Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Une comparaison des estimateurs de la variance pour la poststratification en fonction de totaux de contrôle estimés

par Jill A. Dever et Richard Valliant

Juin 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2010
Vol. 36, N° 1, pp. 49-61
Statistique Canada, N° 12-001-X au catalogue

Une comparaison des estimateurs de la variance pour la poststratification en fonction de totaux de contrôle estimés

Jill A. Dever et Richard Valliant [1]

Résumé

Les méthodes de calage, telles que la poststratification, s'appuient sur de l'information auxiliaire pour accroître l'efficacité des estimations par sondage. L'hypothèse est que les totaux de contrôle, en fonction desquels les poids de sondage sont poststratifiés (ou calés), sont les valeurs de population. Toutefois, les totaux de contrôle sont souvent estimés d'après d'autres enquêtes. De nombreux chercheurs appliquent les estimateurs classiques d'estimation de la variance par poststratification à des situations où les totaux de contrôle sont estimés, supposant donc que toute variance d'échantillonnage supplémentaire associée à ces totaux estimés est négligeable. Le but de l'étude présentée ici est d'évaluer des estimateurs de la variance pour des plans de sondage stratifiés à plusieurs degrés, sous une poststratification en fonction de totaux de contrôle estimés (CE) en utilisant des valeurs de contrôle sans biais par rapport au plan. Nous comparons les propriétés théoriques et empiriques des estimateurs de variance par linéarisation et par le jackknife pour un estimateur poststratifié d'un total de population. Nous donnons des exemples des effets qu'ont sur les variances divers niveaux de précision des totaux de contrôle estimés. Notre étude donne à penser que i) les estimateurs de variance classiques peuvent sous-estimer considérablement la variance théorique et que ii) deux estimateurs de variance par poststratification CE peuvent atténuer le biais négatif.

Mots clés : Poststratification en fonction de totaux de contrôle estimés ; biais de couverture de la base de sondage ; totaux de contrôle estimés d'après un sondage.

1. Introduction

Les estimateurs poststratifiés, et d'autres estimateurs par calage, sont utilisés dans de nombreux types de sondage pour réduire les variances ou pour corriger certains défauts de la base de sondage. À titre d'exemple particulier, mentionnons les grandes enquêtes du gouvernement des États-Unis, telles que la Consumer Expenditure Survey (voir, par exemple, Jayasuriya et Valliant 1996), les enquêtes auprès de populations spéciales, telles que la Survey of Health Related Behaviors among Military Personnel du U.S. Department of Defense (Bray, Hourani, Rae, Dever, Brown, Vincus, Pemberton, Marsden, Faulkner et Vandermaas-Peeler 2003), et une foule d'enquêtes réalisées en dehors des États-Unis, dont l'Enquête sur le commerce de détail du Canada (voir, par exemple, Hidiroglou et Patak 2006), l'Enquête sur la population active de la Suède (Mirza et Hörngren 2002), et la Household Panel Survey du Royaume-Uni (Taylor, Brice, Buck et Prentice-Lane 2007).

Les estimateurs par calage, tels ceux produits par poststratification, sont utilisés pour réduire au minimum les erreurs associées aux bases de sondage incomplètes (par exemple, le sous-dénombrement) ainsi qu'à l'échantillonnage et à la non-réponse (voir, par exemple, Särndal, Swensson et Wretman 1992 ; Lessler et Kalsbeek 1992 ; Kott 2006). Ainsi, les estimations produites d'après le Behavioral Risk Factor Surveillance System (BRFSS), une enquête téléphonique par composition aléatoire (CA) de portée nationale réalisée par les Centers for Disease Control and Prevention (CDC) des États-Unis, sont poststratifiées en fonction des nombres de ménages équipés et non équipés d'un service téléphonique classique à fil (Centers for Disease Control and Prevention 2006). La réduction des erreurs est liée à l'association des totaux de contrôle de population avec le sous-dénombrement de la base de sondage, les profils de non-réponse non-ignorables et la variable d'intérêt (Kim, Li et Valliant 2007).

S'il n'existe pas de totaux de contrôle de population pertinents, de nombreux chercheurs utilisent des totaux de contrôle estimés d'après un sondage et appliquent les formules classiques de calcul de la variance comme si les totaux de contrôle étaient connus sans erreur. Par exemple, Nadimpalli, Judkins et Chu (2004) ont corrigé les poids pour la *National Survey of Parents and Youth de 2003* en fonction du nombre de ménages américains comptant des enfants de 9 à 18 ans estimé d'après la *Current Population Survey* (CPS) en se servant d'un algorithme de ratissage (*raking ratio*) (www.census. gov/cps). Des estimations de la façon dont les personnes vivant aux États-Unis emploient leur temps peuvent être calculées d'après l'*American Time Use Survey* en utilisant des poids qui ont été poststratifiés en fonction d'estimations projetées d'après le recensement décennal des États-Unis (Killion 2006). Plus récemment, aux Pew Research Centers, des chercheurs ont calé des poids pour un ensemble de sondages pré-électoraux réalisés durant la campagne présidentielle de 2008 aux États-Unis

1. Jill A. Dever, RTI International. Courriel : jdever@rti.org ; Richard Valliant, Survey Research Center, University of Michigan and Joint Program in Survey Methodology, University of Maryland. Courriel : rvalliant@survey.umd.edu.

sur des estimations démographiques provenant de la CPS de mars 2007, ainsi que sur des estimations des habitudes d'utilisation du téléphone établies d'après la *National Health Interview Survey* réalisée de juillet à décembre 2007 (Keeter, Dimock et Christian 2008).

Le but de notre étude est d'élaborer et d'évaluer des estimateurs de variance pour des estimations ponctuelles calculées en se servant de poids qui contiennent une correction par poststratification en fonction d'un ensemble de totaux de contrôle estimés par sondage. Nous donnons à la méthodologie qui tient compte correctement des totaux de contrôle estimés le nom de *poststratification en fonction de totaux de contrôle estimés* (*CE*) Dans le présent article, nous nous intéressons tout spécialement à l'estimateur poststratifié en fonction de totaux de contrôle estimés (PSCE) d'un total de population pour des données recueillies selon un plan de sondage stratifié à plusieurs degrés, où les unités d'échantillonnage de premier degré sont sélectionnées *avec remise*. Dans la suite de la présente section, nous passons brièvement en revue le calage des poids et la poststratification. La section 2 contient une définition explicite de l'estimateur PSCE étudié et la section 3 donne une évaluation de ses propriétés de biais. Au moyen d'une évaluation théorique (section 4) et d'une étude par simulation, nous comparons les estimateurs de variance élaborés pour l'estimateur PSCE à un estimateur de variance choisi sous l'hypothèse naïve des « totaux de contrôle de population ». Notre étude porte à la fois sur les estimateurs de variance par linéarisation et par répliques. Nous donnons des exemples des effets qu'ont sur les estimations de la variance divers niveaux de précision des totaux de contrôle estimés. À la section 5, nous décrivons en détail les spécifications de l'étude par simulation et à la section 6, nous résumons les résultats. L'article se termine par un bref résumé et un aperçu des futurs travaux de recherche dans le domaine

Les *estimateurs par calage* (Deville et Särndal 1992), tels qu'un estimateur poststratifié d'un total de population, empruntent de l'information auxiliaire pour accroître l'efficacité des estimations par sondage comparativement aux méthodes de pondération plus simples. Si les variables auxiliaires sont reliées (linéairement) à l'ensemble de variables étudiées clés, les estimateurs par calage peuvent être très efficaces.

La forme générale d'un estimateur par calage *classique* ou *sur totaux de contrôle fixes* se décrit le mieux comme un estimateur à facteur d'extension ou « à pondération linéaire » tel qu'il est décrit dans Estevao et Särndal (2000). Désignons par s l'ensemble d'éléments d'échantillon provenant d'un échantillon probabiliste et par $d_k = 1/\pi_k$ le poids de sondage de l'élément k tel que $\pi_k = \Pr(k \in s)$. Un total de population estimé d'une variable y est donné par $\hat{t}_y = \sum_{k\in s} w_k y_k$, où le poids de calage $(w_k = a_k d_k)$ appliqué au k^{e} élément est défini comme une fonction du poids de sondage, d_k, et d'un facteur de correction par calage, a_k, également appelé poids g (Särndal et coll. 1992). Les poids de calage sont calculés en minimisant une fonction spécifiée qui mesure la distance entre les poids de sondage et les poids de calage sous un ensemble de contraintes définies comme étant :

$$\mathbf{t}_{Ux} = \hat{\mathbf{t}}_{Ax} \qquad (1)$$

où $\mathbf{t}_{Ux} = \sum_{k\in U} \mathbf{x}_k$, le vecteur de totaux de contrôle (dénombrements) de population correspondant aux $G(G \geq 1)$ variables auxiliaires, $\hat{\mathbf{t}}_x = \sum_{k\in s} w_k \mathbf{x}_k$, les totaux de contrôle de population estimés correspondant aux composantes de $\mathbf{t}_{Ux}$, et $\mathbf{x}_k$ est un vecteur de longueur G contenant les valeurs des variables auxiliaires ou d'étalonnage pour l'élément k. Notons que $\mathbf{x}_k$ peut contenir des valeurs un et des valeurs zéro pour indiquer la présence ou l'absence d'une caractéristique donnée (par exemple, âgé de 18 à 25 ans), ou des valeurs plus grandes (par exemple nombre d'enfants). La fonction de distance des moindres carrés généralisés (ou du khi-deux) $\sum_{k\in s_A} (w_k - d_k)^2/c_k d_k$ qui est minimisée sous les contraintes données par (1) est un exemple d'un tel système de calage. Ce système produit une solution analytique appelée estimateur par la régression généralisée (GREG) pour $c_k = 1$ (Deville et Särndal 1992). L'estimateur poststratifié est un cas particulier de l'estimateur GREG.

