θ ou l'estimateur $\hat{\theta}$ comme la somme de produits de moyennes ou de totaux, en utilisant au besoin la linéarisation de Taylor; b) transformer l'expression obtenue à la première étape en une combinaison linéaire de sommes imbriquées et traiter ces sommes imbriquées de façon à obtenir des estimations ou des espérances non biaisées selon le cas; c) transformer les sommes imbriquées résultant de la deuxième étape de nouveau en une somme de produits de moyennes ou de totaux.

La clé de l'automatisation des résultats de la théorie de l'échantillonnage est d'avoir recours à des partitions. En général, les partitions simples comme celles d'un nombre entier aussi bien que les partitions plus complexes comme les partitions complètes sont le résultat de l'application répétée d'une règle fondamentale. Lorsque la règle est identifiée, il

[1] J.E. Stafford et D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London (Ontario), N6A 5B7.

en résulte la possibilité d'automatiser le calcul. Des formules apparemment sans lien peuvent résulter de la même règle fondamentale et un même outil d'algèbre informatique peut servir à mettre en oeuvre de nombreux calculs différents.

La notation utilisée dans le présent exposé est expliquée à la section 2. On trouvera à la section 3 une analyse des opérateurs d'espérance. La concept de partitionnement est passé en revue à la section 4 et on y trouve une règle qui donne lieu à une méthode récursive simple d'énumération des partitions. Il est question à la section 5 des partitions de nombres entiers et de la linéarisation de Taylor. La section 6 indique comment l'énumération des partitions entraîne le calcul automatique des valeurs attendues de produits de moyennes d'échantillon et de statistiques k de même que le développement d'estimateurs non biaisés de produits de moyennes de populations finies et de statistiques k. Dans cette même section, nous appliquons la méthode à l'estimation de quotients et de la régression.

L'automatisation de ces calculs et développements donne lieu à des procédures que l'on peut exécuter instantanément et sans erreur à l'aide d'un ordinateur. De même, le recours à des formules possiblement longues et complexes est éliminé. Bon nombre de calculs algébriques manuels peuvent ainsi être évités. Les codes machine de mise en oeuvre de la méthode décrite ici ont été rédigés en langage symbolique *Mathematica* 2.0 monté sur un IBM Risc 6000 muni de 64 mégaoctets de mémoire vive. Celui-ci est accessible par FTP anonyme au *fisher.stats.uwo.ca*. Même si nous utilisons *Mathematica*, il est sans doute possible de procéder dans d'autres environnements comme *Maple*, *Macsyma* ou *Reduce*. Ainsi, Kendall (1993) décrit un système, mis en oeuvre dans *Reduce*, qui permet d'identifier des expressions invariantes. On trouvera dans Kendall (1993) un aperçu complet de l'algèbre informatique en probabilité et en statistique avant 1991.

2. ÉLÉMENTS DE NOTATION

Soit une population finie de taille N. Une mesure d'intérêt y_j est prise sur chaque unité $j, j \in U = \{1, ..., N\}$. De plus, une variable auxiliaire simple x_j ou possiblement un vecteur $P \times 1$ de variables auxiliaires $\boldsymbol{x}_j$ peut être pris sur les unités. La p-ième entrée de ce vecteur $\boldsymbol{x}_j$ est x_{pj}, où $p = 1, ..., P$. Plusieurs types de paramètres de population finie peuvent être définis sur les mesures y_j, x_j, ou $\boldsymbol{x}_j$ pour $j = 1, ..., N$. Nous notons θ un paramètre d'intérêt de population finie. Il est souvent possible d'exprimer θ sous forme de fonction continue de moyennes de population finie, de moments centrés et de statistiques k. Par souci de simplicité ici nous traitons uniquement de moyennes et de statistiques k. À noter que les variances et covariances de population finie sont également des statistiques k d'ordre deux.

Les éléments de population N ne sont pas tous observés. Supposons qu'un échantillon s de taille n soit choisi dans une population U selon un schéma d'échantillonnage quelconque. Un estimateur de θ, donné par $\hat{\theta}$, est une fonction continue de moyennes d'échantillons et de statistiques k d'échantillons.

Afin d'éviter une bonne partie de la notation encombrante de sommation, nous adaptons la notation des indices de McCullagh (1987) à nos besoins. Pour tout j le vecteur $\boldsymbol{x}_j$ contient des entrées P de façon que chacune de ces variables x puisse être associée à l'un des indices P. Soit $\{i_1, ..., i_m\}$ un sous-ensemble de m de ces indices P. Dans notre adaptation de la notation de McCullagh, $x_{i_1 j}$ représente maintenant ce que nous avons appelé le vecteur $\boldsymbol{x}_j$. Les produits de ces quantités en indice deviennent des tableaux à plusieurs dimensions. Ainsi, le produit $x_{i_1 j} x_{i_2 j} x_{i_3 j}$ est un tableau tridimensionnel de dimension $P \times P \times P$.

Notons M une moyenne de population finie. L'argument de M indique la structure de l'opérande dans la moyenne. Par exemple, $M(y) = \sum_{j \in U} y_j / N$ et $M(yy)$ ou par équivalence $M(y^2) = \sum_{j \in U} y_j^2 / N$.. Dans la notation des indices, par exemple,

$$M(x_{i_1} x_{i_2} x_{i_3}) = \sum_{j \in U} x_{i_1 j} x_{i_2 j} x_{i_3 j} / N \tag{1}$$

est un tableau tridimensionnel. Un élément de ce tableau est la moyenne de produits dans l'une des permutations des éléments P pris trois à la fois dans $\boldsymbol{x}_j$ où jusqu'à trois des éléments peuvent être semblables. Le (p, q, r)-ième élément de ce tableau est $\sum_{j \in U} x_{pj} x_{qj} x_{rj}$ où $p, q, r = 1, ..., P$. La moyenne de l'échantillon est notée m de sorte que, par exemple,

$$m(x_{i_1} x_{i_2} x_{i_3}) = \sum_{j \in s} x_{i_1 j} x_{i_2 j} x_{i_3 j} / n. \tag{2}$$

Afin de réaliser des expansions asymptotiques, puisque la variance d'un estimateur donné $\hat{\theta}$ sera $O(n^{-1})$, nous définissons une variable normalisée pour $\hat{\theta}$: il s'agit de la variable originale $\hat{\theta}$ centrée par rapport à son espérance et réduite par $1/\sqrt{n}$. C'est-à-dire,

$$z(\hat{\theta}) = [\hat{\theta} - E(\hat{\theta})]\sqrt{n}. \tag{3}$$

Au besoin nous utilisons la convention de sommation de McCullagh (1987), où les indices inférieurs répétés comme indices supérieurs indiquent des sommes implicites pour cet indice. Si, par exemple, on suppose que les $\boldsymbol{x}_j$ sont des vecteurs indépendants et distribués de façon identique (i.d.i.) provenant d'une superpopulation infinie, on peut obtenir des moments de superpopulation à plusieurs variables par l'entremise de la fonction génératrice des moments et qui selon cette convention s'exprime comme suit:

$$\mathrm{FGM}(\boldsymbol{t}) = 1 + \sum_{h=1}^{\infty} \mu_{i_1 \cdots i_h} \prod_{j=1}^{h} t^{i_j} / h!, \tag{4}$$

où

$$\mu_{i_1 \cdots i_h} = \frac{\partial^h}{\partial t_{i_1} \cdots \partial t_{i_h}} \mathrm{FGM}(\boldsymbol{t})\big|_{\boldsymbol{t}=0}. \tag{5}$$

Par définition, la relation entre la fonction génératrice des moments et la fonction génératrice des cumulants est déterminée par la règle $\mathrm{FGM}(\boldsymbol{t}) = \exp\{K(\boldsymbol{t})\}$, où

$$\mathrm{K}(\boldsymbol{t}) = \sum_{h=1}^{\infty} \kappa_{i_1 \cdots i_h} \prod_{j=1}^{h} t^{i_j} / h! \tag{6}$$

est la fonction qui produit des cumulants, où

$$\kappa_{i_1 \cdots i_h} = \frac{\partial^h}{\partial t_{i_1} \ldots \partial t_{i_h}} \mathrm{K}(t)\Big|_{t=0}.$$

Les statistiques k de population finie, notées $K(\cdot)$, se définissent comme les estimateurs non biaisés (dans le cadre du modèle de superpopulation i.d.i.) des cumulants de modèle associés. Le nombre d'arguments dans K séparés par des virgules indique l'ordre de la statistique k. Ainsi, la statistique k d'ordre trois $K(x_{i_1}, x_{i_2}, x_{i_3})$ est l'estimation non biaisée basée sur le modèle de (6), où

$$K(x_{i_1}, x_{i_2}, x_{i_3}) = \frac{N}{(N-1)(N-2)} \times \sum_{j \in U} [x_{i_1 j} - M(x_{i_1})][x_{i_2 j} - M(x_{i_2})][x_{i_3 j} - M(x_{i_3})]. \quad (7)$$

Dans le cas à une variable les statistiques k de population finie sont décrites dans Wishart (1952). En particulier $K(y, y)$ et $K(y, y, y)$ dans la notation courante sont K_2 et K_3 dans la notation de Wishart (1952). Les statistiques k de l'échantillon, notées $k(\cdot)$ avec les arguments appropriés, se définissent comme les estimateurs non biaisés dans le cadre d'un sondage aléatoire simple sans remise des statistiques k de population finie associées. Comme dans Wishart (1952) la statistique k de l'échantillon peut être obtenue de la statistique k de la population si l'on remplace N par n et si l'on prend la somme sur les $j \in s$ au lieu de toutes les unités de la population finie. Par exemple,

$$k(x_{i_1}, x_{i_2}, x_{i_3}) = \frac{n}{(n-1)(n-2)} \times \sum_{j \in s} [x_{i_1 j} - m(x_{i_1})][x_{i_2 j} - m(x_{i_2})][x_{i_3 j} - m(x_{i_3})].$$

À noter que s'il n'y a pas de virgule dans la statistique k de la population ou de l'échantillon, le produit des éléments qui apparaissent ensemble est nécessaire. Par exemple, $K(xy)$ est la statistique k de population finie d'ordre un d'une nouvelle variable qui est le produit des mesures x_j et y_j pour $j = 1, \ldots, N$; $K(x, y)$ est une statistique k d'ordre deux, en particulier la covariance de population finie entre x et y.

3. OPÉRATEURS

L'opérateur d'espérance E peut être appliqué directement à toute somme imbriquée de l'échantillon de façon à fournir une somme imbriquée de population finie. De même, un estimateur non biaisé de toute somme imbriquée de population finie est une somme imbriquée de l'échantillon. Pour ce qui est des sommes imbriquées triples, par exemple,

$$E\left[\sum_{J_3 \in s} x_{i_1 j} x_{i_2 k} x_{i_3 l}\right] = \sum_{J_3=1}^{N} \pi_{jkl} x_{i_1 j} x_{i_2 k} x_{i_3 l} \quad (8)$$

et

$$\sum_{J_3=1}^{N} x_{i_1 j} x_{i_2 k} x_{i_3 l} \sim \sum_{J_3 \in s} x_{i_1 j} x_{i_2 k} x_{i_3 l} / \pi_{jkl}, \quad (9)$$

où J_3 est un ensemble d'indices (j, k, l) tel que $j \neq k \neq l$ et où π_{jkl} est une probabilité d'inclusion conjointe. Des expressions parallèles peuvent être établies pour des schémas d'échantillonnage avec remise.

À noter que m sera non biaisé pour le M associé dans le cadre d'un sondage aléatoire simple sans remise. En général, pour tout plan d'échantillonnage de taille fixe n,

$$E[m(x_{i_1} x_{i_2} x_{i_3})] = \frac{N}{n} M(x_{i_1} x_{i_2} x_{i_3} \pi)$$

et

$$M(x_{i_1} x_{i_2} x_{i_3}) \sim \frac{n}{N} m(x_{i_1} x_{i_2} x_{i_3} / \pi)$$

où $M(x_{i_1} x_{i_2} x_{i_3})$ et $m(x_{i_1} x_{i_2} x_{i_3})$ sont définis dans (1) et (2) respectivement.

La détermination de l'espérance d'un estimateur $\hat{\theta}$ ou la détermination d'un estimateur non biaisé pour le paramètre de θ peut être représentée schématiquement sous la forme

$$\sum \prod \rightarrow \sum \sum \rightarrow \sum \prod. \quad (10)$$

où $\sum\prod$ signifie la somme des produits et $\sum\sum$ signifie une somme de sommes imbriquées. Si θ ou $\hat{\theta}$ peut s'exprimer comme une quantité $\sum\prod$, c.-à-d. une somme de produits de moyennes, la détermination d'un estimateur non biaisé de θ ou de moments de $\hat{\theta}$ se réduit à suivre le schéma que l'on trouve dans (10) et à appliquer l'opérateur approprié, comme ceux qui sont donnés dans (8) ou (9), à $\sum\sum$, l'étape moyenne du schéma. Si θ ou $\hat{\theta}$ sont des fonctions continues de moyennes, mais ne se laissent pas exprimer directement comme des quantités $\sum\prod$, il faut une étape initiale avant de pouvoir appliquer le schéma dans (10). Pour $\hat{\theta}$ l'étape initiale consiste à obtenir un développement en série de Taylor. Pour θ l'étape initiale consiste à obtenir une équation d'estimation puis à résoudre cette équation pour le paramètre.

Illustrons le schéma que l'on trouve dans (10) en considérant le cas simple de la détermination de $E[\{m(x_{i_1})\}^2]$ dans le cadre d'un sondage aléatoire simple sans remise. Il s'agit d'abord d'exprimer $\{m(x_{i_1})\}^2$ sous forme de sommes imbriquées. En particulier,

$$\{m(x_{i_1})\}^2 = \frac{1}{n^2} \sum_{j \in s} x_{i_1 j}^2 + \frac{1}{n^2} \sum_{j \neq k \in s} x_{i_1 j} x_{i_1 k}. \quad (11)$$

Nous avons là l'étape $\sum\prod \rightarrow \sum\sum$. L'opérateur d'espérance peut maintenant être appliqué à $\sum\sum$. Lorsqu'on applique les probabilités d'inclusion $\pi_j = n/N$ et $\pi_{jk} = n(n-1)/[N(N-1)]$, l'opération d'espérance sur (11) donne

$$\frac{1}{n^2} \frac{n}{N} \sum_{j=1}^{N} x_{i_1 j}^2 + \frac{1}{n^2} \frac{n(n-1)}{N(N-1)} \sum_{j \neq k=1}^{N} x_{i_1 j} x_{i_1 k}. \quad (12)$$

L'étape $\sum\sum \rightarrow \sum\prod$ est maintenant appliquée. Lorsqu'on exprime la somme imbriquée qui se trouve dans (12) sous

forme de somme de produits, en particulier $\sum_{j\neq k=1}^{N} x_{i_j} x_{i_k} = \sum_{j=1}^{N} x_{i_j} \sum_{j=1}^{N} x_{i_j} - \sum_{j=1}^{N} x_{i_j} x_{i_j}$, la troisième opération donne

$$E[\{(m(x_{i_1})\}^2] = \frac{N(n-1)}{(N-1)n}\{M(x_{i_1})\}^2 + \frac{N-n}{n(N-1)} M(x_{i_1}^2). \quad (13)$$

Dans (13) $M(x_{i_1}) = K(x_{i_1})$ et $M(x_{i_1}^2) = [N/(N-1)]K(x_{i_1}, x_{i_1}) + K(x_{i_1})K(x_{i_1})$ de sorte que (13) peut s'exprimer de nouveau sous la forme

$$E(m(x_{i_1})^2) = \{K(x_{i_1})\}^2 + (N-n)K(x_{i_1}, x_{i_1})/(Nn). \quad (14)$$

De même, si l'on suit le schéma qui se trouve dans (10) les opérations qui permettent de trouver un estimateur non biaisé de, par exemple, $\{M(x_{i_1})\}^2$ ressemblent à (11), (12) et (13). L'estimande $\{M(x_{i_1})\}^2$ est exprimé dans des sommes imbriquées semblables à (11). Ces sommes seront des sommes imbriquées de population finie. Comme pour (12) on applique les probabilités d'inclusion. Dans ce cas, les sommes de population finie sont remplacées par des sommes d'échantillons et l'opérande est divisée par la probabilité d'inclusion appropriée. Enfin, comme pour (13) les sommes imbriquées de l'échantillon qui en résultent s'expriment sous forme de produits de sommes.

Chacune des opérations élémentaires qui servent à obtenir une valeur attendue par l'entremise des équations (11), (13) et (14), ou à obtenir un estimateur non biaisé, peut être effectuée à l'aide de partitions. Ces opérations sont: l'expression de sommes de produits sous forme de sommes imbriquées et inversement et l'expression de moyennes sous formes de statistiques k et inversement.

4. PARTITIONS ET PROCÉDURES FONDAMENTALES

Un aspect central de l'automatisation de tous les calculs algébriques considérés ici est la notion de partition. Le partitionnement comme point de convergence donne l'impression que les méthodes automatisées qui sont présentées ici ne sont rien de plus que la partition de nombres entiers ou la partition d'un ensemble d'indices. Nous pouvons supposer que la partition d'un nombre entier est comprise, mais une partition complète suppose une définition plus poussée.

Soit $I_m = \{i_1, ..., i_m\}$ un ensemble d'indices m. Une partition simple P_m de I_m divise les indices m en $k \leq m$ sous-ensembles ou blocs mutuellement exclusifs et exhaustifs de I_m. Nous écrivons $P_m = (b_1 \mid b_2 \mid ... \mid b_k)$, où les $b_1, ..., b_k$ sont les blocs de I_m. P_m est unique jusqu'aux permutations des indices à l'intérieur des blocs b_i. Le bloc b_i est constitué d'un sous-ensemble des indices de I_m. Des éléments à l'intérieur d'un bloc peuvent se limiter à un ordre alphabétique et les blocs eux-mêmes peuvent être ordonnés de façon que les premiers éléments de chaque bloc soient dans l'ordre alphabétique. On assure ainsi le caractère unique de la partition P_m. Dans ce cas P_m serait une partition ordonnée standard. Le fait d'ordonner les partitions de cette façon n'offre aucun avantage du point de vue des calculs et ne représente donc pas une exigence pour ce qui suit. La partition complète de I_m est l'ensemble $\wp_m$ de toutes les partitions simples P_m de I_m.

Nous pouvons maintenant cerner la partition complète de I_m par voie d'algorithme à l'aide d'une règle d'inclusion-exclusion.

i. Soit $\wp_1 = \{i_1\}$.
ii. Une règle d'inclusion-exclusion détermine la contribution à $\wp_t$ par une partition $P_{t-1} \in \wp_{t-1}$. Dans la partie inclusion de la règle, le nouvel indice i_t est ajouté tour à tour comme un élément à chacun des blocs $b_1, ..., b_k$ qui constituent P_{t-1}. Si P_{t-1} comporte k blocs, on crée k partitions pour $\wp_t$. Dans la partie exclusion de la règle, un nouveau bloc contenant l'indice simple i_t est ajouté à P_{t-1}.

Par exemple, la partition complète de $I_3 = \{i_1, i_2, i_3\}$ est fournie par les étapes

$$\begin{aligned} &\text{i.} \quad \wp_1 = \{(i_1)\} \\ &\text{ii.} \quad \wp_2 = \{(i_1 i_2), (i_1 \mid i_2)\} \\ &\text{iii.} \quad \wp_3 = \{(i_1 i_2 i_3), (i_1 i_2 \mid i_3), (i_1 i_3 \mid i_2), (i_1 \mid i_2 i_3), (i_1 \mid i_2 \mid i_3)\}. \end{aligned} \quad (15)$$

De l'étape (i) à l'étape (ii) la règle d'inclusion donne la partition $(i_1 i_2)$ et la règle d'exclusion donne $(i_1 \mid i_2)$. De l'étape (ii) à l'étape (iii) la règle d'inclusion entraîne la création des partitions $(i_1 i_2 i_3)$, $(i_1 i_3 \mid i_2)$, et $(i_1 \mid i_2 i_3)$. La règle d'exclusion donne les partitions $(i_1 i_2 \mid i_3)$ et $(i_1 \mid i_2 \mid i_3)$. Ce type de construction est facile à automatiser puisqu'il dépend d'une règle simple. On trouvera dans Stafford (1996) des renseignements détaillés sur l'automatisation de la partition des indices en partitions complètes et en partitions d'ensembles complémentaires.

Considérons, par exemple, le problème classique de l'écriture des moments suivant un modèle du vecteur aléatoire x_{i_1} en fonction de ses cumulants. Comme dans (5) nous pouvons identifier le h-ième tableau des moments en différenciant FGM(t) dans (4) h fois et en posant t égal au vecteur zéro. Le résultat est le h-ième coefficient du développement de FGM(t). Par équivalence, on peut appliquer la même opération à $\exp\{K(t)\}$. Dans ce cas, le résultat est une somme qui dépend des coefficients de $K(t)$ dans (6). Par exemple, on peut écrire les trois premiers moments sous forme de cumulants comme suit:

$$\begin{aligned} \mu_{i_1} &= \kappa_{i_1} \\ \mu_{i_1 i_2} &= \kappa_{i_1 i_2} + \kappa_{i_1}\kappa_{i_2} \\ \mu_{i_1 i_2 i_3} &= \kappa_{i_1 i_2 i_3} + \kappa_{i_1 i_2}\kappa_{i_3} + \kappa_{i_1 i_3}\kappa_{i_2} + \kappa_{i_1}\kappa_{i_2 i_3} + \kappa_{i_1}\kappa_{i_2}\kappa_{i_3}. \end{aligned}$$

Dans chaque cas donc le résultat est une somme pour les partitions complètes indiquées dans (15). Ces partitions surviennent car la règle de multiplication pour la différenciation imite la règle d'inclusion-exclusion pour l'énumération de la partition complète.

Le résultat ci-dessus est appliqué à la théorie de l'échantillonnage où nous examinons le problème de la détermination de la valeur espérée d'un produit de sommes d'échantillons. Le calcul nécessite le développement du

produit des sommes de façon à identifier les termes pour lesquels l'opérateur d'espérance de la population finie manifestera un comportement différent à cause de différences dans les valeurs des probabilités d'inclusion et des probabilités d'inclusion conjointes.

Par exemple, le produit des sommes $\sum_{j\in s}x_{i_1j}\sum_{j\in s}x_{i_2j}\sum_{j\in s}x_{i_3j}$ peut s'exprimer sous la forme

$$\sum_{j\in s}x_{i_1j}x_{i_2j}x_{i_3j}+\sum_{j\neq k\in s}x_{i_1j}x_{i_2j}x_{i_3k}+\sum_{j\neq k\in s}x_{i_1j}x_{i_2k}x_{i_3j}$$
$$+\sum_{j\neq k\in s}x_{i_1k}x_{i_2j}x_{i_3j}+\sum_{j\neq k\neq l\in s}x_{i_1j}x_{i_2k}x_{i_3l}. \quad (16)$$

Le résultat correspond à la partition complète des indices $I_3=\{i_1,i_2,i_3\}$ donnés par $\wp_3$ dans (15). L'ordre des partitions dans $\wp_3$ est le même que l'ordre donné pour les termes dans (16). Pour chaque partition dans $\wp_3$, les variables qui se trouvent dans le même bloc comportent le même deuxième indice dans le terme approprié dans (16). Par exemple, la partition $(i_1i_3\,|\,i_2)$ correspond au terme $\sum_{j\neq k\in S}x_{i_1j}x_{i_2k}x_{i_3j}$ dans (16). Chaque terme du résultat peut être identifié par une partition de I_3 et chaque partition détermine le comportement de l'opérateur de la valeur attendue.

En général, nous souhaitons un développement des produits de type $\prod_{r=1}^{m}\sum_{j\in s}x_{i_rj}$, où le produit est pris sur les éléments i_r de l'ensemble d'indices $I_m=\{i_1,\ldots,i_m\}$. Comme dans (16), le produit peut être exprimé en fonction de la partition complète de I_m. Il en est ainsi parce que la règle itérative qui régit le développement d'un produit de sommes imite la règle d'inclusion-exclusion.

Le développement des produits de sommes par l'entremise de partitions est démontrée par induction comme suit. Supposons que le produit des $t-1$ premières sommes puisse être exprimé comme une somme sur la partition complète de l'ensemble d'indices $I_{t-1}=\{i_1,\ldots,i_{t-1}\}$, en particulier

$$\prod_{r=1}^{t-1}\Big(\sum_{j\in s}x_{i_rj}\Big)=\sum_{P_{t-1}\in\wp_{t-1}}X_{P_{t-1}}. \quad (17)$$

Dans (17) le terme $X_{P_{t-1}}$ est la somme identifiée par la partition $P_{t-1}=(b_1\,|\ldots|\,b_k), k=1,\ldots,t-1$. Les blocs b_j indiquent des groupes de variables ayant le même deuxième indice et donc P_{t-1} produit un ensemble $J_k=\{j_1,\ldots,j_k\}$ de deuxièmes indices. On peut exprimer $X_{P_{t-1}}$ sous la forme

$$X_{P_{t-1}}=\sum_{j_1\neq\ldots\neq j_k\in s}\Big(\prod_{j\in J_k}X_{b_j}\Big), \quad (18)$$

où X_{b_j} est un produit de x définis par le bloc b_j ayant tous le même deuxième indice. À titre d'illustration de (18), considérons, par exemple, le troisième terme de (16). Ici $P_{t-1}=(i_1i_3\,|\,i_2)$ et $J_2=\{j,k\}$ de sorte que dans (18) la somme est prise sur $j\neq k\in s$ et les multiplicandes du produit sont $X_{b_j}=x_{i_1j}x_{i_3j}$ et $X_{b_k}=x_{i_2k}$. Pour ce qui est de la discussion générale, lorsque l'un ou l'autre des membres de (17) est multiplié par $\sum_{j\in s}x_{i_tj}$ on obtient le produit des t premières sommes. Le produit $X_{P_{t-1}}\sum_{j\in s}x_{i_tj}$ peut donc s'exprimer sous la forme

$$\sum_{j_1\neq\ldots\neq j_k\in s}\Big(\sum_{l=1}^{k}x_{i_tj_l}\prod_{j\in J_k}X_{b_j}\Big)+\sum_{j_1\neq\ldots\neq j_k\neq j_{k+1}\in s}\Big(\prod_{j\in J_k}X_{b_j}x_{i_tj_{k+1}}\Big). \quad (19)$$

Le premier terme de (19) correspond à la partie inclusion de la règle et le deuxième terme de (19) correspond à la partie exclusion de la règle. Lorsqu'on établit la somme de (19) pour tous les $P_{t-1}\in\wp_{t-1}$, le résultat est une somme pour la partition complète des t premiers indices donnés par I_t, c.-à-d. la somme pour tous les $P_t\in\wp_t$.

Lorsqu'on a procédé au développement du produit des sommes $\prod_{r=1}^{m}\sum_{j\in s}x_{i_rj}$ en une somme de sommes imbriquées, l'opérateur de valeur espérée de la population finie peut être appliqué à chaque terme de façon que l'on obtienne la valeur espérée de ce produit. La valeur attendue dans le cadre d'un sondage aléatoire simple sans remise du produit des sommes donne lieu à une somme pondérée de sommes imbriquées, chaque somme étant prise sur la population finie. Il s'agit ensuite d'évaluer ces sommes imbriquées.

En général, il s'agit d'évaluer la somme imbriquée $\sum_{J_t}Y_{J_t}$ où J_t est l'ensemble d'indices $\{j_1,\ldots,j_t\}$. La somme est prise sur tous les $j_1\neq\ldots\neq j_t$ avec chaque $j_r=1,\ldots,N$. L'opérande Y_{J_t} est le produit $x_{i_1j_1}x_{i_2j_2}\ldots x_{i_tj_t}$. Dans le cas spécial où $t=3$ où $J_3=\{j,k,l\}$ la somme imbriquée peut s'écrire sous la forme de sommes complètes:

$$\sum_{J_3}Y_{jkl}=\sum_{j\neq k\neq l=1}^{N}Y_{jkl}=\sum_{j\neq k\neq l=1}^{N}x_{i_1j}x_{i_2k}x_{i_3l}=$$
$$2\sum_{j=1}^{N}x_{i_1j}x_{i_2j}x_{i_3j}-\sum_{j=1}^{N}x_{i_1j}x_{i_2j}\sum_{j=1}^{N}x_{i_3j}-\sum_{j=1}^{N}x_{i_1j}x_{i_3j}\sum_{j=1}^{N}x_{i_2j}-$$
$$\sum_{j=1}^{N}x_{i_1j}\sum_{j=1}^{N}x_{i_2j}x_{i_3j}+\sum_{j=1}^{N}x_{i_1j}\sum_{j=1}^{N}x_{i_2j}\sum_{j=1}^{N}x_{i_3j}. \quad (20)$$

À noter que les sommes complètes de l'expression à l'extrême droite dans (20) sont le résultat de la partition complète $\wp_3$ dans (15). L'ordre des partitions dans $\wp_3$ est le même que l'ordre des termes à la droite de (20). Les indices inférieurs à la droite de (20) désignent les membres de blocs dans $\wp_3$. Par exemple, la partition $(i_1i_3\,|\,i_2)$ correspond au terme $\sum_{j=1}^{N}x_{i_1j}x_{i_3j}\sum_{j=1}^{N}x_{i_2j}$ dans (20). À noter également dans (20) que la détermination d'une somme imbriquée est rendue complexe par la détermination supplémentaire des coefficients appropriés des sommes complètes.

En général l'évaluation des sommes imbriquées de population finie est le résultat de l'application répétée de la règle

$$\sum_{j_1\neq\ldots\neq j_t=1}^{N}\Big(\prod_{r=1}^{t}x_{i_rj_r}\Big)=\sum_{j_1\neq\ldots\neq j_{t-1}=1}^{N}\Big[\prod_{r=1}^{t-1}x_{i_rj_r}\Big(\sum_{j_t=1}^{N}x_{i_tj_t}\Big)\Big]$$
$$-\sum_{j_1\neq\ldots\neq j_{t-1}=1}^{N}\Big[\sum_{l=1}^{t-1}x_{i_lj_l}\Big(\prod_{r=1}^{t-1}x_{i_rj_r}\Big)\Big]. \quad (21)$$

Cette expression imite la règle d'inclusion-exclusion selon laquelle le premier ensemble de sommes à la droite suit la partie exclusion de la règle et le deuxième ensemble suit la

partie inclusion de la règle. L'application répétée de (21) donne

$$\sum_{j_1=1}^{N}\cdots\sum_{j_t=1}^{N}\left(\prod_{r=1}^{t}x_{i_r j_r}\right)=\sum_{P_t\in\Phi_t}(-1)^{|J_t|-|P_t|}\times\left\{\prod_{b_k\in P_t}\left[(|b_k|-1)!\sum_{j=1}^{N}\left(\prod_{i_r\in b_k}x_{i_r j}\right)\right]\right\}$$

où $|J_t|$, $|P_t|$ et $|b_k|$ sont le nombre d'indices dans J_t, le nombre de blocs dans la partition simple P_t et le nombre d'éléments dans le bloc b_k respectivement.

5. PARTITIONS DE NOMBRES ENTIERS ET LINÉARISATION DE TAYLOR

Supposons que dans le cadre d'un plan d'échantillonnage quelconque un estimateur $\hat{\theta}$ d'un paramètre θ présente un certain intérêt. La méthode décrite dans les sections 2 à 4 peut être utilisée pour les calculs de moments pour $\hat{\theta}$ ou pour trouver des estimateurs non biaisés de ces moments. Ce n'est que dans les cas les plus simples que cette méthode peut être appliquée directement. Typiquement $\hat{\theta}$ doit être linéarisé de façon à devenir une fonction polynomiale de moyennes ou de sommes d'échantillons qui sont des variables aléatoires $O_p(1)$ relativement au plan d'échantillonnage. Lorsque $\hat{\theta}$ a été linéarisé de cette façon, la méthode des sections 2 à 4 est applicable.

L'objectif de la linéarisation est d'écrire $\hat{\theta}$ sous forme de développement asymptotique où les termes se suivent dans l'ordre descendant selon $1/\sqrt{n}$, plus particulièrement

$$\hat{\theta}=\hat{\theta}_0+\hat{\theta}_1/\sqrt{n}+\hat{\theta}_2/n+\ldots, \tag{22}$$

où $\hat{\theta}_i$ est le coefficient du $n^{-i/2}$ terme. Typiquement $\hat{\theta}$ est un produit de quantités qui se prête également à ce type de développement. Par exemple, si la mesure d'intérêt est y et si une variable auxiliaire x est présente, θ pourrait dès lors être $M(y)$ et l'information auxiliaire disponible est $M(x)$ de même que x_j pour $j\in s$. Dès lors $\hat{\theta}=M(x)m(y)/m(x)$, l'estimateur de quotient simple, est un produit de trois quantités $M(x)$, $m(y)$ et $1/m(x)$ ayant toutes des développements asymptotiques individuels. Le développement de $M(x)$ est elle-même. À partir de (3) le développement pour $m(y)$ donne $M(y)+z(m(y))/\sqrt{n}$. Le développement pour $1/m(x)$ est le résultat de (3) puis de l'application d'un développement en séries de Taylor à $[M(x)+z(m(x))/\sqrt{n}]^{-1}$.

En général, on peut trouver toute développement d'une fonction de régularité suffisante si l'on définit des opérateurs pour le développement d'une fonction, par exemple $g(\dot{e})$ où $\dot{e}$ est elle-même un développement. Nous recherchons le développement de fonctions de type

$$g(\dot{e})=\prod_{j=1}^{p}\dot{e}_j. \tag{23}$$

où $\dot{e}_j$ elle-même comporte le développement $\sum_{i=0}^{\infty}e_{ij}n^{-i/2}$. Au moment de linéariser $\hat{\theta}$ l'exigence de base est de définir un opérateur qui reprend $\hat{\theta}_i$ dans (22). L'efficacité de cet opérateur est tirée uniquement d'une règle pour le développement de fonctions du type donné dans (23). Les calculs requis sont des fonctions de partitions de nombres entiers. Par exemple, le terme $1/n$ du développement de $\prod_{j=1}^{3}\dot{e}_j$ est

$$e_{21}e_{02}e_{03}+e_{01}e_{22}e_{03}+e_{01}e_{02}e_{23}+e_{11}e_{12}e_{13}+e_{11}e_{02}e_{13}+e_{01}e_{12}e_{13}. \tag{24}$$

Le fait de rassembler les premiers indices pour chaque terme de la somme permet d'établir une liste dans laquelle chaque élément participe à la somme 2: {(2,0,0), (0,2,0), (0,0,2), (1,1,0), (1,0,1), (0,1,1)}. En notant que le terme d'ordre $n^{-i/2}$ dans tout développement $\dot{e}_j$ est en réalité le $(i+1)$-ième terme de la somme $\sum_{i=0}^{\infty}e_{ij}n^{-i/2}$, nous pouvons modifier la liste tirée de (24) de façon que les entrées identifient la position des termes dans une somme. La modification consiste à ajouter 1 à chaque valeur d'indice de la liste. Dans la liste tirée de (25), il en résulte un regroupement de toutes les partitions du nombre entier 5 en 3 blocs: {(3,1,1), (1,3,1), (1,1,3), (2,2,1), (2,1,2), (1,2,2)}. En général, l'i-ième terme du développement de $\prod_{j=1}^{p}\dot{e}_j$ ou $\dot{e}_j^{\,p}$, où p est un nombre entier positif, est une somme sur toutes les partitions du nombre entier $i+p$ en p blocs. Par conséquent, si l'on utilise cette méthode, il est possible de trouver tout terme du développement de l'estimateur d'un quotient, par exemple.

Illustrons cette technique à l'aide d'une estimation de quotient et de régression. L'estimateur de quotient est donné par

$$M(x)m(y)/m(x) \tag{25}$$

et l'estimateur de régression par

$$k(y)+b[K(x)-k(x)]=k(y)+\frac{k(x,y)}{k(x,x)}[K(x)-k(x)] \tag{26}$$

dans la notation de statistiques k.

Lorsqu'on utilise (3) l'estimateur de quotient (25) peut être exprimé sous la forme

$$M(x)\left[M(y)+\frac{z(y)}{\sqrt{n}}\right]\left[M(x)+\frac{z(x)}{\sqrt{n}}\right]^{-1}. \tag{27}$$

L'expression que l'on trouve dans (27) peut être formulée en fonction de (24) avec $p=3$. Le premier terme de (27) est le développement $\sum_{i=0}^{\infty}e_{i1}n^{-i/2}$ avec $e_{01}=M(x)$ et $e_{11}=e_{21}=\cdots=0$. Le premier terme entre crochets que l'on trouve dans (28) est le développement $\sum_{i=0}^{\infty}e_{i2}n^{-i/2}$ où $e_{02}=M(y)$, $e_{12}=z(m(y))$ et $e_{22}=e_{32}=\cdots=0$. Le deuxième terme entre crochets est le développement $\sum_{i=0}^{\infty}e_{i3}n^{-i/2}$ où $e_{i3}=(-1)^i\{z(m(y))\}^i/\{M(x)\}^{i+1}$. Afin d'obtenir le terme $1/\sqrt{n}$ dans le développement de (27), dans lequel cas $i=1$ et $p=3$, nous devons trouver les partitions de nombres entiers de 4 en blocs de 3. Cela donne les partitions (2,1,1), (1,2,1) et

(1,1,2). Si l'on soustrait 1 de chaque valeur d'indice dans la liste, on obtient la liste (1,0,0), (0,1,0), (0,0,1). Par conséquent le terme requis dans le développement est $(e_{11}e_{02}e_{03} + e_{01}e_{12}e_{03} + e_{01}e_{02}e_{13})/\sqrt{n}$ ou par équivalence $[z(m(y)) - M(y)z(m(x))/M(x)]/\sqrt{n}$. Le terme $1/n$ est obtenu de (24) qui se réduit à

$$[M(y)\{z(x)\}^2/\{M(x)\}^2 - z(x)z(y)/M(x)]/n.$$

L'estimateur de régression qui se trouve dans (26) peut être exprimé sous la forme

$$K(y) + \frac{z(k(y))}{\sqrt{n}} + \left[K(x,y) + \frac{z(k(x,y))}{\sqrt{n}}\right] \times \left[K(x,x) + \frac{z(k(x,x))}{\sqrt{n}}\right]^{-1}\left[\frac{z(k(x))}{\sqrt{n}}\right] \quad (28)$$

à l'aide de (3). Les termes entre crochets dans (28) peuvent donner lieu à un développement semblable à celle qui donne l'estimateur de quotient. Dans ce cas les termes des développements deviennent: $e_{01} = K(x,y)$, $e_{11} = z(k(x,y))$ et $e_{21} = e_{31} = \cdots = 0$; $e_{i2} = (-1)^i\{z(k(x,x))\}^i/\{K(x,x)\}^{i+1}$ pour $i = 0, 1, 2, \ldots$; et $e_{03} = 0$, $e_{13} = z(k(x))$ et $e_{23} = e_{33} = \cdots = 0$. Par conséquent, le terme $1/\sqrt{n}$ du développement des termes entre crochets dans (28) est

$$-\frac{K(x,y)z(k(x))}{K(x,x)\sqrt{n}}$$

et le terme $1/n$ est

$$-\frac{1}{n}\left[\frac{z(k(x,y))}{K(x,x)} - \frac{K(x,y)z(k(x,x))}{K(x,x)^2}\right]z(k(x)).$$

Ceux-ci ont été obtenus à l'aide du même argument utilisé dans l'estimateur de quotient.

6. APPLICATIONS MACHINE AU CALCUL DES VALEURS ESPÉRÉES D'UNE VARIABLE STATISTIQUE ET À L'OBTENTION D'ESTIMATEURS NON BIAISÉS

Puisque l'application machine à la méthode décrite dans les sections 3 à 5 a été effectuée en langage de programmation *Mathematica*, nous présentons une brève description de l'utilisation de *Mathematica*. Nous décrivons ensuite les opérateurs qui ont été élaborés dans *Mathematica* afin de présenter une algèbre informatique destinée à la théorie des enquêtes par échantillonnage.

La programmation *Mathematica* utilise des expressions du type $h[e_1, e_2, \ldots]$ où l'objet h est la tête de l'expression et les e sont les éléments de l'expression. Nous avons élaboré un certain nombre d'expressions machine dans *Mathematica* de type $h[e_1, e_2, \ldots]$ pour des opérateurs que nous appliquons à la mise au point d'une algèbre informatique pour l'échantillonnage. Tous ces opérateurs ont été conçus de façon que leurs arguments puissent prendre la forme de vecteurs aussi bien que de grandeurs scalaires. Il existe quatre opérateurs de base: $EV[\cdot]$ pour la valeur espérée, $Cum[\cdot]$ pour le calcul des cumulants, $UE[\cdot]$ pour l'estimateur non biaisé et $Aexp[\cdot]$ pour le développement asymptotique. Il existe également un opérateur qui sert à passer de la notation axée sur les statistiques k à une notation axée sur des moyennes et inversement.

L'opérateur de valeur espérée $EV[\cdot]$ de la statistique d'échantillonnage combine et exécute dans *Mathematica* les trois opérations de base indiquées dans le schéma de (10). $EV[\cdot]$ contient deux arguments, le premier étant l'expression pour laquelle la valeur espérée doit être obtenue et le deuxième le plan d'échantillonnage qui définit les probabilités d'inclusion. L'application dans *Mathematica* de $EV[\cdot]$ à $m(x_{i_1})m(x_{i_2})m(x_{i_3})$ dans le cadre d'un sondage aléatoire simple sans remise donne

$$K(x_{i_1})K(x_{i_2})K(x_{i_3}) + \frac{(N-n)(K(x_{i_1},x_{i_2})K(x_{i_3})}{Nn} + \frac{K(x_{i_1},x_{i_3})K(x_{i_2}) + K(x_{i_1})K(X_{i_2},x_{i_3}))}{Nn} + \frac{(N^2 - 3Nn + 2n^2)K(x_{i_1},x_{i_2},x_{i_3})}{N^2n^2}$$

dans l'expression la plus simple de la sortie. À noter que le résultat est une fonction de la partition complète de $\{i_1, i_2, i_3\}$. Si l'opérande est changé à $\{m(x_{i_1}) - M(x_{i_1})\} \times \{m(x_{i_2}) - M(x_{i_2})\} \times \{m(x_{i_3}) - M(x_{i_3})\}$, l'application de $EV[\cdot]$ donne

$$\frac{(N^2 - 3Nn + 2n^2)K(x_{i_1},x_{i_2},x_{i_3})}{N^2n^2},$$

que Nath (1968) a obtenu pour des valeurs particulières des indices i_1, i_2 et i_3. En réalité, les résultats que l'on trouve dans Nath (1968, 1969) pour les produits de trois et quatre moyennes et les résultats exacts que l'on trouve dans Raghunandanan et Srinivasan (1973) jusqu'à un produit de huit moyennes peuvent tous être reproduits automatiquement à l'aide du logiciel qui a été mis au point.

Jusqu'à présent, le plan d'échantillonnage utilisé dans chacun des exemples a été un sondage aléatoire simple sans remise. Il est possible d'obtenir des résultats dans le cadre de plans d'échantillonnage généraux. Nous illustrons ces résultats pour l'opérateur $Cum[\cdot]$ qui sert à obtenir les cumulants d'un estimateur. À noter que le deuxième cumulant pour un estimateur est également la variance. L'opérateur $Cum[\cdot]$ a trois arguments. Le premier est une expression de l'estimateur, le deuxième est l'ordre du cumulant et le troisième est le plan d'échantillonnage. Dans le cadre de plans d'échantillonnage généraux les estimateurs peuvent être exprimés sous la forme $\sum\prod$ dans le schéma donné par (10) et on peut procéder au développement de $\sum\prod$ pour obtenir $\sum\sum$, le terme moyen dans (10). Toutefois, il n'existe pas de simplification générale pour l'obtention du terme final dans (10), ce qui se laisse

illustrer avec l'estimateur de Horvitz-Thompson de $\sum\sum$ donné par $(n/N)m(y/\pi)$ dans la notation élaborée ici. L'application de l'opérateur $Cum[\cdot]$ dans le cadre d'un plan d'échantillonnage général en vue de l'obtention du troisième cumulant de l'estimateur de Horvitz-Thompson donne

$$2\frac{\left\{\sum_{i=1}^{N} y_i\right\}^3}{N^3} - 3\frac{\left\{\sum_{i=1}^{N} y_i\right\}\left\{\sum_{i=1}^{N} \frac{y_i^2}{\pi_i}\right\}}{N^3} - 3\frac{\sum_{i=1}^{N} \frac{y_i^3}{\pi_i^2}}{N^3}$$

$$-3\frac{\left\{\sum_{i=1}^{N} y_i\right\}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\pi_{ij} y_i y_j}{(\pi_i \pi_j)}\right\}}{N^3} + 3\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\pi_{ij} y_i y_j^2}{(\pi_i \pi_j^2)}}{N^3}$$

$$+\sum_{i=1}^{N}\sum_{h=1}^{N}\sum_{k=1}^{N} \frac{\frac{\pi_{ijk} y_i y_j y_k}{(\pi_i \pi_j \pi_k)}}{N^3}$$

où, par exemple, le terme π_{ii} est la probabilité d'inclusion simple π_i.

L'opérateur $Aexp[\cdot]$ a deux arguments, c'est-à-dire la fonction pour laquelle le développement est requis et l'ordre du développement. Cet opérateur est utilisé avec les opérateurs $EV[\cdot]$ ou $Cum[\cdot]$ en vue de l'obtention de cumulants ou espérances approximatives. On peut l'illustrer dans le cas de l'estimateur de régression linéaire multiple dans le cadre d'un sondage aléatoire simple sans remise. Lorsqu'il y a q covariables l'estimateur de régression résultant est donné par

$$k(y) + b_{i_1}[K(x^{i_1}) - k(x^{i_1})] \tag{29}$$

à l'aide de la notation des indices et des statistiques k. Dans (29) le coefficient b_{i_1} est le vecteur qui résulte du produit $k(x_{i_1}, y)ik(x^{i_1}, x_{i_2})$ dans la notation des indices, où le tableau $q \times q$ $ik(x_{i_1}, x_{i_2})$ est l'inverse du tableau $q \times q$ donné par $k(x_{i_1}, x_{i_2})$. De même, nous pouvons utiliser $IK(x_{i_1}, x_{i_2})$ pour désigner l'inverse du tableau de population finie $K(x_{i_1}, x_{i_2})$. La dérivation de l'erreur quadratique moyenne de (29) suppose des développements en séries de Taylor des éléments de b_{i_1} suivies des calculs de moments et de la collecte de termes appropriés. La commande *Mathematica* qui permet d'obtenir la variance approximative de (29) est obtenue en appliquant d'abord $Aexp[\cdot]$ à (29) avec 2 pour l'ordre du développement. L'opérateur $Cum[\cdot]$ est ensuite appliqué aux résultats avec les arguments suivants: le résultat du développement asymptotique comme estimateur, le sondage aléatoire simple comme plan et 2 pour l'ordre du cumulant. Cela donne

$$\frac{(N-n)K(y,y)}{Nn} + \frac{(-N+n)K(x_{i_1}, y)K(x_{i_2}, y)IK(x^{i_1}, x^{i_2})}{Nn}$$

dans la notation des indices comme sortie.

L'estimation est réalisée par l'entremise de l'opérateur $UE[\cdot]$ qui comporte deux arguments, l'estimande et le plan d'échantillonnage. Ainsi, l'application de $UE[\cdot]$ à $\{M(x)\}^2$ dans le cadre d'un sondage aléatoire simple donne

$$\frac{(Nn)\{k(x)\}^2 + (N-n)k(x,x)}{Nn}.$$

Si l'estimande ne se laisse pas exprimer comme une somme de sommes imbriquées, mais se laisse plutôt exprimer comme la racine d'une fonction estimative, $UE[\cdot]$ donne un estimateur convergent.

7. DISCUSSION AU SUJET DES TRAVAUX À VENIR

Les éléments de base de l'élaboration d'une algèbre informatique globale destinée à la théorie des enquêtes par échantillonnage ont été établis. Cette algèbre se fonde sur l'énumération de partitions. Les opérations de base dans le cadre de l'énumération des partitions englobent l'évaluation de sommes imbriquées et des développements en séries de Taylor. Ces opérations terminées, on peut calculer les espérances de statistiques d'échantillonnage ou déterminer les estimateurs non biaisés de quantités de population.

La prochaine étape du travail consiste à étendre les résultats à un degré à des sondages à plusieurs degrés et à plusieurs phases. Tant pour les sondages à plusieurs degrés que pour les sondages à plusieurs phases, le problème se réduit à une évaluation informatique de sommes multiples dans le cadre d'un opérateur d'espérance ou à la détermination d'un estimateur non biaisé de sommes multiples de population finie. Le problème des sondages à plusieurs degrés est actuellement à l'étude. On cherche également à étendre l'algèbre à des modèles de superpopulation.

Lorsque l'algèbre de base aura été mise en place, on pourra aisément étudier des problèmes de recherche mettant en cause des formules d'échantillonnage complexes du point de vue algébrique.

REMERCIEMENTS

Les auteurs tiennent à remercier David Andrews de ses discussions utiles. Les auteurs ont bénéficié de subventions du Conseil de recherches en sciences naturelles et en génie du Canada et d'un contrat de recherches de Statistique Canada.

BIBLIOGRAPHIE

ANDREWS, D.F., et STAFFORD, J.E. (1993). Tools for the symbolic computation of asymptotic expansions. *Journal of the Royal Statistical Society (B)*, 55: 613-628.

KENDALL, W.S. (1993). Computer algebra in probability and statistics. *Statistica Neerlandica*, 47, 9-25.

McCULLAGH, P. (1987). *Tensor Methods in Statistics*. New York: Chapman and Hall.

NATH, S.N. (1968). On product moments from a finite universe. *Journal of the American Statistical Association*, 63, 535-541.

NATH, S.N. (1969). More results on product moments from a finite universe. *Journal of the American Statistical Association*, 64, 864-869.

RAGHUNANDANAN, K., et SRINIVASAN, R. (1973). Some product moments useful in sampling theory. *Journal of the American Statistical Association*, 68, 409-413.

STAFFORD, J.E. (1996). A note on symbolic Newton-Raphson, submitted for publication.

STAFFORD, J.E., et ANDREWS, D.F. (1993). A symbolic algorithm for studying adjustments to the profile likelihood. *Biometrika*, 80, 715-730.

WISHART, J. (1952). Moment coefficients of the *k*-statistics in samples from a finite population. *Biometrika*, 39, 1-13.

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 13-24
Statistique Canada

Algorithmes de plan de sondage inverses

SUSAN HINKINS, H. LOCK OH, et FRITZ SCHEUREN[1]

RÉSUMÉ

Dans le travail ordinaire en statistique, l'échantillonnage est souvent exécuté en fonction d'un processus qui choisit des variables aléatoires telles qu'elles sont indépendantes et distribuées de façon identique (IDI). D'importantes techniques comme la régression et l'analyse des tableaux de contingence ont été élaborées largement dans ce contexte IDI, de sorte qu'il faut avoir recours à des rajustements pour les utiliser dans le contexte d'une enquête complexe. Toutefois, au lieu de rajuster l'analyse, les auteurs ont adopté une formulation qui a ceci de nouveau qu'elle prélève un second échantillon dans l'échantillon original. Dans ce second échantillon, le premier ensemble de sélections est inversé de façon à fournir à terme un échantillon aléatoire simple. Bien entendu, il serait inefficace d'utiliser ce processus en deux étapes pour tirer un échantillon aléatoire simple unique d'une enquête complexe normalement beaucoup plus grande, et c'est pourquoi des échantillons aléatoires simples multiples sont prélevés, les auteurs ayant élaboré une façon de fonder sur eux des inférences. Les échantillons originaux ne peuvent pas tous être inversés, mais les auteurs abordent de nombreux cas spéciaux qui couvrent tout un éventail de possibilités.

MOTS CLÉS: Échantillonnage de populations finies; inférence dans les enquêtes complexes; rééchantillonnage.

1. INTRODUCTION

L'évolution des enquêtes par échantillonnage comtemporaines est un phénomène extraordinaire (Bellhouse 1988; Hansen 1987; Kish 1995). La richesse même de cette évolution a peut-être entraîné, par contre, l'isolement des enquêtes par échantillonnage du reste du secteur statistique, l'attention y étant accordée à la richesse des modèles. En effet, il est bien connu que, dans le travail ordinaire en statistique, l'échantillonnage est souvent exécuté en fonction d'un processus qui choisit des variables aléatoires telles qu'elles sont indépendantes et distribuées de façon identique (IDI).

D'importantes techniques comme la régression et l'analyse des tableaux de contingences ont été élaborées largement dans ce contexte IDI, de sorte qu'il faut avoir recours à des rajustements pour les utiliser dans le contexte d'une enquête complexe. Des ouvrages entiers ont même été consacrés à cette question (Skinner, Holt et Smith 1989) et on y a accordé beaucoup de temps et d'efforts dans des logiciels (comme SUDAAN ou WESVAR PC) préparés expressément pour les enquêtes (voir aussi Wolter 1985). Compte tenu de tout ce qui a été accompli déjà, y a-t-il quelque chose de valable à ajouter? Nous proposons une façon de mieux traiter l'«interface» que l'on trouve actuellement entre l'IDI et la statistique des enquêtes.

Le présent exposé est divisé en quatre sections. Cette introduction est la section 1. Dans les sections 2 et 3 on trouve un énoncé général du problème et plusieurs «résolutions» pour quelques-uns des plans les mieux connus. Notre stratégie consiste à rééchantillonner l'échantillon complexe de façon à obtenir une structure des données plus facile à analyser. En particulier, nous abordons l'échantillonnage d'éléments stratifiés, les échantillons en grappes à un et deux degrés, ainsi que le plan important de deux unités primaires d'échantillonnage par strate (section 2). Puisqu'il est peu probable qu'un rééchantillonnage donné contienne toute l'information de l'enquête originale, nous vérifions ce qui se produit lorsque l'échantillon complexe original est rééchantillonné de façon répétée. On trouve également à la section 3 une illustration concrète de nos idées, tirée de notre pratique et fondée sur un échantillon SDR (statistique des revenus) hautement stratifiée de déclarations de revenus des sociétés (p. ex. Hughes, Mulrow, Hinkins, Collins et Uberall 1994). Enfin, à la section 4, nous décrivons quelques applications et les prochaines étapes qui rendront nos idées embryonnaires encore plus utiles.

