Statistique Canada Statistics Canada

TECHNIQUES D'ENQUÊTE

UNE REVUE
DE
STATISTIQUE CANADA

VOLUME 13, NUMÉRO 2
DÉCEMBRE 1987

Canada

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010300617

DATE

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA

DÉCEMBRE 1987

Publication autorisée par
le ministre des Approvisionnements
et Services Canada

©Ministre des Approvisionnements
et Services Canada 1988

Le lecteur peut reproduire sans autorisation des extraits de cette publication à des fins d'utilisation personnelle à condition d'indiquer la source en entier. Toutefois, la reproduction de cette publication en tout ou en partie à des fins commerciales ou de redistribution nécessite l'obtention au préalable d'une autorisation écrite des Services d'édition, Agent de droit d'auteur, Centre d'édition du gouvernement du Canada, Ottawa, Canada K1A 0S9.

Mai 1988

Prix: Canada, $20.00 par année
Autres pays, $23.00 par année

Paiement en dollars canadiens ou l'équivalent

Catalogue 12-001, vol. 13, n° 2

ISSN 0714-0045

Ottawa

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue 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.

COMITÉ DE DIRECTION

Président G.J. Brackstone

Membres B.N. Chinnappa, G.J.C. Hole, C. Patrick, F. Mayda (Directeur de la production), R. Platek, D. Roy, M.P. Singh

COMITÉ DE RÉDACTION

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

Rédacteurs associés

K.G. Basavarajappa, *Statistique Canada*
D.R. Bellhouse, *University of Western Ontario*
L. Biggeri, *Université de Florence*
D. Binder, *Statistique Canada*
E.B. Dagum, *Statistique Canada*
W.A. Fuller, *Iowa State University*
J.F. Gentleman, *Statistique Canada*
D. Holt, *University of Southampton*
G. Kalton, *University of Michigan*
M.N. Murthy, *Applied Statistics Centre, Inde*
W.M. Podehl, *Statistique Canada*
J.N.K. Rao, *Carleton University*
I. Sande, *Statistique Canada*
C.E. Särndal, *Université de Montréal*
F.J. Scheuren, *U.S. Internal Revenue Service*
V. Tremblay, *Statplus, Montréal*
K.M. Wolter, *U.S. Bureau of the Census*

Rédacteurs adjoints

J. Armstrong, J. Gambino et H. Lee, *Statistique Canada*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistique 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 statistique, 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

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociale, Statistique Canada, 4e étage, Édifice Jean-Talon, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 20.00$ par année au Canada, et de 23.00$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit, soit $10.00 (É.-U.) ($14.00 Can.) est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada. Veuillez envoyer votre demande d'abonnement directement à l'organisation.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 13, numéro 2, décembre 1987

TABLE DES MATIÈRES

Dans ce numéro 135

A. GIOMMI
Méthodes non paramétriques pour l'estimation des probabilités de réponse individuelles 137

C.S. WITHERS
Estimations fondées sur des données arrondies aléatoirement 145

G.H. CHOUDHRY et H. LEE
Estimation de la variance pour l'enquête sur la population active du Canada 157

P. ANTOINE, X. BRY, et P.D. DIOUF
La fiche «AGEVEN»: un outil pour la collecte des données rétrospectives 173

Section Spéciale - Méthodes d'estimation et de pondération

K.R. COPELAND, F.K. PEITZMEIER, et C.E. HOY
Méthode alternative pour ajuster les estimations de la Current Population Survey aux chiffres de population 183

C.H. ALEXANDER
Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages 193

G. LEMAÎTRE et J. DUFOUR
Une méthode intégrée de pondération des personnes et des familles 211

H.L. OH et F. SCHEUREN
Variante de la méthode itérative du quotient 221

Communications brèves

S.G. PRABHU-AJGAONKAR
Comparaison de la méthode de Horvitz-Thompson et de la méthode de Hansen-Hurwitz 233

Remerciements 237

Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 135-136
Statistique Canada

Dans ce numéro

Deux nouvelles rubriques paraissent dans la présente livraison de Techniques d'enquête. "**Dans ce numéro**" résume les articles que renferme le numéro, et elle paraîtra régulièrement. La deuxième s'intitule "**Communications brèves**" et elle sera publiée à l'occasion.

Vous trouverez dans ce numéro neuf articles dont quatre consacrés à des **méthodes d'estimation et de pondération**; deux de ces derniers traitent des estimations sur les familles. C'est grâce à l'initiative de Fritz Scheuren et à son aide lors de la rédaction que cette section spéciale paraît dans la présente livraison de Techniques d'enquête.

Les trois premiers articles de la section spéciale traitent (au moins partiellement) des méthodes des moindres carrés pour pondérer des données d'enquête. Il y a là une certaine ironie du sort: dans leur article de 1940, Deming et Stephan ont présenté l'ajustement proportionnel itératif comme une méthode rapide et pratique pour obtenir des approximations des estimations obtenues en minimisant une fonction quadratique des cellules d'un tableau de contingence, sous réserve de restrictions marginales. Cette technique s'est passablement répandue comme mode de pondération des données d'enquête, on l'appelle la "méthode itérative du quotient".

Dans "Méthode alternative pour ajuster les estimations de la Current Population Survey aux chiffres de population", Copeland, Peitzmeier et Hoy comparent un estimateur fondé sur la méthode itérative du quotient et un estimateur fondé sur la méthode des moindres carrés généralisés, sous réserve dans les deux cas des mêmes restrictions marginales. Ils comparent des estimations de caractéristiques individuelles obtenues lors de la Current Population Survey, une enquête sur les ménages exécuté par le Bureau of the Census des États-Unis. Ils notent que les estimations produites par les deux méthodes sont semblables.

La plupart des méthodes actuelles de pondération des données d'enquête sur les ménages donnent des poids différents pour les divers membres d'un ménage donné. Un seul poids par ménage, en plus d'être conceptuellement attrayant, éliminerait les différences répétées et souvent embarrassantes entre les estimations par personne et par famille. Alexander, dans "Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages", considère une catégorie de méthodes fondées sur la notion de "distance minimum conditionnelle" (dont celle des moindres carrés généralisés) qui en fait produisent un seul poids par famille tout en respectant les chiffres de population pour les individus. Les propriétés de ces méthodes en cas de sous-dénombrement sont ensuite étudiées au moyen de quelques modèles de couverture simples.

Lemaître et Dufour, dans "Une méthode intégrée de pondération des personnes et des familles", proposent un estimateur par la régression qui produit également un seul poids par ménage et qui, sous certaines réserves générales, équivaut à l'estimateur par les moindres carrés généralisés. En se servant de données de l'enquête sur la population active du Canada, ils atteignent un degré d'efficacité beaucoup plus élevé dans leurs estimations relatives aux familles et un peu plus élevé dans leurs estimations de personnes, comparativement aux méthodes actuelles.

Dans le dernier article de la section spéciale, intitulé "Variante de la méthode itérative du quotient", Oh et Scheuren décrivent une méthode d'estimation qui ressemble à la méthode itérative usuelle. Leur méthode peut être utilisée lorsque les totaux de la population sont disponibles non seulement pour les marges mais aussi pour les cellules intérieures d'un tableau multidimensionnel. Elle combine l'estimation conventionnelle par le quotient pour les cellules qui contiennent les échantillons de grande taille et l'estimation itérative par le quotient pour les cellules dans lesquelles l'échantillon est petit (ou inexistant). Dans une application concernant un échantillonnage de déclarations d'impôt de sociétés, la méthode Oh-Scheuren a produit des estimations plus efficaces que la méthode itérative du quotient conventionnelle. Les auteurs soulignent le fait qu'il reste encore du travail à faire, notamment comparer leur méthode aux manières conventionnelles de regrouper les données, avant de songer à répandre cette approche.

Les quatre autres articles du présent numéro traitent de l'élaboration et de l'application de méthodes et de procédures relatives aux probabilités de réponse dans le cadre d'une enquête, aux critères d'arrondissement visant à protéger la confidentialité, à la collecte et l'analyse des données aux fins d'enquêtes rétrospectives et à l'estimation de la variance dans l'enquête sur la population active du Canada.

Toute enquête donne lieu à des problèmes de non-réponse. On les résoud ordinairement par imputation ou ajustement, en supposant que la non-réponse est aléatoire à l'intérieur des catégories d'imputation ou d'ajustement. Les estimations ainsi produites sont généralement biaisées si l'hypothèse précitée n'est pas satisfaite. On a proposé diverses méthodes d'estimation des probabilités de réponse au moyen de modèles, notamment la méthode de Cassel, Sarndal et Wretman (CSW), mais ces méthodes ne sont pas efficaces lorsque le modèle supposé est inadéquat. Dans "Méthodes non paramétriques pour l'estimation des probabilités de réponse individuelles", Giommi décrit des méthodes non paramétriques d'estimation des probabilités de réponse avec utilisation de l'information auxiliaire, fournissant ainsi une alternative à l'estimateur de CSW qui résiste bien aux problèmes causés par des modèles de population et de réponse qui ne sont pas valables. Les estimateurs obtenus donnent de bons résultats dans les simulations de Monte Carlo.

L'arrondissement aléatoire est utilisé pour garantir le caractère confidentiel des renseignements sur des personnes dans les agrégats statistiques. Dans le contexte du recensement du Canada de 1971, Nargundkar et Saveland ont élaboré un processus d'arrondissement sans biais dans lequel la valeur attendue des données arrondies est la même que celle des données non arrondies. Fellegi (TE, 1975) a proposé l'arrondissement aléatoire contrôlé qui, en plus d'être sans biais, conserve aux données leur caractère additif. Plusieurs autres articles sur la question ont été publiés depuis, dont celui très récent de Cox (JASA, 1987), qui généralisent et étendent les applications à d'autres domaines. Dans "Estimations fondées sur des données arrondies aléatoirement", Withers donne une expression de la variance des estimations sans biais des probabilités de cellule et présente une comparaison de l'efficacité faisant intervenir les méthodes d'arrondissement utilisées en Australie, au Royaume-Uni, en Nouvelle-Zélande et au Canada. Il étend aussi ses résultats à toute fonction lisse des probabilités de cellule, à appliquer à divers secteurs de la statistique.

Dans "Estimation de la variance pour l'enquête sur la population active du Canada", Choudry et Lee décrivent les études effectuées en vue de choisir un estimateur de la variance des estimations itératives par le quotient provenant de l'enquête sur la population active. Ils donnent les résultats d'une comparaison de trois estimateurs de la variance pour le plan d'échantillonnage par groupe aléatoire (Keyfitz, Rao-Hartley-Cochran et Rao). Même si elle est inférieure aux deux autres méthodes pour ce qui est du biais et de la stabilité, la méthode de Keyfitz est recommandée parce qu'elle est simple à utiliser.

Dans "La fiche AGEVEN: un outil pour la collecte des données rétrospectives", Antoine, Bry et Diouf décrivent les techniques utilisées pour recueillir des données sur la natalité et la mortalité auprès des femmes de Pikine, une banlieue de Dakar, au Sénégal. La démarche rétrospective suivie comporte le placement d'événements observés (principalement les naissances et les décès) dans leur contexte socio-économique; elle permet, selon les auteurs, "de mieux évaluer la relation entre l'insertion urbaine et les changements de comportement démographique". On constate aisément en analysant les données de l'enquête que le taux de mortalité est plus élevé chez les enfants nés dans des villages ruraux que chez ceux qui naissent à Pikine.

Il est bien connu que la stratégie de Hansen et Hurwitz est inférieure à celle de Horvitz et Thompson dans le cas d'un certain nombre de procédures d'échantillonnage avec PSPT (probabilité de sélection proportionnelle à la taille). Dans le dernier article du numéro, dans la section "Communications brèves", Prabhu-Ajgaonkar donne de ces résultats des démonstrations beaucoup plus simples que celles qu'on trouve actuellement dans la littérature spécialisée.

Le rédacteur en chef

Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 137–144
Statistique Canada

Méthodes non paramétriques pour l'estimation des probabilités de réponse individuelles

ANDREA GIOMMI[1]

RÉSUMÉ

S'inspirant de l'approche de Cassel, Särndal et Wretman (1983), l'auteur aborde le problème de la non-réponse dans l'estimation de la moyenne d'une population finie. L'auteur propose tout d'abord deux méthodes très simples pour estimer les probabilités de réponse individuelles; il applique ensuite ces méthodes à un modèle de superpopulation pour construire des estimateurs de la moyenne de la population. Enfin, au moyen d'une étude de Monte Carlo, il fait une première évaluation des propriétés des méthodes proposées. Les résultats de cette évaluation nous éclairent sur l'efficacité de ces méthodes.

MOTS CLÉS: Non-réponse; probabilité de réponse individuelle; méthodes non paramétriques.

1. INTRODUCTION

S'intéressant à l'estimation de la moyenne (ou total, etc.) d'une population finie en situation de non-réponse, Cassel, Särndal et Wretman (1983) ont imaginé une méthode d'estimation très générale qui repose sur la notion fondamentale de probabilité de réponse individuelle (PRI). Ils ont proposé des estimateurs qui sont déterminés en partie par un modèle de superpopulation et en partie par un modèle de réponse, c'est-à-dire un modèle qui reproduit le mécanisme de réponse et qui permet d'estimer la PRI à partir de données d'échantillon. L'estimation de la PRI est le point central de leur théorie. De fait, si le modèle de superpopulation ne convient pas parfaitement, comme cela est souvent le cas, seul un modèle de réponse approprié évitera que les estimateurs soient entachés d'un biais attribuable au plan de sondage. À l'aide d'une étude de Monte Carlo, Giommi (1985a) a montré qu'un modèle de réponse qui offrait une "bonne approximation" du "vrai" modèle de réponse pouvait éliminer virtuellement toute possibilité d'erreur systématique. Toutefois, on peut difficilement dire ce qu'est une bonne approximation et, de toute façon, le choix d'un modèle de réponse, outre qu'il est arbitraire, peut s'avérer contraignant. Une manière naturelle de contourner le problème est d'estimer la PRI par des méthodes non paramétriques. Dans cet article, nous proposons deux méthodes très simples pour estimer la PRI lorsque l'information supplémentaire dont nous disposons (et qui est censée dépendre des réponses obtenues) est représentée par une variable continue unique. Ces méthodes, qui intègrent certains éléments de la théorie d'estimation de noyaux, peuvent être considérées comme un prolongement de la méthode de compensation de la non-réponse, d'usage courant, qui consiste à repondérer les unités au moyen de cellules de correction.

Dans cet article, nous faisons une évaluation empirique de ces méthodes et analysons le biais et l'efficacité des estimateurs correspondants.

2. ESTIMATION DES PROBABILITÉS DE RÉPONSE INDIVIDUELLES

Considérons une population de N unités identifiées k ($k=1, 2, \ldots, N$), et une variable Y à l'étude, dont nous voulons estimer la moyenne $\bar{Y}=\Sigma_k\, y_k/N$ à l'aide d'un échantillon s de

[1] Andrea Giommi, Département de statistique, Université de Florence, Via Curtatone, 1, 50123 Florence, Italie.

n unités prélevé suivant un plan d'échantillonage $p(s)$. Pour l'estimation, nous disposons de renseignements supplémentaires représentés par les valeurs connues x_k $(k=1, \ldots, N)$, d'une variable scalaire continue X (en principe, l'application des méthodes proposées pour le cas à plusieurs variables n'est pas censée poser de problèmes).

Dans l'échantillon, Y est observable uniquement pour un sous-ensemble r de n_r répondants; elle ne peut être observée pour les $n - n_r$ non-répondants. Après l'échantillonnage, l'information disponible peut être représentée de la façon suivante:

$$(k, I_k, I_k y_k, x_k) \qquad k \in s; N, n,$$

où I_k est une variable aléatoire indicatrice telle que $E(I_k) = q_k$ et q_k est la PRI.

Pour estimer q_k, on suppose habituellement un modèle paramétrique (Cassel et coll. 1983) de la forme:

$$q_k = q(\Theta, x_k),$$

où Θ est un paramètre (ou vecteur de paramètres) inconnu et $q(\cdot,\cdot)$ est une fonction à définir. On obtient des estimations de q_k en remplaçant Θ par son estimation $\hat{\Theta}$ dans le modèle paramétrique ci-dessus.

Dans cet article, nous estimons q_k $(k \in r)$ sans utiliser de définition paramétrique de la fonction $q(\cdot,\cdot)$; nous continuons toutefois à supposer que les PRI dépendent des valeurs de x_k. Deux méthodes (méthodes (1) et (2)) sont proposées.

Selon la première méthode, q_k $(k \in r)$ est estimée comme taux de réponse (c'est-à-dire, la proportion de répondants) d'un groupe d'unités centré sur l'unité k et qui correspond à un intervalle approprié de valeurs de x centré sur x_k. En supposant qu'il s'agit d'un intervalle de longueur $2h_k$, q_k est estimée par le ratio suivant:

$$\hat{q}_k = \sum_{j \in r} D(x_k - x_j) \Big/ \sum_{j \in s} D(x_k - x_j), \tag{1}$$

où

$$D(x_k - x_j) = \begin{cases} 1 \text{ si } |x_k - x_j| \leq h_k \\ 0 \text{ dans le cas contraire.} \end{cases}$$

Il est évident que l'estimation $\hat{q}_k$ dépend de h_k ou de h si nous utilisons, comme c'est le cas ici, un intervalle fixe; dans la pratique, la détermination de la valeur de h est un problème majeur.

Selon la seconde méthode, l'estimation de q_k repose sur toutes les unités de l'échantillon et non seulement sur une partie d'entre elles. Cela élimine les inconvénients que pouvait poser une catégorisation des unités répondantes. Autrement dit, on pouvait très difficilement imaginer, pour l'estimation de q_k, que des unités aient une pondération de 1 tandis que d'autres avaient une pondération nulle. Selon la méthode (2), l'estimation est définie par:

$$\hat{q}_k = \sum_{j \in r} D^*(x_k - x_j) \Big/ \sum_{j \in s} D^*(x_k - x_j), \tag{2}$$

où D^* doit être définie. Dans ce cas, chaque valeur x_j contribue au calcul de l'estimation

$\hat{q}_j$ par l'intermédiaire de D^*, qui est une quantité inversement proportionnelle à l'écart $|x_k - x_j|$.

La méthode (2) comporte deux difficultés; il faut en effet définir i) la fonction D^* et ii) les valeurs de ses paramètres. Dans cet article, nous définissons D^* comme une fonction de distribution d'une loi normale:

$$D^*(z) = (h^2 2\pi)^{-1/2} \exp(-z^2/2h^2); \qquad z = x_k - x_j, \tag{3}$$

où l'écart-type, désigné par h, joue le même rôle que le paramètre h de l'expression (1). Lorsque h augmente dans les expressions (1) et (2), $\hat{q}_k$ tend vers la valeur constante n_r/n. Dans (1), il atteint n_r/n lorsque h est au moins égal à l'intervalle des valeurs de x.

Une étude empirique visant à évaluer les propriétés des méthodes proposées a été conçue pour un très grand nombre de valeurs de h. Dans le présent article, nous ne rapportons que les résultats se rattachant à trois de ces valeurs (constantes), soit $h = 1/10$, $3/10$ et $5/10$ de l'intervalle des valeurs de x de l'échantillon. Enfin, soulignons que les équations (1) et (2) ne sont pas que des facteurs de normalisation; chacune d'elles se présente aussi comme le rapport de deux estimateurs noyaux de fonction de densité (selon l'approche de Rosenblatt (1956) pour diverses séries de valeurs de x. Comme le propose Giommi (1985b), on peut donc choisir la valeur de h en tenant compte des propositions contenues dans cette théorie.

3. MODÈLE DE SUPERPOPULATION ET ESTIMATEURS

Pour le choix de l'estimateur de $\bar{Y}$, nous supposons un modèle de superpopulation Φ, où les valeurs de la population y_k, $k = 1, 2, \ldots, N$, sont réputées former un échantillon aléatoire de telle sorte que:

$$\begin{aligned} E_\Phi(Y_k) &= \mu_k = \beta x_k, \\ \mathrm{Var}_\Phi(Y_k) &= \sigma_k^2 = \sigma^2 x_k, \end{aligned} \tag{4}$$

où β et Φ sont inconnues et x_k est la valeur (connue) de la variable auxiliaire X. Il est clair que ce modèle de superpopulation s'applique plus à des variables quantitatives qu'à des variables qualitatives; dans ces conditions, il conviendrait d'utiliser d'autres modèles. Par ailleurs, nous nous limitons ici à l'étude des échantillons aléatoires simples. À la condition que la variance de Y puisse être définie par l'équation en (4), Cassel et coll. (1983) ont montré que l'estimateur:

$$T = \bar{X}\left(\sum_r / y_k / q_k\right) / \left(\sum_r / x_k / q_k\right),$$

où Σ_r indique la sommation par rapport à l'ensemble r et $\bar{X} = \Sigma_k^N x_k / N$, est approximativement non biaisé (grâce au facteur de correction q_k) même si la première équation en (4) ne définit pas réellement la relation entre X et Y. Cette situation peut être observée, par exemple, lorsque le "véritable" modèle comporte une ordonnée à l'origine ou deux coefficients de régression (voir modèle (5) ci-dessous), etc.

Malheureusement, l'estimateur T ne se prête pas à des applications pratiques puisque q_k est inconnue. Nous devons donc évaluer les propriétés de cet estimateur en remplaçant q_k par son estimation calculée à l'aide de la méthode (1) ou (2).

Nous allons analyser ce genre d'estimateur pour les trois valeurs de h choisies. Nous désignons les estimateurs par TD_i et TD_i^* où $i = 1, 3, 5$ (voir le tableau 1).

Tableau 1

Définition des estimateurs

h	Estimateurs Méthode (1)	Méthode (2)
0.1	TD_1	TD_1^*
0.3	TD_3	TD_3^*
0.5	TD_5	TD_5^*

Par ailleurs, nous considérons les estimateurs suivants dans l'étude de Monte Carlo:

$$TC = \bar{X}\left(\sum_s y_k / \sum_s x_k\right) \quad \text{and} \quad TI = \bar{X}\left(\sum_r y_k / \sum_r x_k\right).$$

TC est l'estimateur pour l'échantillon complet, c'est-à-dire l'estimateur par quotient selon l'hypothèse de l'absence de non-réponse, et TI est le même estimateur pour le sous-ensemble des répondants sans compensation de la non-réponse. Soulignons que TI est aussi un estimateur qui découle d'une méthode d'imputation bien connue (par régression) (Cassel et coll. 1983) et qui est égal à TD lorsque h est au moins égal à l'intervalle des valeurs de x. TI n'est approximativement sans biais que si (4) est vraie. Nous verrons que le biais dépend de la différence entre les conditions définies en (4) et celles de la population à l'étude. Comme dans la simulation de la section suivante, le modèle (4) sera un "faux" modèle (c'est-à-dire que les populations étudiées sont définies par des modèles différents de (4)); par ailleurs, la simulation permet de mieux comprendre cette méthode d'imputation très simple et largement répandue.

4. ÉTUDE DE MONTE CARLO

Par une simulation de Monte Carlo, deux populations, POP1 et POP2, ont été produites selon la méthode utilisée par Särndal et Hui (1981). POP1 et POP2 comportent chacune deux strates ($S1$ et $S2$) de 500 unités chacune, et satisfont aux équations suivantes:

$$\begin{aligned} E_\Phi(Y_k) &= \beta_1 x_{k1} + \beta_2 x_{k2}, \\ \mathrm{Var}_\Phi(Y_k) &= \sigma_1^2 x_{k1} + \sigma_2^2 x_{k2}, \end{aligned} \tag{5}$$

où $x_{k1} = x_k \partial_k$ et $x_{k2} = x_k(1 - \partial_k)$, $\partial_k = 1$ si $k \in S1$ et $\partial_k = 0$ si $k \in S2$. La différence entre les modèles (4) et (5) sert à simuler l'une des nombreuses erreurs que l'on peut commettre en définissant le modèle de superpopulation. Les caractéristiques numériques de POP1 et POP2 figurent dans le tableau 2.

Voici en bref les étapes de la simulation:

1) un échantillon aléatoire simple s de n ($n=50$, 100) unités est prélevé dans chaque population;

2) on enregistre ensuite les données de l'échantillon complet puis on crée de la non-réponse à l'aide des deux modèles paramétriques suivants:

Modèle A: $q_k = \exp(-\Theta x_k)$,

Modèle B: $q_k = \Theta_1^{\partial_k} \Theta_2^{1-\partial_k}$; $\quad \partial_k = 1$ (0) si $k \in S1$ ($S2$),

Tableau 2

Propriétés des populations simulées

Population et strates		POP1 Moyenne	E.T.	C.V.	Asym.	POP2 Moyenne	E.T.	C.V.	Asym.
Strate n° 1	x	19.305	12.71	.66	1.30	20.037	14.50	.72	2.25
	y	7.612	5.38	.71	1.62	1.961	2.21	1.13	3.03
Strate n° 2	x	50.325	21.32	.42	.77	49.775	23.28	.47	1.21
	y	30.325	13.38	.44	.72	44.862	21.31	.47	1.04
Total	x	34.815	23.42	.67	.90	34.906	24.44	.70	1.32
	y	18.969	15.26	.80	1.06	23.411	26.25	1.12	1.15

E.T. = Ecart type de la population; Asym. = asymétrie (3-ième moment/(2-ième moment)$^{3/2}$); C.V. = coefficient de variation.

où les paramètres Θ, Θ_1, Θ_2 sont choisis de telle manière que le taux de réponse moyen $\bar{q}$ pour l'ensemble de la population soit de 0.6 ou de 0.7. En pratique, on forme des ensembles de répondants en exécutant un tirage de Bernoulli pour chaque unité $k \in s$ avec une probabilité de résultat favorable q_k (réponse) et une probabilité de résultat défavorable $1-q_k$ (non réponse);

3) on estime la PRI à l'aide des méthodes (1) et (2) et on calcule pour chaque échantillon les valeurs de *TC, TI, TD, TD**;

4) on répète 1000 fois les étapes 1 à 3 et on calcule à la fin le biais, la variance (VAR) et l'erreur quadratique moyenne (EQM) des estimateurs pour chaque combinaison possible des éléments suivants: taille de l'échantillon (50, 100), modèle de réponse (A, B), taux de réponse moyen (0.6, 0.7) et population (POP1, POP2).

Les résultats de la simulation figurent dans les tableaux 3 et 4.

5. RÉSULTATS DE L'ÉTUDE DE MONTE CARLO

L'analyse des tableaux 2 et 3 fait ressortir des points intéressants.

1. Comme prévu, *TC* est approximativement non biaisé dans toutes les simulations.

2. Le biais de *TI* est constamment supérieur à celui de *TD* et de *TD**. Il convient donc, du moins dans les situations qui s'apparentent à celles de la simulation, d'utiliser l'estimateur rajusté de préférence à l'estimateur non-rajusté, qui découle d'une méthode d'imputation par régression.

3. Pour la même valeur de h, le biais de *TD* est toujours moindre que celui de *TD**. Les écarts sont négligeables pour $h=.1$. Lorsque h augmente, *TD** se rapproche de *TI* plus rapidement que *TD*; lorsque $h=.5$, l'écart entre *TD** et *TI* est négligeable sur le plan pratique.

4. La réduction du biais découlant de la substitution de *TD* à *TI* est appréciable dans tous les cas (de 55 à 82% pour le modèle A et de 67 à 92% pour le modèle B). La substitution de *TD** à *TI* entraîne aussi une diminution notable du biais: de 51 à 68% pour le modèle A et de 61 à 84% pour le modèle B.

5. Pour $h=.1$, *TD* et *TD** s'équivalent au point de vue de l'EQM quoique le second soit légèrement plus stable (c'est-à-dire, variance moindre). Pour $h=.3$ et $h=.5$, la moins grande stabilité de *TD* par rapport à *TD** est compensée de façon générale par un biais moindre et cette compensation est telle que *TD* devient préférable à *TD** au point de vue de l'EQM.

6. Les estimateurs corrigés par l'estimation de la PRI ne sont pas très stables mais doivent être utilisés de préférence à *TI* si l'on se fonde sur l'EQM.

7. Comme prévu, le biais dépend directement du taux de non-réponse et de la différence entre le vrai modèle de superpopulation et le modèle hypothétique (c'est-à-dire, le "faux" modèle sur lequel reposent les estimateurs). L'utilisation de deux modèles de réponse (A et B) n'a pas d'incidence notable sur les résultats (voir Giommi (1984) pour les effets de divers modèles).

8. L'élargissement de l'échantillon semble se traduire par une légère diminution du biais pour tous les estimateurs étudiés. *TD* et *TD** sont des exceptions: dans ce cas, la réduction du biais n'est pas attribuable à la modification des conditions expérimentales mais à l'amélioration réelle de l'estimation q_k lorsque n augmente.

En conclusion, nous pouvons affirmer que les deux méthodes proposées dans cet article peuvent servir dans des situations comparables à celles que nous venons d'analyser et qu'il convient d'accorder une certaine préférence à la méthode (1) en raison de son application plus simple. Toutefois, nous n'avons pas résolu le problème de la détermination de la meilleure

Tableau 3

Rendement de divers estimateurs selon le modèle de réponse A

Estimateurs		TC	TI	TD_1	TD_3	TD_5	TD_1^*	TD_3^*	TD_5^*
				Taux de réponse moyen $\bar{q}=.60$					
				POP1					
$n=50$	BIAIS	.015	.861	.349	.420	.669	.380	.620	.765
	VAR	.405	.973	1.115	1.036	1.007	1.041	.995	.989
	EQM	.405	1.714	1.237	1.212	1.455	1.185	1.379	1.574
$n=100$	BIAIS	.007	.805	.164	.323	.610	.227	.544	.686
	VAR	.186	.416	.443	.429	.412	.415	.404	.402
	EQM	.186	1.064	.470	.533	.784	.467	.700	.873
				POP2					
$n=50$	BIAIS	.090	3.125	1.433	1.682	2.544	1.544	2.378	2.887
	VAR	3.952	8.744	9.821	9.823	9.743	9.390	9.233	9.118
	EQM	3.960	18.510	11.874	12.652	16.215	11.774	14.888	17.453
$n=100$	BIAIS	.056	2.959	.749	1.387	2.337	1.004	2.104	2.566
	VAR	1.710	4.144	4.515	5.122	4.819	4.238	4.632	4.518
	EQM	1.713	12.900	5.076	7.046	10.281	5.246	9.059	11.102
				Taux de réponse moyen $\bar{q}=.70$					
				POP1					
$n=50$	BIAIS	.015	.581	.226	.271	.418	.249	.415	.439
	VAR	.405	.765	.794	.750	.738	.754	.752	.753
	EQM	.405	1.103	.845	.823	.913	.816	.924	.946
$n=100$	BIAIS	.007	.531	.099	.205	.396	.143	.357	.457
	VAR	.186	.328	.323	.307	.327	.313	.327	.336
	EQM	.186	.610	.333	.349	.484	.333	.454	.545
				POP2					
$n=50$	BIAIS	.090	2.130	.813	.939	1.542	.887	1.453	1.822
	VAR	3.952	6.996	7.122	6.827	6.991	6.708	6.753	6.871
	EQM	3.960	11.533	7.783	7.709	9.396	7.495	8.864	10.191
$n=100$	BIAIS	.056	1.966	.473	.953	1.541	.658	1.406	1.732
	VAR	1.710	3.071	3.005	3.062	3.027	2.926	3.008	3.040
	EQM	1.713	6.937	3.229	3.970	5.402	3.359	4.985	6.040

Tableau 4

Rendement de divers estimateurs selon le modèle de réponse B

Estimateurs		TC	TI	TD_1	TD_3	TD_5	TD_1^*	TD_3^*	TD_5^*
				Taux de réponse moyen $\bar{q}=.60$					
				POP1					
$n=50$	BIAIS	.015	1.086	.290	.383	.716	.323	.688	.992
	VAR	.405	.966	1.208	1.011	.937	1.050	.907	.928
	EQM	.405	2.145	1.29	1.158	1.450	1.154	1.380	1.912
$n=100$	BIAIS	.007	1.079	.120	.349	.732	.196	.668	.902
	VAR	.186	.422	.513	.429	.420	.447	.401	.403
	EQM	.186	1.586	.527	.551	.956	.485	.847	1.217
				POP2					
$n=50$	BIAIS	.090	4.046	1.362	1.757	2.826	1.562	2.749	3.562
	VAR	3.952	10.285	12.519	12.089	12.010	11.605	11.046	10.994
	EQM	3.960	26.655	14.374	15.176	19.996	14.045	18.603	23.682
$n=100$	BIAIS	.056	3.897	.454	1.531	2.707	.853	2.521	3.284
	VAR	1.710	4.151	5.432	5.121	5.103	4.798	4.541	4.381
	EQM	1.713	19.338	5.638	7.465	12.431	5.525	10.896	15.166
				Taux de réponse moyen $\bar{q}=.70$					
				POP1					
$n=50$	BIAIS	.015	.584	.179	.221	.409	.196	.376	.499
	VAR	.405	.751	.826	.425	.716	.769	.723	.743
	EQM	.405	1.092	.858	.474	.883	.807	.864	.992
$n=100$	BIAIS	.007	.536	.046	.173	.365	.087	.317	.436
	VAR	.186	.307.	318	.295	.295	.299	.295	.302
	EQM	.186	.594	.320	.325	.428	.307	.395	.492
				POP2					
$n=50$	BIAIS	.090	2.057	.682	.891	1.477	.804	1.392	1.822
	VAR	3.952	6.199	6.788	6.165	6.232	6.340	6.093	6.270
	EQM	3.960	10.430	7.253	6.959	8.414	6.986	8.031	9.590
$n=100$	BIAIS	.056	1.918	.157	.755	1.311	.374	1.175	1.562
	VAR	1.710	2.826	2.897	2.884	2.867	2.796	2.836	2.923
	EQM	1.713	6.506	2.922	3.454	4.586	2.936	4.217	5.363

valeur de h (ou de h_k dans le cas général). Notre analyse nous a permis de constater que, selon certaines limites, de faibles valeurs de h réduisent le biais mais réduisent en même temps la stabilité de l'estimateur corrigé. Nous avons pu établir que la valeur optimum de h se situait autour de 0.1 pour notre analyse. Des résultats qui sont tirés de là même simulation mais qui ne sont pas reproduits ici indiquent que si l'on réduit davantage la valeur de h, le biais tend à s'accroître. Ce résultat est prévisible puisque en faisant tendre h vers 0, on obtient une série d'estimations $\hat{q}_k$ $(k=1, \ldots, n)$ égales à 1 ou à 0 selon qu'il s'agit de répondants ou de non-répondants.

6. REMERCIEMENTS

Je tiens à exprimer ma reconnaissance au professeur Luigi Biggeri pour le soutien qu'il m'a accordé au cours de la préparation de cet article. Je veux aussi remercier les arbitres pour les commentaires utiles qu'ils ont exprimés sur la version préliminaire.

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

GIOMMI, A. (1984). On a simple method for estimating individual response probabilities in sampling from finite populations. *Metron*, 42, 185-200.

GIOMMI, A. (1985a). On estimation in nonresponse situations. *Statistica*, 1, 57-63.

GIOMMI, A. (1985b). On the estimation of the individual response probabilities. *Proceedings of the 45th Session of the International Statistical Institute,* Vol. 2 (Contributed Papers), 577-578.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates for the density function. *Annals of Mathematical Statistics*, 27, 832-837.

SARNDAL, C.E., et T.K. HUI (1981). Estimation for nonresponse situations: to what extent must we rely on models? Dans *Current Topics in Survey Sampling,* (éd. D. Krewski, R. Platek et J.N.K. Rao), New York: Academic Press, 227-246.

Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 145–156
Statistique Canada

Estimations fondées sur des données arrondies aléatoirement

C.S. WITHERS[1]

RÉSUMÉ

Dans cet article, nous présentons des méthodes servant à estimer les fonctions des probabilités de cellule associées à un tableau de données multinomiales qui ont été arrondies aléatoirement selon des multiples d'un nombre donné l. Nous montrons que: (i) l'arrondissement aléatoire n'a que des effets de second ordre sur le biais et la variance; (ii) l'utilisation des estimateurs naturels des probabilités de cellule entraîne une très faible perte d'efficacité si la fréquence de la cellule est élevée par rapport à $(l^2 - 1)/(6R)$ où R désigne le nombre de cellules dans le tableau; et (iii) il existe des estimateurs dont le biais est apparemment de taille exponentiellement faible pour les moments de ces estimateurs naturels et les polynomes des probabilités de cellule.

MOTS-CLÉS: Arrondissement aléatoire; réduction du biais; efficacité.

1. INTRODUCTION ET SOMMAIRE

Dans cet article, nous décrivons des méthodes qui permettent d'estimer une fonction des probabilités de cellule associées à un tableau de données multinomiales arrondies aléatoirement. L'arrondissement aléatoire est une méthode très répandue qui sert à préserver le caractère confidentiel des données lorsqu'une valeur contenue dans un tableau (par ex. : 1) risque d'être associée à une personne en particulier, ce qui peut entraîner la violation de confidentialité. Cette méthode consiste à arrondir la fréquence N, d'un tableau au multiple d'un nombre donné l supérieur avec une probabilité (a.p.) α ou au multiple inférieur a.p. $1 - \alpha$, où α est choisie de telle manière que la valeur arrondie M satisfait l'équation suivante

$$E(M \mid N) = N.$$

En d'autres termes, si pour un nombre entier j, $jl \leq N < (j + 1)l$, alors

$$M = \begin{cases} jl \text{ a.p. } 1 - \alpha \\ (j + 1)l \text{ a.p. } \alpha \end{cases} \tag{1.1}$$

où $\alpha = r/l$ et $r = N - jl$.

Le département de Statistique de la Nouvelle-Zélande utilise $l = 3$ Penny comme base d'arrondissement l alors que Statistique Canada utiliserait $l = 5$. Voir Penny et Ryan (1986).

Il ne faut pas confondre l'arrondissement aléatoire avec le groupement ou arrondissement non aléatoire de données d'échantillon au multiple entier de l le plus près (associé avec les corrections de Sheppard pour les moments), ni avec la contamination aléatoire, qui est une autre méthode de confidentialité qui consiste simplement à ajouter à la fréquence N une variable aléatoire indépendante de moyenne 0. (Le principal inconvénient de cette méthode est qu'elle peut produire une fréquence négative). Pour obtenir les titres d'ouvrages traitant

[1] C.S. Withers, Applied Mathematics Division, Department of Scientific and Industrial Research, C.P. 1335, Wellington, Nouvelle-Zélande.

de ces méthodes, voir Gastwirth et coll. (1978) et Kendall et Stuart (1977). On trouvera également une bibliographie sur l'arrondissement aléatoire pour les données multidimensionnelles et les données groupées dans Gastwirth et coll. (1978).

Dans cet article, nous allons nous concentrer sur les problèmes liés à l'estimation d'une fonction de probabilités de celule associées à un tableau de R valeurs arrondies aléatoirement. Pour des raisons de commodité, nous désignons les probabilités de cellule par $p_1, \ldots, p_R$ au lieu de $\{p_{ij}, 1 \leq i \leq I, 1 \leq j \leq J\}$, comme cela se fait normalement pour un tableau $I \times J$.

Ainsi, $1 = \Sigma_1^R p_i$ et $n = \Sigma_1^R N_i$ est la somme des fréquences contenues dans le tableau. Soient $\{M_i\}$ les valeurs arrondies de $\{N_i\}$. Étant donné n, nous supposons que $\{N_i\}$ suit une loi multinomiale avec les paramètres n et $\{p_i\}$. L'hypothèse ci-dessus est vraie pour $p_i = m_i / \Sigma_j m_j$ si $\{N_i\}$ sont non conditionnellement des variables de Poisson indépendantes avec des moyennes $\{m_i\}$.

Deux estimateurs sans biais de p_1 sont définis par:

$$p_1^* = N_1 / n \quad \text{et } \hat{p}_1 = M_1 / n. \tag{1.2}$$

Le premier n'est pas un vrai estimateur puisqu'on ne connaît pas N_1. Le second est l'estimateur naturel. (Nous supposons que n est connu. S'il ne l'est pas nous pouvons sans difficulté le remplacer par $\Sigma_1^R M_i$.) Il existe toutefois d'autres estimateurs sans biais, notamment "l'estimateur complémentaire"

$$\tilde{p}_1 = - \sum_{j \neq 1} M_j / n, \tag{1.3}$$

d'où

$$p_1(\lambda) = (1 - \lambda)\hat{p}_1 + \lambda \tilde{p}_1 \text{ pour n'importe valeur } \lambda \text{ donnée.} \tag{1.4}$$

Il faut alors se demander quelle est la meilleure valeur de λ que l'on peut utiliser et quelle perte d'efficacité doit-on subir si l'on utilise l'estimateur naturel c'est-à-dire $\lambda = 0$. Pour répondre à ces questions, il faut connaître les variances de ces estimateurs. Elles sont définies ci-dessous.