Les méthodes d'estimation de la variance pour l'estimateur poststratifié, et de façon plus générale pour l'estimateur GREG, ont été étudiées abondamment. Binder (1995) démontre les méthodes utilisées pour calculer un estimateur de variance par *linéarisation de Taylor* pour l'estimateur GREG. D'autres références pour l'estimateur de variance par linéarisation sous poststratification (et sous calage de manière plus générale) comprennent Deville, Särndal et Sautory (1993), Demnati et Rao (2004), ainsi que Hidiroglou et Patak (2006). Särndal, Swensson et Wretman (1989) ont élaboré une estimation de la variance approximative par linéarisation de l'estimateur GREG d'un total de population sous forme d'une fonction des résidus de population issus d'un modèle spécifié et des poids de sondage (d_k). Valliant (1993), ainsi que Yung et Rao (1996) ont modifié l'estimateur de variance fondé sur les résidus en multipliant les résidus d'échantillon par les poids de calage $w_k (= a_k d_k)$. Ils ont démontré que cet estimateur révisé, créé en linéarisant l'estimateur jackknife connexe, réduisait le biais associé à la formule originale. Cet estimateur de variance est également discuté dans Särndal et coll. (1992), Stukel, Hidiroglou et Särndal (1996), ainsi que dans le chapitre 11 de Särndal et Lundström (2005). Les propriétés des estimateurs de variance par répliques (c'est-à-dire jackknife et répliques répétées équilibrées) ont été examinées,

par exemple, par Valliant (1993), Rust et Rao (1996), Canty et Davison (1999), Théberge (1999), Rao et Shao (1999), Yung et Rao (1996 ; 2000), et Kott (2006).

Les auteurs des articles susmentionnés émettent l'hypothèse que les totaux de contrôle, en fonction desquels sont ajustées les estimations d'échantillon auxiliaires, sont soit les valeurs réelles de population connues sans erreur, soit des valeurs tirées d'une enquête indépendante, très précise, beaucoup plus importante que l'enquête nécessitant le calage. Dans certains cas, cependant, ces totaux de contrôle sont estimés d'après des enquêtes dont les variances d'échantillonnage ne sont pas négligeables. Par exemple, on a tenté de caler des enquêtes par panel en ligne sur des valeurs tirées d'enquêtes de référence de haute qualité, distinctes, dont la portée n'est pas beaucoup plus grande que celle des enquêtes par panel proprement dites (par exemple, Krotki 2007 ; Terhanian, Bremer, Smith et Thomas 2000).

De nombreux chercheurs appliquent des formules élaborées pour la poststratification classique, même si les totaux de contrôle ont été estimés. L'hypothèse tacite est que toute erreur supplémentaire (variance et biais) associée à ces valeurs de contrôle est négligeable et peut être ignorée. À l'heure actuelle, la validité de cette hypothèse ne peut pas être vérifiée et elle ne pourra l'être que lorsque l'on aura brossé un tableau complet de la poststratification en fonction totaux de contrôle estimés.

2. L'estimateur poststratifié en fonction de totaux de contrôle estimés

Afin de faciliter notre discussion de l'estimateur poststratifié en fonction de totaux de contrôle estimés, nous donnons à l'enquête nécessitant la poststratification le nom d'*enquête analytique* et à la source des totaux de contrôle le nom d'*enquête repère*. En pratique, les totaux de contrôle peuvent être tirés de plus d'une enquête repère. Cependant, pour le développement théorique, nous supposerons qu'une seule de ces enquêtes est utilisée, afin qu'il soit possible d'estimer les variances et les covariances des totaux de contrôle.

Soit U la population finie cible contenant N éléments et $t_y = \sum_{k \in U} y_k$, le total de population d'intérêt d'une variable y. Soit s_A un échantillon aléatoire de taille n_A tiré de la base de sondage U_A pour l'enquête analytique. Un échantillon aléatoire s_B de taille n_B est sélectionné pour l'enquête repère dans la base de sondage correspondante U_B. Nous permettons qu'il soit possible que chacune des bases de sondage, U_A et U_B, ne couvre pas entièrement la population cible U. Cependant, la couverture est traitée comme un événement aléatoire, de sorte que tous les éléments compris dans la population cible ont une probabilité positive d'être couverts par la base de sondage analytique ou par la base de sondage repère.

Dans tout l'exposé, nous adoptons la convention qu'un indice inférieur « A » signifie une association avec l'enquête analytique, telle qu'un paramètre du plan de sondage ou une estimation. Un indice inférieur « B » désigne les quantités associées à l'enquête repère. Ces indices inférieurs sont absents des paramètres associés à la population étudiée, c'est-à-dire t_y.

Sous le plan de sondage stratifié à plusieurs degrés supposé pour l'enquête analytique, $m_{Ah} (m_{Ah} \geq 2)$ unités primaires d'échantillonnage (UPE), désignées par l'indice i, sont sélectionnées *avec remise* parmi un total de M_{Ah} UPE dans la h^{e} strate du plan de sondage ($h = 1, \ldots, H$ avec $H \geq 2$). Nous supposons que les n_{Ahi} éléments, portant chacun l'indice k, sont tirés de N_{Ahi} dans l'UPE hi de façon qu'il soit possible de produire une estimation sans biais du total de l'UPE. Le poids de sondage, d_k, est calculé comme l'inverse de la probabilité d'inclusion inconditionnelle pour $k \in s_{Ahi}$, l'ensemble d'éléments de l'enquête analytique dans la hi^{e} UPE. Donc, n_A, la taille de l'échantillon de l'enquête analytique, est calculée comme $n_A = \sum_{h=1}^{H} \sum_{i=1}^{m_{Ah}} n_{Ahi}$. Dans le cas de l'enquête repère, les éléments sont tirés aléatoirement de la base de sondage correspondante ; aucune spécification explicite n'est faite pour la méthode d'échantillonnage aléatoire.

La poststratification peut être utilisée pour corriger les erreurs d'échantillonnage et de couverture. Par conséquent, nous permettons l'existence d'un sous-dénombrement dans les bases de sondage de l'enquête analytique ainsi que de l'enquête repère. En outre, nous ne tenons pas compte des effets de la non-réponse.

Supposons que la population U peut être divisée en $g = 1, ..., G$ poststrates mutuellement exclusives et exhaustives. Quand le nombre d'éléments dans la population, N_g, est connu pour chaque poststrate, l'estimateur poststratifié classique d'un total pour y est défini par l'expression

$$\hat{t}_{yPS} = \sum_{g=1}^{G} N_g \frac{\hat{t}_{Ayg}}{\hat{N}_{Ag}}, \qquad (2)$$

où y_k est la valeur de la variable d'analyse y pour l'élément k ; $\hat{t}_{Ayg} = \sum_{k \in s_A} \delta_{gk} d_k y_k$, le total de y dans la poststrate g estimé d'après les données de l'enquête analytique ; $\hat{N}_{Ag} = \sum_{k \in s_A} \delta_{gk} d_k$, le total estimé d'après l'enquête analytique dans la poststrate g, et $\delta_{gk} = 1$ si l'élément appartient à la g^{e} poststrate et est égal à zéro autrement. Notons que $\hat{t}_{Ayg}$ peut également s'exprimer sous la forme $\hat{t}_{Ayg} = \sum_{k \in s_{Ag}} d_k y_k$, où s_{Ag} désigne l'ensemble d'éléments de l'enquête analytique dans la poststrate g. Dans cette dernière expression, nous utilisons la notation « chapeau » pour faire la distinction entre un estimateur de population (par exemple $\hat{N}_{Ag}$) et le paramètre de

population connu (par exemple N_g). Si l'on estime le nombre d'éléments dans la poststrate g en posant que $y_k = 1$ dans la formule pour $\hat{t}_{Ayg}$, $\hat{t}_{yPS}$ égale N_g. En ce sens, $\hat{t}_{yPS}$ est poststratifié en fonction des chiffres de population $N_1, ..., N_G$.

Cependant, dans certaines situations, les chiffres de population ne sont pas disponibles et doivent être estimés d'après une enquête repère. Exprimons l'estimateur PSCE d'un total de population d'une variable y sous la forme

$$\hat{t}_{yP} = \sum_{g=1}^{G} \hat{N}_{Bg} \frac{\hat{t}_{Ayg}}{\hat{N}_{Ag}}. \qquad (3)$$

Le nombre d'éléments de la population dans la g^{e} poststrate $(g = 1, \ldots, G)$ estimé d'après l'enquête repère est désigné par $\hat{N}_{Bg} = \sum_{l \in s_{Bg}} w_l$, où s_{Bg} est l'ensemble d'éléments de l'échantillon dans la poststrate g provenant de l'enquête repère et w_l est le poids associé au l^{e} élément. Les facteurs de correction par calage appliqués aux poids de sondage de l'enquête analytique pour $\hat{t}_{yP}$ sont calculés selon l'expression $a_k = \hat{N}_{Bg}/\hat{N}_{Ag}$ pour $k \in s_{Ag}$.

Si nous relions les estimateurs poststratifiés au système de calage décrit à la section précédente, $\hat{\mathbf{t}}_{Ax}$ est un vecteur de longueur G de chiffres de population estimés pour chaque poststrate, tel que $\hat{\mathbf{t}}_{Ax} = (\hat{t}_{Ax1}, \ldots, \hat{t}_{AxG})'$, où $\hat{t}_{Axg} \equiv \hat{N}_{Ag} = \sum_{k \in s_A} d_k \delta_{gk}$ et $x_k \equiv \delta_{gk} = 1$ si l'élément k appartient à la g^{e} poststrate et 0 autrement. Le vecteur $\mathbf{t}_{Ux}$ correspond soit à $\mathbf{N} = (N_1, \ldots, N_G)'$ pour l'estimateur $\hat{t}_{yPS}$ donné par (2), soit à $\hat{\mathbf{N}}_B = (\hat{N}_{B1}, \ldots, \hat{N}_{BG})'$, un vecteur de dimension $G \times 1$ d'estimations de contrôle d'après l'enquête repère pour l'estimateur $\hat{t}_{yP}$ donné par (3).

