

DÉTERMINATION DE TAILLE D'ÉCHANTILLON POUR LES ENQUÊTES POSTCENSITAIRES AUTOCHTONES

Éric Langlet¹

RÉSUMÉ

Les enquêtes post-censitaires tirent leurs échantillons à partir des réponses fournies au questionnaire long du recensement, administré à environ un ménage sur cinq au Canada (partie 2B) sauf dans les régions éloignées et les réserves indiennes où tous les ménages doivent le compléter (partie 2D). Ceci constitue l'échantillon de première phase. Un échantillon stratifié de deuxième phase est ensuite tiré selon les caractéristiques observées à la première phase. Afin d'estimer les tailles d'échantillon requises, une approximation du plan est utilisée pour ensuite effectuer une allocation optimale entre les strates 2B et 2D. De nombreux ajustements sont par la suite effectués sur les tailles obtenues.

MOTS CLÉS: Allocation optimale; chevauchement entre enquêtes; échantillonnage à deux phases; estimation de la variance.

ABSTRACT

The post-censal surveys draw their samples from answers to the Census long form, which is administered to approximately one in five households in Canada (2B component) except in remote areas and Indian reserves where it is administered to all households (2D component). The Census long form sample constitutes the first phase sample. A second phase stratified random sample is then selected according to variables observed in the first phase. Sample sizes are derived by approximating this sampling plan and by using an optimal allocation between the 2B and 2D strata. Several adjustments are then performed on the resulting sample sizes.

KEY WORDS : Optimal allocation; ; Survey overlap; Two-phase sampling; Variance estimation.

1. INTRODUCTION

1.1 Enquêtes postcensitaires autochtones

En 2006, Statistique Canada a mené deux enquêtes postcensitaires autochtones, soit l'Enquête sur les enfants autochtones (EEA) ainsi que l'Enquête sur les peuples autochtones (EAPA). L'EEA est une toute nouvelle enquête dont l'objectif est de donner un portrait du développement de la petite enfance chez les autochtones de 0 à 5 ans. Cette enquête porte sur des domaines spécifiques du développement et du bien-être des jeunes enfants autochtones comme la santé, l'éducation, la langue, la garde des enfants, la nutrition et les étapes du développement chez l'enfant. L'enquête couvre principalement la portion hors réserve des enfants autochtones de 0 à 5 ans, soit une population d'environ 140,000 individus avec un échantillon d'environ 18,000.

L'EAPA de 2001, quant à elle, en est à sa troisième édition après les enquêtes de 1991 et 2001. Cette enquête porte sur le mode de vie et les conditions de vie des peuples autochtones. L'enquête couvre des sujets comme les besoins en matière de santé, la langue, l'emploi, le revenu, l'éducation, le logement et la mobilité. En 2006, l'enquête couvre uniquement la population autochtone vivant hors des réserves indiennes pour les enfants de 6 à 14 ans et les adultes de 15 ans et plus. Le terme « adulte » est utilisé pour les 15 ans et plus, étant donné que ce groupe reçoit un questionnaire différent des enfants de 6 à 14 ans, appelé « questionnaire adulte ». Cette population « hors-réserve » d'environ 1,300,000 individus est couverte à partir d'un échantillon d'environ 61,000 individus. Il est à noter que nous prévoyons couvrir les réserves indiennes avant le recensement de 2011.

¹ Éric Langlet, Édifice R.H.Coats, 15R, 100, promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6, langlet@statcan.ca