Théorème 1.1

Soit

$$\operatorname{var}(\hat{p}_1) = (p_1 - p_1^2)\, n^{-1} + \{(l^2 - 1)/6 + \Delta_n(p_1)\} n^{-2} = v_n(p_1), \tag{1.5}$$

où

$$\Delta_n(p_1) = \sum_{i=0}^{l-1} i(l - i)\, \{P(N_1 \bmod l = i) - l^{-1}\}. \tag{1.6}$$

Aussi,

$$\operatorname{var}(\tilde{p}_1) = (p_1 - p_1^2) n^{-1} + \{(R - 1)(l^2 - 1)/6 + \sum_{j \neq 1} \Delta_n(p_j)\} n^{-2}, \tag{1.7}$$

et

$$\operatorname{var}(p_1(\lambda)) = (p_1 - p_1^2) n^{-1} + \{\alpha(\lambda)(l^2 - 1)/6 + \nabla_n(p)\} n^{-2}, \tag{1.8}$$

où

$$\alpha(\lambda) = (1 - \lambda)^2 + (R - 1)\lambda^2 \tag{1.9}$$

et

$$\nabla_n(p) = (1 - \lambda)^2 \Delta_n(p_1) + \lambda^2 \Sigma_{i \neq 1} \Delta_n(p_i). \tag{1.10}$$

(Les démonstrations se trouvent à la section 2.)

Dans l'annexe A, nous démontrons que pour $0 < p_1 < 1$, $P(N_1 \bmod l = i) - l^{-1} \to 0$ de façon exponentielle lorsque $n \to \infty$, de sorte que $\Delta_n(p_1) \to 0$ de façon exponentielle lorsque $n \to \infty$, et $\nabla_n(p)$ fait de même à la condition que $p_i \neq 0$ pour tous les i.

Comme $\alpha(\lambda)$ est minimisée par $\lambda_R = R^{-1}$ et $\alpha(\lambda_R) = 1 - R^{-1}$, $\mathrm{var}(p_1(\lambda))$ l'est aussi asymptotiquement. Ainsi, la perte d'efficacité causée par l'utilisation de l'estimateur naturel $\hat{p}_1$ au lieu de l'estimateur asymptotiquement optimal sans biais $p_1(\lambda_R)$ lorsque R est grand) est donnée par

$$\{\mathrm{var}(\hat{p}_1) - \mathrm{var}(p_1(\lambda_R))\}/\mathrm{var}(p_1(\lambda_R)) \approx (l^2 - 1)/\{6Rn(p_1 - p_1^2)\}, \qquad (1.11)$$

laquelle valeur est négligeable si $M_1(1 - M_1/n) \approx n(p_1 - p_1^2)$ est élevée par rapport à $(l^2 - 1)/\{6R\}$.

En règle générale, M_1 est une bonne approximation du membre gauche de (1.11). Nous avons ainsi une façon pratique de vérifier l'efficacité des estimateurs naturels. Si une ou plusieurs valeurs $\{p_i\}$ sont nulles, $p_i = 0$, alors $N_i = M_i = 0$, nous disons que $\Sigma_{i \neq 1}$ ne comprend pas les cellules pour lesquelles $p_i = 0$, et que R désigne le nombre de cellules du tableau pour lesquelles $p_i \neq 0$.)

En nous servant de l'équation (1.5) nous pouvons maintenant faire une brève comparaison de l'arrondissement aléatoire et de la contamination aléatoire. Selon nos informations, les organismes statistiques de l'Australie et du Royaume-Uni arrondissent les fréquences de cellule en leur ajoutant 1 avec une probabilité 1/4, 0 avec une probabilité 1/2 et -1 avec une probabilité 1/4, de sorte que

$$\mathrm{var}(\hat{p}_1) = (p_1 - p_1^2)n^{-1} + 1/2n^{-2}.$$

Le coefficient 1/2 passe à 4/3 dans le cas de la Nouvelle-Zélande ($l = 3$) et à 4 pour le Canada ($l = 5$). La méthode de la contamination aléatoire implique une moins grande protection (une variation maximale de 1 par opposition à 2 pour le système de Nouvelle-Zélande et à 4 pour le système canadien) et la possibilité d'une fréquence négative si elle est appliquée à des cellules vides.

Le théorème 1.1 montre que l'arrondissement aléatoire n'a qu'un effet de second ordre sur l'efficacité de l'estimateur de p_1; la variance n'est accrue que par un facteur de l'ordre n^{-2}. La démonstration suivante montre que ce résultat très important est aussi vrai pour l'estimation de n'importe quelle fonction lisse de $\{p_i\}$. Posons $r = R - 1$, $\mathbf{p} = (p_1, \ldots, p_r)$, $\mathbf{N} = (N_1, \ldots, N_r)$, $\mathbf{M} = (M_1, \ldots, M_r)$, $\mathbf{p}^* = \mathbf{N}/n$ et $\hat{\mathbf{p}} = \mathbf{M}/n$. Alors nous avons $\mathrm{cov}(\mathbf{p}^*) = V/n$ où $V = \mathrm{diag}(\mathbf{p} - \mathbf{p}\mathbf{p}')$. Supposons maintenant que nous voulons estimer $f(\mathbf{p})$, une fonction dont les secondes dérivées sont continues.

Ainsi, $\dot{f}(\mathbf{p}) = \partial f(\mathbf{p})/\partial \mathbf{p}$ est une fonction continue $r \times 1$ et $\ddot{f}(\mathbf{p}) = \partial^2 f(\mathbf{p})/\partial \mathbf{p} \partial \mathbf{p}'$ est une fonction continue $r \times r$.

Théorème 1.2. Lorsque $n \to \infty$, $E(f(\mathbf{p}^*))$ et $E(f(\hat{\mathbf{p}}))$ sont toutes deux égales à

$$f(\mathbf{p}) + B(\mathbf{p})n^{-1} + O(n^{-2}) \text{ où } B(\mathbf{p}) = \mathrm{trace}\,(\ddot{f}(\mathbf{p})V/2). \qquad (1.12)$$

De plus, $\mathrm{var}(f(\mathbf{p}^*))$ et $\mathrm{var}(f(\hat{\mathbf{p}}))$ sont toutes deux égales à

$$v(\mathbf{p})n^{-1} + O(n^{-2}) \text{ où } v(\mathbf{p}) = \dot{f}(\mathbf{p})' V \dot{f}(\mathbf{p}). \qquad (1.13)$$

Ce théorème montre que

(a) l'arrondissement aléatoire n'accroît que de $O(n^{-2})$ la variance de l'estimateur naturel de $f(\mathbf{p})$;

(b) l'arrondissement aléatoire n'a qu'un effet du second ordre sur le biais de l'estimateur naturel de $f(\mathbf{p})$.

Selon (1.12), l'estimateur naturel de $f(\mathbf{p})$, $f(\hat{\mathbf{p}})$, a un biais d'ordre n^{-1}. Nous allons voir maintenant comme réduire ce biais à un ordre n^{-2}.

Corollaire 1.1. Si pour une fonction $f_n(\mathbf{p})$, $E(f_n(\mathbf{p}^*)) = f(\mathbf{p}) + O(n^{-2})$, alors $E(f_n(\hat{\mathbf{p}})) = f(\mathbf{p}) + O(n^{-2})$.

Deux estimateurs possible pour $f_n(\hat{\mathbf{p}})$ dont "l'estimateur delta" pour lequel

$$f_n(\mathbf{p}) = f(\mathbf{p}) - \left\{ \sum_{i=1}^{r} f_{ii}(\mathbf{p})p_i - \mathbf{p}'\ddot{f}(\mathbf{p})\mathbf{p} \right\} / (2n), \tag{1.14}$$

où $f_{ii}(\mathbf{p}) = \partial^2 f(\mathbf{p}) / \partial p_i^2$, et "l'estimateur jack-knife" pour lequel

$$f_n(\mathbf{p}) = nf(\mathbf{p}) - (n-1)\bar{f}, \tag{1.15}$$

où

$$\bar{f} = \Sigma_{i=1}^{r} p_i f([(n\mathbf{p} - e_i)/(n-1)]) + (1 - \Sigma_1^r p_i) f([n\mathbf{p}/(n-1)]),$$

$$e_i = \text{le } i\text{-ième vecteur unitaire dans } R^r,$$

$$\text{et } [x] : R^r \to R^r \text{ est définie par } [x]_i = \begin{cases} 0, & x_i < 0 \\ x_i, & 0 \le x_i \le 1. \\ 1, & x_i > 1 \end{cases}$$

Ces estimateurs ont été calculés dans Withers (1987a et 1987b). En particulier, si $f(\mathbf{p})$ est uniquement une fonction de p_1, par exemple $f(\mathbf{p}) = g(p_1)$, alors $f_n(\mathbf{p}) = g(p_1) - \ddot{g}(p_1)(p_1 - p_1^2)/(2n)$ et $\bar{f} = p_1 g([(np_1 - 1)/(n-1)]) + (1 - p_1) g([np_1/(n-1)])$. Si par exemple, $f(\mathbf{p}) = p_1^2$ nous avons, pour l'estimateur delta, l'équation suivante: $f_n(\mathbf{p}) = p_1^2 \{1 - (1 - p_1)/n\}$.

Nous allons maintenant montrer que si $f(\mathbf{p})$ est une fonction polynomiale, il est effectivement possible de trouver un estimateur de $f(\mathbf{p})$ fondé sur l'estimateur naturel avec un biais apparemment de taille exponentiellement faible. Pour cela, nous prenons le cas $f(\mathbf{p}) = p_1^2$.

Théorème 1.3. $\hat{\lambda}_1 = \{\hat{p}_1^2 - n^{-1}\hat{p}_1 - n^{-2}(l^2 - 1)/6\}(1 - n^{-1})^{-1}$ est un estimateur de $\lambda_1 = p_1^2$ avec un biais $\Delta_n(p_1)(n^2 - n)^{-1}$.

De même si $f_n(\mathbf{p})$ est un moment de $\hat{\mathbf{p}}$ il est possible de trouver un estimateur de $f_n(\mathbf{p})$ avec un biais apparemment de taille exponentiellement faible. À des fins d'illustration, nous prenons le cas $f_n(\mathbf{p}) = \text{var}(\hat{p}_1)$.

Théorème 1.4. $\hat{\lambda}_{2n} = n^{-1}(\hat{p}_1 - \hat{\lambda}_1) - n^{-2}(l^2 - 1)/6$ est un estimateur de $\lambda_{2n} = \text{var}(\hat{p}_1)$ avec un biais $-\Delta_n(p_1)(n^2 - n)^{-1}$.

On peut étendre ces résultats à des moments et à des polynômes d'ordre supérieur au moyen de l'expression définie dans l'annexe B pour les moments et les cumulants de $\hat{\mathbf{p}}$. Nous allons maintenant montrer qu'il existe un estimateur sans biais pour le cas particulier où $f(\mathbf{p})$ est collinéaire.

Théorème 1.5. Soit $f_I(\mathbf{p}) = \Pi_{i=1}^{I} p_i$ où $1 \le I \le R$ et

$$a_{nI} = n^{-I} n!/(n-I)! = (1 - n^{-1})(1 - 2n^{-1})\ldots(1 - \{I - 1\}n^{-1}). \tag{1.16}$$

Alors $$E(f_l(\hat{\mathbf{p}})) = E(f_l(\mathbf{p}^*)) = f_l(\mathbf{p})a_{nl}. \tag{1.17}$$

$f_l(\hat{\mathbf{p}})/a_{nl}$ est donc un estimateur sans biais de $f(\mathbf{p})$.

Corollaire 1.2. $\mathrm{cov}(\hat{p}_1,\hat{p}_2) = -p_1p_2/n$. L'estimateur sans biais correspondant est $-\hat{p}_1\hat{p}_2/(n-1)$. D'une manière plus générale, pour $1 \leq I \leq R$, $\mathrm{E}(\Pi_{i=1}^{I}(\hat{p}_i - p_i)) = c_{nI}\Pi_{i=1}^{I}p_i$ avec l'estimateur sans biais $(\Pi_{i=1}^{I}\hat{p}_i)a_{nI}/c_{nI}$ où $c_{nI} = \Sigma_{j=0}^{I}(-1)^{I-j}\binom{I}{j}a_{nj}$. (On obtient le même résultat si on remplace $\hat{\mathbf{p}}$ par $\mathbf{p}^*$.)

L'équation (1.16) nous permet de calculer des estimateurs sans biais pour d'autres polynômes spéciaux en $\mathbf{p}$ tels que p_1^2, $p_1p_2(p_1 + p_2)$et $\Sigma_{i=1}^{R}p_i^3$ - mais non pour des polynômes de la forme $p_1^2p_2$ ou p_1^3.

Corollaire 1.3. Pour $1 \leq I < R$ un estimateur sans biais de

$$f_I(\mathbf{p}) \sum_{1}^{I} p_i \text{ est } f_I(\hat{\mathbf{p}}^1)\left\{1 - In^{-1} - \sum_{I+1}^{R}\hat{p}_i\right\}/a_{n,I+1}. \tag{1.18}$$

En particulier, $\hat{p}_1(\tilde{p}_1 - n^{-1})(1 - n^{-1})^{-1}$ est un estimateur sans biais de p_1^2. (1.19)

Nous tenons à souligner que les résultats de cette étude reposent sur l'hypothèse que les fréquences de tableaux sont des variables de Poisson indépendantes ou à tout le moins, des variables multinomiales, étant donné le total. Le modèle de Poisson et le modèle multinomial sont tous deux intéressants parce qu'ils sont d'interprétation simple et parce que la somme de variables de Poisson est elle-même une variable de Poisson. Cependant, la somme de variables multinomiales n'est multinomiale que si les probabilités de cellule $\mathbf{p}$ sont les mêmes pour toutes les variables. Cela donne à penser que les modèles multinomiaux peuvent produire des conclusions moins exactes si les populations étudiées sont constituées de deux groupes non-homogènes ou plus.

2. PREUVES

Preuve du théorème 2.1. Posons $r = N_1 \bmod l$. Alors l'équation (1.1) est vraie pour $N = N_1$, $M = M_1$ avec $jl = N - r$ et

$$E(M_1^2 \mid r) = (N_1 - r)^2(1 - r/l) + (N_1 - r + l)^2 r/l = N_1^2 + lr - r^2.$$

Par conséquent,

$$E(\hat{p}_1^2) = E(p_1^{*2}) + n^{-2}A_n(p_1), \tag{2.1}$$

où

$$A_n(p_1) = E(M_1^2 - N_1^2) = E(lr - r^2) = \sum_{i=0}^{l-1}(li - i^2)P(N = i)$$

$$= (l^2 - 1)/6 + \Delta_n(p_1)$$

puisque

$$l^{-1} \sum_{i=0}^{l-1} i(l - i) = (l^2 - 1)/6. \tag{2.2}$$

Or,

$$E(p_1^{*2}) = p_1^2 + (p_1 - p_1^2)n^{-1}, \tag{2.3}$$

ce qui nous amène à l'équation (1.5). Par ailleurs, $\tilde{p}_1 = \hat{p}_1 - \sum (M_j - N_j)/n$,

alors

$$E(\tilde{p}_1^2) = E(\hat{p}_1^2) - 2n^{-2} \sum E(M_1(M_j - N_j)) + n^{-2} \sum E((M_i - N_i)(M_j - N_j))$$

$$= E(\hat{p}_1^2) - 2n^{-2}A_n(p_1) + n^{-2} \sum A_n(p_i)$$

puisque $E(\Pi_i f_i(M_i)|\{N_i\}) = \Pi_i E(f_i(M_i)|N_i)$. (2.4)

Par conséquent $\mathrm{var}(\tilde{p}_1) = (p_1 - p_1^2)n^{-1} + n^{-2}\Sigma_{i \neq 1} A_n(p_i)$ ce qui vérifie l'équation (1.7).

Par ailleurs

$$E(\hat{p}_1\tilde{p}_1) = p_1 - n^{-2} \sum_{i \neq 1} E(M_1M_i) = p_1 - \sum_{i \neq 1} E(p_1^* p_i^*)$$

$$= p_1 - \sum_{i \neq 1} p_1p_i(1 - n^{-1}) = p_1 - p_1(1 - p_1)\ (1 - n^{-1}),$$

alors,

$$\mathrm{cov}(\hat{p}_1, \tilde{p}_1) = (p_1 - p_1^2)n^{-1}. \tag{2.5}$$

Par conséquent, $\mathrm{var}(p_1(\lambda)) = (p_1 - p_1^2)n^{-1} + \{(1 - \lambda)^2 A_n(p_1) + \lambda^2 \Sigma_{i \neq 1} A_n(p_i)\}n^{-2}$ ce qui vérifie (1.8).

Preuve du théorème 1.2. Cette démonstration a été faite pour $\mathbf{p}^*$ dans Withers (1987a). En outre, comme $\ddot{f}$ est finie dans un voisinage de $\mathbf{p}$,

$$f(\hat{\mathbf{p}}) = f(\mathbf{p}^*) + (\hat{\mathbf{p}} - \mathbf{p}^*)'\dot{f}(\mathbf{p}^*) + O(|\hat{\mathbf{p}} - \mathbf{p}^*|^2).$$

$$E((\hat{\mathbf{p}} - \mathbf{p}^*)|N) = 0,\ E((\hat{p}_1 - p_i^*)^2|N) = 2n^{-2}I(N_1 \bmod l \neq 0),$$

où $I(A) = 1$ ou 0 selon que A est vrai ou faux, autrement dit, $I(\cdot)$ est la fonction indicatrice. Par conséquent $\mathrm{E}(f(\hat{\mathbf{p}})) = \mathrm{E}(f(\mathbf{p}^*)) + O(n^{-2})$ et $\mathrm{var}(f(\hat{\mathbf{p}})) = \mathrm{var}(f(\mathbf{p}^*)) + O(n^{-2})$.

Preuve du théorème 1.3. Découle directement de (2.1) et de (2.3).

Preuve du théorème 1.4. Découle de (2.1) et de (1.5).

Preuve du théorème 1.5. La première équation de l'expression (1.16) découle de (2.4) tandis que la seconde découle du théorème multinomial. Le corollaire 1.2 s'ensuit automatiquement.

Preuve du corollaire 1.3. Selon (1.16), pour $1 \leq I < i \leq R$ nous avons

$$E(f_I(\hat{\mathbf{p}})\hat{p}_i) = f_I(\mathbf{p})p_i a_{n,I+1}$$

alors

$$E(f_I(\hat{\mathbf{p}}) \sum_{I+1}^{R} \hat{p}_i / a_{n,I+1}) = f_I(\mathbf{p})(1 - \Sigma_1^I p_i)$$

$$= E(f_I(\hat{\mathbf{p}}) / a_{nI}) - f_I(\mathbf{p}) \Sigma_1^I p_i.$$

REMERCIEMENTS

Je tiens à remercier Peter McGavin pour les calculs de l'annexe A.

ANNEXE A

Pour une fonction lisse f on devrait avoir:

$$E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \ldots, N_s = j_s \bmod l)) \to f(\mathbf{p})l^{-s} \tag{A.1}$$

lorsque $n \to \infty$ pourvu que $0 < p_i < 1$ pour $1 \leq i \leq s \leq R$.

Si $E(f(\hat{\mathbf{p}})) = f(\mathbf{p})$, on devrait observer une vitesse de convergence exponentielle, c'est-à-dire $O(e^{-\lambda n})$ pour certaine valeur de $\lambda > 0$. Si $f(\hat{\mathbf{p}})$ est biaisé, le biais est alors $O(n^{-1})$, et cette valeur devrait représenter une vitesse de convergence pour (A.1). En règle générale, la convergence cesse lorsque $\mathbf{p}$ approche les limites de $[0,1]^r$, puisque

$$E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \ldots, N_s = j_s \bmod l))$$

$$= \begin{cases} f(\mathbf{p})I(j_1 = j_2 = \ldots = j_s = 0) & \text{si } \mathbf{p} = 0 \\ f(\mathbf{p})I(j_1 = n \bmod l) & \text{si } p_1 = 1. \end{cases}$$

Pour vérifier ces hypothèses, nous avons considéré le cas où $s = 1$, $l = 3$, $j = 0$ et les fonctions (a) $f(\mathbf{p}) = 1$, (b) $f(\mathbf{p}) = p_1$, et (c) $f(\mathbf{p}) = \exp(p_1)$. Les calculs ont été effectués à quadruple précision sur un VAX11/780, ce qui a permis de calculer

$$\Delta = E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \ldots, N_s = j_s \bmod l)) - f(\mathbf{p})l^{-s}$$

à une précision de 112 bits, soit près de 34 décimales. Les figures 1a, 1b et 1c décrivent la relation entre Δ et p_1 pour $n = 6, 18, 54$. Puisque $n \bmod 3 = 0$, Δ est symétrique autour de $p_1 = 1/2$ pour (a).

Insert Figure 1 (a), 1(b)

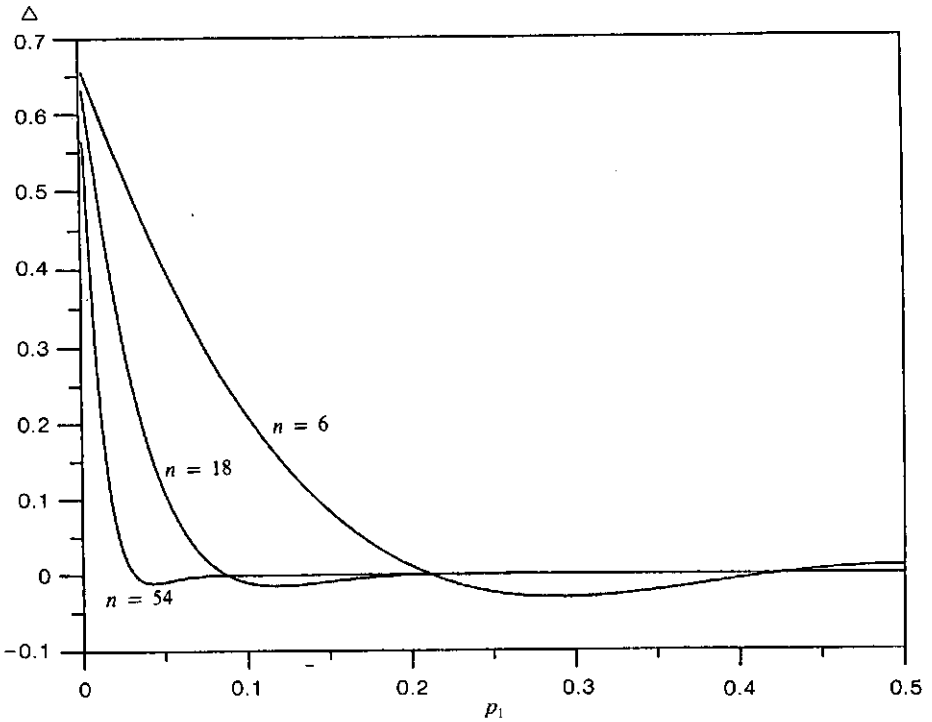

Figure 1a. Confirmation de l'hypothèse (A.1) lorsque $f(\mathbf{p}) = 1$.

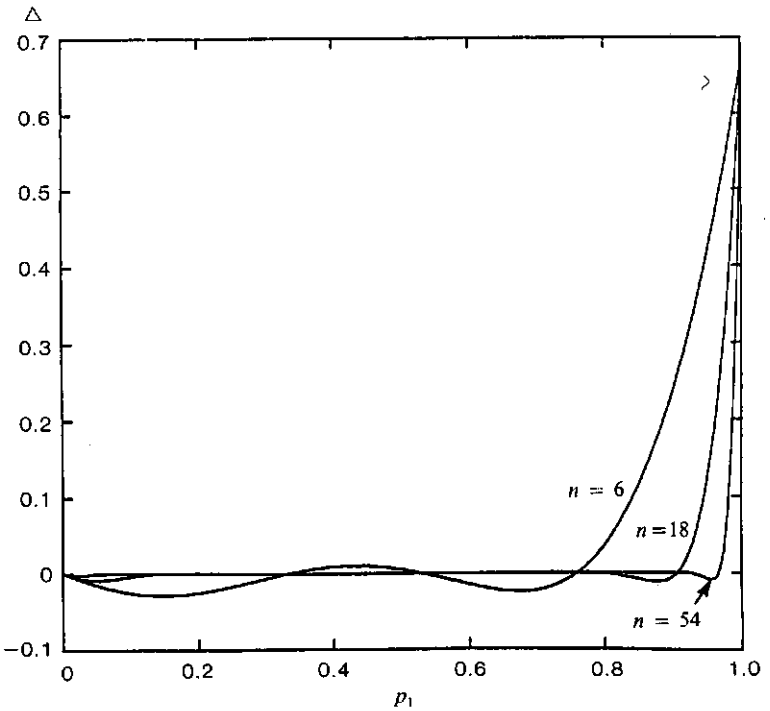

Figure 1b. Confirmation de l'hypothèse (A.1) lorsque $f(\mathbf{p}) = p_1$.

Insert Figure 1(c), 2(a)

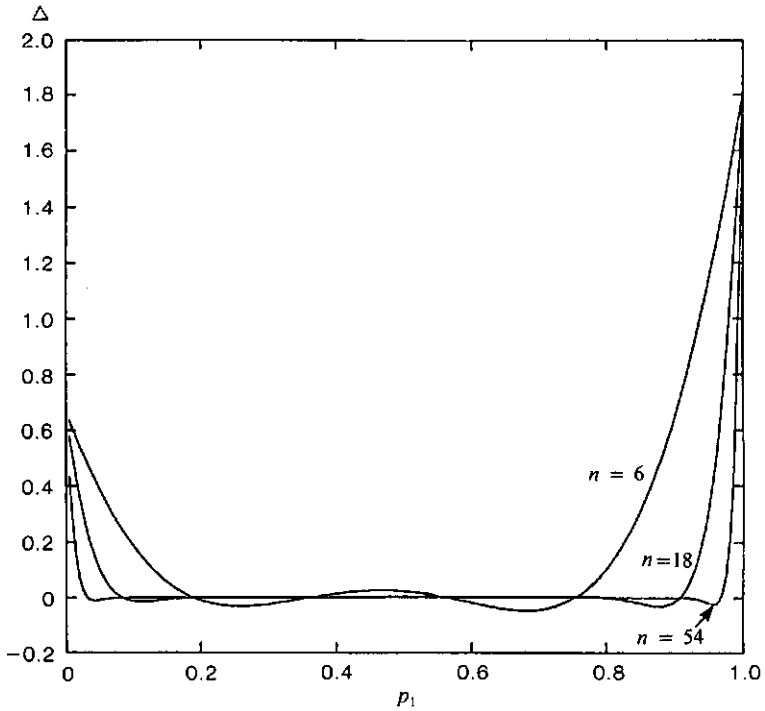

Figure 1c. Confirmation de l'hypothèse (A.1) lorsque $f(\mathbf{p}) = \exp(p_1)$.

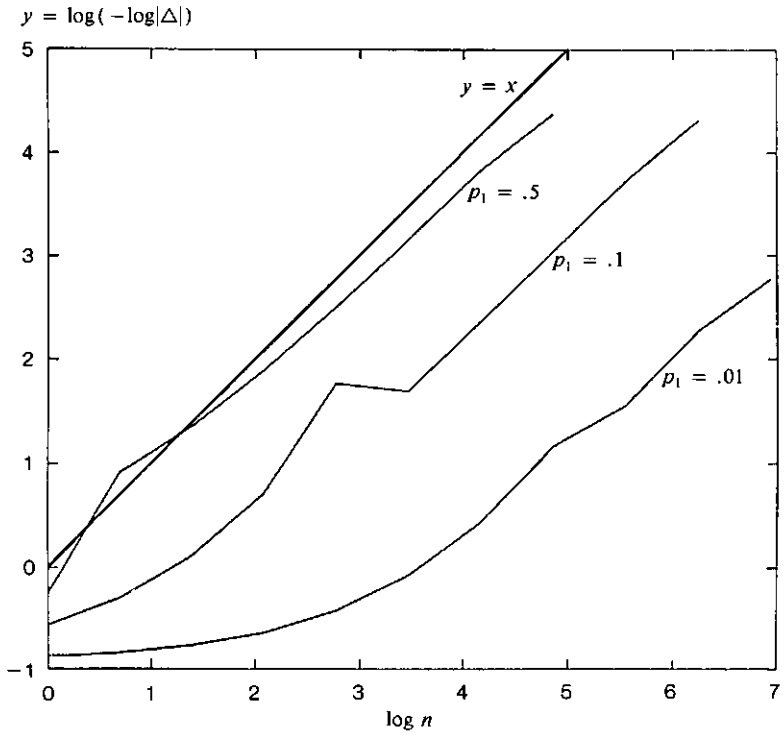

Figure 2a. Confirmation de la convergence exponentielle en (A.1) lorsque $f(\mathbf{p}) = 1$.

insert Figure 2 (b), 3

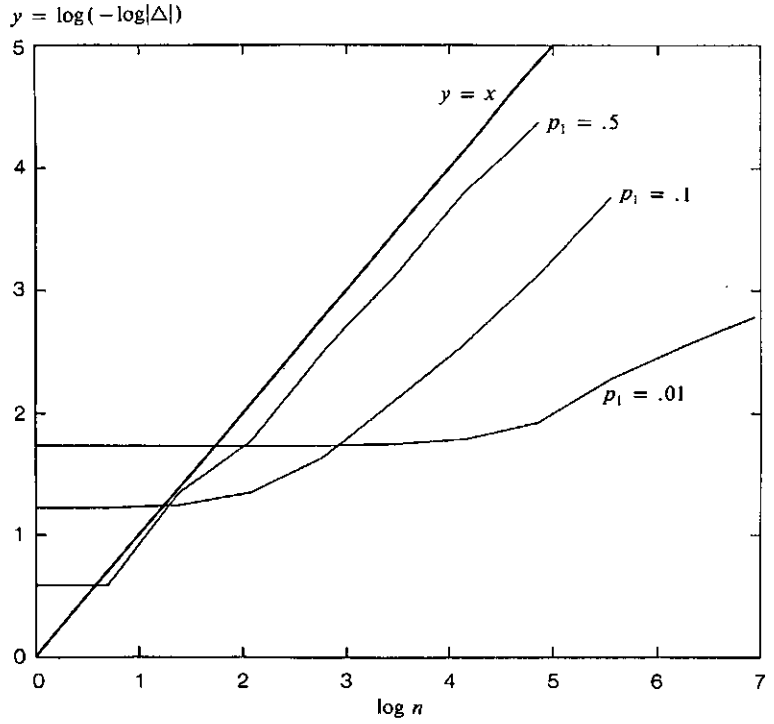

Figure 2b. Confirmation de la convergence exponentielle en (A.1) lorsque $f(\mathbf{p}) = p_1$.

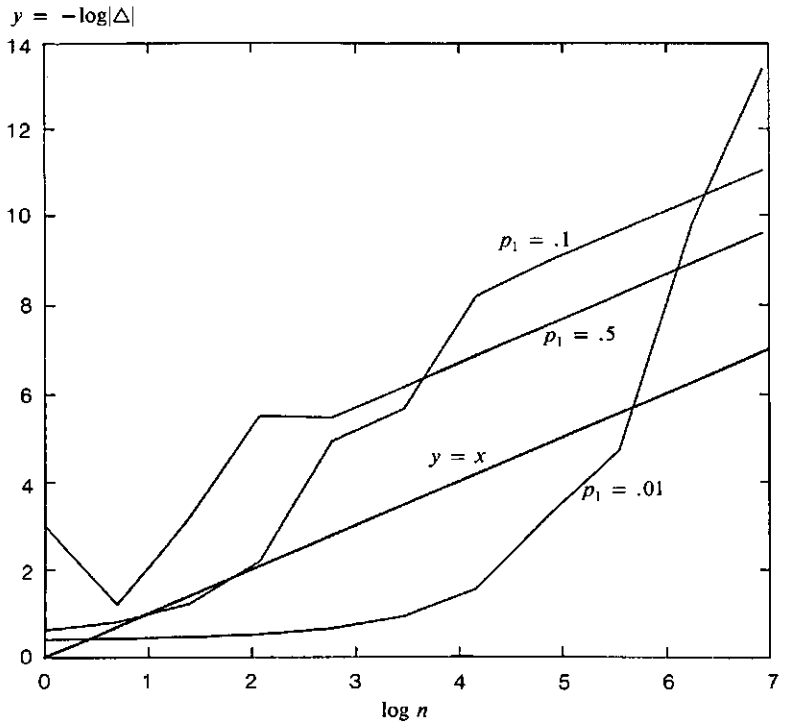

Figure 3. Confirmation de la convergence en (A.1) à un rythme de $\sim n^{-1}$ lorsque $f(\mathbf{p}) = \exp(p_1)$.

Puisque $\Delta = 2/3f(0)$ à $p_1 = 0$, et est égale à 2/3, 0 et 2/3 pour (a), (b) et (c) respectivement. La convergence cesse à $p_1 = 0$ pour (a) et (c), mais non pour (b). Pour $n = 18$, Δ est déjà à peu près nul pour les valeur de p_1 situées dans l'intervalle (.2, .8) pour la fonction (a) et pour les valeurs de p_1 situées dans l'intervalle (.1, .8) pour les (b) et (c). Pour $n = 54$, ces intervalles deviennent (.1, .9) pour (a), (.02, .95) pour (b), et (.07, .95) pour (c).

Les figures 2a et 2b décrivent la relation entre $Y = \log(-\log|\Delta|)$ et $X = \log(n)$ pour (a) $f(\mathbf{p}) = 1$ et (b) $f(\mathbf{p}) = p_1$. Comme prévu abstraction faite des petites valeurs de n, les courbes sont à peu près parallèles à la courbe $Y = X$ (sauf en ce qui concerne la fonction (b) pour $p = .01$), ce qui est conforme à la relation $\Delta = O(e^{-\lambda n})$ pour certaine valeur de $\lambda > 0$. Les courbes ne sont pas lisses parce que Δ n'a été calculé qu'en fonction de n puissance de $2(n = 2^i$ pour $0 \leq i \leq 7)$.

La figure 3 décrit la relation entre $Y = -\log|\Delta|$ et $X = \log(n)$ pour (c) $f(\mathbf{p}) = \exp(p_1)$. Pour les valeurs supérieures de n, les courbes sont parallèles à la courbe $Y = X$ pour $p_1 = .5$ et .1, ce qui est conforme à $\Delta = O(n^{-1})$; pour $p_1 = 0.1$ toutefois, la courbe ne reflète pas du tout une relation linéaire; le rythme d'accroissement est beaucoup plus prononcé. Les graphiques confirment de façon générale nos hypothèses sur la vitesse de convergence de (A.1). Si nous voulions vérifier ces hypothèses par des méthodes analytiques, il nous faudrait recourir à la théorie des nombres.

ANNEXE B

Dans cette annexe, nous comparons les moments et les cumulants de $\mathbf{p}^* = N/n$ et $\hat{\mathbf{p}} = M/n$. Posons $q_1 = 1 - p_1$, $n_i = N_i \bmod l$, et $m_i(j) = E(p_1^{*i} I(n_1 = j)) - p_1^i/l$ lorsque $n \to \infty$, en supposant que $p_1 \neq 0$ or 1.
Par des calculs élémentaires, nous obtenons

$$\mu(\hat{\mathbf{p}}) = \mu(\mathbf{p}^*) = \mathbf{p},$$

$$\mu_2(\hat{p}_1) = \mu_2(p_1^*) + M_{22}n^{-2} = p_1q_1n^{-1} + O(n^{-2}),$$

où

$$M_{22} = A_n(p_1) = \sum_{i=0}^{l-1} i(l-i)m_0(i) - (l^2-1)/6$$

lorsque $n \to \infty$,

$$\mu_3(\hat{p}_1) = \mu_3(p_1^*) + 3n^{-2}\sum_{j=0}^{l-1}(lj - j^2)\{m_2(j) - 2p_1m_1(j) + p_1^2m_0(j)\}$$

$$+ n^{-3}\sum_{j=0}^{l-1} a_{jl}m_0(j)$$

$$= \mu_3(p_1^*) + o(n^{-2}) = p_1q_1(1-2p_1)n^{-2} + o(n^{-2}),$$

et

$$a_{jl} = -j^3(1 - j/l) + (l-j)^3 j/l.$$

De même $\mu_4(\hat{p}_1)$ peut s'écrire $\mu_4(p_1^*) + \Sigma_2^4 M_{4i}n^{-i} = O(n^{-2})$ et $\kappa_4(\hat{p}_1)$ peut s'écrire $\Sigma_2^4 k_{4i}n^{-i}$ où $k_{42} = M_{42}$ ne tend pas vers 0 lorsque $n \to \infty$. D'où $\kappa_4(\hat{p}_1) \sim n^{-2}$, et non

n^{-3}. Par conséquent, $\hat{\mathbf{p}}$ ne satisfait pas à l'hypothèse de Cornish-Fisher selon laquelle $\kappa_r(\hat{\mathbf{p}}) = O(n^{1-r})$ pour $r \geq 1$: voir par exemple Kendall et Stuart (1977).

On peut aussi déterminer les moments et les cumulants à l'aide de la f.g.m. (fonction génératrice des moments), que nous définissons ci-sessous.

$$E(\exp(t_1 M_1/n) \mid N_1) = \exp(t_1 N_1/n) S(t_1, n_1),$$

où

$$S(t_1, n_1) = (1 - n_1/l)\exp(-n_1 t_1/n) + (n_1/l)\exp(l - n_1)t_1/n.$$

Par (2.4), la f.g.m. est

$$E(\exp(t'\hat{\mathbf{p}})) = E(\exp(t'\mathbf{N}/n))S(t) \text{ où } S(t) = \Pi_1^l S(t_i, n_i).$$

De plus, à $t = 0$, $S_1 = 0$ et $S_{ij\ldots} = 0$ si l'indice inférieur ne se repète pas. Par exemple, si nous posons

$$S = S(t),\ \partial_i = \partial/\partial t_i,\ S_i = \partial_i S,\ S_{ij} = \partial_i \partial_j S,\ \ldots$$

nous obtenons

$$E(\hat{p}_1^2 \exp(t'\hat{\mathbf{p}})) = E(\exp(t'\mathbf{N}/n)\{p_1^{*2} S + 2p_1^* S_1 + S_{11}\}),$$

$$E(\hat{p}_1^2 \hat{p}_2^2 \exp(t'\hat{\mathbf{p}})) = E(\exp(t'\mathbf{N}/n)\{p_1^{*2}(p_2^{*2} S + 2p_2^* S_2 + S_{22}) +$$

$$2p_1^*(p_2^{*2} S_1 + 2p_2^* S_{12} + S_{122}) + (p_2^{*2} S_{11} + 2p_2^* S_{112} + S_{1122})\}).$$

Ainsi $E(\hat{p}_1^2) = E\{p_1^{*2} + S_{11}(0)\}$ et

$$E(\hat{p}_1^2 \hat{p}_2^2) = E\{p_1^{*2} p_2^{*2} + p_1^{*2} S_{22}(0) + p_2^{*2} S_{11}(0) + S_{1122}(0)\}.$$

où $S_{ii}(0) = S_{11}(0, n_i) = n^{-2}(l - n_i)n_i = n^{-2}\Sigma_{k=0}^{l-1}(l - k)kI(n_i = k)$ et $S_{1122}(0) = S_{11}(0)\, S_{22}(0)$. Nous pouvons simplifier davantage en utilisant $N_2 \mid N_1 \sim Bi(\theta, n - N_1)$ où $\theta = p_2/(1 - p_1)$. La f.g.m. multinomiale donne

$$E(p_1^{*2} p_2^{*2}) = n^{-4} p_1 p_2 \{(n)_4 p_1 p_2 + (n)_3(p_1 + p_2) + (n)_2\}$$

où $(n)_i = n!/(n - i)! = n(n - i)\ldots(n - i + 1)$.

BIBLIOGRAPHIE

GASTWIRTH, J.L., KRIEGE, A.M., et RUBIN, D.B. (1978). Statistical analyses from summary data and their impact on the issue of confidentiality. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 183-188.

KENDALL, M.G., et STUART, A. (1977). *The Advanced Theory of Statistics, Volume* 7. London: Griffin.

