Université de Montréal

Estimation simplifiée de la variance dans le cas de l'échantillonnage à deux phases

par
Audrey Béliveau

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en statistique

Août, 2011

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Estimation simplifiée de la variance dans le cas de l'échantillonnage à deux phases

présenté par:

Audrey Béliveau

a été évalué par un jury composé des personnes suivantes:

Pierre Lafaye de Micheaux,	président-rapporteur
David Haziza,	directeur de recherche
Pierre Duchesne,	membre du jury

Mémoire accepté le: 20 décembre 2011

RÉSUMÉ

Dans ce mémoire, nous étudions le problème de l'estimation de la variance pour les estimateurs par double dilatation et de calage pour l'échantillonnage à deux phases. Nous proposons d'utiliser une décomposition de la variance différente de celle habituellement utilisée dans l'échantillonnage à deux phases, ce qui mène à un estimateur de la variance simplifié. Nous étudions les conditions sous lesquelles les estimateurs simplifiés de la variance sont valides. Pour ce faire, nous considérons les cas particuliers suivants : (1) plan de Poisson à la deuxième phase, (2) plan à deux degrés, (3) plan aléatoire simple sans remise aux deux phases, (4) plan aléatoire simple sans remise à la deuxième phase. Nous montrons qu'une condition cruciale pour la validité des estimateurs simplifiés sous les plans (1) et (2) consiste à ce que la fraction de sondage utilisée pour la première phase soit négligeable (ou petite). Nous montrons sous les plans (3) et (4) que, pour certains estimateurs de calage, l'estimateur simplifié de la variance est valide lorsque la fraction de sondage à la première phase est petite en autant que la taille échantillonnale soit suffisamment grande. De plus, nous montrons que les estimateurs simplifiés de la variance peuvent être obtenus de manière alternative en utilisant l'approche renversée (Fay, 1991 et Shao et Steel, 1999). Finalement, nous effectuons des études par simulation dans le but d'appuyer les résultats théoriques.

Mots clés : Échantillonnage à deux phases, estimateur par double dilatation, estimateurs de calage, estimation de la variance, approche renversée, étude par simulation.

ABSTRACT

In this thesis we study the problem of variance estimation for the double expansion estimator and the calibration estimators in the case of two-phase designs. We suggest to use a variance decomposition different from the one usually used in two-phase sampling, which leads to a simplified variance estimator. We look for the necessary conditions for the simplified variance estimators to be appropriate. In order to do so, we consider the following particular cases : (1) Poisson design at the second phase, (2) two-stage design, (3) simple random sampling at each phase, (4) simple random sampling at the second phase. We show that a crucial condition for the simplified variance estimator to be valid in cases (1) and (2) is that the first phase sampling fraction must be negligible (or small). We also show in cases (3) and (4) that the simplified variance estimator can be used with some calibration estimators when the first phase sampling fraction is negligible and the population size is large enough. Furthermore, we show that the simplified estimators can be obtained in an alternative way using the reversed approach (Fay, 1991 and Shao and Steel, 1999). Finally, we conduct some simulation studies in order to validate the theoretical results.

Keywords: Two-phase sampling, double expansion estimator, calibration estimators, variance estimation, reversed approach, simulation study.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

REMERCIEMENTS

Mes premiers remerciements vont à mon directeur de recherche, David Haziza, avec qui ce fut un plaisir de travailler sur ce mémoire. Je n'aurais pas pu terminer ce mémoire aussi rapidement sans son soutien, sa disponibilité et ses qualités indéniables de chercheur telles l'intuition ainsi qu'une bonne organisation des idées. Je le remercie également pour le temps consacré à la relecture de ce mémoire et pour toutes les opportunités qu'il m'a offertes lors de mes études à l'université de Montréal.

J'aimerais également remercier toutes les personnes qui ont égayé mes journées au département : je pense entre autres à Caroline, Fabiola, Loredana et Pierre-Luc. J'ai également une pensée pour Mylène Bédard, qui m'a offert ma première expérience de recherche en statistique, et pour Christian Léger, qui m'a soutenu dans mon processus d'application au doctorat et grâce à qui une nouvelle aventure excitante m'attend.

Je tiens également à remercier mes parents pour leur générosité matérielle et temporelle ainsi que pour leur confiance dans nos actions et nos choix d'avenir qui m'ont permis d'entreprendre mes études universitaires sans tracas. Merci aussi à toi Olivier pour tous les services qui m'ont permis d'améliorer ma productivité, mais surtout et simplement pour le fait que tu es un merveilleux compagnon et que tu me fais découvrir tant de choses.

Enfin, je suis grandement reconnaissante pour l'appui financier octroyé tout au long de ma maîtrise par le Conseil de recherches en sciences naturelles et en génie du Canada et par le Fonds québécois de la recherche sur la nature et les technologies.

Un dernier mot également pour remercier les membres du jury, Pierre Duchesne et Pierre Lafaye de Micheaux pour leurs commentaires judicieux.

CHAPITRE 1

INTRODUCTION

Les enquêtes sont d'une importance capitale dans la prise de décision de nos gouvernements. Elles permettent de quantifier plusieurs aspects d'une population : son économie, sa culture, son éducation, sa santé, etc. L'enquête est une activité planifiée et méthodique qui requiert généralement la disponibilité d'une base de sondage, i.e. une liste permettant d'identifier et de rejoindre les unités de la population cible. Idéalement, cette liste contient des caractéristiques supplémentaires sur les unités, par exemple, l'âge, le sexe, la province, qui pourront être utiles aux statisticiens à plusieurs étapes de l'enquête. Les statisticiens d'enquête participent à plusieurs étapes de la planification de l'enquête, notamment à l'élaboration du plan de sondage qui indique la façon dont les unités de la population seront sélectionnées. En effet, la grande majorité des sondages de nos jours sont de type probabiliste, c'est-à-dire que la sélection des unités dans la population se fait de façon aléatoire. Les statisticiens d'enquête interviennent également dans la conception du questionnaire et dans le choix des méthodes d'interview (téléphone, en personne, par la poste ou autre). Une fois les données recueillies, le statisticien participe à leur vérification et corrige les erreurs à l'aide de méthodes d'imputation lorsque cela s'avère nécessaire. Enfin, les statisticiens choisissent les méthodes d'estimation des paramètres d'intérêt au niveau de la population à l'aide des données recueillies ; c'est à cette étape que nous nous intéressons dans ce mémoire.

L'échantillonnage doit être distingué de la statistique classique où l'on fait généralement l'hypothèse que les variables observées sont aléatoires et proviennent d'une population hypothétique *infinie*. En effet, dans le contexte des enquêtes, les données recueillies proviennent de populations *finies*. Les variables d'intérêt sont alors traitées comme fixes (plutôt qu'aléatoires) et le mécanisme aléatoire se situe plutôt au niveau du tirage des unités : le fait d'être sélectionné ou non dans l'échantillon est aléatoire.

Le concept de pondération en est également un très important en échantillonnage. En effet, puisque seul un échantillon de la population finie est observé, un poids est généralement associé à chaque unité échantillonnée. Ce poids, qui peut être interprété comme un indicateur du nombre d'unités que celle-ci représente dans la population, est déterminé à partir du plan de sondage utilisé et, dans certains cas, de l'information auxiliaire disponible. Les statisticiens d'enquête utilisent généralement des systèmes de poids assurant un estimateur (asymptotiquement) non-biaisé sous le plan de sondage de la caractéristique d'intérêt au niveau de la population. Dans ce mémoire, nous nous intéressons à l'estimation du total d'une variable d'intérêt dans la population. Par exemple, on pourrait estimer le total du revenu annuel des entreprises au Canada ou le nombre total de chômeurs dans la population.

Afin de pouvoir utiliser des plans de sondage efficaces, les statisticiens d'enquête ont besoin d'information auxiliaire disponible sur la base de sondage. Par exemple, il est coutumier de stratifier la population en sous-groupes homogènes (les strates) au moyen de l'information auxiliaire disponible. Ce plan de sondage se révèle plus efficace que le plan aléatoire simple sans remise si les strates sont homogènes. Mais, dans certaines situations, il arrive que peu ou pas d'information auxiliaire pertinente soit disponible sur la base de sondage. Dans ce cas, on peut avoir recours à l'échantillonnage à deux phases. C'est à ce type d'échantillonnage que nous nous intéressons dans ce mémoire. Celui-ci consiste à recueillir de l'information auxiliaire peu coûteuse pour un échantillon de la population (première phase), duquel on tire un sous-échantillon (deuxième phase). En fait, pour les unités appartenant à l'échantillon de première phase, il est habituel de collecter de l'information auxiliaire peu coûteuse grâce à laquelle l'échantillon de deuxième phase est tiré. L'efficacité du plan de sondage en est ainsi accrue.

Deux estimateurs d'un total pouvant être utilisés dans le cas des plans de sondage à deux phases sont l'estimateur par double dilatation et l'estimateur de calage, qui s'écrivent comme une somme pondérée de la variable d'intérêt au niveau de l'échantillon. Le premier utilise un système de pondération basé uniquement sur les probabilités d'inclusion

des unités dans l'échantillon. Le second intègre en plus de l'information auxiliaire dans le but d'obtenir un estimateur plus efficace. Quelque soit l'estimateur choisi, l'estimation de la variance de l'estimateur peut cependant poser problème. En effet, il peut être difficile d'obtenir des estimateurs de variance puisqu'ils requièrent généralement les probabilités d'inclusion jointes dont le calcul peut être très fastidieux (voire impossible). Il nous apparaît donc utile de proposer des estimateurs de variance simplifiés qui auront l'avantage de ne pas dépendre des probabilités d'inclusion jointes et qui pourront être obtenus au moyen d'un logiciel d'estimation de variance pour les plans de sondage à une phase. Ce dernier aspect est particulièrement important en pratique.

La structure de ce mémoire est la suivante : au chapitre 2, nous présentons des notions qui préparent à l'objet principal du mémoire tels le plan de sondage à deux phases, l'estimateur par double dilatation, l'estimateur de calage ainsi que leurs estimateurs de variance usuels ; au chapitre 3, nous présentons des estimateurs simplifiés de la variance et étudions les conditions sous lesquelles ils sont valides ; finalement, au chapitre 4, nous présentons les résultats de quelques études par simulation.

CHAPITRE 2

NOTIONS PRÉLIMINAIRES D'ÉCHANTILLONNAGE

Considérons une population finie U de N éléments indexés par i $(1 \leq i \leq N)$. De cette population, un échantillon $s \subseteq U$ de n unités est tiré selon une méthode de sélection aléatoire. L'ensemble de tous les 2^N échantillons possibles est dénoté par Ω et inclut l'ensemble vide ainsi que U lui-même. Soit $p(\cdot)$ une fonction telle que $p(s)$ désigne la probabilité qu'un échantillon $s \in \Omega$ donné soit tiré. Puisque $p(s)$ est une distribution de probabilité, les propriétés suivantes sont satisfaites :

i. $p(s) \geq 0, \forall s \in \Omega$;

ii. $\sum_{s \in \Omega} p(s) = 1$.

Le couple $\big(\Omega, p(s)\big)$ est appelé le plan de sondage. Au plan de sondage sont associées les quantités $\pi_i = \mathrm{P}(i \in s)$ et $\pi_{ij} = \mathrm{P}(i \in s, j \in s)$, désignant respectivement la probabilité d'inclusion de l'unité i dans l'échantillon et la probabilité d'inclusion conjointe des unités i et j dans l'échantillon. On posera $\pi_{ii} = \pi_i$.

Soit y une variable d'intérêt recueillie pour toutes les unités échantillonnées. Dans ce mémoire, nous supposons que les erreurs non dues à l'échantillonnage (par exemple, les erreurs de non-réponse ou les erreurs de couverture) sont négligeables. En pratique, on peut vouloir estimer divers paramètres de la population tels le total ou la moyenne d'une variable d'intérêt ou encore un quantile, un ratio de deux totaux ou un coefficient de corrélation ou de régression. On peut également être intéressé à produire des estimations au niveau de domaines (sous-populations) particuliers. Dans ce mémoire, nous nous intéressons à l'estimation du total de la variable d'intérêt y dans la population :

$$t_y = \sum_{i \in U} y_i.$$

2.1 Quelques plans de sondage simples

Dans cette section, nous présentons quelques plans de sondage, que nous allons ensuite considérer au chapitre 3.

2.1.1 Échantillonnage aléatoire simple sans remise

Le plan de sondage aléatoire simple sans remise est l'un des plus simples. Ce plan de sondage considère tous les échantillons possibles de n unités distinctes. De plus, il accorde à chaque échantillon la même probabilité d'être tiré. Il y a donc $\binom{N}{n}$ échantillons possibles et la probabilité de sélection d'un échantillon $s \in \Omega$ de taille n est donnée par :

$$p(s) = \frac{1}{\binom{N}{n}}.$$

Pour le plan de sondage aléatoire simple sans remise, les probabilités d'inclusion sont données par $\pi_i = n/N$ et $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$ si $i \neq j$. Le plan de sondage aléatoire simple sans remise est dit à taille fixe puisque les échantillons tirés sont obligatoirement de taille n.

2.1.2 Échantillonnage de Poisson

Le plan de Poisson est un plan à taille aléatoire. Ce plan de sondage consiste à effectuer, pour chaque unité i de la population, une expérience de Bernoulli avec probabilité de succès π_i. Un succès correspond alors à la sélection de l'unité i dans l'échantillon. La sélection des unités dans l'échantillon est donc indépendante d'une unité à l'autre. Ainsi, la probabilité de sélection d'un échantillon $s \in \Omega$ est donnée par

$$p(s) = \prod_{i \in s} \pi_i \prod_{i \in U \setminus s} (1 - \pi_i).$$

Notons qu'il y a un total de 2^N échantillons possibles. De plus, on a $\pi_{ij} = \pi_i \pi_j$ lorsque $i \neq j$ étant donné l'indépendance dans la sélection des unités.

Si on pose $\pi_i = \pi$ pour tout $i \in U$, le plan de Poisson est également appelé plan de

Bernoulli.

2.1.3 Échantillonnage stratifié

L'échantillonnage stratifié consiste à partitionner la population en H strates $U_1, \ldots, U_H$ de tailles $N_1, \ldots, N_H$ respectivement. Dans chaque strate U_h, $h = 1, \ldots, H$, un échantillon s_h de taille n_h est tiré selon un plan de sondage $p_h(s_h)$. Ces échantillons sont tirés indépendamment d'une strate à l'autre. Il en résulte un échantillon $s = \bigcup_{h=1}^{H} s_h$ de taille $n = \sum_{h=1}^{H} n_h$. Compte tenu de la propriété d'indépendance, la probabilité de sélection d'un échantillon s est donnée par

$$p(s) = \prod_{h=1}^{H} p_h(s_h).$$

De par cette même propriété, on a $\pi_{ij} = \pi_i \pi_j$ lorsque i et j sont dans des strates différentes. Autrement, si i et j sont dans la même strate, les probabilités π_i et π_{ij} sont propres au plan de sondage utilisé dans la strate.

2.2 Estimateur par dilatation

L'estimateur par dilatation du total t_y est donné par :

$$\hat{t}_{y\pi} = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i \in s} d_i y_i, \tag{2.1}$$

où $d_i = {\pi_i}^{-1}$ désigne le poids de sondage de l'unité i. Cet estimateur est également connu sous le nom d'*estimateur d'Horvitz-Thompson* (Horvitz et Thompson, 1952).

Nous allons aborder les propriétés de l'estimateur par dilatation sous le plan de sondage p. Les opérateurs $\mathrm{E}_p(.)$ et $\mathrm{V}_p(.)$ seront utilisés pour désigner respectivement l'espérance et la variance sous le plan de sondage p.

Proposition 1. L'estimateur par dilatation $\hat{t}_{y\pi}$ est sans biais par rapport au plan de sondage p pour le total t_y.

Démonstration. Afin de le démontrer, il suffit d'introduire une variable indicatrice de sélection I_i prenant la valeur 1 si $i \in s$ et 0, sinon. Ainsi,

$$\mathrm{E}_p(\hat{t}_{y\pi}) = \mathrm{E}_p\left(\sum_{i\in U}\frac{1}{\pi_i}y_i I_i\right) = \sum_{i\in U}\frac{1}{\pi_i}y_i\,\mathrm{E}_p(I_i) = \sum_{i\in U} y_i = t_y.$$

□

Proposition 2. La variance de l'estimateur (2.1) par rapport au plan de sondage p est donnée par

$$\mathrm{V}_p\left(\hat{t}_{y\pi}\right) = \sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}. \tag{2.2}$$

Démonstration. Définissons l'opérateur $\mathrm{Cov}_p(\cdot,\cdot)$ qui désigne la covariance par rapport au plan de sondage p. Ainsi,

$$\begin{aligned}
\mathrm{V}_p(\hat{t}_{y\pi}) &= \sum_{i\in U}\mathrm{V}_p\left(\frac{1}{\pi_i}y_i I_i\right) + \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\mathrm{Cov}_p\left(\frac{1}{\pi_i}y_i I_i, \frac{1}{\pi_j}y_j I_j\right)\\
&= \sum_{i\in U}\frac{1}{\pi_i^2}y_i^2\,\mathrm{V}_p(I_i) + \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\mathrm{Cov}_p(I_i,I_j)\\
&= \sum_{i\in U}\frac{y_i^2}{\pi_i^2}\pi_i(1-\pi_i) + \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{ij}-\pi_i\pi_j)\\
&= \sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}.
\end{aligned}$$

Notons que $\mathrm{V}_p(I_i) = \pi_i(1-\pi_i)$ et $\mathrm{Cov}(I_i,I_j) = \pi_{ij}-\pi_i\pi_j$ si $i \neq j$. □

Proposition 3. Si $\pi_{ij} > 0$ pour tout (i,j), un estimateur sans biais de la variance (2.2) est donné par

$$\widehat{\mathrm{V}}_p\left(\hat{t}_{y\pi}\right) = \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}. \tag{2.3}$$

Démonstration. Nous allons montrer que $\widehat{\mathrm{V}}_p(\hat{t}_{y\pi})$ donné par (2.3) est sans biais pour $\mathrm{V}_p(\hat{t}_{y\pi})$.

$$\begin{aligned}
\mathrm{E}_p\left(\widehat{\mathrm{V}}_p(\hat{t}_{y\pi})\right) &= \mathrm{E}_p\left(\sum_{i\in U}\sum_{j\in U}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}I_iI_j\right)\\
&= \sum_{i\in U}\sum_{j\in U}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\mathrm{E}_p(I_iI_j)\\
&= \sum_{i\in U}\sum_{j\in U}\left\{\pi_{ij}-\pi_i\pi_j\right\}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\\
&= \mathrm{V}_p(\hat{t}_{y\pi}).
\end{aligned}$$

Notons que $\mathrm{E}_p(I_iI_j)=\pi_{ij}$ pour tout $i,j\in U$. □

Si le plan de sondage est un plan aléatoire simple sans remise, l'expression de la variance (2.2) devient

$$\mathrm{V}_p(\hat{t}_{y\pi}) = N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}, \tag{2.4}$$

où $S_y^2 \equiv \sum_{i\in U}(y_i-\bar{y}_U)^2/(N-1)$ et $\bar{y}_U \equiv t_y/N$. La variance de l'estimateur $\hat{t}_{y\pi}$ sera donc petite si la fraction de sondage n/N est grande et/ou la taille de l'échantillon n est grande et/ou si la dispersion de la variable dans la population est petite. De plus, l'expression de l'estimateur de variance (2.3) devient

$$\widehat{\mathrm{V}}_p(\hat{t}_{y\pi}) = N^2\left(1-\frac{n}{N}\right)\frac{s_y^2}{n}, \tag{2.5}$$

où $s_y^2 \equiv \sum_{i\in s}(y_i-\bar{y}_s)^2/(n-1)$ et $\bar{y}_s \equiv \sum_{i\in s} y_i/n$.

Si le plan de sondage est un plan de Poisson, l'expression de la variance (2.2) devient

$$\mathrm{V}_p(\hat{t}_{y\pi}) = \sum_{i\in U}\frac{1-\pi_i}{\pi_i}y_i^2 \tag{2.6}$$

et l'expression de l'estimateur de variance (2.3) devient

$$\widehat{\mathrm{V}}_p(\hat{t}_{y\pi}) = \sum_{i\in s} \frac{1-\pi_i}{\pi_i^2} y_i^2. \tag{2.7}$$

2.3 Estimateur de calage

Il est possible d'utiliser de l'information auxiliaire à l'étape de l'estimation dans le but d'améliorer la précision de l'estimation. Le calage permet d'incorporer une information auxiliaire aux estimateurs. Soit $\boldsymbol{x}_i = (x_{1i},...,x_{Ji})'$ un vecteur de dimension J des variables auxiliaires recueillies pour l'unité $i \in s$ et dont le total au niveau de la population, $\boldsymbol{t}_x = \sum_{i\in U} \boldsymbol{x}_i$, est connu. L'estimateur de calage s'exprime comme une somme pondérée des y_i :

$$\hat{t}_{yC} = \sum_{i\in s} w_i y_i, \tag{2.8}$$

où w_i désigne le poids de calage associé à l'unité i.

L'objectif est de déterminer des poids de calage w_i qui satisfont la contrainte

$$\hat{t}_{xC} = \sum_{i\in s} w_i \boldsymbol{x}_i = \boldsymbol{t}_x, \tag{2.9}$$

signifiant que lorsque le système de pondération $\{w_i; i \in s\}$ est appliqué au vecteur des variables auxiliaires $\boldsymbol{x}$, on retrouve le vecteur des vrais totaux dans la population $\boldsymbol{t}_x$. Le respect de la contrainte (2.9) garantit une certaine cohérence entre les estimations issues de l'enquête et les totaux connus au niveau de la population.

Les poids de calage w_i sont choisis de façon à minimiser la distance avec les poids d_i qui sont utilisés dans l'estimateur par dilatation (2.1) qui, rappelons-le, a la propriété d'être sans biais. La distance entre les poids w_i et d_i est mesurée par une fonction $G(w_i/d_i)$ qui satisfait aux conditions suivantes :

i. $G(w_i/d_i) \geq 0$ et $G(1) = 0$;

ii. $\frac{\partial G(w_i/d_i)}{\partial w_i}$ existe, est continue et bijective (donc inversible) ;

iii. $G(w_i/d_i)$ est strictement convexe.

Plus précisément, les poids w_i sont obtenus en minimisant

$$\sum_{i\in s} \frac{d_i G(w_i/d_i)}{q_i}, \tag{2.10}$$

sous la contrainte (2.9), où q_i représente un poids associé à l'unité i. Comme nous le verrons à la section 2.3.2, le coefficient q_i est lié à la structure de variance du modèle reliant la variable d'intérêt y au vecteur des variables auxiliaires $\boldsymbol{x}$.

En utilisant la méthode des multiplicateurs de Lagrange, nous cherchons les poids w_i qui minimisent

$$\sum_{i\in s} \frac{d_i G(w_i/d_i)}{q_i} - \boldsymbol{\lambda}' \left(\sum_{i\in s} w_i \boldsymbol{x}_i - \boldsymbol{t}_x \right), \tag{2.11}$$

où $\boldsymbol{\lambda}$ est un vecteur de J multiplicateurs de Lagrange et l'opérateur « $'$ » désigne la transposée. En dérivant (2.11) par rapport à w_i, on obtient :

$$w_i = d_i F(q_i \boldsymbol{\lambda}' \boldsymbol{x}_i), \tag{2.12}$$

où $F(.)$ représente la fonction inverse de $G'(.) \equiv \frac{\partial G(.)}{\partial .}$. On remarque que les poids de calage w_i s'écrivent comme le produit du poids de sondage d_i et d'un facteur d'ajustement $F_i = F(q_i \boldsymbol{\lambda}' \boldsymbol{x}_i)$. Il reste à déterminer $\boldsymbol{\lambda}$ en substituant les poids w_i de (2.12) dans la contrainte (2.9) qui devient

$$\sum_{i\in s} d_i F(q_i \boldsymbol{\lambda}' \boldsymbol{x}_i) \boldsymbol{x}_i = \boldsymbol{t}_x. \tag{2.13}$$

Dans certains cas, le système d'équations (2.13) possède une solution explicite pour $\boldsymbol{\lambda}$. Dans le cas contraire, un algorithme de localisation des zéros d'une fonction tel que la méthode de Newton-Raphson peut être utilisé.

Lorsque la fonction de distance utilisée est celle des moindres carrés généralisés

$$G(w_i/d_i) = \frac{1}{2}\left(\frac{w_i}{d_i} - 1\right)^2, \tag{2.14}$$

on a $F(u) = 1 + u$ et les poids w_i en (2.12) deviennent

$$w_i = d_i(1 + q_i \boldsymbol{\lambda}' \boldsymbol{x}_i). \tag{2.15}$$

En utilisant (2.15) dans (2.9), on obtient une solution explicite pour $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = \left(\sum_{i \in s} d_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi}). \tag{2.16}$$

Finalement, en combinant (2.15) et (2.16), l'estimateur de calage (2.8) s'écrit comme

$$\hat{t}_{yC} = \hat{t}_{y\pi} + \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi}\right)' \widehat{\boldsymbol{B}}, \tag{2.17}$$

où

$$\widehat{\boldsymbol{B}} = \left(\sum_{i \in s} d_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i \in s} d_i q_i \boldsymbol{x}_i y_i\right). \tag{2.18}$$

Comme nous le verrons à la section 2.3.2, l'estimateur (2.17) peut être obtenu alternativement en supposant un modèle de régression reliant la variable d'intérêt y au vecteur des variables auxiliaires $\boldsymbol{x}$.

Une autre fonction de distance fréquemment utilisée en pratique est la distance *raking ratio*, définie par

$$G(w_i/d_i) = \left(\frac{w_i}{d_i}\right) \log\left(\frac{w_i}{d_i}\right) - \left(\frac{w_i}{d_i}\right) + 1.$$

Dans ce cas, on a $F(u) = e^u$ et l'équation de calage (2.13) s'écrit comme

$$\sum_{i \in s} d_i e^{(q_i \boldsymbol{\lambda}' \boldsymbol{x}_i)} \boldsymbol{x}_i = \boldsymbol{t}_x. \tag{2.19}$$

La résolution du système (2.19) requiert un algorithme de localisation des zéros d'une fonction comme l'algorithme de Newton-Raphson déjà mentionné.

Remarque 2.1. *La distance des moindres carrés généralisés peut mener à des poids de calage w_i négatifs. La méthode du raking ratio, quant à elle, garantit que les poids w_i sont positifs. Par contre, certains poids peuvent être très grands, ce qui tend à accroître l'instabilité de l'estimateur de calage résultant.*

Remarque 2.2. *Deville et Särndal (1992) ont proposé plusieurs autres fonctions de distance, dont certaines permettent de contrôler la dispersion des poids de calage.*

2.3.1 Propriétés de l'estimateur de calage

Les propriétés de l'estimateur de calage peuvent être étudiées en utilisant le cadre asymptotique décrit à l'annexe I.2, où la taille de la population N est considérée comme fixe. Sous certaines conditions de régularité, Deville et Särndal (1992) ont montré que

$$\frac{1}{N}\left(\hat{t}_{yC} - \hat{t}_{y\pi}\right) = O_p(1/\sqrt{n}).$$

Autrement dit, l'estimateur de calage $\hat{t}_{yC}$ est convergent sous le plan de sondage. Ils ont également montré, sous certaines conditions de régularité, que la variance de $\hat{t}_{yC}$ peut être approximée par

$$\mathrm{V}_p\left(\hat{t}_{yC}\right) \approx \sum_{i\in U}\sum_{j\in U}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{E_i}{\pi_i}\frac{E_j}{\pi_j}, \tag{2.20}$$

où $E_i = y_i - \boldsymbol{x}_i'\boldsymbol{B}$ et

$$\boldsymbol{B} = \left(\sum_{i\in U} q_i\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i\in U} q_i\boldsymbol{x}_i y_i\right).$$

Comme nous le verrons à la section 2.3.2, le coefficient $\boldsymbol{B}$ peut être vu comme l'estimateur des moindres carrés généralisés du vecteur des paramètres dans un modèle de régression reliant la variable d'intérêt y au vecteur des variables auxiliaires $\boldsymbol{x}$. De plus,

E_i peut être vu comme un résidu de la régression. La variance (2.20) sera petite lorsque les résidus E_i sont petits, ce qui survient lorsque le modèle ajuste bien les données.

Le résultat (2.20) est remarquable car il stipule que, quelque soit la fonction de distance G utilisée, la variance de l'estimateur $\hat{t}_{yC}$ est approximativement égale à celle de l'estimateur de calage dans le cas où G est donnée par la distance des moindres carrés généralisés. En effet, les propriétés de l'estimateur (2.17) peuvent être étudiées à l'aide d'une linéarisation permettant d'exprimer asymptotiquement l'estimateur (2.17) comme une fonction linéaire d'estimateurs de totaux. À cette fin, un développement en série de Taylor (Woodruff, 1971) ou la méthode de Demnati-Rao (Demnati et Rao, 2004) peuvent être utilisés mais demandent d'évaluer une série de dérivées partielles. Ici, nous choisissons de présenter la méthode de linéarisation automatisée de Esteavo et Särndal (2006) qui est plus directe. On écrit

$$\begin{aligned}\hat{t}_{yC} - t_y &= (\hat{t}_{y\pi} - t_y) + \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi}\right)'(\widehat{\boldsymbol{B}} - \boldsymbol{B} + \boldsymbol{B}), \\ &= (\hat{t}_{y\pi} - t_y) + \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi}\right)'\boldsymbol{B} + \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi}\right)'(\widehat{\boldsymbol{B}} - \boldsymbol{B}). \end{aligned} \tag{2.21}$$

Ainsi, sous les conditions usuelles (Isaki et Fuller, 1982),

$$N^{-1}\left(\hat{t}_{y\pi} - t_y\right) = O_p(1/\sqrt{n}),$$

$$N^{-1}\left(\hat{\boldsymbol{t}}_{x\pi} - \boldsymbol{t}_x\right) = O_p(1/\sqrt{n}),$$

$$\widehat{\boldsymbol{B}} - \boldsymbol{B} = O_p(1/\sqrt{n}),$$

le dernier terme de (2.21) est négligeable par rapport aux deux premiers et on a

$$\frac{1}{N}\left(\hat{t}_{yC} - t_y\right) = \frac{1}{N}\sum_{i\in s} d_i E_i - \frac{1}{N}\sum_{i\in U} E_i + O_p(n^{-1}). \tag{2.22}$$

En ignorant les termes d'ordre $O_p(n^{-1})$ dans (2.22) la variance de $\hat{t}_{yC}$ est obtenue comme

suit :

$$\begin{aligned} \mathrm{V}_p\left(\hat{t}_{yC}\right) &\approx \mathrm{V}_p\left(\sum_{i\in s} d_i E_i\right), \\ &= \sum_{i\in U}\sum_{j\in U}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{E_i}{\pi_i}\frac{E_j}{\pi_j}, \end{aligned}$$

et correspond bien à l'expression (2.20).

Remarque 2.3. *La variance asymptotique (2.20) peut être estimée sans biais par :*

$$\sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{E_i}{\pi_i}\frac{E_j}{\pi_j}. \tag{2.23}$$

Toutefois, cette dernière quantité ne peut pas être calculée en pratique, car le calcul des E_i requiert que les valeurs de $\boldsymbol{x}_i$ et y_i soient disponibles pour toutes les unités de la population. L'estimateur usuel approximativement sans biais de la variance asymptotique (2.20) est obtenu en remplaçant dans (2.23) *les quantités E_i par les quantités $e_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{B}}$, qui peuvent être calculées à partir des valeurs recueillies au niveau de l'échantillon. L'estimateur est donné par*

$$\widehat{\mathrm{V}}_p\left(\hat{t}_{yC}\right) = \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{e_i}{\pi_i}\frac{e_j}{\pi_j}.$$

Remarque 2.4. *Dans le cas où q_i est de la forme ${q_i}^{-1} = \boldsymbol{\alpha}'\boldsymbol{x}_i$ pour un vecteur $\boldsymbol{\alpha}$ connu, on peut montrer que*

$$\sum_{i\in U} E_i = \sum_{i\in s} d_i e_i = 0;$$

voir annexe I.1 pour la preuve.

2.3.2 Construction de l'estimateur de calage par la régression généralisée

Dans cette section, nous présentons une manière alternative d'obtenir l'estimateur de calage (2.17) en utilisant une approche par la régression. Supposons le modèle de

superpopulation suivant :

$$m : y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{2.24}$$

où $\boldsymbol{\beta}$ est un vecteur de J paramètres inconnus. Soit $\mathrm{E}_m(.)$ et $\mathrm{V}_m(.)$ les opérateurs désignant l'espérance et la variance sous le modèle m. On suppose que $\mathrm{E}_m\left(\varepsilon_i\right) = 0, \mathrm{E}_m\left(\varepsilon_i\varepsilon_j\right) = 0$ pour $i \neq j$ et $\mathrm{V}_m\left(\varepsilon_i\right) = \sigma^2 c_i$. Le coefficient c_i associé à l'unité i est supposé connu. Le modèle (2.24) suppose que la population U à l'étude est tirée d'une superpopulation infinie hypothétique.