1.2 Plan d'échantillonnage

La population cible de ces deux enquêtes correspond aux individus s'identifiant comme autochtone ou ceux ayant une origine autochtone (qu'ils s'identifient ou non comme autochtone). Nous définissons cette population à partir de quatre questions filtres présentes sur le formulaire long du recensement. Deux versions principales existent de ce formulaire long, soit les formulaires 2B et 2D. Le 2D est administré dans les réserves indiennes et les régions éloignées du Canada à toute la population, tandis que le 2B est administré partout ailleurs sur une base échantillonnale systématique d'un ménage sur cinq à l'intérieur de chaque unité de collecte (UC). Une UC est une petite unité géographique comprenant habituellement de 250 à 500 ménages, correspondant à la charge de travail d'un énumérateur. Nous définissons les domaines d'estimation visés par ces deux enquêtes, à partir du croisement des régions géographiques, des groupes autochtones (Indien d'Amérique du Nord, Métis et Inuit) et des groupes d'âges (0-5 ans, 6-14 ans et 15 ans et plus). À noter que la classification des individus selon ces caractéristiques peut varier entre le recensement et l'enquête.

Une fois les réponses aux questionnaires longs obtenues (1^{ière} phase du plan), nous stratifions la population cible des enquêtes selon les domaines d'estimation prévus par l'enquête et sous-stratifions selon le type de régions (régions recevant le 2B et régions recevant le 2D). Nous tirons ensuite un échantillon aléatoire simple (EAS) d'autochtones à l'intérieur de chacune de ces strates. Dans les régions 2B, le plan correspond donc à un échantillon à deux phases où les unités d'échantillonnage sont différentes aux deux phases (ménages à la 1^{ière} phase et individus à la 2^{ième} phase) et où la stratification est différente à chacune des deux phases (UC à la 1^{ière} phase et domaines d'estimation à la 2^{ième} phase). Dans les régions 2D, le plan correspond donc à un EAS à une seule phase d'autochtones dans chaque domaine d'estimation.

Nous présentons la méthode d'allocation à la section 2. Par la suite, nous décrivons à la section 3 les divers ajustements aux tailles d'échantillon initiales obtenues. La section 4 couvrira des méthodes d'estimation de la variance envisagées pour ce plan d'échantillonnage.

2. MÉTHODE D'ALLOCATION

Dans chaque domaine d'estimation, l'objectif de l'allocation est d'estimer une proportion minimale d'autochtones avec une caractéristique d'intérêt, P , selon une précision donnée par un coefficient de variation, CV . Dans la méthode utilisée, nous supposons que dans les régions 2B, la première phase (plan d'échantillonnage du 2B) nous donne l'équivalent d'un EAS d'autochtones dans chaque domaine d'estimation visé. Étant donné que dans ces régions 2B, un EAS d'autochtones est tiré dans chaque domaine d'estimation visé à la 2^{ième} phase, nous pouvons approximer le plan à deux phases par un plan EAS à une seule phase. Nous apportons une correction à l'aide d'un effet de plan, $deff$, de 1.2 pour tenir compte de cette approximation. Cette valeur choisie de 1.2 représente d'avantage une valeur intuitive qu'une valeur fondée sur une étude rigoureuse. Nous effectuons présentement une étude à ce sujet.

2.1 Allocation optimale

Ignorons la non-réponse pour l'instant. Dénotons l'estimateur d'une proportion, P , dans une population de taille N selon un plan stratifié simple par

$$\hat{P}_{st} = \sum_{h=1}^L \frac{N_h}{N} \hat{P}_h = \sum_h W_h \hat{P}_h = \sum_h W_h \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}, \quad (2.1)$$

où $y_{hi}=1$ si l'unité i de la strate h a la caractéristique d'intérêt et 0 sinon, et où n_h est la taille d'échantillon tirée dans la strate h de taille N_h .

Les formules qui vont suivre peuvent se retrouver à partir de Cochran (1977), en substituant les moyennes par des proportions. La variance corrigée par un effet de plan de cet estimateur est

$$V(\hat{P}_{st}) = deff \sum_h \frac{W_h^2 P_h Q_h}{n_h} \left(1 - \frac{n_h}{N_h}\right), \quad P_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}, \quad Q_h = (1 - P_h). \quad (2.2)$$

