Université de Montréal

Estimation de la variance en présence de données imputées pour des plans de sondage à grande entropie

par

Audrey-Anne Vallée

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

juillet 2014

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé

Estimation de la variance en présence de données imputées pour des plans de sondage à grande entropie

présenté par

Audrey-Anne Vallée

a été évalué par un jury composé des personnes suivantes :

Christian Léger
(président-rapporteur)

David Haziza
(directeur de recherche)

Alejandro Murua
(membre du jury)

Mémoire accepté le :
14 juillet 2014

SOMMAIRE

Les travaux portent sur l'estimation de la variance dans le cas d'une non-réponse partielle traitée par une procédure d'imputation. Traiter les valeurs imputées comme si elles avaient été observées peut mener à une sous-estimation substantielle de la variance des estimateurs ponctuels. Les estimateurs de variance usuels reposent sur la disponibilité des probabilités d'inclusion d'ordre deux, qui sont parfois difficiles (voire impossibles) à calculer. Nous proposons d'examiner les propriétés d'estimateurs de variance obtenus au moyen d'approximations des probabilités d'inclusion d'ordre deux. Ces approximations s'expriment comme une fonction des probabilités d'inclusion d'ordre un et sont généralement valides pour des plans à grande entropie. Les résultats d'une étude de simulation, évaluant les propriétés des estimateurs de variance proposés en termes de biais et d'erreur quadratique moyenne, seront présentés.

Mots clés : Imputation, non-réponse, estimation de la variance, entropie d'un plan de sondage, probabilités d'inclusion d'ordre deux.

SUMMARY

Variance estimation in the case of item nonresponse treated by imputation is the main topic of this work. Treating the imputed values as if they were observed may lead to substantial under-estimation of the variance of point estimators. Classical variance estimators rely on the availability of the second-order inclusion probabilities, which may be difficult (even impossible) to calculate. We propose to study the properties of variance estimators obtained by approximating the second-order inclusion probabilities. These approximations are expressed in terms of first-order inclusion probabilities and are usually valid for high entropy sampling designs. The results of a simulation study evaluating the properties of the proposed variance estimators in terms of bias and mean squared error will be presented.

Keywords : Imputation, nonresponse, variance estimation, entropy of a sampling design, second-order inclusion probabilities.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

REMERCIEMENTS

Je tiens particulièrement à remercier mon directeur, David Haziza. Ce fut un réel plaisir et privilège d'apprendre de son expérience en échantillonnage. Je le remercie de m'avoir si bien accompagnée tout au long de l'élaboration de ce mémoire. Ses encouragements continuels, sa patience, sa disponibilité et son travail méticuleux ont permis la réussite de cette formidable année universitaire. Je le remercie aussi pour le soutien qu'il m'apporte dans la poursuite de ma carrière.

Un gros merci à ma famille qui a su m'encourager dans mes accomplissements. En particulier, je remercie mes parents, qui sont pour moi des modèles de vie. Leur amour et leur soutien me stimulent à me dépasser autant dans mes choix personnels, académiques et de carrière. Merci aussi à Julien, qui m'a appuyé et qui a embelli mes journées tout au long de la rédaction de ce mémoire. Merci de m'avoir autant encouragée, réconfortée et d'avoir été présent en tout temps.

J'ai aussi une pensée pour mes collègues Paule-Marjolaine, Janie, Alexandre et Gabrielle, avec qui j'ai vécu des expériences inoubliables. Ils ont su égayer mes journées et agrémenter mes études. Ce fut un plaisir d'être en votre compagnie tout au long de ce cheminement.

INTRODUCTION

Les enquêtes statistiques jouent un rôle primordial dans la société. Les organismes statistiques (Statistique Canada par exemple) réalisent des enquêtes sur plusieurs sphères de la population canadienne telles la santé, l'éducation, l'environnement, l'économie, la culture, etc. Les statisticiens participent à plusieurs étapes des enquêtes statistiques. Lorsque les obectifs et la population cible d'une enquête sont clairement énoncés, il faut définir la base de sondage. Cette dernière permet d'identifier les unités de la population d'enquête et les moyens de communication avec celles-ci. Une fois la base de sondage bien définie, les statisticiens déterminent la méthode de sélection d'un échantillon. À chaque unité de la base de sondage, on assigne une probabilité d'inclusion et l'échantillon est tiré aléatoirement à partir de cette base de sondage, selon une procédure d'échantillonnage bien définie respectant les probabilités d'inclusion. Les statisticiens participent ensuite à la conception d'un questionnaire et au choix de la méthode de collecte de données. Il peut y avoir, par exemple, des entrevues en face à face, au téléphone ou des questionnaires envoyés par la poste. Une fois les données collectées, les statisticiens vérifient et corrigent les bases de données en imputant les valeurs manquantes. Lorsque la base de données est complète, les statisticiens procèdent à l'estimation de paramètres d'intérêt, comme le total d'une variable d'intérêt.

Lors d'une enquête, les estimations sont exposées à plusieurs sources d'erreurs : les erreurs dues à l'échantillonnage et les erreurs non dues à l'échantillonage. Les erreurs dues à l'échantillonnage sont causées par le fait que la variable d'intérêt est mesurée sur une partie de la population seulement, au lieu de la population au complet. Parmi les erreurs non dues à l'échantillonnage, on retrouve entre autres les erreurs de couverture, de mesure et de non-réponse. Nous sommes en présence d'erreurs de couverture lorsque la base de sondage ne contient pas tous les individus de la population cible et/ou elle contient des individus qui ne font pas partie de la population cible. L'erreur de mesure est la différence entre les valeurs inscrites dans la base de données et les vraies valeurs. Dans ce mémoire, on

s'intéresse particulièrement aux erreurs de non-réponse, au traitement de la non-réponse par imputation et à l'estimation de la variance en présence de données imputées.

La non-réponse est inévitable dans les enquêtes menées par les organismes statistiques comme Statistique Canada. Les statisticiens d'enquête distinguent la non-réponse totale de la non-réponse partielle. La première survient lorsqu'aucune information n'est collectée sur une unité, alors que certaines variables d'intérêt, mais pas toutes, sont manquantes dans le cas de la deuxième. Les travaux de recherche portent sur l'estimation de la variance dans le cas d'une non-réponse partielle traitée par une procédure d'imputation. L'imputation consiste à remplacer une valeur manquante par une valeur artificielle construite au moyen d'informations auxiliaires disponibles pour toutes les unités échantillonnées. Bien entendu, le fait de traiter les valeurs imputées comme si elles avaient été observées peut mener à une sous-estimation substantielle de la variance des estimateurs ponctuels, particulièrement si le taux de réponse est faible. Dans la littérature, on retrouve plusieurs méthodes d'estimation de la variance. Cependant, quel que soit la méthode, les estimateurs de variance reposent sur la disponibilité des probabilités d'inclusion d'ordre deux dans l'échantillon, qui sont parfois difficiles (voire impossibles) à calculer. Dans ce mémoire, nous développons des estimateurs de variance obtenus au moyen d'approximations des probabilités d'inclusion d'ordre deux. Ces approximations, qui s'expriment comme une fonction des probabilités d'inclusion d'ordre un, sont valides pour des plans à grande entropie tels que le plan de Poisson conditionnel et le plan de Rao-Sampford. En pratique, les fichiers de données produits par des organismes de statistique ne contiennent habituellement pas les probabilités d'inclusion d'ordre deux. Seules les probabilités d'ordre un y sont présentes. Pouvoir estimer la variance des estimateurs en ne requérant que les probabilités d'inclusion d'ordre un représente donc un avantage dans la pratique. Une vaste étude de simulation sera effectuée afin d'évaluer empiriquement les propriétés des estimateurs de variance proposés. Les résultats obtenus seront utiles aux praticiens car les estimateurs de variance proposés pourront être facilement obtenues au moyen de procédures informatiques usuelles.