Si les variables d'intérêt étaient connues pour toutes les unités de la population (cas d'un recensement), on pourrait estimer $\boldsymbol{\beta}$ par son estimateur des moindres carrés pondérés :

$$\boldsymbol{B}_G = \left(\sum_{i\in U} {c_i}^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i\in U} {c_i}^{-1}\boldsymbol{x}_i y_i\right) \tag{2.25}$$

et les résidus de la régression seraient

$$E_{Gi} = y_i - \boldsymbol{x}_i'\boldsymbol{B}_G.$$

Toutefois, l'estimateur $\boldsymbol{B}_G$ en (2.25) ne peut être calculé car la variable d'intérêt y n'est observée que pour $i \in s$. Un estimateur de $\boldsymbol{B}_G$ basé sur les valeurs de l'échantillon est donné par

$$\widehat{\boldsymbol{B}}_G = \left(\sum_{i\in s} {c_i}^{-1} d_i\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i\in s} {c_i}^{-1} d_i\boldsymbol{x}_i y_i\right).$$

L'estimateur $\widehat{\boldsymbol{B}}_G$ est asymptotiquement sans biais pour $\boldsymbol{B}$ par rapport au plan de sondage. Les résidus sont donnés par

$$e_{Gi} = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{B}}_G.$$

On obtient un estimateur du total en décomposant t_y comme

$$t_y = \sum_{i\in U} y_i$$

$$= \sum_{i \in U} \boldsymbol{x}_i' \boldsymbol{\beta} + \sum_{i \in U} \varepsilon_i,$$

puis en estimant chaque terme séparément, ce qui mène à

$$\hat{t}_{yG} = \sum_{i \in U} \boldsymbol{x}_i' \widehat{\boldsymbol{B}}_G + \sum_{i \in s} d_i e_{Gi}. \tag{2.26}$$

L'estimateur $\hat{t}_{yG}$ en (2.26) est connu sous le nom d'estimateur par la régression généralisée ou GREG (*Generalized REGression estimator*). L'estimateur (2.26) peut être réécrit comme la somme de l'estimateur de dilatation de t_y et d'un terme d'ajustement. En effet,

$$\begin{aligned}
\hat{t}_{yG} &= \sum_{i \in U} \boldsymbol{x}_i' \widehat{\boldsymbol{B}}_G + \sum_{i \in s} d_i e_{Gi} \\
&= \sum_{i \in U} \boldsymbol{x}_i' \widehat{\boldsymbol{B}}_G + \sum_{i \in s} d_i y_i - \sum_{i \in s} d_i \boldsymbol{x}_i' \widehat{\boldsymbol{B}}_G \\
&= \hat{\boldsymbol{t}}_x' \widehat{\boldsymbol{B}}_G + \hat{t}_{y\pi} - \hat{\boldsymbol{t}}_{x\pi}' \widehat{\boldsymbol{B}}_G \\
&= \hat{t}_{y\pi} + \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi}\right)' \widehat{\boldsymbol{B}}_G.
\end{aligned}$$

L'estimateur (2.26) est donc équivalent à l'estimateur de calage (2.17), qui est basé sur la fonction de distance des moindres carrés, en posant $q_i = {c_i}^{-1}$.

Remarque 2.5. *L'estimateur par la régression généralisée est asymptotiquement sans biais, peu importe si le modèle (2.24) est bien spécifié. Sa variance asymptotique est donnée par (2.20) en posant* $q_i = {c_i}^{-1}$.

Remarque 2.6. *Un cas particulier de l'estimateur (2.26) lorsqu'il n'y a qu'une seule variable auxiliaire* x_i *et que* $c_i = x_i$ *est l'estimateur par le ratio,*

$$\hat{t}_{yR} = \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}} t_x.$$

Un cas particulier de l'estimateur par le ratio est l'estimateur de Hájek,

$$\hat{t}_{yH} = \frac{\hat{t}_{y\pi}}{\widehat{N}_\pi} N, \tag{2.27}$$

qui est obtenu en posant $x_i = 1$ *pour tout* $i \in U$, *où* $\widehat{N}_\pi = \sum_{i \in s} d_i$.

Ces deux estimateurs sont des cas particuliers d'estimateurs de calage qui satisfont la condition $q_i^{-1} = \boldsymbol{\alpha}' \boldsymbol{x}_i$ *(voir remarque 2.4) en posant* $\boldsymbol{\alpha} = 1$.

Remarque 2.7. *Si le vecteur des variables auxiliaires contient une ordonnée à l'origine, i.e.* $\boldsymbol{x}_i = \big(1, x_{1i}, \ldots, x_{iJ}\big)'$ *et que* $q_i = 1$ *pour tout* $i \in U$ *alors cette situation est un cas particulier pour lequel* $q_i^{-1} = \boldsymbol{\alpha}' \boldsymbol{x}_i$ *en posant* $\boldsymbol{\alpha} = \big(1, 0, \ldots, 0\big)'$.

2.4 Échantillonnage à deux phases

En pratique, il arrive que peu ou pas d'information auxiliaire pertinente ne soit disponible sur la base de sondage. En l'absence d'information auxiliaire, il est donc difficile d'utiliser des plans de sondage efficaces s'appuyant sur une information auxiliaire appropriée ; par exemple un plan stratifié ou proportionnel à la taille (PPT).

Dans une telle situation, il est possible d'utiliser un plan de sondage à deux phases. D'abord, une première phase permet de recueillir de l'information auxiliaire peu dispendieuse pour les unités d'un premier échantillon $s_1 \subseteq U$ de taille n_1 sélectionné selon un plan de sondage $p_1(s_1)$. Définissons $\boldsymbol{I}_1 = (I_{11}, \ldots, I_{1N})'$ le vecteur des variables indicatrices de sélection des unités dans l'échantillon s_1. Puis, en seconde phase, les variables d'intérêt sont recueillies pour les unités d'un sous-échantillon $s_2 \subseteq s_1 \subseteq U$ de taille n_2 tiré selon un plan de sondage $p_2(s_2 \mid \boldsymbol{I}_1)$ généralement plus complexe car pouvant utiliser l'information auxiliaire recueillie à la première phase. Définissons également $\boldsymbol{I}_2 = (I_{21}, \ldots, I_{2N})'$ le vecteur des variables indicatrices de sélection des unités dans l'échantillon s_2.

Les probabilités d'inclusion de l'unité i dans chacun des échantillons seront notées $\pi_{1i} = \mathrm{P}(I_{1i} = 1)$, la probabilité que l'unité i soit dans l'échantillon s_1 et $\pi_{2i}(\boldsymbol{I}_1) = \mathrm{P}(I_{2i} = 1 \mid \boldsymbol{I}_1)$, la probabilité que l'unité i soit dans l'échantillon s_2 conditionnelle aux unités sélectionnées à la première phase. De même, les probabilités d'inclusion conjointe des unités i et j dans chacun des échantillons seront notées $\pi_{1ij} = \mathrm{P}(I_{1i} = 1, I_{1j} = 1)$ et $\pi_{2ij}(\boldsymbol{I}_1) = \mathrm{P}(I_{2i} = 1, I_{2j} = 1 \mid \boldsymbol{I}_1)$. Définissons également $\pi_i^* = \pi_{1i}\pi_{2i}(\boldsymbol{I}_1)$ et $\pi_{ij}^* = \pi_{1ij}\pi_{2ij}(\boldsymbol{I}_1)$. Il est important de remarquer que les quantités $\pi_{2i}(\boldsymbol{I}_1)$ et $\pi_{2ij}(\boldsymbol{I}_1)$ sont des variables aléatoires puisque le plan de sondage en deuxième phase dépend généralement de l'échantillon observé en première phase. Finalement, on définit les quantités $d_{1i} = \pi_{1i}^{-1}$, $d_{2i}(\boldsymbol{I}_1) = \pi_{2i}(\boldsymbol{I}_1)^{-1}$ et $d_i^* = \pi_i^{*-1}$.

2.4.1 Invariance et indépendance

Deux concepts importants en présence d'échantillonnage à deux phases sont l'invariance et l'indépendance. La propriété d'invariance requiert que la sélection à la deuxième phase ne dépende pas du résultat de la sélection à la première phase, ce qui se traduit par

$$p_2(\boldsymbol{I}_2 \mid \boldsymbol{I}_1) = p_2(\boldsymbol{I}_2).$$

Lorsqu'il y a invariance, on a $\pi_{2i}(\boldsymbol{I}_1) = \pi_{2i}$ et $\pi_{2ij}(\boldsymbol{I}_1) = \pi_{2ij}$. Des exemples de plans de sondages de deuxième phase respectant la propriété d'invariance sont : un plan aléatoire simple sans remise à la deuxième phase pour lequel la fraction de sondage n_2/n_1 ne dépend pas du résultat de l'échantillonnage de première phase ou encore un plan de Poisson où les π_i sont tous égaux (plan de Bernouilli) et indépendants du résultat de l'échantillonnage de première phase.

Pour sa part, la propriété d'indépendance requiert, qu'à la deuxième phase, les unités soient sélectionnées de manière à ce que les variables indicatrices I_{2i} et I_{2j} soient indépendantes si $i \neq j$. Lorsqu'il y a indépendance, on a $\pi_{2ij}(\boldsymbol{I}_1) = \pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)$ lorsque $i \neq j$. Un exemple de plan de sondage de deuxième phase respectant la propriété d'indépendance est le plan de Poisson.

Le plan de sondage à deux degrés qui sera décrit à la section suivante satisfait à la fois à la propriété d'invariance et d'indépendance.

2.4.2 Échantillonnage à deux degrés

Un cas particulier de l'échantillonnage à deux phases est l'échantillonnage à deux degrés, qui est fréquemment utilisé en pratique dans le contexte des enquêtes auprès des ménages. L'échantillonnage à deux degrés peut être utilisé lorsqu'il n'existe pas de base de sondage des éléments sur lequel les mesures seront prises, mais qu'il existe une base de sondage pour des regroupements de ces éléments. Prenons l'exemple d'une enquête s'intéressant aux élèves de niveau secondaire de la province du Québec. Il est fort probable qu'il n'existe pas une base de sondage des élèves à l'échelle de la province. On pourrait en construire une mais cela s'avérerait prohibitif. Par contre, s'il existe une base de sondage des écoles de la province, cette dernière peut être utilisée afin de tirer préalablement un échantillon d'écoles, qui seront contactées afin que dans chaque école, un échantillon d'élèves puisse être sélectionné. Dans cet exemple, les écoles sont appelées les unités primaires d'échantillonnage (UPE) ou les grappes et les élèves sont appelés les unités secondaires d'échantillonnage (USE).

Nous allons maintenant décrire l'échantillonnage à deux degrés de manière générale et introduire une notation qui lui est propre et qui sera utilisée à la section 3.1.2. Considérons une population composée de N UPE, $U_1, \ldots, U_N$ de taille $M_1, \ldots, M_N$ USE respectivement. Le premier degré consiste à tirer à partir de cette population un échantillon, s, de taille n UPE selon un plan de sondage donné. Puis dans chaque UPE g de s, on tire un échantillon, s_g, de taille m_g USE selon un plan de sondage sonné. Soit y_{gi} la valeur d'une variable d'intérêt y pour le i^{e} élément dans la g^{e} UPE et soit s' l'échantillon d'UPE pour lesquelles au moins une USE a été sélectionnée.

On désignera par $\pi_{1g} = \mathrm{P}(g \in s)$ et $\pi_{1gh} = \mathrm{P}(g \in s, h \in s)$ les probabilités d'inclusion simple et double des UPE au premier degré. On désignera également par $\pi_{2gi} = \mathrm{P}(i \in s_g)$

et $\pi_{2gihj} = \mathrm{P}(i \in s_g, j \in s_h)$ les probabilités d'inclusion simple et double des USE au second degré.

Remarque 2.8. *Le plan de sondage à deux degrés satisfait habituellement les propriétés d'indépendance et d'invariance. En effet, les échantillons s_g sont généralement tirés indépendamment d'une grappe à l'autre. De plus, la stratégie d'échantillonnage utilisée au deuxième degré ne dépend généralement pas de ce qui a été obtenu au premier degré.*

Remarque 2.9. *L'échantillonnage à deux degrés peut être vu comme un cas particulier de l'échantillonnage à deux phases. En effet, l'ensemble des USE tirées lors du tirage des UPE au premier degré peut être vu comme l'échantillon de première phase. L'ensemble des USE tirées au deuxième degré peut être vu comme l'échantillon de deuxième phase obtenu au moyen d'un plan stratifié pour lequel les strates sont définies par les UPE tirées au premier degré.*

2.5 Estimateur par double dilatation

Afin d'estimer le total $t_y = \sum_{i \in U} y_i$ dans le cadre d'un plan à deux phases, un estimateur, appelé estimateur par double dilatation, est donné par :

$$\hat{t}^*_{y\pi} = \sum_{i \in s_2} \frac{1}{\pi^*_i} y_i = \sum_{i \in s_2} d^*_i y_i. \tag{2.28}$$

Proposition 4. L'estimateur par double dilatation est sans biais pour t_y sous le plan, c'est-à-dire $\mathrm{E}_p\left(\hat{t}^*_{y\pi}\right) \equiv \mathrm{E}_1 \mathrm{E}_2\left(\hat{t}^*_{y\pi} \mid \boldsymbol{I}_1\right) = t_y$, où $\mathrm{E}_1(\cdot)$ et $\mathrm{E}_2(\cdot \mid \boldsymbol{I}_1)$ désignent respectivement l'espérance par rapport au plan de sondage $p_1(s_1)$ et $p_2(s_2 \mid \boldsymbol{I}_1)$.

Démonstration.

$$\begin{aligned}
\mathrm{E}_p\left(\hat{t}^*_{y\pi}\right) &= \mathrm{E}_1 \mathrm{E}_2\left(\sum_{i \in U} d_{1i} d_{2i}(\boldsymbol{I}_1) y_i I_{1i} I_{2i} \,\middle|\, \boldsymbol{I}_1\right) \\
&= \mathrm{E}_1\left(\sum_{i \in U} d_{1i} d_{2i}(\boldsymbol{I}_1) y_i I_{1i} \mathrm{E}_2(I_{2i} \mid \boldsymbol{I}_1)\right) \\
&= \mathrm{E}_1\left(\sum_{i \in U} d_{1i} y_i I_{1i}\right)
\end{aligned}$$

$$= \sum_{i \in U} d_{1i} y_i \mathrm{E}_1(I_{1i})$$
$$= t_y.$$

□

Le développement de l'expression de la variance de l'estimateur passe par la décomposition classique de l'erreur totale, $\hat{t}^*_{y\pi} - t_y$, en la somme des erreurs dues à chaque phase :

$$\hat{t}^*_{y\pi} - t_y = \underbrace{\left\{\hat{t}^1_{y\pi} - t_y\right\}}_{\substack{\text{erreur due à} \\ \text{la première phase}}} + \underbrace{\left\{\hat{t}^*_{y\pi} - \hat{t}^1_{y\pi}\right\}}_{\substack{\text{erreur due à} \\ \text{la deuxième phase}}},$$

où

$$\hat{t}^1_{y\pi} = \mathrm{E}_2\left(\hat{t}^*_{y\pi} \mid \boldsymbol{I}_1\right) = \sum_{i \in s_1} d_{1i} y_i$$

constitue l'estimateur qui aurait été utilisé s'il n'y avait qu'une seule phase d'échantillonnage. Dénotons par $\mathrm{V}_1(\cdot)$ et $\mathrm{V}_2(\cdot \mid \boldsymbol{I}_1)$ la variance par rapport au plan de sondage $p_1(\boldsymbol{I}_1)$ et $p_2(s_2 \mid \boldsymbol{I}_1)$, respectivement. La variance de $\hat{t}^*_{y\pi}$ peut s'écrire comme la somme de deux termes :

$$\begin{aligned}
\mathrm{V}_p\left(\hat{t}^*_{y\pi}\right) &= \mathrm{E}_p\left(\hat{t}^*_{y\pi} - t_y\right)^2 \\
&= \mathrm{E}_1\,\mathrm{E}_2\left(\hat{t}^*_{y\pi} - t_y \,\middle|\, \boldsymbol{I}_1\right)^2 \\
&= \mathrm{E}_1\,\mathrm{E}_2\left(\left\{\hat{t}^1_y - t_y\right\} + \left\{\hat{t}^*_{y\pi} - \hat{t}^1_y\right\} \,\middle|\, \boldsymbol{I}_1\right)^2 \\
&= \mathrm{E}_1\,\mathrm{E}_2\left(\hat{t}^1_y - t_y \,\middle|\, \boldsymbol{I}_1\right)^2 + \mathrm{E}_1\,\mathrm{E}_2\left(\hat{t}^*_{y\pi} - \hat{t}^1_y \,\middle|\, \boldsymbol{I}_1\right)^2 \\
&= \mathrm{V}_1\,\mathrm{E}_2\left(\hat{t}^*_{y\pi} \,\middle|\, \boldsymbol{I}_1\right) + \mathrm{E}_1\,\mathrm{V}_2\left(\hat{t}^*_{y\pi} \,\middle|\, \boldsymbol{I}_1\right), \qquad (2.29)
\end{aligned}$$

où

$$\mathrm{V}_1\,\mathrm{E}_2\left(\hat{t}^*_{y\pi} \,\middle|\, \boldsymbol{I}_1\right) = \mathrm{V}_1\left(\hat{t}_{y1}\right) = \sum_{i \in U} \sum_{j \in U} (\pi_{1ij} - \pi_{1i}\pi_{1j}) \frac{y_i}{\pi_{1i}} \frac{y_j}{\pi_{1j}}$$

et

$$\mathrm{E}_1\,\mathrm{V}_2\left(\hat{t}^*_{y\pi} \mid s_1\right) = \mathrm{E}_1\left(\sum_{i \in s_1} \sum_{j \in s_1} \left\{\pi_{2ij}(s_1) - \pi_{2i}(s_1)\pi_{2j}(s_1)\right\} \frac{y_i}{\pi^*_i} \frac{y_j}{\pi^*_j}\right),$$

en utilisant un développement analogue à celui dans la preuve (2.2) de la variance de $\hat{t}_{y\pi}$. Ces deux quantités représentent respectivement la contribution de la première et de la deuxième phase d'échantillonnage à la variance totale.

Remarque 2.10. *L'estimateur usuel sans biais de* $\mathrm{V}_p\left(t^*_{y\pi}\right)$ *est donné par*

$$\widehat{\mathrm{V}}_p\left(\hat{t}^*_{y\pi}\right) = \sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi^*_{ij}}\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1j}} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{2ij}(\boldsymbol{I}_1)-\pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)}{\pi_{2ij}(\boldsymbol{I}_1)}\frac{y_i}{\pi^*_i}\frac{y_j}{\pi^*_j}, \tag{2.30}$$

où les premier et deuxième termes estiment sans biais respectivement les premier et deuxième termes de la variance (2.29)*.*

2.6 Estimateurs de calage pour des plans à deux phases

Dans cette section, nous présentons des estimateurs de calage pour les plans de sondage à deux phases. Ces estimateurs ont été traités, entre autres, par Dupont (1995), Hidiroglou et Särndal (1998) et Esteavo et Särndal (2006).

Dans le contexte de l'échantillonnage à deux phases, nous distinguons deux niveaux d'information auxiliaire :

(1) les variables auxiliaires dont le total est connu au niveau de la population ;

(2) les variables auxiliaires disponibles pour toutes les unités à la première phase uniquement.

Soit $\boldsymbol{x}_{1i}$ le vecteur des J_1 variables auxiliaires de type (1) et soit $\boldsymbol{x}_{2i}$ le vecteur des J_2 variables auxiliaires de type (2).

2.6.1 Information auxiliaire de type (1) seulement

Dans cette section, nous considérons le cas où seulement de l'information auxiliaire de type (1) est utilisée au moment de l'estimation. L'estimateur de calage s'exprime comme une somme pondérée des y_i :

$$\hat{t}^*_{yC1} = \sum_{i\in s_2} w^*_i y_i, \tag{2.31}$$

où w_i^* désigne le poids de calage associé à l'unité i.

Les poids de calage w_i^* permettent de satisfaire la contrainte suivante sur l'information auxiliaire disponible au niveau de la population :

$$\sum_{i \in s_2} w_i^* \boldsymbol{x}_{1i} = \sum_{i \in U} \boldsymbol{x}_{1i}. \tag{2.32}$$

Soit $G(w_i^*/d_i^*)$ la fonction qui mesure la distance entre le poids w_i^* et le poids d_i^*. Les poids w_i^* sont choisis de façon à minimiser l'expression :

$$\sum_{i \in s_2} \frac{d_i^* G(w_i^*/d_i^*)}{q_i},$$

sous la contrainte (2.32), où q_i représente un poids associé à l'unité i. Par un développement analogue à celui dans le cas de l'échantillonnage à une seule phase (voir section 2.3), on obtient

$$w_i^* = d_i^* F(q_i \boldsymbol{\lambda}' \boldsymbol{x}_{1i}), \tag{2.33}$$

où $F(.)$ représente la fonction inverse de $G'(.) \equiv \frac{\partial G(.)}{\partial .}$ et où $\boldsymbol{\lambda}'$ est tel que la contrainte (2.32) est satisfaite. On remarque que le poids w_i^* s'écrit comme le produit du poids de sondage d_i^* et d'un facteur d'ajustement $F_i = F(q_i \boldsymbol{\lambda}' \boldsymbol{x}_{1i})$. On peut déterminer $\boldsymbol{\lambda}'$ en solutionnant l'équation suivante obtenue en remplaçant les w_i^* de (2.33) dans la contrainte (2.32) :

$$\sum_{i \in s_2} d_i^* F(q_i \boldsymbol{\lambda}' \boldsymbol{x}_{1i}) \boldsymbol{x}_{1i} = \sum_{i \in U} \boldsymbol{x}_{1i}. \tag{2.34}$$

Lorsque G est donnée par la distance des moindres carrés généralisés (voir expression 2.14), l'estimateur de calage est donné par

$$\hat{t}_{yC1}^* = \sum_{i \in U} \boldsymbol{x}_{1i}' \widehat{\boldsymbol{B}}_1 + \sum_{i \in s_2} d_i^* e_{1i}, \tag{2.35}$$

où $e_{1i} = y_i - \boldsymbol{x}'_{1i}\widehat{\boldsymbol{B}}_1$ et

$$\widehat{\boldsymbol{B}}_1 = \left(\sum_{i\in s_2} d_i^* q_i \boldsymbol{x}_{1i}\boldsymbol{x}'_{1i}\right)^{-1}\left(\sum_{i\in s_2} d_i^* q_i \boldsymbol{x}_{1i} y_i\right).$$

Cet estimateur peut également être construit par la régression généralisée comme à la section 2.3.2. De plus, la variance asymptotique de l'estimateur (2.31) est la même que celle de l'estimateur (2.35). Elle peut être obtenue à partir du développement en série de Taylor de (2.35) :

$$\hat{t}^*_{yC1} \approx \sum_{i\in U} \boldsymbol{x}'_{1i}\boldsymbol{B}_1 + \sum_{i\in s_2} d_i^* E_{1i},$$

où $E_{1i} = y_i - \boldsymbol{x}'_{1i}\boldsymbol{B}_1$ et

$$\boldsymbol{B}_1 = \left(\sum_{i\in U} q_i \boldsymbol{x}_{1i}\boldsymbol{x}'_{1i}\right)^{-1}\left(\sum_{i\in U} q_i \boldsymbol{x}_{1i} y_i\right).$$

Ainsi, la variance de l'estimateur (2.31) est donnée par

$$\begin{aligned}
\mathrm{V}_p\left(\hat{t}^*_{yC1}\right) &= \mathrm{V}_1\,\mathrm{E}_2\left(\hat{t}^*_{yC1}\mid \boldsymbol{I}_1\right) + \mathrm{E}_1\,\mathrm{V}_2\left(\hat{t}^*_{yC1}\mid \boldsymbol{I}_1\right)\\
&\approx \mathrm{V}_1\left(\sum_{i\in s_1} d_{1i}E_{1i}\right) + \mathrm{E}_1\,\mathrm{V}_2\left(\sum_{i\in s_2} d_i^* E_{1i}\,\middle|\,\boldsymbol{I}_1\right)\\
&= \sum_{i\in U}\sum_{j\in U}(\pi_{1ij}-\pi_{1i}\pi_{1j})\frac{E_{1i}}{\pi_{1i}}\frac{E_{1j}}{\pi_{1j}}\\
&\quad + \mathrm{E}_1\left(\sum_{i\in s_1}\sum_{j\in s_1}\left\{\pi_{2ij}(\boldsymbol{I}_1)-\pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)\right\}\frac{E_{1i}}{\pi_i^*}\frac{E_{1j}}{\pi_j^*}\right). \qquad (2.36)
\end{aligned}$$

Remarque 2.11. *L'estimateur usuel approximativement sans biais de la variance asymptotique (2.36) est obtenu en estimant sans biais chaque terme et en remplaçant les quantités E_{1i} par les quantités e_{1i} qui sont connues au niveau de s_2. Il est donné par*

$$\widehat{\mathrm{V}}_p\left(\hat{t}^*_{yC1}\right) = \sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi^*_{ij}}\frac{e_{1i}}{\pi_{1i}}\frac{e_{1j}}{\pi_{1j}} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{2ij}(\boldsymbol{I}_1)-\pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)}{\pi_{2ij}(\boldsymbol{I}_1)}\frac{e_{1i}}{\pi_i^*}\frac{e_{1j}}{\pi_j^*}. \qquad (2.37)$$

Remarque 2.12. *Dans le cas où q_i est tel que ${q_i}^{-1} = \boldsymbol{\alpha}'\boldsymbol{x}_{1i}$ pour un certain vecteur $\boldsymbol{\alpha}$ de constantes connues, on peut montrer que*

$$\sum_{i\in U} E_{1i} = \sum_{i\in s_2} d_i^* e_{1i} = 0. \tag{2.38}$$

Ce résultat aura un intérêt particulier au chapitre 3.

Remarque 2.13. *Un estimateur par le ratio peut être obtenu comme cas particulier de l'estimateur de calage (2.35) avec $\boldsymbol{x}_{1i} = x_{1i}$ et en posant ${q_i}^{-1} = x_{1i}$:*

$$\hat{t}_{yR1}^* = \frac{\hat{t}_{y\pi}^*}{\hat{t}_{x1\pi}^*} t_{x1}. \tag{2.39}$$

Un cas particulier de $\hat{t}_{yR1}^*$ est l'estimateur de Hájek, obtenu en posant $x_{1i} = 1$

$$\hat{t}_{yH1}^* = \frac{\hat{t}_{y\pi}^*}{\widehat{N}_\pi^*} N,$$

où $\widehat{N}_\pi^* = \sum_{i\in s_2} d_i^*$. Ces deux estimateurs satisfont à la condition ${q_i}^{-1} = \boldsymbol{\alpha}'\boldsymbol{x}_{1i}$ (voir remarque 2.12) en posant $\boldsymbol{\alpha} = 1$.

2.6.2 Information auxiliaire de type (2) seulement

Dans cette section, nous considérons le cas où de l'information auxiliaire de type (2) seulement est utilisée au moment de l'estimation. Le tableau 2.1 résume la disponibilité de l'information auxiliaire dans cette situation.

Tableau 2.1 – Disponibilité de l'information pour les plans à deux phases avec information auxiliaire de type (2) seulement

Ensemble d'unités	Information disponible
Population	Aucune
Échantillon de première phase	$\{\boldsymbol{x}_{2i} : i \in s_1\}$
Échantillon de deuxième phase	$\{(\boldsymbol{x}_{2i}, y_i) : i \in s_2\}$

Cette situation est assez fréquente en pratique, car l'échantillonnage à deux phases est utile lorsque peu ou pas d'information auxiliaire est disponible sur la base de sondage.