Dans chaque domaine d'estimation, nous utilisons une allocation optimale entre les parties 2B et 2D si cette allocation est réalisable. Il n'y a donc que deux strates par domaine d'estimation mais nous présentons quand même ici les formules pour un nombre quelconque de strates. Le problème consiste à minimiser le coût $C = C_0 + \sum_h C_h n_h$ pour une variance fixe, $V(\hat{P}_{st})$, ou de façon équivalente, pour un coefficient de variation fixe, $CV(\hat{P}_{st}) = \sqrt{V(\hat{P}_{st})}/P$. Nous obtenons la taille d'échantillon, n , à tirer dans tout le domaine selon la formule

$$n = \frac{deff \sum_h (W_h \sqrt{P_h Q_h} \sqrt{c_h}) \sum_h (W_h \sqrt{P_h Q_h} / \sqrt{c_h})}{V(\hat{P}_{st}) + \frac{1}{N} deff \sum_h W_h P_h Q_h}. \quad (2.3)$$

Si nous supposons un coût unitaire par strate, C_h , identique pour les deux strates (2B vs. 2D), la valeur maximale de n est atteinte si les proportions, P_h , sont les mêmes dans les deux strates. Ce faisant, nous nous couvrons contre le pire cas, soit le cas où la stratification n'apporte aucun gain en termes de précision. Dans ce cas, la taille, n , se réduit à

$$n = \frac{Q deff}{P \times CV^2(\hat{P}_{st}) + Q deff / N}, \quad Q = (1 - P), \quad (2.4)$$

soit la formule d'allocation pour un plan EAS corrigée par un effet de plan, $deff$. Une fois la taille, n , déterminée, nous obtenons les tailles optimales à prendre dans chaque strate selon l'expression

$$n_h = \frac{n N_h \sqrt{P_h Q_h} / \sqrt{C_h}}{\sum_h N_h \sqrt{P_h Q_h} / \sqrt{C_h}} = \frac{n N_h}{N}, \quad (2.5)$$

si les C_h et P_h sont égaux dans chaque strate, ce qui correspond à l'allocation proportionnelle.

2.2 Allocation alternative

L'allocation proportionnelle (2.5) n'est pas toujours réalisable si le nombre de formulaires longs disponibles dans la strate 2B n'est pas assez élevé pour un domaine d'estimation particulier, une fois que nous appliquons les taux de réponse attendus, r_h , aux tailles obtenues, n_h , dans chaque strate. Nous supposons ici que la probabilité de répondre est constante dans chaque strate. Si l'allocation proportionnelle n'est pas réalisable, nous devons prendre tous les 2B disponibles et prendre autant de 2D que nécessaire pour atteindre le CV désiré. Partant de la formule de variance (2.2), en tenant compte du taux de réponse, r_h , dans chaque strate et en posant $P_h = P \forall h = 1, \dots, L$, nous obtenons

$$V(\hat{P}_{st}) = \frac{PQ}{N^2} deff \sum_h \frac{N_h^2}{r_h n_h} \left(\frac{N_h - r_h n_h}{N_h} \right) = \frac{PQ}{N^2} deff \sum_h N_h \left(\frac{N_h}{r_h n_h} - 1 \right). \quad (2.6)$$

En termes de CV plutôt que de variance, nous obtenons

$$CV^2(\hat{P}_{st}) = \frac{V(\hat{P}_{st})}{P^2} = \frac{Q deff}{PN^2} \sum_{h=1}^L N_h \left(\frac{N_h}{r_h n_h} - 1 \right). \quad (2.7)$$

Dans notre cas, nous avons deux strates, la strate $h=1$ étant la strate 2D et la strate $h=2$ étant la strate 2B. L'expression devient

$$CV^2(\hat{P}_{st}) = \frac{Q deff}{PN^2} \left\{ N_1 \left(\frac{N_1}{r_1 n_1} - 1 \right) + N_2 \left(\frac{N_2}{r_2 n_2} - 1 \right) \right\}. \quad (2.8)$$