L'estimateur $\hat{t}_{yP}$ peut être exprimé en notation matricielle sous la forme $\hat{t}_{yP} = \hat{\mathbf{N}}'_B \hat{\bar{\mathbf{Y}}}_A$, où $\hat{\bar{\mathbf{Y}}}_A = (\hat{\mathbf{N}}_A)^{-1} \hat{\mathbf{t}}_{Ay}$, un vecteur de dimension $G \times 1$ d'estimations d'après l'enquête analytique de la forme $\hat{\bar{\mathbf{Y}}}_A = [\hat{t}_{A1}/\hat{N}_{A1}, \ldots, \hat{t}_{AG}/\hat{N}_{AG}]'$; $\hat{\mathbf{N}}_A = \text{diag}(\hat{N}_{A1}, \ldots, \hat{N}_{AG})$, une matrice diagonale de totaux de poststrate estimés d'après l'enquête analytique, et $\hat{\mathbf{t}}_{Ay} = [\hat{t}_{A1}, \ldots, \hat{t}_{AG}]'$ est un vecteur de dimension $G \times 1$ de totaux de poststrate pour la variable de résultat estimés d'après l'enquête analytique. Les autres variables associées à la notation matricielle ont été définies plus haut.

Une correction par poststratification efficace peut réduire le biais dans les estimations ponctuelles résultantes, et réduit ou accroît à peine la variance comparativement à la pondération non corrigée. Cet effet est bien connu dans le cas de la poststratification classique ; aux sections suivantes, nous présentons l'évaluation comparative dans la situation où les valeurs de contrôle sont estimées.

3. Biais dans la PSCE pour un total de population

La poststratification classique a la réputation de réduire le biais associé à une base de sondage incomplète. Cette réduction est la plus fructueuse quand les poststrates sont formées de telle manière que soit presque nulle la corrélation intra-poststrate de y_k avec la probabilité que le k^{e} élément soit inclus dans la base de sondage (Kim, Li et Valliant 2007).

Pour évaluer le biais (inconditionnel) par rapport au plan de sondage de $\hat{t}_{yP}$, nous devons tenir compte de la propriété aléatoire de quatre composantes, à savoir les plans de sondage de l'enquête analytique et de l'enquête repère et les propensions des bases de sondage correspondantes à couvrir la population. Suivant les travaux de Kim, Li et Valliant (2007, équation 2), le biais approximatif par rapport au plan de $\hat{t}_{yP}$ en tant qu'estimateur du total de population $t_y = \sum_{k \in U} y_k$ se calcule comme il suit

$$\begin{aligned} \text{Biais}(\hat{t}_{yP}) &= E(\hat{t}_{yP}) - t_y \\ &\cong \sum_{g=1}^{G} \left[t_{yg} \left\{ \frac{N_{Bg}}{N_g} - 1 \right\} + N_{Bg} \text{Cov}(y_g, \phi_{Ag}) \bar{\phi}_{Ag}^{-1} \right] \end{aligned} \qquad (4)$$

où N_g est la taille de population pour l'ensemble d'éléments U_g dans la poststrate g ; $N_{Bg} = E(\hat{N}_{Bg})$, la valeur prévue des estimations pour la poststrate sous le plan de l'enquête repère ; $\text{Cov}(y_g, \phi_{Ag}) = N_g^{-1} \sum_{k \in U_g} (y_k - \bar{y}_g)(\phi_{Ak} - \bar{\phi}_{Ag})$, la covariance de population entre la variable de résultat (y_k) et les propensions à la couverture (ϕ_{Ak}) dans la strate g ; $\bar{y}_g = t_{yg}/N_g$, la g^{e} moyenne de poststrate de y ; $t_{yg} = \sum_{k \in U_g} y_k$, le total de population de y dans la poststrate g ; et $\bar{\phi}_{Ag} = N_{Ag}/N_g$, la propension moyenne à la couverture dans la poststrate sous le plan de l'enquête analytique avec $N_{Ag} = E(\hat{N}_{Ag})$. Notons que le total de population peut également être exprimé sous la forme $t_y = \sum_g t_{yg}$.

Les composantes du biais sont nulles uniquement sous certaines conditions. *i*) Si $N_{Bg} = N_g$ pour tout g (c'est-à-dire aucune erreur de couverture dans la base de sondage repère), le biais dépend uniquement de l'association entre la variable et les propensions à la couverture, $\text{Cov}(y_g, \phi_{Ag})$. La valeur de $\text{Biais}(\hat{t}_{yP})$ se réduit alors à la formule donnée dans Kim, Li et Valliant (2007, équation 2) pour l'estimateur poststratifié classique, $\hat{t}_{yPS}$. *ii*) Si les probabilités de couverture sont constantes dans chaque poststrate (c'est-à-dire $\phi_{Ak} = \bar{\phi}_{Ag}$, $k \in U_g$ pour tout g), la deuxième composante du biais est nulle. Uniquement si *les deux* conditions sont satisfaites pouvons-nous dire que $\hat{t}_{yP}$ est approximativement sans biais. D'aucuns soutiendront peut-être que l'on pourrait former une combinaison « parfaite » de poststrates telle que les composantes positives et négatives s'annulent. Cependant, nous pensons que la probabilité qu'une telle combinaison existe est tellement faible qu'elle est virtuellement impossible.

Ayant examiné le biais, nous présentons une évaluation de la variance de $\hat{t}_{yP}$. Pour certains estimateurs, la

contribution du biais (élevé au carré) à l'erreur quadratique moyenne (EQM) totale est faible relativement à la variance.

4. Estimation de la variance pour la PSCE

Des estimateurs de variance ont été élaborés pour la poststratification classique et sont disponibles dans les logiciels conçus pour l'analyse des données d'enquête, comme R® (R Development Core Team 2009), SAS® (SAS Institute Inc. 2004), Stata® (StataCorp 2010) et SUDAAN® (Research Triangle Institute 2008). Cependant, peu de travaux portant sur l'estimation de la variance dans le cas de la poststratification en fonction de totaux de contrôle estimés ont été effectués.

Dans les sous-sections qui suivent, nous présentons quatre estimateurs de variance CE pour $\hat{t}_{yP}$ qui tiennent compte de la variance dans les totaux de contrôle après avoir défini la variance d'échantillonnage de population. Ces estimateurs comprennent un estimateur de variance par linéarisation qui vient d'être développé et trois estimateurs de variance par la méthode du jackknife avec suppression d'une UPE. Dans le cas du jackknife avec suppression d'une UPE, les répliques sont créées en supprimant séquentiellement une UPE et en corrigeant les poids pour les UPE restantes dans la strate correspondante du plan de sondage. Cela donne un total de $m_A = \sum_{h=1}^{H} m_{Ah}$ répliques calculé par sommation du nombre d'UPE de l'enquête analytique par strate (m_{Ah}) sur les H strates $(h = 1, ..., H)$.

Un estimateur de variance efficace reproduira la variance d'échantillonnage de population correspondante espérée. La variance d'échantillonnage de population approximative (ou asymptotique) de $\hat{t}_{yP} = \hat{\mathbf{N}}_B' \hat{\bar{\mathbf{Y}}}_A$ a la forme suivante :

$$\begin{aligned} \mathrm{AV}(\hat{t}_{yP}) &= \mathbf{N}_B' \mathbf{V}_A \mathbf{N}_B + 2\bar{\mathbf{Y}}_A' \mathrm{Cov}(\hat{\mathbf{N}}_B, \hat{\bar{\mathbf{Y}}}_A)\mathbf{N}_B + \bar{\mathbf{Y}}_A' \mathbf{V}_B \bar{\mathbf{Y}}_A \\ &= \mathbf{N}_B' \mathbf{V}_A \mathbf{N}_B + \bar{\mathbf{Y}}_A' \mathbf{V}_B \bar{\mathbf{Y}}_A \end{aligned} \quad (5)$$

où $\mathbf{N}_B = E(\hat{\mathbf{N}}_B)$, un vecteur de valeurs prévues pour les dénombrements de poststrate repères dans les G poststrates ; $\hat{\mathbf{N}}_B = (\hat{N}_{B1}, \ldots, \hat{N}_{BG})'$ est un vecteur de longueur G de totaux de contrôle exprimés d'après l'enquête repère ; $\bar{\mathbf{Y}}_A$ est un vecteur de longueur G de composantes de la population de la forme $\bar{y}_{Ag} = t_{Ayg}/N_{Ag}$; $\mathbf{V}_A$ est la matrice de (variance-)covariance des composantes estimées du vecteur $\bar{\mathbf{Y}}_A$; et $\mathbf{V}_B$ est la matrice de covariance des G estimations de contrôle d'après l'enquête repère $\hat{\mathbf{N}}_B$. La première composante, $\mathbf{N}_B' \mathbf{V}_A \mathbf{N}_B$, est la variance approximative de l'estimateur poststratifié classique, $\hat{t}_{yPS}$, c'est-à-dire que les estimations repères sont traitées comme si elles étaient fixes. La composante $\bar{\mathbf{Y}}_A' \mathbf{V}_B \bar{\mathbf{Y}}_A$, est la variance associée aux estimations repères conditionnellement à l'échantillon de l'enquête analytique, c'est-à-dire la composante de la variance de poststratification CE. Parce que nous supposons que les enquêtes analytiques et repères sont indépendantes, la covariance des estimations d'après les deux enquêtes est, par définition, nulle. Donc, la composante $\mathrm{Cov}(\hat{\mathbf{N}}_B, \hat{\bar{\mathbf{Y}}}_A)$ dans (5) est éliminé de l'expression.

Krewski et Rao (1981), Rao et Wu (1985) et d'autres ont démontré la convergence asymptotique des estimateurs de variance par linéarisation et par le jackknife pour des fonctions non linéaires. Cependant, cet examen doit être étendu à la poststratification en fonction de totaux de contrôle estimés (CE). Nous discutons de l'ensemble d'estimateurs de variance CE pour la variance d'échantillonnage de la population identifiés plus bas ou élaboré pour notre étude. Nous avons calculé les estimateurs d'échantillon en substituant les estimations d'échantillon aux paramètres de variance correspondants. Nous commençons par évaluer un estimateur de variance poststratifié classique ou naïf qui ne tient pas compte de la variance dans les totaux de contrôle estimés.