Chapitre 1

INTRODUCTION À L'ÉCHANTILLONNAGE ET À L'ESTIMATION DE LA VARIANCE

1.1. Notions d'échantillonnage

Considérons une population finie U de taille N. On cherche à estimer le total dans la population d'une variable y, noté $Y = \sum_{i \in U} y_i$. De cette population, on tire un échantillon $s \subseteq U$ de taille n, selon un certain plan de sondage. Un plan de sondage est une loi de probabilité $p(.)$ faisant correspondre à chaque échantillon $s \subset \Omega$ une probabilité d'être tiré $p(s)$, où Ω est l'ensemble de tous les échantillons possibles s. La loi de probabilité $p(s)$ satisfait :

(i) $p(s) \geq 0, \quad \forall s \subset \Omega\,;$

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

À chaque individu i de la population U correspond une variable indicatrice de sélection définie par

$$\delta_i = \begin{cases} 1 & \text{si } i \text{ est dans l'échantillon } s, \\ 0 & \text{sinon.} \end{cases}$$

On définit la probabilité d'inclusion d'ordre un de l'unité i selon

$$\pi_i = \mathrm{P}(i \in s) = \mathrm{P}(\delta_i = 1) = \sum_{\substack{s \in \Omega \\ s \ni i}} p(s).$$

On supposera que $\pi_i > 0$ pour tout $i \in U$. La probabilité d'inclusion d'ordre deux, π_{ij}, est la probabilité que les individus i et j soient tous deux présents dans l'échantillon tiré. Elle est définie selon

$$\pi_{ij} = \mathrm{P}(i \in s, j \in s) = \mathrm{P}(\delta_i = 1, \delta_j = 1) = \sum_{\substack{s \in \Omega \\ s \ni (i,j)}} p(s).$$

Notons que $\pi_{ij} = \pi_i$ si $i = j$.

On distingue les plans de sondage à taille fixe de ceux à taille aléatoire. Pour un plan à taille fixe, on utilisera la notation n pour désigner la taille de l'échantillon, alors que l'on utilisera la notation n_s pour un plan à taille aléatoire. Pour un plan à taille fixe, on a

$$n = \sum_{i \in U} \pi_i, \tag{1.1.1}$$

en notant que $n = \sum\limits_{i \in U} \delta_i$ et que

$$\mathrm{E}_p(\delta_i) = \pi_i, \tag{1.1.2}$$

où $\mathrm{E}_p(.)$ désigne l'espérance par rapport au plan de sondage. De même, pour un plan à taille aléatoire, on a

$$\mathrm{E}_p(n_s) = \sum_{i \in U} \pi_i. \tag{1.1.3}$$

Notons que $\mathrm{E}_p(n_s)$ représente la taille espérée de l'échantillon.

Soient $\mathrm{V}_p(.)$ et $\mathrm{Cov}_p(.)$ la variance et la covariance par rapport au plan de sondage, respectivement.

Proposition 1.1.1. *Les probabilités d'inclusion respectent les propriétés suivantes pour un certain plan de sondage $p(s)$:*

(i) $\mathrm{V}_p(\delta_i) = \pi_i(1 - \pi_i)$ *;*

(ii) $\mathrm{E}_p(\delta_i \delta_j) = \pi_{ij}$ *;*

(iii) $\mathrm{Cov}_p(\delta_i, \delta_j) = \pi_{ij} - \pi_i \pi_j,\ i \neq j.$

Proposition 1.1.2. *Pour un plan de sondage $p(s)$ à taille fixe, on a :*

(i) $\sum\limits_{\substack{j \in U \\ j \neq i}} \pi_{ij} = (n-1)\pi_i,\ \forall\ i \in U;$

(ii) $\sum\limits_{i \in U} \sum\limits_{\substack{j \in U \\ j \neq i}} \pi_{ij} = n(n-1).$

L'information sur y n'étant recueillie que pour les individus présents dans l'échantillon, on estimera le total par l'estimateur d'Horvitz-Thompson :

$$\widehat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}. \tag{1.1.4}$$

Proposition 1.1.3. *Si $\pi_i > 0$ pour tout $i \in U$, l'estimateur d'Horvitz-Thompson (1.1.4) est sans biais pour Y par rapport au plan de sondage ; i.e.,* $\mathrm{E}_p(\widehat{Y}_{HT}) = Y$.

Démonstration. Commençons par écrire

$$\widehat{Y}_{HT} = \sum_{i \in U} \frac{y_i}{\pi_i} \delta_i.$$

On a alors que

$$\mathrm{E}_p(\widehat{Y}_{HT}) = \mathrm{E}_p\left(\sum_{i\in U}\frac{y_i}{\pi_i}\delta_i\right) = \sum_{i\in U}\frac{y_i}{\pi_i}\mathrm{E}_p(\delta_i) = Y.$$

□

Proposition 1.1.4. *La variance par rapport au plan de sondage de* $\widehat{Y}_{HT}$ *est donnée par*

$$\mathrm{V}_p(\widehat{Y}_{HT}) = \sum_{i\in U}\sum_{j\in U}\Omega_{ij}\, y_i\, y_j, \tag{1.1.5}$$

avec $\Omega_{ij} = \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}$.

Démonstration.

$$\begin{aligned}
\mathrm{V}_p(\widehat{Y}_{HT}) &= \mathrm{V}_p\left(\sum_{i\in U}\frac{y_i}{\pi_i}\delta_i\right)\\
&= \sum_{i\in U}\frac{y_i^2}{\pi_i^2}\mathrm{V}_p(\delta_i) + \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\mathrm{Cov}_p(\delta_i,\delta_j)\\
&= \sum_{i\in U}\frac{y_i^2}{\pi_i^2}\pi_i(1-\pi_i) + \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{ij}-\pi_i\pi_j)\\
&= \sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\\
&= \sum_{i\in U}\sum_{j\in U}\Omega_{ij}\, y_i\, y_j,
\end{aligned}$$

en notant que

$$\Omega_{ii} = \frac{\pi_i(1-\pi_i)}{\pi_i^2}.$$

□

L'information sur y n'étant disponible que pour les individus de l'échantillon, il est impossible de calculer cette variance. Autrement dit, la variance (1.1.5) est elle-même un paramètre de la population finie que l'on va devoir estimer. Un estimateur de $\mathrm{V}_p(\widehat{Y}_{HT})$ est l'estimateur de variance d'Horvitz-Thompson :

$$\widehat{\mathrm{V}}_{HT} = \sum_{i\in s}\sum_{j\in s}\frac{\Omega_{ij}}{\pi_{ij}}\, y_i\, y_j. \tag{1.1.6}$$

Cet estimateur peut être utilisé lorsque le plan de sondage est à taille fixe ou aléatoire.

Proposition 1.1.5. *L'estimateur (1.1.6) est sans biais pour la variance (1.1.5) par rapport au plan de sondage, pourvu que* $\pi_{ij} > 0$ *pour tout* $i \neq j$, *i.e.,* $\mathrm{E}_p(\widehat{\mathrm{V}}_{HT}) = \mathrm{V}_p(\widehat{Y}_{HT})$.

Démonstration.

$$\begin{aligned}
\mathrm{E}_p(\widehat{\mathrm{V}}_{HT}) &= \mathrm{E}_p\left\{\sum_{i\in s}\sum_{j\in s}\frac{\Omega_{ij}}{\pi_{ij}}\,y_i\,y_j\right\}\\
&= \mathrm{E}_p\left\{\sum_{i\in U}\sum_{j\in U}\frac{\Omega_{ij}}{\pi_{ij}}\,y_i\,y_j\delta_i\delta_j\right\}\\
&= \sum_{i\in U}\sum_{j\in U}\frac{\Omega_{ij}}{\pi_{ij}}\,y_i\,y_i\,\mathrm{E}_p(\delta_i\delta_j)\\
&= \sum_{i\in U}\sum_{j\in U}\frac{\Omega_{ij}}{\pi_{ij}}\,y_i\,y_j\,\pi_{ij}\\
&= \sum_{i\in U}\sum_{j\in U}\Omega_{ij}\,y_i\,y_j.
\end{aligned}$$

□

Pour un plan à taille fixe, la variance (1.1.5) peut également s'écrire comme

$$\mathrm{V}_p(\widehat{Y}_{HT}) = -\frac{1}{2}\sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2. \tag{1.1.7}$$

L'expression (1.1.7) suggère un autre estimateur de la variance : l'estimateur de Sen-Yates-Grundy donné par

$$\widehat{\mathrm{V}}_{SYG} = -\frac{1}{2}\sum_{i\in s}\sum_{j\in s}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2. \tag{1.1.8}$$

Proposition 1.1.6. *L'estimateur (1.1.8) est sans biais pour la variance (1.1.7) par rapport au plan de sondage, pourvu que* $\pi_{ij} > 0$ *pour tout* $i \neq j$, *i.e.,* $\mathrm{E}_p(\widehat{\mathrm{V}}_{SYG}) = \mathrm{V}_p(\widehat{Y}_{HT})$.