L'estimateur de calage s'écrit sous la forme

$$\hat{t}^*_{yC1} = \sum_{i \in s_2} w^*_i y_i, \tag{2.40}$$

où les poids de calage w^*_i satisfont la contrainte suivante :

$$\sum_{i \in s_2} w^*_i \boldsymbol{x}_{1i} = \sum_{i \in s_1} d_{1i} \boldsymbol{x}_{1i}. \tag{2.41}$$

Soit $G(w^*_i/d^*_i)$ la fonction qui mesure la distance entre le poids w^*_i et le poids d^*_i. Les poids w^*_i sont choisis de façon à minimiser l'expression :

$$\sum_{i \in s_2} \frac{d^*_i G(w^*_i/d^*_i)}{q_i},$$

sous la contrainte (2.41) et où q_i représente un poids associé à l'unité i. Par un développement analogue à celui dans le cas de l'échantillonnage à une seule phase (voir section 2.3), on obtient

$$w^*_i = d^*_i F(q_i \boldsymbol{\lambda}' \boldsymbol{x}_{2i}), \tag{2.42}$$

où $F(.)$ représente la fonction inverse de $G'(.) \equiv \frac{\partial G(.)}{\partial .}$ et où $\boldsymbol{\lambda}'$ est tel que la contrainte (2.41) est satisfaite. On remarque que le poids obtenu w^*_i s'écrit comme le produit du poids de sondage d^*_i et d'un facteur d'ajustement $F_i = F(q_i \boldsymbol{\lambda}' \boldsymbol{x}_{2i})$. On peut déterminer $\boldsymbol{\lambda}'$ en solutionnant l'équation suivante obtenue en remplaçant les w^*_i de (2.42) dans la contrainte (2.41) :

$$\sum_{i \in s_2} d^*_i F(q_i \boldsymbol{\lambda}' \boldsymbol{x}_{2i}) \boldsymbol{x}_{2i} = \sum_{i \in s_1} d_{1i} \boldsymbol{x}_{2i}. \tag{2.43}$$

Lorsque G est donnée par la distance des moindres carrés généralisés (voir expression 2.14), l'estimateur de calage est donné par

$$\hat{t}^*_{yC2} = \sum_{i \in s_1} d_{1i} \boldsymbol{x}'_{2i} \widehat{\boldsymbol{B}}_2 + \sum_{i \in s_2} d^*_i e_{2i}, \tag{2.44}$$

où $e_{2i} = y_i - \boldsymbol{x}'_{2i}\widehat{\boldsymbol{B}}_2$ et

$$\widehat{\boldsymbol{B}}_2 = \left(\sum_{i\in s_2} d_i^* q_i \boldsymbol{x}_{2i}\boldsymbol{x}'_{2i}\right)^{-1}\left(\sum_{i\in s_2} d_i^* q_i \boldsymbol{x}_{2i} y_i\right).$$

Cet estimateur peut également être construit par la régression généralisée comme à la section 2.3.2. De plus, la variance asymptotique de l'estimateur (2.40) est la même que celle de l'estimateur (2.44). Elle peut être obtenue à partir du développement en série de Taylor de (2.44) :

$$\hat{t}^*_{yC2} \approx \sum_{i\in s_1} d_{1i}\boldsymbol{x}'_{2i}\boldsymbol{B}_2 + \sum_{i\in s_2} d_i^* E_{2i},$$

où $E_{2i} = y_i - \boldsymbol{x}'_{2i}\boldsymbol{B}_2$ et

$$\boldsymbol{B}_2 = \left(\sum_{i\in U} q_i \boldsymbol{x}_{2i}\boldsymbol{x}_{2i}{}'\right)^{-1}\left(\sum_{i\in U} q_i \boldsymbol{x}_{2i} y_i\right).$$

Sa variance est donnée par

$$\begin{aligned}
\mathrm{V}_p\left(\hat{t}^*_{yC2}\right) &= \mathrm{V}_1\,\mathrm{E}_2\left(\hat{t}^*_{yC2}\,|\,\boldsymbol{I}_1\right) + \mathrm{E}_1\,\mathrm{V}_2\left(\hat{t}^*_{yC2}\,|\,\boldsymbol{I}_1\right)\\
&\approx \mathrm{V}_1\left(\sum_{i\in s_1} d_{1i} y_i\right) + \mathrm{E}_1\,\mathrm{V}_2\left(\sum_{i\in s_2} d_i^* E_{2i}\,\middle|\,\boldsymbol{I}_1\right)\\
&= \sum_{i\in U}\sum_{j\in U}(\pi_{1ij} - \pi_{1i}\pi_{1j})\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1j}}\\
&\quad + \mathrm{E}_1\left(\sum_{i\in s_1}\sum_{j\in s_1}\left\{\pi_{2ij}(\boldsymbol{I}_1) - \pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)\right\}\frac{E_{2i}}{\pi_i^*}\frac{E_{2j}}{\pi_j^*}\right). \qquad (2.45)
\end{aligned}$$

Remarque 2.14. *L'estimateur usuel approximativement sans biais de la variance asymptotique (2.45) est obtenu en estimant sans biais chaque terme et en remplaçant les quantités E_{2i} par les quantités e_{2i} qui sont connues au niveau de s_2. Il est donné par*

$$\widehat{\mathrm{V}}_p\left(\hat{t}^*_{yC2}\right) = \sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi^*_{ij}}\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1j}} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{2ij}(\boldsymbol{I}_1) - \pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)}{\pi_{2ij}(\boldsymbol{I}_1)}\frac{e_{2i}}{\pi_i^*}\frac{e_{2j}}{\pi_j^*}. \qquad (2.46)$$

Remarque 2.15. *Un autre estimateur sans biais de la variance (2.45) peut être obtenu en utilisant le fait que* $y_i = \boldsymbol{x}'_{2i}\boldsymbol{B}_2 + E_{2i}$*, où* $E_{2i} = y_i - \boldsymbol{x}'_{2i}\boldsymbol{B}_2$*. Ainsi, la composante de la variance due à l'échantillonnage de première phase peut s'exprimer comme*

$$\begin{aligned}
\mathrm{V}_1\,\mathrm{E}_2(\hat{t}^*_{yC2}) &\approx \sum_{i\in U}\sum_{j\in U}(\pi_{1ij}-\pi_{1i}\pi_{1j})\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1j}} \\
&= \sum_{i\in U}\sum_{j\in U}(\pi_{1ij}-\pi_{1i}\pi_{1j})\frac{\boldsymbol{x}'_{2i}\boldsymbol{B}_2}{\pi_{1i}}\frac{\boldsymbol{x}'_{2j}\boldsymbol{B}_2}{\pi_{1j}} + 2\sum_{i\in U}\sum_{j\in U}(\pi_{1ij}-\pi_{1i}\pi_{1j})\frac{\boldsymbol{x}'_{2i}\boldsymbol{B}_2}{\pi_{1i}}\frac{E_{2j}}{\pi_{1j}} \\
&\quad + \sum_{i\in U}\sum_{j\in U}(\pi_{1ij}-\pi_{1i}\pi_{1j})\frac{E_{2i}}{\pi_{1i}}\frac{E_{2j}}{\pi_{1j}} \qquad (2.47)
\end{aligned}$$

Un estimateur approximativement sans biais sous le plan de l'expression (2.47) est donné par

$$\begin{aligned}
&\sum_{i\in s_1}\sum_{j\in s_1}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1ij}}\frac{\boldsymbol{x}'_{2i}\widehat{\boldsymbol{B}}_2}{\pi_{1i}}\frac{\boldsymbol{x}'_{2j}\widehat{\boldsymbol{B}}_2}{\pi_{1j}} + 2\sum_{i\in s_1}\sum_{j\in s_2}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1ij}}\frac{\boldsymbol{x}'_{2i}\widehat{\boldsymbol{B}}_2}{\pi_{1i}}\frac{e_{2j}}{\pi^*_j} \\
&+ \sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi^*_{ij}}\frac{e_{2i}}{\pi_{1i}}\frac{e_{2j}}{\pi_{1j}}
\end{aligned}$$

en notant que $\boldsymbol{x}'_{2i}\widehat{\boldsymbol{B}}_2$ *est disponible pour les unités de* s_1*. Ainsi, un estimateur approximativement sans biais de* $V(\hat{t}^*_{yG})$ *est donné par*

$$\begin{aligned}
\widehat{V}(\hat{t}^*_{yG})^{alt} &= \sum_{i\in s_1}\sum_{j\in s_1}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1ij}}\frac{\boldsymbol{x}'_{2i}\widehat{\boldsymbol{B}}_2}{\pi_{1i}}\frac{\boldsymbol{x}'_{2j}\widehat{\boldsymbol{B}}_2}{\pi_{1j}} + 2\sum_{i\in s_1}\sum_{j\in s_2}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1ij}}\frac{\boldsymbol{x}'_{2i}\widehat{\boldsymbol{B}}_2}{\pi_{1i}}\frac{e_{2j}}{\pi^*_j} \\
&\quad + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi^*_{ij}}\frac{e_{2i}}{\pi_{1i}}\frac{e_{2j}}{\pi_{1j}} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{2ij}(\boldsymbol{I}_1)-\pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)}{\pi_{2ij}(\boldsymbol{I}_1)}\frac{e_{2i}}{\pi^*_i}\frac{e_{2j}}{\pi^*_j}. \qquad (2.48)
\end{aligned}$$

L'utilisation d'information auxiliaire à l'étape de l'estimation de la variance a été étudiée entre autres par Axelson (1998) et Hidiroglou, Rao et Haziza (2009). Ainsi, nous nous attendons à ce que l'estimateur (2.48) *soit plus efficace que l'estimateur usuel* (2.46) *puisqu'il utilise davantage d'information auxiliaire.*

Remarque 2.16. *Dans le cas où q_i est de la forme ${q_i}^{-1} = \boldsymbol{\alpha}'\boldsymbol{x}_{2i}$, on peut montrer que*

$$\sum_{i\in U} E_{2i} = \sum_{i\in s_2} d_i^* e_{2i} = 0. \tag{2.49}$$

Ce résultat aura un intérêt particulier au chapitre 3.

Remarque 2.17. *Si seulement une variable auxiliaire x_2 est utilisée et en posant ${q_i}^{-1} = x_{2i}$, on obtient un estimateur de type ratio :*

$$\hat{t}_{yR2}^* = \frac{\hat{t}_{y\pi}^*}{\hat{t}_{x2\pi}^*}\hat{t}_{x2}^1. \tag{2.50}$$

2.6.3 Information auxiliaire de types (1) et (2)

Dans cette section, nous considérons le cas où de l'information auxiliaire de type (1) et (2) est utilisée au stade de l'estimation. Posons $\boldsymbol{x}_i = \left(\boldsymbol{x}_{1i}', \boldsymbol{x}_{2i}'\right)'$ le vecteur des $J = J_1 + J_2$ variables auxiliaires disponibles pour tout $i \in s_1$. Le tableau 2.2 résume l'information disponible pour les unités.

Tableau 2.2 – Disponibilité de l'information pour les plans à deux phases avec information auxiliaire de type (1) et (2)

Ensemble d'unités	Information disponible
Population	$\{\boldsymbol{x}_{1i} : i \in U\}$ ou $\boldsymbol{t}_{x_1}$
Échantillon de première phase	$\{\boldsymbol{x}_i : i \in s_1\}$
Échantillon de deuxième phase	$\{(\boldsymbol{x}_i, y_i) : i \in s_2\}$

Dans ce qui suit, nous allons généraliser la théorie présentée dans Hidiroglou et Särndal (1998) qui n'ont traité que le cas des fonctions de distance de type moindres carrés généralisés. L'estimateur de calage s'écrit

$$\hat{t}_{yC}^* = \sum_{i\in s_2} w_i^* y_i, \tag{2.51}$$

où w_i^* désigne le poids de calage associé à l'unité i. Nous utilisons la méthode « top-down » (Esteavo et Särndal, 2006) pour laquelle le calage s'effectue en deux temps : au niveau de la population, puis au niveau de l'échantillon de première phase. Il est à noter

qu'il existe d'autres méthodes tout aussi défendables qui ne mènent généralement pas au mêmes poids de calage comme la méthode « bottom up » à deux temps ou la méthode en un temps. Celles-ci sont présentées dans Esteavo et Särndal (2006).

Dans un premier temps, on cherche des poids intermédiaires w_{1i} permettant de satisfaire la contrainte suivante sur l'information auxiliaire de type (1) disponible au niveau de la population :

$$\sum_{i \in s_1} w_{1i} \boldsymbol{x}_{1i} = \sum_{i \in U} \boldsymbol{x}_{1i}. \tag{2.52}$$

Soit $G_1(w_{1i}/d_i^*)$ la fonction qui mesure la distance entre le poids w_{1i} et le poids d_i^* utilisé dans l'estimateur de double dilatation. Les poids w_{1i} sont choisis de façon à minimiser l'expression :

$$\sum_{i \in s_1} \frac{d_{1i} G_1(w_{1i}/d_{1i})}{q_{1i}},$$

sous la contrainte (2.52) et où q_{1i} représente un poids associé à l'unité i. Par un développement analogue à celui dans le cas de l'échantillonnage à une seule phase (voir section 2.3), on obtient

$$w_{1i} = d_{1i} F_1(q_{1i} \boldsymbol{\lambda}_1' \boldsymbol{x}_{1i}), \tag{2.53}$$

où $F_1(.)$ représente la fonction inverse de $G_1'(.) \equiv \frac{\partial G_1(.)}{\partial .}$ et où $\boldsymbol{\lambda}_1'$ est tel que la contrainte (2.52) est satisfaite. On remarque que le poids obtenu w_{1i} s'écrit comme le produit du poids de sondage d_i^* et d'un facteur d'ajustement $F_{1i} = F_1(q_{1i} \boldsymbol{\lambda}_1' \boldsymbol{x}_{1i})$. On peut déterminer $\boldsymbol{\lambda}_1'$ en solutionnant l'équation suivante obtenue en remplaçant les w_{1i} de (2.53) dans la contrainte (2.52) :

$$\sum_{i \in s_1} d_{1i} F_1(q_{1i} \boldsymbol{\lambda}_1' \boldsymbol{x}_{1i}) \boldsymbol{x}_{1i} = \sum_{i \in U} \boldsymbol{x}_{1i}. \tag{2.54}$$

Dans un deuxième temps, on cherche des poids w_i^* permettant de satisfaire la contrainte suivante sur l'information auxiliaire disponible au niveau de l'échantillon de première phase :

$$\sum_{i \in s_2} w_i^* \boldsymbol{x}_i = \sum_{i \in s_1} w_{1i} \boldsymbol{x}_i. \tag{2.55}$$

On considère la fonction de distance $G_2(w_i^*/w_{1i} d_{2i}(\boldsymbol{I}_1))$ pour laquelle on voudra mini-

miser

$$\sum_{i\in s_2} \frac{w_{1i}d_{2i}(\boldsymbol{I}_1)G_2(w_i^*/w_{1i}d_{2i}(\boldsymbol{I}_1))}{q_i},$$

sous la contrainte (2.55) et où q_i représente un poids associé à l'unité i. Par un développement analogue à celui dans le cas de l'échantillonnage à une seule phase (voir section 2.3), on obtient

$$\begin{aligned} w_i^* &= w_{1i}d_{2i}F_2(q_i\boldsymbol{\lambda}_2'\boldsymbol{x}_i), \\ &= d_i^*F_1(q_{1i}\boldsymbol{\lambda}_1'\boldsymbol{x}_{1i})F_2(q_i\boldsymbol{\lambda}_2'\boldsymbol{x}_i), \end{aligned} \tag{2.56}$$

où $F_2(.)$ représente la fonction inverse de $G_2'(.) \equiv \frac{\partial G_2(.)}{\partial .}$. On remarque que les poids de calage w_i^* s'écrivent comme le produit du poids de sondage d_i^* et d'un facteur d'ajustement $F_i^* = F_1(q_{1i}\boldsymbol{\lambda}_1'\boldsymbol{x}_{1i})F_2(q_i\boldsymbol{\lambda}_2'\boldsymbol{x}_i)$. On peut déterminer $\boldsymbol{\lambda}_2'$ en solutionnant l'équation suivante obtenue en remplaçant les w_{1i} et w_i^* de (2.53) et (2.56) dans la contrainte (2.55) :

$$\sum_{i\in s_2} d_i^*F_1(q_{1i}\boldsymbol{\lambda}_1'\boldsymbol{x}_{1i})F_2(q_i\boldsymbol{\lambda}_2'\boldsymbol{x}_i)\boldsymbol{x}_i = \sum_{i\in s_1} d_{1i}F_1(q_{1i}\boldsymbol{\lambda}_1'\boldsymbol{x}_{1i})\boldsymbol{x}_i, \tag{2.57}$$

où $\boldsymbol{\lambda}_1'$ est le résultat de la solution de l'équation (2.54) précédemment trouvée.

Un cas particulier important de l'estimateur de calage à deux phases est obtenu lorsque la fonction de distance des moindres carrés est utilisée à chacune des deux phases de calage :

$$G_1(w_{1i}/d_i) = \frac{1}{2}\left(\frac{w_{1i}}{d_i} - 1\right)^2$$

et

$$G_2(w_i^*/w_{1i}) = \frac{1}{2}\left(\frac{w_i^*}{w_{1i}} - 1\right)^2.$$

Ceci est le cas particulier traité dans Hidiroglou et Sarndal (1998). Les équations (2.54) et (2.57) possèdent alors une solution explicite et l'estimateur de calage est donné par

$$\hat{t}_{yC}^* = \hat{t}_y^* + \left(\boldsymbol{t}_{x_1} - \hat{\boldsymbol{t}}_{x_1}^1\right)'\widehat{\boldsymbol{B}}_1 + \left(\hat{\boldsymbol{t}}_x^1 - \hat{\boldsymbol{t}}_x^*\right)'\widehat{\boldsymbol{B}}, \tag{2.58}$$

où

$$\widehat{\boldsymbol{B}}_1 = \left(\sum_{i \in s_1} d_{1i} q_{1i} \boldsymbol{x}_{1i} \boldsymbol{x}_{1i}' \right)^{-1} \left\{ \sum_{i \in s_1} d_{1i} q_{1i} \boldsymbol{x}_{1i} \boldsymbol{x}_i \widehat{\boldsymbol{B}} + \sum_{i \in s_2} d_i^* q_{1i} \boldsymbol{x}_{1i} (y_i - \boldsymbol{x}_i \widehat{\boldsymbol{B}}) \right\} \tag{2.59}$$

et

$$\widehat{\boldsymbol{B}} = \left(\sum_{i \in s_2} d_i^* q_i \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \left(\sum_{i \in s_2} d_i^* q_i \boldsymbol{x}_i y_i \right). \tag{2.60}$$

Les expressions $\widehat{\boldsymbol{B}}_1$ et $\widehat{\boldsymbol{B}}$ estiment les paramètres de modèles linéaires lorsque les variables q_{1i} et q_i désignent l'inverse de la variance des résidus de modèles de régression ; respectivement un modèle de régression reliant la variable d'intérêt y au vecteur des variables auxiliaires $\boldsymbol{x}_1$ et un autre reliant la variable d'intérêt y au vecteur des variables auxiliaires $\boldsymbol{x}$. La construction de l'estimateur (2.58) par la régression sera traitée à la section 2.6.3.2.

2.6.3.1 Propriétés de l'estimateur de calage à deux phases

Les propriétés de l'estimateur (2.58) peuvent être étudiées à l'aide de la linéarisation :

$$\begin{aligned} \hat{t}_{yC}^* - t_y &\approx \left(\hat{t}_y^* - t_y\right) + \left(\boldsymbol{t}_{x_1} - \hat{\boldsymbol{t}}_{x_1}^1\right)' \boldsymbol{B}_1 + \left(\hat{\boldsymbol{t}}_x^1 - \hat{\boldsymbol{t}}_x^*\right)' \boldsymbol{B} \\ &= \left\{ \sum_{i \in s_1} d_{1i} E_{1i} - \sum_{i \in U} E_{1i} \right\} + \left\{ \sum_{i \in s_2} d_i^* E_i - \sum_{i \in s_1} d_{1i} E_i \right\}, \end{aligned}$$

où $E_{1i} = y_i - \boldsymbol{x}_{1i}' \boldsymbol{B}_1$ et $E_i = y_i - \boldsymbol{x}_i' \boldsymbol{B}$ peuvent être vus comme les résidus de modèles de régression.

La variance de l'estimateur (2.58) est donnée asymptotiquement par :

$$\begin{aligned} \mathrm{V}_p\left(\hat{t}_{yC}^*\right) &= \mathrm{V}_1 \mathrm{E}_2\left(\hat{t}_{yC}^* \,\middle|\, \boldsymbol{I}_1\right) + \mathrm{E}_1 \mathrm{V}_2\left(\hat{t}_{yC}^* \,\middle|\, \boldsymbol{I}_1\right), \\ &\approx \mathrm{V}_1 \left(\sum_{i \in s_1} d_{1i} E_{1i} \right) + \mathrm{E}_1 \mathrm{V}_2 \left(\sum_{i \in s_2} d_i^* E_i \,\middle|\, \boldsymbol{I}_1 \right), \\ &= \sum_{i \in U} \sum_{j \in U} \left(\pi_{1ij} - \pi_{1i}\pi_{1j}\right) \frac{E_{1i}}{\pi_{1i}} \frac{E_{1j}}{\pi_{1j}} \end{aligned}$$

$$+\mathrm{E}_1\left\{\sum_{i\in s_1}\sum_{j\in s_1}\left(\pi_{2ij}(\boldsymbol{I}_1)-\pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)\right)\frac{E_i}{\pi_i^*}\frac{E_j}{\pi_j^*}\right\}. \tag{2.61}$$

La variance (2.61) sera petite lorsque les résidus E_{1i} et les E_i sont petits, ce qui survient lorsque les modèles de régression ajustent bien les données. D'ailleurs, on s'attend à ce que les résidus E_i soient plus petits que les résidus E_{1i} puisque l'information contenue dans le vecteur $\boldsymbol{x}$ est habituellement plus riche que celle contenue dans le vecteur $\boldsymbol{x}_1$.

Remarque 2.18. *L'estimateur usuel approximativement sans biais de* (2.61) *est obtenu en estimant sans biais chaque terme et en remplaçant les quantités E_{1i} et E_i par les quantités e_{1i} et e_i qui sont connues au niveau de s_2. Il est donné par*

$$\widehat{\mathrm{V}}_p\left(\hat{t}_{yG}^*\right)=\sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{ij}^*}\frac{e_{1i}}{\pi_{1i}}\frac{e_{1j}}{\pi_{1j}}+\sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{2ij}(\boldsymbol{I}_1)-\pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)}{\pi_{2ij}(\boldsymbol{I}_1)}\frac{e_i}{\pi_i^*}\frac{e_j}{\pi_j^*}, \tag{2.62}$$

où $e_{1i}=y_i-\boldsymbol{x}_{1i}'\widehat{\boldsymbol{B}}_1$ et $e_i=y_i-\boldsymbol{x}_i'\widehat{\boldsymbol{B}}$.

Remarque 2.19. *Un autre estimateur sans biais de la variance* (2.61) *peut être obtenu en utilisant le fait que $E_{1i}=y_i-\boldsymbol{x}_{1i}'\boldsymbol{B}_1=\boldsymbol{x}_i'\boldsymbol{B}-\boldsymbol{x}_{1i}'\boldsymbol{B}_1+E_i$ où $E_i=y_i-\boldsymbol{x}_i'\boldsymbol{B}$. Ainsi, la composante de la variance due à l'échantillonnage de première phase peut s'exprimer comme*

$$\begin{aligned}\mathrm{V}_1\,\mathrm{E}_2(\hat{t}_{yC}^*)&\approx\sum_{i\in U}\sum_{j\in U}(\pi_{1ij}-\pi_{1i}\pi_{1j})\frac{E_{1i}}{\pi_{1i}}\frac{E_{1j}}{\pi_{1j}},\\&=\sum_{i\in U}\sum_{j\in U}(\pi_{1ij}-\pi_{1i}\pi_{1j})\frac{\boldsymbol{x}_i'\boldsymbol{B}-\boldsymbol{x}_{1i}'\boldsymbol{B}_1}{\pi_{1i}}\frac{\boldsymbol{x}_j'\boldsymbol{B}-\boldsymbol{x}_{1j}'\boldsymbol{B}_1}{\pi_{1j}}\\&+2\sum_{i\in U}\sum_{j\in U}(\pi_{1ij}-\pi_{1i}\pi_{1j})\frac{\boldsymbol{x}_i'\boldsymbol{B}-\boldsymbol{x}_{1i}'\boldsymbol{B}_1}{\pi_{1i}}\frac{E_j}{\pi_{1j}}+\sum_{i\in U}\sum_{j\in U}(\pi_{1ij}-\pi_{1i}\pi_{1j})\frac{E_i}{\pi_{1i}}\frac{E_j}{\pi_{1j}}.\end{aligned} \tag{2.63}$$

Un estimateur approximativement sans biais sous le plan de l'expression (2.63) est

donné par :

$$\sum_{i\in s_1}\sum_{j\in s_1}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1ij}}\frac{\boldsymbol{x}_i'\widehat{\boldsymbol{B}}-\boldsymbol{x}_{1i}'\widehat{\boldsymbol{B}}_1}{\pi_{1i}}\frac{\boldsymbol{x}_j'\widehat{\boldsymbol{B}}-\boldsymbol{x}_{1j}'\widehat{\boldsymbol{B}}_1}{\pi_{1j}}+2\sum_{i\in s_1}\sum_{j\in s_2}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1ij}}\frac{\boldsymbol{x}_i'\widehat{\boldsymbol{B}}-\boldsymbol{x}_{1i}'\widehat{\boldsymbol{B}}_1}{\pi_{1i}}\frac{e_j}{\pi_j^*}$$
$$+\sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{ij}^*}\frac{e_i}{\pi_{1i}}\frac{e_j}{\pi_{1j}}$$

en notant que $\boldsymbol{x}_i'\boldsymbol{B}-\boldsymbol{x}_{1i}'\boldsymbol{B}_1$ *est disponible pour les unités de* s_1*. Ainsi, un estimateur approximativement sans biais de* (2.20) *est donné par*

$$\begin{aligned}\widehat{V}(\hat{t}_{yC}^*)^{alt}=&\sum_{i\in s_1}\sum_{j\in s_1}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1ij}}\frac{\boldsymbol{x}_i'\widehat{\boldsymbol{B}}-\boldsymbol{x}_{1i}'\widehat{\boldsymbol{B}}_1}{\pi_{1i}}\frac{\boldsymbol{x}_j'\widehat{\boldsymbol{B}}-\boldsymbol{x}_{1j}'\widehat{\boldsymbol{B}}_1}{\pi_{1j}}\\&+2\sum_{i\in s_1}\sum_{j\in s_2}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{1ij}}\frac{\boldsymbol{x}_i'\widehat{\boldsymbol{B}}-\boldsymbol{x}_{1i}'\widehat{\boldsymbol{B}}_1}{\pi_{1i}}\frac{e_j}{\pi_j^*}\\&+\sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{1ij}-\pi_{1i}\pi_{1j}}{\pi_{ij}^*}\frac{e_i}{\pi_{1i}}\frac{e_j}{\pi_{1j}}+\sum_{i\in s_2}\sum_{j\in s_2}\frac{\pi_{2ij}(\boldsymbol{I}_1)-\pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)}{\pi_{2ij}(\boldsymbol{I}_1)}\frac{e_i}{\pi_i^*}\frac{e_j}{\pi_j^*}.\end{aligned}\tag{2.64}$$

Cet estimateur de variance aura un intérêt particulier au chapitre 3.

Remarque 2.20. *Dans le cas où* q_{1i} *est tel que* $q_{1i}^{-1}=\boldsymbol{\alpha}'\boldsymbol{x}_{1i}$ *pour un vecteur* $\boldsymbol{\alpha}$ *de constantes connues, on peut montrer que*

$$\sum_{i\in U}E_{1i}=\sum_{i\in s_2}d_i^*e_{1i}=0.$$

De la même façon, on montre que si q_i *est tel que* ${q_i}^{-1}=\boldsymbol{\alpha}'\boldsymbol{x}_i$ *pour un vecteur* $\boldsymbol{\alpha}$ *de constantes connues, alors*

$$\sum_{i\in U}E_i=\sum_{i\in s_2}d_i^*e_i=0.\tag{2.65}$$

Ce dernier résultat sera d'un intérêt particulier au chapitre 3.

Remarque 2.21. *Notons qu'en présence d'information auxiliaire de type (1) seulement, la variance de l'estimateur* $\hat{t}_{yC1}^*$ *donnée par* (2.36) *et l'estimateur de variance donné par* (2.37) *peuvent être obtenus comme cas particulier de la variance* (2.61) *et de l'estima-*

teur de variance (2.62) *en posant* $\boldsymbol{x} = \boldsymbol{x}_1$ *et* $\boldsymbol{x}_2 = \boldsymbol{0}$. *En effet, dans ce cas on a* $E_i = E_{1i}$ *et* $e_i = e_{1i}$.

Notons également qu'en présence d'information auxiliaire de type (2) seulement, la variance de l'estimateur $\hat{t}^*_{yC2}$ *donnée par* (2.45) *et les estimateurs de variance donnés par* (2.46) *et* (2.48) *peuvent être obtenus comme cas particulier de la variance* (2.61) *et des estimateurs de variance* (2.62) *et* (2.64) *en posant* $\boldsymbol{x} = \boldsymbol{x}_2$ *et* $\boldsymbol{x}_1 = \boldsymbol{0}$. *En effet, dans ce cas on a* $E_{1i} = y_i$, $E_i = E_{2i}$, $e_{1i} = y_i$ *et* $e_i = e_{2i}$.

2.6.3.2 Construction de l'estimateur de calage à deux phases par la régression généralisée

Une manière alternative d'obtenir l'estimateur (2.58) est d'utiliser une approche par la régression en deux étapes, en supposant deux modèles.

À la première étape, on utilise l'information auxiliaire disponible pour les unités de U, c'est-à-dire le vecteur $\boldsymbol{x}_1$. On suppose également pour l'instant que la variable d'intérêt est connue pour toutes les unités dans s_1.

Un modèle de superpopulation traduisant la relation entre y et $\boldsymbol{x}_1$ est donné par

$$m_1 : y_i = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1 + \varepsilon_{1i},$$

où $\boldsymbol{\beta}_1$ est un vecteur de J paramètres inconnus, et ε_{1i} est une variable aléatoire telle que $\mathrm{E}_{m_1}\left(\varepsilon_{1i}\right) = 0$, $\mathrm{E}_{m_1}\left(\varepsilon_{1i}\varepsilon_{1j}\right) = 0$ pour $i \neq j$ et $\mathrm{V}_{m_1}\left(\varepsilon_{1i}\right) = \sigma_1^2 c_{1i}$. Le coefficient c_{1i} associé à l'unité i est supposé connu. Si les valeurs de la variable d'intérêt étaient connues pour toutes les unités de la population, c'est-à-dire dans le cas d'un recensement, l'estimateur

des moindres carrés pondérés de $\boldsymbol{\beta}_1$ serait

$$\boldsymbol{B}_{G1} = \left(\sum_{i \in U} c_{1i}^{-1} \boldsymbol{x}_{1i} \boldsymbol{x}'_{1i} \right)^{-1} \left(\sum_{i \in U} c_{1i}^{-1} \boldsymbol{x}_{1i} y_i \right). \tag{2.66}$$

D'autre part, si les valeurs de la variable d'intérêt étaient connues pour toutes les unités dans s_1, un estimateur approximativement sans biais de $\boldsymbol{B}_{G1}$ sous le plan p_1 serait donné par

$$\tilde{\boldsymbol{B}}_{G1} = \left(\sum_{i \in s_1} d_{1i} c_{1i}^{-1} \boldsymbol{x}_{1i} \boldsymbol{x}'_{1i} \right)^{-1} \left(\sum_{i \in s_1} d_{1i} c_{1i}^{-1} \boldsymbol{x}_{1i} y_i \right).$$

Afin d'obtenir l'expression de l'estimateur de t_y par la régression, on commence par exprimer t_y comme

$$t_y = \sum_{i \in U} (\mathbf{x}'_{1i} \boldsymbol{\beta}_1 + \varepsilon_{1i}) = \sum_{i \in U} \boldsymbol{x}'_{1i} \boldsymbol{\beta}_1 + \sum_{i \in U} \varepsilon_{1i}.$$

Un estimateur approximativement sans biais de t_y serait

$$\tilde{t}_y = \sum_{i \in U} \boldsymbol{x}'_{1i} \tilde{\boldsymbol{B}}_{G1} + \sum_{i \in s_1} d_{1i} \left(y_i - \boldsymbol{x}'_{1i} \tilde{\boldsymbol{B}}_{G1} \right). \tag{2.67}$$

Bien sûr, les valeurs de la variable d'intérêt ne sont connues que pour les unités de l'échantillon s_2 et (2.67) ne peut pas être calculé en pratique. C'est pourquoi, une deuxième étape de régression est nécessaire.

À la deuxième étape, on utilise l'information auxiliaire disponible au niveau de s_1, c'est-à-dire le vecteur $\boldsymbol{x}$. Le modèle de superpopulation qui traduit la relation entre y et $\boldsymbol{x}$ est donné par

$$m : y_i = \boldsymbol{x}'_i \boldsymbol{\beta} + \varepsilon_i,$$

où $\boldsymbol{\beta}$ est un vecteur de J paramètres inconnus et ε_i est une variable aléatoire satisfaisant $\mathrm{E}_m(\varepsilon_i) = 0$, $\mathrm{E}_m(\varepsilon_i \varepsilon_j) = 0$ pour $i \neq j$ et $\mathrm{V}_m(\varepsilon_i) = \sigma^2 c_i$. Le coefficient c_i associé à l'unité i est supposé connu. Si les valeurs de la variable d'intérêt étaient connues pour toutes les unités de la population, c'est-à-dire dans le cas d'un recensement, l'estimateur

des moindres carrés pondérés de $\boldsymbol{\beta}$ serait

$$\boldsymbol{B}_G = \left(\sum_{i \in U} c_i^{-1} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \left(\sum_{i \in U} c_i^{-1} \boldsymbol{x}_i y_i \right).$$

D'autre part, si les valeurs de la variable d'intérêt étaient connues pour toutes les unités dans s_1, un estimateur de $\boldsymbol{B}_G$ serait donné par

$$\tilde{\boldsymbol{B}}_G = \left(\sum_{i \in s_1} d_{1i} c_i^{-1} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \left(\sum_{i \in s_1} d_{1i} c_i^{-1} \boldsymbol{x}_i y_i \right).$$

Mais puisque les valeurs de la variable d'intérêt ne sont connues que pour les unités dans s_2, on estimera $\tilde{\boldsymbol{B}}_G$ par

$$\widehat{\boldsymbol{B}}_G = \left(\sum_{i \in s_2} d_i^* c_i^{-1} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \left(\sum_{i \in s_2} d_i^* c_i^{-1} \boldsymbol{x}_i y_i \right).$$

On peut maintenant estimer la quantité $\tilde{t}_y$ en exprimant d'abord $\sum_{i \in s_1} d_{1i} y_i$ comme

$$\begin{aligned} \sum_{i \in s_1} d_{1i} y_i &= \sum_{i \in s_1} d_{1i} \left(\boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i \right) \\ &= \sum_{i \in s_1} d_{1i} \boldsymbol{x}_i' \boldsymbol{\beta} + \sum_{i \in s_1} d_{1i} \varepsilon_i. \end{aligned}$$

Donc, un estimateur de $\sum_{i \in s_1} d_{1i} y_i$ est donné par

$$\sum_{i \in s_1} d_{1i} \boldsymbol{x}_i' \widehat{\boldsymbol{B}}_G + \sum_{i \in s_2} d_i^* \left(y_i - \boldsymbol{x}_i' \widehat{\boldsymbol{B}}_G \right).$$

Finalement, un estimateur de t_y est donné par

$$\hat{t}_{yG}^* = \sum_{i \in U} \boldsymbol{x}_{1i}' \widehat{\boldsymbol{B}}_{G1} + \sum_{i \in s_1} d_{1i} \left(\boldsymbol{x}_i' \widehat{\boldsymbol{B}}_G - \boldsymbol{x}_{1i}' \widehat{\boldsymbol{B}}_{G1} \right) + \sum_{i \in s_2} d_i^* \left(y_i - \boldsymbol{x}_i' \widehat{\boldsymbol{B}}_G \right), \tag{2.68}$$

où $\widehat{\boldsymbol{B}}_{G1}$ est un estimateur approximativement sans biais pour $\tilde{\boldsymbol{B}}_{G1}$ donné par

$$\widehat{\boldsymbol{B}}_{G1} = \left(\sum_{i \in s_1} d_{1i} {c_{1i}}^{-1} \boldsymbol{x}_{1i} \boldsymbol{x}'_{1i} \right)^{-1} \left(\sum_{i \in s_1} d_{1i} {c_{1i}}^{-1} \boldsymbol{x}_{1i} \boldsymbol{x}'_i \widehat{\boldsymbol{B}}_G + \sum_{i \in s_2} d^*_i q_{1i} \boldsymbol{x}_{1i} (y_i - \boldsymbol{x}'_i \widehat{\boldsymbol{B}}_G) \right),$$

où

$$\left(\sum_{i \in s_1} d_{1i} {c_{1i}}^{-1} \boldsymbol{x}_{1i} \boldsymbol{x}'_i \widehat{\boldsymbol{B}}_G + \sum_{i \in s_2} d^*_i q_{1i} \boldsymbol{x}_{1i} (y_i - \boldsymbol{x}'_i \widehat{\boldsymbol{B}}_G) \right)$$

est un estimateur par la régression de

$$\sum_{i \in s_1} d_{1i} c_{1i}^{-1} \boldsymbol{x}_{1i} y_i = \sum_{i \in s_1} d_{1i} c_{1i}^{-1} \boldsymbol{x}_{1i} \boldsymbol{x}'_i \boldsymbol{B}_G + \sum_{i \in s_1} d_{1i} c_{1i}^{-1} \boldsymbol{x}_{1i} \big(y_i - \boldsymbol{x}'_i \boldsymbol{B}_G \big).$$

Avec un peu de calculs, on peut montrer que l'estimateur (2.68) est identique à l'estimateur de calage (2.58) lorsque l'on pose $q_{1i} = {c_{1i}}^{-1}$ et $q_i = {c_i}^{-1}$.