4.1 Un estimateur de variance classique pour la poststratification CE (naïf)

Divers estimateurs de variance ont été élaborés pour les estimateurs par poststratification. Dans toutes les méthodes, les totaux de contrôle sont supposés fixes et connus sans erreur. Par conséquent, $\bar{\mathbf{Y}}_A' \mathbf{V}_B \bar{\mathbf{Y}}_A$, la deuxième composante (positive) dans l'expression (5), est nulle parce que $\mathbf{V}_B = \mathbf{0}$ par hypothèse. L'estimateur de variance par linéarisation prend la forme

$$\mathrm{var}_{\mathrm{Naïf}}(\hat{t}_{yP}) = \hat{\mathbf{N}}_B' \hat{\mathbf{V}}_A \hat{\mathbf{N}}_B \quad (6)$$

où $\hat{\mathbf{N}}_B$ est le vecteur des G estimations repères des totaux de contrôle, et $\hat{\mathbf{V}}_A$ est la matrice de covariance estimée des estimations $\hat{\bar{\mathbf{Y}}}_A = (\hat{t}_{Ay1}/\hat{N}_{A1}, \ldots, \hat{t}_{AyG}/\hat{N}_{AG})$. Comme la deuxième composante dans la deuxième ligne de (5) n'est pas estimée, toute formule de variance élaborée pour la poststratification classique sous-estimera, par définition, la variance d'échantillonnage de la population. Cependant, si les estimations repères sont très précises, la contribution de la composante de variance par poststratification CE à l'estimation globale pourrait être négligeable. Donc, la différence entre les estimations pour la poststratification classique et la poststratification CE sera, dans ces situations, également négligeable.

4.2 Linéarisation par série de Taylor (STCE)

Un estimateur de variance par linéarisation pour $\hat{t}_{yP}$ prend la forme :

$$\mathrm{var}_{\mathrm{STCE}}(\hat{t}_{yP}) = \hat{\mathbf{N}}_B' \hat{\mathbf{V}}_A \hat{\mathbf{N}}_B + \hat{\bar{\mathbf{Y}}}_A' \hat{\mathbf{V}}_B \hat{\bar{\mathbf{Y}}}_A \quad (7)$$

où $\hat{\mathbf{V}}_B$ est la matrice de covariance repère estimée pour l'ensemble des G totaux de contrôle. Les autres termes sont définis pour l'expression (6). La formule STCE est une

fonction de la variance sous la poststratification classique et d'un terme d'accroissement additif associé à la variation dans les totaux de contrôle, c'est-à-dire $\text{var}_{\text{STCE}}(\hat{t}_{yP}) = \text{var}_{\text{Naïf}}(\hat{t}_{yP}) + \hat{\tilde{\mathbf{Y}}}_A' \hat{\mathbf{V}}_B \hat{\tilde{\mathbf{Y}}}_A$.

Idéalement, le fichier d'analyse de l'enquête repère serait disponible pour calculer les valeurs de $\hat{\mathbf{V}}_B$. Toutefois, les chercheurs pourraient devoir se servir d'estimations publiées pour les totaux de contrôle marginaux seulement, c'est-à-dire des estimations ponctuelles et de variance pour une seule caractéristique au lieu des dénombrements et des estimations de covariance pour un ensemble de caractéristiques. Nous discutons plus en détail des incidences lorsque l'information est limitée à la section 4.4.

4.3 Méthode jackknife à deux phases de Fuller (F2CE)

Isaki, Tsay et Fuller (2004) ont appliqué un estimateur de variance jackknife par suppression d'une unité à deux phases élaboré par Fuller (1998) à une situation de poststratification CE. Le principe qui sous-tend la méthode de Fuller (F2CE) consiste à prendre une décomposition spectrale (valeur propre) de la matrice de covariance repère $(\hat{\mathbf{V}}_B)$, à construire des corrections des valeurs repères qui sont une fonction des valeurs propres et des vecteurs propres résultants, puis à ajouter les corrections au vecteur des totaux de contrôle repères $(\hat{\mathbf{N}}_B)$ pour créer un ensemble de répliques des totaux de contrôle. Un sous-ensemble choisi aléatoirement des m_A répliques est poststratifié en fonction des G répliques des totaux de contrôle construites, où le nombre total d'UPE doit être égal ou supérieur au nombre de poststrates, c'est-à-dire $m_A \geq G$. Spécifiquement, le total de contrôle repère pour la r^{e} réplique est défini comme étant

$$\hat{\mathbf{N}}_{B(r)} = \hat{\mathbf{N}}_B + c_h \hat{\mathbf{z}}'_{(r)} \tag{8}$$

où $\hat{\mathbf{z}}'_{(r)} = \delta_{(r)} \sum_{g=1}^{G} \delta_{g|(r)} \hat{\mathbf{z}}'_g$; $c_h = \sqrt{m_{Ah}/(m_{Ah}-1)}$, une constante reliée à la méthode jackknife avec suppression d'une unité d'estimation de la variance ; $\delta_{(r)}$ est un indicateur zéro/un qui identifie les G (parmi les m_A) répliques choisies aléatoirement pour recevoir une correction ; $\delta_{g|(r)} = 1$ si la g^{e} composante de la décomposition de la covariance repère est choisie aléatoirement pour la tâche sachant que la réplique r est sélectionnée pour la correction ; et $\hat{\mathbf{z}}_g = \hat{\mathbf{q}}_g \sqrt{\hat{\lambda}_g}$, une fonction d'un vecteur propre $(\hat{\mathbf{q}}_g)$ et de la valeur propre associée $(\hat{\lambda}_g)$ où $\hat{\mathbf{V}}_B = \sum_{g=1}^{G} \hat{\mathbf{z}}_g \hat{\mathbf{z}}'_g$, par définition. Donc, sachant que $\delta_{(r)} = 1$ pour une réplique particulière, un indicateur unique $\delta_{g|(r)}$ doit alors être égal à un ; cependant, si $\delta_{(r)} = 0$, *tous* les indicateurs $\delta_{g|(r)}$ sont nuls.

Le jackknife avec suppression d'une unité peut prendre de multiples formes selon la valeur de centrage. Nous choisissons l'estimateur de variance un peu prudent centré autour de l'estimation en échantillon complet pour notre étude (v_4 dans Wolter 2007, section 4.5). L'estimateur de variance par le jackknife avec suppression d'une unité, $\text{var}_{\text{F2CE}}(\hat{t}_{yP})$, se calcule comme il suit sous la méthode de Fuller pour un plan d'échantillonnage stratifié à plusieurs degrés.

$$\begin{aligned}\text{var}_{\text{F2CE}}(\hat{t}_{yP}) &= \sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\ddot{t}_{yP(r)} - \hat{t}_{yP})^2 \\ &= \sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{t}_{yP(r)} - \hat{t}_{yP} + c_h \hat{\mathbf{z}}'_{(r)} \hat{\mathbf{B}}_{A(r)})^2 \end{aligned} \tag{9}$$

où les termes de (9) sont définis ci-après. Notons que l'association de la r^{e} réplique à une strate particulière du plan de sondage est définie d'après l'appartenance de l'UPE éliminée à la strate. Dans (9), les estimations répétées sont définies comme étant $\hat{t}_{Ayg(r)} = \sum_h \sum_{i \in s_{Ah}} d_{i(r)} \sum_{k \in s_{Ahi}} \delta_{gk} d_k y_k$ et $\hat{N}_{Ag(r)} = \sum_h \sum_{i \in s_{Ah}} d_{i(r)} \sum_{k \in s_{Ahi}} \delta_{gk} d_k$, où les poids de sous-échantillonnage des UPE sont calculés comme il suit

$$d_{i(r)} = \begin{cases} 0 & \text{si } r = i,\ i \in s_{Ah} \\ 1 & \text{si } h \neq h' \text{ pour } r \in s_{Ah} \text{ et } i \in s_{Ah'} \\ m_{Ah}/(m_{Ah}-1) & \text{si } r \neq i \text{ mais } h = h'. \end{cases} \tag{10}$$

Les autres termes de (9) sont $\hat{\mathbf{B}}_{A(r)} = \hat{t}_{Ayg(r)} / \hat{N}_{Ag(r)}$, la moyenne estimée de la variable de résultat dans la poststrate g et la réplique r ;

$$\ddot{t}_{yP(r)} = \sum_{g=1}^{G} \hat{N}_{Bg(r)} (\hat{t}_{Ayg(r)} / \hat{N}_{Ag(r)}), \tag{11}$$

une fonction des estimations répétées avec $\hat{N}_{Bg(r)}$ définie comme étant la g^{e} composante dans l'expression (8) ; $\hat{t}_{yP(r)}$ est l'estimation répétée sous la poststratification classique, à savoir $\sum_{g=1}^{G} \hat{N}_{Bg} (\hat{t}_{Ayg(r)} / \hat{N}_{Ag(r)})$; et $\hat{t}_{yP}$ est le total estimé donné dans l'expression (3), calculé à partir du fichier de l'échantillon complet. Élever au carré les termes de (9) donne une composante de variance conditionnellement aux totaux de contrôle repères, une composante due à la variabilité des totaux de contrôle repères et un terme croisé de degré plus faible dont l'espérance est approximativement nulle. L'espérance sous le plan de l'estimateur de variance jackknife résultant est asymptotiquement équivalente à $\text{AV}(\hat{t}_{yP})$ dans (5) uniquement si les composantes respectives sont calculées avec des valeurs provenant d'estimateurs convergents sous le plan. Fuller (1998) a également démontré que la variance jackknife des totaux de contrôle répliqués, $\text{var}_{\text{F2CE}}(\hat{\mathbf{N}}_B)$, reproduit la matrice de covariance repère estimée $\hat{\mathbf{V}}_B$ pour chaque échantillon.