Ici, n_2 est connu puisque nous prenons tous les 2B mais nous obtenons $r_2 n_2$ répondants. En exprimant n_1 en fonction de n_2 ,

$$\frac{N_1}{r_1 n_1} = \frac{\frac{PN^2 CV^2 (\hat{p}_{st})}{Q deff} - N_2 \left(\frac{N_2}{r_2 n_2} - 1 \right)}{N_1} + 1,$$

et

$$n_1^{-1} = r_1 \left(\frac{PN^2 CV^2 (\hat{p}_{st})}{N_1^2 Q deff} - \frac{N_2}{N_1^2} \left(\frac{N_2}{r_2 n_2} - 1 \right) + \frac{1}{N_1} \right). \quad (2.9)$$

Si $n_1 > N_1$, le CV désiré ne peut être atteint et celui-ci doit être augmenté.

2.3 Sélection de l'échantillon en deux vagues

Étant donné les retards importants dans la collecte du recensement, nous avons dû tirer l'échantillon de chaque enquête en deux vagues afin de respecter la date de début de collecte de chaque enquête. L'idée était de choisir un premier échantillon dans des régions où la base du recensement était presque complète et de commencer la collecte dans ces régions. Nous avons par la suite tiré un 2^{ième} échantillon une fois la base du recensement complétée. Ce 2^{ième} échantillon couvrait tous les formulaires des régions non couvertes à la vague 1 de même que les nouveaux formulaires des régions couvertes à la vague 1, ajoutés à la base depuis la sélection du 1^{ier} échantillon.

À la première vague, nous avons dû estimer les totaux de population à partir d'une base incomplète. Nous avons pondéré les effectifs de population cible dans chaque UC par l'inverse du pourcentage estimé de formulaires longs présents sur la base dans cette UC au moment de la sélection de l'échantillon. Une fois la taille de population cible estimée dans chaque domaine d'estimation, nous avons déterminé une fraction de sondage selon la méthode d'allocation décrite ci-haut. Pour la vague 1, nous avons appliqué cette fraction de sondage aux unités présentes sur la base à ce moment-là. Nous avons par la suite utilisé les mêmes fraction de sondage pour les unités s'étant ajoutées à la vague 2 tombant dans les régions couvertes à la vague 1. À la vague 2, les régions non couvertes à la vague 1 ne posaient aucun problème spécifique parce que toutes les données étaient présentes sur la base à ce moment-là.

3. AJUSTEMENTS DE TAILLES D'ÉCHANTILLON

Dans les formules présentées à la section 2, les taux de réponse, r_h , tiennent non seulement compte de la non-réponse mais également d'autres formes de perte d'échantillon. Parmi ces autres formes de perte, mentionnons ce que nous appelons les *faux positifs*. Les *faux positifs* sont des individus déclarés comme étant autochtones au recensement qui n'ont pas par la suite été déclarés comme étant autochtones à l'enquête. L'unité d'échantillonnage de l'enquête étant l'individu et non le ménage, nous avons parfois échantillonné plusieurs individus dans un même ménage. Nous avons apporté une limite de ce nombre pour chaque enquête et nous avons retiré certaines unités après échantillonnage. Nous avons également éliminé a posteriori une partie du chevauchement entre ces deux enquêtes et l'Enquête longitudinale sur les enfants et les jeunes, étant donné le contenu corrélé entre ces enquêtes. Une autre composante importante de réduction de taille d'échantillon est celle due au chevauchement avec les autres enquêtes postcensitaires.

