Estimation de la variance par le bootstrap avec remise pour les enquêtes auprès des ménages. Principes, exemples et mise en oeuvre

Pascal Bessonneau, Gwennaëlle Brilhaut, Guillaume Chauvet, Cédric Garcia

Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Estimation de la variance par le bootstrap avec remise pour les enquêtes auprès des ménages Principes, exemples et mise en œuvre

par Pascal Bessonneau, Gwennaëlle Brilhaut, Guillaume Chauvet et Cédric Garcia

Date de diffusion : le 6 janvier 2022

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, décembre 2021
Vol. 47, n° 2, pp. 339-375
Statistique Canada, n° 12-001-X au catalogue

Estimation de la variance par le bootstrap avec remise pour les enquêtes auprès des ménages
Principes, exemples et mise en œuvre

Pascal Bessonneau, Gwennaëlle Brilhaut, Guillaume Chauvet et Cédric Garcia[1]

Résumé

L'estimation de la variance est un problème difficile dans les enquêtes, car plusieurs facteurs non négligeables contribuent à l'erreur d'enquête totale, notamment l'échantillonnage et la non-réponse totale. Initialement conçue pour saisir la variance des statistiques non triviales à partir de données indépendantes et identiquement distribuées, la méthode bootstrap a depuis été adaptée de diverses façons pour tenir compte des éléments ou facteurs propres à l'enquête. Dans l'article, nous examinons l'une de ces variantes, le bootstrap avec remise. Nous considérons les enquêtes auprès des ménages, avec ou sans sous-échantillonnage de personnes. Nous rendons explicites les estimateurs de la variance que le bootstrap avec remise vise à reproduire. Nous expliquons comment le bootstrap peut servir à tenir compte de l'effet de l'échantillonnage, du traitement de la non-réponse et du calage sur l'erreur d'enquête totale. Par souci de clarté, les méthodes proposées sont illustrées au moyen d'un exemple traité en fil rouge. Elles sont évaluées dans le cadre d'une étude par simulations et appliquées au Panel Politique de la Ville (PPV) français. Deux macros SAS pour exécuter les méthodes bootstrap sont également élaborées.

Mots-clés : Bootstrap; calage; estimation de variance; non-réponse totale.

1. Introduction

L'estimation de la variance est un problème difficile dans les enquêtes. Les poids finaux utilisés à l'étape de l'estimation comprennent plusieurs traitements statistiques, notamment la correction de la non-réponse totale et le calage, dont l'effet sur la variance doit être évalué. Le bootstrap est un instrument utile, qui permet de créer les poids dits bootstrap publiés avec l'ensemble de données de l'enquête. Ces poids peuvent servir à calculer de façon répétée la version bootstrap du paramètre d'intérêt, ce qui donne un estimateur de la variance ou un intervalle de confiance basés sur des simulations. L'intérêt pour les utilisateurs est le fait qu'aucune information autre que les poids bootstrap n'est requise pour l'estimation de la variance. En particulier, il n'est pas nécessaire de décrire de façon exhaustive le plan de sondage initial et le processus d'estimation, ce qui serait le cas dans le cadre d'une approche analytique où l'estimateur de la variance doit être mis au point. Ainsi, un même ensemble de poids bootstrap sert à obtenir l'estimation de la variance, que les paramètres d'intérêt soient des totaux, des médianes ou des coefficients de régression. Même si l'on dispose de la description complète du plan de sondage et du processus d'estimation, l'approche analytique pose des problèmes pour des paramètres importants pour lesquels l'estimation de la variance par linéarisation n'est pas simple; voir par exemple Shao (1994) pour les L-statistiques, et Shao et Rao (1993) pour les proportions de faible revenu.

La littérature sur le bootstrap dans l'échantillonnage d'enquête est abondante; on trouve par exemple des revues détaillées dans Rao et Wu (1988), Rao, Wu et Yue (1992), Shao et Tu (1995, chapitre 6), Davison et Hinkley (1997, section 3.7), Davison et Sardy (2007), Chauvet (2007) et Mashreghi, Haziza et

1. Pascal Bessonneau, Ined; Gwennaëlle Brilhaut, Ined; Guillaume Chauvet, Ensai (Irmar), Campus de Ker Lann, Bruz, France. Courriel : guillaume.chauvet@ensai.fr; Cédric Garcia, Université Gustave Eiffel.

Léger (2016). L'une de ces techniques est le dit *rescaled bootstrap* (bootstrap rééchelonné) proposé par Rao et Wu (1988), qui peut se résumer comme suit. Premièrement, à l'intérieur de chaque échantillon au premier degré S_h de taille n_h sélectionné dans la strate h, un échantillon aléatoire simple avec remise de taille m_h est sélectionné, ce qui donne les poids bootstrap initiaux. Ensuite, ces poids peuvent être rééchelonnés de façon à reproduire un estimateur de la variance sans biais pour l'estimation d'un total (cas linéaire). Comme l'expliquent Rao et Wu (1988), le bootstrap rééchelonné peut être appliqué à divers plans de sondage, y compris l'échantillonnage à deux degrés et l'échantillonnage avec ou sans remise au premier degré. Toutefois, il n'est pas facile de tenir compte de certaines caractéristiques pratiques des enquêtes, comme le traitement de la non-réponse totale. Cette question est examinée par Yeo, Mantel et Liu (1999) et Girard (2009). Un sujet connexe est traité dans Kim, Navarro et Fuller (2006), qui se penchent sur l'estimation de la variance par répliques pour un échantillonnage à deux phases.

L'application du bootstrap Rao-Wu dans le cas particulier où les tailles de rééchantillonnage sont $m_h = n_h - 1$ donne ce qu'on appelle bootstrap des unités primaires d'échantillonnage (UPE) ou bootstrap avec remise (McCarthy et Snowden, 1985). Le bootstrap avec remise est assez simple à mettre en œuvre, notamment parce qu'il suffit de rééchantillonner les unités primaires d'échantillonnage, et non les unités finales. La prise en compte du traitement de la non-réponse et du calage est assez naturelle, comme l'explique le présent article. Une des propriétés importantes de toute méthode bootstrap consiste à reproduire (au moins approximativement) un estimateur de la variance connu dans le cas linéaire, que nous appelons estimateur repère de la variance. Pour le bootstrap avec remise, il est possible d'énoncer précisément cet estimateur repère de la variance à toute étape de la méthode, ce qui est utile afin de comprendre le fonctionnement de la méthode pour évaluer l'erreur d'enquête totale. Le bootstrap avec remise donne une estimation prudente de la variance, en ce sens que la variance d'échantillonnage au premier degré est surestimée si les plans de sondage utilisés à l'intérieur des strates au premier degré sont plus efficaces que l'échantillonnage multinomial, ce que nous supposons vrai dans le présent article. Il s'agit donc d'une méthode prudente de production des intervalles de confiance. Le biais positif de l'estimateur de la variance bootstrap devrait être négligeable lorsque les taux de sondage au premier degré à l'intérieur des strates sont négligeables, ce qui est souvent le cas dans les enquêtes téléphoniques. De plus, si l'enquête est répétée au fil du temps, il est probable que la contribution de la variance due à l'échantillonnage au premier degré s'estompe, tandis que la variance attribuable à l'attrition et à la non-réponse totale augmente.

Notre article, qui porte sur le bootstrap avec remise, se veut axé sur les utilisateurs. C'est pourquoi nous ne proposons pas de modifications particulières du bootstrap avec remise. Nous expliquons plutôt comment appliquer cette méthode bootstrap pour tenir compte de l'échantillonnage, du traitement de la non-réponse et du calage et, ce faisant, quel est l'estimateur de la variance que nous cherchons à reproduire lors de l'estimation d'un total. Nous donnons des exemples en fil rouge pour illustrer comment les poids bootstrap sont calculés dans des cas simples. Deux macros SAS mettant en œuvre les méthodes bootstrap proposées sont présentées, évaluées au moyen d'une étude par simulations et illustrées sur un ensemble réel de données d'enquête, tirées du Panel Politique de la Ville.

Pour simplifier la présentation, notre terminologie est celle des enquêtes auprès des ménages, qui sont également la motivation première de notre article. Nous examinons deux cas : premièrement, quand un échantillon de ménages seulement est sélectionné et deuxièmement, quand un sous-échantillon de personnes est sélectionné dans les ménages sélectionnés. Malgré cette terminologie particulière, notre démarche est générale et peut être appliquée à toute autre situation où l'enquête est effectuée par sondage à un degré (premier cas) ou par sondage à deux degrés (deuxième cas).

Nous nous intéressons plus particulièrement aux enquêtes téléphoniques auprès des ménages, largement utilisées à l'Institut national d'études démographiques (Ined) français au cours des dernières décennies. À l'origine, un échantillon de numéros de téléphone était sélectionné dans un registre de numéros de téléphone fixes, et plus récemment, les numéros de téléphone utilisés dans l'enquête sont générés de façon aléatoire pour tenir compte des ménages non couverts dans les registres (numéros de téléphone non répertoriés et numéros de téléphone cellulaire). À la deuxième étape, des personnes sont sélectionnées au sein des ménages, au moyen de méthodes de sélection classiques (par exemple individu Kish). Les sondages téléphoniques ont prouvé leur efficacité, en particulier pour des sujets sensibles comme la sexualité, la violence ou les dépendances. Parmi les exemples d'enquêtes réalisées par l'Ined, citons l'enquête nationale sur les violences faites aux femmes en France (Enveff) en 2000, l'enquête Violences et rapports de genre en 2015 et 2018 (Virage et Virage Dom, respectivement), ou l'enquête nationale sur le contexte de la sexualité en France en 2006. Le même protocole sera probablement utilisé dans un proche avenir pour des enquêtes aux sujets similaires, comme l'enquête sur la sexualité des jeunes adultes ou celle sur le contrôle des naissances, qui doivent commencer entre 2021 et 2023.

L'article est organisé comme suit. À la section 2, nous définissons nos principales notations et nous considérons l'estimation d'un total en tenant compte de l'échantillonnage, de la non-réponse totale et du calage. Nous traitons à la section 2.1 la situation où un échantillon de ménages seulement est sélectionné (scénario à un degré) et à la section 2.2 le cas où des personnes sont sous-échantillonnées au sein des ménages (scénario à deux degrés). La méthode bootstrap de base est décrite à la section 3 : le scénario à un degré est examiné aux sections 3.1 et 3.2, et le scénario à deux degrés est examiné aux sections 3.3 et 3.4. Nous expliquons à la section 3.5 comment on peut appliquer la procédure bootstrap élémentaire pour obtenir un estimateur de la variance ou un intervalle de confiance. Les méthodes bootstrap proposées sont évaluées à la section 4 au moyen d'une étude par simulations. À la section 5, nous illustrons les méthodes au moyen d'un échantillon de ménages et d'individus du Panel Politique de la Ville français. Des conclusions sont données à la section 6. Les estimateurs repères de la variance pour l'échantillon de personnes sont présentés à l'annexe A. Le programme SAS qui a servi à réaliser l'estimation de la variance bootstrap est présenté aux annexes B et C. Ces programmes SAS peuvent être mis à disposition par l'auteur correspondant sur demande.

2. Notation et estimation

Dans la présente section, nous définissons nos principales notations et décrivons le processus d'échantillonnage et d'estimation. Nous examinons d'abord à la section 2.1 le cas où un échantillon de

ménages seulement est sélectionné, et nous décrivons le processus d'estimation qui comprend traitement de la non-réponse totale et calage. Nous indiquons dans chaque cas l'estimateur repère de la variance considéré, c'est-à-dire l'estimateur de la variance que nous cherchons à reproduire pour l'estimation d'un total au moyen de la méthode bootstrap proposée à la section 3. Le cas où des personnes sont sous-échantillonnées dans les ménages est traité à la section 2.2. Les estimateurs repères de la variance pour ce deuxième cas sont donnés à l'annexe A.

2.1 Scénario avec échantillon de ménages seulement

Nous considérons une estimation pour une population U_{hh} de ménages. Nous désignons par y_k la valeur prise par une variable d'intérêt pour le ménage k. Nous nous intéressons à l'estimation du total

$$Y_{hh} = \sum_{k \in U_{hh}} y_k. \tag{2.1}$$

2.1.1 Plan de sondage

Nous supposons qu'un échantillon S_{hh} est sélectionné dans U_{hh} au moyen d'un plan de sondage à un degré stratifié. La population U_{hh} est divisée en H strates $U_{hh}^1, \ldots, U_{hh}^H$, les échantillons $S_{hh}^1, \ldots, S_{hh}^H$ y sont sélectionnés indépendamment, et l'échantillon S_{hh} est l'union de ces échantillons. Soit π_k la probabilité d'inclusion d'un ménage donné k. Le poids de sondage est

$$d_k = \frac{1}{\pi_k}. \tag{2.2}$$

En cas de réponse complète, l'estimateur de Y_{hh} est

$$\hat{Y}_{hh} = \sum_{k \in S_{hh}} d_k y_k. \tag{2.3}$$

Nous considérons comme un estimateur repère de la variance

$$v_{\text{mult}}(\hat{Y}_{hh}) = \sum_{h=1}^{H} \left[\frac{n_h}{n_h - 1} \sum_{k \in S_{hh}^h} \left(d_k y_k - \frac{1}{n_h} \sum_{k' \in S_{hh}^h} d_{k'} y_{k'} \right)^2 \right], \tag{2.4}$$

avec n_h la taille de l'échantillon S_{hh}^h. Cet estimateur de la variance est sans biais si les échantillons sont sélectionnés dans les strates par échantillonnage multinomial (Tillé, 2011, section 5.4), aussi appelé échantillonnage avec remise. Il est conservatif si les plans d'échantillonnage utilisés dans les strates sont plus efficaces que l'échantillonnage multinomial (Särndal, Swensson et Wretman, 1992, section 4.6), ce que nous supposons vrai dans le reste de l'article. Le biais positif de cet estimateur de la variance devrait être négligeable quand les taux de sondage à l'intérieur des strates sont eux-mêmes négligeables, ce qui est souvent le cas dans les enquêtes téléphoniques. Les résultats de notre étude par simulations en témoignent, voir la section 4.