À l'heure actuelle, aucun logiciel n'existe pour calculer l'estimateur F2CE. Les six étapes nécessaires pour calculer $\text{var}_{\text{F2CE}}(\hat{t}_{yP})$ en utilisant tout progiciel programmable approprié sont les suivantes :

1. Calculer l'estimation en échantillon complet $\hat{t}_{yP}$ en utilisant l'expression (3).
2. Déterminer les G valeurs propres $\hat{\lambda}_g$ et vecteurs propres $\hat{\mathbf{q}}_g$ pour $\hat{\mathbf{V}}_B$, et calculer les corrections répliquées $\hat{\mathbf{z}}_g = \hat{\mathbf{q}}_g \sqrt{\hat{\lambda}_g}$. Concaténer la matrice $G \times G$ des $\hat{\mathbf{z}}_g$ avec une matrice $G \times (m_A - G)$ de zéros, et trier aléatoirement les colonnes. Désigner par $\hat{\mathbf{Z}}$ cette nouvelle matrice $G \times m_A$.
3. Calculer un vecteur de longueur m_A dont les valeurs sont égales à $c_h = \sqrt{m_{Ah}/(m_{Ah}-1)}$, classé de $h = 1$ à H. Peupler chaque ligne d'une matrice $G \times m_A$, appelée $\mathbf{C}$, avec ce vecteur, c'est-à-dire que les valeurs de ligne sont répliquées. Le vecteur de longueur m_A des poids de strate jackknife, $\mathbf{W}_R$, est créé avec des composantes égales à $(m_{Ah}-1)/m_{Ah}$ où l'UPE supprimée est extraite de la strate h.
4. Calculer le produit de Hadamard (ou élément par élément) (Searle 1982, page 49) de $\hat{\mathbf{Z}}$ et le désigner par $\hat{\mathbf{Z}} \bullet \mathbf{C}$. Répliquer le vecteur $\hat{\mathbf{N}}_B$ dans les colonnes d'une matrice $G \times m_A$ et l'ajouter à $\hat{\mathbf{Z}} \bullet \mathbf{C}$. Cette nouvelle matrice $G \times m_A$, appelée $\hat{\mathbf{N}}_{BR}$, contient les totaux de contrôle repères répliqués discutés dans l'expression (8) pour chacune des m_A répliques.
5. Calculer les estimations répliquées $\hat{\bar{y}}_{Ag(r)} = \hat{t}_{Ayg(r)}/\hat{N}_{Ag(r)}$ en supprimant l'une après l'autre une UPE du fichier de l'échantillon de l'enquête analytique, en corrigeant les poids pour les UPE restantes (valeurs $\mathbf{W}_R$) et en sommant les valeurs pondérées pour le numérateur et le dénominateur dans la poststrate g. Désigner par $\hat{\bar{\mathbf{Y}}}_R$ la matrice $G \times m_A$ résultante.
6. Calculer les m_A estimations répliquées, $\ddot{t}_{yP(r)}$, en commençant par multiplier les éléments $\hat{\mathbf{N}}_{BR}$ par $\hat{\bar{\mathbf{Y}}}_R$, puis en faisant la somme des valeurs de ligne dans une colonne. Puis, soustraire $\hat{t}_{yP}$ de chacune des m_A valeurs et élever les termes au carré, multiplier par les corrections des poids de sous-échantillonnage des UPE spécifiés en (10) et calculer la somme sur les m_A estimations. La valeur résultante est la variance estimée en utilisant la méthode de Fuller, $\text{var}_{\text{F2CE}}(\hat{t}_{yP})$.

4.4 Méthode jackknife de Nadimpalli-Judkins-Chu (NJCCE)

Nadimpalli et coll. (2004) ont élaboré un estimateur de variance jackknife avec suppression d'une unité qui perturbe aléatoirement les totaux de contrôle pour l'ensemble complet de répliques au lieu de corriger uniquement un sous-échantillon de répliques comme dans la méthode F2CE. Pour l'enquête repère, les totaux de contrôle répliqués ont la forme suivante :

$$\hat{\mathbf{N}}_{B(r)} = \hat{\mathbf{N}}_B + c_h R_h \hat{\mathbf{S}}_B \boldsymbol{\eta}_{(r)} \qquad (12)$$

où $c_h = \sqrt{m_{Ah}/(m_{Ah}-1)}$, comme pour la méthode F2CE ; $R_h = \sqrt{1/(H\, m_{Ah})}$, une fonction du nombre total de strates (H) et d'UPE (m_{Ah}) de l'enquête analytique ; $\hat{\mathbf{S}}_B$ est une matrice *diagonale* des erreurs-types estimées pour les totaux de contrôle repères ; et $\boldsymbol{\eta}_{(r)}$ est un vecteur de longueur G de valeurs générées aléatoirement pour chaque réplique à partir de la loi normale standard. Les autres termes sont spécifiés pour la méthode F2CE après l'expression (8). Notons que les estimations de covariance incluses dans l'estimateur F2CE, c'est-à-dire les valeurs hors diagonales de $\hat{\mathbf{V}}_B$, sont fixées à zéro pour l'estimateur NJCCE.

L'estimateur de variance jackknife avec suppression d'une unité correspondant du total poststratifié se calcule comme il suit :

$$\begin{aligned}\text{var}_{\text{NJCCE}}(\hat{t}_{yP}) &= \sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\ddot{t}_{yP(r)} - \hat{t}_{yP})^2 \\ &= \sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{t}_{yP(r)} - \hat{t}_{yP} \\ &\qquad + c_h R_h \boldsymbol{\eta}'_{(r)} \hat{\mathbf{S}}_B \hat{\mathbf{B}}_{A(r)})^2, \qquad (13)\end{aligned}$$

où $\ddot{t}_{yP(r)}$ est calculé comme il est décrit pour l'estimateur F2CE en (11), mais avec $\hat{N}_{Bg(r)}$ défini par la g^{e} composante dans (12). Contrairement à la méthode F2CE, la variance d'échantillon des totaux de contrôle répliqués NJCCE donnée en (12) reproduit l'*espérance* de la matrice de covariance repère $\mathbf{V}_B$ uniquement si les termes de covariance sont réellement nuls (voir l'annexe A pour des détails). Si $\mathbf{V}_B$ n'est pas diagonale, $\text{var}_{\text{NJCCE}}$ ne passe pas ce test.

L'utilisation de la méthode NJCCE serait plausible dans deux cas : *i*) la matrice de covariance repère complète pour les totaux de contrôle n'est pas disponible (par exemple, estimations tirées d'un rapport précédant) ou *ii*) les termes de covariance sont négatifs de sorte que les valeurs résultantes définies par (12) donneraient lieu à des estimations de variance prudentes. La matrice diagonale pour $\hat{\mathbf{S}}_B$ serait correcte si les dénombrements de poststrate estimés étaient vraiment non corrélés. Cependant, cette situation est peu probable, à cause de la structure multinomiale de $\hat{\mathbf{N}}_B$. Étant donné les conditions établies pour la méthode NJCCE, l'espérance de l'estimateur de variance *n*'approximera *pas* $\text{AV}(\hat{t}_{yP})$ donné par (5) ; le terme de biais est relié à la différence entre les espérances sous le plan de $\hat{\mathbf{S}}_B^2$ et de $\mathbf{V}_B$.

4.5 Méthode du jackknife normal multivariée (MVCE)

La méthode normale multivariée (MVCE) est une généralisation de la méthode NJCCE qui, autant que nous sachions, est exposée pour la première fois dans le présent

article. La méthode MVCE utilise la matrice de covariance complète $\hat{\mathbf{V}}_B$ et s'appuie sur la théorie des grands échantillons de sorte que les corrections des totaux de contrôle peuvent être modélisées comme étant issues d'une loi normale multivariée (NMV) à G dimensions. Pour la méthode MVCE, les totaux de contrôle répliqués ont la forme

$$\hat{\mathbf{N}}_{B(r)} = \hat{\mathbf{N}}_B + c_h R_h \hat{\boldsymbol{\varepsilon}}_{(r)} \qquad (14)$$

où $\hat{\boldsymbol{\varepsilon}}_{(r)}$ est un vecteur de longueur G de variables aléatoires tel que $\hat{\boldsymbol{\varepsilon}}_{(r)} \overset{\text{i.i.d.}}{\sim} \text{NMV}_G(\mathbf{0}, \hat{\mathbf{V}}_B)$; $c_h = \sqrt{m_{Ah}/(m_{Ah}-1)}$; et $R_h = \sqrt{1/(H\, m_{Ah})}$.

L'estimateur de variance jackknife avec suppression d'une unité pour la méthode MVCE se calcule comme il suit

$$\begin{aligned}\text{var}_{\text{MVCE}}(\hat{t}_{yP}) &= \sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\ddot{t}_{yP(r)} - \hat{t}_{yP})^2 \\ &= \sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{t}_{yP(r)} - \hat{t}_{yP} \\ &\qquad + c_h R_h \hat{\boldsymbol{\varepsilon}}'_{(r)} \hat{\mathbf{B}}_{A(r)})^2, \qquad (15)\end{aligned}$$

où $\ddot{t}_{yP(r)}$ est calculé comme il est décrit pour la méthode F2CE en (11), mais avec $\hat{N}_{Bg(r)}$ défini par la g^{e} composante dans (14). Contrairement à la méthode de Fuller, $\text{var}_{\text{MVCE}}(\hat{\mathbf{N}}_B) \neq \hat{\mathbf{V}}_B$; à la place, la méthode MVCE doit s'appuyer sur les propriétés de l'estimateur fondées sur le plan de sondage. L'espérance sous le plan de cet estimateur est évaluée par rapport à la loi NMV conditionnellement aux estimations repères (E_ε), puis par rapport au plan de l'enquête repère (E_B). Comme nous le montrons à l'annexe B.1,

$$E_B[E_\varepsilon(\text{var}_{\text{MVCE}}(\hat{\mathbf{N}}_B)|B)] = E_B(\hat{\mathbf{V}}_B). \qquad (16)$$

Si $\hat{\mathbf{V}}_B$ est un estimateur approximativement sans biais de $\mathbf{V}_B$, la matrice de covariance de population est reproduite en appliquant cette méthode.