NARGUNDAR, M.S., et SAVELAND, W. (1972). Random rounding to prevent statistical disclosures. *Proceedings of the Social Statistics Section, American Statistical Association,* 382-385.

PENNY, R., et RYAN, M. (1986). A problem associated with random-rounding. *New Zealand Statistician,* 21, 43-52.

WITHERS, C.S. (1987a). Bias reduction by Taylor series. *Communications in Statistics - Theory and Methods* (en voie de rédaction).

WITHERS, C.S. (1987b). Jackknifing binomials and multinomials. Document non-publié, Department of Scientific and Industrial Research.

Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 157–172
Statistique Canada

Estimation de la variance pour l'enquête sur la population active du Canada

G.H. CHOUDHRY et H. LEE[1]

RÉSUMÉ

Au moyen d'une étude de Monte Carlo, les auteurs évaluent le biais et la stabilité de divers estimateurs de la variance pour le plan de sondage à deux degrés fondé sur la méthode des groupes aléatoires (Rao et coll. 1962) dans le contexte de l'enquête sur la population active du Canada. Ils se servent de la méthode de linéarisation de Taylor pour déterminer la formule de la variance se rapportant à la technique d'estimation itérative par le quotient. Enfin, ils analysent les propriétés de cette formule par une simulation de Monte Carlo.

MOTS CLÉS: Estimateur de la variance de Keyfitz; estimateur itératif par le quotient; linéarisation de Taylor; simulation de Monte Carlo.

1. INTRODUCTION

L'enquête sur la population active du Canada (EPA), la plus vaste enquête mensuelle sur les ménages menée par Statistique Canada, sert à produire des estimations de diverses caractéristiques de la population active aux niveaux national, provincial et infraprovincial. Elle repose sur un plan de sondage stratifié à plusieurs degrés avec six groupes de renouvellement (Platek et Singh 1976).

L'échantillon de l'EPA est remanié après chaque recensement décennal de la population. Dans le cadre du remaniement entamé après le recensement de 1981, un programme de recherche intensif a été mis en oeuvre pour examiner diverses méthodes d'échantillonnage, d'estimation et de collecte des données (Singh et Drew 1981). La méthode d'estimation par le quotient pour domaines post-stratifiés, utilisée dans l'ancien plan, a fait place à la technique d'estimation itérative par le quotient afin d'accroître la fiabilité des données infraprovinciales. Dans cet article, nous nous intéressons plus particulièrement aux méthodes d'estimation de la variance.

La méthode d'estimation de la variance utilisée dans l'ancienne version de l'EPA reposait sur la généralisation de la méthode de Keyfitz (Keyfitz 1957) proposée par Woodruff (Woodruff 1971), selon laquelle les estimations par le quotient pour les domaines post-stratifiés étaient soumises à une linéarisation de Taylor (Platek et Singh 1976). Nous désignerons cette méthode comme la méthode de Keyfitz (voir Platek et Singh 1976).

Le plan de sondage de l'EPA renferme trois genres de secteurs, soit les secteurs auto-représentatifs (AR), qui sont composés des grandes villes, les secteurs non auto-représentatifs (NAR), qui correspondent aux petits centres urbains et aux régions rurales, et les secteurs spéciaux, qui comprennent les établissements militaires, les institutions et les régions éloignées. En ce qui concerne les secteurs NAR et les secteurs spéciaux, nous avons choisi d'appliquer une version modifiée de la méthode de Keyfitz, qui fait intervenir l'estimation itérative par le quotient.

Par ailleurs, pour ce qui est des secteurs AR, nous avons analysé deux estimateurs de la variance définis respectivement par Rao, Hartley et Cochran (1962) et par Rao (1975) pour le plan de sondage à deux degrés fondé sur la méthode des groupes aléatoires et les avons

[1] G.H. Choudhry et H. Lee, Division des méthodes d'enquêtes sociales, Statistique Canada, 4-ième étage, Immeuble Jean-Talon, Parc Tunney, Ottawa (Ontario), K1A 0T6.

comparés aux estimateurs de la méthode de Keyfitz à l'aide d'une simulation de Monte Carlo. Nous avons comparé le biais et la stabilité de ces divers estimateurs dans deux situations différentes: avec et sans correction par le quotient. Nous avons aussi étudié l'effet d'un accroissement du nombre d'échantillons répétés sur l'estimateur de la variance de Keyfitz. La section suivante donne un compte rendu de ces analyses. Compte tenu des résultats de l'analyse, nous avons étendu la méthode de Keyfitz aux secteurs AR.

Dans la section 3, nous établissons la formule de la variance de Keyfitz pour les estimations itératives par le quotient pour tous les genres de secteurs de l'EPA et analysons cette formule par une étude de Monte Carlo. Enfin, dans la section 4, nous tirons les conclusions qui s'imposent.

2. ESTIMATION DE LA VARIANCE POUR LE PLAN DE SONDAGE DES SECTEURS AR

2.1 Plan de sondage des secteurs AR

Pour les secteurs AR, le plan de sondage de l'EPA est un plan à deux degrés fondé sur la méthode des groupes aléatoires (Rao et coll. 1962) avec probabilité de sélection des unités primaires (UPE) proportionnelle à la taille (PPT) et échantillonnage systématique des logements au second degré de telle sorte que le plan devient un plan à auto-pondération. Supposons qu'une strate donnée contient N UPE et définissons respectivement x_j et M_j, $j = 1, 2, \ldots, N$, comme la taille de la j-ième UPE de la strate et son nombre de logements. Définissons $1/W$ comme le taux de sondage dans la strate, W étant un nombre entier, et n comme le nombre d'UPE qui doivent être prélevées dans la strate. Les N UPE contenues dans la strate sont réparties aléatoirement en n groupes de sorte que le i-ième groupe aléatoire contient N_i UPE et $\Sigma_{i=1}^{n} N_i = N$.

Définissons

$$p_j = \frac{x_j}{\sum_{t=1}^{N} x_t}, \quad j = 1, 2, \ldots, N,$$

et

$$\begin{aligned} \delta_{ij} &= 1 \text{ si la } j\text{-ième UPE est incluse dans le } i\text{-ième groupe} \\ &= 0 \text{ dans le cas contraire.} \end{aligned}$$

Alors, $\pi_i = \Sigma_{j=1}^{N} \delta_{ij} p_j$ est la taille relative du i-ième groupe aléatoire.

Définissons maintenant W_{ij}, l'intervalle d'échantillonnage pour l'échantillonnage systématique. Posons $a_{ij} = \delta_{ij} W p_j / \pi_i$ et $r_{ij} = a_{ij} - [a_{ij}]$ où $[a]$ est le plus grand nombre entier égal ou inférieur à a. Sans limiter la généralité de ce qui précède, nous pouvons supposer que les éléments de l'ensemble $\{r_{ij}, j = 1, 2, \ldots, N\}$ sont exprimés par ordre décroissant. Alors, W_{ij} est défini par

$$\begin{aligned} W_{ij} &= [a_{ij}] + 1, \quad j = 1, 2, \ldots, R \\ &= [a_{ij}], \quad j = R + 1, \ldots, N \end{aligned}$$

où $R = \Sigma_{j=1}^{N} r_{ij}$. Puis, par définition, $\Sigma_{j=1}^{N} W_{ij} = W$ pour le i-ième groupe aléatoire, $i = 1, 2, \ldots, n$.

Comme W_{ij} est l'intervalle d'échantillonnage défini pour l'échantillonnage systématique de logements dans la grappe prélevée dans le i-ième groupe aléatoire, on le définit comme un nombre entier pour simplifier les opérations.

Une UPE est prélevée avec probabilité proportionnelle à W_{ij} dans chacun des n groupes aléatoires. L'UPE j prélevée dans le i-ième groupe aléatoire est sous-échantillonnée systématiquement suivant un taux de sondage $1 / W_{ij}$. Donc, le taux de sondage global dans chacun des n groupes aléatoires est $1 / W$, de sorte que le plan devient un plan à auto-pondération avec un poids de base de W. On attribue à chaque groupe aléatoire un numéro de renouvellement de 1 à 6. Le nombre de groupes aléatoires n est habituellement un multiple de six de sorte qu'il y a le même nombre de groupes aléatoires pour chaque groupe de renouvellement.

Comme une seule UPE est échantillonnée dans chaque groupe aléatoire, nous désignons par $1 / W_i$ le taux de sous-échantillonnage dans l'UPE prélevée dans le i-ième groupe aléatoire et par m_i le nombre de logements échantillonnés du groupe aléatoire i.

2.2 Estimateurs de la variance

Supposons que nous voulons connaître le total d'une caractéristique y pour la strate. Soit y_{jk} la valeur de la caractéristique pour le k-ième logement de la j-ième UPE, où $k = 1, 2, \ldots, M_j$. Nous pouvons alors estimer le total $Y = \Sigma_{j=1}^{N} \Sigma_{k=1}^{M_j} y_{jk}$ par $\hat{Y} = W \Sigma_{i=1}^{n} y_i$, où y_i est la somme des valeurs de la caratéristique y pour les m_i logements échantillonnés dans l'UPE qui a elle-même été prélevée dans le i-ième groupe, $i = 1, 2, \ldots, n$. Nous considérons ci-dessous divers estimateurs de la variance du total estimé $\hat{Y}$:

(1) Estimateur de la variance de Keyfitz (1957)

Cet estimateur était utilisé dans l'ancien plan de sondage avec deux pseudo-échantillons répétés qui étaient formés, dans un cas, des groupes de renouvellement à numéro impair et dans l'autre cas, des groupes à numéro pair. Si on ne tient pas compte de la correction pour population finie (CPF), la formule de la variance est

$$\hat{V}_1(\hat{Y}) = W^2 \left(\sum_o y_i - \sum_e y_i \right)^2, \tag{2.1}$$

où $\sum_o$ désigne la sommation pour tous les groupes à numéro impair et $\sum_e$ désigne la sommation pour tous les groupes à numéro pair. Par ailleurs, on peut utiliser l'estimateur de la variance Keyfitz généralisé pour $n(\geq 2)$ échantillons répétés, qui est défini par

$$\hat{V}_2(\hat{Y}) = W^2 \frac{n}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2, \tag{2.2}$$

où $\bar{y} = (1/n) \Sigma_{i=1}^{n} y_i$. Dans ce cas, chaque UPE ou groupe de renouvellement est considéré comme un échantillon répété. On s'est intéressé à l'estimateur $\hat{V}_2$ parce qu'on croyait qu'il pouvait être plus efficace (stable) que $\hat{V}_1$ en raison du plus grand nombre de degrés de liberté.

(2) Estimateur de la variance de Rao, Hartley et Cochran (1962)

La formule de cet estimateur repose sur l'hypothèse que le nombre d'unités secondaires m_i qui doivent être sélectionnées dans le i-ième groupe est fixe pour $i = 1, 2, \ldots, n$ et que ces unités sont sélectionnées suivant un échantillonnage aléatoire simple (EAS). L'estimateur de la variance est défini par

$$\hat{V}_3(\hat{Y}) = A \sum_1^n \pi_i \left(\frac{M_i y_i}{m_i p_i} - \hat{Y}\right)^2 + \sum_1^n \frac{\pi_i}{p_i} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2, \tag{2.3}$$

où

$$A = \frac{\sum_1^n N_i^2 - N}{N^2 - \sum_1^n N_i^2}, \tag{2.4}$$

$$s_i^2 = \frac{1}{m_i - 1} \sum_{k=1}^{m_i} (y_{ik} - \bar{y}_i)^2. \tag{2.5}$$

M_i est le nombre de logements contenus dans l'UPE sélectionnée dans le i-ième groupe et parmi lesquels m_i logements sont sélectionnés suivant un échantillonnage systématique. Toutefois, l'estimation de la variance est calculée suivant l'hypothèse d'un EAS. La valeur de y pour le k-ième logement tiré de l'UPE qui a elle-même été sélectionnée dans le i-ième groupe est y_{ik} et $\bar{y}_i = y_i / m_i$.

Comme $\pi_i / p_i = W / W_i$ et $M_i / m_i = W_i$, (ces égalités ne sont pas rigoureuses en raison de l'utilisation de nombres entiers pour W_i), la formule (2.3) peut être exprimée comme étant

$$\hat{V}_3(\hat{Y}) = A \sum_1^n \pi_i \left(W \frac{y_i}{\pi_i} - \hat{Y}\right)^2 + W \sum_1^n \left(1 - \frac{m_i}{M_i}\right) M_i s_i^2. \tag{2.6}$$

(3) Estimateur de la variance de Rao (1975)

En ce qui concerne cet estimateur, on suppose que m_i unités secondaires sont prélevées suivant un EAS; toutefois, comme il s'agit d'un plan à auto-pondération, la taille m_i de l'échantillon au second degré est considérée comme une variable aléatoire. L'estimateur de la variance est défini par

$$\hat{V}_4(\hat{Y}) = A \sum_1^n \pi_i \left(W \frac{y_i}{\pi_i} - \hat{Y}\right)^2 + \sum_1^n \left\{\frac{\pi_i^2}{p_i^2} - A\left(\frac{\pi_i}{p_i^2} - \frac{\pi_i^2}{p_i^2}\right)\right\} \frac{M_i^2 s_i^2}{m_i} - \sum_1^n \frac{\pi_i}{p_i} M_i s_i^2. \tag{2.7}$$

où A est défini en (2.4) et s_i^2 est défini en (2.5). Après quelques simplifications, nous pouvons réécrire (2.7) de la façon suivante:

$$\hat{V}_4(\hat{Y}) = \hat{V}_3(\hat{Y}) + W^2 \sum_1^n m_i s_i^2 \left\{\left(1 - \frac{W_i}{W}\right) - A\left(\frac{1}{\pi_i} - 1\right)\right\}. \tag{2.8}$$

Nous constatons que la formule de la variance comporte un terme additionnel, qui peut être positif ou négatif, lorsqu'on suppose que la taille de l'échantillon au second degré est aléatoire.

Tableau 1

Strates utilisées pour l'étude de Monte Carlo

Strate	Nombre de logements	Nombre d'UPE	Nombre d'UPE sélectionnés	Taille d'échantillon espérées
1	737	49	6	29.5
2	490	33	4	19.6
3	745	45	6	29.8
4	720	34	6	28.8
5	621	37	6	24.8
6	630	38	6	25.2
7	503	31	4	20.1
8	340	23	4	13.6
9	472	33	4	18.9
10	468	33	4	18.7
11	367	28	4	14.7
12	390	23	4	15.6
13	626	36	6	25.0
14	650	39	6	26.0
15	350	22	4	14.0
16	736	46	6	29.4
17	573	35	6	22.9
18	773	48	6	30.9
19	866	64	8	34.6
Total	11,057	697	100	442.3

2.3 Étude de Monte Carlo

Afin d'évaluer le biais des quatre estimateurs définis ci-dessus et leur stabilité relative, nous avons effectué une étude de Monte Carlo avec 19 strates de l'enquête sur la population active de la région métropolitaine de recensement (RMR) de Halifax en nous servant de données du recensement de 1981. Pour les besoins de cette étude, nous nous sommes servis des données de l'échantillon de recensement auquel était destiné le questionnaire complet, en l'occurrence un échantillon systématique de 20% sélectionné dans les secteurs de dénombrement. Nous avons fixé le taux de sondage, $1/W$, à 0.04 pour que la taille espérée de l'échantillon soit la même que dans la version remaniée de l'EPA. Nous avons déterminé un nombre pair de groupes aléatoires dans chaque strate de manière que la taille espérée de l'échantillon dans ces groupes soit le plus près possible de 4.5 pour assurer la conformité avec le plan actuel de l'EPA. Le tableau 1 donne les 19 strates choisies pour l'étude ainsi que le nombre d'UPE, le nombre d'UPE sélectionnés, le nombre de logements, la taille espérée de l'échantillon et les totaux correspondants pour les 19 strates. Mille échantillons ont été produits individuellement dans chacune des 19 strates par la méthode de Monte Carlo; on s'est servi à cette fin du plan de sondage décrit dans la sous-section 2.1 (plan fondé sur la méthode des groupes aléatoires).

Soit $\hat{Y}_{ht}$ l'estimation du total Y_h pour la strate h, tirée du t-ième échantillon de Monte Carlo, $h=1, 2, \ldots, 19$, et $t=1, 2, \ldots, 1{,}000$. De même, $\hat{V}_{jht}$, $j=1, 2, 3, 4$ désigne tour à tour les quatre estimateurs de la variance de $\hat{Y}_{ht}$.

Maintenant, définissons

$$Y = \sum_{h=1}^{19} Y_h,$$

$$\hat{Y}_t = \sum_{h=1}^{19} \hat{Y}_{ht},$$

$$\hat{V}_{jt} = \sum_{h=1}^{19} \hat{V}_{jht}, \quad j = 1, 2, 3, 4,$$

où $t = 1, 2, \ldots, 1000$.

$\hat{Y}_t$ est l'estimation du total Y établie à partir du t-ième échantillon de Monte Carlo et $\hat{V}_{jt}$, $j = 1, 2, 3, 4$ désigne tour à tour les quatre estimateurs de la variance correspondants.

L'espérance mathématique et la variance de Monte Carlo, désignées respectivement par E^* et V^*, sont définies ainsi pour T échantillons de Monte Carlo:

$$E^*(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \hat{\theta}_t,$$

$$V^*(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} [\hat{\theta}_t - E^*(\hat{\theta})]^2,$$

où $\hat{\theta}$ est un estimateur du paramètre inconnu θ et $\hat{\theta}_t$ est l'estimation tirée du t-ième échantillon. De ces définitions nous déduisons la variance de Monte Carlo de l'estimateur $\hat{Y}$, $V^*(\hat{Y})$, de même que l'espérance mathématique et la variance de Monte Carlo de l'estimateur de la variance $\hat{V}_j$, soit $E^*(\hat{V}_j)$ et $V^*(\hat{V}_j)$ respectivement, pour $j = 1, 2, 3, 4$.

Définissons maintenant le biais de l'estimateur de la variance $\hat{V}_j$ par l'expression suivante:

$$B_j = E^*(\hat{V}_j) - V^*(\hat{Y}),$$

et le biais en pourcentage par:

$$PB_j = 100 \frac{B_j}{V^*(\hat{Y})}, \quad j = 1, 2, 3, 4.$$

Alors, l'erreur quadratique moyenne (EQM) de $\hat{V}_j$ est définie par:

$$MSE_j = V^*(\hat{V}_j) + B_j^2, \quad j = 1, 2, 3, 4.$$

Nous définissons l'efficacité de $\hat{V}_j$ par rapport à l'estimateur de la variance de Keyfitz avec deux échantillons répétés (c'est-à-dire, $\hat{V}_1$) par l'expression suivante:

$$\text{Eff. Rel.}(\hat{V}_j \text{ vs. } \hat{V}_1) = (EQM_1 / EQM_j)^{1/2}, \quad j = 2, 3, 4.$$

Dans cette étude, nous considérons trois caractéristiques de la population active: personnes occupées, personnes en chômage et personnes actives. Les tableaux 2A et 3A donnent respectivement les biais et les efficacités relatives des estimateurs de la variance pour les trois caractéristiques. En ce qui concerne le biais, nous constatons que l'estimateur 1 ressemble à l'estimateur 2 tandis que l'estimateur 3 ressemble à l'estimateur 4. Les estimateurs 1 et 2 ont des biais positifs très élevés, notamment en ce qui concerne les personnes occupées et les personnes actives, tandis que les estimateurs 3 et 4 ont des biais relativement faibles. Au point de vue de l'efficacité, les estimateurs 3 et 4 s'équivalent et sont très supérieurs aux estimateurs 1 et 2. De plus, l'estimateur 2 est plus efficace que l'estimateur 1.

Nous avons aussi évalué ces estimateurs pour des estimations par le quotient basé sur la population de l'ensemble des strates. Par ailleurs, nous avons calculé de tels estimateurs, désignés par $\hat{V}_j^{(R)}$, $j = 1, 2, 3, 4$, pour chaque échantillon de Monte Carlo en appliquant la méthode de linéarisation de Taylor, puis nous avons déterminé le biais (en pourcentage) de ces quatre estimateurs (Tableau 2B) et l'efficacité des trois derniers par rapport au premier (Tableau 3B).

Nous constatons que les estimateurs 1 et 2 ont un biais beaucoup moins élevé pour des estimations par le quotient, surtout en ce qui a trait aux personnes occupées et aux personnes actives. Les estimateurs 3 et 4 ont aussi un biais moindre en ce qui concerne ces deux caractéristiques mais un biais à peu près inchangé pour ce qui est des personnes en chômage. Bien que les biais des quatre estimateurs soient faibles, seul le biais de l'estimateur 3 pour les personnes actives s'est révélé non significatif à un seuil de 5%. En ce qui concerne les écarts observés entre les biais, seuls ceux se rapportant aux estimateurs 1 et 2 se sont avérés non significatifs à un seuil de 5% pour les trois caractéristiques.

Par ailleurs, nous nous sommes servis des quatre estimateurs étudiés pour déterminer les intervalles de confiance (IC) de 95% pour les estimations par le quotient de chaque échantillon de Monte Carlo. Le taux de couverture des intervalles correspondait à la proportion des IC qui renfermaient la valeur réelle du total de la caractéristique. Les résultats figurent dans le Tableau 4 et montrent que les quatre estimateurs produisent des résultats très satisfaisants pour toutes les caractéristiques. Comme les taux de couverture indiqués dans le Tableau 4 se rapprochent sensiblement du seuil de confiance théorique, les faibles biais observés dans le Tableau 2B n'ont plus vraiment d'importance. Nous pouvons en conclure qu'en ce qui concerne le biais, les quatre estimateurs de la variance ne sont pas très différents les uns des autres pour des estimations par le quotient. Par ailleurs, l'efficacité relative des estimateurs 3 et 4 n'est plus que très légèrement supérieure à celle de l'estimateur 2, peu importe la caractéristique. En l'occurence, ces trois estimateurs ont une efficacité relative supérieure à 2 pour les personnes occupées et les personnes actives. Pour ce qui a trait aux personnes en chômage, l'efficacité relative est quelque peu inférieure dans les trois cas, les valeurs se situant entre 1.5 et 1.8, ce qui est très comparable aux rapports observés pour la même caractéristique dans le cas des estimations non corrigées. Il convient de souligner ici que l'estimateur 1 est calculé avec 19 degrés de liberté (soit 1 par strate). Par contre, les trois autres estimateurs sont calculés avec 81 degrés de liberté puisque chaque UPÉ est un échantillon répété. Nous pouvons en conclure qu'une augmentation du nombre d'échantillons répétés aura pour effet d'accroître sensiblement la stabilité de l'estimateur de la variance de Keyfitz pour les estimations par le quotient et de la rendre comparable à celle des deux autres estimateurs (voir Tableau 3B).

2.4 Estimateur de la variance de Keyfitz avec 2 échantillons répétés par rapport à 6 échantillons répétés pour l'EPA

Les résultats de l'étude de Monte Carlo ont montré que la méthode de Keyfitz se compare avantageusement, au point de vue du biais et de l'efficacité, aux autres méthodes d'estimation de la variance pour des estimations par le quotient lorsque chaque méthode utilise le même nombre d'échantillons répétés. En outre, la méthode de Keyfitz a l'avantage d'être simple tandis qu'estimer la variance de variations ou de moyennes à l'aide des autres méthodes est un exercice très complexe. On a donc décidé d'étendre la méthode de Keyfitz aux secteurs AR. Dans le but d'accroître l'efficacité de cette méthode, on a remplacé les deux échantillons répétés de l'ancien plan de sondage par 6 groupes de renouvellement qui tiennent lieu d'échantillons répétés. Au départ, on craignait principalement que cette substitution se traduise par une forte hausse de l'estimation de la variance imputable au biais de renouvellement.

Tableau 2A

Biais (en pourcentage) des estimateurs de la variance des estimations de totaux des caractéristiques de la population active sans correction par le quotient

Caractéristique	Biais en pourcentage			
	$\hat{V}_1$	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
Personnes occupées	23.4	24.5	−4.7	−6.3
Personnes en chômage	6.3	6.6	3.7	1.2
Personnes actives	24.2	25.2	−5.1	−6.7

Tableau 2B

Biais (en pourcentage) des estimateurs de la variance des estimations de totaux des caractéristiques de la population active avec correction par le quotient

Caractéristique	Biais en pourcentage			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
Personnes occupées	3.7	4.3	−1.1	−3.1
Personnes en chômage	5.3	5.5	4.0	1.4
Personnes actives	4.5	5.0	−0.5	−2.5

Tableau 3A

Efficacité de $\hat{V}_2$, $\hat{V}_3$ et $\hat{V}_4$ par rapport à $\hat{V}_1$
(Eff. rel. de $\hat{V}_j = [EQM(\hat{V}_1)/EQM(\hat{V}_j)]^{1/2}$, $j = 2,3,4$

Caractéristique	Efficacité relative		
	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
Personnes occupées	1.51	3.22	3.11
Personnes en chômage	1.52	1.71	1.76
Personnes actives	1.49	3.24	3.12

Tableau 3B

Efficacité de $\hat{V}_2^{(R)}$, $\hat{V}_3^{(R)}$ et $\hat{V}_4^{(R)}$ par rapport à $\hat{V}_1^{(R)}$
(Eff. rel. de $\hat{V}_j^{(R)} = [EQM(\hat{V}_1^{(R)})/EQM(\hat{V}_j^{(R)})]^{1/2}$, $j = 2,3,4$

Caractéristique	Efficacité relative		
	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
Personnes occupées	2.13	2.59	2.52
Personnes en chômage	1.57	1.71	1.76
Personnes actives	2.08	2.56	2.51

Tableau 4

Taux de couverture des intervalles de confiance de 95% pour les estimations de totaux des caractéristiques de la population active avec correction par le quotient

Caractéristique	Taux de couverture			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
Personnes occupées	93.6	95.4	94.6	94.2
Personnes en chômage	94.3	95.1	95.3	95.0
Personnes actives	93.2	95.3	94.6	94.2

Nous avons analysé cet aspect de la question pour les trois caractéristiques de la population active en calculant, à l'aide de la formule élaborée dans la section 3, les estimations de la variance pour 2 et pour 6 échantillons répétés fondés sur les données de l'EPA pour une période de 24 mois (mars 1985 - février 1987). Nous avons ensuite calculé, pour chaque plan, la moyenne et l'écart type (ET) des 24 estimations pour chaque caractéristique de la population active. Enfin, nous avons établi le rapport des moyennes et des écarts types calculés pour chaque plan (2 échantillons/6 échantillons) pour 24 régions métropolitaines de recensement (RMR) et avons fait la moyenne des rapports pour ces 24 régions. Les résultats pertinents figurent dans le Tableau 5. Nous pouvons en tirer les conclusions suivantes:

(i) Le fait d'utiliser 6 échantillons répétés au lieu de 2 a très peu d'effet sur les variances; l'utilisation de groupes de renouvellement influe donc peu sur le biais des estimations de la variance.

(ii) Comme prévu, les variances sont plus stables avec 6 échantillons répétés qu'avec 2 et les résultats ne sont pas très différents de ceux de l'étude de Monte Carlo (voir la première colonne du tableau 3B).

En conclusion, l'utilisation de 6 échantillons répétés accroît sensiblement l'efficacité de la méthode de Keyfitz sans influer de façon notable sur le biais.

3. ESTIMATION DE LA VARIANCE POUR DES ESTIMATIONS ITÉRATIVES PAR LE QUOTIENT

3.1 Estimations itératives par le quotient dans l'EPA

L'ancien plan de sondage de l'EPA utilisait l'estimation par le quotient pour domaines post-stratifiés. Le sous-poids, qui est le résultat de la correction du poids de base en fonction de la non-réponse, était rajusté par le quotient en fonction d'estimations auxiliaires (ou externes) de la population cible de l'EPA pour 38 post-strates selon l'âge et le sexe au niveau provincial. La population cible de l'EPA comprend toutes les personnes âgées de 15 ans et plus, sauf les membres des forces armées, les pensionnaires d'institution et les personnes vivant dans les réserves indiennes.

L'estimation par le quotient avait pour effet de rehausser sensiblement la qualité des données provinciales alors que les données infraprovinciales demeuraient peu fiables. Afin d'accroître la fiabilité des données infraprovinciales, surtout au niveau des régions économiques (RE) et des RMR, on a appliqué une technique d'estimation itérative par le quotient, qui permet de corriger simultanément des estimations aux niveaux provincial et infraprovincial.

Tableau 5

Comparaison des estimations de la variance pour les secteurs AR par suite de l'utilisation de 2 et de 6 échantillons répétés par strate selon des données de l'EPA pour des RMR mars 1985 - février 1987

Caractéristique	Rapport moyen des moyennes des variances (2 échantillons/ 6 échantillons)	Rapport moyen des E.T. des variances (2 échantillons/ 6 échantillons)
Personnes occupées	0.997	1.813
Personnes en chômage	0.995	1.515
Personnes actives	1.003	1.833

Note: Pour chaque RMR, on a calculé les moyennes et les écarts types des estimations de la variance pour 2 et pour 6 échantillons répétés, à partir de données s'étendant sur une période de 24 mois. On a ensuite calculé, pour chaque RMR, le rapport (2 échantillons/6 échantillons) des moyennes des variances et le rapport des E.T. des variances. Les rapports moyens qui figurent dans le tableau sont la moyenne des rapports des 24 RMR.

Les itérations se traduisent en une suite de corrections: premièrement, on rajuste le sous-poids en fonction de la population infraprovinciale (c'est-à-dire, population de RMR ou de régions autres qu'une RMR dans les RE), puis on multiplie le poids ainsi obtenu par le facteur de correction âge-sexe pour la province (le remaniement de l'échantillon a fait passer le nombre de groupes d'âge-sexe de 38 à 24). On répète l'opération une fois pour obtenir une deuxième paire de poids. Il convient de souligner que lorsqu'on définit des cellules de correction pour les RE, on ne tient pas compte des RMR contenues dans ces régions de sorte que les cellules de correction au niveau infraprovincial s'excluent mutuellement. Désignons par W_0 le sous-poids et par (W_1, W_2) et (W_3, W_4) les deux paires de poids tirées respectivement de la première et de la seconde itération. On utilise W_4 pour estimer les caractéristiques de la population active. À cause de l'ordre des corrections, la somme des valeurs de W_4 pour les groupes d'âge-sexe de niveau provincial est égale aux estimations auxiliaires des groupes correspondants mais la somme des mêmes valeurs au niveau infraprovincial (RE et RMR) n'est pas tout à fait égale aux estimations auxiliaires correspondantes. Les écarts sont toutefois très faibles.

Les bases de secteurs spéciaux, qui sont composées des établissements militaires, des institutions et des régions éloignées, ne respectent pas en général les divisions des RE et des RMR et, de ce fait, ne sont pas traitées de la même façon dans l'itération que les bases de secteurs réguliers. Chaque secteur spécial constitue une strate dans chaque province, les seules exceptions étant les régions éloignées du Québec et de l'Alberta, qui sont sous-stratifiées. Les RE et les RMR qui sont utilisées pour la base de secteurs spéciaux sont dites "désignées". Les enregistrements d'un secteur spécial contenus dans le fichier de l'échantillon sont reproduits pour chaque RE ou RMR désignée avec des sous-poids rajustés en fonction de la proportion de la population du secteur contenue dans la RE ou la RMR en question. On applique ensuite la méthode itérative de la façon définie précédemment.

3.2 Formule de la variance pour des estimations par le quotient calculées en une seule itération

Dans cette sous-section, nous déterminons la formule de la variance pour des estimations par le quotient calculées en une seule itération. Pour cela, nous procédons essentiellement à une application répétée de l'approximation de la série de Taylor aux estimations itératives par le quotient jusqu'à ce que nous obtenions une forme linéaire de sous-poids. Nous appliquons ensuite la formule d'itération à la manière de Woodruff (1971). Arora et Brackstone (1977a,b) et Brackstone et Rao (1979) ont aussi recours à l'application répétée de l'approximation de la série de Taylor pour déterminer la formule de la variance des estimations

itératives par le quotient pour l'échantillonnage aléatoire simple d'unités ou de grappes. Nous utilisons cette méthode pour l'échantillonnage stratifié à plusieurs degrés avec PPT en nous fondant sur l'approche de Woodruff.

Soient $Y^{(0)}$, $Y^{(1)}$, $Y^{(2)}$ les estimations d'une caractéristique y de la population active dans une province, ces estimations étant fondées respectivement sur W_0, W_1, et W_2. Les chiffres entre parenthèses correspondent aux indices inférieurs de W.

Nous pouvons alors exprimer $Y^{(2)}$ de la façon suivante:

$$Y^{(2)} = \sum_a \frac{Y_a^{(1)}}{P_a^{(1)}} P_a \tag{3.1}$$

où $Y_a^{(1)}$ = estimation de la caractéristique y pondérée avec W_1 pour le groupe d'âge-sexe a dans la province;

$P_a^{(1)}$ = estimation de la population pondérée avec W_1 pour le groupe d'âge-sexe a dans la province;

P_a = estimation auxiliaire (ou externe) de la population pour le groupe d'âge-sexe a dans la province.

Posons $F_a = Y_a^{(1)}/P_a^{(1)}$. L'approximation de Taylor du premier ordre pour F_a à $(E(Y_a^{(1)}), E(P_a^{(1)}))$ est

$$F_a \doteq \frac{E(Y_a^{(1)})}{E(P_a^{(1)})} + \frac{1}{E(P_a^{(1)})}\left\{Y_a^{(1)} - E(Y_a^{(1)})\right\} - \frac{E(Y_a^{(1)})}{\{E(P_a^{(1)})\}^2}\left\{P_a^{(1)} - E(P_a^{(1)})\right\}$$

où E désigne l'espérance mathématique.

Nous pouvons alors formuler une approximation de Taylor pour la variance de $Y^{(2)}$:

$$V(Y^{(2)}) = V\left(\sum_a F_a P_a\right) \doteq V\left\{\sum_a \frac{P_a}{E(P_a^{(1)})}(Y_a^{(1)} - R_{Ya}^{(1)} P_a^{(1)})\right\} \tag{3.2}$$

où

$$R_{Ya}^{(1)} = \frac{E(Y_a^{(1)})}{E(P_a^{(1)})}.$$

Or, il est possible d'exprimer les estimations $Y_a^{(1)}$ et $P_a^{(1)}$, fondées sur W_1, en fonction des estimations pondérées selon W_0:

$$\begin{aligned} Y_a^{(1)} &= \sum_s \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_s, \\ P_a^{(1)} &= \sum_s \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_s, \end{aligned} \tag{3.3}$$

où s désigne une RMR ou une RE, ou encore la partie d'une RE qui ne correspond pas à des RMR, et P_s est la population de la région infraprovincial s. En remplaçant $Y_a^{(1)}$ et $P_a^{(1)}$ dans l'équation (3.2) par les équations (3.3) et en appliquant l'approximation de Taylor du premier ordre aux rapports des estimations pondérées en fonction de W_0, nous obtenons:

$$V(Y^{(2)}) \doteq V\left[\sum_a \frac{P_a}{E(P_a^{(1)})} \sum_s \frac{P_s}{E(P_s^{(0)})} \left\{\left(Y_{sa}^{(0)} - R_{Ysa}^{(0)} P_s^{(0)}\right) - R_{Ya}^{(1)}\left(P_{sa}^{(0)} - R_{Psa}^{(0)} P_s^{(0)}\right)\right\}\right], \tag{3.4}$$

où

$$R_{Ysa}^{(0)} = \frac{E(Y_{sa}^{(0)})}{E(P_s^{(0)})} \text{ et } R_{Psa}^{(0)} = \frac{E(P_{sa}^{(0)})}{E(P_s^{(0)})}.$$

L'équation (3.4) peut être réécrite en fonction d'estimations pour échantillons répétés. Définissons

$$\begin{aligned} Z_{Yshia}^{(0)} &= \frac{P_a}{E(P_a^{(1)})} \frac{P_s}{E(P_s^{(0)})} (Y_{shia}^{(0)} - R_{Ysa}^{(0)} P_{shi}^{(0)}), \\ Z_{Pshia}^{(0)} &= \frac{P_a}{E(P_a^{(1)})} \frac{P_s}{E(P_s^{(0)})} (P_{shia}^{(0)} - R_{Psa}^{(0)} P_{shi}^{(0)}), \end{aligned} \tag{3.5}$$

où h désigne une strate de s et i désigne un échantillon répété dans h.

Alors, nous pouvons réécrire (3.4) en modifiant l'ordre des sommations:

$$\begin{aligned} V(Y^{(2)}) &\doteq V\left\{\sum_s \sum_{h \in s} \sum_{i=1}^{n_h} \sum_a \left(Z_{Yshia}^{(0)} - R_{Ya}^{(1)} Z_{Pshia}^{(0)}\right)\right\} \\ &= V\left(\sum_s \sum_{h \in s} \sum_{i=1}^{n_h} D_{shi}^{(0)}\right) \end{aligned} \tag{3.6}$$

où

$$D_{shi}^{(0)} = \sum_a \left(Z_{Yshia}^{(0)} - R_{Ya}^{(1)} Z_{Pshia}^{(0)}\right).$$

Pour les strates de secteurs réguliers, les variables $(\Sigma_{i=1}^{n_h} D_{pshi}^{(0)})$ sont indépendantes parce qu'elles reposent sur des sous-poids. Toutefois, en ce qui concerne les strates de secteurs spéciaux, ces variables sont fortement corrélées parce que les mêmes enregistrements sont attribués à toutes les régions infraprovinciales désignées.

Nous pouvons réécrire (3.6) comme suit:

$$\begin{aligned} V(Y^{(2)}) &\doteq V\left(\sum_{h \in s} \sum_h \sum_{i=1}^{n_h} D_{shi}^{(0)}\right) \\ &= V\left(\sum_h \sum_{i=1}^{n_h} \sum_{s \ni h} D_{shi}^{(0)}\right) \end{aligned} \tag{3.7}$$

où $\Sigma_{s \ni h}$ désigne la sommation pour toutes les régions infraprovinciales qui renferment la strate h. S'il s'agit d'une strate d'un secteur régulier, la sommation $(\Sigma_{s \ni h})$ est superflue puisque la strate ne se trouve que dans une seule région infraprovincial. En revanche, la strate

d'un secteur spécial peut être incluse dans plus d'une région infraprovinciale et la sommation ($\Sigma_{s \ni h}$) sert à additionner toutes les valeurs ($D_{shi}^{(0)}$) comprises dans cette strate.

Posons

$$D_{hi}^{(0)} = \sum_{s \ni h} D_{shi}^{(0)}.$$

Alors, (3.7) devient

$$V(Y^{(2)}) \doteq V\left(\sum_{h} \sum_{i=1}^{n_h} D_{hi}^{(0)}\right) \tag{3.8}$$

Les variables ($\Sigma_i D_{hi}^{(0)}$) sont indépendantes puisqu'elles reposent sur des sous-poids. En ne tenant pas compte de la CPF (correction pour population finie), nous pouvons estimer la variance à l'aide de la formule suivante:

$$\hat{V}(Y^{(2)}) \doteq \sum_{h} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (D_{hi}^{(0)} - \bar{D}_h^{(0)})^2 \tag{3.9}$$

où

$$\bar{D}_h^{(0)} = \frac{1}{n_h} \sum_{i=1}^{n_h} D_{hi}^{(0)}.$$

Or, cette expression implique des espérances mathématiques qui sont inconnues. Nous pouvons obtenir une approximation assez juste de la variance en remplaçant les espérances mathématiques par leurs estimations; ainsi, de l'équation (3.9) nous déduisons la forme finale de $\hat{V}$:

$$\hat{V} \doteq \sum_{h} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (D_{hi}^{(2)} - \bar{D}_h^{(2)})^2 \tag{3.10}$$

où

$$D_{hi}^{(2)} = \sum_{s \ni h} D_{shi}^{(2)},$$

$$\bar{D}_h^{(2)} = \frac{1}{n_h} \sum_{i=1}^{n_h} D_{hi}^{(2)},$$

$$D_{shi}^{(2)} = \sum_{a} \left(Z_{Yshia}^{(2)} - R_{Ya}^{(2)} Z_{Pshia}^{(2)}\right),$$

$$Z_{Yshia}^{(2)} = \frac{P_a}{P_a^{(1)}} \frac{P_s}{P_s^{(0)}} \left(Y_{shia}^{(0)} - \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)}\right)$$

$$= Y_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} Y_{sa}^{(2)},$$

$$Z_{Pshia}^{(2)} = \frac{P_a}{P_a^{(1)}} \frac{P_s}{P_s^{(0)}} \left(P_{shia}^{(0)} - \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)}\right)$$

$$= P_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} P_{sa}^{(2)},$$

et

$$R_{Ya}^{(2)} = \frac{Y_a^{(1)}}{P_a^{(1)}} = \frac{P_a}{P_a^{(1)}} \frac{Y_a^{(1)}}{P_a} = \frac{Y_a^{(2)}}{P_a}.$$

La formule (3.10) donne la variance des estimations de caractéristiques de la population active pondérées en fonction de W_2 et nécessite l'utilisation des deux poids W_0 et W_2.