2.1.2 Traitement de la non-réponse

En pratique, l'échantillon S_{hh} est sujet à la non-réponse totale, ce qui mène à l'observation d'un sous-échantillon de répondants seulement. Nous désignons par r_k l'indicateur de réponse d'un ménage k et par p_k la probabilité de réponse du ménage k. Nous supposons que les ménages répondent indépendamment les uns des autres. De plus, nous supposons que la non-réponse totale est traitée par la méthode des groupes de réponse homogènes (GRH), populaire en pratique (par exemple Brick, 2013; Juillard et Chauvet, 2018). Dans ce cadre, on suppose que l'échantillon S_{hh} peut être divisé en C GRH notés $S_{1,hh},\ldots,S_{C,hh}$, de sorte que la probabilité de réponse p_k soit constante à l'intérieur d'un GRH.

Pour $c=1,\ldots,C$, p_c désigne la probabilité de réponse commune à l'intérieur du GRH $S_{c,hh}$. Elle est estimée par

$$\hat{p}_c = \frac{\sum_{k\in S_{c,hh}} \theta_k r_k}{\sum_{k\in S_{c,hh}} \theta_k}, \tag{2.5}$$

avec θ_k un certain poids donné au ménage k. Le choix $\theta_k = 1$ conduit à estimer p_c par le taux de réponse non pondéré dans le GRH. Le choix $\theta_k = d_k$ conduit à estimer p_c par le taux de réponse dans le GRH, pondéré par les poids d'échantillonnage (par exemple Kott, 2012).

La prise en compte des probabilités de réponse estimées donne les poids corrigés pour la non-réponse

$$d_{rk} = \frac{d_k}{\hat{p}_{c(k)}}, \tag{2.6}$$

avec $c(k)$ le GRH du ménage k. L'estimateur de Y_{hh} ajusté pour tenir compte de la non-réponse est

$$\hat{Y}_{r,hh} = \sum_{k\in S_{r,hh}} d_{rk} y_k. \tag{2.7}$$

Construit à partir de l'estimateur de la variance multinomial de (2.4) et de la linéarisation des estimateurs repondérés pour tenir compte de la non-réponse totale (Kim et Kim, 2007, section 2), notre estimateur repère de la variance est

$$v_{\text{mult}}(\hat{Y}_{r,hh}) = \sum_{h=1}^{H}\left[\frac{n_h}{n_h-1}\sum_{k\in S_{hh}^h}\left(d_k u_{1k} - \frac{1}{n_h}\sum_{k'\in S_{hh}^h} d_{k'} u_{1k'}\right)^2\right], \tag{2.8}$$

avec

$$u_{1k} = \theta_k \pi_k \bar{y}_{rc(k)} + \frac{r_k}{\hat{p}_{c(k)}}\left\{y_k - \theta_k \pi_k \bar{y}_{rc(k)}\right\},$$

et

$$\bar{y}_{rc} = \frac{\sum_{k \in S_{c,hh}} d_k r_k y_k}{\sum_{k \in S_{c,hh}} \theta_k r_k}.$$

Il s'agit d'un estimateur conservatif pour la variance asymptotique de $\hat{Y}_{r,hh}$. Une hypothèse essentielle à cet égard est que les indicateurs de réponse r_k sont mutuellement indépendants.

2.1.3 Calage

Enfin, les poids ajustés pour tenir compte de la non-réponse sont calés sur les totaux auxiliaires connus dans la population. Par souci de simplicité, nous décrivons seulement l'estimateur par la régression généralisée (GREG, Särndal et coll., 1992, chapitre 6). Soit x_k le vecteur des variables de calage au niveau du ménage, et X_{hh} le total sur la population U_{hh}. Pour l'échantillon $S_{r,hh}$, cela donne les poids linéaires calés

$$w_k = d_{rk}\left(1 + x_k^{\mathrm{T}} \lambda_{hh}\right),$$

avec

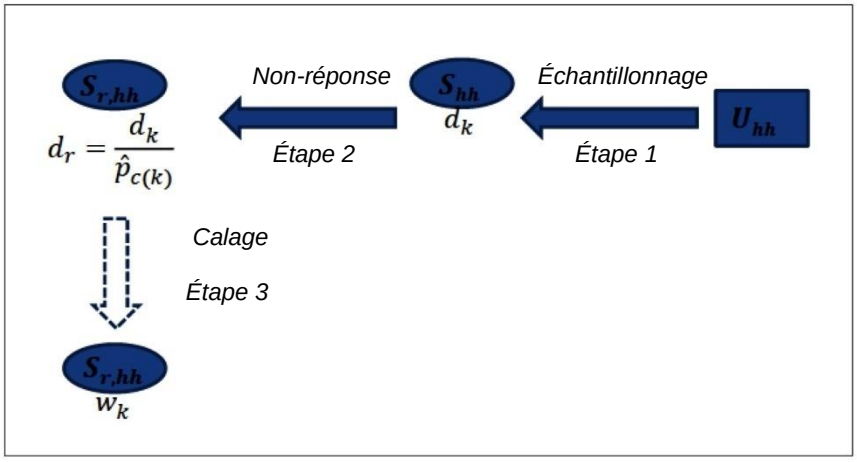
$$\lambda_{hh} = \left(\sum_{k \in S_{r,hh}} d_{rk} x_k x_k^{\mathrm{T}}\right)^{-1} \left(X_{hh} - \hat{X}_{r,hh}\right), \tag{2.9}$$

et où $\hat{X}_{r,hh}$ est l'estimateur de X_{hh}, obtenu en insérant x_k dans (2.7). L'estimateur calé est

$$\hat{Y}_{\mathrm{cal},hh} = \sum_{k \in S_{r,hh}} w_k y_k. \tag{2.10}$$

Les étapes d'échantillonnage et d'estimation sont résumées dans la figure 2.1.

Figure 2.1 Étapes d'échantillonnage et d'estimation pour un échantillon de ménage.



Construit à partir de la linéarisation des estimateurs repondérés pour tenir compte de la non-réponse totale et du calage (Kim et Kim, 2007, section 5), notre estimateur repère de la variance est

$$v_{\text{mult}}(\hat{Y}_{\text{cal},hh}) = \sum_{h=1}^{H}\left[\frac{n_h}{n_h-1}\sum_{k\in S_{hh}^h}\left(d_k u_{2k} - \frac{1}{n_h}\sum_{k'\in S_{hh}^h} d_{k'}u_{2k'}\right)^2\right], \tag{2.11}$$

avec

$$u_{2k} = \theta_k \pi_k \bar{e}_{rc(k)} + \frac{r_k}{\hat{p}_{c(k)}}\left\{e_k - \theta_k \pi_k \bar{e}_{rc(k)}\right\},$$

et

$$\bar{e}_{rc} = \frac{\sum_{k\in S_{c,hh}} d_k r_k e_k}{\sum_{k\in S_{c,hh}} \theta_k r_k},$$

où nous supposons que

$$e_k = y_k - \hat{B}_{r,hh}^{\mathrm{T}} x_k \quad \text{avec} \quad \hat{B}_{r,hh} = \left(\sum_{k\in S_{r,hh}} d_{rk} x_k x_k^{\mathrm{T}}\right)^{-1} \sum_{k\in S_{r,hh}} d_{rk} x_k y_k \tag{2.12}$$

désigne les résidus de régression estimés de la variable d'intérêt sur les variables de calage. Il s'agit d'un estimateur conservatif pour la variance asymptotique de $\hat{Y}_{\text{cal},hh}$.

2.1.4 Calcul des poids des ménages dans un exemple

Voici un petit exemple pour fixer les idées. Examinons une population U_{hh} composée de $N_{hh} = 100$ ménages. Nous supposons sans perte de généralité qu'une seule strate est utilisée et qu'un échantillon de $n_{hh} = 10$ ménages est sélectionné.

L'échantillon est $S = \{A, B, \ldots, J\}$. Les probabilités d'inclusion des unités sélectionnées sont (supposons)

$$\pi_A = \pi_B = \pi_C = \pi_D = \frac{1}{4} \quad \text{et} \quad \pi_E = \pi_F = \pi_G = \pi_H = \pi_I = \pi_J = \frac{1}{16}, \tag{2.13}$$

ce qui donne les poids de sondage

$$d_A = d_B = d_C = d_D = 4 \quad \text{et} \quad d_E = d_F = d_G = d_H = d_I = d_J = 16. \tag{2.14}$$

Parmi les 10 ménages sélectionnés, 7 seulement sont sondés en raison de la non-réponse. On en tient compte au moyen de la méthode des GRH, avec deux groupes : les unités A, B, F et J dans le premier, et les unités C, D, E, G, H et I dans le second. Les unités B, C et G sont des non-répondants. Dans chaque GRH, nous calculons les probabilités de réponse estimées, pondérées par les poids de sondage $(\theta_k = d_k)$. Cela donne

$$\hat{p}_1 = \frac{\sum_{k \in S_{1,hh}} d_k r_k}{\sum_{k \in S_{1,hh}} d_k} = \frac{d_A + d_F + d_J}{d_A + d_B + d_F + d_J} = \frac{9}{10},$$
$$\hat{p}_2 = \frac{d_D + d_E + d_H + d_I}{d_C + d_D + d_E + d_G + d_H + d_I} = \frac{13}{18}. \quad (2.15)$$

On obtient les poids tenant compte de la non-réponse pour les répondants en divisant les poids d'échantillonnage par les probabilités de réponse estimées. Cela donne les poids

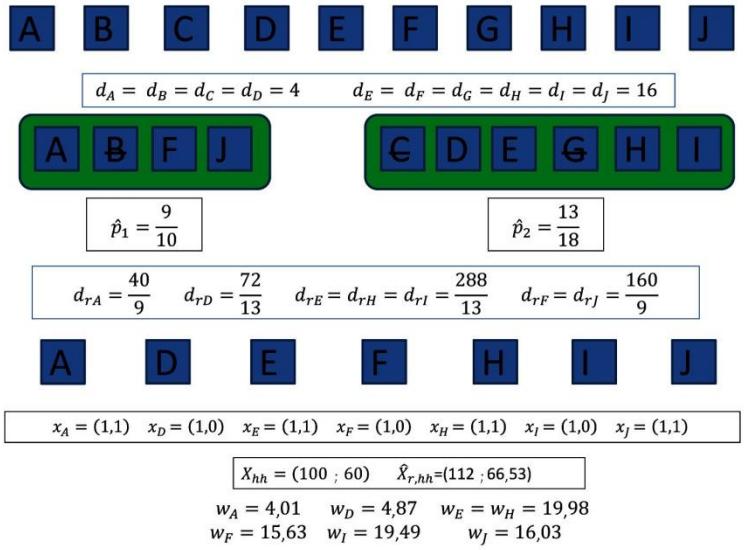
$$d_{rA} = \frac{40}{9} \quad d_{rD} = \frac{72}{13} \quad d_{rE} = d_{rH} = d_{rI} = \frac{288}{13} \quad d_{rF} = d_{rJ} = \frac{160}{9}. \quad (2.16)$$

Enfin, les poids sont calés pour qu'ils permettent de reproduire exactement la taille de la population $N_{hh} = 100$ et à un total auxiliaire $X_{1,hh} = 60$. Notons qu'en utilisant l'échantillon de répondants, nous obtenons $\hat{N}_{r,hh} = 112$ et $\hat{X}_{1r,hh} = 66{,}53$. Les poids calés sont

$$w_A = 4{,}01; \quad w_D = 4{,}87; \quad w_E = w_H = 19{,}98;$$
$$w_F = 15{,}63; \quad w_I = 19{,}49; \quad w_J = 16{,}03. \quad (2.17)$$

Les étapes d'échantillonnage et d'estimation sont résumées dans la figure 2.2.

Figure 2.2 Étapes d'estimation de la pondération des ménages.



2.2 Scénario avec échantillon de ménages et de personnes

Nous nous intéressons à la population U_{ind} de personnes associées à la population U_{hh} de ménages considérée à la section 2.1. Si nous désignons par y_l la valeur prise par une variable d'intérêt pour la personne l, le paramètre d'intérêt est

$$Y_{\text{ind}} = \sum_{l \in U_{\text{ind}}} y_l. \tag{2.18}$$

2.2.1 Plan de sondage

Dans tout ménage échantillonné $k \in S_{hh}$, un sous-échantillon $S_{\text{ind},k}$ de personnes est sélectionné, et l'échantillon S_{ind} est l'union de ces échantillons. Nous désignons par $\pi_{l|k}$ la probabilité d'inclusion conditionnelle de la personne l à l'intérieur du ménage k. Le poids de sondage conditionnel de l est

$$d_{l|k} = \frac{1}{\pi_{l|k}} \quad \text{pour tout} \quad l \in k, \tag{2.19}$$

et le poids de sondage non conditionnel est

$$d_l = d_{l|k} \times d_k \quad \text{pour tout} \quad l \in k. \tag{2.20}$$

En cas de réponse complète, l'estimateur de Y_{ind} est

$$\hat{Y}_{\text{ind}} = \sum_{k \in S_{hh}} d_k \sum_{l \in S_{\text{ind},k}} d_{l|k} y_l = \sum_{k \in S_{\text{ind}}} d_l y_l. \tag{2.21}$$

On obtient l'estimateur repère de la variance pour $\hat{Y}_{\text{ind}}$ à partir de (2.4), en remplaçant y_k par

$$\hat{y}_k = \sum_{l \in S_{\text{ind},k}} d_{l|k} y_l. \tag{2.22}$$

2.2.2 Traitement de la non-réponse

En tenant compte de la non-réponse des ménages, les poids des personnes sont

$$d_{rl} = d_{rk(l)} d_{l|k(l)} \quad \text{avec } k(l) \text{ le ménage contenant } l, \tag{2.23}$$

avec d_{rk} le poids du ménage k corrigé pour tenir compte de la non-réponse totale (voir l'équation (2.6)) et $d_{l|k}$ le poids d'échantillonnage conditionnel de la personne l dans le ménage k (voir l'équation (2.19)). Soit

$$S_{r,\text{ind}} = \bigcup_{k \in S_{r,hh}} S_{\text{ind},k} \tag{2.24}$$

l'ensemble de toutes les personnes échantillonnées à l'intérieur des ménages répondants.