Remarque 2.22. *Un estimateur alternatif pour $\tilde{\boldsymbol{B}}_{G1}$ est donné par*

$$\widehat{\boldsymbol{B}}^{alt}_{G1} = \left(\sum_{i \in s_2} d^*_i {c_{1i}}^{-1} \boldsymbol{x}_{1i} \boldsymbol{x}'_{1i} \right)^{-1} \left(\sum_{i \in s_2} d^*_i {c_{1i}}^{-1} \boldsymbol{x}_{1i} y_i \right).$$

Dans ce cas, l'estimateur (2.68) *correspond à l'estimateur par la régression généralisée pour les plans à deux phases présenté dans Särndal, Swensson et Wretman (1992). Cet estimateur est asymptotiquement sans biais même si les modèles m_1 et m sont mal spécifiés. De plus, sa variance est donnée par l'expression* (2.61) *en posant $q_{1i} = {c_{1i}}^{-1}$ et $q_i = {c_i}^{-1}$.*

CHAPITRE 3

ESTIMATION SIMPLIFIÉE DE LA VARIANCE POUR LES PLANS À DEUX PHASES

Dans ce chapitre, nous traitons de la problématique de l'utilisation des estimateurs de variance en pratique. Nous suggérons des estimateurs simplifiés de la variance qui présentent l'avantage de pouvoir être obtenus au moyen d'un logiciel d'estimation de variance pour les plans de sondage à une phase. Nous considérons d'abord le cas de l'estimation de la variance pour l'estimateur par double dilatation, puis le cas des estimateurs de calage.

3.1 Estimation simplifiée de la variance de l'estimateur par double dilatation

Rappelons que l'estimateur usuel sans biais de $\mathrm{V}\left(\hat{t}_{y\pi}^{*}\right)$ est donné par :

$$\begin{aligned}\widehat{\mathrm{V}}\left(\hat{t}_{y\pi}^{*}\right) &= \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{1ij}}{\pi_{2ij}(\boldsymbol{I}_1)}y_i y_j + \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{2ij}(\boldsymbol{I}_1)\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1j}}, \\ &\equiv \widehat{\mathrm{V}}_1\left(\hat{t}_{y\pi}^{*}\right) + \widehat{\mathrm{V}}_2\left(\hat{t}_{y\pi}^{*}\right),\end{aligned} \tag{3.1}$$

où

$$\Delta_{1ij} = \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1ij}\pi_{1i}\pi_{1j}}$$

et

$$\Delta_{2ij}(\boldsymbol{I}_1) = \frac{\pi_{2ij}(\boldsymbol{I}_1) - \pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)}{\pi_{2ij}(\boldsymbol{I}_1)\pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)}.$$

L'utilisation de l'estimateur (3.1) n'est pas aisée en pratique car il requiert les probabilités d'inclusion jointes π_{1ij} et $\pi_{2ij}(\boldsymbol{I}_1)$. Lorsque le plan de deuxième phase est complexe, il est difficile (voire impossible) d'obtenir les probabilités d'inclusion jointes $\pi_{2ij}(\boldsymbol{I}_1)$. De plus, l'estimateur (3.1) requiert un logiciel spécialisé conçu pour effectuer l'estimation de variance dans les plans de sondage à deux phases.

Il nous apparaît donc utile de proposer des estimateurs de variance simplifiés qui auront l'avantage de ne pas dépendre des probabilités d'inclusion jointes $\pi_{2ij}(\boldsymbol{I}_1)$ et qui pourront être obtenus au moyen d'un logiciel d'estimation de variance pour les plans de sondage à une phase.

Un premier pas en ce sens a été effectué par Haziza et Beaumont (2005) qui ont proposé une décomposition alternative de l'estimateur de variance en effectuant les substitutions

$$\frac{1}{\pi_{2ij}(\boldsymbol{I}_1)} = \frac{1}{\pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)} - \Delta_{2ij}(\boldsymbol{I}_1)$$

et

$$\frac{1}{\pi_{1i}\pi_{1j}} = \frac{1}{\pi_{1ij}} + \Delta_{1ij}$$

dans le premier et deuxième terme respectivement de l'expression (3.1) ; voir aussi Singh (2008). Ainsi, l'estimateur $\widehat{\mathrm{V}}\big(\hat{t}^*_{y\pi}\big)$ peut s'écrire sous la forme alternative suivante :

$$\widehat{\mathrm{V}}\big(\hat{t}^*_{y\pi}\big) = \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{1ij}\frac{y_i}{\pi_{2i}(\boldsymbol{I}_1)}\frac{y_j}{\pi_{2j}(\boldsymbol{I}_1)} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{2ij}(\boldsymbol{I}_1)}{\pi_{1ij}}y_i y_j \tag{3.2}$$

$$\equiv \widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{y\pi}\big) + \widehat{\mathrm{V}}_2^R\big(\hat{t}^*_{y\pi}\big). \tag{3.3}$$

Notons que le terme $\widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{y\pi}\big)$ ne dépend pas de $\pi_{2ij}(\boldsymbol{I}_1)$ et qu'il peut être calculé à partir des logiciels d'estimation pour des plans à une phase. En effet, on peut écrire $\widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{y\pi}\big)$ sous la forme

$$\widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{y\pi}\big) = \sum_{i\in s_1}\sum_{j\in s_1}\Delta_{1ij}z_i z_j, \tag{3.4}$$

où

$$z_i = \frac{y_i}{\pi_{2i}(\boldsymbol{I}_1)}I_{2i}. \tag{3.5}$$

Cette expression correspond à l'estimateur de variance de l'estimateur par dilatation (une seule phase d'échantillonnage) pour une variable d'intérêt z_i donnée par (3.5).

La question qui se pose à ce stade est donc : quand l'estimateur $\widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{y\pi}\big)$ donné par

(3.4) est-il un bon estimateur de la variance totale $\mathrm{V}\left(\hat{t}^*_{y\pi}\right)$? Autrement dit, sous quelle(s) condition(s) le terme $\widehat{\mathrm{V}}_2^R\left(\hat{t}^*_{y\pi}\right)$ est-il négligeable ? Haziza et Beaumont (2005) ont trouvé de telles conditions dans le cas d'un plan de Poisson à la deuxième phase et dans le cas du plan à deux degrés. Aux sections 3.1.1 et 3.1.2, nous exposons leurs résultats, puis à la section 3.1.3, nous considérons le cas d'un plan de sondage aléatoire simple sans remise à chaque phase et finalement, à la section 3.1.4, nous considérons le cas plus général d'un plan quelconque à la première phase et d'un plan aléatoire simple sans remise à la deuxième phase.

La contribution de $\widehat{\mathrm{V}}_2^R\left(\hat{t}^*_{y\pi}\right)$ à la variance totale peut être mesurée par

$$C_2\left(\hat{t}^*_{y\pi}\right) \equiv \frac{\widehat{\mathrm{V}}_2^R\left(\hat{t}^*_{y\pi}\right)}{\widehat{\mathrm{V}}\left(\hat{t}^*_{y\pi}\right)} = \frac{\widehat{\mathrm{V}}_2^R\left(\hat{t}^*_{y\pi}\right)}{\widehat{\mathrm{V}}_1\left(\hat{t}^*_{y\pi}\right) + \widehat{\mathrm{V}}_2\left(\hat{t}^*_{y\pi}\right)}.$$

Dans ce qui suit, nous considérons également la quantité

$$C_2^R\left(\hat{t}^*_{y\pi}\right) \equiv \frac{\widehat{\mathrm{V}}_2^R\left(\hat{t}^*_{y\pi}\right)}{\widehat{\mathrm{V}}_2\left(\hat{t}^*_{y\pi}\right)}$$

qui constitue une borne supérieure puisqu'on a

$$\left|C_2\left(\hat{t}^*_{y\pi}\right)\right| \leq \left|C_2^R\left(\hat{t}^*_{y\pi}\right)\right| \tag{3.6}$$

lorsque $\widehat{\mathrm{V}}_1\left(\hat{t}^*_{y\pi}\right) \geq 0$ et $\widehat{\mathrm{V}}_2\left(\hat{t}^*_{y\pi}\right) \geq 0$.

3.1.1 Cas d'un plan de Poisson utilisé à la deuxième phase

Le plan de Poisson est décrit dans la section 2.1.2. Dans ce cas, on écrit $C_2^R\left(\hat{t}^*_{y\pi}\right)$ comme

$$C_2^R\left(\hat{t}^*_{y\pi}\right) = \sum_{i\in s_2}\sum_{j\in s_2} \Omega_{ij}(\boldsymbol{I}_1)\frac{\pi_{1i}\pi_{1j}}{\pi_{1ij}}, \tag{3.7}$$

où

$$\Omega_{ij}(\boldsymbol{I}_1) = \frac{\Delta_{2ij}(\boldsymbol{I}_1)\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1j}}}{\sum_{k\in s_2}\sum_{l\in s_2}\Delta_{2kl}(\boldsymbol{I}_1)\frac{y_k}{\pi_{1k}}\frac{y_l}{\pi_{1l}}}$$

et

$$\sum_{i\in s_2}\sum_{j\in s_2}\Omega_{ij}(\boldsymbol{I}_1) = 1.$$

Puisque la sélection des unités se fait de façon indépendante à la deuxième phase, on a $\pi_{2ij}(\boldsymbol{I}_1) = \pi_{2i}(\boldsymbol{I}_1)\pi_{2j}(\boldsymbol{I}_1)$ pour $i \neq j$ et donc $\Delta_{2ij}(\boldsymbol{I}_1) = 0$ pour $i \neq j$. En notant que $\Omega_{ii}(\boldsymbol{I}_1) > 0$ pour tout i, on obtient à partir de (3.7) :

$$\begin{aligned}
|C_2^R(\hat{t}_{y\pi}^*)| &= \sum_{i\in s_2}\Omega_{ii}(\boldsymbol{I}_1)\frac{\pi_{1i}^2}{\pi_{1i}} \\
&= \sum_{i\in s_2}\Omega_{ii}(\boldsymbol{I}_1)\pi_{1i} \\
&\leq \sum_{i\in s_2}\Omega_{ii}(\boldsymbol{I}_1)\max(\pi_{1i}) \\
&= \max(\pi_{1i})\sum_{i\in s_2}\Omega_{ii}(\boldsymbol{I}_1) \\
&= \max(\pi_{1i}). \qquad (3.8)
\end{aligned}$$

Ainsi, sous la condition usuelle $\max(\pi_{1i}) = O(n_1/N)$, il suffit que la fraction de sondage à la première phase soit négligeable pour que la contribution de $\widehat{\mathrm{V}}_2^R(\hat{t}_{y\pi}^*)$ soit négligeable. Ce résultat est important dans un contexte de non-réponse totale. En effet, le mécanisme de non-réponse est souvent modélisé au moyen d'un plan de Poisson à la deuxième phase.

Remarque 3.1. *Dans le cas particulier où $\pi_{1i} = \pi_1$ pour tout $i \in U$ (par exemple, un plan aléatoire simple sans remise ou un plan Bernoulli à la première phase), on obtient*

$$\begin{aligned}
C_2^R(\hat{t}_{y\pi}^*) &= \sum_{i\in s_2}\Omega_{ii}(\boldsymbol{I}_1)\frac{\pi_1^2}{\pi_1} \\
&= \pi_1\sum_{i\in s_2}\Omega_{ii}(\boldsymbol{I}_1) \\
&= \pi_1.
\end{aligned}$$

Ainsi, la contribution de $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{y\pi}^*\right)$ *sera négligeable si* π_1 *est négligeable.*

3.1.2 Cas d'un plan à deux degrés

Considérons le cas de l'échantillonnage à deux degrés décrit à la section 2.4.2. Dans la notation de l'échantillonnage à deux phases, la borne supérieure sur la contribution du terme $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{y\pi}^*\right)$ est donnée par

$$\begin{aligned} C_2^R\left(\hat{t}_{y\pi}^*\right) &= \frac{\widehat{\mathrm{V}}_2^R\left(\hat{t}_{y\pi}^*\right)}{\widehat{\mathrm{V}}_2\left(\hat{t}_{y\pi}^*\right)} \\ &= \frac{\sum_{i\in s_2}\sum_{j\in s_2}\Delta_{2ij}\frac{y_i y_j}{\pi_{1ij}}}{\sum_{i\in s_2}\sum_{j\in s_2}\Delta_{2ij}\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1j}}}. \end{aligned}$$

En utilisant la notation propre à l'échantillonnage à deux degrés introduite à la section 2.4.2, on exprime C_2^R de la façon suivante :

$$C_2^R\left(\hat{t}_{y\pi}^*\right) = \frac{\sum_{g\in s'}\sum_{h\in s'}\sum_{i\in s_g}\sum_{j\in s_h}\Delta_{2gihj}\frac{y_{gi}y_{hj}}{\pi_{1gh}}}{\sum_{g\in s'}\sum_{h\in s'}\sum_{i\in s_g}\sum_{j\in s_h}\Delta_{2gihj}\frac{y_{gi}}{\pi_{1gi}}\frac{y_{hj}}{\pi_{1hj}}},$$

où

$$\Delta_{2gihj} = \frac{\pi_{2gihj} - \pi_{2gi}\pi_{2hj}}{\pi_{2gihj}\pi_{2gi}\pi_{2hj}}.$$

Puisque la sélection des unités se fait généralement de façon indépendante entre les UPE, on a $\pi_{2gihj} = \pi_{2gi}\pi_{2hj}$ lorsque $g \neq h$, et dans ce cas, $\Delta_{2gihj} = 0$. Alors,

$$C_2^R\left(\hat{t}_{y\pi}^*\right) = \frac{\sum_{g\in s'}\sum_{i\in s_g}\sum_{j\in s_g}\Delta_{2gigj}\frac{y_{gi}y_{gj}}{\pi_{1g}}}{\sum_{g\in s'}\sum_{i\in s_g}\sum_{j\in s_g}\Delta_{2gigj}\frac{y_{gi}}{\pi_{1g}}\frac{y_{gj}}{\pi_{1g}}},$$

en notant que $\pi_{1gg} = \pi_{1g}$. On peut alors écrire

$$C_2^R\left(\hat{t}_{y\pi}^*\right) = \sum_{g\in s'}\frac{\nu_g}{\sum_{g\in s'}\nu_g}\pi_{1g},$$

où

$$\nu_g = \sum_{i \in s_g} \sum_{j \in s_g} \Delta_{2gigj} \frac{y_{gi}}{\pi_{1g}} \frac{y_{gj}}{\pi_{1g}}$$

est une estimation de la variance de $\hat{t}_g^* = \sum_{i \in s_g} y_{gi}/(\pi_{1g}\pi_{2gi})$ conditionnelle à $\boldsymbol{I}_1$. En effet, on a

$$\mathrm{V}_2\left(\hat{t}_g^* \mid \boldsymbol{I}_1\right) = \sum_{i \in U_g} \sum_{j \in U_g} \Delta_{2gigj} \pi_{2gigj} \frac{y_{gi}}{\pi_{1g}} \frac{y_{gj}}{\pi_{1g}}$$

et $\mathrm{E}_2\left(\nu_g \mid \boldsymbol{I}_1\right) = \mathrm{V}_2\left(\hat{t}_g^* \mid \boldsymbol{I}_1\right)$.

Finalement, on obtient

$$\begin{aligned} |C_2^R(\hat{t}_{y\pi}^*)| &\leq \sum_{g \in s'} \frac{|\nu_g|}{\sum_{g \in s'} |\nu_g|} \max(\pi_{1g}) \\ &= \max(\pi_{1g}). \end{aligned} \tag{3.9}$$

Si $\max(\pi_{1g}) = O(n/N)$ alors la contribution de $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{y\pi}^*\right)$ est négligeable si n/N, la fraction de sondage au premier degré, est négligeable. Il est à noter que la propriété d'invariance du plan de sondage à deux degrés n'a pas été utilisée pour démontrer ce résultat. Par conséquent, l'estimateur simplifié $\widehat{\mathrm{V}}_1^R\left(\hat{t}_{y\pi}^*\right)$ reste valide même si les probabilités de sélection à la deuxième phase dépendent de l'échantillon sélectionné à la première phase.

Il est facile de montrer que dans le cas d'un plan à deux degrés, le terme $\widehat{\mathrm{V}}_1^R\left(\hat{t}_{y\pi}^*\right)$ est équivalent à l'estimateur proposé dans Särndal, Swensson et Wretman (1992) (équation (4.3.17), p.139) donné par

$$\sum_{g \in s'} \sum_{h \in s'} \Delta_{1gh} \hat{t}_g \hat{t}_h, \tag{3.10}$$

où

$$\Delta_{1gh} = \frac{\pi_{1gh} - \pi_{1g}\pi_{1h}}{\pi_{1gh}\pi_{1g}\pi_{1h}}$$

et

$$\hat{t}_g = \sum_{i \in s_g} \frac{y_{gi}}{\pi_{2gi}}.$$

En effet, en utilisant la notation du plan à deux degrés, $\widehat{\mathrm{V}}_1^R\left(\hat{t}_{y\pi}^*\right)$ donné par (3.4) s'exprime

de la façon suivante :

$$\widehat{\mathrm{V}}_1^R\left(\hat{t}_{y\pi}^*\right) = \sum_{g\in s'}\sum_{h\in s'}\sum_{i\in s_g}\sum_{j\in s_h}\Delta_{1gh}\frac{y_{gi}}{\pi_{2gi}}\frac{y_{hj}}{\pi_{2hj}}. \tag{3.11}$$

L'équivalence des expressions (3.10) et (3.11) est triviale. Cet estimateur présente un biais négatif signifiant qu'il sous-estime généralement la variance. Toutefois la sous-estimation n'est pas importante dans plusieurs cas comme lorsque les probabilité d'inclusion à la première phase sont petites.

3.1.3 Cas d'un plan aléatoire simple sans remise utilisé aux deux phases

Supposons qu'un échantillon aléatoire simple sans remise de n_1 parmi N unités est tiré à la première phase, suivi du tirage d'un second échantillon aléatoire simple sans remise de n_2 parmi n_1 unités à la deuxième phase. Le plan de sondage aléatoire simple sans remise a été décrit à la section 2.1.1. Dénotons par $f_1 \equiv n_1/N$ et $f_2 \equiv n_2/n_1$ les fractions de sondage à la première et à la deuxième phase, respectivement. Dans le cas d'un plan aléatoire simple sans remise aux deux phases, on a

$$\pi_{1i} = n_1/N = f_1, \tag{3.12}$$

$$\pi_{2i} = n_2/n_1 = f_2, \tag{3.13}$$

$$\pi_{1ij} = \frac{n_1(n_1-1)}{N(N-1)} = \frac{f_1(Nf_1-1)}{(N-1)} \text{ pour } i \neq j \tag{3.14}$$

et

$$\pi_{2ij} = \frac{n_2(n_2-1)}{n_1(n_1-1)} = \frac{f_2(Nf_1f_2-1)}{(Nf_1-1)} \text{ pour } i \neq j. \tag{3.15}$$

En utilisant (3.12) à (3.15), on peut calculer les expressions exactes de $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{y\pi}^*\right)$, $\widehat{\mathrm{V}}_2\left(\hat{t}_{y\pi}^*\right)$ et $\widehat{\mathrm{V}}\left(\hat{t}_{y\pi}^*\right)$ qui sont données par :

$$\begin{aligned}\widehat{\mathrm{V}}_2^R\left(\hat{t}_{y\pi}^*\right) &= \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{2ij}}{\pi_{1ij}}y_iy_j \\ &= Nf_1^{-1}f_2^{-1}(1-f_2)\left(1-N^{-1}f_1^{-1}\right)^{-1}\end{aligned}$$

$$\times \left\{ \left[f_1 + N^{-1} \left(f_1^{-1} f_2^{-1} - f_2^{-1} - 1 \right) \right] s_{2y}^2 - (1 - f_1) \bar{y}_2^2 \right\}, \tag{3.16}$$

$$\begin{aligned} \widehat{\mathrm{V}}_2\left(\hat{t}_{y\pi}^*\right) &= \sum_{i \in s_2} \sum_{j \in s_2} \Delta_{2ij} \frac{y_i}{\pi_{1i}} \frac{y_j}{\pi_{1j}} \\ &= N^2 \left(1 - \frac{n_1}{n_2} \right) \frac{s_{2y}^2}{n_2} \\ &= N f_1^{-1} f_2^{-1} (1 - f_2) s_{2y}^2 \end{aligned} \tag{3.17}$$

et

$$\begin{aligned} \widehat{\mathrm{V}}\left(\hat{t}_{y\pi}^*\right) &= N^2 \left(1 - \frac{n_2}{N} \right) \frac{s_{2y}^2}{n_2} \\ &= N f_1^{-1} f_2^{-1} (1 - f_1 f_2) s_{2y}^2, \end{aligned} \tag{3.18}$$

où $\bar{y}_2 \equiv \sum_{i \in s_2} y_i / n_2$ représente la moyenne de la variable d'intérêt dans l'échantillon s_2 et $s_{2y}^2 \equiv \sum_{i \in s_2} (y_i - \bar{y}_2)^2 / (n_2 - 1)$ représente la variance de la variable d'intérêt dans s_2 (voir preuves à l'annexe I.3).

Dans les deux prochaines sections, nous obtenons les expressions de $C_2^R\left(\hat{t}_{y\pi}^*\right)$ et $C_2\left(\hat{t}_{y\pi}^*\right)$ et énonçons les conditions sous lesquelles ces quantités sont petites.

3.1.3.1 Étude de la quantité $C_2^R\left(\hat{t}_{y\pi}^*\right)$

À partir de (3.16) et (3.17), on peut calculer la contribution $C_2^R\left(\hat{t}_{y\pi}^*\right)$ qui est donnée par :

$$\begin{aligned} C_2^R\left(\hat{t}_{y\pi}^*\right) &= \frac{\widehat{\mathrm{V}}_2^R\left(\hat{t}_{y\pi}^*\right)}{\widehat{\mathrm{V}}_2\left(\hat{t}_{y\pi}^*\right)} \\ &= (1 - N^{-1} f_1^{-1})^{-1} \left\{ f_1 + N^{-1} \left(f_1^{-1} f_2^{-1} - f_2^{-1} - 1 \right) - (1 - f_1) \frac{1}{\mathrm{cv}_2(y)^2} \right\}, \end{aligned} \tag{3.19}$$

où $\mathrm{cv}_2(y) \equiv s_{2y}^2/\bar{y}_2$ désigne le coefficient de variation calculé au moyen des unités dans s_2.

La contribution de $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{y\pi}^*\right)$ donnée par (3.19) n'est, en général, pas négligeable comme nous le montrons dans une étude empirique présentée au chapitre 4. Cependant, un cas retiendra notre attention : celui pour lequel $\bar{y}_2 = 0$. Dans ce cas, on a $\mathrm{cv}_2(y) = \infty$ et $1/\mathrm{cv}_2(y) = 0$. Le tableau 3.1 exhibe les valeurs de $C_2^R\left(\hat{t}_{y\pi}^*\right)$ lorsque $1/\mathrm{cv}_2(y) = 0$ pour différentes valeurs de f_1 et f_2 et quatre valeurs de N. On constate que lorsque la taille de la population, N, est suffisamment grande, on a $C_2^R\left(\hat{t}_{y\pi}^*\right) \approx f_1$. On constate également, dans le cas contraire, que pour f_1 fixé les plus petites valeurs de $C_2^R\left(\hat{t}_{y\pi}^*\right)$ sont associées à une grande fraction de sondage f_2.

$f_1(\%)$	$f_2(\%)$	$C_2^R(\%)$ $N = 10^3$	$C_2^R(\%)$ $N = 10^4$	$C_2^R(\%)$ $N = 10^5$	$C_2^R(\%)$ $N = 10^6$
5	5	43.8	8.8	5.4	5.0
5	10	24.4	6.9	5.2	5.0
5	25	12.8	5.8	5.1	5.0
5	50	8.9	5.4	5.0	5.0
10	5	28.2	11.8	10.2	10.0
10	10	19.1	10.9	10.1	10.0
10	25	13.6	10.4	10.0	10.0
10	50	11.8	10.2	10.0	10.0
20	5	28.4	20.8	20.1	20.0
20	10	24.0	20.4	20.0	20.0
20	25	21.6	20.2	20.0	20.0
20	50	20.8	20.1	20.0	20.0

Tableau 3.1 – Valeurs de $C_2^R\left(\hat{t}_{y\pi}^*\right)$ selon les fractions de sondage pour différentes tailles de population lorsque $1/\mathrm{cv}_2(y) = 0$.

En fait, à partir de (3.19), on peut montrer que si $1/\mathrm{cv}_2(y) = 0$ et $N \to \infty$, alors

$$C_2^R\left(\hat{t}_{y\pi}^*\right) \to f_1 \tag{3.20}$$

en considérant que f_1 et f_2 sont fixes. Rappelons que $|C_2(\hat{t}^*_{y\pi})| \leq |C_2^R(\hat{t}^*_{y\pi})|$. Il découle de (3.20) que $|C_2(\hat{t}^*_{y\pi})| \leq f_1$ lorsque l'on pose $1/\mathrm{cv}_2(y) = 0$ et $N = \infty$. Dans ce cas, une condition suffisante afin que $\widehat{\mathrm{V}}_2^R(\hat{t}^*_{y\pi})$ soit négligeable est que la fraction de sondage à la première phase soit petite. Ce résultat sera particulièrement important lorsque nous étudierons les estimateurs simplifiés de la variance des estimateurs de calage car ces derniers peuvent souvent s'écrire asymptotiquement en fonction de résidus dont la moyenne est égale à 0 (voir section 3.2).

3.1.3.2 Étude de la quantité $C_2(\hat{t}^*_{y\pi})$

À partir de (3.16) et (3.18), on peut calculer la contribution $C_2(\hat{t}^*_{y\pi})$ qui est donnée par :

$$\begin{aligned} C_2(\hat{t}^*_{y\pi}) &= \frac{\widehat{\mathrm{V}}_2^R(\hat{t}^*_{y\pi})}{\widehat{\mathrm{V}}(\hat{t}_{y\pi})} \\ &= (1-N^{-1}f_1^{-1})^{-1}(1-f_2)(1-f_1f_2)^{-1} \\ &\times \left\{ f_1 + N^{-1}\left(f_1^{-1}f_2^{-1} - f_2^{-1} - 1\right) - (1-f_1)\frac{1}{\mathrm{cv}_2(y)^2} \right\}. \end{aligned} \tag{3.21}$$

En supposant que f_1 et f_2 sont fixes et que $N \to \infty$, on a, à partir de l'expression (3.21) :

$$C_2(\hat{t}^*_{y\pi}) \to (1-f_2)(1-f_1f_2)^{-1}\left\{ f_1 - (1-f_1)\frac{1}{\mathrm{cv}_2(y)^2} \right\}. \tag{3.22}$$

Le comportement de (3.22) est illustré à la figure 3.1 pour différentes valeurs de $\mathrm{cv}_2(y)$. Les plages de valeurs de f_1 et f_2 colorées représentent les valeurs permettant d'obtenir $|C_2(\hat{t}^*_{y\pi})| \leq 5\%$.

On remarque que lorsque $\mathrm{cv}_2(y)$ est petit, le choix des fractions de sondage aux première et deuxième phases est relativement restreint. Autrement dit, il est difficile de trouver des couples (f_1, f_2) pour lesquels la contribution de $\widehat{\mathrm{V}}_2^R(\hat{t}^*_{y\pi})$ est négligeable. Par contre, lorsque le $\mathrm{cv}_2(y)$ s'accroît, il devient de plus en plus facile de trouver des situations pour lesquelles $\widehat{\mathrm{V}}_2^R(\hat{t}^*_{y\pi})$ est négligeable. Lorsque $N = \infty$ et $1/\mathrm{cv}_2(y) = 0$ on

peut montrer que

$$0 \leq C_2\left(\hat{t}^*_{y\pi}\right) \leq f_1$$

signifiant qu'une fraction de sondage de première phase négligeable est une condition suffisante pour que la contribution de $\widehat{\mathrm{V}}_2^R\left(\hat{t}^*_{y\pi}\right)$ à l'estimateur de variance soit négligeable. Une fraction de sondage élevée à la deuxième phase peut également contribuer à faire diminuer $|C_2\left(\hat{t}^*_{y\pi}\right)|$.

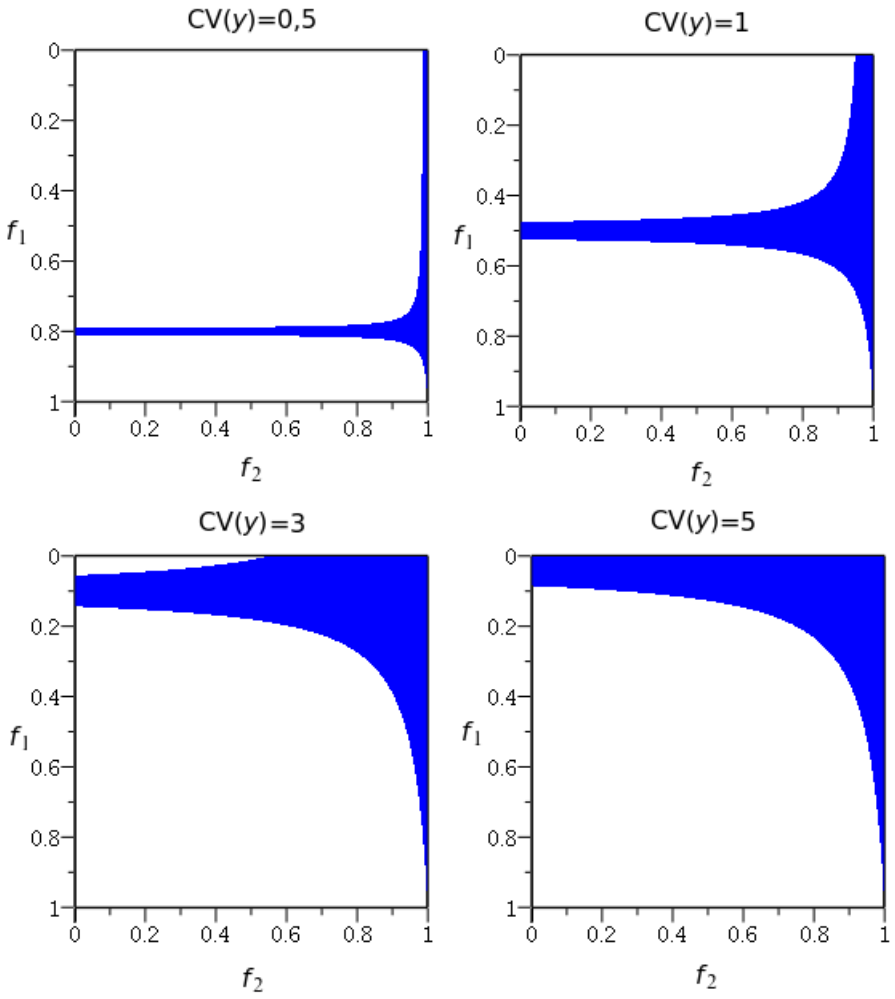

Figure 3.1 – Fractions de sondage permettant $|C_2\left(\hat{t}^*_{y\pi}\right)| \leq 5\%$ lorsque $N = \infty$.

Il peut également être intéressant de regarder le comportement de l'expression exacte de $C_2\left(\hat{t}^*_{y\pi}\right)$ en (3.21) pour différentes valeurs de N tel qu'illustré sur les graphiques suivants.