3.3 Application de la formule de la variance établie dans la sous-section 3.2 aux estimations par le quotient calculées en deux itérations

L'application répétée de la méthode de linéarisation de Taylor peut servir à déterminer la formule de la variance pour des estimations par le quotient calculées en deux itérations. La formule obtenue de cette façon est toutefois très complexe. Nous avons supposé que la formule de la variance pour des estimations par le quotient calculées en une seule itération produirait une estimation assez juste de la variance des estimations calculées en deux itérations. Cette hypothèse repose sur le fait que les poids ne subissent que de légères variations après la première itération. Or, la formule de la variance pour les estimations calculées en une seule itération utilise la paire de poids (W_0, W_2). Nous avons donc décidé de substituer (W_0, W_4) à (W_0, W_2) après avoir constaté que l'utilisation de W_4 ne changeaient rien aux coefficients de variation (CV) des estimations des caractéristiques de la population active fondées sur W_4. La formule de la variance qui utilise la paire de poids (W_0, W_4) sera désignée comme l'estimateur de la variance pour une seule itération.

Nous avons vérifié notre hypothèse à l'aide d'une simulation de Monte Carlo. Nous avons utilisé à cette fin les données du recensement de 1981 pour la province de Nouvelle-Écosse. Le plan de sondage de l'EPA a été simulé à toutes les étapes de l'échantillonnage et un total de 1 000 échantillons de Monte Carlo ont été sélectionnés individuellement. Pour chacun de ces échantillons, nous avons calculé les valeurs suivantes pour les trois caractéristiques de la population active aux niveaux provincial et infraprovincial:

1. Estimation par le quotient calculées en deux itérations $Y^{(4)}$.
2. Estimation de la variance $\hat{V}(Y^{(4)})$, à l'aide de l'estimateur de la variance pour une seule itération, et estimation du CV correspondant.
3. Intervalle de confiance de 95% (c'est-à-dire, $Y^{(4)} \pm 1.96\sqrt{\hat{V}(Y^{(4)})}$).

À la fin de la simulation, nous avons fait la moyenne des 1 000 estimations de CV et avons comparé cette moyenne au CV de Monte Carlo, qui est très près de la valeur réelle. Les résultats sont reproduits dans le tableau 6A. L'écart est inférieur à 8% dans les 21 cas (3 caractéristiques pour chacune des 7 régions) et inférieur à 4% dans 13 cas.

Nous avons aussi déterminé la proportion des intervalles de confiance qui renferment la valeur réelle de la caractéristique. Les résultats sont reproduits dans le tableau 6B. Les taux de couverture pour les "personnes occupées" et les "personnes actives" sont, en règle générale, très près du seuil théorique tandis que ceux pour les "personnes en chômage" sont quelque peu inférieurs mais encore acceptables.

Tableau 6A

CV moyens obtenus par l'estimateur de la variance pour une seule itération et CV de Monte Carlo

Caractéristique	RE 210	RE 220	RE 230	RE 240	RE 250	RMR Halifax	Province (N.-É.)
				CV moyens			
Personnes occupées	3.52	3.46	3.14	3.05	1.96	2.01	1.08
Personnes en chômage	10.36	12.28	13.13	13.43	10.35	10.55	5.27
Personnes actives	2.98	3.17	2.85	2.73	1.77	1.83	0.91
				CV de Monte Carlo			
Personnes occupées	3.48	3.35	2.95	2.86	1.97	1.99	1.11
Personnes en chômage	10.90	12.71	13.28	13.37	11.12	11.31	5.59
Personnes actives	2.76	3.08	2.76	2.53	1.72	1.74	0.92

Tableau 6B

Taux de couverture des intervalles de confiance de 95% établis à l'aide de l'estimateur de la variance pour une seule itération

Caractéristique	RE 210	RE 220	RE 230	RE 240	RE 250	RMR Halifax	Province (N.-É.)
Personnes occupées	94.5	92.8	94.0	94.7	94.7	94.9	92.5
Personnes en chômage	92.1	90.7	91.4	91.8	92.7	92.7	93.1
Personnes actives	96.2	93.0	93.6	95.2	95.2	96.0	94.0

Nous avons aussi constaté que l'estimation par le quotient calculée en deux itérations est presque non biaisée, le biais maximum dans les 21 cas étant de 0.35%.

4. CONCLUSIONS

En ce qui a trait aux caractéristiques de la population active considérées dans cette étude, nous avons vu que la méthode de Keyfitz appliquée à des estimations sans correction par le quotient (dans ce cas la méthode de Keyfitz est réduite à une méthode répétitive) produit des estimations qui ont un biais positif très élevé et qui sont peu efficaces tandis que les autres méthodes appliquées au même genre d'estimations produisent des estimations qui sont plus efficaces et dont le biais est négligeable.

Toutefois, lorsqu'appliquées à des estimations par le quotient, toutes les méthodes sans exception produisent des estimations qui ont un biais négligeable. Nous avons aussi montré que l'on pouvait accroître sensiblement l'efficacité de la méthode de Keyfitz au point de la rendre comparable à celle des autres méthodes en augmentant le nombre d'échantillons répétés. Nous avons vu à l'aide de données authentiques de l'EPA que le fait d'utiliser 6 groupes de renouvellement au lieu de 2 pseudo-échantillons répétés dans la formule de la variance de Keyfitz n'introduisait pas de biais de renouvellement. Comme en font foi les résultats de l'étude de Monte Carlo, l'estimateur de la variance pour une seule itération, déduit de la méthode de Keyfitz par une linéarisation de Taylor, produit des estimations assez justes de la variance des estimations par le quotient calculées en deux itérations et est aussi caractérisé par de bons taux de couverture.

REMERCIEMENTS

Les auteurs tiennent à exprimer leur reconnaissance aux deux arbitres ainsi qu'au rédacteur en chef et à un rédacteur associé pour leurs commentaires utiles sur la version préliminaire de cet article.

BIBLIOGRAPHIE

ARORA, H.R., et BRACKSTONE, G.J. (1977a). An investigation of the properties of raking ratio estimators: I with simple random sampling. *Techiques d'enquête*, 3, 62-83.

ARORA, H.R., et BRACKSTONE, G.J. (1977b). An investigation of the properties of raking ratio estimators: II with cluster sampling. *Techniques d'enquête*, 3, 232-252.

BRACKSTONE, G.J., et RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Série C, 41, 97-114.

KEYFITZ, N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Journal of the American Statistical Association,* 52, 503-510.

PLATEK, R., et SINGH, M.P. (1976). Méthodologie de l'enquête sur la population active du Canada. No. 71-526 au catalogue, Statistique Canada.

RAO, J.N.K. (1975). Unbiased variance estimation for multi-stage designs. *Sankhyā*, Série C, 37, 133-139.

RAO, J.N.K., HARTLEY, H.O., et COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society,* 24, 482-490.

SINGH, M.P., et DREW, J.D. (1981). Redesigning continuous surveys in a changing environment. *Techniques d'enquête,* 7, 44-73.

WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association,* 66, 411-414.

Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 173–181
Statistique Canada

La fiche «AGEVEN»: un outil pour la collecte des données rétrospectives

PHILIPPE ANTOINE, XAVIER BRY et PAP DEMBA DIOUF[1]

RÉSUMÉ

La fiche «AGEVEN» permet, grâce à sa simplicité d'emploi, une meilleure datation des événements et d'opérer un classement respectif des événements démographiques (naissances et décès), des changements matrimoniaux et des changements de lieux de résidence. Les données obtenues servent à reconstituer avec précision les conditions socio-économiques au moment où se produisent les événements démographiques étudiés.

MOTS CLÉS: Enquête rétrospective; biographies; enquête démographique.

1. INTRODUCTION

Deux grandes méthodes de collecte sont à la disposition du démographe pour recueillir des données afférentes au mouvement naturel (natalité et mortalité): l'observation suivie et le questionnaire rétrospectif. La méthode d'observation suivie (suivre un même échantillon de population pendant un intervalle de temps relativement long) est en théorie celle qui donne les résultats les plus précis. Elle présente toutefois certains inconvénients. Les coûts d'enquête sont élevés du fait des nombreux passages nécessaires à l'observation. Les délais d'obtention des résultats demeurent relativement longs. Enfin, en milieu urbain, l'application de la méthode se heurte à l'extrême mobilité de la population ce qui entraîne une déperdition importante de l'échantillon, comme celle rencontrée lors des enquêtes mortalité infantile et juvénile (EMIJ) de l'IFORD (Scott 1985, Fargues 1985).

La méthode rétrospective donne des résultats moins fiables car elle fait davantage appel à la mémoire des enquêtes. Cependant elle porte sur une période totale d'observation en général plus longue que celle des enquêtes longitudinales mises en place ces dernières années dans les pays africains. Les risques d'omission des événements demeurent élevés et leur datation reste imprécise; enfin, en milieu urbain, le recueil de la vie passée mêle des événements qui se sont déroulés dans la ville objet de l'enquête et d'autres, plus anciens, qui se sont produits dans d'autres lieux de résidence (urbain ou rural).

Désirant connaître la mortalité et la fécondité différentielles à Pikine, dans la banlieue de Dakar, et souhaitant obtenir rapidement des résultats assez fiables, notre choix s'est porté sur une méthode de collecte permettant de reconstituer avec précision les facteurs de risques de mortalité infanto-juvénile au moment du décès pour chacun des enfants des femmes soumises à l'enquête. L'enquête a été réalisée conjointement par la Direction de la Statistique du Sénégal et l'Orstom (Antoine et Diouf 1986). L'enquête de terrain s'est déroulée de mars à mai 1986. Les premiers résultats étaient disponibles dès septembre 1986. Cette méthode se distingue donc de la méthode rétrospective la plus employée qui ne prend en compte que les caractéristiques socio-économiques et culturelles des femmes au moment de l'enquête, alors que celles-ci ont pu considérablement évoluer au cours de leur vie féconde (amélioration ou dégradation des conditions d'habitat, changement dans la situation matrimoniale, dans l'activité...). Cette méthode permet de mieux évaluer les interférences entre l'insertion

[1] Philippe Antoine, démographe et Xavier Bry, statisticien à l'ORSTOM BP 1386 Dakar Sénégal; Pap Demba Diouf, démographe à la direction de la Statistique BP 116 Dakar Sénégal.

urbaine et les changements de comportements démographiques. Ces objectifs devaient guider notre stratégie de collecte:
- obtenir une liste complète des événements observés (principalement naissances et décès)
- dater le plus précisément possible ces événements
- replacer les événements dans leur contexte socio-économique du moment (situation matrimoniale, situation professionnelle du mari et de la femme, conditions de vie).

2. RECUEIL ET DATATION DES ÉVÉNEMENTS DÉMOGRAPHIQUES

Réussir une enquête rétrospective nécessite en particulier d'établir une biographie la plus précise possible (concernant le domaine étudié) de chaque personne soumise à l'enquête; il faut donc trouver le moyen d'aider la personne interrogée à replacer dans le temps les événements vécus dans le passé.

Plusieurs améliorations méthodologiques ont déjà été proposées. Ainsi Ferry (1977) a utilisé un "fichier événement" qui repose sur l'affectation d'une fiche à chaque événement. Selon l'auteur "toute l'originalité de la méthode consiste, avec la personne enquêtée elle-même, à mettre de l'ordre dans la succession des événements (Grossesses, unions et ruptures d'union, lieux d'habitation, etc....) et à les situer les uns par rapport aux autres (...). La technique consiste alors à retrouver avec la personne enquêtée la succession, la logique, les interférences et finalement la biographie individuelle". Toutefois, cette méthode est relativement complexe et nécessite le maniement de nombreuses fiches sur le terrain et lors de l'exploitation.

Un autre mode de classement et de datation des événements a été retenu lors de l'enquête sénégalaise sur la fécondité de 1978, il s'agit du graphique «AGEVEN» (âge à l'événement). L'utilisation du graphique «AGEVEN» pour l'enquête sénégalaise visait deux buts:
- permettre une meilleure estimation de l'âge des femmes et celui de leurs enfants grâce à une datation relativement précise;
- permettre une bonne estimation de la fécondité en faisant l'historique des maternités de toutes les femmes.

Le graphique «AGEVEN» de l'enquête sénégalaise fécondité (figure 1) se présente sous la forme de deux courbes. La courbe de droite figurant la ligne de vie de la femme (courbe L.V.) est représentée par un intervalle gradué en trimestres qui, permettait de situer dans l'année les événements affectant la femme. La courbe de gauche appelée D.E. donnait la durée écoulée entre le moment de l'événement et la date de l'enquête. Aussi, à chaque année sur la courbe L.V. correspondait un âge sur la courbe D.E., et inversement. Ce graphique, repris par l'enquête ivoirienne sur la fécondité, apparaît surtout comme un instrument de datation des événements.

3. L'UTILISATION DE LA FICHE «AGEVEN» LORS DE L'ENQUÊTE MENÉE À PIKINE

Nous avons cherché à combiner certains des avantages de chacune de ces méthodes de collecte: d'une part la simplicité d'emploi du graphique «AGEVEN» dans la datation des événements, et d'autre part la possibilité de saisir plusieurs types d'événements et leur classement les uns par rapport aux autres comme le permet le fichier événement. Nous avons donc systématisé cette fiche «AGEVEN» en distinguant le repérage des événements démographiques (naissances, décès), les changements de situation matrimoniale, et les changements de lieu de résidence. Par commodité nous avons retenu le même vocable pour désigner notre fiche, que celui l'employé pour dénommer le grapique utilisé lors de l'enquête sénégalaise

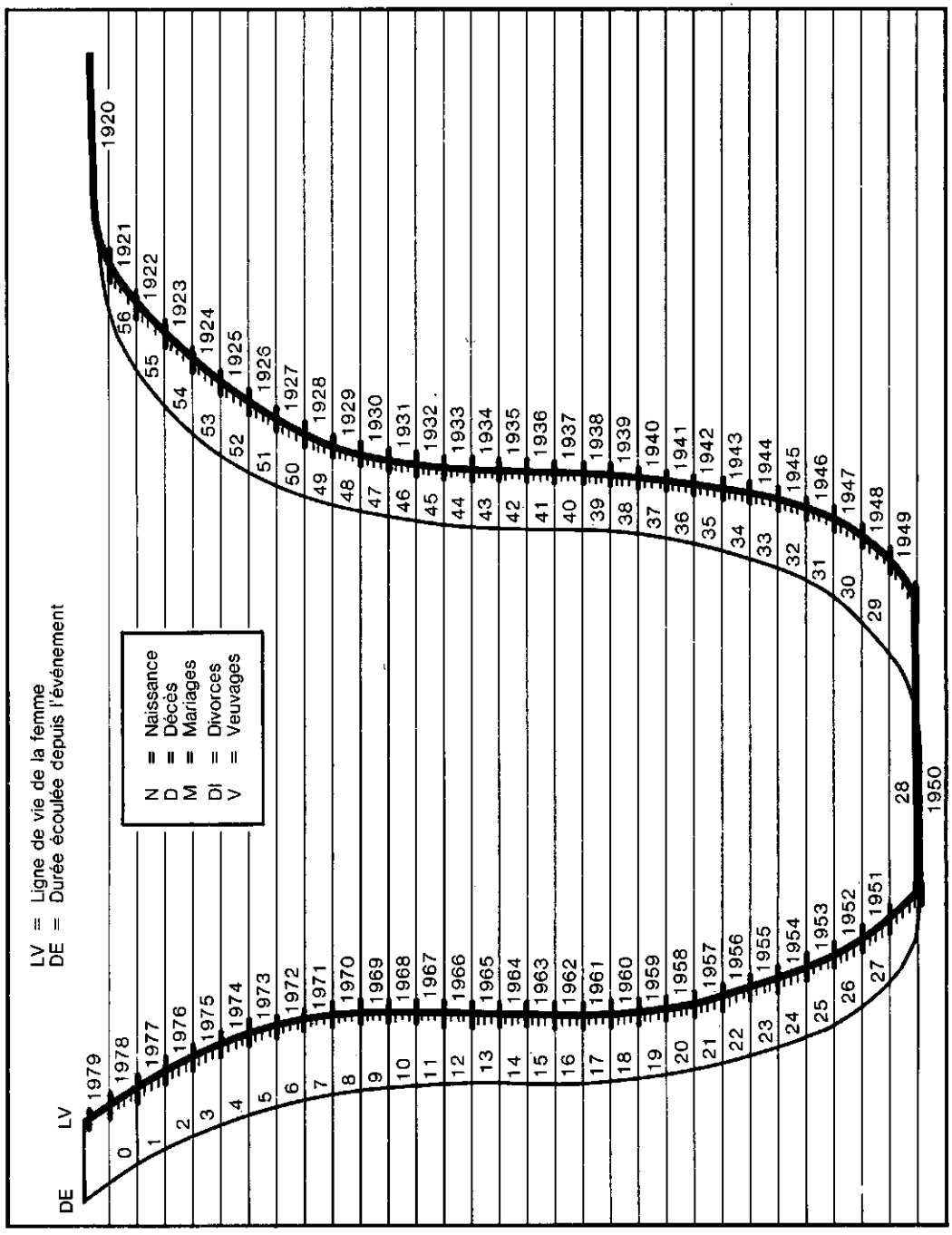

Figure 1. Graphique AGEVEN de l'enquête Sénégalaise sur la fécondité

sur la fécondité. Le nom est le même, les possibilités d'utilisation diffèrent. La fiche «AGEVEN» dont un exemple est présenté à la figure 2 comprend trois colonnes:

- la première concerne les événements démographiques (naissance (N); décès (DCD); avortement (AV); fausse couche (FC); mort-né (MN)). Chaque événement, naissance ou décès doit être suivi de son rang chronologique, des nom et prénom de l'enfant et éventuellement de la date précise;
- la seconde colonne concerne les événements matrimoniaux et le rang de chacun des partenaires ou conjoints (mariages (M), divorces (D), veuvages (V), le rang des différents pères indicé P1, P2, ..., Pn).
- la troisième colonne permet de noter le lieu de résidence lors de chacun des événements démographiques et matrimoniaux. Cette colonne permet d'établir ensuite l'itinéraire migratoire des femmes et de déterminer leur date d'installation à Pikine.

La fiche «AGEVEN» est un outil méthodologique qui remplit plusieurs fonctions:

- repérer les événements dans le temps;
- aider la femme à situer temporellement des événements dont elle a oublié la date;
- s'assurer de l'exhaustivité des événements démographiques vécus par la femme interrogée;
- repérer les changements de résidence et de localiser les événements;
- vérifier la cohérence des événements entre eux.

Le déroulement de l'entretien comprenait deux phases, l'une concernant le ménage, l'autre les femmes de 15-49 ans. Le questionnaire "ménage", au niveau duquel sont répertoriées toutes les personnes résidentes ou non du ménage, s'intéresse en particulier à la filiation des personnes recensées, à leur lien de parenté avec le chef de ménage ou de noyau, à leur sexe, à leur situation matrimoniale, à leur date de naissance ou leur âge. Le questionnaire "femme" concerne l'ensemble des femmes, résidentes et présentes dans le ménage, âgées de 15-49 ans. La fiche «AGEVEN» est utilisée au moment de remplir ce questionnaire.

Pour transcrire les informations obtenues sur cette fiche l'enquêteur peut prendre différents points de référence (la date de naissance de la femme; la date de la première naissance ...), et reconstitue avec la participation de l'enquêtée toute sa ligne de vie, c'est-à-dire tous les autres événements qui ont affecté sa vie tels mariage, divorce, et les différents accouchements. Cette opération se décompose comme suit:

1. Après enregistrement de la première naissance vivante, d'enquêteur demande à l'enquêtée de lui communiquer toutes les naissances vivantes qui ont suivi dans un ordre croissant, que l'enfant soit décédé ou non, qu'il vive dans le ménage ou ailleurs.
2. Ensuite l'enquêteur enregistre ces naissances sur la fiche en se servant des pièces officielles qui lui sont présentées. Les pièces officielles existaient dans la plupart des cas pour les enfants surtout s'ils étaient nés dans la région de Dakar. Pour l'âge des mères, par contre, comme pour certaines autres naissances d'enfants, l'enquêteur devrait se servir d'éléments provenant du calendrier historique pour déterminer les dates (mois et années).

La fiche «AGEVEN» permet de situer les événements soit selon l'âge de la femme au moment de l'événement, ou la durée écoulée depuis l'événement, soit selon la date de cet événement. Tout écart important entre deux naissances, ou toute incohérence entre deux événements est ainsi repéré plus facilement au cours de l'entretien avec la femme.

À l'aide de la fiche «AGEVEN», on peut également vérifier la cohérence des événements. Ainsi, deux enfants ne peuvent naître à moins de neuf mois d'intervalle; ou bien une femme ne peut déclarer s'être mariée à l'âge de 12 ans, avoir eu son premier enfant à 14 ans en 1970 et déclarer être né en 1950. Dans ce cas, il est fort probable que la date de naissance de la femme soit erronée et il faudra la corriger.

La fiche permet d'enregistrer aussi bien des événements donnés avec une date précise, que des événements dont on peut simplement donner l'âge (tel enfant a aujourd'hui 10 ans; je

Année	Âge	Événements démographiques	Événements matrimoniaux	Lieux de résidence
1986	mars			
1985	0			
1984	1	N_3 Aminata 18-12-84	P_2	Pikine
1983	2			
1982	3		M_2	Pikine
1981	4			
1980	5			
1979	6	D_1 Ibrahima à 4 ans	DI_1	Pikine
1978	7	N_2 Abdoul 5-01-78	P_1	Dakar
1977	8			
1976	9			
1975	10	N_1 Ibrahima	P_1	Dakar
1974	11			
1973	12		M_1	Thiès
1972	13			
1971	14			
1970	15			
1969	16			
1968	17			
1967	18			
1966	19			
1965	20			
1964	21			
1963	22			
1962	23			Kaolack
1961	24			
1960	25			
1959	26			
1958	27			

AGEVEN BNR/ORSTOM

CODES

N_i + date + nom + lieu = naissance enfant n° i
D_i + date + nom + lieu = décès enfant n° i
M_i + date + nom + lieu = mariage n° i
DI_i + date + nom = divorce conjoint n° i
V_i + date + nom = veuvage conjoint n° i
FC = fausse couche A = avortement MN = mort-né
P_i = père n° i

IDENTIFIANT

Nom:
Ilot:
Concession:
Ménage:
N° de la femme:

Année	Âge	Événements démographiques	Événements matrimoniaux	Lieux de résidence
1936	49			
1937	48			
1938	47			
1939	46			
1940	45			
1941	44			
1942	43			
1943	42			
1944	41			
1945	40			
1946	39			
1947	38			
1948	37			
1949	36			
1950	35			
1951	34			
1952	33			
1953	32			
1954	31			
1955	30			
1956	29	NF Awa		
1957	28			Kaolack

Figure 2. Exemple d'utilisation de la fiche «AGEVEN».

me suis mariée il y a 15 ans). Enfin, grâce à cette fiche des événements datés de façon imprécise, peuvent être situés: tel enfant est né entre celui né le 10-2-74 et celui né en 1978. Il est fort probable que l'enfant soit alors né durant l'année 1976. Cette fiche nécessite que l'enquêteur porte un regard critique sur l'enchaînement des événements et qu'il cherche à la compléter au maximum en s'assurant de la fiabilité et de la cohérence des réponses portées. Cela n'est possible qu'en instaurant un dialogue confiant avec la personne enquêtée.

Après avoir enregistré toutes les naissances vivantes déclarées par l'enquêtée, l'enquêteur s'intéresse alors aux intervalles intergénésiques. Il arrive que certains de ces événements ne soient pas rapportés dans les premières réponses, mais l'utilisation de la fiche «AGEVEN» permet à l'enquêteur de mieux dépister les événements omis. Aussi, interroge-t-il l'enquêtée sur ce qui s'est passé, à chaque fois qu'il s'aperçoit que l'intervalle entre deux naissances vivantes est supérieur à deux ans. Les réponses fournies par l'enquêtée permettent ainsi de relever les avortements, mort-nés, naissances suivies de décès, voire d'obtenir des informations sur les pratiques contraceptives. Ce point ne faisait pas partie des objectifs de l'enquête, mais le dialogue qui se noue autour de l'«AGEVEN» peut cependant permettre d'approfondir des questions ayant trait à la planification des naissances.

Chacun de ces événements étaient mis en relation avec le lieu, la situation matrimoniale et le partenaire de la femme au moment de l'événement. Après avoir relevé tous les événements ayant affecté la femme, l'enquêteur devait estimer de facon beaucoup plus précise la date de naissance de la mère. En effet, au moment de remplir le questionnaire "ménage", l'enquêteur avait déjà inscrit la date de naissance de la mère qui lui avait été communiquée soit par la femme ou par le chef de ménage. Cette fois, seul avec l'enquêtée et en possession des événements ayant affecté la femme, il était à même de fournir le meilleur âge possible de l'enquêtée.

Prenons l'exemple d'une femme. Awa est née en 1956 à Kaolack: elle déclare avoir trois enfants, Ibrahima, qui aurait aujourd'hui 10 ans, né à Dakar, décédé à l'âge de 4 ans à Pikine; Abdoul, né le 5 janvier 1978 à Dakar; et Aminata née le 18 décembre 1984 à Pikine. Cette femme s'est mariée une première fois à l'âge de 17 ans à Thiès. Elle a divorcé en 1979 (elle résidait alors à Pikine). Elle s'est remariée en 1982 et résidait à cette époque à Pikine (figure 2). En menant l'interview l'enquêteur constatera un écart de presque 7 ans entre Abdoul et Aminata. Il devra insister pour savoir si durant cet intervalle d'autres naissances ou d'autres grossesses ne se sont pas produites. Dans ce cas précis, le divorce et le remariage trois ans plus tard peuvent expliquer cet écart, mais il faut s'assurer auprès de la femme qu'un tel écart ne masque pas des événements démographiques.

La forme interactive prise par l'interview semble favoriser le dialogue avec la personne enquêtée et améliore le contact entre enquêteur et enquêté(e), malheureusement trop souvent basé sur le doute de l'enquêteur et la méfiance de l'enquêté(e) (Bonnet 1984). Au fur et à mesure que l'enquêteur ou l'enquêtrice poursuit son investigation de nouveaux événements sont déclarés. Lorsque l'on demande s'il n'y a pas un événement entre deux naissances à plus de deux ans d'intervalle, la personne interrogée s'étonne et formule l'un ou l'autre type de réponse: si elle n'a connu aucun événement "Pourquoi demandez-vous celà?". Par contre, si un événement omis existe, bien souvent elle questionne "Qui est-ce qui te l'a dit?", en ayant l'impression que l'enquêteur sait déjà quelque chose. La fiche «AGEVEN» apparaît comme un instrument de "divination" tels les cauris. Parfois, l'entretien revêt un aspect ludique, la personne interrogée est contente de pouvoir remettre de l'ordre dans le déroulement d'événements passés; une femme à la vie génésique et matrimoniale complexe a même voulu une copie de sa fiche «AGEVEN». Comme toutes les enquêtes l'utilisation de cette fiche se heurte à certaines difficultés. Il est parfois difficile ou délicat de s'isoler avec la personne enquêtée et bien souvent les femmes sont gênées si la fiche met en évidence des événements concernant un partenaire précédant le mari actuel.

À l'usage cette fiche est incomplète car il manque une question permettant d'éviter les risques de confusion entre mort-né et enfant décédé juste après la naissance, confusion entretenue par les réponses en wolof. En effet, dans cette langue, il est aussi difficile de faire la distinction entre fausse couche et avortement qu'entre mort-né et décédé juste après la naissance. Certains termes ou mots français ne sont pas directement traduisibles en wolof. C'est ainsi que pour mort-né, il n'y a pas une question qui permette à elle seule d'obtenir la réponse souhaitée. Il faut donc au moins deux questions pour avoir l'information désirée. Pour ce qui est des intervalles intergénésiques plus ou moins longs, les enquêteurs posent la question suivante par exemple: "Lou am dikhane té Moussa ak Ali?": ("qu'est-ce qu'il y a eu entre Moussa et Ali?"). À cette interrogation, les femmes comprennent à juste titre mort-nés, avortements, fausses couches, etc. Pour avoir une réponse satisfaisante des précisions sont nécessaires: "Dikhane té Moussa ak Ali, amo fi dom diou dé guinaw bou mou indé bakhane?" ("entre Moussa et Ali, n'auriez-vous pas un enfant décédé après qu'il ait manifesté un signe quelconque de vie?"). La confusion est surtout entretenue par le fait que la distinction entre fausse couche et mort-né n'est pas toujours évidente et que l'enfant n'a de prénom qu'après une semaine. Aussi, selon certaines ethnies, ce n'est qu'à partir de ce moment qu'il est vraiment pris en compte. Une colonne précisant si l'enfant a ou nom crié au moment de la naissance aurait donc été la bienvenue.

La fiche «AGEVEN» de l'enquête Pikine a certes donné des informations plus satisfaisantes que le graphique utilisé lors de l'enquête sénégalaise fécondité de par la nature et la quantité des informations recueillies. Par contre elle n'a pas pour autant éliminé l'inconvénient de la tendance à arrondir en années les intervalles intergénésiques (environ 37% des intervalles) et plus particulièrement les intervalles de deux ans, qui représentent environ 20% des intervalles observés entre les naissances successives. De plus, cette technique se révèle insuffisante pour recenser toutes les issues de grossesses des jeunes filles ayant déjà été enceintes mais n'ayant eu aucune naissance vivante. Certains biais, certes classique en démographie, subsistent donc et cette méthode ne disperse d'une grande vigilence sur le terrain.

4. LA TRANSCRIPTION DE LA FICHE «AGEVEN» SUR LE QUESTIONNAIRE ET SON EXPLOITATION INFORMATIQUE

Le questionnaire concernant la vie génésique des femmes a été conçu de façon à permettre la meilleure transcription possible des informations recueillies sur la fiche «AGEVEN». Dans un premier temps, les caractéristiques propres à chacun des enfants sont notées dans l'ordre chronologique des naissances, ainsi que les dates de décès s'il y a lieu. Dans un second temps, l'enquêteur devait relever la situation matrimoniale au moment de chacun des événements afin de suivre les changements éventuels de conjoint. Ensuite étaient prises en compte l'évolution de la situation socio-économique du père, celle de la mère, ainsi que l'évolution des conditions d'habitat et les différents lieux de résidence. L'ensemble de l'enquête comprenait d'autres questionnaires concernant les caractéristiques du ménage, des individus, et des femmes soumises à l'observation.

Le mode de recueil de l'information permet deux types d'analyse. L'une concerne l'analyse classique de la mortalité par génération et par sous-population (selon le quartier, le type de logement, etc...). Mais l'intérêt majeur de cette étude est de permettre l'analyse de la mortalité (et de la fécondité) en tenant compte des comportements migratoires et de l'évolution des conditions socio-économiques des femmes soumises à l'enquête. Grâce à cette méthode, la mortalité n'est plus interprétée en fonction des seules conditions socio-économiques au moment de l'enquête, mais au contraire, elles est rapportée aux conditions réellement vécues au moment de l'événement, et l'on peut de cette façon mieux appréhender les différences propres aux conditions de vie en milieu urbain (Pikine dans notre cas).

Tableau 1

Mortalité selon le lieu de naissance (en pour mille)

	Pikine	Dakar	Autres Villes	Rural	Total	Test Pkn-Rural
Infantile	52	57	45	114	58	-6,586**
Junévile	55	62	90	156	68	-10,093**
Effectif	5155	1513	644	704	8016	

Suivant le lieu de naissance de l'enfant, des niveaux différents de mortalité ont été relevés. De nombreuses mères observées à l'enquête, sont des migrantes provenant d'une autre ville ou d'un village de l'intérieur du pays. Parmi leurs enfants, ceux nés en milieu rural présentent des risques de mortalité bien plus grands que ceux nés dans l'agglomération dakaroise.

Le quotient de mortalité juvénile (entre 1 et 4 ans) traduit bien les risques liés aux disparités socio-économiques. Le risque de mourir entre 1 et 4 ans est 2,84 fois plus élevé pour les enfants nés dans les villages que pour ceux nés à Pikine. Le "test Z" montre que la différence entre les deux quotients (quotient de mortalité Pikine et quotient de mortalité rural) est significative. On test l'hypothèse que le quotient de mortalité des enfants nés à Pikine est le même que celui des enfants nés en milieu rural. Les tailles d'échantillon étant assez élevées, l'approximation par une loi normale est justifiée. Le test basé sur la statistique Z admet la loi de probabilité d'une normale centrée réduite sous l'hypothèse d'égalité des quotients de mortalité. Le signe "**" indique une différence significative au niveau de $\alpha = 0.05$. Un recueil rétrospectif classique sans distinction du lieu de naissance de l'enfant nous aurait amené à assimiler les naissances hors Pikine à celles de l'agglomération observée et induit un niveau plus élevé de mortalité (mortalité juvénile de 68 pour mille au lieu de 55 pour mille).

Par ailleurs, une seconde analyse est possible pour chacune des femmes observées; on peut constituer un fichier biographique simplifié, dont les étapes successives sont délimitées par les naissances. Une relation est ainsi établie entre les événements matrimoniaux, les changements de résidence, et les données génésiques. On peut également reconstituer les principales étapes du cheminement migratoire depuis la naissance du premier enfant, ou bien depuis le mariage. Les données longitudinales recueillies de cette façon se prêtent bien aux méthodes récentes d'analyse des interférences entre phénomènes (Courgeau et Lelievre (1986), Cox et Oakes (1984)).

5. CONCLUSION

L'information recueillie pour chacune des variables est très succincte mais elle doit permettre de dégager quelques différences importantes, et de recueillir des informations sur les conditions de vie à la naissance et au moment du décès. La méthodologie de collecte retenue est adaptée à la collecte des données concernant la vie génésique des femmes et le devenir de leurs enfants. L'intérêt majeur de la fiche «AGEVEN» réside dans la facilité apportée au repérage dans le temps de certains événements et dans le classement des événements les uns par rapport aux autres, en gardant la possibilité d'insérer les événements omis au fur et à mesure de la conversation. La souplesse d'utilisation de la fiche «AGEVEN» nous amène à suggérer se transposition à d'autres domaines comme celui des biographies professionnelles, ou des itinéraires migratoires en mettant en parallèle localité de résidence, profession, situation matrimoniale, situation familiale, conditions de logement... Le domaine de l'analyse

des biographies en démographie a fait l'objet de nombreuses recherches méthodologiques (Courgeau 1984; Haeringer 1972; Riandey 1985). Notre méthode se veut uniquement un outil simple et fiable d'aide à la collecte des données. À chaque utilisateur de retenir les variables à classer dans le temps sur la fiche «AGEVEN», et une fois la trame biographique collectée approfondir le(s) domaine(s) étudié(s) à l'aide du questionnaire.

REMERCIEMENTS

Les auteurs tiennent à remercier les arbitres pour leurs commentaires utiles.

BIBLIOGRAPHIE

ANTOINE, Ph., et DIOUF, P.D. (1986). Changements démographiques en milieu urbain. Communication présentée au Séminaire sur la mortalité au Sénégal. Dakar.

BONNET, D. (1984). Occultation, omissions. Quelques problèmes soulevés par l'enquête quantitative en matière de santé. *Medicus Mundi,* 11.

COURGEAU, D. (1984). Relations entre cycle de vie et migrations. *Population,* 39, 483-513.

COURGEAU, D., et LELIEVRE, E. (1986). Nuptialité et agriculture. *Population,* 41, 303-326.

COX, R., et OAKES, D. (1984). *Analysis of Survival Data.* London: Chapman & Hall.

DIRECTION DE LA STATISTIQUE, 1981: *Enquête Sénégalaise sur la Fécondité, 1978 - Rapport National d'Analyse,* 1.

FARGUES, Ph. (1985). L'évaluation du niveau de la mortalité à partir des données des enquêtes EMIJ. *Les enquêtes sur la mortalité infantile et juvénile (EMIJ),* 1, 60-84.

FERRY, B. (1977). Le fichier événement. Une nouvelle méthode d'observation rétrospective. Dans *l'Observation démographique dans les pays à statistiques déficientes.* Liege, Belgique: Ordina Editions, 137-150.

HAERINGER, Ph. (1972). Méthodes de recherche sur les migrations africaines. Un modèle d'interview biographique et sa transcription synoptique. *Cahiers ORSTOM,* 9, 439-453.

SCOTT, Ch. (1985). Les problèmes de déperdition dans les enquêtes suivies. Dans *Les enquêtes sur la mortalité infantile et juvénile (EMIJ),* 1, 44-47.

RIANDEY, B. (1985). L'enquête "biographie familiale professionnelle et migratoire" (INED, 1981). Le bilan de la collecte. Dans *Migrations internes, collecte des données et méthode d'analyse.* Université de Louvain, 117-134.

Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 183–191
Statistique Canada

Méthode alternative pour ajuster les estimations de la Current Population Survey aux chiffres de population

K.R. COPELAND, F.K. PEITZMEIER, et C.E. HOY[1]

RÉSUMÉ

La Current Population Survey utilise la méthode itérative du quotient dans l'estimation par stratification a posteriori pour redresser les estimations démographiques d'échantillon en fonction d'estimations démographiques fondées sur le recensement. Dans cet article, les auteurs proposent une deuxième méthode, fondée sur les moindres carrés généralisés, et la comparent à la méthode actuelle.

MOTS CLÉS: Moindres carrés généralisés; stratification a posteriori; estimation itérative du quotient.

1. INTRODUCTION

La Current Population Survey (CPS) produit des estimations de la population active pour l'ensemble de la population civile hors institutions d'âge actif aux États-Unis à partir d'un échantillon aléatoire mensuel à plusieurs degrés de quelque 60,000 logements. Chaque mois, un échantillon avec renouvellement composé de 8 panels (appelés groupes de renouvellement) de logements est interviewé afin de recueillir des données démographiques et des données sur l'activité pour tous les occupants adultes civils des logements échantillonnés.

Les estimations, sous-agrégées selon des caractéristiques démographiques, sont publiées mensuellement. Des estimations concernant d'autres sous-agrégats de la population (états, familles, anciens combattants, salariés, inactifs, etc.) sont aussi publiées mensuellement, trimestriellement ou annuellement.

On calcule le poids des personnes de l'échantillon en utilisant la probabilité de sélection, un facteur de compensation de la non-réponse et un facteur de correction sous forme de quotient afin de réduire la composante de la variance attribuable au prélèvement d'unités primaires d'échantillonnage. On corrige ensuite les poids ainsi obtenus au moyen d'une méthode d'estimation par stratification a posteriori de manière à rapprocher les estimations démographiques de l'enquête d'estimations démographiques supplémentaires. Enfin, les poids ainsi corrigés sont utilisés dans une méthode d'estimation composite, puis désaisonnalisés afin d'obtenir des estimations nationales (Hanson 1978).

Pour certains sous-domaines de population (familles, salariés, inactifs, revenus de la famille et anciens combattants), le calcul d'estimations détaillées requiert des poids d'échantillon déterminés par des méthodes de correction qui viennent s'ajouter à l'estimation par stratification a posteriori.

La méthode des moindres carrés généralisés (MCG) pourrait peut-être remplacer l'estimation par stratification a posteriori ou servir à intégrer les diverses méthodes de correction de la CPS. On a déjà proposé l'application des MCG dans l'enquête sur les dépenses des consommateurs et des analyses ont été faites à ce sujet (Zieschang 1986).

Nous nous proposons dans cet article de comparer la méthode d'estimation par stratification a posteriori utilisée actuellement dans la CPS (méthode itérative du quotient) et la méthode MCG en nous fondant sur les données de la CPS pour deux mois particuliers (juillet 1983

[1] K.R. Copeland, F.K. Peitzmeier et C.E. Hoy, Division of Statistical Methods, Office of Employment and Unemployment Statistics, Bureau of Labor Statistics, Washington, D.C. 20212, États-Unis.

et juillet 1984). Afin d'apprécier les différences qui pouvaient exister entre les deux méthodes pour cette application, nous nous sommes intéressés aussi bien aux macro-données qu'aux micro-données.