2. ÉNONCÉ DU PROBLÈME ET «RÉSOLUTIONS» POSSIBLES

2.1 Raisonnement et stratégie de base

À supposer que nous voulions appliquer une procédure IDI à un échantillon d'enquête complexe. À supposer, également, que nous voulions jeter un regard neuf sur la façon de «résoudre» le problème d'interface qui survient parce que le plan d'enquête n'est pas du type IDI. Comment procéder? Il existe une expression courante qui semble résumer notre stratégie:

Si vous n'avez qu'un marteau, tout problème devient un clou.

À titre d'échantillonneurs, nous avons un marteau et c'est le processus d'échantillonnage. Pouvons-nous transformer le problème d'interface des enquêtes en un clou dont nous pourrions nous occuper à l'aide d'un autre plan de sondage?

[1] Susan Hinkins, Internal Revenue Service, Bozman, MT, U.S.A.; H. Lock Oh, Internal Revenue Service, Washington, DC, U.S.A.; Fritz Scheuren, Ernest and Young, 1402 Ruffner Rd., Alexandria, VA 22302 U.S.A.

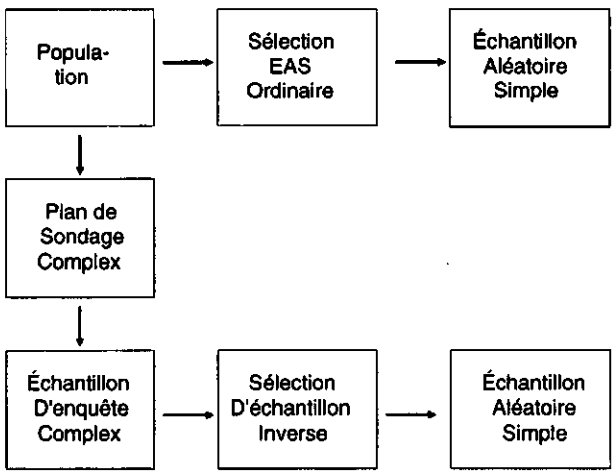
Nous sommes d'avis qu'il y a parfois lieu de répondre à cette question par un «Oui». Nous appelons ce deuxième plan de sondage un «algorithme de plan de sondage inverse», d'où le titre du présent exposé.

Un schéma peut nous aider à visualiser l'algorithme (voir la figure 1). Dans le diagramme, deux stratégies d'échantillonnage sont comparées, toutes deux donnant des échantillons aléatoires simples d'une population:

(1) Le premier plan (rangée du haut) fait appel à un processus conventionnel de sélection directe d'échantillons aléatoires simples (EAS) (p. ex., Cochran 1977), de sorte que tous les échantillons possibles d'une taille donnée comportent la même probabilité de sélection. (Ce type de plan est souvent peu commode ou inefficace ou les deux à la fois, et c'est pourquoi on ne les utilise à peu près jamais, même s'il en est question dans les manuels.)

(2) Le deuxième plan représente un processus en deux étapes. La première étape consiste à échantillonner la population d'une façon complexe qui s'appuie soigneusement sur la nature de la population et des besoins de la clientèle, les ressources de la clientèle étant utilisées sobrement (c'est là le domaine par excellence des concepteurs d'enquête).

(3) Notre formulation a ceci de nouveau qu'elle prélève un second échantillon (peut-être complexe?) qui inverse le premier ensemble de sélections de façon à fournir à terme un échantillon aléatoire simple. Bien entendu, il serait inefficace d'utiliser ce processus en deux étapes pour tirer un échantillon aléatoire simple unique d'une enquête complexe normalement beaucoup plus grande, et c'est pourquoi nous proposons la création d'échantillons aléatoires simples multiples, nos inférences étant fondées sur ceux-ci.



Bien qu'il soit possible d'élaborer, la nature fondamentale de algorithmes dont nous parlons devrait maintenant être évidente. Ils peuvent ne comporter que quatre étapes de base:

(1) Inverser, dans la mesure du possible, le plan complexe existant de façon à pouvoir créer des sous-échantillons aléatoires simples (à un degré d'approximation utile).

(2) Possiblement, appliquer le logiciel statistique conventionnel directement au sous-échantillon, puisque cela est désormais approprié.

(3) Répéter le sous-échantillonnage et l'analyse conventionnelle des étapes (1) et (2) de façon répétée.

(4) Retenir, dans la mesure du possible, le caractère du paradigme de randomisation original en utilisant la distribution des résultats du sous-échantillon comme base d'inférence (au lieu de l'échantillon complexe original).

Soulignons ce que cette stratégie est et n'est pas: tout d'abord, elle est très exigeante en traitement informatique, se fondant sur des calculs bon marché et même très bon marché. Deuxièmement, elle suppose l'existence d'algorithmes inverses pratiques (ce qui n'est pas toujours le cas). Troisièmement, elle suppose que la robustesse originale de l'échantillon tout entier peut être captée si l'on prélève un nombre suffisant de sous-échantillons, de façon qu'il n'y ait aucune perte appréciable d'efficacité. Quatrièmement, même si elle ressemble à la méthode bootstrap (Efron 1979), il ne s'agit pas de bootstrap. Il n'y a aucune intention d'imiter les sélections originales, comme il faudrait le faire pour bien utiliser la méthode bootstrap (p. ex. McCarthy et Snowden 1985; Rao et Wu 1988). Tout au contraire; notre but ici est de créer, à partir du plan original, un ensemble de sous-échantillons totalement différent et plus facilement analysable.

2.2 Définition d'un algorithme de sondage inverse

Supposons que nous voulons tirer un échantillon aléatoire simple, sans remise, d'une population finie de taille N. Supposons, également, que la population ne soit plus disponible pour l'échantillonnage, mais que nous possédions un échantillon choisi à même cette population à l'aide d'un plan de sondage D; notons cet échantillon S_D. Soit S_m, un deuxième échantillon de taille m qui pourrait être tiré de la population. Un algorithme de sondage inverse doit décrire la façon de choisir un échantillon de S_D de façon que pour tout échantillon S_m donné

$$\Pr(\text{sélection } S_m \mid S_D) * \Pr(S_m \subset S_D) = \frac{1}{\binom{N}{m}}. \qquad (1)$$

La première étape consiste à calculer la probabilité qu'un échantillon S_m arbitraire mais fixe soit contenu dans l'échantillon S_D. Bien entendu, il existe des contraintes quant à la taille de l'échantillon aléatoire simple (EAS) qui peut être prélevé de cette façon; la probabilité que S_D contienne S_m ne peut pas être zéro. Par conséquent, la taille de l'EAS ne peut certainement pas être supérieure à celle de l'échantillon S_D original et, en effet, la taille de l'EAS doit généralement être bien inférieure à celle de l'échantillon complexe original.

Il s'agit donc de trouver un algorithme général capable de tirer un EAS d'un échantillon S_D donné selon une probabilité conditionnelle correcte. Il est également nécessaire de s'assurer que l'on utilise des fonctions de probabilité valables. Les sous-sections qui suivent illustrent les algorithmes de sondage inverse pour quelques-uns des plans

de sondage les plus courants: stratifié, en grappes, à degrés multiples et à degrés multiples stratifiés. On y trouve également un exemple d'un algorithme inverse qui à prime abord ne semble pas possible.

2.3 Inversion d'un échantillon stratifié

La présente sous-section introduit un algorithme inverse pour un échantillon stratifié à quatre strates. L'algorithme permet de généraliser pour tout nombre de strates. Nous avons un échantillon stratifié ayant des tailles d'échantillon fixes n_h dans chaque strate h, et des tailles de population de strate connues, $N_1 + N_2 + N_3 + N_4 = N$. Puisqu'un échantillon donné de taille arbitraire m de la population risque d'être contenu entièrement dans une même strate, l'échantillon aléatoire simple le plus grand qui puisse être tiré d'un échantillon stratifié est de taille $m = \min\{n_h\}$.

Pour un échantillon donné S_m, notons (x_1, x_2, x_3, x_4) le nombre d'unités dans chaque strate. Chaque x_i se situe entre 0 et m et $x_1 + x_2 + x_3 + x_4 = m$. La probabilité que S_m soit contenu dans l'échantillon stratifié est égale au nombre d'échantillons stratifiés qui contiennent ces unités m divisé par le nombre total d'échantillons stratifiés possibles, c'est-à-dire

$$\Pr(S_m \subset S_D) = \frac{\binom{N_1-x_1}{n_1-x_1}\binom{N_2-x_2}{n_2-x_2}\binom{N_3-x_3}{n_3-x_3}\binom{N_4-x_4}{n_4-x_4}}{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\binom{N_4}{n_4}}. \quad (2)$$

L'algorithme qui permet de tirer un EAS de l'échantillon stratifié comporte les trois étapes que voici:

(1) Déterminer la taille des EAS qu'il s'agit de choisir: $m \le \min\{n_h\}$.

(2) Établir une réalisation $\{m_1, ..., m_4\}$ à partir d'une distribution hypergéométrique, à probabilités

$$\Pr(m_1 = i_1, m_2 = i_2, ..., m_4 = i_4) = \frac{\binom{N_1}{i_1}\binom{N_2}{i_2}\binom{N_3}{i_3}\binom{N_4}{i_4}}{\binom{N}{m}} \quad (3)$$

où $i_1 + i_2 + i_3 + i_4 = m$ et $0 \le i_1 \le m$, $0 \le i_2 \le m$, $0 \le i_3 \le m$, $0 \le i_4 \le m$.

(3) Dans chaque strate h, choisir un échantillon aléatoire simple de taille m_h, sans remise, à partir des unités d'échantillonnage n_h.

La probabilité conditionnelle de sélection de l'échantillon S_m, étant donné qu'il est contenu dans l'échantillon stratifié, est donc

$$\frac{\binom{N_1}{x_1}\cdots\binom{N_4}{x_4}}{\binom{N}{m}} \frac{1}{\binom{n_1}{x_1}\cdots\binom{n_4}{x_4}}. \quad (4)$$

La probabilité de sélection d'un échantillon S_m donné quelconque à l'aide de l'algorithme inverse est le produit des deux probabilités données dans les équations (2) et (4). On peut montrer aisément que ce produit est égal à

$$\frac{1}{\binom{N}{m}}.$$

Par conséquent, cette procédure reproduit inconditionnellement un mécanisme d'échantillonnage aléatoire simple, c.-à-d. lorsqu'il est pris sur tous les échantillons stratifiés possibles. À noter que si l'on veut tirer tous les EAS possibles de cette population, il faut répéter toute la séquence à partir du choix d'un échantillon stratifié et poursuivant la démarche jusqu'aux étapes 1 - 3.

2.4 Inversion d'un échantillon en grappes à un degré

Dans la présente sous-section, nous envisageons trois cas spéciaux. Pour commencer, nous examinons des échantillons en grappes dont les grappes sont de taille égale. Nous abordons ensuite le cas plus fréquent où les grappes sont de taille inégale. Dans l'un et l'autre contextes, nous supposons que les grappes sont prélevées à l'aide d'un mécanisme d'échantillonnage aléatoire simple et sans remise. Le troisième cas étudié est celui de grappes inégales échantillonnées à l'aide d'un mécanisme de probabilité proportionnelle à la taille (PPT). Dans ce dernier cas, nous supposons un échantillonnage avec remise.

2.4.1 Échantillonnage en grappes à un degré à tailles de grappes égales et à probabilités égales

Supposons que nous ayons une population de N grappes où toutes les grappes sont de taille M, k d'entre elles étant choisies par un mécanisme d'échantillonnage aléatoire simple sans remise.

Si l'on veut construire un algorithme inverse, il faut décider ce que pourrait être le plus grand sous-échantillon d'éléments. Il est évident que le plus grand EAS d'éléments que l'on peut choisir est k. Soit dit en passant, la taille des grappes n'est pas une contrainte pour ce qui est de la taille du sous-échantillon.

Pour un échantillon S_k donné, notons q le nombre de grappes représenté dans S_k; $0 < q \le k$. Dès lors la probabilité que S_k soit contenu dans l'échantillon en grappes est égale au nombre d'échantillons en grappes qui contiennent ces grappes q divisé par le nombre total d'échantillons en grappes possibles, c'est-à-dire

$$\Pr(S_k \subset S_D) = \frac{\binom{N-q}{k-q}}{\binom{N}{k}}. \quad (5)$$

Comme pour l'échantillon stratifié, l'algorithme détermine d'abord le nombre d'unités à choisir dans chaque grappe, $(m_1, m_2, ..., m_k)$. La distribution de probabilités à utiliser pour choisir les m_i est

$$\Pr(m_1=i_1,\ldots,m_k=i_k)=\frac{\binom{M}{i_1}\cdots\binom{M}{i_k}}{\binom{NM}{k}} * \frac{N(N-1)\ldots(N-q+1)}{k(k-1)\ldots(k-q+1)} \quad (6)$$

où $0 \le i_j \le k$, $i_1 + i_2 + \ldots + i_k = k$, et q est le nombre de i_j non zéro. Ainsi, avec $M = 100$, $N = 6$, $k = 3$ on a

$$\Pr(m_1=1,m_2=0,m_3=2) = \frac{\binom{100}{1}\binom{100}{0}\binom{100}{2}}{\binom{600}{3}} * \frac{6*5}{3*2}$$

$$\Pr(m_1=3,m_2=0,m_3=0) = \frac{\binom{100}{3}}{\binom{600}{3}} * \frac{6}{3}.$$

Une fois les m_i déterminés, un échantillon aléatoire simple de taille m_i est choisi dans la grappe $i, i = 1, 2, \ldots, k$. Par conséquent, la probabilité conditionnelle de sélection de S_k est

$$\Pr(\text{sélection } S_k \mid S_D) = \frac{1}{\binom{NM}{k}} * \frac{N(N-1)\ldots(N-q+1)}{k(k-1)\ldots(k-q+1)}. \quad (7)$$

La probabilité de sélection d'un échantillon S_k particulier est obtenue en multipliant l'équation (5) par l'équation (7). Il est facile de vérifier que cette opération donne la probabilité correcte de sélection d'un EAS.

Contrairement à l'exemple stratifié, où la fonction de sélection des valeurs de m_i était une fonction de probabilité connue, il n'est pas immédiatement évident que l'équation (6) décrit une distribution de probabilités. Puisque les valeurs établies à l'aide de cette fonction sont toutes non négatives, il suffit de montrer que leur somme égale un sur l'ensemble des valeurs possibles. Le premier facteur de l'équation se présente sous forme d'une distribution hypergéométrique, sauf que le numérateur se limite à k parmi les N grappes, tandis que le dénominateur reflète les N grappes totales. Il est utile de définir une partition de k comme combinaison de nombres entiers positifs dont l'addition donne k, sans égard à l'ordre. Ainsi, les partitions de $k = 3$ sont {3}, {1,2} et {1,1,1}. Puisque les grappes sont toutes de la même taille, M, tous les motifs de sélection qui correspondent à la même partition sont également probables. Prenons, à titre d'exemple, $N = 6$ et $k = 3$. Dans la pleine distribution hypergéométrique, à taille de grappes égale, chacune des combinaisons ci-dessous a les mêmes chances de se produire:

(0,0,0,0,1,2), (0,0,0,0,2,1), (0,0,0,1,2,0), ..., (2,1,0,0,0,0).

Le nombre total de combinaisons de ce genre est $N(N-1)\ldots(N-q+1)$, où q est la taille de la partition, c'est-à-dire le nombre de valeurs (non zéro) dans la partition. Dans l'exemple ci-dessus, $q = 2$. Pour une partition donnée, si le nombre de valeurs non zéro peut seulement être placé dans k cellules particulières, on a dès lors $k(k-1)\ldots(k-q+1)$ ordonnancements de ce genre. Par conséquent, la sommation de la distribution sur toutes les valeurs de $(i_1, \ldots, i_k)$ peut se faire en prenant d'abord la somme sur toutes les partitions de k et ensuite pour chaque partition, quitte à prendre la somme sur tous les ordonnancements possibles de cette partition en k cellules. Puisque tous les ordonnancements associés à une partition particulière ont les mêmes chances de se produire, il en résulte une sommation qui correspond à la sommation de la distribution hypergéométrique sur l'espace correct et, par conséquent, la somme de l'expression (6) donne un.

La distribution de probabilités requise pour ce plan de sondage simple en grappes (équation 6) est appréciablement plus difficile à produire que la distribution hypergéométrique dans le cas de l'échantillon stratifié. Toutefois, à mesure que la fraction de sondage k/N diminue, la probabilité est souvent contenue dans deux des partitions seulement: $q = k$ et $q = k - 1$. (Ces probabilités sont calculées à l'annexe.) La probabilité peut même être concentrée uniquement dans le motif avec $q = k$ (on trouvera également à l'annexe un cas spécial de ce phénomène).

Compte tenu des résultats de l'annexe, il est peut-être possible de se rapprocher de l'inverse exact en choisissant un cas dans chaque grappe, en faisant appel à un sondage systématique dans l'échantillon en grappes original. Cette stratégie a une valeur réelle car les calculs de distribution de probabilités deviennent peu maniables à mesure que le nombre de grappes dans l'échantillon augmente. Pour qu'un inverse systématique réussisse, toutefois, l'«étape» doit bien sûr être au moins aussi grande que M sinon plus grande, suivant le nombre de grappes dans la population. Si l'on voulait répéter ce genre de sous-échantillonnage pour chaque inverse systématique d'échantillon, les unités à l'intérieur de chaque grappe seraient ordonnées de nouveau de façon aléatoire avant la prochaine sélection et les grappes triées de nouveau, également de façon aléatoire, après quoi on aurait un autre départ aléatoire avant de passer de nouveau à travers l'échantillon original.

2.4.2 Échantillonnage en grappes à un degré à grappes inégales et à probabilités égales

L'algorithme de sondage inverse pour un échantillon de grappes de taille égale ne se laisse pas généraliser aisément lorsqu'on prélève un échantillon de grappes de taille inégale. Il en est ainsi même s'il semble facile de généraliser cette stratégie d'une façon évidente. En particulier, il ne semble pas difficile de généraliser la méthode précédente de façon que l'on puisse multiplier les «probabilités» et produire la probabilité de sélection «correcte», c'est-à-dire

$$\frac{1}{\binom{M_+}{k}}, \quad \text{où} \quad M_+ = \sum_1^N M_i. \quad (8)$$

Toutefois, le fait de généraliser en fonction de tailles de grappes inégales M_i en choisissant m_i sous forme de

$$\Pr(m_1=i_1,\ldots,m_k=i_k)=\frac{\binom{M_1}{i_1}\cdots\binom{M_k}{i_k}}{\binom{\sum_1^N M_i}{k}}*\frac{N(N-1)\ldots(N-q+1)}{k(k-1)\ldots(k-q+1)} \quad (9)$$

n'entraîne pas une distribution de probabilités valable. Soit dit en passant, nous supposerons de nouveau que les grappes originales subissent un échantillonnage à probabilités égales et sans remise, comme à la sous-section 2.4.1. Ensuite (sous-section 2.4.3), nous examinerons des échantillons originaux qui se prêtent à une forme quelconque de sélection PPT (probabilité proportionnelle à la taille).

Pour constater qu'il n'est pas facile de simplement généraliser l'équation (6) sous la forme de l'équation (9), considérons le contre-exemple suivant où la «probabilité» calculée à l'aide de (9) est supérieure à un. Supposons $N = 4$ et des tailles de grappes $M_1 = 4$, $M_2 = 6$, $M_3 = 8$, et $M_4 = 10$. Supposons également que nous prélevons un échantillon en grappes avec $k = 2$ et que, simplement par hasard, les deux grappes retenues sont les plus grandes, c'est-à-dire $M_3 = 8$ et $M_4 = 10$. Il est évident qu'avec ces sélections, l'équation (9) permettrait de produire une probabilité de sélection d'une unité dans chaque grappe qui serait supérieure à un.

Cette difficulté peut-elle être surmontée? Oui, bien que non pas, sans doute, de façon tout à fait satisfaisante. Une méthode consiste à employer une distribution hypergéométrique qui suppose des grappes qui sont toutes aussi grandes que la plus grande grappe de la population. L'inconvénient, c'est que la taille de l'échantillon inverse obtenue n'est plus fixe, et que le sous-échantillon qui en résulte n'est un EAS que conditionnellement, compte tenu de la taille de l'échantillon atteinte, désignée par exemple k_0. Autrement dit, pour un échantillon donné de taille k_0, $k_0 \le k$, tous les échantillons de taille k_0 ont les mêmes chances d'être choisis à l'aide de l'algorithme inverse.

Soit M_* la taille maximale des grappes, $M_* = \text{Max}\{M_1, M_2, \ldots, M_N\}$. Il s'agit de créer une population en remplissant chacune des grappes originales d'unités «fictives» ou de paramètres substituables, $j = M_i + 1, M_i + 2, \ldots, M_*$. Dès lors, grâce à une méthode semblable à celle de Lahiri (1951) pour l'échantillonnage PPT, l'algorithme inverse permet de choisir des unités de la population constituées de N grappes de taille M_* chacune, puis d'écarter tout élément qui ne se trouve pas dans la «sous-population» constituée des grappes originales de taille M_i.

Plus particulièrement, étant donné un échantillon en grappes constitué de k grappes, il s'agit de choisir le vecteur $\boldsymbol{m}$ de la distribution de probabilités

$$\Pr(m_1 = i_1, \ldots, m_k = i_k) = \frac{\binom{M_*}{i_1}\binom{M_*}{i_2}\cdots\binom{M_*}{i_k}}{\binom{NM_*}{k}} * \frac{N(N-1)\ldots(N-q+1)}{k(k-1)\ldots(k-q+1)} \quad (10)$$

où la somme des constituants de $\boldsymbol{m}$ est k et q des constituants m_i ne sont pas zéro. Nous avons là une distribution de probabilités convenable. Étant donné la valeur choisie de m_i, il s'agit de choisir un échantillon aléatoire de m_i unités de la grappe i, où la grappe contient M_i unités de la population et $M_* - M_i$ «paramètres substituables». On écarte toutes les unités choisies qui sont des paramètres substituables, dans l'ensemble de $j = M_i + 1, M_i + 2, \ldots, M_*$. Par conséquent, la taille finale de l'échantillon n'est pas nécessairement égale à k, mais peut lui être inférieure, par exemple k_0.

L'échantillon qui en résulte est conditionnellement un EAS de la population, en ce sens que pour une valeur donnée de k_0, tous les échantillons de taille k_0 ont les mêmes chances d'être choisis à l'aide de cet algorithme inverse. Pour le constater, il suffit de continuer à considérer le problème comme un cas de sous-population, P, de N grappes de taille M_i, $i = 1, \ldots, N$, à l'intérieur d'une population P_* de N grappes, chacune de taille M_*. À noter que, pour tout échantillon, S_*, de taille k tiré de la population P_*, la probabilité de sélection de S_* à l'aide de l'algorithme inverse est la suivante

$$\frac{1}{\binom{NM_*}{k}}. \quad (11)$$

Si $k_0 = k$, c'est là la probabilité de sélection de cet échantillon à l'aide de l'algorithme inverse. Pour une valeur $k_0 < k$, fixe, notons S_0 tout échantillon donné de taille k_0 contenu dans P. Nous pouvons produire un échantillon S_* contenant S_0 en commençant par S_0 et en y ajoutant $k - k_0$ éléments des $N^*M_* - M_*$ paramètres substituables dans P_*. Le nombre d'échantillons S_* de ce type, entraînant une sélection de S_0, est le suivant:

$$\binom{NM_* - M_*}{k - k_0} \quad \text{où} \quad M_* = \sum_{i=1}^{N} M_i. \quad (12)$$

Par conséquent, la probabilité de sélection de S_0 à l'aide de l'algorithme inverse est égale à la probabilité de sélection de S_* à l'aide de l'algorithme inverse, donnée dans (11), avec sommation sur tous les échantillons S_* construits selon la description ci-dessus, où le nombre d'échantillons de ce genre est donné par (12). Cette probabilité est égale à:

$$\frac{\binom{NM_+ - M_+}{k - k_0}}{\binom{NM_+}{k}}$$

et tous les échantillons de taille k_0 ont la même chance d'être choisis à l'aide de l'algorithme inverse.

Il existe malheureusement une probabilité positive que cette stratégie entraîne la sélection d'un échantillon sans éléments. Cela pourrait se produire s'il y avait une différence importante dans la taille des grappes. Toutefois, lorsque le nombre de grappes k dans l'échantillon original est grand, il est peu probable que cela soit un problème.

Comme dans le cas des tailles de grappes égales, on a accès à une approximation en utilisant un sous-échantillon systématique comme inverse. Cette fois-ci, nous souhaitons avoir un pas d'échantillonnage au moins aussi grand que la taille de grappe maximale. Soit dit en passant, le recours à un inverse systématique offrirait l'avantage d'un meilleur contrôle de la taille réelle du sous-échantillon prélevé.

2.4.3 Échantillonnage en grappes à un degré, à grappes inégales, à probabilités inégales

Si l'on choisit un échantillon de k grappes avec PPT, il est possible qu'il existe un algorithme inverse. Supposons que les échantillons sont choisis avec remise d'une population constituée de N grappes, la taille des grappes $M_1, M_2, ..., M_N$. étant inégale. Supposons également que la mesure de la taille est soit égale à M_i ou proportionnelle à M_i. Dès lors, à chaque prélèvement, on a:

$$\Pr(\text{sélection grappe } j) = \frac{M_j}{M_+} \quad \text{où } M_+ = \sum_{i=1}^{N} M_i. \tag{13}$$

Enfin, puisque l'on prélève un échantillon à un degré, une fois la grappe j choisie toutes les unités M_j de cette grappe sont comprises dans l'échantillon.

Un algorithme inverse dans ce cas-ci devrait entraîner un EAS avec remise. Autrement dit, pour tout vecteur $\boldsymbol{S}$ résultant de k sélections indépendantes de la population, la probabilité de sélection du vecteur ordonné est la suivante:

$$\Pr(\text{sélection } \boldsymbol{S}) = \left(\frac{1}{M_+}\right)^k. \tag{14}$$

Un algorithme inverse consiste à simplement choisir au hasard une unité de chaque grappe dans l'échantillon en grappes. Puisque les grappes ont été choisies avec remise, il convient de considérer les grappes prélevées comme étant ordonnées, dans l'ordre de leur sélection, ou dans un ordre fixe quelconque. Si donc la population contient 20 grappes, un échantillon en grappes possible de taille $k = 5$ est (7, 5, 7, 18, 6), *etc.*

La population est constituée de M_+ unités, notées $u_1, u_2, ..., u_{M+}$. Soit S, un échantillon donné avec remise, $S = (s_1, s_2, ..., s_k)$, soit $\boldsymbol{c} = (c_1, c_2, ..., c_k)$ la grappe associée pour chaque unité. Par exemple, supposons la population suivante:

Unités	Grappe
1	$u_1\ u_2\ u_3\ u_4$
2	$u_5\ u_6\ u_7\ u_8$
3	$u_9\ u_{10}\ u_{11}$
4	$u_{12}\ u_{13}\ u_{14}$
5	$u_{15}\ u_{16}\ u_{17}$
6	$u_{18}\ u_{19}\ u_{20}$

et $k = 3$. Dès lors l'échantillon $(s_1 = u_2, s_2 = u_4, s_3 = u_{17})$ correspond à $\boldsymbol{c} = (1, 1, 5)$. L'échantillon $(s_1 = u_{18}, s_2 = u_{19}, s_3 = u_{18})$ correspond à $\boldsymbol{c} = (6, 6, 6)$. À noter que ce deuxième échantillon ne peut être choisi que si la grappe 6 est la seule grappe choisie dans l'échantillon en grappes.

Pour un échantillon S donné de taille k, et pour le vecteur $\boldsymbol{c}$ correspondant de membres, de la grappe, la probabilité inconditionnelle de sélection de S à l'aide de l'algorithme inverse est la suivante:

$$\Pr(\text{sélec. } S \mid \text{échan. en gr. } \boldsymbol{c}) * \Pr(\text{sélec. } \boldsymbol{c}) = \left(\prod_{i=1}^{k} \frac{1}{M_{c(i)}}\right)\left(\prod_{i=1}^{k} \frac{M_{c(i)}}{M_+}\right) \tag{15}$$

qui est égale à la probabilité souhaitée, équation (14).

À noter que ce même algorithme inverse est valable lorsque k grappes sont choisies selon une probabilité proportionnelle à la taille avec remise, mais qu'un échantillon de taille fixe m est choisi (échantillon aléatoire simple sans remise) de la grappe retenue, en supposant que $M_i > m$ pour toutes les grappes i.

2.4.4 Quelques remarques au sujet des plans de sondage à un degré

Nous avons vu que, avec un peu de prudence, il est possible de construire des algorithmes inverses pour plusieurs cas particuliers dans lesquels l'échantillon original comporte un plan en grappes à un degré. Deux de nos résultats s'appliquent à des échantillons en grappes tirés à probabilités égales sans remise. Le troisième est un plan de type probabilité proportionnelle à la taille avec remise.

Il est même possible qu'un inverse systématique pratique soit valable à titre d'approximation de l'algorithme inverse correct lorsque nous avons un échantillon en grappes. L'approximation est valable lorsqu'on utilise un échantillon aléatoire simple avec remise qui «se rapproche» d'un échantillon aléatoire simple sans remise, c'est-à-dire, dans notre notation, lorsque k/NM est très petit de façon que $1/(NM - k + 1)$ soit à peu près égal à $1/NM$. Tout semble donc être intuitivement uniforme pour l'ensemble des cas étudiés.

De nombreux plans en grappes ne se rapportent à aucun des cas particuliers examinés. Pour quelques-uns d'entre eux, nous supposons qu'il n'existe peut-être pas d'algorithmes inverses exacts. En particulier, le cas général d'un échantillonnage de type probabilité proportionnelle à la taille sans remise semble être un tel cas, y compris la variante souvent utilisée d'une probabilité proportionnelle à la taille sans remise systématique. Cela peut être ou ne pas être un problème pour les praticiens qui utilisent souvent l'hypothèse (normalement) prudente qu'il s'agit d'un échantillonnage avec remise; dans ce cas il existerait un algorithme inverse selon le même ordre d'approximation supposé dans l'estimation des variances.

2.5 Plans de sondage en grappes à plusieurs degrés

Qu'en est-il des plans de sondage à plusieurs degrés? Peut-on les inverser? Dans certains cas, nous croyons que la réponse est oui. Trois plans de sondage seront considérés: (1) un plan à deux degrés avec échantillonnage aléatoire simple aux degrés 1 et 2 (sous-section 2.5.1); (2) un plan faisant appel à un échantillonnage de type probabilité proportionnelle à la taille pour le premier degré et un échantillonnage aléatoire simple pour le deuxième (sous-section 2.5.2); (3) le plan de sondage très important à degrés multiples stratifiés avec deux unités primaires d'échantillonnage par strate, qui mérite au moins un bref aperçu.

Comme nous le verrons, il est possible d'élargir assez facilement les résultats stratifiés et à un degré. Pour le confirmer, notre stratégie de base consiste à appliquer de façon répétée les méthodes déjà décrites.

2.5.1 Plans de sondage à plusieurs degrés avec échantillonnage aléatoire simple pour les deux degrés

Supposons, tout d'abord, qu'à l'origine un échantillon aléatoire simple (EAS) de k grappes, toutes de taille M, a été prélevé au premier degré et qu'un sous-échantillon aléatoire simple de taille «r» a été prélevé au deuxième degré, dans chaque grappe retenue au premier degré.

Comme antérieurement, notre échantillon inverse ne peut pas être plus grand que k. Supposons d'abord que $1/(NM - k + 1)$ est à peu près égal à $1/NM$ et nous pouvons dès lors utiliser un algorithme inverse de type échantillon aléatoire simple avec remise, puisque EASar et EASsr sont très proches l'une de l'autre. À l'aide des résultats de la sous-section 2.4.3, nous prélevons un EASar de k grappes et ensuite, dans chaque grappe retenue, nous prenons une observation au hasard. Une autre possibilité serait, comme à la sous-section 2.4.1, de déterminer d'abord le nombre d'unités à choisir dans chaque grappe, $(m_1, m_2, ..., m_k)$. Une fois les m_i déterminés, un échantillon aléatoire simple sans remise de taille m_i est choisi dans la grappe $i, i = 1, 2, ..., k$. Il peut s'agir d'un résultat presque exact, sauf qu'il est possible que l'échantillon inverse de deuxième degré de taille m_i soit plus grand que l'échantillon original de deuxième degré de taille «r». Si cela se produit, nous pouvons tout de même faire appel aux résultats de la sous-section 2.4.2 et prélever notre échantillon de deuxième degré avec des «paramètres substituables». Dans ce deuxième cas, l'échantillon réel obtenu ne serait plus fixe, mais il resterait conditionnellement un échantillon aléatoire simple. Si les grappes de premier degré sont de taille inégale mais prélevées avec remise, nous pouvons encore une fois avoir recours à la stratégie utilisée à la sous-section 2.4.2 et créer des «paramètres substituables». La taille des échantillons est aléatoire et nous n'obtenons un inverse de type EAS que conditionnellement.

Une autre façon de procéder est de noter que l'échantillon aléatoire simple le plus grand qui peut être choisi à l'aide d'un algorithme inverse est de taille $k_0 = \min\{k, r\}$. Il s'agit d'abord de déterminer le nombre d'unités à choisir dans chaque grappe, $(m_1, m_2, ..., m_k)$, où la somme des m_i doit maintenant être k_0 au lieu de k. Une fois les m_i déterminés, un échantillon aléatoire simple de taille m_i est choisi dans la grappe $i, i = 1, 2, ..., k$. La distribution de probabilités à utiliser pour choisir les m_i est la suivante:

$$\Pr(m_1 = i_1, ..., m_k = i_k) = \frac{\binom{M}{i_1} ... \binom{M}{i_k}}{\binom{NM}{k_0}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)}$$

où $0 \le i_j \le k_0$, $i_1 + i_2 + ... + i_k = k_0$, et q est le nombre de i_j non zéro.

Notons enfin, pour les grappes de taille tant égale qu'inégale, qu'il semble exister la possibilité d'un inverse systématique approximatif, moyennant bien sûr, plus ou moins, les mêmes réserves décrites ci-dessus.

2.5.2 Plans de sondage à plusieurs degrés avec échantillonnage PPT au premier degré et EAS au deuxième degré

Encore une fois, notre échantillon inverse ne peut pas être plus grand que k. Il est évident qu'une façon de construire un inverse serait d'utiliser les résultats de la sous-section 2.4.3. Plus particulièrement, nous prélèverions un échantillon aléatoire simple avec remise de k grappes et, ensuite, une observation au hasard dans chaque grappe retenue. Il est possible que d'autres algorithmes inverses existent également. Un inverse systématique semble raisonnable, pourvu que la probabilité de sélection de la même grappe plus d'une fois soit faible à presque nulle.

2.5.3 Plans de sondage à plusieurs degrés stratifiés avec deux unités primaires d'échantillonnage par strate

Est-il possible d'inverser des plans à deux UPÉ (unité primaire d'échantillonnage)? Notre réponse est oui, si les sélections intra-strate se font de l'une ou l'autre façon discutée en détail antérieurement. C'est essentiellement le seul cas que nous abordons.

Compte tenu de nos résultats des sous-sections 2.3 et 2.4, il est évident que pour qu'il existe un inverse, la taille de l'échantillon m ne saurait être supérieure à $m = 2$. Suivant

l'échantillonnage à l'intérieur de chaque strate, nous pourrions employer un ou plusieurs des inverses exacts ou approximatifs de façon à obtenir deux sélections EAS à l'intérieur de chaque strate. Afin d'obtenir un échantillon aléatoire simple global, nous utiliserions l'algorithme inverse de la sous-section 2.3 pour ces deux sélections et, en fin de compte, nous n'aurions que deux sélections globalement.

2.5.4 Quelques remarques au sujet des plans de sondage à plusieurs degrés

Dans la présente sous-section, nous avons rapidement abordé quelques plans de sondage à plusieurs degrés et nous avons fourni des inverses exacts ou approximatifs. Les résultats ont été obtenus grâce à des résultats antérieurs des sous-sections 2.3 et surtout 2.4. Bien entendu, de nombreux plans de sondage à plusieurs degrés ne correspondent à aucun des cas spéciaux examinés, notamment ceux qui comportent des sélections systématiques au dernier degré.

De nombreux lecteurs se demanderont peut-être comment une méthode qui choisit uniquement un échantillon de taille deux (comme nous l'avons fait à la sous-section 2.5.3) peut être de quelque utilité pratique. La prochaine section apportera peut-être des éclaircissements.

3. RÉÉCHANTILLONNAGE EN VUE D'UNE PUISSANCE ACCRUE

3.1 Contexte général

Le prélèvement d'un échantillon aléatoire simple unique et plus petit à partir d'un échantillon plus complexe et plus grand peut convenir dans certaines situations. Toutefois, pour la plupart des usagers, la perte de puissance entre l'estimation qui se fonde sur l'échantillon complexe et l'estimation qui se fonde sur un échantillon aléatoire simple est inacceptable.

Afin d'augmenter la puissance de notre stratégie, nous avons tout naturellement considéré des techniques de rééchantillonnage. La taille des échantillons aléatoires simples qu'il est possible de prélever est limitée, mais nous pouvons répéter le processus. En répétant la procédure de sous-échantillonnage dans son ensemble, nous pouvons produire g échantillons aléatoires simples ayant chacun la taille m, où chaque échantillon aléatoire simple est choisi indépendamment à même l'échantillon original global. Chaque répétition doit inclure toutes les étapes de la procédure de sous-échantillonnage. Ainsi, dans le cas stratifié, la taille des sous-échantillons de strate doit être redéfinie à l'aide de la distribution hypergéométrique.

Dans la présente section, nous indiquons des conditions dans lesquelles la précision des estimations fondées sur des échantillons aléatoires simples multiples peut être rapprochée arbitrairement de la précision des estimations originales. Pour commencer, définissons notre notation.

Soit D, tout plan de sondage inversible (p. ex. un plan du type abordé à la section 2). Soit T, la quantité de la population qui nous intéresse (p. ex. un total de population); soit T_D, un estimateur non biaisé de T calculé à partir de l'échantillon $\boldsymbol{S}_D$. Supposons g échantillons aléatoires simples qui sont prélevés indépendamment de l'échantillon $\boldsymbol{S}_D$ donné et notons t_i l'estimateur de l'i-ième échantillon aléatoire simple. On peut dès lors montrer que

$$\text{si } E(t_i \mid \boldsymbol{S}_D) = T_D$$

$$\text{alors } \operatorname{Var}\left(\frac{1}{g}\sum_{i=1}^{g} t_i\right) = \operatorname{Var}(T_D) + \frac{1}{g}(\operatorname{Var}(t_1) - \operatorname{Var}(T_D)).$$

Preuve: Puisque les g répétitions du processus d'échantillonnage aléatoire simple sont conditionnellement indépendantes, nous avons

$$\text{pour } i \neq j,\ E(t_i t_j \mid \boldsymbol{S}_D) = T_D^2.$$

Par conséquent, inconditionnellement, pour i non égal à j,

$$\begin{aligned}\operatorname{Cov}(t_i, t_j) &= E(t_i t_j) - T^2 \\ &= \operatorname{Var}(T_D).\end{aligned}$$

Et le résultat suit directement.

Quelques-unes des conditions de cette preuve peuvent être assouplies; si T_D est biaisé, on peut obtenir des résultats semblables pour l'EQM au lieu de la variance. Toutefois, la condition

$$E(t_i \mid \boldsymbol{S}_D) = T_D$$

est nécessaire. Or cette condition n'est pas satisfaite pour les estimateurs de rapport. Par contre, si la condition est satisfaite séparément pour le numérateur et pour le dénominateur de l'estimateur de rapport, et si la taille finale de l'échantillon combiné est suffisamment grande pour qu'une approximation de la série de Taylor soit acceptable, il est possible de trouver des résultats semblables pour des approximations de la variance pour des rapports de la façon normale. Soit dit en passant, même dans le plan de sondage de deux unités primaires d'échantillonnage par strate, cette stratégie est valable, pourvu que nous puissions obtenir une estimation non biaisée de chaque échantillon individuel de taille 2. Pour des estimations de totaux, ce peut être le cas, à supposer que l'on puisse construire un inverse à chaque degré d'échantillonnage.

3.2 Estimation de l'erreur d'échantillonnage pour des moyennes ou des totaux

Par voie de rééchantillonnage, il est possible d'obtenir presque la même précision que l'estimateur du plan original. Mais puisque les échantillons aléatoires simples rééchantillonnés ne sont que conditionnellement indépendants, l'estimation de l'erreur-type n'est pas aussi simple que si un seul échantillon aléatoire simple avait été prélevé. Toutefois, l'estimation demeure relativement simple.

Soit S^2, la variance de la population pour la variable X; soit T, son total de population. Pour les moyennes, les totaux et les variances d'échantillons calculés à partir des échantillons aléatoires simples produits, notons

$$t_{..} = \frac{1}{g}\sum_{j=1}^{g} t_j = \frac{1}{g}\sum_{j=1}^{g} N\bar{x}_j = \frac{1}{g}\sum_{j=1}^{g}\frac{N}{m}\sum_{i=1}^{m} x_{ji}$$

$$s_j^2 = \left(\frac{1}{m-1}\right)\sum_{i=1}^{m}(x_{ji} - \bar{x}_j)^2$$

$$s_*^2 = \left(\frac{1}{gm-1}\right)\sum_{j=1}^{g}\sum_{i=1}^{m}(x_{ji} - \bar{x}_{..})^2$$

où $\bar{x}_{..} = \frac{t_{..}}{N} = \frac{1}{gm}\sum_j\sum_i x_{ji}$.

À noter que la variance de l'échantillon fondée sur toutes les gm unités peut s'exprimer comme suit:

$$s_*^2 = \frac{1}{mg-1}\left[(m-1)\sum_{j=1}^{g} s_j^2 + \frac{m}{N^2}\sum_{j=1}^{g}(t_j - T)^2 - \frac{mg}{N^2}(t_{..} - T)^2\right].$$

Par conséquent

$$E(s_*^2) = \frac{1}{mg-1}\left[g(m-1)S^2 + \frac{m}{N^2}\sum_{j=1}^{g}\mathrm{Var}(t_j) - \frac{mg}{N^2}\mathrm{Var}(t_{..})\right].$$

La réécriture donne

$$\mathrm{Var}(t_{..}) = N^2\left(\frac{m-1}{m}\right)S^2 + \left(\frac{1}{g}\right)\sum_{j=1}^{g}\mathrm{Var}(t_j) - N^2\left(\frac{mg-1}{mg}\right)E(s_*^2).$$

Donc, en remplaçant S^2 et $\mathrm{Var}(t_j)$ par des estimations non biaisées et en remplaçant $E(s_*^2)$ par s_*^2, nous pouvons produire des estimations à peu près non biaisées de $\mathrm{Var}(t_{..})$.

Il est peut-être bon de souligner que ce résultat n'exige pas que l'usager connaisse quoi que ce soit au sujet du plan de sondage original. Si l'on indique aux usagers comment inverser le plan de sondage original, ils pourront, à l'aide de sous-échantillonnages répétés, presque atteindre l'efficacité du plan de sondage original et déterminer aisément les erreurs d'échantillonnage appropriées. Une condition s'applique à ce résultat, à savoir que la taille du sous-échantillon doit être telle que $m \geq 2$. Soit dit en passant, pour $m = 2$, l'expression de la variance devient:

$$\mathrm{Var}(t_{..}) = \frac{N^2}{2}S^2 + \left(\frac{1}{g}\right)\sum_{j=1}^{g}\mathrm{Var}(t_j) - N^2\left(\frac{2g-1}{2g}\right)E(s_*^2).$$

Par conséquent, comme ci-dessus, il serait possible de construire un estimateur de la variance pour des plans de sondage à deux unités primaires d'échantillonnage par strate.

3.3 Illustration SDR

Nous examinons ici un exemple d'algorithme inverse et son fonctionnement. Notre point de départ est l'échantillon des sociétés SDR (statistique des revenus). Comme nous l'avons noté antérieurement, l'échantillon SDR comporte essentiellement un plan de sondage de type échantillon aléatoire simple stratifié et il est donc possible de l'inverser (sous-section 2.2).

Nous sommes d'avis que de nombreux usagers SDR trouveront un échantillon aléatoire simple inverse complet plus utile et plus facile à utiliser que la base de données d'échantillons stratifiés au complet. Un objectif provisoire pourrait être de leur fournir un ensemble d'échantillons aléatoires simples. Un système plus souple serait de fournir le logiciel interactif permettant à l'usager de désigner les échantillons aléatoires simples qui l'intéressent, choisis à même la base de données tout entière.

Dans nos simulations, nous avons utilisé quatre des strates dans l'échantillon SDR des déclarations des sociétés, c'est-à-dire les strates représentant les plus petites sociétés ordinaires (Hughes, Mulrow et coll., 1994). Comme l'indique le tableau 1, l'échantillon stratifié (de quatre strates) comportait 15,618 unités et le plus grand échantillon aléatoire simple qui peut être choisi est $m = 2,224$. Le tableau montre également la taille des populations et la variance estimative de la variable Actif total, à l'intérieur de chaque strate.

Tableau 1
Taille de la population de sociétés et de l'échantillon, ainsi que la variance estimative des strates, pour quatre strates SDR

Strate (h)	N_h	n_h	S_h^2 (en milliers)
1	1,376,801	3,889	222,808
2	552,909	2,224	670,162
3	678,371	4,005	12,796,578
4	436,023	5,500	14,984,753

La variable Actif total a été utilisée parce qu'il s'agit de la principale variable de stratification; par conséquent, la perte de précision qu'entraînerait l'ignorance de la stratification serait relativement grande. En effet, c'est ce que l'on a observé.

On trouvera ci-dessous le rapport de la variance du total estimatif en fonction de g échantillons aléatoires simples, de 2,224 chacun, divisé par la variance du total fondée sur l'échantillon stratifié. Le tableau présente des valeurs de g entre 1 et 1,000. Par exemple, si un seul échantillon aléatoire simple est choisi, la variance du total estimatif est 29 fois plus grande que la variance du total stratifié.

g	Augmentation relative de la variance
1	29.31
2	15.16
10	3.83
100	1.28
500	1.06
1000	1.03

À force de rééchantillonner de 500 à 1,000 fois, on a réduit la variance au même ordre de grandeur que l'échantillon stratifié. Même avec 100 sous-échantillons, nous avons de bons résultats ici, ce qui indique que le recours à un algorithme inverse pourrait donner de bons résultats pour ce genre de strate. Il n'est pas recommandé pour autant qu'un algorithme inverse soit utilisé en général avec un si faible rééchantillonnage. Sans doute, dans des populations très asymétriques, il en faudrait un bien plus grand nombre.

4. APPLICATIONS POSSIBLES ET PROCHAINES ÉTAPES

Dans le présent exposé, nous avons montré qu'il existe des algorithmes de plan de sondage inverses dans certains cas spéciaux. Nous n'avons toujours pas de résultat général, à supposer qu'il en existe un. Il s'agit là clairement d'un élément du problème qu'il y a lieu d'approfondir. Comme la plupart des outils, un algorithme de sondage inverse n'est pas nécessairement le meilleur choix dans certains cas; ce n'est peut-être même pas un choix raisonnable dans certaines circonstances. Il existe toutefois des applications dans lesquelles il semble offrir des avantages et il y a donc lieu de le considérer. Dans la présente section, nous décrivons brièvement des domaines dans lesquels cette méthode pourrait être utile et nous abordons certaines limites et certains problèmes qui subsistent.

Perspective axée sur la clientèle – Il est bon de souligner que notre stratégie est axée sur la clientèle. Même s'il n'était pas possible de les justifier pour d'autres raisons, on pourrait envisager les algorithmes inverses dans le cadre de la «réinvention» (p. ex. Osborne et Gaebler 1992). À l'heure actuelle, de nombreuses enquêtes complexes d'envergure ne sont peut-être pas suffisamment utiles pour la société, pour des raisons de sous-analyse grave ou même d'analyse fautive:

- À long terme, nous devons chercher à rendre la clientèle actuelle et éventuelle plus sensible aux enquêtes et à la quantification en général, par exemple à l'aide de la nouvelle série *What Is a Survey*? (sous la direction de Scheuren 1995).
- À court terme, nous devons commencer là où se trouve notre clientèle, en tenant compte du rôle souvent minime que jouent les données d'enquête dans leur processus de prise de décisions. Il y a certainement lieu de songer à réduire le fardeau cognitif que notre clientèle doit porter pour utiliser nos «produits» d'enquête complexes.

Choix de possibilités – Les gens sont de plus en plus sensibilisés aux faiblesses du paradigme de randomisation traditionnel (p. ex. Särndal et Swensson 1993). Mentionnons en particulier tout le travail que nous devons accomplir pour corriger les erreurs non imputables à l'échantillonnage. Cet aspect est exprimé dans Rao et Shao (1993). En reprenant les rajustements possibles pour ces erreurs non imputables à l'échantillonnage dans le cadre d'un échantillonnage aléatoire simple, nous pourrions peut-être même progresser davantage.

Depuis des décennies, les praticiens des enquêtes élaborent des plans de sondage extrêmement complexes pour ensuite en tirer des estimations d'intervalle de confiance et des estimations ponctuelles efficaces. Que savons-nous réellement des distributions que nos estimateurs d'échantillon produisent lorsque la taille utile des échantillons est petite à modérée? Serons-nous en mesure de bénéficier pleinement de la «révolution de visualisation» qui se produit actuellement (p. ex. Cleveland 1993)? Et en particulier en présence d'erreurs non imputables à l'échantillonnage? Nous devrions peut-être procéder de façon à toujours examiner les distributions. Le recours à un algorithme d'échantillonnage inverse serait une possibilité parmi d'autres (voir aussi Pfeffermann et Nathan 1985). De toute façon, des outils de visualisation plus puissants pour les enquêtes complexes pourraient aider même les vétérans parmi nous à approfondir leurs intuitions et à mieux les rattacher à la population particulière à l'étude. Évidemment, les efforts de visualisation entraînent également une baisse des prix d'utilisation des données d'enquête.

Un problème épineux qui se prêterait peut-être à un algorithme de sondage inverse est le cas où nous avons un plan de sondage à deux unités primaires d'échantillonnage par strate, avec L strates, où L est petit, par exemple moins de 30. Supposons également que, pour quelques-unes des variables de l'enquête, la stratification et la répartition en grappes n'ont pas d'importance, c'est-à-dire que l'effet du plan est $\delta = 1$, approximativement. Pour ces variables, ne serait-il pas possible de rendre la stabilité de l'estimation de la variance plus grande à l'aide de la méthode de rééchantillonnage qu'à l'aide de la méthode à répétition compensée que l'on applique normalement à l'estimation de la variance?

Un autre exemple que nous examinons est le cas où l'utilisateur s'intéresse à des tests d'indépendance en tableaux 2×2, fondés sur des données d'échantillon stratifié (Hinkins, Oh et Scheuren 1995). Pour la statistique du test du chi carré, nous en sommes à la comparaison de nos résultats à la stratégie proposée par Scheuren (1972) et Fellegi (1980). Il semble jusqu'à présent que la puissance de notre méthode soit comparable à celle de ces stratégies plus familières (comme on pourrait s'y attendre, par exemple, de Westfall et Young, 1993). C'est peut-être un cas où le travail supplémentaire exigé par l'algorithme d'échantillonnage inverse offrirait de réels avantages, en plus de simplement rendre le recours à des outils familiers plus facile pour les usagers, ceux-ci pouvant examiner la distribution et non pas seulement une valeur p.

Exemples de problèmes qui subsistent – Nous présentons ici des exemples de problèmes qui subsistent relativement à notre algorithme inverse. Ainsi, que se passe-t-il lorsque nous ignorons la taille de la population? Que se passe-t-il lorsque la population comporte plus d'une unité élémentaire, les personnes par exemple pour une analyse, les ménages pour une autre, les quartiers pour une troisième? Il existe des réponses pour ces difficultés, mais elles nous semblent ponctuelles. Dans de nombreuses enquêtes, par exemple, nous devinons la valeur N et nous utilisons cette valeur dans la stratification a posteriori. Ce degré d'approximation serait peut-être acceptable pour un inverse. Quant au problème des unités d'analyse multiples, nous pourrions mener plusieurs

inverses. Cette façon de procéder semble extrêmement gauche.

Nous avons indiqué que, dans certains cas, il n'est peut-être pas trop difficile de rééchantillonner plusieurs fois à l'aide de l'algorithme inverse afin d'atteindre une efficacité raisonnable. Mais que dire du cas où l'usager d'un échantillon stratifié s'intéresse à des sous-populations? Si les domaines d'intérêt sont en réalité les strates, il n'est pas avantageux pour l'usager d'utiliser les échantillons aléatoires simples produits à l'aide de l'algorithme inverse. Si les domaines d'intérêt chevauchent les strates et s'ils sont petits, le nombre d'échantillons requis pour l'algorithme inverse risque d'être très grand si l'on veut maintenir une estimation raisonnable pour les domaines.

Enfin, mentionnons brièvement un autre problème que nous avons examiné. De nombreux plans de sondage à plusieurs degrés retiennent réellement une seule unité primaire d'échantillonnage par strate. Les strates sont alors appariées à des fins d'estimation de la variance. Nous avons déjà noté qu'il existe un inverse pour cette approximation que l'on peut rendre à peu près aussi valable que l'est cette approximation à l'origine. Y a-t-il moyen d'obtenir une meilleure approximation à l'aide de la stratégie inverse directement?

Derniers mots – Notre profession évolue appréciablement. La révolution mondiale de la qualité a certainement eu des répercussions (Mulrow et Scheuren 1996). Nous assistons à une refonte de la façon de mener des enquêtes, de la conception à la collecte des données et à l'utilisation de ces données par la clientèle. Le présent exposé représente peut-être un modeste apport à cet égard.

REMERCIEMENTS

Nous tenons à remercier les arbitres et le rédacteur adjoint de leurs remarques pertinentes et de leur dévouement. Le document original que nous avons présenté n'était qu'une ébauche de l'article publié. Nous tenons également à remercier Phil Kott qui a décrit nos travaux en cours lors de différentes réunions de la Washington Statistical Society.