Sous la méthode à deux phases de Fuller, $\text{Var}[\text{var}_{\text{F2CE}}(\hat{\mathbf{N}}_B)] = \text{Var}(\hat{\mathbf{V}}_B)$ parce que $\text{var}_{\text{F2CE}}(\hat{\mathbf{N}}_B) = \hat{\mathbf{V}}_B$. Pour comparer davantage les méthodes F2CE et MVCE, notons que, si nous définissons $y_k = 1$ dans l'enquête analytique, $\hat{t}_{yP} = \mathbf{1}'\hat{\mathbf{N}}_B$. Comme nous le montrons à l'annexe B.2,

$$\text{Var}[\text{var}_{\text{MVCE}}(\mathbf{1}'\hat{\mathbf{N}}_B)] =$$

$$\text{Var}_B[\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1}] + \frac{2}{H\bar{m}_A^*}[E_B(\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1})^2] > \text{Var}_B[\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1}] \quad (17)$$

où $\bar{m}_A^*$ est la moyenne harmonique des tailles des échantillons d'UPE par strate dans l'enquête analytique. Cela donne à penser que les espérances en grand échantillon de var_{F2CE} et de var_{MVCE} sont similaires, quoiqu'en pratique, l'estimateur MVCE est vraisemblablement plus variable que l'estimateur F2CE. Nous examinons cette question au moyen d'une étude par simulation décrite à la section suivante.

5. Description de l'étude par simulation

Nous complétons l'évaluation théorique des cinq estimateurs de variance dont il était question à la section précédente par l'analyse de résultats de simulation.

5.1 Paramètres de simulation

La population sur laquelle porte la simulation est un sous-ensemble aléatoire du fichier à grande diffusion de la National Health Interview Survey (NHIS) de 2003 contenant les enregistrements obtenus pour 21 664 adultes. Nous avons réparti ces enregistrements en 25 strates, contenant chacune six UPE. Nous avons tiré les échantillons dans cette « population » selon un plan d'échantillonnage à deux degrés. Nous avons d'abord sélectionné deux UPE *avec remise* en utilisant des probabilités proportionnelles au nombre total d'adultes (PPT) dans l'UPE. Dans chaque UPE échantillonnée, nous avons sélectionné des échantillons aléatoires simples de $(n_{Ahi} =)$ 20 et 40 personnes *sans remise*, ce qui a donné des tailles totales d'échantillon de 1 000 et de 2 000, respectivement. Pour notre étude, nous avons considéré deux tailles d'échantillon intra-UPE afin d'évaluer les effets des composantes de la variance d'enquêtes analytiques plus petites, calculées en augmentant n_A, sur la variance de $\hat{t}_{yP}$. Pour chaque combinaison d'échantillons au niveau de l'UPE et au niveau de la personne (c'est-à-dire 50 UPE et soit 1 000 ou 2 000 personnes), nous avons tiré 4 000 échantillons de simulation. Nous avons calculé les estimations des totaux de population et des variances connexes pour deux variables binaires de la NHIS : NOTCOV = 1 indique qu'un adulte *n'était pas* couvert par une asssurance-maladie au cours des 12 mois qui ont précédé l'interview de la NHIS (environ 17 % de la population) et PDMED12M = 1 indique qu'un adulte *avait retardé* des soins médicaux à cause de leur coût au cours des 12 mois qui ont précédé l'interview (environ 7 % de la population). Nous n'avons pas tenu compte de la non-réponse dans la présente étude par simulation afin de réduire au minimum les facteurs susceptibles d'avoir une incidence sur nos comparaisons. (Nota : Les questions de l'interview pour ces variables figurent dans le questionnaire de base sur la famille au ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Survey_Questionnaires/NHIS/2003/qfamilyx.pdf. Nous nous sommes servis des réponses aux questions FHI.070 et FAU.010/FAU.020 pour générer les variables NOTCOV et PDMED12M, respectivement).

La poststratification peut réduire légèrement les variances. Toutefois, dans les enquêtes-ménages, cette technique est principalement utilisée pour corriger le sous-dénombrement

de la base de sondage, ainsi que d'autres problèmes inhérents aux sondages. Chacun des 4 000 échantillons de simulation a été sélectionné de façon à imiter, pour l'enquête analytique, une base de sondage présentant des différences de sous-dénombrement, comme celles utilisée pour de nombreuses enquêtes téléphoniques. Seize $(G = 16)$ cellules de post-stratification ont été définies par croisement d'une variable d'âge à huit niveaux avec le sexe. Les taux de couverture pour les 16 cellules ont été créés en se fondant sur les moyennes de population pour chaque groupe d'âge selon le sexe et varient en valeur de 0,5 à 0,9. Un taux de couverture égal à 1 indiquerait une couverture complète. Avant la sélection de chaque échantillon, la base de sondage a été définie comme un sous-échantillon aléatoire stratifié de la population complète de 21 664. Par exemple, 90 % de la population masculine de 65 à 69 ans ont été sélectionnés aléatoirement pour faire partie de la base de sondage pour les simulations de NOTCOV. Ce processus de définition d'un sous-ensemble de la population pour former la base de sondage a été exécuté indépendamment pour chaque échantillon et pour chaque variable de résultat.

Nous soupçonnons que les chercheurs fondent sur la précision des totaux de contrôle leur décision d'utiliser un estimateur de variance par poststratification classique ou par poststratification CE. Nous avons calculé la matrice de covariance repère $(\hat{\mathbf{V}}_B)$ d'après le fichier de données à grande diffusion complet de la NHIS (92 148 enregistrements) et corrigé proportionnellement les valeurs afin de refléter une taille d'échantillon comparable à notre population de simulation $(N = 21\,664)$. Les valeurs hors diagonale de $\hat{\mathbf{V}}_B$ varient de -0,05 à 0,75 avec une valeur moyenne de 0,22. Partant de cette matrice, nous avons calculé quatre matrices de covariance pour la simulation en divisant la matrice originale par les facteurs de correction 1,0, 3,6, 18 et 72. Les corrections reflètent les enquêtes repères avec une taille effective d'échantillon approximative de 21 700, 6 000 $(\approx 21\,700/3{,}6)$, 1 200 et moins de 500, respectivement.

Nous avons exécuté la simulation au moyen du logiciel R® (Lumley 2009 ; R Development Core Team 2009) étant donné ses très grandes capacités d'analyse de données d'enquête et son efficacité pour les analyses simulées. Nous avons produit un code pour calculer les estimations de la variance par linéarisation et par répliques pour l'estimateur poststratifié CE dont il est question plus haut parce que le code pertinent n'existe pas à l'heure actuelle.

5.2 Critère d'évaluation

Nous avons comparé les résultats empiriques pour les cinq estimateurs de variance discutés à la section précédente (naïf, STCE, F2CE, NJCCE et MVCE) en utilisant trois mesures sur l'ensemble des $j = 1, \ldots, 4\,000$ échantillons de simulation et les deux variables de résultat (NOTCOV et PDMED12M). Les mesures comprennent *i*) le biais relatif en pourcentage estimé de l'estimateur de variance, $(1/4\,000\sum_j \text{var}(\hat{t}_{yP_j}) - \text{eqm})/\text{eqm}$ où $\text{var}(\hat{t}_{yP_j})$ est l'une des cinq estimations de variance évaluées pour l'échantillon j et eqm est l'erreur quadratique moyenne de $\hat{t}_{yP}$ définie ci-après, *ii*) le taux de couverture de l'intervalle de confiance à 95 %, $1/4\,000\sum_j I(|\hat{z}_j| \le z_{1-\alpha/2})$ où $\hat{z}_j = (\hat{t}_{yP_j} - t_y)/\sqrt{\text{var}(\hat{t}_{yP_j})}$ et *iii*) l'écart-type des erreurs-types estimées, calculé comme la racine carrée de $1/(4\,000-1)$ $\sum_j(\sqrt{\text{var}(\hat{t}_{yP_j})} - 1/4\,000\sum_j\sqrt{\text{var}(\hat{t}_{yP_j})})^2$. Le biais relatif et la racine carrée de l'erreur quadratique moyenne de nos estimateurs ponctuels sont calculés sous la forme $1/4\,000\sum_j(\hat{t}_{yP_j} - t_y)/t_y$ et $\sqrt{\text{eqm}} = \sqrt{1/4\,000\sum_s(\hat{t}_{yP_j} - t_y)^2}$, respectivement.

6. Résultats de l'étude par simulation

6.1 Estimateur ponctuel

Afin de justifier le recours à la poststratification, nous avons d'abord évalué l'estimation d'Horvitz-Thompson $(\sum_{s_A} d_k y_k)$ pour les deux variables de résultat. Cet estimateur a la réputation d'être sans biais par rapport au plan de sondage dans des conditions parfaites. Le biais relatif en pourcentage indique que l'estimateur HT présente un biais négatif, sous-estimant le total de population de 38 % pour la variable NOTCOV et de 41 % pour la variable PDMED12M. Ces grandes valeurs indiquent qu'une certaine correction est nécessaire pour ces niveaux non négligeables de biais. Pour l'estimateur poststratifié $\hat{t}_{yP}$, le biais relatif en pourcentage est nettement plus faible, cet estimateur présentant un biais positif n'excédant pas 2 % pour les deux variables de résultat.

6.2 Estimateurs de la variance

Complétant l'évaluation théorique exposée à la section 4, un estimateur de la variance efficace doit produire des résultats empiriques dont le *biais relatif en pourcentage* est quasi nul ou légèrement positif pour une mesure prudente (voir la section 5.2 pour la formule du biais relatif en pourcentage).

Les biais relatifs en pourcentage produits par notre étude par simulation sont présentés au tableau 1. Les estimations du biais sont plus grandes pour les estimateurs naïf et NJCEE de la variance que pour les autres estimateurs fondés sur des totaux de contrôle estimés (CE) pour toutes nos simulations. Les estimations pour l'estimateur STCE sont un peu plus faibles que celles calculées pour les estimateurs F2CE et MVCE pour des enquêtes repères relativement petites. Cependant, les différences sont négligeables quand la taille de l'enquête repère augmente.