En 2006, Statistique Canada a mené cinq enquêtes postcensitaires à partir de cinq échantillons indépendants, étant donné les problèmes associés à la création d'échantillons coordonnés. Nous avons apporté certaines contraintes pour réduire le chevauchement une fois les échantillons tirés. Nous avons limité le nombre d'enquêtes pour un même ménage à deux et nous avons limité le nombre d'entrevues pour un même ménage à quatre. Si le ménage était choisi pour une seule enquête, nous avons limité le nombre d'entrevues à trois. Dans le cas d'un ménage tiré pour plus de deux enquêtes, nous avons tiré deux enquêtes de façon équiprobable entre les enquêtes. Une fois le nombre d'enquêtes réduit à au plus deux par ménage, nous avons par la suite réduit le nombre d'entrevues pour chaque enquête si ceci était encore nécessaire. Nous avons utilisé un tirage proportionnel à la taille pour réduire le nombre d'entrevues d'une même enquête. La mesure de taille utilisée était la fraction de sondage de 2^{ième} phase correspondant à la strate de chaque individu. L'idée était de

donner une probabilité d'inclusion plus grande à un individu provenant d'une strate nécessitant une grande fraction de sondage qu'à un individu tombant dans une strate nécessitant une plus faible fraction de sondage. L'augmentation du CV est en effet plus grande pour le retrait d'une unité dans la première catégorie que le retrait d'une unité dans la seconde.

Toutes ces formes de pertes d'échantillon ont été estimées soit selon l'EAPA de 2001 soit par simulation. Nous avons, par exemple, estimé les taux de faux positifs dans chacune des strates à partir de l'EAPA de 2001. Pour ce qui est du chevauchement, nous l'avons estimé en générant cinq échantillons préliminaires pour les cinq enquêtes. Nous avons par la suite augmenté les tailles d'échantillon en conséquence.

4. MÉTHODES ENVISAGÉES D'ESTIMATION DE LA VARIANCE

Nous considérons plusieurs méthodes pour le calcul de la variance. Une possibilité est de se servir de la linéarisation de Taylor qui est la méthode utilisée dans le Système Généralisé d'Estimation (SGE) de Statistique Canada (Statistics Canada, 2005), par exemple. Pour notre problème, ce système présente plusieurs limites. En particulier, pour l'échantillonnage à deux phases, les unités d'échantillonnage à chacune des deux phases doivent être les mêmes. En réalité, nous échantillonnons des ménages à la 1^{ière} phase et des individus à la 2^{ième} phase. Il faudrait donc émettre des hypothèses simplificatrices sur le plan d'échantillonnage de 1^{ière} phase. Même avec ces hypothèses simplificatrices, il semble que la capacité de mémoire du système soit pour l'instant dépassée par la taille de nos échantillons de 1^{ière} et 2^{ième} phase dans les grandes provinces.

Une alternative est d'utiliser une méthode de rééchantillonnage, tel que le bootstrap qui est très populaire auprès des utilisateurs. En faisant la même hypothèse simplificatrice que dans la méthode d'allocation, à savoir que dans les régions 2B, nous pouvons approximer la plan stratifié à deux phases par un EAS stratifié à une seule phase, nous pourrions utiliser une méthode bootstrap standard pour des plans à une seule phase, tel que le bootstrap de Rao-WU par exemple (Rao et Wu, 1988). Dans ce cas, les individus échantillonnés à la 2^{ième} phase seraient rééchantillonnés avec remise dans chaque strate de 2^{ième} phase.

Une autre possibilité serait d'utiliser un bootstrap pour plans à deux phases avec rééchantillonnage des unités de 1^{ière} phase. Il s'agirait ici de tirer un échantillon de ménages avec remise dans chaque strate de 1^{ière} phase (UC). À noter que cette méthode supposerait que le plan de 1^{ière} phase est un échantillon stratifié de ménages par UC et non un échantillon stratifié systématique par UC. Cette hypothèse aurait tendance à sous-estimer la variance s'il existe un lien entre l'ordre des ménages dans l'UC et la caractéristique à l'étude. En effet, il est possible que les autochtones soient regroupés dans un secteur spécifique de l'UC. Kott et Stukel (1997) proposent une méthode jackknife pour un plan d'échantillonnage à deux phases appropriée pour l'estimateur de développement repondéré. Cette méthode pourrait vraisemblablement s'adapter au bootstrap. Kim, Navaro et Fuller (2006) proposent une généralisation de la méthode pour différentes méthodes d'estimation de variance par rééchantillonnage, dont le bootstrap. Leur méthode est appropriée non seulement pour l'estimateur de développement repondéré, mais aussi pour l'estimateur de développement double. Cependant, toutes ces méthodes proposées émettent l'hypothèse que la fraction de sondage de 1^{ière} phase est négligeable. Dans le cas contraire, une composante importante de la variance pour plans à deux phases manque, entraînant ainsi une sous-estimation. L'ampleur de cette sous-estimation sera d'autant plus forte que la fraction de sondage de 1^{ière} phase est grande et que la fraction de sondage de 2^{ième} phase est faible. Dans le cas qui nous préoccupe, la fraction de sondage de 1^{ière} phase est loin d'être négligeable puisque celle-ci est d'environ 20% pour la majorité des régions.

