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

Article

L'effet des ajustements pour la non-réponse sur l'estimation de la variance

par David Haziza, Katherine Jenny Thompson et Wesley Yung

Juin 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2010
Vol. 36, N° 1, pp. 39-48
Statistique Canada, N° 12-001-X au catalogue

L'effet des ajustements pour la non-réponse sur l'estimation de la variance

David Haziza, Katherine Jenny Thompson et Wesley Yung [1]

Résumé

Dans le cas de nombreux sondages, des procédures d'ajustement des poids sont utilisées pour réduire le biais de non-réponse. Ces ajustements s'appuient sur les données auxiliaires disponibles. Le présent article traite de l'estimation de la variance par la méthode du jackknife pour les estimateurs qui ont été corrigés de la non-réponse. En suivant l'approche inversée d'estimation de la variance proposée par Fay (1991), ainsi que par Shao et Steel (1999), nous étudions l'effet dû au fait de ne pas recalculer l'ajustement des poids pour la non-réponse dans chaque réplique jackknife. Nous montrons que l'estimateur de variance jackknife « simplifié » résultant a tendance à surestimer la variance réelle des estimateurs ponctuels dans le cas de plusieurs procédures d'ajustement des poids utilisées en pratique. Ces résultats théoriques sont confirmés au moyen d'une étude par simulation dans laquelle nous comparons l'estimateur de variance jackknife simplifié à l'estimateur de variance jackknife complet obtenu en recalculant l'ajustement des poids pour la non-réponse dans chaque réplique jackknife.

Mots clés : Calage ; ajustement pour la non-réponse ; non-réponse totale ; estimateur de la variance par la méthode du jackknife ; estimateur de la variance par linéarisation.

1. Introduction

La non-réponse totale, c'est-à-dire la situation où, pour une unité échantillonnée, les données manquent pour toutes les variables étudiées ou l'information utilisable n'est pas suffisante, est inévitable dans les sondages. Pour résoudre ce problème, les non-répondants sont supprimés du fichier de données et les poids de sondage des répondants sont ajustés pour tenir compte des unités supprimées. L'objectif principal d'une méthode d'ajustement des poids consiste à réduire le biais de non-réponse qui survient quand les répondants et les non-répondants diffèrent en ce qui concerne les variables étudiées. Pour bien réduire le biais, il est essentiel d'utiliser de l'information auxiliaire puissante, disponible pour les répondants ainsi que les non-répondants.

Dans le présent article, nous considérons l'estimation de la variance par la méthode du jackknife en présence de non-réponse totale. Cette méthode d'estimation de la variance est très répandue en pratique en raison de ses propriétés théoriques et de la simplicité des calculs. Contrairement aux méthodes de linéarisation de Taylor, la méthode du jackknife ne nécessite pas le calcul individuel de chaque paramètre d'intérêt ni des probabilités d'inclusion de deuxième ordre qui sont parfois difficiles à obtenir dans les enquêtes complexes. En cas d'utilisation d'un estimateur jackknife de variance dans le contexte de la non-réponse, certains ont soulevé la question de savoir s'il faut ou non produire des répliques de l'ajustement pour la non-réponse (par exemple, Valliant 2004). Dans le présent article, nous considérons deux estimateurs de variance par le jackknife, à savoir i) un estimateur de variance jackknife *complet* qui recalcule le facteur d'ajustement pour la non-réponse dans chaque réplique jackknife et ii) un estimateur de variance jackknife *simplifié*, qui ne le fait pas. Ce deuxième estimateur est commode en pratique, mais, autant que nous sachions, ses propriétés théoriques n'ont pas été complètement étudiées dans la littérature. Des considérations tenant à la production ont tendance à dicter l'usage d'un estimateur de variance jackknife simplifié, car dans le contexte de l'échantillonnage stratifié, l'estimateur de variance jackknife complet peut demander assez bien de temps et de ressources informatiques, surtout quand le programme d'enquête utilise un grand nombre de cellules de pondération. Selon certaines études menées récemment par le U.S. Census Bureau (Thompson 2005, ainsi que Ozcoskun, Thompson et Williams 2005), les différences sont négligeables entre les estimations de variance obtenues en utilisant une méthode d'ajustement des poids avec répliques complètes et celles obtenues en utilisant une méthode « simplifiée » avec un estimateur de variance par le jackknife stratifié, le jackknife avec suppression d'un groupe ou la méthode des demi-échantillons modifiés.

Deux catégories de procédures d'ajustement sont utilisées couramment en pratique. La première, appelée *pondération par la propension à la non-réponse* (PPN), consiste à modéliser d'abord les propensions à répondre, puis à utiliser l'inverse des propensions estimées comme facteurs de correction de la pondération. Les propensions à répondre estimées sont habituellement obtenues en ajustant un modèle paramétrique (par exemple, modèle de régression

1. David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, H3C 3J7, Canada. Courriel : David.haziza@umontreal.ca ; Katherine Jenny Thompson, U.S. Census Bureau, Washington, DC 20233. Courriel : Katherine.J.Thompson@census.gov ; Wesley Yung, Statistique Canada, Ottawa (Ontario) K1A 0T6. Courriel : wesley.yung@statcan.gc.ca.

logistique) ou un modèle non paramétrique (par exemple Da Silva et Opsomer 2006). Un cas particulier de PPN, qui est très répandu en pratique, consiste à d'abord répartir les répondants et les non-répondants en classes de pondération, puis à ajuster les poids de sondage des répondants par l'inverse des taux de réponse dans chaque classe. Ces classes sont formées en se basant sur l'information auxiliaire recueillie pour toutes les unités comprises dans l'échantillon ; voir, par exemple, Eltinge et Yansaneh (1997), ainsi que Little (1986). La deuxième catégorie de procédures d'ajustement, appelée *pondération par calage pour la non-réponse* (PCN), peut être considérée comme une extension de l'approche du calage (Deville et Särndal 1992) adaptée au contexte de la non-réponse totale. Le lecteur est invité à consulter Särndal et Lundström (2005), Kott (2006), ainsi que Brick et Montaquila (2008) pour un survol complet de la PPN et de la PCN. Dans certaines situations, la PPN et la PCN mènent au même estimateur, par exemple, l'estimateur ajusté par la fréquence dans la cellule présenté plus loin (voir l'expression (1.4)). Dans le présent article, nous nous concentrons sur la pondération par calage pour la non-réponse (PCN). L'estimation de la variance dans le contexte de la PPN a été étudiée récemment par Kim et Kim (2007).

Considérons une population finie U de taille N. L'objectif est d'estimer le total de population $Y = \sum_{i \in U} y_i$ d'une variable d'intérêt y. Supposons qu'un échantillon aléatoire s de taille n est sélectionné dans U conformément à un plan donné $p(s)$. Dans le cas de données complètes, un estimateur de base de Y est l'estimateur à facteur d'extension bien connu donné par

$$\hat{Y}_\pi = \sum_{i \in s} d_i y_i \tag{1.1}$$

où $d_i = 1/\pi_i$ désigne le poids de sondage relié à l'unité i et $\pi_i = P(i \in s)$ désigne la probabilité d'inclusion de premier ordre dans l'échantillon. En présence de non-réponse totale, un sous-ensemble seulement de s est observé et le calcul de $\hat{Y}_\pi$ selon (1.1) est impossible.