Tableau 1
Estimations du biais relatif en pourcentage pour cinq estimateurs de la variance selon la variable de résultat et la taille relative de l'enquête repère par rapport à l'enquête analytique

Variable de résultat	Estimateur de la variance	Taille relative (n_A = 1 000)				Taille relative (n_A = 2 000)			
		0,3	1,2	6,0	21,7	0,2	0,6	3,0	10,8
NOTCOV	Naïf	-50,3	-23	-10,7	-9,2	-56,0	-31	-14,2	-12,2
	STCE	-4,5	-4,5	-6,1	-7,7	-0,2	-8,4	-8,2	-10,1
	F2CE	-4,7	-4,6	-5,8	-7,5	0,1	-8,2	-8,3	-10,1
	NJCCE	-36,7	-17,1	-8,9	-8,2	-40	-24,2	-11,9	-11,1
	MVCE	-4,3	-4,1	-6,0	-7,5	-0,2	-8,1	-8,1	-10,0
PDMED12M	Naïf	-34,4	-14,5	-5,7	-3,9	-48,1	-23,4	-10	-10,1
	STCE	-3,3	-3,7	-2,7	-2,6	-4,7	-6,4	-5,1	-7,8
	F2CE	-3,5	-3,5	-2,4	-2,3	-4,6	-6,8	-5,2	-7,8
	NJCCE	-24,5	-10,5	-4,0	-2,7	-35,1	-17,6	-7,6	-8,4
	MVCE	-3,0	-3,3	-2,4	-2,2	-4,3	-6,3	-5,0	-7,7

Comme prévu, l'estimateur poststratifié classique (naïf) est celui dont le biais est le plus négatif parmi les estimateurs comparés. Quand l'enquête repère est de plus petite portée que l'enquête analytique (et par conséquent produit des estimations moins précises que cette dernière), l'estimateur naïf présente un biais négatif allant jusqu'à 56 %. Le niveau de biais s'améliore à mesure que la taille relative de l'enquête repère augmente ; toutefois, l'estimateur naïf produit encore, au mieux, une sous-estimation de 4 %. L'estimateur NJCCE donne d'un peu meilleurs résultats que l'estimateur naïf, quoique le biais (-2,7 à -40 %) demeure plus grand que pour les autres estimateurs CE de la variance, pour lesquels il varie de -10,1 à 0,1 %.

Pour une petite enquête repère relativement à la taille de l'enquête analytique (c'est-à-dire la taille relative moins 1), les niveaux de biais (absolus) augmentent spectaculairement pour les estimateurs naïf et NJCCE. Nous constatons l'effet opposé pour les autres estimateurs CE de la variance. La composante de la variance associée à l'enquête repère, c'est-à-dire $\hat{\bar{\mathbf{Y}}}_A' \hat{\mathbf{V}}_B \hat{\bar{\mathbf{Y}}}_A$ donnée pour var_{STCE} dans (7), devient le terme dominant dans les estimateurs CE de la variance à mesure que la précision des estimations fondées sur l'enquête repère diminue. Donc, la composante de la variance repère corrige dans une certaine mesure la sous-estimation associée à la composante de la variance analytique. D'autres travaux de recherche sont nécessaires pour déterminer s'il existe un seuil auquel peut avoir lieu ce genre de compensation du biais. Le biais négatif global de nos estimations est comparable au biais des estimateurs de la variance par linéarisation présenté dans un autre contexte par Rao et Wu (1985, section 4) et par Wu (1985). Cependant, les travaux devront se poursuivre afin de déterminer comment réduire au minimum la sous-estimation.

Notons que les tailles relatives de 21,7 quand $n_A = 1\,000$ et de 10,8 quand $n_A = 2\,000$ impliquent toutes deux des tailles d'échantillon de l'enquête repère d'environ 21 600. Donc, la composante d'ordre $O(M^2/m_B)$ de la variance, $\bar{\mathbf{Y}}_A' \mathbf{V}_B \bar{\mathbf{Y}}_A$, est plus importante pour les estimations du tableau 1 fondées sur $n_A = 2\,000$. Cela donne lieu à des biais relatifs plus grands dans ces estimations que ceux produits sous $n_A = 1\,000$, même si la taille de l'échantillon de l'enquête analytique est plus grand.

Les tendances qui se dégagent pour le biais relatif en pourcentage sont reflétées par les *taux de couverture des intervalles de confiance à 95 %* pour les totaux estimés, mais qui ne sont pas présentés par souci de concision. Les estimateurs naïf et NJCCE sont plus susceptibles de donner lieu à des taux de couverture des intervalles de confiance inférieurs à 95 %. Ces taux s'approchent du niveau approprié à mesure que la précision des estimations fondées sur l'enquête repère s'améliore. Toutefois, les autres estimateurs CE de la variance ont des taux de couverture proches des niveaux acceptables quelle que soit la taille relative des enquêtes et, par conséquent, sont plus robustes.

Jusqu'à présent, la discussion donne à penser que les différences théoriques, ainsi qu'empiriques, entre les méthodes STCE, F2CE et MVCE sont minimales. Enfin, nous examinons l'*écart-type des erreurs-types* (*e.-t.*) *estimées* pour tenter de distinguer les estimateurs. Un examen de cette variabilité peut donner une idée de la stabilité (empirique) des estimateurs de la variance, c'est-à-dire qu'un estimateur de la variance instable pourrait produire une estimation de la variance médiocre en fonction des nuances d'un échantillon particulier. Le tableau 2 donne l'accroissement relatif en pourcentage des écarts-types pour les méthodes F2CE et MVCE, toutes deux comparativement à la méthode STCE.

La variation des estimations MVCE de la variance est appréciablement plus grande que celle produite par la méthode F2CE, mais uniquement pour des enquêtes de référence relativement petites. L'écart augmente à mesure que la taille de l'échantillon de l'enquête analytique augmente. Ce qui laisse entendre qu'on pourrait préférer la méthode F2CE à la méthode MVCE étant donné la stabilité accrue des estimations de la variance. Toutefois, nous poursuivons ces travaux en vue de déterminer le seuil auquel l'instabilité peut avoir une incidence sur les estimations.

Tableau 2
Accroissement en pourcentage de l'instabilité des estimations de la variance comparativement à la méthode STCE selon la variable de résultat et la taille relative de l'enquête repère

		Taille relative (n_A = 1 000)				Taille relative (n_A = 2 000)			
Variable de résultat	Estimateur de la variance	0,3	1,2	6,0	21,7	0,2	0,6	3,0	10,8
NOTCOV	F2CE	12,0	5,5	2,3	0,2	15,1	8,4	2,1	0,6
	MVCE	21,2	7,4	1,8	0,3	30,8	8,5	2,4	0,7
PDMED12M	F2CE	7,7	3,8	1,1	0,4	12,0	6,3	2,1	0,7
	MVCE	11,5	4,0	0,9	0,5	22,6	7,6	2,2	1,1

7. Conclusion et futurs travaux

Les travaux théoriques et analytiques exposés dans le présent article appuient l'idée qu'une nouvelle méthodologie est nécessaire pour traiter la poststratification lorsque l'on utilise des totaux de contrôle estimés, c'est-à-dire la poststratification en fonction de totaux de contrôle estimés (CE). Les estimateurs classiques de la variance peuvent sous-estimer fortement la variance d'échantillonnage de la population, ce qui risque, par exemple, de donner lieu à des décisions incorrectes pour les tests d'hypothèse et à des répartitions non optimales de l'échantillon quand le plan de sondage est mis en œuvre ultérieurement.

L'estimateur de la variance par linéarisation CE var_{STCE} donné par l'expression (7) est prometteur pour la poststratification CE. Cet estimateur réduit particulièrement bien le biais relatif en pourcentage observé pour l'estimateur naïf de la variance donné par (6) quand l'enquête repère est petite comparativement à l'enquête analytique. L'estimateur de la variance par répliques var_{F2CE} donné par (9) est recommandé spécifiquement pour des études nécessitant des poids de répliques, par exemple quand les fichiers d'analyse à grande diffusion sont diffusés sans information sur le plan de sondage pour accroître la protection des données confidentielles et de la vie privée des répondants. L'estimateur par répliques de rechange var_{MVCE} donne aussi de bons résultats et est un peu plus facile à appliquer que var_{F2CE}.

La mise en œuvre des estimateurs de la variance recommandés requiert des programmes informatiques spécialisés, parce que les fonctions requises ne sont pas disponibles à l'heure actuelle dans les logiciels standard. L'estimateur par linéarisation pourrait être plus approchable, parce que sa mise en œuvre comporte une modification en fonction des estimations de la variance disponibles, par exemple $\text{var}_{\text{STCE}}(\hat{t}_{y\text{PSCE}}) = \text{var}_{\text{Naïf}}(\hat{t}_{y\text{PSCE}}) + \hat{\bar{\mathbf{Y}}}_A' \hat{\mathbf{V}}_B \hat{\bar{\mathbf{Y}}}_A$. Nous donnons une discussion étape par étape des procédures requises pour l'estimateur var_{F2CE} (voir la section 4.3) pour faciliter la création du programme informatique.

Des extensions des présents travaux de recherche qui seront présentées à une date ultérieure comprennent une généralisation au calage linéaire, à d'autres statistiques, y compris la moyenne estimée par le ratio et à l'estimation par domaine. Nous cherchons aussi à savoir s'il est possible de dégager des valeurs seuils qui déterminent *i*) quand les différences entre les estimations classique et CE de la variance sont négligeables et *ii*) quand les totaux de contrôle repères sont trop imprécis pour être utilisés pour le calage. Nous prévoyons aussi étudier les incidences théoriques des erreurs de mesure dans les enquêtes analytiques ainsi que dans les enquêtes repères.

Remerciements

Les présents travaux ont été exécutés dans le cadre de la thèse de doctorat de la première auteure au Joint Program in Survey Methodology, à l'Université du Maryland. Elle remercie les membres de son comité, Richard Valliant, Phillip Kott, Frauke Kreuter, Stephen Miller et Paul Smith, de leur encadrement. Les auteurs remercient également le rédacteur associé et les examinateurs de leurs commentaires constructifs qui ont permis d'éclaircir l'exposé.