Les personnes dans $S_{r,\text{ind}}$ sont elles-mêmes sujettes à la non-réponse, bien qu'habituellement, on s'attende à ce que ce soit dans une moindre mesure. Cela conduit à l'observation d'un sous-échantillon de répondants $S_{rr,\text{ind}}$ seulement. Nous désignons par r_l l'indicateur de réponse et par p_l la probabilité de réponse de la personne l. Nous supposons que les personnes répondent indépendamment les unes des autres. De plus, nous supposons que cette non-réponse est traitée au moyen de la méthode des GRH :

l'échantillon $S_{r,\text{ind}}$ peut être divisé en D GRH notés $S_{r1,\text{ind}},\ldots,S_{rD,\text{ind}}$ de telle sorte que la probabilité de réponse p_l soit constante à l'intérieur d'un GRH.

Pour $d=1,\ldots,D$, p_d désigne la probabilité de réponse commune à l'intérieur du GRH $S_{rd,\text{ind}}$. Elle est estimée par

$$\hat{p}_d = \frac{\sum_{l\in S_{rd,\text{ind}}}\theta_l r_l}{\sum_{l\in S_{rd,\text{ind}}}\theta_l}, \tag{2.25}$$

avec θ_l un certain poids donné à la personne l. Le choix $\theta_l=1$ conduit à estimer p_d par le taux de réponse non pondéré dans le GRH. Le choix $\theta_l=d_l$ conduit à estimer p_d par le taux de réponse dans le GRH, pondéré par les poids d'échantillonnage de la personne. Le choix $\theta_l=d_{rl}$ conduit à estimer p_d par le taux de réponse dans le GRH, pondéré par les poids d'échantillonnage de la personne corrigés pour tenir compte de la non-réponse totale du ménage. Nous comparons ces différents choix dans l'étude par simulations présentée à la section 4.

La prise en compte des probabilités de réponse estimées donne les poids individuels corrigés pour tenir compte de la non-réponse de la personne et du ménage.

$$d_{rrl} = \frac{d_{rl}}{\hat{p}_{d(l)}} \quad \text{avec } d(l) \text{ le ménage contenant } l. \tag{2.26}$$

L'estimateur de Y_{ind} ajusté pour tenir compte de la non-réponse de la personne et du ménage est

$$\hat{Y}_{rr,\text{ind}} = \sum_{l\in S_{rr,\text{ind}}} d_{rrl}y_l. \tag{2.27}$$

2.2.3 Calage

Soit z_l le vecteur des variables de calage au niveau de la personne, et Z_{ind} le total sur la population U_{ind}. Pour l'échantillon $S_{rr,\text{ind}}$, cela donne les poids linéaires calés

$$w_l = d_{rrl}\left(1+z_l^{\top}\lambda_{\text{ind}}\right),$$

avec

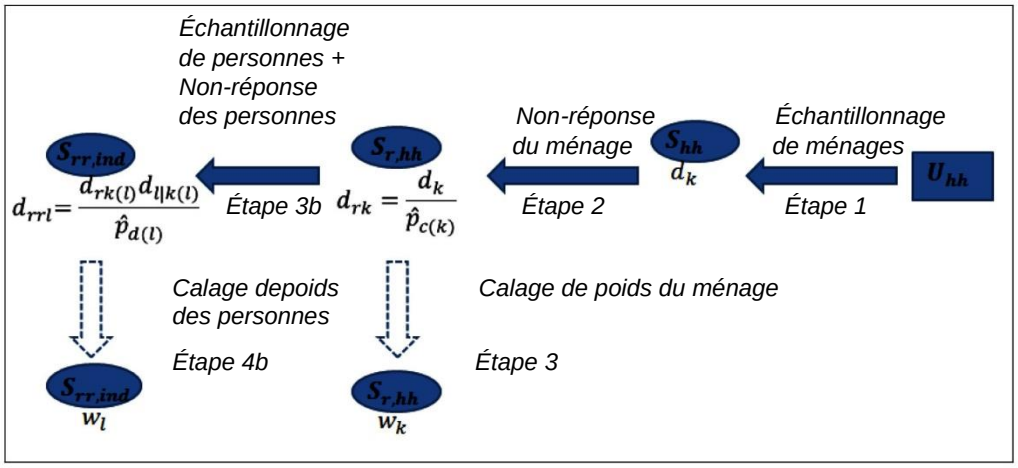
$$\lambda_{\text{ind}} = \left(\sum_{l\in S_{rr,\text{ind}}} d_{rrl}z_l z_l^{\top}\right)^{-1}\left(Z_{\text{ind}} - \hat{Z}_{rr,\text{ind}}\right), \tag{2.28}$$

et où $\hat{Z}_{rr,\text{ind}}$ est l'estimateur de Z_{ind}, obtenu en insérant z_l dans (2.27). L'estimateur calé est

$$\hat{Y}_{\text{cal,ind}} = \sum_{l\in S_{rr,\text{ind}}} w_l y_l. \tag{2.29}$$

Les étapes d'échantillonnage et d'estimation sont résumées dans la figure 2.3.

Figure 2.3 Étapes d'échantillonnage et d'estimation pour un échantillon de ménages avec sous-échantillonnage de personnes.



2.2.4 Calcul des poids des personnes dans un exemple

Nous poursuivons avec l'exemple présenté à la section 2.1.4. Rappelons que l'échantillon des ménages répondants est $S_{r,hh} = \{A, D, E, F, H, I, J\}$. L'ensemble de toutes les personnes à l'intérieur des ménages répondants se présente comme suit (supposons) :

$$\underbrace{(i_1, i_2, i_3)}_{A} \quad \underbrace{(i_4)}_{D} \quad \underbrace{(i_5, i_6)}_{E} \quad \underbrace{(i_7, i_8, i_9)}_{F} \quad \underbrace{(i_{10}, i_{11})}_{H} \quad \underbrace{(i_{12})}_{I} \quad \underbrace{(i_{13})}_{J}. \tag{2.30}$$

Nous supposons que le plan de sondage consiste à sélectionner exactement une personne à l'intérieur de chaque ménage. L'ensemble $S_{r,\text{ind}}$ de toutes les personnes échantillonnées à l'intérieur des ménages répondants est

$$S_{r,\text{ind}} = \{i_1, i_4, i_6, i_8, i_{11}, i_{12}, i_{13}\}. \tag{2.31}$$

À partir des équations (2.23) et (2.16), les poids des personnes corrigés pour tenir compte de la non-réponse des ménages sont par conséquent

$$d_{r1} = \frac{40}{3},\ d_{r4} = \frac{72}{13},\ d_{r6} = \frac{576}{13},\ d_{r8} = \frac{160}{3},\ d_{r11} = \frac{576}{13},\ d_{r12} = \frac{288}{13},\quad d_{r13} = \frac{160}{9}. \tag{2.32}$$

Parmi ces sept personnes sélectionnées, quatre seulement sont sondées en raison de la non-réponse, prise en compte au moyen de la méthode des groupes de réponse homogènes (GRH). Nous supposons qu'il y a deux GRH : les unités i_1, i_6, i_8 et i_{11} dans le premier, et les unités i_4, i_{12} et i_{13} dans le second. Les unités i_4, i_{11} et i_{13} sont des non-répondants. À l'intérieur de chaque GRH, nous calculons les probabilités de réponse estimées non pondérées $(\theta_l = 1)$. Cela donne

$$\hat{p}_1 = \frac{\sum_{l \in S_{r1,\text{ind}}} r_l}{\sum_{l \in S_{r1,\text{ind}}} 1} = \frac{3}{4},$$
$$\hat{p}_2 = \frac{\sum_{l \in S_{r2,\text{ind}}} r_l}{\sum_{l \in S_{r2,\text{ind}}} 1} = \frac{1}{3}. \tag{2.33}$$

On obtient les poids tenant compte de la non-réponse des personnes ou des ménages pour les répondants en divisant les poids de (2.32) par les probabilités de réponse estimées. Cela donne les poids

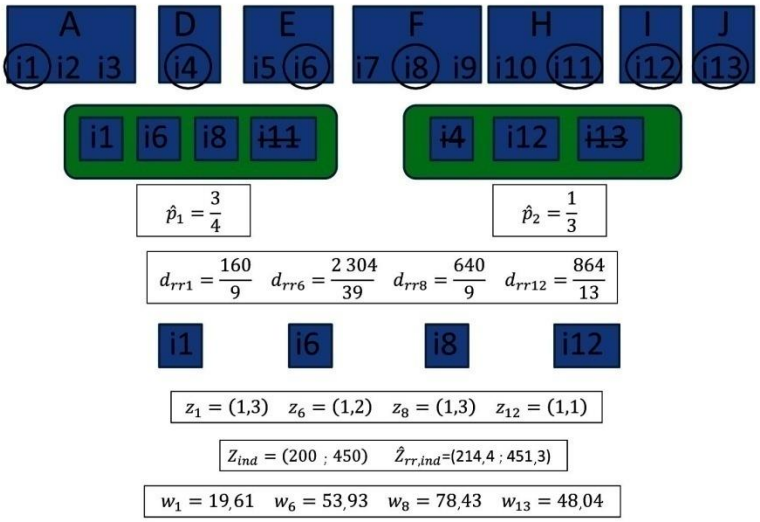
$$d_{rr1} = \frac{160}{9}, \quad d_{rr6} = \frac{2\,304}{39}, \quad d_{rr8} = \frac{640}{9}, \quad d_{rr12} = \frac{864}{13}. \tag{2.34}$$

Enfin, les poids sont calés pour qu'ils permettent de reproduire exactement la taille de la population $N_{\text{ind}} = 200$ et le total auxiliaire $Z_{1,\text{ind}} = 450$. Notons qu'en utilisant l'échantillon de répondants, nous obtenons $\hat{N}_{r,\text{ind}} = 214{,}4$ et $\hat{Z}_{1r,\text{ind}} = 451{,}3$. Les poids calés sont

$$w_1 = 19{,}61; \quad w_6 = 53{,}93; \quad w_8 = 78{,}43; \quad w_{13} = 48{,}04. \tag{2.35}$$

Les étapes d'échantillonnage et d'estimation sont résumées dans la figure 2.4.

Figure 2.4 Étapes d'estimation de la pondération des personnes.



3. Estimation de la variance bootstrap

Nous commençons à la section 3.1 par la description de l'étape élémentaire de la méthode bootstrap quand on sélectionne seulement un échantillon de ménages. Nous l'illustrons dans la section 3.2 sur l'exemple présenté à la section 2.1.4. La méthode bootstrap en cas d'échantillonnage de personnes à l'intérieur des ménages est décrite à la section 3.3, et elle est illustrée à la section 3.4. Dans la section 3.5,

nous expliquons comment l'étape élémentaire de la méthode bootstrap proposée sert à effectuer l'estimation de la variance et produire des intervalles de confiance.

3.1 Étape élémentaire du bootstrap pour les ménages

Au moyen du bootstrap avec remise, nous tirons d'abord à l'intérieur de l'échantillon initial S_{hh}^h sélectionné dans la strate U_{hh}^h un rééchantillon avec remise S_{hh*}^h de $n_h - 1$ ménages, avec probabilités égales. Notons que le rééchantillonnage est effectué sur l'unité d'échantillonnage (un ménage) plutôt que sur l'unité finale d'observation (une personne), ce qui est essentiel pour saisir correctement la variance due à l'échantillonnage. En particulier, cette méthode bootstrap permet de saisir la variance due à l'échantillonnage du second degré (sélection des personnes) sans rééchantillonner les unités finales dans le processus de bootstrap. Pour tout $k \in S_{hh}^h$, nous définissons le facteur d'ajustement de repondération

$$G_k = \frac{n_h}{n_h - 1} \times m_k, \tag{3.1}$$

avec m_k le nombre de fois que le ménage k est sélectionné dans le rééchantillon S_{hh*}^h, également appelé la multiplicité. Il faut noter qu'une unité $k \in S_{hh}^h$ peut ne pas apparaître dans le rééchantillon, auquel cas cette unité a une multiplicité nulle; un exemple est donné à la section 3.2. Les facteurs d'ajustement de la repondération G_k sont utilisés pour obtenir les poids bootstrap qui tiennent compte du plan d'échantillonnage, de la non-réponse totale et du calage, comme le décrit l'algorithme 1. Les étapes font référence à la figure 2.1. Le rééchantillonnage présenté dans l'algorithme 1 est ensuite répété B fois indépendamment pour l'estimation de la variance ou pour produire un intervalle de confiance, voir l'algorithme 3 à la section 3.5.

Algorithme 1. Calcul des poids bootstrap des ménages tenant compte de la non-réponse et du calage

- Étape 1 : nous tenons compte de l'échantillonnage des ménages en calculant, pour tout $k \in S_{hh}$, le poids d'échantillonnage bootstrap

$$d_{k*} = G_k d_k. \tag{3.2}$$

  La version bootstrap de l'estimateur en présence de réponse complète donnée dans (2.3) est

$$\hat{Y}_{hh*} = \sum_{k \in S_{hh}} d_{k*} y_k. \tag{3.3}$$

- Étape 2 : nous tenons compte de la non-réponse totale des ménages en calculant les probabilités estimées bootstrap à l'intérieur des GRH.

$$\hat{p}_{c*} = \frac{\sum_{k \in S_{c,hh}} G_k \theta_k r_k}{\sum_{k \in S_{c,hh}} G_k \theta_k}, \tag{3.4}$$

et nous calculons les poids bootstrap corrigés pour tenir compte de la non-réponse

$$d_{rk*} = \frac{d_{k*}}{\hat{p}_{c(k)*}}, \tag{3.5}$$

avec $c(k)$ le GRH contenant le ménage k. La version bootstrap de l'estimateur corrigé pour tenir compte de la non-réponse totale donnée dans (2.7) est

$$\hat{Y}_{r,hh*} = \sum_{k \in S_{r,hh}} d_{rk*} y_k. \tag{3.6}$$

- Étape 3 : nous tenons compte du calage en calant les poids d_{rk*} sur les totaux X_{hh}. Cela donne les poids bootstrap calés

$$w_{k*} = d_{rk*}\left(1 + x_k^{\mathrm{T}} \lambda_{hh*}\right), \tag{3.7}$$

avec

$$\lambda_{hh*} = \left(\sum_{k \in S_{r,hh}} d_{rk*} x_k x_k^{\mathrm{T}} \right)^{-1} \left(X_{hh} - \hat{X}_{r,hh*} \right)$$

et

$$\hat{X}_{r,hh*} = \sum_{k \in S_{r,hh}} d_{rk*} x_k.$$

La version bootstrap de l'estimateur calé donnée dans (2.10) est

$$\hat{Y}_{\mathrm{cal},hh*} = \sum_{k \in S_{r,hh}} w_{k*} y_k. \tag{3.8}$$

Le traitement de la non-réponse totale dans le processus bootstrap mérite quelques explications. Premièrement, notre approche est conditionnelle aux indicateurs de réponse r_k. Contrairement aux indicateurs d'appartenance de l'échantillon qui sont traités par bootstrap à l'étape 1 de l'algorithme 1, les indicateurs de réponse demeurent fixes dans le processus bootstrap. Cela est dû au fait que nous cherchons à reproduire un estimateur de la variance qui considère l'échantillon S_{hh} comme étant sélectionné avec remise, et que dans ce cas, il n'est pas nécessaire d'appliquer la technique bootstrap aux r_k. Deuxièmement, la prise en compte de la non-réponse à l'étape 2 de l'algorithme 1 est réalisée conditionnellement sur les GRH : nous n'appliquons pas de bootstrap au processus menant à la construction des GRH (sur le sujet, voir par exemple Girard, 2009; Haziza et Beaumont, 2017). Enfin, le bootstrap des probabilités de réponse tel qu'il est décrit dans l'équation (3.4) tient compte de l'estimation des probabilités de réponse p_c. En d'autres termes, nous utilisons dans chaque rééchantillonnage les

mêmes GRH que ceux déterminés à partir de l'échantillon, mais les ajustements pour tenir compte de la non-réponse dans les GRH sont basés sur le contenu du rééchantillonnage. Cela est illustré dans l'exemple présenté à la section 3.2. Si nous n'appliquons pas de bootstrap aux probabilités de réponse et que nous insérons directement dans l'équation (3.5) les probabilités estimées à l'origine $\hat{p}_c$, alors les probabilités de réponse sont traitées comme si elles étaient connues, ce qui entraîne habituellement une surestimation de la variance (Beaumont, 2005; Kim et Kim, 2007).