ANNEXE

Supposons un échantillon de k grappes d'une population de N grappes, dans laquelle chaque grappe compte le même nombre d'unités, M. Dans l'algorithme de sondage inverse, la première étape consiste à choisir le vecteur $(m_1, m_2, ..., m_k)$ qui contient le nombre d'unités à choisir de chaque grappe. Notons q le nombre de valeurs non zéro de m_i. La probabilité de sélection du motif avec $q = k$, c'est-à-dire le motif avec $m_i = 1$, pour tous les $i = 1, 2, ..., k$, est la suivante:

$$\Pr(q = k) = M^{k-1}\frac{(N-1)(N-2)...(N-k+1)}{(NM-1)(NM-2)...(NM-k+1)}.$$

Appelons cette probabilité P_1. Si $NM >>> k$ il est possible d'établir une approximation de P_1 à l'aide de

$$\prod_{i=1}^{k-1}\frac{(N-i)}{N} = \frac{(N-1)(N-2)...(N-k+1)}{N^{k-1}}.$$

Considérons ensuite la partition de k correspondant à $q = k - 1$; cela correspond exactement à une partition de k, c'est-à-dire $\{1,1, ..., 1,2\}$. Il existe $k(k-1)$ motifs également probables de $(m_1, ..., m_k)$ avec $q = k - 1$. La probabilité de sélection d'un vecteur $\boldsymbol{m}$ avec $q = k - 1$ est la suivante:

$$\Pr(q = k-1) = \frac{k(k-1)(M-1)}{2M(N-k+1)}P_1.$$

Par conséquent, il n'est pas difficile de calculer la probabilité que le $\boldsymbol{m}$ choisi compte soit $q = k$ ou $q = k - 1$. Le tableau ci-dessous donne quelques exemples des deux valeurs de M.

Tableau A
$\Pr(q = k - 1 \text{ ou } q = k)$

k	N	$M = 10$	$M = 100$
4	8	.92	.90
4	20	.99	.98
10	20	.38	.34
10	30	.63	.59
10	50	.83	.80
10	200	.99	.98
50	500	.35	.30
50	1000	.70	.66
50	5000	.98	.98

Pour k petit, il n'est pas difficile de calculer l'entière distribution des probabilités nécessaires à la production de $\boldsymbol{m}$. Mais à mesure que k augmente, le nombre de partitions augmente et ce calcul devient difficile ou du moins fastidieux. Pour $k = 4$, il n'existe que 4 partitions; pour $k = 10$ il existe 39 partitions possibles. On peut voir dans le tableau A que, à mesure que l'échantillon en grappes devient «plus grand», si le taux d'échantillonnage est assez faible, c'est-à-dire si $k << N$, on pourrait se limiter à calculer les probabilités pour ces deux partitions de façon à inverser approximativement l'échantillon en grappes. Pour $k = 10$ et $N = 200$, ces deux partitions tiennent compte essentiellement de toute la distribution des probabilités.

La probabilité de sélection d'une seule unité par grappe $(q = k)$ est plus faible que les valeurs du tableau A; donc, pour utiliser un inverse systématique, nous voudrions $k <<< N$. On peut l'obtenir dans certains contextes lorsque le nombre de grappes est grand et que nous sommes prêts à prendre k très petit, quitte à rééchantillonner l'enquête originale de façon répétée comme il a été décrit à la section 3.

Pour illustrer, supposons un échantillon de taille k_0 où, bien entendu, $k_0 < k$, de façon qu'un inverse soit possible. De plus, afin de vérifier si un inverse systématique pourrait

fonctionner, notons $k_0 <<< N$. C'est le cas que nous illustrons dans le tableau B. Dans le tableau B, nous avons limité notre attention à une seule valeur de N, N =5,000 grappes, bien que les résultats puissent être élargis facilement.

Tableau B
Pr{l'échantillon inverse choisit le motif (1,1, ..., 1)}

k_0	k_0/N	M = 10	M = 100
2	.0004	0.9998	0.9998
5	.001	0.9982	0.998
10	.002	0.9919	0.9911
20	.004	0.9663	0.9627
30	.006	0.9245	0.9166
40	.008	0.8687	0.8553
50	.01	0.8015	0.7821

Évidemment, à mesure que k/N devient plus petit, un échantillon systématique représente un meilleur inverse approximatif. Seule l'expérience pourrait confirmer si l'approximation à k_0 = 20 et k_0/N = 0,004, par exemple, est adéquate. Nous estimons qu'elle pourrait l'être, surtout du fait que l'utilisation d'un inverse systématique entraîne normalement des calculs de la variance plus prudents (car typiquement la corrélation intra-grappe [$\rho > 0$] est supérieure à 0).

BIBLIOGRAPHIE

BELLHOUSE, D. (1988). A brief history of random sampling methods. *Handbook of Statistics*, 6, 1-14.

CLEVELAND, W. (1993). *Visualizing Data*. Summit, NJ: Hobart Press.

COCHRAN, W. (1977). *Sampling Techniques*. New York: Wiley.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics, 7, 139-172.*

FELLEGI, I. (1980). Approximate tests of independence et goodness of fit based on multistage samples. *Journal of the American Statistical Association*, 75, 261-268.

HANSEN, M. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2, 162-179.

HINKINS, S., OH, H.L., et SCHEUREN, F. (1995). Using an Inverse Algorithm for Testing of Independence Based on Stratified Samples. George Washington University, Rapport Technique.

HUGHES, S., MULROW, J., HINKINS, S., COLLINS, R., et UBERALL, B. (1994). Section 3, *Statistics of Income – 1991, Corporation Income Tax Returns*, 9-17. Washington, DC: Internal Revenue Service.

KISH, L. (1995). The Hundred Years Wars of Survey Sampling. Centennial representative Sampling Conference, Rome, le 31 mai 1995.

LAHIRI, D. (1951). A method for sample selection providing unbiased ratio estimates, *Bulletin de l'Institut International de Statistique*, 34, 72-86.

McCARTHY, P., et SNOWDEN, C. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics*. Series 2, No. 95, DHHS Pub. No. (PHS) 85-1369. Washington, DC: Public Health Service.

MULROW, J., et SCHEUREN, F. (1996). Measuring to improve quality and productivity in a processing environment. *Data Quality*, 2, 11-20.

OSBORNE, D., et GAEBLER, T. (1992). *Reinventing Government*. New York: Plume.

PFEFFERMANN, D., et NATHAN, G. (1985). Problems in model identification based on data from complex samples. *Bulletin de l'Institut International de Statistique*, 68.

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

SÄRNDAL, C.-E., et SWENSSON, B. (1993). Présentation à la Washington Statistical Society sur la nature changeante du paradigme de l'échantillonnage.

SCHEUREN, F. (1972). Topics in Multivariate Finite Population Sampling and Data Analysis. George Washington University Dissertation de doctorat.

SCHEUREN, F. (Éd.) (1995). *What is a Survey?* tiré d'une série de brochures publiée par l'American Statistical Association pour accroître la connaissance des enquêtes.

SKINNER, C., HOLT, D., et SMITH, T., (Éds.) (1989). *Analysis of Complex Surveys*. New York: Wiley.

WESTFALL, P., et YOUNG, S. (1993). *Resampling-Based Multiple Testing*. New York: Wiley.

WOLTER, K. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 25-35
Statistique Canada

Sélection des variables pour l'estimation par régression dans le cas des populations finies

PEDRO L. D. NASCIMENTO SILVA et CHRIS J. SKINNER[1]

RÉSUMÉ

Les auteurs examinent la sélection des variables auxiliaires pour l'estimation par régression des paramètres des populations finies dans le cas d'un plan de sondage aléatoire simple. Ce problème fondamental que posent les méthodes d'échantillonnage fondé sur un modèle ou assisté par un modèle prend une importance d'ordre pratique quand le nombre de variables disponibles est grand. Les auteurs élaborent une méthode consistant à minimiser un estimateur de l'erreur quadratique moyenne, puis, la comparent à d'autres en utilisant un ensemble fixe de variables auxiliaires, un test de signification classique, une méthode de réduction du nombre de conditions et une méthode de régression ridge. Selon les résultats de l'étude, la méthode proposée est efficace. Les auteurs soulignent que la méthode de sélection des variables influe sur les propriétés des estimateurs types de la variance, ce qui entraîne par conséquent un problème d'estimation de la variance.

MOTS CLÉS: Informations auxiliaires; étalonnage; enquêtes par sondage; sélection d'un sous-ensemble; régression ridge.

1. INTRODUCTION

L'estimation par régression est une méthode utilisée fréquemment dans le cas des enquêtes par sondage pour intégrer des informations auxiliaires sur la population $\bar{X}$ (Cochran 1977, chap. 7). Dans le cas élémentaire où on connaît la moyenne de la population d'un vecteur de variables x_i et qu'on effectue un échantillonnage aléatoire simple, l'estimateur de régression de la moyenne de la population $\bar{Y}$ d'une variable étudiée y_i prend la forme

$$\bar{y}_r = \bar{y} + (\bar{X} - \bar{x})'b \qquad (1)$$

où $\bar{y}$ et $\bar{x}$ sont respectivement les moyennes d'échantillon de y_i et de x_i, et où b est le vecteur d'échantillon des coefficients de régression linéaire de y_i sur x_i.

L'estimation par régression est utile pour au moins trois raisons. Premièrement, la méthode est souple. En principe, on peut intégrer dans $\bar{X}$ n'importe quel nombre de moyennes de population de variables continues ou binaires. Plus précisément, la stratification a posteriori se dégage comme un cas particulier (Särndal, Swensson et Wretman 1992, sec. 7.6). La méthode s'étend aussi au traitement de plans de sondage complexes. Deuxièmement, l'estimation par régression présente certaines propriétés d'optimisation de l'efficacité. À cet égard, consulter, par exemple, Isaki et Fuller (1982, Theorem 3). Troisièmement, $\bar{y}_r$ présente la propriété d'«étalonnage» voulant que, si y_i est une des variables de x_i de sorte que $\bar{Y}$ est connue, alors $\bar{y}_r = \bar{Y}$ (Deville et Särndal 1992).

Dans le présent document, nous cherchons à déterminer comment sélectionner les variables x utilisées dans l'estimateur de régression. Cette question présente un intérêt pour deux raisons. La première, d'ordre pratique, tient simplement au fait que, dans certaines circonstances, le nombre de variables x_i peut être très grand. Par exemple, dans le cas du recensement de la population de plusieurs pays, on enregistre les valeurs de certaines variables au moyen d'un «questionnaire abrégé» rempli par tous les membres de la population et les valeurs d'autres variables, au moyen d'un «questionnaire détaillé» rempli par un échantillon de la population. Donc, on connaît les moyennes de population des variables visées par le questionnaire abrégé, ainsi que leurs carrés, leurs cubes, leurs produits et ainsi de suite. Habituellement, on connaît aussi la définition des petites régions. Donc, la dimension de x_i en tant que vecteur contenant des fonctions des variables visées par le questionnaire abrégé ainsi que des variables fictives représentant chaque petite région pourrait facilement être de l'ordre de plusieurs milliers. Le cas échéant, la sélection de variables x devient une nécessité pratique.

La deuxième raison est surtout fondamentale dans le cas d'une méthode d'échantillonnage assisté par modèle ou fondé sur un modèle. Dans le contexte de l'estimation par régression, on peut décrire ces méthodes comme suit. Pour commencer, on choisit un modèle de régression possédant un «bon pouvoir de prédiction», de sorte que l'estimateur de régression ait une «bonne efficacité». Puis, on adopte une méthode d'inférence fondée sur le plan de sondage dans le cas de la méthode assistée par modèle (Särndal et coll. 1992) ou une méthode de prédiction basée sur un modèle, dans le cas de la méthode fondée sur un modèle. Si de nombreux articles ont été publiés sur le problème d'inférence, les chercheurs semblent avoir accordé officiellement très peu d'attention au problème de sélection du modèle. En pratique, il semble que les efforts se limitent à choisir les variables x «principales» qui

[1] Pedro L. D. Nascimento Silva, IBGE-Departamento de Metodologia, Avenida Chile 500, Rio de Janeiro-RJ, Brasil; et Professor Chris J. Skinner, Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, UK.

expliquent «la plus grande partie» du R^2 de l'échantillon (consulter Särndal et coll. 1992, sec. 7.9.1). Pourtant, un plus grand soutien théorique semble nécessaire, particulièrement quand le nombre de variables x est grand.

Une raison supplémentaire d'examiner plus formellement le problème de sélection des variables tient à ce qu'une telle étude apporterait des éclaircissements quant à l'incidence éventuelle de la sélection des variables sur l'inférence. On reconnaît de longue date le fait que la sélection d'estimateurs d'après l'échantillon peut avoir des effets sur les propriétés des estimateurs choisis (Hansen et Tepping 1969, App.), mais rares sont les études qui visent à déterminer quels pourraient être ces effets.

Dans le présent article, nous examinons une méthode de sélection des variables visant à minimiser l'erreur quadratique moyenne de $\bar{y}_r$. Toutefois, à la section 2, nous commençons par étudier la corrélation entre l'erreur quadratique moyenne de $\bar{y}_r$ et le nombre de variables x, puis, à la section 3, nous considérons d'autres estimateurs de l'erreur quadratique moyenne de $\bar{y}_r$. Enfin, à la section 4, nous présentons certaines méthodes de sélection des variables fondées sur ces estimateurs.

Nous comparons notre méthode de sélection des variables à quatre méthodes existantes. En premier lieu, nous considérons la méthode habituelle consistant à utiliser un sous-ensemble fixe de variables auxiliaires, indépendamment de l'échantillon observé. Puis, nous examinons une «méthode de réduction du nombre de conditions» inspirée des travaux de Bankier (1990), selon laquelle on élimine des variables auxiliaires afin de réduire le nombre de conditions d'une matrice donnée de produits croisés des variables x.

En troisième lieu, nous suivons Bardsley et Chambers (1984) et considérons une méthode de régression ridge. Plutôt que d'inclure une sélection de variables, cette méthode vise à résoudre le problème éventuel de multicollinéarité de l'estimateur de régression grâce à une modification de l'estimateur, en tenant compte de certaines erreurs d'étalonnage. La méthode de régression ridge et la méthode de réduction du nombre de conditions présentent toute deux l'avantage de ne pas obliger à préciser la variable dépendante y, car elles visent à produire un ensemble unique de poids «d'étalonnage» applicable à toutes les variables étudiées. Cependant, leur application ne garantit pas que l'efficacité augmentera. Dans les tableaux présentés à la section 6, les résultats obtenus par ces méthodes sont séparés des autres par une ligne pour indiquer qu'ils sont différents.

En quatrième lieu, nous envisageons la sélection des variables en s'appuyant sur les critères des tests de signification classiques. Selon nous, dans l'ensemble, l'objectif de la sélection des variables lors d'estimation des paramètres de populations finies par régression est assez différent de celui de l'estimation des paramètres ou de la prédiction des valeurs de y dans le cas des observations isolées produites par la régression classique (Miller 1990). Néanmoins, il nous paraît souhaitable de prendre cette dernière comme point de comparaison.

À la section 5, nous examinons les propriétés de l'estimateur de régression après sélection des variables d'après les variances estimées. À la section 6, nous décrivons une étude empirique effectuée en vue de comparer les méthodes de sélection des variables que nous proposons aux méthodes concurrentes décrites plus haut. Cette étude porte sur les données d'un recensement pilote effectué dans la municipalité de Limeira, au Brésil, en prévision du Recensement de la population du Brésil de 1991. À la section 7, nous présentons nos conclusions et certaines orientations que pourraient prendre les travaux de recherche futurs.

2. RELATION ENTRE LA VARIANCE DE L'ESTIMATEUR DE RÉGRESSION ET LE NOMBRE DE VARIABLES x

Commençons par définir certaines conventions de notation. Représentons par $U = \{1,...,N\}$ une population finie de N éléments distincts et par $s \subset U$ un échantillon de n éléments distincts tirés de U selon un plan d'échantillonnage aléatoire simple sans remise. Supposons que $x_i = (x_{i1},...,x_{iq})'$ est le vecteur $q \times 1$ des variables auxiliaires associées au i-ième élément de la population. Nous supposons qu'on connaît les valeurs des $x_i (i \in s)$ de l'échantillon, ainsi que le vecteur des moyennes de la population $\bar{X} = N^{-1}\sum_{i\in U} x_i$. Le vecteur des moyennes de l'échantillon est représenté par $\bar{x} = n^{-1}\sum_{i\in s} x_i$.

Représentons par y_i la valeur de la variable étudiée y pour le i-ième élément de la population et supposons qu'on n'observe les valeurs de y_i que pour $i \in s$. Nous visons à estimer la moyenne de la population $\bar{Y} = N^{-1}\sum_{i\in U} y_i$.

L'estimateur de régression de $\bar{Y}$ est donné par $\bar{y}_r$ dans l'équation (1), où $\bar{y} = n^{-1}\sum_{i\in s} y_i$, $b = \hat{S}_x^{-1}\hat{S}_{xy}$, $\hat{S}_x = n^{-1}\sum_{i\in s} (x_i - \bar{x})(x_i - \bar{x})'$, et $\hat{S}_{xy} = n^{-1}\sum_{i\in s}(x_i - \bar{x})(y_i - \bar{y})$.

Cet estimateur peut être amené par le modèle linéaire sous-jacent

$$y_i = \beta_0 + x_i'\beta + \varepsilon_i \qquad (2)$$

où les ε_i sont des perturbations indépendantes dont les moyennes sont nulles et dont la variance commune est σ^2, puisque nous pouvons écrire $\bar{y}_r = \hat{\beta}_0 + \bar{X}'\hat{\beta}$, où $\hat{\beta}_0 = \bar{y} - \bar{x}'b$ et $\hat{\beta} = b$ représentent les estimateurs par les moindres carrés de β_0 et β, respectivement. En vertu de ce modèle, la variance de $\bar{y}_r - \bar{Y}$ conditionnelle à x_i peut s'écrire

$$\mathrm{Var}_M(\bar{y}_r - \bar{Y} \mid x_i) = \sigma^2 n^{-1}[1 - n/N + (\bar{X} - \bar{x})'\hat{S}_x^{-1}(\bar{X} - \bar{x})]. \qquad (3)$$

On peut interpréter le dernier terme comme l'effet de l'estimation de β au moyen de b. À mesure que le nombre q de variables x augmente, on peut s'attendre à ce que la variance résiduelle σ^2 diminue, alors que le terme $(\bar{X} - \bar{x})'\hat{S}_x^{-1}(\bar{X} - \bar{x})$ pourrait augmenter à mesure que $\hat{S}_x^{-1}$ devient plus instable. Une autre façon d'interpréter ce terme consiste à représenter $\bar{y}_r$ comme un estimateur pondéré $\bar{y}_r = n^{-1}\sum_{i\in s} g_i y_i$ où $g_i = 1 + (\bar{X} - \bar{x})'\hat{S}_x^{-1}(x_i - \bar{x})$ Alors, nous pouvons écrire (3) comme suit

$$\mathrm{Var}_M(\bar{y}_r - \bar{Y} \mid x_i) = \sigma^2 n^{-1}(1 - n/N + c_g^2) \qquad (4)$$

où c_g est le coefficient de variation des g_i de l'échantillon.

Pour étudier la corrélation prévue entre c_g^2 et q, nous étendons maintenant le modèle en supposant que les variables x_i sont indépendantes et distribuées identiquement selon la loi normale. Notant que $(\bar{x} - \bar{X})$ et $\hat{S}_x$, sont des quantités indépendantes et que $E_M(\bar{y}_r - \bar{Y} \mid x_i) = 0$, nous obtenons la variance non conditionnelle

$$\begin{aligned}&\mathrm{Var}_M(\bar{y}_r - \bar{Y})\\ &= \sigma^2 n^{-1}\{1 - n/N + \mathrm{tr}[E_M[(\bar{X} - \bar{x})(\bar{X} - \bar{x})']E_M(\hat{S}_x^{-1})]\}\\ &= \sigma^2 n^{-1}(1 - n/N)[1 + q/(n - q - 2)] \qquad (5)\end{aligned}$$

en nous appuyant sur le fait que $n^{-1}\hat{S}_x^{-1}$ est caractérisé par une distribution de Wishart inverse (Mardia, Kent et Bibby 1979, p. 69 et 85). Ce résultat est également vérifié pour les grandes valeurs de n, même si les conditions normales ne sont pas respectées, en ce sens que $[1 - n/N + c_g^2]/(1 - n/N)[1 + q/(n - q - 2)]$ continue de converger vers 1 quand n augmente et que q est constant (dans des conditions faibles).

L'expression (5) rend la corrélation avec q explicite. À mesure que q augmente, nous pouvons nous attendre à ce que σ^2 diminue, mais que $E_M(c_g^2)$ augmente. La diminution de σ^2 deviendra faible après l'inclusion de quelques variables x importantes, donc la variance commencera à augmenter au moment où le nombre de variables x représentera une fraction non négligeable de la taille de l'échantillon.

Les résultats (4) et (5) se fondant sur de fortes hypothèses de modélisation ne nous fournissent que la motivation. Dans le cas général, $\bar{x} - \bar{X} = O_p(n^{-1/2})$ (distribution de randomisation avec conditions types de régularité), de sorte que le dernier terme de (3) est d'ordre $O_p(n^{-2})$. Quand le modèle (2) ne doit pas être vérifié, on obtient une approximation asymptotique de deuxième ordre plus générale de l'erreur quadratique moyenne de $\bar{y}_r$ correspondant au plan en généralisant le théorème 4.1 de Deng et Wu (1987). Le lecteur trouvera des précisions dans Silva (1996).

Notre objectif consistant à élaborer une méthode de sélection des variables qui réduise au minimum l'erreur quadratique moyenne estimée de $\bar{y}_r$ nous allons maintenant examiner les estimateurs de cette erreur quadratique moyenne.

3. ESTIMATION DE L'ERREUR QUADRATIQUE MOYENNE DE L'ESTIMATEUR DE RÉGRESSION MULTIPLE

Nous obtenons un estimateur simple de l'erreur quadratique moyenne de $\bar{y}_r$ en généralisant l'expression (7.29) de Cochran (1977, p. 195) au cas de plusieurs variables auxiliaires, soit:

$$v_s = \frac{1 - f}{n}\hat{S}_e \qquad (6)$$

où $\hat{S}_e = (n - q - 1)^{-1}\sum_{i\in s}\hat{e}_i^2$ et où $\hat{e}_i = (y_i - \bar{y}) - (x_i - \bar{x})'b$.

Toutefois, cet estimateur ne tient pas compte de l'élément $O(n^{-2})$ de l'erreur quadratique moyenne. Donc, en guise de deuxième estimateur de l'erreur quadratique moyenne, nous généralisons l'estimateur v_d étudié par Deng et Wu (1987) au cas où q est général. Il s'agit d'un cas particulier de l'estimateur G_2 de la variance fondé sur un modèle et résistant au biais qui a été proposé au départ par Royall et Cumberland (1978) pour traiter le cas où les variances résiduelles produites par le modèle (2) sont constantes. Cet estimateur est donné par

$$v_d = \frac{1 - f}{n(n - 1)}\sum_{i\in s}\alpha_i\hat{e}_i^2 \qquad (7)$$

où

$$\alpha_i = (g_i^2 - 2g_i f + f)/\{(1 - f)[1 - (x_i - \bar{x})'\hat{S}_x^{-1}(x_i - \bar{x})/(n - 1)]\}.$$

Nous supposions au départ que v_d serait un estimateur de deuxième ordre non biaisé, comme Deng et Wu (1987, éq. 4.4) le démontre dans le cas où $q = 1$. Malheureusement, cette hypothèse n'est pas confirmée dans le cas général où $q > 1$, mais on peut s'attendre à ce que le biais entachant v_d soit plus petit que celui de v_s, comme l'indiquent les expressions du biais de deuxième ordre obtenues pour v_s et v_d par Silva (1996).

Le problème que pose l'utilisation de v_d comme estimateur de la variance tient à ce qu'il est difficile de généraliser son application aux plans de sondage complexes. Donc, nous envisageons comme troisième estimateur de la variance une version modifiée d'un estimateur proposé par Särndal, Swensson et Wretman (1989), définie par:

$$v_g = \frac{1 - f}{n(n - q - 1)}\sum_{i\in s} g_i^2\hat{e}_i^2. \qquad (8)$$

En principe, cet estimateur devrait se comporter de façon similaire à v_d puisque $\alpha_i = g_i^2 + O_p(n^{-1/2})$. Conformément à la terminologie utilisée par Särndal et coll. (1992, p. 232), les g_i sont les poids g appropriés dans les conditions d'échantillonnage aléatoire simple si on adopte (2) comme modèle sous-jacent. L'expression (8) diffère de l'estimateur correspondant proposé par Särndal et coll. (1989, exemple 4.4) en ce sens que nous utilisons le dénominateur $(n - q - 1)$ plutôt que le dénominateur original $(n - 1)$.

4. MÉTHODES DE SÉLECTION DES VARIABLES

Nous considérons deux méthodes fondamentales de sélection des variables. Premièrement, nous examinons une méthode englobant *tous les sous-ensembles* qui consiste à calculer un des estimateurs de l'erreur quadratique moyenne v_s, v_d ou v_g de la section 3 pour les 2^q sous-ensembles possibles des q variables auxiliaires (incluant toujours la coordonnée à l'origine) et à choisir le sous-ensemble qui correspond à l'estimation la plus faible de l'erreur quadratique moyenne. Cette méthode comporte manifestement de longs calculs si q est grand. Donc, nous considérons une deuxième méthode, appelée méthode de *sélection progressive*, qui consiste à

choisir au départ la moyenne de l'échantillon comme estimateur, puis à ajouter la variable qui minimise l'estimation de l'erreur quadratique moyenne. La procédure est répété jusqu'à ce que l'estimation de l'erreur quadratique moyenne commence à augmenter, le sous-ensemble de variables qui produit l'estimation minimale de l'erreur quadratique moyenne étant sélectionné à ce moment-là.

Nous pouvons comparer ces deux méthodes à une autre, inspirée des travaux de Bankier et de ses associés (voir Bankier (1990) et Bankier, Rathwell et Majkowski (1992) que nous appelons *méthode de réduction du nombre de conditions*. Avant de la décrire, notons que l'on peut aussi exprimer l'estimateur de régression donné en (1) par

$$\bar{y}_r = [n\bar{y} + (N\bar{X}^* - n\bar{x}^*)'(X_s^{*\prime}X_s^*)^{-1}X_s^{*\prime}y_s]/N \qquad (9)$$

où X_s^* est la matrice $n \times (q+1)$ pour laquelle $x_i^{*\prime} = (1, x_{i1}, ..., x_{iq})' = (1 : x_i')$ représente la i-ième rangée, $\bar{x}^* = (1 : \bar{x}')'$ et $\bar{X}^* = (1 : \bar{X}')'$ représentent, respectivement, les vecteurs des moyennes de l'échantillon et de la population des x_i^* et y_s est le vecteur $n \times 1$ des observations de la variable dépendante dans l'échantillon.

Donc, l'estimateur de régression dépend de l'inversion de la matrice $X_s^{*\prime}X_s^*$ de produits croisés, matrice dont les conditions deviennent parfois inappropriées et qui, par conséquent, gonfle la variance de l'estimateur de régression.

Bankier (1990) propose une méthode en deux étapes pour calculer les estimateurs de régression des moyennes (ou des totaux) en vertu de laquelle on élimine certaines colonnes de la matrice de données auxiliaires X_s^* afin de réduire le nombre de conditions de la matrice de produits croisés $X_s^{*\prime}X_s^*$ et d'éviter ainsi les situations indésirables (poids négatifs ou aberrants, caractéristiques rares ou dépendance linéaire exacte entre colonnes). Bankier et ses collaborateurs (1992) décrivent en détail l'application de la méthode aux données du Recensement de la population du Canada de 1991. Il convient de souligner que, si elle intègre la sélection des variables, la méthode élaborée par Bankier et ses collaborateurs ne vise pas à être efficace pour une variable observée particulière. Elle est axée principalement sur l'étalonnage, avec, en parallèle, la fourniture d'un ensemble unique de poids applicable à toutes les variables observées.

Nous pouvons décrire notre méthode de réduction du nombre de conditions au moyen de l'algorithme ci-après, qui se fonde sur l'élimination à rebours des variables auxiliaires qui produisent un grand nombre de conditions pour la matrice de produits croisés CP = $X_s^{*\prime}X_s^*$, au lieu de l'ajout progressif de variables décrit par Bankier et coll. (1992).

1) Calculer la matrice de produits croisés CP = $X_s^{*\prime}X_s^*$ en tenant compte de toutes les colonnes existantes au départ (sous-ensemble saturé).
2) Calculer la forme canonique d'Hermite de CP, soit H (voir Rao 1973, p. 18), et déceler les singularités en examinant les éléments diagonaux de H. Tout élément diagonal nul de H indique que les colonnes correspondantes de $X_s^{*\prime}X_s^*$ (et X_s^*) dépendent linéairement d'autres colonnes (voir Rao 1973, p. 27). Éliminer chacune de ces colonnes en supprimant les rangées et colonnes correspondantes de $X_s^{*\prime}X_s^*$.
3) Après avoir éliminé toute colonne linéairement dépendante, calculer le nombre de conditions $c = \lambda_{max}/\lambda_{min}$ de la matrice CP réduite, où λ_{max} et λ_{min} représentent la plus grande et la plus petite valeurs propres de CP, respectivement. Si $c < L$ est une valeur précise, arrêter et utiliser toutes les variables auxiliaires restantes.
4) Sinon, procéder à l'élimination à rebours comme suit. Pour chaque k, supprimer les k-ième rangée et colonne de CP, et calculer de nouveau les valeurs propres et le nombre de conditions de la matrice réduite. Calculer les réductions du nombre de conditions $r_k = c - c_k$, où c_k est le nombre de conditions après avoir éliminé les k-ième rangée et colonne de CP. Déterminer $r_{max} = \max_k(r_k)$ et $k_{max} = \{k : r_{max} = r_k\}$, et éliminer la colonne k_{max} en supprimant les rangée et colonne k_{max} de CP. Poser $c = c_{k_{max}}$ et itérer tant que $c \geq L$ et que $q \geq 2$, en commençant chaque nouvelle itération par la matrice CP réduite résultant de la précédente.

Une autre méthode que nous étudions est l'estimation par régression ridge de Bardsley et Chambers (1984). Au lieu de se fonder sur la sélection de sous-ensembles des variables auxiliaires existantes, elle s'appuie sur le relâchement des propriétés d'étalonnage de l'estimateur de régression en vue d'obtenir des estimations plus stables. L'estimateur de régression ridge est représenté par

$$\bar{y}_{BC} = [n\bar{y} + (N\bar{X}^* - n\bar{x}^*)'(\lambda C^{-1} + X_s^{*\prime}X_s^*)^{-1}X_s^{*\prime}y_s]/N \quad (10)$$

où λ est un paramètre scalaire de «ridging» et où C est une matrice diagonale de coefficients de «coût» associés aux erreurs d'étalonnage tolérées quand on estime les totaux des variables auxiliaires au moyen de $\bar{y}_{BC}$.

Bardsley et Chambers (1984) laissent entendre qu'on pourrait utiliser la spécification de la matrice C pour contrôler l'influence de chaque variable auxiliaire sur l'estimateur résultant de la moyenne de la variable dépendante, donc imiter le procédé de sélection des sous-ensembles.. Comme dans le cas du paramètre de «ridging» λ, ils proposent de choisir la valeur la plus petite, de façon qu'aucun poids de cas implicites ne soit inférieur à $1/N$ (ou à 1 pour les totaux estimés).

5. PROPRIÉTÉS DES ESTIMATEURS DE RÉGRESSION APRÈS LA SÉLECTION DES VARIABLES

Dans le cas des méthodes fondamentales de sélection des variables, nous considérons un ensemble de stratégies d'estimation $S = \{(\bar{y}_r^\gamma, v^\gamma); \gamma \in \Gamma\}$, où $\bar{y}_r^\gamma$ et v^γ représentent l'estimateur de régression et un estimateur de sa variance, respectivement, pour un sous-ensemble γ des q variables auxiliaires existantes, et où Γ représente l'ensemble des sous-ensembles. La procédure de sélection des variables consiste

à choisir un sous-ensemble γ^* dans Γ conformément à une règle déterminée par les données et par S, l'estimateur ponctuel résultant étant $\bar{y}_r^{\gamma^*}$.

Pour chaque sous-ensemble déterminé γ, il s'ensuit, dans des conditions de régularité type (Isaki et Fuller 1982), que $\bar{y}_r^{\gamma}$ est convergent pour la moyenne de la population $\bar{Y}$, autrement dit que $\bar{y}_r^{\gamma} - \bar{Y} = o_p(1)$. Alors, pour une valeur donnée $\delta > 0$, $|\bar{y}_r^{\gamma} - \bar{Y}| > \delta$ implique que $|\bar{y}_r^{\gamma} - \bar{Y}| > \delta$ pour certains sous-ensembles γ. Nous pouvons donc écrire

$$\Pr(|\bar{y}_r^{\gamma^*} - \bar{Y}| > \delta) \leq \sum_{\gamma \in \Gamma} \Pr(|\bar{y}_r^{\gamma} - \bar{Y}| > \delta) \qquad (11)$$

et, comme Γ est fini, le membre droit de l'équation (11) converge vers zéro et il s'ensuit que $\bar{y}_r^{\gamma^*}$ est également convergent.

Cependant, la distribution de $\bar{y}_r^{\gamma^*}$ dépendra de la règle de sélection de façon complexe. Consulter Grimes et Sukhatme (1980) pour un examen de l'efficacité de $\bar{y}_r^{\gamma^*}$ dans le cas le plus simple où il existe seulement deux estimateurs possibles, à savoir un estimateur de régression à une seule variable x et un estimateur de différence (dont la moyenne est un cas particulier), et où les variables suivent conjointement une distribution normale.

Contrairement à $\bar{y}_r^{\gamma^*}$, il n'y a aucune raison que v^{γ^*} soit convergent pour $\mathrm{Var}(\bar{y}_r^{\gamma^*})$, même si v^{γ} est convergent pour $\mathrm{Var}(\bar{y}_r^{\gamma})$ pour chaque sous-ensemble γ déterminé. En particulier, nous pouvons nous attendre à ce que v^{γ^*} sous-estime $\mathrm{Var}(\bar{y}_r^{\gamma^*})$ si la règle de sélection est telle que v^{γ^*} représente le minimum de v^{γ}. Cet effet est semblable à la surestimation bien connue de R^2 après la sélection de sous-ensembles dans le cas de la régression linéaire multiple type (Miller 1990, p. 7 à 10).

6. ÉTUDE EN SIMULATION

Nous présentons ici une petite étude en simulation effectuée pour évaluer la performance des diverses méthodes de sélection des variables envisagées. Nous avons choisi comme population pour la simulation un ensemble de données comprenant 426 enregistrements obtenus pour les chefs de ménage qui ont répondu au questionnaire détaillé utilisé durant l'essai de recensement de population effectué en 1988 à Limeira, dans l'État de São Paulo, au Brésil.

Cet essai a été effectué à titre d'enquête pilote durant la préparation du Recensement de la population du Brésil de 1991. Le test comprenait deux vagues de collecte de données. Durant la première, chaque recenseur a visité tous les logements occupés dans une région de recensement donnée (une région comptant de 200 à 300 ménages, en moyenne) et a rempli un questionnaire abrégé. Ce dernier contenait quelques questions sur les caractéristiques du ménage et sur chaque membre du ménage (sexe, âge, lien avec le chef de ménage et niveau d'alphabétisme). Le questionnaire comprenait aussi, à l'intention des chefs de ménage uniquement, une question sur la scolarité et une autre sur le revenu mensuel total. Le revenu mensuel total déclaré par les chefs de ménage ne fournit qu'une approximation du revenu réel, en raison des limitations de la méthode d'interview suivie durant la première vague de collecte de données.

Puis, une deuxième vague de collecte de données a été entreprise dans chaque région de recensement. Les mêmes recenseurs ont rendu visite à un échantillon de un ménage sur dix (sélectionné systématiquement d'après la liste des logements occupés établie durant la première vague de collecte de données) pour recueillir des renseignements au moyen d'un questionnaire détaillé contenant toutes les questions du questionnaire abrégé et de nombreuses autres.

La taille de la population étudiée était de l'ordre de 44,000 ménages comptant en tout 188,000 personnes. La taille de l'échantillon correspondait à 10 % environ de la taille de la population. Pour limiter le coût du traitement informatique, nous avons utilisé, pour notre étude en simulation, une sous-population comprenant tous les enregistrements d'échantillon pour 426 chefs de ménage vivant dans 20 des 170 régions de recensement. Nous avons choisi ces enregistrements comme population pour la simulation parce qu'ils contiennent tous les renseignements détaillés fournis par le questionnaire utilisé pour interviewer les ménages faisant partie de l'échantillon, ainsi que les renseignements obtenus par personne interposée au moyen du questionnaire abrégé durant la première vague d'interviews.

Nous avons étudié le revenu mensuel total, tel qu'obtenu au moyen du questionnaire détaillé, à titre de variable dépendante principale (y), ainsi que 11 variables auxiliaires possibles, à savoir:

x_1 = indicateur du sexe du chef de ménage égal masculin;
x_2 = indicateur de l'âge du chef de ménage inférieur ou égal à 35;
x_3 = indicateur de l'âge du chef de ménage supérieur à 35 et inférieur ou égal à 55;
x_4 = nombre total de pièces dans le logement;
x_5 = nombre total de salles de bains dans le logement;
x_6 = indicateur de propriété du logement;
x_7 = indicateur de ce que le type de logement est une maison;
x_8 = indicateur de la possession d'au moins une voiture par le ménage;
x_9 = indicateur de la possession d'un téléviseur couleur par le ménage;
x_{10} = nombre d'années d'études du chef de ménage;
x_{11} = approximation du revenu mensuel total du chef de ménage.

À partir de ces 11 variables, nous avons construit deux ensembles distincts de variables auxiliaires pour nos simulations. Nous avons défini le premier ensemble en prenant cinq variables auxiliaires, nommément $x_1,...,x_4$ et x_{11} possédant un pouvoir de prédiction de y raisonnable, particulièrement à cause de la présence de l'approximation du revenu x_{11}. Le deuxième ensemble que nous avons examiné contenait dix variables auxiliaires, nommément $x_1,...,x_{10}$, ayant, à cause de l'exclusion de x_{11}, un pouvoir de prédiction plus faible que l'ensemble précédent. À titre de référence, nous présentons au tableau 3 en annexe la matrice de corrélation de la population pour la variable observée y et pour les 11 variables auxiliaires de la population.

Puis, nous avons sélectionné 1,000 échantillons de taille 100 à partir de cette population de simulation par échantillonnage aléatoire simple sans remise.

Avant d'examiner les résultats détaillés de la simulation, examinons quels avantages pourrait offrir la sélection des variables à la lumière de la discussion motivante fondée sur les modèles de la section 2. Rappelons que, comme le montre l'équation (4), si on se base sur le modèle (2), un terme c_g^2 gonfle la variance conditionnelle de $\bar{y}_r$ à cause de l'estimation de β. Nous avons évalué la distribution de c_g^2 sur les 1,000 échantillons, quand on considère cinq et dix variables auxiliaires. Dans le cas de cinq variables auxiliaires, la valeur de la médiane de c_g^2 est 0.036, celle du quartile supérieur est 0.056 et celle du maximum est 0.255. Ces résultats concordent grossièrement avec l'équation (5) qui implique que, en vertu du modèle, la valeur attendue de c_g^2 est $(1-n/N)q/(n-q-2) = 0.041$. Il convient de souligner que la forte variation de c_g^2 d'un échantillon à l'autre donne à penser qu'il pourrait être judicieux d'adopter une méthode consistant à sélectionner un ensemble distinct de variables pour chaque échantillon. La variation de c_g^2 est encore plus forte dans le cas de dix variables auxiliaires, la médiane étant de 0.078, le quartile supérieur de 0.107 et le maximum de 0.329, résultats qui concordent également grossièrement avec la valeur de 0.087 prévue par le modèle, conformément à l'équation (5). Cette interprétation dépend manifestement de la validité du modèle (2), laquelle est douteuse pour ces données, mais elle donne à penser que la sélection des variables permettrait de réaliser des gains d'efficacité.

Une autre façon d'évaluer les gains d'efficacité éventuels dus à la sélection des variables consiste à calculer des approximations de la variance de l'estimateur de régression en choisissant divers sous-ensembles des variables auxiliaires disponibles et en se servant de tous les enregistrements obtenus pour la population. La figure 1 représente graphiquement l'approximation que donne une version pour population finie de l'équation (5) calculée pour des sous-ensembles croissants des dix variables auxiliaires, où la variable ajoutée à chaque étape est celle qui produit la diminution la plus forte de l'approximation. Ces valeurs de l'approximation type de premier ordre axée sur le plan de sondage $(1-f)S_e/n$ sont également représentées graphiquement à titre de référence, quoique, comme nous l'avons déjà souligné, cette approximation soit monotone, n'augmentant pas quand on ajoute de nouvelles variables auxiliaires. Les estimations par simulation de l'erreur quadratique moyenne de l'estimateur de régression correspondant à chaque sous-ensemble sont également représentées graphiquement. Le graphique montre clairement que si on choisit d'utiliser un estimateur de régression type contenant un ensemble fixe de variables auxiliaires, le sous-ensemble contenant cinq prédicteurs est le meilleur choix dans le cas de l'approximation normale de la variance fondée sur l'expression (5), tandis que le sous-ensemble saturé est plus indiqué dans le cas de l'approximation de la variance fondée sur le plan de sondage. Le graphique montre aussi que les estimations par simulation de l'erreur quadratique moyenne concordent davantage avec le modèle d'approximation normal qu'avec l'approximation type de premier ordre, particulièrement pour les grands sous-ensembles de variables auxiliaires. Nous obtenons des résultats comparables quand nous calculons les approximations correspondantes de la variance pour l'ensemble de cinq variables auxiliaires.

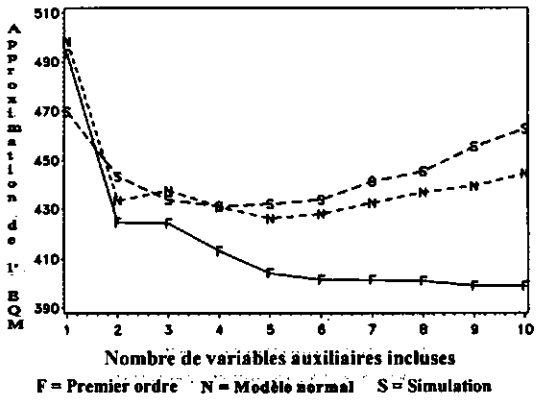


Figure 1. Approximations et estimations par simulation de l'EQM de l'estimateur de régression calculée, dans le cas d'une population finie, pour des sous-ensembles croissants de dix variables auxiliaires.

Donc, la distribution de c_g^2 obtenue par simulation de même que les approximations pour une population finie de la variance de l'estimateur de régression indiquent qu'on pourrait réaliser des gains d'efficacité en sélectionnant les variables pour cette population. Afin de déterminer si cette constatation s'applique à nos données, nous décrivons maintenant en détail l'étude en simulation.

Pour chaque échantillon répété (représenté par s) et pour chacun des deux sous-ensembles de variables auxiliaires considérés, nous avons calculé les estimations de la moyenne de population du revenu mensuel total, ainsi que les estimations correspondantes de la variance, en appliquant plusieurs stratégies d'estimation. Chaque stratégie est définie par la combinaison d'une méthode de sélection d'un sous-ensemble, d'un estimateur de la moyenne et d'un estimateur de la variance correspondante. Voici la liste des stratégies étudiées.

SM) Estimateur de la moyenne de l'échantillon, sans variables auxiliaires $(\bar{y}, v_s)$. Cette stratégie représente la norme à laquelle seront comparées toutes les autres.

Fs) Sélection progressive de variables auxiliaires avec $(\bar{y}_r, v_s)$.

Fd) Sélection progressive de variables auxiliaires avec $(\bar{y}_r, v_d)$.

Fg) Sélection progressive de variables auxiliaires avec $(\bar{y}_r, v_g)$.

Bs) Sélection du meilleur sous-ensemble possible de variables auxiliaires avec $(\bar{y}_r, v_s)$.

Bd) Sélection du meilleur des sous-ensemble possible de variables auxiliaires avec $(\bar{y}_r, v_d)$.

Bg) Sélection du meilleur sous-ensemble possible des variables auxiliaires avec $(\bar{y}_r, v_g)$.

FI) Sous-ensemble fixé de variables auxiliaires avec $(\bar{y}_r, v_s)$.

SS) Sous-ensemble saturé de variables auxiliaires avec $(\bar{y}_r, v_s)$.

FR) Sélection progressive d'un sous-ensemble au moyen de SAS PROC REG, avec $(\bar{y}_r, v_s)$.

CN) Sélection d'un sous-ensemble par réduction du nombre de conditions avec $(\bar{y}_r, v_s)$.

RI) Estimateur de régression ridge avec un sous-ensemble saturé de variables auxiliaires et un estimateur de variance, que nous représentons par v_{DC}, proposé par Dunstan et Chambers 1986), $(\bar{y}_{BC}, v_{DC})$.

Les stratégies Fs à Bg sont des variantes de deux méthodes que nous avons proposées pour sélectionner des sous-ensembles qui découlent de l'utilisation des trois estimateurs de l'erreur quadratique moyenne examinés à la section 3. Les stratégies FI et SS se fondent sur le même ensemble de variables auxiliaires quel que soit l'échantillon sélectionné. Dans le cas de SS, le sous-ensemble saturé englobant toutes les variables auxiliaires disponibles est utilisé constamment. Dans le cas de FI, nous avons choisi un sous-ensemble à partir de l'ensemble de cinq variables auxiliaires (x_1, x_4, x_{11} choisies) ou de dix variables auxiliaires ($x_1, x_2, x_5, x_8, x_{10}$ choisies) considéré, en appliquant une méthode de régression progressive type à l'ensemble de données sur la population. Puis, pour chaque échantillon, nous nous sommes servis du sous-ensemble sélectionné, d'où le nom de stratégie du «sous-ensemble fixé» pour FI. En pratique, cette stratégie n'est pas applicable, car on ne dispose pas des données sur la population pour la variable dépendante, mais nous la considérons théoriquement en tant que «meilleur scénario possible» dans le cadre de la méthode traditionnelle.

Dans le cas de la stratégie FR, nous utilisons «naïvement» SAS PROC REG pour exécuter la sélection progressive type d'un sous-ensemble pour chaque échantillon. La *valeur prédictive p* utilisée pour décider si une nouvelle variable devrait être incluse est la valeur implicite du programme, à savoir 0.50. Pour plus de détails, consulter SAS (1990, p. 1397).

Dans le cas de la sélection de sous-ensembles en vue de réduire le nombre de conditions CN, nous fixons à 1,000 la valeur du paramètre L qui contrôle la méthode. Pour la stratégie d'estimation par régression ridge RI, nous fixons à 1 la valeur de tous les coefficients de coût associés aux erreurs d'étalonnage qui entachent les diverses variables. Après avoir choisi la valeur de λ garantissant qu'aucun poids ne soit inférieur à $1/N$, nous rééchelonnons les poids de sorte que leur somme soient exactement 1, pour être certains que l'étalonnage soit exact au moment de l'estimation de la taille de la population.

Quelle que soit la stratégie utilisée, nous représentons par $\bar{y}(s)$ et $v[\bar{y}(s)]$ les estimations de la moyenne de population et de son erreur quadratique moyenne pour l'échantillon s, respectivement. Pour chaque stratégie, nous résumons les résultats de la simulation en calculant les estimations du biais, de l'erreur quadratique moyenne (EQM) et de la moyenne des estimations de l'erreur quadratique moyenne (MEEQM) pour l'ensemble des 1 000 échantillons répétés. Ces mesures sont données respectivement par

$$\text{BIAIS} = \sum_s [\bar{y}(s) - \bar{Y}]/1{,}000 \qquad (12)$$

$$\text{EQM} = \sum_s [\bar{y}(s) - \bar{Y}]^2/1{,}000 \qquad (13)$$

$$\text{MEEQM} = \sum_s v[\bar{y}(s)]/1{,}000. \qquad (14)$$

Pour chaque stratégie, nous calculons également une mesure de l'efficacité en divisant l'erreur quadratique moyenne obtenue pour la simulation correspondante par l'erreur quadratique moyenne obtenue pour la simulation selon la stratégie de la moyenne de l'échantillon (stratégie SM) et en multipliant le résultat par 100. Nous calculons aussi, pour chaque stratégie les taux de couverture empiriques pour les intervalles de confiance de 95% fondés sur la théorie normale asymptotique. Ces taux, exprimés en pourcentage, sont présentés dans les dernières colonnes des tableaux 1 et 2.

Le Tableau 1 montre les résultats de la simulation visant à estimer la moyenne de la variable dépendante en se fondant sur l'ensemble formé des cinq variables auxiliaires ($x_1 - x_4, x_{11}$) ayant le pouvoir de prédiction le plus grand. Dans ce cas, l'utilisation de l'estimateur de régression améliore considérablement la précision de chaque stratégie d'estimation étudiée, sauf celle consistant à sélectionner un sous-ensemble par réduction du nombre de conditions (CN). Le biais est négligeable (inférieur à 1% en ce qui concerne le biais relatif absolu) pour toutes les stratégies d'estimation (la moyenne de y est 194.34), sauf, peut-être, la stratégie RI, pour laquelle on observe un léger biais.

Les résultats sont les mêmes pour les stratégies fondées sur la sélection progressive d'un sous-ensemble (Fs-Fg) et pour les stratégies correspondantes fondées sur la sélection à partir de tous les sous-ensembles possibles (Bs-Bg). Par conséquent, il est préférable d'appliquer les méthodes de sélection progressives, qui sont plus rapides et moins coûteuses.

Parmi les stratégies fondées sur la sélection progressive d'un sous-ensemble, Fd et Fg (avec v_d et v_g comme estimateur de l'erreur quadratique moyenne, respectivement) sont les plus efficaces et produisent des résultats fort semblables. Il convient aussi de souligner que Fd et Fg donnent de meilleurs résultats que FI et SS, à savoir les stratégies axées sur l'estimateur de régression s'appuyant sur un sous-ensemble fixe de cinq variables auxiliaires pour chaque échantillon. Cette observation est vraie dans le cas du sous-ensemble saturé (SS), ainsi que dans celui du sous-ensemble fixe choisi d'après des renseignements tirés de l'ensemble de la population (FI). Donc, il est possible d'obtenir de meilleurs résultats que ceux donnés par la méthode traditionnelle, c'est-à-dire l'utilisation d'un estimateur de régression avec un sous-ensemble fixé de variables auxiliaires, en adoptant une méthode adaptative consistant à choisir le «meilleur» estimateur de régression (sous-ensemble) pour chaque échantillon, du moins quand la variable dépendante étudiée est celle considérée pour la sélection du sous-ensemble. C'est la forte variation de la valeur c_g^2 d'un échantillon à l'autre, situation dans laquelle

Tableau 1
Biais, erreur quadratique moyenne, moyenne des estimations de l'erreur quadratique moyenne et efficacité de diverses stratégies d'estimation de la moyenne de la variable dépendante y quand on dispose de cinq variables auxiliaires ($x_1 - x_4, x_{11}$)

Stratégie d'estimation		BIAIS	EQM	MEEQM	Efficacité rapport à SM (%)	Couverture[1] empirique (%)
SM)	Moyenne de l'échantillon $(\bar{y},v_s)$	0.25	620.09	619.05	100.00	91.8
Fs)	Progressive $(\bar{y}_r,v_s)$	0.4	233.78	239.62	37.7	82.7
Fd)	Progressive $(\bar{y}_r,v_d)$	-1.25	188.08	196.88	30.33	82.0
Fg)	Progressive $(\bar{y}_r,v_g)$	-1.28	188.38	192.73	30.38	81.1
Bs)	Meilleur $(\bar{y}_r,v_s)$	0.44	236.9	239.49	38.2	82.7
Bd)	Meilleur $(\bar{y}_r,v_d)$	-1.22	190.52	196.84	30.72	82.0
Bg)	Meilleur $(\bar{y}_r,v_g)$	-1.24	190.83	192.71	30.77	81.1
FI)	Fixe $(\bar{y}_r,v_s)$	0.29	227.90	241.24	36.75	83.3
SS)	Saturé $(\bar{y}_r,v_s)$	0.3	233.58	242.32	37.67	82.5
FR)	PROC REG $(\bar{y}_r,v_s)$	0.38	235.86	240.26	38.04	82.5
CN)	Réd. nbre. Cond. $(\bar{y}_r,v_s)$	0.34	507.33	483.63	81.82	89.8
RI)	Ridge $(\bar{y}_{BC},v_{DC})$	2.12	304.95	257.07	49.18	82.5

[1] Couverture nominale de 95%.

résultats devraient, en principe, être meilleurs si on choisit un plus petit nombre de variables x pour les échantillons affichant les valeurs de c_g^2 les plus grandes, qui a laissé entrevoir cette propriété.

La comparaison avec la stratégie adaptative FR, basée sur la sélection type de sous-ensembles offerte par PROC REG de SAS, montre que le fait d'utiliser pour critère un estimateur approprié de l'erreur quadratique moyenne de l'estimateur de régression améliore les résultats. L'efficacité de FR est semblable à celle des méthodes traditionnelles basées sur un sous-ensemble fixe (FI-SS).

La faible efficacité de la méthode de sélection d'un sous-ensemble par réduction du nombre de conditions (CN) comparativement à celle de toutes les méthodes basées sur l'estimateur de régression est un résultat plus frappant, mais toutefois pas inattendu, car la méthode ne tient pas compte de la variable dépendante. Cette observation donne du poids à l'argument selon lequel, si la moyenne d'une variable dépendante particulière est la principale cible d'inférence, il convient d'en tenir compte quand on sélectionne les variables auxiliaires à utiliser avec l'estimateur de régression.

Dans le cas de l'ensemble de cinq variables auxiliaires, nous notons aussi que, pour chaque échantillon, la première variable éliminée pour réduire le nombre de conditions est l'approximation du revenu (x_{11}). Cette situation est due au fait que les valeurs propres (donc, le nombre de conditions) de la matrice CP dépendent des unités de mesure des variables auxiliaires. Puisque toutes les autres variables auxiliaires sont des dénombrements d'une sorte ou d'une autre, l'approximation du revenu est la variable affichant, de loin, la variance la plus forte. Son exclusion de chaque échantillon explique, en partie, la performance médiocre de cette méthode, puisqu'il s'agit du meilleur prédicteur unique de la variable dépendante.

Les travaux de Bankier ne font pas ressortir cette difficulté, car, dans le cas des données d'échantillon du Recensement de la population du Canada de 1991 auxquelles il applique sa méthode, toutes les variables auxiliaires étudiées sont des dénombrements de personnes, de familles ou de ménages, exprimées dans des unités comparables.