Démonstration.

$$\begin{aligned}
\mathrm{E}_p(\widehat{\mathrm{V}}_{SYG}) &= \mathrm{E}_p\left\{-\frac{1}{2}\sum_{i\in s}\sum_{j\in s}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\right\}\\
&= \mathrm{E}_p\left\{-\frac{1}{2}\sum_{i\in U}\sum_{j\in U}\delta_i\delta_j\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\right\}\\
&= -\frac{1}{2}\sum_{i\in U}\sum_{j\in U}\mathrm{E}_p\left(\delta_i\delta_j\right)\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\\
&= -\frac{1}{2}\sum_{i\in U}\sum_{j\in U}\pi_{ij}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\\
&= -\frac{1}{2}\sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2.
\end{aligned}$$

□

L'estimateur $\widehat{\mathrm{V}}_{SYG}$ est préférable à l'estimateur $\widehat{\mathrm{V}}_{HT}$ dans certaines situations. Considérons un plan à taille fixe pour lequel $y_i = c\pi_i$, où c est une constante. Dans ce cas, on a

$$\widehat{Y}_{HT} = Y, \text{ pour tout } s \in \Omega.$$

L'estimateur d'Horvitz-Thompson estime ainsi parfaitement le total Y. Par conséquent, on a

$$\mathrm{V}_p\left(\widehat{Y}_{HT}\right) = 0.$$

Examinons le comportement de $\widehat{\mathrm{V}}_{HT}$ et de $\widehat{\mathrm{V}}_{SYG}$ dans une telle situation. D'une part, en posant $y_i/\pi_i = c$ dans (1.1.8), on a bien

$$\widehat{\mathrm{V}}_{SYG} = 0, \ \forall \ s \ \in \ \Omega.$$

Dans ce cas, $\mathrm{V}_p\left(\widehat{\mathrm{V}}_{SYG}\right) = 0$. D'autre part, en posant $y_i/\pi_i = c$ dans (1.1.6), on obtient

$$\widehat{\mathrm{V}}_{HT} = c^2 \sum_{i\in s}\sum_{i\in s} \frac{\Omega_{ij}}{\pi_{ij}} \neq 0,$$

en général. On a donc $\mathrm{V}_p\left(\widehat{\mathrm{V}}_{HT}\right) > 0$. Notons également que l'estimateur $\widehat{\mathrm{V}}_{SYG}$ est non négatif si la condition suivante est satisfaite :

$$\pi_{ij} - \pi_i\pi_j < 0, \, i \neq j. \tag{1.1.9}$$

Cette condition est souvent appelée la condition de Sen-Yates-Grundy. L'estimateur $\widehat{\mathrm{V}}_{HT}$, quant à lui, peut être négatif même lorsque cette condition est satisfaite. En effet, l'estimateur $\widehat{\mathrm{V}}_{HT}$ étant sans biais, il n'a d'autre choix que de prendre des valeurs négatives. La plupart des plans de sondage rencontrés en pratique satisfont à la condition de Sen-Yates-Grundy.

1.2. Quelques plans de sondage

Nous présentons quelques plans de sondage qui seront utiles dans ce mémoire.

1.2.1. Plan aléatoire simple sans remise

Un des plans de sondage les plus simples est l'échantillonnage aléatoire simple sans remise (EASSR). Ce plan consiste à tirer un nombre fixe d'unités, de manière à ce que chaque sous-ensemble possible de n unités ait la même probabilité d'être pigé que n'importe quel autre sous-ensemble de n unités. Sachant qu'il y a $\binom{N}{n}$ échantillons possibles, chaque échantillon s de taille n a une probabilité

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

d'être tiré. Les probabilités d'inclusion d'ordre un sont égales à

$$\pi_i = \frac{n}{N}, \text{ pour tout } i.$$

Les probabilités d'inclusion d'ordre deux sont égales à

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}, \ i \neq j.$$

1.2.2. Plan de Bernoulli

Un plan de sondage simple, qui n'est pas à taille fixe, est le plan de Bernoulli (BE). Pour chaque unité de la population, on effectue une expérience de Bernoulli avec une probabilité $\pi \in (0,1)$. Si l'expérience est un succès, l'unité est sélectionnée dans l'échantillon s, sinon elle est rejetée. La probabilité de tirer l'échantillon s de taille n_s est

$$p(s) = \pi^{n_s}(1-\pi)^{N-n_s}. \tag{1.2.1}$$

Il y a 2^N échantillons possibles. Les probabilités d'inclusion d'ordre un sont égales à

$$\pi_i = \pi.$$

Les tirages étant indépendants, les probabilités d'inclusion d'ordre deux sont

$$\pi_{ij} = \pi^2, \ i \neq j.$$

La taille de l'échantillon n_s est une variable aléatoire suivant une loi binomiale de paramètres N et π, de telle sorte que $\mathrm{E}(n_s) = N\pi$ et $\mathrm{V}(n_s) = N\pi(1-\pi)$.

1.2.3. Plan de Bernoulli conditionnel

Le plan de Bernoulli conditionnel correspond au plan de Bernoulli avec une taille échantillonnale fixée n. Un échantillon de Bernoulli de taille n_s est tiré en répétant une expérience de Bernoulli pour chaque individu avec une probabilité $\pi \in (0,1)$. Si l'échantillon tiré est de taille $n_s = n$, alors l'échantillon est conservé. Si $n_s \neq n$, un nouvel échantillon de Bernoulli est tiré, et ce, jusqu'à l'obtention d'un échantillon de taille $n_s = n$. La probabilité de tirer l'échantillon s est

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

Ce plan correspond donc à l'EASSR.

1.2.4. Plan de Poisson

Le plan de Poisson est une généralisation du plan de Bernoulli, avec des probabilités d'inclusion d'ordre un inégales. Soient $\pi_1, \pi_2, ..., \pi_N$, les probabilités d'inclusion d'ordre un. Pour chaque unité i de la population, on effectue une expérience de Bernoulli avec probabilité π_i. La probabilité de tirer l'échantillon s est

$$p(s) = \prod_{k \in s} \pi_k \times \prod_{k \in U \setminus s} (1 - \pi_k).$$

La probabilité d'inclusion d'ordre un de l'unité i est donc π_i et la probabilité d'inclusion d'ordre deux pour les unités i et j est $\pi_{ij} = \pi_i \pi_j,\ i \neq j$. La taille n_s de l'échantillon s est aléatoire. Son espérance est

$$\mathrm{E}_p(n_s) = \sum_{i \in U} \pi_i$$

et sa variance est

$$\mathrm{V}_p(n_s) = \sum_{i \in U} \pi_i (1 - \pi_i).$$

1.2.5. Plan de Poisson conditionnel

Le plan de Poisson conditionnel correspond au plan de Poisson avec une taille d'échantillon fixée n. Un échantillon de Poisson de taille n_s est tiré en répétant une expérience de Bernoulli pour chaque individu i avec une probabilité $\pi_i \in (0,1)$. Si l'échantillon tiré est de taille $n_s = n$, alors l'échantillon est conservé. Si $n_s \neq n$, un nouvel échantillon est tiré selon un plan de Poisson, et ce, jusqu'à l'obtention d'un échantillon de taille $n_s = n$. La probabilité de tirer l'échantillon s est

$$p(s|n_s = n) = \frac{\prod_{k \in s} \pi_k \times \prod_{k \in U \setminus s} (1 - \pi_k)}{\sum_{s \in S_n} \prod_{k \in s} \pi_k \times \prod_{k \in U \setminus s} (1 - \pi_k)},$$

où S_n est l'ensemble de tous les échantillons possibles de taille n. Les probabilités d'inclusion d'ordre un, nommées $\widetilde{\pi}_i$, sont obtenues de manière itérative et elles sont un ajustement des probabilités π_i, utilisées dans l'algorithme. Les probabilités d'inclusion d'ordre deux π_{ij} sont également obtenues de manière itérative. Chen et coll. (1994) et Deville (2000) ont proposé des algorithmes pour calculer $\widetilde{\pi}_i$ et π_{ij}. D'autre part, lorsque $\pi_i = \pi$, le plan de Poisson conditionnel coïncide avec le plan aléatoire simple sans remise. Notons que le plan de Poisson conditionnel est également appelé le plan réjectif (Hájek, 1964).