2. MÉTHODE D'ESTIMATION PAR STRATIFICATION À POSTERIORI UTILISÉE ACTUELLEMENT DANS LA CPS

Dans la CPS, on utilise actuellement la méthode itérative du quotient (MIQ) pour corriger les poids d'échantillon à l'intérieur d'un groupe de renouvellement de manière à ajuster les estimations démographiques de l'échantillon à des estimations démographiques supplémentaires dans trois catégories distinctes (état, âge/sexe/origine ethnique, âge/sexe/origine raciale).

Deming et Stephan (1940) ont été les premiers à proposer l'application de la MIQ pour la correction des données de tableaux en remplacement de la méthode des moindres carrés. On a montré que la MIQ produisait de meilleurs estimateurs asymptotiquement normaux (MEAN) dans un échantillonnage aléatoire simple et qu'elle minimisait, comme nous le verrons dans la sous-section 4.2, les corrections apportées aux poids d'échantillon (Ireland et Kullback 1968). De plus, même si elle produit des estimations biaisées, la MIQ peut parfois réduire l'erreur quadratique moyenne des estimations d'enquête. C'est le rôle que l'on attribue à la MIQ dans la CPS (Hanson 1978).

En ce qui a trait à la CPS, la MIQ vise à corriger les chiffres d'échantillon $\{n_{ijk}\}$ découlant des étapes de pondération antérieures afin de produire des chiffres d'échantillon rajustés $\{\tilde{n}_{ijk}\}$ à la condition que

$$\text{(A)} \qquad \sum_{j,k} \tilde{n}_{ijk} = m_{i..}$$

$$\text{(B)} \qquad \sum_{i,k} \tilde{n}_{ijk} = m_{.j.}$$

$$\text{(C)} \qquad \sum_{i,j} \tilde{n}_{ijk} = m_{..k}$$

soient satisfaites simultanément,

où i = état ($i = 1, \ldots, 51$),
j = âge/sexe/origine ethnique ($j = 1, \ldots, 16$),
k = âge/sexe/origine raciale ($k = 1, \ldots, 70$),
$m_{i..}$ = estimation supplémentaire relative à l'état,
$m_{.j.}$ = estimation supplémentaire relative à la combinaison âge/sexe/origine ethnique,
$m_{..k}$ = estimation supplémentaire relative à la combinaison âge/sexe/origine raciale.

La MIQ permet de corriger proportionnellement, en se servant de quotients, les données d'échantillon selon les trois catégories de critères (c'est-à-dire, état, âge/sexe/origine ethnique et âge/sexe/origine raciale) par étapes successives.

(1) Correction par quotient selon l'état:

$$n_{ijk}^{(1,1)} = (m_{i..}/n_{i..})\, n_{ijk} = a_i^{(1)}\, n_{ijk}.$$

2) Correction par quotient selon la combinaison âge/sexe/origine ethnique:

$$n_{ijk}^{(1,2)} = (m_{.j.}/n_{.j.}^{(1,1)})\, n_{ijk}^{(1,1)} = b_j^{(1)}\, n_{ijk}^{(1,1)}$$
$$= a_i^{(1)}\, b_j^{(1)}\, n_{ijk}.$$

3) Correction par quotient selon la combinaison âge/sexe/origine raciale:

$$n_{ijk}^{(1,3)} = (m_{..k}/n_{..k}^{(1,2)})\, n_{ijk}^{(1,2)} = d_k^{(1)}\, n_{ijk}^{(1,2)}$$
$$= a_i^{(1)}\, b_j^{(1)}\, d_k^{(1)}\, n_{ijk},$$

où $n_{i..}$ = total de ligne pour l'échantillon
$n_{.j.}$ = total de colonne pour l'échantillon
$n_{..k}$ = total de couche pour l'échantillon.

L'exécution des trois étapes ci-dessus correspond à une itération de la méthode itérative du quotient. On reprend les trois étapes en substituant à chaque fois la valeur de $n_{ijk}^{(h,3)}$ (donnée-échantillon corrigée issue de la troisième étape de la h-ième itération) à n_{ijk} dans l'étape 1) jusqu'à ce qu'on ait effectué 6 itérations. (Le nombre d'itérations effectuées dans la CPS a été établi en fonction des propriétés de convergence de la MIQ pour cette enquête et des gains relatifs réalisés selon le nombre d'itérations.) La valeur finale $\{n_{ijk}^{(6,3)}\}$ est définie comme $\{\tilde{n}_{ijk}\}$.

Le facteur de correction des poids d'échantillon pour les enregistrements de la cellule $\{ijk\}$ est

$$F_{ijk} = n_{ijk}^{(6,3)}/n_{ijk}$$
$$= \prod_{h=1}^{6} a_i^{(h)}\, b_j^{(h)}\, d_k^{(h)}.$$

Pour obtenir les poids corrigés, on multiplie les poids d'échantillon observés avant l'application de la MIQ par le facteur F_{ijk} approprié.

3. APPLICATION DES MCG DANS LA CPS

La méthode des moindres carrés généralisés (MCG) permet de corriger les poids d'échantillon découlant des étapes de pondération antérieures en minimisant les carrés pondérés des corrections, sous réserve que les poids corrigés satisfassent à une série de contraintes "de contrôle" linéaires. C'est exactement ce à quoi Deming et Stephan cherchaient à s'attaquer lorsqu'ils ont mis au point la MIQ. Comme celle-ci, la méthode MCG produit des MEAN dans certaines conditions, en l'occurrence lorsque toutes les cellules sont non nulles (Neyman 1949). Par définition, la méthode MCG minimise les corrections apportées aux poids d'échantillon par l'intermédiaire d'une mesure de proximité (voir sous-section 4.2).

En ce qui a trait à la CPS, chaque dimension qui définit une série de chiffres de population dans la stratification a posteriori courante définira aussi une série de contraintes linéaires pour la méthode MCG. La fonction à minimiser est

$$f(\underline{F}) = (\underline{F} - \underline{P})'\, P_0^{-1}\, (\underline{F} - \underline{P})$$
$$= \sum_i (W_{2i} - W_{1i})^2 / W_{1i},$$

à la condition que $X'\underline{F} = \underline{N}$,

où $\underline{F}$ = vecteur ($n \times 1$) des poids finals calculés (W_{2i}) pour chacune des n personnes de l'échantillon,

$\underline{P}$ = vecteur ($n \times 1$) des poids des personnes de l'échantillon avant la stratification a posteriori (W_{1i}),

$\underline{P}_0$ = matrice diagonale ($n \times n$) dont la diagonale est formée des W_{1i},

$\underline{X}$ = matrice descriptive ($n \times k$), où les lignes correspondent aux personnes de l'échantillon et les colonnes aux cellules de contrôle. La matrice est formée de zéros (0) et de uns (1), qui indiquent les catégories de contrôle appropriées pour chacune des n personnes de l'échantillon,

$\underline{N}$ = vecteur ($k \times 1$) des estimations démographiques supplémentaires, qui correspond aux colonnes de X. Ces estimations servent aussi à la MIQ dans la CPS.

Les colonnes de X doivent être linéairement indépendantes pour qu'il existe un inverse de la matrice $(X' P_0 X)$. Lors de la formation des matrice X et $\underline{N}$ pour la CPS, les 137 cellules de contrôle utilisées pour la MIQ (état, âge/sexe/origine ethnique, âge/sexe/origine raciale) ont été réduites à $k = 132$ cellules qui étaient linéairement indépendantes.

La seule solution de $X'\underline{F} = \underline{N}$ qui minimise $f(\underline{F})$ est (voir Luery 1986)

$$\underline{F} = \underline{P} + P_0 X (X' P_0 X)^{-1} (\underline{N} - X' \underline{P})$$

Bien qu'il ne soit pas nécessaire que les éléments de $\underline{F}$ soient positifs, ils l'étaient tous dans ce cas sans que l'on ait eu à poser des conditions additionnelles. Huang et Fuller (1978) et Zieschang (1986), entre autres, analysent des méthodes qui permettent d'obtenir des poids non négatifs dans ce contexte.

4. RÉSULTATS

4.1 Macro-données

a. Estimations

Les estimations de la population active ont été établies pour plusieurs groupes démographiques pour juillet 1983 et juillet 1984 au moyen des poids finals calculés par la MIQ et les MCG. On a aussi calculé les erreurs types pour les deux méthodes au moyen d'un estimateur de la méthode des groupes aléatoires défini comme suit (Wolter 1985):

$$\sum_{k=1}^{8} (8Y_k - \hat{Y})^2/56,$$

où Y_k = somme des poids des enregistrements de l'échantillon pour le k-ième groupe de renouvellement avec la caractéristique Y,

$\hat{Y}$ = somme des Y_k.

On a utilisé cet estimateur de variance, même s'il ne tient pas compte du plan de sondage à plusieurs degrés de la CPS, parce que le fichier de micro-données à grande diffusion de la CPS ne contenait aucune information sur le plan de sondage.

On a calculé des différences relatives pour les estimations de niveau et les estimations d'erreur type. La formule de la différence relative était

$$(Y_{GLS} - Y_{RRE}) / Y_{RRE},$$

où Y_{MIQ} = estimation de Y fondée sur les poids calculés selon la MIQ,

Y_{MCG} = estimation de Y fondée sur les poids calculés selon la méthode des MCG.

Tableau 1

Estimations de la population active selon la combinaison sexe/origine raciale ou ethnique

		1983				1984			
		MCG		(MCG-MIQ)/ MIQ		MCG		(MCG-MIQ)/ MIQ	
		Total (milliers)	S.E. (milliers)	Total (%)	S.E. (%)	Total (milliers)	S.E. (milliers)	Total (%)	S.E. (%)
Total									
Total	Occupées	103516	403	0.00	−0.14	107535	352	−0.01	1.12
	En chômage	10669	221	−0.04	−0.75	8765	118	−0.06	−0.21
	Taux	*9.34%*	*0.19%*	*−0.04*	*−0.56*	*7.54%*	*0.09%*	*−0.05*	*0.27*
	Inactives	59938	373	0.01	−0.68	60080	419	0.02	0.41
Personnes de race blanche	Occupées	91338	344	0.00	−0.33	94417	274	0.00	0.70
	En chômage	7928	236	0.00	−0.27	6282	120	0.00	−0.14
	Taux	*7.99%*	*0.23%*	*0.00*	*−0.26*	*6.24%*	*0.10%*	*0.00*	*−0.16*
	Inactives	51915	340	0.00	−0.36	51700	358	0.00	0.39
Personnes de race noire	Occupées	9871	69	0.06	−3.44	10371	98	0.02	0.17
	En chômage	2434	68	−0.12	−1.07	2202	60	−0.03	1.41
	Taux	*19.78%*	*0.55%*	*−0.14*	*−1.60*	*17.51%*	*0.42%*	*−0.04*	*1.49*
	Inactives	6628	26	−0.04	−1.47	6765	109	−0.02	0.09
Personnes d'origine hispanique	Occupées	6132	73	−0.03	−0.59	6607	102	−0.03	1.90
	En chômage	920	79	−0.05	−0.29	786	70	−0.08	−0.03
	Taux	*13.04%*	*1.10%*	*−0.02*	*−0.33*	*10.63%*	*0.96%*	*−0.05*	*0.35*
	Inactives	3760	31	0.05	−0.39	3786	73	0.04	1.02
Hommes									
Total	Occupées	58985	147	0.00	−1.58	61045	188	0.00	1.74
	En chômage	5980	134	−0.05	−0.88	4682	79	−0.02	0.77
	Taux	*9.20%*	*0.19%*	*−0.05*	*−0.79*	*7.12%*	*0.11%*	*−0.02*	*1.30*
	Inactives	17495	178	0.01	−1.81	17840	214	0.02	0.64
De race blanche	Occupées	52674	482	0.00	0.42	54261	111	0.00	0.34
	En chômage	4484	131	0.01	−0.49	3394	93	0.01	−0.12
	Taux	*7.84%*	*0.21%*	*0.00*	*−0.47*	*5.89%*	*0.15%*	*0.01*	*−0.13*
	Inactives	14985	160	−0.02	−0.40	15077	150	0.00	0.16
De race noire	Occupées	5047	56	0.07	−1.70	5263	84	0.01	−0.50
	En chômage	1300	45	−0.20	−1.87	1137	33	0.08	1.12
	Taux	*20.49%*	*0.71%*	*−0.21*	*−2.02*	*17.76%*	*0.51%*	*0.05*	*0.94*
	Inactives	2097	40	−0.04	−0.13	2236	88	−0.07	−0.48
D'origine hispanique	Occupées	3781	48	0.01	−0.86	4064	79	−0.02	1.29
	En chômage	534	45	−0.16	−0.83	451	41	−0.05	0.51
	Taux	*12.38%*	*0.99%*	*−0.15*	*−0.89*	*9.99%*	*0.95%*	*−0.03*	*0.66*
	Inactives	981	42	0.00	−0.42	964	57	0.07	1.40
Femmes									
Total	Occupées	44531	320	−0.01	−0.01	46490	194	−0.01	1.48
	En chômage	4689	107	−0.04	−0.19	4083	88	−0.10	−1.22
	Taux	*9.53%*	*0.23%*	*−0.03*	*−0.02*	*8.07%*	*0.16%*	*−0.09*	*−0.80*
	Inactives	42443	287	0.01	−0.26	42240	217	0.02	0.34
De race blanche	Occupées	38664	315	0.00	−0.29	40156	191	0.00	0.66
	En chômage	3444	115	−0.01	0.16	2888	68	0.00	−0.32
	Taux	*8.18%*	*0.28%*	*−0.01*	*0.11*	*6.71%*	*0.15%*	*0.00*	*−0.34*
	Inactives	36929	283	0.01	−0.32	36623	214	0.00	0.53
De race noire	Occupées	4824	57	0.05	0.56	5108	50	0.02	1.69
	En chômage	1134	46	−0.02	0.07	1065	46	−0.14	−0.62
	Taux	*19.03%*	*0.80%*	*−0.06*	*0.08*	*17.25%*	*0.67%*	*−0.13*	*−0.63*
	Inactives	4531	24	−0.04	2.99	4529	59	0.01	1.49
D'origine hispanique	Occupées	2350	44	−0.08	−0.46	2543	38	−0.05	3.04
	En chômage	385	41	0.10	0.51	335	34	−0.13	−0.62
	Taux	*14.08%*	*1.46%*	*0.16*	*0.57*	*11.64%*	*1.18%*	*−0.07*	*−0.11*
	Inactives	2778	33	0.07	−0.87	2822	27	0.03	0.13

Comme l'indiquent les données du tableau 1, aucune différence ou tendance notable ne se dégage des estimations pondérées de la population active ou des estimations de l'erreur type calculées par l'une et l'autre des deux méthodes lorsque ces estimations sont sous-agrégées en fonction du sexe et de l'origine raciale ou ethnique.

En ce qui concerne les estimations de la population active selon la combinaison sexe/origine raciale ou ethnique, les différences relatives estimées (en valeur absolue) étaient toutes inférieures à 0.3% (ce qui est bien au-dessous des CV estimés de chaque estimation). Pour la majorité de ces estimations, notamment celles ayant trait à l'ensemble des personnes de sexe masculin ou de sexe féminin ou aux personnes de race blanche, la différence relative en valeur absolue était inférieure à 0.1%.

Pour un bon nombre de caractéristiques, le signe de la différence relative a changé de 1983 à 1984; l'écart entre les estimations produites par les deux méthodes ne semble donc pas suivre une tendance particulière.

En ce qui concerne les estimations de l'erreur type pour les estimations nationales de la population active, les différences relatives en valeur absolue étaient toutes inférieures à 1.9% pour la population totale, à 0.7% pour les personnes de race blanche, à 3.5% pour les personnes de race noire et à 3.1% pour les personnes d'origine hispanique.

b. Indices de l'effet du nombre de mois d'inclusion dans l'échantillon

De nombreuses études montrent que les estimations calculées à l'aide des poids finals de la CPS sont entachées d'un biais relatif qui varie selon le nombre de mois d'inclusion du groupe de renouvellement dans l'échantillon (Bailar 1975). Des indices de l'effet du nombre de mois d'inclusion dans l'échantillon

$$I_k = (8Y_k / \hat{Y}) \times 100,$$

ont été calculés pour juillet 1983 et juillet 1984 d'après les estimations calculées par la MIQ et les MCG.

Dans les deux cas, on a obtenu des indices virtuellement identiques pour la population active selon l'origine raciale, selon le sexe et selon l'origine ethnique.

4.2 Micro-données

a. Correction des poids d'échantillon

La MIQ aussi bien que les MCG minimisent une mesure de proximité des poids d'échantillon (poids initial vs poids corrigé). Dans le cas de la MIQ, la mesure est (Ireland et Kullback 1968)

$$M_A = \sum_i W_{2i} \ln(W_{2i} / W_{1i}).$$

Pour les MCG, la mesure est (Luery 1986)

$$M_B = \sum_i (W_{2i} - W_{1i})^2 / W_{1i}$$

où W_{1i} = poids de l'enregistrement i de l'échantillon avant correction,
W_{2i} = poids de l'enregistrement i de l'échantillon après correction.

La comparaison des mesures de proximité (voir le tableau 2) donne des résultats intéressants et parfois étonnants. Ainsi, la MIQ a produit des valeurs moindres pour les deux mesures. La méthode des MCG avait tendance à produire des valeurs moindres pour certains sous-groupes, notamment les personnes de race noire et les personnes d'origine hispanique. Il convient de souligner que l'écart entre les valeurs produites par les deux méthodes était presque toujours inférieur à 1%.

Tableau 2

Comparaison des mesures de proximité fondée sur 8 groupes de renouvellement pour chaque année (nombre de GR pour lesquels MIQ < MCG)

	M_A		M_B	
	1983	1984	1983	1984
Total	8	8	4	7
De race blanche	7	7	3	4
De race noire	3	3	1	1
D'origine hispanique	0	0	0	0
Hommes	2	7	1	5
Femmes	8	8	8	8

Bien que M_B devrait normalement être minimisée par la méthode des MCG, la valeur de M_B fondée sur les poids calculés selon cette méthode pour l'échantillon total était plus élevée que la valeur de M_B fondée sur les poids calculés selon la MIQ dans 11 cas sur 16.

En cherchant à expliquer cette contradiction apparente, nous avons remarqué qu'après six itérations, les estimations produites par la MIQ ne concordaient toujours pas avec les chiffres de population pour l'âge, le sexe et l'origine ethnique. Le degré de "non-convergence" est toutefois *très faible* (moins de 1.0% pour toutes les catégories de contrôle). Cependant, compte tenu de l'écart entre les valeurs de M_B fondées respectivement sur la MIQ et les MCG, il suffirait de modifier les poids calculés selon la MIQ dans une proportion de 0.1 à 0.2% pour obtenir le résultat inverse. En soumettant la MIQ à 15 itérations, qui n'ont pas suffi néanmoins à assurer la convergence, nous avons pu constater que les résultats observés pour M_B pouvaient être attribuables à la légère non-convergence de la MIQ. (Il convient de souligner que la méthode des MCG minimise M_B pour la série de méthodes de correction qui produisent des estimations conformes aux chiffres de population. Comme les estimations produites par la MIQ lors de la CPS ne concordent pas avec les chiffres de population, elle ne fait pas partie de la catégorie précitée.)

Bien qu'une méthode de correction comme la MIQ ou les MCG puisse minimiser une mesure de proximité pour l'échantillon global, il n'est pas dit qu'elle fera la même chose pour des sous-agrégats de l'échantillon qui ont été comparés à des chiffres de population (par exemple, personnes de race noire, personnes d'origine hispanique, personnes de sexe masculin). Compte tenu de l'utilisation de chiffres de population et du fait que la mesure globale de proximité est minimisée, il semblerait souhaitable d'avoir une méthode de correction qui produise aussi de faibles mesures de proximité pour les sous-agrégats. La méthode des MCG a produit de telles mesures dans presque tous les groupes de renouvellement pour les personnes d'origine hispanique, dans de nombreux groupes de renouvellement pour les personnes de race noire et dans plusieurs groupes de renouvellement pour les personnes de race blanche et de sexe masculin.

b. Comparaison des corrections

La MIQ et la méthode MCG définissent toutes deux des facteurs de correction pour les cellules formées par l'intersection des contraintes marginales. Le même facteur s'applique à tous les enregistrements d'une cellule. Afin de comparer les corrections apportées aux poids par l'une ou l'autre des deux méthodes, nous avons comparé les facteurs définis par chacune des méthodes pour chaque enregistrement de l'échantillon à l'aide du rapport suivant:

$$RRE/GLS = [(W_{2i}/W_{1i})_{RRE}] / [(W_{2i}/W_{1i})_{GLS}] .$$

Tableau 3

Comparaison des corrections apportées par la MIQ et les MCG, 1984

Catégorie de contrôle	Intervalle de taux de couverture	Proportion de l'échantillon total	Proportion de l'échantillon pour laquelle MIQ/MCQ <0.95 ou >1.05	Proportion des enregistrements dans l'intervalle pour lesquels MIQ/MCQ <0.95 ou >1.05
Âge/sexe/ origine raciale	<0.7	0.007	0.057	0.219
	0.7-0.8	0.022	0.116	0.136
	0.8-0.9	0.241	0.147	0.019
	0.9-1.1	0.699	0.504	0.019
	1.1-1.2	0.021	0.069	0.084
	>1.2	0.010	0.106	0.275
Âge/sexe/ origine ethnique	<0.7	0.010	0.078	0.198
	0.7-0.8	0.014	0.032	0.058
	0.8-0.9	0.106	0.135	0.033
	0.9-1.1	0.869	0.741	0.022
	1.1-1.2	0.001	0.007	0.202
	>1.2	0.001	0.007	0.373
État	<0.7	0.056	0.068	0.031
	0.7-0.8	0.111	0.180	0.042
	0.8-0.9	0.278	0.325	0.030
	0.9-1.1	0.479	0.342	0.018
	1.1-1.2	0.026	0.009	0.009
	<1.2	0.049	0.077	0.040

Ce rapport décrit la relation entre les corrections apportées au poids d'une personne de l'échantillon par la MIQ d'une part et par la méthode MCG d'autre part. Pour les besoins de la comparaison, nous dirons qu'un rapport inférieur à 0.95 ou supérieur à 1.05 représente un écart significatif entre les corrections apportées par l'une et l'autre des deux méthodes.

Pour chaque série de chiffres de population, nous avons calculé des ratios E/C (taux de couverture), où E est l'estimation d'échantillon fondée sur les poids des personnes de l'échantillon avant la stratification a posteriori et C est le chiffre de population.

Pour chaque catégorie de contrôle (état, âge/sexe/origine ethnique, âge/sexe/origine raciale), nous avons classé les enregistrements de l'échantillon selon le taux de couverture. Le tableau 3 donne la répartition de l'échantillon selon l'intervalle de taux de couverture et la valeur du rapport *MIQ/MCG*, de même que la proportion des enregistrements dans chaque intervalle de taux de couverture pour lesquels le rapport *MIQ/MCG* est inférieur à 0.95 ou supérieur à 1.05.

Les données de ce tableau indiquent que pour chaque catégorie de contrôle, la MIQ et la méthode MCG étaient plus susceptibles de produire des corrections divergentes dans le cas d'enregistrements tirés de groupes de population qui avaient été surdénombrés ou sousdénombrés jusqu'à un certain point dans l'enquête (c'est-à-dire pour lesquels le taux de couverture n'était pas près de 1) que dans le cas d'enregistrements tirés de groupes de population qui avaient été dénombrés normalement.

4.3 Ressources informatiques

L'application des deux méthodes a été exécutée sur un Système 370 d'IBM au National Institute of Health au moyen de PROC MATRIX du système SAS. La préparation des fichiers

et la pondération ont exigé un temps d'utilisation du processeur environ trois fois plus long dans le cas de la méthode MCG que dans le cas de la MIQ. La méthode MCG a nécessité aussi plus de temps pour la mise en mémoire des fichiers. (Les matrices utilisées pour la CPS ont des dimensions très appréciables, par exemple $\underline{P}$, P_0, X, et $\underline{N}$ comptent environ 14,000 lignes pour chaque groupe de renouvellement.)

5. RÉSUMÉ ET CONCLUSIONS

Cet article avait pour but de comparer la MIQ et la méthode MCG dans le cadre de la CPS tant au point de vue des macro-données qu'au point de vue des micro-données.

En ce qui concerne les macro-données, les deux méthodes produisent des estimations comparables.

Les mesures de proximité ont permis de constater que les corrections apportées aux poids d'échantillon pour respecter les contraintes de contrôle étaient légèrement inférieures dans le cas de la MIQ. En revanche, celle-ci avait tendance à produire des mesures de proximité légèrement plus élevées (par rapport aux MCG) pour les sous-agrégats de minorités ethniques. Là où les deux méthodes diffèrent le plus l'une de l'autre a trait aux corrections qui sont apportées aux groupes de population qui sont surdénombrés ou sousdénombrés.

Cette analyse nous permet de déduire que la MIQ exige moins de temps machine que la méthode MCG pour la correction du second degré de la CPS.

REMERCIEMENTS

Les auteurs tiennent à exprimer leur reconnaissance à Fritz Scheuren, qui a révisé la version originale de cet article, de même qu'aux arbitres et au rédacteur associé qui, par leurs commentaires très utiles, ont contribué à améliorer la qualité de cet article.

BIBLIOGRAPHIE

BAILAR, B. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

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

HANSON, R.H. (1978). The Current Population Survey design and methodology. Technical Paper 40, U.S. Bureau of the Census.

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

IRELAND, C.T., et KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.

LUERY, D. (1986). Weighting sample survey data under linear constraints on the weights. *Proceedings of the Section on Social Statistics, American Statistical Association*, 325-350.

NEYMAN, J. (1949). Contribution to the Theory of the X^2 Test. Dans *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*, (éd. J. Neyman), Berkeley: University of California Press, 239-273.

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

ZIESCHANG, K.D. (1986). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 64-71.

Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 193–209
Statistique Canada

Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages

CHARLES H. ALEXANDER[1]

RÉSUMÉ

Nous analysons une série de méthodes fondées sur la notion de "distance minimum conditionnelle" et qui assurent la concordance des poids des ménages avec les données supplémentaires concernant l'effectif de diverses cellules d'âge × race × sexe. Les poids conditionnels se rapprochent autant que possible des poids initiaux, qui sont fondés sur l'inverse de la probabilité de sélection. Cette série de méthodes comprend la méthode itérative du quotient et la méthode des moindres carrés généralisés de même que la méthode multinomiale du maximum de vraisemblance, où les cellules de la distribution représentent des catégories de ménages. À l'aide de modèles d'observation simples, nous étudions les propriétés de ces méthodes dans une situation de sous-dénombrement systématique des catégories de ménages. Après une comparaison avec la méthode de la personne principale, nous concluons à la nécessité de mieux connaître la nature du sous-dénombrement avant de décider de la méthode la plus appropriée (méthodes fondées sur la distance minimum conditionnelle ou méthode de la personne principale).

MOTS CLÉS: Pondération; information supplémentaire; méthode itérative du quotient; méthode de la personne principale; champ d'enquête.

1. INTRODUCTION

La post-stratification est souvent utilisée pour rajuster les poids de sondage en fonction de données supplémentaires sur le nombre d'unités de certaines catégories dans la population. Par exemple, on peut obtenir des estimations indépendantes de la population pour diverses cellules de post-stratification portant sur l'âge, l'origine raciale et le sexe en rajustant les chiffres du recensement en fonction des variations démographiques observées depuis le recensement. Ces estimations indépendantes sont souvent appelées "chiffres de population". Avant la stratification a posteriori, chaque personne (ou ménage) de l'échantillon a déjà un poids qui correspond habituellement à l'inverse de sa probabilité de sélection. Dans la post-stratification, chaque poids est multiplié par un facteur de correction sous forme de ratio qui s'applique à chaque cellule de sorte que la somme des poids corrigés pour une cellule est égale à son chiffre de population. Cette correction est particulièrement importante lorsqu'il y a un sous-dénombrement systématique des ménages ou des personnes à l'intérieur des ménages.

Dans la plupart des enquêtes démographiques du U.S. Census Bureau, la post-stratification sert à attribuer des poids aux personnes de l'échantillon mais non aux ménages, car il est plus difficile d'obtenir des estimations indépendantes pour les ménages. Pour attribuer des poids aux ménages dans ces enquêtes, on utilise plutôt une version de la méthode de la personne principale. Selon la version fondamentale de cette méthode, on attribue aux ménages un poids équivalent à celui attribué par post-stratification à la personne principale du ménage. Dans la section 2, nous exposons les règles qui permettent de définir la personne principale d'un ménage. En utilisant les poids attribués aux personnes après la post-stratification, la méthode de la personne principale permet d'incorporer les estimations indépendantes du nombre de personnes dans les poids attribués aux ménages.

[1] Charles H. Alexander, Statistical Methods Division, Bureau of the Census, Washington, D.C. 20233, U.S.A.

Toutefois, cette méthode a un inconvénient majeur. En effet, lorsqu'on se sert des poids de ménage obtenus par cette méthode pour calculer des estimations pondérées du nombre de personnes dans chaque cellule de post-stratification, chaque personne étant affectée du poids du ménage auquel elle appartient, ces estimations ne correspondent pas aux chiffres de population utilisés dans la post-stratification. C'est pourquoi on s'est intéressé à des méthodes de pondération des ménages qui produisent des poids à partir desquels on obtient nécessairement des estimations de population conformes aux chiffres de population.

Le présent article vise à exposer une série de méthodes de pondération des ménages qui produisent des poids à partir desquels on obtient nécessairement des estimations conformes aux chiffres de population des diverses cellules. Le principe de base de ces méthodes est de déterminer des poids de ménage qui satisfont aux conditions et qui se rapprochent le plus possible du vecteur de poids attribués initialement aux ménages. Les méthodes analysées dans ce document sont des façons de mesurer la distance entre le vecteur de poids initial et le vecteur de poids corrigé.

Dans la section 2, nous décrivons six méthodes de pondération fondées sur la notion de "distance minimum conditionnelle" et une version de la méthode de la personne principale. Trois des six méthodes précitées ont déjà fait l'objet d'analyses; les autres sont présentées pour donner une image plus complète de la question. La section 3 porte sur le calcul des poids. Dans la section 4, nous voyons comment la composition du ménage influe sur la correction des poids. Dans la section 5, nous analysons les résultats et présentons des exemples qui peuvent aider à comprendre le mécanisme de ces méthodes. Enfin dans la section 6, nous proposons des sujets de recherche.

Cet article n'est pas le premier du genre. Luery (1986) propose d'appliquer la série de méthodes fondées sur la notion de distance minimum conditionnelle à la pondération des ménages. Poursuivant le travail de Luery, Zieschang (1986a) propose d'appliquer l'une de ces méthodes, la méthode des moindres carrés généralisés, à la pondération dans les enquêtes sur les dépenses des consommateurs aux États-Unis. Une autre méthode de la série est la "méthode de l'information minimum discriminante", connue aussi sous le nom d'"estimation itérative par la méthode du quotient" ou simplement "méthode itérative du quotient". Oh et Scheuren (1978a) traitent particulièrement de l'application de cette méthode à la pondération des ménages et enrichissent la bibliographie déjà très abondante sur cette méthode d'estimation et les méthodes connexes. Deming et Stephan (1940) comptent parmi les premiers qui ont associé la méthode itérative du quotient à la notion de distance minimum conditionnelle. Les principes fondamentaux de cette approche sont analysés dans Ireland et Kullback (1968). Brackstone et Rao (1979) ont bien décrit l'application de cette méthode à la correction des poids de sondage. La série de méthodes comprend aussi deux fonctions de critère qui se rattachent à la méthode multinomiale du maximum de vraisemblance. La relation entre cette méthode et la méthode itérative du quotient a fait l'objet de nombreuses études; voir, par exemple, Bishop, Fienberg et Holland (1976). Fienberg (1986) souligne que les critères de distance analysés dans le présent article peuvent être considérés comme des cas spéciaux d'une famille de fonctions paramétriques analysées dans Cressie et Read (1984).

2. MÉTHODES FONDÉES SUR LA NOTION DE DISTANCE MINIMUM CONDITIONNELLE

2.1 Méthodes fondées sur les poids des ménages

Considérons un échantillon de K ménages, dont les poids initiaux sont définis par le vecteur $\underline{S} = (S_1, \ldots, S_K)$. Dans cet article, S_k désigne l'inverse de la probabilité de sélection du k-ième ménage; dans certaines applications, le poids initial peut inclure d'autres facteurs de correction, par exemple des facteurs de non-réponse.

Supposons qu'il y a J cellules de post-stratification et que l'on connaît le nombre de personnes (N_j) dans chacune. Par exemple, dans le cas de l'enquête sur les dépenses des consommateurs aux États-Unis, il y a $J = 48$ cellules formées par la combinaison des deux sexes, de deux origines raciales (noir, non noir) et de douze catégories d'âge. Les personnes de moins de 14 ans ne sont pas visées par cette enquête. Les chiffres de population pour ces cellules sont définis en l'occurrence par un vecteur $\underline{N} = (N_1, \ldots, N_J)$.

La composition des ménages de l'échantillon est définie par une matrice $A = (a_{kj})$, où a_{kj} est égal au nombre de personnes du k-ième ménage qui se trouvent dans la j-ième cellule de post-stratification. La sommation des cellules de post-stratification pour le k-ième ménage donne $a_{k.}$, soit le nombre total de personnes que compte le ménage. Le vecteur $(a_{k1}, \ldots, a_{kJ})$ décrit la composition du ménage k. Par exemple, un vecteur $(2, 1, 0, 0, \ldots, 0)$ signifie que deux personnes du ménage se trouvent dans la première cellule et une troisième dans la seconde cellule.

Étant donné les poids initiaux $\underline{S}$, l'estimation pondérée du nombre de personnes dans la cellule j sera $\hat{N}_j = \Sigma_k a_{kj} S_k$ ou, de façon générale, $\underline{\hat{N}} = A'\underline{S}$.

En règle générale, $\underline{\hat{N}} \neq \underline{N}$, autrement dit, l'estimation pondérée initiale du nombre de personnes dans les cellules de post-stratification peut différer de l'effectif réel de la cellule.

Nous cherchons ici à définir un nouveau vecteur de poids $\underline{W} = (W_1, \ldots, W_K)$ pour les ménages de l'échantillon de sorte que $\underline{N} = A'\underline{W}$ ou

$$\sum_k a_{kj} W_k = N_j \quad \text{pour } j = 1, \ldots, J. \tag{1}$$

L'équation (1) n'a pas nécessairement une solution unique. Les méthodes fondées sur la notion de distance minimum conditionnelle consistent essentiellement à choisir un vecteur $\underline{W}$ de manière à minimiser la distance $D(\underline{W},\underline{S})$ entre les vecteurs $\underline{W}$ et $\underline{S}$ à condition que l'équation (1) soit respectée. Ainsi, en respectant la condition d'équivalence entre les poids corrigés et les totaux de contrôle, nous nous trouvons à modifier le moins possible les poids initiaux $\underline{S}$. Il convient de souligner que pour certaines valeurs $N_1, \ldots, N_J$, il peut être impossible de calculer un vecteur de poids $\underline{W}$ qui satisfasse à la condition (1). En pratique toutefois, cette éventualité ne semble pas poser de problème si l'échantillon est suffisamment grand pour inclure un nombre raisonnable d'unités des diverses catégories de ménages, puisque les totaux de contrôle $\underline{N}$ sont tirés de la population réelle et que par conséquent, on peut les considérer comme "possibles".

Il y a plusieurs façons de mesurer la différence entre deux vecteurs. Nous allons étudier trois critères de distance $D(\underline{W},\underline{S})$ qui correspondent respectivement, au niveau des ménages, à une fonction objective des moindres carrés généralisés (MCG-M), à une fonction d'information minimum discriminante (IMD-M) et à une fonction estimatrice du maximum de vraisemblance (EMV-M). Les critères sont:

$$\text{MCG - M:} \qquad \sum_k (W_k - S_k)^2 / S_k, \tag{2a}$$

$$\text{MDI - H:} \qquad (S_. - W_.) + \sum_k W_k \ln(W_k / S_k), \tag{2b}$$

$$\text{MLE - H:} \qquad (W_. - S_.) - \sum_k S_k \ln(W_k / S_k). \tag{2c}$$

Dans cet article, les points au bas des lettres désignent la sommation par rapport à un indice inférieur.

Dans chaque cas, $D(\underline{W},\underline{S})$ est non négatif et est égal à zéro si et seulement si $\underline{W} = \underline{S}$. On peut le vérifier, de la façon habituelle, en examinant les dérivées partielles du premier et du second ordre de chaque expression par rapport à W_k.

Dans la section 3, nous étudierons des algorithmes permettant de calculer $\underline{W}$ de manière à minimiser ces trois critères tout en respectant la condition (1) au niveau d'approximation voulu.

2.2 Méthodes fondées sur les poids attribués aux personnes

La question peut être envisagée sous un autre angle; cette approche amène toutefois une modification légère mais importante des trois critères de distance. Les critères modifiés sont définis par les expressions (5a), (5b) et (5c) ci-dessous. Bien que ces critères servent à déterminer des poids pour les ménages, ils sont le produit d'une méthode qui, au départ, cherche à définir des poids pour les personnes. En conséquence, nous devons tout d'abord déterminer, pour les personnes, les poids qui se rapprochent le plus possible des poids initiaux des ménages auxquels appartiennent ces personnes, à condition que l'estimation pondérée du nombre de personnes dans chaque cellule de post-stratification égale le chiffre de population. Désignons les membres du k-ième ménage par $i = 1, \ldots, a_k$. et définissons S_{ki} comme le poids initial de la i-ème personne dans le k-ième ménage; remarquez que $S_{ki} = S_k$.

Soit b_{kij} une variable indicatrice (0-1) qui montre si la i-ème personne du k-ième ménage se trouve dans la j-ième cellule de post-stratification. Alors, la condition pour qu'il y ait correspondance avec les chiffres de population est

$$\sum_k \sum_i b_{kij} W_{ki} = N_j. \tag{3}$$

Les trois critères pour la pondération des personnes seraient

$$\sum_k \sum_i (W_{ki} - S_{ki})^2 / S_{ki}, \tag{4a}$$

$$S_{..} - W_{..} + \sum_k \sum_i W_{ki} \ln(W_{ki}/S_{ki}), \tag{4b}$$

$$W_{..} - S_{..} - \sum_k \sum_i S_{ki} \ln(W_{ki}/S_{ki}). \tag{4c}$$

Ces critères peuvent servir à déterminer des poids pour les personnées. De fait, le critère (4c) produit des poids de post-stratification qui servent à la pondération des personnes dans l'enquête sur les dépenses des consommateurs (voir Alexander 1986). Cependant, nous cherchons ici à déterminer des poids pour les ménages. Nous pouvons déterminer de tels poids à l'aide des fonctions de critère ci-dessus si nous posons comme condition additionnelle que tous les membres d'un même ménage aient la même pondération. Par conséquent, posons $W_{ki} = W_k$ pour $i = 1, \ldots, a_k$. Sous cette condition, (3) devient

$$N_j = \sum_k \left(\sum_i b_{kij} \right) W_{ki} = \sum_k a_{kj} W_k,$$

qui est identique à l'équation (1) de la section 2.1. Les critères de distance (4a), (4b) et (4c) s'écrivent maintenant:

$$\text{MCG:} \quad \sum_k a_{k.} \, (W_k - S_k)^2 / S_k, \tag{5a}$$

$$\text{IMD-P:} \quad \sum_k a_{k.} \, S_k - \sum_k a_{k.} \, W_k + \sum_k a_{k.} W_k \ln \, (W_k / S_k), \tag{5b}$$

$$\text{EMV-P:} \quad \sum_k a_{k.} \, W_k - \sum_k a_{k.} \, S_k - \sum_k a_{k.} \, S_k \ln \, (W_k / S_k). \tag{5c}$$

Les critères sont devenus des sommations en fonction des ménages mais la taille du ménage $a_{k.}$ y a été incluse dans le but de mesurer la distance entre le vecteur de poids initial et le vecteur de poids corrigé. Nous verrons que ces critères présentent des avantages par rapport à la méthode plus directe qui a servi à établir (2a), (2b) et (2c).

2.3 Méthode de la personne principale

Dans la version de base de la méthode de la personne principale, le poids de post-stratification attribué à la "personne principale" du ménage sert de poids pour le ménage. Pour déterminer la personne principale, il faut tout d'abord déterminer la personne repère du ménage. Pour l'intervieweur, la personne repère sera celle dont le nom figure à côté de la question suivante: "Donnez-moi tout d'abord le nom du propriétaire ou du locataire du logement".

On définit les rapports entre les membres du ménage en fonction de leur rapport avec la personne repère. Pour les besoins de la cause, la notion de "personne repère" est substituée à celle de "chef de ménage".