Annexe A

Calcul de $\text{var}_{\text{NJCCE}}(\hat{\mathbf{N}}_B)$

Pour les calculs qui suivent, soit E_ε l'espérance par rapport à une loi normale standard. Tous les autres termes sont définis dans le corps de l'exposé.

$$\text{var}_{\text{NJCCE}}(\hat{\mathbf{N}}_B) = \sum_{h=1}^{H} \frac{m_{Ah} - 1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{\mathbf{N}}_{B(r)} - \hat{\mathbf{N}}_B)(\hat{\mathbf{N}}_{B(r)} - \hat{\mathbf{N}}_B)'$$

$$= \frac{1}{H} \hat{\mathbf{S}}_B \left(\sum_{h=1}^{H} \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} \boldsymbol{K}_{(r)} \right) \hat{\mathbf{S}}_B$$

où $\boldsymbol{K}_{(r)} = \boldsymbol{\eta}_{(r)} \boldsymbol{\eta}_{(r)}'$, une matrice produit vectoriel de dimension $G \times G$ des valeurs normales standard et $\hat{\mathbf{S}}_B^2 = \text{diag}(\hat{\mathbf{V}}_B)$. Parce que $E_\varepsilon(\boldsymbol{K}_{(r)}) = \mathbf{I}_G$, une matrice identité de dimension G, nous avons $E_\varepsilon[\text{var}_{\text{NJCCE}}(\hat{\mathbf{N}}_B)] = \text{diag}(\hat{\mathbf{V}}_B)$. D'où, $\text{var}_{\text{NJCEC}}(\hat{\mathbf{N}}_B)$ ne reproduit pas l'espérance de $\hat{\mathbf{V}}_B$.

Annexe B

Évaluation de l'estimateur MVCE

Pour les calculs qui suivent, soit E_B et Var_B l'espérance et la variance par rapport au plan d'échantillonnage de l'enquête repère. En outre, soit E_ε et Var_ε l'espérance et la variance par rapport à la loi normale multivariée à G dimensions, $\text{NMV}_G(\mathbf{0}, \hat{\mathbf{V}}_B)$. Tous les autres termes sont définis dans le corps de l'exposé.

B.1 : Calcul de $E[\text{var}_{\text{MVEC}}(\hat{\mathbf{N}}_B)]$ donnée dans (15)

En utilisant l'expression (14) et $c_h^2 = m_{Ah}/(m_{Ah}-1)$,

$$
\begin{aligned}
E[\text{var}_{\text{MVCE}}(\hat{\mathbf{N}}_B)] &= E_B\left[E_\varepsilon\left(\sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \right.\right. \\
&\qquad \left.\left. \sum_{r=1}^{m_{Ah}} (\hat{\mathbf{N}}_{B(r)} - \hat{\mathbf{N}}_B)(\hat{\mathbf{N}}_{B(r)} - \hat{\mathbf{N}}_B)' \,\middle|\, B \right) \right], \\
&= \frac{1}{H} E_B\left[\sum_{h=1}^{H} \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} E_\varepsilon(\hat{\boldsymbol{\varepsilon}}_{(r)} \hat{\boldsymbol{\varepsilon}}'_{(r)} \,|\, B) \right] \\
&= \frac{1}{H} \sum_{h=1}^{H} \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} E_B(\hat{\mathbf{V}}_B) = E_B(\hat{\mathbf{V}}_B).
\end{aligned}
$$

B.2 : Calcul de $\text{Var}[\text{var}_{\text{MVCE}}(\hat{\mathbf{N}}_B)]$ donnée dans (15)

Quand $y_k = 1$ de sorte que $\hat{t}_{yP} = \mathbf{1}'\hat{\mathbf{N}}_B$, $\text{var}_{\text{MVCE}}(\mathbf{1}'\hat{\mathbf{N}}_B) = H^{-1}\sum_{h=1}^{H} m_{Ah}^{-1} \sum_{r=1}^{m_{Ah}} \mathbf{1}'\hat{\boldsymbol{\varepsilon}}_{(r)}\hat{\boldsymbol{\varepsilon}}'_{(r)}\mathbf{1}$. En utilisant la formule pour la variance d'une forme quadratique (Searle 1982, section 13.5), nous avons

$$
\begin{aligned}
&\text{Var}[\text{var}_{\text{ECMV}}(\mathbf{1}'\hat{\mathbf{N}}_B)] \\
&\quad = \text{Var}_B\left[\frac{1}{H} \sum_{h=1}^{H} \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} E_\varepsilon(\mathbf{1}'\hat{\boldsymbol{\varepsilon}}_{(r)}\hat{\boldsymbol{\varepsilon}}'_{(r)}\mathbf{1} \,|\, B) \right] \\
&\quad + E_B\left[\frac{1}{H^2} \sum_{h=1}^{H} \frac{1}{m_{Ah}^2} \sum_{r=1}^{m_{Ah}} \text{Var}_\varepsilon(\mathbf{1}'\hat{\boldsymbol{\varepsilon}}_{(r)}\hat{\boldsymbol{\varepsilon}}'_{(r)}\mathbf{1} \,|\, B) \right] \\
&\quad = \text{Var}_B\left[\frac{1}{H} \sum_{h=1}^{H} \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} \mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1} \right] \\
&\quad + E_B\left[\frac{1}{H^2} \sum_{h=1}^{H} \frac{1}{m_{Ah}} \{2tr(\mathbf{1}\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1}\mathbf{1}'\hat{\mathbf{V}}_B)\} \right] \\
&\quad = \text{Var}_B[\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1}] + \frac{2}{H\bar{m}_A^*}[E_B(\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1})^2],
\end{aligned}
$$

où $\bar{m}_A^* = (H^{-1}\sum_{h=1}^{H} m_{Ah}^{-1})^{-1}$ est la moyenne harmonique de de m_{Ah}.

Bibliographie

Binder, D.A. (1995). Méthodes de linéarisation pour les échantillons à une et deux phases : une approche de type « recette ». *Techniques d'enquête*, 22, 1, 17-22.

Bray, R., Hourani, L., Rae, K., Dever, J., Brown, J., Vincus, A., Pemberton, M., Marsden, M., Faulkner, D. et Vandermaas-Peeler, R. (2003). 2002 Department of Defense Survey of Health Related Behaviors Among Military Personnel. Rapport technique RTI/7841/006-FR, U.S. Department of Defense préparé par RTI International. URL http://dodwws.rti.org/2002WWFinalReportComplete05-04.pdf.

Canty, A.J., et Davison, A.C. (1999). Resampling-based variance estimation for Labour Force Surveys. *The Statistician*, 48, 379-391.

Centers for Disease Control and Prevention (2006). Technical Information and Data for the Behavioral Risk Factor Surveillance System (BRFSS) – BRFSS Weighting Formula. Atlanta, Georgia : U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 11 septembre 2006.

Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 1, 17-27.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 88(423), 1013-1020.

Estevao, V.M., et Särndal, C.-E. (2000). A Functional form approach to calibration. *Journal of Official Statistics*, 16(4), 379-399.

Fuller, W.A. (1998). Replication variance estimation for the two-phase samples. *Statistica Sinica*, 8, 1153-1164.

Hidiroglou, M.A., et Patak, Z. (2006). Raking ratio estimation: An application to the Canadian Retail Trade Survey. *Journal of Official Statistics*, 22(1), 71-80.

Isaki, C.T., Tsay, J.H. et Fuller, W.A. (2004). Pondération de données d'échantillon reposant sur des contrôles indépendants. *Techniques d'enquête*, 30, 1, 39-49.

Jayasuriya, B.R., et Valliant, R. (1996). Application de l'estimation par régression restreinte dans une enquête-ménage. *Techniques d'enquête*, 22, 2, 127-138.

Keeter, S., Dimock, M. et Christian, L. (2008). Calling Cell Phones in '08 Pre-Election Polls. Nouvelle diffusion (18 décembre 2008) : Pew Research Center for the People & the Press. URL http://people-press.org/reports/pdf/cell-phone-commentary.pdf.

Killion, R.A. (2006). Weighting Specifications for The American Time Use Survey (ATUS) for 2006. U.S. Bureau of the Census, Internal Memo (Doc.#ATUS-16).

Kim, J.J., Li, J. et Valliant, R. (2007). Regroupement de cellules lors de la poststratification. *Techniques d'enquête*, 33, 2, 157-170.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 2, 149-160.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9(5), 1010-1019.

Krotki, K. (2007). Combining RDD and Web Panel Surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association (sous presse).

Lessler, J.T., et Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. New York : John Wiley & Sons, Inc.

Lumley, T. (2009). Survey: Analysis of complex survey samples. R package version 3.19. University of Washington : Seattle.

Mirza, H., et Hörngren, J. (2002). The Sampling and the Estimation Procedure in the Swedish Labour Force Survey. Rapport technique, Statistics Sweden, Stockholm : Sweden.

Nadimpalli, V., Judkins, D. et Chu, A. (2004). Survey Calibration to CPS Household Statistics. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 4090-4094.

R Development Core Team (2009). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Disponible au : http://www.R-project.org.

Rao, J.N.K., et Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86(2), 403-415.

Rao, J.N.K., et Wu, C.F.J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80(391), 620-630.

Research Triangle Institute (2008). *SUDAAN Language Manual*. Release 10.0, Research Triangle Park, NC : Research Triangle Institute.

Rust, K.F., et Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. England : John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. et Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3), 527-537.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag, Inc.

SAS Institute Inc. (2009). *SAS/STAT® 9.2 User's Guide*. Cary, NC : SAS Institute Inc.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics*. New York : John Wiley & Sons, Inc.

StataCorp (2010). *Stata Statistical Software*: *Release 11*. Survey Data, College Station, TX : StataCorp LP.

Stukel, D.M., Hidiroglou, M.A. et Särndal, C.-E. (1996). Estimation de la variance des estimateurs de calage : comparaison des méthodes du jackknife et de la linéarisation de Taylor. *Techniques d'enquête*, 22, 2, 117-126.

Taylor, M.F., Brice, J., Buck, N. et Prentice-Lane, E. (2007). *British Household Panel Survey User Manual Volume A*: *Introduction, Technical Report and Appendices*. University of Essex, Colchester.

Terhanian., G., Bremer, J., Smith, R. et Thomas, R. (2000). *Correcting Data from Online Survey for the Effects of Nonrandom Selection and Nonrandom Assignment*. Papier de recherche : Harris Interactive.

Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94(446), 635-644.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.

Wolter, K.M. (2007). *Introduction to Variance Estimation*. New York : Springer Science+Business Media, LLC.

Wu, C.F.J. (1985). Variance estimation for the combined ratio and combined regression estimators. *Journal of the Royal Statistical Society*, Séries B, 47(1), 147-154.

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

Yung, W., et Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95(451), 903-915.