1.2.6. Plan de Poisson séquentiel

Le plan de Poisson séquentiel (Ohlsson, 1998) est un plan à taille fixe n intimement lié au plan de Poisson. On commence par générer N variables aléatoires $x_i \sim \mathcal{U}[0,1]$. On calcule ensuite la variable u_i telle que

$$u_i = nx_i/\pi_i.$$

L'échantillon est composé des n unités ayant les plus petites valeurs de u_i. Les probabilités d'inclusion d'ordre un ne sont pas exactement π_i et les probabilités d'inclusion d'ordre deux sont complexes à calculer.

1.2.7. Plan de Rao-Sampford

Le plan de Rao-Sampford est un plan à taille fixe (Rao, 1965; Sampford, 1967). Pour obtenir un échantillon, un premier individu est tiré avec une probabilité p_i, où $p_1, ..., p_N$ sont telles que $\sum\limits_{i \in U} p_i = 1$. Ensuite, pour les $n-1$ autres unités de l'échantillon, l'unité i est tirée avec remise et avec une probabilité proportionnelle à $\frac{p_i}{1-np_i}$. Dans l'échantillon obtenu, si un individu est sélectionné plus d'une fois, l'échantillon est rejeté. Les étapes sont répétées jusqu'à l'obtention d'un échantillon de taille n dont tous les individus sont distincts. La probabilité d'inclusion d'ordre un est

$$\pi_i = np_i.$$

Les probabilités d'inclusion d'ordre deux sont telles que $\pi_{ij} > 0$ et satisfont la condition (1.1.9). Sampford (1967) propose une formule itérative pour calculer les probabilités d'inclusion d'ordre deux.

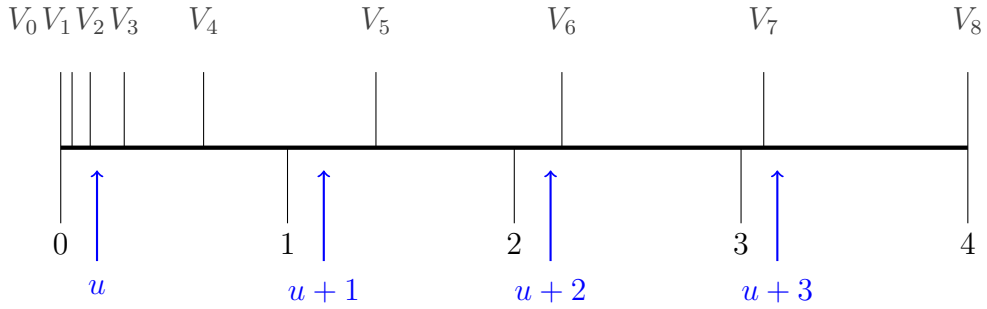
1.2.8. Plan systématique ordonné

Le plan systématique ordonné est un plan à taille fixe. Ce plan est souvent utilisé pour sa simplicité et sa facilité d'application. Les données sont premièrement ordonnées en ordre croissant de probabilité d'inclusion. Ensuite un échantillon systématique est tiré. Pour ce faire, définissons la valeur

$$V_i = \sum_{l=1}^{i} \pi_l, \ \forall \ i \in U,$$

avec $V_0 = 0$ et $V_N = \sum\limits_{i \in U} \pi_i = n$. On génère une variable uniforme u, telle que $u \sim \mathcal{U}[0,1]$. La première unité sélectionnée est l'unité i_1 telle que $V_{i_1-1} \leq u < V_{i_1}$. Les autres unités sélectionnées i_k sont telles que $V_{i_k-1} \leq u + k - 1 < V_{i_k}$, avec $k = 2, ..., n$.

FIGURE 1.1. Exemple : représentation de l'échantillon tiré



La probabilité de tirer l'échantillon s dépend des probabilités d'inclusion d'ordre un et de la distribution de la loi uniforme $\mathcal{U}[0,1]$. La probabilité d'inclusion d'ordre un de l'unité i est π_i. La probabilité d'inclusion d'ordre deux est

$$\pi_{ij} = min\left\{max(0,\pi_i - \delta_{ij}),\pi_j\right\} + min\left\{\pi_i,max(0,\delta_{ij} + \pi_j - 1)\right\},$$

pour $i < j$ et avec $\delta_{ij} = V_{ij} - \lfloor V_{ij} \rfloor$ et $V_{ij} = \sum_{l=i}^{j-1} \pi_l$. Dans ce plan, certaines probabilités d'inclusion d'ordre deux sont nulles. Dans un contexte d'estimation de la variance, cela représente un inconvénient important.

Illustrons le plan de sondage systématique par un exemple. Supposons que $N = 8$ et que $n = 4$. Les probabilités d'inclusion d'ordre un et les valeurs V_i sont données dans le tableau 1.1. Supposons ensuite que la valeur aléatoire uniforme est $u = 0{,}16$. La sélection de l'échantillon est illustrée à la figure 1.1. L'échantillon tiré est $s = \{3,5,6,8\}$.

TABLEAU 1.1. Exemple : probabilités d'inclusion et probabilités cumulatives

i	0	1	2	3	4	5	6	7	8	Total
π_i	0	0,05	0,08	0,15	0,35	0,76	0,82	0,89	0,90	4
V_i	0	0,05	0,13	0,28	0,63	1,39	2,21	3,10	4	

1.2.9. Plan systématique randomisé

Le plan systématique randomisé est une variante du plan systématique ordonné défini à la section 1.2.8. Les unités sont d'abord triées dans un ordre aléatoire. Ensuite, un échantillon systématique est tiré selon la méthode décrite dans la section 1.2.8. La probabilité de tirer l'échantillon s est complexe et dépend des probabilités d'inclusion d'ordre un et d'une partie combinatoire pour l'effet de la randomisation. La probabilité d'inclusion d'ordre deux est la moyenne, pour toutes les permutations possibles, des probabilités d'inclusion d'ordre deux, selon un plan systématique ordonné. Le calcul des probabilités d'inclusion d'ordre deux relève donc d'un problème complexe de combinatoire. Notons que l'ensemble de

tous les échantillons possible est obtenu en exhibant toutes les permutations possibles de la population et, pour chaque permutation, en dressant la liste de tous les échantillons systématiques possibles. Par conséquent, les probabilités d'inclusion d'ordre deux sont généralement strictement positives, bien qu'il ne soit pas impossible d'observer $\pi_{ij} = 0$ pour certains couples (i,j), voir Tillé (2006).

1.3. Plans à grande entropie

L'entropie d'un plan de sondage $p(s)$ est définie selon

$$\mathrm{I}(p) = -\sum_{s\in\Omega} p(s) log\, p(s). \tag{1.3.1}$$

Un plan de sondage à grande entropie est caractérisé par une grande difficulté à prédire l'échantillon sélectionné, donc par un haut niveau d'incertitude face aux individus qui seront tirés.

Proposition 1.3.1. *Le plan qui maximise l'entropie sur l'ensemble* $\Omega = \{s|s \subseteq U\}$ *est le plan de Bernoulli avec* $\pi = 1/2$.