Contrairement aux valeurs propres de la matrice CP, l'estimateur de régression ne dépend ni de l'emplacement ni de la transformation d'échelle des variables auxiliaires. Donc, pour éviter que la méthode du nombre de conditions dépende arbitrairement des unités des variables auxiliaires, il est naturel de commencer par normaliser ces variables, puis de calculer le nombre de conditions de la matrice de corrélation de l'échantillon $\hat{R}_x$ plutôt que de la matrice $X_s^{*\prime}X_s^*$. Cependant, lors de l'essai de cette méthode, même le choix de valeurs modestes pour L (100) n'a abouti à l'élimination d'aucune variable auxiliaire. Par conséquent, l'ensemble saturé a été utilisé chaque fois, réduisant ainsi CN à SS.

La stratégie fondée sur l'estimateur de régression ridge (RI) donne de moins bons résultats que celle basée sur le sous-ensemble saturé (SS) en ce qui concerne l'efficacité. En outre, elle produit une estimation de l'erreur quadratique moyenne entachée d'un certain biais. Cette perte d'efficacité tient au fait que tous les poids doivent être supérieurs ou égaux à $1/N$, condition imposée uniquement dans ce cas-ci. En revanche, cette stratégie est beaucoup plus efficace que celle consistant à réduire le nombre de conditions (CN).

En ce qui concerne les taux de couverture empiriques, seule la stratégie de réduction du nombre de conditions CN donne des résultats proches de ceux obtenus par la méthode de la moyenne de l'échantillon (SM), chacune produisant un légère sous-couverture. Toutes les autres méthodes basées sur l'estimation par régression donne des taux de couverture comparables, nettement inférieurs à la cible de 95%.

Les résultats de la simulation effectuée avec l'ensemble de dix variables auxiliaires ($x_1 - x_{10}$) sont présentés au tableau 2 ci-dessous. Comme prévu, ces résultats indiquent que les méthodes basées sur l'estimateur de régression sont plus efficaces que la méthode de la moyenne de l'échantillon. Cependant, les gains d'efficacité ne sont pas aussi importants que ceux indiqués au tableau 1, dans le cas de cinq variables auxiliaires ayant un plus grand pouvoir explicatif. Comme auparavant, les méthodes adaptatives basées sur la sélection progressive d'un sous-ensemble donnent des résultats comparables à ceux des méthodes fondées sur la sélection du meilleur sous-ensemble parmi tous les sous-ensembles possibles. De nouveau, les stratégies adaptatives utilisant v_d ou v_g comme estimateur de l'erreur quadratique moyenne sont un peu plus efficaces que les stratégies correspondantes basées sur v_s, quoique, dans ce cas, au prix d'une sous-couverture plus grande des intervalles de confiance nominaux de 95%.

Les stratégies adaptatives d'estimation les plus efficaces (Fd, Fg, Bd et Bg) mènent à une moyenne de la population et à une erreur quadratique moyenne entachées d'un biais non négligeable. En revanche, les stratégies FI et SS ne produisent aucun biais significatif pour la moyenne, même si l'estimation de l'erreur quadratique moyenne est entachée d'un certain biais dans le cas de la stratégie SS. Il convient de souligner le biais négatif important qui entachent les estimateurs de l'erreur quadratique moyenne, lequel est illustré par les différences entre les valeurs des colonnes intitulées EQM et MEEQM dans le tableau 2. Le pire biais semble être produit par les stratégies Fd, Fg, Bd et Bg. Viennent ensuite les stratégies Fs et Bs, puis les stratégies SS, FR et CN pour lesquelles le biais n'est pas si mauvais.

La comparaison de Fd et de Fg à CN montre que les deux premières méthodes aboutissent à un gain d'efficacité comparativement à la méthode de réduction du nombre de conditions, au prix d'une certaine augmentation du biais qui entache aussi bien l'estimateur de la moyenne que celui de l'erreur quadratique moyenne. Donc, même quand le pouvoir de prédiction des variables auxiliaires disponibles est grand, il est possible d'adopter une stratégie plus efficace que la stratégie CN.

Le choix d'un sous-ensemble fixe inapproprié (par exemple, le sous-ensemble saturé utilisé pour la stratégie SS) pourrait aboutir à des résultats médiocres au chapitre de l'efficacité et biaiser dans une certaine mesure l'estimation de l'erreur quadratique moyenne. Cependant, si, par exemple, on utilisait v_d au lieu de v_s comme estimateur de l'erreur quadratique moyenne dans le cas de la stratégie SS, on ne noterait aucun biais apparent (la MEEQM observée dans ce cas est 459.67, donc une valeur beaucoup plus proche de celle de l'estimation de l'erreur quadratique moyenne de la simulation, soit 462.71).

De nouveau, l'estimateur de régression ridge donne des résultats légèrement inférieurs à ceux de la stratégie du sous-ensemble saturé (SS), mais, cette fois-ci, sans aucun biais évident entachant l'estimation de la moyenne ou de l'erreur quadratique moyenne. Une fois de plus, la régression ridge s'avère plus efficace que la stratégie de réduction du nombre de conditions CN, mais l'écart entre les deux méthodes est plus faible ici. Elle donne également de bons résultats en ce qui concerne la couverture empirique de l'intervalle de confiance.

La stratégie FR donne de nouveau des résultats comparables à ceux des stratégies à sous-ensemble fixe FI et SS, donc, est surpassée par les stratégies fondée sur un estimateur de l'erreur quadratique moyenne de l'estimateur de régression tel que v_d ou v_g.

Tableau 2

Biais, erreur quadratique moyenne, moyenne des estimations de l'erreur quadratique moyenne et efficacité de diverses stratégies d'estimation de la moyenne de la variable dépendante y quand on dispose de dix variables auxiliaires ($x_1 - x_{10}$)

Stratégie d'estimation		BIAIS	EQM	MEEQM	Efficacité par rapport à SM (%)	Couverture[1] empirique (%)
SM)	Moyenne de l'échantillon $(\bar{y}, v_s)$	0.25	620.09	619.05	100.00	91.8
Fs)	Progressive $(\bar{y}_r, v_s)$	0.06	468.46	397.99	75.55	86.7
Fd)	Progressive $(\bar{y}_r, v_d)$	-8.12	434.27	338.90	70.03	81.7
Fg)	Progressive $(\bar{y}_r, v_g)$	-7.90	433.71	328.46	69.94	81.6
Bs)	Meilleur $(\bar{y}_r, v_s)$	-0.00	466.16	397.59	75.18	86.6
Bd)	Meilleur $(\bar{y}_r, v_d)$	-7.90	434.54	336.88	70.08	81.5
Bg)	Meilleur $(\bar{y}_r, v_g)$	-7.60	433.26	326.05	69.87	81.6
FI)	Fixe $(\bar{y}_r, v_s)$	0.45	490.49	461.86	79.10	89.0
SS)	Saturé $(\bar{y}_r, v_s)$	-0.20	462.71	413.17	74.62	86.9
FR)	PROC REG $(\bar{y}_r, v_s)$	-0.07	466.13	399.34	75.17	86.4
CN)	Réd. Nbre. Cond. $(\bar{y}_r, v_s)$	3.49	562.91	450.36	90.78	87.3
RI)	Ridge $(\bar{y}_{BC}, v_{DC})$	1.05	480.18	472.82	77.44	89.4

[1] Couverture nominale de 95%.

Les résultats donnent à penser que, quand on estime la moyenne de la population d'une variable dépendante unique, les méthodes adaptatives proposées combinant l'estimateur de régression et une certaine forme de sélection d'un sous-ensemble de variables s'appuyant sur un estimateur approprié de l'erreur quadratique moyenne produisent un gain utile d'efficacité comparativement aux méthodes concurrentes. Cependant, ces stratégies risquent d'introduire un certain biais quand le pouvoir de prédiction des variables auxiliaires disponibles est faible et les estimateurs correspondants de l'erreur quadratique moyenne peuvent être biaisés considérablement, situation qui se traduit par une mauvaise couverture.

7. CONCLUSIONS ET ORIENTATIONS FUTURES

Nos résultats laissent entendre que, dans le cas de l'estimation par régression, on peut réaliser des gains d'efficacité en adoptant une méthode de sélection des variables fondées sur un des estimateurs de l'erreur quadratique moyenne v_d ou v_g. Dans le cas de la méthode de régression type, et compte tenu des renseignements limités fournis par la simulation, on ne dispose que de peu d'indices permettant de choisir entre ces deux estimateurs.

Les méthodes de sélection progressives d'un sous-ensemble de variables sont aussi efficaces que celles basées sur l'examen de tous les sous-ensembles possibles qui demandent beaucoup plus de calculs. Nos résultats indiquent également qu'il est possible d'améliorer la méthode de sélection d'un sous-ensemble par réduction du nombre de conditions quand on étudie une variable dépendante particulière.

Un des problèmes que pose la stratégie de sélection des variables tient au fait que l'estimation connexe de la variance est probablement biaisée pour l'estimation de l'erreur quadratique moyenne globale de l'estimateur de régression après la sélection des variables, donc, menant à une mauvaise couverture des intervalles de confiance types. Il convient donc de poursuivre les travaux en vue d'étudier d'autres méthodes d'estimation de la variance.

Le présent article porte principalement sur l'utilisation de l'estimation par régression en vue de réduire la variance d'échantillonnage dans le contexte de l'échantillonnage classique. En pratique, on utilise largement l'estimation par régression pour corriger les biais causés par les erreurs non dues à l'échantillonnage. Dans de telles applications, la question de savoir combien de variables auxiliaires il faut utiliser est également importante. Certaines variables pourraient être incluses pour des raisons indépendantes de l'erreur d'échantillonnage, par exemple, parce qu'on sait qu'elles sont des déterminants importants de la non-réponse. Néanmoins, la variance d'échantillonnage pourrait augmenter à mesure que croît le nombre de variables auxiliaires et, selon nous, il serait utile d'établir une règle de décision pour limiter le nombre de variables auxiliaires employées en tenant compte de la variance d'échantillonnage. En cas de biais non dû à l'échantillonnage, l'écart entre $\bar{x}$ et $\bar{X}$ sera généralement donné par $O_p(1)$ plutôt que par $O_p(n^{-1/2})$ et, par conséquent, les résultats présentés ici ne sont pas directement applicables. Il conviendra donc d'étudier l'extension de notre méthode à ce cas.

D'autres travaux doivent aussi être effectués pour étendre notre méthode aux plans de sondage complexes. Une méthode que l'on pourrait envisager pour les estimateurs généraux de régression, méthode considérée, par exemple, par Särndal et coll. (1992, sec. 6.4), consisterait à remplacer les poids g_i par les poids «généralisés», décrits par Särndal et coll. (1992, éq. 6.5.9), et à fonder la sélection des variables sur la minimisation de la version généralisée de v_g donnée par Särndal et coll. (1992, éq. 6.6.4).

Tableau 3
Matrice de corrélation pour les variables utilisées dans l'étude en simulation basée sur les données du Recensement de la population de 1988

Variable	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	0.23										
x_2	-0.04	0.2									
x_3	0.17	0.07	-0.40								
x_4	0.47	0.13	-0.15	0.12							
x_5	0.48	0.09	-0.11	0.15	0.83						
x_6	0.05	-0.09	-0.32	-0.03	0.22	0.20					
x_7	-0.17	0.01	-0.12	-0.01	-0.17	-0.31	0.16				
x_8	0.38	0.29	0.07	0.17	0.44	0.41	0.13	-0.20			
x_9	0.20	0.08	-0.06	0.04	0.30	0.25	0.16	-0.13	0.37		
x_{10}	0.43	0.23	0.33	0.17	0.39	0.39	-0.10	-0.30	0.49	0.26	
x_{11}	0.78	0.23	-0.00	0.22	0.54	0.54	0.01	-0.19	0.41	0.21	0.49

REMERCIEMENTS

Pedro L.D. Nascimento Silva exprime sa gratitude à CVCP-UK, CNPq-Brasil et IBGE-Brasil pour leur appui financier. Les auteurs remercient Ray Chambers, Danny Pfeffermann, Jon Rao, Michael Bankier et deux arbitres anonymes pour leurs commentaires. Michael Bankier a également offert un appui précieux en fournissant la documentation et le logiciel concernant sa méthode GLSEP.

BIBLIOGRAPHIE

BANKIER, M.D. (1990). Two Step Generalized Least Squares Estimation. Ottawa: Statistique Canada, Division des méthodes d'enquêtes sociales, rapport interne.

BANKIER, M.D., RATHWELL, S., et MAJKOWSKI, M. (1992). Two Step Generalized Least Squares Estimation in the 1991 Canadian Census. Ottawa: Statistique Canada, document de travail, DMES, 92-007E.

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

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

DENG, L.Y., et WU, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82, 568-576.

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

DUNSTAN, R., et CHAMBERS, R.L. (1986). Model-based confidence intervals in mltipurpose surveys. *Applied Statistics*, 35, 276-280.

GRIMES, J.E., et SUKHATME, B.V. (1980). A regression-type estimator based on preliminary test of significance. *Journal of the American Statistical Association*, 75, 957-962.

HANSEN, M.H., et TEPPING, B.J. (1969). Progress and problems in survey methods and theory illustrated by the work of the United States Bureau of the Census. *New Developments in Survey Sampling*, (N.L. Johnson et H. Smith Jr., Éds.). New York: John Wiley & Sons.

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

MARDIA, K.V., KENT, J.T., et BIBBY, J.M. (1979). *Multivariate Analysis*. London: Academic Press.

MILLER, A.J. (1990). *Subset Selection in Regression*. London: Chapman and Hall.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications* (2ième éd.). New York: John Wiley & Sons.

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

SAS INSTITUTE INC. (1990). *SAS/STAT User's Guide* (Version 6, Vol. 2, 4ième éd.). Cary, NC: SAS Institute Inc.

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

SILVA, P.L.D.N. (1996). Some Asymptotic Results on the Mean Squared Error of the Regression Estimator Under Simple Random Sampling Without Replacement. Southampton: University of Southampton, Centre for Survey Data Analysis Technical Report 96-2.

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 37-45
Statistique Canada

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

JOHN L. ELTINGE et IBRAHIM S. YANSANEH[1]

RÉSUMÉ

Les auteurs décrivent certaines méthodes diagnostiques simples utilisées pour guider la construction de cellules de correction pour la non-réponse. S'inspirant des travaux de Little (1986), ils étudient la construction de cellules de correction par regroupement d'unités d'échantillonnage selon la probabilité estimée de réponse ou selon la réponse estimée aux questions de l'enquête. Ils examinent plus particulièrement l'évaluation de la sensibilité des estimations corrigées de la moyenne à la variation de k, c'est-à-dire le nombre de cellules utilisées, le dépistage de cellules particulières qui nécessitent une mise au point supplémentaire, la comparaison des estimations corrigées et non corrigées de la moyenne et la comparaison des estimations obtenues au moyen des cellules fondées sur la probabilité estimée de réponse, d'une part, et sur la réponse estimée aux questions, d'autre part. Les auteurs justifient les méthodes proposées et les illustrent par une application à l'estimation du revenu moyen des unités de la U.S. Consumer Expenditure Survey.

MOTS CLÉS: Données incomplètes; données manquantes; quasi-randomisation; propension à répondre; analyse de sensibilité; correction par pondération.

1. INTRODUCTION

1.1 Énoncé du problème

Les analystes d'enquête recourent souvent à la construction de cellules de correction pour tenir compte de la non-réponse. L'idée générale consiste à définir d'abord des groupes, ou «cellules», d'unités d'échantillonnage qu'on estime présenter à peu près la même probabilité de réponse ou produire à peu près la même valeur pour une question particulière, comme celle sur le revenu, puis, à corriger par pondération ou à effectuer une simple imputation «hot-deck» dans chaque cellule de correction. L'estimateur corrigé obtenu d'une moyenne ou d'un total de la population est alors entaché d'un biais dû à la non-réponse approximativement nul, à condition que les covariances intracellulaires entre les réponses aux questions et les probabilités de réponse soient approximativement nulles.

Certains travaux antérieurs sur la correction pour la non-réponse consistaient à créer des cellules de correction en groupant des variables de classification démographiques ou géographiques simples. Cependant, Little (1986) et d'autres chercheurs ont étudié la construction de cellules par groupement direct d'unités d'échantillonnage selon la probabilité estimée de réponse ou selon la valeur estimée des réponses. Dans le présent article, nous examinons certaines méthodes diagnostiques simples qui facilitent l'application de ces idées à la création de cellules. Nous nous concentrons surtout sur la sensibilité des résultats au nombre de cellules utilisées, au dépistage de cellules particulières qui nécessitent une mise au point supplémentaire, à la comparaison des estimations corrigées et non corrigées de la moyenne et à la comparaison des estimations obtenues d'après les cellules fondées sur les probabilités estimées, d'une part, et sur les réponses estimées aux questions, d'autre part. Nous illustrons ces méthodes diagnostiques grâce aux données sur le revenu collectées dans le cadre de la U.S. Consumer Expenditure Survey.

1.2 Notation, biais dû à la non-réponse et cellules de correction

Représentons par U une population donnée de taille N et les questions d'enquête par Y_i, $i \in U$; et considérons l'estimation de la moyenne de la population $\bar{Y} = N^{-1}\sum_{i\in U} Y_i$. Tirons un échantillon s de taille n de la population U et représentons par π_i la probabilité que l'unité i soit incluse dans l'échantillon.

Supposons que la non-réponse satisfait le modèle de quasi-randomisation qui suit (Oh et Scheuren 1983). Posons que R_i est une variable indicatrice égale à 1 si l'unité d'échantillonnage i choisie est un répondant et égale à 0, autrement. Enfin, supposons que les R_i sont des variables aléatoires de Bernoulli (η_i) mutuellement indépendantes, pour lesquelles les probabilités fixes de réponse η_i peuvent varier d'une unité à l'autre. En outre, définissons les poids de sondage $\lambda_i = \pi_i^{-1}$ et la réponse moyenne non corrigée pondérée selon le plan de sondage

$$\hat{\bar{Y}}_1 \stackrel{\text{def}}{=} \left(\sum_{i\in s}\lambda_i R_i\right)^{-1}\sum_{i\in s}\lambda_i R_i Y_i. \qquad (1.1)$$

À cause d'écarts entre les η_i, l'estimateur non corrigé $\hat{\bar{Y}}_1$ est entaché d'un biais dû à la non-réponse approximativement égal à $N^{-1}\bar{\eta}^{-1}\sum_{i\in U}\eta_i(Y_i - \bar{Y})$, où $\bar{\eta} = N^{-1}\sum_{i\in U}\eta_i$ et où les espérances mathématiques sont déterminées à la fois sur le plan de sondage original et sur le modèle de quasi-randomisation. Pour réduire ce biais, on partage souvent la population

[1] John L. Eltinge, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A.; Ibrahim S. Yansaneh, Westat, 1650 Research Blvd., Rockville, MD 20850-3195, U.S.A.

en k «cellules de correction» U_h, et l'échantillon s, en groupes correspondants s_h, puis on utilise l'estimateur corrigé

$$\hat{\bar{Y}}_k \stackrel{\text{def}}{=} \sum_{h=1}^{k} w_h \bar{Y}_{hR}, \tag{1.2}$$

où $w_h = (\sum_{i\in s} \lambda_i)^{-1} \sum_{i\in s_h} \lambda_i$ et $\bar{Y}_{hR} = (\sum_{i\in s_h} \lambda_i R_i)^{-1} \sum_{i\in s_h} \lambda_i R_i Y_i$. Soulignons que, si $k = 1$, alors les estimateurs (1.1) et (1.2) sont identiques. Pour une discussion générale des méthodes basées sur les cellules de correction, consulter, par exemple, Cassel, Särndal et Wretman (1983), Oh et Scheuren (1983), et Kalton et Maligalig (1991).

L'estimateur corrigé $\hat{\bar{Y}}_k$ est entaché d'un biais résiduel dû à la non-réponse approximativement égal à

$$N^{-1} \sum_{h=1}^{k} \bar{\eta}_h^{-1} \sum_{i\in U_h} (\eta_i - \bar{\eta}_h)(Y_i - \bar{Y}_h), \tag{1.3}$$

où N_h représente le nombre d'unités dans U_h et où $(\bar{\eta}_h, \bar{Y}_h) = N_h^{-1} \sum_{i\in U_h} (\eta_i, Y_i)$. Par conséquent, on préfère construire des cellules telles que la covariance entre η_i et Y_i est approximativement nulle dans chaque cellule. En pratique, on s'efforce de le faire en construisant des cellules qui sont approximativement homogènes en regard des probabilités de réponse η_i ou des réponses aux questions Y_i, ou des deux. Dans certains cas, on définit *a priori* des ensembles «naturels» de cellules grâce à des combinaisons de variables de classification connues tant pour les répondants que pour les non-répondants. Par exemple, Ezzati et Khare (1992) utilisent 72 cellules définies selon l'âge, la race, la religion, la situation d'urbanisation et la taille du ménage pour apporter des corrections pour la non-réponse à une partie des données de la National Health and Nutrition Examination Survey. Toutefois, fort souvent, en pratique, la liste des variables susceptibles d'être utilisées pour construire les cellules est assez longue, ce qui peut produire un nombre considérable de cellules contenant peu de répondants, voire aucun. Par conséquent, plusieurs auteurs ont mis au point des méthodes permettant de déterminer les variables de classification les moins importantes et à regrouper les cellules de correction peu peuplées de façon à ce que chaque cellule retenue soit raisonnablement homogène. Consulter, par exemple, Tremblay (1986), Lepkowski, Kalton et Kasprzyk (1989), Kalton et Maligalig (1991), Goskel, Judkins et Mosher (1991) et la discussion connexe sur le regroupement des strates a posteriori de Little (1993). De surcroît, les méthodes axées sur la création de cellules de correction sont apparentées à d'autres, comme les méthodes de correction fondées sur la régression [consulter Rao (1996, section 2.4) et les auteurs qu'il cite] et la méthode itérative généralisée (Deville, Särndal et Sautory 1993).

1.3 Cellules de correction fondées sur l'estimation de la propension à répondre ou sur les réponses prévues

Comme, en principe, les cellules de correction sont approximativement homogènes, on pourrait argumenter que de telles cellules définissent implicitement un modèle pour les valeurs de η_i ou de Y_i, ou des deux. Une modélisation plus explicite mène à deux méthodes connexes de création des cellules. En premier lieu, représentons par X_i un vecteur de variables auxiliaires observées pour les unités d'échantillonnage i répondantes ainsi que non répondantes et servons-nous des valeurs observées dans l'échantillon (R_i, X_i) pour ajuster un modèle pour $\eta_i = \eta(X_i)$ par régression linéaire, logistique ou probit. Puis, construisons les cellules de l'échantillon s_h en groupant les unités d'échantillonnage selon la probabilité estimée de réponse $\hat{\eta}_i$. Comme deuxième solution, considérons la régression des réponses Y_i en fonction d'un vecteur auxiliaire X_i pour produire les réponses estimées $\hat{Y}_i$ pour les unités d'échantillonnage répondantes ainsi que non répondantes. Puis, construisons les cellules de l'échantillon s_h en groupant les unités selon les valeurs de $\hat{Y}_i$.

Ces deux méthodes ont été proposées par Little (1986) comme extension des travaux de Rosenbaum et de Rubin (1983, 1984) sur les scores de propension calculés d'après des données d'observation. Consulter aussi David, Little, Samuhel et Triest (1983). Au départ, les méthodes ont été élaborées dans le contexte d'un modèle, mais elles s'étendent directement au cadre de référence actuel. Selon Little (1986), l'utilisation de cellules fondées soit sur les valeurs de $\hat{\eta}_i$, soit sur celles de $\hat{Y}_i$, pourrait réduire le biais dû à la non-réponse et celle des cellules fondées sur $\hat{Y}_i$ permettrait aussi de contrôler la variance. En outre, dans certains cas, la méthode des cellules fondées sur $\hat{\eta}_i$ et sur $\hat{Y}_i$ est plus souple que celle des cellules définies *a priori*. De surcroît, les cellules de correction fondées sur $\hat{Y}_i$ sont reliées conceptuellement aux notions de stratification optimale (consulter, par exemple, Cochran 1977, sections 5A.7 et 5A.8).

Little (1986) ne propose pas de règle particulière pour déterminer la construction des cellules. Cependant, en s'inspirant des travaux connexes sur les données d'observation effectuées par Cochran (1968) et par Rosenbaum et Rubin (1984), on pourrait envisager de répartir les unités en cellules définies en se fondant sur les quantiles $k^{-1}j$ estimés des populations de $\hat{\eta}_i$ ou de $\hat{Y}_i$ où $j = 1, 2, ..., k-1$. Cette méthode des quantiles égaux permet de contrôler dans une certaine mesure le nombre prévu de répondants dans chaque cellule. En outre, la lecture des deux références susmentionnées donne à penser que, pour un ensemble donné de prédicteurs X_i, on peut réaliser la plupart de la réduction faisable du biais grâce à un nombre relativement faible de cellules, disons $k = 5$. Une étude de cas effectuée par Czajka, Hirabayashi, Little et Rubin (1992) comprend la construction de $k = 6$ cellules de correction fondées sur les $\hat{\eta}_i$ dans chacune de plusieurs strates en appliquant des règles un peu plus complexes que la règle des quantiles égaux considérée ici. Cependant, il ne faut pas surinterpréter le fait qu'un petit nombre de cellules pourrait être adéquat. Par exemple, si on omettait un régresseur important, les estimateurs corrigés d'après les cellules pourraient être entachés d'un biais résiduel important, quel que soit le nombre utilisé de cellules fondées sur la probabilité de réponse ou sur la réponse estimée.

Enfin, on peut remplacer la correction par pondération par l'imputation. Par exemple, la simple imputation «hot-deck»

permet de remplacer une valeur manquante dans une cellule de correction donnée en sélectionnant au hasard des répondants repères dans la même cellule. Parallèlement à (1.1) et à (1.2), l'estimateur résultant de la moyenne est $\hat{\bar{Y}}_{imp} = (\sum_{i\in s}\lambda_i)^{-1}\sum_{i\in s}\lambda_i Y_i^*$, où Y_i^* représente une valeur observée ou imputée, selon le cas. En pratique, on s'appuie souvent sur la correction par pondération pour tenir compte de la non-réponse des unités et sur l'imputation, pour tenir compte de la non-réponse à une question. Cependant, pour un ensemble donné de cellules, l'estimateur ponctuel corrigé par pondération (1.2) et l'estimateur par imputation $\hat{\bar{Y}}_{imp}$ sont entachés du même biais approximatif (1.3). Par souci de simplicité, la suite du présent article portera avant tout sur la correction par pondération, mais il ne faut pas perdre de vue que, pour un ensemble donné de cellules, le problème de réduction du biais est le même qu'on se serve de ces cellules pour effectuer la correction par pondération ou par simple imputation «hot-deck».

1.4 Plan de l'article

Nous examinons dans le présent article certains détails de l'application des méthodes de construction de cellules fondées sur la probabilité estimée de réponse et sur la réponse estimée. Nous accordons une attention particulière aux méthodes diagnostiques permettant de déceler les problèmes que pose un ensemble particulier de cellules, et nous justifions et illustrons ces méthodes en décrivant en détail leur application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. À la section 2, nous donnons certains renseignements généraux sur le problème de la non-réponse concernant le revenu. À la section 3, nous décrivons et appliquons plusieurs méthodes diagnostiques, y compris la comparaison des estimations $\hat{\bar{Y}}_k$ et des erreurs-types pour plusieurs valeurs de k (section 3.1), l'évaluation partielle du biais intracellulaire (section 3.2.1), l'évaluation de la largeur des cellules comparativement à la précision des estimations $\hat{\eta}_i$ (section 3.2.2) et la comparaison des estimations corrigées et non corrigées de la moyenne $\hat{\bar{Y}}_k$ et $\hat{\bar{Y}}_1$ (section 3.3). À la section 4, nous montrons qu'on peut appliquer des méthodes diagnostiques semblables à la correction des cellules fondées sur les chiffres prévus de revenu $\hat{Y}_i$, et nous comparons les estimations du revenu moyen calculé d'après les cellules fondées sur les probabilités estimées, d'une part, et sur le revenu estimé, d'autre part. À la section 5, nous résumons les idées principales qui sous-tendent le présent article et mentionnons certains domaines dans lesquels il conviendrait de poursuivre les travaux.

2. NON-RÉPONSE CONCERNANT LE REVENU DANS LE CAS DE LA U.S. CONSUMER EXPENDITURE SURVEY

2.1 Consumer Expenditure Survey, méthodes de pondération et estimation de la variance

La U.S. Consumer Expenditure Survey (CE) est une enquête à plan d'échantillonnage stratifié à plusieurs degrés avec renouvellement effectuée par le Census Bureau pour le Bureau of Labor Statistics (BLS). Les éléments de l'échantillon sont des «unités de consommation», grossièrement équivalentes aux ménages. Durant l'enquête, on demande à chaque unité d'échantillonnage sélectionnée de participer à cinq interviews. La méthode actuelle de pondération de la CE tient compte des probabilités de sélection initiale, de la correction pour la non-interview, de la stratification a posteriori fondée sur plusieurs variables démographiques et de mises au point supplémentaires; consulter Zieschang (1990) et le United States Bureau of Labor Statistics (1992). En raison de la complexité des travaux de pondération de la CE, le BLS a décidé d'utiliser des estimateurs de la variance fondés sur des méthodes de pseudorépétition à 44 échantillons répétés. Cette pseudorépétition est approximativement équivalente à la répétition compensée type (Wolter 1985, chapitre 3). Toutes les erreurs-types mentionnées ici sont fondées sur la méthode de pseudorépétition, toutes les étapes supplémentaires d'estimation des paramètres et de correction de la pondération étant exécutées séparément pour chaque répétition.

2.2 Non-réponse concernant le revenu

En général, on estime que la correction pour la non-interview incluse dans la méthode de pondération de la CE tient compte comme il convient de la non-réponse d'une unité, comme l'absence de contact ou le refus de participer à une interview particulière. Donc, nous ne nous pencherons plus sur la question de la non-réponse d'une unité ici. Cependant, le BLS craint que les estimations du revenu moyen soient entachées d'un biais dû à la non-réponse partielle aux questions sur le revenu de la CE. Suivent certains renseignements généraux.

Les données détaillées sur le revenu sont collectées durant les deuxième et cinquième interviews de la CE et sont utilisées pour produire des estimations du revenu moyen des unités de consommation (U.S. Bureau of Labor Statistics, 1991) et d'autres paramètres. Les données sur le revenu de la CE sont collectées grâce à un ensemble complexe de questions et le taux de non-réponse à ces questions est relativement élevé. Pour donner une indication sommaire de la réponse ou de la non-réponse à l'ensemble complet de questions sur le revenu, le BLS classifie chaque unité de consommation qui participe à la seconde ou à la cinquième interview comme déclarant complètement ou incomplètement leur revenu. La définition officielle de l'«unité déclarant complètement son revenu» est relativement compliquée; pour une discussion détaillée, consulter Garner et Blanciforti (1994). La méthode appliquée à l'heure actuelle par le BLS pour estimer le revenu moyen consiste à utiliser la réponse moyenne non corrigée $\hat{\bar{Y}}_1$ définie par (1.1), où les R_i représentent les indicateurs de déclaration complète du revenu, Y_i représente le revenu et les poids λ_i correspondent à ceux décrits à la section 2.1. La moyenne pondérée $\hat{\bar{Y}}_1$ est calculée d'après les données de la deuxième ainsi que de la cinquième interview pour une période de référence précise, mais ne s'appuie pas directement sur la structure par panel

des données de la CE. Conformément à cette approche, nous ne ferons la distinction entre les données de la deuxième et de la cinquième interviews dans le présent article que pour construire les modèles de $\hat{\eta}_i$ et $\hat{Y}_i$.

Ici, nous avons utilisé les données des rapports de la deuxième et de la cinquième interviews pour toutes les unités de consommation pour lesquelles une deuxième interview était prévue en 1990. Les données de la deuxième interview se rapportent à 5,125 unités et celles de la cinquième, à 5,093 unités. Pour chaque unité interviewée (ayant déclaré complètement ou incomplètement son revenu), nous avons tiré des enregistrements du BLS des données sur un grand nombre de variables démographiques et de variables de dépenses que nous avons utilisées comme variables auxiliaires dans les travaux de modélisation décrits aux sections 3 et 4 ci-après. À la deuxième ainsi qu'à la cinquième interview, environ 14% des unités de consommation interviewées ont déclaré incomplètement leur revenu.

3. CELLULES FONDÉES SUR LES PROBABILITÉS ESTIMÉES DE RÉPONSE

Nous examinons pour commencer la construction de cellules de correction fondées sur les probabilités estimées de réponse. Nous avons ajusté séparément les modèles de régression logistique utilisés pour calculer la probabilité qu'une unité déclare complètement son revenu $\eta_i = \eta(X_i)$ aux données de la deuxième et de la cinquième interviews décrites à la section 2. Les détails de l'ajustement des modèles, y compris l'estimation des paramètres et le calcul des erreurs-types, sont décrits dans Yansaneh et Eltinge (1993). Nous avons calculé toutes les estimations de la variance par la méthode de pseudorépétition décrite à la section 2. Nous nous sommes servis des modèles résultant des ajustements finals pour estimer, pour chaque unité ayant participé à la deuxième et à la cinquième interviews, les probabilités de déclarer complètement le revenu $\hat{\eta}_i$. Conformément à la stratégie décrite à la section 1.3, nous avons regroupé les unités selon la valeur de $\hat{\eta}_i$ en k cellules dont nous avons défini les limites par la méthode des quantiles égaux.

3.1 Analyse initiale de la sensibilité au nombre choisi de cellules

Les trois premières colonnes du tableau 1 donnent les estimations ponctuelles corrigées $\hat{\bar{Y}}_k$ du revenu moyen et les erreurs-types associées pour plusieurs valeurs de k. La comparaison de ces estimations ponctuelles indique dans quelle mesure les estimations corrigées sont sensibles au choix d'une valeur particulière de k. Pour $k \geq 5$, les estimations ponctuelles présentées sont relativement stables, variant de \$32,630 à \$32,664. Cette observation concorde avec l'idée énoncée à la section 1.3 selon laquelle $k = 5$ cellules pourrait fournir la plupart de la réduction effective du biais produite par une méthode donnée de construction des cellules; consulter Rosenbaum et Rubin (1984, section 1 et appendice A) pour certains renseignements mathématiques connexes.

En outre, soulignons que pour $k \geq 3$, l'erreur-type qui entache $\hat{\bar{Y}}_k$ est également assez stable, variant de \$508 à \$530. Cette observation contredit en partie l'idée générale selon laquelle le choix du nombre approprié de cellules s'appuie sur un compromis entre le biais et la variance. En ce qui concerne l'ensemble de données examinées ici, il semble que la réduction effective du biais se produise assez rapidement (disons, pour $k = 5$,), alors qu'une augmentation considérable de la variance ne survient qu'après qu'on ait dépassé la valeur $k = 20$. Ce résultat n'est pas irréaliste, puisque, même pour $k = 20$, le nombre de réponses sur le revenu par cellule demeure assez grand (variant de 461 à 569), donc ne donne pas lieu au problème général de l'estimateur instable associé à un nombre croissant de cellules peu peuplées. Par contre, le compromis entre le biais et la variance pourrait poser des problèmes plus graves pour des valeurs assez faibles de k dans le cas d'applications où la taille effective de l'échantillon est plus petite, comme l'estimation de petites sous-populations.

Tableau 1
Estimations corrigées du revenu moyen quand les limites des cellules sont déterminées d'après les quantiles des probabilités estimées de réponse

Nombre de cellules	Estimation ponctuelle	Erreur-type	$ET(\hat{\bar{Y}}_k - \hat{\bar{Y}}_1)$	Ratio $EQM(\hat{\bar{Y}}_k)$
Non corrigée ($k = 1$)	32,967	569	S/O	S/O
$k = 3$ cellules	32,736	530	112	1.30
$k = 4$ cellules	32,779	518	122	1.28
$k = 5$ cellules	32,630	523	138	1.53
$k = 6$ cellules	32,664	515	122	1.51
$k = 10$ cellules	32,640	514	116	1.58
$k = 15$ cellules	32,638	515	118	1.58
$k = 20$ cellules	32,634	508	118	1.63

3.2 Deux méthodes diagnostiques simples applicables aux cellules

Pour compléter l'analyse de sensibilité qui précède, il est utile d'examiner certains ensembles de cellules de correction plus en détail. Représentons par $C_1 = \{s_1, ..., s_k\}$ un ensemble donné de cellules de correction à examiner, comme les cellules créées par la division en quantiles égaux pour $k = 3$ ou $k = 5$ décrites à la section 3.1. Nous pouvons perfectionner les cellules contenues dans l'ensemble C_1 en effectuant une division en quantiles égaux pour une valeur plus grande de k ou en divisant directement une ou plusieurs cellules de l'ensemble C_1. Ce perfectionnement pourrait être utile quand les observations empiriques indiquent que 1) l'estimateur de la moyenne de la cellule $\bar{Y}_{hR}$ risque d'être fortement biaisé ou que 2) une cellule est large comparativement à la précision avec laquelle les valeurs η_i sont estimées. Nous décrivons aux sous-sections 3.2.1 et 3.2.2 deux méthodes diagnostiques simples qui permettent de résoudre les problèmes (1) et (2), respectivement. Dans chaque sous-section, la méthode

diagnostique proposée mène au dépistage des «cellules problématiques» éventuelles et à la construction d'un ensemble plus perfectionné de cellules de correction que nous appelons C_2. La comparaison des estimations de $\bar{Y}$ fondées sur C_1 et C_2 permet alors de décider quel est l'ensemble de cellules de correction fondées sur $\hat{\eta}_i$ qui convient le mieux.

3.2.1 Évaluation du biais à l'intérieur des cellules

Comme nous l'avons fait remarquer à la section 1.2, un estimateur corrigé $\hat{\bar{Y}}_k$ donné réduit, mais n'élimine pas complètement, le biais dû à la non-réponse et le biais résiduel de $\hat{\bar{Y}}_k$ dépend du biais qui entache les estimations de la moyenne intracellulaire $\bar{Y}_{hR}$. Considérons l'autre estimateur de la moyenne intracellulaire

$$\bar{Y}_{h\eta} = \left(\sum_{i \in s_h} \hat{\eta}_i^{-1} \lambda_i R_i \right)^{-1} \sum_{i \in s_h} \hat{\eta}_i^{-1} \lambda_i R_i Y_i. \qquad (3.1)$$

Si les estimations $\hat{\eta}_i$ étaient égales aux probabilités réelles de réponse η_i, alors (3.1) serait un estimateur approximativement non biaisé de la moyenne réelle de la sous-population $\bar{Y}_h$. Le cas échéant, un estimateur du biais intracellulaire $E(\bar{Y}_{hR} - \bar{Y}_h)$ serait $\hat{B}_h = \bar{Y}_{hR} - \bar{Y}_{h\eta}$, et l'estimateur correspondant du biais global $E(\hat{\bar{Y}}_k - \bar{Y})$ serait $\hat{B} = (\sum_{h=1}^{k} \sum_{j \in s_h} \lambda_j)^{-1} \sum_{h=1}^{k} (\sum_{j \in s_h} \lambda_j) \hat{B}_h$.

Puisque les valeurs de $\hat{\eta}_i$ sont sujettes à des erreurs d'estimation, les termes $\hat{B}_h$ et $\hat{B}$ ne donnent qu'une indication partielle des problèmes éventuels de biais. Par exemple, une grande valeur de $\hat{B}_h$ pourrait être le reflet d'un biais important entachant $\bar{Y}_{hR}$ ou de biais entachant l'autre estimateur $\bar{Y}_{h\eta}$, à cause des erreurs $\hat{\eta}_i - \eta_i$; lire les commentaires de mise en garde de Little (1986, p. 146) concernant l'utilisation directe des poids $\hat{\eta}_i^{-1}$ pour établir l'estimation corrigée de $\bar{Y}$. Donc, si on observe une grande valeur de $\hat{B}_h$, cela vaut la peine d'envisager la mise au point de la cellule h, mais la décision finale quant à l'utilisation de l'ensemble perfectionné de cellules ainsi obtenu dépendra du fait que cet ensemble produit ou non une estimation nettement différente de la moyenne globale $\bar{Y}$.

Nous présentons aux tableaux 2 et 3 les valeurs de $\hat{B}_h$, les erreurs-types associées et les valeurs de la statistique t pour les cellules formées par division en quantiles égaux pour $k = 3$ et $k = 5$, respectivement. Soulignons que, dans le cas où $k = 3$, le test diagnostic s'appliquant à $\hat{B}_h$ indique une contribution éventuelle au biais pour la cellule de rang le plus bas. Cette observation concorde avec l'idée énoncée à la section 3.1 selon laquelle $k = 3$ cellules pourrait ne pas donner une correction satisfaisante pour la non-réponse. En outre, la valeur correspondante de $\hat{B}$ est 111, avec une erreur-type de 75; cette valeur de $\hat{B}$ est très proche de la différence $\hat{\bar{Y}}_3 - \hat{\bar{Y}}_5 = 106$ des estimations $\hat{\bar{Y}}_3$ et $\hat{\bar{Y}}_5$ du tableau 1.

À la lumière des résultats qui précèdent, nous avons divisé en deux la cellule à faible $\hat{\eta}_i$ du cas où $k = 3$. Nous avons déterminé les limites supérieures des deux nouvelles cellules (soit, $h = 1'$ et $h = 1''$) grâce aux quantiles estimés 0.167 et 0.333 de la population de $\hat{\eta}_i$. Les valeurs résultantes de $\hat{B}_h$ et les erreurs-types sont 90 et 197 pour la cellule $1'$, et -42 et 79, pour la cellule $1''$. En outre, l'ensemble ainsi perfectionné de quatre cellules donne $\hat{B} = 30$, avec une erreur-type de 75, et l'estimation corrigée de $\bar{Y}$, égale à \$32,652, ainsi que l'erreur-type, égale à \$518, sont proches des valeurs obtenues par la méthode de division en quantiles égaux pour $k = 5$.

Tableau 2
Statistiques $\hat{B}_h$ cellulaires pour les cellules fondées sur les probabilités, $k = 3$

h	$\hat{B}_h$	et$(\hat{B}_h)$	$t = \hat{B}_h/\text{et}(\hat{B}_h)$
1	269	136	1.98
2	-19	43	-0.44
3	84	45	1.87

Tableau 3
Statistiques $\hat{B}_h$ cellulaires pour les cellules fondées sur les probabilités, $k = 5$

h	$\hat{B}_h$	et$(\hat{B}_h)$	$t = \hat{B}_h/\text{et}(\hat{B}_h)$
1	96	217	0.44
2	-72	116	-0.62
3	-52	56	-0.93
4	-16	27	-0.59
5	98	50	1.96

Contrairement aux résultats obtenus pour $k = 3$, les valeurs de $\hat{B}_h$ obtenues pour $k = 5$ posent assez peu de problèmes, sauf, éventuellement, dans le cas de la cellule $h = 5$, pour laquelle la valeur de la statistique t est égale à 1.96. Pour $k = 5$, la valeur de $\hat{B}$ est 11, avec une erreur-type de 93. Une division supplémentaire de la cellule $h = 5$ n'a pas modifié notablement l'estimation de $\bar{Y}$ ou de l'erreur-type associée. Les valeurs de $\hat{B}_h$ résultant de la division des cellules correspondant à une valeur de k plus grande par la méthode des quantiles égaux présentent encore moins de signes de l'existence d'un biais intracellulaire. Par exemple, pour $k = 6$, la statistique t est inférieure ou égale à 1.65 pour les valeurs de $\hat{B}_h$ de chacune des six cellules et, pour $k = 10$, la statistique t est inférieure ou égale à 1.54 pour les valeurs de $\hat{B}_h$ de chacune des dix cellules.

3.2.2 Comparaison entre la largeur des cellules et la précision des estimations η_i

La comparaison de la largeur des cellules de correction à la largeur des intervalles de confiance associés aux probabilités de réponse η_i fournit une autre méthode diagnostique pour repérer les cellules problématiques éventuelles. Premièrement, représentons par $a_h = (\sum_{i \in s_h} \lambda_i R_i)^{-1} \sum_{i \in s_h} \lambda_i$, le facteur de correction pour la non-réponse appliqué aux unités répondantes de la cellule h. Deuxièmement, conformément aux résultats types de la régression logistique, notons qu'un intervalle de confiance d'environ 95% pour η_i est

$$(LB_i, UB_i) = ([1 + \exp\{-X_i'\hat{\theta} + 1.96 D_i^{1/2}\}]^{-1}$$

$$[1 + \exp\{-X_i'\hat{\theta} - 1.96 D_i^{1/2}\}]^{-1}),$$

où $\hat{\theta}$ est le vecteur des estimations des paramètres de régression logique, où $D_i = X_i' \hat{V}_\theta X_i$, et où $\hat{V}_\theta$ est la matrice de covariance estimée d'après la pseudorépétition pour $\hat{\theta}$. Représentons par $\bar{d}_h$ la moyenne de l'échantillon pondérée en fonction de λ_i des largeurs des intervalles de confiance $UB_i - LB_i$ pour les unités i de la cellule h, et comparons $\bar{d}_h$ à la largeur de la cellule h. Si la cellule h est relativement large, tant en valeur absolue que comparativement à $\bar{d}_h$, alors, la division de cette cellule peut produire de nouvelles cellules ayant des facteurs de pondération a_h nettement différents. Inversement, si $\bar{d}_h$ est beaucoup plus grand que la largeur de la cellule h, alors, les écarts entre les $\hat{\eta}_i$ dans cette cellule peuvent résulter davantage de l'erreur d'estimation que d'écarts entre les η_i réels. Le cas échéant, une division supplémentaire de la cellule modifiera vraisemblablement peu les facteurs de pondération a_h; et, par conséquent, l'estimateur de $\bar{Y}$ corrigé pour la non-réponse obtenu variera assez peu.

Nous présentons aux tableaux 4 et 5 les limites des cellules, les largeurs des cellules, $\bar{d}_h$, ainsi que les valeurs de a_h pour $k = 5$ et $k = 10$, respectivement. Pour $k = 5$, la largeur des cellules 2 à 5 n'est pas grande comparativement aux valeurs de $\bar{d}_h$. Essentiellement, chacune de ces cellules est divisée en deux pour produire le cas où $k = 10$ cellules. Les paires résultants de a_h pour $k = 10$ sont assez proches des valeurs correspondantes de a_h dans les cellules 2 à 5 pour $k = 5$.

Tableau 4
Limites des cellules fondées sur la probabilité estimée de réponse, largeur moyenne des cellules, largeur des intervalles de confiance et facteur de correction pour la non-réponse, $k = 5$

h	Limite inférieure	Limite Supérieure	Largeur de la cellule	$\bar{d}_h$	a_h
1	0.384	0.810	0.426	0.197	1.35
2	0.810	0.861	0.051	0.139	1.20
3	0.861	0.894	0.033	0.110	1.13
4	0.894	0.924	0.030	0.088	1.08
5	0.924	0.994	0.070	0.067	1.07

En revanche, pour $k = 5$, la cellule 1 est plus de deux fois plus large que $\bar{d}_1$. Quand $k = 10$, cette cellule est divisée en cellules plus petites ayant des facteurs de pondération corrigés pour la non-réponse a_h légèrement différents, soit 1.45 et 1.27, respectivement. Cependant, les estimations correspondantes de la moyenne cellulaire sont assez proches, à savoir $\bar{Y}_{1R}$ = \$24,045 et $\bar{Y}_{2R}$ = \$24,582 pour $k = 10$. Donc, dans cet exemple, les estimations corrigées pour la non-réponse $\hat{\bar{Y}}_5$ et $\hat{\bar{Y}}_{10}$ sont relativement proches, parce que quatre des cinq divisions cellulaires entraînent une variation assez faible des poids et que la cinquième produit des cellules dont la moyenne est similaire.

Tableau 5
Limites des cellules fondées sur la probabilité estimée de réponse, largeur des cellules, largeur moyenne des intervalles de confiance et facteur de correction pour la non-réponse, $k = 10$

h	Limite inférieure	Limite supérieure	Largeur de la cellule	$\bar{d}_h$	a_h
1	0.384	0.762	0.378	0.220	1.45
2	0.762	0.810	0.048	0.174	1.27
3	0.810	0.840	0.030	0.146	1.21
4	0.840	0.861	0.021	0.132	1.19
5	0.861	0.878	0.017	0.111	1.14
6	0.878	0.894	0.016	0.108	1.11
7	0.894	0.908	0.014	0.093	1.09
8	0.908	0.924	0.016	0.083	1.08
9	0.924	0.944	0.020	0.072	1.08
10	0.944	0.994	0.050	0.062	1.06

Enfin, les facteurs a_h du tableau 5 indiquent que les taux moyens de réponse dans le cas de $k = 10$ cellules se situent dans une fourchette raisonnable, allant de $(1.45)^{-1} = 0.69$ à $(1.06)^{-1} = 0.94$. Certains ensembles de données sur la non-réponse sont caractérisés par une fourchette plus large, donc plus susceptibles de produire des écarts plus prononcés après la division des cellules. Inversement, d'autres sont caractérisés par une distribution plus serrée des probabilités de réponse, donc moins susceptibles d'être affectés notablement par la division des cellules.

3.3 Comparaison des estimations corrigées d'après les cellules aux estimations non corrigées

Pour conclure l'évaluation des cellules fondées sur les $\hat{\eta}_i$, nous comparons les estimations corrigées $\hat{\bar{Y}}_k$ aux estimations non corrigées $\hat{\bar{Y}}_1$. Premièrement, le tableau 1 indique que, pour les valeurs déclarées de $k \geq 5$, les différences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ sont supérieures ou égales à \$303. Deuxièmement, pour $k \geq 5$, les erreurs-types estimées des différences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ sont toutes inférieures ou égales à \$138, et les valeurs correspondantes de la statistique t sont toutes supérieures à 2.44. Donc, pour $k = 5$ par exemple, un test formel de l'hypothèse $H_0: E(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_5) = 0$ mènerait au rejet de cette dernière pour les seuils de signification types, ce qui signifie que la méthode des cellules de correction produit une modification importante de l'estimation du revenu moyen.

De surcroît, une comparaison grossière de l'efficacité de $\hat{\bar{Y}}_1$ et $\hat{\bar{Y}}_k$ se dégage du ratio des erreurs quadratiques moyennes estimées

$$\hat{\gamma}_k = \{\hat{V}(\hat{\bar{Y}}_k)\}^{-1}[\hat{V}(\hat{\bar{Y}}_1) + \max\{0, (\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)^2 - \hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)\}]$$

où $\hat{V}(\hat{\bar{Y}}_1)$, $\hat{V}(\hat{\bar{Y}}_k)$, et où $\hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$ sont les estimations de la variance basée sur la pseudorépétition pour les moyennes indiquées. Pour interpréter ce ratio, supposons pour le moment que $\hat{\bar{Y}}_k$ est un estimateur approximativement non biaisé de $\bar{Y}$. Alors, $\hat{\gamma}_k$ est un estimateur de l'erreur quadratique moyenne de l'estimateur non corrigé $\hat{\bar{Y}}_1$, comparativement à l'erreur quadratique moyenne de $\hat{\bar{Y}}_k$.

Conséquemment, $\hat{\gamma}_k$ reflète la perte d'efficacité résultant de l'utilisation de l'estimateur biaisé, non corrigé $\hat{\bar{Y}}_1$ au lieu de l'estimateur corrigé, non biaisé $\hat{\bar{Y}}_k$. Cependant, cette interprétation doit être considérée avec prudence, puisqu'elle dépend de la supposition que $\hat{\bar{Y}}_k$ est un estimateur approximativement non biaisé de $\bar{Y}$, et, puisque les $\hat{\gamma}_k$ sont des fonctions des termes aléatoires $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$, $\hat{V}(\hat{\bar{Y}}_1)$, $\hat{V}(\hat{\bar{Y}}_k)$, et $\hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$.

Comme l'a suggéré un arbitre, on pourrait aussi considérer le ratio des erreurs quadratiques moyennes

$$\{\hat{V}(\hat{\bar{Y}}_\eta)\}^{-1}[\hat{V}(\hat{\bar{Y}}_k) + \max\{0, (\hat{\bar{Y}}_k - \hat{\bar{Y}}_\eta)^2 - \hat{V}(\hat{\bar{Y}}_k - \hat{\bar{Y}}_\eta)\}]$$

où $\hat{\bar{Y}}_\eta$ est égal à l'expression (1.1) avec λ_i remplacé par $(\hat{\eta}_i)^{-1}\lambda_i$. Cette approche équivaudrait à comparer chaque estimation $\hat{\bar{Y}}_k$ fondée sur les cellules à $\hat{\bar{Y}}_\eta$. Cette démarche est appropriée si $\hat{\bar{Y}}_\eta$ est approximativement non biaisé, mais l'absence de biais peut être problématique dans certains cas; à cet égard, consulter Little (1986, p. 146).

La dernière colonne du tableau 1 donne les ratios estimés $\hat{\gamma}_k$ pour des valeurs particulières de k. Pour $k \geq 5$, chaque $\hat{\gamma}_k$ présenté est plus grand que 1.5. Enfin, soulignons que chaque estimation corrigée $\hat{\bar{Y}}_k$ est inférieure à l'estimation non corrigée $\hat{\bar{Y}}_1$. Cette situation tient au fait que, pour une valeur donnée de k, les cellules associées aux probabilités de réponse les plus grandes ont tendance à produire une estimation plus grande de la moyenne $\bar{Y}_{hR}$. Par exemple, pour $k = 5$, les valeurs de $\bar{Y}_{hR}$ sont \$24,333, \$33,729, \$33,398, \$34,620 et \$37,057 pour $h = 1$ (cellule à $\hat{\eta}_i$ la plus faible) à $h = 5$ (cellule à $\hat{\eta}_i$ la plus élevée), respectivement.

4. CELLULES FONDÉES SUR LES VALEURS ESTIMÉES DU REVENU

Les idées diagnostiques générales décrites à la section 3 s'appliquent également aux cellules fondées sur $\hat{Y}_i$. Pour illustrer ce point, nous ajustons séparément des équations de régression pondérée où Y_i = représente le revenu déclaré par les répondants à la deuxième et à la cinquième interviews. Yansaneh et Eltinge (1993) décrivent en détail les calculs, y compris l'estimation des paramètres et des erreurs-types. Nous nous sommes servi des modèles de régression résultants pour calculer les estimations du revenu $\hat{Y}_i$ pour les unités qui ont déclaré complètement et incomplètement leur revenu, puis nous avons regroupé les unités en cellules d'après leur valeur de $\hat{Y}_i$, les limites des cellules étant déterminées par la méthode des quantiles égaux.