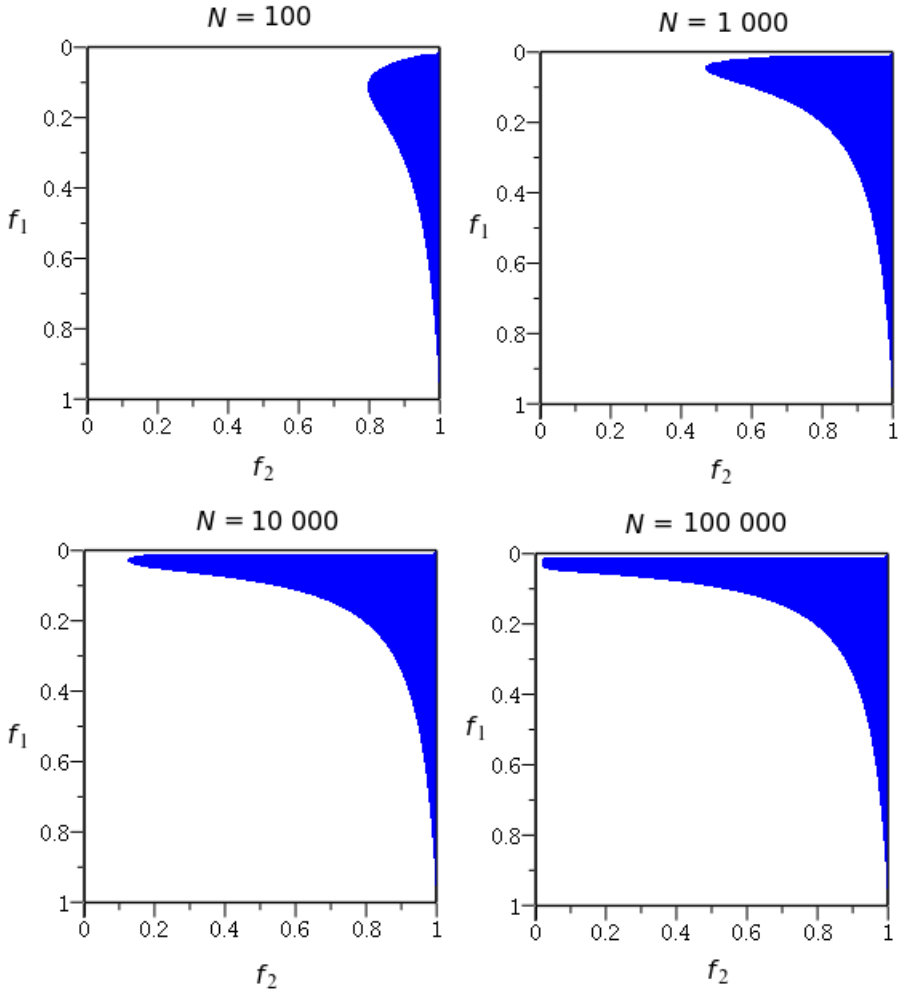

Figure 3.2 – Fractions de sondage permettant $|C_2\left(\hat{t}^*_{y\pi}\right)| \leq 5\%$ lorsque $\mathrm{cv}_2(y) = \infty$.

3.1.4 Cas d'un plan aléatoire simple sans remise à la deuxième phase

Lorsqu'un plan aléatoire simple sans remise est utilisé à la deuxième phase avec un plan quelconque à la première phase, on peut montrer que (voir développements à l'annexe I.4) :

$$\frac{\mathrm{E}_1\,\mathrm{E}_2\left(\widehat{\mathrm{V}}_2^R\left(\hat{t}_{y\pi}^*\right)|\boldsymbol{I}_1\right)}{\mathrm{E}_1\,\mathrm{E}_2\left(\widehat{\mathrm{V}}_2\left(\hat{t}_{y\pi}^*\right)|\boldsymbol{I}_1\right)} \approx \frac{-\left(\sum_{i\in U} y_i\right)^2 + n_1 \sum_{i\in U} y_i^2}{-\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}y_i y_j - \left(\sum_{i\in U} y_i\right)^2 + n_1 \sum_{i\in U}\frac{y_i^2}{\pi_{1i}}}. \tag{3.23}$$

En supposant les conditions de régularité

$$\sum_{i\in U} y_i^2 = O(N),$$

$$\sum_{i\in U} y_i = O(N),$$

$$\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}y_i y_j = O(N^2/n_1)$$

et

$$\sum_{i\in U}\frac{y_i^2}{\pi_{1i}} = O(N^2/n_1),$$

le numérateur est $O(N^2)$ et le dénominateur est $O(N^2)$ (voir annexe I.5 pour une justification des conditions de régularité). Le ratio est donc $O(1)$ si le rapport du dénominateur sur N^2 est borné loin de zéro. En général, si les ordres supposés ne surévaluent pas trop le véritable comportement asymptotique des suites, on ne peut pas conclure que le ratio (3.42) est négligeable et on ne s'attend pas à ce que la composante $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{y\pi}^*\right)$ le soit.

Toutefois, si on a $\sum_{i\in U} y_i = 0$ (ce cas survient dans le contexte des estimateurs de calage), le numérateur est $O(Nn_1)$ et le dénominateur est $O(N^2)$ et le ratio (3.42) est $O(n_1/N)$. Ainsi, si la fraction de sondage à la première phase n_1/N est négligeable, le ratio est petit.

3.2 Estimation simplifiée de la variance de l'estimateur de calage pour les plans à deux phases

Haziza et Beaumont (2005) se sont intéressés à l'estimation simplifiée de la variance pour l'estimateur de Hájek donné par (2.27), qui est un cas particulier d'un estimateur de calage (voir section 2.6.1). Toutefois, le problème de l'estimation simplifiée de la variance pour l'ensemble des estimateurs de calage n'a, à notre connaissance, toujours pas été traité dans la littérature. Dans cette section, nous présentons des estimateurs de la variance simplifié pour les estimateurs de calage, en distinguant bien les trois cas de disponibilité de l'information auxiliaire présentés précédemment aux sections 2.6.1 à 2.6.3. De plus, nous expliquons sous quelles conditions ces estimateurs de la variance simplifiés sont valides pour différents plans de sondage.

3.2.1 Information auxiliaire de type (1) seulement

Dans cette section, nous considérons le cas de l'estimateur de calage lorsque l'information auxiliaire est disponible au niveau de la population tel que décrit à la section 2.6.1. Rappelons que l'estimateur de calage est donné par (2.31), sa variance asymptotique par (2.36) et un estimateur de la variance par (2.37) dont l'expression est reportée ici :

$$\begin{aligned}\widehat{\mathrm{V}}_p\left(\hat{t}_{yC1}^*\right) &= \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{1ij}}{\pi_{2ij}(\boldsymbol{I}_1)}e_{1i}e_{1j} + \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{2ij}(\boldsymbol{I}_1)\frac{e_{1i}}{\pi_{1i}}\frac{e_{1j}}{\pi_{1j}} \\ &\equiv \widehat{\mathrm{V}}_1\left(\hat{t}_{yC1}^*\right) + \widehat{\mathrm{V}}_2\left(\hat{t}_{yC1}^*\right),\end{aligned} \tag{3.24}$$

où $e_{1i} = y_i - \boldsymbol{x}_{1i}'\hat{\boldsymbol{B}}_1$ et

$$\hat{\boldsymbol{B}}_1 = \left(\sum_{i\in s_2} d_i q_i \boldsymbol{x}_{1i}\boldsymbol{x}_{1i}'\right)^{-1}\left(\sum_{i\in s_2} d_i q_i \boldsymbol{x}_{1i} y_i\right).$$

L'expression (3.24) est identique à l'expression de l'estimateur de la variance de l'estimateur par double dilatation donné par (3.1) en remplaçant y par e_1. Ainsi, on obtient

un estimateur simplifié de la variance de la même façon qu'à la section 3.1 :

$$\begin{aligned}\widehat{\mathrm{V}}\big(\hat{t}^*_{yC1}\big) &= \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{1ij}\frac{e_{1i}}{\pi_{2i}(\boldsymbol{I}_1)}\frac{e_{1j}}{\pi_{2j}(\boldsymbol{I}_1)} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{2ij}(\boldsymbol{I}_1)}{\pi_{1ij}}e_{1i}e_{1j} \\ &\equiv \widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{yC1}\big) + \widehat{\mathrm{V}}_2^R\big(\hat{t}^*_{yC1}\big).\end{aligned} \tag{3.25}$$

qui est identique à l'estimateur simplifié de la variance de l'estimateur par double dilatation donné par (3.2) en remplaçant y par e_1.

Le terme $\widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{yC1}\big)$ peut s'écrire sous la forme

$$\widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{yC1}\big) = \sum_{i\in s_1}\sum_{j\in s_1}\Delta_{1ij}z_iz_j, \tag{3.26}$$

où

$$z_i = \frac{e_{1i}}{\pi_{2i}(\boldsymbol{I}_1)}I_{2i} \tag{3.27}$$

et peut ainsi être calculé facilement en utilisant les procédures d'estimation de la variance pour les plans à une phase.

Afin d'étudier les conditions sous lesquelles le terme $\widehat{\mathrm{V}}_2^R\big(\hat{t}^*_{yC1}\big)$ est négligeable, on utilise le fait que les quantités $\widehat{\mathrm{V}}_1\big(\hat{t}^*_{yC1}\big)$, $\widehat{\mathrm{V}}_2\big(\hat{t}^*_{yC1}\big)$, $\widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{yC1}\big)$ et $\widehat{\mathrm{V}}_2^R\big(\hat{t}^*_{yC1}\big)$ sont identiques aux quantités $\widehat{\mathrm{V}}_1\big(\hat{t}^*_{y\pi}\big)$, $\widehat{\mathrm{V}}_2\big(\hat{t}^*_{y\pi}\big)$, $\widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{y\pi}\big)$ et $\widehat{\mathrm{V}}_2^R\big(\hat{t}^*_{y\pi}\big)$, respectivement, lorsque l'on remplace y par e_1. De ce fait, les quantités $C_2\big(\hat{t}^*_{yC1}\big)$ et $C_2^R\big(\hat{t}^*_{yC1}\big)$ sont identiques aux quantités $C_2\big(\hat{t}^*_{y\pi}\big)$ et $C_2^R\big(\hat{t}^*_{y\pi}\big)$, respectivement, lorsque l'on remplace y par e_1.

Ainsi, dans le cas où un plan de Poisson est utilisé à la deuxième phase, on obtient

$$\big|C_2^R\big(\hat{t}^*_{yC1}\big)\big| \leq \max\big(\pi_{1i}\big)$$

puisque le résultat (3.8) ne dépend pas de la variable d'intérêt y. En supposant que $\max\big(\pi_{1i}\big) = O(n_1/N)$, le terme $\widehat{\mathrm{V}}_2^R$ est négligeable lorsque la fraction de sondage à la première phase, $f_1 = n_1/N$, est négligeable.

Dans le cas où un plan à deux degrés est utilisé, on obtient

$$\left|C_2^R\left(\hat{t}_{yC1}^*\right)\right| \leq \max\left(\pi_{1g}\right)$$

puisque le résultat (3.9) ne dépend pas de la variable d'intérêt y. En supposant que $\max\left(\pi_{1g}\right) = O(n/N)$, le terme $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC1}^*\right)$ est négligeable lorsque la fraction de sondage au premier degré, n/N, est négligeable.

Dans le cas où un plan aléatoire simple sans remise est utilisé aux deux phases, on obtient

$$C_2^R\left(\hat{t}_{yC1}^*\right) \to f_1$$

lorsque $N \to \infty$ et $\mathrm{cv}_2^{-1}(e_1) = 0$ (voir résultat (3.20)). Sous ces conditions, on obtient également

$$C_2\left(\hat{t}_{yC1}^*\right) \to (1-f_2)(1-f_1f_2)^{-1}f_1, \tag{3.28}$$

(voir résultat (3.22)). Or, il est facile d'avoir $\mathrm{cv}_2^{-1}(e_1) = 0$ en choisissant des estimateurs de calage qui satisfont à $q_i = \boldsymbol{\alpha}'\boldsymbol{x}_{1i}$ pour un certain vecteur $\boldsymbol{\alpha}$ de constantes connues (voir résultat (2.38)). L'estimateur par le ratio et l'estimateur de Hájek en sont des exemples. Ainsi, lorsque $N \to \infty$ et $1/\mathrm{cv}_2(e_1) = 0$, le terme $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC1}^*\right)$ est négligeable lorsque la fraction de sondage à la première phase, $f_1 = n_1/N$, est négligeable. De plus, la contribution du terme $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC1}^*\right)$ décroît à mesure que f_2 croît pour une valeur fixée de f_1.

Finalement, lorsqu'un plan aléatoire simple sans remise est utilisé à la première phase avec un plan quelconque à la deuxième phase, on a que

$$\frac{\mathrm{E}_1\mathrm{E}_2\left(\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC1}^*\right)|\boldsymbol{I}_1\right)}{\mathrm{E}_1\mathrm{E}_2\left(\widehat{\mathrm{V}}_2\left(\hat{t}_{yC1}^*\right)|\boldsymbol{I}_1\right)} \approx \frac{-\left(\sum_{i\in U}E_{1i}\right)^2 + n_1\sum_{i\in U}E_{1i}^2}{-\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}E_{1i}E_{1j} - \left(\sum_{i\in U}E_{1i}\right)^2 + n_1\sum_{i\in U}\frac{E_{1i}^2}{\pi_{1i}}}. \tag{3.29}$$

Cette expression est identique à (3.42) en remplaçant y_i par E_{1i}. Lorsqu'on utilise des estimateurs de calage satisfaisant à la contrainte $q_i = \boldsymbol{\alpha}'\boldsymbol{x}_{1i}$ pour un certain vecteur $\boldsymbol{\alpha}$ de

constantes connues, on a $\sum_{i\in U} E_{1i} = 0$ et il s'ensuit

$$\frac{\mathrm{E}_1\,\mathrm{E}_2\left(\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC1}^*\right)|\boldsymbol{I}_1\right)}{\mathrm{E}_1\,\mathrm{E}_2\left(\widehat{\mathrm{V}}_2\left(\hat{t}_{yC1}^*\right)|\boldsymbol{I}_1\right)} \approx \frac{n_1\sum_{i\in U} E_{1i}^2}{-\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}E_{1i}E_{1j} + n_1\sum_{i\in U}\frac{E_{1i}^2}{\pi_{1i}}}. \tag{3.30}$$

En supposant les conditions de régularité

$$\sum_{i\in U} E_{1i}^2 = O(N),$$

$$\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}E_{1i}E_{1j} = O(N^2/n_1)$$

et

$$\sum_{i\in U}\frac{E_{1i}^2}{\pi_{1i}} = O(N^2/n_1),$$

le numérateur est $O(Nn_1)$ et le dénominateur est $O(N^2)$ et donc le ratio est $O(n_1/N)$ si le rapport du dénominateur sur N^2 est borné loin de zéro. Ainsi, si la fraction de sondage à la première phase n_1/N est négligeable, on s'attend à ce que la composante $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC1}^*\right)$ de l'estimateur de variance soit négligeable.

3.2.2 Information auxiliaire de type (2) seulement

Dans cette section, nous considérons le cas de l'estimateur de calage lorsque l'information auxiliaire est disponible au niveau de l'échantillon de première phase s_1 tel que décrit à la section 2.31. Nous cherchons un estimateur simplifié de la variance de l'estimateur de calage (2.44). Nous considérons l'estimateur de variance alternatif (2.48) dont l'expression est reportée ici

$$\widehat{\mathrm{V}}\left(\hat{t}_{yC2}^*\right)^{\mathrm{alt}} = \widehat{\mathrm{V}}_1\left(\hat{t}_{yC2}^*\right) + \widehat{\mathrm{V}}_2\left(\hat{t}_{yC2}^*\right), \tag{3.31}$$

où

$$\widehat{\mathrm{V}}_1\left(\hat{t}_{yC2}^*\right) = \sum_{i\in s_1}\sum_{j\in s_1}\Delta_{1ij}\boldsymbol{x}_{2i}'\hat{\boldsymbol{B}}_2\boldsymbol{x}_{2j}'\hat{\boldsymbol{B}}_2 + 2\sum_{i\in s_1}\sum_{j\in s_2}\Delta_{1ij}\boldsymbol{x}_{2i}'\hat{\boldsymbol{B}}_2\frac{e_{2j}}{\pi_{2j}\left(\boldsymbol{I}_1\right)} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{1ij}}{\pi_{2ij}\left(\boldsymbol{I}_1\right)}e_{2i}e_{2j}$$

et

$$\widehat{\mathrm{V}}_2\big(\hat{t}^*_{yC2}\big) = \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{2ij}\big(\boldsymbol{I}_1\big)\frac{e_{2i}}{\pi_{1i}}\frac{e_{2j}}{\pi_{1j}}.$$

Comme à la section 3.1, on peut utiliser les équivalences

$$\frac{1}{\pi_{2ij}\big(\boldsymbol{I}_1\big)} = \frac{1}{\pi_{2i}\big(\boldsymbol{I}_1\big)}\frac{1}{\pi_{2j}\big(\boldsymbol{I}_1\big)} - \Delta_{2ij}\big(\boldsymbol{I}_1\big)$$

et

$$\frac{1}{\pi_{1i}}\frac{1}{\pi_{1j}} = \frac{1}{\pi_{1ij}} + \Delta_{1ij}$$

afin d'écrire

$$\begin{aligned}&\sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{1ij}}{\pi_{2ij}\big(\boldsymbol{I}_1\big)}e_{2i}e_{2j} + \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{2ij}\big(\boldsymbol{I}_1\big)\frac{e_{2i}}{\pi_{1i}}\frac{e_{2j}}{\pi_{1j}}\\ &= \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{1ij}\frac{e_{2i}}{\pi_{2i}\big(\boldsymbol{I}_1\big)}\frac{e_{2j}}{\pi_{2j}\big(\boldsymbol{I}_1\big)} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{2ij}\big(\boldsymbol{I}_1\big)}{\pi_{1ij}}e_{2i}e_{2j}. \qquad (3.32)\end{aligned}$$

En utilisant l'équivalence (3.32) dans l'équation (3.31) on obtient

$$\widehat{\mathrm{V}}\big(\hat{t}^*_{yC2}\big)^{\mathrm{alt}} = \widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{yC2}\big) + \widehat{\mathrm{V}}_2^R\big(\hat{t}^*_{yC2}\big), \qquad (3.33)$$

où

$$\widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{yC2}\big) = \sum_{i\in s_1}\sum_{j\in s_1}\Delta_{1ij}\boldsymbol{x}_i'\hat{\boldsymbol{B}}_2\boldsymbol{x}_{2j}'\hat{\boldsymbol{B}} + 2\sum_{i\in s_1}\sum_{j\in s_2}\Delta_{1ij}\boldsymbol{x}_{2i}'\hat{\boldsymbol{B}}_2\frac{e_{2j}}{\pi_{2j}\big(\boldsymbol{I}_1\big)} + \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{1ij}\frac{e_{2i}}{\pi_{2i}\big(\boldsymbol{I}_1\big)}\frac{e_{2j}}{\pi_{2j}\big(\boldsymbol{I}_1\big)}$$

et

$$\widehat{\mathrm{V}}_2^R\big(\hat{t}^*_{yC2}\big) = \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{2ij}\big(\boldsymbol{I}_1\big)}{\pi_{1ij}}e_{2i}e_{2j}.$$

Notons que le terme $\widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{yC2}\big)$ peut s'écrire comme

$$\widehat{\mathrm{V}}_1^R\big(\hat{t}^*_{yC2}\big) = \sum_{i\in s_1}\sum_{j\in s_1}\Delta_{1ij}z_iz_j, \qquad (3.34)$$

où

$$z_i = \boldsymbol{x}_{2i}'\hat{\boldsymbol{B}}_{\boldsymbol{2}} + \frac{e_{2i}}{\pi_{2i}(\boldsymbol{I}_1)} I_{2i}. \tag{3.35}$$

Ainsi, dans le cas où un plan de Poisson est utilisé à la deuxième phase, on obtient encore

$$\left|C_2^R\left(\hat{t}_{yC2}^*\right)\right| \leq \max\left(\pi_{1i}\right)$$

puisque le résultat (3.8) ne dépend pas de la variable d'intérêt y. En supposant que $\max\left(\pi_{1i}\right) = O(n_1/N)$, le terme $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC2}^*\right)$ est négligeable lorsque la fraction de sondage à la première phase, $f_1 = n_1/N$, est négligeable.

Dans le cas où un plan à deux degrés est utilisé, on obtient encore

$$\left|C_2^R\left(\hat{t}_{yC2}^*\right)\right| \leq \max\left(\pi_{1g}\right)$$

puisque le résultat (3.9) ne dépend pas de la variable d'intérêt y. En supposant que $\max\left(\pi_{1g}\right) = O(n/N)$, le terme $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC2}^*\right)$ est négligeable lorsque la fraction de sondage au premier degré, n/N, est négligeable.

Dans le cas où un plan aléatoire simple sans remise est utilisé aux deux phases, on obtient

$$C_2^R\left(\hat{t}_{yC2}^*\right) \to f_1$$

lorsque $N \to \infty$ et $\mathrm{cv}_2^{-1}(e_2) = 0$ (voir résultat (3.20)). Or, il est facile d'avoir $\mathrm{cv}_2^{-1}(e_2) = 0$ en choisissant des estimateurs de calage qui satisfont à $q_i = \boldsymbol{\alpha}'\boldsymbol{x}_{2i}$ pour un certain vecteur $\boldsymbol{\alpha}$ de constantes connues (voir résultat (2.49)). L'estimateur par le ratio en est un exemple. Ainsi, lorsque $N \to \infty$ et $\mathrm{cv}_2^{-1}(e_2) = 0$, le terme $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC2}^*\right)$ est négligeable lorsque la fraction de sondage à la première phase, $f_1 = n_1/N$, est négligeable.

Finalement, lorsqu'un plan aléatoire simple sans remise est utilisé à la première phase avec un plan quelconque à la deuxième phase, on a que

$$\frac{\mathrm{E}_1\,\mathrm{E}_2\left(\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC2}^*\right)|\boldsymbol{I}_1\right)}{\mathrm{E}_1\,\mathrm{E}_2\left(\widehat{\mathrm{V}}_2\left(\hat{t}_{yC2}^*\right)|\boldsymbol{I}_1\right)} \approx \frac{-\left(\sum_{i\in U} E_{2i}\right)^2 + n_1 \sum_{i\in U} E_{2i}^2}{-\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}E_{2i}E_{2j} - \left(\sum_{i\in U} E_{2i}\right)^2 + n_1 \sum_{i\in U}\frac{E_{2i}^2}{\pi_{1i}}}. \tag{3.36}$$

Cette expression est identique à (3.42) en remplaçant y_i par E_{2i}. Lorsqu'on utilise des estimateurs de calage satisfaisant à la contrainte $q_i = \boldsymbol{\alpha}' \boldsymbol{x}_{2i}$ pour un certain vecteur $\boldsymbol{\alpha}$ de constantes connues, on a $\sum_{i\in U} E_{2i} = 0$ et donc

$$\frac{\mathrm{E}_1 \mathrm{E}_2 \left(\widehat{\mathrm{V}}_2^R \left(\hat{t}_{yC2}^*\right) | \boldsymbol{I}_1\right)}{\mathrm{E}_1 \mathrm{E}_2 \left(\widehat{\mathrm{V}}_2 \left(\hat{t}_{yC2}^*\right) | \boldsymbol{I}_1\right)} \approx \frac{n_1 \sum_{i\in U} E_{2i}^2}{-\sum_{i\in U} \sum_{j\in U} \Delta_{1ij} \pi_{1ij} E_{2i} E_{2j} + n_1 \sum_{i\in U} \frac{E_{2i}^2}{\pi_{1i}}}. \tag{3.37}$$

En supposant les conditions de régularité

$$\sum_{i\in U} E_{2i}^2 = O(N),$$

$$\sum_{i\in U} \sum_{j\in U} \Delta_{1ij} \pi_{1ij} E_{2i} E_{2j} = O(N^2/n_1),$$

et

$$\sum_{i\in U} \frac{E_{2i}^2}{\pi_{1i}} = O(N^2/n_1)$$

le numérateur est $O(Nn_1)$ et le dénominateur est $O(N^2)$ et donc le ratio est $O(n_1/N)$ si le rapport du dénominateur sur N^2 est borné loin de zéro. Ainsi, si la fraction de sondage à la première phase n_1/N est négligeable, on s'attend à ce que la composante $\widehat{\mathrm{V}}_2^R \left(\hat{t}_{yC2}^*\right)$ de l'estimateur de variance soit négligeable.

3.2.3 Information auxiliaire de types (1) et (2)

Dans cette section, nous considérons le cas de l'estimateur de calage lorsque l'information auxiliaire est disponible aux deux niveaux, s_1 et U, tel que décrit à la section 2.6.3. Nous cherchons un estimateur simplifié de la variance de l'estimateur de calage (2.51). Pour ce faire, nous considérons l'estimateur de variance alternatif (2.64) dont l'expression est reportée ici

$$\widehat{\mathrm{V}}\left(\hat{t}_{yC}^*\right)^{\mathrm{alt}} = \widehat{\mathrm{V}}_1\left(\hat{t}_{yC}^*\right) + \widehat{\mathrm{V}}_2\left(\hat{t}_{yC}^*\right), \tag{3.38}$$

où

$$\widehat{\mathrm{V}}_1\big(\hat{t}_{yC}^*\big) = \sum_{i\in s_1}\sum_{j\in s_1}\Delta_{1ij}(\boldsymbol{x}_i'\hat{\boldsymbol{B}}-\boldsymbol{x}_{1i}'\hat{\boldsymbol{B}}_1)(\boldsymbol{x}_j'\hat{\boldsymbol{B}}-\boldsymbol{x}_{1j}'\hat{\boldsymbol{B}}_1)+2\sum_{i\in s_1}\sum_{j\in s_2}\Delta_{1ij}(\boldsymbol{x}_i'\hat{\boldsymbol{B}}-\boldsymbol{x}_{1i}'\hat{\boldsymbol{B}}_1)\frac{e_j}{\pi_{2j}\big(\boldsymbol{I}_1\big)} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{1ij}}{\pi_{2ij}\big(\boldsymbol{I}_1\big)}e_ie_j$$

et

$$\widehat{\mathrm{V}}_2\big(\hat{t}_{yC}^*\big) = \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{2ij}\big(\boldsymbol{I}_1\big)\frac{e_i}{\pi_{1i}}\frac{e_j}{\pi_{1j}}.$$

Comme à la section 3.1, on peut utiliser les équivalences

$$\frac{1}{\pi_{2ij}\big(\boldsymbol{I}_1\big)} = \frac{1}{\pi_{2i}\big(\boldsymbol{I}_1\big)}\frac{1}{\pi_{2j}\big(\boldsymbol{I}_1\big)} - \Delta_{2ij}\big(\boldsymbol{I}_1\big)$$

et

$$\frac{1}{\pi_{1i}}\frac{1}{\pi_{1j}} = \frac{1}{\pi_{1ij}} + \Delta_{1ij}$$

afin d'écrire

$$\begin{aligned}&\sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{1ij}}{\pi_{2ij}\big(\boldsymbol{I}_1\big)}e_ie_j + \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{2ij}\big(\boldsymbol{I}_1\big)\frac{e_i}{\pi_{1i}}\frac{e_j}{\pi_{1j}}\\ &= \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{1ij}\frac{e_i}{\pi_{2i}\big(\boldsymbol{I}_1\big)}\frac{e_j}{\pi_{2j}\big(\boldsymbol{I}_1\big)} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{2ij}\big(\boldsymbol{I}_1\big)}{\pi_{1ij}}e_ie_j. \qquad (3.39)\end{aligned}$$

En utilisant l'équivalence (3.39) dans l'équation (3.38) on obtient

$$\widehat{\mathrm{V}}(\hat{t}_{yC}^*)^{\mathrm{alt}} = \widehat{\mathrm{V}}_1^R\big(\hat{t}_{yC}^*\big) + \widehat{\mathrm{V}}_2^R\big(\hat{t}_{yC}^*\big), \qquad (3.40)$$

où

$$\widehat{\mathrm{V}}_1^R\big(\hat{t}_{yC}^*\big) = \sum_{i\in s_1}\sum_{j\in s_1}\Delta_{1ij}(\boldsymbol{x}_i'\hat{\boldsymbol{B}}-\boldsymbol{x}_{1i}'\hat{\boldsymbol{B}}_1)(\boldsymbol{x}_j'\hat{\boldsymbol{B}}-\boldsymbol{x}_{1j}'\hat{\boldsymbol{B}}_1)+2\sum_{i\in s_1}\sum_{j\in s_2}\Delta_{1ij}(\boldsymbol{x}_i'\hat{\boldsymbol{B}}-\boldsymbol{x}_{1i}'\hat{\boldsymbol{B}}_1)\frac{e_j}{\pi_{2j}\big(\boldsymbol{I}_1\big)} + \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{1ij}\frac{e_i}{\pi_{2i}\big(\boldsymbol{I}_1\big)}\frac{e_j}{\pi_{2j}\big(\boldsymbol{I}_1\big)}$$

et

$$\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC}^*\right) = \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{2ij}\left(\boldsymbol{I}_1\right)}{\pi_{1ij}}e_i e_j.$$

Notons que le terme $\widehat{\mathrm{V}}_1^R\left(\hat{t}_{yC}^*\right)$ peut s'écrire

$$\widehat{\mathrm{V}}_1^R\left(\hat{t}_{yC}^*\right) = \sum_{i\in s_1}\sum_{j\in s_1}\Delta_{1ij}z_i z_j$$

où

$$z_i = \boldsymbol{x}_i'\hat{\boldsymbol{B}} - \boldsymbol{x}_{1i}'\hat{\boldsymbol{B}}_1 + \frac{e_i}{\pi_{2i}\left(\boldsymbol{I}_1\right)}I_{2i}. \tag{3.41}$$

Ainsi, dans le cas où un plan de Poisson est utilisé à la deuxième phase, on obtient encore

$$\left|C_2^R\left(\hat{t}_{yC}^*\right)\right| \leq \max\left(\pi_{1i}\right)$$

puisque le résultat (3.8) ne dépend pas de la variable d'intérêt y. En supposant que $\max\left(\pi_{1i}\right) = O(n_1/N)$, le terme $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC}^*\right)$ est négligeable lorsque la fraction de sondage à la première phase, $f_1 = n_1/N$, est négligeable.

Dans le cas où un plan à deux degrés est utilisé, on obtient encore

$$\left|C_2^R\left(\hat{t}_{yC}^*\right)\right| \leq \max\left(\pi_{1g}\right)$$

puisque le résultat (3.9) ne dépend pas de la variable d'intérêt y. En supposant que $\max\left(\pi_{1g}\right) = O(n/N)$, le terme $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC}^*\right)$ est négligeable lorsque la fraction de sondage au premier degré, n/N, est négligeable.

Dans le cas où un plan aléatoire simple sans remise est utilisé aux deux phases, on obtient

$$C_2^R\left(\hat{t}_{yC}^*\right) \to f_1$$

lorsque $N \to \infty$ et $\mathrm{cv}_2^{-1}(e) = 0$ (voir résultat (3.20)). Or, il est facile d'avoir $\mathrm{cv}_2^{-1}(e) = 0$ en choisissant des estimateurs de calage qui satisfont à $q_i = \boldsymbol{\alpha}'\boldsymbol{x}_i$ pour un certain vecteur $\boldsymbol{\alpha}$ de constantes connues (voir résultat (2.65)). Ainsi, lorsque $N \to \infty$ et $\mathrm{cv}_2^{-1}(e) = 0$, le terme $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC}^*\right)$ est négligeable lorsque la fraction de sondage à la première phase,

$f_1 = n_1/N$, est négligeable.

Finalement, lorsqu'un plan aléatoire simple sans remise est utilisé à la première phase avec un plan quelconque à la deuxième phase, on a que

$$\frac{\mathrm{E}_1\,\mathrm{E}_2\left(\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC}^*\right)|\boldsymbol{I}_1\right)}{\mathrm{E}_1\,\mathrm{E}_2\left(\widehat{\mathrm{V}}_2\left(\hat{t}_{yC}^*\right)|\boldsymbol{I}_1\right)} \approx \frac{-\left(\sum_{i\in U} E_i\right)^2 + n_1 \sum_{i\in U} E_i^2}{-\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}E_iE_j - \left(\sum_{i\in U} E_i\right)^2 + n_1\sum_{i\in U}\frac{E_i^2}{\pi_{1i}}}. \tag{3.42}$$

Cette expression est identique à (3.42) en remplaçant y_i par E_i. Lorsqu'on utilise des estimateurs de calage satisfaisant à la contrainte $q_i = \boldsymbol{\alpha}'\boldsymbol{x}_i$ pour un certain vecteur $\boldsymbol{\alpha}$ de constantes connues, on a $\sum_{i\in U} E_i = 0$ et donc

$$\frac{\mathrm{E}_1\,\mathrm{E}_2\left(\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC}^*\right)|\boldsymbol{I}_1\right)}{\mathrm{E}_1\,\mathrm{E}_2\left(\widehat{\mathrm{V}}_2\left(\hat{t}_{yC}^*\right)|\boldsymbol{I}_1\right)} \approx \frac{n_1 \sum_{i\in U} E_i^2}{-\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}E_iE_j + n_1\sum_{i\in U}\frac{E_i^2}{\pi_{1i}}}. \tag{3.43}$$

En supposant les conditions de régularité

$$\sum_{i\in U} E_i^2 = O(N),$$

$$\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}E_iE_j = O(N^2/n_1),$$

et

$$\sum_{i\in U}\frac{E_i^2}{\pi_{1i}} = O(N^2/n_1),$$

le numérateur est $O(Nn_1)$ et le dénominateur est $O(N^2)$ et donc le ratio est $O(n_1/N)$ si le rapport du dénominateur sur N^2 est borné loin de zéro. Ainsi, si la fraction de sondage à la première phase n_1/N est négligeable, on s'attend à ce que la composante $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{yC}^*\right)$ de l'estimateur de variance soit négligeable.