Démonstration. Notons $\Omega = \{s_1,...,s_{2^N}\}$, l'ensemble de tous les échantillons possibles. On veut maximiser l'entropie (1.3.1) avec la contrainte

$$\sum_{t=1}^{2^N} p(s_t) = 1, \tag{1.3.2}$$

On fait donc face à une maximisation sous contrainte. On cherche à maximiser la fonction lagrangienne

$$\mathcal{L}\left(p(s_1),...,p(s_{2^N}),\lambda\right) = -\sum_{t=1}^{2^N} p(s_t)\ \log\ p(s_t) + \lambda\left\{\sum_{t=1}^{2^N} p(s_t) - 1\right\},$$

où λ désigne un multiplicateur de Lagrange. On a

$$\frac{\partial\mathcal{L}\left(p(s_1),...,p(s_{2^N}),\lambda\right)}{\partial\, p(s_k)} = -\sum_{t=1}^{2^N} \frac{\partial\, p(s_t)\ \log\ p(s_t)}{\partial\, p(s_k)} + \lambda\sum_{t=1}^{2^N} \frac{\partial\, p(s_t)}{\partial\, p(s_k)} = 0,$$

ou

$$-\left\{\log\ p(s_k) + \frac{p(s_k)}{p(s_k)}\right\} + \lambda = 0,$$

ou encore

$$\lambda - 1 - \log\, p(s_k) = 0,$$

ce qui conduit à

$$p(s_k) = e^{\lambda - 1}, \tag{1.3.3}$$

pour $k = 1,...,2^N$. En insérant (1.3.3) dans (1.3.2), on obtient que

$$\sum_{t=1}^{2^N} e^{\lambda-1} = 2^N e^{\lambda-1} = 1$$

et

$$e^{\lambda-1} = \frac{1}{2^N}.$$

Par (1.3.3), on a que

$$p(s_t) = \frac{1}{2^N},\ \text{pour } s_t \in \Omega. \tag{1.3.4}$$

La loi de probabilité d'un plan de Bernoulli (1.2.1), avec $\pi = \frac{1}{2}$, conduit à

$$p(s_t) = \left(\frac{1}{2}\right)^{n_s} \left(1 - \frac{1}{2}\right)^{N-n_s} = \left(\frac{1}{2}\right)^N,$$

pour $s_t \in \Omega$. Donc le plan de sondage (1.3.4) correspond à un plan de Bernoulli avec $\pi = \frac{1}{2}$. □

Proposition 1.3.2. *Le plan qui maximise l'entropie sur l'ensemble* $\Omega = \{s|s \subseteq U\}$ *avec une taille fixe* n *est l'échantillonnage aléatoire simple sans remise.*

Démonstration. Notons $\Omega = \{s_1,...,s_{\binom{N}{n}}\}$, l'ensemble de tous les échantillons possibles de taille fixe n. On veut maximiser l'entropie (1.3.1) avec la contrainte

$$\sum_{t=1}^{\binom{N}{n}} p(s_t) = 1, \tag{1.3.5}$$

où $s_t \in \Omega$, l'ensemble de tous les échantillons de taille n. On veut maximiser la fonction lagrangienne

$$\mathcal{L}\left(p(s_1),...,p(s_{\binom{N}{n}})\right) = -\sum_{t=1}^{\binom{N}{n}} p(s_t)\, \log\, p(s_t) + \lambda \left\{ \sum_{t=1}^{\binom{N}{n}} p(s_t) - 1 \right\}.$$

On trouve, de manière semblable à la démonstration précédente, que

$$p(s_t) = e^{\lambda - 1}. \tag{1.3.6}$$

En insérant (1.3.6) dans (1.3.5), on obtient que

$$\sum_{t=1}^{\binom{N}{n}} e^{\lambda-1} = 1,$$

alors

$$e^{\lambda-1}\binom{N}{n} = 1,$$

ce qui conduit à

$$e^{\lambda-1} = \frac{1}{\binom{N}{n}}.$$

On a donc

$$p(s_t) = \frac{1}{\binom{N}{n}} \text{ pour } s_t \in \Omega,$$

qui correspond au plan aléatoire simple sans remise. □

Remarque 1.3.1. *Dans la classe des plans à probabilités inégales avec le jeu de probabilités $\pi_1, ..., \pi_N$, le plan de sondage maximisant l'entropie est le plan de Poisson. Dans la classe des plans de sondage à probabilités inégales à taille fixe avec le jeu de probabilités π_1,...,π_N, le plan à entropie maximale est le plan de Poisson conditionnel.*

Remarque 1.3.2. *Le plan de Rao-Sampford, le plan systématique randomisé et le plan de Poisson séquentiel font partie de la classe des plans à grande entropie. On dira qu'un plan de sondage à taille fixe $p(.)$ est à grande entropie si son entropie est proche de celle du plan de Poisson conditionel (CPS) lorsque $\sum_{i\in U} \pi_i(1-\pi_i) \to \infty$. Plus précisément, un plan de sondage à taille fixe est à grande entropie si la divergence du Kullback-Leibler, $K(p, p_{CPS})$,*

$$K(p, p_{CPS}) = \sum_{s\in\Omega} \log p(s) \log\left(\frac{p(s)}{p_{CPS}(s)}\right)$$

tend vers 0 lorsque $\sum_{i\in U} \pi_i(1-\pi_i) \to \infty$; voir Berger (1998).

1.4. Approximation de la variance

Les estimateurs de la variance (1.1.6) et (1.1.8) présentent deux inconvénients importants. En effet, pour certains plans de sondage et lorsque la taille de l'échantillon est grande, il peut être très complexe, voire impossible, de calculer les probabilités d'inclusion d'ordre deux, π_{ij}. De plus, les estimateurs de la variance s'expriment comme une double somme, ce qui peut conduire à des estimateurs de variance instables. Dans le cas des plans de sondage à grande entropie, il est possible d'approximer les probabilités d'inclusion d'ordre deux en fonction des probabilités d'inclusion d'ordre un, qui elles, sont toujours disponibles. Plusieurs

approximations des π_{ij} ont été proposées dans la litérature. Certaines d'entres elles sont présentées dans cette section.

1.4.1. Approximation de Hájek

Une approximation des probabilités d'inclusion d'ordre deux bien connue est celle de Hájek :

$$\pi_{ij} - \pi_i\pi_j \approx -\frac{\pi_i(1-\pi_i)\pi_j(1-\pi_j)}{D}, \tag{1.4.1}$$

où $D = \sum_{i\in U} \pi_i(1-\pi_i)$. En insérant l'approximation (1.4.1) dans la forme Sen-Yates-Grundy de la variance donnée par (1.1.7), on obtient une expression de la variance approximative, notée V_{HA}.

Proposition 1.4.1. *En insérant (1.4.1) dans (1.1.7), on obtient*

$$\mathrm{V}_{HA} = \sum_{i\in U} \pi_i(1-\pi_i)\left\{\frac{y_i}{\pi_i} - \frac{1}{D}\sum_{j\in U}(1-\pi_j)y_j\right\}^2. \tag{1.4.2}$$

Démonstration. On a

$$\begin{aligned}
\mathrm{V}_p(\widehat{Y}_{HT}) &\approx \frac{1}{2}\sum_{i\in U}\sum_{j\in U}\frac{\pi_i(1-\pi_i)\pi_j(1-\pi_j)}{D}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \\
&= \frac{1}{2D}\sum_{i\in U}\sum_{j\in U}\pi_i(1-\pi_i)\pi_j(1-\pi_j)\left(\frac{y_i^2}{\pi_i^2} + \frac{y_j^2}{\pi_j^2} - 2\frac{y_iy_j}{\pi_i\pi_j}\right) \\
&= \sum_{i\in U}(1-\pi_i)\frac{y_i^2}{\pi_i} - \frac{1}{D}\sum_{i\in U}(1-\pi_i)y_i\sum_{j\in U}(1-\pi_j)y_j \\
&= \sum_{i\in U}\pi_i(1-\pi_i)\left\{\frac{y_i^2}{\pi_i^2} - \frac{1}{D}\frac{y_i}{\pi_i}\sum_{j\in U}(1-\pi_j)y_j\right\} \\
&= \sum_{i\in U}\pi_i(1-\pi_i)\left\{\frac{y_i}{\pi_i} - \frac{1}{D}\sum_{j\in U}(1-\pi_j)y_j\right\}^2.
\end{aligned}$$

□

1.4.2. Approximations de Brewer et Donadio

Brewer et Donadio (2003) ont considéré plusieurs approximations des π_{ij}. Toutes les approximations s'expriment sous la forme

$$\pi_{ij} - \pi_i\pi_j \approx \frac{1}{2}\pi_i\pi_j(c_i + c_j - 2), \tag{1.4.3}$$

où c_i est une constante choisie.