Discutons maintenant de l'estimation de la variance bootstrap pour les estimateurs calés, comme cela est fait à l'étape 3 de l'algorithme 1 où l'étape de calage est réalisée sur le total réel de la population X_{hh}. Si l'on suit le principe bootstrap selon lequel l'échantillon S_{hh} est à l'échantillon bootstrap S_{hh*} ce que la population U_{hh} est à l'échantillon S_{hh}, il pourrait sembler plus intuitif de caler plutôt les totaux estimés $\hat{X}_{hh}$ obtenus en insérant x_k dans l'équation (2.3). Les deux démarches semblent valides pour ce qui est de l'estimation de la variance bootstrap pour l'estimateur calé $\hat{Y}_{\text{cal},hh}$, mais les variables de calage x_k peuvent être sujettes à la non-réponse sur l'échantillon S_{hh}, ce qui rend l'estimateur $\hat{X}_{hh}$ impossible à calculer, alors que le total X_{hh} est connu à partir d'une source extérieure.

3.2 Exemple de calcul des poids bootstrap des ménages

Nous poursuivons avec l'exemple présenté à la section 2.1.4. On réalise le bootstrap en sélectionnant d'abord un rééchantillon de $n_{hh} - 1 = 9$ ménages, avec remise et probabilités égales, parmi les ménages initialement échantillonnés. Dans cet exemple, nous supposons que le ménage A est sélectionné trois fois, que le ménage G est sélectionné deux fois et que les ménages D, E, H et I sont sélectionnés une fois. Au moyen de l'équation (3.2), on obtient les poids d'échantillonnage bootstrap

$$d_{A*} = \frac{40}{3} \quad d_{D*} = \frac{40}{9} \quad d_{E*} = d_{H*} = d_{I*} = \frac{160}{9} \quad d_{G*} = \frac{320}{9}. \tag{3.9}$$

Les poids d'échantillonnage bootstrap sont corrigés pour tenir compte de la non-réponse de la même façon que dans la correction initiale de la non-réponse, au moyen des mêmes GRH et des probabilités estimées pondérées. Dans ce cas, le premier GRH contient uniquement l'unité A qui est un répondant, de sorte que $\hat{p}_{1*} = 1$. Le second GRH contient D, E, G (non-répondant), H et I. Cela donne

$$\hat{p}_{2*} = \frac{d_{D*} + d_{E*} + d_{H*} + d_{I*}}{d_{D*} + d_{E*} + d_{G*} + d_{H*} + d_{I*}} = \frac{13}{21}, \tag{3.10}$$

et les poids bootstrap corrigés pour tenir compte de la non-réponse

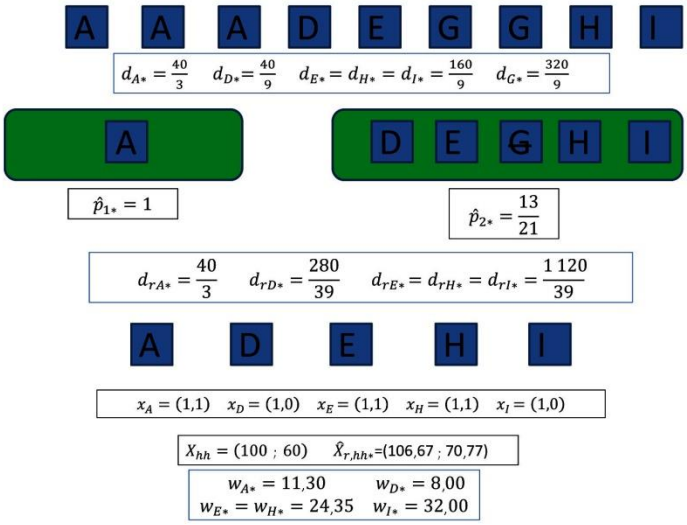
$$d_{rA*} = \frac{40}{3} \quad d_{rD*} = \frac{280}{39} \quad d_{rE*} = d_{rH*} = d_{rI*} = \frac{1\,120}{39}. \tag{3.11}$$

Enfin, les poids sont calés pour qu'ils soient appariés à la taille de la population $N_{hh} = 100$ et au total auxiliaire $X_{1,hh} = 60$. Cela donne les poids bootstrap calés

$$w_{A*} = 11{,}30 \quad w_{D*} = 8{,}00 \quad w_{E*} = w_{H*} = 24{,}35 \quad w_{I*} = 32{,}00. \tag{3.12}$$

Le calcul des poids bootstrap est résumé à la figure 3.1.

Figure 3.1 Calcul des poids bootstrap des ménages.



3.3 Calcul des poids bootstrap pour les personnes

Le calcul des poids bootstrap tenant compte du plan d'échantillonnage, de la non-réponse des ménages et des personnes et du calage est décrit dans l'algorithme 2. Les étapes font référence à la figure 2.3. En plus des étapes de bootstrap de l'algorithme 1, notons que l'algorithme 2 implique le calcul bootstrap des probabilités de réponse individuelles uniquement. Ajoutons que le sous-échantillonnage des personnes à l'intérieur des ménages n'a pas besoin d'être traité par bootstrap, comme nous l'indiquons à la section 3.1.

Algorithme 2. Calcul des poids individuels bootstrap tenant compte de la non-réponse des ménages, de la non-réponse des personnes et du calage

- Exécuter les étapes 1 et 2 de l'algorithme 1. Les poids bootstrap des ménages corrigés pour tenir compte de la non-réponse sont d_{rk}^{*}, selon l'équation (3.5).
- Étape 3b : nous tenons d'abord compte de l'échantillonnage des personnes en calculant les poids bootstrap individuels corrigés pour tenir compte de la non-réponse totale du ménage.

$$d_{rl*} = d_{rk(l)*} d_{l|k(l)} \quad \text{avec } k(l) \text{ le ménage contenant } l. \tag{3.13}$$

Nous tenons ensuite compte de la non-réponse totale des personnes. Nous calculons les probabilités estimées bootstrap à l'intérieur des GRH.

$$\hat{p}_{d*} = \frac{\sum_{l \in S_{rd,\text{ind}}} G_{k(l)} \theta_l r_l}{\sum_{l \in S_{rd,\text{ind}}} G_{k(l)} \theta_l}. \tag{3.14}$$

Nous calculons les poids bootstrap des personnes avec correction pour tenir compte de la non-réponse du ménage ou d'une personne, à savoir :

$$d_{rrl*} = \frac{d_{rl*}}{\hat{p}_{d(l)*}}, \tag{3.15}$$

avec $d(l)$ le GRH contenant la personne l. La version bootstrap de l'estimateur corrigé pour tenir compte de la non-réponse totale donnée dans (2.27) est

$$\hat{Y}_{rr,\text{ind}*} = \sum_{l \in S_{rr,\text{ind}}} d_{rrl*} y_l. \tag{3.16}$$

- Étape 4b : nous tenons compte du calage en calant les poids d_{rrl*} sur les totaux Z_{ind}. Cela donne les poids bootstrap calés

$$w_{l*} = d_{rrl*}\left(1 + z_l^{\top} \lambda_{\text{ind}*}\right), \tag{3.17}$$

avec

$$\lambda_{\text{ind}*} = \left(\sum_{k \in S_{rr,\text{ind}}} d_{rrl*} z_l z_l^{\top} \right)^{-1} \left(Z_{\text{ind}} - \hat{Z}_{rr,\text{ind}*} \right)$$

et

$$\hat{Z}_{rr,\text{ind}*} = \sum_{l \in S_{rr,\text{ind}}} d_{rrl*} z_l.$$

La version bootstrap de l'estimateur calé donnée dans (2.29) est

$$\hat{Y}_{\text{cal,ind}*} = \sum_{l \in S_{rr,\text{ind}}} w_{l*} y_l. \tag{3.18}$$

3.4 Exemple de calcul des poids bootstrap des personnes

Nous poursuivons avec l'exemple présenté à la section 3.2. L'échantillon bootstrap de ménages est constitué de A (trois fois), G (deux fois), et D, E, H et I (une fois). En raison de la non-réponse des ménages, nous observons A, D, E, H et I seulement. À partir de (2.30), on obtient l'échantillon bootstrap de personnes

$$S_{r,\text{ind}*} = \{i_1, i_4, i_6, i_{11}, i_{12}\}. \tag{3.19}$$

Les poids bootstrap des ménages corrigés pour tenir compte de la non-réponse totale sont donnés dans l'équation (3.11). À partir de l'équation (3.13), les poids bootstrap des personnes ajustés pour tenir compte de la non-réponse des ménages sont

$$d_{r1*} = 40 \quad d_{r4*} = \frac{280}{39} \quad d_{r6*} = \frac{2\ 240}{39} \quad d_{r11*} = \frac{2\ 240}{39} \quad d_{r12*} = \frac{1\ 120}{39}. \tag{3.20}$$

Ces poids bootstrap sont corrigés pour tenir compte de la non-réponse des personnes de la même façon que dans la correction initiale de la non-réponse individuelle, au moyen des mêmes GRH et des probabilités estimées non pondérées. Cependant, nous devons tenir compte dans ces probabilités de la multiplicité m_k et du facteur d'ajustement de la repondération G_k, voir l'équation (3.1). Dans notre cas, le premier GRH contient les personnes i_1, i_6 et i_{11}, et i_{11} est un non-répondant. La personne i_1 appartient au ménage A, qui a été sélectionné trois fois $(m_A = 3)$ dans l'échantillon bootstrap. La personne i_6 appartient au ménage E, et la personne i_{11} appartient au ménage H, qui ont tous deux été sélectionnés une fois dans l'échantillon bootstrap $(m_E = m_H = 1)$. Le calcul est semblable pour le second GRH et donne

$$\hat{p}_{1*} = \frac{G_A + G_E}{G_A + G_E + G_H} = \frac{4}{5},$$
$$\hat{p}_{2*} = \frac{G_I}{G_D + G_I} = \frac{1}{2}, \tag{3.21}$$

et les poids bootstrap des personnes corrigés pour tenir compte de la non-réponse du ménage ou de la personne sont donnés par

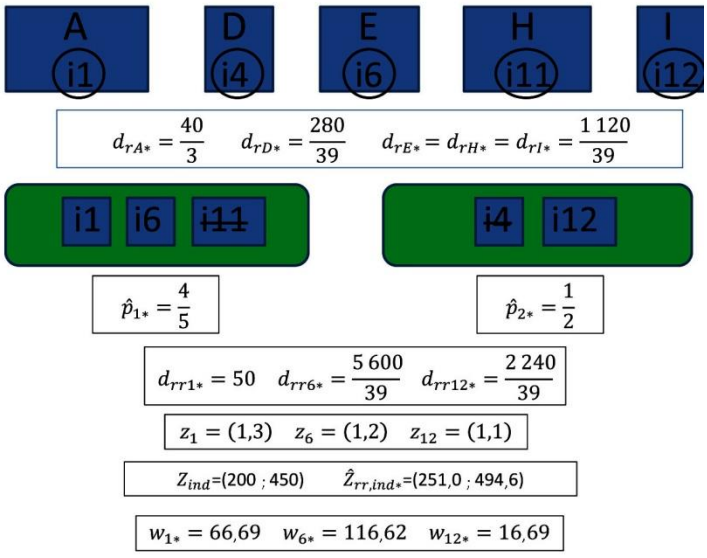
$$d_{rr1*} = 50 \quad d_{r6*} = \frac{5\ 600}{39} \quad d_{r12*} = \frac{2\ 240}{39}. \tag{3.22}$$

Enfin, les poids sont calés pour qu'ils permettent de reproduire exactement la taille de la population $N_{\text{ind}} = 200$ et le total auxiliaire $Z_{1,\text{ind}} = 450$. Cela donne les poids bootstrap calés

$$w_{1*} = 66{,}69 \quad w_{6*} = 116{,}62 \quad w_{12*} = 16{,}69. \tag{3.23}$$

Le calcul des poids bootstrap des personnes est résumé à la figure 3.2.

Figure 3.2 Calcul des poids bootstrap des personnes.



3.5 Estimation de la variance bootstrap et intervalles de confiance

Dans la présente section, nous nous intéressons aux paramètres qui peuvent être écrits comme des fonctions lisses de totaux. Nous expliquons comment l'étape élémentaire de la méthode bootstrap proposée sert à effectuer l'estimation de la variance et produire des intervalles de confiance. Par souci de concision, nous nous concentrons sur les paramètres définis sur la population de ménages U_{hh}. Le traitement des paramètres d'intérêt dans la population de personnes U_{ind} est semblable.