Afin de définir un estimateur de Y ajusté pour la non-réponse, nous supposons qu'il existe un vecteur de variables auxiliaires $\mathbf{x}$ pour toutes les unités échantillonnées (répondants et non-répondants), de sorte que le vecteur des totaux estimés, $\hat{\mathbf{X}}_\pi = \sum_{i \in s} d_i \mathbf{x}_i$, est disponible. Nous supposons aussi qu'il existe un vecteur de variables instrumentales $\mathbf{z}$, de même dimension que $\mathbf{x}$, pour les répondants. Soit r_i un indicateur de réponse relié à l'unité i tel que $r_i = 1$ si l'unité i est une unité répondante et $r_i = 0$, autrement. Pour estimer Y, nous considérons les estimateurs par calage de la forme

$$\hat{Y}_{\text{CAL}} = \sum_{i \in s} w_i r_i y_i, \tag{1.2}$$

où $w_i = d_i g_i$ et g_i est un facteur d'ajustement de la pondération pour la non-réponse relié à l'unité i et donné par

$$g_i = 1 + (\hat{\mathbf{X}}_\pi - \hat{\mathbf{X}}_r)' \hat{\mathbf{T}}_r^{-1} \mathbf{z}_i, \tag{1.3}$$

où $\hat{\mathbf{X}}_r = \sum_{i \in s} d_i r_i \mathbf{x}_i$ et $\hat{\mathbf{T}}_r = \sum_{i \in s} d_i r_i \mathbf{z}_i \mathbf{x}_i'$. Quand $\mathbf{z}_i = \mathbf{x}_i/\nu_i$, où ν_i est une constante connue, l'estimateur (1.3) est identique à l'estimateur *InfoS* donné dans Särndal et Lundström (2005, équation 7.15). Les propriétés de l'estimateur (1.2) ont été étudiées, entre autres, par Deville (2002), Sautory (2003), Särndal et Lundström (2005), et Kott (2006).

Ici, nous évaluons les propriétés (par exemple biais et variance) de $\hat{Y}_{\text{CAL}}$ en utilisant l'approche du modèle de non-réponse sous lequel l'inférence est faite par rapport à la loi conjointe induite par le plan d'échantillonnage et par le mécanisme de non-réponse, $q(\mathbf{r} \mid \mathbf{I})$, où $\mathbf{I} = (I_1, ..., I_N)'$ est le vecteur des indicateurs de sélection dans l'échantillon tels que $I_i = 1$ si l'unité i est sélectionnée dans l'échantillon et $I_i = 0$, autrement et $\mathbf{r} = (r_1, ..., r_N)'$ est le vecteur des indicateurs de réponse. Soit $p_i = P(r_i = 1 \mid \mathbf{I}, I_i = 1)$ la probabilité de réponse pour l'unité i. Nous supposons que $p_i > 0$ pour tout i et que les unités répondent indépendamment les unes des autres ; autrement dit, $p_{ij} = P(r_i = 1, r_j = 1 \mid \mathbf{I}, I_i = 1, I_j = 1, i \neq j) = p_i p_j$.

L'estimateur $\hat{Y}_{\text{CAL}}$ est asymptotiquement sans biais pour le total réel Y si i) $p_i^{-1} = 1 + \boldsymbol{\lambda}' \mathbf{z}_i$ pour tout $i \in U$, où $\boldsymbol{\lambda}$ est un vecteur de constantes inconnues ou ii) $y_i = \mathbf{x}_i' \boldsymbol{\beta}$ pour tout $i \in U$, où $\boldsymbol{\beta}$ est un vecteur de constantes ; voir Särndal et Lundström (2005, chapitre 9.5). Si la condition (i) est satisfaite, l'estimateur ponctuel $\hat{Y}_{\text{CAL}}$ est asymptotiquement sans biais pour Y quelle que soit la variable d'intérêt y estimée. En outre, il découle de (ii) que $\hat{Y}_{\text{CAL}}$ présente un petit biais si les résidus $E_i = y_i - \mathbf{x}_i' \mathbf{B}$ sont faibles, où $\mathbf{B} = (\sum_{i \in U} \mathbf{z}_i \mathbf{x}_i')^{-1} \sum_{i \in U} \mathbf{z}_i y_i$. Par conséquent, le biais de l'estimateur $\hat{Y}_{\text{CAL}}$ est petit si le vecteur $\mathbf{x}$ explique la variable d'intérêt y. Dans le cas de plusieurs variables d'intérêt, notons que le vecteur $\mathbf{x}$ pourrait expliquer convenablement une variable d'intérêt donnée, mais ne pas y être relié du tout, auquel cas certaines estimations pourraient être biaisées. Nous supposons que $\hat{Y}_{\text{CAL}}$ est asymptotiquement sans biais pour Y, de sorte que la question du biais des estimateurs étudiés ne se posera pas dans la suite de l'exposé.

Nous considérons trois cas particuliers de (1.2) qui présentent un intérêt pratique (voir également Kalton et Flores-Cervantes 2003). Premièrement, soit $\boldsymbol{\delta} = (\delta_{i1}, ..., \delta_{ic}, ..., \delta_{iC})'$ un vecteur de dimension C d'indicateurs de classe de pondération attachés à l'unité i tels que $\delta_{ic} = 1$ si l'unité i appartient à la classe c et $\delta_{ic} = 0$, autrement pour $c = 1, ..., C$. Si $\mathbf{x}_i = \mathbf{z}_i = \boldsymbol{\delta}_i$, le facteur d'ajustement g_i donné par (1.3) se réduit à $g_i = \hat{N}_c/\hat{N}_{rc}\, \delta_{ic}$, où $\hat{N}_c = \sum_{i \in s} d_i \delta_{ic}$ et $\hat{N}_{rc} = \sum_{i \in s} d_i r_i \delta_{ic}$. Autrement dit, le facteur

d'ajustement de la pondération pour la non-réponse dans une cellule de pondération est calculé en divisant le nombre, pondéré par les poids de sondage, d'unités échantillonnées dans la cellule de pondération par le nombre, pondéré par les poids de sondage, d'unités répondantes dans la cellule de pondération. Nous donnons à cette procédure d'ajustement des poids le nom de procédure d'*ajustement par la fréquence*. Il s'ensuit que l'estimateur (1.2) se réduit à l'estimateur ajusté par la fréquence.