Proposition 1.4.2. *En insérant (1.4.3) dans (1.1.7), on obtient*

$$\mathrm{V}_B = \sum_{i\in U} \pi_i(1-\pi_i c_i)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2. \tag{1.4.4}$$

Démonstration. On peut écrire (1.1.7) comme

$$\begin{aligned}\mathrm{V}_p(\widehat{Y}_{HT}) &= \frac{1}{2}\sum_{i\in U}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \\ &= \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2 \\ &\quad - \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)\left(\frac{y_j}{\pi_j} - \frac{Y}{n}\right). \end{aligned} \tag{1.4.5}$$

En utilisant la Proposition 1.1.2 dans le premier terme de (1.4.5), on a que

$$\begin{aligned}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2 &= \sum_{i\in U}\pi_i(n-\pi_i)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2 - \sum_{i\in U}\pi_i(n-1)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2 \\ &= \sum_{i\in U}(\pi_i - \pi_i^2)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2. \end{aligned} \tag{1.4.6}$$

En insérant l'approximation (1.4.3) dans le deuxième terme de (1.4.5), on obtient

$$\begin{aligned}\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)\left(\frac{y_j}{\pi_j} - \frac{Y}{n}\right) &\approx -\sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\pi_i\pi_j\frac{2-c_i-c_j}{2}\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)\left(\frac{y_j}{\pi_j} - \frac{Y}{n}\right) \\ &= \sum_{i\in U}\pi_i^2(1-c_i)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2. \end{aligned} \tag{1.4.7}$$

En additionnant (1.4.6) et (1.4.7), on obtient

$$\mathrm{V}_p(\widehat{Y}_{HT}) \approx \sum_{i\in U}\pi_i(1-\pi_i c_i)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2.$$

□

Brewer et Donadio (2003) considèrent plusieurs choix de c_i :

(i) $c_i = \frac{n-1}{n-\pi_i}$;

(ii) $c_i = \frac{n-1}{n-\frac{1}{n}\sum_{i\in U}\pi_i^2}$;

(iii) $c_i = \frac{n-1}{n-2\pi_i-\frac{1}{n}\sum\limits_{i\in U}\pi_i^2}$;

(iv) $c_i = \frac{n-1}{n}\left(1+\frac{2\pi_i}{n}-\frac{1}{n^2}\sum\limits_{i\in U}\pi_i^2\right)$.

1.4.3. Estimateur par calage dans un plan de Poisson

Dans le cas d'un plan de Poisson séquentiel décrit dans la section 1.2.6, Ohlsson (1998) a proposé d'approximer la stratégie consistant en un plan de Poisson séquentiel et l'estimateur d'Horvitz-Thompson par la stratégie consistant en un plan de Poisson et l'estimateur par calage suivant :

$$\widehat{Y}_c = \frac{n}{n_s}\widehat{Y}_{HT} = \frac{\sum\limits_{i\in U}\pi_i}{\sum\limits_{i\in s}\frac{1}{\pi_i}\pi_i}\widehat{Y}_{HT}. \tag{1.4.8}$$

L'estimateur (1.4.8) fait bien partie de la classe des estimateurs par calage car si l'on remplace y_i par π_i dans (1.4.8), on a $\widehat{Y}_c = \sum_{i\in U}\pi_i$. Autrement dit, l'estimateur (1.4.8) est calé sur le total des π_i dans la population. L'estimateur (1.4.8) étant un estimateur de type ratio, il est biaisé. Par contre, son biais est asymptotiquement négligeable lorsque n est grand. De plus, sa variance ne s'obtient pas facilement, auquel cas on a recours à un développement par séries de Taylor du premier ordre. Sous certaines conditions de régularité, on a

$$\begin{aligned}
\widehat{Y}_c &= Y + \frac{\partial\,\widehat{Y}_c}{\partial\,\widehat{Y}_{HT}}\bigg|_{\substack{\widehat{Y}_{HT}=Y\\ n_s=n}}(\widehat{Y}_{HT}-Y) + \frac{\partial\,\widehat{Y}_c}{\partial\,n_s}\bigg|_{\substack{\widehat{Y}_{HT}=Y\\ n_s=n}}(n_s-n) + O_p\left(\frac{1}{n}\right)\\
&= Y + \frac{n}{n_s}\bigg|_{\substack{\widehat{Y}_{HT}=Y\\ n_s=n}}(\widehat{Y}_{HT}-Y) - \frac{\widehat{Y}_{HT}}{n_s^2}n\bigg|_{\substack{\widehat{Y}_{HT}=Y\\ n_s=n}}(n_s-n) + O_p\left(\frac{1}{n}\right)\\
&= Y + \sum_{i\in s}\frac{z_i}{\pi_i} + O_p\left(\frac{1}{n}\right),
\end{aligned}$$

où $z_i = y_i - R\pi_i$ et $R = \frac{\sum_{i\in U} y_i}{\sum_{i\in U}\pi_i} = \frac{Y}{n}$. En négligeant les termes d'ordre supérieur (ce qui est approprié lorsque la taille de l'échantillon est suffisamment grande), on peut approximer la variance de l'estimateur (1.4.8) par (1.1.5) en remplaçant y_i par z_i, ce qui conduit à

$$\begin{aligned}
\mathrm{V}_p(\widehat{Y}_c) &= \mathrm{V}_p(\widehat{Y}_c - Y)\\
&\approx \mathrm{V}_p\left(\sum_{i\in s}\frac{z_i}{\pi_i}\right)\\
&= \sum_{i\in U}\sum_{j\in U}\Omega_{ij}\,z_i\,z_j.
\end{aligned}$$

Proposition 1.4.3. *Pour un plan de Poisson, on peut approximer la variance de l'estimateur (1.4.8) par*

$$\mathrm{V}_p\left(\widehat{Y}_c\right) \approx \sum_{i\in U} \pi_i(1-\pi_i)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2. \tag{1.4.9}$$

Démonstration.

$$\begin{aligned}
\mathrm{V}_p\left(\widehat{Y}_c\right) &\approx \sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j)\frac{z_i}{\pi_i}\frac{z_j}{\pi_j} \\
&= \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}(\pi_i\pi_j-\pi_i\pi_j)\frac{z_i}{\pi_i}\frac{z_j}{\pi_j} + \sum_{i\in U}(\pi_i-\pi_i^2)\frac{z_i^2}{\pi_i^2} \\
&= \sum_{i\in U}\frac{(1-\pi_i)}{\pi_i}z_i^2 \\
&= \sum_{i\in U}\pi_i(1-\pi_i)\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2,
\end{aligned}$$

en notant que $R = Y/n$. □

Ohlsson (1998) propose d'approximer la variance de $\widehat{Y}_{HT}$ dans le cas d'un plan de poisson séquentiel par (1.4.9). Le plan de Poisson séquentiel étant un plan à grande entropie, on conjecture que la méthode d'Ohlsson (1998) peut s'appliquer aux autres plans à probabilités inégales ayant une grande entropie (par exemple, le plan de Poisson conditionnel, le plan de Rao-Sampford et le plan systématique randomisé).

Remarque 1.4.1. *En posant $c_i = 1$ dans (1.4.4), on retrouve (1.4.9). Autrement dit, l'approximation (1.4.9) est obtenue en approximant π_{ij} par $\pi_i\pi_j$; voir l'expression (1.4.3).*

1.4.4. Forme générale

Il est intéressant de noter que les variances approximatives (1.4.2), (1.4.4) et (1.4.9) peuvent toutes s'écrire sous une forme générale :

$$\mathrm{V}_{APP}\left(\widehat{Y}_{HT}\right) = \sum_{i\in U}\phi_i\left(\frac{y_i}{\pi_i} - B\right)^2, \tag{1.4.10}$$

où

$$B = \frac{\sum\limits_{i\in U} a_i y_i}{\sum\limits_{i\in U} a_i\pi_i}$$

et ϕ_i et a_i sont des coefficients donnés. En spécifiant les valeurs ϕ_i et a_i de manière appropriée, on retrouve les trois approximations de la variance (1.4.2), (1.4.4) et (1.4.9), (voir Tableau 1.2).

TABLEAU 1.2. Composantes ϕ_i et a_i des trois approximations de la variance

Coefficients	**Expression** (1.4.2)	**Expression** (1.4.4)	**Expression** (1.4.9)
ϕ_i	$\pi_i(1-\pi_i)$	$\pi_i(1-c_i\pi_i)$	$\pi_i(1-\pi_i)$
a_i	$(1-\pi_i)$	1	1

La forme commune (1.4.10) présente plusieurs avantages et permet de régler les inconvénients de la variance (1.1.5) :

(i) l'utilisation des probabilités d'inclusion d'ordre deux π_{ij} n'est plus nécessaire ;

(ii) la variance (1.4.10) consiste en une simple sommation.