Nous présentons au tableau 6 les résultats de l'analyse fondamentale de sensibilité et de la mesure de l'efficacité pour les cellules fondées sur les $\hat{Y}_i$; sa présentation est la même que celle du tableau 1. Les résultats de l'analyse de sensibilité sont qualitativement similaires, mais non identiques, à ceux présentés pour les cellules fondées sur les $\hat{\eta}_i$. Au cours de travaux supplémentaires non décrits en détail ici, nous avons examiné la division de chaque cellule fondée sur les $\hat{Y}_i$ en quantiles égaux. Pour $k \geq 4$, les estimations résultantes de la moyenne et les erreurs-types connexes ne diffèrent pas notablement de celles présentées au tableau 6.

Tableau 6
Estimations corrigées du revenu moyen quand les limites des cellules sont déterminées d'après les quantiles du revenu estimé

Méthode de correction	Estimation ponctuelle	Erreur-type	$ET(\hat{\bar{Y}}_k - \hat{\bar{Y}}_1)$	Ratio EQM
Non corrigée ($k = 1$)	32,967	569	S/O	S/O
$k = 3$ cellules	32,512	509	106	2.01
$k = 4$ cellules	32,468	512	108	2.14
$k = 5$ cellules	32,473	511	115	2.12
$k = 6$ cellules	32,492	508	117	2.08
$k = 10$ cellules	32,488	510	119	2.07
$k = 15$ cellules	32,478	504	124	2.16
$k = 20$ cellules	32,495	513	124	2.02

Les deux dernières colonnes du tableau 6 permettent de comparer $\hat{\bar{Y}}_k$ à l'estimation non corrigée $\hat{\bar{Y}}_1$. Pour $k \geq 4$, les différences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ sont supérieures ou égales à \$472, et les erreurs-types estimées, inférieures ou égales à \$124. Les valeurs de la statistique t associée sont toutes supérieures à 3.80. En outre, les ratios des erreurs quadratiques moyennes estimées $\hat{\gamma}_k$ sont tous plus grand que 2.0.

Qui plus est, les cellules fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$ produisent des estimations corrigées légèrement différente du revenu moyen, mais les écarts observés ne sont pas statistiquement significatifs pour les seuils de signification α ordinaires. Par exemple, pour $k = 5$, l'écart entre les estimations fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$ est égal à \$32,630\$ - \$32,473 = \$157, l'erreur-type est égale à \$122 et la statistique t est égale à 1.29. Pareillement, pour $k = 10$, la différence entre les estimations fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$ est de \$152, avec une erreur-type de \$104. Donc, les données permettent peu de faire la distinction entre les résultats obtenus par les deux méthodes générales de construction de cellules.

Enfin, soulignons qu'un ensemble donné de cellules fondées sur les $\hat{Y}_i$ est fondamentalement lié à une variable Y particulière, comme le revenu de l'unité de consommation. Par conséquent, cet ensemble de cellules ne donnera pas nécessairement de bons résultats pour l'estimation de la moyenne d'une variable Y différente.

5. DISCUSSION

5.1 Résumé des méthodes

Nous examinons dans le présent article certaines méthodes diagnostiques simples permettant de construire des cellules de correction pour la non-réponse. Nous pouvons résumer comme suit la méthodologie.

1. D'après des travaux de modélisation préliminaires et les variables auxiliaires étudiées X_i, calculer la probabilité estimée de réponse $\hat{\eta}_i$ pour chaque unité d'échantillonnage (répondants et non-répondants).

2. Construire k cellules de correction dont les limites sont déterminées par les quantiles $k^{-1}j$ estimés de la population de $\hat{\eta}_i$, $j = 1,2,...,k-1$. Calculer l'estimation de la moyenne corrigée résultante, $\hat{\bar{Y}}_k$.
3. Répéter (2) pour plusieurs nombre entiers $k > 1$. À mesure que k augmente, repérer le point où $\hat{\bar{Y}}_k$ devient à peu près constante. Compte tenu des résultats de Rosenbaum et Rubin (1984) et des résultats empiriques exposés ici, les valeurs de k proches de 5 pourraient présenter un intérêt particulier.
4. Utiliser des méthodes diagnostiques simples (par exemple $\hat{B}_h$ et $\bar{d}_h$ décrits à la section 3.2) pour repérer les cellules de corrections obtenues par division en quantiles égaux posant éventuellement des problèmes. Si l'application de la méthode diagnostique indique que certaines cellules sont problématiques, essayer d'effectuer une mise au point supplémentaire de ces cellules. Calculer les estimations de $\bar{Y}$ d'après les ensembles perfectionnés de cellules et comparer les nouvelles estimations aux valeurs de $\hat{\bar{Y}}_k$ obtenue au point 3.
5. Évaluer l'effet global de la correction en comparant les différences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ aux erreurs-types $se(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$ et en calculant les ratios des erreurs quadratiques moyennes estimées $\hat{\gamma}_k$.
6. Répéter les étapes (1) à (5), au besoin, pour les cellules de correction fondées sur les $\hat{Y}_i$. Comparer les estimations finales de $\bar{Y}$ obtenues par les méthodes fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$.

5.2 Domaines dans lesquels il faut poursuivre les travaux

Les résultats de la présente étude donnent à penser qu'il pourrait être utile de poursuivre les travaux dans deux domaines. Premièrement, le problème que pose la non-réponse aux questions sur le revenu dans le cas de la CE est similaire à celui que pose la non-réponse dans le cas de plusieurs autres enquêtes à grande échelle. Cependant, comme pour toute étude de cas, on devrait éviter de surgénéraliser les résultats empiriques présentés ici. Il serait utile d'appliquer les méthodes diagnostiques que nous décrivons aux problèmes que posent d'autres estimations (par exemple, moyennes croisées) ou à des ensembles de données sur la non-réponse présentant des caractéristiques légèrement différentes (par exemple, taille effective de l'échantillon plus grande ou plus petite, ou distribution plus étalée ou plus serrée des estimations des $\hat{\eta}_i$). Cet exercice apporterait des éclaircissements sur les caractéristiques opératoires des méthodes de construction de cellules fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$ dans la pratique. Deuxièmement, il serait intéressant d'étendre l'étude à des problèmes à plusieurs variables (par exemple, rapport entre les données sur le revenu de la deuxième et de la cinquième interview de la CE).

REMERCIEMENTS

Les auteurs remercient Richard Dietz, Thesia Garner, Paul Hsen, Eva Jacobs, Geoffrey Paulin, Stuart Scott et Stephanie Shipp pour les nombreuses discussions fructueuses sur la Consumer Expenditure Survey, et Wayne Fuller, Steve Miller, Geoff Paulin, Stuart Scott, trois arbitres, et le rédacteur, pour leurs commentaires précieux sur des versions antérieures de l'article. Les présents travaux ont été effectués pendant un séjour des auteurs au Bureau of Labor Statistics organisé dans le cadre du ASA/NSF/BLS Research Fellow Program et financé grâce à une bourse de la National Science Foundation (SES-9022443). Les travaux de recherche de Eltinge ont également été financés en partie par une bourse du National Institutes of Health (CA 57030-04). Les opinions exprimées dans le présent article sont celles des auteurs et ne représentent pas nécessairement les politiques du Bureau of Labor Statistics.

BIBLIOGRAPHIE

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*, (Vol. 3), (Éds. W.G. Madow, I. Olkin, et D. Rubin). New York: Academic Press, 143-160.

COCHRAN, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 205-213.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley.

CZAJKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., et RUBIN, D.B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business and Economic Statistics*, 10, 117-131.

DAVID, M.H., LITTLE, R.J.A., SAMUHEL, M., et TRIEST, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 168-173.

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

EZZATI, T., et KHARE, M. (1992). Nonresponse adjustments in a national health survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 339-344.

GARNER, T.I., et BLANCIFORTI, L.A. (1994). Household income reporting: An analysis of U.S. Consumer Expenditure Survey data. *Journal of Official Statistics* 10, 69-91.

GOKSEL, H., JUDKINS, D.R., et MOSHER, W.D. (1991). Nonresponse adjustments for a telephone follow-up to a national in-person survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 581-586.

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.

LEPKOWSKI, J., KALTON, G., et KASPRZYK, D. (1989). Weighting adjustments for partial nonresponse in the 1984 SIPP panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 296-301.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 139-157.

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

OH, H.L., et SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete Data in Sample Surveys*, (Vol. 2), (Éds. W.G. Madow, I. Olkin et D.B. Rubin). New York: Academic Press, 143-184.

RAO, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

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

ROSENBAUM, P.R., et RUBIN, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524.

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

UNITED STATES BUREAU OF LABOR STATISTICS (1991). News: Consumer Expenditures in 1990. Publication USDL 91-607, United States Department of Labor, Washington, DC.

UNITED STATES BUREAU OF LABOR STATISTICS (1992). BLS Handbook of Methods. Bulletin 2414, United States Department of Labor, Washington, DC.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

YANSANEH, I.S., et ELTINGE, J.L. (1993). Construction of adjustment cells based on surrogate items or estimated response propensities. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 538-543.

ZIESCHANG, K.D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 47-59
Statistique Canada

Estimation de la variance des mesures de l'inégalité et de la polarisation du revenu – Étude empirique

MILORAD S. KOVAČEVIĆ et WESLEY YUNG[1]

RÉSUMÉ

Les mesurent de l'inégalité et de la polarisation du revenu sont essentielles à l'étude de nombreux dossiers économiques et sociaux. La plupart de ces mesures étant des fonctions non linéaires de la fonction de distribution et(ou) des quantiles, on ne peut exprimer leur variance au moyen d'une formule simple et on doit recourir aux méthodes d'estimation de la variance approximative. Dans le présent article, on résume plusieurs méthodes appliquées à l'estimation de la variance de six mesures particulières de l'inégalité et de la polarisation du revenu et on étudie empiriquement leur performance grâce à une étude en simulation fondée sur l'Enquête canadienne sur les finances des consommateurs. Les résultats indiquent que, pour les mesures étudiées, la méthode bootstrap et celle des équations d'estimation donnent de nettement meilleurs résultats que les autres.

MOTS CLÉS: Indice de Gini; ordonnée de la courbe de Lorenz; proportion de faible revenu; indice de polarisation; part du quantile; estimation de la variance par rééchantillonnage; méthode de linéarisation.

1. INTRODUCTION

Les analyses de la répartition du revenu sont des éléments fondamentaux de l'examen de questions socioéconomiques importantes, dont la grandeur de l'inégalité, la pauvreté ou la grandeur de la classe moyenne. De nombreux articles statistiques et économétriques ont été publiés sur le sujet, particulièrement sur les diverses mesures de l'inégalité du revenu et sur leurs propriétés (Sen (1973); Kakwani (1980); Nygård et Sandström (1981)). Cependant, les auteurs s'efforcent rarement de produire des données sur la variabilité d'échantillonnage des estimations utilisées pour déterminer la grandeur de l'inégalité ou de la polarisation. Or, on doit s'appuyer sur ce genre d'information pour i) déterminer la précision des estimations obtenues à partir de données d'enquête et ii) faire des inférences statistiques officielles sur la répartition du revenu, particulièrement quand on effectue des comparaisons régionales ou chronologiques.

Les mesures de l'inégalité et de la polarisation du revenu étant des paramètres de population finie qui s'expriment sous forme de fonctions des valeurs ordonnées de la population, on ne peut calculer leur variance au moyen d'une formule simple et on doit recourir aux méthodes d'estimation de la variance approximative. En général, les inférences concernant ces mesures, fondées sur un plan d'échantillonnage complexe, englobent une estimation ponctuelle et des intervalles de confiance. Nous étudions ici l'estimation de la variance de certaines de ces mesures, dont les quantiles, le seuil de faible revenu, la proportion de faible revenu, l'ordonné de la courbe de Lorenz, la parts du quantile, l'indice de Gini et l'indice de polarisation.

Dans tout l'article, nous supposons que la population observée est une population finie invariable, autrement dit que chaque unité de la population est associée à un nombre réel, fixe, mais inconnu, à savoir la valeur du revenu gagné par l'unité. Nous supposons aussi que la population est subdivisée en L strates, et nous représentons par N_h le nombre d'unités primaires d'échantillonnage (UPÉ) dans la h-ième strate. Au premier degré d'échantillonnage, nous tirons $n_h(\geq 2)$ UPÉ de la strate h (de façon indépendante d'une strate à l'autre). Enfin, nous supposons qu'on effectue un sous-échantillonnage des UPÉ échantillonnées pour s'assurer que les estimations des totaux des UPÉ, $Y_{hc}, c = 1, ..., n_h$; $h = 1, ..., L$ ne soient pas biaisées. La (hci)-ième unité finale d'échantillonnage est liée à la valeur observée de la variable étudiée, y_{hci}, ainsi qu'au poids d'échantillonnage w_{hci}. Nous représentons par $\sum_s = \sum_h \sum_c \sum_i$ la sommation multiple sur toutes les unités finales de l'échantillon tenant compte de tous les degrés d'échantillonnage.

Après avoir passé en revue les définitions fondamentales des mesures étudiées, nous présentons à la section 2 leur estimation ponctuelle conformément au plan de sondage proposé. À la section 3, qui traite de l'estimation de la variance de ces mesures, nous passons en revue les méthodes existantes et nous en décrivons cinq en détail, à savoir la méthode jackknife, les méthodes du groupement de demi-échantillons compensés et du groupement répété de demi-échantillons compensés, la méthode bootstrap et la méthode de linéarisation fondée sur des équations d'estimation. À la section 4, nous décrivons l'étude en simulation basée sur les données de l'Enquête canadienne sur les finances des consommateurs de 1988. Cette étude empirique a pour objectif de comparer les méthodes d'estimation de la variance d'un certain nombre de mesures de l'inégalité du revenu. Nous présentons, résumons et interprétons divers résultats. Enfin, à la section 5, nous présentons nos conclusions.

[1] Milorad S. Kovačević, méthodologiste principale, Division des méthodes d'enquêtes des ménages, et Wesley Yung, méthodologiste sénior, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Immeuble R.H. Coats, Parc Tunney, Ottawa (Ontario) Canada, K1A 0T6.

2. ESTIMATION DES MESURES DE L'INÉGALITÉ DU REVENU

Les moyens les plus simples de mesurer l'inégalité entre deux distributions consistent à comparer les fonctions de distribution cumulatives (FDC) ou les quantiles de ces distributions. Nous commencerons par définir la FDC et les quantiles d'une population finie. Les autres mesures examinées dans le présent article, qui sont des fonctions de la FDC ou un nombre déterminé de quantiles, seront présentées à la section 2.1.

Soit une variable Y de la population finie $U = \{1, ..., N\}$. Nous définissons la FDC de cette variable comme étant

$$F_N(y) = \sum_{i \in U} I\{Y_i \le y\} \frac{1}{N},$$

où $I\{a\}$ est une fonction indicatrice dont la valeur est 1 si a est vrai et 0 autrement. Nous représentons l'estimateur sans biais dû au plan de sondage de $F_N(y)$ par

$$\tilde{F}(y) = \sum_{i \in s} I\{y_i \le y\} \frac{w_i}{N}$$

où les poids d'échantillonnage, w_i, sont calculés d'après le plan de sondage et sont égaux à l'inverse des probabilités d'inclusion de premier ordre. Toutefois, cet estimateur pourrait ne pas être une FDC, puisque $\tilde{F}(\infty) = \hat{N}/N$ n'est pas nécessairement égale à 1. Donc, nous préférons utiliser l'estimateur éventuellement entaché d'un biais dû au plan de sondage, soit:

$$\hat{F}(y) = \sum_{i \in s} I\{y_i \le y\} w_i \Big/ \sum_{i \in s} w_i = \sum_{i \in s} I\{y_i \le y\} \bar{w}_i, \qquad (2.1)$$

où $\bar{w}_i = w_i / \sum_s w_i$, $i \in s$. L'estimateur (2.1) n'est entaché d'aucun biais dû au plan de sondage quand $\sum_s w_i = N$, autrement dit si on effectue un échantillonnage aléatoire simple ou qu'on étalonne les poids, w_i, d'après des totaux connus de la population. En général, on applique à l'estimateur (2.1) des poids finals, ordinairement obtenus après stratification à posteriori, correction pour la non-réponse, certains étalonnages itératifs, etc. Dans le présent article, nous examinons uniquement le cas où les poids sont établis d'après le plan de sondage.

Passons maintenant aux quantiles. Nous représentons les quantiles d'une population finie par la fonction

$$\xi_N(p) = \inf_{i \in U} \{Y_i \mid F_i \ge p\} \quad \text{pour} \quad 0 < p \le 1,$$

où $F_i = F_N(Y_i)$. Nous estimons les quantiles de la population d'après les quantiles de l'échantillon, soit

$$\hat{\xi}_p = \inf_{i \in s} \{y_i \mid \hat{F}_i \ge p\} \quad \text{pour} \quad 0 < p \le 1,$$

où $\hat{F}_i = \hat{F}(y_i)$. Si un paramètre est une fonction des quantiles, disons $\theta_N = g(\xi_N)$ avec $\xi_N = \{\xi_N(p_1), ..., \xi_N(p_k)\}$, alors son estimateur est $\hat{\theta} = g(\hat{\xi})$ où $\hat{\xi} = (\hat{\xi}_{p_1}, ..., \hat{\xi}_{p_k})$.

2.1 Mesures de l'inégalité et de la polarisation du revenu en tant que paramètres d'une population finie

Dans la présente section, nous présentons certaines mesures de l'inégalité et de la polarisation du revenu utilisées fréquemment, à savoir le seuil de faible revenu, la proportion de faible revenu, la courbe de Lorenz et les statistiques connexes, les parts des quantiles, l'indice de Gini et, enfin, la courbe de polarisation et l'indice de polarisation. Nous nous limiterons ici à les présenter brièvement. Pour plus de précisions, le lecteur consultera Nygård et Sandström (1981) et Wolfson (1994).

Nous définissons le *seuil de faible revenu*, ou *seuil de pauvreté*, comme une fonction de la médiane, $\lambda_\alpha = \alpha\, \xi_N(0.5)$, où $0 < \alpha \le 1$ est une constante donnée et où $\xi_N(0.5)$ est la médiane de la population finie. Son estimation est simplement $\hat{\lambda}_\alpha = \alpha\, \hat{\xi}_{0.5}$.

La *proportion de faible revenu (PFR)* est la proportion d'unités (particuliers, familles, ménages) de la population qui se situe sous le seuil de faible revenu λ_α et est représentée par $\Lambda_\alpha = F_N(\lambda_\alpha)$. L'estimateur de la proportion de faible revenu englobe à la fois l'estimateur de la fonction de distribution et celui du seuil de faible revenu, $\hat{\Lambda}_\alpha = \hat{F}(\hat{\lambda}_\alpha) = \sum_s I\{y_{hci} \le \alpha\, \hat{\xi}_{0.5}\} \bar{w}_{hci}$.

L'*ordonnée de la courbe de Lorenz (OCL)* d'une population finie précise la part du revenu que reçoit le percentile $100p$ le plus pauvre de la population. On la définit comme une fonction de p $(0 \le p \le 1)$ qui représente simplement le revenu cumulatif en fonction de la part de la population. En tant que paramètre, nous l'exprimons sous la forme

$$L(p) = \frac{1}{\mu_Y} \int_0^p \xi_q \, dq$$

où μ_Y est la moyenne de la population et où ξ_q est la fonction de quantile. Dans le cas d'une grande population sans double, l'expression ci-dessus est représentée approximativement par

$$L_N(p) \approx \sum_U \frac{I\{F_i \le p\}\, Y_i}{\mu_N} \frac{1}{N}$$

et estimée par

$$\hat{L}(p) = \sum_s \frac{I\{\hat{F}_{hci} \le p\}\, y_{hci}}{\hat{\mu}} \bar{w}_{hci}$$

où $\hat{\mu} = \sum_s \bar{w}_{hci} y_{hci}$ et $\hat{F}_{hci} = \hat{F}(y_{hci})$.

Par *part du quantile (PQ)*, nous entendons la proportion du revenu total que partage la population associée à un intervalle de quantile $[\xi_{p_1}, \xi_{p_2})$, soit:

$$Q_N(p_1, p_2) \approx \sum_U \frac{I\{ep_1 \le F_i \le p_2\}\, Y_i}{\mu_N} \frac{1}{N} = L_N(p_2) - L_N(p_1).$$

Pour $0 \le p_1 < p_2 \le 1$ nous estimons la part du quantile en remplaçant les paramètres par leur valeur estimée.

La mesure la plus utilisée de l'inégalité globale de la répartition du revenu, à savoir l'*indice de Gini*, correspond, par définition, à la surface comprise entre la courbe de Lorenz et l'axe de 45°, normalisée de façon à ce que sa valeur soit comprise entre 0 et 1, c'est-à-dire $G = 1 - 2\int_0^1 L(p)\,dp$. Dans le cas d'une population finie, l'estimation de l'indice de Gini prend la forme

$$\hat{G} = \sum_s \frac{[2\hat{F}_{hci} - 1]y_{hci}}{\hat{\mu}} \tilde{w}_{hci}.$$

Le lecteur qui souhaite obtenir plus de précisions sur l'indice de Gini devrait consulter Nygärd et Sandström (1985).

De façon analogue à la courbe de Lorenz et à l'indice de Gini, Foster et Wolfson (1992) définissent la *courbe de polarisation* par la relation

$$B(p) = \int_{0.5}^{p} \frac{F^{-1}(q) - \xi_{0.5}}{\xi_{0.5}}\,dq,$$

qui, dans le cas d'une population finie, prend la forme

$$B(p) = \begin{cases} 0.5 - p - \dfrac{1}{\xi_{0.5}} \sum\limits_U I\{p < F_i < 0.5\} Y_i \dfrac{1}{N}, & 0 < p \le 0.5, \\ 0.5 - p + \dfrac{1}{\xi_{0.5}} \sum\limits_U I\{0.5 \le F_i < p\} Y_i \dfrac{1}{N}, & 0.5 < p \le 1. \end{cases}$$

La courbe de polarisation montre, pour tout percentile de population, la mesure dans laquelle le revenu s'écarte de la médiane. La surface située au-dessous de la courbe de polarisation donne une mesure sommaire de la polarisation. La version normalisée, de façon à ce que sa valeur soit comprise entre 0 et 1, appelée *indice de polarisation* (*IP*), prend la forme

$$IP_N = \sum_U \frac{[2 - 2I\{F_i \le 0.5\} - 2F_i]Y_i}{\xi_N(0.5)} \frac{1}{N}$$

où $\xi_N(0.5), \mu_N, F_i$ ont été définis plus haut. On obtient l'estimation de l'indice de polarisation en remplaçant les paramètres par leurs valeurs estimées dans cette équation.

3. ESTIMATION DE LA VARIANCE

L'estimation de la variance de statistiques non lissées, comme les quantiles, et des fonctions axées sur les quantiles, comme la proportion de faible revenu ou l'indice de polarisation, n'est pas chose facile, surtout quand on ne peut soutenir l'hypothèse de l'échantillonnage aléatoire simple et qu'il faut tenir compte d'un plan de sondage complexe. Dans la première partie de la présente section, nous passerons en revue certains résultats de l'estimation de la variance des quantiles qui permettront de mieux comprendre ultérieurement l'estimation complexe de la variance des mesures de l'inégalité du revenu, ainsi que les résultats de l'estimation de la variance de certaines mesures, dont les ordonnées de la courbe de Lorenz. Dans la deuxième partie, nous décrivons les méthodes d'estimation de la variance faisant l'objet de la présente étude.

Woodruff (1952) a proposé une méthode permettant de calculer les intervalles de confiance de quantiles distincts. À partir de ces intervalles, Francisco et Fuller (1986) et Rao et Wu (1987) ont établi les estimateurs de la variance. Bien que l'estimateur dépende du niveau de confiance, Rao et Wu (1987) ont déterminé sa convergence asymptotique pour tout seuil de signification α. Au moyen de simulations de Monte Carlo, ils ont étudié les écarts-types des quantiles des échantillons en grappes estimés. Leurs résultats donnent à penser que le choix d'un intervalle de confiance de 95% pour déduire l'écart-type donne de bons résultats. Binder (1991) obtient une forme comparable de l'estimateur de la variance en appliquant la méthode de linéarisation.

L'augmentation de la puissance informatique a rendu très courante l'utilisation d'estimateurs jackknife de la variance dans le cas des fonctions lissées des totaux et des moyennes. L'application de la théorie asymptotique type à la médiane d'une distribution à densité continue bornée, f, montre que $nE(\hat{\xi}_{0.5} - \xi_{0.5})^2 - 1/[4f^2(\xi_{0.5})]$ quand $n \to \infty$. Efron (1979) fait remarquer que l'application de la méthode jackknife à la médiane de l'échantillon produit une estimation de la variance asymptotiquement non convergente puisque

$$n \operatorname{var}_{JK}(\hat{\xi}_{0.5}) \to \frac{1}{4f^2(\xi_{0.5})} [\chi_2^2/2]^2$$

où la moyenne de $[\chi_2^2/2]^2$ est égale à 2 et sa variance est égale à 20, ce qui signifie que l'estimateur jackknife de la variance a tendance à surestimer par 100%, en moyenne, la variance asymptotique correcte. Kovar (1987) confirme empiriquement l'incohérence des estimateurs jackknife avec suppression d'une unité pour un plan d'échantillonnage stratifié. Au moyen d'une simulation portant sur une population stratifiée, il montre que les estimateurs jackknife avec suppression d'une unité (il en examine six) donnent de piètres résultats, surestimant la variance réelle de 30 à 70% dans le cas d'un plan comptant deux unités par strate et des résultats encore pires dans le cas d'un plan comptant cinq unités par strate. Néanmoins, Shao et Wu (1989) montrent que, dans certaines conditions, la méthode jackknife avec suppression d'une unité présente des propriétés asymptotiques désirables pour l'estimation de la variance de statistiques non lissées. Ce résultat a encouragé Rao, Wu et Yue (1992) à appliquer la méthode jackknife avec suppression d'une UPÉ à l'échantillonnage stratifié à plusieurs degrés. Grâce à une étude en simulation limitée, il montre que le biais ainsi que le biais relatif de l'estimateur de la variance de la médiane diminuent à mesure que la taille de la grappe augmente quand la corrélation à l'intérieur de la grappe est invariable.

La méthode bootstrap d'estimation de la variance, mentionnée pour la première fois par Efron (1979), donne des résultats convergents dans le cas d'observations indépendantes, distribuées de façon identique (consulter aussi Babu

(1986)). Rao et Wu (1988) proposent une méthode bootstrap modifiée pour l'estimation de la variance dans le cas de plans d'échantillonnage stratifiés. Kovar (1987) et Kovar et coll. (1988) indiquent que la méthode donne de bons résultats pour la médiane quand la taille de l'échantillon auquel on l'applique est $n_h^* = n_h - 1$.

Dans le cas de la méthode d'estimation de la variance par groupement de demi-échantillons compensés (GDÉC), les grappes échantillonnées dans chaque strate sont réparties au hasard en deux groupes (demi-échantillons) auxquels on applique la méthode de répétition compensée. Rao et Shao (1996) montrent que cette méthode est asymptotiquement incorrecte en ce sens que la distribution du facteur pivot t associé ne converge pas vers une distribution normale réduite et que les intervalles de confiance connexes sont asymptotiquement incorrects. Pour surmonter cette difficulté, ils proposent de répéter indépendamment le groupement T fois puis de calculer la moyenne des T estimations résultantes de la variance. Il montre qu'un tel estimateur est asymptotiquement correct pour un plan d'échantillonnage aléatoire stratifié quand $\min n_h \to \infty$ et $T \to \infty$. Grâce à une petite simulation, ils montrent que la méthode donne de bons résultats pour des valeurs de T aussi petites que 15 dans le cas d'estimateurs lissés. La méthode du groupement répété de demi-échantillons compensées (GRDÉC) produit un meilleur estimateur de la variance de la médiane de la population que les méthodes jackknife et GDÉC en ce sens que le biais relatif et le coefficient de variation sont plus faibles. Récemment, McCarthy (1993) a examiné et comparé diverses méthodes d'estimation de la variance de la médiane fondées sur l'échantillonnage aléatoire simple sans remise d'une population finie. Son étude englobe la plupart des méthodes de rééchantillonnage.

Bien que les méthodes de linéarisation, qui sont précieuses en statistiques non linéaire, soient difficiles à appliquer aux quantiles, puisqu'il faut estimer la densité, Binder (1991), Binder et Kovačević (1995) et Kovačević et Binder (1997) ont obtenu des estimateurs convergents de la variance de certaines mesures non lissées de l'inégalité et de la polarisation du revenu en appliquant une méthode de linéarisation à des équations d'estimation. Le calcul des estimateurs obtenus selon cette méthode est plus simple que celui des estimateurs fondés sur le rééchantillonnage, mais nécessite un calcul théorique.

Plusieurs auteurs ont étudié l'estimation de la variance de l'indice de Gini en supposant que les conditions d'échantillonnage aléatoires simples étaient respectées (Glasser 1962; Sendler 1979; Sandström, Wretman et Walden 1985; et Yitzhaki 1991). Dans le cas d'un plan d'échantillonnage complexe, Love et Wolfson (1976) proposent une méthode «grossière de répétition de demi-échantillons». Sandström, Wretman et Walden (1988) comparent les méthodes de calcul de la variance approximative à la méthode jackknife avec suppression d'une unité d'échantillonnage dans le cas de trois plans d'échantillonnage, dont deux complexes.

L'estimation de la variance des ordonnées de la courbe de Lorenz et des parts de quantile correspondantes a reçu moins d'attention. Le calcul de leur variance asymptotique est assez complexe. À cet égard, on mentionnera les percées effectuées par Beach et Davidson (1983) et par Beach et Kaliski (1986). Leurs travaux ont pour cadre de référence le modèle de la superpopulation en vertu duquel on considère les poids de sondage constants quand on construit les estimations. Le fait que cette méthode soit fondée sur un modèle pourrait restreindre son application à des données tirées d'enquêtes par sondage pour lesquelles on juge le plan d'échantillonnage significatif.

Dans les sous-sections qui suivent, nous passons en revue les méthodes d'estimation de la variance utilisées dans le cadre de la présente étude.

3.1 Méthode Jackknife avec suppression d'une UPÉ

La méthode repose sur l'exclusion séquentielle (suppression) d'une UPÉ à la fois du calcul de l'estimation. Après la suppression, on modifie les poids des unités retenues dans l'échantillon de sorte que les poids supprimés soient compensés et que la FDC estimée à partir de l'échantillon résiduel ait les mêmes propriétés que la FDC originale. Représentons par $\hat{F}_{(gj)}(y)$ l'estimation de la FDC fondée sur un échantillon dont on a supprimé la gj-ième UPÉ, c'est-à-dire

$$\hat{F}_{(gj)}(y) = \hat{G}_{(gj)}(y) / \hat{N}_{(gj)}$$

où

$$\hat{G}_{(gj)}(y) = \sum_{h \neq g} \sum_{c} \sum_{i} w_{hci} I\{y_{hci} \leq y\} + \frac{n_g}{n_g - 1} \sum_{c \neq j} \sum_{i} w_{gci} I\{y_{gci} \leq y\}$$

et

$$\hat{N}_{(gj)} = \sum_{h \neq g} \sum_{c} \sum_{i} w_{hci} + \frac{n_g}{n_g - 1} \sum_{c \neq j} \sum_{i} w_{gci}.$$

L'estimateur jackknife «avec suppression d'une UPÉ» de la variance de $\hat{F}(y)$ est

$$v_{J1}(\hat{F}(y)) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{F}_{(gj)}(y) - \hat{F}(y))^2.$$

On peut établir la convergence asymptotique de $v_{J1}(\hat{F}(y))$ en se servant des résultats de Krewski et de Rao (1981).

Par souci de commodité, nous notons que l'on peut écrire toutes les mesures considérées ici sous la forme générale

$$\theta_N = \sum_U J(F_N, Y_i, \boldsymbol{\beta}) \frac{1}{N},$$

où $J(\cdot)$ est une fonction à valeur réelle qui dépend éventuellement du paramètre dérangeant $\boldsymbol{\beta}$. Le paramètre de population finie θ_N est alors estimé par

$$\hat{\theta} = \sum_s J(\hat{F}, y_{hci}, \hat{\boldsymbol{\beta}}) \bar{w}_{hci} \qquad (3.1)$$

où $\hat{\boldsymbol{\beta}}$ représente le vecteur estimé des paramètres dérangeants et où $\tilde{w}_{hci}$ sont les poids normalisés. Selon cette formule générale, l'estimation d'une mesure de l'inégalité du revenu calculée d'après l'échantillon après suppression de gjUPÉ est donnée par

$$\hat{\theta}_{(gj)} = \sum_s J(\hat{F}_{(gj)}, y_{hci}, \hat{\boldsymbol{\beta}}_{(gj)})\, \tilde{w}_{hci(gj)}$$

où $\hat{F}_{(gj)}$ et $\hat{\boldsymbol{\beta}}_{(gj)}$ sont les valeurs de la fonction de distribution et du paramètre dérangeant estimées d'après l'échantillon dont on a supprimé la gj-ième UPÉ et où

$$\tilde{w}_{hci(gj)} = \begin{cases} w_{hci}/\hat{N}_{(gj)}, & \text{si } h \neq g, \\ \dfrac{n_g}{n_g - 1} w_{gci}/\hat{N}_{(gj)}, & \text{si } h = g, c \neq j, \\ 0, & \text{si } h = g,\ c = j. \end{cases}$$

L'estimateur jackknife «avec suppression d'une UPÉ» résultant de la variance $\hat{\theta}$ est

$$v_{J1}(\hat{\theta}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\theta}_{(gj)} - \hat{\theta})^2. \qquad (3.2)$$

Si on remplace $\hat{\theta}$ par $\hat{\theta}_{\cdot} = \sum_g \sum_j \hat{\theta}_{(gj)}/n$, on obtient une variante de l'estimation jackknife de la variance. Nous la représentons par $v_{J2}(\hat{\theta})$. Manifestement, $v_{J2}(\hat{\theta}) \leq v_{J1}(\hat{\theta})$. Krewski et Rao (1981) ont démontré la convergence de (3.2) pour les statistiques lissées.

Dans le cas de l'estimation de la variance des quantiles et de la fonction de quantiles, nous commençons par calculer les quantiles d'après l'échantillon dont on a supprimé la gj-ième UPÉ,

$$\hat{\xi}_{(gj)}(p) = \inf\{y_{hci} \mid \hat{F}_{(gj)}(y_{hci}) \geq p, hci \in s \setminus (gj)\},$$

puis nous calculons $\hat{\theta}_{(gj)} = g(\hat{\xi}_{(gj)})$, et enfin nous utilisons l'équation (3.2) pour obtenir l'estimateur jackknife de la variance.

3.2 Méthode du groupement de demi-échantillons compensés (GDÉC) et méthode du groupement répété de demi-échantillons compensés (GRDÉC)

Originellement, la méthode des demi-échantillons compensés a été proposée pour les plans de sondage produisant deux grappes par strate. Or, le cas qui nous intéresse est celui où les strates comptent plus de deux grappes. Ordinairement, dans cette situation, on répartit les grappes (unités de premier degré) de chaque strate en deux groupes. Nous explorons l'idée formulée par Wu (1991) et nous simplifions son application en vue d'estimer la variance de la FDC. En premier lieu, dans chaque strate $h, (h = 1, ..., L)$, nous groupons les UPÉ au hasard en deux moitiés, h_1 et h_2, contenant $m_{h_1} = [n_h/2]$ et $m_{h_2} = n_h - m_{h_1}$ UPÉ, respectivement. Si nous posons que l'indicateur de groupe est

$$\delta_h^{(r)} = \begin{cases} 1, & h_1 \in r \\ -1, & h_2 \in r \end{cases}$$

où $r = 1, ..., R$ représente un demi-échantillon (échantillon répété), les demi-échantillons sont compensés sur les groupes si $\sum_{r=1}^{R} \delta_h^{(r)} = 0$ et $\sum_{r=1}^{R} \delta_h^{(r)} \delta_{h'}^{(r)} = 0, (h \neq h')$. L'utilisation d'une matrice de Hadamard d'ordre $R\,(L + 1 \leq R \leq L + 4)$ permet d'obtenir un ensemble minimal de demi-échantillons compensés.

L'estimateur de la fonction de distribution fondé sur le r-ième demi-échantillon est

$$\hat{F}^{(r)}(y) = \frac{\hat{G}^{(r)}(y)}{\hat{N}^{(r)}}$$

où

$$\hat{G}^{(r)}(y) = \sum_h \sum_c A_{hc}^{(r)} \sum_i w_{hci}\, I\{y_{hci} \leq y\}, \hat{N}^{(r)} = \sum_h \sum_c A_{hc}^{(r)} \sum_i w_{hci}$$

et où $A_{hc}^{(r)}$ représente le modificateur de poids qui demeure constant pour toutes les grappes d'un même demi-échantillon. Nous supposons que les poids de toutes les unités (ménages) d'une grappe sont rééchelonnés de façon égale par le modificateur $A_{hc}^{(r)}$.

Dans le cas de la méthode GDÉC type, quand n_h est pair, nous utilisons

$$A_{hc}^{(r)} = \begin{cases} 1 + \delta_h^{(r)}, & c \in h_1, \\ 1 - \delta_h^{(r)}, & c \in h_2 \end{cases} \qquad (3.3)$$

ce qui signifie que les poids sont modifiés par un facteur 2 ou un facteur 0 selon qu'une unité figure ou non dans l'échantillon répété. Dans les cas où n_h est impair, diverses modifications ont été envisagées [consulter Shao (1993) et Sitter (1993)].

La méthode que nous appliquons se fonde sur le plan type de rééchantillonnage avec répétition compensée et sur une variante de la méthode de rééchelonnage proposée par Shao (1993), soit

$$A_{hc}^{(r)} = \begin{cases} 1 + (1 - a_h)\delta_h^{(r)}, & c \in h_1; \\ 1 - (1 - b_h)\delta_h^{(r)}, & c \in h_2. \end{cases}$$

Le fait de maintenir la taille de l'échantillon de strate dans tout demi-échantillon répété signifie que

$$\sum_{c \in h1} [1 + (1 - a_h)\delta_h^{(r)}] + \sum_{c \in h2} [1 - (1 - b_h)\delta_h^{(r)}] = n_h,$$

ce qui donne

$$A_{hc}^{(r)} = \begin{cases} 1 + (1 - a_h)\,\delta_h^{(r)}, & c \in h_1; \\ 1 - (1 - a_h)\dfrac{m_{h_1}}{m_{h_2}}\delta_h^{(r)}, & c \in h_2. \end{cases} \qquad (3.4)$$

Pour être certain que les poids modifiés ne soient pas négatifs, a_h devrait satisfaire $0 \leq a_h < 1$. Quand n_h est pair, nous aimerions que (3.4) se réduise à (3.3). Conformément à l'idée de Shao (1993), nous voulons que l'estimateur de la

variance selon la méthode GDÉC concorde avec un estimateur convergent de la variance dans le cas des statistiques linéaires. Donc, nous établissons pour les facteurs de perturbation $1 - a_h$: particuliers aux strates les contraintes suivantes:

Pour toutes les valeurs de h: (i) $0 < 1 - a_h \leq 1$; (ii) $(1 - a_h)^2 (m_{h_1}/m_{h_2})^2 \approx 1$; (iii) $(1 - a_h)^2 m_{h_1}/m_{h_2} \approx 1$. Pour les n_h pairs, nous posons simplement $1 - a_h = 1$. Cependant, maintenir la contrainte $1 - a_h = 1$ pour les valeurs impaires de n_h exclurait toute contribution des grappes du premier demi-échantillon quand $\delta_h^{(r)} = -1$, [voir l'équation (3.4)]. Aux fins de l'étude en simulation, nous choisissons

$$1 - a_h = \sqrt{\frac{n_h}{2\, m_{h2}}} \tag{3.5}$$

qui se réduit à 1 quand n_h est pair. Quand la taille de l'échantillon de strate est impaire, l'expression est égale à $\sqrt{1 - 1/(n_h + 1)}$. Dans notre simulation, très peu de strates ont un n_h impair et nous obtenons $v_{GB1}(\hat{\mu}_Y) = v_{GB2}(\hat{\mu}_Y) \approx v_L(\hat{\mu}_Y)$, où $\hat{\mu}_Y$ est la moyenne de l'échantillon et où $v_L(\hat{\mu}_Y)$ est l'estimateur de la variance de linéarisation utilisé couramment. Cependant, nous sommes d'avis qu'il faut poursuivre les travaux visant à modifier la méthode GDÉC de façon à pouvoir traiter un grand nombre de strates contenant un nombre impair d'UPÉ.

À l'instar de la méthode jackknife, l'estimation de la mesure de l'inégalité du revenu calculée à partir du r-ième demi-échantillon est donné par $\hat{\theta}^{(r)} = \sum_s J(\hat{F}^{(r)}, y_{hci}, \hat{\beta}^{(r)}) \tilde{w}_{hci}^{(r)}$ où $\hat{\beta}^{(r)}$ est une estimation du paramètre dérangeant fondée sur le r-ième demi-échantillon et où $\tilde{w}_{hci}^{(r)} = \tilde{w}_{hci} A_{hc}^{(r)}$. L'estimateur résultant de la variance de $\hat{\theta}$ selon la méthode GDÉC prend la forme

$$v_{GB1}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2. \tag{3.6}$$

Si nous répétons T fois le groupement aléatoire des unités dans chaque strate, que nous calculons chaque fois $v_{GB1}(\hat{\theta})$ et que nous calculons la moyenne pour les T répétitions, nous obtenons l'estimateur de la variance selon la méthode du groupement répété des demi-échantillons compensés (GRDÉC), soit

$$v_{RG1}(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} v_{GB1_t}(\hat{\theta}).$$

En remplaçant $\hat{\theta}$ par $\hat{\theta} = \sum_r \hat{\theta}^{(r)}/R$, nous obtenons une variante de l'estimateur GDÉC (et GRDÉC) que nous dénotons $v_{GB2}(\hat{\theta})$ (et $v_{RG2}(\hat{\theta})$).

Inutile de dire que, pour étalonner les poids, il faut les modifier comme il convient pour chaque répétition GDÉC selon la même méthode avec demi-échantillon compensé.

3.3 Méthode bootstrap

Nous avons également examiné les résultats obtenus en appliquant la méthode bootstrap pour estimer la variance de diverses statistiques du revenu. Nous avons adopté pour l'échantillon en grappes stratifié le plan de rééchantillonnage bootstrap préconisé par Rao, Wu et Yue (1992). Brièvement, il s'agit de sélectionner indépendamment dans chaque strate un échantillon aléatoire simple de $n_h - 1$ grappes avec remise (pour les n_h grappes). On obtient le poids bootstrap, w_{hci}^*, en modifiant le poids w_{hci} original de la façon suivante:

$$w_{hci}^* = A_{hc} w_{hci}$$

où

$$A_{hc} = \frac{n_h}{n_h - 1} m_{hc}^*$$

et où m_{hc}^* est le nombre de fois que la hc-ième grappe est sélectionnée. Il convient de souligner que $\sum_c m_{hc}^* = n_h - 1$. La méthode est répétée indépendamment B fois; pour chaque échantillon bootstrap, nous calculons $\hat{\theta}^* = \sum_s J(\hat{F}^*, y_{hci}, \hat{\beta}^*) \tilde{w}_{hci}^*$, où $\hat{\beta}^*$ est une estimation du paramètre dérangeant fondée sur l'échantillon bootstrap et où $\tilde{w}_{hci}^* = w_{hci}^*/\sum_s w_{hci}^*$. L'estimation bootstrap de la variance de $\hat{\theta}$ est donnée par

$$v_{B1}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_{(b)}^* - \hat{\theta})^2.$$

On obtient une autre estimation de la variance en remplaçant $\hat{\theta}$ par la moyenne des répétitions bootstrap.

3.4 Linéarisation par la méthode des équations d'estimation

Contrairement aux méthodes de rééchantillonnage, la méthode des équations d'estimation (ÉE) de Binder (Binder, 1991; Binder et Patak 1994; Binder et Kovačević 1995), ne demande pas énormément de calcul. Cette méthode, fondée sur la linéarisation, produit des formules de variance asymptotique faciles à programmer malgré leur apparence complexe.

En appliquant les méthodes des équations d'estimation décrites par Binder et Patak (1994), Binder et Kovačević (1995) et Kovačević et Binder (1997), nous obtenons pour les estimateurs de la variance approximative des mesures étudiées du revenu des expressions telles que

$$v_{EE} = \sum_h \frac{n_h}{n_h - 1} \sum_c \left(u_{hc}^* - \bar{u}_h^* \right)^2 \tag{3.7}$$

où, $u_{hc}^* = \sum_i \tilde{w}_{hci} u_{hci}^*$, $\bar{u}_h^* = \sum_c u_{hc}^*/n_h$ et $\tilde{w}_{hci}$ est un poids normalisé. Pour obtenir plus de précisions sur la méthode des équations d'estimation, en particulier sur la relation entre les variables aléatoires u_{hci}^* et la fonction J, le lecteur devrait consulter Binder et Kovačević (1995). Pour les mesures du revenu étudiées ici, les variables aléatoires u_{hci}^* figurent dans le tableau 1.

Dans les cas de la proportion de faible revenu et de l'indice de polarisation, les expressions représentant les variables aléatoires u_{hci}^* dépendent de l'estimation de la fonction de densité correspondant à la médiane, $\hat{f}(\hat{\xi}_{0.5})$, et à la moitié de la médiane, $\hat{f}(\hat{\xi}_{0.5}/2)$. Binder et Kovačević (1995) décrivent une méthode appropriée pour estimer ces quantités.

Tableau 1
Définition de la variable aléatoire u^*_{hci} pour la méthode des équation d'estimation

Mesure	u^*_{hci}
Indice de Gini	$2[\hat{A}(y_{hci})y_{hci} + \hat{B}(y_{hci}) - \hat{\mu}(\hat{G}+1)/2]/\hat{\mu}$ où $A(y) = \hat{F}(y) - \frac{\hat{G}+1}{2}$ et $B(y) = \sum_s w_{hcj} y_{hcj} I\{y_{hcj} \geq y\}$.
Courbe de Lorenz	$[(y_{hci} - \hat{\xi}_p) I\{y_{hci} \leq \hat{\xi}_p\} + p\hat{\xi}_p - y_{hci}\hat{L}(p)]/\hat{\mu}$
Part du quantile	$\frac{1}{\hat{\mu}}[(y_{hci} - \hat{\xi}_{p_2}) I\{y_{hci} \leq \hat{\xi}_{p_2}\} - (y_{hci} - \hat{\xi}_{p_1}) I\{y_{hci} \leq \hat{\xi}_{p_1}\} + p_2\hat{\xi}_{p_2} - p_1\hat{\xi}_{p_1} - y_{hci}\hat{Q}(p_1, p_2)]$
Quantile	$-[I\{y \leq \hat{\xi}_p\} - p]/\hat{f}(\hat{\xi}_p)$, $\hat{f}(.)$ is the finite population density estimator
Proportion de faible revenu	$-\frac{\hat{f}(\hat{\xi}_{0.5}/2)}{2\hat{f}(\hat{\xi}_{0.5})}[I\{y_{hci} \leq \hat{\xi}_{0.5}\} - 1/2] + [I\{y_{hci} \leq \hat{\xi}_{0.5}/2\} - \hat{\Lambda}_{0.5}]$
Indice de polarisation	$\frac{2}{\hat{\xi}_{0.5}}[(\hat{\xi}_{0.5} - y_{hci})(I\{y_i \leq \hat{\xi}_{0.5}\} - 0.5) - (A(y_{hci})y_{hci} + B(y_{hci}) - (\hat{G}+1)\hat{\xi}_{0.5}/2 + \hat{G}y_{hci}/2)] + \frac{P\hat{I}}{\hat{\xi}_{0.5}\hat{f}(\hat{\xi}_{0.5})}(I\{y_{hci} \leq \hat{\xi}_{0.5}\} - 0.5) - P\hat{I}$

4. ÉTUDE EN SIMULATION

4.1 Données et conception de l'étude en simulation

La population sous-jacente à l'étude est celle de l'échantillon tiré en Ontario à l'occasion de l'Enquête canadienne sur les finances des consommateurs (EFC) de 1988. L'EFC est un complément annuel de l'Enquête mensuelle sur la population active. La population étudiée contient 7,474 ménages répartis entre 525 UPÉ tirées de 40 strates. Originellement, l'échantillon de l'Ontario a été sélectionné à partir de 91 strates que nous avons regroupées pour former des strates suffisamment grandes. Pour chaque ménage, nous disposions d'une valeur non négative du revenu annuel total. La courbe de répartition du revenu obtenue pour cette micro-population était fortement étalée vers la droite, les coefficients d'asymétrie et d'aplatissement étant égaux à 4.5 et 89.5, respectivement. Nous avons calculé les valeurs réelles des paramètres étudiés (mesures de l'inégalité et de la polarisation du revenu) pour cette population. Suivant la méthode de Neyman, nous avons répartit 108 grappes (UPÉ) entre les 40 strates. Puis, selon un plan d'échantillonnage par grappes à un degré, nous avons sélectionné avec probabilité proportionnelle à la taille et avec remise, des échantillons de strate dont la taille variait de 2 à 6 grappes. Enfin, dans une grappe sélectionnée, nous avons dénombré tous les ménages (6 à 20).

Les mesures du revenu sur lesquelles porte l'étude sont l'indice de Gini, la proportion de faible revenu, l'indice de polarisation, un ensemble de parts du quantile, un ensemble d'ordonnées de la courbe de Lorenz et les quantiles correspondants. Nous avons utilisé comme approximation de l'erreur quadratique moyenne des estimations de ces mesures la valeur de l'erreur quadratique moyenne empirique (EMSE pour *empirical mean squared error*), calculée sur 10,000 échantillons indépendants tirés conformément au plan d'échantillonnage décrit plus haut. Nous avons utilisé ces EMSE en remplacement des erreurs quadratiques moyennes «réelles» pour la comparaison avec les variances estimées.

Pour chacun des 10,000 échantillons, en plus de l'estimation des paramètres, nous avons estimé la variance d'échantillonnage en appliquant diverses méthodes, à savoir la méthode jackknife avec suppression d'une UPÉ (JK), la méthode du groupement de demi-échantillons compensés (GDÉC) et celle du groupement répété de demi-échantillons compensés (GRDÉC), la méthode bootstrap (BS) et la méthode de linéarisation basée sur des équations d'estimation (ÉE). Pour toutes les méthodes de rééchantillonnage, nous nous sommes servis de deux estimateurs distincts, l'un correspondant à l'estimation d'après l'échantillon complet et l'autre à la moyenne de toutes les répétitions. Le calcul des estimateurs de la variance est fondé sur 108 itérations dans le cas de la méthode jackknife et sur 100 itérations dans le cas de la méthode bootstrap. Les méthodes GDÉC et GRDÉC se fondent sur 44 échantillons répétés compensés obtenus à partir d'une matrice de Hadamard de 44 sur 44, avec en plus 3 répétitions pour la méthode GRDÉC, ce qui donne en tout, dans ce cas, 132 demi-échantillons répétés. Il convient de souligner que le nombre d'itérations jackknife dépend du nombre de grappes dans l'échantillon, donc n'est pas arbitraire. Pareillement, dans le cas de la méthode GDÉC, le nombre de strates détermine le nombre d'échantillons répétés. Pour que le nombre d'échantillons répétés soit comparable pour toutes les méthodes, nous avons décidé d'effectuer 100 (≈ 108) itérations bootstrap et 3 répétitions du GDÉC pour obtenir 132 échantillons répétés dans le cas du GRDÉC.

Afin d'évaluer l'exactitude et la précision des méthodes étudiées, nous avons calculé leur biais relatif (*rel. bias*) et leur variance relative (*rel. var.*), c'est-à-dire leur instabilité, sur les $A = 10{,}000$ simulations, soit

$$\text{rel.bias}(v_M) = \frac{\sum_a v_M(a)/A - \text{EMSE}}{\text{EMSE}}$$

$$\text{rel.var.}(v_M) = \frac{\sqrt{\sum_a [v_M(a) - \text{EMSE}]^2/A}}{\text{EMSE}}$$

Pour évaluer l'efficacité des intervalles de confiance fondés sur la loi normale, nous avons calculé les taux empiriques de couverture pour des niveaux de confiance nominaux de $100(1 - \alpha)\% = 90, 95$ et 99%,

$$\text{cov. prob.}\,(v_M) = \frac{\sum_a I\{|\hat{\theta}_a - \theta| \Big/ \sqrt{v_M(a)} \le z_{\alpha/2}\}}{A},$$

où $z_{\alpha/2}$ représente le $\alpha/2$-ième percentile centré réduit supérieur. Nous avons également calculé les taux d'erreur pour les intervalles latéraux supérieur et inférieur de la façon suivante,

$$\text{err_}L(v_M) = \frac{\sum_a I\{(\hat{\theta}_a - \theta) \Big/ \sqrt{v_M(a)} < -z_{\alpha/2}\}}{A}$$

$$\text{err_}U(v_M) = \frac{\sum_a I\{(\hat{\theta}_a - \theta) \Big/ \sqrt{v_M(a)} > z_{\alpha/2}\}}{A}.$$

Nous allons maintenant résumer le grand ensemble de résultats de la simulation séparément pour chaque mesure de l'inégalité du revenu.

4.2 Sommaire des résultats

Indice de Gini

Dans le cas de l'indice de Gini, la précision de toutes les méthodes d'estimation de la variance est comparable, le biais relatif, très faible, variant de -2.2% à -0.6%. De tous les estimateurs, c'est l'estimateur GRDÉC qui est entaché du biais relatif le plus faible.

Les divers estimateurs ont tous à peu près la même stabilité, de l'ordre de 87% à 99%. Les méthodes fondées sur le groupement de demi-échantillons compensés (GDÉC et GRDÉC) donnent des résultats un peu moins bons que les autres.

Pour l'intervalle de confiance de 95%, la probabilité de couverture varie de 92.6 (pour le GDÉC) à 93.9 (pour le GRDÉC). Le taux nominal de 2.5 % est une sous-estimation du taux d'erreur pour l'intervalle latéral inférieur dans le cas de toutes les méthodes étudiées. Selon nos résultats, le taux d'erreur pour l'intervalle inférieur varie de 4.6% à 5.4%, donc est plus de 100% plus élevé que le taux nominal de 2.5%. Inversement, le taux nominal est une surestimation du taux d'erreur pour l'intervalle supérieur dans le cas de toutes les méthodes (voir le tableau 2). Pour les intervalles de confiance de 90% et de 99%, les probabilités de couverture que nous avons calculées varient de 87.2 (pour le GDÉC) à 88.5 (pour le GRDÉC) et de 97.7 (pour le GDÉC) à 98.5 (pour le GRDÉC), respectivement. Pareillement, la comparaison des taux d'erreur obtenus pour les intervalles bilatéraux aux taux nominaux de 5% et de 1% indique la même tendance que pour le taux de 2.5%.