$$\hat{Y}_{\text{freq.}} = \sum_{c=1}^{C} \frac{\hat{N}_c}{\hat{N}_{rc}} \hat{Y}_{rc}, \tag{1.4}$$

où

$$\hat{Y}_{rc} = \sum_{i \in s} d_i \, r_i \, \delta_{ic} \, y_i.$$

Le deuxième cas particulier de (1.2) repose sur l'hypothèse qu'une variable continue x est disponible pour toutes les unités échantillonnées. Soit $\mathbf{x}_i = (\delta_{i1} x_i, ..., \delta_{ic} x_i, ..., \delta_{iC} x_i)'$ et $\mathbf{z}_i = \boldsymbol{\delta}_i$. Dans ce cas, le facteur d'ajustement g_i donné par (1.3) se réduit à $g_i = \hat{X}_c / \hat{X}_{rc} \, \delta_{ic}$ si l'unité i appartient à la classe c, où $\hat{X}_c = \sum_{i \in s} d_i \, \delta_{ic} \, x_i$ et $\hat{X}_{rc} = \sum_{i \in s} d_i \, r_i \, \delta_{ic} \, x_i$. Ici, le facteur d'ajustement de la pondération pour la non-réponse pour une classe de pondération c est égal à la somme des données auxiliaires pondérées par les poids de sondage pour les unités dans la cellule de pondération divisée par la somme des données auxiliaires pondérées par les poids de sondage pour toutes les unités répondantes dans la cellule de pondération. Nous donnons à cette procédure d'ajustement des poids le nom de procédure d'*ajustement par le ratio*. L'estimateur (1.2) se réduit à l'estimateur ajusté par le ratio

$$\hat{Y}_{\text{ratio}} = \sum_{c=1}^{C} \frac{\hat{X}_c}{\hat{X}_{rc}} \hat{Y}_{rc}. \tag{1.5}$$

Notons que l'estimateur ajusté par la fréquence (1.4) est un cas particulier de l'estimateur ajusté par le ratio quand $x_i = 1$ pour toutes les unités de population échantillonnées.

Enfin, si $\mathbf{x}_i = \mathbf{z}_i = (\delta_{i1}, ..., \delta_{ic}, ..., \delta_{iC}, \delta_{i1} x_i, ..., \delta_{ic} x_i, ..., \delta_{iC} x_i)'$, nous obtenons un autre cas particulier de (1.2). Dans ce cas, le facteur d'ajustement g_i donné par (1.3) se réduit à

$$g_i = \hat{N}_c \left[1 + (\bar{x}_c - \bar{x}_{rc}) \frac{(x_i - \bar{x}_{rc})}{\sum_{i \in s} r_i \, \delta_{ic} \, (x_i - \bar{x}_{rc})^2} \right],$$

si l'unité i appartient à la classe c, où $\bar{x}_c = \hat{X}_c / \hat{N}_c$ et $\bar{x}_{rc} = \hat{X}_{rc} / \hat{N}_{rc}$. Nous donnons à cette procédure d'ajustement des poids le nom de procédure d'*ajustement par régression linéaire simple*. L'estimateur (1.2) se réduit à l'estimateur ajusté par régression linéaire simple

$$\hat{Y}_{\text{regls}} = \sum_{c=1}^{C} \hat{N}_c \, [\hat{Y}_{rc} + (\bar{x}_c - \bar{x}_{rc}) \, \hat{B}_{rc}], \tag{1.6}$$

où

$$\hat{B}_{rc} = \frac{\sum_{i \in s} d_i \, r_i \, \delta_{ci} (x_i - \bar{x}_{rc})(y_i - \bar{y}_{rc})}{\sum_{i \in s} d_i \, r_i \, \delta_{ic} (x_i - \bar{x}_{rc})^2}.$$

Les estimateurs (1.4) à (1.6) reposent sur une certaine forme d'ajustement de la pondération dans les classes. Tous sont asymptotiquement sans biais pour Y si les probabilités de réponse des unités sont égales à l'intérieur des classes (c'est-à-dire que le mécanisme de non-réponse est uniforme à l'intérieur des classes). Cette condition est un cas particulier de la condition (i) discutée plus haut.

Dans le présent article, nous montrons que l'estimateur de variance jackknife simplifié, dans lequel les facteurs d'ajustement sont traités comme étant fixes, a tendance à surestimer la variance réelle de $\hat{Y}_{\text{CAL}}$, au moins dans certains cas simples. Nous prolongeons les travaux de recherche antérieurs de Thompson et Yung (2006) qui ont établi les expressions de la version par linéarisation des estimateurs de variance jackknife complet ainsi que simplifié, et évalué ces expressions empiriquement en utilisant des données provenant de l'Annual Capital Expenditures Survey (ACES) réalisée par le U.S. Census Bureau. Il est intéressant de noter que, dans le contexte de la pondération par la propension à la non-réponse, Kim et Kim (2007) ont montré que traiter les probabilités estimées de réponse comme si elles étaient fixes produit une surestimation de la variance réelle quand les poids d'échantillonnage ne sont pas utilisés dans l'estimation de ces probabilités. Beaumont (2005) a obtenu des résultats comparables dans le contexte de l'imputation quand les probabilités de réponse sont estimées en utilisant un modèle de régression logistique.

À la section 2, nous discutons des estimateurs de variance jackknife complet et simplifié, et montrons que l'estimateur simplifié est asymptotiquement biaisé. À la section 3, nous évaluons la gravité de ce biais pour deux plans d'échantillonnage utilisés fréquemment. À la section 4, nous présentons les résultats d'une étude par simulation comparant les estimateurs de variance jackknife complet et simplifié. Enfin, nous concluons par certaines observations générales à la section 5.

2. Estimation jackknife de la variance

Habituellement, l'estimation de la variance dans le contexte de la non-réponse est effectuée en utilisant le cadre à deux phases, qui consiste à voir la non-réponse comme une deuxième phase d'échantillonnage. Ici, nous considérons plutôt le cadre inversé qui a été proposé par Fay (1991) et perfectionné par Shao et Steel (1999). Ce cadre,

qui offre une base théorique pour l'étude des propriétés des estimateurs jackknife de variance, peut être décrit comme il suit : premièrement, en appliquant le mécanisme de non-réponse, la population U est divisée aléatoirement en une population de répondants U_r et une population de non-répondants U_m. Puis, sachant (U_r, U_m), l'échantillon aléatoire s est sélectionné conformément au plan d'échantillonnage choisi. La variance totale de $\hat{Y}_{\mathrm{CAL}}$ peut être exprimée sous la forme

$$V(\hat{Y}_{\mathrm{CAL}}) = E_q V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r}) + V_q E_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r}), \qquad (2.1)$$

où $E_p(.)$ et $V_p(.)$ désignent l'espérance et la variance par rapport au plan d'échantillonnage et $E_q(.)$ et $V_q(.)$ désignent l'espérance et la variance par rapport au mécanisme de non-réponse, $q(\mathbf{r} \mid \mathbf{I})$.

À la présente section, nous nous penchons sur l'échantillonnage aléatoire simple stratifié, qui est le plan habituellement utilisé dans les enquêtes-entreprises. Sous ce plan d'échantillonnage, la population U est partitionnée en L strates $U_1, ..., U_L$ de taille $N_1, ..., N_L$, respectivement. Un échantillon aléatoire simple sans remise s_h, de taille n_h, est tiré de la strate h, $h = 1, ..., L$. Chaque échantillon de strate est sélectionné indépendamment et nous supposons que $n_h \geq 2$ pour tout h. Dans ce contexte, le poids de sondage de l'unité i dans la strate h est $d_{hi} = N_h/n_h$. Un estimateur de variance jackknife complet de $\hat{Y}_{\mathrm{CAL}}$ sous échantillonnage aléatoire simple stratifié s'obtient de la façon suivante :

i) supprimer l'unité (gj) de l'échantillon, $g = 1, ..., L$; $j = 1, ..., n_h$;

ii) ajuster les poids de sondage d_{hi} pour obtenir les poids jackknife $d_{hi(gj)}$, où $d_{hi(gj)}$ est donné par