Supposons que y_k est un vecteur de taille q de variables d'intérêt, et que nous nous intéressons à un paramètre $\beta_{hh} = f(Y_{hh})$ utilisant une fonction connue et lisse $f: \mathbb{R}^q \to \mathbb{R}$. En cas de réponse complète, l'estimateur par substitution de β_{hh} est

$$\hat{\beta}_{hh} = f(\hat{Y}_{hh}), \tag{3.24}$$

voir, par exemple, Deville (1999). En cas de non-réponse totale au niveau du ménage, l'estimateur de β_{hh} corrigé pour tenir compte de la non-réponse totale est

$$\hat{\beta}_{r,hh} = f(\hat{Y}_{r,hh}), \tag{3.25}$$

et l'estimateur calé de β_{hh} est

$$\hat{\beta}_{\text{cal},hh} = f(\hat{Y}_{\text{cal},hh}). \tag{3.26}$$

Dans chaque cas, on obtient un estimateur de la variance bootstrap en appliquant un grand nombre de fois (disons B) l'étape de base de la méthode bootstrap dans l'algorithme 1, puis en calculant la dispersion des estimateurs bootstrap. Cela est résumé dans l'algorithme 3.

Algorithme 3. Estimation de la variance bootstrap pour l'estimation de la population des ménages

1. Répéter B fois la procédure bootstrap décrite dans l'algorithme 1. Soit $\hat{Y}^b_{hh*}$, $\hat{Y}^b_{r,hh*}$ et $\hat{Y}^b_{\text{cal},hh*}$ les estimateurs bootstrap des totaux calculés sur le b-ème échantillon. De plus, désignons par $\hat{\beta}^b_{hh*}$, $\hat{\beta}^b_{r,hh*}$ et $\hat{\beta}^b_{\text{cal},hh*}$ les estimateurs bootstrap associés à β_{hh}.
2. L'estimateur de la variance bootstrap pour $\hat{\beta}_{hh}$ est

$$\hat{V}_{\text{boot}}(\hat{\beta}_{hh}) = \frac{1}{B-1}\sum_{b=1}^{B}\left\{\hat{\beta}^b_{hh*} - \frac{1}{B}\sum_{b'=1}^{B}\hat{\beta}^{b'}_{hh*}\right\}^2, \tag{3.27}$$

et de même pour $\hat{\beta}_{r,hh}$ et $\hat{\beta}_{\text{cal},hh}$.

L'estimateur de la variance bootstrap peut servir à calculer un intervalle de confiance reposant sur la normalité avec un niveau ciblé $1-2\alpha$. Par exemple, l'intervalle de confiance quand on utilise l'estimateur en présence de réponse complète $\hat{\beta}_{hh}$ est

$$\mathrm{IC}_{\mathrm{nor}}(\beta_{hh}) = \left[\hat{\beta}_{hh} \pm u_{1-\alpha} \left\{ \hat{V}_{\mathrm{boot}}(\hat{\beta}_{hh}) \right\}^{0,5} \right], \tag{3.28}$$

avec $u_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la distribution normale standard. On s'attend à ce que cet intervalle de confiance soit conservatif, puisque la méthode bootstrap proposée l'est.

Nous examinons aussi les intervalles de confiance bootstrap (aussi dits élémentaires) dites du percentile et du percentile inverse. Ils peuvent être calculés directement à partir des poids bootstrap et sont par conséquent intéressants du point de vue des utilisateurs des données, contrairement aux méthodes nécessitant une grande puissance de calcul comme le bootstrap t (par exemple Davison et Hinkley, 1997; Shao et Tu, 1995). Pour $\hat{\beta}_{hh}$, l'intervalle de confiance percentile est obtenu au moyen de la distribution de $\hat{\beta}_{hh*}$ comme approximation de la distribution de $\hat{\beta}_{hh}$. Cette méthode utilise les estimations bootstrap ordonnées $\hat{\beta}_{hh*}^{(1)}, \ldots, \hat{\beta}_{hh*}^{(B)}$ pour former l'intervalle de confiance

$$\mathrm{IC}_{\mathrm{per}}(\beta_{hh}) = \left[\hat{\beta}_{hh*}^{(L)}, \hat{\beta}_{hh*}^{(U)} \right], \tag{3.29}$$

avec le niveau ciblé $1-2\alpha$, où $L = \alpha B$ et $U = (1-\alpha)B$. On obtient l'intervalle de confiance percentile inverse en considérant la distribution de $(\hat{\beta}_{hh*} - \hat{\beta}_{hh})$ comme une approximation de la distribution de $(\hat{\beta}_{hh} - \beta_{hh})$. Cela donne l'intervalle de confiance

$$\mathrm{IC}_{\mathrm{rev}}(\beta_{hh}) = \left[2\hat{\beta}_{hh} - \hat{\beta}_{hh*}^{(U)}, 2\hat{\beta}_{hh} - \hat{\beta}_{hh*}^{(L)} \right]. \tag{3.30}$$

Les propriétés de l'estimateur de la variance bootstrap et des trois intervalles de confiance sont évaluées dans l'étude par simulations effectuée à la section 4 pour l'estimation d'un total.

Le choix du nombre B de rééchantillonnages constitue un problème pratique important. Girard (2009) propose d'envisager plusieurs tailles de rééchantillonnage possibles (par exemple en augmentant B par un incrément de 100) et de représenter graphiquement les estimateurs de la variance bootstrap en fonction de B. La valeur pour laquelle cet estimateur de la variance commence à se stabiliser est alors retenue. Il s'agit d'une méthode simple, mais qui peut nécessiter une solution de compromis si différentes variables d'intérêt donnent différentes valeurs de stabilisation. Beaumont et Patak (2012) proposent de choisir B de telle sorte qu'avec une forte probabilité, la longueur de l'intervalle de confiance bootstrap donnée dans (3.28) soit proche de la longueur de l'intervalle de confiance obtenu au moyen d'un estimateur de la variance analytique. En supposant que, conditionnellement à l'échantillon initial, l'estimateur bootstrap normalisé du total est normalement distribué, ils établissent que la valeur B peut être déterminée à partir de la distribution d'une variable du chi deux (Beaumont et Patak, 2012, équation 10). Il est intéressant d'observer que la valeur obtenue ne dépend pas de la variable d'intérêt. À partir de ces résultats, ils proposent d'utiliser une valeur B qui ne soit pas inférieure à 750 et une valeur plus grande si l'hypothèse de normalité de l'estimateur bootstrap peut ne pas se vérifier. Nous avons utilisé $B = 1\,000$ dans l'étude par simulations présentée dans la section suivante. Pour les enquêtes devant répondre à plusieurs besoins analytiques – allant de paramètres de population simples à complexes et à diverses tailles de domaine – la

sélection d'au moins 1 000 répliques est la norme compte tenu des ressources informatiques disponibles à l'heure actuelle.

4. Étude par simulations

Afin d'évaluer la méthode bootstrap proposée, nous avons mené une étude par simulations sur une population artificielle. Nous générons d'abord une population U_{hh} contenant N_{hh} = 100 000 ménages, avec quatre variables auxiliaires $x_1, \ldots, x_4$ générées à partir d'une distribution gamma avec les paramètres de forme et d'échelle 2 et 5. À l'intérieur de la population, nous générons trois variables d'intérêt $y_1, \ldots, y_3$ conformément aux modèles suivants :

$$\begin{aligned} y_{1k} &= 10 + x_{1k} + x_{2k} + \sigma_\varepsilon \varepsilon_k, \\ y_{2k} &= 10 + x_{1k} + x_{3k} + \sigma_\varepsilon \varepsilon_k, \\ y_{3k} &= 10 + x_{3k} + x_{4k} + \sigma_\varepsilon \varepsilon_k, \end{aligned} \tag{4.1}$$

où ε_k est généré selon une distribution normale standard. Nous utilisons $\sigma_\varepsilon = 10$, ce qui donne un coefficient de détermination d'environ 0,50 pour chaque modèle. Les variables auxiliaires $1, x_{1k}, x_{2k}$ sont utilisées comme variables de calage au niveau du ménage dans cette étude par simulations. Les trois variables d'intérêt correspondent donc à des cas où le modèle de calage est correctement spécifié (y_1), en partie correctement spécifié (y_2), ou incorrectement spécifié (y_3). La population U_{hh} est répartie aléatoirement dans cinq groupes de réponse homogènes (GRH) de tailles égales. La probabilité de réponse p_c dans le GRH c est égale à 0,5 pour le premier groupe, à 0,6 pour le deuxième groupe, ..., et à 0,9 pour le cinquième groupe, ce qui donne un taux de réponse moyen de 70 % pour les ménages.

À l'intérieur de chaque ménage k, nous générons N_k individus, où $N_k - 1$ est généré selon une distribution de Poisson avec le paramètre 1, ce qui donne un nombre moyen de 2 individus par ménage. Dans la population correspondante U_{ind}, nous générons quatre variables auxiliaires $z_1, \ldots, z_4$ avec les paramètres de forme et d'échelle 2 et 0,5. De plus, nous générons trois variables d'intérêt y_4, y_5, y_6 selon les modèles suivants

$$\begin{aligned} y_{4l} &= 5 + 0{,}5z_{1l} + 0{,}5z_{2l} + \sigma_\eta \eta_l, \\ y_{5l} &= 5 + 0{,}5z_{1l} + 0{,}5z_{3l} + \sigma_\eta \eta_l, \\ y_{6l} &= 5 + 0{,}5z_{3l} + 0{,}5z_{4l} + \sigma_\eta \eta_l, \end{aligned} \tag{4.2}$$

où η_l est généré selon une distribution normale standard. Nous utilisons $\sigma_\eta = 0{,}4$, ce qui donne un coefficient de détermination d'environ 0,6 pour chaque modèle. Les variables auxiliaires $1, z_{1l}, z_{2l}$ sont utilisées comme variables de calage au niveau de la personne dans cette étude par simulations. Les trois variables d'intérêt correspondent donc à un cas où le modèle de calage est correctement spécifié (y_4), en partie correctement spécifié (y_5), ou incorrectement spécifié (y_6).

La population U_{ind} est divisée en cinq GRH comme suit. Les personnes qui sont seules dans leur ménage forment un GRH distinct, avec une probabilité de réponse de 1. La justification de ce choix est

que, dans ce cas, la personne équivaut en quelque sorte à son ménage et que la non-réponse est modélisée au niveau du ménage. Parmi le reste des personnes vivant dans un ménage k comprenant $N_k = 2$ personnes ou plus, les variables z_1 et z_2 servent à former quatre GRH de taille approximativement égale. La probabilité de réponse p_d varie de 0,80 à 0,95 dans les quatre GRH restants. Cela donne un taux de réponse global d'environ 90 % pour les personnes.

Dans la population U_{hh}, nous sélectionnons un échantillon S_{hh} de $n_{hh} = 1\,000$ ménages par échantillonnage aléatoire simple sans remise. Notons que comme le taux d'échantillonnage est faible (1 %), l'échantillonnage aléatoire simple n'est pas très différent qu'il soit avec ou sans remise, et le biais des estimateurs de la variance bootstrap devrait être faible dans cette configuration. La non-réponse est générée selon le modèle des GRH pour les ménages, ce qui donne un échantillon $S_{r,hh}$ de ménages répondants. Les probabilités de réponse estimées $\hat{p}_c$ sont obtenues à partir de l'équation (2.5), avec un poids égal $\theta_k = 1$. Dans chaque $k \in S_{r,hh}$, un individu Kish est sélectionné aléatoirement avec des probabilités égales, ce qui donne un échantillon de personnes $S_{r,\text{ind}}$. Dans $S_{r,\text{ind}}$, la non-réponse est générée selon le modèle des GRH pour les individus, ce qui donne un échantillon $S_{rr,\text{ind}}$ de personnes répondantes. Les probabilités de réponse estimées $\hat{p}_d$ sont obtenues à partir de l'équation (2.25), de trois manières possibles : soit avec des poids égaux $\theta_l = 1$, soit avec les poids d'échantillonnage $\theta_l = d_l$, soit avec les poids des personnes corrigés pour tenir compte de la non-réponse du ménage $\theta_l = d_{rl}$.

Les étapes d'échantillonnage et de non-réponse sont répétées $R = 1\,000$ fois. Sur chaque échantillon S_{hh}, nous calculons l'estimateur en présence de réponse complète donné dans (2.3), et sur chaque échantillon $S_{r,hh}$, nous calculons l'estimateur ajusté pour tenir compte de la non-réponse $\hat{Y}_{r,hh}$ donné dans (2.7) et l'estimateur $\hat{Y}_{\text{cal},hh}$ donné dans (2.10) avec l'ensemble des variables de calage $x_k = (1, x_{1k}, x_{2k})^\top$. Sur chaque échantillon $S_{rr,\text{ind}}$, nous calculons l'estimateur ajusté pour tenir compte de la non-réponse $\hat{Y}_{rr,\text{ind}}$ donné dans (2.27) et l'estimateur $\hat{Y}_{\text{cal,ind}}$ donné dans (2.29) avec l'ensemble des variables de calage $z_l = (1, z_{1l}, z_{2l})^\top$. Pour ces cinq estimateurs, nous calculons la racine de l'erreur quadratique moyenne normalisée

$$\text{REQMN}(\hat{Y}) = 100 \times \frac{\sqrt{\text{EQM}(\hat{Y})}}{Y}, \tag{4.3}$$

avec $\text{EQM}(\hat{Y})$ une approximation basée sur la simulation de l'erreur quadratique moyenne de $\hat{Y}$, obtenue à partir de l'exécution indépendante de 10 000 simulations.