Dans l'ensemble, il est difficile de dire laquelle des méthodes d'estimation de la variance de l'indice de Gini étudiées ici est la meilleure, puisqu'elles produisent toutes des résultats comparables. L'application des méthodes à demi-échantillons compensés représente un léger compromis entre l'exactitude et la stabilité, puisque ces méthodes produisent les estimations les plus exactes, mais

Tableau 2
Valeurs des statistiques calculées pour évaluer les estimateurs de la variance de l'indice de Gini

		Jackknife		GDÉC		GRDÉC		Bootstrap		Équations d'estimation
		v_{J1}	v_{J2}	v_{GB1}	v_{GB2}	v_{RG1}	v_{RG2}	v_{B1}	v_{B2}	v_{EE}
Biais relatif (%)		-1.3	-1.3	-0.9	-1.1	-0.6	-0.7	-1.2	-2.2	-1.5
Variation relative (%)		87.1	87.1	99.4	99.2	95.2	95.1	88.5	87.6	87.0
Probabilité de couverture (95%)		93.8	93.8	92.6	92.6	93.9	93.9	93.5	93.4	93.7
Taux d'erreur des intervalles latéraux (2.5%)	Sup.	4.8	4.8	5.4	5.4	4.6	4.6	5.0	5.1	4.9
	Inf.	1.4	1.4	2.0	2.0	1.5	1.5	1.5	1.5	1.4

Tableau 3
Valeurs des statistiques calculées pour évaluer les estimateurs de la variance de la proportion de faible revenu

		Jackknife		GDÉC		GRDÉC		Bootstrap		Équations d'estimation
		v_{J1}	v_{J2}	v_{GB1}	v_{GB2}	v_{RG1}	v_{RG2}	v_{B1}	v_{B2}	v_{EE}
Biais relatif (%)		76.9	58.4	25.8	21.0	26.8	21.9	8.9	3.8	1.1
Variation relative (%)		113.1	81.0	62.5	61.0	40.8	39.5	35.1	33.5	31.0
Probabilité de couverture (95%)		97.4	96.9	94.6	94.1	96.2	95.7	93.9	93.3	93.2
Taux d'erreur des intervalles latéraux (2.5%)	Sup.	2.1	2.6	3.3	3.5	2.4	2.6	4.6	5.0	5.0
	Inf.	0.5	0.6	2.0	2.4	1.4	1.7	1.5	1.7	1.7

Tableau 4
Valeurs des statistiques calculées pour évaluer les estimateurs de la variance de l'indice de polarisation

		Jackknife		GDÉC		GRDÉC		Bootstrap		Équations d'estimation
		v_{J1}	v_{J2}	v_{GB1}	v_{GB2}	v_{RG1}	v_{RG2}	v_{B1}	v_{B2}	v_{EE}
Biais relatif (%)		95.4	56.5	13.9	11.2	14.7	12.1	6	2.9	4.2
Variation relative (%)		138.7	78.5	77.5	75.9	60	58.6	48.4	47	50
Probabilité de couverture (95%)		98.6	98	94.2	93.8	95.4	95.2	95	94.7	94.4
Taux d'erreur des intervalles latéraux (2.5%)	Sup.	0.7	0.8	2.2	2.4	1.4	1.4	1.8	2	2
	Inf.	0.8	1.1	3.6	3.9	3.2	3.4	3.2	3.4	3.6

parallèlement, les moins stables, de la variance. Les probabilités empiriques de couverture sont aussi très semblables pour tous les estimateurs. Les valeurs obtenues pour les taux d'erreur latéraux donnent à penser que l'utilisation d'intervalles de confiance asymétriques serait plus appropriée.

Proportion de faible revenu (PFR)

Toutes les méthodes étudiées ont tendance à produire une valeur surestimée de la variance de la PFR. Cependant, la grandeur de la surestimation varie fortement selon la méthode, l'écart allant de 1.1% pour la méthode des équations d'estimation (ÉE) à 76.9% pour la méthode jackknife (JK1). Des méthodes de rééchantillonnage, c'est la méthode bootstrap qui donne les meilleurs résultats, le biais relatif étant de 8.9% pour l'estimateur BS1 et de 3.8% pour l'estimateur BS2.

L'estimation de la variance de la proportion de faible revenu selon la méthode jackknife est très instable. Les estimateurs GRDÉC sont également moins stables. Les estimateurs bootstrap et ÉE donnent des résultats comparables, la variation relative allant de 31% à 45%.

Dans le cas de l'intervalle de confiance de 95%, les taux de couverture calculés d'après les estimations de la variance selon la méthode jackknife sont supérieurs aux taux nominal, soit 97.4% et 96.9%, à cause de la surestimation de la variance. Les autres méthodes produisent des taux de couverture un peu plus faibles que le taux nominal. Le calcul des taux d'erreur précisant les intervalles latéraux montre que toutes les méthodes aboutissent, pour l'intervalle inférieur, à un taux supérieur au taux nominal, ce qui laisse entendre que la proportion de faible revenu est caractérisée par une distribution asymétrique, fortement allongée vers la droite. Les taux de couverture et les taux d'erreur latéraux que nous obtenons pour les intervalles de confiance de 90% et de 99% suivent exactement la même tendance.

Dans l'ensemble, les méthodes bootstrap et ÉE sont nettement supérieures aux autres méthodes envisagées pour estimer la variance de la proportion de faible revenu.

Indice de polarisation

Les statistiques calculées pour évaluer les estimateurs de la variance de l'indice de polarisation indiquent que l'efficacité des diverses méthodes appliquées concordent fortement avec celle observée pour l'estimation de la variance de la proportion de faible revenu. De nouveau, ce sont les méthodes bootstrap et ÉE qui donnent les meilleurs résultats.

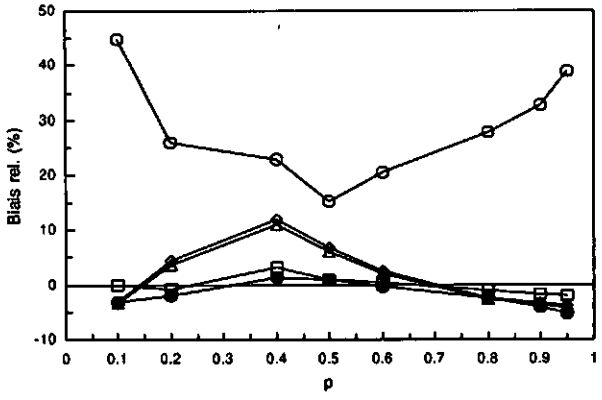
Ordonnées de la courbe de Lorenz et parts des quantiles

Les résultats complets obtenus pour les ordonnées de la courbe de Lorenz et pour les parts des quantiles figurent dans Kovačević, Yung et Pandher (1995). Nous présentons ici un résumé graphique des résultats (figures 1a à 1c). La méthode jackknife (les deux estimateurs) mène à une surestimation significative de la variance de toutes les ordonnées de la courbe de Lorenz (OCL) et de toutes les parts de quantile (PQ) envisagées. Pour l'OCL, le biais relatif de l'estimateur JK1 varie de 15% à 45% et celui de l'estimateur JK2, de 9% à 27%. Le biais relatif est plus faible au milieu de l'intervalle ($0 \leq p \leq 1$) et presque trois fois plus grand dans les queues (pour les grandes et les petites valeurs de p). Le biais relatif de l'estimateur JK1 est pratiquement 50% plus grand que celui de l'estimateur JK2 pour l'OCL. L'écart est attribuable à la différence significative entre l'estimation de l'OCL à partir de l'échantillon complet, d'une part, et la moyenne calculée pour les itérations jackknife, d'autre part.

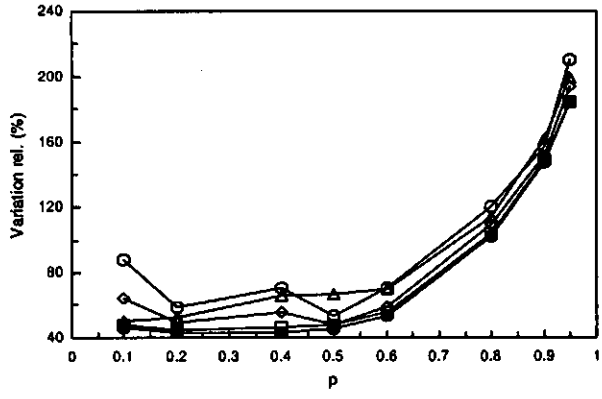
Nous obtenons des résultats similaires en ce qui concerne les estimateurs jackknife de la variance de la part du quantile, la variance étant surestimée de 26% à 237%, selon la part de population. La surestimation la plus importante se manifeste au milieu de l'intervalle. De nouveau, le biais relatif de l'estimateur JK1 surpasse d'environ 75% celui de l'estimateur JK2.

Le biais relatif observé pour les autres méthodes est très petit. Cependant, aucune tendance nette ne se dégage quant à la direction du biais – il est parfois positif, mais souvent négatif. La méthode bootstrap et la méthode des équations d'estimation donnent de meilleurs résultats que les autres, particulièrement autour de l'ordonnée de la courbe de Lorenz correspondant à $p = 0.5$ (voir figure 2a). Pour rendre l'illustration graphique plus claire, nous ne présentons pas les estimateurs jackknife dans les graphiques 2a et 2b.

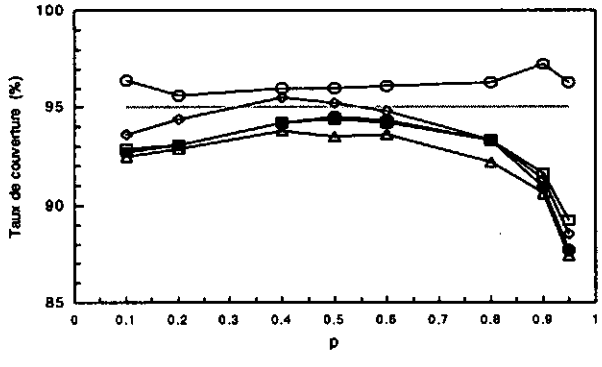
Les résultats de l'estimation de la variance de la part du quantile sont similaires. La méthode bootstrap et celle des équations d'estimation sont celles qui fournissent les estimations les plus précises de la variance de l'ordonnée de la courbe de Lorenz et de la part du quantile. Pour l'ordonnée



a) Biais relatif



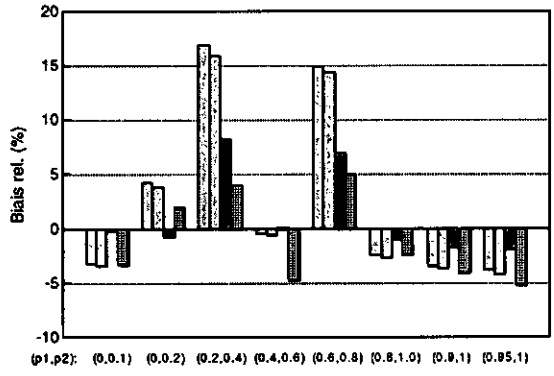
b) Variation relative



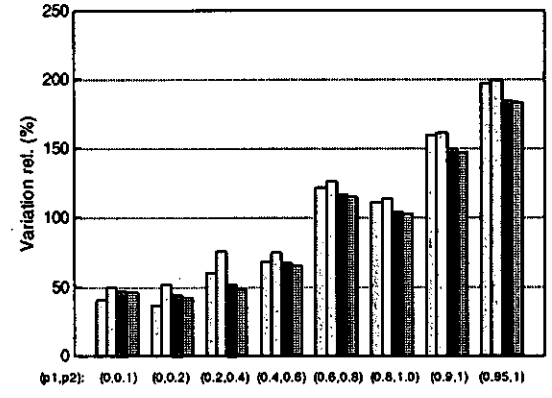
c) Taux de couverture (pour le taux nominal de 95%)

○-JK1 ◇-GBHS1 △-RGBHS1 □-BS1 ●-EE

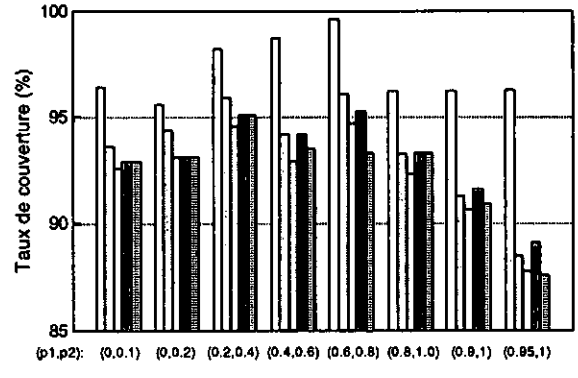
Figure 1. Propriétés des estimateurs de la variance de l'ordonnée de la courbe de Lorenz



a) Biais relatif
(les estimateurs jackknife ne sont pas représentés)



b) Variation relative
(les estimateurs jackknife ne sont pas représentés)



c) Taux de couverture (pour l'intervalle nominal de 95%)

JK1 GBHS1 RGBHS1 BS1 EE

Figure 2: Propriétés des estimateurs de la variance de la part du quantile

de la courbe de Lorenz, le biais relatif varie de -2% à +3% dans le cas de la méthode bootstrap et de -5% à +1% dans le cas de la méthode des équations d'estimation. Pour la part du quantile, le biais relatif varie de -3% à +8% pour les estimations obtenues par la méthode bootstrap, et de -3% à +5% pour celles obtenues par la méthode des équations d'estimation.

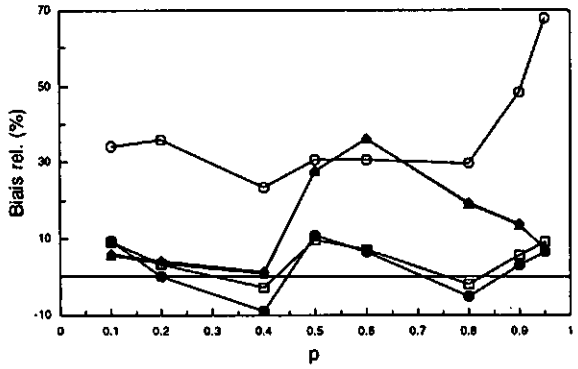
En ce qui concerne la stabilité des estimateurs de la variance, notre étude indique que les diverses méthodes donnent des résultats comparables, celle des équations d'estimation étant toutefois légèrement supérieure aux autres. Elle montre aussi qu'il existe une corrélation directe manifeste entre la mesure relative de la variation et la valeur de p.

Si nous comparons les méthodes d'après les propriétés de couverture des estimateurs de la variance de l'ordonnée de la courbe de Lorenz et de la part du quantile, nous constatons que, pour l'intervalle de confiance nominal de 95%, la méthode du jackknife produit des taux de couverture empiriques variant de 94.5% à 96.5% pour l'ordonnée de la courbe de Lorenz et de 94.5% à 99% pour la part du quantile. Les autres méthodes donnent des résultats comparables, les taux de couverture variant de 88% à 94%. Aussi bien dans le cas de l'ordonnée de la courbe de Lorenz que dans celui de la part du quantile, la couverture est meilleure quand la valeur de p est faible (voir figure 1c). Contrairement aux résultats obtenus pour l'indice de Gini, le taux d'erreur calculé pour l'intervalle latéral inférieur est environ deux fois plus grand que celui calculé pour l'intervalle latéral supérieur dans le cas de toutes les méthodes, pour l'ordonnée de la courbe de Lorenz ainsi que pour la part du quantile. Les résultats sont similaires pour les intervalles de confiance de 90% et de 99%.

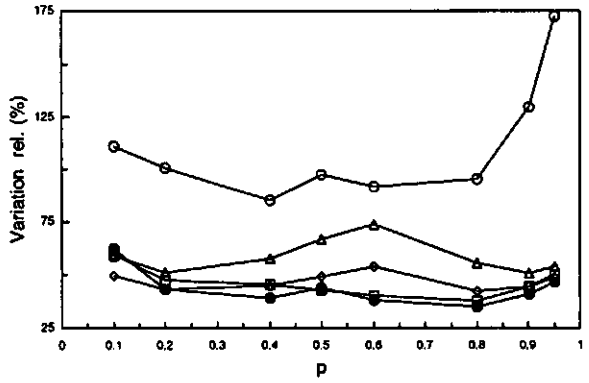
Nos résultats empiriques donnent à penser que la méthode jackknife ne convient pas pour estimer la variance de l'ordonnée de la courbe de Lorenz ou de la part du quantile, particulièrement pour les grandes et les petites valeurs de p. Les méthodes du groupement de demi-échantillons compensés et du groupement répété de demi-échantillons compensés donnent de meilleurs résultats. Toutefois, le choix le plus judicieux est celui de la méthode des équations d'estimation ou de la méthode bootstrap.

Quantiles

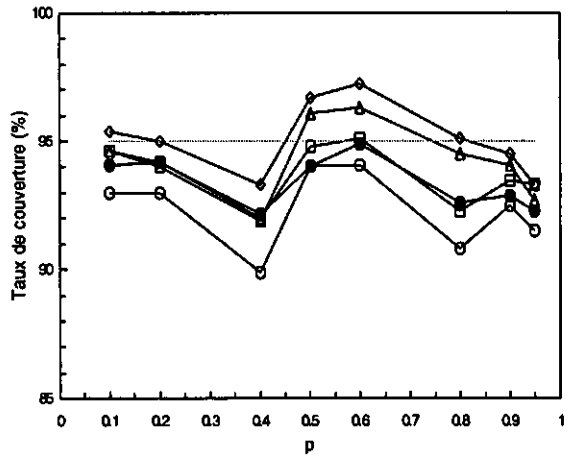
Les résultats obtenus pour les quantiles sont présentés dans Kovačević, Yung et Pander (1995) et résumés graphiquement ici. Le biais relatif de l'estimation de la variance des quantiles selon la méthode jackknife varie de 23% à 67% pour l'estimateur JK1 et de 17% à 52% pour l'estimateur JK2. Les surestimations les plus importantes sont obtenues pour les variances de $\hat{\xi}_{0.90}$ et $\hat{\xi}_{0.95}$. Le tableau est nettement différent pour les méthodes GRDÉC et GDÉC. Tandis que la variance de la médiane calculée selon ces méthodes est surestimée de 27%, l'estimation de la variance des quantiles latéraux est très précise, le biais relatif variant de 3% à 7%. Les autres méthodes étudiées donnent également de beaucoup meilleurs résultats pour les quantiles latéraux et des résultats raisonnablement meilleurs pour la médiane et les quantiles qui l'entourent. Plus précisément, la méthode bootstrap et celle des équations d'estimation produisent les estimations entachées des biais relatifs les plus petits, sans que se dégage toutefois une tendance nette quant à la direction du biais. Pour les estimateurs bootstrap, le biais relatif se situe dans l'intervalle (-5%, +9%) et pour les équations d'estimation, dans l'intervalle (-8%, +9%) (voir le figure 3a).



a) Biais relatif



b) Variation relative



c) Taux de couverture (pour l'intervalle nominal de 95%)

o-JK1 ◊-GBHS1 ▵-RGBHS1 □-BS1 •-ÉE

Figure 3. Propriétés des estimateurs de la variance des quantiles

Tableau 5
Classements des méthodes selon le biais relatif, la stabilité relative et la probabilité empirique de couverture

	Jackknife	GDÉC	GRDÉC	Bootstrap	ÉE (Taylor)	Meilleures méthodes
Indice de Gini	Toutes les méthodes donnent des résultats comparables					–
Quantiles	5, 5, 5	3, 4, 4	4, 3, 1	1, 2, 3	2, 1, 2	ÉE, BS
Ordonnées de la courbe	5, 5, 5	3, 4, 4	4, 3, 1	1, 2, 3	2, 1, 2	ÉE, BS
Parts de quantiles	5, 5, 5	3, 4, 4	4, 3, 1	1, 2, 2	2, 1, 3	BS, ÉE
Proportion de faible	5, 5, 5	3, 4, 2	4, 3, 1	2, 2 ,3	1, 1, 4	ÉE, BS
Indice de polarisation	5, 5, 5	3, 4, 4	4, 3, 2	2, 1, 1	1, 2, 3	BS, ÉE

Les estimateurs jackknife sont les moins stables. La stabilité des estimateurs GRDÉC, bootstrap et ÉE est comparable et, quand on fait la moyenne pour l'ensemble des quantiles, environ trois fois plus grande que celle des estimateurs jackknife. C'est autour de la médiane que nous avons obtenu la stabilité la plus forte (voir le figure 3b).

En général, la probabilité de couverture calculée pour les quantiles est inférieure à la valeur nominale pour toutes les méthodes étudiées, à quelques exceptions près pour les méthodes GDÉC et GRDÉC (voir figure 3c). En revanche, la comparaison des taux d'erreur des intervalles latéraux semble indiquer que toutes les méthodes donnent des résultats semblables, le taux calculé pour l'intervalle latéral supérieur (droit) étant plus élevé pour les quantiles inférieurs ($p = 0.1, 0.2$); pour d'autres, c'est la situation inverse que l'on observe, la queue inférieure (gauche) étant plus importante. Les résultats obtenus pour les intervalles de confiance de 90% et de 99% sont similaires.

Les résultats de cette étude empirique confirment qu'il faut éviter d'utiliser la méthode jackknife pour estimer la variance des quantiles. Dans le cas de la variance de la médiane en particulier, le choix le meilleur semble être celui de la méthode des équations d'estimation ou de la méthode bootstrap. Pour d'autres quantiles, la méthode du GRDÉC donne également de très bons résultats.

Nos résultats sont condensés dans le tableau 5 où nous classons le biais relatif, la variation relative et la probabilité de couverture obtenus pour les méthodes étudiées selon une échelle allant de 1 à 5 (1 = résultat le meilleur). Pour les méthodes de rééchantillonnage, nous avons calculé la moyenne des valeurs obtenues pour les deux estimateurs. Pour les quantiles, l'ordonnée de la courbe de Lorenz et la part du quantile, nous avons calculé la moyenne des valeurs obtenues pour l'ensemble des p. Dans la dernière colonne, nous indiquons les deux meilleures méthodes.

5. DISCUSSION ET CONCLUSION

La méthode de linéarisation au moyen des équations d'estimation est celle qui donne le meilleur résultat dans l'ensemble, produisant le biais relatif le plus petit, la variation relative la plus faible et d'assez bonnes propriétés de couverture. Vient ensuite la méthode bootstrap, qui est la meilleure des méthodes de rééchantillonnage envisagées. Les méthodes GRDÉC et GDÉC donnent des résultats aussi bons pour l'ordonnée de la courbe de Lorenz, la part du quantile et certains quantiles, en ce sens qu'elles produisent des estimations dont le faible biais relatif et la stabilité relative sont du même ordre que dans le cas de la méthode bootstrap. La méthode jackknife donne des résultats médiocres pour toutes les mesures de l'inégalité du revenu, sauf l'indice de Gini.

Il est bien connu que l'estimation de la variance selon la méthode jackknife donne de mauvais résultats dans le cas de fonctions non lissées. Le degré de lissage de la fonction J définie en (3.1) est un déterminant essentiel des propriétés asymptotiques de l'estimateur de sa variance. Si nous classons les mesures de l'inégalité du revenu dans la catégorie «lisses» ou «non lisses» d'après les fonctions J, nous constatons que le seul estimateur lisse étudié ici est l'indice de Gini. Il n'est donc pas surprenant que ce soit la seule mesure du revenu pour laquelle la méthode jackknife donne de bons résultats. Néanmoins, avant de décider d'utiliser l'estimateur jackknife de la variance, il convient de vérifier que les hypothèses en vertu desquelles la méthode est valide sont satisfaites.

Si l'objectif est de fournir une méthode d'estimation de la variance pour la longue liste de statistiques du revenu, notre étude empirique montre que la méthode bootstrap est la méthode par échantillonnage la plus satisfaisante et que la linéarisation au moyen d'équations d'estimation est non seulement la méthode la meilleure, mais aussi celle demandant le moins de calculs, même si elle nécessite certains travaux algébriques préparatoires, qui diffèrent pour chaque mesure du revenu.

Il convient de souligner que l'étude empirique se fonde sur un plan d'échantillonnage par grappes à un degré, les grappes étant sélectionnées avec probabilité proportionnelle à la taille, de sorte qu'il n'est pas tenu compte de la variabilité à l'intérieur de la grappe. Certaines études de faible portée aboutissent à des observations similaires au sujet des méthodes étudiées ici dans le cas de plans d'échantillonnage à deux degrés (voir Binder et Kovačević, 1995, et Kovačević et Binder, 1997

REMERCIEMENTS

Les auteurs remercient G.S. Pandher pour sa participation fructueuse au démarrage du projet, J. Gambino pour sa lecture minutieuse d'une version antérieure de l'article, ainsi que H. Mantel, rédacteur adjoint, les arbitres anonymes et le

rédacteur, pour leurs commentaires précieux qui ont permis d'améliorer considérablement cet article.

BIBLIOGRAPHIE

BABU, G.J. (1986). A note on bootstrapping the variance of sample quantile. *Annals of the Institute of Statistical Mathematics*,38-A, 439-443.

BEACH, C.M., et DAVIDSON, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies*, 50, 723-735.

BEACH, C.M., et KALISKI, S.F. (1986). Lorenz curve inference with sample weights: an application to the distribution of unemployment experience. *Applied Statistics*, 35, 38-45.

BINDER, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 34-42.

BINDER, D.A., et KOVAČEVIĆ, M.S. (1995). Estimation de l'inégalité du revenu d'après les données d'enquête: application de la méthode des équations d'estimation. *Techniques d'enquête*, 21, 151-159.

BINDER, D.A., et PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

EFRON, B. (1979). Bootstrap method: another look at the jackknife. *Annals of Statistics* 7, 1-26.

FOSTER, J.E., et WOLFSON, M.C. (1992). Polarization and the decline of the middle class: Canada and the U.S. (Manuscrit).

FRANCISCO, C.A., et FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

GLASSER, G.J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, 57, 648-654.

KAKWANI, N.C. (1980). *Income Inequality and Poverty*. Washington, D.C.: World Bank.

KOVAČEVIĆ, M.S., YUNG, W., et PANDHER, G.S. (1995). Estimating the Sampling Variances of Income Inequality and Polarization – An Empirical Study. Direction de la méthodologie, Document de travail, HSMD-95-007-E. Statistique Canada.

KOVAČEVIĆ, M.S., et BINDER, D.A. (1997). Variance estimation for measures of income inequality and polarization- the estimating equations approach. (A paraître dans *Journal of Official Statistics*).

KOVAR, J.G. (1987). Variance Estimation of Medians in Stratified Samples. Direction de la méthodologie, document de travail, BSMD-87-004-E. Statistique Canada.

KOVAR, J.G., RAO, J.N.K., et WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics* 16, 25-45.

KREWSKI, D., 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.

LOVE, R., et WOLFSON, M.C. (1976). Income inequality: statistical methodology and Canadian illustrations. Ottawa, Statistique Canada.

McCARTHY, P.J. (1993). Standard error and confidence interval estimation for the median. *Journal of Official Statistics*, 9, 673-689.

NYGÄRD, F., et SANDSTRÖM, A. (1981). *Measuring Income Inequality*. Stockholm: Almqvist & Wiksell International.

NYGÄRD, F., et SANDSTRÖM, A. (1985). The estimation of the Gini and the entropy inequality parameters in finite populations. *Journal of Official Statistics*, 1, 399-412.

RAO, J.N.K., et SHAO, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.

RAO, J.N.K., et WU, C.F.J. (1987). Methods for standard errors and confidence intervals from survey data: Some recent work. *Proceedings of the 46th Session of International Statistical Institute*, 3, 5-19.

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

RAO, J.N.K., WU, C.F.J., et YUE, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 225-234.

SANDSTRÖM, A., WRETMAN, J.H., et WALDÉN, B. (1985). Variance estimators of the Gini coefficient, simple random sampling. *Metron*, 43, 41-70.

SANDSTRÖM, A., WRETMAN, J.H., et WALDÉN, B. (1988). Variance estimators of the Gini coefficient – probability sampling. *Journal of Business and Economic Statistics*, 6, 113-119.

SEN, A.K. (1973). *On Economic Inequality*. London: Oxford University Press.

SENDLER, W. (1979). On statistical inference in concentration measurement. *Metrika*, 26, 119-122.

SHAO, J. (1993). Balanced repeated replication. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 544-549.

SHAO, J., et WU, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.

SITTER, R.R. (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, 80, 211-221.

WOLFSON, M.C. (1994). When inequalities diverge. *American Economic Review*, 84, 353-358.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

WU, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, 78, 181-188.

YITZHATI, S. (1991). Calculating jackknife variance estimators for parameters of the Gini method. *Journal of Business and Economic Statistics*, 9, 235-239.

Techniques d'enquête, Juin 1997
Vol. 23, n° 1, pp. 61-68
Statistique Canada

Estimation des variables instrumentales des flux bruts en présence d'erreur de mesure

K. HUMPHREYS et C. J. SKINNER[1]

RÉSUMÉ

Les auteurs étudient le problème qui consiste à estimer les taux de transition au moyen des données d'une enquête longitudinale, lorsqu'il existe des erreurs de classification. Ils examinent les approches faisant appel à des données auxiliaires sur les taux de classification erronée et ainsi que d'autres approches pour modéliser l'erreur de mesure. À partir de variables instrumentales nominales, ils suggèrent comment identifier et estimer les modèles qui comprennent des variables de ce genre en considérant un modèle à structure latente restreinte. Enfin, ils étudient les propriétés numériques des estimateurs implicites des variables instrumentales pour les taux des flux grâce aux données de l'étude par panel sur la dynamique de revenu.

MOTS CLÉS: Structure latente; longitudinal; erreur de classification; taux de transition.

1. INTRODUCTION

Un des principaux avantages d'une enquête longitudinale est qu'elle permet d'estimer les flux bruts, par exemple le nombre de chômeurs qui trouvent un emploi (lire Hogue et Flaim 1986). Quand on estime ces flux, le biais introduit par l'erreur de mesure pose un problème majeur. En effet, si les erreurs de classification peuvent avoir tendance à s'annuler lorsqu'on estime des proportions transversales (Chua et Fuller 1987), on ne peut en dire autant quand il s'agit de flux longitudinaux.

La première tentative en vue de résoudre le problème de l'erreur de mesure devrait clairement être de réduire cette erreur dans les méthodes de mesure de l'enquête. Biemer, Groves, Lyberg, Mathiowetz et Sudman (1991) examinent les approches pertinentes, aussi ne les reprendrons-nous pas ici. Néanmoins, aussi bonnes que puissent être les méthodes d'enquête, l'erreur de mesure reste inévitable et on devra en compenser les effets à l'analyse.

En général, les méthodes utilisées pour compenser l'erreur de mesure reposent sur un modèle hypothétique quelconque du processus d'erreur. Plusieurs auteurs décrivent divers modèles qu'on examinera à la partie 2. Pour les identifier et les estimer, on a habituellement besoin d'informations complémentaires, par exemple les renseignements recueillis lors des réinterviews (lire Meyer 1988). Les réinterviews coûtant cher et leur but étant rarement d'estimer les paramètres de la distribution des erreurs de mesure dans la pratique (Forsman et Schreiner 1991), d'autres méthodes s'avèrent nécessaires en l'absence de données issues des réinterviews. Quand l'erreur de mesure porte sur des variables continues et qu'on manque d'informations auxiliaires sur la distribution de cette erreur, on recourt couramment à l'estimation des variables instrumentales (lire Fuller 1987, partie 1.4). Par variable instrumentale, on entend une variable faisant partie des données d'enquête associées à la variable réelle, comprenant une erreur de mesure, mais non corrélée avec cette dernière. Les variables instrumentales et les hypothèses qui s'y rattachent nous fournissent l'information susceptible de remplacer celle des réinterviews et de faciliter l'identification et l'estimation des paramètres du modèle, parmi lesquels se retrouve la variable réelle. Dans cet article, nous verrons comment adapter la méthode d'estimation des variables instrumentales pour estimer les flux entre états discrets. On se rend compte que les modèles à structure latente (lire Bartholomew 1987, Ch. 2) procurent un cadre général dans lequel les hypothèses sur la variable instrumentale correspondent à certaines restrictions imposées aux paramètres du modèle. Notre approche rappelle donc celles qui imposent des restrictions aux modèles à structure latente (van de Pol et de Leeuw 1986; van de Pol et Langeheine 1990).

2. MODÈLES

Nous n'étudierons que le cas à deux possibilités, $t = 1$ et $t = 2$. Soit r, le nombre d'états dans lequel une personne peut être classée à chaque occasion. Appelons les états classés à $t = 1$ et $t = 2$ X et Y, respectivement, et les états véritables, x et y. On suppose l'existence d'un modèle où la valeur vectorielle de (X, Y, x, y) est le résultat indépendant d'un vecteur aléatoire commun de distribution $\text{pr}(X = i, Y = j, x = u, y = v)$.

La première hypothèse sur cette distribution, formulée par divers auteurs (lire Abowd et Zellner 1985; Poterba et Summers 1986 et Chua et Fuller 1987), et que nous reprendrons ici, est que les erreurs de classification effectuées aux deux occasions sont conditionnellement indépendantes, l'état véritable étant connu, c'est-à-dire

$$\text{pr}(X = i, Y = j \mid x = u, y = v) = \text{pr}(X = i \mid x = u, y = v)\,\text{pr}(Y = i \mid x = u, y = v). \quad \text{(A1)}$$

[1] K. Humphreys, Department of Psychology, Stockholm University, S-106 91 Stockholm, Sweden; C. J. Skinner, Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, United Kingdom.

Pareille hypothèse se retrouve couramment dans les modèles généraux à variables latentes (lire Anderson 1959). Elle semble raisonnable quand les méthodes de mesure de l'enquête sont indépendantes aux deux occasions. Si on obtient X rétrospectivement de la même interview qui a servi à mesurer Y, il est raisonnable de croire que la tendance des répondants à fournir des réponses exagérément cohérentes lors d'une interview entraînera une association positive entre les erreurs de classification. Marquis et Moore (1990) le prouvent avec l'enquête sur le revenu et la participation aux programmes. La possibilité d'une hétérogénéité individuelle dans les probabilités d'erreur de classification (certains répondants pouvant se montrer plus fiables que d'autres) est une deuxième raison pour questionner l'hypothèse d'indépendance conditionnelle. On lira à ce sujet Skinner et Torelli (1993), et Singh et Rao (1995). À la partie 4, nous accepterons l'hétérogénéité en supposant seulement que le modèle se confirme dans les cellules d'un tableau des variables observées.

La seconde hypothèse fondamentale est que l'erreur de classification ne dépend que de l'état réel courant, si bien que

$$\text{pr}(X = i \mid x = u, y = v) = \text{pr}(X = i \mid x = u) = K_{xiu}, \quad \text{(A2)}$$

par exemple, et

$$\text{pr}(Y = j \mid x = u, y = v) = \text{pr}(Y = j \mid y = v) = K_{yjv}, \text{ par exemple.}$$

K_{xiu} et K_{yjv} définissent les matrices $r \times r$ des erreurs de classification $\boldsymbol{K}_x = [K_{xiu}]$ et $\boldsymbol{K}_y = [K_{yiv}]$. Soit $\boldsymbol{P}$, la matrice $r \times r$ ayant pour ij-ième élément $\text{pr}(X = i, Y = j)$ et Π, la matrice $r \times r$ dont le uv-ième élément est $\text{pr}(x = u, y = v)$. On obtient l'équation matricielle

$$\boldsymbol{P} = \boldsymbol{K}_x \Pi \boldsymbol{K}_y'. \quad (1)$$

La matrice Π renferme les paramètres auxquels on s'intéresse alors que c'est la matrice $\boldsymbol{P}$ qu'on peut estimer à partir des valeurs X et Y venant de l'échantillon. Si on dispose des estimations complémentaires $\boldsymbol{K}_x$ et $\boldsymbol{K}_y$ et qu'il ne s'agit pas de matrices singulières, il est possible de résoudre l'équation (1) pour estimer Π. Par ailleurs, si on peut vérifier l'état réel grâce à une réinterview, on peut estimer directement $\boldsymbol{K}_x$ et $\boldsymbol{K}_y$ (Abowd et Zellner 1985). D'un autre côté, si on n'obtient qu'une nouvelle classification indépendante à la réinterview, on ne pourra estimer que les matrices interview-réinterview

$$\boldsymbol{K}_x \Delta_x \boldsymbol{K}_x' \text{ et } \boldsymbol{K}_y \Delta_y \boldsymbol{K}_y'$$

où $\Delta_x = \text{diag}[\text{pr}(x = u)]$ et $\Delta_y = \text{diag}[\text{pr}(y = v)]$ (Chua et Fuller 1987). Chaque matrice interview-réinterview est symétrique et la somme de leurs éléments donne un. Elles ne renferment donc que $r(r+1)/2 - 1$ éléments d'information «indépendants». Puisque chaque colonne de la matrice $\boldsymbol{K}$ et la diagonale de la matrice Δ ont pour somme un, le nombre de paramètres inconnus à chaque occasion est $r(r-1) + r - 1 = r^2 - 1$. Par conséquent, la différence entre le nombre de paramètres et le nombre d'éléments d'information est $r^2 - 1 - r(r+1)/2 + 1 = r(r-1)/2$ à chaque occasion, et le modèle est mal identifié lorsque $r \geq 2$. Chua et Fuller (1987) proposent qu'on ajoute une hypothèse naturelle afin de faciliter l'identification, à savoir supposer que les erreurs de mesure ne sont pas biaisées ou que

$$\text{pr}(x = i) = \text{pr}(X = i), \text{pr}(y = i) = \text{pr}(Y = i)\, i = 1, \ldots, r. \quad (2)$$

Dans ce cas, les valeurs faussement positives et faussement négatives ont tendance à s'annuler lors de l'estimation transversale des proportions. La nouvelle hypothèse réduit le nombre de paramètres de $r - 1$ à chaque occasion. Malgré cela, le modèle reste mal identifié lorsque $r \geq 3$, et Chua et Fuller (1987) sont forcés d'introduire de nouvelles hypothèses.

Voyons maintenant comment identifier le modèle sans les données d'une réinterview. Dans une régression linéaire simple où la covariable comporte une erreur de mesure la covariable, l'approche de la variable instrumentale (Fuller 1987, partie 1.4) suppose l'existence d'une variable instrumentale «observée» W, corrélée à la covariable mais indépendante de l'erreur de mesure et de l'erreur de l'équation de régression. Nous avons élargi cette hypothèse au cadre actuel en faisant en sorte que W soit une *variable instrumentale* si elle n'est pas indépendante de x et si

W et (X, Y) sont conditionnellement indépendants étant donné (x, y), (A3)

si W et y sont conditionnellement indépendants étant donné x. (A4)

En général, nous laisserons W être une variable nominale comprenant un nombre arbitraire s de catégories même si, en pratique, $s = r$ puisqu'on désire que W soit étroitement lié à x. Une possibilité spécifique consiste à faire correspondre W à l'état classé au temps $t - 1$. L'usage de la valeur déphasée d'une «covariable» comme variable instrumentale remonte aux débuts des travaux sur l'estimation des variables instrumentales (lire Reiersol 1941; Durbin 1954). Dans le cas présent, l'hypothèse A4 se confirme si les états réels suivent un processus de Markov et si les erreurs de classification sont conditionnellement indépendantes comme dans A1.

On peut représenter le modèle issu des hypothèses (A1) à (A4) par le graphique d'indépendance conditionnelle de la figure 1. Chaque point représente une variable. Les points ne sont pas reliés si les variables correspondantes présentent une indépendance conditionnelle avec les autres variables.

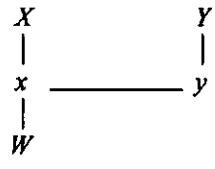


Figure 1. Représentation graphique de l'indépendance conditionnelle dans le modèle de base.

Le modèle illustre un modèle restreint à structure latente (Goodman 1974) où les variables X, Y et W observées sont conditionnellement indépendantes étant donné les variables latentes x et y, bref sont indépendantes dans les r^2 classes latentes définies par les couples (x, y). Ce modèle comprend

$2(r-1)r^2 + (s-1)r^2 + (r^2-1)$ paramètres donnés par les $(r-1)r^2$ paramètres $\text{pr}(X = i \mid x = u, y = v)$, les $(r-1)r^2$ paramètres $\text{pr}(Y = j \mid x = u, y = v)$, les $(s-1)r^2$ paramètres $\text{pr}(W = k \mid x = u, y = v)$ et les $r^2 - 1$ paramètres libres $\text{pr}(x = u, y = v)$. Ces paramètres sont soumis aux $2r(r-1)^2$ restrictions de (A2) et aux $(s-1)r(r-1)$ restrictions que suppose (A4). Nous nous attaquerons d'abord au cas $r = 2$. Dans ce cas, $4s + 7$ paramètres sont sujets à $2s + 2$ restrictions, ce qui laisse $2s + 5$ paramètres libres

$$\left\{K_{x2u}, K_{y2u}, \varphi_{2u}, \ldots, \varphi_{su}, \theta_u, \pi;\ u = 1,2\ v = 1,2\right\},$$

où $\varphi_{ku} = \text{pr}(W = k \mid x = u)$, $\theta_u = \text{pr}(y = 2 \mid x = u)$, et $\pi = \text{pr}(x = 2)$. Le nombre probable de cellules «libres» dans la grille de X par Y par W est $r^2 s - 1$, ou $4s - 1$ quand $r = 2$. Une condition essentielle à l'identification quand $r = 2$ est donc que $4s - 1 \geq 2s + 5$ ou $s \geq 3$. Malheureusement, cette condition ne suffit pas. Soit

$$R_u = \text{pr}(Y = 2 \mid x = u) = \sum_{v=1}^{2} K_{y2v}\, \theta_u^{v-1}(1 - \theta_u)^{2-v}, \qquad (3)$$

alors

$$\text{pr}(X = i, Y = j, W = k) = \sum_{u=1}^{2} K_{xiu}\varphi_{ku} R_u^{j-1}(1 - R_u)^{2-j}\, \pi^{u-1}(1 - \pi)^{2-u}. \qquad (4)$$

Les $4s - 1$ probabilités des cellules libres sont donc déterminées par seulement les $2s + 3$ paramètres

$$\left\{K_{x2u}, \varphi_{2u}, \ldots, \varphi_{su}, R_u, \pi;\ u = 1, 2\right\}$$

qui ne seront identifiés que si $4s - 1 \geq 2s + 3$ ou $s \geq 2$. En réalité, cette condition suffit pour identifier les paramètres, sauf dans quelques combinaisons exceptionnelles. (Lire Madansky (1960) pour le cas où $s = 2$, et Goodman (1974) pour le cas général où $s \geq 2$.)

Toutefois, bien que les $2s + 3$ paramètres qui précèdent soient généralement identifiés lorsque $s \geq 2$, on ne peut déterminer les quatre paramètres $K_{y21}, K_{y22}, \theta_1$ et θ_2, car ils ne sont associés qu'à deux paramètres connus, R_1 et R_2, au moyen de l'équation (3). Les paramètres principaux qui nous intéressent, notamment θ_1 et θ_2, restent mal identifiés, peu importe la valeur de s.

Pour les raisons qui précèdent, il faut imposer au moins deux autres restrictions au modèle si l'on veut identifier θ_1 et θ_2. À l'instar de Chua et Fuller (1987), on pourrait supposer que les erreurs de mesure ne sont pas biaisées comme dans (2), ce qui impose deux contraintes:

$$\pi = K_{x21}(1 - \pi) + K_{x22}\pi \qquad (5)$$

$$\theta_1(1 - \pi) + \theta_2 \pi = R_1(1 - \pi) + R_2\pi. \qquad (6)$$

Malheureusement, la première contrainte ne s'applique qu'aux paramètres déjà identifiés lorsque $s \geq 2$, si bien qu'on ne peut identifier θ_1 et θ_2. Nous formulerons donc une hypothèse supplémentaire, c'est-à-dire que le processus d'erreur est constant dans le temps, de sorte que

$$K_{xiu} = K_{yiu} = K_{iu}, \text{ par exemple, pour } i, u = 1,2,\ldots,r. \qquad (A5)$$

Pareille hypothèse semble élémentaire et naturelle si on applique la même procédure de mesure dans le cadre de l'enquête d'année en année. L'hypothèse élimine le problème de sous-identification du cas $r = 2$ examiné précédemment, puisqu'en identifiant $K_{xiu} = K_{iu}$ et R_u, on peut déterminer θ_u à partir de (3)

$$\theta_u = (R_u - K_{21})/(K_{22} - K_{21}) \qquad (7)$$

(en excluant le cas ordinaire où les variables quantifiées sont indépendantes des variables réelles, si bien que $K_{22} = K_{21}$).

En résumé, quand les hypothèses (A1) à (A5) se confirment et que $r = 2$, le modèle comporte $2s + 3$ paramètres libres $\{K_{2u}, \varphi_{2u}, \ldots, \varphi_{su}, \theta_u, \pi;\ u = 1,2\}$ que l'on peut identifier si $s \geq 2$, sauf dans les cas exceptionnels comme ceux analysés par Madansky (1960).

Enfin, revenons au cas général r. Puisque (A5) impose $(r-1)r$ restrictions, il s'ensuit qu'on compte $2(r-1)r^2 + (s-1)r^2 + (r^2-1) - [2r(r-1)^2 + (s-1)r(r-1)] - (r-1)$ $r = 2r^2 + sr - 2r - 1$ paramètres libres. Il y a donc $r^2 s - 1$ cellules libres probables dans le tableau de X par Y par W et, dans l'ensemble, on pourra identifier le modèle si $r(r-1)(s-2) \geq 0$. La condition relative à l'identification des paramètres reste donc $s \geq 2$, pour une valeur quelconque de $r \geq 2$. On peut dire d'autre part que

$$R_{ju} = \text{Pr}(Y = j \mid x = u) = \sum_{v=1}^{r} K_{jv}\theta_{uv}$$

où $\theta_{uv} = \text{pr}(y = v \mid x = u)$. Par conséquent, il est possible de déterminer θ_{uv} à partir de R_{ju} et K_{jv}, également identifiés, pourvu que la matrice $[K_{iu}]$ ne soit pas singulière. Pour le cas général r, le modèle est identifié en vertu des hypothèses (A1) à (A5), sauf dans les cas exceptionnels examinés par Goodman (1974).

3. ESTIMATION

Supposons que pour un échantillon de taille n on retrouve le chiffre n_{ijk} dans les cellules du tableau de contingence $r \times r \times s$ de $X \times Y \times W$, et que les cellules aient une distribution multinomiale selon les paramètres n et $p_{ijk} = \text{pr}(X = i, Y = j, W = k)$. La vraisemblance logarithmique implicite est

$$l = \sum_i \sum_j \sum_k n_{ijk}\, \log p_{ijk}.$$

En vertu d'un plan d'échantillonnage complexe, on peut considérer que n_{ijk} est un chiffre pondéré, ce qui débouche sur une pseudo-vraisemblance logarithmique (Skinner 1989). Les estimateurs des paramètres obtenus lorsqu'on maximise l porteront le nom d'estimateurs de *variables instrumentales* (VI).

Dans le reste de l'article, nous n'examinerons que le cas $r = s = 2$ en vertu du quel le modèle vient d'être identifié (sauf pour les valeurs aberrantes des paramètres). On pourrait établir que $p_{ijk} = n_{ijk}/n$ et résoudre les équations (6) et (7) pour les paramètres inconnus. Si elles se retrouvent dans les limites autorisées, c'est-à-dire dans la fourchette [0,1], les solutions résultantes correspondront aux estimations des VI. Dans la pratique cependant, on se rend compte que les échantillons de taille moyenne débouchent sur des solutions souvent irréalisables. De plus, les calculs sont complexes. Nous avons jugé plus facile de maximiser l directement au moyen des méthodes numériques de l'ensemble GAUSS (Edlefsen et Jones 1984) ou d'ensembles qui conviennent aux modèles à structure latente reposant sur l'algorithme EM, tel PANMARK (van de Pol, Langeheine et de Jong 1991). Un ensemble à structure latente permettrait d'ajuster un modèle à deux classes sans restriction, puis d'estimer θ_1 et θ_2 avec (7). Rien ne garantira néanmoins que les estimations résultantes tombent dans la fourchette [0,1] acceptable. S'ajouterait en outre la complication qu'on doive établir l'erreur-type de l'estimation de θ_1 et de θ_2 à partir de la matrice des covariances des estimations de $(R_1, R_2, K_{21}, K_{22})$. Il nous a semblé plus commode d'ajuster directement le modèle comme un modèle restreint à structure latente. Un autre avantage de cette approche est qu'elle s'étend naturellement à l'ajustement de modèles analogues d'un sous-groupe à l'autre, sous réserve de la contrainte éventuelle que certains paramètres restent constants entre les sous-groupes. Nous y reviendrons à la partie 4.

Avec les hypothèses multinomiales, il se pourrait que l'erreur-type repose sur la dérivée seconde du logarithme du rapport de vraisemblance obtenu pour l'estimation des VI. Pareille approche devient toutefois problématique quand la valeur maximale de l se trouve à la limite de l'espace paramétrique. Une solution consiste simplement à prendre les valeurs des paramètres telles que connues, à la limite. Le risque est qu'on sous-estime l'incertitude. Baker et Laird (1988) examinent deux autres approches au calcul de l'intervalle estimatif des paramètres dans des circonstances de ce genre: une méthode bootstrap et la méthode du profil de vraisemblance. La première méthode suppose le prélèvement réitéré d'échantillons multinomiaux où p_{ijk} est égal à n_{ijk}/n, et l'inscription de la distribution des paramètres estimatifs pour une suite d'échantillons bootstrap. La méthode du profil de vraisemblance donne les intervalles estimatifs d'un paramètre sous la forme des jeux de valeurs du paramètre non rejetées par un test du ratio de vraisemblance. Ces méthodes sont illustrées à la fin de la partie 4.

4. EXEMPLES NUMÉRIQUES

Aux fins d'illustration numérique, nous nous servirons des données du sous-échantillon à probabilité identique de l'étude américaine par panel sur la dynamique du revenu (EPDR) (lire Hill 1992). Nous examinerons deux états, occupé et non occupé, codés 1 et 2 respectivement, donc limiterons encore une fois notre attention au cas binaire. Pour plus de commodité, nous ne tiendrons pas compte des non-réponses et prendrons l'échantillon de 5,357 personnes de 18 à 64 ans de 1986 avec les valeurs entières pour les variables situation d'emploi en 1985, en 1986 et en 1987, possession d'une voiture, âge, sexe et degré de scolarité.

Les propriétés de l'estimateur VI sont évaluées de deux façons. Tout d'abord, à la partie 4.1, on compare le biais et l'erreur-type de l'estimateur VI à ceux de l'estimateur "non rajusté" des variables instrumentales hypothétiques, pour diverses associations avec x. Deuxièmement, à la partie 4.2, on étudie les conséquences de l'utilisation de différentes variables réelles de l'EPDR comme variables instrumentales.

4.1 Propriétés du biais et de l'erreur-type des estimateurs des variables instrumentales hypothétiques

Les paramètres qui nous intéressent le plus sont les probabilités conjointes $\text{pr}(x = i, y = j)$ ou les probabilités conditionnelles $\text{pr}(y = j \mid x = i)$ qui en dérivent. Les estimateurs simples "non rajustés" des paramètres reposent sur les proportions de l'échantillon correspondantes des variables X et Y classées et ont pour espérance $\text{pr}(X = i, Y = j)$, pour un échantillonnage multinomial. Puisque $\text{Pr}(X = i, Y = j)$ diffère généralement de $\text{pr}(x = i, y = j)$, les estimateurs non rajustés sont habituellement biaisés. En supposant que les hypothèses (A1) à (A5) du modèle se vérifient, les estimateurs VI de $\text{pr}(x = i, y = j)$ ne seront pas biaisés de façon asymptotique, même avec une variance supérieure à celle des estimateurs non rajustés. Nous essayerons ici de déterminer l'importance du compromis entre le biais des estimateurs non rajustés et le relèvement de la variance des estimateurs VI. On supposera que les hypothèses du modèle (A1) à (A5) tiennent et que l'échantillon est assez important pour que l'estimateur VI ne présente pas de biais.

Dans le cadre de notre analyse numérique, il est préférable d'utiliser certaines valeurs "réalistes" des paramètres. Celles-ci viennent de l'arrondissement des estimations du flux annuel survenu entre 1986 et 1987 selon les analyses de la partie 4.2 (voir le tableau 3). Les valeurs des cinq paramètres libres du modèle indépendant de W qui ont été retenues sont $K_{21} = 0.03$, $K_{22} = 0.94$, $\text{pr}(x = 2) = \pi = 0.22$, $\text{pr}(y = 2, x = 1) = \theta_1(1 - \pi) = 0.03$ et $\text{pr}(y = 2, x = 2) = \theta_2 \pi = 0.19$. Les valeurs des deux autres paramètres libres $\varphi_{11} = \text{pr}(W = 1 \mid x = 1)$ et $\varphi_{12} = \text{pr}(W = 1 \mid x = 2)$ apparaissent dans des colonnes différentes du tableau 1. La statistique V de Cramér, qui définit le lien entre deux variables binaires en répartissant essentiellement la valeur chi carré sur un intervalle [0,1], résume la robustesse du lien entre les variables W et x. Pour chaque valeur possible des paramètres, le tableau 1 donne l'erreur-type estimative des estimateurs VI pour un échantillon $n = 5{,}357$ de l'EPDR. Le tableau 1 indique aussi le biais et l'erreur-type de l'estimateur non rajusté pour les mêmes valeurs paramétriques $K_{21}, K_{22}, \pi, \theta_1$ et θ_2, avec un échantillon de taille identique.

Pour montrer comment on calcule le biais des estimateurs non rajustés, considérons $\text{pr}(x = 1, y = 1)$. L'espérance de l'estimateur non rajusté du paramètre, soit $\text{pr}(X = 1, Y = 1)$, qu'on obtient à partir des valeurs $K_{21}, K_{22}, \pi, \theta_1$ et θ_2 et des

hypothèses (A1) à (A5), est 0.71. On doit la comparer à la valeur hypothétique de 0.75 pour $\text{pr}(x = 1, y = 1)$. Le biais est donc de $0.71 - 0.75 = -0.04$. Tel qu'indiqué précédemment, on suppose que le biais des estimateurs VI est égal à zéro. L'erreur-type des estimateurs non rajustés vient de la formule binomiale habituelle. Par exemple, l'erreur-type de l'estimateur non rajusté de $\text{pr}(x = 1, y = 1)$ est $\sqrt{0.71 \times 0.29/5{,}357} = 0.0062$, où 0.71 est la valeur de $\Pr(X = 1, Y = 1)$. On obtient l'erreur-type des estimateurs VI en prenant la valeur inverse de la matrice d'information prévue, donnée par $n \sum p_{ijk} H_{ijk}$, où H_{ijk} est la matrice 7×7 des dérivées secondes de $\log p_{ijk}$ pour les sept paramètres libres. Après différenciation, on donne aux paramètres leur valeur hypothétique, tel qu'indiqué précédemment. Notons que l'erreur-type obtenue grâce à la matrice d'information multinomiale est sans doute sous-estimée étant donné le plan d'échantillonnage complexe de l'EPDR.