$$d_{hi(gj)} = \begin{cases} 0 & \text{si } (hi) = (gj) ; \\ \dfrac{n_g}{n_g - 1} d_{gi} & \text{si } h = g,\ i \neq j \ ; \\ d_{hi} & \text{autrement} \end{cases}$$

iii) calculer l'estimateur $\hat{Y}_{\mathrm{CAL}(gj)}$ de la même façon que $\hat{Y}_{\mathrm{CAL}}$ avec les poids jackknife $d_{hi(gj)}$ au lieu des poids de sondage d_{hi} ; c'est-à-dire $\hat{Y}_{\mathrm{CAL}(gj)} = \sum_{(hi)\in s} w_{hi(gj)}\, r_{hi}\, y_{hi}$, où $w_{hi(gj)} = d_{hi(gj)}\, g_{hi(gj)}$ avec $g_{hi(gj)} = 1 + (\hat{\mathbf{X}}_{\pi(gj)} - \hat{\mathbf{X}}_{r(gj)})' \hat{\mathbf{T}}_{r(gj)}^{-1} \mathbf{z}_{hi}$, $\hat{\mathbf{X}}_{\pi(gj)} = \sum_{i\in s} d_{hi(gj)} \mathbf{x}_{hi}$, $\hat{\mathbf{X}}_{r(gj)} = \sum_{(hi)\in s} d_{hi(gj)} r_{hi} \mathbf{x}_{hi}$ et $\hat{\mathbf{T}}_{r(gj)} = \sum_{(hi)\in s} d_{hi(gj)}\, r_{hi}\, \mathbf{z}_{hi}\, \mathbf{x}'_{hi}$;

iv) replacer l'unité supprimée à l'étape (i) dans l'échantillon ;

v) répéter les étapes (i) à (iv) pour toutes les unités (gj), $g = 1, ..., L$; $j = 1, ..., n_h$.

Notons que les facteurs d'ajustement pour la non-réponse g_{hi} sont recalculés dans chaque réplique. Nous obtenons ainsi l'estimateur de variance jackknife complet

$$v_{JF} = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j\in s_h} (\hat{Y}_{\mathrm{CAL}(gj)} - \hat{Y}_{\mathrm{CAL}})^2. \qquad (2.2)$$

L'estimateur de variance v_{JF} est un estimateur du premier terme du deuxième membre de (2.1), $E_q V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$. Ce terme représente la variance sous le plan de sondage que nous aurions obtenue si les unités répondantes avaient été sélectionnées selon un plan d'échantillonnage aléatoire simple stratifié avec remise, ou, de façon équivalente, si les fractions d'échantillonnage dans les strates, (n_h / N_h) étaient négligeables. Autrement dit, l'estimateur de variance jackknife complet (2.2) est un estimateur de la variance d'échantillonnage conditionnellement au vecteur des indicateurs de réponse $\mathbf{r}$. Par conséquent, v_{JF} est asymptotiquement sans biais et convergent pour $E_q V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$ sous échantillonnage aléatoire simple stratifié avec remise, indépendamment de la validité des hypothèses sous-jacentes. Notons que, puisque v_{JF} est l'estimateur d'une variance d'échantillonnage, nous pouvons l'obtenir facilement en utilisant un logiciel conçu pour l'estimation jackknife de la variance en présence de données complètes. En d'autres termes, aucun logiciel spécialisé n'est nécessaire. Mentionnons aussi que le deuxième terme du membre de droite de (2.1), $V_q E_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$, n'est pas pris en compte. Donc, l'estimateur de variance jackknife complet ne reflète pas le deuxième terme de (2.1). Toutefois, la contribution de ce terme à la variance totale est négligeable si les fractions d'échantillonnage dans les strates, n_h / N_h, sont négligeables. Donc, v_{JF} est asymptotiquement sans biais et convergent pour la variance totale, $V(\hat{Y}_{\mathrm{CAL}})$. C'est-à-dire que $E_{pq}(v_{JF}) \approx V(\hat{Y}_{\mathrm{CAL}})$. Puisque l'objectif de l'étude est de comparer les estimateurs jackknife complet et simplifié, dans la suite de l'exposé, nous supposons que les fractions d'échantillonnage dans les strates sont négligeables et nous nous concentrons sur les estimations des totaux, de sorte que nous pouvons omettre l'estimation du deuxième terme de (2.1). Nous constatons que, même si le deuxième terme n'est pas négligeable, nos comparaisons sont valides, car l'estimateur complet et l'estimateur simplifié produisent tous deux une sous-estimation de la variance totale qui correspond au même terme.

Un estimateur de variance jackknife simplifié de $\hat{Y}_{\mathrm{CAL}}$ est donné par

$$v_{JS} = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j\in s_h} (\hat{Y}^*_{\mathrm{CAL}(gj)} - \hat{Y}_{\mathrm{CAL}})^2, \qquad (2.3)$$

où $\hat{Y}^*_{\mathrm{CAL}(gj)} = \sum_{(hi)\in s} d_{hi(gj)}\, g_{hi}\, r_{hi}\, y_{hi}$. Notons que les facteurs d'ajustement des poids pour la non-réponse g_{hi} ne sont pas recalculés dans chaque réplique jackknife. Autrement dit, les facteurs g_{hi} sont traités comme des constantes, ce qui est inapproprié puisqu'ils dépendent de l'échantillon et de l'ensemble de répondants. Par conséquent, en général,

nous avons $E_{pq}(v_{JS}) \neq V(\hat{Y}_{\mathrm{CAL}})$, et l'estimateur de variance simplifié, v_{JS}, est biaisé.

Pour étudier l'ampleur du biais de v_{JS}, nous considérons la différence entre les deux estimateurs jackknife de la variance, $D = v_{JS} - v_{JF}$. Puisque l'estimateur de variance v_{JF} est un estimateur asymptotiquement sans biais du terme $V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$, il est asymptotiquement équivalent à un estimateur de variance obtenu en utilisant un développement en série de Taylor de premier degré. L'estimateur de variance résultant, désigné par $\tilde{v}_{JF}$, est l'estimateur de variance jackknife par linéarisation étudié par Yung et Rao (2000). De même, l'estimateur de variance jackknife simplifié v_{JS} est asymptotiquement équivalent à un estimateur de $V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$ obtenu en traitant les facteurs d'ajustement des poids pour la non-réponse g_{hi} comme des constantes. Nous désignons cet estimateur de variance par $\tilde{v}_{JS}$. La quantité D peut donc être approximée par $\tilde{D} = \tilde{v}_{JS} - \tilde{v}_{JF}$. Pour que cette approximation soit valide, nous supposons que le nombre de répondants est grand.