Pour ces cinq estimateurs, nous calculons aussi les estimateurs de la variance bootstrap obtenus par l'application de l'algorithme 3 avec $B = 1\,000$. Pour mesurer le biais d'un estimateur de la variance $v(\hat{Y})$, nous utilisons le biais relatif Monte Carlo en pourcentage

$$\text{BR}\left\{v(\hat{Y})\right\} = 100 \times \frac{R^{-1}\sum_{c=1}^{R} v_c(\hat{Y}_c) - \text{EQM}(\hat{Y})}{\text{EQM}(\hat{Y})}, \tag{4.4}$$

où $v_c(\hat{Y}_c)$ représente l'estimateur de la variance dans le c -ème échantillon. Comme mesure de stabilité de $v(\hat{Y})$, nous utilisons la stabilité relative

$$\mathrm{SR}\left\{v(\hat{Y})\right\} = 100 \times \frac{\left[R^{-1}\sum_{c=1}^{R}\left\{v_c(\hat{Y}_c) - \mathrm{EQM}(\hat{Y})\right\}^2\right]^{1/2}}{\mathrm{EQM}(\hat{Y})}. \quad (4.5)$$

De plus, nous calculons les taux de couverture de l'intervalle de confiance associé au bootstrap par la méthode du percentile, la méthode du percentile inverse et à l'intervalle de confiance reposant sur la normalité, avec un taux d'erreur nominal unitaléral de 2,5 % dans chaque queue de distribution.

Les résultats sont présentés au tableau 4.1 pour l'estimation de la population des ménages. La racine de l'erreur quadratique moyenne normalisée de l'estimateur calé $\hat{Y}_{\mathrm{cal},hh}$ est plus petite quand les variables de calage sont explicatives pour la variable d'intérêt, comme cela était attendu. Nous observons un biais légèrement positif de l'estimateur de la variance bootstrap pour l'estimateur en présence de réponse complète $\hat{Y}_{hh}$, mais presque aucun biais pour les estimateurs repondérés $\hat{Y}_{r,hh}$ et $\hat{Y}_{\mathrm{cal},hh}$. L'estimateur de la variance bootstrap est un peu moins stable avec les estimateurs repondérés, ce qui est probablement attribuable à la variabilité supplémentaire associée à la correction de la non-réponse totale. En ce qui concerne les intervalles de confiance, nous constatons que les taux de couverture sont correctement respectés dans tous les cas et pour les trois méthodes étudiées.

Nous passons maintenant au résultat sur la population de personnes, qui est présenté au tableau 4.2. Nous observons que le biais relatif de l'estimateur de la variance bootstrap est très petit dans tous les cas. Le choix des poids θ_k utilisés dans l'estimation des probabilités de réponse ne semble pas avoir d'effet sur la racine de l'erreur quadratique moyenne normalisée des estimateurs, mais l'utilisation des poids $\theta_l = d_{rl}$ ajustés pour tenir compte de la non-réponse des ménages donne des estimateurs de la variance légèrement plus stables pour $\hat{Y}_{rr,\mathrm{ind}}$. Les taux de couverture sont à peu près respectés dans tous les cas.

Tableau 4.1
Coefficient de variation de l'estimateur du total, Biais relatif et Stabilité relative de l'estimateur de la variance bootstrap et Taux d'erreur nominaux unilatéraux pour l'estimation de la variance bootstrap par la méthode des centiles et du bootstrap de base pour 3 variables de la population des ménages

					Bootstrap par la méthode des centiles			Bootstrap de base			Reposant sur la normalité		
		REQMN	BR	SR	L	U	L+U	L	U	L+U	L	U	L+U
$\hat{Y}_{hh}$	y_1	1,47	2,48	7,2	2,2	3,1	5,3	2,1	3,3	5,4	2,2	3,2	5,4
	y_2	1,48	0,73	6,6	2,6	3,3	5,9	2,7	3,4	6,1	2,6	3,2	5,8
	y_3	1,48	1,11	6,6	2,6	2,7	5,3	2,7	3,0	5,7	2,4	2,7	5,1
$\hat{Y}_{r,hh}$	y_1	1,82	0,42	8,7	2,4	2,4	4,8	2,3	2,7	5,0	2,3	2,6	4,9
	y_2	1,83	-0,76	8,2	2,7	2,8	5,5	2,5	3,0	5,5	2,2	2,7	4,9
	y_3	1,82	0,72	8,4	2,8	2,1	4,9	2,8	2,2	5,0	2,8	1,9	4,7
$\hat{Y}_{\mathrm{cal},hh}$	y_1	1,29	1,27	8,3	2,4	2,7	5,1	2,8	2,8	5,6	2,8	2,7	5,5
	y_2	1,58	-0,55	8,2	2,5	3,5	6,0	2,8	3,9	6,7	2,8	3,6	6,4
	y_3	1,82	0,49	8,4	2,9	1,8	4,7	3,0	2,2	5,2	2,9	2,0	4,9

Tableau 4.2
Coefficient de variation de l'estimateur du total, Biais relatif et Stabilité relative de l'estimateur de la variance bootstrap et Taux d'erreur nominaux unilatéraux pour l'estimation de la variance bootstrap par la méthode des centiles et du bootstrap de base pour 3 variables de la population des personnes

					Bootstrap par la méthode des centiles			Bootstrap de base			Reposant sur la normalité		
		REQMN	BR	SR	L	U	L+U	L	U	L+U	L	U	L+U
		Poids égaux $\theta_l = 1$											
$\hat{Y}_{rr,\text{ind}}$	y_4	2,01	0,31	9,6	2,0	3,2	5,2	1,9	3,3	5,2	1,9	3,0	4,9
	y_5	2,02	-0,17	9,6	2,4	3,4	5,8	2,2	3,7	5,9	2,3	3,5	5,8
	y_6	2,02	-0,24	9,6	2,2	3,3	5,5	2,0	3,7	5,7	2,0	3,2	5,2
$\hat{Y}_{\text{cal,ind}}$	y_4	0,29	1,72	10,8	2,1	2,4	4,5	2,1	2,3	4,4	2,1	2,2	4,3
	y_5	0,39	1,04	11,3	2,3	2,5	4,8	2,3	2,5	4,8	2,2	2,4	4,6
	y_6	0,47	1,90	11,2	2,8	2,1	4,9	2,2	2,5	4,7	2,3	2,0	4,3
		Poids d'échantillonnage $\theta_l = d_l$											
$\hat{Y}_{rr,\text{ind}}$	y_4	2,00	-0,08	9,5	1,8	3,8	5,6	1,7	3,8	5,5	1,7	3,4	5,1
	y_5	2,00	0,14	9,4	1,9	3,3	5,2	2,2	3,5	5,7	1,8	3,5	5,3
	y_6	1,99	0,61	9,3	1,7	3,2	4,9	1,7	3,4	5,1	1,7	3,2	4,9
$\hat{Y}_{\text{cal,ind}}$	y_4	0,29	-0,57	10,3	2,9	2,4	5,3	3,3	2,2	5,5	3,0	2,3	5,3
	y_5	0,39	0,40	11,6	2,4	3,2	5,6	2,7	3,3	6,0	2,3	3,2	5,5
	y_6	0,47	-0,05	11,2	2,3	2,2	4,5	1,8	2,3	4,1	1,8	2,3	4,1
		Poids corrigés pour tenir compte de la non-réponse des ménages $\theta_l = d_{rl}$											
$\hat{Y}_{rr,\text{ind}}$	y_4	1,99	-0,71	8,9	2,5	2,3	4,8	2,6	2,7	5,3	2,5	2,4	4,9
	y_5	1,99	-0,82	8,9	3,1	2,2	5,3	2,9	2,5	5,4	2,5	2,2	4,7
	y_6	1,99	-0,26	9,1	3,1	2,3	5,4	3,0	3,0	6,0	2,9	2,5	5,4
$\hat{Y}_{\text{cal,ind}}$	y_4	0,29	1,70	10,6	2,7	3,4	6,1	2,6	3,3	5,9	2,5	3,3	5,8
	y_5	0,39	1,38	11,3	2,1	2,7	4,8	2,2	3,0	5,2	1,7	3,0	4,7
	y_6	0,47	0,61	10,9	2,5	2,8	5,3	2,3	3,0	5,3	2,3	2,8	5,1

5. Application sur le Panel Politique de la Ville français

Dans cette section, nous présentons une illustration de la méthodologie proposée sur un panel français concernant la politique de la ville. Le plan de sondage et les étapes d'estimation pour l'échantillon des ménages sont brièvement décrits à la section 5.1, et trois intervalles de confiance bootstrap possibles sont calculés. La macro SAS élaborée en vue de mettre en œuvre la méthodologie proposée pour l'échantillonnage à un degré est donnée à l'annexe B, accompagnée d'un petit exemple. Les étapes supplémentaires d'échantillonnage et d'estimation pour l'échantillon de personnes sont décrites à la section 5.2, et trois intervalles de confiance bootstrap possibles sont calculés. La macro SAS élaborée en vue de mettre en œuvre la méthodologie proposée pour l'échantillonnage à deux degrés est donnée à l'annexe C, accompagnée d'un petit exemple.

5.1 Échantillon de ménages

Le Panel Politique de la Ville (PPV) est une enquête en quatre vagues, menée entre 2011 et 2014 par le secrétariat général du Comité interministériel des villes (SGCIV) de France. L'enquête visait à recueillir des informations sur la sécurité, l'emploi, la précarité, la scolarité et la santé de personnes vivant dans les

zones urbaines sensibles (ZUS). Nous nous intéressons uniquement à la vague de 2011. Un échantillon de ménages est sélectionné, et toutes les personnes vivant dans les ménages sélectionnés sont sondées en théorie.

L'échantillon des ménages est obtenu par échantillonnage à deux degrés, voir par exemple Chauvet (2015); Chauvet et Vallée (2018). Premièrement, la population des quartiers est divisée en quatre strates, et un échantillon global de $n_I = 40$ quartiers est sélectionné au moyen d'un échantillonnage à probabilités proportionnelles à la taille à l'intérieur des strates. Un échantillon de ménages est ensuite sélectionné au deuxième degré dans chaque quartier sélectionné au moyen d'un échantillonnage aléatoire simple, de sorte que les probabilités d'inclusion finale des ménages sont approximativement égales à l'intérieur des strates (plan d'échantillonnage auto-pondéré). Aux fins de l'illustration, la sélection à deux degrés des ménages n'est pas prise en compte ici, et l'échantillon des ménages est considéré comme étant directement sélectionné au moyen d'un échantillonnage aléatoire simple stratifié.

L'échantillon comprend 2 971 ménages, mais en raison de la non-réponse totale, 1 256 ménages seulement sont observés. La non-réponse est prise en compte au moyen de groupes de réponse homogènes, définis en fonction de cinq variables auxiliaires : période de construction du logement, type de logement (appartement/maison), nombre de pièces, habitation à loyer modique (oui/non) et région. En utilisant une régression logistique et la méthode des scores (par exemple Haziza et Beaumont, 2007), nous obtenons huit groupes de réponse homogènes. Les cinq variables auxiliaires ayant servi à la définition des GRH sont également utilisées pour le calage.

Nous nous intéressons à quatre variables catégorielles liées à la sécurité, à l'urbanisme et à la mobilité résidentielle. La variable y_1 donne la réputation perçue du quartier (bonne, passable, mauvaise, sans opinion). La variable y_2 indique si un membre du ménage a été témoin de trafic (jamais, rarement, parfois, sans opinion). La variable y_3 indique si des travaux routiers importants ont été réalisés dans le quartier au cours des 12 derniers mois (oui, non, sans opinion). La variable y_4 indique si le ménage a l'intention de quitter le quartier au cours des 12 prochains mois (certainement/probablement, certainement pas, probablement pas, sans opinion). Pour chaque catégorie g de chaque variable y, nous nous intéressons à la proportion

$$\beta_{g,hh} = \frac{\sum_{k \in U_{hh}} 1(y_k = g)}{N_{hh}}, \tag{5.1}$$

avec N_{hh} le nombre total de ménages. L'estimateur de β_g ajusté pour tenir compte de la non-réponse est

$$\hat{\beta}_{\mathrm{gr},hh} = \frac{\sum_{k \in S_{r,hh}} d_{rk} 1(y_k = g)}{\sum_{k \in S_{r,hh}} d_{rk}}, \tag{5.2}$$

voir l'équation (2.7). L'estimateur calé de β_g est

$$\hat{\beta}_{\mathrm{gcal},hh} = \frac{\sum_{k \in S_{r,hh}} w_k 1(y_k = g)}{\sum_{k \in S_{r,hh}} w_k}, \tag{5.3}$$

voir l'équation (2.10).

Pour chaque proportion, nous donnons l'intervalle de confiance reposant sur la normalité en utilisant l'estimateur de la variance bootstrap, la méthode du percentile ou la méthode du percentile inverse, voir la section 3.5. Nous utilisons le bootstrap avec remise présenté dans l'algorithme 1 avec $B = 1\,000$ rééchantillonnages. Les résultats avec un taux d'erreur nominal unilatéral de 2,5 % sont présentés dans le tableau 5.1. Les trois intervalles de confiance sont très similaires dans tous les cas.