Si la personne repère est un homme marié qui vit avec sa femme, celle-ci est reconnue comme la personne principale. Autrement, la personne repère devient la personne principale. Ces critères reposent sur l'idée que la personne principale doit être une personne qui est peu susceptible d'échapper à l'enquête s'il y avait sous-dénombrement à l'intérieur du ménage. En règle générale, les femmes échappent moins souvent aux enquêtes que les hommes. En outre, le propriétaire ou le locataire d'un logement peut difficilement échapper à l'enquête.

Le principe de base de la méthode de la personne principale veut qu'il n'y ait qu'une personne principale par ménage. Ainsi, l'estimation du nombre de personnes principales suffit pour estimer le nombre de ménages. Cette méthode est utilisée dans la U.S. National Crime Survey. D'autres enquêtes, comme la U.S. Consumer Expenditure Survey ou la Current Population Survey, comportent des mesures de redressement additionnelles fondées sur des hypothèses concernant le sous-dénombrement de personnes principales à l'intérieur des ménages par comparaison au sous-dénombrement d'autres personnes dans la même cellule de post-stratification (Alexander 1986).

Il est difficile de construire un modèle théorique de la méthode de la personne principale parce que la définition de la personne repère est quelque peu arbitraire. Dans les exemples fictifs de la section 5, nous allons utiliser une version simplifiée de la méthode de la personne principale, selon laquelle la personne principale est le membre du ménage dont la cellule de post-stratification présente le meilleur taux de dénombrement, c'est-à-dire celle dont le facteur de post-stratification est le plus près de un. Scheuren (1981) applique une version semblable.

Cette version simplifiée de la méthode de la personne principale sera représentée de la façon suivante. Pour le k-ième ménage de l'échantillon, soit $j(k)$ la cellule de post-stratification à laquelle appartient la personne principale de ce ménage. Alors, le poids attribué à cette personne est

$$W_k = S_k (N_{j(k)} / \hat{N}_{j(k)}).$$

3. CALCUL DES POIDS

Les deux méthodes des moindres carrés, MCG-M et MCG-P, ont des expressions en forme analytique pour $\underline{W}$, à condition qu'il existe une solution pour l'équation (1). En ce qui concerne la méthode MCG-M, l'équation des poids corrigés est

$$\underline{W} = \underline{S} + MA(A'MA)^{-1}(\underline{N} - A'\underline{S}) \tag{6}$$

où $\underline{S} = (S_1, \ldots, S_K)$, $\underline{N} = (N_1, \ldots, N_J)$, A est la matrice (a_{kj}) et M est la matrice diagonale $K \times K$, dont la diagonale principale est constituée des éléments de $\underline{S}$. En ce qui a trait à la méthode MCG-P, les poids $\underline{W}$ sont aussi définis par l'équation (6) sauf que M est une matrice diagonale $K \times K$ dont la diagonale principale est constituée des valeurs $S_1/a_{1.}, \ldots, S_K/a_{K.}$.

L'inconvénient de la solution (6) pour l'une ou l'autre des méthodes précitées est qu'elle peut produire des poids négatifs, ce qui est déroutant au point de vue théorique et inacceptable au point de vue pratique. Il est habituellement possible d'introduire de nouvelles conditions pour faire en sorte que les poids soient positifs. Zieschang (1986a) et Huang et Fuller (1978) montrent comment. Toutefois, l'introduction de nouvelles conditions élimine l'avantage que présente une expression en forme analytique simple.

La méthode itérative du quotient (IMD-P) a déjà servi à la pondération des ménages (voir, par exemple, Oh et Scheuren (1978a)). Une méthode connexe, qui a été largement expérimentée, est décrite dans Pugh, Tyler et George (1976), ceux-ci s'étant fondés sur l'approche de Stephan (1942). S'inspirant de Darroch et Ratcliff (1972), Luery (1986) définit un algorithme d'itération qui s'avère convergent lorsqu'il existe une solution à l'équation (1). Nous décrivons ici cette méthode étant donné que le processus itératif est facile à interpréter. Le processus débute par les poids à l'étape 0.

$$W_k(0) = S_k(N_{.}/\hat{N}_{.})$$

Autrement dit, le poids initial S_k est multiplié par un indice de correction global qui correspond au quotient de la population connue $N_{.}$ par la population totale initiale pondérée. Pour les itérations suivantes, la correction est définie par

$$W_k(i) = W_k(i-1) \prod_j \left(N_j / \sum_s a_{sj} W_s(i-1) \right)^{a_{kj}/a_{k.}}$$

Vous remarquerez que $W_k(i-1)$ est multiplié par la moyenne géométrique des facteurs de stratification a posteriori qui s'appliquent aux membres du k-ième ménage, ces facteurs étant calculés à l'aide des poids de l'itération $i-1$.

Les trois autres méthodes (IMD-M, EMV-M et EMV-P) ont fait l'objet de peu d'analyses. Les algorithmes d'itération décrits ci-dessous se sont avérés efficaces dans des exemples fictifs simples comme ceux présentés dans la section 5. Dans chaque cas, on peut déterminer à l'aide des multiplicateurs de Lagrange un système d'équations que les poids doivent satisfaire pour minimiser le critère de distance assujetti aux conditions. Il n'est pas possible de résoudre directement ces équations mais s'il se trouve une méthode itérative qui produit des solutions de la forme voulue, ces solutions minimisent le critère de distance. Si les algorithmes sont convergents, les solutions satisferont les équations. Toutefois, on n'a pu vérifier de façon générale la convergence de ces algorithmes. L'approche de Haber et Brown (1986) pourrait être une solution alternative dans le cas des critères du "maximum de vraisemblance". Fagan et Greenberg (1985) proposent également des solutions à cet égard.

3.1 Solution pour IMD-M

L'équation des poids est

$$W_k = S_k \prod_j \gamma_j a_{kj} \tag{7}$$

à condition que (1) soit respectée. S'il est possible de déterminer des valeurs, $\gamma_1, \ldots, \gamma_J$ de telle sorte que les poids calculés selon (7) satisfassent l'équation (1), alors ces poids minimisent (2b) à condition que (1) soit respectée. Nous décrivons ci-dessous un algorithme d'itération qui produit un vecteur de poids $\underline{W}$.

Posons $W_k(0) = S_k$ et $\gamma_{(0)} = 1$. Puis, à la i-ème itération, posons

$$\gamma_j(i) = \gamma_j(i-1) \left[1 - (\hat{N}_j(i-1) - N_j) / \sum_s a_{sj}^2 W_s(i-1) \right],$$

où $\hat{N}_j(i-1) = \sum_s a_{kj} W_s(i-1)$. Enfin, posons $W_k(i) = S_k \prod_j (\gamma_j(i))^{a_{kj}}$.

3.2 Solution pour EMV-M

La solution a la forme

$$W_k = S_k / \left(1 + \sum_j \gamma_j a_{kj} \right),$$

à condition que (1) soit respectée.

Une solution itérative est

$$W_k(0) = S_k \qquad \text{et} \qquad \gamma_j(0) = 0,$$

$$\gamma_j(i) = \gamma_j(i-1) + (\hat{N}_j(i-1) - N_j) / \left(\sum_s (a_{sj} W_s(i-1))^2 / S_k \right),$$

$$W_k(i) = S_k / \left(1 + \sum_j \gamma_j(i) a_{kj} \right).$$

3.3 Solution pour EMV-P

La solution a la forme

$$W_k = S_k / \left(\sum_j \gamma_{kj} a_{kj} / a_{k.} \right),$$

à condition que (1) soit respectée.

Une solution itérative est

$$W_k(0) = S_k \qquad \text{et} \qquad \gamma_j(0) = 1,$$

$$\gamma_j(i) = \gamma_j(i-1) \hat{N}_j(i-1) / N_j,$$

$$W_k(i) = S_k / \left(\sum_j \gamma_j(i) a_{kj} / a_{k.} \right).$$

4. RÔLE DU "MODE DE COMPOSITION" D'UN MÉNAGE

En ce qui concerne les six méthodes fondées sur la notion de distance minimum conditionnelle, le rapport du poids initial d'un ménage à son poids corrigé dépend de la répartition des membres de ce ménage dans les diverses cellules de post-stratification. Avant de poursuivre cette analyse, nous devons définir ici le "mode de composition" d'un ménage. Deux ménages d'un échantillon (par exemple k et m) auront le même mode de composition s'ils comptent exactement le même nombre de personnes dans chacune des cellules de post-stratification, c'est-à-dire si

$$a_{kj} = a_{mj} \text{ pour } j = 1, \ldots, J. \tag{8}$$

On pourrait avoir, par exemple, un ménage constitué d'un homme de race blanche âgé de 35 à 39 ans et d'une femme de race blanche âgée de 30 à 34 ans. Il convient de souligner que le mode de composition ne dépend pas des liens de parenté.

Le rapport du poids corrigé au poids initial, W_k / S_k, est identique pour tous les ménages qui ont le même mode de composition. Autrement dit, si k et m satisfont l'équation (8), alors $W_k / S_k = W_m / S_m$. Ireland et Scheuren (1975) ont appliqué cette relation. Alexander et Roebuck (1986) en font la démonstration.

Une conséquence avantageuse de cette relation est que la catégorie de ménages peut remplacer le ménage comme unité d'analyse dans le calcul des poids pour les méthodes fondées sur la distance minimum conditionnelle. Un exemple simple nous permettra de mieux saisir les conséquences de ces observations. Supposons qu'il y a deux cellules de post-stratification: $j = 1$ pour le femmes et $j = 2$ pour les hommes. L'échantillon est formé de K ménages. Pour le ménage k, le vecteur (a_{k1}, a_{k2}) définit le nombre de femmes et d'hommes qui constituent le ménage. Un vecteur (2,1) signifie que le ménage compte deux personnes de sexe féminin et une de sexe masculin.

En pratique, un ménage ne peut dépasser une certaine taille et le nombre de catégories de ménages est aussi limité. Pour les besoins de la cause, supposons qu'aucun ménage ne compte plus de trois personnes. Nous avons donc $T = 9$ modes de composition de ménage qui correspondent aux vecteurs (1,0), (0,1), (2,0), (1,1), (0,2), (2,1), (1,2), (3,0), (0,3). Ces modes seront désignés dans l'ordre par $t = 1, \ldots, 9$. On leur attribuera aussi, dans l'ordre, les codes mnémotechniques F, H, FF, FH, HH, FFH, FHH, FFF, HHH. Le tableau 1 contient des données d'échantillon et des totaux de contrôle hypothétiques. Précisons que S est le poids total initial attribué aux ménages de la catégorie t.

Les corrections de poids selon les méthodes fondées sur la distance minimum conditionnelle peuvent être effectuées à l'aide des poids attribués aux modes de composition des ménages $S_1, \ldots, S_9$ sans tenir compte des poids des ménages proprements dits. Nous pouvons calculer les poids corrigés $W_1, \ldots, W_9$ au moyen des algorithmes présentés dans la section 3 en effectuant la sommation en fonction de t plutôt qu'en fonction de k. Ainsi, le poids corrigé d'un ménage de catégorie t est défini comme le produit de W_t / S_t par le poids initial de ce ménage. (Malgré le risque de confusion, nous avons choisi de désigner le poids des ménages par S_k et le poids total d'une catégorie de ménages par S_t pour souligner le fait que les formules définies dans les sections 2 et 3 s'appliquent aussi bien aux catégories de ménages qu'aux ménages proprement dits. Au moment des calculs, le contexte permet de faire la distinction entre S_k et S_t.)

Le fait de pouvoir utiliser des catégories de ménages comme unités d'analyse au lieu des ménages proprement dits est très avantageux pour la présentation d'exemples simples. Même lorsque l'analyse s'étend aux 48 cellules de post-stratification, il peut être encore avantageux d'utiliser les catégories de ménages. Malgré qu'il existe en théorie un nombre indéfini de

catégories de ménages, le nombre de catégories incluses dans un échantillon ne peut jamais dépasser la taille de cet échantillon et est, le plus souvent, beaucoup moindre. C'est ce qu'Ireland et Scheuren (1975) ont constaté pour des cellules de ménages connexes. Le simple fait de réduire le volume des calculs en combinant les poids des ménages à personne unique d'une même catégorie peut avoir des avantages; c'est ce qu'a fait le U.S. Bureau of Labor Statistics en appliquant la méthode des moindres carrés généralisés à l'enquête sur les dépenses des consommateurs.

La version simplifiée de la méthode de la personne principale dépend aussi uniquement du mode de composition du ménage. Si deux ménages ont le même mode de composition, leurs personnes principales se trouveront dans la même cellules de post-stratification, celle dont le facteur de stratification a posteriori est le plus près de 1. On utiliserait donc le même facteur de correction pour les deux ménages. Selon la version fondamentale de la méthode de la personne principale, la personne principale est en partie définie en fonction de celui ou celle qui se trouve être la personne repère. Dans ce cas, le facteur de correction n'est donc pas entièrement déterminé par le mode de composition du ménage.

Il convient de souligner que la méthode EMV-M équivaut à calculer des estimations multinomiales du maximum de vraisemblance (moyennant le respect de la condition (1)) de p_t, $t = 1, \ldots, T$, où p_t est la proportion de ménages de la catégorie t dans la population. La méthode EMV-P peut être interprétée de la même manière. Aucun de ces modèles, qui se rattachent aussi aux méthodes MCG et IMD correspondantes, n'admet le sous-dénombrement systématique.

5. ANALYSE DES MÉTHODES

Nous allons tout d'abord formuler des hypothèses sur les propriétés des méthodes fondées sur la distance minimum conditionnelle en nous servant des résultats de la section 4 et nous allons poursuivre avec des exemples fictifs simples qui, de façon générale, semblent appuyer les hypothèses.

La première hypothèse est que les méthodes EMV-M, MCG-M et IMD-M tendront à produire des résultats similaires et que les méthodes EMV-P, MCG-P et IMD-P tendront à se rapprocher l'une de l'autre, du moins pour de grands échantillons. Cette hypothèse repose sur le fait que les estimateurs de ces méthodes sont tous de meilleurs estimateurs asymptotiques normaux selon le modèle d'échantillonnage multinomial pertinent, où les cellules représentent les catégories de ménages. En ce qui concerne les échantillons de taille faible ou moyenne, on peut s'attendre à des écarts plus marqués entre les résultats des diverses méthodes, surtout s'il y a un grand nombre de modes de composition des ménages, ce qui crée des échantillons de faible taille dans les "cellules" du modèle multinomial.

Les exemples présentés plus loin tendent à confirmer cette hypothèse; les méthodes pour les ménages, tout comme celles pour les personnes, produisent toutes des résultats très comparables, même lorsque les données hypothétiques s'ajustent mal au modèle. Toutefois, comme ces exemples ne concernent qu'un petit nombre de catégories de ménages et de cellules de post-stratification, ils servent plus à illustrer l'hypothèse qu'à la vérifier.

La seconde hypothèse repose sur une analyse qui vise à déterminer le genre de modèles d'échantillonnage où l'on peut appliquer plus particulièrement les fonctions estimatrices du maximum de vraisemblance ou des approximations asymptotiques de celles-ci. Dans ce genre de modèles, on suppose que le dénombrement est complet. On suppose aussi une distribution correspondant à des probabilités qui représentent les proportions réelles dans la population et qui sont conformes aux totaux de contrôle "réels" utilisés dans l'équation (1). Selon ce genre de modèles, les estimations initiales de l'échantillon se rapprocheraient des totaux

de contrôle pour des échantillons suffisamment grands. Cela ne pourrait être le cas s'il y avait un fort taux de sous-dénombrement dans la base de sondage. D'ailleurs, le sous-dénombrement est une justification importante de la post-stratification. La question du sous-dénombrement peut prendre une importance particulière dans les enquêtes par téléphone qui ne prévoient pas de base de sondage supplémentaire pour les ménages qui n'ont pas le téléphone. Si aucune formule particulière n'est prévue pour compenser la non-réponse attribuable à la non-interview (par exemple, refus ou incapacité de répondre), la non-réponse peut être alors une autre source de sous-dénombrement.

Compte tenu de ces observations, nous pouvons maintenant formuler la seconde hypothèse: en l'absence de toute forme de compensation, les méthodes fondées sur la distance minimum conditionnelle peuvent s'avérer inefficace devant le sous-dénombrement systématique même pour de grands échantillons. Dans les modèles où on suppose un dénombrement complet, les méthodes produisent des résultats optimaux; il serait raisonnable de penser qu'elles produisent des résultats inférieurs lorsque le dénombrement n'est plus complet.

Les exemples que nous produisons ci-dessous confirment en partie cette hypothèse. Selon certaines hypothèses concernant le sous-dénombrement, les méthodes fondées sur la distance minimum conditionnelle ne sont pas aussi efficaces que la version simplifiée de la méthode de la personne principale. Selon d'autres hypothèses, certaines de ces méthodes peuvent être assez efficaces. Nous devons en conclure qu'il faut mieux connaître la nature du sous-dénombrement avant de juger si l'une ou l'autre de ces méthodes est supérieure à la méthode de la personne principale. Oh et Scheuren (1978b) soulèvent le même genre de question à propos de l'erreur quadratique moyenne de l'estimateur de la méthode itérative du quotient lorsqu'il y a sous-dénombrement.

Tableau 1

Sous-dénombrement des ménages:
Description de la population et de l'échantillon

			Poids total (W_t) pour les méthodes:			
Catégorie et description	Population réelle	Poids total initial	MCG-M	IMD-M	EMV-M	MCG-P IMD-P EMV-P Personne principale
1: F	25,000	22,500	23,785	23,745	23,704	25,000
2: H	15,000	13,500	14,120	14,097	14,075	15,000
3: FF	7,000	6,300	7,020	7,016	7,013	7,000
4: FH	40,000	36,000	39,708	39,672	39,632	40,000
5: HH	5,000	4,500	4,913	4,906	4,900	5,000
6: FFH	12,000	10,800	12,529	12,506	12,594	12,000
7: FHH	12,000	10,800	12,408	12,428	12,449	12,000
8: FFF	0	0	0	0	0	0
9: HHH	0	0	0	0	0	0
Total	116,000	104,400	114,483	114,370	114,367	116,000

Chiffres de population:	Nombre de femmes	=	115,000
	Nombre d'hommes	=	101,000
Comptes initiaux pondérés:	Femmes	=	103,500
	Hommes	=	90,900

Nous présenterons deux exemples qui reproduisent deux formes de sous-dénombrement contraires. Dans le premier exemple (sous-dénombrement de ménages), nous supposons qu'il y a un taux uniforme de sous-dénombrement des ménages de 10% mais qu'il n'y a pas de sous-dénombrement à l'intérieur des ménages. Dans le second exemple (sous-dénombrement à l'intérieur des ménages), nous supposons qu'il y a un taux de sous-dénombrement des hommes de 10% attribuable au sous-dénombrement à l'intérieur des ménages qui comptent des personnes des deux sexes et au sous-dénombrement des ménages qui ne comptent que des personnes de sexe masculin. Pour les ménages formés d'une seule personne, le "sous-dénombrement à l'intérieur des ménages" signifie que le ménage au complet échappe à l'enquête.

Le tableau 1 décrit le premier exemple. On constatera que chaque catégorie de ménages dans l'échantillon est sous-représentée dans une proportion de 10%. Dans le cas d'un échantillon suffisamment grand, cette sous-représentation serait sûrement attribuable au sous-dénombrement systématique plutôt qu'à l'erreur d'échantillonnage. En appliquant les méthodes fondées sur la distance minimum conditionnelle et la méthode de la personne principale, nous obtenons pour chaque catégorie de ménages les poids totaux corrigés qui figurent dans les quatre dernières colonnes du tableau.

On remarquera que les méthodes MCG-P, IMD-P et EMV-P produisent toutes des poids équivalents à la population réelle. Il s'agit donc de poids "non biaisés".Comme toutes les personnes ont un facteur du second degré de 1/.9, la méthode de la personne principale produit aussi des poids équivalents à la population réelle. Les autres méthodes, MCG-M, IMD-M et EMV-M, produisent toutes des poids beaucoup trop faibles pour les ménages formés d'une seule personne et des poids beaucoup trop élevés pour les ménages formés de trois personnes. Cela est compréhensible intuitivement. Puisque ces trois méthodes n'admettent pas le sous-dénombrement systématique et qu'elles doivent expliquer la sous-représentation de personnes par l'erreur d'échantillonnage, l'explication la plus plausible est que l'échantillon comportait par hasard un nombre de ménages de trois personnes inférieur à la moyenne. Le rendement supérieur de EMV-P est compréhensible puisque cette méthode repose sur un modèle d'échantillonnage multinomial, selon lequel les personnes sont échantillonnées sans égard aux ménages.

Au point de vue pratique, cet exemple déprécie largement les méthodes MCG-M, IMD-M et EMV-M. Un taux de sous-dénombrement uniforme n'empêchera pas ces méthodes de fausser la distribution de la taille des ménages. Chose encore plus grave toutefois, la distorsion créée par l'application de ces méthodes dans l'exemple est contraire au fait généralement admis en ce qui concerne les différences de taux de dénombrement des ménages, c'est-à-dire que les ménages de faible taille sont plus susceptibles d'échapper à l'enquête que les ménages de grande taille et que, par conséquent, ils devraient avoir des poids relativement plus élevés et non des poids relativement moins élevés.

Le deuxième exemple met l'accent sur le sous-dénombrement des hommes à l'intérieur des ménages. Le cas est plus complexe que dans l'exemple précédent parce que la composition apparente d'un ménage peut être différente de sa composition réelle. Par exemple, un ménage composé d'un homme et d'une femme peut sembler être un ménage d'une seule personne. Nous allons différencier la composition réelle de la composition apparente en modifiant la notation utilisée jusqu'à maintenant. Par exemple, un ménage de type FH où l'homme n'est pas dénombré sera désigné par F[H]. Un ménage désigné par [H] ou [HH] aura tout simplement été omis. Le tableau 2 contient les données hypothétiques. La population réelle est la même que celle de l'exemple précédent.

On observe un taux de sous-dénombrement des hommes de 10% à cause du sous-dénombrement à l'intérieur des ménages et du sous-dénombrement des ménages composés uniquement de personnes de sexe masculin. Chaque homme a 10% de chances d'échapper à l'enquête.

Tableau 2

Sous-dénombrement à l'intérieur des ménages:
Description de la population et de l'échantillon

Catégorie de ménages (composition réelle)	Catégorie de ménages (composition apparente)	Population réelle	Poids totaux initiaux
1: F	F	25,000	25,000
2: H	H	13,500	13,500
	[H]	1,500	0
3: FF	FF	7,000	7,000
4: FH	FH	36,000	36,000
	F[H]	4,000	4,000
5: HH	HH	4,500	4,500
	[HH]	500	0
6: FHH	FFH	10,800	10,800
	FF[H]	1,200	1,200
7: FHH	FHH	10,800	10,800
	FH[H]	1,200	1,200
8: FFF	FFF	0	0
9: HHH	HHH	0	0
		116,000	114,000
Chiffres de population:	Nombre de femmes	115,000	
	Nombre d'hommes	101,000	
Comptes initiaux pondérés:	Femmes	115,000	
	Hommes	90,900	

Aucune des colonnes de chiffres du tableau 2 n'est observée puisqu'il n'y a pas de comptage de contrôle de ménages. En outre, on ne connaît pas le mode de composition réel des ménages auxquels appartiennent les unités de l'échantillon. Ainsi, les ménages de type F[H] (catégorie 4) passent pour des ménages de type F (catégorie 1). On observe plutôt les poids totaux initiaux des ménages qui semblent avoir un mode de composition donné. Le tableau 3 donne les poids corrigés calculés selon trois méthodes: EMV-M, EMV-P et personne principale. Comme les méthodes MCG-M et IMD-M produisent des résultats comparables à ceux de EMV-M et que les méthodes MCG-P et IMD-P produisent les mêmes résultats que EMV-P, nous n'avons pas cru nécessaire de reproduire tous ces résultats.

Les trois dernières colonnes du tableau 3 contiennent les poids totaux corrigés qui ont été calculés pour chaque catégorie de ménages (composition réelle) à l'aide des méthodes EMV-M et EMV-P et de la méthode de la personne principale. On constate que les poids obtenus par la méthode de la personne principale pour chaque catégorie "réelle" correspondent aux chiffres de la population réelle qui figurent dans la troisième colonne du tableau 1. C'est pourquoi on peut dire que la méthode de la personne principale produit des poids non biaisés.

Cet exemple reprend les hypothèses qui sous-tendent la version simplifiée de la méthode de la personne principale. Il est donc normal, en l'occurrence, que cette méthode produise des résultats très satisfaisants. Il va de soi que les ménages de type [H] ou [HH] qui échappent entièrement à l'enquête n'ont pas de poids; en contrepartie, le poids des ménages de

Tableau 3

Sous-dénombrement à l'intérieur des ménages:
Catégories et poids observés,
avec les poids corrigés obtenus selon les trois méthodes

Catégorie de ménages	Poids total initial	Poids attribué à la catégorie de ménages (composition apparente)			Poids attribué à la catégorie de ménages (composition réelle)		
		EMV-M	EMV-P	Personne principale	EMV-M	EMV-P	Personne principale
F	29,000	27,450	26,973	29,000	23,664	23,253	25,000
H	13,500	14,997	16,338	15,000	14,997	16,338	15,000
FF	8,200	7,368	7,626	8,200	6,290	6,510	7,000
FH	37,200	38,887	39,128	37,200	41,419	41,586	40,000
HH	4,500	5,623	5,446	5,000	5,623	5,446	5,000
FFH	10,800	10,661	10,885	10,800	11,739	12,001	12,000
FHH	10,800	12,605	11,878	10,800	13,859	13,140	12,000
FFF	0	0	0	0	0	0	0
HHH	0	0	0	0	0	0	0
Total	114,000	117,591	118,274	116,000	117,591	118,274	116,000

type H ou HH qui sont dénombrés est augmenté en conséquence. Le nombre total pondéré de ménages obtenu par la méthode de la personne principale est égal au nombre de ménages dans la population.

Dans cet exemple, les méthodes fondées sur la distance minimum conditionnelle surestiment le nombre total de ménages mais pondèrent trop faiblement les ménages qui ne comptent pas de personne de sexe masculin. De façon générale, ces méthodes produisent des poids trop élevés pour les ménages qui comptent des personnes de sexe masculin.

Il ne faudrait pas en conclure que la méthode de la personne principale est toujours supérieure aux méthodes fondées sur la distance minimum conditionnelle lorsqu'il y a sous-dénombrement à l'intérieur des ménages. Suivant d'autres hypothèses sur le dénombrement, elle pourrait ne pas être aussi efficace. De fait, il existe différentes versions de la méthode de la personne principale selon les enquêtes où elle est appliquée, chaque enquête ayant ses propres hypothèses sur le dénombrement. Soulignons aussi qu'il est possible de combiner la méthode de la personne principale avec la méthode itérative du quotient; à ce sujet, voir Scheuren (1981).

Même dans cet exemple, les poids biaisés qui sont produits par les méthodes fondées sur la distance minimum conditionnelle pourraient être utiles pour l'estimation de certaines caractéristiques. Si les ménages où des hommes ne sont pas dénombrés tendent à minimiser la variable d'intérêt, un poids trop élevé pour ces ménages pourrait néanmoins compenser le biais de réponse lié au sous-dénombrement à l'intérieur des ménages.

L'exemple le plus frappant de cette utilité est l'estimation du nombre total d'hommes; dans ce cas, les poids calculés selon EMV-M et EMV-P correspondent aux totaux de contrôle, ce qui n'est pas le cas des poids calculés selon la méthode de la personne principale. Toutefois, les poids biaisés ne sont pas souhaitables lorsqu'il s'agit de caractéristiques qui font peu souvent l'objet d'erreurs de déclaration attribuables au sous-dénombrement de personnes de sexe masculin (par exemple, mode d'occupation: locataire/propriétaire). L'efficacité des méthodes de pondération dans des cas comme ceux exposés ici dépend incontestablement de la nature du sous-dénombrement et de sa relation avec la variable estimée. Alexander et Roebuck (1986) traitent plus en détail le sujet et fournissent d'autres exemples.

En attendant d'autres recherches sur la couverture des enquêtes et ses effets sur la pondération, quelles recommandations pouvons-nous faire? En ce qui concerne les méthodes fondées sur la distance minimum conditionnelle, MCG-M, IMD-M et EMV-M semblent présenter peu d'intérêt à cause de l'irrégularité de leurs résultats dans le contexte d'un sous-dénombrement uniforme des ménages, ceci en dépit du fait que dans un contexte de dénombrement complet, EMV-M semble reposer sur un modèle plus sensible que EMV-P puisqu'en l'occurrence, la dernière unité d'échantillonnage est le ménage et non la personne.

La possibilité de poids négatifs nous amène à nous interroger sur l'utilité de la méthode MCG-P, malgré que certaines applications (voir, par exemple, Zieschang 1986b) comportent très peu de poids négatifs de sorte qu'on pourrait leur substituer des poids positifs sans que cela ait d'effet notable sur les estimations. Restent les méthodes IMD-P et EMV-P. Les résultats observés ne permettent pas vraiment de déterminer laquelle de ces méthodes est préférable. Si l'on s'en tient uniquement au calcul, on est porté à choisir la méthode itérative du quotient (IMD-P). Selon les quelques résultats obtenus avec les algorithmes de la section 3, la convergence est plus lente pour les méthodes EMV que pour les méthodes IMD. En outre, de nombreuses recherches ont été faites dans le but de trouver des moyens d'accroître l'efficacité de la méthode itérative du quotient dans des applications à grande échelle (voir, par exemple, Ireland et Scheuren 1975). Compte tenu des remarques précédentes, la méthode itérative du quotient (IMD-P) est celle qui semble offrir les meilleures perspectives parmi les méthodes fondées sur la distance minimum conditionnelle.

Contrairement à la méthode de la personne principale, les méthodes fondées sur la distance minimum conditionnelle produisent des poids de ménages qui correspondent aux totaux de contrôle pour les personnes. Cependant, il n'est pas du tout démontré que les méthodes fondées sur la distance minimum conditionnelle sont plus efficaces que la méthode de la personne principale en ce qui concerne le redressement en fonction du sous-dénombrement. Le sous-dénombrement est un aspect fondamental du problème de la pondération dans les enquêtes. La méthode de la personne principale est un palliatif du sous-dénombrement, qui repose sur des hypothèses très simplistes concernant la couverture des enquêtes. En revanche, nous avons vu dans la section 4 que les méthodes fondées sur la distance minimum conditionnelle pouvaient être considérées comme des estimateurs "optimaux" (c'est-à-dire, estimateurs du maximum de vraisemblance ou l'équivalent asymptotique) dans les modèles où l'on pose comme hypothèse le dénombrement complet. Il s'agit donc de choisir entre une solution optimale appliquée à un faux problème et un palliatif de ce qui peut être ou non le vrai problème. Ces observations indiquent qu'il faut manifestement pousser la recherche.

6. QUELQUES SUJETS DE RECHERCHE

6.1 Totaux de contrôle pour les ménages

S'il existait des estimations supplémentaires du nombre de ménages des diverses catégories, on pourrait appliquer les méthodes courantes de post-stratification aux estimations des ménages. On étudie actuellement la possibilité d'établir des chiffres de population de ménages selon la taille dans le cadre d'un projet qui vise à mettre à jour les données du recensement de 1980 (Das Gupta et coll. 1986). Les chiffres de population de ménages nous permettraient d'envisager d'une toute autre façon le problème de la pondération des ménages.

Les chiffres de population de ménages ne devraient pas exclure pour autant l'utilisation des chiffres de population de personnes. Les chiffres de population de ménages ne sont pas susceptibles de fournir des renseignements détaillés sur l'âge, l'origine raciale et le sexe des membres des ménages. Se servant d'une estimation du nombre total de ménages, Scheuren (1981) applique la méthode itérative du quotient pour comparer simultanément les estimations à des totaux de contrôle indépendants pour les personnes et les ménages. Zieschang

(1986b) montre comment apporter des corrections similaires à l'aide de la méthode des moindres carrés généralisés.

Les chiffres de population de ménages réunissent sans aucun doute toutes les caractéristiques nécessaires pour tenir compte des variations de taux de couverture des diverses catégories de ménages. Ils peuvent néanmoins poser des problèmes en ce qui concerne le sous-dénombrement à l'intérieur des ménages puisque leur application peut fausser le calcul de la taille réelle des ménages, ce qui conduirait à une mauvaise classification des ménages de l'échantillon à l'étape de la post-stratification.

6.2 Études relatives à la couverture

On peut évaluer assez bien le taux de couverture des personnes en comparant les estimations initiales de l'enquête (N_j) aux totaux de contrôle N_j. Il est difficile de déterminer dans quelle proportion le sous-dénombrement est attribuable à un sous-dénombrement des ménages ou à un sous-dénombrement des personnes à l'intérieur des ménages. On pourrait obtenir des renseignements supplémentaires en comparant les estimations pondérées initiales des ménages aux chiffres de population de ménages, lorsque ces chiffres seraient disponibles. Pour le moment, on pourrait comparer, pour chaque catégorie de ménages, les estimations d'enquête de 1980 et les chiffres correspondants du recensement de la même année.

Cependant, les renseignements supplémentaires ne nous permettent pas de faire parfaitement la distinction entre le sous-dénombrement de ménages et le sous-dénombrement à l'intérieur des ménages; pour cela, nous devons absolument poser de nouvelles hypothèses. Alexander et Roebuck (1986) proposent certaines solutions préliminaires pour ajuster des données d'enquête et de recensement à une série de modèles de dénombrement. Wolter (1986) décrit d'autres façons de construire un modèle de dénombrement.

6.3 Estimation de variances

En ce qui concerne la plupart des méthodes fondées sur la distance minimum conditionnelle, on n'a pas cherché à mettre au point des méthodes d'estimation de la variance des estimateurs pondérés. Seule la méthode itérative du quotient a fait l'objet de telles études; voir Arora et Brackstone (1977), Bankier (1978) et Fan et coll. (1981).

Dans tous les cas, des méthodes d'itération pourraient servir à estimer la variance des estimateurs pondérés. Ce genre de méthodes s'est avéré raisonnablement efficace dans des conditions assez générales; voir, par exemple, Krewski et Rao (1981). Il reste à déterminer si ces conditions peuvent s'appliquer aux méthodes fondées sur la distance minimum conditionnelle.

6.4 Applications

Zieschang (1986b) a appliqué les méthodes des moindres carrés généralisés à l'enquête sur les dépenses des consommateurs des États-Unis. Scheuren (1981) décrit une application à grande échelle de la méthode itérative du quotient à la pondération des ménages. Les algorithmes du maximum de vraisemblance (EMV-M et EMV-P) n'ont pas été soumis à des applications de ce genre. S'ils devaient servir à la pondération dans les enquêtes, on pourrait devoir se livrer à des recherches en vue d'accroître leur applicabilité.

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance à Michael J. Roebuck pour l'aide qu'il lui a apportée dans cette étude, de même qu'au rédacteur associé et aux arbitres pour leurs commentaires utiles.

L'auteur tient également à exprimer sa gratitude à Brenda Kelly pour son assiduité dans la préparation de la copie dactylographiée.

BIBLIOGRAPHIE

ALEXANDER, C.H. (1986). The present Consumer Expenditure Surveys weighting method. In *Population Controls in Weighting Sample Units,* Section 7, Washington D.C.: U.S. Bureau of Labor Statistics, 1-32.

ALEXANDER, C.H., et ROEBUCK, M.J. (1986). Comparison of alternative methods for household estimation. *Proceedings of the Section on Survey Research, American Statistical Association,* 54-64.

ARORA, H.R., et BRACKSTONE, G.J. (1977). An investigation of the properties of raking ratio estimates: II. With cluster sampling. *Survey Methodology,* 4, 232-252.

BANKIER, M.D. (1978). An estimate of the efficiency of raking ratio estimators under sample random sampling. *Survey Methodology,* 4, 115-124.

BRACKSTONE, G.J., et RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā,* Série C, 41, 97-114.

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

CRESSIE, N., et READ, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society,* Série B, 46, 440-464.

DARROCH, J.N., et RATCLIFF, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics,* 63, 1470-1480.

DAS GUPTA, P., GIBSON, C., HERRIOT, R.A., LAMAS, E., et ZITTER M. (1986). New approaches to estimating households and their characteristics for states and counties. Article présenté au congrès annuel du Population Association of America de 1986.

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

FAGAN, J.T., et GREENBERG, B. (1985). Algorithms for making tables additive: raking, maximum likelihood, and minimum chi-square. Statistical Research Division Report Series No. Census/SRD/RR-85/12, U.S. Bureau of the Census,

FAN, M.C., WOLTMAN, H.F., MISKURA, S.M., et THOMPSON, J.H. (1981). 1980 census variance estimation procedure. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 176-181.

FIENBERG, S.E. (1986). Comments on some estimation problems in the Consumer Expenditure Surveys. Dans *Population Controls in Weighting Sample Units.* Section 5, Washington, D.C.: U.S. Bureau of Labour Statistics, 1-12.

HABER, M., et BROWN, M.B. (1986). Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints, *Journal of the American Statistical Association,* 81, 477-482.

HUANG, E.T., et FULLER, W. (1978). Nonnegative regression estimation for sample survey data. *American Statistical Association Proceedings of Social Statistics Section,* 300-305.

IRELAND, C.T., et SCHEUREN, F.J. (1975). The rake's progress, Dans *Computer Programs for Contingency Table Analysis,* Washington, D.C.: The George Washington University, 155-216.

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.

LUERY, D.M. (1986). Weighting sample survey data under linear constraints on the weights. *Proceedings of the Social Statistics Section, American Statistical Association,* 325-330.

OH, H.L., et SCHEUREN, F.J. (1978a). Multivariate raking ratio estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 716-722.

OH, H.L., et SCHEUREN, F.J. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 723-725.

PUGH, R.E., TYLER, B.S., et GEORGE, S. (1976). Computer-based procedure for N-dimensional adjustment of data - NJUST. Staff Paper No. 24, U.S. Social Security Administration.

SCHEUREN, F.J. (1981). Methods of estimation for the 1973 exact match study. Dans *Studies from Interagency Data Linkages, Report No. 10,* U.S. Department of Health and Human Services, U.S. Social Security Administration, 9-122.

STEPHAN, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics,* 13, 166-178.

WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association,* 81, 338-346.

ZIESCHANG, K.D. (1986a). Generalized least squares: an alternative to principal person weighting. *Population Controls in Weighting Sample Units,* Section 2, Washington, D.C.: U.S. Bureau of Labor Statistics, 1-41.

ZIESCHANG, K.D. (1986b). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 64-71.

Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 211–220
Statistique Canada

Une méthode intégrée de pondération des personnes et des familles

G. LEMAÎTRE et J. DUFOUR[1]

RÉSUMÉ

Les enquêtes sur les ménages utilisent habituellement des méthodes d'estimation distinctes pour les caractéristiques des personnes et celles des familles. Les auteurs proposent dans cet article une méthode intégrée de pondération et définissent à cette fin un estimateur par les moindres carrés. Ils montrent que cet estimateur est sans biais à certaines conditions générales. Au moyen de données de l'enquête sur la population active du Canada, ils calculent les variances de cet estimateur et montrent qu'elles se comparent avantageusement aux variances calculées pour les méthodes usuelles.

MOTS CLÉS: Estimation pour les familles; pondération des familles; pondération par les moindres carrés.

1. INTRODUCTION

De nombreuses enquêtes sur les ménages comportent souvent une étape de stratification a posteriori par laquelle on contraint les estimations de population de l'enquête, normalement classées selon l'âge et le sexe, à respecter des totaux supplémentaires tirés de sources démographiques. Pour faciliter la totalisation, on attribue habituellement à chaque répondant un poids qui est égal au produit de l'inverse du taux de sondage par un facteur de compensation de la non-réponse par un facteur de correction âge et le sexe sous forme de quotient. On calcule ensuite l'estimation pour une caractéristique donnée en additionnant les poids attribués à tous les répondants de l'échantillon qui présentent cette caractéristique. En règle générale, les membres d'un même ménage n'ont pas le même poids. Lorsqu'il s'agit d'estimer les caractéristiques de personnes, cela ne pose pas de problème particulier; en revanche, lorsqu'il s'agit de produire des estimations relatives aux ménages ou aux familles, il n'est pas évident lequel de ces poids convient d'utiliser dans les circonstances, s'il y a lieu effectivement d'en choisir un.