En notant que $\mathrm{Biais}(v_{JF}) = E_{pq}(v_{JF}) - V(\hat{Y}_{\mathrm{CAL}}) \approx 0$, il s'ensuit que le biais de v_{JS}, $\mathrm{Biais}(v_{JS}) = E_{pq}(v_{JS}) - V(\hat{Y}_{\mathrm{CAL}})$, peut être approximé par $E_{pq}(D) \approx E_{pq}(\tilde{D})$. Soit $v(y)$ l'estimateur de variance de l'estimateur sur données complètes (1.1). En utilisant un développement en série de Taylor de premier degré, nous pouvons montrer qu'un estimateur de $V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$ est donné par

$$\tilde{v}_{JF} = v(\hat{\xi}) \tag{2.4}$$

où

$$\hat{\xi}_{hi} = \mathbf{x}'_{hi}\,\hat{\mathbf{B}}_r + g_{hi}\,r_{hi}\,e_{hi},$$

avec $e_{hi} = (y_{hi} - \mathbf{x}'_{hi}\hat{\mathbf{B}}_r)$ et $\hat{\mathbf{B}}_r = \hat{\mathbf{T}}_r^{-1}\sum_{(hi)\in s} d_{hi}\,r_{hi}\,\mathbf{z}_{hi}\,y_{hi}$. Par ailleurs, traiter les facteurs g_{hi} comme des constantes implique que $\hat{Y}_{\mathrm{CAL}}$ est linéaire en les poids de sondage d_{hi}. Il s'ensuit que $\tilde{v}_{JS}$ est donné par

$$\tilde{v}_{JS} = v(\psi), \tag{2.5}$$

où $\psi_{hi} = g_{hi}\,r_{hi}\,y_{hi}$.

Par exemple, pour un plan d'échantillonnage à taille fixe ou aléatoire, un estimateur de variance possible est

$$\tilde{v}_{JF} = \sum_{i\in s}\sum_{j\in s} \Delta_{ij}\,\hat{\xi}_i\,\hat{\xi}_j,$$

où $\Delta_{ij} = (\pi_{ij} - \pi_i\pi_j)/\pi_{ij}\pi_i\pi_j$ et π_{ij} est la probabilité d'inclusion de deuxième ordre des unités i et j. Notons que $\pi_{ii} = \pi_i$. De même, nous avons

$$\tilde{v}_{JS} = \sum_{i\in s}\sum_{j\in s} \Delta_{ij}\,\psi_i\,\psi_j.$$

3. Biais de v_{JS} dans certains cas particuliers

3.1 Échantillonnage aléatoire simple sans remise

À la présente section, nous supposons que l'échantillon s a été sélectionné selon un plan d'échantillonnage aléatoire simple sans remise. Nous supposons également que la fraction d'échantillonnage n/N est négligeable et que le nombre de répondants r est grand. Enfin, nous supposons qu'il n'existe qu'une seule classe de pondération. Bien que la situation susmentionnée ne soit pas réaliste en pratique, elle donne une certaine idée du biais asymptotique de v_{JS}.

Dans le cas de l'estimateur ajusté par le ratio (1.5), nous pouvons montrer que $\tilde{D}$ est donné approximativement par

$$\begin{aligned}\tilde{D} = \frac{N^2}{r}\left(1-\frac{r}{n}\right)\Bigg\{&\left(\frac{\bar{x}}{\bar{x}_r}\right)^2 (s_{yr}^2 - s_{er}^2) \\ &+ 2\left(\frac{\bar{x}}{\bar{x}_r}\right)\hat{R}_r\left[\left(\frac{\bar{x}}{\bar{x}_r}\right)-1\right]\frac{s_{exr}}{n} \\ &+ \hat{R}_r^2\left[\left(\frac{\bar{x}}{\bar{x}_r}\right)^2 s_{xr}^2 - s_x^2\right] + \left(\frac{\bar{x}}{\bar{x}_r}\right)^2 \bar{y}_r^2\Bigg\},\end{aligned} \tag{3.1}$$

où $(\bar{x}_r, \bar{y}_r) = 1/r\sum_{i\in s} r_i\,(x_i, y_i)$ désigne la moyenne des répondants pour les variables x et y respectivement, et r est le nombre de répondants, $\hat{R}_r = \bar{y}_r/\bar{x}_r$, $s_{xr}^2 = 1/(r-1)\sum_{i\in s} r_i\,(x_i - \bar{x}_r)^2$, $s_x^2 = 1/(n-1)\sum_{i\in s}(x_i - \bar{x})^2$ avec $\bar{x} = 1/n\sum_{i\in s} x_i$, $s_{er}^2 = 1/(r-1)\sum_{i\in s} r_i\,(y_i - \hat{R}_r x_i)^2$ et $s_{exr} = 1/(r-1)\sum_{i\in s} r_i\,(y_i - \hat{R}_r x_i)\,x_i$. Si nous supposons en outre que toutes les unités ont la même probabilité de réponse (c'est-à-dire un mécanisme de réponse uniforme), nous avons $\bar{x}/\bar{x}_r \xrightarrow{p} 1$ et $s_{xr}^2/s_x^2 \xrightarrow{p} 1$. Dans ce cas, le biais asymptotique de v_{JS} est donné par

$$\begin{aligned}\mathrm{Biais}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \frac{N^2}{E_{pq}(r)}\left(1 - E_{pq}\left(\frac{r}{n}\right)\right) \\ &\quad S_y^2\left(\frac{1}{\mathrm{CV}(y)^2} + 2\frac{\mathrm{CV}(x)}{\mathrm{CV}(y)}\rho_{xy} - \frac{\mathrm{CV}(x)^2}{\mathrm{CV}(y)^2}\right),\end{aligned} \tag{3.2}$$

où $\mathrm{CV}(x) = S_x/\bar{X}$ et $\mathrm{CV}(y) = S_y/\bar{Y}$ désignent les coefficients de variation de population pour les variables x et y, respectivement avec $S_y^2 = 1/(N-1)\sum_{i\in U}(y_i - \bar{Y})^2$ et $\bar{Y} = 1/N\sum_{i\in U} y_i$, S_x^2 et $\bar{X}$ sont définis de la même façon, et ρ_{xy} désigne le coefficient de corrélation en population finie pour les variables x et y. De (3.2) il découle que le biais asymptotique de v_{JS} est non négatif si, et uniquement si

$$B_0 < \frac{\bar{Y}}{2}\left(\frac{1+\mathrm{CV}(x)^2}{\mathrm{CV}(x)^2}\right) \tag{3.3}$$

à condition que $0 < E_{pq}(r/n) < 1$, où $B_0 = \bar{Y} - B_1\bar{X}$ est l'ordonnée à l'origine en population finie de la droite des moindres carrés lorsque l'on fait la régression de y sur x avec

$$B_1 = \frac{\sum_{i\in U}(x_i - \bar{X})\,(y_i - \bar{Y})}{\sum_{i\in U}(x_i - \bar{X})^2}.$$