Une autre approche bootstrap consiste à créer une population totale artificielle à partir de l'échantillon plutôt que de retirer directement des unités de l'échantillon. C'est l'approche suggérée par Gross (1980), approche appelée « bootstrap sans remise ». Nous devons d'abord recréer la population cible de 1^{ière} phase à partir des unités tirées à la 2^{ième} phase. Pour une strate donnée, si nous avons échantillonné 30 unités sur 60, chaque unité de 2^{ième} phase devrait être répétée deux fois. Cette population cible artificielle de 1^{ière} phase est ensuite combinée à l'échantillon réel de 1^{ière} phase de ménages ne contenant pas d'unités de notre population cible (ménages non autochtones). On doit recréer par la suite, toute la population de ménages autochtones et non autochtones, au moins pour les UC contenant des ménages autochtones dans notre population cible artificielle de 1^{ière} phase. Nous devons nous servir des poids de 1^{ière} phase pour ce faire. Nous pourrions aussi recréer les ménages non autochtones des UC ne contenant pas de ménages autochtones mais ces ménages n'interviendront pas dans les calculs de toute manière.

Dans le cas où l'inverse de la fraction de sondage n'est pas entier, différentes méthodes sont possibles pour créer cette population artificielle, dont certaines sont mentionnées dans Sitter (1992). Il propose lui-même une extension du « bootstrap sans remise » pour certains plans de sondage dont l'échantillonnage stratifié et l'échantillonnage à deux degrés. À notre connaissance, aucune méthode n'a cependant été proposée pour l'échantillonnage à deux phases.

Une fois la population totale recréée de cette façon, nous pouvons répéter le processus d'échantillonnage à deux phases un grand nombre de fois pour produire des échantillons bootstrap. Nous tirons d'abord un échantillon de ménages à l'intérieur de chaque UC. Il est probable que nous devions simplifier le plan de 1^{ière} phase et tirer un échantillon aléatoire simple plutôt qu'un échantillon aléatoire systématique de ménages à cette étape. Par la suite, pour chaque échantillon bootstrap de 1^{ière} phase obtenu, nous stratifions la population cible observée et nous tirons un échantillon de 2^{ième} phase de la même façon que nous avons tiré l'échantillon maître. Il n'est cependant pas certain qu'une telle méthode puisse fonctionner pour l'échantillonnage à deux phases.

Nous nous proposons donc d'étudier ces méthodes de même que des alternatives possibles et de comparer ces méthodes entre elles.

RÉFÉRENCES

Cochran, W.G. (1977). *Sampling Techniques* (3rd. ed.). New York: John Wiley.

Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Kim, J.K., Navarro, A., et Fuller, W.A. (2006). Replication Variance Estimation for Two-Phase Stratified Sampling. *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 312-320.

Kott, P.S., et Stukel, D.M. (1997). La méthode du jackknife convient-elle à un échantillon à deux phases? *Techniques d'enquête*, vol. 23, no. 2, pp. 89-98.

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

Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *La revue canadienne de statistique*, vol. 20, no. 2, pp. 35-154.

Statistics Canada (2005). *GES v4.3 User Guide*. Statistics Canada document, June 2005, pp. 344-349.