Tableau 5.1
Estimation des proportions marginales avec trois intervalles de confiance pour quatre variables d'intérêt

	Réputation perçue du statut du quartier							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
	Bonne	Passable	Mauvaise	Sans opinion	Bonne	Passable	Mauvaise	Sans opinion
Estimation	0,217	0,225	0,531	0,027	0,217	0,224	0,532	0,027
IC Norm.	[0,194;0,241]	[0,201;0,249]	[0,503;0,559]	[0,018;0,036]	[0,193;0,240]	[0,200;0,248]	[0,504;0,560]	[0,018;0,036]
IC. Centiles	[0,195;0,241]	[0,201;0,251]	[0,504;0,558]	[0,019;0,036]	[0,193;0,240]	[0,201;0,251]	[0,505;0,560]	[0,019;0,036]
IC de base	[0,193;0,240]	[0,200;0,249]	[0,503;0,557]	[0,018;0,035]	[0,193;0,240]	[0,198;0,248]	[0,504;0,559]	[0,018;0,035]
	Témoin de trafic							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
	Jamais	Rarement	Parfois	Sans opinion	Jamais	Rarement	Parfois	Sans opinion
Estimation	0,599	0,065	0,155	0,181	0,606	0,065	0,156	0,173
IC Norm.	[0,571;0,627]	[0,050;0,079]	[0,135;0,175]	[0,161;0,201]	[0,581;0,632]	[0,050;0,079]	[0,135;0,176]	[0,159;0,188]
IC Centiles	[0,572;0,628]	[0,050;0,080]	[0,134;0,175]	[0,161;0,201]	[0,582;0,633]	[0,051;0,080]	[0,134;0,175]	[0,160;0,188]
IC de base	[0,570;0,626]	[0,049;0,078]	[0,136;0,176]	[0,161;0,201]	[0,579;0,630]	[0,049;0,078]	[0,136;0,177]	[0,159;0,187]
	Travaux routiers dans le quartier							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
	Oui	Non	Sans opinion		Oui	Non	Sans opinion	
Estimation	0,471	0,495	0,034		0,470	0,496	0,034	
IC Norm.	[0,444;0,498]	[0,468;0,523]	[0,024;0,044]		[0,443;0,496]	[0,469;0,523]	[0,024;0,045]	
IC Centiles	[0,442;0,496]	[0,469;0,524]	[0,025;0,045]		[0,440;0,495]	[0,470;0,524]	[0,025;0,045]	
IC de base	[0,445;0,500]	[0,466;0,522]	[0,023;0,043]		[0,444;0,499]	[0,468;0,522]	[0,024;0,044]	
	Intention de quitter le quartier							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
	Cert./Prob.	Prob. non	Cert. non	Sans opinion	Cert./Prob.	Prob. non	Cert. non	Sans opinion
Estimation	0,286	0,130	0,548	0,036	0,287	0,131	0,546	0,036
IC Norm.	[0,260;0,312]	[0,111;0,149]	[0,520;0,576]	[0,025;0,047]	[0,261;0,313]	[0,112;0,150]	[0,518;0,573]	[0,025;0,047]
IC Centiles	[0,260;0,313]	[0,111;0,149]	[0,521;0,576]	[0,026;0,047]	[0,261;0,313]	[0,113;0,151]	[0,520;0,574]	[0,026;0,048]
IC de base	[0,259;0,312]	[0,111;0,149]	[0,520;0,575]	[0,025;0,046]	[0,261;0,313]	[0,111;0,149]	[0,517;0,572]	[0,025;0,047]

5.2 Échantillon de personnes

L'échantillon des ménages répondants comprend 3 098 personnes qui sont théoriquement sondées, mais en raison de la non-réponse totale, nous observons un sous-ensemble de 2 804 répondants individuels seulement. La non-réponse est prise en compte au moyen de groupes de réponse homogènes, définis en fonction de huit variables auxiliaires : trois au niveau de la personne (sexe, âge, nationalité) et cinq au niveau du logement (période de construction du logement, type de logement, nombre de pièces, habitation à loyer modique ou non, région). En utilisant une régression logistique et la méthode des scores, nous obtenons huit groupes de réponse homogènes. Les trois variables auxiliaires de personnes ayant servi à la définition des GRH sont également utilisées pour le calage.

Nous considérons trois variables d'intérêt. La variable y_5 est quantitative et donne le nombre d'enfants. La variable y_6 indique si la personne a un ou plusieurs emplois (un, plusieurs, aucun, aucune réponse). La variable y_7 indique si la personne bénéficie d'une couverture médicale complémentaire complète (oui, non, pas de réponse). Pour la variable y_5, nous calculons l'estimateur du total ajusté pour

tenir compte de la non-réponse et l'estimateur calé donnés dans les équations (2.27) et (2.29), respectivement. Pour les deux autres variables d'intérêt et pour chaque catégorie g, nous nous intéressons à la proportion

$$\beta_{g,\text{ind}} = \frac{\sum_{l \in U_{\text{ind}}} 1(y_k = g)}{N_{\text{ind}}}, \tag{5.4}$$

avec N_{ind} le nombre total d'individus. L'estimateur de $\beta_{g,\text{ind}}$ ajusté pour tenir compte de la non-réponse est

$$\hat{\beta}_{\text{grr,ind}} = \frac{\sum_{l \in S_{rr,\text{ind}}} d_{rrl} 1(y_l = g)}{\sum_{l \in S_{rr,\text{ind}}} d_{rrl}}, \tag{5.5}$$

voir l'équation (2.27). L'estimateur calé de $\beta_{g,\text{ind}}$ est

$$\hat{\beta}_{\text{gcal,ind}} = \frac{\sum_{l \in S_{rr,\text{ind}}} w_l 1(y_l = g)}{\sum_{l \in S_{rr,\text{ind}}} w_l}, \tag{5.6}$$

voir l'équation (2.29).

Pour chaque paramètre, nous donnons l'intervalle de confiance reposant sur la normalité en utilisant l'estimateur de la variance bootstrap, la méthode du percentile et la méthode du percentile inverse. Nous utilisons le bootstrap avec remise présenté dans l'algorithme 2 avec $B = 1\,000$ rééchantillonnages. Les résultats avec un taux d'erreur nominal unilatéral de 2,5 % sont présentés au tableau 5.2. Les trois intervalles de confiance sont très similaires dans tous les cas.

Tableau 5.2
Estimation des proportions marginales avec trois intervalles de confiance pour quatre variables d'intérêt

	Nombre d'enfants							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
Estimation ($\times 10^6$)	4,40				4,39			
IC Norm.	[4,15;4,64]				[4,21;4,58]			
IC Centiles	[4,16;4,65]				[4,21;4,58]			
IC de base	[4,14;4,63]				[4,20;4,57]			
	La personne a-t-elle plusieurs emplois ?							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
	Un	Plusieurs	Aucun	Aucune réponse	Un	Plusieurs	Aucun	Aucune réponse
Estimation	0,304	0,016	0,372	0,308	0,305	0,016	0,372	0,307
IC Norm.	[0,286;0,323]	[0,011;0,021]	[0,352;0,392]	[0,290;0,326]	[0,285;0,325]	[0,011;0,021]	[0,350;0,394]	[0,283;0,332]
IC Centiles	[0,287;0,323]	[0,011;0,021]	[0,351;0,393]	[0,289;0,326]	[0,284;0,325]	[0,011;0,020]	[0,351;0,393]	[0,284;0,333]
IC de base	[0,286;0,322]	[0,011;0,020]	[0,351;0,393]	[0,289; 0,326]	[0,285;0,325]	[0,011;0,020]	[0,352;0,393]	[0,282;0,330]
	Couverture médicale complémentaire complète							
	Estimateur ajusté pour non-réponse				Estimateur par calage			
	Oui	Non	Aucune réponse		Oui	Non	Aucune réponse	
Estimation	0,122	0,626	0,252		0,122	0,627	0,251	
IC Norm.	[0,106;0,137]	[0,603;0,650]	[0,234;0,270]		[0,105;0,138]	[0,604;0,650]	[0,227;0,275]	
IC Centiles	[0,105;0,137]	[0,603;0,651]	[0,235;0,269]		[0,105;0,138]	[0,604;0,650]	[0,230;0,276]	
IC de base	[0,106;0,138]	[0,602;0,649]	[0,235;0,269]		[0,105;0,138]	[0,605;0,651]	[0,227;0,273]	

6. Conclusion et travaux futurs

Dans le présent article, nous avons expliqué comment il est possible d'appliquer le bootstrap avec remise aux enquêtes auprès des ménages, afin de tenir compte de toute la variabilité du processus d'échantillonnage, y compris l'échantillonnage et la non-réponse, ainsi qu'à des ajustements a posteriori comme le calage. Les méthodes ont été illustrées au moyen d'un exemple simple permettant de les exposer clairement, évaluées par une étude par simulations et appliquées à un panel français sur la politique de la ville. Afin de faciliter la mise en œuvre de la méthode par les utilisateurs, nous avons élaboré deux macros SAS que l'auteur correspondant peut mettre à disposition sur demande.

Les résultats de l'étude par simulations montrent que les estimateurs de la variance bootstrap et les trois intervalles de confiance bootstrap fonctionnent correctement dans le cas d'une petite fraction de sondage. Si la fraction de sondage est plus grande, l'estimateur de la variance bootstrap est connu pour être conservatif, et l'intervalle de confiance reposant sur la normalité devrait donc l'être également. Toutefois, les propriétés de couverture des deux autres intervalles de confiance dans ce contexte demeurent floues. Il serait intéressant d'approfondir cette question.

Dans l'article, nous nous sommes concentrés sur l'application du bootstrap pour l'estimation de la variance, après que les ajustements statistiques (traitement de la non-réponse totale et calage) ont été effectués par le méthodologiste de l'enquête. Le bootstrap peut également être utilisé a priori comme outil de diagnostic servant à évaluer la pertinence d'éventuels ajustements statistiques. Par exemple, il peut être tentant d'utiliser un grand nombre de groupes de réponse homogènes (GRH) pour corriger la non-réponse totale, afin de réduire le biais de non-réponse. Cependant, cela peut entraîner une variabilité accrue des estimateurs repondérés. Le bootstrap peut servir à évaluer plusieurs ensembles possibles de GRH, par exemple en produisant des histogrammes des ajustements de non-réponse bootstrap ou des estimateurs bootstrap corrigés pour tenir compte de la non-réponse totale, afin de donner une idée de la stabilité de l'estimation avec un ensemble possible de GRH. Cela est utile pour trouver un compromis entre le biais et la variance. Cette méthode mentionnée dans Girard (2009) est une question importante qu'il faudrait approfondir.

Nous avons étudié une situation où l'enquête n'est effectuée qu'une seule fois. Si nous souhaitons effectuer des estimations longitudinales, les unités sont généralement suivies dans le temps. Si nous nous intéressons aussi aux estimations transversales réalisées à plusieurs reprises, des échantillons supplémentaires sont sélectionnés aux vagues postérieures et mélangés avec l'échantillon initial. L'estimation de la variance bootstrap dans le contexte des enquêtes longitudinales est une question très importante qui devrait faire l'objet d'autres travaux.

Remerciements

Les auteurs remercient chaleureusement le rédacteur adjoint et les deux examinateurs pour leurs nombreuses suggestions sur une version précédente de l'article, qui ont permis de considérablement l'améliorer.

Annexe

A. Estimateurs de la variance repère pour l'échantillon de personnes

Nous considérons d'abord l'estimateur $\hat{Y}_{\text{ind}}$ dans l'équation (2.21) que nous utilisons en cas de réponse complète. L'estimateur repère de la variance est

$$v_{\text{mult}}(\hat{Y}_{\text{ind}}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{k \in S_{hh}^h} \left(d_k \hat{y}_k - \frac{1}{n_h} \sum_{k' \in S_{hh}^h} d_{k'} \hat{y}_{k'} \right)^2, \tag{A.1}$$

avec

$$\hat{y}_k = \sum_{l \in S_{\text{ind},k}} d_{l|k} y_l.$$

Nous considérons maintenant l'estimateur $\hat{Y}_{rr,\text{ind}}$ donné dans l'équation (2.27), qui est ajusté pour tenir compte de la non-réponse des ménages et des personnes. L'estimateur repère de la variance est

$$v_{\text{mult}}(\hat{Y}_{\text{ind}}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{k \in S_{hh}^h} \left(d_k v_{1k} - \frac{1}{n_h} \sum_{k' \in S_{hh}^h} d_{k'} v_{1k'} \right)^2, \tag{A.2}$$

où

$$v_{1k} = \hat{u}_{1k} + u_{3k},$$

où la première variable linéarisée $\hat{u}_{1k}$ est semblable à celle donnée dans l'équation (2.8), tandis que la deuxième variable linéarisée u_{3k} tient compte de l'estimation des probabilités de réponse des personnes. Nous avons pour la première variable linéarisée

$$\hat{u}_{1k} = \theta_k \pi_k \bar{\hat{y}}_{rc(k)} + \frac{r_k}{\hat{p}_{c(k)}} \left\{ \hat{y}_{r,k} - \theta_k \pi_k \bar{\hat{y}}_{rc(k)} \right\},$$

et

$$\bar{\hat{y}}_{rc} = \frac{\sum_{k \in S_{c,hh}} d_k r_k \hat{y}_{r,k}}{\sum_{k \in S_{c,hh}} \theta_k r_k},$$

et

$$\hat{y}_{r,k} = \sum_{l \in S_{\text{ind},k}} \frac{d_{l|k} r_l}{\hat{p}_l} y_l, \tag{A.3}$$

et pour la deuxième variable linéarisée

$$u_{3k} = \frac{r_k}{d_k} \sum_{l \in S_{\text{ind},k}} \theta_l \left(1 - \frac{r_l}{\hat{p}_l} \right) \bar{y}_{rrd(l)}, \tag{A.4}$$

avec

$$\bar{y}_{rrd} = \frac{\sum_{l \in S_{rd,\text{ind}}} d_{rl} r_l y_l}{\sum_{l \in S_{rd,\text{ind}}} \theta_l r_l}. \tag{A.5}$$

Considérons maintenant l'estimateur calé $\hat{Y}_{\text{cal,ind}}$ donné dans l'équation (2.29). L'estimateur repère de la variance est le même que celui donné dans l'équation (A.2) pour $\hat{Y}_{rr,\text{ind}}$, en remplaçant la variable y_l par les résidus de régression estimés de la variable d'intérêt sur les variables de calage, à savoir

$$e_l = y_l - \hat{B}_{rr,\text{ind}}^{\top} z_l \quad \text{avec} \quad \hat{B}_{rr,\text{ind}} = \left(\sum_{l \in S_{rr,\text{ind}}} d_{rrl} z_l z_l^{\top} \right)^{-1} \sum_{l \in S_{rr,\text{ind}}} d_{rrl} z_l y_l. \tag{A.6}$$

B. Programme SAS pour échantillonnage à un degré

Dans cette section, nous présentons la macro SAS élaborée pour mettre en œuvre la méthodologie proposée en cas d'échantillonnage des ménages seulement (échantillonnage à un degré). Le paramétrage du programme SAS pour le calcul des poids bootstrap est présenté à la section B.1. Par souci de clarté, un petit exemple est présenté à la section B.2.

B.1 Programme de calcul des poids bootstrap

Les paramètres liés à la base de données sont :

- `BASE` : bibliothèque contenant la table SAS avec la liste des unités échantillonnées. La valeur par défaut est `BASE=WORK`.
- `ECHMEN` : table SAS contenant la liste des unités échantillonnées dans la population. Les non-répondants doivent aussi être inclus dans cette table.

Les paramètres liés au bootstrap sont :

- `ITBOOT` : nombre d'itérations bootstrap. La valeur par défaut est `ITBOOT=1000`.