3.3 Justification par l'approche renversée

Nous considérons l'approche renversée proposée par Fay (1991) et développée par Shao et Steel (1999) dans le cadre de la non-réponse dans les enquêtes. En fait, la si-

tuation qui prévaut en présence de non réponse peut être vue comme un échantillonnage à deux phases. À la première phase, un échantillon de la population est tiré selon un certain plan de sondage. L'ensemble (aléatoire) des répondants, quant à lui, peut être vu comme un échantillon de deuxième phase, généré au moyen du mécanisme (inconnu) de non-réponse. Il est à noter que dans un cadre de non-réponse, la propriété d'invariance est généralement satisfaite, car le mécanisme de non-réponse est généralement indépendant du résultat de l'échantillonnage de première phase. L'approche renversée suggère de renverser l'ordre des deux phases : d'abord, en appliquant le mécanisme de non-réponse, la population est divisée en une population de répondants et une population de non-répondants. Ensuite, un sous-échantillon est tiré de la population ainsi divisée, selon le plan de sondage choisi.

L'approche renversée peut également s'avérer utile pour obtenir des estimateurs de la variance dans le contexte de l'échantillonnage à deux phase. Considérons, par exemple, l'estimateur par double dilatation donné par (2.28). Sa variance peut être exprimée comme suit :

$$\mathrm{V}_p\left(\hat{t}_{y\pi}\right) = \mathrm{E}_2\,\mathrm{V}_1\left(\hat{t}_{y\pi} \,\middle|\, \boldsymbol{I}_2\right) + \mathrm{V}_2\,\mathrm{E}_1\left(\hat{t}_{y\pi} \,\middle|\, \boldsymbol{I}_2\right), \tag{3.44}$$

où $\mathrm{E}_1(\cdot \,|\, \boldsymbol{I}_2)$ et $\mathrm{E}_2(\cdot)$ désignent respectivement l'espérance par rapport au plan de sondage $p_1(s_1 \,|\, \boldsymbol{I}_2)$ et $p_2(s_2)$. Notons que la variance $\mathrm{V}_p\left(\hat{t}_{y\pi}\right)$ est identique à celle obtenue dans la section 2.5. Autrement dit, l'approche renversée fournit une décomposition alternative à celle obtenue sous l'approche usuelle. Une différence importante réside dans le fait que, contrairement à l'approche usuelle, l'approche renversée peut être utilisée uniquement si la propriété d'invariance est satisfaite, ce que nous supposerons dans le reste de la présente section. Dans les sections 3.3.1 et 3.3.2, nous montrons que les estimateurs simplifiés de la variance obtenus au chapitre 4 pour les estimateurs par double dilatation et de calage peuvent être obtenus en utilisant l'approche renversée.

3.3.1 Cas de l'estimateur par double dilatation

Rappelons que l'estimateur par double dilatation est donné par

$$\hat{t}^*_{y\pi} = \sum_{i\in s_2} \frac{1}{\pi^*_i} y_i = \sum_{i\in s_2} d^*_i y_i.$$

Sa variance est donnée par (3.44), où

$$\begin{aligned}
\mathrm{E}_2\,\mathrm{V}_1\left(\hat{t}^*_{y\pi} \,\middle|\, \boldsymbol{I}_2\right) &= \mathrm{E}_2\,\mathrm{V}_1\left(\sum_{i\in U} d^*_i y_i I_{1i} I_{2i} \,\middle|\, \boldsymbol{I}_2\right) \\
&= \mathrm{E}_2\left(\sum_{i\in U}\sum_{j\in U} \Delta_{1ij}\pi_{1ij}\frac{y_i I_{2i}}{\pi_{2i}}\frac{y_j I_{2j}}{\pi_{2j}}\right)
\end{aligned}$$

et

$$\begin{aligned}
\mathrm{V}_2\,\mathrm{E}_1\left(\hat{t}^*_{y\pi} \,\middle|\, \boldsymbol{I}_2\right) &= \mathrm{V}_2\,\mathrm{E}_1\left(\sum_{i\in U} d^*_i y_i I_{1i} I_{2i} \,\middle|\, \boldsymbol{I}_2\right) \\
&= \mathrm{V}_2\left(\sum_{i\in U} \frac{1}{\pi_{2i}} y_i I_{2i}\right) \\
&= \sum_{i\in U}\sum_{j\in U} \Delta_{2ij}\pi_{2ij} y_i y_j.
\end{aligned}$$

Un estimateur sans biais au sens $\mathrm{E}_p\left(\cdot\right) \equiv \mathrm{E}_2\,\mathrm{E}_1\left(\cdot \,\middle|\, \boldsymbol{I}_2\right)$ de la variance (3.44) est obtenu en estimant sans biais chaque terme :

$$\sum_{i\in s_2}\sum_{j\in s_2} \Delta_{1ij}\frac{y_i}{\pi_{2i}}\frac{y_j}{\pi_{2j}} + \sum_{i\in s_2}\sum_{j\in s_2} \frac{\Delta_{2ij}}{\pi_{1ij}} y_i y_j. \tag{3.45}$$

L'estimateur de variance (3.45) est identique à l'estimateur simplifié de la variance (3.2).

3.3.2 Cas de l'estimateur de calage pour les plans à deux phases

Dans cette section, nous montrons comment obtenir les estimateurs de la variance simplifiés des estimateurs de calage obtenus dans chacune des trois situations de disponibilité de l'information auxiliaire à partir de l'approche renversée.

3.3.2.1 Information auxiliaire de type (1)

Rappelons que le développement en série de Taylor de l'estimateur $\hat{t}^*_{yC1}$ mène à :

$$\hat{t}^*_{yC1} \approx \sum_{i\in U} \boldsymbol{x}'_{1i}\boldsymbol{B}_1 + \sum_{i\in s_2} d^*_i E_{1i},$$

où $E_{1i} = y_i - \boldsymbol{x}'_{1i}\boldsymbol{B}_1$ et

$$\boldsymbol{B}_1 = \left(\sum_{i\in U} q_i \boldsymbol{x}_{1i}\boldsymbol{x}'_{1i}\right)^{-1}\left(\sum_{i\in U} q_i \boldsymbol{x}_{1i} y_i\right).$$

On a

$$\begin{aligned}
\mathrm{V}_p\left(\hat{t}^*_{yC1}\right) &= \mathrm{E}_2\,\mathrm{V}_1\left(\hat{t}^*_{yC1}\,\middle|\,\boldsymbol{I}_2\right) + \mathrm{V}_2\,\mathrm{E}_1\left(\hat{t}^*_{yC1}\,\middle|\,\boldsymbol{I}_2\right)\\
&\approx \mathrm{E}_2\left(\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}\frac{E_{1i}I_{2i}}{\pi_{2i}}\frac{E_{1j}I_{2j}}{\pi_{2j}}\right) + \mathrm{V}_2\left(\sum_{i\in U}\frac{1}{\pi_{2i}}E_{1i}I_{2i}\right)\\
&= \mathrm{E}_2\left(\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}\frac{E_{1i}I_{2i}}{\pi_{2i}}\frac{E_{1j}I_{2j}}{\pi_{2j}}\right) + \sum_{i\in U}\sum_{j\in U}\Delta_{2ij}\pi_{2ij}E_{1i}E_{1j}.
\end{aligned}$$

Ce développement est identique au développement dans le cas de l'estimateur par double dilatation en remplaçant y_i par E_{1i}. Un estimateur sans biais de $\mathrm{V}_p\left(\hat{t}^*_{yC1}\right)$ est obtenu en estimant sans biais chaque terme et en remplaçant E_{1i} par e_{1i} :

$$\sum_{i\in s_2}\sum_{j\in s_2}\Delta_{1ij}\frac{e_{1i}}{\pi_{2i}}\frac{e_{1j}}{\pi_{2j}} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{2ij}}{\pi_{1ij}}e_{1i}e_{1j}. \tag{3.46}$$

L'expression (3.46) est identique à l'expression (3.25).

3.3.2.2 Information auxiliaire de type (2)

Rappelons que le développement en série de Taylor de l'estimateur $\hat{t}^*_{yC2}$ mène à :

$$\hat{t}^*_{yC2} \approx \sum_{i\in s_1} d_{1i}\boldsymbol{x}'_{2i}\boldsymbol{B}_2 + \sum_{i\in s_2} d^*_i E_{2i},$$

où $E_{2i} = y_i - \boldsymbol{x}'_{2i}\boldsymbol{B}_2$ et

$$\boldsymbol{B}_2 = \left(\sum_{i\in U} q_i \boldsymbol{x}_{2i}\boldsymbol{x}'_{2i}\right)^{-1}\left(\sum_{i\in U} q_i \boldsymbol{x}_{2i} y_i\right).$$

On a

$$\mathrm{V}_p\left(\hat{t}^*_{yC2}\right) = \mathrm{E}_2\,\mathrm{V}_1\left(\hat{t}^*_{yC2} \,\middle|\, \boldsymbol{I}_2\right) + \mathrm{V}_2\,\mathrm{E}_1\left(\hat{t}^*_{yC2} \,\middle|\, \boldsymbol{I}_2\right),$$

où

$$\begin{aligned}\mathrm{E}_2\,\mathrm{V}_1\left(\hat{t}^*_{yC2} \,\middle|\, \boldsymbol{I}_2\right) &\approx \mathrm{E}_2\,\mathrm{V}_1\left(\sum_{i\in U} d_{1i} I_{1i}\left\{\boldsymbol{x}'_{2i}\boldsymbol{B}_2 + d_{2i}E_{2i}I_{2i}\right\} \,\middle|\, \boldsymbol{I}_2\right)\\ &= \mathrm{E}_2\left(\sum_{i\in U}\sum_{j\in U} \Delta_{1ij}\pi_{1ij}\left\{\boldsymbol{x}'_{2i}\boldsymbol{B}_2 + d_{2i}E_{2i}I_{2i}\right\}\left\{\boldsymbol{x}'_{2j}\boldsymbol{B}_2 + d_{2j}E_{2j}I_{2j}\right\}\right)\end{aligned}$$

et

$$\begin{aligned}\mathrm{V}_2\,\mathrm{E}_1\left(\hat{t}^*_{yC2} \,\middle|\, \boldsymbol{I}_2\right) &\approx \mathrm{V}_2\,\mathrm{E}_1\left(\sum_{i\in U} d_{1i} I_{1i}\left\{\boldsymbol{x}'_{2i}\boldsymbol{B}_2 + d_{2i}E_{2i}I_{2i}\right\} \,\middle|\, \boldsymbol{I}_2\right)\\ &= \mathrm{V}_2\left(\sum_{i\in U}\left\{\boldsymbol{x}'_{2i}\boldsymbol{B}_2 + d_{2i}E_{2i}I_{2i}\right\}\right)\\ &= \mathrm{V}_2\left(\sum_{i\in U} d_{2i}E_{2i}I_{2i}\right)\\ &= \sum_{i\in U}\sum_{j\in U}\Delta_{2ij}\pi_{2ij}E_{2i}E_{2j}.\end{aligned}$$

Un estimateur sans biais de $\mathrm{V}_p\left(\hat{t}^*_{yC2}\right)$ est obtenu en estimant sans biais chaque terme et en remplaçant E_{2i} par e_{2i} :

$$\sum_{i\in s_1}\sum_{j\in s_1}\Delta_{1ij}\left\{\boldsymbol{x}'_{2i}\hat{\boldsymbol{B}}_2 + d_{2i}e_{2i}I_{2i}\right\}\left\{\boldsymbol{x}'_{2j}\hat{\boldsymbol{B}}_2 + d_{2j}e_{2j}I_{2j}\right\} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{2ij}}{\pi_{1ij}}e_{2i}e_{2j}. \tag{3.47}$$

L'expression (3.47) est identique à l'expression (3.33).

3.3.2.3 Information auxiliaire de types (1) et (2)

Rappelons que le développement en série de Taylor de l'estimateur $\hat{t}^*_{yC}$ mène à :

$$\hat{t}^*_{yC} \approx \sum_{i\in U} \boldsymbol{x}'_{1i}\boldsymbol{B}_1 + \sum_{i\in s_1} d_{1i}\left(\boldsymbol{x}'_i\boldsymbol{B} - \boldsymbol{x}'_{1i}\boldsymbol{B}_1\right) + \sum_{i\in s_2} d^*_i E_i,$$

où $E_{1i} = y_i - \boldsymbol{x}'_{1i}\boldsymbol{B}_1$ et $E_i = y_i - \boldsymbol{x}'_i\boldsymbol{B}$ et

$$\boldsymbol{B}_1 = \left(\sum_{i\in U} q_{1i}\boldsymbol{x}_{1i}\boldsymbol{x}'_{1i}\right)^{-1}\left(\sum_{i\in U} q_{1i}\boldsymbol{x}_{1i}y_i\right)$$

et

$$\boldsymbol{B} = \left(\sum_{i\in U} q_i\boldsymbol{x}_i\boldsymbol{x}'_i\right)^{-1}\left(\sum_{i\in U} q_i\boldsymbol{x}_i y_i\right).$$

On a

$$\mathrm{V}_p\left(\hat{t}^*_{yC}\right) = \mathrm{E}_2\,\mathrm{V}_1\left(\hat{t}^*_{yC} \,\middle|\, \boldsymbol{I}_2\right) + \mathrm{V}_2\,\mathrm{E}_1\left(\hat{t}^*_{yC} \,\middle|\, \boldsymbol{I}_2\right),$$

où

$$\begin{aligned}
\mathrm{E}_2\,\mathrm{V}_1\left(\hat{t}^*_{yC} \,\middle|\, \boldsymbol{I}_2\right) &\approx \mathrm{E}_2\,\mathrm{V}_1\left(\sum_{i\in U} d_{1i}I_{1i}\left\{\boldsymbol{x}'_i\boldsymbol{B} - \boldsymbol{x}'_{1i}\boldsymbol{B}_1 + d_{2i}E_iI_{2i}\right\} \,\middle|\, \boldsymbol{I}_2\right) \\
&= \mathrm{E}_2\left(\sum_{i\in U}\sum_{j\in U} \Delta_{1ij}\pi_{1ij}\left\{\boldsymbol{x}'_i\boldsymbol{B} - \boldsymbol{x}'_{1i}\boldsymbol{B}_1 + d_{2i}E_iI_{2i}\right\}\left\{\boldsymbol{x}'_j\boldsymbol{B} - \boldsymbol{x}'_{1j}\boldsymbol{B}_1 + d_{2j}E_jI_{2j}\right\}\right)
\end{aligned}$$

et

$$\begin{aligned}
\mathrm{V}_2\,\mathrm{E}_1\left(\hat{t}^*_{yC} \,\middle|\, \boldsymbol{I}_2\right) &\approx \mathrm{V}_2\,\mathrm{E}_1\left(\sum_{i\in U} d_{1i}I_{1i}\left\{\boldsymbol{x}'_i\boldsymbol{B} - \boldsymbol{x}'_{1i}\boldsymbol{B}_1 + d_{2i}E_iI_{2i}\right\} \,\middle|\, \boldsymbol{I}_2\right) \\
&= \mathrm{V}_2\left(\sum_{i\in U}\left\{\boldsymbol{x}'_i\boldsymbol{B} - \boldsymbol{x}'_{1i}\boldsymbol{B}_1 + d_{2i}E_iI_{2i}\right\}\right) \\
&= \mathrm{V}_2\left(\sum_{i\in U} d_{2i}E_iI_{2i}\right) \\
&= \sum_{i\in U}\sum_{j\in U} \Delta_{2ij}\pi_{2ij}E_iE_j.
\end{aligned}$$

Un estimateur sans biais de $\mathrm{V}_p\left(\hat{t}_{yC}^*\right)$ est obtenu en estimant sans biais chaque terme et en remplaçant E_{1i} par e_{1i} de même que E_i par e_i :

$$\sum_{i\in s_1}\sum_{j\in s_1}\Delta_{1ij}\left\{\boldsymbol{x}_i'\hat{\boldsymbol{B}}-\boldsymbol{x}_{1i}'\hat{\boldsymbol{B}}_1+d_{2i}e_iI_{2i}\right\}\left\{\boldsymbol{x}_j'\hat{\boldsymbol{B}}-\boldsymbol{x}_{1j}'\hat{\boldsymbol{B}}_1+d_{2i}e_jI_{2j}\right\}+\sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{2ij}}{\pi_{1ij}}e_ie_j. \quad (3.48)$$

L'expression (3.48) est identique à l'expression (3.40).

CHAPITRE 4

ÉTUDES PAR SIMULATION

Dans ce chapitre, nous effectuons des études par simulation afin d'étudier la performance des estimateurs ponctuels et des estimateurs de variance présentés aux chapitres 2 et 3 et d'appuyer les résultats théoriques développés. Dans la première étude, nous considérons le cas où un plan aléatoire simple sans remise est utilisé aux deux phases. Dans la seconde étude, nous considérons le cas où un plan de Bernoulli est utilisé à la deuxième phase. Finalement, dans la troisième étude, nous traitons le cas du plan à deux degrés. Nous allons d'abord décrire les études par simulation, puis nous présenterons et discuterons les résultats.

4.1 Description des études par simulation

Dans le cadre de chacune des études, nous générons quelques populations à partir de modèles statistiques. Dans chaque population donnée, un grand nombre d'échantillons, K, est tiré, selon un plan à deux phases. À partir de chaque échantillon, nous calculons les valeurs des estimateurs par double dilatation et/ou de type calage ainsi que les estimateurs de variance correspondants. Dans les sections 4.1.1, 4.1.2 et 4.1.3, nous décrivons en détail chacune des trois études par simulation. Puis, dans la section 4.1.4, nous présentons les mesures Monte Carlo utilisées afin d'évaluer la performance des estimateurs.

4.1.1 Étude 1 : plan aléatoire simple sans remise aux deux phases

Pour cette étude sont générées trois populations de taille 10 000 constituées de deux variables : une variable d'intérêt y et une variable auxiliaire x. Nous générons d'abord 10 000 réalisations de la variable auxiliaire x à partir d'une distribution beta de paramètres $\alpha = 10$ et $\beta = 3$. Puis, à partir des réalisations de la variable x, nous générons les

réalisations de la variable y en utilisant le modèle

$$y_i = 2x_i + \varepsilon_i,$$

où les ε_i sont des variables aléatoires indépendantes de loi normale d'espérance nulle et de variance σ^2. La valeur du paramètre σ^2, qui varie selon la population, est déterminée de façon à ce que le coefficient de corrélation, ρ, entre x et y soit de $0,5$, $0,7$ et $0,9$ dans la première, deuxième et troisième population, respectivement.

Dans une population donnée, nous sélectionnons $K = 25\ 000$ échantillons selon le plan à deux phases suivant. À la première phase, nous tirons un échantillon aléatoire simple sans remise de la population avec une fraction de sondage f_1. À la deuxième phase, nous tirons un échantillon aléatoire simple sans remise du premier échantillon avec une fraction de sondage f_2. Nous considérons plusieurs valeurs pour les fractions de sondage de première et de deuxième phase : $f_1 = 5$ %, 10 %, 20 % et $f_2 =$ 5 %, 10 %, 25 % et 50 %. À partir de chaque échantillon, nous calculons les estimateurs $\hat{t}^*_{y\pi}$, $\hat{t}^*_{yR1}$ et $\hat{t}^*_{yR2}$ donnés par (2.28), (2.39) et (2.50) ainsi que les estimateurs de variance $\widehat{\mathrm{V}}\left(\hat{t}^*_{y\pi}\right)$, $\widehat{\mathrm{V}}\left(\hat{t}^*_{yR1}\right)$ et $\widehat{\mathrm{V}}\left(\hat{t}^*_{yR2}\right)^{\mathrm{alt}}$ donnés par (3.2), (3.25) et (3.33) et les estimateurs de variance simplifiés $\widehat{\mathrm{V}}^R_1\left(\hat{t}^*_{y\pi}\right)$, $\widehat{\mathrm{V}}^R_1\left(\hat{t}^*_{yR1}\right)$ et $\widehat{\mathrm{V}}^R_1\left(\hat{t}^*_{yR2}\right)$ donnés par (3.4), (3.26) et (3.34).

4.1.2 Étude 2 : plan Bernoulli à la deuxième phase

Pour cette deuxième étude par simulation, nous utilisons les même populations que dans l'étude précédente. Dans une population donnée, nous sélectionnons $K = 25\ 000$ échantillons à partir de chaque population selon le plan à deux phases suivant. À la première phase, nous tirons un échantillon aléatoire simple sans remise de la population avec une fraction de sondage f_1. À la deuxième phase, nous tirons un échantillon Bernouilli de paramètre π_2. Nous considérons plusieurs valeurs pour f_1 et π_2 : $f_1 = 5$ %, 10 %, 20 % et $\pi_2 = 5$ %, 10%, 25 % et 50 %. À partir de chaque échantillon, nous calculons les estimateurs $\hat{t}^*_{y\pi}$, $\hat{t}^*_{yR1}$ et $\hat{t}^*_{yR2}$ donnés par (2.28), (2.39) et (2.50) ainsi que les

estimateurs de variance $\widehat{\mathrm{V}}\left(\hat{t}_{y\pi}^{*}\right)$, $\widehat{\mathrm{V}}\left(\hat{t}_{yR1}^{*}\right)^{\mathrm{alt}}$ et $\widehat{\mathrm{V}}\left(\hat{t}_{yR2}^{*}\right)^{\mathrm{alt}}$ donnés par (3.2), (3.25) et (3.33) et les estimateurs de variance simplifiés $\widehat{\mathrm{V}}_1^R\left(\hat{t}_{y\pi}^{*}\right)$, $\widehat{\mathrm{V}}_1^R\left(\hat{t}_{yR1}^{*}\right)$ et $\widehat{\mathrm{V}}_1^R\left(\hat{t}_{yR2}^{*}\right)$ donnés par (3.4), (3.26) et (3.34).

4.1.3 Étude 3 : plan à deux degrés

Dans le cadre de cette troisième étude, nous générons deux populations de taille $N = 500$ UPES. Dans chaque UPE, U_g, nous générons le nombre d'USE M_g selon le modèle

$$M_g = 10 + W,$$

où $W \sim Bin(25, 0.6)$. Chaque USE est constituée d'une variable d'intérêt y dont la i^{e} valeur dans l'UPE U_g est générée d'après le modèle

$$y_{gi} = 200 + \alpha_g + \varepsilon_{gi},$$

où les α_g sont des variables aléatoires indépendantes de loi normale d'espérance nulle et de variance σ_α^2 et les ε_{gi} sont des variables aléatoires indépendantes de loi normale d'espérance nulle et de variance σ_ε^2. Les valeurs de σ_α^2 et σ_ε^2 sont déterminées de façon à ce que le coefficient de corrélation intra-classe, donné par

$$\mathrm{ICC} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}, \tag{4.1}$$

soit égal à 5 % dans la première population et à 20 % dans la seconde.

Nous sélectionnons, dans chaque population, $K = 25\,000$ échantillons selon le plan à deux degrés suivant. Au premier degré, nous tirons un échantillon sans remise avec des probabilités d'inclusion inégales à l'aide de la méthode de Rao-Sampford (voir Rao, 1965 et Sampford, 1967). Nous considérons les valeurs de la fraction de sondage au premier degré $f_1 = n/N$ suivantes : 2 % et 10 %. Au deuxième degré, nous tirons un échantillon aléatoire simple sans remise de $m_g = 2,\ 5$ ou 10 USES dans chaque UPE.

À partir de chaque échantillon, nous calculons l'estimateur $\hat{t}^*_{y\pi}$ donné par (2.28), l'estimateur de variance $\widehat{\mathrm{V}}\left(\hat{t}^*_{y\pi}\right)$ donné par (3.2) ainsi que l'estimateur de variance simplifié $\widehat{\mathrm{V}}^R_1\left(\hat{t}^*_{y\pi}\right)$ donné par (3.4).

4.1.4 Mesures Monte Carlo

Les mesures Monte Carlo ont pour but d'évaluer la performance d'un estimateur dans le cadre d'une étude par simulation. Soit $\widehat{\theta}$ un estimateur d'un paramètre θ. On peut approximer l'espérance de $\widehat{\theta}$ par son espérance Monte Carlo qui est donnée par

$$\mathrm{E}_{\mathrm{MC}}(\widehat{\theta}) = \frac{1}{K}\sum_{k=1}^{K}\widehat{\theta}_k, \tag{4.2}$$

où $\widehat{\theta}_k$ est l'estimateur de θ calculé à partir de l'échantillon k. La loi des grands nombres garantit que $\mathrm{E}_{\mathrm{MC}}(\widehat{\theta})$ ne s'éloigne pas trop de $\mathrm{E}(\widehat{\theta})$ pourvu que K soit suffisamment grand. Comme mesure du biais de $\widehat{\theta}$, nous utilisons son biais relatif Monte Carlo donné par

$$\mathrm{RB}_{\mathrm{MC}}(\widehat{\theta}) = \frac{\mathrm{E}_{\mathrm{MC}}(\widehat{\theta}) - \theta}{\theta}, \tag{4.3}$$

où $\mathrm{E}_{\mathrm{MC}}(\widehat{\theta})$ est donné par (4.2).

La variance de l'estimateur $\widehat{\theta}$ peut être approximée par la variance Monte Carlo

$$\mathrm{V}_{\mathrm{MC}}(\widehat{\theta}) = \frac{1}{K}\sum_{k=1}^{K}\left[\widehat{\theta}_k - \mathrm{E}_{\mathrm{MC}}(\widehat{\theta})\right]^2. \tag{4.4}$$

Une autre mesure de l'efficacité de l'estimateur $\widehat{\theta}$, liée à la variance (4.4), est donnée par le coefficient de variation Monte Carlo :

$$\mathrm{CV}_{\mathrm{MC}}(\widehat{\theta}) = \frac{\sqrt{\mathrm{V}_{MC}(\widehat{\theta})}}{\theta}. \tag{4.5}$$

La probabilité de couverture d'un intervalle de confiance de niveau $1-\alpha$ pour le para-

mètre θ peut être approximée par

$$\mathrm{PC}_{\mathrm{MC}} = \frac{1}{K}\sum_{k=1}^{K} A_k,$$

où A_k est une variable indicatrice telle que $A_k = 1$, si l'intervalle de confiance pour le k^{e} échantillon contient le paramètre θ, et $A_k = 0$, sinon. Notons que l'intervalle de confiance de niveau 95 % pour le paramètre θ sous l'échantillon k est donné par

$$\hat{\theta}_k \pm 1{,}96\sqrt{\widehat{\mathrm{V}}(\widehat{\theta}_k)},$$

où $\widehat{\mathrm{V}}(\widehat{\theta}_k)$ est un estimateur de la variance de $\widehat{\theta}_k$.

Dans chacune des trois études par simulation, nous calculons des mesures Monte Carlo. Dans ce qui suit, nous utilisons la notation générique $\hat{t}_y$ pour désigner un des estimateurs ponctuels : $\hat{t}^*_{y\pi}$, $\hat{t}^*_{yC1}$ ou $\hat{t}^*_{yC2}$ selon le contexte. Nous utilisons également la notation générique $\widehat{\mathrm{V}}$ afin de désigner un estimateur de variance.

Afin d'étudier le comportement des estimateurs ponctuels en termes de biais, nous utilisons le biais relatif Monte Carlo donné par (4.3) en remplaçant $\widehat{\theta}$ par $\hat{t}_y$ et θ par t_y. De manière similaire, nous évaluons l'efficacité en calculant le coefficient de variation Monte Carlo à partir de la formule (4.5) en remplaçant $\widehat{\theta}$ par $\hat{t}_y$ et θ par t_y.

Afin d'étudier le problème de l'estimation de la variance, nous calculons le biais relatif Monte Carlo des estimateurs de variance simplifiés et non simplifiés en remplaçant $\widehat{\theta}$ par $\widehat{\mathrm{V}}$ et θ par $\mathrm{V}(\hat{t}_y)$ dans la formule (4.3). Nous calculons également l'espérance Monte Carlo des contributions $C_2(\cdot)$ et $C_2^R(\cdot)$ dans l'expression simplifiée de l'estimateur de variance à partir de la formule (4.2) en remplaçant $\widehat{\theta}$ par $C_2(\cdot)$ et $C_2^R(\cdot)$.

4.2 Résultats et discussion

Dans les paragraphes qui suivent, nous commentons uniquement les tableaux de résultats pour la population avec le coefficient de corrélation $\rho = 0,7$ et celle avec ICC $= 5$ %. Les tableaux des résultats pour les autres populations se trouvent à l'annexe II. Toutes les simulations ont été effectuées avec le logiciel SAS version 9.2 pour Windows.

Les tableaux 4.1, 4.2 et 4.3 présentent la performance des estimateurs ponctuels pour chacune des études pas simulation. Dans les trois cas, les résultats suggèrent que tous les estimateurs considérés ne présentent pratiquement pas de biais, ce qui est cohérent avec les résultats du chapitre 2. En effet, le biais relatif est dans tous les cas inférieur ou égal à $0,1$ %. On remarque également que, pour une valeur fixe de f_1, l'efficacité des estimateurs ponctuels augmente lorsque la fraction de sondage à la deuxième phase augmente. Par exemple, dans le tableau 4.3, lorsque $f_1 = 10$ % et $m_g = 2$, on a $\mathrm{CV}_{\mathrm{MC}}\left(\hat{t}^*_{y\pi}\right) = 2,3$ % alors que lorsque $m_g = 10$, on a $\mathrm{CV}_{\mathrm{MC}}\left(\hat{t}^*_{y\pi}\right) = 1,6$ %. De plus, dans les deux premières études par simulation, on remarque que l'estimateur $\hat{t}^*_{yR1}$ est toujours au moins aussi efficace que l'estimateur $\hat{t}^*_{yR2}$, qui lui est toujours plus efficace que l'estimateur $\hat{t}^*_{y\pi}$. Par exemple, dans le tableau 4.1, lorsque $f_1 = 10$ % et $f_2 = 25$ %, on a $\mathrm{CV}_{\mathrm{MC}}\left(\hat{t}^*_{yR1}\right) = 0,6$ %, $\mathrm{CV}_{\mathrm{MC}}\left(\hat{t}^*_{yR2}\right) = 0,7$ % et $\mathrm{CV}_{\mathrm{MC}}\left(\hat{t}^*_{y\pi}\right) = 1,1$ %. Il n'est pas surprenant que les estimateurs de calage soient plus efficaces que l'estimateur par double dilatation car ceux-ci incorporent de l'information auxiliaire fortement liée à la variable d'intérêt y. De plus, l'information auxiliaire utilisée par $\hat{t}^*_{yR1}$ est plus riche que celle utilisée par $\hat{t}^*_{yR2}$ car elle est connue au niveau de la population plutôt qu'au niveau de l'échantillon s_1.

		$\hat{t}^*_{y\pi}$		$\hat{t}^*_{yR1}$		$\hat{t}^*_{yR2}$	
f_1 (%)	f_2 (%)	RB_{MC} (%)	CV_{MC} (%)	RB_{MC} (%)	CV_{MC} (%)	RB_{MC} (%)	CV_{MC} (%)
5	5	0,0	3,5	0,0	1,9	0,0	2,0
5	10	0,0	2,5	0,0	1,4	0,0	1,5
5	25	0,0	1,6	0,0	0,9	0,0	1,1
5	50	0,0	1,1	0,0	0,6	0,0	0,9
10	5	0,0	2,4	0,0	1,3	0,0	1,4
10	10	0,0	1,7	0,0	1,0	0,0	1,0
10	25	0,0	1,1	0,0	0,6	0,0	0,7
10	50	0,0	0,8	0,0	0,4	0,0	0,6
20	5	0,0	1,7	0,0	1,0	0,0	1,0
20	10	0,0	1,2	0,0	0,7	0,0	0,7
20	25	0,0	0,8	0,0	0,4	0,0	0,5
20	50	0,0	0,5	0,0	0,3	0,0	0,4

Tableau 4.1 – Performance des estimateurs ponctuels dans le cadre de l'étude par simulation avec un plan aléatoire simple sans remise aux deux phases pour la population avec $\rho = 0,7$.