L'expression (3.2) montre clairement que le biais de v_{JS} augmente si i) le taux de réponse prévu $E_{pq}(r/n)$ diminue, ii) ρ_{xy} augmente, iii) $\mathrm{CV}(y)$ diminue ou iv) $\mathrm{CV}(x)$ augmente. En outre, il découle de (3.3) que v_{JS} surestime la variance réelle quand l'ordonnée à l'origine B_0 n'est pas trop grande. Le tableau 1 illustre la relation entre $\mathrm{CV}(x)$ et la condition (3.3). Par exemple, quand $\mathrm{CV}(x) = 0$, v_{JS} surestime toujours la variance réelle puisque, dans ce cas, la condition (3.3) se réduit à la condition $B_0 < \infty$, qui est toujours satisfaite. Ce résultat n'est pas étonnant, car si $\mathrm{CV}(x) = 0$, les valeurs de x sont toutes égales et l'estimateur ajusté par le ratio (1.5) est identique à l'estimateur ajusté par la fréquence (1.4). Comme nous le discutons plus loin, v_{JS} surestime toujours la variance réelle dans ce cas. Quand $\mathrm{CV}(x)$ est grand (par exemple $\mathrm{CV}(x) = 2$), v_{JS} surestime la variance réelle si, et uniquement si, $B_0 < 0{,}625\bar{Y}$. Cette dernière condition est satisfaite si l'ordonnée à l'origine n'est pas « trop loin » de l'origine. Par conséquent, si la relation entre y et x passe par l'origine (c'est-à-dire si le modèle du ratio est vérifié), l'estimateur de variance simplifié surestimera la variance réelle. Par contre, si l'estimateur ajusté par le ratio est utilisé quand le modèle du ratio n'est pas vérifié, par exemple quand $B_0 \geq 0{,}625\bar{Y}$, l'estimateur de variance simplifié v_{JS} sous-estimera la variance réelle. En conclusion, nous pouvons nous attendre à ce que v_{JS} surestime la variance réelle quand nous utilisons une méthode d'ajustement par le ratio, à moins que le modèle du ratio soit extrêmement mal spécifié pour les données disponibles, ce qui pourrait se produire, par exemple, si les variables y et x sont négativement corrélées.

Tableau 1
Relation entre $\mathrm{CV}(x)$ et la condition (3.3)

$\mathrm{CV}(x)$	$\frac{\bar{Y}}{2}\left(\frac{1+\mathrm{CV}(x)^2}{\mathrm{CV}(x)^2}\right)$
0	∞
0,1	50,5 $\bar{Y}$
0,5	2,5 $\bar{Y}$
1	2 $\bar{Y}$
1,5	0,722 $\bar{Y}$
2	0,625 $\bar{Y}$

Si nous considérons maintenant l'estimateur ajusté par la fréquence (1.4), en posant $x_i = 1$ pour tout i dans (3.1), nous obtenons

$$\tilde{D} = \frac{N^2}{r}\left(1 - \frac{r}{n}\right)\bar{y}_r^2. \qquad (3.4)$$

Il découle de (3.4) que le biais relatif de v_{JS}, $\mathrm{BR}(v_{JS}) = \mathrm{Biais}(v_{JS})/V(\hat{Y}_{\mathrm{CAL}})$, peut être approximé par $E_{pq}(R\tilde{D})$ où $R\tilde{D} = \tilde{D}/\tilde{v}_{JF}$. Sous un mécanisme de non-réponse uniforme, de simples opérations algébriques mènent à

$$\mathrm{BR}(v_{JS}) \approx E_{pq}(R\tilde{D}) \approx \left(1 - E_{pq}\left(\frac{r}{n}\right)\right)\frac{1}{\mathrm{CV}(y)^2}. \qquad (3.5)$$

L'expression (3.5) montre que, dans le cas de l'estimateur ajusté par la fréquence (1.4), v_{JS} surestime toujours la variance réelle. L'ampleur de la surestimation augmente à mesure que le taux de réponse prévu $E_{pq}(r/n)$ diminue ou que $\mathrm{CV}(y)$ diminue. Par exemple, si le taux de réponse prévu est égal à 70 % et que $\mathrm{CV}(y) = 1$, nous avons $E_{pq}(R\tilde{D}) = 1{,}3$, de sorte que la valeur de l'estimateur de variance jackknife simplifié, v_{JS}, est en moyenne 30 % plus élevée que la variance réelle de $\hat{Y}_{\mathrm{CAL}}$. Par ailleurs, si le taux de réponse est égal à 70 % et que $\mathrm{CV}(y) = 0{,}5$, nous avons $E_{pq}(R\tilde{D}) = 5{,}3$, auquel cas la surestimation est considérable.

Enfin, penchons-nous sur le cas de l'estimateur ajusté par régression linéaire simple (1.6). Sous un mécanisme de non-réponse uniforme, nous pouvons montrer que le biais asymptotique de v_{JS} est donné par

$$\begin{aligned}\mathrm{Biais}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \frac{N^2}{E_{pq}(r)}\left(1 - E_{pq}\left(\frac{r}{n}\right)\right) \\ &\quad S_y^2\left(\frac{1}{\mathrm{CV}(y)^2} + \rho_{xy}^2\right) \geq 0. \qquad (3.6)\end{aligned}$$

De (3.6), il découle que v_{JS} surestime toujours la variance réelle dans le cas de l'estimateur ajusté par régression linéaire simple (1.6). Le biais (3.6) augmente si i) le taux de réponse prévu diminue, ii) ρ_{xy}^2 augmente ou iii) $\mathrm{CV}(y)$ diminue.

3.2 Échantillonnage aléatoire simple stratifié : les classes de pondération coïncident avec les strates

À la présente section, nous supposons que les classes de pondération coïncident avec les strates originales du plan de sondage. Cette situation n'est pas inhabituelle en pratique, surtout dans les enquêtes-entreprises. Si les strates sont telles que les unités qu'elles contiennent ont à peu près la même propension à répondre (c'est-à-dire réponse uniforme à l'intérieur de la strate), les expressions pour le biais de v_{JS} s'obtiennent facilement à partir des expressions (3.2), (3.4) et (3.6).

Dans le cas de l'estimateur ajusté par le ratio, l'expression (3.2) peut être étendue facilement au cas de l'échantillonnage aléatoire simple stratifié pour obtenir

$$\begin{aligned}\text{Biais}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \sum_{h=1}^{L} \frac{N_h^2}{E_{pq}(r_h)}\left(1 - E_{pq}\left(\frac{r_h}{n_h}\right)\right) \\ &\quad S_{yh}^2\left(\frac{1}{\text{CV}_h(y)^2} + 2\frac{\text{CV}_h(x)}{\text{CV}_h(y)}\rho_{hxy} - \frac{\text{CV}_h(x)^2}{\text{CV}_h(y)^2}\right), \quad (3.7)\end{aligned}$$

où les quantités r_h, $\text{CV}_h(x)$, $\text{CV}_h(y)$, S_{yh}^2 et ρ_{hxy} correspondent à r, $\text{CV}(x)$, $\text{CV}(y)$, S_y^2 et ρ_{xy} calculés dans chaque strate.

Pour l'estimateur ajusté par la fréquence, l'expression (3.4) peut être étendue facilement au cas de l'échantillonnage aléatoire simple stratifié pour obtenir

$$\begin{aligned}\text{Biais}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \sum_{h=1}^{L} \frac{N_h^2}{E_{pq}(r_h)}\left(1 - E_{pq}\left(\frac{r_h}{n_h}\right)\right)\frac{S_{yh}^2}{\text{CV}_h(y)^2}. \quad (3.8)\end{aligned}$$

Enfin, pour l'estimateur ajusté par régression linéaire simple, l'expression (3.6) peut être étendue facilement au cas de l'échantillonnage aléatoire simple stratifié pour obtenir

$$\begin{aligned}\text{Biais}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \sum_{h=1}^{L} \frac{N_h^2}{E_{pq}(r_h)}\left(1 - E_{pq}\left(\frac{r_h}{n_h}\right)\right) \\ &\quad S_{yh}^2\left(\frac{1}{\text{CV}_h(y)^2} + \rho_{hxy}^2\right). \quad (3.9)\end{aligned}$$

Des expressions (3.7) à (3.9), il découle qu'il faut faire preuve d'une certaine prudence lorsque l'on utilise l'estimateur de variance jackknife simplifié. En effet, même si le biais de cet estimateur est faible dans chaque strate, la somme de ces biais au niveau de la population peut être considérable s'ils ont tous la même direction.

