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

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

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

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

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 ($D^{(0)}_{shi}$) comprises dans cette strate.

Posons

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

Alors, (3.7) devient

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

Les variables ($\Sigma_i D^{(0)}_{hi}$) 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:

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

où

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

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 $\hat{V}$:

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

où

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

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

$$D^{(2)}_{shi} = \sum_a \left(Z^{(2)}_{Yshia} - R^{(2)}_{Ya} Z^{(2)}_{Pshia}\right),$$

$$Z^{(2)}_{Yshia} = \frac{P_a}{P^{(1)}_a} \frac{P_s}{P^{(0)}_s} \left(Y^{(0)}_{shia} - \frac{Y^{(0)}_{sa}}{P^{(0)}_s} P^{(0)}_{shi}\right)$$

$$= Y^{(2)}_{shia} - \frac{P^{(0)}_{shi}}{P^{(0)}_s} Y^{(2)}_{sa},$$

$$Z^{(2)}_{Pshia} = \frac{P_a}{P^{(1)}_a} \frac{P_s}{P^{(0)}_s} \left(P^{(0)}_{shia} - \frac{P^{(0)}_{sa}}{P^{(0)}_s} P^{(0)}_{shi}\right)$$

$$= P^{(2)}_{shia} - \frac{P^{(0)}_{shi}}{P^{(0)}_s} P^{(2)}_{sa},$$

et

$$R^{(2)}_{Ya} = \frac{Y^{(1)}_a}{P^{(1)}_a} = \frac{P_a}{P^{(1)}_a} \frac{Y^{(1)}_a}{P_a} = \frac{Y^{(2)}_a}{P_a}.$$

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.