		$\hat{t}^*_{y\pi}$		$\hat{t}^*_{yR1}$		$\hat{t}^*_{yR2}$	
f_1 (%)	f_2 (%)	RB_{MC} (%)	CV_{MC} (%)	RB_{MC} (%)	CV_{MC} (%)	RB_{MC} (%)	CV_{MC} (%)
5	5	0,1	19,8	0,0	2,0	0,0	2,1
5	10	0,1	13,7	0,0	1,4	0,0	1,5
5	25	0,1	7,9	0,0	0,9	0,0	1,1
5	50	0,1	4,6	0,0	0,6	0,0	0,9
10	5	0,0	14,1	0,0	1,4	0,0	1,4
10	10	0,0	9,7	0,0	1,0	0,0	1,0
10	25	0,1	5,6	0,0	0,6	0,0	0,7
10	50	0,0	3,2	0,0	0,4	0,0	0,6
20	5	0,1	9,9	0,0	1,0	0,0	1,0
20	10	0,1	6,8	0,0	0,7	0,0	0,7
20	25	0,0	3,9	0,0	0,4	0,0	0,5
20	50	0,0	2,3	0,0	0,3	0,0	0,4

Tableau 4.2 – Performance des estimateurs ponctuels dans le cadre de l'étude par simulation avec un plan Bernoulli à la deuxième phase pour la population avec $\rho = 0,7$.

f_1 $(\%)$	m_g	$\mathrm{RB}_{\mathrm{MC}}\left(\hat{t}^*_{y\pi}\right)$ $(\%)$	$\mathrm{CV}_{\mathrm{MC}}\left(\hat{t}^*_{y\pi}\right)$ $(\%)$
2	2	0,0	5,2
2	5	0,0	4,1
2	10	0,0	3,6
10	2	0,0	2,3
10	5	0,0	1,8
10	10	0,0	1,6

Tableau 4.3 – Performance des estimateurs ponctuels dans le cadre de l'étude par simulation avec un plan à deux degrés pour la population avec ICC $= 5$ %.

Les tableaux 4.4, 4.5 et 4.6 exhibent le comportement des estimateurs de variance non simplifiés dans chacune des étude par simulation. Dans tous les scénarios, le biais de l'estimateur de variance est petit, ce qui est cohérent avec les résultats vus au chapitre 2. Par exemple, dans le tableau 4.5, lorsque $f_1 = 5$ % et $f_2 = 10$ %, on a $\mathrm{RB}_{\mathrm{MC}}\left[\widehat{\mathrm{V}}\left(\hat{t}^*_{y\pi}\right)\right] = -1,2$ %, $\mathrm{RB}_{\mathrm{MC}}\left[\widehat{\mathrm{V}}\left(\hat{t}^*_{yR1}\right)\right] = -3,2$ % et $\mathrm{RB}_{\mathrm{MC}}\left[\widehat{\mathrm{V}}\left(\hat{t}^*_{yR2}\right)\right] = -2,8$ %. Les pourcentages de couverture Monte Carlo sont tous entre $91,5$ % et $95,1$ % et sont plus près de 95 % lorsque f_1 est grande. Cela s'explique par le fait que les conditions du théorème limite central sont mieux respectées lorsque la taille d'échantillon est grande.

		$\widehat{\mathrm{V}}\left(\hat{t}^*_{y\pi}\right)$		$\widehat{\mathrm{V}}\left(\hat{t}^*_{yR1}\right)$		$\widehat{\mathrm{V}}\left(\hat{t}^*_{yR2}\right)^{\mathrm{alt}}$	
f_1 (%)	f_2 (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)
5	5	0,4	93,6	-0,6	93,8	-0,8	93,8
5	10	-0,1	94,3	-0,5	94,4	-0,2	94,8
5	25	-0,5	94,6	-0,5	94,7	0,7	94,8
5	50	1,0	95,0	0,2	94,8	1,2	95,1
10	5	1,0	94,5	1,6	94,7	1,7	94,7
10	10	-0,1	94,7	0,6	94,8	0,5	94,8
10	25	-2,0	94,8	-1,3	94,6	-0,6	94,7
10	50	0,9	95,0	0,2	95,0	0,8	94,9
20	5	1,0	94,8	-1,5	94,5	-1,0	94,5
20	10	0,4	94,9	0,0	94,8	0,2	95,0
20	25	0,2	94,9	0,4	94,9	1,5	95,1
20	50	-0,3	95,0	1,4	95,1	0,2	95,0

Tableau 4.4 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R(\,\cdot\,) + \widehat{\mathrm{V}}_2^R(\,\cdot\,)$ dans le cadre de l'étude par simulation avec un plan aléatoire simple sans remise aux deux phases pour la population avec $\rho = 0,7$.

		$\widehat{\mathrm{V}}\left(\hat{t}^*_{y\pi}\right)$		$\widehat{\mathrm{V}}\left(\hat{t}^*_{yR1}\right)$		$\widehat{\mathrm{V}}\left(\hat{t}^*_{yR2}\right)^{\mathrm{alt}}$	
f_1	f_2	$\mathrm{RB}_{\mathrm{MC}}$	$\mathrm{PC}_{\mathrm{MC}}$	$\mathrm{RB}_{\mathrm{MC}}$	$\mathrm{PC}_{\mathrm{MC}}$	$\mathrm{RB}_{\mathrm{MC}}$	$\mathrm{PC}_{\mathrm{MC}}$
(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
5	5	0,0	93,9	-8,3	91,6	-7,6	92,2
5	10	-1,2	94,5	-3,2	93,4	-2,8	94,1
5	25	0,4	95,1	-1,5	94,4	-0,8	94,5
5	50	-0,6	94,9	0,2	95,0	-0,4	94,8
10	5	-0,8	94,4	-3,5	93,5	-3,4	93,6
10	10	-1,7	95,0	0,8	94,5	0,9	94,6
10	25	-0,1	95,0	-0,1	94,8	-0,5	94,7
10	50	0,4	94,9	0,1	95,0	0,7	95,1
20	5	1,0	94,8	-0,9	94,3	-1,2	94,4
20	10	0,9	94,8	-1,8	94,7	-1,5	94,8
20	25	0,7	95,0	0,6	94,9	1,4	94,9
20	50	-0,1	95,1	1,3	94,7	-0,5	94,9

Tableau 4.5 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R(\cdot)+\widehat{\mathrm{V}}_2^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan Bernoulli à la deuxième phase pour la population avec $\rho = 0,7$.

f_1	m_g	$\mathrm{RB}_{\mathrm{MC}}\left(\widehat{\mathrm{V}}\left(\hat{t}^*_{y\pi}\right)\right)$	$\mathrm{CV}_{\mathrm{MC}}\left(\widehat{\mathrm{V}}\left(\hat{t}^*_{y\pi}\right)\right)$
(%)		(%)	(%)
2	2	0,0	91,5
2	5	0,0	91,7
2	10	-0,4	91,6
10	2	2,0	94,5
10	5	0,6	93,9
10	10	-0,2	93,4

Tableau 4.6 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R\left(\hat{t}^*_{y\pi}\right)+\widehat{\mathrm{V}}_2^R\left(\hat{t}^*_{y\pi}\right)$ dans le cadre de l'étude par simulation avec un plan à deux degrés pour la population avec ICC $= 5$ %.

Les tableaux 4.7, 4.8 et 4.9 présentent la performance des estimateurs de variance simplifiés pour les trois études par simulation. On remarque que dans la première étude, où un plan aléatoire simple sans remise a été utilisé aux deux phases, l'estimateur simplifié de la variance de l'estimateur par double dilatation présente un biais relatif Monte Carlo important (entre 1451 % et 3000 %) et que les pourcentages de couverture Monte Carlo associés sont tous de 100 %. Cela confirme que le terme $\widehat{\mathrm{V}}_2^R\left(\hat{t}_{y\pi}^*\right)$ n'est pas négligeable dans cette situation et que $\widehat{\mathrm{V}}_1^R\left(\hat{t}_{y\pi}^*\right)$ surestime considérablement la variance, menant à des intervalles de confiance trop larges.

En dehors de l'estimateur simplifié de la variance de l'estimateur par double dilatation dans la première étude, tous les autres estimateurs simplifiés de la variance se comportent raisonnablement. D'une part, ils présentent des valeurs raisonnables de biais relatif Monte Carlo. Par exemple, dans le tableau 4.8, lorsque $f_1 = 10$ % et $f_2 = 25$ %, on a $\mathrm{RB}_{\mathrm{MC}}\left[\widehat{\mathrm{V}}_1^R\left(\hat{t}_{y\pi}^*\right)\right] = -10,0$ %, $\mathrm{RB}_{\mathrm{MC}}\left[\widehat{\mathrm{V}}\left(\hat{t}_{yR1}^*\right)\right] = -7,8$ % et $\mathrm{RB}_{\mathrm{MC}}\left[\widehat{\mathrm{V}}\left(\hat{t}_{yR2}^*\right)\right] = -5,5$ %. D'ailleurs, on remarque que le biais Monte Carlo est moins important pour les petites valeurs de f_1. Par exemple, dans le tableau 4.9, lorsque $f_1 = 2$ % et $m_g = 2$ %, on a $\mathrm{RB}_{\mathrm{MC}}\left[\widehat{\mathrm{V}}\left(\hat{t}_{y\pi}^*\right)\right] = -1,2$ % et lorsque $f_1 = 10$ %, on a $\mathrm{RB}_{\mathrm{MC}}\left[\widehat{\mathrm{V}}\left(\hat{t}_{y\pi}^*\right)\right] = -4,3$ %. En général, celui-ci est également moins important lorsque la fraction de sondage à la deuxième phase augmente, pour une valeur de f_1 fixée. D'autre part, les pourcentages de couverture Monte Carlo se situent tous entre $90,9$ % et $94,9$ %. D'ailleurs, on remarque que les pourcentages de couverture Monte Carlo sont tous légèrement inférieurs à ceux leur correspondant dans l'un des tableaux 4.4, 4.5 et 4.6.

		$\widehat{\mathrm{V}}_1^R\left(\hat{t}_{y\pi}^*\right)$		$\widehat{\mathrm{V}}_1^R\left(\hat{t}_{yR1}^*\right)$		$\widehat{\mathrm{V}}_1^R\left(\hat{t}_{yR2}^*\right)$	
f_1 (%)	f_2 (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)
5	5	3000	100	-9,0	92,6	-8,3	92,7
5	10	2830	100	-6,7	93,6	-5,4	94,1
5	25	2372	100	-4,8	94,2	-2,2	94,5
5	50	1624	100	-2,6	94,5	-0,1	94,9
10	5	2859	100	-9,9	93,2	-8,7	93,4
10	10	2690	100	-9,4	93,5	-7,8	93,7
10	25	2232	100	-9,2	93,6	-5,8	94,1
10	50	1574	100	-5,2	94,3	-1,8	94,7
20	5	2541	100	-21,2	91,4	-19,1	91,8
20	10	2418	100	-18,7	92,1	-15,7	92,7
20	25	2073	100	-15,6	92,8	-9,5	93,8
20	50	1451	100	-9,9	93,7	-5,4	94,3

Tableau 4.7 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R\left(\cdot\right)$ dans le cadre de l'étude par simulation avec un plan aléatoire simple sans remise aux deux phases pour la population avec $\rho = 0,7$.

		$\widehat{\mathrm{V}}_1^R\left(\hat{t}_{y\pi}^*\right)$		$\widehat{\mathrm{V}}_1^R\left(\hat{t}_{yR1}^*\right)$		$\widehat{\mathrm{V}}_1^R\left(\hat{t}_{yR2}^*\right)$	
f_1 (%)	f_2 (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)
5	5	-4,9	93,4	-12,7	90,9	-11,5	91,6
5	10	-6,1	93,9	-7,6	92,8	-6,4	93,5
5	25	-4,6	94,5	-5,2	93,9	-3,2	94,3
5	50	-5,4	94,3	-2,4	94,7	-1,6	94,6
10	5	-10,7	93,0	-12,7	92,2	-11,7	92,4
10	10	-11,5	93,7	-8,4	93,3	-6,7	93,7
10	25	-10,0	93,7	-7,8	93,7	-5,5	94,2
10	50	-9,4	93,7	-5,1	94,3	-1,8	94,8
20	5	-19,1	91,7	-19,9	91,5	-18,6	91,9
20	10	-19,2	92,0	-19,8	91,6	-16,7	92,6
20	25	-19,3	92,3	-15,3	92,7	-9,4	93,6
20	50	-19,6	92,0	-10,0	93,5	-6,0	94,3

Tableau 4.8 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R\left(\cdot\right)$ dans le cadre de l'étude par simulation avec un plan Bernoulli à la deuxième phase pour la population avec $\rho = 0,7$.

f_1 (%)	m_g	$\mathrm{RB_{MC}}\left(\widehat{\mathrm{V}}^1\left(\hat{t}^*_{y\pi}\right)\right)$ (%)	$\mathrm{CV_{MC}}\left(\widehat{\mathrm{V}}^1\left(\hat{t}^*_{y\pi}\right)\right)$ (%)
2	2	-1,2	91,3
2	5	-0,6	91,5
2	10	-0,7	91,5
10	2	-4,3	93,6
10	5	-3,1	93,3
10	10	-2,0	93,2

Tableau 4.9 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R\left(\hat{t}^*_{y\pi}\right)$ dans le cadre de l'étude par simulation avec un plan à deux degrés pour la population avec ICC $= 5$ %.

Les tableaux 4.10, 4.11 et 4.12 présentent l'espérance Monte Carlo des mesures de contribution de $\widehat{\mathrm{V}}_2^R(\,\cdot\,)$ à l'estimateur de variance dans chacune des études par simulation. Dans la première étude, il n'est pas surprenant de constater que l'espérance Monte Carlo des mesures de contribution $C_2(\hat{t}_{y\pi}^*)$ et $C_2^R(\hat{t}_{y\pi}^*)$ est importante (respectivement, entre -3301 % et -1566 % et entre -3466 % et -2625 %). Toutefois, en ce qui a trait aux estimateurs par le ratio, les mesures de contribution $\mathrm{E_{MC}}\left[C_2^R(\,\cdot\,)\right]$ concordent avec les valeurs exactes figurant dans le tableau 3.1. On observe d'ailleurs $\mathrm{E_{MC}}\left[C_2^R(\,\cdot\,)\right] \approx f_1$ (voir résultat (3.20)). Dans cette première étude, on observe également pour tous les estimateurs de variance que pour f_1 fixé, $\left|\mathrm{E_{MC}}\left[C_2^R(\cdot)\right]\right|$ et $\left|\mathrm{E_{MC}}\left[C_2(\cdot)\right]\right|$ diminuent lorsque f_2 augmente.

Dans la seconde étude, on observe $\mathrm{E_{MC}}\left[C_2^R(\cdot)\right] = f_1$ pour les trois estimateurs de variance, ce qui concorde avec la remarque 3.1.

Dans la troisième étude, on observe $\mathrm{E_{MC}}\left[C_2^R(\cdot)\right] \approx f_1$.

À l'exception du cas de l'estimateur $\hat{t}_{y\pi}^*$ dans l'étude 1, on observe dans tous les cas que $\mathrm{E_{MC}}\left[C_2(\,\cdot\,)\right]$ est inférieur à $\mathrm{E_{MC}}\left[C_2^R(\,\cdot\,)\right]$ (voir expression (3.6)) et que la différence est d'autant plus importante que la fraction de sondage à la deuxième phase est grande pour un f_1 fixé. Ceci est particulièrement notable pour les estimateurs de calage. Par exemple, dans le tableau 4.11 pour l'estimateur $\widehat{\mathrm{V}}(\hat{t}_{yR2}^*)$, lorsque $f_1 = 10$ % et $f_2 = 5$ %, on observe $\mathrm{E_{MC}}\left[C_2^R\left(\widehat{\mathrm{V}}(\hat{t}_{yR2}^*)\right)\right] = 8{,}6$ % et $\mathrm{E_{MC}}\left[C_2\left(\widehat{\mathrm{V}}(\hat{t}_{yR2}^*)\right)\right] = 10$ % alors que lorsque $f_2 = 50$ % on observe $\mathrm{E_{MC}}\left[C_2^R\left(\widehat{\mathrm{V}}(\hat{t}_{yR2}^*)\right)\right] = 2{,}5$ % et $\mathrm{E_{MC}}\left[C_2\left(\widehat{\mathrm{V}}(\hat{t}_{yR2}^*)\right)\right] = 10$ %.

En comparant les résultats obtenus pour la population avec $\rho = 0{,}7$ à ceux des populations avec $\rho = 0{,}5$ et $\rho = 0{,}9$ (voir annexe II.1 et II.2), on note que, dans l'étude 1, l'estimateur simplifié de la variance de l'estimateur par double dilatation n'est valides dans aucune des populations. De plus, les valeurs de $\mathrm{E_{MC}}\left(C_2^R(\hat{t}_{yR1}^*)\right)$, $\mathrm{E_{MC}}\left(C_2(\hat{t}_{yR1}^*)\right)$ et $\mathrm{E_{MC}}\left(C_2^R(\hat{t}_{yR2}^*)\right)$ sont identiques dans les trois populations, mais celles de $\mathrm{E_{MC}}\left(C_2(\hat{t}_{yR2}^*)\right)$ sont plus petites lorsque ρ est grand. En ce qui a trait à l'étude par simulation 2, les va-

leurs de $\mathrm{E_{MC}}\left(C_2^R(\,\cdot\,)\right)$ et $\mathrm{E_{MC}}\left(C_2\left(\hat{t}_{yR1}^*\right)\right)$ sont identiques dans les trois populations, les valeurs de $\mathrm{E_{MC}}\left[C_2(\hat{t}_{y\pi}^*)\right]$ sont très similaires et les valeurs de $\mathrm{E_{MC}}\left(C_2\left(\hat{t}_{yR2}^*\right)\right)$ sont plus petites lorsque ρ est grand. En comparant les résultats obtenus pour la population avec ICC $= 5$ % à ceux de la population avec ICC $= 20$ % (voir annexe II.3), on note que les valeurs de $\mathrm{E_{MC}}\left[C_2^R\left(\hat{t}_{y\pi}^*\right)\right]$ sont très similaires dans les deux populations et que les valeurs de $\mathrm{E_{MC}}\left[C_2\left(\hat{t}_{y\pi}^*\right)\right]$ sont plus petites dans la population avec le plus grand ICC.

Enfin, les études par simulation confirment parfaitement la théorie présentée au chapitre 3.

		$\hat{t}_{y\pi}^*$		$\hat{t}_{yR1}^*$		$\hat{t}_{yR2}^*$	
f_1 (%)	f_2 (%)	$\mathrm{E_{MC}}\left(C_2\right)$ (%)	$\mathrm{E_{MC}}\left(C_2^R\right)$ (%)	$\mathrm{E_{MC}}\left(C_2\right)$ (%)	$\mathrm{E_{MC}}\left(C_2^R\right)$ (%)	$\mathrm{E_{MC}}\left(C_2\right)$ (%)	$\mathrm{E_{MC}}\left(C_2^R\right)$ (%)
5	5	-3301	-3466	8,4	8,8	7,5	8,8
5	10	-2971	-3285	6,2	6,9	5,1	6,9
5	25	-2429	-3198	4,4	5,8	2,8	5,8
5	50	-1621	-3162	2,8	5,4	1,3	5,4
10	5	-2969	-3109	11,3	11,8	10,2	11,8
10	10	-2756	-3032	9,9	10,9	8,2	10,9
10	25	-2301	-2992	8,0	10,4	5,2	10,4
10	50	-1566	-2976	5,4	10,2	2,6	10,2
20	5	-2576	-2685	20,0	20,8	18,3	20,8
20	10	-2436	-2652	18,7	20,4	15,8	20,4
20	25	-2079	-2633	15,9	20,2	10,8	20,2
20	50	-1459	-2625	11,2	20,1	5,6	20,1

Tableau 4.10 – Contribution de $\widehat{\mathrm{V}}_2^R(\,\cdot\,)$ dans le cadre de l'étude par simulation avec un plan aléatoire simple sans remise aux deux phases pour la population avec $\rho = 0,7$.

		$\hat{t}^*_{y\pi}$		$\hat{t}^*_{yR1}$		$\hat{t}^*_{yR2}$	
f_1 (%)	f_2 (%)	$\mathrm{E_{MC}}(C_2)$ (%)	$\mathrm{E_{MC}}(C_2^R)$ (%)	$\mathrm{E_{MC}}(C_2)$ (%)	$\mathrm{E_{MC}}(C_2^R)$ (%)	$\mathrm{E_{MC}}(C_2)$ (%)	$\mathrm{E_{MC}}(C_2^R)$ (%)
5	5	5,0	5	4,8	5	4,2	5
5	10	5,0	5	4,5	5	3,7	5
5	25	4,9	5	3,8	5	2,4	5
5	50	4,9	5	2,6	5	1,2	5
10	5	10,0	10	9,5	10	8,6	10
10	10	10,0	10	9,1	10	7,5	10
10	25	9,9	10	7,7	10	5,0	10
10	50	9,7	10	5,3	10	2,5	10
20	5	20,0	20	19,2	20	17,5	20
20	10	19,9	20	18,4	20	15,5	20
20	25	19,8	20	15,8	20	10,7	20
20	50	19,5	20	11,1	20	5,5	20

Tableau 4.11 – Contribution de $\widehat{\mathrm{V}}_2^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan Bernoulli à la deuxième phase pour la population avec $\rho = 0,7$.

f_1 (%)	m_g	$\mathrm{E_{MC}}(C_2)$ (%)	$\mathrm{E_{MC}}(C_2^R)$ (%)
2	2	1,6	2,0
2	5	0,9	2,0
2	10	0,4	2,0
10	2	6,5	10,1
10	5	4,0	10,1
10	10	1,9	10,2

Tableau 4.12 – Contribution de $\widehat{\mathrm{V}}_2^R\left(\hat{t}^*_{y\pi}\right)$ dans le cadre de l'étude par simulation avec un plan à deux degrés pour la population avec ICC $= 5$ %.

CHAPITRE 5

CONCLUSION

Dans ce mémoire, nous avons proposé des estimateurs simplifiés de la variance des estimateurs par double dilatation et de calage utilisés dans le cas des plans de sondages à deux phases. Ces estimateurs simplifiés présentent l'avantage de pouvoir être calculés à partir de logiciels pour les plans à une phase.

Nous avons étudié les conditions de validité de ces estimateurs simplifiés de la variance. Nous avons conclu que lorsqu'un plan de Poisson est utilisé à la deuxième phase, les estimateurs simplifiés sont valides lorsque la fraction de sondage à la première phase est négligeable. Nous avons conclu, dans le cas d'un plan à deux degrés, que les estimateurs simplifiés sont valides lorsque la fraction de sondage au premier degré est négligeable. Ce résultat est en accord avec celui de Singh (2008) qui propose d'ailleurs un plan de sondage hybride permettant de toujours satisfaire la propriété d'indépendance. Finalement, nous avons conclu, lorsqu'un plan aléatoire simple sans remise est utilisé à la deuxième phase (ou aux deux phases), que l'estimateur simplifié de la variance de l'estimateur par double dilatation n'est généralement pas valide mais que l'estimateur simplifié de la variance de certains estimateurs de calage, comme les estimateurs par le ratio ou l'estimateur de Hájek, est valide lorsque la taille échantillonnale est suffisamment grande et que la fraction de sondage à la première phase est négligeable.

De plus, nous avons montré, en supposant la propriété d'invariance des plans à deux phases, que les estimateurs simplifiés peuvent être obtenus au moyen de l'approche renversée proposée par Fay (1991) et Shao et Steel (1999).

Enfin, les études par simulation que nous avons effectuées ont corroboré les résultats théoriques.

BIBLIOGRAPHIE

[1] Axelson, M. (1998). Variance Estimation for the Generalised Regression Estimator Under Two-Phase Sampling - A Modified Approach. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 85-89.

[2] Demnati A. et Rao J.N.K. (2004). Linearization Variance Estimators for Survey Data. *Survey Methodology*, 30, 17-26.

[3] Deville, J.C. et Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, 376-382.

[4] Dupont, F. (1995). Alternative Adjustments Where There Are Several Levels of Auxiliary Information. *Survey Methodology*, 21, 125-135.

[5] Esteavo V.M. et Särndal C.-E. (2006). Survey Estimates by Calibration on Complex Auxiliary Information. *International Statistical Review*, 74, 127-147.

[6] Fay, R. E. (1991). A Design-Based Perspective on Missing Data Variance. *Proceeding of the 1991 Annual Research Conference*, US Bureau of the Census, 429-440.

[7] Haziza, D. et Beaumont, J-F. (2005). Estimation simplifiée de la variance dans le cas de l'échantillonnage à deux phases. *Méthodes d'enquêtes et sondages*, sous la direction de Lavallée, P. et Rivest, L.-P., Dunod, 372-377.

[8] Hidiroglou M.A. et Särndal C.-E. (1998). Use of Auxiliary Information for Two-Phase Sampling. *Survey Methodology*, 24, 11-20.

[9] Hidiroglou, M.A., Rao, J.N.K. et Haziza, D. (2009). Variance Estimation in Two Phase Sampling. *Australian and New Zealand Journal of Statistics*, 51, 127-141.

[10] Horvitz D.G. et Thompson D.J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47, 663-685.

[11] Isaki, C.T. et Fuller, W.A. (1982). Survey Design Under the Regression Superpopulation Model. *Journal of the American Statistical Association*, 77, 89-96.

[12] Kim, J.K., Navarro, A. et Fuller W.A. (2006). Replication Variance Estimation for Two-Phase Stratified Sampling. *Journal of the American Statistical Association*, 101, 312-320.

[13] Rao, J.N.K. (1965). On two Simple Schemes of Unequal Probability Sampling Without Replacement. *Journal of Indian Statistical Association*, 3, 173-180.

[14] Sampford, M.R. (1967). On Sampling Without Replacement with Unequal Probabilities of Selection. *Biometrika*, 54, 499-513.

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

[16] Shao, J. and Steel, P. (1999). Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions. *Journal of the American Statistical Association*, 94, 254-265.

[17] Singh, A.C. (2008). Single Phase Simplified Variance Estimation Approach to Two Phase - Stage Hybrid Designs. *2008 JSM Proceedings of Section on Survey Research Methods*.

[18] Woodruff, R. (1971). A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of the American Statistical Association*, 66, 411-414.

Annexe I

Compléments théoriques

I.1 Preuve de la remarque 2.4

Nous montrons dans un premier temps que, lorsque $q_i^{-1} = \boldsymbol{\alpha}'\boldsymbol{x}_i$, on a $\sum_{i\in s} d_i e_i = 0$. D'abord notons que $1 = q_i\boldsymbol{\alpha}'\boldsymbol{x}_i$. On a

$$\begin{aligned}
\sum_{i\in s} d_i e_i &= \sum_{i\in s} d_i\left(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{B}}\right) \\
&= \sum_{i\in s} d_i q_i \boldsymbol{\alpha}'\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{B}}\right) \\
&= \boldsymbol{\alpha}'\left[\sum_{i\in s} d_i q_i \boldsymbol{x}_i y_i - \sum_{i\in s} d_i q_i \boldsymbol{x}_i\boldsymbol{x}_i'\hat{\boldsymbol{B}}\right] \\
&= \boldsymbol{\alpha}'\left[\sum_{i\in s} d_i q_i \boldsymbol{x}_i y_i - \sum_{i\in s} d_i q_i \boldsymbol{x}_i y_i\right] \\
&= 0.
\end{aligned}$$

De façon analogue, on peut montrer que $\sum_{i\in U} E_i = 0$ en remplaçant s par U, $\hat{\boldsymbol{B}}$ par $\boldsymbol{B}$, e_i par E_i et d_i par 1 dans la démarche précédente.

I.2 Cadre asymptotique

Le cadre asymptotique fréquemment utilisé en échantillonnage pour étudier les propriétés des estimateurs prend en compte le fait que la taille de la population N est fixe. On considère donc une suite de populations finies $\{U_\nu\}_{\nu=1}^{\infty}$ telles que $U_1 \subseteq U_2 \subseteq \ldots$ dont les tailles sont croissantes : $N_\nu > N_{\nu-1}$. On considère également la suite d'échantillons $\{s_\nu \subseteq U_\nu\}_{\nu=1}^{\infty}$ tirés par un plan de sondage $p_\nu(s_\nu)$. Les tailles d'échantillon sont aussi croissantes : $n_\nu > n_{\nu-1}$. Sous ce cadre asymptotique, il est sous-entendu que $N_\nu \to \infty$ et $n_\nu \to \infty$ lorsque $\nu \to \infty$.

Définition 1. Soit $\{X_\nu\}_{\nu=1}^{\infty}$ une suite de variables aléatoires et $\{h_\nu\}_{\nu=1}^{\infty}$ une suite de

nombre réels tels que $h_\nu > 0$ pour $h_\nu = 1, \ldots, \infty$. On écrit

$$X_\nu = O_p\big(h_\nu\big), \nu \to \infty \tag{I.1}$$

si et seulement si pour tout nombre réel $\varepsilon > 0$, il existe un nombre réel $M_\varepsilon > 0$ et un entier naturel ν_0 tel que $\mathrm{P}(|X_\nu| \geq M_\varepsilon h_\nu) \leq \varepsilon$ pour tout $\nu > \nu_0$.

Définition 2. Soit $\{x_\nu\}_{\nu=1}^{\infty}$ et $\{h_\nu\}_{\nu=1}^{\infty}$ des suites de nombres réels tels que $h_\nu > 0$ pour $\nu = 1, \ldots, \infty$.

On écrit

$$x_\nu = O\big(h_\nu\big), \nu \to \infty \tag{I.2}$$

si et seulement si il existe un nombre réel $M > 0$ et un entier naturel ν_0 tel que $|x_\nu| \leq M h_\nu$ pour tout $\nu > \nu_0$.

I.3 Calcul des quantités $\widehat{\mathrm{V}}_2^R\big(\hat{t}_{y\pi}^*\big)$, $\widehat{\mathrm{V}}_2\big(\hat{t}_{y\pi}^*\big)$ et $\widehat{\mathrm{V}}\big(\hat{t}_{y\pi}\big)$ dans le cas d'un plan aléatoire simple sans remise aux deux phases

On a

$$\begin{aligned}
\widehat{\mathrm{V}}_2^R\big(\hat{t}_{y\pi}^*\big) &= \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{2ij}}{\pi_{1ij}}y_i y_j \\
&= \sum_{i\in s_2}\sum_{\substack{j\in s_2\\ j\neq i}}\frac{\Delta_{2ij}}{\pi_{1ij}}y_i y_j + \sum_{i\in s_2}\frac{\Delta_{2ii}}{\pi_{1i}}y_i^2 \\
&= \frac{N(N-1)}{n_1(n_1-1)}\sum_{i\in s_2}\sum_{\substack{j\in s_2\\ j\neq i}}\Delta_{2ij}y_i y_j + \frac{N}{n_1}\sum_{i\in s_2}\Delta_{2ii}y_i^2 \\
&= \frac{N(N-1)}{n_1(n_1-1)}\left[\sum_{i\in s_2}\sum_{j\in s_2}\Delta_{2ij}y_i y_j - \sum_{i\in s_2}\Delta_{2ii}y_i^2\right] + \frac{N}{n_1}\sum_{i\in s_2}\Delta_{2ii}y_i^2 \\
&= \frac{N(N-1)}{n_1(n_1-1)}\sum_{i\in s_2}\sum_{j\in s_2}\Delta_{2ij}y_i y_j - \frac{N}{n_1}\left(\frac{N-1}{n_1-1}-1\right)\sum_{i\in s_2}\Delta_{2ii}y_i^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{N(N-1)}{n_1(n_1-1)} n_1^2 \left(1-\frac{n_2}{n_1}\right)\frac{s_2^2}{n_2} - \frac{N}{n_1}\left(\frac{N-n_1}{n_1-1}\right)\sum_{i\in s_2}\left(1-\frac{n_2}{n_1}\right)\left(\frac{n_1}{n_2}\right)^2 y_i^2 \\
&= \frac{N(N-1)}{(n_1-1)} n_1 \left(1-\frac{n_2}{n_1}\right)\frac{s_2^2}{n_2} - \frac{Nn_1}{n_2^2}\left(\frac{N-n_1}{n_1-1}\right)\left(1-\frac{n_2}{n_1}\right)\sum_{i\in s_2} y_i^2 \\
&= \frac{N(N-1)}{(n_1-1)} n_1 \left(1-\frac{n_2}{n_1}\right)\frac{s_2^2}{n_2} - \frac{Nn_1}{n_2^2}\left(\frac{N-n_1}{n_1-1}\right)\left(1-\frac{n_2}{n_1}\right)\left[(n_2-1)s_2^2 + n_2\bar{y}_2^2\right] \\
&= \frac{Nn_1}{(n_1-1)n_2}\left(1-\frac{n_2}{n_1}\right)s_2^2\left[(N-1) - \frac{(N-n_1)(n_2-1)}{n_2}\right] \\
&- \frac{Nn_1}{n_2}\left(\frac{N-n_1}{n_1-1}\right)\left(1-\frac{n_2}{n_1}\right)\bar{y}_2^2 \\
&= \frac{Nn_1}{(n_1-1)n_2^2}\left(1-\frac{n_2}{n_1}\right)s_2^2\left(N-n_2+n_1n_2-n_1\right) - \frac{Nn_1}{n_2}\left(\frac{N-n_1}{n_1-1}\right)\left(1-\frac{n_2}{n_1}\right)\bar{y}_2^2 \\
&= \frac{N^2n_1}{(n_1-1)n_2}\left(1-\frac{n_2}{n_1}\right)\left(\frac{1}{Nn_2}\left\{N-n_2+n_1n_2-n_1\right\}s_{2y}^2 + \frac{1}{N}\{N-n_1\}\bar{y}_2^2\right) \\
&= \frac{N^2n_1}{(n_1-1)n_2}\left(1-\frac{n_2}{n_1}\right)\left(N^{-1}\left\{\frac{N}{n_2}-1+n_1-\frac{n_1}{n_2}\right\}s_{2y}^2 + \left\{1-\frac{n_1}{N}\right\}\bar{y}_2^2\right) \\
&= Nf_1^{-1}f_2^{-1}(1-f_2)\left(1-N^{-1}f_1^{-1}\right)^{-1} \\
&\times \left\{\left[f_1 + N^{-1}\left(f_1^{-1}f_2^{-1} - f_2^{-1} - 1\right)\right]s_{2y}^2 - (1-f_1)\bar{y}_2^2\right\},
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\mathrm{V}}_2\left(\hat{t}_{y\pi}^*\right) &= \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{2ij}\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1j}} \\
&= \left(\frac{N}{n_1}\right)^2\sum_{i\in s_2}\sum_{j\in s_2}\Delta_{2ij}y_iy_j \\
&= \left(\frac{N}{n_1}\right)^2 n_1^2\left(1-\frac{n_2}{n_1}\right)\frac{s_{2y}^2}{n_2} \\
&= N^2\left(1-\frac{n_2}{n_1}\right)\frac{s_{2y}^2}{n_2} \\
&= Nf_1^{-1}f_2^{-1}(1-f_2)s_{2y}^2
\end{aligned}
$$

et

$$\begin{aligned}\widehat{\mathrm{V}}\left(\hat{t}^{*}_{y\pi}\right) &= N^2\left(1-\frac{n_2}{N}\right)\frac{s^2_{2y}}{n_2}\\ &= Nf_1^{-1}f_2^{-1}(1-f_1f_2)s^2_{2y}\end{aligned}$$

puisque le tirage d'un échantillon aléatoire simple sans remise de n_1 parmi N unités suivit d'un échantillon aléatoire simple sans remise de n_2 parmi les n_1 unités résultant du premier tirage est équivalent à un seul tirage aléatoire simple sans remise de n_2 parmi N unités.