Examinons maintenant le comportement de la forme générale (1.4.10) pour les approximations (1.4.2), (1.4.4) et (1.4.9), pour un plan aléatoire simple sans remise, c'est-à-dire lorsque la probabilité d'inclusion d'ordre un vaut $\pi_i = n/N$. En insérant $\pi_i = n/N$ dans (1.4.10), on trouve

$$\begin{aligned} \mathrm{V}_{APP}\left(\widehat{Y}_{HT}\right) &= \sum_{i\in U}\phi_i\left(\frac{N}{n}y_i - \frac{\sum\limits_{i\in U}a_iy_i}{\sum\limits_{i\in U}a_i}\right)^2 \\ &= \frac{N^2}{n^2}\sum_{i\in U}\phi_i\left(y_i - \frac{\sum\limits_{i\in U}a_iy_i}{\sum\limits_{i\in U}a_i}\right)^2. \end{aligned} \tag{1.4.11}$$

Peu importe le choix de a_i (soit 1, soit $1-\pi_i = 1-n/N$), on a

$$\frac{\sum\limits_{i\in U}a_iy_i}{\sum\limits_{i\in U}a_i} = \overline{Y}, \tag{1.4.12}$$

la moyenne de la variable dans la population. En insérant (1.4.12) dans (1.4.11), on trouve

$$\mathrm{V}_{APP}\left(\widehat{Y}_{HT}\right) = \frac{N^2}{n^2}\sum_{i\in U}\phi_i\left(y_i-\overline{Y}\right)^2. \tag{1.4.13}$$

Lorsque

$$\phi_i = \pi_i(1-\pi_i) = \frac{n}{N}\left(1-\frac{n}{N}\right),$$

on a (voir Tableau 1.2)

$$\begin{aligned} \mathrm{V}_{APP}\left(\widehat{Y}_{HT}\right) &= \frac{N^2}{n^2}\frac{n}{N}\left(1-\frac{n}{N}\right)\frac{N-1}{N-1}\sum_{i\in U}\left(y_i-\overline{Y}\right)^2 \\ &= \frac{N-1}{N}\left\{N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}\right\}, \end{aligned} \tag{1.4.14}$$

où

$$S_y^2 = \frac{1}{N-1}\sum_{i\in U}\left(y_i - \overline{Y}\right)^2.$$

Le terme entre accolades dans (1.4.14) représente la variance de $\widehat{Y}_{HT}$ dans le cas d'un plan aléatoire simple sans remise. Lorsque $\phi_i = \pi_i(1-c_i\pi_i)$ et que $c_i = \frac{n-1}{n-\pi_i}$, on obtient

$$\begin{aligned}\phi_i &= \frac{n}{N}\left(1 - \frac{n-1}{n-n/N}\frac{n}{N}\right)\\ &= \frac{n}{N-1}\left(1-\frac{n}{N}\right).\end{aligned}$$

Il découle que

$$\begin{aligned}\mathrm{V}_p\left(\widehat{Y}_{HT}\right) &= \frac{N^2}{n^2}\frac{n}{N-1}\left(1-\frac{n}{N}\right)\sum_{i\in U}\left(y_i-\overline{Y}\right)^2\\ &= N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}. \qquad (1.4.15)\end{aligned}$$

L'approximation de la variance (1.4.4) avec $c_i = \frac{n-1}{n-\pi_i}$ se simplifie alors exactement à la variance de l'EASSR. De plus, en négligeant le facteur $(N-1)/N$ (ce qui est approprié lorsque la taille de la population est grande), les approximations de la variance (1.4.2) et (1.4.9) se simplifient aussi à la variance de l'estimateur de Horvitz-Thompson dans le cas du plan aléatoire simple sans remise.

Remarque 1.4.2. *Dans cette section, nous avons présenté certaines approximation des π_{ij} retrouvées dans la litérature. D'autres approximations ont été examinées dans la litérature ; voir Haziza et coll. (2008) pour une discussion de ces approximations.*

1.4.5. Estimation et autres approximations

Un estimateur général de la variance approximative (1.4.10) est

$$\widehat{\mathrm{V}}_{Gen} = \sum_{i\in s}\varphi_i\left(\frac{y_i}{\pi_i} - \widehat{B}\right)^2, \qquad (1.4.16)$$

où

$$\widehat{B} = \frac{\sum_{i\in s}\alpha_i\frac{y_i}{\pi_i}}{\sum_{i\in s}\alpha_i}$$

et φ_i et α_i sont des coefficients choisis en fonction de l'approximation des probabilités π_{ij} utilisée. Notons que ces coefficients sont différents des coefficients ϕ_i et a_i de l'approximation (1.4.10). Les approximations des probabilités π_{ij} vues dans cette section mènent à l'estimateur de la variance (1.4.16). Pour les autres

approximations existantes, le tableau 1.3 présente les coefficients φ_i et α_i correspondant à ces approximations.

L'estimateur (1.4.16) présente les mêmes avantages que la forme générale de la variance (1.4.10). En effet, l'estimateur (1.4.16) ne requiert pas les probabilités d'inclusion d'ordre deux, il s'exprime comme une simple somme. De plus, le coefficient φ_i étant toujours positif (voir Tableau 1.3), l'estimateur $\widehat{\mathrm{V}}_{Gen}$ est toujours positif. De plus, lorsque l'on utilise la probabilité d'inclusion d'ordre un $\pi_i = n/N$, l'estimateur (1.4.16) se simplifie pour donner

$$\widehat{\mathrm{V}} = N^2 \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n},$$

avec

$$s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y})^2,$$

qui est l'estimateur usuel de la variance dans le cas de l'échantillonnage aléatoire simple sans remise.

1.5. Estimateur du total de Hájek

Lorsque la variable d'intérêt n'est pas proportionnelle aux probabilités d'inclusion, l'estimateur d'Horvitz-Thompson (1.1.4) peut s'avérer grandement inefficace. En effet, considérons le cas où $y_i = c$ pour tout $i \in U$. L'estimateur d'Horvitz-Thompson conduit à

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = c \sum_{i \in s} \frac{1}{\pi_i} = c\hat{N}_{HT},$$

où $\hat{N}_{HT} = \sum_{i \in s} \frac{1}{\pi_i}$ est un estimateur de type Horvitz-Thompson de la taille de la population N. On a alors $\mathrm{V}_p(\widehat{Y}_{HT}) > 0$. Un estimateur réagissant beaucoup mieux à cette situation est celui de Hájek :

$$\hat{Y}_{HA} = N \frac{\hat{Y}_{HT}}{\hat{N}_{HT}}, \tag{1.5.1}$$

Dans le cas $y_i = c$, on a bien $\hat{Y}_{HA} = Y$ pour tout $s \subseteq \Omega$. Il en découle que $\mathrm{V}_p(\widehat{Y}_{HA}) = 0$, ce qui est préférable à l'estimateur (1.1.4).