Pour estimer les caractéristiques de familles, on pourrait recourir à l'estimation par le quotient en se servant de données supplémentaires sur les familles et les personnes. Cependant, on dispose rarement de données supplémentaires fiables et récentes sur les familles. À cause d'événements tels qu'une naissance, un décès, un mariage, un divorce ou le départ d'un membre du ménage ou l'arrivée d'un nouveau membre, des caractéristiques comme la taille de la famille varient d'un recensement à l'autre de façon moins prévisible qu'une caractéristique telle que l'âge. Les dossiers administratifs qui nous renseignent sur les variations postcensitaires de la population (c'est-à-dire, registres des naissances, des décès et des migrations) ne disent rien sur les variations touchant les ménages. Le registre des naissances, par exemple, n'indique pas la taille des familles où un enfant est né. Les dossiers fiscaux peuvent combler en partie cette lacune (voir Auger 1987) mais ils ne s'étendent pas à toute la population et ne renferment pas des données suffisamment récentes pour produire des estimations courantes. En l'absence de données supplémentaires sur les familles, on s'est mis à utiliser les poids attribués aux personnes pour estimer les caractéristiques de familles. Pour diverses raisons, cette méthode n'est pas la solution idéale. Dans cet article, nous proposons une méthode d'estimation qui permet de calculer un poids unique pour le ménage, lequel poids peut servir aussi bien à l'estimation pour les personnes qu'à l'estimation pour les familles.

[1] G. Lemaître et J. Dufour, Division des méthodes d'enquêtes sociales, Statistique Canada, 4-ième étage, Immeuble Jean Talon, Parc Tunney, Ottawa (Ontario) K1A 0T6.

On a déjà proposé des méthodes qui visaient à déterminer un poids unique pour chaque ménage; on insistait alors sur l'utilisation de données supplémentaires sur les personnes pour améliorer les estimations relatives aux familles. Oh et Scheuren (1978) ont proposé une technique d'estimation itérative "multidimensionnelle" fondée sur la méthode du quotient, qui consiste à corriger successivement par la méthode du quotient les estimations de population des strates formées a posteriori en utilisant les facteurs de correction calculés pour chaque strate, puis à répéter l'opération jusqu'à ce qu'il y ait convergence. Les corrections calculées à chaque étape sont appliquées aux ménages qui comptent des membres dans la strate (formée a posteriori) qui fait l'objet de la correction. Zieschang (1986) a proposé une méthode des moindres carrés généralisés (MCG) selon laquelle on minimise la somme des carrés pondérés des corrections apportées aux poids du plan de sondage (ou poids initiaux) en respectant une série de contraintes linéaires. Alexander (1987) analyse plusieurs méthodes de pondération fondées sur la minimisation d'une fonction de distance, sujette à certaines contraintes, y compris la méthode MCG, et les évalue par rapport au sous-dénombrement dans les enquêtes. Bien que les méthodes ci-dessus aient été proposées initialement dans le but d'améliorer les estimations relatives aux familles, les poids découlant des divers estimateurs peuvent très bien servir à estimer les caractéristiques de personnes. Cet article préconise l'application d'une méthode d'estimation adaptée aussi bien aux personnes qu'aux familles. Dans la section 2, nous analysons les faiblesses des méthodes d'estimation actuellement en usage pour ce qui a trait aux caractéristiques des personnes et des familles. La section suivante sert à définir un estimateur fondé sur un modèle et inspiré d'une méthode de pondération généralisée mise de l'avant par Bethléhem et Keller (1987). La section 4 présente des résultats empiriques tirés de l'enquête sur la population active du Canada. Enfin, la section 5 expose des projets de recherche supplémentaires.

2. MÉTHODES D'ESTIMATION COURANTES

La plupart des enquêtes sur les ménages ont pour objectif principal de produire des estimations pour les caractéristiques de personnes, notamment les caractéristiques de l'activité. Si l'unité d'échantillonnage ultime de ces enquêtes est le ménage, c'est surtout pour des raisons d'économie et de commodité. Bien que les premières étapes de la pondération (compensation de la non-réponse, correction en fonction des régions rurales et urbaines, etc.) tiennent compte du ménage en tant qu'unité, ce fait n'est pas reconnu dans l'étape finale (c'est-à-dire que l'on ne tient pas compte du fait que les membre d'un ménage sont échantillonnés intégralement). De façon plus particulière, les biais de couverture qui pourraient être liés à l'unité échantillonnée ne sont pas directement pris en considération ou compensés dans l'estimation. On suppose donc que le sous-dénombrement est "ignorable" au sens de Rubin (1976); autrement dit, l'étape de l'estimation traite de façon identique toutes les personnes comprises dans une strate formée a posteriori selon les critères d'âge et de sexe, peu importe que ces personnes vivent seules ou qu'elles appartiennent à un ménage formé de plusieurs personnes. Toutefois, une étude sur la non-réponse dans l'enquête sur la population active (Paul et Lawes 1983) a démontré que les ménages de faible taille, particulièrement les ménages sans enfant, tendent à être sous-représentés dans l'échantillon. Bien qu'il n'existe pas d'étude comparable pour les ménages oubliés dans l'enquête sur la population active, des études portant sur le sous-dénombrement des ménages privés dans le recensement ont montré que les ménages non-dénombrés sont effectivement de plus petite taille en moyenne que les ménages dénombrés (Gosselin et Théroux 1980). Une méthode qui suppose que les ménages sont absents de façon aléatoire peut introduire un biais dans les estimations de la population active pour les personnes, surtout si la répartition des membres des ménages de faible taille par rapport aux caractéristiques de l'activité est différente de celle des membres des ménages de taille plus élevée, toutes choses étant égales par ailleurs. Intuitivement, une méthode d'estimation qui tiendrait compte (ne serait-ce qu'indirectement) du fait que les ménages de faible taille sont

moins exposés à la non-réponse et au sous-dénombrement que les ménages de taille plus élevée pourrait corriger en partie cette faiblesse de l'échantillon.

Faute de pouvoir incorporer des données supplémentaires sur les ménages ou les familles dans une méthode de pondération appropriée afin d'attribuer un poids bien défini aux familles, beaucoup de méthodes actuellement en usage attribuent à la famille le poids de la "personne principale" de cette famille. Dans l'enquête sur la population active du Canada, la personne principale est le conjoint de sexe féminin, si elle est présente, sinon c'est le chef de ménage. Comme ces méthodes supposent que les ménages non-dénombrés sont absents de façon aléatoire, les estimations produites à l'aide du poids de la personne principale tendent à surestimer les familles de taille plus élevée et à sous-estimer les personnes seules. En outre, on peut estimer de nombreuses caractéristiques (par exemple, population, revenu) à l'aide du poids attribué aux personnes ou du poids attribué aux familles et en règle générale, les deux séries d'estimations diffèreront entre elles, parfois de façon substantielle. Il est vrai que même dans des conditions d'échantillonnage et d'interview idéales, avec un taux de non-réponse ou de sous-dénombrement uniforme, il y aura toujours des écarts entre les estimations fondées sur les familles et celles fondées sur les personnes pour une même caractéristique. Toutefois, si l'échantillon est suffisamment grand, les différences devraient être faibles. Dans les conditions réelles d'enquête, qui ne sont pas des conditions idéales, les différences sont parfois trop fortes pour que l'on puisse les expliquer simplement par la variabilité d'échantillonnage. En appliquant une méthode d'estimation qui attribue un poids unique au ménage, lequel poids concorde avec les chiffres de population supplémentaires lorsqu'il est utilisé comme poids de personne, nous nous trouvons à résoudre le problème des deux systèmes d'estimation parallèles. C'est ce à quoi nous nous attachons dans la section suivante en définissant un estimateur particulier.

3. ESTIMATEUR PROPOSÉ

Nous allons commencer par définir une méthode de pondération généralisée fondée sur des modèles linéaires de Bethlehem et Keller (1987) et, comme eux, nous allons l'appliquer tout d'abord à l'estimation fondée sur les personnes. Nous modifions ensuite la méthode de manière qu'elle produise des poids de ménages qui conviennent à l'estimation des caractéristiques de personnes. Nous allons ici nous inspirer largement de l'article de Bethlehem et Keller.

Supposons une population cible d'enquête formée de N unités, un N-vecteur Y contenant les valeurs d'une variable cible et une matrice X de dimension $N \times p$ contenant les variables auxiliaires définies pour chaque unité de la population cible. On suppose que les chiffres de population pour chaque varaible auxiliaire sont connus et on les désigne collectivement par le p-vecteur x. Dans notre modèle, x sera constitué des totaux de groupes d'âge-sexe. S'il y a corélation entre les variables auxiliaires et la variable cible, les valeurs de $E = Y - XB$ varieront moins que les valeurs de la variable cible Y pour un p-vecteur B approprié. En appliquant les moindres carrés ordinaires à toutes les unités de la population cible, on obtient

$$B = (X'X)^{-1}X'Y, \tag{3.1}$$

pourvu que X soit une matrice à rang complet. Une estimation de B pour l'échantillon est définie par

$$\hat{B} = (X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}TY, \tag{3.2}$$

où T est une matrice diagonale dont le i-ième élément est égal à 1 si la i-ième unité de la population est incluse dans l'échantillon, et est égal à 0 dans le cas contraire, et $E(T) = \Pi$.

Il est possible de montrer que $\hat{B}$ sera approximativement non biaisé pour de grands échantillons. Or, le paramètre d'intérêt n'est pas B mais le chiffre de population y. Si nous posons $\hat{y} = \hat{B}'x$, $\hat{y}$ sera un estimateur approximativement non biaisé de y à la condition que $B'x = y$, ou que la somme des résidus pour le modèle de population $Y = XB + E$ soit égale à 0. Cette condition sera respectée si le N-vecteur formé de chiffres 1 est dans l'espace délimité par les colonnes de X et, en particulier, si les variables auxiliaires X renferment un ensemble exhaustif de variables indicatrices qui s'excluent mutuellement (pour les groupes d'âge-sexe par exemple).

Si nous écrivons $\hat{y} = \hat{B}'x = Y'\Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}x$, nous constatons que l'estimateur définit implicitement un N-vecteur de poids donné par

$$W = \Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}x,$$

qui ne dépend pas de la variable cible faisant l'objet de l'estimation. Si nous utilisons les poids pour produire des estimations pour les variables auxiliaires, nous obtenons $X'W = x$. Les poids donnent en effet les totaux de population espérés. De plus, si X est constituée entièrement d'un ensemble exhaustif de variables indicatrices qui s'excluent mutuellement, l'estimateur de régression y équivaudra à l'estimateur régulier de la stratification a posteriori. Pour plus de renseignements, voir Bethlehem et Keller (1987).

Il peut être utile de souligner que selon cette méthode, le poids d'une personne quelconque échantillonnée i peut être défini en règle générale de la façon suivante:

$$W_i = \sum_j \frac{x_{ij}b_j}{\pi_i}, \tag{3.3}$$

où $(b_1, \ldots, b_p) = (X'\Pi^{-1}TX)^{-1}x$ et π_i est la probabilité de sélection de la personne i. Cela donne à penser que l'on peut modifier la méthode d'estimation décrite ci-dessus de manière à obtenir les poids voulus en définissant les variables auxiliaires de la même façon pour tous les membres du ménage. Une manière simple de le faire est de définir des variables auxiliaires pour le ménage en remplaçant, par exemple, les variables correspondantes pour les personnes par la moyenne du ménage. De façon plus formelle, soit Z une matrice $N \times p$ définie pour une personne i $(i = 1, \ldots, N)$ appartenant au ménage h $(h = 1, \ldots, H)$ par

$$Z_{ij} = \frac{U_{hj}}{n_h},$$

où U_{hj} est le total pour la caractéristique j dans le ménage h, c'est-à-dire $U_{hj} = \Sigma_k X_{kj}$, où la sommation porte sur tous les membres k du ménage h, n_h = taille du ménage h, et $\Sigma_h n_h = N$. Définissons Y comme un N-vecteur de valeurs d'une variable cible arbitraire définie pour les *personnes*. Comme pour l'estimation relative aux personnes, nous utilisons le modèle de population $Y = ZC + E$ et appliquons les moindres carrés aux données de l'échantillon pour obtenir une estimation

$$\hat{C} = (Z'\Pi^{-1}TZ)^{-1}Z'\Pi^{-1}TY. \tag{3.4}$$

Nous posons $\hat{y} = \hat{C}'x$, où x est encore le vecteur des chiffres de population pour les variables auxiliaires. $\hat{y}$ sera un estimateur approximativement non biaisé de y pourvu que le N-vecteur formé de chiffres 1 se trouve dans l'espace délimité par les colonnes de Z. Comme pour (3.3), le poids d'une personne quelconque échantillonnée dans le ménage h sera défini par

$$W_h = \sum_j \frac{U_{hj}c_j}{\pi_h n_h}. \tag{3.5}$$

Comme tous les membres du ménage se rattachent au même vecteur ligne de Z et que la probabilité de sélection du premier ordre est la même pour tous, ils auront tous le même poids. De plus, lorsqu'on se sert du oids du ménage pour els personnes, on obtient des résultats compatibles avec les chiffres de population supplémentaires. Bien que cette méthode puisse produire des poids négatifs (si la valeur de quelques-uns des c_j's est inférieure à zéro), les ménage dont le poids est modifié le plus par cette méthode sont en règle générale des ménages dont la composition est peu commune et que l'on retrouve rarement dans l'échantillon et dans la population en général, du moins en ce qui concerne les échantillons qui ne sont pas exposés à un taux de non-réponse ou de sous-dénombrement élevé. La méthode proposée a servi récemment à pondérer des données de l'enquête sur la population active portant sur une période de 24 mois; à cette occasion, un seul ménage a reçu un poids négatif, faible par surcroît. Les poids négatifs font problème parce qu'il est difficile de leur attacher le sens que l'on attribue normalement aux poids, c'est-à-dire nombre de personnes ou de ménages dans la population en général représentés par une personne ou un ménage particuliers échantillonnés. Cependant, selon la formule décrite ci-dessus, les poids finals ne sont définis qu'implicitement et on pourrait de fait considérer que ces poids ne sont qu'un moyen commode de produire des estimations. En pratique, il est peu probable qu'une estimation significative de la valeur d'une caractéristique d'intérêt devienne négative sous l'effet de quelques poids négatifs. En revanche, c'est tout autre chose de vouloir faire comprendre à un utilisateur perplexe la notion de poids négatif.

La variance de l'estimateur $\hat{y} = \hat{C}'x$ défini dans le présent article peut être calculée au moyen des méthodes décrites dans Fuller (1975). De plus, il est possible de montrer que cet estimateur équivaut aux estimateurs de la méthode MCG proposée par Zieschang (1986) et Alexander (1987) lorsque l'espace délimité par les variables auxiliaires Z renferme un vecteur de chiffres 1. Wright (1983) décrit d'autres propriétés de ce genre d'estimateur.

4. RÉSULTATS EMPIRIQUES

L'enquête sur la population active du Canada est une enquête mensuelle à échantillon avec renouvellement qui s'adresse à quelque 48,000 ménages répartis dans tout le pays (voir Platek et Singh 1976 et Singh, Drew et Choudhry 1984). Une fois échantillonnés, les ménages demeurent dans l'échantillon six mois consécutifs avant d'être remplacés. Les strates géographiques primaires sont les dix provinces. La taille des échantillons peut varier de 1,500 ménages, à l'Île-du-Prince-Édouard (la plus petite province), à environ 9,000 en Ontario (province la plus populeuse). L'enquête permet de recueillir des données sur la situation des répondants vis-à-vis de l'activité pendant une semaine de référence donnée à chaque mois et aboutit à la publication de toute une série d'estimations relatives au marché du travail dans le pays.

Les données d'une des enquêtes mensuelles ont servi à faire une évaluation préliminaire de l'estimateur proposé. Nous avons choisi à cette fin l'enquête de mai 1981 pour pouvoir comparer les résultats à ceux du recensement de 1981 tenu à peu près à cette date. Bien que nous ayons utilisé jusqu'ici les termes "ménage" et "famille" indistinctement, l'utilisateur s'intéresse plus souvent aux estimations touchant la "famille économique", laquelle est constituée de tous les membres d'un ménage qui sont apparentés par le sang, par alliance ou par adoption. Or, l'unité échantillonnée (en l'occurrence, le ménage) se prête mieux théoriquement à la pondération. Néanmoins, les résultats empiriques que nous exposons ici reposent sur des estimations qui ont trait aux familles économiques. Dans cette évaluation, nous nous sommes intéressés aussi bien aux caractéristiques des personnes (situation vis-à-vis de l'activité) qu'à celles des familles (nombre de familles économiques et de personnes seules). La pondération par les moindres carrés a porté sur deux séries de groupes d'âge-sexe formés par intervalle de cinq ans, les personnes de 70 ans et plus étant groupées selon le sexe. Les enfants de 0 à 14 ans ont été exclus de la première série de groupes (24 au total) pour que l'on puisse comparer l'estimateur proposé à un estimateur régulier de stratification a posteriori

fondé sur les personnes en utilisant la même information supplémentaire. La seconde série comprenait aussi les enfants répartis en six groupes d'âge-sexe et a servi uniquement à la pondération par les moindres carrés puisque la pondération des enfants par la stratification a posteriori régulière n'a aucun effet sur la pondération des personnes de 15 ans et plus.

Bien que tous les estimateurs étudiés soient approximativement non biaisés pour les estimations des caractéristiques de personnes, les hypothèses concernant la nature du sous-dénombrement et de la non-réponse varient de l'un à l'autre (la méthode de compensation de la non-réponse utilisée dans l'enquête sur la population active suppose que les ménages non répondants sont absents de façon aléatoire à l'intérieur d'une région géographique). L'estimateur par stratification a posteriori suppose implicitement que les différences de taux de non-réponse ou de sous-dénombrement dépendent uniquement de l'âge et du sexe et que, par conséquent, elles peuvent être éliminées au moyen d'une estimation fondée sur les personnes et appuyée de l'information supplémentaire portant sur ces caractéristiques. Dans la pondération par les moindres carrés, le poids d'une personne dépendra de la composition âge-sexe du ménage (sans enfant dans un cas et avec des enfants dans l'autre). Ainsi, toutes choses étant égales par ailleurs, la correction appliquée au poids initial d'une personne appartenant à un groupe d'âge-sexe exosé à un fort taux de sous-dénombrement devrait être plus élevée si cette personne demeure seule que si elle demeure avec des personnes qui appartiennent à des groupes d'âge-sexe bien représentés par l'échantillon.

Tableau 1

Nombre de personnes occupés et en chômage, nombre de familles économiques et de personnes seules, enquête sur la population active, mai 1981 (en milliers)

	Estimateur[a]	Occupées	En chômage	Familles économiques	Personnes seules
Canada	A	11,094	850	6,424	2,432
	B	11,090	850	6,446	2,442
	C	11,120	851	6,410	2,495
Région de l'Atlantique	A	819	102	563	156
	B	819	102	570	154
	C	821	102	569	156
Québec	A	2,725	304	1,723	587
	B	2,724	304	1,725	596
	C	2,735	305	1,714	614
Ontario	A	4,198	274	2,325	863
	B	4,200	273	2,325	861
	C	4,211	273	2,310	881
Région des Prairies	A	2,074	83	1,078	506
	B	2,072	84	1,089	510
	C	2,074	83	1,085	517
Colombie-Britannique	A	1,277	88	735	319
	B	1,276	88	738	321
	C	1,280	88	734	327

[a] A = stratification a posteriori/personne principale, B = moindres carrés (enfants exclus de la pondération) et C = moindres carrés (enfants inclus dans la pondération).

Comme il existe des chiffres de population supplémentaires selon l'âge et le sexe pour chaque province, nous avons établi des estimations pour chacune des provinces. Cependant, dans les tableaux qui suivent, les provinces de moindre importance sont fondues en deux groupes.

Tableau 2

Répartition des écarts en pourcentage entre les poids finals et les poids initiaux, estimateur par stratification a posteriori et estimateur par les moindres carrés, enquête sur la population active, mai 1981

Écart en pourcentage	Pourcentage de l'échantillon		
	Stratification a posteriori	Moindres carrés	Moindres carrés (avec enfants)
> −30%	0.0	0.1	0.2
−30 à −20%	0.0	0.5	0.9
−20 à −10%	0.6	3.0	5.3
−10 à 0%	23.9	20.4	27.1
0 à 10%	53.9	44.6	37.3
10 à 20%	20.6	26.3	21.6
20 à 30%	0.6	4.4	6.2
30 à 40%	0.1	0.4	0.9
40 à 50%	0.0	0.0	0.2
< 50%	0.0	0.0	0.2

Note: La taille de l'échantillon est N = 159014.

Tableau 3

Efficacité estimée de l'estimateur par les moindres carrées par rapport à l'estimateur par stratification a posteriori: nombre de personnes occupées et en chômage, nombre de familles économiques et de personnes seules, enquête sur la population active, mai 1981

	Estimateur[a]	Personnes occupées	Personnes en chômage	Familles économiques	Personnes seules
Canada	B	1.044	0.999	1.565	1.038
	C	1.066	0.999	1.616	1.036
Région de l'Atlantique	B	1.110	0.977	1.266	0.998
	C	1.193	0.992	1.567	1.070
Québec	B	1.059	1.005	1.553	1.020
	C	1.063	0.992	1.582	0.992
Ontario	B	1.028	1.011	1.825	1.064
	C	1.059	1.010	1.828	1.037
Région des Prairies	B	1.001	1.009	1.205	1.009
	C	1.072	1.066	1.420	1.134
Colombie-Britannique	B	1.038	0.964	1.248	1.048
	C	1.053	0.978	1.203	1.045

[a] B = moindres carrés (enfants exclus de la pondération) et C = moindres carrés (enfants inclus dans la pondération).

De façon générale, les trois estimateurs, et plus particulièrement A et B, ne produisent pas des estimations très différentes les unes des autres. Le fait d'inclure les enfants dans la pondération semble se traduire par des estimations légèrement plus élevées pour les personnes occupées et les personnes seules et des estimations légèrement moins élevées pour les familles économiques au niveau national et dans les provinces plus importantes (comparer ces résultats à ceux de Scheuren et coll. 1981). Ces chiffres sont conformes aux prévisions bien qu'il subsiste des écarts notables par rapport aux résultats du recensement, qui indiquent (arrondi au millier près) 6,369,000 familles économiques et 2,583,000 personnes seules au pays. Nous pouvons en conclure que même si l'estimateur par les moindres carrés produit des chiffres qui nous rapprochent plus de la réalité (lorsqu'on tient compte des enfants), il nous faudra de l'information supplémentaire précise et récente pour effacer le biais résiduel.

L'efficacité de l'estimateur par les moindres carrés n'est pas tout à fait évidente. Certes, si nous devions faire une prévision en nous fondant sur les résultats ci-dessus, nous serions portés à dire que l'estimateur par les moindres carrés sera aussi efficace que l'estimateur par stratification a posteriori en constatant la similitude des estimations produites par les deux genres d'estimateur. Par ailleurs, on devrait s'attendre à des gains d'efficacité pour les estimations relatives aux familles économiques parce que l'estimateur par les moindres carrés fait intervenir les chiffres de population supplémentaires dans le calcul du poids des ménages. Cependant, on ne peut attribuer un poids unique au ménage sans une redistribution des poids.

Comme l'indique le tableau 2, la dispersion est un peu plus forte pour les poids calculés par les moindres carrés que pour ceux calculés par les méthodes ordinaires de stratification a posteriori. Le fait de tenir compte des enfants crée une dispersion encore plus forte. Le degré de dispersion des poids reflète essentiellement la disparité qui existe entre la composition de l'échantillon et celle de la population en général au point de vue de l'âge, du sexe et de la taille des ménages. Comme le fait de vouloir calculer un poids unique pour tout le ménage ajoute une contrainte à la méthode d'estimation, on pourrait s'attendre que les variances en subissent quelque peu les conséquences, surtout si aucune autre information supplémentaire n'est utilisée pour l'estimation. Toutefois, la réalité est quelque peu différente.

Les variances de l'estimateur par stratification a posteriori ont été estimées au moyen de la méthode de Keyfitz (1957), où les échantillons répétés étaient des UPÉ (unités primaires d'échantillonnage) ou des UPÉ regroupées. Les variances de l'estimateur par les moindres carrés ont été estimées à l'aide de la méthode décrite dans Fuller (1975). Pour obtenir une meilleure comparabilité, on a calculé les variances pour plusieurs caractéristiques estimées par stratification a posteriori au moyen de la méthode de Fuller et on les a comparées aux variances obtenues par la méthode de Keyfitz. Les deux séries d'estimations étaient très comparables dans tous les cas (1 à 2 pour cent d'écart).

Le tableau 3 donne l'efficacité estimée de l'estimateur par les moindres carrés par rapport à l'estimateur par stratification a posteriori pour les caractéristiques étudiées dans le tableau 1. Les gains d'efficacité pour les estimations relatives aux familles économiques sont substantiels. Les estimations relatives aux personnes occupées et aux personnes seules semblent aussi être légèrement plus efficaces; toutefois, les réductions de la variance pour ces naractéristiques sont faibles, sauf pour les personnes occupées dans la région de l'Atlantique, surtout quand les enfants figurent dans la pondération. Il est intéressant de constater que la taille moyenne d'un ménage dans la région Atlantique est supérieure à ce qu'elle est ailleurs au Canada, bien que l'influence que cela pourrait avoir sur les estimations de personnes occupées n'est pas tout à fait claire. Pour ce qui a trait aux personnes en chômage, la méthodes des moindres carrés ne change essentiellement rien aux variances. Il est permis de croire que ces résultats seront observés de façon générale, c'est-à-dire pour des caractéristiques quelconques. Bien que le critère du poids unique par ménage soit restrictif pour les estimations des caractéristiques de personnes, l'estimateur par les moindres carrés semble résoudre cette difficulté grâce aux variables "explicatives" additionnelles du modèle linéaire, c'est-à-dire les moyennes de toutes les variables auxiliaires des ménages. Les résultats préliminaires

rapportés ci-dessus donnent à penser qu'il serait possible d'intégrer les deux genres d'estimation (personnes et familles) sans que cela ne réduise pour autant l'efficacité des estimations relatives aux personnes.

5. PROJETS DE RECHERCHE

On procède actuellement à une étude empirique plus poussée des propriétés de l'estimateur par les moindres carrés, où l'on s'intéresse particulièrement à l'évolution chronologique des estimations et à leur efficacité, pour un plus grand nombre de caractéristiques, par rapport aux estimations produites par la méthode itérative du quotient utilisée actuellement dans l'enquête sur la population active. Les résultats ci-dessus donnent à penser que la composition âge-sexe d'un ménage a un "pouvoir explicatif" au moins aussi grand que celui du groupe d'âge-sexe même, du moins en ce qui concerne certaines caractéristiques de personnes. Il sera intéressant de voir si les efficacités relatives seront aussi satisfaisantes pour des caractéristiques qui ont une plus forte corrélation avec l'âge et le sexe. Par ailleurs, même si en pratique les poids négatifs sont rares, il faudra penser à élaborer une méthode qui permettra de traiter ces poids le cas échéant. On pourrait, par exemple, les considérer comme des valeurs aberrantes ou peut être les prévenir en limitant la valeur des modifications de poids (Zieschang 1987). Enfin, il serait utile de définir explicitement le modèle de sous-dénombrement sur lequel repose l'estimateur par les moindres carrés pour que l'on puisse faire une évaluation du modèle proprement dit.

REMERCIEMENTS

L'auteur tient à remercier F. Scheuren pour ses commentaires et suggestions lors de la rédaction de cet article.

BIBLIOGRAPHIE

ALEXANDER, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête,* 13, 193-210.

AUGER, E. (1987). Family data from the canadian personal income tax file. Dans *Statistics of Income and Related Administrative Record Research: 1986-1987,* (éd. W. Alvey et B. Kilss), Washington, D.C.: Internal Revenue Service, 177-184.

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

FULLER, W.A. (1975). Regression analysis for sample surveys. *Sankhyā,* C37, 117-132.

GOSSELIN, J.-F., et THÉROUX, G. (1980). Recensement du Canada de 1976 Qualité des données Série I; Sources d'erreurs - Couverture. N° 99-840 au répertoire, Statistique Canada.

KEYFITZ, N. (1957). Estimates of sampling variance where two units are selected from each stratum, *Journal of the American Statistical Association,* 52, 503-510.

OH, H.L., et SCHEUREN, F. (1978). Multivariate raking ratio estimation in the 1973 exact match study. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 716-722.

PAUL, E.C., et LAWES, M. (1982). Caractéristiques des ménages répondants et non répondants dans l'enquête sur la population active du Canada. *Techniques d'enquête,* 8, 48-85.

PLATEK, R., et SINGH, M.P. Méthodologie de l'enquête sur la population active du Canada, N° 71-526 au répertoire, Statistique Canada.

RUBIN, D.B. (1976). Inference on missing data. *Biometrika,* 63, 581-592.

SINGH, M.P., DREW, J.D., et CHOUDHRY G.H. (1984). Remaniement de l'enquête sur la population active au Canada à partir des résultats du recensement de 1981. *Techniques d'enquête,* 10, 139-154.

SCHEUREN, F., OH, H.L., VOGEL, L., et YUSKAVAGE, R. (1981). studies from interagency data linkages, report No. 10: methods of estimation for the 1973 exact match study. U.S. Department of Health and Human Services, Social Security Administration, SSA Publication No. 13-11750.

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

ZIESCHANG, K.D. (1986). A generalized least quares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

ZIESCHANG, K.D. (1987). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. Document non publié, U.S. Bureau of Labour Statistics.

Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 221–232
Statistique Canada

Variante de la méthode itérative du quotient

H. LOCK OH et FRITZ SCHEUREN[1]

RÉSUMÉ

Les auteurs présentent une méthode d'estimation par le quotient qui est une combinaison de la méthode classique et de la méthode itérative du quotient et qui est utilisée lorsque les fréquences de population N_{ij} d'un tableau de contingence sont connues mais qu'un certain nombre des fréquences observées n_{ij} sont faibles ou nulles. Ils décrivent la façon dont cette méthode a évolué dans le cadre du programme de la statistique du revenu des sociétés (Corporate Statistics of Income Program - CSIP) au cours des dernières années. Enfin, ils envisagent d'autres modifications pour l'avenir et en font l'analyse.

MOTS CLÉS: Méthode itérative du quotient; estimation par le quotient; biais et variance conditionnels.

1. INTRODUCTION

La technique d'estimation itérative par le quotient, ou méthode itérative du quotient, est couramment utilisée dans les enquêtes par sondage. Ses applications varient selon la nature du plan de sondage, la quantité d'information supplémentaire disponible et l'importance des erreurs d'observation (attribuables à la non-réponse ou au sous-dénombrement).

Deming et Stephan (1940) ont été les premiers à proposer l'utilisation de la méthode itérative du quotient pour établir une correspondance entre les chiffres de population et des données d'échantillon du recensement de la population des États-Unis de 1940. Ils n'ont pas tardé par la suite à approfondir leurs propres recherches (Deming 1943; Stephan 1942). Depuis lors, plusieurs autres ont redéfini à leur façon cette méthode (Fienberg 1970) peut-être parce que intuitivement, l'algorithme d'itération utilisé paraît intéressant.

Cette méthode a été aussi l'objet de nombreux perfectionnements. Par exemple, Ireland et Kullback (1968) ont fait d'importantes recherches sur la convergence de l'algorithme. Comme on pouvait s'y attendre, les spécialistes de Statistique Canada et du U.S. Bureau of the Census se sont penchés longuement sur l'application du processus itératif dans les recensements et les sondages, particulièrement lorsque le processus ne va pas jusqu'à la convergence complète. (voir par exemple Brackstone et Rao 1979; Fan *et coll.* 1981). Oh et Scheuren (1978a) donnent une bibliographie assez complète concernant la recherche statistique qui s'est faite sur l'itération avant 1978.

Habituellement, lorsqu'on discute de la méthode itérative du quotient, on suppose que deux séries (ou plus) de fréquences marginales, disons $N_{i.}$ et $N_{.j}$, sont connues mais que les fréquences de cellules N_{ij} ne peuvent être estimées qu'à partir de l'échantillon. Lorsque les N_{ij} sont aussi connues, on choisit spontanément l'estimateur par le quotient usuel avec des poids N_{ij}/n_{ij}, à moins que la taille des échantillons correspondants n_{ij} ne soit "trop faible".

Le présent article décrit une méthode hybride qui fait intervenir la méthode classique d'estimation par le quotient et la méthode itérative du quotient dans les cas où les fréquences de cellules N_{ij} sont connues mais que certaines des fréquences observées n_{ij} sont faibles (ou nulles). La section 2 renferme une description de la méthode. Dans la section 3, nous commentons certains résultats empiriques produits lors de l'application de la méthode dans le cadre du programme de la statistique du revenu des sociétés (Corporate Statistics of Income Program - CSIP). Dans la section 4, nous faisons une brève rétrospective de la question et proposons des sujets de recherche.

[1] H. Lock Oh et Fritz Scheuren, Statistics of Income Division, Internal Revenue Service, 1111 Constitution Avenue N.W., Washington, D.C. 20224, États-Unis.

2. ESTIMATION ITÉRATIVE PAR LE QUOTIENT

2.1 Considérations d'ordre général

L'estimation itérative par le quotient suppose habituellement que deux (ou plus de deux) fréquences marginales – par exemple, $N_{i.}$ et $N_{.j}$ – sont connues mais que les fréquences de cellules N_{ij} ne peuvent être estimées qu'à partir de l'échantillon par $\tilde{N}_{ij}$; nous avons ci-dessous une représentation graphique du problème (Deming 1943):

$$
\begin{array}{c|cccccc|c}
 & 1 & 2 & & & S & & \\
\hline
1 & N_{11} & N_{12} & & \cdots & N_{1S} & & N_{1.} \\
2 & N_{21} & N_{22} & & \cdots & N_{2S} & & N_{2.} \\
\vdots & \vdots & \vdots & & \boxed{N_{ij}} & \vdots & & \vdots\, N_{i.} \\
R & N_{R1} & N_{R2} & & \cdots & N_{RS} & & N_{R.} \\
\hline
 & N_{.1} & N_{.2} & \cdots & N_{.j} & \cdots & N_{.S} & N
\end{array}
$$

où $i = l, \ldots, R$ et $j = l, \ldots, S$. Le tableau de l'échantillon correspondant est

$$
\begin{array}{c|cccccc|c}
 & 1 & 2 & & & S & & \\
\hline
1 & n_{11} & n_{12} & & \cdots & n_{1S} & & n_{1.} \\
2 & n_{21} & n_{22} & & \cdots & n_{2S} & & n_{2.} \\
\vdots & \vdots & \vdots & & \boxed{n_{ij}} & \vdots & & \vdots\, n_{i.} \\
R & n_{R1} & n_{R2} & & \cdots & n_{RS} & & n_{R.} \\
\hline
 & n_{.1} & n_{.2} & \cdots & n_{.j} & \cdots & n_{.S} & n\ .
\end{array}
$$

Lorsqu'on applique l'algorithme d'itération à un échantillonnage aléatoire simple, on commence par poser

$$\tilde{N}_{ij} = \frac{N}{n} n_{ij}, \tag{2.1}$$

puis on corrige proportionnellement les valeurs $\tilde{N}_{ij}$ de telle sorte que les équations

$$\sum_{j}^{S} \tilde{N}_{ij} = N_{i.} \tag{2.2}$$

et

$$\sum_{i}^{R} \tilde{N}_{ij} = N_{.j} \tag{2.3}$$

soient satisfaites. Chaque étape de l'algorithme débute par les résultats de l'étape précédente, les valeurs $\widetilde{N}_{ij}$ ne cessant de varier; le processus se termine lorsqu'un nombre déterminé d'étapes ont été réalisées ou lorsque les expressions (2.2) et (2.3) sont satisfaites simultanément au degré d'approximation voulu. (Voir Oh et Scheuren (1983) pour plus de détails; voir Ireland et Scheuren (1975) pour l'application de cet algorithme à des tableaux multidimensionnels et l'étude des questions relatives aux ressources informatiques.)

Par une application de la théorie de l'information minimum discriminante (Kullback 1968), il est possible de montrer (voir Ireland et Kullback 1968) qu'étant donné certaines conditions de régularité si on ne connaît que les valeurs $N_{i.}$ et $N_{.j}$, les $\widetilde{N}_{ij}$ calculés par un processus convergent sont asymptotiquement non biaisés, suivent une distribution normale et ont une variance minimum (c'est-à-dire, meilleurs estimateurs asymptotiquement normaux ou MEAN). Ces résultats théoriques justifient en partie l'estimateur itératif d'une caractéristique générale d'enquête Y_{ijk} (par exemple, revenu ou actif) lorsque nous cherchons à estimer le total de population

$$Y = \sum_{i}^{R} \sum_{j}^{S} \sum_{k}^{N_{ij}} Y_{ijk} \tag{2.4}$$

au moyen, par exemple, de la statistique

$$\widetilde{Y} = \sum_{i}^{R} \sum_{j}^{S} \frac{\widetilde{N}_{ij}}{n_{ij}} \left(\sum_{k}^{n_{ij}} Y_{ijk} \right). \tag{2.5}$$

Lors du traitement des données d'enquête, on attribue habituellement un poids

$$\widetilde{W}_{ij} = \frac{\widetilde{N}_{ij}}{n_{ij}} \tag{2.6}$$

à chaque enregistrement du fichier. Il convient de souligner qu'une des caractéristiques de l'algorithme itératif est que $\widetilde{N}_{ij}$ est nécessairement égal à 0 si $n_{ij} = 0$. Pour des raisons de commodité, nous poserons $\widetilde{W}_{ij} = 0$ dans de telles circonstances.

Dans le reste de cette section, nous allons nous intéresser surtout aux propriétés conditionnelles des divers estimateurs étudiés. Comme le font valoir Holt et Smith (1979) et Rao (1985), cette approche présente un intérêt considérable. (Soit dit en passant, soulignons que Brackstone et Rao (1979), entre autres, ont analysé le comportement conditionnel de l'estimateur itératif. Cependant, les conditions étaient définies en fonction des fréquences marginales de l'échantillon $n_{i.}$ et $n_{.j}$.)

2.2 Biais conditionnel

À la suite de Oh et Scheuren (1983), nous allons nous arrêter aux propriétés conditionnelles de $\widetilde{Y}$, étant donné $\underline{n} = (n_{11}, n_{12} \ldots, n_{RS})$. En particulier, définissons $\bar{Y}_{ij}$ comme la moyenne de la population pour le ij-ième sous-groupe. L'espérance mathématique conditionnelle de $\widetilde{Y}$ est donc

$$\mathrm{E}(\widetilde{Y} \mid \underline{n}) = \sum_{i}^{R} \sum_{j}^{S} \widetilde{N}_{ij} \bar{Y}_{ij} = Y + \sum_{i}^{R} \sum_{j}^{S} (\bar{Y}_{ij} - \bar{Y}) (\widetilde{N}_{ij} - N_{ij}). \tag{2.7}$$

Par conséquent $\widetilde{Y}$ est conditionnellement biaisé, l'importance du biais variant selon la structure de la population et selon que le processus d'itération converge ou non. (Évidemment, lorsque le processus est convergent, $E(\widetilde{N}_{ij}) = N_{ij}$ asymptotiquement et non conditionnellement.)

En utilisant les formules usuelles de l'analyse de variance (voir Scheffé 1959), nous avons

$$(\bar{Y}_{ij} - \bar{Y}) = (\bar{Y}_{i.} - \bar{Y}) + (\bar{Y}_{.j} - \bar{Y}) + (\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}); \tag{2.8}$$

le biais conditionnel, étant donné $\underline{n}$, peut donc être exprimé comme suit:

$$\text{Biais}(\tilde{Y} \mid \underline{n}) = \sum_{i}^{R} (\bar{Y}_{i.} - \bar{Y})(\tilde{N}_{i.} - N_{i.}) + \sum_{j}^{S} (\bar{Y}_{.j} - \bar{Y})(\tilde{N}_{.j} - N_{.j})$$

$$+ \sum_{i}^{R} \sum_{j}^{S} (\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y})(\tilde{N}_{ij} - N_{ij}). \tag{2.9}$$

Si le processus est convergent, les deux premiers termes de la formule du biais conditionel deviennent nuls. Pour que le troisième terme de la formule devienne nul, qu'il y ait convergence ou non, il suffit que les Y_{ij} soient telles qu'il n'y ait aucune interaction. Dans les grandes enquêtes portant sur de nombreuses variables, cette hypothèse est irréaliste; dans la pratique toutefois, l'interaction est souvent un élément mineur de la décomposition de Y_{ij}; par conséquent, l'estimateur itératif par le quotient peut, dans beaucoup de cas, ne comporte qu'un faible biais même pour des échantillons de taille moyenne.