L'erreur-type de l'estimateur VI suit une tendance évidente puisqu'elle augmente à mesure que le lien entre W et x s'affaiblit. La hausse est assez semblable pour tous les paramètres, par exemple le ratio de V = 0.20 contre V = 1.00 se situe entre 3 et 4 pour l'ensemble des paramètres. Dans tous les cas, l'erreur-type de l'estimateur VI est supérieure à celle de l'estimateur non rajusté. La perte d'efficacité entre l'estimateur VI "idéal" (association parfaite entre W et x) et l'estimateur rajusté se situe entre ces deux pôles. Grosso modo, la perte est plus importante pour les paramètres conditionnels que pour les paramètres inconditionnels. Cette perte d'efficacité peut être vue comme la conséquence du rajustement résultant de l'erreur de mesure de y, laquelle reste nécessaire même quand W mesure x à la perfection. En vertu d'une telle interprétation, il est plausible que l'efficacité des paramètres conditionnels diminue davantage, car ils "dépendent moins" des paramètres de la distribution des x marginaux que les données sur W permettent d'estimer.

Pour étudier le compromis entre le biais de l'estimateur non rajusté et la plus forte variance de l'estimateur VI, nous avons calculé la taille minimale n de l'échantillon nécessaire pour que l'EQM de l'estimateur VI soit inférieure à celle de l'estimateur non rajusté. Pour les plans d'échantillonnage complexes, on suppose que la taille de l'échantillon correspond à la taille de l'échantillon après avoir prix en compte l'effet de plan. Le tableau 2 indique la valeur minimale de l'échantillon pour des forces d'association variables entre W et x. Sans erreur de classification, toutes les entrées se trouvent à l'infini puisque l'efficacité des estimateurs non rajustés est toujours supérieure à celle des estimateurs VI. Pour un nombre hypothétique d'erreurs, délimité par $K_{21} = 0.03$ et $K_{12} = 0.06$, la taille de l'échantillon requise augmente rapidement à mesure que V diminue. La variation dans les rangées du tableau 1 et la variation du biais des estimateurs non rajustés expliquent en partie l'écart entre les rangées du tableau 2. Le biais de l'estimateur non rajusté de $\text{pr}(x = 2, y = 2)$ est donc relativement faible, ce qui se traduit par les fortes valeurs de la rangée correspondante, au tableau 2. Soulignons que la valeur de 1 pour $\text{pr}(x = 2, y = 1)$ et pour le facteur V de Cramér résulte du fait que les deux estimateurs ont la même erreur-type (voir le tableau 1), si bien que le biais des estimateurs non rajustés suppose que l'EQM de l'estimateur VI est plus faible pour $n \geq 1$.

La principale conclusion qu'on retire du tableau 2 reste que dans un certain nombre de situations pratiques, l'estimation des VI s'avère utile pourvu que les hypothèses du modèle soient valables, même s'il faut légèrement augmenter la taille de l'échantillon afin d'accepter des plans d'échantillonnage complexes.

Tableau 1
Biais et erreur-type sous diverses VI hypothétiques

		Valeur des paramètres hypothétiques de l'estimateur VI						
$\text{pr}(W = 1 \mid x = 1)$		1.0	0.1	0.1	0.1	0.3	0.1	0.5
$\text{pr}(W = 1 \mid x = 2)$		0.0	0.9	0.7	0.5	0.7	0.3	0.3
V de Cramér		1.0	0.74	0.59	0.42	0.34	0.24	0.17
		Valeur des paramètres hypothétiques de l'estimateur VI						
Paramètre estimatif	Bias (× 100) de l'estimateur non rajusté	Estimateur non rajusté	Estimateur VI					
$\text{pr}(x = 1, y = 1)$	-4.0	0.62	0.68	0.75	0.88	1.13	1.16	1.82	2.05
$\text{pr}(x = 1, y = 2)$	3.0	0.32	0.39	0.43	0.51	0.64	0.69	1.03	1.24
$\text{pr}(x = 2, y = 1)$	3.0	0.32	0.32	0.37	0.44	0.57	0.66	0.95	1.27
$\text{pr}(x = 2, y = 2)$	-2.0	0.51	0.59	0.65	0.73	0.89	1.06	1.42	1.99
$\text{pr}(y = 1 \mid x = 1)$	-3.9	0.37	0.55	0.55	0.64	0.81	0.88	1.30	1.58
$\text{pr}(y = 1 \mid x = 2)$	12.4	0.60	1.40	1.63	1.95	2.56	2.90	4.30	5.55

Nota: 1 = occupé, 2 = non occupé, $n = 5{,}357$; échantillonnage multinomial; biais des estimateurs VI égal à zéro.

Tableau 2
Taille de l'échantillon nécessaire pour que l'EQM de l'estimateur VI soit inférieure à celle de l'estimateur non rajusté (échantillonnage multinomial)

	Valeur présumée du V de Cramér hypothétique pour les estimateurs VI						
	1.0	0.74	0.59	0.42	0.34	0.24	0.17
Paramètre estimé	Taille n de l'échantillon requise						
$\text{pr}(x = 1, y = 1)$	28	59	132	300	320	971	1273
$\text{pr}(x = 1, y = 2)$	31	50	91	184	219	573	843
$\text{pr}(x = 2, y = 1)$	1	20	51	129	198	476	811
$\text{pr}(x = 2, y = 2)$	112	227	366	720	1184	2397	5070
$\text{pr}(y = 1 \mid x = 1)$	42	60	97	183	219	541	818
$\text{pr}(y = 1 \mid x = 2)$	57	81	121	216	281	633	1061

4.2 Résultats avec les variables instrumentales réelles

Les résultats de la partie qui précède reposaient sur des variables instrumentales hypothétiques. Pour illustrer de façon plus réaliste notre propos, nous envisagerons maintenant des variables instrumentales réelles éventuelles. La principal difficulté consiste à trouver une variable W qui respecte les hypothèses (A3) et (A4). Il est apparemment plus facile de trouver une variable qui satisfera l'hypothèse (A3) que (A4), essentiellement parce qu'il est plausible que beaucoup de variables quantifiables sans erreur se plient à l'hypothèse (A3). Des variables sans lien avec un changement de la situation d'emploi, donc qui respectent l'hypothèse (A4), paraissent plus difficiles à trouver.

Aux fins d'illustration, nous avons envisagé deux possibilités. Tout d'abord, nous avons retenu la possession d'une voiture pour W (W = 2 si l'intéressé possède une voiture, W = 1 dans le cas contraire). Il se peut que la mesure de cette variable présente une erreur, mais a priori, il est raisonnable de supposer que l'erreur en question n'est pas liée à l'erreur de mesure concernant la situation d'emploi. Ainsi, lorsqu'ils analysent les erreurs résultant de l'inscription de la possession d'une voiture lors du recensement britannique de 1981, Britton et Birch (1985, p. 67) concluent que la principale difficulté associée au faible nombre d'aberrations dérive des véhicules hors d'usage ou disponibles de façon temporaire - ceux loués, par exemple - mais on peut au moins supposer qu'en réalité, ces erreurs présentent peu de liens avec le genre d'erreurs qui survient quand on enregistre la situation d'emploi. Parallèlement, la possession d'une voiture pourrait servir de valeur de remplacement à la situation sociale ou économique associée à un changement de la situation d'emploi, de sorte que l'hypothèse (A4) semble plus douteuse. Dans notre exemple, nous supposons toutefois qu'il y a vérification de (A3) et (A4).

Comme deuxième exemple, nous avons pris pour W la situation d'emploi déphasée en 1985. Dans ce cas, la difficulté est que l'hypothèse (A4) suppose effectivement un processus de Markov pour l'historique de l'emploi et que les taux de transition sont identiques alors qu'en réalité, ils varient d'une personne à l'autre, ce qui infirme l'hypothèse (A4) (lire van de Pol et Langeheine 1990). Pour pouvoir s'écarter de l'hypothèse (A4), il faut désagréger l'échantillon en 16 groupes, définis par classement croisé entre l'âge (quatre groupes), le sexe et le degré de scolarité (études collégiales ou non). On suppose ensuite que le modèle s'applique à l'intérieur d'un sous-groupe et on se sert des tests du rapport de vraisemblance pour déterminer quels paramètres restent constants d'un sous-groupe à l'autre. Ces tests ne donnent qu'une idée très générale, car ils ne tiennent pas compte du plan d'échantillonnage de l'EPDR. On n'a pas relevé de preuve sensible d'écart dans les probabilités d'erreur de classification K_{ij} d'un sous-groupe à l'autre. Il n'existe pas non plus d'indication sensible d'écart dans $\Pr(W \mid x$, sous-groupe) entre les deux sous-groupes du degré de scolarité, pour chacun des huit sous-groupes définis par l'âge et le sexe. En supposant l'égalité des paramètres, on obtient une valeur chi carré non significative du rapport de vraisemblance de la qualité de l'ajustement égale à 52.9 pour 46 dl (46 correspondant au nombre de cellules $= 16 \times 8 = 128$, moins les deux paramètres K_{ij}, moins $16 \times 4 = 64$ $\text{pr}(x, y,$sous-groupe) paramètres moins $8 \times 2 = 16$ pr ($W \mid x,$sous-groupe) paramètres). Lorsqu'on combine les paramètres estimatifs du modèle désagrégé, on obtient l'estimation du flux global $\text{pr}(x,y)$.

Le tableau 3 présente l'estimation des paramètres principaux pour les deux sortes de variables instrumentales et la version désagrégée du second choix. On remarque d'abord que l'erreur-type de l'estimateur VI reposant sur la possession d'une voiture est relativement élevée. Le tableau 1 le laissait prévoir, car x et W sont faiblement liés (le V de Cramér est de 0.12). Malgré cela, les rajustements résultants, qui relèvent la valeur estimative des entrées diagonales, sont plausibles et les intervalles de confiance issus de cet estimateur VI semblent plus réalistes que ceux de l'estimateur non rajusté.

L'erreur-type de la deuxième variable instrumentale choisie est plus faible, ainsi qu'on pouvait s'y attendre puisque le lien avec X est plus fort (V de Cramér de 0.73). En

réalité, l'erreur-type n'est pas beaucoup plus importante que celle de l'estimateur non rajusté. L'intervalle de confiance (de deux erreurs-types) ne chevauche plus l'intervalle correspondant de l'estimateur non rajusté des quatre paramètres.

Tableau 3
Estimation non rajustée et estimation VI selon les données de l'EPDR

Paramètre	Estimation non rajustée	Estimations VI		
		VI = possession d'une voiture	VI = emploi déphasé	VI = emploi déphasé (Désagrégé)
$pr(x = 1, y = 1)$	0.719	0.773	0.766	0.757
	(0.006)	(0.033)	(0.008)	(0.007)
$pr(x = 1, y = 2)$	0.055	0.011	0.017	0.025
	(0.003)	(0.020)	(0.005)	(0.003)
$pr(x = 2, y = 1)$	0.061	0.018	0.024	0.032
	(0.003)	(0.019)	(0.004)	(0.003)
$pr(x = 2, y = 2)$	0.166	0.198	0.193	0.186
	(0.005)	(0.027)	(0.007)	(0.006)

Nota: Erreurs-types en vertu d'une hypothèse multinomiale entre parenthèses. Désagrégation selon l'âge (4 groupes), le sexe et le degré de scolarité (2 groupes).

Ainsi qu'on l'a vu précédemment, on peut remettre en question l'hypothèse (A4) pour la variable correspondant à l'emploi déphasé. La version désagrégée de l'estimateur pose des hypothèses "plus faibles", n'exigeant seulement que (A4) se confirme à l'intérieur d'un sous-groupe. Les estimations qui en résultent se rapprochent passablement de l'estimateur VI original et leur erreur-type est légèrement plus faible, peut-être parce qu'on s'est servi des données complémentaires sur le sexe, l'âge et le degré de scolarité (voir l'analyse, plus loin). Fait intéressant, la désagrégation atténue les effets du rajustement d'une valeur relativement faible dans chaque cas. Il se pourrait qu'en s'écartant de (A4), on ait tendance à trop rajuster l'estimateur VI et que l'approche par désagrégation retenue ici contribue à atténuer le biais et permette d'évaluer la sensibilité des résultats selon la spécification du modèle, quand la désagrégation repose sur d'autres variables.

Ainsi qu'on a pu le lire à la partie 3, les estimations VI se retrouvent souvent à la limite de l'intervalle [0,1]. En réalité, parmi les analyses présentées au tableau 3, seule l'analyse par désagrégation donne des estimations limites. Pour les 64 paramètres $pr(x = i, y = j$, sous-groupe) de $i, j = 1, 2$, sous-groupe $= 1, ..., 16$, cinq estimations se retrouvent à la limite (contre aucune pour les 18 autres paramètres $pr(W = 1 \mid X = 1)$, et le reste). L'erreur-type indiquée au tableau 3 suppose que les paramètres sont connus, si bien qu'on pourrait sous-estimer l'incertitude de l'estimation des paramètres agrégés $pr(x = i, y = j)$.

Le tableau 4 donne d'autres estimations de l'erreur-type pour un sous-groupe, soit les hommes de 26 à 35 ans sans études collégiales. L'estimation de $pr(x = 1, y = 2)$ et celles qui en dérivent, notamment $pr(y = 1 \mid x = 1)$, se situent à la limite. Comme au tableau 3, l'estimation "type" de l'erreur-type repose sur la matrice des observations recueillies et on considère que les paramètres estimés à la limite sont connus. Les estimations bootstrap de l'erreur-type (pour 10,000 répétitions) se rapprochent beaucoup des estimations types pour les paramètres dont l'estimation ne se retrouve pas à la limite. Ainsi on ne possède pas d'estimation type de l'erreur-type pour l'estimation VI de $pr(x = 1, y = 2)$ située à la limite. En fait, estimer l'écart-type de la distribution de l'échantillon aurait peu de sens dans un tel cas. Il serait plus sensé d'établir un intervalle de confiance unilatéral ce qu'on peut faire avec la méthode du profil de vraisemblance, qui donne [0.016], ou avec la méthode bootstrap du percentile qui donne [0.009]. Les intervalles correspondant pour $pr(y = 1 \mid x = 1)$ sont [0.983,1] et [0.990,1].

Tableau 4
Autres estimations de l'erreur-type pour les hommes de 26 à 35 ans sans études collégiales

Paramètre	Estimations VI	Erreur-type estimative	
		Type	Bootstrap
$pr(W = 1 \mid x = 1)$	0.947	0.011	0.011
$pr(W = 1 \mid x = 2)$	0.107	0.089	0.091
$pr(X = 1 \mid x = 1)$	0.969	0.006	0.007
$pr(X = 1 \mid x = 2)$	0.084	0.088	0.075
$pr(x = 1, y = 1)$	0.953	0.011	0.012
$pr(x = 1, y = 2)$	0	*	*
$pr(x = 2, y = 1)$	0.006	0.007	0.006
$pr(x = 2, y = 2)$	0.041	0.012	0.011
$pr(x = 1)$	0.953	0.011	0.011
$pr(y = 1 \mid x = 1)$	1	*	*
$pr(y = 1 \mid x = 2)$	0.128	0.139	0.117

Nota: $n = 455$; les estimateurs «types» reposent sur la matrice des informations recueillies en vertu de laquelle on présume que les paramètres à la limite sont connus; 10,000 répétitions pour la méthode bootstrap; hypothèses multinomiales.

5. CONCLUSION

L'erreur de mesure peut introduire un biais important dans l'estimation type des taux de transition issus de données longitudinales. Si on peut estimer les taux d'erreur de classification de façon indépendante, il est possible de recourir à diverses méthodes de rajustement. Le présent article montre comment ajuster l'erreur de mesure par l'estimation des variables instrumentales en l'absence de telles données.

Comme c'est le cas pour la méthode d'estimation classique des variables instrumentales, la principale difficulté consiste à trouver une variable qui satisfera les conditions d'une variable instrumentale. En outre, même quand ces conditions sont satisfaites, il est bon que la variable présente un lien robuste avec la situation réelle si on veut obtenir des valeurs raisonnablement précises. Lorsqu'on trouve une variable de

ce genre, la méthode d'estimation des variables instrumentales peut donner de bons résultats.

REMERCIEMENTS

Les auteurs remercient Wayne Fuller de qui vient l'idée à la base de cet article. Les recherches n'auraient pu être menées à bien sans l'aide fournie par l'Economic and Social Research Council (subvention H519 25 5005) dans le cadre de son programme Analysis of Large and Complex Datasets.

BIBLIOGRAPHIE

ABOWD, J.M., et ZELLNER, A. (1985). Estimating gross labor force flows. *Journal of Business and Economic Statistics*, 3, 254-283.

ANDERSON, T.W. (1959). Some scaling models and estimation procedures in the latent class model. Probability and Statistics, (Éd. U. Grenander). Stockholm: Wiksell and Almquist.

BAKER, S.G., et LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

BARTHOLOMEW, D.J. (1987). *Latent Variable Models and Factor Analysis*. London: Griffin.

BIEMER, P.P., GROVES, R.M., LYBERG, L.E., MATHIOWETZ, N.A., et SUDMAN, S. (1991). *Measurement Errors in Surveys* New York: Wiley.

BRITTON, M., et BIRCH, F. (1985). *1981 Census Post-Enumeration Survey*. London: Her Majesty's Stationery Office.

CHUA, T., et FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.

DURBIN, J. (1954). Errors in variables. *Revue de l'Institut International de Statistique*, 22, 23-31.

EDLEFSEN, L.E., et JONES, S.D. (1984). Reference Guide to GAUSS. Applied Technical Systems.

FORSMAN, G., et SCHREINER, I. (1991). The design and analysis of reinterview: an overview. Dans *Measurement Errors in Surveys*. (Éds., Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. et Sudman, S.). New York: Wiley.

FULLER, W.A. (1987). *Measurement Error Models*. New York: Wiley.

GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215-231.

HILL, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage.

HOGUE, C.R., et FLAIM, P.O. (1986). Measuring gross flows in the labor force: an overview of a special conference. *Journal of Business and Economic Statistics*, 41, 111-21.

MADANSKY, A. (1960). Determinental methods in latent class analysis. *Psychometrika*, 25, 183-198.

MARQUIS, K.H., et MOORE, J.C. (1990). Measurement errors in the Survey of Income and Program Participation (SIPP): Program Reports. *Proceedings of the 1990 Annual Research Conference*. US Bureau of the Census, 721-745.

MEYER, B.D. (1988). Classification-error models and labor-market dynamics. *Journal of Business and Economic Statistics*, 6, 385-390.

POTERBA, J.M., et SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 1319-1338.

REIERSOL, D. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica*, 9, 1-24.

SINGH, A.C., et RAO, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 90, 478-488.

SKINNER, C.J. (1989). Domain means, regression and multivariate analysis. Dans *Analysis of Complex Surveys*, (Ch. 3) (Éds. Skinner, C.J., Holt, D., et Smith, T.M.F). Chichester: Wiley.

SKINNER, C.J., et TORELLI, N. (1993). Measurement error and the estimation of gross flows from longitudinal economic data. *Statistica*, 53, 391-405.

VAN DE POL, F., et DE LEEUW, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods and Research*, 15, 118-141.

VAN DE POL, F., et LANGEHEINE, R. (1990). Mixed Markov latent class models. Dans *Sociological Methodology 1990*, (Éd. C.C. Clogg). Oxford: Basil Blackwell, 213-247.

VAN DE POL, F., LANGEHEINE, R., et DE JONG, W. (1991). PANMARK User Manual. Panel analysis using Markov chains. Version 2.2. Netherlands Central Bureau of Statistics.

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 69-80
Statistique Canada

Suréchantillonnage géographique dans les enquêtes démographiques aux États-Unis

JOSEPH WAKSBERG, DAVID JUDKINS et JAMES T. MASSEY[1]

RÉSUMÉ

Aux États-Unis, les enquêtes démographiques polyvalentes comptent souvent parmi leurs principaux objectifs la production d'estimations de petites sous-populations définies, par exemple, selon la race, l'origine ethnique et le revenu. Le suréchantillonnage géographique est l'une des techniques fréquemment envisagées pour accroître la fiabilité des statistiques sur ces petites sous-populations (ou domaines), dans la mesure où l'information sur les îlots ou les groupes d'îlots du Bureau of the Census permet d'identifier les régions où se concentrent les sous-populations auxquelles on s'intéresse. Les auteurs passent en revue les questions relatives au suréchantillonnage géographique des régions, parallèlement à la présélection des ménages en vue d'améliorer la précision des estimations sur les petits domaines. Ils présentent les résultats d'une évaluation empirique de la réduction de la variance obtenue par suréchantillonnage géographique et évaluent la robustesse de l'efficacité de l'échantillonnage dans le temps, à mesure que les données servant à la stratification perdent de leur actualité. L'article aborde aussi la question du suréchantillonnage simultané de plusieurs petites sous-populations.

MOTS CLÉS: Échantillonnage; stratification; populations rares.

1. INTRODUCTION

Ceux qui parrainent un grand nombre de vastes enquêtes démographiques polyvalentes souhaitent des analyses distinctes sur des sous-populations définies selon la race, l'origine ethnique et le revenu. En général, les échantillons équiprobables ne sont pas suffisants pour permettre une analyse assez précise de ces domaines, si bien qu'on doit procéder à un suréchantillonnage quelconque. Cette exigence soulève des problèmes méthodologiques intéressants, car aucun registre de la population américaine n'autorise le prélèvement d'échantillons stratifiés en fonction des domaines désirés. Le Bureau of the Census garde bien des listes des logements qui identifient les sous-populations désirées, mais les chercheurs de l'extérieur n'y ont pas accès. Quand l'enquête exige une interview personnelle, les chercheurs externes sont donc contraints à recourir aux techniques d'échantillonnage aréolaire. Le Bureau lui-même procède parfois à un suréchantillonnage géographique, car ses listes ne sont mises à jour que tous les dix ans. Nous décrirons des méthodes de suréchantillonnage efficaces pour les domaines précités, dans le contexte de l'échantillonnage aréolaire.

Les données du recensement décennal américain sur la concentration de diverses caractéristiques démographiques sont largement diffusées pour de petites unités géographiques; ainsi, on possède des données sur la race et l'origine ethnique pour chaque îlot et des données sur le revenu pour chaque groupe d'îlots. (Par «îlot», on entend une zone indivisible, circonscrite des quatre côtés par une route. Les groupes d'îlots sont constitués de plusieurs îlots contigus.) On peut recourir à ces données pour améliorer à peu de frais la précision des statistiques sur des sous-populations rares en suréchantillonnant les îlots ou les groupes d'îlots dans lesquels se concentrent les membres des groupes en question et en abandonnant ou en sous-échantillonnant les personnes qui n'appartiennent pas aux groupes rares recherchés. C'est à Kish (1965, partie 4.5) qu'on doit la théorie générale applicable aux plans d'échantillonnage de ce genre. Waksberg (1973) a fait une présentation personnelle de cette théorie en l'illustrant d'exemples tirés du recensement de 1960. Kalton et Anderson (1986) ont donné d'autres exemples pertinents et analysé les méthodes connexes, ce qu'on retrouve également dans un article rédigé par Kalton dans United Nations (1993). Dans le présent article, nous étendrons les illustrations antérieures à d'autres domaines, actualiserons les résultats à 1990 et évaluerons de façon empirique la robustesse de ces méthodes dans le temps.

Nous commencerons par passer brièvement en revue les aspects associés à l'élimination et au sous-échantillonnage des personnes qui ne font pas partie des groupes visés. Ensuite, nous examinerons la théorie de la répartition optimale en vertu de laquelle les strates sont définies d'après la densité des populations rares, et nous appliquerons cette théorie à plusieurs populations rares. La partie principale de l'article consiste toutefois en une évaluation empirique de la réduction de la variance qui résulte du suréchantillonnage géographique de diverses minorités et populations rares, ainsi que de la robustesse de telles réductions dans le temps. Enfin, nous aborderons les problèmes particuliers que pose l'échantillonnage simultané de plusieurs populations rares, avant d'établir nos conclusions.

2. COÛT DE L'ENQUÊTE ET DÉCISION D'EFFECTUER UNE PRÉSÉLECTION

Soit U, l'univers visé, par exemple les particuliers ou les ménages pour lesquels on dispose d'une base de sondage, et

[1] Joseph Waksberg, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.; David Judkins, Research Triangle Institute, 5901-B Peachtree-Dunwoody Road, Suite 500, Atlanta, GA 30325, U.S.A.; James T. Massey, autrefois de Westat Inc., maintenant décédé.

soit D, une petite sous-population présentant un intérêt particulier, par exemple les personnes de race noire que l'on ne peut distinguer du reste de U lors de l'échantillonnage. Soit Y, la valeur vectorielle des caractéristiques auxquelles on s'intéresse, par exemple le revenu annuel, la situation d'emploi et le nombre de consultations médicales de l'année antérieure. Certaines enquêtes n'ont d'autre objectif qu'estimer la distribution de Y dans D. Dans une enquête de ce genre, les éléments U-D identifiés au moment de la présélection des membres de U seront supprimés de l'échantillon. Un questionnaire général peu coûteux permet d'effectuer cette présélection et d'établir qui recevra le questionnaire intégral.

On désire parfois estimer la distribution de Y simultanément dans D et dans U. Dans ce cas, on gardera au moins quelques membres U-D découverts lors de l'interview de présélection afin de leur faire passer l'interview complète. Lorsqu'on recourt au suréchantillonnage géographique, l'échantillon initial comprendra un suréchantillon des membres de U-D habitant dans les régions à forte concentration de population D. Même lorsqu'on s'intéresse à l'univers U-D, on s'efforce habituellement d'éviter le suréchantillonnage de U-D là où la population D est très concentrée car la variation de la probabilité de sélection de U-D entraîne des effets exagérés du plan d'échantillonnage sur les statistiques relatives à U et à U-D. À cause des effets plus marqués du plan d'échantillonnage, la taille de l'échantillon supplémentaire de U-D n'entraîne souvent qu'une réduction négligeable de la variance des statistiques relatives à U-D. On préfère généralement utiliser les fonds supplémentaires destinés aux interviews additionnelles auprès des membres de la population U-D pour accroître la taille de l'échantillon initial.

Il est assez facile d'établir des méthodes de sous-échantillonnage qui aboutiront à l'obtention d'un échantillon équiprobable de U-D. Le sous-échantillonnage peut-être centralisé après la présélection, ou l'intervieweur peut s'en charger avant de laisser le ménage sélectionné, en se renseignant sur la composition du ménage. En effet, on a mis au point des techniques qui facilitent considérablement le sous-échantillonnage par l'intervieweur (Waksberg et Mohadjer 1991). Il n'est pas nécessaire d'apprendre à ce dernier comment effectuer un tirage aléatoire. Lors des sondages papier et crayon, l'intervieweur reçoit des instructions sur les aspects qui devront être abordés avec tel ou tel ménage avant l'interview, d'une maison à l'autre. Ces instructions sont randomisées centralement avant la présélection en vue de donner les taux de sondage souhaités. Avec l'IPAO, il est possible de programmer le sous-échantillonnage pour qu'il s'effectue automatiquement sur l'ordinateur portatif; c'est l'ordinateur qui signale à l'intervieweur quel ménage répondra au questionnaire complet et quel ménage sera rejeté consécutivement au sous-échantillonnage.

La décision de garder tous les membres de U-D échantillonnés ou de procéder à un sous-échantillonnage dépend de la taille relative de U et de U-D, de la précision des données qu'on requiert sur les deux populations et du coût relatif de l'interview intégrale et de l'interview de présélection. Soit $c\star$, le coût variable associé à l'échantillonnage d'un membre de U ainsi qu'à la collecte et au traitement de ses données et soit c', le coût variable associé à l'échantillonnage, à la sélection et à l'abandon d'un membre de U. L'équation $c = c\star/c'$ exprime le rapport entre le coût d'une interview complète et celui de l'interview de présélection. Lorsque c est beaucoup plus grand que 1, et que l'enquête porte sur U-D on devrait envisager le sous-échantillonnage même si, ce faisant, l'enquête gagne en complexité. L'interview intégrale étant par définition plus longue que l'interview de présélection, c devrait toujours être légèrement supérieur à 1. Avec les enquêtes par panel et les enquêtes longitudinales, il conviendrait d'inclure le coût des interviews subséquentes à $c\star$, de sorte que le coût de l'interview complète pourrait devenir un très grand multiple du coût de l'interview de présélection c'est-à-dire, $c >> 1$. La même remarque s'applique aux enquêtes qui nécessitent le prélèvement de spécimens, et des travaux de laboratoire onéreux, ainsi qu'aux enquêtes qui font appel à des spécialistes coûteux (un médecin par exemple) pour la collecte des données primaires. Pour ce genre d'enquête, nous recommanderions fortement de ne pas recourir seulement au suréchantillonnage géographique mais de le combiner à la présélection et au sous-échantillonnage. Quand l'enquête se fait de porte à porte et ne comporte qu'une interview donnée par un intervieweur moyen (en mesure de poser des questions et d'enregistrer les réponses mais pas de procéder à une évaluation technique ou anthropologique), c se situe souvent entre 3 et 5, valeur suffisante dans de nombreux cas pour justifier le recours au sous-échantillonnage de U-D dans les régions suréchantillonnées.

3. STRATIFICATION

Même s'il est impossible de distinguer D de U au moment de l'échantillonnage, on suppose qu'on possède certains renseignements sur la distribution de D et de U dans une série d'entités géographiques. Aux États-Unis, les îlots ou groupes d'îlots (GI) constituent des éléments géographiques naturels et le recensement décennal nous renseigne sur eux. (Avant le recensement de 1990, il n'y avait pas d'ilôts dans les régions rurales; pour le suréchantillonnage, on se servait d'unités plus importantes baptisées «secteurs de dénombrement».) Le Bureau of the Census diffuse largement les données sur la composition raciale et ethnique des îlots, ainsi que des données cartographiques qui permettent aux organismes de sondage d'identifier ces îlots lors des années subséquentes. Les données sur le revenu ne sont fournies qu'au niveau des GI.

Il est courant de stratifier les îlots ou les GI en fonction de la concentration locale de D. Ainsi, les îlots où D représente moins de 10% de la population de l'îlot pourraient constituer une strate. D'autres strates pourraient correspondre à 30% et à 60% de la population, ce qui donnerait un total de quatre strates. Le nombre optimal de strates ou de points de découpage n'a guère fait l'objet d'études empiriques. En général, le plan d'échantillonnage est d'autant plus efficace qu'il compte plus de strates, mais il arrive un moment où les complexités opérationnelles soulevées par un grand nombre de strates dépassent les gains d'efficacité. Une certaine sagesse remontant à Kish (1965), veut que la stratification

donne l'essentiel des gains quand on se limite à un nombre assez faible de strates.

4. RÉPARTITION OPTIMALE POUR UN DOMAINE

L'objectif consiste à adapter les formules générales pour la répartition optimale d'un échantillon stratifié, afin de les appliquer à la réduction de la variance attribuable au suréchantillonnage géographique. Le développement mathématique correspond essentiellement à celui de Kish (1965), selon la notation proposée par Kalton dans United Nations (1993). Supposons que la population soit divisée en un certain nombre de strates, comme nous l'avons vu précédemment. Soit N, la taille de la population totale et N_h, la taille de la population de la h-ième strate. Soit P_h, la proportion de la h-ième strate représentant les membres de D et P, la proportion globale de la population que constitue D. On peut se servir du recensement décennal le plus récent pour estimer P_h et P, ou recourir à un autre sondage important plus récent donnant les codes des îlots et/ou des GI pour chaque ménage ou particulier échantillonné, de sorte que le couplage avec le dernier recensement décennal permettra d'affecter le ménage ou le particulier à une strate.

Nous supposerons que c est constant d'une strate à l'autre, même si ce n'est pas toujours le cas. Ainsi, procéder à des interviews dans les îlots où l'on trouve une grande concentration d'Amérindiens, d'Esquimaux ou d'Aléoutes signifie presque toujours se rendre dans des endroits éloignés où le transport s'avère difficile. Toutefois, même l'estimation d'une moyenne nationale pour c soulève des difficultés dans la plupart des enquêtes. En général, on ne pourra donc pas obtenir d'estimations par strates.

On suppose aussi que la distribution de Y dans D est constante d'une strate à l'autre. Plus précisément, on présume que

$$E(Y \mid D \text{ et } h) \equiv E(Y \mid D) \quad \text{et que}$$

$$\text{Var}(Y \mid D \text{ et } h) \equiv \text{Var}(Y \mid D),$$

où la valeur et la variance attendues se rapportent à la population, pas au plan d'échantillonnage. Cette hypothèse laisse habituellement à désirer, mais puisque les éléments du vecteur des caractéristiques auxquelles on s'intéresse se comportent souvent différemment d'une strate à l'autre, il est inutile d'essayer d'être plus précis. Enfin, on suppose que les fractions d'échantillonnage sont assez petites dans toutes les strates pour qu'on néglige les facteurs de correction de la population finie.

Des hypothèses qui précèdent, il ressort que la fraction d'échantillonnage optimale de la h-ième strate de l'enquête dans laquelle on délaisse tous les éléments U-D sélectionnés est

$$f_h = k\sqrt{\frac{P_h}{P_h(c-1)+1}}, \qquad (1)$$

où k représente une constante déterminée par les exigences de précision ou les contraintes budgétaires. (Voir une des sources précitées pour une preuve de (1). Cette règle est une application de la répartition de Neyman.) Si $c = 1$ (c'est-à-dire si la présélection coûte aussi cher que l'interview), on peut simplifier l'équation qui précède en $f_h \propto \sqrt{P_h}$, ce qui peut déboucher sur une répartition fort différente d'un échantillon équiprobable pour toutes les strates. Si la présélection est beaucoup moins onéreuse que l'interview (c'est-à-dire si $c >> 1$) et que D n'est pas extrêmement rare (à savoir P_h n'approche pas de zéro), cette relation donne des fraction d'échantillonnage assez uniformes ce qui se traduit par une répartition proportionnelle à la population.

Étant donné un budget fixe B, on calcule k grâce à l'équation

$$B = \sum_h N_h f_h c' \left[P_h c + (1 - P_h)\right]. \qquad (2)$$

Pour obtenir un échantillon aléatoire simple n du domaine D, il suffit de sélectionner un échantillon de taille n/P, qui coûtera en tout

$$B = ncc' + \left(\frac{n}{P} - n\right)c'. \qquad (3)$$

En établissant l'égalité des coûts, on peut calculer la constante de proportionnalité de (1), qui devient:

$$k = \frac{n\left(c - 1 + \frac{1}{P}\right)}{\sum_h N_h P_h \sqrt{c - 1 + \frac{1}{P_h}}}. \qquad (4)$$

Pour calculer de façon réaliste les avantages d'une telle répartition, on doit admettre que les estimations de P_h utilisées pour établir la répartition seront quelque peu désuètes au moment où se déroulera l'enquête. Soit A_h, la proportion réelle de D dans la h-ième strate lors de l'échantillonnage et de la collecte des données. On suppose que P n'a pas changé, même si la distribution entre les strates varie selon A_h. En établissant $NP = N_D$ et $N_D A_h = N_{dh}$, on se rend compte que la taille réelle de l'échantillon, n_D, pour D correspond à

$$n_D = \sum_h NPA_h f_h. \qquad (5)$$

Selon Kish (1965), la variance de cet échantillon sera plus élevée que celle d'un échantillon aléatoire simple de même taille, pour D. Le facteur d'extension de la variance ou l'effet du plan d'échantillonnage associé aux différents taux de sondage entre les strates est donné par l'équation bien connue

$$deff = \left(\sum_h A_h f_h\right)\left(\sum_h A_h / f_h\right). \qquad (6)$$

Par conséquent, la taille *utile* de l'échantillon associé au suréchantillonnage géographique sera

$$\frac{n_D}{deff} = \frac{NP}{\left(\sum_h A_h / f_h\right)}. \quad (7)$$

Si on remplace les formules (1) et (4) dans (7), on obtient

$$\frac{n_D}{deff} = \frac{n\left(c - 1 + \frac{1}{P}\right)}{\left(\sum_h A_h \sqrt{c - 1 + \frac{1}{P_h}}\right)\left(\sum_h \frac{N_h P_h}{NP} \sqrt{c - 1 + \frac{1}{P_h}}\right)}. \quad (8)$$

Cette dernière formule nous permet de comparer la variance d'une statistique arbitraire sur le domaine D après suréchantillonnage géographique à la variance de la même statistique résultant d'un échantillonnage aléatoire simple pour le même coût total B. On peut récrire la formule (8) de façon algébrique afin d'établir la part de la variance de l'échantillonnage aléatoire simple supprimée par le suréchantillonnage géographique

$$\frac{\frac{\sigma^2}{n} - \frac{\sigma^2 deff}{n_D}}{\frac{\sigma^2}{n}} =$$

$$1 - \frac{\left(\sum_h A_h \sqrt{c - 1 + \frac{1}{P_h}}\right)\left(\sum_h \frac{N_h P_h}{NP} \sqrt{c - 1 + \frac{1}{P_h}}\right)}{\left(c - 1 + \frac{1}{P}\right)}. \quad (9)$$

Il peut sans aucun doute arriver que la réduction soit négative, bref que l'échantillonnage aléatoire simple ait une variance inférieure pour le même coût. Pareille situation survient le plus souvent avec une strate où $NPA_h >> N_h P_h$, bref une strate qu'on croyait contenir une très petite partie de D mais qui en inclut en réalité une forte proportion. Soulignons que si $P_h \equiv P$, il ne faut s'attendre à aucune réduction de la variance consécutivement au suréchantillonnage géographique. De plus, à mesure que c s'approche de l'infini pour une valeur P fixe (ce qui correspond à une présélection de moins en moins chère par rapport à l'interview intégrale), la réduction de la variance approche la valeur nulle. Étant donné les complications supplémentaires qu'entraîne un échantillon stratifié, on en déduit que pour une valeur c élevée et une valeur P moyenne, la personne qui conçoit le plan d'échantillonnage devrait envisager un échantillonnage aléatoire simple au lieu de la stratification. La valeur du suréchantillonnage augmente à mesure que P s'approche de zéro, que c approche de 1, et que D se concentre dans une strate. Lorsque la sous-population D à laquelle on s'intéresse se retrouve dans une seule strate, l'efficacité de l'échantillon s'accroît car il y a de moins en moins de membres de D dans les autres strates, où la différence est plus importante. La partie qui suit évalue de façon empirique la réduction potentielle de la variance attribuable au suréchantillonnage géographique dans diverses conditions, pour plusieurs domaines démographiques.

5. ÉVALUATION EMPIRIQUE

Il est très difficile d'évaluer l'équation (9) pour les domaines auxquels on s'intéresse. On peut extraire des données sur P_h des bandes sommaires du recensement décennal diffusées par le Bureau of the Census au niveau de l'îlot, du groupe d'îlots et du district de dénombrement. Il est donc possible de définir des strates raisonnables et d'évaluer les équations (1) à (4). En supposant que P_h est statique dans le temps, on pourrait aussi évaluer les autres équations. Néanmoins, les Américains ont tendance à déménager souvent, ce qui modifie la composition raciale et ethnique de nombreux îlots (Judkins, Massey et Waksberg 1992). Quand les membres de D déménagent dans une région où ils étaient peu nombreux avant, les avantages du suréchantillonnage géographique s'amenuisent. Puisqu'on ne souhaite pas surestimer les avantages de la méthode, nous en avons cherché une autre afin d'obtenir une estimation raisonnable de A_h à certains points dans le temps après le recensement. Le couplage des données d'îlots ou de GI entre deux recensements consécutifs pourrait constituer une solution, mais il est irréalisable. Jusqu'à présent, les îlots ont été définis et étiquetés indépendamment d'un recensement à l'autre, et on n'a pas essayé de conserver les définitions en prévision d'une analyse longitudinale. C'est pourquoi il faut recourir à d'autres sources pour estimer A_h.

Les microdonnées issues des enquêtes-ménages que poursuit présentement le Bureau of the Census s'avèrent une bonne source d'information sur A_h, pour analyser les avantages de l'échantillonnage géographique à l'égard des populations de race noire et d'origine hispanique. Plus précisément, nous nous sommes servis des données du 1988 National Health Interview Survey (NHIS) effectuée par interview. Des employés du Bureau of the Census nous ont préparé une bande spéciale donnant le code du groupe d'îlots ou du district de dénombrement de 1980 pour tous les ménages ou presque interviewés lors de l'enquête de 1988, pour les habitations construites avant 1980. (Celles construites dans les années 80 auraient été échantillonnées à partir des permis de construction plutôt que par échantillonnage aérolaire. Pour ces cas, les étiquettes des îlots et les groupes d'îlots ne sont pas attribuées aux logements échantillonnés à cause de difficultés techniques.) Les données de l'ENS de 1988 ont ensuite été appariées aux fichiers sommaires du recensement de 1980 par groupe d'îlots ou district de dénombrement, ce qui a permis la répartition des ménages de l'ENS en strates définies selon la concentration de personnes de race noire et d'origine hispanique en 1980. Des facteurs de pondération ont ensuite servis à estimer la distribution de différentes sous-populations dans les strates. (On a supposé que les habitations construites dans les années 80 se trouvaient dans la strate où les sous-populations rares comptaient le moins de représentants.) Des opérations analogues auraient pu être effectuées pour les natifs d'Asie ou

du Pacifique, les Amérindiens, les Esquimaux, les Aléoutes et les personnes à faible revenu, mais on ne l'a pas fait.

Les tableaux et les graphiques qu'on trouvera dans le reste de l'article présentent les données de diverses sources à divers moments dans le temps. Il convient de se rappeler que les données dont on s'est servi pour créer les strates ne sont pas nécessairement identiques à celles utilisées pour répartir l'échantillon. Les données qui ont permis d'évaluer l'échantillon peuvent venir d'un troisième point dans le temps ou d'une troisième source. Le présent article examine les combinaisons suivantes:

Étiquette	Source des données (stratification)	Source des données (répartition)	Source des données (évaluation)
80/80/80 GI	Recensement de 1980 (GI)	Recensement de 1980	Recensement de 1980
80/80/88 GI	Recensement de 1980 (GI)	Recensement de 1980	ENS 1988
80/88/88 GI	Recensement de 1980 (GI)	ENS 1988	ENS 1988
90/90/90 GI	Recensement de 1990 (GI)	Recensement de 1990	Recensement de 1990
90/90/90 îlot	Recensement de 1990 (îlot)	Recensement de 1990	Recensement de 1990

6. SURÉCHANTILLONNAGE DE LA POPULATION NOIRE

Le tableau 1 présente différents aspects de la ségrégation résidentielle de la population noire aux États-Unis, aspects importants à connaître lorsqu'on entreprend une enquête sur la population. Bien que la proportion de noirs habitant dans des groupes d'îlots à forte concentration de population noire (60% et plus) ait diminué entre 1980 et 1990, la ségrégation demeure très forte au sein de ce groupe. Les colonnes indiquant la population en 1988 revêtent une importance particulière, car elles illustrent la dynamique des données stratifiées dans le temps. En 1988, la proportion de noirs qui vivaient dans un groupe d'îlots où l'on comptait moins de 10% de noirs en 1980 avait doublé, de 9.7% à 20.5%. Ce phénomène a d'importantes répercussions sur l'efficacité du suréchantillonnage géographique, comme on le verra plus loin. Il est aussi intéressant de noter que la population totale des groupes d'îlots à forte concentration de personnes de race noire (c'est-à-dire où l'on recensait plus de 60% de noirs) en 1980 a diminué d'environ 2 millions d'habitants entre 1980 et 1988. La tendance résulte donc en partie de l'abandon de quelques vieux logements ou quartiers. La concentration est plus élevée au niveau de l'îlot qu'à celui du groupe d'îlots, en 1990, ainsi qu'on s'y attendait. (Les données sur les îlots ne sont plus disponibles pour les États-Unis à partir de 1980.) Quoique l'échantillonnage d'îlots coûte légèrement plus cher que l'échantillonnage de groupes d'îlots (à cause du plus grand nombre d'îlots et de la nécessité de prendre certaines dispositions quand l'îlot compte moins d'habitants que la taille idéale de la grappe), elle permet de mieux cibler le domaine auquel on s'intéresse.

La figure 1 résume les effets des données sur la densité de population qui apparaissent au tableau 1 à l'égard du suréchantillonnage de la population noire. On constate l'effet sensible de c sur l'efficacité du suréchantillonnage géographique. Quand c dépasse 20, la meilleure façon d'échantillonner la population noire consiste vraisemblablement à sélectionner un échantillon équiprobable.

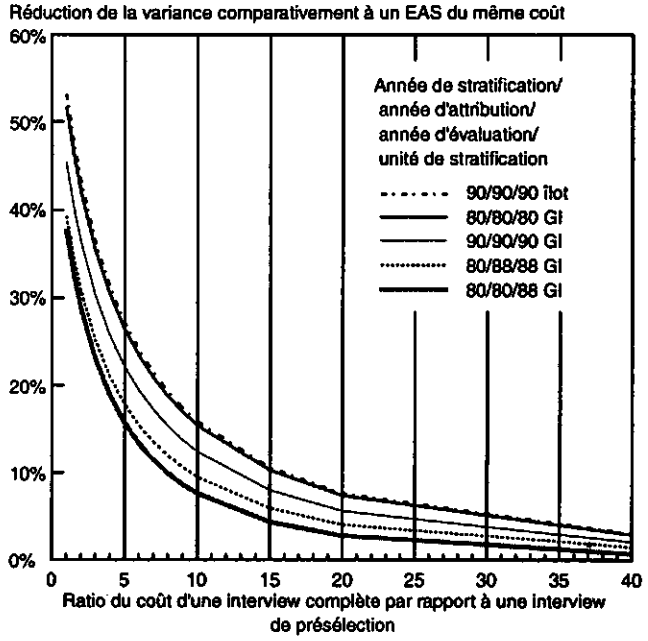


Figure 1. Réduction de la variance par suréchantillonnage géographique de la population noire

La figure montre aussi le risque qu'on court en ne se fiant qu'aux données stratifiées pour évaluer les avantages du suréchantillonnage géographique. La ligne 80/80/80 indique la réduction de la variance qu'on pourrait obtenir si la distribution de la population noire ne variait pas dans le temps selon les strates de densité définies au moyen des données sur les groupes d'îlots de 1980. La ligne 80/80/88 révèle la réduction réelle de la variance obtenue en 1988 grâce aux mêmes strates et à la même répartition. À $c = 5$, la variance diminue de 26% avec une distribution statique, alors qu'elle ne baisse que de 16% si la distribution change. Nous avons essayé de déterminer si la répartition de l'échantillon entre les anciennes strates, en fonction de nouvelles données sur la distribution réduirait la variance davantage en 1988. La réponse est oui, mais pas de beaucoup. La ligne 80/88/88 signale la réduction de la variance réalisable quand on applique la distribution de 1988 aux strates de 1980 et s'en sert pour une enquête en 1988. À $c = 5$, la nouvelle répartition réduit la variance de 18%, modeste amélioration par rapport à la diminution de 16% que permet la répartition selon l'ancienne distribution. Nous en avons conclu que la principale difficulté résidait dans l'ancienne stratification. En 1988, un si grand nombre de noirs avaient quitté les groupes

Tableau 1
Ségrégation résidentielle de la population noire

Strate de densité (pourcentage de Noirs dans l'unité de stratification l'année de la stratification)	Pourcentage de Noirs vivant dans la strate l'année indiquée				Pourcentage de la population totale vivant dans la strate l'année indiquée			
Année de mesure	1980	1988	1990	1990	1980	1988	1990	1990
Année de stratification	1980	1980	1990	1990	1980	1980	1990	1990
Unité de stratification	GI/DD	GI/DD	GI	Îlot	GI/DD	GI/DD	GI	Îlot
< 10%	9.7	20.5	12.0	8.5	78.2	81.4	75.7	77.5
10-30%	13.5	13.2	16.8	13.9	8.9	7.1	11.4	9.6
30-60%	18.9	20.4	20.3	16.2	5.1	5.1	5.7	4.5
60-100%	57.9	45.9	51.0	61.4	7.8	6.4	7.2	8.4
Population totale (milliers)	26,495	29,380	29.986	29,986	226,546	240,876	248,710	248,710
Pourcentage de Noirs aux États-Unis à l'année de la mesure	11.7	12.0	12.1	12.1				

Sources: Recensement décennal de 1980 (totalisation Westat)
Enquête nationale sur la santé de 1988 (totalisation Westat)
Recensement décennal de 1990 (totalisation Westat)

d'îlots où existait une forte concentration de population noire en 1980 pour d'autres groupes d'îlots que la réduction de la variance réalisable par suréchantillonnage avait presque diminué de moitié. La migration des noirs vers des groupes d'îlots où les gens de cette race étaient moins nombreux en 1980 augmente le nombre de noirs de l'échantillon auxquels est associé un facteur de pondération important, ce qui accroît la variabilité des poids, donc la variance. Quoi qu'il en soit, la réduction de la variance qu'indique la ligne 80/80/88 pour les valeurs $c < 10$ suffit indubitablement pour être jugée utile.

Lorsqu'on examine les données de 1990 à la figure 1, on constate que la ligne 90/90/90 des GI reste toujours quelques points sous la ligne 80/80/80, signe que le suréchantillonnage géographique des groupes d'îlots est un peu moins utile dans les années 90 que dans les années 80. On le doit à la légère diminution de la ségrégation des Américains de race noire en 1990, comparativement à la situation qui prévalait en 1980, ainsi qu'on l'a déjà souligné. D'un autre côté, la ligne 90/90/90 pour les ilôts est presque identique à la ligne 80/80/80. Le suréchantillonnage géographique des îlots devrait donc donner d'aussi bons résultats dans les années 90 que le suréchantillonnage des groupes d'îlots dans les années 80. Bien qu'on n'ait pas encore recueilli de données sur la distribution de la population noire à la fin des années 90 entre les strates de densité de 1990, il est raisonnable de croire que la migration s'est poursuivie, bref que les gains signalés par les lignes de 1990 soient plus faibles (conformément à la tendance générale que révèle la ligne 80/80/88), quand on établit des prévisions sur les économies envisageables pour la fin des années 90 et les premières années qui suivront l'an 2000.

7. SURÉCHANTILLONNAGE DES PERSONNES D'ORIGINE HISPANIQUE

Le tableau 2 présente divers aspects de la ségrégation résidentielle des personnes d'ascendance hispanique aux États-Unis qu'il est important de connaître lorsqu'on élabore une enquête sur la population. Plusieurs points valent la peine

Tableau 2
Ségrégation résidentielle des personnes d'ascendance hispanique

Strate de densité (pourcentage de personnes d'ascendance hispanique dans l'unité de stratification l'année de la stratification)	Pourcentage de personnes d'ascendance hispanique vivant dans la strate l'année indiquée				Pourcentage de la population totale vivant dans la strate l'année indiquée			
Année de mesure	1980	1988	1990	1990	1980	1988	1990	1990
Année de stratification	1980	1980	1990	1990	1980	1980	1990	1990
Unité de stratification	GI/DD	GI/DD	GI	Îlot	GI/DD	GI/DD	GI	Îlot
< 5%	14.8	29.3	10.6	6.6	76.8	79.8	68.4	68.9
5-10%	9.6	9.5	8.7	8.1	8.8	7.7	10.9	10.3
10-30%	22.6	21.2	22.8	22.1	8.5	7.4	11.8	11.5
30-60%	23.1	18.8	24.1	23.3	3.5	3.0	5.1	4.9
60-100%	30.0	21.2	33.9	39.8	2.4	2.0	3.8	4.4
Population totale (milliers)	14,609	19,393	22,354	22,354	226,546	240,876	248,710	248,710
Pourcentage de personnes d'ascendance hispanique aux États-Unis l'année de mesure	6.4	8.1	9.0	9.0				

Sources: Recensement décennal de 1980 (totalisation Westat)
Enquête nationale sur la santé de 1988 (totalisation Westat)
Recensement décennal de 1990 (totalisation Westat)

d'être mentionnés. D'abord il semble qu'à l'inverse des noirs, la ségrégation ait légèrement augmenté pour les membres de ce groupe, entre 1980 et 1990. Néanmoins, il existe des tendances analogues pour les deux populations. Ainsi, en 1980, 30% de la population d'origine hispanique vivait dans des groupes d'îlots comprenant au moins 60% de membres du même groupe. En 1988, ces groupes d'îlots ne comprenaient plus qu'environ 21% de personnes d'ascendance hispanique. En revanche, la proportion de personnes d'origine hispanique habitant un groupe d'îlots qui comptait moins de 5% de personnes de même ascendance en 1980 est passée de 15% à 29% entre 1980 et 1988. Cette évolution traduit un déplacement des personnes d'origine hispanique entre les régions, et un accroissement de la population hispanique aux États-Unis. La restratification de la population hispanique à partir des données de 1990 fait ressortir des tendances similaires à celles révélées par la distribution de 1980.

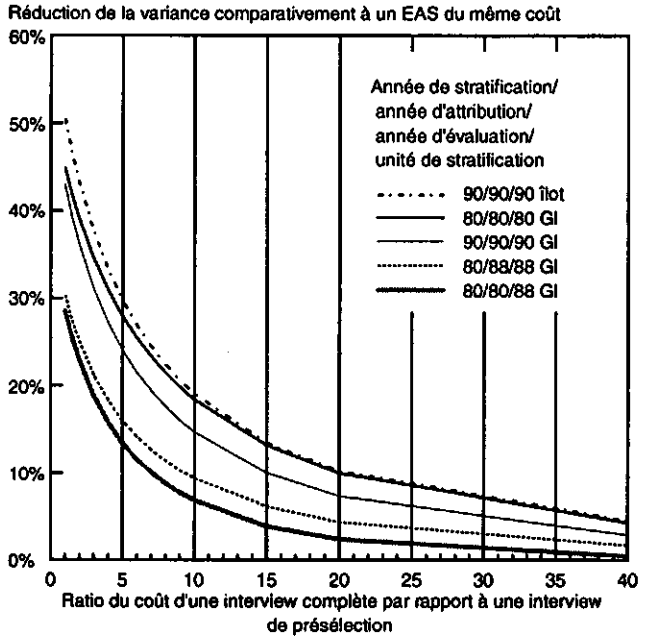


Figure 2. Réduction de la variance par suréchantillonnage géographique des personnes d'ascendance hispanique

La figure 2 donne un aperçu des répercussions de cette ségrégation sur les plans de suréchantillonnage. Les courbes suivent la même tendance générale que celles de la population noire et le suréchantillonnage géographique paraît utile pour les valeurs $c < 10$. Encore une fois cependant, il convient de se rappeler les effets de la migration sur la réduction de la variance. L'écart entre les lignes 80/80/80 et 80/80/88 est plus grand pour la population hispanique que pour la population noire, surtout aux valeurs $c < 5$. Pour l'instant, on ne dispose pas d'une base assez solide pour dire que la situation des années 80 se répétera dans les années 90.