4. Étude par simulation

Une étude par simulation nous a permis de comparer les propriétés statistiques des estimateurs de variance jackknife simplifié et complet sous diverses conditions. Nous avons généré cinq populations stratifiées différentes contenant chacune 30 000 unités et deux variables. D'abord, nous avons tiré les valeurs de x d'une loi gamma dont les paramètres étaient α et λ. Puis, sachant les valeurs de x, nous avons généré les valeurs de y selon le modèle suivant :

$$y_{hi} = \beta_0 + \beta_1 x_{hi} + \varepsilon_{hi},$$

où $\varepsilon_{hi} \sim N(0, \sigma_{\varepsilon h}^2)$. Nous avons fixé les valeurs de la variance et de $\sigma_{\varepsilon h}^2$ de telle façon que le coefficient de corrélation (désigné par ρ_{xy}) entre x_{hi} et y_{hi} soit égal à 0,7 dans toutes les populations. Chaque population a été stratifiée en trois strates contenant chacune 10 000 unités. Les paramètres des populations simulées sont présentés au tableau 2.

La population 1 correspond très bien au modèle du ratio avec une ordonnée à l'origine nulle dans toutes les strates. La population 2 possède un terme d'ordonnée à l'origine non négligeable dans chacune des trois strates. La population 3 est une combinaison des populations 1 et 2, où le modèle du ratio est bien ajusté pour les strates 2 et 3, mais non pour la strate 1. La population 4 est semblable à la population 1, excepté que les unités des strates 1 et 2 ont 70 % de chance de déclarer une valeur nulle. Cette population est destinée à imiter la situation de l'Annual Capital Expenditures Survey (ACES) du U.S. Census Bureau, qui a motivé la présente étude. L'ACES emploie un estimateur de variance jackknife simplifié qui, selon les études empiriques, donne des résultats qui s'approchent de ceux de l'estimateur de variance jackknife complet. Sa population est caractérisée par de nombreuses valeurs nulles déclarées pour les dépenses en immobilisations par la majorité des petites et moyennes entreprises échantillonnées, la majorité des dépenses déclarées étant fournie par les grandes entreprises. Nous avons généré la population 5 afin de montrer que, pour l'estimateur ajusté par le ratio, l'estimateur simplifié peut effectivement présenter un biais négatif quand le modèle du ratio est spécifié incorrectement (démontré dans l'expression (3.3) pour un échantillonnage aléatoire simple). Pour cette population, le terme d'ordonnée à l'origine est fortement significatif dans toutes les strates.

Tableau 2
Paramètres de population

Population	β_0 (À l'intérieur de la strate)			β_1 (À l'intérieur de la strate)			α	λ	CV(x)	CV(y)
	1	2	3	1	2	3				
1	0	0	0	2	4	6	4	5	50 %	76 %
2	120	240	360	2	4	6	4	5	50 %	44 %
3	120	0	0	2	4	6	4	5	50 %	51 %
4	0	0	0	2	4	6	4	5	50 %	134 %
5	50	200	300	0,5	1	2	4	5	200 %	63 %

De chaque population, nous avons tiré 5 000 échantillons aléatoires simples stratifiés de taille 300 (100 unités par strate). Dans chaque échantillon, nous avons produit une non-réponse en utilisant un mécanisme de réponse uniforme à l'intérieur de chaque strate, avec les probabilités de réponse égales à 60 % dans la strate 1, 70 % dans la strate 2

et 90 % dans la strate 3. Ce profil de réponse n'est pas inhabituel dans les enquêtes-entreprises où un suivi plus important est effectué pour les unités de moyenne et de grande taille (strates 2 et 3).

Pour chaque échantillon, nous avons calculé les estimateurs ajustés par la fréquence et par le ratio, donnés respectivement par (1.4) et (1.5), en utilisant les strates comme classes de pondération. Nous avons estimé la variance des estimateurs ponctuels au moyen de v_{JF} et de v_{JS}, donnés respectivement par (2.2) et (2.3). Comme mesure du biais d'un estimateur de variance v, nous avons utilisé le biais relatif Monte Carlo en pourcentage donné par

$$\mathrm{BR}_{\mathrm{MC}}(v) = \frac{1}{5\,000}\sum_{t=1}^{5\,000}\frac{v^{(t)} - \mathrm{EQM}_{\mathrm{MC}}(\hat{Y}_{\mathrm{CAL}})}{\mathrm{EQM}_{\mathrm{MC}}(\hat{Y}_{\mathrm{CAL}})}\times 100,$$

où $v^{(t)}$ est l'estimation de la variance obtenue à partir du t^{e} échantillon, et $\mathrm{EQM}_{\mathrm{MC}}(\hat{Y}_{\mathrm{CAL}})$ est l'erreur quadratique moyenne (EQM) Monte Carlo définie par

$$\mathrm{EQM}_{\mathrm{MC}}(\hat{Y}_{\mathrm{CAL}}) = \frac{1}{50\,000}\sum_{t=1}^{50\,000}(\hat{Y}_{\mathrm{CAL}}^{(t)} - Y)^2,$$

où $\hat{Y}_{\mathrm{CAL}}^{(t)}$ est l'estimation (corrigée par le ratio ou par la fréquence) de Y pour le t^{e} échantillon. Le tableau 3 donne le biais relatif Monte Carlo en pourcentage pour les estimateurs ajustés par la fréquence ainsi que par le ratio.

Tableau 3
Biais relatif Monte Carlo en pourcentage pour les estimateurs de variance jackknife simplifié et complet

Population	Estimateur ajusté par la fréquence		Estimateur ajusté par le ratio	
	$\mathrm{BR}_{\mathrm{MC}}(v_{JS})$	$\mathrm{BR}_{\mathrm{MC}}(v_{JF})$	$\mathrm{BR}_{\mathrm{MC}}(v_{JS})$	$\mathrm{BR}_{\mathrm{MC}}(v_{JF})$
1	57,3 %	1,1 %	80,5 %	-0,3 %
2	877,1 %	0,4 %	364,7 %	0,5 %
3	220,7 %	0,6 %	185,9 %	-0,2 %
4	21,6 %	0,6 %	29,1 %	1,4 %
5	266,4 %	0,2 %	-67,2 %	5,0 %

Comme prévu, l'estimateur simplifié surestime l'EQM Monte Carlo de l'estimateur ajusté par la fréquence pour toutes les populations. La surestimation varie d'environ 20 % pour la population 4 à plus de 800 % pour la population 2. L'expression (3.8) montre que le biais de v_{JS} dépend du taux de réponse et de $\bar{y}_h^2$. Pour la population 2, le terme d'ordonnée à l'origine est grand et accroît la valeur de $\mathrm{CV}_h(y)$ dans toutes les strates, ce qui à son tour accroît le biais de v_{JS}. La population 3 est similaire à la population 2, excepté que le terme d'ordonnée à l'origine n'est grand que pour la première strate. Comme prévu, le biais de v_{JS} dans cette population est compris entre ceux des populations 1 et 2. La population 4 est celle qui imite la population de l'ACES avec les valeurs remplacées par zéro pour certaines unités dans les strates 1 et 2. Le biais relatif Monte Carlo de 21,6 % émane, en grande partie, de la troisième strate où aucune unité n'a été remplacée par une valeur nulle (ce fait peut être constaté en utilisant l'expression (3.8)). Par comparaison, pour chacune des cinq populations, l'estimateur de variance jackknife complet suit très bien l'EQM Monte Carlo, le biais relatif absolu étant inférieur à 1,1 %.