I.4 Calcul de l'espérance de $\widehat{\mathrm{V}}_2^R\left(\hat{t}^{*}_{y\pi}\right)$ et $\widehat{\mathrm{V}}_2\left(\hat{t}^{*}_{y\pi}\right)$ dans le cas d'un plan aléatoire simple sans remise à la première phase

$$\begin{aligned}\mathrm{E}_1\,\mathrm{E}_2\left[\widehat{\mathrm{V}}_2^R\left(\hat{t}^{*}_{y\pi}\right)|\boldsymbol{I}_1\right] &\approx \sum_{i\in U}\sum_{j\in U}\Delta_{2ij}\pi_{2ij}y_iy_j\\ &= -\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\frac{1}{n_1-1}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}y_iy_j+\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\sum_{i\in U}y_i^2\\ &= -\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\frac{1}{n_1-1}\sum_{i\in U}\sum_{j\in U}y_iy_j+\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\frac{n_1}{n_1-1}\sum_{i\in U}y_i^2\\ &= -\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\frac{1}{n_1-1}\left(\sum_{i\in U}y_i\right)^2+\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\frac{n_1}{n_1-1}\sum_{i\in U}y_i^2\end{aligned}$$

$$\begin{aligned}\mathrm{E}_1\,\mathrm{E}_2\left[\widehat{\mathrm{V}}_2\left(\hat{t}^{*}_{y\pi}\right)|\boldsymbol{I}_1\right] &\approx \sum_{i\in U}\sum_{j\in U}\Delta_{2ij}\pi_{2ij}\pi_{1ij}\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1j}}\\ &= -\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\frac{1}{n_1-1}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}y_iy_j+\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\sum_{i\in U}\frac{y_i^2}{\pi_{1i}}\end{aligned}$$

$$
\begin{aligned}
&= -\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\frac{1}{n_1-1}\sum_{i\in U}\sum_{j\in U}\left\{\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}-1+1\right\}y_i y_j \\
&+\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\frac{1}{n_1-1}\sum_{i\in U}\frac{y_i^2}{\pi_{1i}}+\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\sum_{i\in U}\frac{y_i^2}{\pi_{1i}} \\
&= -\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\frac{1}{n_1-1}\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}y_i y_j \\
&-\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\frac{1}{n_1-1}\left(\sum_{i\in U}y_i\right)^2+\frac{n_1}{n_2}\left(1-\frac{n_2}{n_1}\right)\frac{n_1}{n_1-1}\sum_{i\in U}\frac{y_i^2}{\pi_{1i}}
\end{aligned}
$$

I.5 Justification de conditions de régularité

On a $\sum_{i\in U} y_i = O(N)$ puisque

$$
\begin{aligned}
\sum_{i\in U} y_i &\leq \sum_{i\in U}\max(y_i) \\
&= N\max(y_i).
\end{aligned}
$$

On a $\sum_{i\in U} y_i^2 = O(N)$ puisque

$$
\begin{aligned}
\sum_{i\in U} y_i^2 &\leq \sum_{i\in U}\max(y_i^2) \\
&= N\max(y_i^2).
\end{aligned}
$$

On a $\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}y_i y_j = O(N^2/n_1)$ puisque

$$
\begin{aligned}
\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1ij}y_i y_j &= \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\left\{\pi_{1ij}-\pi_{1i}\pi_{1j}\right\}\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1i}}+\sum_{i\in U}\left(\frac{1}{\pi_{1i}}-1\right)y_i^2 \\
&\leq \left|\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\left\{\pi_{1ij}-\pi_{1i}\pi_{1j}\right\}\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1i}}\right|+\sum_{i\in U}\left\{\max(d_{1i})-1\right\}y_i^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \left|\pi_{1ij}-\pi_{1i}\pi_{1j}\right| \frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1i}} + \{\max(d_{1i})-1\}\sum_{i\in U} y_i^2 \\
&\leq \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \max\left|\pi_{1ij}-\pi_{1i}\pi_{1j}\right| \max(d_{1i})\max(d_{1j}) y_i y_j \\
&+ \{\max(d_{1i})-1\}\sum_{i\in U} y_i^2 \\
&= \max\left|\pi_{1ij}-\pi_{1i}\pi_{1j}\right| \max(d_{1i})\max(d_{1j}) \left\{ \left(\sum_{i\in U} y_i\right)^2 - \sum_{i\in U} y_i^2 \right\} \\
&+ \{\max(d_{1i})-1\}\sum_{i\in U} y_i^2,
\end{aligned}
$$

en supposant que $\max(d_{1i}) = O(N/n_1)$ et $\max\left|\pi_{1ij}-\pi_{1i}\pi_{1j}\right| = O(n_1/N^2)$. Ces conditions sont respectées, par exemple, dans le cas d'un plan aléatoire simple sans remise à la première phase, car on a

$$
\begin{aligned}
\max(d_{1i}) &= N/n_1 \\
&= O(N/n_1)
\end{aligned}
$$

et

$$
\begin{aligned}
\max\left|\pi_{1ij}-\pi_{1i}\pi_{1j}\right| &= \left|\frac{n_1}{N}\frac{n_1-1}{N-1} - \frac{n_1}{N}\frac{n_1}{N}\right| \\
&= \frac{n_1}{N}\left|\frac{n_1-1}{N-1} - \frac{n_1}{N}\right| \\
&= \frac{n_1}{N}\left|\frac{N-n_1}{N(N-1)}\right| \\
&= \frac{n_1}{N}\frac{1}{N-1}\left(1-\frac{n_1}{N}\right) \\
&\leq \frac{n_1}{N}\frac{1}{N-1} \\
&\approx O(n_1/N^2).
\end{aligned}
$$

On a $\sum_{i\in U} \frac{y_i^2}{\pi_{1i}} = O(N^2/n_1)$ puisque

$$\begin{aligned}\sum_{i\in U} \frac{y_i^2}{\pi_{1i}} &\leq \sum_{i\in U} \max(y_i^2)\max(d_{1i}) \\ &= N\max(y_i^2)\max(d_{1i}),\end{aligned}$$

en supposant que $\max(d_{1i}) = O(N/n_1)$.

Annexe II

Résultats de simulation supplémentaires

II.1 Étude 1 : plan aléatoire simple sans remise aux deux phases

II.1.1 Résultats pour la population avec $\rho = 0,5$

		$\hat{t}^*_{y\pi}$		$\hat{t}^*_{yR1}$		$\hat{t}^*_{yR2}$	
f_1 (%)	f_2 (%)	RB_{MC} (%)	CV_{MC} (%)	RB_{MC} (%)	CV_{MC} (%)	RB_{MC} (%)	CV_{MC} (%)
5	5	0,0	4,2	0,0	2,9	0,0	3,0
5	10	0,0	2,9	0,0	2,1	0,0	2,2
5	25	0,0	1,9	0,0	1,3	0,0	1,5
5	50	0,0	1,3	0,0	0,9	0,0	1,1
10	5	0,0	2,9	0,0	2,1	0,0	2,1
10	10	0,0	2,1	0,0	1,5	0,0	1,5
10	25	0,0	1,3	0,0	0,9	0,0	1,0
10	50	0,0	0,9	0,0	0,6	0,0	0,8
20	5	0,0	2,1	0,0	1,5	0,0	1,5
20	10	0,0	1,5	0,0	1,0	0,0	1,1
20	25	0,0	0,9	0,0	0,6	0,0	0,7
20	50	0,0	0,6	0,0	0,4	0,0	0,5

Tableau II.1 – Performance des estimateurs ponctuels dans le cadre de l'étude par simulation avec un plan aléatoire simple sans remise aux deux phases pour la population avec $\rho = 0,5$.

		$\widehat{\mathrm{V}}\left(\hat{t}^{*}_{y\pi}\right)$		$\widehat{\mathrm{V}}\left(\hat{t}^{*}_{yR1}\right)$		$\widehat{\mathrm{V}}\left(\hat{t}^{*}_{yR2}\right)^{\mathrm{alt}}$	
f_1 (%)	f_2 (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)
5	5	0,4	93,8	0,8	94,0	1,1	94,1
5	10	0,0	94,4	1,0	94,5	0,3	94,5
5	25	-0,2	94,8	-2,0	94,4	-2,2	94,5
5	50	1,1	94,7	0,6	94,8	0,4	95,0
10	5	0,8	94,3	-0,3	94,3	-0,1	94,5
10	10	0,3	94,9	0,2	94,7	0,4	94,8
10	25	-0,8	94,8	-0,1	94,7	-0,6	94,8
10	50	0,3	94,9	1,6	95,1	0,1	94,9
20	5	0,5	94,9	1,5	94,8	1,3	94,9
20	10	1,1	95,0	-0,9	94,8	-0,3	94,9
20	25	0,3	95,0	0,6	95,1	0,2	95,2
20	50	2,1	95,2	-0,1	94,9	0,8	95,2

Tableau II.2 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R(\cdot)+\widehat{\mathrm{V}}_2^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan aléatoire simple sans remise aux deux phases pour la population avec $\rho = 0,5$.

		$\widehat{\mathrm{V}}_1^R\left(\hat{t}^{*}_{y\pi}\right)$		$\widehat{\mathrm{V}}_1^R\left(\hat{t}^{*}_{yR1}\right)$		$\widehat{\mathrm{V}}_1^R\left(\hat{t}^{*}_{yR2}\right)$	
f_1 (%)	f_2 (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)
5	5	2085	100	-7,7	93,0	-7,0	93,1
5	10	1978	100	-5,3	93,7	-5,4	93,8
5	25	1654	100	-6,2	93,9	-5,7	94,1
5	50	1133	100	-2,1	94,5	-1,4	94,8
10	5	1981	100	-11,6	92,9	-10,9	93,1
10	10	1881	100	-9,7	93,5	-8,7	93,8
10	25	1573	100	-8,1	93,7	-7,0	93,9
10	50	1087	100	-3,9	94,4	-3,5	94,5
20	5	1752	100	-18,8	92,0	-18,1	92,0
20	10	1693	100	-19,4	92,1	-17,6	92,4
20	25	1441	100	-15,4	92,9	-12,9	93,5
20	50	1034	100	-11,2	93,5	-7,0	94,3

Tableau II.3 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan aléatoire simple sans remise aux deux phases pour la population avec $\rho = 0,5$.

		$\hat{t}^*_{y\pi}$		$\hat{t}^*_{yR1}$		$\hat{t}^*_{yR2}$	
f_1 $(\%)$	f_2 $(\%)$	$\mathrm{E_{MC}}\left(C_2\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2^R\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2^R\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2^R\right)$ $(\%)$
5	5	-2284	-2398	8,4	8,8	8,0	8,8
5	10	-2069	-2288	6,2	6,9	5,7	6,9
5	25	-1688	-2222	4,4	5,8	3,5	5,8
5	50	-1129	-2202	2,8	5,4	1,9	5,4
10	5	-2055	-2152	11,3	11,8	10,8	11,8
10	10	-1917	-2108	9,9	10,9	9,1	10,9
10	25	-1600	-2081	8,0	10,4	6,5	10,4
10	50	-1088	-2068	5,4	10,2	3,6	10,2
20	5	-1783	-1858	20,0	20,8	19,2	20,8
20	10	-1691	-1841	18,7	20,4	17,3	20,4
20	25	-1442	-1827	15,9	20,2	13,1	20,2
20	50	-1013	-1823	11,2	20,1	7,7	20,1

Tableau II.4 – Contribution de $\widehat{\mathrm{V}}_2^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan aléatoire simple sans remise aux deux phases pour la population avec $\rho = 0,5$.

II.1.2 Résultats pour la population avec $\rho = 0,9$

		$\hat{t}^*_{y\pi}$		$\hat{t}^*_{yR1}$		$\hat{t}^*_{yR2}$	
f_1 (%)	f_2 (%)	RB_{MC} (%)	CV_{MC} (%)	RB_{MC} (%)	CV_{MC} (%)	RB_{MC} (%)	CV_{MC} (%)
5	5	0,0	3,0	0,0	1,0	0,0	1,2
5	10	0,0	2,1	0,0	0,7	0,0	0,9
5	25	0,0	1,3	0,0	0,4	0,0	0,8
5	50	0,0	0,9	0,0	0,3	0,0	0,7
10	5	0,0	2,1	0,0	0,7	0,0	0,8
10	10	0,0	1,5	0,0	0,5	0,0	0,6
10	25	0,0	1,0	0,0	0,3	0,0	0,5
10	50	0,0	0,7	0,0	0,2	0,0	0,5
20	5	0,0	1,5	0,0	0,5	0,0	0,6
20	10	0,0	1,1	0,0	0,3	0,0	0,4
20	25	0,0	0,7	0,0	0,2	0,0	0,4
20	50	0,0	0,5	0,0	0,1	0,0	0,3

Tableau II.5 – Performance des estimateurs ponctuels dans le cadre de l'étude par simulation avec un plan aléatoire simple sans remise aux deux phases pour la population avec $\rho = 0,9$.

		$\widehat{\mathrm{V}}\left(\hat{t}^*_{y\pi}\right)$		$\widehat{\mathrm{V}}\left(\hat{t}^*_{yR1}\right)$		$\widehat{\mathrm{V}}\left(\hat{t}^*_{yR2}\right)^{\mathrm{alt}}$	
f_1 (%)	f_2 (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)
5	5	-0,8	93,5	0,2	93,8	-0,4	94,4
5	10	0,7	94,4	0,9	94,5	0,7	94,7
5	25	0,2	94,7	0,7	94,8	-0,2	94,9
5	50	-0,1	94,8	-0,9	94,5	-0,3	94,7
10	5	0,3	94,2	0,5	94,5	1,6	94,9
10	10	0,7	94,8	1,3	94,9	0,6	95,0
10	25	-0,6	94,8	-2,0	94,7	-0,4	94,9
10	50	0,2	95,0	-0,8	94,6	-0,2	94,8
20	5	-0,7	94,6	0,8	94,8	0,1	94,8
20	10	-0,7	94,7	-0,3	94,9	-0,4	94,9
20	25	0,5	95,1	0,1	94,9	0,1	95,0
20	50	1,2	95,2	0,7	94,8	-0,8	94,9

Tableau II.6 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R(\cdot)+\widehat{\mathrm{V}}_2^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan aléatoire simple sans remise aux deux phases pour la population avec $\rho = 0,9$.

		$\widehat{\mathrm{V}}_1^R\left(\hat{t}^*_{y\pi}\right)$		$\widehat{\mathrm{V}}_1^R\left(\hat{t}^*_{yR1}\right)$		$\widehat{\mathrm{V}}_1^R\left(\hat{t}^*_{yR2}\right)$	
f_1 (%)	f_2 (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)
5	5	3904	100	-8,2	92,7	-6,3	93,7
5	10	3766	100	-5,4	93,7	-2,8	94,3
5	25	3146	100	-3,7	94,2	-1,6	94,7
5	50	2116	100	-3,7	94,2	-0,8	94,6
10	5	3747	100	-10,8	93,0	-6,6	93,8
10	10	3569	100	-8,8	93,5	-4,9	94,3
10	25	2984	100	-9,8	93,7	-3,0	94,6
10	50	2059	100	-6,2	94,0	-1,2	94,6
20	5	3297	100	-19,4	92,0	-14,6	92,8
20	10	3153	100	-19,0	92,1	-11,2	93,4
20	25	2744	100	-15,8	92,7	-5,5	94,3
20	50	1945	100	-10,6	93,5	-3,1	94,7

Tableau II.7 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan aléatoire simple sans remise aux deux phases pour la population avec $\rho = 0,9$.

		$\hat{t}^*_{y\pi}$		$\hat{t}^*_{yR1}$		$\hat{t}^*_{yR2}$	
f_1 $(\%)$	f_2 $(\%)$	$\mathrm{E_{MC}}\left(C_2\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2^R\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2^R\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2^R\right)$ $(\%)$
5	5	-4354	-4572	8,4	8,8	5,8	8,8
5	10	-3920	-4334	6,2	6,9	3,4	6,9
5	25	-3197	-4210	4,4	5,8	1,4	5,8
5	50	-2138	-4170	2,8	5,4	0,5	5,4
10	5	-3916	-4102	11,3	11,8	8,0	11,8
10	10	-3629	-3991	9,9	10,9	5,5	10,9
10	25	-3030	-3939	8,0	10,4	2,6	10,4
10	50	-2063	-3920	5,4	10,2	1,0	10,2
20	5	-3400	-3543	20,0	20,8	14,7	20,8
20	10	-3213	-3498	18,7	20,4	10,9	20,4
20	25	-2741	-3472	15,9	20,2	5,6	20,2
20	50	-1925	-3465	11,2	20,1	2,3	20,1

Tableau II.8 – Contribution de $\widehat{\mathrm{V}}_2^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan aléatoire simple sans remise aux deux phases pour la population avec $\rho = 0,9$.

II.2 Étude 2 : plan Bernoulli à la deuxième phase

II.2.1 Résultats pour la population avec $\rho = 0,5$

		$\hat{t}^*_{y\pi}$		$\hat{t}^*_{yR1}$		$\hat{t}^*_{yR2}$	
f_1 (%)	f_2 (%)	RB_{MC} (%)	CV_{MC} (%)	RB_{MC} (%)	CV_{MC} (%)	RB_{MC} (%)	CV_{MC} (%)
5	5	0,0	19,9	0,0	3,0	0,0	3,1
5	10	0,1	13,8	0,0	2,1	0,0	2,2
5	25	0,1	8,0	0,0	1,3	0,0	1,4
5	50	0,1	4,7	0,0	0,9	0,0	1,1
10	5	0,1	14,2	0,0	2,1	0,0	2,1
10	10	0,0	9,8	0,0	1,5	0,0	1,5
10	25	0,1	5,6	0,0	0,9	0,0	1,0
10	50	0,0	3,3	0,0	0,6	0,0	0,8
20	5	0,0	9,9	0,0	1,5	0,0	1,5
20	10	0,1	6,8	0,0	1,0	0,0	1,1
20	25	0,0	4,0	0,0	0,6	0,0	0,7
20	50	0,0	2,3	0,0	0,4	0,0	0,5

Tableau II.9 – Performance des estimateurs ponctuels dans le cadre de l'étude par simulation avec un plan Bernoulli à la deuxième phase pour la population avec $\rho = 0,5$.

		$\widehat{\mathrm{V}}(\hat{t}^*_{y\pi})$		$\widehat{\mathrm{V}}(\hat{t}^*_{yR1})$		$\widehat{\mathrm{V}}(\hat{t}^*_{yR2})^{\mathrm{alt}}$	
f_1 (%)	f_2 (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)
5	5	0,1	94,0	-8,2	91,4	-7,6	91,8
5	10	-1,0	94,7	-4,8	93,3	-4,3	93,5
5	25	0,4	95,0	2,5	95,0	2,0	94,9
5	50	-0,4	95,0	-0,8	94,8	-1,4	94,9
10	5	-0,8	94,6	-2,2	93,5	-1,9	93,5
10	10	-1,1	95,0	-1,1	94,1	-1,5	94,2
10	25	-0,1	95,0	0,7	95,0	0,9	94,8
10	50	0,5	95,2	1,5	95,0	0,2	95,0
20	5	1,1	94,8	-3,8	93,8	-3,8	93,9
20	10	1,7	94,8	-0,7	94,7	-0,6	94,7
20	25	0,4	94,8	-2,3	94,5	-2,7	94,4
20	50	-0,9	94,9	-0,5	94,8	0,2	94,9

Tableau II.10 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R(\cdot)+\widehat{\mathrm{V}}_2^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan Bernoulli à la deuxième phase pour la population avec $\rho = 0,5$.

		$\widehat{\mathrm{V}}_1^R(\hat{t}^*_{y\pi})$		$\widehat{\mathrm{V}}_1^R(\hat{t}^*_{yR1})$		$\widehat{\mathrm{V}}_1^R(\hat{t}^*_{yR2})$	
f_1 (%)	f_2 (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)
5	5	-4,9	93,3	-12,5	90,8	-11,8	91,3
5	10	-6,0	93,9	-9,1	92,8	-8,3	93,0
5	25	-4,6	94,5	-1,4	94,6	-1,1	94,6
5	50	-5,2	94,3	-3,3	94,5	-3,1	94,8
10	5	-10,7	93,1	-11,6	92,2	-10,9	92,4
10	10	-11,0	93,6	-10,1	92,9	-9,7	93,3
10	25	-10,0	93,8	-7,1	94,0	-5,4	94,1
10	50	-9,2	93,9	-3,9	94,4	-3,4	94,7
20	5	-19,1	91,9	-22,2	90,8	-21,6	91,0
20	10	-18,6	92,2	-18,9	92,2	-17,5	92,3
20	25	-19,5	92,0	-17,8	92,2	-15,4	92,5
20	50	-20,1	91,9	-11,6	93,4	-7,5	93,9

Tableau II.11 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan Bernoulli à la deuxième phase pour la population avec $\rho = 0,5$.

		$\hat{t}^*_{y\pi}$		$\hat{t}^*_{yR1}$		$\hat{t}^*_{yR2}$	
f_1 $(\%)$	f_2 $(\%)$	$\mathrm{E_{MC}}\left(C_2\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2^R\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2^R\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2\right)$ $(\%)$	$\mathrm{E_{MC}}\left(C_2^R\right)$ $(\%)$
5	5	5,0	5	4,8	5	4,5	5
5	10	5,0	5	4,5	5	4,1	5
5	25	4,9	5	3,8	5	3,0	5
5	50	4,8	5	2,6	5	1,7	5
10	5	10,0	10	9,5	10	9,1	10
10	10	10,0	10	9,1	10	8,3	10
10	25	9,9	10	7,7	10	6,2	10
10	50	9,6	10	5,3	10	3,6	10
20	5	20,0	20	19,2	20	18,4	20
20	10	19,9	20	18,4	20	16,9	20
20	25	19,8	20	15,8	20	13,0	20
20	50	19,4	20	11,1	20	7,7	20

Tableau II.12 – Contribution de $\widehat{\mathrm{V}}_2^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan Bernoulli à la deuxième phase pour la population avec $\rho = 0,5$.

II.2.2 Résultats pour la population avec $\rho = 0,9$

		$\hat{t}^*_{y\pi}$		$\hat{t}^*_{yR1}$		$\hat{t}^*_{yR2}$	
f_1 (%)	f_2 (%)	RB_{MC} (%)	CV_{MC} (%)	RB_{MC} (%)	CV_{MC} (%)	RB_{MC} (%)	CV_{MC} (%)
5	5	0,0	19,7	0,0	1,0	0,0	1,2
5	10	0,1	13,7	0,0	0,7	0,0	0,9
5	25	0,1	7,8	0,0	0,4	0,0	0,8
5	50	0,1	4,6	0,0	0,3	0,0	0,7
10	5	0,0	14,0	0,0	0,7	0,0	0,8
10	10	0,0	9,7	0,0	0,5	0,0	0,7
10	25	0,0	5,6	0,0	0,3	0,0	0,5
10	50	0,0	3,2	0,0	0,2	0,0	0,5
20	5	0,0	9,8	0,0	0,5	0,0	0,6
20	10	0,1	6,8	0,0	0,3	0,0	0,4
20	25	0,0	3,9	0,0	0,2	0,0	0,4
20	50	0,0	2,3	0,0	0,1	0,0	0,3

Tableau II.13 – Performance des estimateurs ponctuels dans le cadre de l'étude par simulation avec un plan Bernoulli à la deuxième phase pour la population avec $\rho = 0,9$.

		$\widehat{\mathrm{V}}\left(\hat{t}^*_{y\pi}\right)$		$\widehat{\mathrm{V}}\left(\hat{t}^*_{yR1}\right)$		$\widehat{\mathrm{V}}\left(\hat{t}^*_{yR2}\right)^{\mathrm{alt}}$	
f_1 (%)	f_2 (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)
5	5	0,5	94,2	-7,4	91,6	-5,5	93,0
5	10	-1,2	94,7	-4,2	93,1	-2,7	94,4
5	25	1,0	95,0	-1,7	94,6	0,6	95,1
5	50	-0,9	94,9	-0,7	94,9	-0,6	94,7
10	5	-0,6	94,4	-1,9	93,7	-1,4	94,2
10	10	-1,6	94,8	-2,3	94,2	-2,4	94,7
10	25	-0,4	94,9	-1,1	94,6	1,2	95,0
10	50	0,6	94,9	0,9	94,9	-0,9	94,7
20	5	1,1	94,7	-2,2	94,1	-1,5	94,7
20	10	1,3	94,9	-0,2	94,6	-1,0	94,9
20	25	0,3	95,0	-1,0	94,7	0,5	95,2
20	50	0,3	95,0	-2,0	94,9	-1,9	94,7

Tableau II.14 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R(\cdot)+\widehat{\mathrm{V}}_2^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan Bernoulli à la deuxième phase pour la population avec $\rho = 0,9$.

		$\widehat{\mathrm{V}}_1^R\left(\hat{t}^*_{y\pi}\right)$		$\widehat{\mathrm{V}}_1^R\left(\hat{t}^*_{yR1}\right)$		$\widehat{\mathrm{V}}_1^R\left(\hat{t}^*_{yR2}\right)$	
f_1 (%)	f_2 (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)	$\mathrm{RB}_{\mathrm{MC}}$ (%)	$\mathrm{PC}_{\mathrm{MC}}$ (%)
5	5	-4,5	93,6	-11,8	90,9	-8,6	92,6
5	10	-6,1	94,1	-8,5	92,5	-5,0	94,0
5	25	-4,0	94,5	-5,4	94,2	-0,6	94,9
5	50	-5,8	94,3	-3,2	94,6	-1,1	94,7
10	5	-10,5	93,2	-11,2	92,6	-8,1	93,4
10	10	-11,4	93,5	-11,2	92,9	-7,3	94,1
10	25	-10,3	93,7	-8,7	93,7	-1,3	94,7
10	50	-9,3	93,8	-4,4	94,5	-1,9	94,6
20	5	-19,1	91,9	-21,0	91,0	-15,4	92,9
20	10	-18,9	92,1	-18,5	92,0	-11,6	93,4
20	25	-19,6	92,2	-16,6	92,5	-5,1	94,6
20	50	-19,4	91,8	-12,9	93,2	-4,1	94,4

Tableau II.15 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan Bernoulli à la deuxième phase pour la population avec $\rho = 0,9$.

		$\hat{t}^*_{y\pi}$		$\hat{t}^*_{yR1}$		$\hat{t}^*_{yR2}$	
f_1 (%)	f_2 (%)	$\mathrm{E_{MC}}(C_2)$ (%)	$\mathrm{E_{MC}}(C_2^R)$ (%)	$\mathrm{E_{MC}}(C_2)$ (%)	$\mathrm{E_{MC}}(C_2^R)$ (%)	$\mathrm{E_{MC}}(C_2)$ (%)	$\mathrm{E_{MC}}(C_2^R)$ (%)
5	5	5,0	5	4,8	5	3,2	5
5	10	5,0	5	4,5	5	2,4	5
5	25	5,0	5	3,8	5	1,2	5
5	50	4,9	5	2,6	5	0,5	5
10	5	10,0	10	9,5	10	6,7	10
10	10	10,0	10	9,1	10	5,0	10
10	25	9,9	10	7,7	10	2,5	10
10	50	9,8	10	5,3	10	1,0	10
20	5	20,0	20	19,2	20	14,0	20
20	10	20,0	20	18,4	20	10,6	20
20	25	19,9	20	15,8	20	5,5	20
20	50	19,6	20	11,1	20	2,3	20

Tableau II.16 – Contribution de $\widehat{\mathrm{V}}_2^R(\cdot)$ dans le cadre de l'étude par simulation avec un plan Bernoulli à la deuxième phase pour la population avec $\rho = 0,9$.

II.3 Étude 3 : plan à deux degrés

II.3.1 Résultats pour la population avec ICC = 20 %

f_1 (%)	m_g	$\mathrm{RB}_{\mathrm{MC}}\left(\hat{t}^*_{y\pi}\right)$ (%)	$\mathrm{CV}_{\mathrm{MC}}\left(\hat{t}^*_{y\pi}\right)$ (%)
2	2	0,0	4,9
2	5	0,0	4,0
2	10	0,0	3,7
10	2	0,0	2,1
10	5	0,0	1,8
10	10	0,0	1,6

Tableau II.17 – Performance des estimateurs ponctuels dans le cadre de l'étude par simulation avec un plan à deux degrés pour la population avec ICC = 20 %.

f_1 (%)	m_g	$\mathrm{RB}_{\mathrm{MC}}\left(\widehat{\mathrm{V}}\left(\hat{t}^*_{y\pi}\right)\right)$ (%)	$\mathrm{CV}_{\mathrm{MC}}\left(\widehat{\mathrm{V}}\left(\hat{t}^*_{y\pi}\right)\right)$ (%)
2	2	0,7	91,8
2	5	0,2	91,6
2	10	0,9	91,7
10	2	1,0	94,5
10	5	-0,3	93,8
10	10	-0,4	93,7

Tableau II.18 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R\left(\hat{t}^*_{y\pi}\right)+\widehat{\mathrm{V}}_2^R\left(\hat{t}^*_{y\pi}\right)$ dans le cadre de l'étude par simulation avec un plan à deux degrés pour la population avec ICC = 20 %.

f_1 (%)	m_g	$\mathrm{RB}_{\mathrm{MC}}\left(\widehat{\mathrm{V}}^1\left(\hat{t}^*_{y\pi}\right)\right)$ (%)	$\mathrm{CV}_{\mathrm{MC}}\left(\widehat{\mathrm{V}}^1\left(\hat{t}^*_{y\pi}\right)\right)$ (%)
2	2	-0,3	91,6
2	5	-0,3	91,5
2	10	0,7	91,7
10	2	-4,1	93,7
10	5	-3,0	93,4
10	10	-1,6	93,5

Tableau II.19 – Performance de l'estimateur de variance $\widehat{\mathrm{V}}_1^R\left(\hat{t}^*_{y\pi}\right)$ dans le cadre de l'étude par simulation avec un plan à deux degrés pour la population avec ICC $= 20$ %.

f_1 (%)	m_g	$\mathrm{E}_{\mathrm{MC}}\left(C_2\right)$ (%)	$\mathrm{E}_{\mathrm{MC}}\left(C_2^R\right)$ (%)
2	2	1,3	2,0
2	5	0,7	2,0
2	10	0,3	2,0
10	2	5,3	10,0
10	5	2,8	10,1
10	10	1,3	10,2

Tableau II.20 – Contribution de $\widehat{\mathrm{V}}_2^R\left(\hat{t}^*_{y\pi}\right)$ dans le cadre de l'étude par simulation avec un plan à deux degrés pour la population avec ICC $= 20$ %.