2.3 Variance conditionnelle

De nombreuses études ont été faites sur la variance conditionnelle et non conditionnelle de l'estimateur itératif par le quotient (voir, par exemple, Binder 1983; Causey 1972; Bankier 1986; Fan *et coll.* 1981; Brackstone et Rao 1979). Dans un de nos ouvrages antérieurs (voir sous-section 3.2), nous avons utilisé des méthodes de répétées d'échantillons (Leszcz, Oh et Scheuren 1983). Ces méthodes (qui équivalaient à définir les conditions en fonction des fréquences marginales de l'échantillon) se sont avérées coûteuses, peu flexibles et relativement instables, ce qui nous a amené à aborder le problème de l'estimation de la variance conditionnelle d'une manière moins élaborée (bien que les conditions étaient plus strictes).

Afin de justifier l'approche que nous utilisons, considérons la variance conditionnelle de $\tilde{Y}$, étant donné $\underline{n}$. En poussant un peu plus loin la recherche entreprise par Oh et Scheuren (1983), il est possible de montrer que

$$\text{Var}(\tilde{Y} \mid \underline{n}) = \sum_{i}^{R} \sum_{j}^{S} n_{ij} (\tilde{W}_{ij})^2 \left(1 - \frac{n_{ij}}{N_{ij}}\right) V_{ij} \tag{2.10}$$

où les V_{ij} sont les variances de la population du ij-ième sous-groupe; de plus, si $N_{ij} = 0$ ou 1, nous posons $V_{ij} = 0$ (nous appliquons aussi la règle $0/0 = 0$ dans l'équation (2.10)).

L'équation (2.10) est valide, peu importe que le processus d'itération soit convergent ou non. Malgré cela, cette expression a fait l'objet de peu d'analyses parce qu'on ne peut l'adapter facilement à l'estimation de la variance conditionnelle. Cet inconvénient est surtout attribuable à notre incapacité de calculer des estimateurs stables de V_{ij} lorsque les n_{ij} sont faibles. Pour surmonter la difficulté, nous nous sommes mis à étudier des méthodes de regroupement fondées sur la valeur du poids d'itération. Nous commençons par définir $\tilde{W}_{ij}$ comme une approximation de N_{ij}/n_{ij}, ce qui donne

$$\text{Var}(\tilde{Y} \mid \underline{n}) \cong \sum_{i}^{R} \sum_{j}^{S} n_{ij} \tilde{W}_{ij} (\tilde{W}_{ij} - 1) V_{ij}. \tag{2.11}$$

Or, si les valeurs de $\tilde{W}_{ij}$ sont classées par ordre croissant et qu'elles s'avèrent se situer dans un court intervalle, le fait de calculer la moyenne de ces poids pour des groupes (ordonnés)

d'au moins 25 observations chacun ($n_g \geq 25$) chacun modifiera très peu la valeur de l'équation (2.11). En revanche, cela nous permettra de calculer des estimations par la stratification a posteriori de V_{ij} fondées sur le regroupement. C'est la méthode que nous avons utilisée dans la section 3.

Nous nous permettons ici de faire une dernière remarque. La méthode que nous proposons dans cet article est stable et implique des calculs relativement simples. Cependant, le nombre limité d'applications ne permet pas de conclure à l'utilité de la méthode et bien que nous estimions que cette méthode vaut la peine d'être étudiée, nous n'en préconisons aucunement l'application générale pour l'instant.

2.4 Version modifiée de la méthode itérative du quotient

Comme nous l'avons déjà signalé, les $\widetilde{N}_{ij}$ sont des MEAN étant donné des conditions relativement générales. Cela ne veut pas dire pour autant que $\widetilde{Y}$ aura toutes les propriétés d'un MEAN. En effet, s'il n'existe pas une forte corrélation entre la caractéristique Y et les variables utilisées lors de l'estimation itérative par le quotient, la variance de l'estimateur $\widetilde{Y}$ pourrait en subir plus fortement les conséquences que celle d'un estimateur ordinaire par le quotient

$$\widetilde{\widetilde{Y}} = \sum_{j}^{S} \left(\frac{N_{.j}}{n_{.j}}\right)\left(\sum_{i}^{R} \sum_{k}^{n_{ij}} Y_{ijk}\right). \tag{2.12}$$

Nous savons d'après l'expérience qu'un même échantillon peut subir des effets tant négatifs que positifs. La tâche du statisticien consiste en quelque sorte à conserver les effets positifs tout en réduisant au minimum les effets négatifs.

Il ne semble pas y avoir de solution générale pour résoudre ce dilemme, bien que nous ayons obtenu des résultats modestes dans notre modèle d'application avec deux méthodes qui débordent peut-être le cadre de notre analyse (voir sous-sections 3.2 et 3.3 pour les résultats).

Habituellement, lorsqu'on discute de la méthode itérative du quotient on suppose que les fréquences marginales $N_{i.}$ et $N_{.j}$ sont connues et que les fréquences de cases N_{ij} ne peuvent être estimées qu'à partir de l'échantillon. Dans notre modèle d'application, nous connaissons les fréquences N_{ij} et avons recours à l'estimation itérative par le quotient pour traiter systématiquement les cellules du tableau dont les n_{ij} sont faibles. Nous pouvions évidemment utiliser les méthodes classiques de regroupement (voir, par exemple, Cochran 1977; Fuller 1966) mais celles-ci ne semblaient pas convenir pour des raisons que nous verrons plus tard.

On peut reconnaître que la méthode itérative du quotient est un moyen satisfaisant de traiter les cellules à faible fréquence dans ce contexte mais que dire des cellules à fréquence plus élevée? Dans ce cas, il serait sûrement préférable d'utiliser l'estimateur ordinaire par le quotient. Si tel était le cas, le biais conditionnel pour ces cellules "à forte fréquence" serait nul mais quel effet cela aurait-il sur les autres cellules? Ce raisonnement nous a amené à adopter une méthode d'estimation hybride selon laquelle on utilise l'estimateur ordinaire par le quotient pour les cellules dont les n_{ij} sont élevées. Ces cellules sont ensuite extraites des tableaux de la population et de l'échantillon et les cellules restantes du tableau de l'échantillon sont soumises à l'estimation itérative par le quotient afin de rajuster leurs fréquences en fonction des fréquences marginales corrigées.

En ce qui a trait aux cellules restantes à fréquence très faible, on a appliqué une seconde méthode visant à réduire les effets négatifs que pouvait avoir l'estimation itérative par le quotient sur certaines variables. Nous avons restreint le processus d'itération de manière que les poids $\widetilde{W}_{ij}$ ne s'écartent pas "trop" du poids initial. (À propos, cette forme de restriction est courante dans l'estimation classique par le quotient, voir par exemple Hanson 1978).

La version restreinte de la méthode itérative du quotient est comparable à la méthode utilisée lorsque seulement une cellule du tableau de l'échantillon renferme une fréquence "élevée".

Elle est comparable en ce sens que nous bornons le $\widetilde{W}_{ij}$ se rapportant à la cellule qui doit être restreinte, puis extrayons respectivement des tableaux de la population et de l'échantillon la fréquence estimée $\widetilde{N}_{ij} = \widetilde{W}_{ij}\, n_{ij}$ et la fréquence n_{ij} pour la cellule en question, pour enfin corriger les observations restantes.

Ces "solutions" partielles soulèvent trois problèmes particuliers. Il y a tout d'abord la définition plutôt arbitraire d'une cellule à fréquence "élevée" et d'un coefficient de pondération qui s'écarte "trop" de sa valeur initiale. En deuxième lieu, si nous étions disposés à utiliser l'estimation ordinaire par le quotient pour les cellules à fréquence "élevée", il faudrait savoir pourquoi ne pouvions-nous pas utiliser les méthodes classiques de regroupement pour les cellules restantes. Le troisième problème a trait aux propriétés de la convergence de l'algorithme itératif lorsque nous utilisons la méthode hybride. Par exemple, les études qui ont été faites sur l'estimation itérative par le quotient montrent très clairement que les tableaux qui renferment un trop grand nombre de zéros seront très instables et que le processus d'itération pourrait ne pas converger (voir, par exemple, Oh et Scheuren 1978b; Ireland et Scheuren 1975). Ce problème revêt un intérêt particulier puisque les modifications proposées ont justement pour but d'introduire des zéros dans le tableau. Si ces zéros sont *mal distribués*, cela pourrait avoir des conséquences très néfastes pour le taux de convergence et même pour la qualité des estimateurs. Nous nous sommes donc proposé au départ d'appliquer la méthode hydride relativement peu souvent. Le présent article n'a pas pour but de résoudre les problèmes évoqués ci-dessus (si ceux-ci peuvent être effectivement résolus.) Néanmoins, nous y reviendrons à la section 3, où nous les réexaminerons en fonction du modèle d'application utilisé, et à la section 4, où nous proposerons des sujets de recherche.

3. ESTIMATION ITÉRATIVE PAR LE QUOTIENT DANS LE CADRE DU PROGRAMME DE LA STATISTIQUE DU REVENU DES SOCIÉTÉS

3.1. Contexte

Depuis plus de 70 ans, l'Internal Revenue Service des États-Unis produit chaque année des statistiques fondées sur les déclarations d'impôt des sociétés. De fait, les données des sociétés constituent l'élément essentiel du programme de la statistique du revenu, qui est l'appellation donnée à l'ensemble des séries de données non administratives produites par l'Internal Revenue Service à l'intention du public.

Jusqu'en 1951, les statistiques des sociétés étaient fondées sur un dénombrement complet des déclarations produites. Depuis lors, on utilise un échantillon aléatoire stratifié formé d'environ 90,000 déclarations annuellement (environ 3,000,000 déclarations sont faites chaque année). L'actif et le revenu sont les principales variables de stratification (Jones et McMahon 1984). La stratification par industrie a été longtemps envisagée mais la qualité des codes d'industrie inscrits par les contribuables ne semblait pas assez bonne pour justifier un usage généralisé de ce mode de stratification. Au niveau d'agrégation le plus détaillé, par exemple, 20% et plus des codes inscrits par les contribuables sont modifiés durant le traitement statistique. Néanmoins, vu l'importance des statistiques des industries, il semblait encore légitime de tenter d'utiliser les données administratives des industries pour stratifier l'échantillon a posteriori et c'est ce qu'on a essayé de faire pendant de nombreuses années (voir, par exemple, Westat Inc. 1974; Leszcz, Oh et Scheuren 1983).

Dans une étude pilote sur la stratification a posteriori réalisée par Westat au début des années 1970, on a réussi à réduire de façon notable les erreurs types pour un certain nombre de variables, notamment les recettes totales (pour lesquelles on a observé une réduction de l'erreur type d'environ 12%). Toutefois, des hausses d'erreur type ont été observées pour des variables qui ne sont pas liées étroitement à l'industrie (par exemple, dividendes versés aux actionnaires) mais il s'agissait de hausses mineures. En ce qui concerne les cellules à faible

fréquence, Westat a utilisé les méthodes classiques de regroupement pour combiner des strates formées a posteriori selon l'industrie à l'intérieur des strates d'échantillon existantes. On a continué à se préoccuper de la qualité des données administratives au niveau de l'industrie, particulièrement en ce qui a trait aux cases à faible fréquence; quoiqu'il en soit, la méthode Westat n'a jamais été mise en application à cause de questions opérationnelles plus pressantes.

Une importante série de compressions budgétaires imposées durant la période 1980-1982 ont entraîné une modification des plans de sondage et des méthodes d'estimation pour presque tous les projets qui constituent le programme de la statistique du revenu (voir, par exemple, Hinkins et Scheuren 1986; Scheuren, Schwartz et Kilss 1984); on a notamment réduit la taille des échantillons pour l'enquête sur les sociétés et cette mesure, quoique partiellement annulée ultérieurement, a contribué à rouvrir le débat sur la stratification a posteriori par industrie.

L'application de la méthode itérative du quotient semblait présenter plus d'intérêt que les travaux de Westat parce qu'on n'était pas tant préoccupé par la qualité des fréquences marginales pour l'industrie que par celle des fréquences de cellules. En outre, il n'était plus nécessaire de mettre en application une méthode de regroupement.

3.2 Version modifiée de la méthode itérative du quotient - résultats préliminaires

Lorsque nous avons appliqué la version pure de la méthode itérative du quotient à l'échantillon de l'année d'imposition 1979, nos principaux clients nous ont fait part de leurs inquiétudes à ce sujet. Ils craignaient surtout que des facteurs de correction élevés puissent avoir un effet négatif sur certaines statistiques. Nous-mêmes, à la lecture des résultats, craignions de ne pas avoir réalisé nos objectifs en ce qui concerne les combinaisons "industrie-strate d'échantillon" pour lesquelles le nombre d'observations de l'échantillon était élevé. En conséquence, ces résultats n'ont jamais été utilisés et les données de l'année d'imposition 1979 ont été publiées sur la base d'un échantillonnage stratifié normal (NORM).

Les recherches se sont poursuivies toutefois et en 1983, on en est arrivé à comparer entre elles les erreurs quadratiques moyennes de six variantes de la méthode itérative du quotient et à les comparer aux résultats de nos études antérieures (Leszcz, Oh et Scheuren 1983). Ces variantes comportaient trois méthodes "pures":

PRRE: Méthode itérative du quotient "classique" avec convergence (Deming et Stephan 1940).

PRRE (200): Correction par le quotient pour les cellules avec échantillons de 200 déclarations ou plus et méthode itérative du quotient "classique" pour les cellules restantes avec convergence.

PRRE (400): Correction par le quotient pour les cellules avec échantillons de 400 déclarations ou plus et méthode itérative du quotient "classique" pour les cellules restantes avec convergence.

En outre, il y avait trois versions de l'estimation itérative par le quotient bornée, où les bornes étaient ($\sqrt{2/3}$, $\sqrt{3/2}$). Ces trois versions étaient:

BRRE: Estimation itérative par le quotient bornée (2 cycles).

BRRE (200): Correction par le quotient pour les cellules avec échantillons de taille $n = 200$ et estimation itérative par le quotient bornée (2 cycles) pour les cellules restantes.

BRRE (400): Correction par le quotient pour les cellules avec échantillons de taille $n = 400$ et estimation itérative par le quotient bornée (2 cycles) pour les cellules restantes.

En ce qui concerne l'estimation itérative par le quotient bornée, nous n'étions pas sûrs au départ que la convergence complète était possible; nous avons donc effectué une simplification et n'avons soumis les équations conditionnelles (2.2) et (2.3) qu'à deux cycles.

Pour comparer les erreurs quadratiques moyennes (EQM), nous avons tiré des pseudo-demi-échantillons répétés conçus de la même façon que l'échantillon général. Cela s'est fait

en cinq étapes: 1) formation des demi-échantillons; 2) classification des données des demi-échantillons selon deux critères - selon la strate d'échantillon originale et selon l'industrie principale (strate formée a posteriori); 3) calcul d'une série de poids pour chaque demi-échantillon et pour chaque estimateur; 4) calcul de l'estimation de certaines unités par l'application de poids à des données d'échantillon pour chaque demi-échantillon et 5) calcul de l'EQM fondé sur les variations des estimations produites par chaque demi-échantillon. Pour des raisons monétaires, nous n'avons utilisé que 14 ensembles de demi-échantillons.

Les résultats de cette comparaison sont présentés sommairement dans le tableau 1; les chiffres concernant le nombre de déclarations sont ce à quoi on s'attendait. Des réductions de près de 100% ont été observées pour les estimations PRRE, PRRE(200), et PRRE(400). Le fait d'appliquer les bornes $\sqrt{2/3}$ et $\sqrt{3/2}$ et de se limiter à deux cycles s'est traduit par des réductions moins fortes mais tout de même appréciables. Le tableau 1 indique aussi des réductions de l'EQM pour les recettes totales, une variable clé, mais ces réductions n'ont pas l'ampleur des premières.

Nous avons inclus dans le tableau 1 les résultats relatifs au crédit d'impôt à l'emploi pour illustrer l'ambivalence qui peut exister dans le cas des variables qui ne sont pas liées étroitement à l'industrie. En particulier, nous observons des hausses (modestes) de l'erreur quadratique moyenne pour cette variable du fait, probablement, que le crédit d'impôt à l'emploi est moins dépendant de la classification des industries utilisée dans cette étude.

En ce qui concerne les recettes totales, il convient de souligner que les réductions de l'erreur quadratique moyenne (attribuables au remplacement de l'estimation initiale (NORM) par l'estimation itérative par le quotient se comparent toutes avantageusement aux résultats de l'étude pilote de Westat. Même si cette constatation est encourageante, la situation a beaucoup évolué entre le moment où Westat a fait connaître les résultats de son étude et la parution de l'ouvrage de Leszcz, Oh et Scheuren (1983). Si nous voulions faire une comparaison vraiment révélatrice, à laquelle nous n'avons pas songé jusqu'à maintenant, il suffirait d'appliquer les méthodes classiques de regroupement et notre version modifiée de la méthode itérative du quotient *à la même série de données.*

Nous nous permettons ici une dernière remarque à propos du tableau 1; les chiffres de ce tableau ont été établis à partir des données administratives des industries qui ont servi à la stratification a posteriori. Comme la classification des industries n'est pas la même pour les besoins administratifs et les besoins statistiques, les chiffres du tableau sont susceptibles d'exagérer les réductions observées pour les statistiques publiées puisque le code d'une forte proportion des entités (plus de 20%) est modifié au cours du traitement de l'échantillon des sociétés.

3.3 Version modifiée de la méthode itérative du quotient - résultats courants

Depuis l'année d'imposition 1980, nous produisons et publions régulièrement les statistiques des sociétés à l'aide de la méthode itérative du quotient bornée BRRE (200) (U.S. Department of Treasury 1984). Pour les années d'imposition 1983 et suivantes, nous avons fait les modifications décrites dans la sous-section 2.3 de manière à pouvoir calculer des variances conditionnelles approximatives. L'année d'imposition 1984 a été la première pour laquelle de telles variances ont été publiées (U.S. Department of Treasury 1987). De plus, afin de vérifier les résultats antérieurs, nous avons comparé pour l'année d'imposition 1984 la variance conditionnelle de la version modifiée de la méthode itérative du quotient avec la variance qui aurait été estimée si nous avions eu recours à l'échantillonnage stratifié normal. Avant d'analyser les résultats de cette comparaison, il serait utile de fournir quelques précisions sur le modèle utilisé pour 1984.

Dans un ouvrage antérieur (Leszcz, Oh et Scheuren 1983) et pour 1984, l'univers des formules 1120 et 1120S des déclarations des sociétés a été réparti entre 58 grandes branches d'activité. Pour 1984, chaque branche d'activité a été subdivisée en 14 strates d'échantillon à chacune des deux années durant lesquelles l'échantillon devait être choisi. Quelques-unes des principales branches d'activité étaient si clairsemées que nous avons réduit sans tarder le

Tableau 1

Réduction de l'erreur quadratique moyenne (EQM) en pourcentage de l'EQM correspondante pour l'échantillonnage stratifié normal

Estimateur	Nombre de déclarations	Recettes totales	Crédit d'impôt à l'emploi
Estimateurs de la méthode itérative du quotient "pure":			
PRRE	98.6	8.3	−3.09
PRRE (400)	98.6	9.2	−3.09
PRRE (200)	98.6	11.9	−3.09
Estimateurs de la méthode itérative du quotient bornée:			
BRRE	74.0	13.8	+1.09
BRRE (400)	73.4	15.6	+1.09
BRRE (200)	72.3	17.4	+1.09

Remarque: Les pourcentages indiqués sont des moyennes simples des réductions (en pourcentage) observées dans les 56 grandes branches d'activité ayant servi à la stratification a posteriori. On remarquera que les pourcentages indiqués dans la partie supérieur de la première colonne (nombre de déclarations) *ne sont pas tout à fait* égaux à 100 parce que l'estimation itérative par le quotient s'est appliquée à toutes les sociétés et que les N_{ij} et n_{ij} ont été définies en conséquence tandis que le tableau porte uniquement sur les sociétés actives (environ 90%). Les pourcentages indiqués dans la partie inférieure de la première colonne (nombre de déclarations) diffèrent l'un de l'autre et des pourcentages indiqués dans la partie supérieure de cette colonne parce que l'estimation itérative ne s'est pas faite jusqu'à la convergence. Cela n'est plus le cas depuis l'année d'imposition 1985.

nombre de ces industries à 56 par la technique du regroupement. Malgré cela, nous avions encore un tableau de taille très appréciables (1,568 cellules).

Il convient de souligner que l'univers comptait 414 cellules à fréquence nulle à l'origine et que l'échantillon en a généré 125 autres. Avant l'estimation itérative par le quotient, nous avons extrait 96 cellules qui renfermaient 200 observations de l'échantillon ou plus et les avons corrigées séparément par la méthode du quotient. (En tout, 57% de l'échantillon de formules 1120 et 1120S a été corrigé de cette façon). Enfin, il a fallu borner 73 autres cellules au cours de l'estimation itérative par le quotient. Cela signifie que 708 cellules en tout (45%) ont été considérées comme des cellules à fréquence nulle durant l'estimation itérative par le quotient.

On a amorcé l'estimation itérative par le quotient en appliquant l'estimateur stratifié normal à chaque cellule du tableau. On a utilisé deux séries de contraintes marginales: 1) par industrie et période d'échantillonnage et 2) par strate d'échantillon et période d'échantillonnage. Dans les données publiées pour 1984 et dans les comparaisons faites ici, l'estimation itérative par le quotient n'a pas été effectuée jusqu'à la convergence; le processus s'est limité en effet à deux cycles. (À ce propos, l'incertitude qui entourait le biais conditionnel de cette méthode nous a amené à effectuer l'estimation itérative par le quotient jusqu'à la convergence pour les données de l'échantillon de 1985.)

Les travaux de 1984 ont eu pour effet de réduire sensiblement l'erreur type des fréquences pour l'ensemble des industries et pour chacune d'elles, les réductions étant toutefois environ la moitié moindres que celles indiquées dans le tableau 1. On a aussi observé des réductions modérées dans le cas des recettes totales (8.7%) tandis que pour de nombreuses variables comme le crédit d'impôt à l'emploi et le bénéfice net, l'erreur type globale est demeurée à peu près la même (voir U.S. Department of Treasury 1987, pour plus de renseignements). Comme nous l'avons déjà souligné, ces différences de résultats peuvent être attribuées en partie à la nature des relations conditionnelles utilisées (Holt et Smith 1979). Les résultats initiaux dépendaient des relations conditionnelles basées sur les fréquences marginales de l'échantillon $n_{i.}$ et $n_{.j}$ tandis que les résultats ultérieurs reposaient sur des relations conditionnelles plus complexes.

Nous cherchons encore d'autres raisons qui pourraient expliquer pourquoi les réductions d'erreur type observées en 1984 sont moins fortes que celles observées dans les études antérieures. Une des raisons qui nous viennent immédiatement à l'esprit est la manière dont

nous avons groupé les données des cellules à fréquence moins élevée, y compris le calcul de la moyenne des facteurs de pondération $\bar{W}_{ij}$ qui en résulte et l'estimation des V_{ij} fondée sur le regroupement de cases. Le fait d'avoir totalisé les données selon la classification industrielle destinée à des fins statistiques au lieu de celle destinée à des fins administratives, comme c'est le cas dans le tableau 1, peut aussi expliquer largement ces différences.

4. CONCLUSIONS ET SUJETS DE RECHERCHE

4.1 Généralités

La version modifiée de la méthode itérative du quotient pour l'échantillon de sociétés représente sûrement un progrès par rapport à l'ancienne méthode de l'échantillonnage stratifié normal. Cependant, il y a encore un certain nombre de points qui nous agacent. Par exemple, il y aurait lieu d'approfondir le lien qui existe entre les méthodes classiques de regroupement et la version modifiée de la méthode itérative du quotient. Toutefois, il ne suffit pas pour cela d'examiner les modifications que l'on peut apporter aux méthodes d'estimation. Il semble tout aussi important de se pencher sur le plan de sondage fondamental. Enfin, l'approximation de la variance utilisée dans le présent article mérite d'être approfondie. Le fait d'avoir opté pour la stabilité et la simplicité des calculs peut nous avoir coûté cher. Comme nous l'avons dit plus tôt, les ouvrages de statistique regorgent de bonnes solutions qui méritent d'être comparées sous tous leurs aspects avec ce qui se fait présentement.

4.2 Questions relatives à l'estimation

Intuitivement, nous sommes fortement tentés de mettre au point une méthode de stratification a posteriori qui nous permettrait de modifier *graduellement* les relations conditionnelles basées sur les fréquences marginales pour qu'elles soient basées également sur les fréquences de cellules. L'inconvénient de la méthode que nous utilisons actuellement est qu'elle a un aspect *temporaire*. De fait, cette méthode n'est qu'un pis-aller en attendant que nous puissions accroître la qualité des données administratives des industries. Notre principal souci est de réduire la variation de réponse qui découle des erreurs de déclaration ou de traitement. Même si nous ne réussissons pas à accroître directement la qualité des données administratives, nous pouvons atténuer les effets de l'erreur de réponse en examinant les tableaux selon l'industrie et la strate d'échantillon sur plusieurs années. C'est ce que nous projetons de faire et c'est ce qui nous permettra peut-être de mieux intégrer d'une part la méthode itérative du quotient et d'autre part l'estimation par la stratification a posteriori fondée sur le regroupement de cellules.

4.3 Questions relatives au plan de sondage

Des données administratives de meilleure qualité pour les industries ont une utilité évidente dans la conception du plan de sondage. À l'heure actuelle, les coefficients de variation diffèrent très largement d'une industrie à l'autre et les industries de moindre importance sont très faiblement représentées. La stratification a posteriori, aussi poussée soit-elle, ne peut corriger à elle seule la situation. Une meilleure répartition de l'échantillon entre les industries, et dans le temps, semble compter parmi les mesures les plus urgentes (voir Hinkins, Jones et Scheuren 1987).

REMERCIEMENTS

Comme c'est le cas pour presque tous les travaux de recherche appliquée, les auteurs tiennent à remercier les nombreuses personnes qui ont rendu possible cet article. Ils veulent tout d'abord souligner la contribution notable d'Homer Jones, responsable de l'échantillon des

sociétés, qui a été habilement secondé par Richard Collins. Nat Shaifer a fourni des données éclairantes sur le degré de comparabilité de la classification industrielle utilisée à des fins administratives et de celle utilisée à des fins statistiques. Mike Leszcz a joué un rôle de premier plan en exposant les résultats préliminaires de notre étude à l'assemblée de l'American Statistical Association de 1983 tenue à Toronto. Les auteurs tiennent aussi à exprimer leur reconnaissance à Rod Little, qui par des échanges a contribué à améliorer la qualité de cet article, et aux arbitres, qui ont apporté de nombreuses suggestions utiles. Enfin, Bettye Jamerson a secondé les auteurs dans la préparation du manuscrit.

BIBLIOGRAPHIE

BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association,* 81, 1074-1079.

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review,* 51, 279-292.

BRACKSTONE, G.J., et RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā,* Sér. C, 41, 97-114.

CAUSEY, B.D. (1972). Sensitivity of raked contingency table totals to changes in problem conditions. *Annals of Mathematical Statistics,* 43, 656-658.

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

DEMING, W.E. (1943). *Statistical Adjustment of Data.* New York: Dover.

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

FAN, M.C., WOLTMAN, H.F., MISKURA, S.M., et THOMPSON, J.H. (1981). 1980 Census variance estimation procedure. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 176-181.

FIENBERG, S.E. (1970). An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics,* 907-917.

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

HANSON, R. (1978). The Current Population Survey: design and methodology. Technical Paper No. 40, U. S. Bureau of the Census.

HINKINS, S., JONES. H., et SCHEUREN, F. (1987). Mise à jour des probabilités de sélection des déclarations d'impôt dans le cadre du programme de la statistique du revenu des sociétés. Communication présentée au Symposium international sur les utilisations statistiques des données administratives, Ottawa.

HINKINS, S., et SCHEUREN, F. (1986). L'imputation par la méthode "hot deck" appliquée à un plan d'échantillonnage à deux degrés. *Techniques d'enquête,* 12, 181-195.

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

IRELAND, C.T., et KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika,* 55, 179-188.

IRELAND, C.T., et SCHEUREN, F.J. (1975). The rake's progress. Dans *Computer Programs for Contingency Table Analysis, Washington, D.C.:* The George Washington University, 155-216.

JONES, H., et MCMAHON, P.B. (1984). Sampling corporation income tax returns for Statistics of Income, 1951 to present. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 269-274.

KULLBACK, S. (1968). *Information Theory and Statistics.* New York: Dover.

LESZCZ, M., OH, H.L., et SCHEUREN, F. (1983). Modified raking estimation in the corporate SOI program. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 107-111.

OH, H.L., et SCHEUREN, F.J. (1978a). Multivariate raking ratio estimation in the 1973 exact match study. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 716-722.

OH, H.L., et SCHEUREN, F.J. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 723-728.

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

RAO, J.N.K. (1985). Inférence conditionnelle dans les enquêtes par sondage. *Techniques d'enquête,* volume 11, 15-31.

SCHEFFÉ, H. (1959). *The Analysis of Variance.* 106-119. New York: John Wiley.

SCHEUREN, F., SCHWARTZ, O., et KILSS, B. (1984). Statistics from individual income tax returns: quality issues and budget cut impact. *Review of Public Data Use,* 12, 55-67.

STEPHAN, F.F. (1942). Iterative method of adjustment sample frequency tables when expected margins are known. *Annals of Mathematical Statistics,* 13, 166-178.

U. S. DEPARTMENT OF TREASURY (1984). Internal Revenue Service, *Statistics of Income - 1980 Corporation Income Tax Returns,* publication n° 16, U.S. Department of Treasury.

U. S. DEPARTMENT OF TREASURY (1987). Internal Revenue Service, *Statistics of Income - 1984 Corporation Income Tax Returns,* publication n° 16, U.S. Department of Treasury.

WESTAT, INC. (1974). Results of a study to improve sampling efficiency of statistics of corporation income. Bethesda, Maryland, (non publié).

Techniques d'enquête, décembre 1987
Vol. 13, n° 2, pp. 233–236
Statistique Canada

Comparaison de la méthode de Horvitz-Thompson et de la méthode de Hansen-Hurwitz

S.G. PRABHU-AJGAONKAR[1]

RÉSUMÉ

La méthode de Hansen-Hurwitz (1943) est réputée moins efficace que la méthode de Horvitz-Thompson (1952) qui est liée à un certain nombre de méthodes d'échantillonnage avec PIPT (probabilité d'inclusion proportionnelle à la taille). Le présent article démontre de façon simple la supériorité de la seconde méthode et présente donc un intérêt au point de vue pédagogique.

MOTS CLÉS: Méthodes d'échantillonnage; probabilité d'inclusion proportionnelle à la taille; forme quadratique définie positive.

1. INTRODUCTION

Soit U une population finie constituée de N unités identifiables $[U_1, U_2, \ldots, U_N]$. Deux nombres X_i et Y_i se rattachent à U_i, la i-ième unité de la population; les valeurs de X_i sont connues tandis que celles de Y_i sont inconnues mais fixes. De façon générale, X_i est une mesure de la taille de U_i qui est fortement correlé avec Y_i.

Si nous utilisons la méthode de Hansen-Hurwitz (1943) pour estimer le total de la population $T_y = Y_1 + Y_2 + \ldots + Y_N$, nous devons choisir avec remise n unités de population avec probabilité proportionnelle à X_i et utiliser l'estimateur sans biais

$$t_{HH} = \frac{1}{n} \sum_{r=1}^{n} \frac{y_r}{p_r},$$

où $p_r = X_r / T_x$, $T_x = X_1 + X_2 + \ldots + X_N$, et y_r $(r=1, 2, \ldots, n)$ est le résultat du r-ième prélèvement. Étant donné que $\Sigma Z_i = 0$, il est facile de montrer que

$$\mathrm{Var}(t_{HH}) = \sum_{i=1}^{N} \frac{Z_i^2}{np_i}, \tag{1}$$

où $Z_i = Y_i - p_i T_y$, $i=1, 2, \ldots, N$.

Dans le cas d'un échantillonnage sans remise, Horvitz et Thompson (1952) ont proposé l'estimateur sans biais

$$t_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i},$$

[1] S.G. Prabhu-Ajgaonkar, Department of Mathematics and Statistics, Marathwada University, Aurangabad 431004, Inde.

où π_i $(i=1, 2, \ldots, N)$ désigne la probabilité d'inclusion de la i-ième unité de population U_i dans l'échantillon. En outre, lorsque π_i est proportionnelle à X_i, nous parlons d'une méthode d'échantillonnage avec PIPT. Pour ce genre de méthode,

$$\mathrm{Var}(t_{HT}) = \sum_{i=1}^{N} \frac{Z_i^2}{np_i} + \sum_{i \neq j=1}^{N} Z_i Z_j \frac{\pi_{ij}}{n^2 p_i p_j} \tag{2}$$

où Z_i est défini en (1) et π_{ij} $(i \neq j=1, 2, \ldots, N)$ est la probabilité (conjointe) d'inclusion des i-ième et j-ième unités de population dans l'échantillon. Lorsqu'on définit une méthode d'échantillonnage avec PIPT, π_{ij} peut être exprimée de façon plus simple.

Par (1) et (2),

$$\phi = \mathrm{Var}(t_{HT}) - \mathrm{Var}(t_{HH}) = \sum_{i \neq j=1}^{N} Z_i Z_j \frac{\pi_{ij}}{n^2 p_i p_j}. \tag{3}$$

2. COMPARAISON DES MÉTHODES

Pour estimer T_y, Midzuno (1952), Sen (1952) et Sankaranarayanan (1969) ont proposé des méthodes d'échantillonnage avec PIPT fondées sur l'estimateur d'Horvitz-Thompson t_{HT}. La méthode de Midzuno et Sen est applicable si

$$p_i = \frac{X_i}{T_x} > \frac{n-1}{n(N-1)}, \; i=1, \ldots, N, \tag{4}$$

tandis que celle de Sankaranarayanan est applicable si

$$\sum_{j \in s} p_j > (n-1)/(N-1) \text{ pour tous } s \in S,$$

qui est une condition moins stricte.

Pour les deux méthodes, les probabilités conjointes d'inclusion sont définies par

$$\pi_{ij} = \frac{n(n-1)}{N-2} \left(p_i + p_j - \frac{1}{N-1}\right).$$

Ainsi, par l'équation (3),

$$\phi = \frac{n(n-1)}{n^2(N-2)} \left[\sum_{i=1}^{N} \frac{Z_i^2}{p_i}\left(2 - \frac{1}{(N-1)p_i}\right) + \frac{1}{(N-1)} \left(\sum_{i=1}^{N} \frac{Z_i}{p_i} \right)^2 \right]. \tag{5}$$

L'expression ci-dessus est non-négative si

$$P_i > \frac{1}{2(N-1)}, \; i=1, 2, \ldots, N,$$

et dans ce cas la méthode de Horvitz-Thompson est supérieure à la méthode de Hansen-Hurwitz. La condition définie en (4) a été établie pour la première fois par Rao (1963) lorsque $n=2$ et que la méthode de Midzuno-Sen est utilisée, mais il est intéressant de constater par l'équation (5) que la condition ne change pas même pour des valeurs de n supérieures à 2.

Chaudhuri (1975) et Mukhopadhyay (1975) ont, chacun de leur côté, obtenu les résultats ci-dessus pour la méthode de Midzuno-Sen.

Brewer (1963), Rao (1965) et Durbin (1967) ont proposé diverses méthodes d'échantillonnage avec PIPT pour $n=2$ avec les mêmes probabilités d'inclusion,

$$\pi_{ij} = \frac{2p_i p_j}{1+k}\left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right) \text{ où } k = \sum_{i=1}^{N} \frac{p_i}{1-2p_i}.$$

Ces méthodes ne sont pas assujetties aux conditions qui s'appliquent aux méthodes précédentes. Par (3),

$$\phi = \frac{1}{1+k}\sum_{i=1}^{N} \frac{Z_i^2}{1-2p_i} \geq 0,$$

de sorte que la méthode de Horvitz-Thompson est de nouveau supérieure à la méthode de Hansen-Hurwitz.

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance au rédacteur en chef, M.P. Singh, et à l'arbitre pour leurs nombreux commentaires utiles.

BIBLIOGRAPHIE

BREWER, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.

CHAUDHURI, A. (1975). On some properties of the sampling scheme due to Midzuno. *Bulletin of Calcutta Statistical Association*, 23, 1-19.

DURBIN, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, 16, 152-164.

HANSEN, M.H., et HURWITZ, W.N. (1943). On the theory of sampling from finite population. *Annals of Mathematical Statistics*, 14, 333-362.

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

MIDZUNO, H. (1952). On the sampling system with probability proportionate to sum of sizes. *Annals of Institute of Statistical Mathematics*, 3, 99-107.

MUKHOPADHYAY, P. (1975). PPS sampling schemes to base HTE. *Bulletin of Calcutta Statistical Association*, 23, 21-44.

RAO, J.N.K. (1963). On two systems of unequal probability sampling without replacement. *Annals of Institute of Statistical Mathematics*, 15, 67-72.

RAO, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

SANKARANARAYANAN, K. (1969). An IPPS sampling scheme using Lahiri's method of selection. *Journal of the Indian Society of Agricultural Statistics*, 21, 58-66.

SEN, A.R. (1952). Further developments of the theory and application of primary sampling units with special reference to the North Carolina agricultural population. Thèse de doctorat, North Carolina State College, Raleigh.

REMERCIEMENTS

La revue Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article entre le premier janvier 1987 et le 31 janvier 1988. Une astérisque indique que la personne a participé plus d'une fois.

D.W. Anderson, *National Institute of Health*
J. Armstrong, *Statistique Canada*
H.R. Arora, *Transports Canada*
M. Bankier, *Statistique Canada*
*K.G. Basavarajappa, *Statistique Canada*
G. Brackstone, *Statistique Canada*
D. Bellhouse, *University of Western Ontario*
L. Biggeri, *Université de Florence*
*D.A. Binder, *Statistique Canada*
R.D. Burgess, *Statistique Canada*
C.-M. Cassel, *Statistics Sweden*
N. Chinnappa, *Statistique Canada*
*G.H. Choudhry, *Statistique Canada*
D. Dodds, *Statistique Canada*
D. Drew, *Statistique Canada*
M. Eagen, *Goss, Gilroy and Associates*
J.L. Eltinge, *Iowa State University*
I.P. Fellegi, *Statistique Canada*
W.A. Fuller, *Iowa State University*
J.F. Gentleman, *Statistique Canada*
E. Gbur, *University of Arkansas*
*G.B. Gray, *Statistique Canada*
M.A. Hidiroglou, *Statistique Canada*
D. Holt, *University of Southampton*
S. Ingram, *Statistique Canada*
W.D. Kalsbeek, *University of North Carolina*
*G. Kalton, *University of Michigan*
N.J. Kirkendall, *Alexandria, Virginia*
G. Kriger, *Statistique Canada*
H. Lee, *Statistique Canada*
J.M. Lepkowski, *University of Michigan*
*S. Kumar, *Statistique Canada*
R. Lachapelle, *Statistique Canada*
E. Langlet, *Statistique Canada*
S. Michaud, *Statistique Canada*
D.G. Paton, *Statistique Canada*
M. Podehl, *Statistique Canada*
*J.N.K. Rao, *Carleton University*
P.S.R.S. Rao, *University of Rochester*
G. Sande, *Statistique Canada*
*I. Sande, *Statistique Canada*
C.E. Särndal, *Université de Montréal*
*F. Scheuren, *U.S. Internal Revenue Service*
E.A. Schillmoeller, *Nielsen Media Research*
M. Sheridan, *Statistique Canada*
A.C. Singh, *Statistique Canada*
K.P. Srinath, *Statistique Canada*
V. Tremblay, *Statplus*
A. van Baaren, *Statistique Canada*
*K.M. Wolter, *U.S. Bureau of the Census*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1987: B. Babcock (Rédaction), C. VanBastelaar (Photocomposition), G. Gaulin (Services aux auteurs) et M. Haight (Services de traduction).

On tient à remercier le personnel des Divisions des méthodes d'enquêtes sociales et des méthodes d'enquêtes-entreprises qui ont aidé à la correction et à la vérification. Finalement on désire exprimer notre reconnaissance à J. Clarke, E. Corriveau, J. Dufresne, M. Kent, C. Larabie, D. Lemire et N. Smalldridge pour leur apport à la coordination, la dactylographie et à la rédaction.

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. 10, n° 2) 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; 1, l).

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.