8. SURÉCHANTILLONNAGE DES AUTRES MINORITÉS RACIALES

Les tableaux 3 et 4 présentent les données sur la ségrégation des natifs d'Asie et du Pacifique, d'une part, et des Amérindiens, des Esquimaux et des Aléoutes, d'autre part. Les figures 3 et 4 décrivent les implications correspondantes sur le suréchantillonnage de ces minorités. Les données de 1980 et de 1988 n'ont pas été reportées au tableau pour cette analyse, car les données de 1990 ne laissent pas entrevoir un suréchantillonnage bon marché des populations concernées, même avec une stratification selon la densité de population. La réduction en pourcentage de la variance est fort importante; elle dépasse celle obtenue avec les populations noire et hispanique, car la présélection nécessaire est beaucoup plus importante. Toutefois, le fait qu'il s'agisse de très petites populations aux États-Unis signifie qu'on aura toujours besoin d'une vaste présélection pour disposer d'un échantillon de taille suffisante au moment de l'interview. Ainsi, avec un ratio de coût égal à trois, il faudrait sélectionner 61,000 personnes (soit environ 24,000 ménages) pour obtenir un échantillon d'Amérindiens, d'Esquimaux et d'Aléoutes ayant la même précision qu'un échantillon aléatoire simple (théorique) de 1,000 personnes, de cette population. (Bien sûr, pour présélectionner 24,000 ménages, on devrait choisir un nombre supérieur d'unités de logement, de manière à tenir compte des logements vacants et des non-réponses). Les chiffres correspondant pour les natifs d'Asie et du Pacifique sont de 18,000 personnes, ou environ 7,000 ménages.

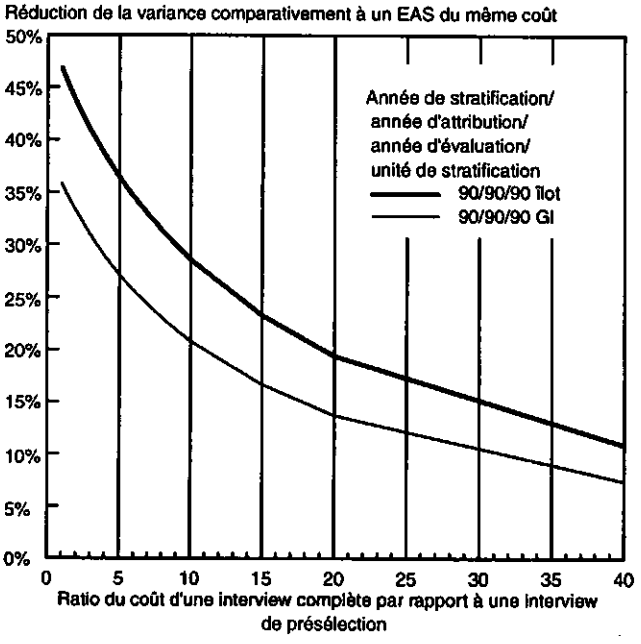


Figure 3. Réduction de la variance par suréchantillonnage géographique des natifs d'Asie et du Pacifique

Tableau 3
Ségrégation résidentielle des natifs d'Asie et du Pacifique

Strate de densité (pourcentage de natifs d'Asie et du Pacifique dans l'îlot ou le groupe d'îlots en 1990)	Pourcentage de natifs d'Asie et du Pacifique vivant dans la strate en 1990		Pourcentage de la population totale vivant dans la strate en 1990	
Unité de stratification:	GI	Îlot	GI	Îlot
< 5%	30.5	19.4	86.4	85.2
5-10%	17.2	17.7	7.2	7.4
10-30%	27.8	32.1	5.0	5.7
30-60%	14.6	18.0	1.0	1.3
60-100%	9.8	13.0	0.4	0.5
Population totale (milliers)	6,968	6,968	248,710	248,710
Pourcentage de natifs d'Asie et du Pacifique aux États-Unis l'année de mesure	2.8	2.8		

Source: Recensement décennal de 1990 (totalisation Westat)

Tableau 4
Ségrégation résidentielle des Amérindiens, des Esquimaux et des Aléoutes

Strate de densité (pourcentage d'Amérindiens, d'Esquimaux et d'Aléoutes dans l'unité de stratification en 1990)	Pourcentage d'Amérindiens, d'Esquimaux et d'Aléoutes vivant dans la strate en 1990		Pourcentage de la population totale vivant dans la strate en 1990	
Unité stratification:	GI	Îlot	GI	Îlot
< 5%	50.3	34.6	98.3	97.4
5-10%	7.4	12.1	0.8	1.4
10-30%	12.4	15.9	0.6	0.8
30-60%	6.0	7.7	0.1	0.1
60-100%	23.8	29.6	0.2	0.2
Population totale (milliers)	1,793	1,793	248,710	248,710
Pourcentage d'Amérindiens, d'Esquimaux et d'Aléoutes aux États-Unis l'année de mesure	0.7	0.7		

Source: Recensement décennal de 1990 (totalisation Westat)

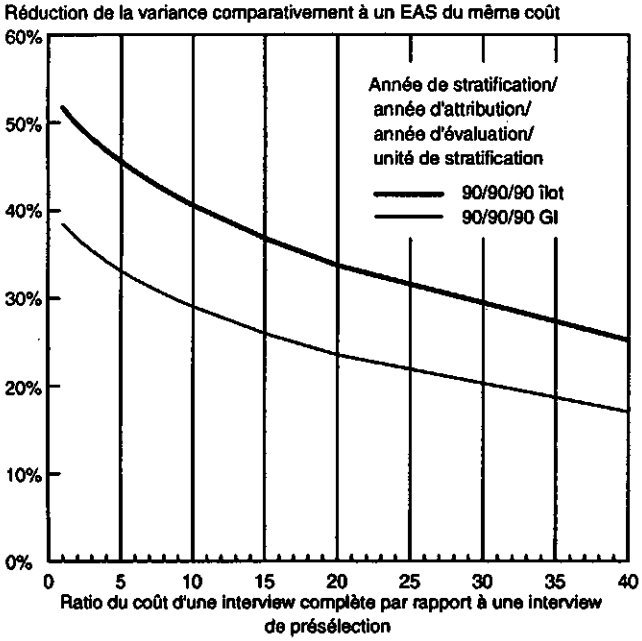


Figure 4. Réduction de la variance par suréchantillonnage géographique des Amérindiens, des Esquimaux et des Aléoutes

9. SURÉCHANTILLONNAGE DES DÉMUNIS

Le tableau 5 montre la distribution de la population défavorisée par groupe d'îlots, classée selon la population de personnes à faible revenu dans les GI, en 1990. Le nombre de GI dans chaque classe varie avec la définition de «faible revenu». Les chiffres qui apparaissent dans le tableau correspondent au pourcentage de personnes défavorisées dans chaque classe. Le tableau 5 illustre une distribution relativement uniforme de la pauvreté entre les différentes classes, pour les trois définitions retenues en 1990. Les données (non indiquées) du recensement de 1970 et du Current Population Survey révèlent que la ségrégation des personnes sous le seuil de pauvreté s'est accrue entre 1970 et 1990 (Waksberg 1995), mais elle reste beaucoup moins importante que la ségrégation des groupes ethniques et raciaux. La concentration est légèrement plus élevée pour le groupe de personnes sous la marque de 150% que pour les groupes nés des deux autres définitions mais, même pour ce groupe, elle demeure considérablement inférieure à celle observée pour les groupes raciaux et ethniques. Comme on peut le constater, avec cette définition, 25% seulement environ des démunis vivent dans des GI comptant 50% ou plus de démunis. Les pourcentages correspondants sont de 19% pour les personnes sous la marque de 125% de pauvreté et de 13% seulement pour celles sous la marque de 100% de pauvreté. Pareille distribution

signifie que le suréchantillonnage des ménages de la strate où on note un pourcentage relativement élevé de personnes à faible revenu ne donnera pas de résultats vraiment supérieurs au suréchantillonnage et à la présélection de la base de sondage complète, à moins que le coût de l'interview intégrale ne dépasse que légèrement celui de l'interview de présélection.

La figure 5 indique le ratio de la variance entre l'échantillon optimal et un échantillon aléatoire simple du même coût, pour les statistiques se rapportant aux défavorisés. Fait intéressant, en dépit de la plus forte concentration qu'entraîne la définition la plus large de «faible revenu», la réduction de la variance attribuable au suréchantillonnage géographique est plus importante avec la définition la plus étroite, car celle-ci exige une présélection accrue, si bien qu'on a plus à gagner d'une stratégie d'échantillonnage qui atténue la présélection. Dans les trois cas, le suréchantillonnage semble déboucher sur des avantages modérés quand c est inférieur à 3 ou 4, soit une réduction de 10% ou de 15% de la variance. Lorsque c atteint 10, les gains sont minimes et le suréchantillonnage des GI n'offre virtuellement aucun avantage avec une concentration élevée de démunis, pour les valeurs de c égales ou supérieures à 20. Bien sûr, il faut aussi tenir compte de la migration, mais nous n'avons pu obtenir les données voulues. La migration devrait se traduire par une réduction de la variance presque certainement plus faible que celle indiquée au tableau. En outre, les données sur le revenu du recensement de 1990 reposent sur un échantillon d'un sixième. La taille de l'échantillon dans un groupe d'îlots typique est légèrement inférieure à 100 ménages. La classification des îlots d'après la proportion de personnes à faible revenu entraîne donc un degré de confusion appréciable et beaucoup de groupes d'îlots ne se retrouveront pas dans la catégorie que leur attribuent les données du recensement, mais dans une classe voisine, ce qui atténuera encore plus la réduction de la variance que l'on peut espérer avec le suréchantillonnage géographique. Pour ces raisons, il est peu probable que le suréchantillonnage géographique débouche sur des gains d'efficacité. En réalité, la variance pourrait augmenter vers le milieu de la décennie ou plus tard. Une étude connexe, inédite, effectuée par Waksberg en 1989 aboutit à des résultats similaires quand on envisage de fusionner les données sommaires sur le revenu selon le code zip aux banques de numéros de téléphone utilisées lors de l'échantillonnage par composition aléatoire. Les gains réalisables grâce à la stratification semblent fort limités.

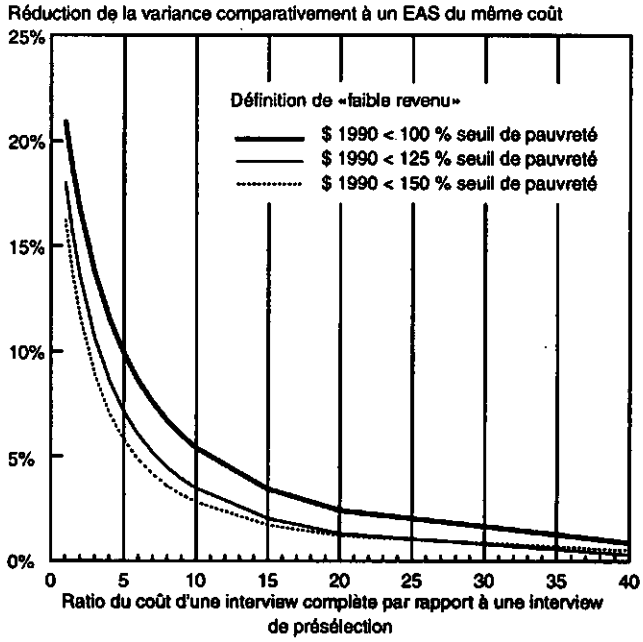


Figure 5. Réduction de la variance par suréchantillonnage géographique des démunis

Tableau 5
Ségrégation résidentielle des démunis

Strate de densité (pourcentage de démunis dans le groupe d'îlots en 1990 selon différentes définitions de «faible revenu»)	Pourcentage des démunis vivant dans la strate en 1990			Pourcentage de la population totale vivant dans la strate en 1990		
Définition de «faible revenu»:	$ < pauvreté	$ < 125% pauvreté	$ < 150% pauvreté	$ < pauvreté	$ < 125% pauvreté	$ < 150% pauvreté
< 5%	5.8	3.2	1.8	33.3	22.4	15.4
5-10%	12.3	8.3	5.7	22.1	19.7	16.7
10-20%	24.8	21 .0	16.8	22.8	25.2	24.8
20-30%	19.8	20.2	19.2	10.7	14.4	16.8
30-40%	14.3	15.9	17.0	5.4	8.1	10.7
40-50%	10.0	12.2	13.7	2.9	4.8	6.7
50-100%	13.0	19.3	25.7	2.8	5.4	8.8
Population totale (milliers)	31,797	42,316	52,521	248,710	248,710	248,710
Pourcentage de démunis aux États-Unis l'année de mesure	12.8	17.0	21.1			

Source: Recensement décennal de 1990 (totalisation Westat de STF-3)

Tableau 6
Ségrégation résidentielle des démunis selon la race et l'ascendance

Strate de densité (taux de pauvreté des personnes de la race ou de l'ascendance indiquée dans le groupe d'îlots en 1990)	Pourcentage de personnes de la race ou de l'ascendance indiquée dont le revenu était inférieur au seuil de pauvreté et qui vivaient dans la strate en 1990			Strate de densité (pourcentage de la minorité indiquée dans l'îlot en 1990)	Pourcentage de personnes de la race ou de l'ascendance indiquée dont le revenu était inférieur au seuil de pauvreté et qui vivaient dans la strate en 1990		
	Domaine				Domaine		
	Noirs	Hispaniques	Autres		Noirs	Hispaniques	Autres
< 5%	0.6	0.6	10.4	< 5%	4.0	4.6	n/a
5-10%	2.2	2.4	19.6	5-10%	3.7	5.1	n/a
10-20%	8.8	11.0	32.6	10-30%	13.2	19.9	n/a
20-30%	13.8	17.0	18.1	30-60%	19.0	25.5	n/a
30-40%	17.0	19.3	9.0	60-100%	60.0	44.8	n/a
40-50%	17.3	17.7	4.6				
50-100%	40.4	32.0	5.6				
Population totale (milliers)	8,557	5,536	17,975	Population totale (milliers)	8,557	5,536	17,975

Source: Recensement décennal de 1990 (totalisation Westat de STF-3)

L'analyse de tableaux plus détaillés (non présentés) révèle que l'efficacité est à peu près la même pour différentes ventilations géographiques, c'est-à-dire par État, par petite ou grande MSA, par noyau urbain, par banlieue ou par région non métropolitaine. Les conclusions tirées de cette analyse s'appliqueront donc approximativement aux enquêtes infranationales.

Le suréchantillonnage géographique demeure toutefois un instrument d'une très grande efficacité pour les populations noires et hispaniques à faible revenu. Comme on peut le constater au tableau 6, les personnes de race noire et celles d'ascendance hispanique qui vivent dans la pauvreté se regroupent beaucoup, à l'inverse des démunis des autres races. Le côté gauche du tableau 6 donne la distribution des défavorisés de race noire, d'origine hispanique et des autres groupes entre les strates de densité définies selon l'indice de pauvreté par rapport à la sous-population sur laquelle on se penche. En prenant un exemple du côté gauche, on se rend compte qu'en 1990, 32% des défavorisés d'origine hispanique vivaient dans un groupe d'îlots où l'indice de pauvreté pour les membres de ce groupe dépassait 50%. Le côté droit présente la distribution des défavorisés de race noire et d'ascendance hispanique pour les strates de densité définies d'après la concentration locale de noirs ou de personnes d'origine hispanique, sans égard au revenu. En prenant un exemple du côté droit, on note qu'en 1990, 44.8% des démunis d'origine hispanique habitaient un groupe d'îlots où les gens de la même ascendance représentaient plus de 60% de la population. Ces chiffres laissent entendre que plus de 90% des défavorisés de race noire et d'origine hispanique vivaient à un endroit où la concentration du groupe racial ou ethnique pertinent dépassait la moyenne. Bref, une stratégie prévoyant le suréchantillonnage des îlots à forte population de personnes de race noire ou d'ascendance hispanique débouchera automatiquement sur un nombre disproportionné de défavorisés. Par ailleurs, on ne retrouve presque aucun démuni de race noire ou d'origine hispanique dans les régions où ces deux groupes indiquent un faible indice de pauvreté. Il s'agit d'un contraste marqué avec les tendances relevées pour les défavorisés qui ne sont ni de race noire, ni d'ascendance hispanique. Apparemment, beaucoup de défavorisés de race blanche mais d'origine non hispanique vivent en voisinage avec d'autres personnes de race blanche mais nanties, peut-être parce que pour elles, la pauvreté a tendance à être un phénomène transitoire ou parce qu'il s'agit de personnes à la retraite qui avaient acheté leur maison à un moment où elles se trouvaient dans une meilleure situation.

10. SURÉCHANTILLONNAGE SIMULTANÉ DE PLUSIEURS DOMAINES RACIAUX-ETHNIQUES

En général, le suréchantillonnage géographique s'avère aussi facile et aussi efficace pour plusieurs domaines raciaux-ethniques que pour un seul. De fait, les taux d'échantillonnage optimaux pour les strates où se concentre chaque sous-population à laquelle on s'intéresse sont à peu près les mêmes que si on se penchait sur un seul domaine. Le taux de présélection global augmente néanmoins, car le nombre de régions à taux d'échantillonnage élevé augmentera avec le nombre de domaines. Ces deux observations résultent du chevauchement restreint entre les régions où il y a forte ségrégation des minorités raciales et ethniques à l'étude.

Le tableau 7 fournit quelques données sur la question, du recensement décennal de 1990. Les seules sous-populations pour lesquelles on remarque un chevauchement important dans les régions à forte concentration sont les personnes d'ascendance hispanique et les natifs d'Asie et du Pacifique. Toutefois, même dans ce cas, le chevauchement ne s'effectue que dans un sens. Comme on compte beaucoup plus de personnes d'origine hispanique que de ressortissants d'Asie et des îles du Pacifique aux États-Unis, la proportion de personnes d'ascendance hispanique habitant les îlots où les

Tableau 7
Mélange résidentiel des minorités

Strate de densité (pourcentage de la minorité indiquée dans l'îlot en 1990)	Pourcentage de Noirs vivant dans la strate en 1990			Pourcentage de personnes d'ascendance hispanique vivant dans la strate en 1990			Pourcentage de natifs d'Asie et du Pacifique vivant dans la strate en 1990			Pourcentage d'Amérindiens, d'Esquimaux et d'Aléoutes vivant dans la strate en 1990		
	Domaine de stratification			Domaine de stratification			Domaine de stratification			Domaine de stratification		
	Hispaniques	Natifs d'Asie et du Pacifique	Amérindiens, Esquimaux, et Aléoutes	Noirs	Natifs d'Asie et du Pacifique	Amérindiens, Esquimaux et Aléoutes	Noirs	Hispaniques	Amérindiens, Esquimaux et Aléoutes	Noirs	Hispaniques	Natifs d'Asie et du Pacifique
< 10%	79.2	95.4	99.6	73.4	86.3	99.1	78.9	59.2	99.6	85.9	81.4	95.1
10-30%	12.7	3.8	0.3	15.5	10.7	0.8	15.2	26.9	0.4	8.2	12.3	3.9
30-60%	5.8	0.7	0.0	7.4	2.5	0.1	4.2	10.8	0.0	3.3	4.5	0.8
60-100%	2.2	0.1	0.0	3.6	0.5	0.1	1.6	3.2	0.0	2.5	1.8	0.2

Sources: Recensement décennal de 1990 (totalisation Westat)

natifs d'Asie et du Pacifique représentent plus de 10% de la population n'est que de 13.7%, alors que le pourcentage de natifs d'Asie et du Pacifique dans les îlots où la population de personnes d'origine hispanique dépasse 10% de la population s'élève à 40.8%. Sur le plan pratique, ce chevauchement présente sans doute peu d'importance, on aurait besoin d'un échantillon de présélection très élevé (dans les régions très concentrées et à l'extérieur de celles-ci) pour trouver un nombre suffisant de natifs d'Asie et du Pacifique en vue de respecter les besoins de précision modérés, pareil échantillon ferait ressortir un nombre suffisant de personnes d'origine hispanique sans qu'on doive recourir à une répartition disproportionnée de l'échantillon entre les îlots où se concentrent davantage les personnes d'ascendance hispanique.

11. CONCLUSIONS

Le suréchantillonnage géographique reposant sur les données issues du recensement décennal le plus récent est une stratégie utile lorsqu'on veut accroître la précision des statistiques sur la population de race noire et d'origine hispanique dans les enquêtes-ménages entreprises aux États-Unis, pourvu que le coût de l'interview complète soit de 5 à 10 fois moindre que celui de l'interview de présélection. La même stratégie donne aussi de bons résultats lorsqu'il s'agit d'améliorer la précision des statistiques sur les natifs d'Asie et du Pacifique ou sur les Amérindiens, les Esquimaux et les Aléoutes, même lorsque le rapport entre le coût de l'interview intégrale et le coût de l'interview de présélection est très élevé.

Quoi qu'il en soit, ces constatations ne signifient pas qu'on puisse réaliser une enquête à un coût raisonnable en vue d'obtenir simultanément des statistiques très précises sur toutes les sous-populations désirées et de garder le degré de précision souhaité pour la population globale. La majorité des enquêtes démographiques réclament une précision raisonnable pour les domaines spécialisés et la population, en général. En déplaçant une portion de l'interview intégrale de la population blanche d'ascendance non hispanique vers les autres sous-populations, on devrait réduire la précision des statistiques sur la population totale. Il vaut habituellement la peine de parvenir à un compromis entre la précision des données sur les sous-populations et sur la population totale. Par cela, nous voulons simplement faire ressortir que le suréchantillonnage géographique ne met pas fin à la nécessité de sélectionner de très gros échantillons et d'entreprendre de nombreuses interviews de présélection quand on s'efforce de recueillir des statistiques précises sur des domaines rares, au coût le plus bas. Les statistiques précises sur des sous-populations rares continueront d'être dispendieuses qu'on recoure ou non au suréchantillonnage géographique.

En ce qui concerne les enquêtes sur les défavorisés, le suréchantillonnage géographique ne donne lieu qu'à de légers gains, et cela uniquement quand le coût de l'interview complète dépasse de plusieurs fois celui de l'interview de présélection et de l'abandon d'un ménage. Les gains devraient pour la plupart disparaître avec la détérioration due au passage du temps. En réalité, vers le milieu de la décennie ou un peu plus tard, une fois que les données du recensement seront sérieusement périmées, il se peut fort bien que le suréchantillonnage géographique entraîne une perte d'efficacité au lieu d'une amélioration, à cause de la migration des démunis et de l'erreur d'échantillonnage qui résulte de la quantification de la pauvreté au niveau des groupes d'îlots. Néanmoins, le suréchantillonnage géographique reste un instrument utile quand on s'intéresse surtout aux personnes défavorisées de race noire ou d'origine hispanique.

REMERCIEMENTS

Ce projet de recherche a été entrepris par Westat Inc. dans le cadre de l'entente 200-89-7021 parrainée par le National Center for Health Statistics et les Centers for Disease Control and Prevention. David Judkins et James Massey ont prêté leur concours au projet pendant la période où ils travaillaient respectivement à Westat et au NCHS. Les auteurs tiennent à remercier John Edmonds et Robert Dymowski, de Westat, pour leur participation à la programmation, ainsi que les arbitres pour leurs commentaires et observations sur une version antérieure de cet article.

BIBLIOGRAPHIE

JUDKINS, D., MASSEY, J., et WAKSBERG, J. (1992). Patterns of residential concentration by race and Hispanic origin. *Proceedings of the Social Statistics Section, American Statistical Association*, 51-60.

KALTON, G., et ANDERSON, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A*, 149, 1, 65-82.

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

MASSEY, J., JUDKINS, D., et WAKSBERG. J. (1993). Collecting health data on minority populations in a national survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 75-84.

UNITED NATIONS (1993). *Sampling Rare and Elusive Populations*. Department for Economic and Social Information and Policy Analysis, Statistical Division, National Household Survey Capability Programme. New York.

WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section, American Statistical Association*, 429-434.

WAKSBERG, J. (1995). Distribution of poverty in Census block groups (BG's) and implications for sample design. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 497-502.

WAKSBERG, J., et MOHADJER, L. (1991). Automation of within-household sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 350-355.

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 81-83
Statistique Canada

Nouvel estimateur de l'erreur-type pour les groupes aléatoires

WILLARD C. LOSINGER[1]

RÉSUMÉ

La méthode utilisée pour estimer l'erreur-type des données des recensements décennaux des États-Unis de 1970 à 1990 donne des résultats disparates. Ainsi, on obtient des erreurs-types différentes pour les réponses «oui» et «non» à la même variable binomiale, alors que les deux estimations devraient être identiques. Quand la plupart des personnes répondent d'une façon à une question binomiale et quelques autres donnent la réponse contraire, l'erreur-type est beaucoup plus élevée pour la réponse la plus fréquente. D'autre part, lorsque les personnes interrogées fournissent toutes la même réponse, l'erreur-type n'est pas égale à zéro et reste fort élevée. Signaler les effets moyens du plan d'échantillonnage pondérés selon le nombre de répondants qui présentent des caractéristiques particulières ne fait qu'aggraver le problème. L'auteur propose une solution de rechange à la méthode d'estimation de l'erreur-type des groupes aléatoires utilisée dans le cadre du recensement des États-Unis.

MOTS CLÉS: Recensement; estimation de la variance; groupes aléatoires; effet du plan d'échantillonnage.

1. INTRODUCTION

Lors du recensement décennal de 1990, tous les répondants ont été priés de fournir des renseignements sur certains éléments de donnée (baptisés «données intégrales»). La plupart des personnes ont inscrit l'information requise sur le questionnaire abrégé du recensement. Par ailleurs, dans le cadre d'un échantillonnage systématique (allant du huitième à la demie, mais se situant en moyenne au sixième), on a demandé aux répondants des précisions sur d'autres éléments de donnée (données d'échantillon), au moyen du questionnaire détaillé du recensement.

Au lieu d'estimer l'erreur-type de toutes les données échantillonnées diffusées, le Census Bureau a préparé des tableaux indiquant les effets généralisés du plan d'échantillonnage. L'utilisateur devait calculer l'erreur-type en supposant un échantillonnage aléatoire simple (soit avec la formule normalisée, soit à partir d'un tableau) et un taux d'échantillonnage d'une personne sur six pour toutes les données d'échantillon. Il suffisait ensuite de multiplier le résultat obtenu par l'effet généralisé du plan d'échantillonnage (présenté dans un autre tableau). Le tableau en question énumérait les effets généralisés du plan d'échantillonnage sur l'élément de donnée concerné et la proportion de personnes ou d'unités de logements dans l'échantillon (on trouvera au tableau 1 les effets du plan d'échantillonnage diffusés pour les données d'échantillon du Vermont en 1990). Ainsi, pour toutes les estimations se rapportant à la profession qui ont été rendues publiques, l'utilisateur a pu constater que le plan d'échantillonnage avait quatre effets généralisés, soit un pour chaque taux d'échantillonnage applicable aux personnes, dans le rapport. Pour connaître l'erreur-type relative au nombre d'enseignants dans un rapport, l'utilisateur devait multiplier l'erreur-type de l'échantillonnage aléatoire simple (en supposant un taux d'échantillonnage d'une personne sur six, selon la formule ou d'après le tableau des erreurs-types) par l'effet du plan d'échantillonnage sur les éléments de donnée se rapportant à la profession qui correspondait au taux d'échantillonnage indiqué. Le nombre estimatif d'enseignants et l'erreur-type permettaient ensuite d'établir un intervalle de confiance. On trouvera d'autres explications sur l'usage du tableau des effets du plan d'échantillonnage à la partie «Accuracy of the Data» des produits de données (U.S. Bureau of the Census 1993, par exemple).

2. ESTIMATION DE L'ERREUR-TYPE

On s'est servi de la méthode des groupes aléatoires pour estimer l'erreur-type des données échantillonnées lors du recensement. Les États-Unis ont été divisés en un peu plus de 60,000 zones distinctes baptisées «régions de pondération» pour lesquelles on a calculé un poids d'échantillonnage. Dans chaque région de pondération, les unités d'échantillonnage (unités de logement ou personnes vivant dans un logement collectif) ont été systématiquement réparties entre 25 groupes aléatoires. On pensait qu'ainsi chaque groupe aléatoire respecterait à peu près le même plan d'échantillonnage que l'échantillon original (Wolter 1985).

On a estimé séparément le total de chacun des 1,804 éléments de donnée d'échantillon pour chaque groupe aléatoire en multipliant la population pondérée de l'élément de donnée échantillonné du groupe par 25. L'erreur-type des groupes aléatoires a ensuite été déterminée pour chaque élément de donnée dont le nombre total de personnes présentant un caractéristique particulière avait été estimé à partir des données échantillonnées au moyen des 25 estimations différentes du total venant des groupes aléatoires:

$$S_{RG} = \sqrt{(1 - n/N)\sum_{i=1}^{25} \frac{(\hat{Y}_i - \hat{Y})^2}{24}}$$

[1] Willard C. Losinger, U.S. Department of Agriculture: APHIS:VS, CEAH, 555 South Howes Street, Suite 200, Fort Collins, CO 80521, U.S.A.

Tableau 1
Effects du plan d'échantillonnage du recensement américain de 1990 indiqués pour les données échantillonnées au Vermont

Caratéristique	Pourcentage de personnes ou d'unités de logement dans l'échantillon			
	< 15%	15 - 30%	30 - 45%	≥ 45%
Âge	1.2	1.0	0.6	0.5
Sexe	1.2	1.0	0.6	0.5
Race	1.2	1.0	0.6	0.5
Ascendance hispanique (peu importe la race)	1.2	1.0	0.6	0.5
État civil	1.1	0.9	0.6	0.5
Type de ménage et relation	1.2	1.0	0.6	0.5
Enfants	2.5	2.2	1.3	1.2
Limitation de mobilité et incapacité liée au travail	1.2	1.0	0.6	0.5
Origines	1.8	1.5	1.0	0.8
Lieu de naissance	1.9	1.6	1.0	0.9
Citoyenneté	1.7	1.4	1.0	0.8
Lieu de rédience en 1985	1.9	1.7	1.0	0.9
Année d'entrée	1.3	1.0	0.6	0.5
Langue parlée à domicile et aptitude à parler l'anglais	1.6	1.3	0.9	0.7
Scolarité	1.3	1.1	0.6	0.5
Inscription à l'école	1.6	1.4	1.0	0.8
Type de résidence (urbaine/rurale)	1.7	1.7	1.4	1.4
Type de ménage	1.2	1.0	0.6	0.5
Type de famille	1.1	1.0	0.6	0.5
Logement collectif	1.0	1.1	0.9	0.8
Type de sous-famille et présence d'enfants	1.1	0.9	0.5	0.5
Situation d'emploi	1.2	1.0	0.6	0.5
Secteur d'activité	1.2	1.0	0.6	0.5
Profession	1.2	1.0	0.6	0.5
Type de travailleur	1.2	1.0	0.6	0.5
Heures de travail par semaine et semaines de travail en 1989	1.4	1.2	0.7	0.6
Nombre de travailleurs	1.3	1.1	0.7	0.6
Lieu de travail	1.4	1.2	0.8	0.6
Moyen de transport au lieu de travail	1.4	1.2	0.7	0.6
Durée du déplacement	1.3	1.1	0.6	0.5
Usage d'un véhicule privé	1.4	1.2	0.7	0.6
Heure de départ pour le travail	1.2	1.0	0.6	0.5
Type de revenu en 1989	1.3	1.1	0.6	0.5
Revenu du ménage en 1989	1.1	1.0	0.6	0.5
Revenu de la famille en 1989	1.1	1.0	0.6	0.5
Indice de pauvreté en 1989	1.5	1.2	0.7	0.7
Indice de pauvreté en 1989	1.1	0.9	0.5	0.5
Forces armées et status d'ancien combattant	1.4	1.1	0.7	0.6

Source: U.S. Bureau of the Census (1993). 1990 Census of Population: Social and Economic Characteristics: Vermont. Rapport numéro 1990 CP-2-47. Page C-11.

où n correspond au nombre non pondéré de personnes de l'échantillon dans la région de pondération, N représente la population recensée dans la région de pondération, $\hat{Y}_i$ indique le total estimatif de l'élément de donnée obtenu en multipliant la population pondérée de l'élément en question du i-ième groupe aléatoire par 25 et $\hat{Y}$ représente le chiffre de population pondéré de l'élément de donnée (à savoir l'estimation de l'échantillon) pour la région de pondération.

L'erreur-type a été calculée selon la méthode d'échantillonnage aléatoire simple et un taux d'échantillonnage d'une personne sur six comme suit:

$$S_{SRS} = \sqrt{5\ \hat{Y}\ (1 - \hat{Y}/N)}.$$

Cette équation dérive des formules habituelles qu'on trouve dans Cochran (1977).

On a mesuré l'effet du plan d'échantillonnage sur chaque élément de donnée dans la région de pondération sous la forme du ratio entre S_{RG} et S_{SRS}:

$$F = \frac{S_{RG}}{S_{SRS}}.$$

Afin de produire un rapport sur les données échantillonnées pour l'État, on a établi les effets moyens du plan d'échantillonnage sur chaque élément de donnée pour l'ensemble des régions de pondération de l'État. On a ensuite calculé l'effet généralisé du plan d'échantillonnage sur chaque élément de donnée (par exemple, tous les éléments de donnée se rapportant à la profession). L'effet généralisé du plan d'échantillonnage a été pondéré en faveur des éléments de donnée pour lesquels la population estimative était plus élevée. Un document du Census Bureau (U.S. Bureau of the Census 1991) donne des précisions sur la plupart des méthodes suivies. On a recouru à la même approche fondamentale pour les produits de données des recensements de 1970 et de 1980.

3. EXEMPLE HYPOTHÉTIQUE DE GROUPES ALÉATOIRES

Le tableau 2 donne un exemple hypothétique de données qu'on aurait pu obtenir grâce à la méthode des groupes aléatoires. La population pondérée de Blancs et de Noirs est indiquée pour les 25 groupes aléatoires d'une région de pondération du Vermont. Dans cette région de pondération fictive, on ne compte aucune personne d'une autre race.

En supposant un échantillonnage simple aléatoire, l'erreur-type est la même pour les deux groupes (ainsi qu'on pourrait s'y attendre avec une variable binomiale). Toutefois, la valeur de S_{RG} est beaucoup plus élevée pour la population blanche que pour la population noire. En outre, l'effet du plan d'échantillonnage est près de cinq fois plus élevé pour les Blancs que pour les Noirs. Puisque l'effet généralisé du plan d'échantillonnage sur les groupes d'éléments de donnée a été pondéré en faveur des éléments de donnée pour lesquels la population estimative était plus élevée, l'effet généralisé du plan d'échantillonnage obtenu pour la race était très important au Vermont.

Les produits de données sur le recensement de 1990 des États-Unis présentaient fréquemment les données sur la race. Puisqu'on a demandé à chaque répondant d'indiquer sa race (il s'agissait d'un élément de donnée couvrant la totalité des personnes recensées) et puisque la méthode de pondération dont le Census Bureau s'est servi a contraint les estimations selon la race à correspondre au chiffre de population global du recensement selon la race, on aurait sans doute dû considéré que les erreurs-types de l'estimation de la race étaient égales à zéro. Toutefois, on a continué de diffuser les effets généralisés du plan d'échantillonnage selon la race, même si des constantes ont été arbitrairement établies dans tous les rapports (on ne les a pas calculés au moyen de la méthode indiquée).

Tableau 2
Exemple hypothétique de données qu'on aurait pu obtenir au moyen de la méthode des groupes aléatoires pour estimer l'erreur-type des données échantillonnées au recensement. Pour une région de pondération du Vermont, on a demandé aux répondants de préciser leur race. Quelques personnes (110) sont noires; la plupart (2,518) sont blanches. On suppose un taux d'échantillonnage d'une personne sur six ($N = 2{,}628$, $n = 438$).

Groupe aléatoire	Population pondéré de Noirs*	Population pondéré de Blancs*	Total du chiffre de la population pondérée #
1	10	90	100
2	0	100	100
3	0	110	110
4	0	140	140
5	5	70	75
6	8	50	58
7	12	103	115
8	20	60	80
9	0	65	65
10	0	100	100
11	0	125	125
12	0	130	130
13	10	90	100
14	0	100	100
15	0	110	110
16	0	140	140
17	5	70	75
18	8	52	60
19	12	103	115
20	20	160	180
21	0	65	65
22	0	100	100
23	0	125	125
24	0	130	130
25	0	130	130
Somme des chiffres pondérés ($\hat{Y}$)	110	2,518	2,628
S_{RG}	145.98	687.96	
S_{SRS}	22.96	22.96	
F	6.36	29.96	

* Les 25 permiers chiffres de la colonne correspondent à la valeur X_i du i-ième groupe aléatoire, selon la méthode modifiée des groupes aléatoires. En multipliant ce chiffre par 25, on obtient $\hat{Y}_i$ selon la méthode des groupes aléatoires utilisée par le U.S. Bureau of the Census.

\# Les 25 premiers chiffres de cette colonne représentent W_i, selon la méthode modifiée des groupes aléatoires.

4. NOUVELLE APPROCHE À LA MÉTHODE DES GROUPES ALÉATOIRES

On obtiendrait des résultats beaucoup plus satisfaisants dans l'estimation de l'erreur-type en modifiant légèrement la méthode des groupes aléatoires (essentiellement en recourant à une technique d'estimation par ratio). Au lieu d'utiliser $\hat{Y}_i$ tel que défini ci-dessus pour estimer le total du i-ième groupe aléatoire, on pourrait recourir à l'équation:

$$\hat{L}_i = NX_i/W_i$$

où X_i représente la population pondérée pour l'élément de donnée du i-ième groupe aléatoire, W_i correspond au chiffre pondéré de population du i-ième groupe aléatoire et N indique la population recensée dans la région de pondération. L'erreur-type estimative obtenue avec la méthode modifiée des groupes aléatoires serait donc

$$S_L = \sqrt{(1 - n/N)\sum_{i=1}^{25} \frac{(\hat{L}_i - \hat{Y})^2}{24}}.$$

Grâce à cette méthode, la valeur de S_L pour les Noirs et les Blancs dans la région de pondération hypothétique du tableau 1 est égale à 160.78 (ce qui s'approche de la valeur de S_{RG} pour les Noirs). Bref, l'erreur-type est la même pour les deux réponses à la question binomiale. D'autre part, si toutes les unités échantillonnées fournissent la même réponse, S_L est égal à zéro, alors que S_{RG} n'obtient la valeur nulle que lorsque chaque groupe aléatoire a la même population pondérée.

Cette nouvelle méthode d'estimation de l'erreur-type pourrait s'avérer utile aux chercheurs qui n'ont pas accès à l'un des nombreux logiciels dont on peut désormais se servir pour effectuer des estimations à partir des données d'échantillon (par exemple SUDAAN, STATA, PC-CARP, VPLX, *etc.*). Par ailleurs, le U.S. Bureau of the Census devrait envisager de modifier sa méthode d'estimation de l'erreur-type pour les données d'échantillon en prévision du recensement de l'an 2000. Enfin, compte tenu de l'intérêt qu'on porte présentement à la gestion de la qualité au Census Bureau, ce dernier pourrait vouloir sonder les utilisateurs de produits de données afin d'établir l'utilité de la présentation actuelle des erreurs-types (par le truchement des effets du plan d'échantillonnage) et demander à un certain nombre d'entre eux de participer à un exercice ayant pour but d'améliorer la manière dont les erreurs-types seront présentées au prochain recensement.

BIBLIOGRAPHIE

COCHRAN, W. G. (1977). *Sampling Techniques* (3ième édition). John Wiley & Sons: New York.

WOLTER, K. M. (1985). *Introduction to Variance Estimation.* Springer-Verlag: New York.

U.S. BUREAU OF THE CENSUS (1991). Computer Specifications for the 1990 Decennial Census Variance Estimation Operation. STSD Decennial Census Memorandum Series #Z-65.

U.S. BUREAU OF THE CENSUS (1993). Appendix C. Accuracy of the Data. Pp. C-1 à C-11 dans 1990 Census of Population: Social and Economic Characteristics: Vermont. Bureau of the Census Document 1990 CP-2-47.

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 85-86
Statistique Canada

Dérivation simple de l'estimateur de régression par linéarisation

KEES ZEELENBERG[1]

RÉSUMÉ

L'auteur explique comment recourir au calcul matriciel pour simplifier la dérivation de l'estimateur du coefficient de régression et de l'estimateur de régression par linéarisation.

MOTS CLÉS: Calcul matriciel; estimateur de régression; développement de Taylor.

1. INTRODUCTION

On calcule souvent la variance des statistiques non linéaires obtenues par échantillonnage au moyen d'une méthode d'approximation linéaire faisant appel à l'expansion de Taylor. La variance de l'estimateur général du coefficient de régression et de l'estimateur de régression en est un exemple. Habituellement, la linéarisation exige des dérivations complexes. Le présent article a pour but de montrer comment le calcul matriciel peut simplifier les dérivations de ce genre, au point où l'on pourrait dériver l'expansion de Taylor de l'estimateur du coefficient de régression en une ligne plutôt qu'en près d'une page ainsi que le proposent Särndal et ses collaborateurs (1992, p. 205-206). Il convient néanmoins de souligner que le calcul matriciel exige certains préparatifs, inutiles avec les méthodes classiques. Quoi qu'il en soit, ces préparatifs peuvent être considérés comme un placement puisqu'une fois connus, on peut en faire un usage fructueux dans de nombreuses applications. Après la rédaction de cet article, Binder (1996) en a publié un autre présentant des techniques analogues pour dériver les variances par linéarisation. Le présent article peut être considéré comme une note pédagogique expliquant comment utiliser les différentielles.

2. DIFFÉRENTIELLES MATRICIELLES

2.1 Introduction

Nous recourrons au calcul matriciel en faisant appel aux différentielles selon la description de Magnus et Neudecker (1988). Cette méthode de calcul s'écarte quelque peu des méthodes habituelles, qui reposent plus sur les dérivées que les différentielles. C'est pourquoi nous commencerons par examiner brièvement la définition et les propriétés des différentielles (lire Zeelenberg 1993, pour plus d'explications). Nous définirons d'abord les différentielles à l'égard des fonctions vectorielles avant de généraliser la définition aux fonctions matricielles.

2.2 Fonctions vectorielles

Soit f, une fonction de l'ensemble ouvert $S \subset \mathbb{R}^m$ à $\mathbb{R}^n$ et x_0, un point de S. La fonction f est *dérivable* à x_0 s'il existe une vraie matrice $n \times m$ A, dépendant de x_0, de sorte que pour chaque valeur $u \in \mathbb{R}^m$ pour laquelle $x_0 + u \in S$,

$$f(x_0 + u) = f(x_0) + A_{x_0} u + o(u), \qquad (1)$$

où $o(u)$ est une fonction selon laquelle $\lim_{|u| \to 0} |o(u)|/|u| = 0$. La matrice A s'appelle *dérivée première* de f à x_0 et est notée $Df(x_0)$ ou $\partial f/\partial(x')|_{x=x_0}$. La dérivée Df est égale à la matrice des dérivées partielles, c'est-à-dire $Df(x)_{ij} = \partial f_i/\partial x_j$. La fonction linéaire $df_{x_0}: \mathbb{R}^m \mapsto \mathbb{R}^n$ définie par $df_{x_0}: u \mapsto A_{x_0} u$ est la *dérivée* de f à x_0. On écrit dx plus souvent que u, de sorte que $df_{x_0}(dx) = A_{x_0} dx$. D'après (1), on se rend compte que la différentielle correspond à la partie linéaire de la fonction, qu'on peut aussi exprimer par

$$y - y_0 = A_{x_0}(x - x_0),$$

ou $y_0 = f(x_0)$. La différentielle d'une fonction est donc la linéarisation de cette fonction, c'est-à-dire l'équation de l'hyperplan qui coupe l'origine parallèlement à l'hyperplan tangent à la courbe de f à x_0. La fonction linéarisée est donc

$$f(x) \doteq f(x_0) + A_{x_0}(x - x_0). \qquad (2)$$

Par ailleurs, si B est une matrice telle que $df_{x_0}(dx) = B dx$, alors B est la dérivée de f à x_0 et comprend les dérivées partielles de f à x_0. Cette relation biunivoque entre différentielles et dérivées s'avère fort utile, car les différentielles sont faciles à manipuler.

Enfin, on omet habituellement l'indice 0 dans x_0, ce qui donne $df = A_x dx$.

2.3 Fonctions matricielles

Une fonction matricielle F d'un ensemble ouvert $S \subset \mathbb{R}^{m \times n}$ à $\mathbb{R}^{p \times q}$ est dérivable si vec F est dérivable. La dérivée DF correspond à la dérivée de vec F par rapport à vec X et est notée ∂ vec F/∂(vec X)'. La différentielle dF représente la fonction matricielle définie par vec $dF_{X_0}(U) = A_{X_0}$ vec U.

[1] Kees Zeelenberg, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands.

2.4 Propriétés des différentielles

Soit A, une matrice de constantes; F et G, des fonctions matricielles dérivables, et α, une valeur scalaire réelle. De ce qui précède, on peut facilement prouver les propriétés suivantes:

$$dA = 0, \tag{3}$$

$$d(\alpha F) = \alpha\, dF, \tag{4}$$

$$d(F + G) = dF + dG, \tag{5}$$

$$d(FG) = (dF)G + F(dG), \tag{6}$$

$$dF^{-1} = -F^{-1}(dF)F^{-1}. \tag{7}$$

Pour prouver la dernière propriété, il suffit de prendre la différentielle de $FF^{-1} = I$ et de la restructurer.

3. LINÉARISATION DE L'ESTIMATEUR DU COEFFICIENT DE RÉGRESSION

L'estimateur π (estimateur de Horvitz-Thompson) du coefficient de régression d'une population finie (lire Särndal et coll. 1992, partie 5.10) est

$$\hat{B} = \hat{T}^{-1}\hat{t}, \tag{8}$$

où

$$\hat{T} = \sum_{k \in s} \frac{x_k x_k'}{\pi_k},$$

$$\hat{t} = \sum_{k \in s} \frac{x_k y_k}{\pi_k},$$

y_k est la variable à laquelle on s'intéresse pour chaque k, x_k est le vecteur de chaque k avec les variables auxiliaires, π_k est la probabilité d'inclusion de chaque k et s représente l'échantillon.

Lorsqu'on prend la différentielle totale de (8) au moyen des propriétés (6) et (7) et l'évalue au point où $\hat{T} = T$, $\hat{t} = t$, on obtient

$$d\hat{B} = -T^{-1}(d\hat{T})T^{-1}t + T^{-1}(d\hat{t}). \tag{9}$$

Étant donné le lien entre les différentielles et l'approximation linéaire signalé à l'équation (2), il s'ensuit que (9) correspond à la linéarisation de l'estimateur du coefficient de régression:

$$\hat{B} \doteq B - T^{-1}(\hat{T}-T)T^{-1}t + T^{-1}(\hat{t}-t) = B + T^{-1}(\hat{t}-\hat{T}B),$$

où $B = T^{-1}t$.

4. LINÉARISATION DE L'ESTIMATEUR DE RÉGRESSION

L'estimateur de régression d'un chiffre de population est (lire Särndal et coll. 1992, partie 6.6)

$$\hat{t}_{yr} = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})'\hat{B}, \tag{10}$$

où $\hat{t}_{y\pi}$ est l'estimateur π de la variable à laquelle on s'intéresse, t_x est le vecteur du chiffre de population pour les variables auxiliaires, $\hat{t}_{x\pi}$ est le vecteur des estimateurs π des variables auxiliaires et $\hat{B}$ est l'estimateur du coefficient de régression des variables auxiliaires de la variable à laquelle on s'intéresse. Lorsqu'on calcule le différentielle totale de (10) grâce aux propriétés (3) et (6) et l'évalue au point où $\hat{t}_{y\pi} = t_y$, $\hat{t}_{x\pi} = t_x$, et $\hat{B} = B$, on obtient l'approximation linéaire de l'estimateur de régression

$$d\hat{t}_{yr} = d\hat{t}_{y\pi} - (d\hat{t}_{x\pi})'B,$$

si bien que

$$\hat{t}_{yr} \doteq t_y + \hat{t}_{y\pi} - t_y + (t_x - \hat{t}_{x\pi})'B = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})'B.$$

Notons qu'il n'est pas nécessaire de linéariser l'estimateur du coefficient de régression B pour linéariser l'estimateur de régression.

REMERCIEMENTS

L'auteur tient à remercier Jeroen Pannekoek, Jos de Ree, Robbert Renssen, deux arbitres et un rédacteur associé pour leurs précieux commentaires. Le point de vue exprimé dans cet article n'engage que l'auteur et ne reflète pas nécessairement les politiques du Statistics Netherland.

BIBLIOGRAPHIE

BINDER, D.A. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases: Une approche de type «recette». *Techniques d'enquête*, 22, 17-22.

MAGNUS, J.R., et NEUDECKER, H. (1988). *Matrix Differential Calculus*. New York: Wiley.

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

ZEELENBERG, C. (1993). A Survey of Matrix Differentiation. Research Paper, Department of Statistical Methods. Voorburg: Statistics Netherlands.

CONTENTS TABLE DES MATIÈRES

Volume 25, No. 1, March/mars 1997

J.N.K. RAO
Developments in sample survey theory: an appraisal

T.M. Fred SMITH
Social surveys and social science

Feifang HU
The asymptotic properties of the maximum relevance weighted likelihood estimators

R.R. SITTER and J.N.K. RAO
Imputation for missing values and corresponding variance estimation

Patrick J. FARRELL, Brenda MacGIBBON and Thomas J. TOMBERLIN
Bootstrap adjustments for empirical Bayes interval estimates of small area proportions

D.A.S. FRASER, N. REID and A. WONG
Simple and accurate inference for the mean of the gamma model

Jianguo SUN and David E. MATTHEWS
A random-effect regression model for medical follow-up studies

Philippe CAPÉRAÀ and Ana Isabel Garralda GUILLEM
Taux de résistance des tests de rang d'indépendance

Volume 25, No. 2, June/juin 1997

X. Joan HU and Jerald F. LAWLESS
Pseudolikelihood estimation in a class of problems with response-related missing covariates

Irwin GUTTMAN and George D. PAPANDONATOS
A Bayesian approach to a reliability problem: theory, analysis and interesting numerics

R.J. OHARA HINES
Fitting generalized linear models to retrospectively sampled clusters with categorical responses

R.R. SITTER and I. FAINARU
Optimal designs for the logit and probit models for binary data

Boxin TANG and C.F.J. WU
A method for constructing supersaturated designs and its $E(s^2)$ optimality

Shu YAMADA and Dennis K.J. LIN
Supersaturated design including an orthogonal base

A.G. BENN and R.J. KULPBERGER
Integrated marked Poisson processes with application to image correlation spectroscopy

Khalid El HIMDI and Roch ROY
Tests for the non-correlation of two multivariate ARMA time series

John J. SPINELLI and Michael A. STEPHENS
Cramér-von Mises tests of fit for the Poisson distribution

Thomas W. O'GORMAN
An adaptive test for the one-way layout

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 emphized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 12, Number 4, 1996

Derivation and Properties of the X11ARIMA and Census X11 Linear Filters
Estela Bee Dagum, Norma Chhab, and Kim Chiu 329

Correcting Unit Nonresponse via Response Modeling and Raking in the California Tobacco Survey
Charles C. Berry, Shirley W. Cavin, and John P. Pierce 349

Multiple Workloads per Stratum Designs
Lynn Weidmann and Lawrence R. Ernst 365

Neural Network Imputation Applied to the Norwegian 1990 Population Census Data
Svein Nordbotten 385

Modeling Income in the U.S. Consumer Expenditure Survey
Geoffrey D. Paulin and Elizabeth M. Sweet 403

The Survey Reinterview: Respondent Perceptions and Response Strategies
Johnny Blair and Seymour Sudman 421

Corrigendum 427

Book Reviews 429

Editorial Collaborators 441

Index to Volume 12, 1996 445

Volume 13, Number 1, 1997

Who Lives Here? Survey Undercoverage and Household Roster Questions
Roger Tourangeau, Gary Shapiro, Anne Kearney, and Lawrence Ernst 1

Suggestive Interviewer Behaviour in Surveys: An Experimental Study
Johannes H. Smit, Wil Dijkstra, and Johannes van der Zouwen 19

Effects of Post-Stratification on the Estimates of the Finnish Labour Force Survey
Kari Djerf 29

Variance Estimation for Measures of Income Inequality and Polarization - The Estimating Equations Approach
Milorad S. Kovačević and David A. Binder 41

Issues in the Use of a Plant-Capture Method for Estimating the Size of the Street Dwelling Population
Elizabeth Martin, Eugene Laska, Kim Hopper, Morris Meisner, and Joe Wanderling 59

A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data
Stephen E. Fienberg, Udi E. Makov, and Ashish P. Sanil 75

Book Review 91

In Other Journals 101

Volume 13, Number 2, 1997

Evaluation of a Reconstruction of the Adjusted 1990 Census for Florida
Michael M. Meyer and Joseph B. Kadane 103

Individual Diaries and Expense Documents in the Italian Consumer Expenditure Survey
Carlo Filippucci and Maria Rosaria Ferrante 113

Testing of Distribution Functions from Complex Sample Surveys
Abba M. Krieger and Danny Pfeffermann 123

Estimating Consumer Price Indices for Small Reference Populations
Martin Boon and Jan de Haan 143

Cognitive Dynamics of Proxy Responding: The Diverging Perspectives of Actors and Observers
Norbert Schwarz and Tracy Wellens 159

Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality
Bärbel Knäuper, Robert F. Belli, Daniel H. Hill, and A. Regula Herzog 181

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.

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 suivants:

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.