Si nous examinons l'estimateur ajusté par le ratio, nous voyons que, de nouveau, l'estimateur de variance jackknife suit relativement bien l'EQM Monte Carlo pour toutes les populations, le biais relatif absolu étant inférieur à 5 %. Par contre, le biais relatif de l'estimateur simplifié varie de -67 % à 364 %. L'examen de l'expression (3.7) montre que, pour un taux de réponse fixe, le biais dépend de $\mathrm{CV}_h(y)$, $\mathrm{CV}_h(x)$ et ρ_{hxy}. Étant donné l'importance des termes d'ordonnée à l'origine dans la deuxième population, les valeurs de $\bar{y}_h$ sont grandes et les $\mathrm{CV}_h(y)$ correspondants sont plus faibles que pour les autres populations. Donc, le dernier terme de l'expression (3.7) est assez grand et le biais relatif résultant de v_{JS} est également grand. Nous faisons la même constatation pour la population 3, mais à un degré moindre, puisque seule la première strate possède un terme d'ordonnée à l'origine. Nous observons l'effet opposé dans la population 4, où l'introduction de valeurs nulles accroît significativement $\mathrm{CV}_h(y)$, ce qui à son tour réduit le biais relatif Monte Carlo en pourcentage de l'estimateur simplifié.

Des simulations supplémentaires ont été exécutées en utilisant certaines populations décrites au tableau 2, mais en faisant varier les taux de réponse. Les résultats ne sont pas présentés ici car ils correspondaient à ceux attendus. Autrement dit, le biais de l'estimateur simplifié diminuait à mesure que le taux de réponse augmentait (tous les autres paramètres étant maintenus constants). L'estimateur jackknife complet continuait de très bien suivre l'EQM Monte Carlo.

5. Conclusion

Dans le présent article, nous avons évalué théoriquement et empiriquement un estimateur de variance jackknife simplifié dans lequel les facteurs d'ajustement pour la non-réponse ne sont pas recalculés dans chaque réplique jackknife, en étudiant en particulier trois procédures distinctes d'ajustement de la pondération pour la non-réponse. Nous avons montré, dans le contexte de l'échantillonnage aléatoire simple stratifié, que l'estimateur de variance jackknife simplifié a tendance à surestimer la variance réelle des estimateurs. Toutefois, dans le contexte de la procédure d'ajustement par le ratio, cet estimateur pourrait sous-estimer la variance réelle si le modèle du ratio n'est pas adapté aux données dont on dispose.

L'une des justifications de l'utilisation d'une procédure simplifiée dans une méthode d'estimation de la variance par rééchantillonnage consiste à gagner du temps et à économiser les ressources informatiques. S'il s'agit de considérations vraiment importantes et que le programme obtient systématiquement des taux de réponse élevés pour les unités dans toutes les cellules de pondération, alors que la production de répliques de la procédure d'ajustement des poids présente des avantages théoriques manifestes, les avantages pratiques pourraient être peu nombreux, voir inexistants. Cela dit, les conditions d'équivalence « pratique » entre les estimateurs de variance par les méthodes complète et simplifiée sont extrêmement contraignantes et nous avons démontré que de faibles variations des conditions relatives aux données sous-jacentes peuvent facilement donner lieu à une violation de ces conditions d'équivalence. Si le jackknife pose réellement un problème en ce qui concerne les ressources informatiques, les auteurs recommandent l'approche de l'estimation jackknife par linéarisation de la variance, qui a les mêmes propriétés asymptotiques que le jackknife complet, mais dont les calculs sont rapides et dont le temps inactif du système est « gratuit » (en ce qui concerne la mise en mémoire des répliques). Voir Thompson et Yung (2006) pour des expressions de l'estimateur de variance jackknife par linéarisation pour les estimateurs ajustés par la fréquence ainsi que par le ratio. Étant donné ces alternatives viables, nous déconseillons d'utiliser un estimateur de variance à procédure simplifiée.

Remerciements

Le présent rapport est publié en vue d'informer les parties intéressées des travaux de recherche en cours et de favoriser la discussion de ces travaux. Toutes opinions exprimées concernant les problèmes statistiques, méthodologiques ou opérationnels sont celles des auteurs et ne reflètent pas forcément celles du U.S. Census Bureau. Les auteurs remercient le rédacteur associé, deux examinateurs anonymes, Samson Adeshiyan, Patrick Cantwell, Carol Caldwell, Michael Hidiroglou, Rita Petroni, Mark Sands et Jun Shao de leurs commentaires constructifs concernant des versions antérieures du présent article. Les travaux de David Haziza ont été financés par des bourses du Conseil de recherches en sciences naturelles et en génie du Canada.

Bibliographie

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society B*, 67, 445-458.

Brick, M.J., et Montaquila, J.M. (2009). Nonresponse and weighting. Dans le *Handbook of Statistics*, *Sample Surveys*: *Design*, *Methods and Applications*, (Éds., C.R. Rao et D. Pfeffermann), 29A, 163-185.

Da Silva, D.N., et Opsomer, J.D. (2006). A kernel smoothing method to adjust for unit nonresponse in sample surveys. *Canadian Journal of Statistics*, 34, 563-579.

Deville, J.-C. (2002). La correction de la non-réponse par calage généralisé. *Actes des Journées de Méthodologie Statistique*, Insee.

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

Eltinge, J.L., et Yansaneh, I.S. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. *Techniques d'enquête*, 23, 37-45.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference*, *US Bureau of the Census*, 429-440.

Kalton, G., et Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.

Kim, J.K., et Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 501-514.

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

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

Ozcoskun, L., Thompson, K.J. et Williams, Q. (2005). Investigation of balanced repeated replication (BRR) variance estimation for the Survey of Residential Alterations and Repairs (SORAR). *Proceedings of the Federal Committee on Statistical Methodology*, Office of Management and Budget.

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

Sautory, O. (2003). CALMAR 2 : une nouvelle version du programme CALMAR de redressement d'échantillon par calage. *Recueil* : *Symposium 2003*, *Défis reliés à la réalisation d'enquêtes pour la prochaine décennie*, Ottawa, Canada.

Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 93, 254-265.

Thompson, K.J. (2005). An empirical investigation into the effects of replicate reweighting on variance estimates for the annual capital expenditures survey. *Proceedings of the Federal Committee on Statistical Methodology*, Office of Management and Budget.

Thompson, K.J., et Yung, W. (2006). To Replicate (A weight adjustment procedure) or not to replicate? An analysis of the variance estimation effects of a shortcut procedure using the stratified jackknife. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3772-3779.

Valliant, R. (2004). The effect of multiple weighting steps on variance estimation. *Journal of Official Statistics*, 20, 1-18.

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, 903-915.