Les paramètres liés aux variables nécessaires dans la table SAS sont :

- `IDMEN` : liste des variables identifiant l'unité statistique. Elles doivent être de type caractère.
- `STMEN` : liste des variables de stratification utilisées pour la sélection de l'échantillon.
- `DMEN` : poids d'échantillonnage.
- `RMEN` : indicateur de réponse (1 pour un répondant, 0 pour un non-répondant).
- `DRMEN` : poids d'échantillonnage, corrigé pour tenir compte de la non-réponse. Les valeurs sont nécessaires uniquement pour les répondants.
- `DCMEN` : poids calé. Les valeurs sont nécessaires uniquement pour les répondants.
- `GRHMEN` : liste des variables identifiant les groupes de réponse homogènes.
- `WGRHMEN` : pondération utilisée dans le calcul des probabilités de réponse à l'intérieur des GRH.
  - Avec `WGRHMEN=0`, les taux de réponse ne sont pas pondérés. Il s'agit de la valeur par défaut.
  - Avec `WGRHMEN=1`, les taux de réponse sont pondérés par les poids de sondage.

- XMENQUANT : liste des variables quantitatives utilisées dans le calage. Les valeurs sont nécessaires uniquement pour les répondants.
- XMENQUALI : liste des variables qualitatives utilisées dans le calage. Les valeurs sont nécessaires uniquement pour les répondants.

Les paramètres liés à la sortie sont :

- SORT_MEN : table SAS contenant les poids d'échantillonnage bootstrap WB_D1,...,WB_D&ITBOOT pour l'échantillon entier.
- SORT_RMEN : table SAS contenant les poids bootstrap WB_N1,...,WB_N&ITBOOT corrigés pour tenir compte de la non-réponse, et les poids bootstrap WB_C1,...,WB_C&ITBOOT corrigés pour tenir compte de la non-réponse et du calage, pour le sous-échantillon de répondants.

B.2 Petit exemple

Considérons l'exemple traité à la section 2.1.4. L'échantillon se présente comme suit :

```
data ech;
input idm$ stmen$ dmen rmen GRHmen$ drmen dcmen x0 x1;
cards;
A  1  4   1  aa  4.44   4.01   1  1
B  1  4   0  aa  .      .      .  .
C  1  4   0  bb  .      .      .  .
D  1  4   1  bb  5.54   4.87   1  0
E  1  16  1  bb  22.15  19.98  1  1
F  1  16  1  aa  17.78  15.63  1  0
G  1  16  0  bb  .      .      .  .
H  1  16  1  bb  22.15  19.98  1  1
I  1  16  1  bb  22.15  19.49  1  0
J  1  16  1  aa  17.78  16.03  1  1
;run;
```

Nous pouvons obtenir $B = 1\,000$ poids bootstrap comme suit. Étant donné que WGRHMEN=1, on suppose que quand la non-réponse totale a été initialement corrigée par la méthode des GRH, les taux de réponse à l'intérieur des GRH ont été pondérés par les poids d'échantillonnage.

```
%BOOTUP_1DEG(BASE=work,ECHMEN=ech,
          ITBOOT=1000,
          IDMEN=idm,STMEN=stmen,DMEN=dmen,
          RMEN=rmen,DRMEN=drmen,DCMEN=dcmen,GRHMEN=GRHmen,WGRHMEN=1,
          XMENQUANT=x0 x1,XMENQUALI=,
          SORT_MEN=ech_boot,SORT_RMEN=echr_boot);
```

C. Programme SAS pour l'échantillonnage à deux degrés

Dans cette section, nous présentons la macro SAS élaborée pour mettre en œuvre la méthodologie proposée en cas d'échantillonnage des ménages et un sous-échantillonnage de personnes (échantillonnage à deux degrés). Le paramétrage du programme SAS pour le calcul des poids bootstrap est présenté à la section C.1. Par souci de clarté, un petit exemple est présenté à la section C.2.

C.1 Programme de calcul des poids bootstrap

La macro SAS `%BOOTUP_2DEG` permet de calculer les poids bootstrap pour une enquête auprès des ménages avec sous-échantillonnage de personnes, et de tenir compte de la correction de la non-réponse totale au moyen de groupes de réponse homogènes, et du calage des poids, à la fois pour les ménages et les personnes.

Les paramètres avec un signe d'égalité sont obligatoires. Toutes les variables d'identification doivent être de type caractère.

Les paramètres liés à la base de données sont :

- `BASE` : bibliothèque contenant les tables SAS `ECHMEN` et `ECHIND`. La valeur par défaut est `BASE=WORK`.
- `BASESOR` : bibliothèque contenant la sortie. La valeur par défaut est `BASESOR=WORK`.
- `ECHMEN=` : table SAS contenant la liste des ménages échantillonnés dans la population. Les non-répondants au niveau ménage doivent aussi être inclus dans cette table.
- `ECHIND=` : table SAS contenant la liste des personnes échantillonnées dans tous les ménages répondants. Les personnes non répondantes doivent aussi être incluses dans cette table.

Les paramètres liés au bootstrap sont :

- `ITBOOT` : nombre d'itérations bootstrap. La valeur par défaut est `ITBOOT=1000`.

Les paramètres liés aux variables nécessaires dans la table SAS du ménage `ECHMEN` sont :

- `IDMEN=` : liste des variables identifiant le ménage. Cette variable est requise dans `ECHMEN` et `ECHIND`.
- `STMEN` : liste des variables de stratification utilisées pour la sélection de l'échantillon.
- `DMEN` : poids d'échantillonnage du ménage.
- `RMEN` : indicateur de réponse du ménage (1 pour un répondant, 0 pour un non-répondant).
- `DRMEN` : poids d'échantillonnage du ménage, corrigé pour tenir compte de la non-réponse. Les valeurs sont nécessaires uniquement pour les répondants.
- `DCMEN` : poids calé. Les valeurs sont nécessaires uniquement pour les répondants.
- `GRHMEN` : liste des variables identifiant les groupes de réponse homogènes pour les ménages.
- `WGRHMEN` : pondération utilisée dans le calcul des probabilités de réponse à l'intérieur des GRH :

  - Avec WGRHMEN=0, les taux de réponse ne sont pas pondérés. Il s'agit de la valeur par défaut.
  - Avec WGRHMEN=1, les taux de réponse sont pondérés par les poids de sondage DMEN.
- XMENQUANT : liste des variables quantitatives utilisées dans le calage. Les valeurs sont nécessaires uniquement pour les répondants.
- XMENQUALI : liste des variables qualitatives utilisées dans le calage. Les valeurs sont nécessaires uniquement pour les répondants.

Les paramètres liés aux variables nécessaires dans la table SAS des personnes ECHIND sont :

- ID_IND= : liste des variables identifiant la personne (variable de caractères).
- R_IND : indicateur de réponse de la personne (1 pour un répondant, 0 pour un non-répondant).
- DR_IND : poids de la personne, corrigé pour tenir compte à la fois de la non-réponse totale du ménage et de la personne. Les valeurs sont nécessaires uniquement pour les répondants.
- DC_IND : poids calé. Les valeurs sont nécessaires uniquement pour les répondants.
- PIKSACI= : probabilité d'inclusion conditionnelle de la personne à l'intérieur de son ménage.
- GRH_IND : liste des variables identifiant les groupes de réponse homogènes.
- WGRH_IND : pondération utilisée dans le calcul des probabilités de réponse à l'intérieur des GRH :
  - Avec WGRH_IND=0, les taux de réponse ne sont pas pondérés. Il s'agit de la valeur par défaut.
  - Avec WGRH_IND=1, les taux de réponse sont pondérés par les poids de sondage des personnes.
  - Avec WGRH_IND=2, les taux de réponse sont pondérés par les poids des personnes, ajustés en fonction de la non-réponse totale du ménage.
- XINDQUANT : liste des variables quantitatives utilisées dans le calage. Les valeurs sont nécessaires uniquement pour les répondants.
- XINDQUALI : liste des variables qualitatives utilisées dans le calage. Les valeurs sont nécessaires uniquement pour les répondants.

Les paramètres liés à la sortie sont :

- SORT_MEN : table SAS contenant tous les ménages échantillonnés et les poids d'échantillonnage bootstrap WB_D1,...,WB_D&ITBOOT pour l'ensemble de l'échantillon.
- SORT_RMEN : table SAS contenant tous les ménages répondants et les poids bootstrap
  - WB_N1,...,WB_N&ITBOOT corrigés pour tenir compte de la non-réponse,
  - WB_C1,...,WB_C&ITBOOT corrigés pour tenir compte de la non-réponse et du calage.
- SORT_RIND : table SAS contenant toutes les personnes répondantes à l'intérieur des ménages répondants, et les poids bootstrap
  - WB_N1,...,WB_N&ITBOOT corrigés pour tenir compte de la non-réponse du ménage,
  - WB_NN1,...,WB_NN&ITBOOT corrigés pour tenir compte à la fois de la non-réponse du ménage et de la non-réponse des personnes,
  - WB_C1,...,WB_C&ITBOOT corrigés pour tenir compte de la non-réponse et du calage.

C.2 Petit exemple

Considérons l'exemple traité à la section 2.2.4. L'échantillon des ménages et l'échantillon des personnes sont les suivants :

```
data echmen;
input idm$ stmen$ dmen rmen GRHmen$ drmen dcmen x0 x1;
cards;
A  1  4   1  aa  4.44   4.01   1  1
B  1  4   0  aa  .      .      .  .
C  1  4   0  bb  .      .      .  .
D  1  4   1  bb  5.54   4.87   1  0
E  1  16  1  bb  22.15  19.98  1  1
F  1  16  1  aa  17.78  15.63  1  0
G  1  16  0  bb  .      .      .  .
H  1  16  1  bb  22.15  19.98  1  1
I  1  16  1  bb  22.15  19.49  1  0
J  1  16  1  aa  17.78  16.03  1  1
;run;


data echind;
input idm$ idi$ piksaci dr1_ind rind GRHind$ phat_ind dr2_ind xi1 xi2
dc_ind;
cards;
A   i01   0.34   13.06   1   g1   0.75   17.41   1   3   19.61
D   i04   1.00   5.54    0   g2   0.33   .       .   .   .
E   i06   0.34   65.15   1   g1   0.75   86.86   1   2   53.93
F   i08   0.33   53.88   1   g1   0.75   71.84   1   3   78.43
H   i11   0.50   44.30   0   g1   0.75   .       .   .   .
I   i13   1.00   22.15   1   g2   0.33   67.12   1   1   48.04
J   i14   1.00   17.78   0   g2   0.33   .       .   .   .
;run;
```

Nous pouvons obtenir $B = 1\,000$ poids bootstrap comme suit. Étant donné que `WGRHMEN=1`, on suppose que quand la non-réponse totale des ménages a été initialement corrigée par la méthode des GRH, les taux de réponse à l'intérieur des GRH ont été pondérés par les poids d'échantillonnage. Étant donné que `WGRH_IND=0`, on suppose que quand la non-réponse totale des personnes a été initialement corrigée par la méthode des GRH, les taux de réponse à l'intérieur des GRH n'étaient pas pondérés.

```
%bootup_2deg(BASE=work,BASESOR=work,ECHMEN=echmen,ECHIND=echind,
  ITBOOT=1000,
  IDMEN=idm,STMEN=stmen,DMEN=dmen,RMEN=rmen,DRMEN=drmen,GRHMEN=GRHmen,W
  GRHMEN=0,
  DCMEN=dcmen,XMENQUANT=x0 x1,XMENQUALI=,
  ID_IND=idi,R_IND=rind,DR_IND=dr2_ind,PIKSACI=piksaci,GRH_IND=GRHind,W
  GRH_IND=0,
  DC_IND=dc_ind,XINDQUANT=xi1 xi2,XINDQUALI=,
  SORT_MEN=sort_men,SORT_RMEN=sort_rmen,
  SORT_RIND=sort_rind);
```

Bibliographie

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3), 445-458.

Beaumont, J.-F., et Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *Revue Internationale de Statistique*, 80(1), 127-148.

Brick, J.M. (2013). Unit non-response and weighted adjustments: A critical review. *Journal of Official Statistics*, 29(3), 329-353.

Chauvet, G. (2007). *Méthodes de Bootstrap en population finie*. Thèse de doctorat, University of Rennes 2.

Chauvet, G. (2015). Coupling methods for multistage sampling. *The Annals of Statistics*, 43(6), 2484-2506.

Chauvet, G., et Vallée, A.-A. (2018). Inference for two-stage sampling designs with application to a panel for urban policy. *arXiv preprint arXiv:1808.09758*.

Davison, A.C., et Hinkley, D.V. (1997). Bootstrap methods and their application, volume 1, *Cambridge Series in Statistical and Probabilistic Mathematics*.

Davison, A.C., et Sardy, S. (2007). Resamping variance estimation in surveys with missing data. *Journal of Official Statistics*, 23(3), 371-386.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 2, 219-230. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1999002/article/4882-fra.pdf.

Girard, C. (2009). The Rao-Wu rescaling bootstrap: from theory to practice. Dans *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, pages 2-4. Citeseer.

Haziza, D., et Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *Revue Internationale de Statistique*, 75(1), 25-43.

Haziza, D., et Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32(2), 206-226.

Juillard, H., et Chauvet, G. (2018). Estimation de la variance sous non-réponse monotone pour une enquête par panel. *Techniques d'enquête*, 44, 2, 295-317. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018002/article/54952-fra.pdf.

Kim, J.K., et Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics/La revue canadienne de statistique*, 35(4), 501-514.

Kim, J.K., Navarro, A. et Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Aassociation*, 101(473), 312-320.

Kott, P.S. (2012). Pourquoi les poids de sondage devraient être intégrés dans la correction de la non-réponse totale fondée sur des groupes de réponse homogènes. *Techniques d'enquête*, 38, 1, 103-107. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012001/article/11689-fra.pdf.

Mashreghi, Z., Haziza, D. et Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10, 1-52.

McCarthy, P., et Snowden, C. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*, (95), 1-23.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401), 231-241.

Rao, J.N.K., Wu, C.F.J. et Yue, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 2, 225-234. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1992002/article/14486-fra.pdf.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics.

Shao, J. (1994). *L*-Statistics in complex survey problems. *The Annals of Statistics*, 22(2), 946-967.

Shao, J., et Rao, J. (1993). Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankhyā: The Indian Journal of Statistics, Series B*, 393-414.

Shao, J., et Tu, D.S. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics.

Tillé, Y. (2011). *Sampling Algorithms*. Springer.

Yeo, D., Mantel, H. et Liu, T.-P. (1999). Bootstrap variance estimation for the national population health survey. Dans *Proceedings of the Survey Research Methods Section*, American Statistical Association. Citeseer.