L'estimateur (1.5.1) étant un estimateur de type ratio, son biais est asymptotiquement nul. On approximera sa variance à l'aide d'un développement par séries de Taylor du premier ordre. On obtient

$$\widehat{Y}_{HA} - Y = \sum_{i \in s} \frac{E_i}{\pi_i} + O_p\left(\frac{1}{n}\right),$$

TABLEAU 1.3. Coefficients φ_i et α_i pour différentes approximations

Estimateur	**Notation**	φ_i	α_i
Berger	$\widehat{\mathrm{V}}_B$	$\frac{n}{n-1}(1-\pi_i)\frac{\sum_{j\in s}(1-\pi_j)}{\sum_{j\in U}\pi_j(1-\pi_j)}$	φ_i
Brewer 1	$\widehat{\mathrm{V}}_{B1}$	$\frac{n}{n-1}(1-\pi_i)$	1
Brewer 2	$\widehat{\mathrm{V}}_{B2}$	$\frac{n}{n-1}\left(1-\pi_i+\frac{\pi_i}{n}+\frac{1}{n^2}\sum_{k\in U}\pi_k^2\right)$	1
Brewer 3	$\widehat{\mathrm{V}}_{B3}$	$\frac{n}{n-1}\left(1-\pi_i-\frac{\pi_i}{n}+\frac{1}{n^2}\sum_{k\in U}\pi_k^2\right)$	1
Brewer 4	$\widehat{\mathrm{V}}_{B4}$	$\frac{n}{n-1}\left(1-\pi_i-\frac{\pi_i}{n-1}+\frac{1}{n(n-1)}\sum_{k\in U}\pi_k^2\right)$	1
Deville 1	$\widehat{\mathrm{V}}_{D1}$	$(1-\pi_i)\left[1-\sum_{i\in s}\left\{\frac{(1-\pi_j)}{\sum_{k\in s}\pi_k(1-\pi_k)}\right\}^2\right]^{-1}$	φ_i
Deville 2	$\widehat{\mathrm{V}}_{D2}$	$(1-\pi_i)\left[1-\sum_{i\in s}\left\{\frac{(1-\pi_j)}{\sum_{k\in s}\pi_k(1-\pi_k)}\right\}^2\right]^{-1}$	1
Hajek	$\widehat{\mathrm{V}}_H$	$\frac{n}{n-1}(1-\pi_i)$	φ_i
Hartley-Rao	$\widehat{\mathrm{V}}_{HR}$	$\frac{n}{n-1}\left(1-\pi_i-\frac{1}{n}\sum_{k\in s}\pi_k+\frac{1}{n}\sum_{k\in U}\pi_k^2\right)$	1
Poisson	$\widehat{\mathrm{V}}_{PO}$	$(1-\pi_i)$	1
Rosen	$\widehat{\mathrm{V}}_R$	$\frac{n}{n-1}(1-\pi_i)$	$\frac{(1-\pi_i)\log(1-\pi_i)}{\pi_i}$

où $E_i = y_i - \overline{Y}$. En négligeant les termes d'ordre supérieur (ce qui est approprié lorsque la taille de l'échantillon est suffisamment grande) on peut approximer la variance de l'estimateur (1.5.1) comme suit :

$$\begin{aligned}\mathrm{V}_p(\widehat{Y}_{HA}) &= \mathrm{V}_p(\widehat{Y}_{HA} - Y)\\ &\approx \mathrm{V}_p\left(\sum_{i\in s}\frac{E_i}{\pi_i}\right)\\ &= \sum_{i\in U}\sum_{j\in U}\Omega_{ij}\,E_i\,E_j.\end{aligned}$$

Pour un plan à taille fixe ou aléatoire, on peut estimer cette variance par

$$\widehat{\mathrm{V}}^*_{HT} = \sum_{i\in s}\sum_{j\in s}\frac{\Omega_{ij}}{\pi_{ij}}\,\widehat{E}_i\,\widehat{E}_j, \tag{1.5.2}$$

où $\widehat{E}_i = y_i - \widehat{Y}_{HT}/\widehat{N}_{HT}$. Notons que l'on pourrait remplacer $\widehat{N}_{HT}$ par N dans l'expression de $\widehat{E}_i$. Pour un plan à taille fixe, on peut également utiliser l'estimateur de variance

$$\widehat{\mathrm{V}}^*_{SYG} = -\frac{1}{2}\sum_{i\in s}\sum_{j\in s}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{\widehat{E}_i}{\pi_i}-\frac{\widehat{E}_j}{\pi_j}\right)^2. \tag{1.5.3}$$

Encore une fois, on fait face aux mêmes problèmes : d'une part, les estimateurs de variance (1.5.2) et (1.5.3) requièrent les probabilités d'inclusion d'ordre deux et, d'autre part, ils s'expriment comme une double somme. Les estimateurs (1.5.2) et (1.5.3) peuvent être vus comme des estimateurs de variance correspondant à l'estimateur d'Horvitz-Thompson (1.1.4) de la variable d'intérêt $\widehat{E}$. La forme générale de la variance et les approximations des probabilités π_{ij} présentées à la section 1.4 peuvent donc être utilisées. Il suffit d'utiliser la variable d'intérêt $\widehat{E}$ au lieu de la variable y. Ainsi, l'estimateur approximatif de la variance dans le cas de $\widehat{Y}_{HA}$ s'écrit comme

$$\widehat{\mathrm{V}}^*_{Gen} = \sum_{i\in s}\varphi_i\left(\frac{\widehat{E}_i}{\pi_i}-\widehat{B}^*\right)^2, \tag{1.5.4}$$

où

$$\widehat{B}^* = \frac{\sum\limits_{i\in s}\alpha_i\frac{\widehat{E}_i}{\pi_i}}{\sum\limits_{i\in s}\alpha_i}$$

et φ_i et α_i sont les coefficients choisis en fonction de l'approximation des probabilités π_{ij} utilisée ; voir Tableau 1.3.

Chapitre 2

LA NON-RÉPONSE

2.1. Introduction à la non-réponse

La non-réponse est un phénomène inévitable dans les enquêtes et elle peut avoir des effets indésirables dans le cadre d'une inférence statistique. On distingue deux types de non-réponse, soient la non-réponse totale et la non-réponse partielle.

La non-réponse totale est l'absence de réponse à toutes les variables pour un individu. Ce type de non-réponse survient lorsque, par exemple, un interviewer ne parvient pas à établir le contact avec un individu sélectionné dans l'échantillon, ou encore, lorsque ce dernier décide de ne pas répondre à l'enquête. La non-réponse totale est généralement traitée par repondération. Dans ce cas, les individus non-répondants sont éliminés du jeu de données et le poids des individus répondants est augmenté afin de compenser pour l'élimination des non-répondants.

La non-réponse partielle est l'absence de réponse à certains items (mais pas tous) chez un individu. L'absence de réponse peut être causée par plusieurs facteurs. Par exemple, le répondant pourrait avoir mal compris une question ou la trouver indiscrète et ne pas vouloir y répondre. Des valeurs manquantes peuvent également résulter d'une incohérence dans les données. Par exemple, une personne pourrait indiquer que son état civil est «marié» et indiquer que son âge est «9 ans». Dans les enquêtes de Statistique Canada, un tel enregistrement est qualifié d'incohérent puisqu'il faut avoir au moins 15 ans pour être éligible au statut «marié». Dans ce cas, au moins une des deux valeurs observées (âge et état civil) sera éliminée et imputée. Un autre exemple de valeur inutilisable est une donnée irréaliste. Par exemple, une personne pourrait indiquer dans une enquête qu'elle est âgée de 200 ans. Ce genre de valeur erronnée est généralement éliminée du jeu de données et remplacée par une valeur manquante. Les valeurs manquantes d'un jeu de données sont généralement traitées par imputation.

L'imputation consiste à affecter une valeur artificielle à un individu ayant une valeur manquante à un item. L'imputation permet d'avoir un jeu de données complet. Elle peut permettre de diminuer le biais de non-réponse si les valeurs imputées sont proches des valeurs réelles. Plusieurs méthodes d'imputation sont utilisées en pratique. Certaines sont présentées à la section 2.3. Notons que les valeurs imputées, étant artificielles, peuvent donner une fausse impression de précision. À l'étape de l'inférence statistique et plus particulièrement à l'étape de l'estimation de la variance, il est important de ne pas traiter les données imputées comme des données observées. Les estimateurs du total Y et de variance présentés au Chapitre 1 dans le contexte d'un jeu de données complet doivent être adaptés pour prendre en compte la non-réponse.

2.1.1. Effets de la non-réponse sur les estimations

La non-réponse peut avoir plusieurs effets négatifs sur les estimations. Elle occasionne généralement un biais appelé biais de non-réponse. Ce dernier peut être particulièrement important lorsqu'il y a une grande différence entre les répondants et les non-répondants en termes des variables d'intérêt. Par exemple, une personne avec un salaire élevé peut être moins portée à indiquer son revenu dans une enquête qu'une personne avec un salaire plus modeste. Ainsi, les non-répondants à cette enquête auront majoritairement des salaires élevés. La moyenne des répondants, comme estimateur de la moyenne de la population, exhibera donc un biais négatif. Un autre effet de la non-réponse concerne la diminution de la taille de l'échantillon. Par exemple, supposons que l'on tire un échantillon s de taille $n = 1000$ dans une population U de taille $N = 5000$ et que l'on observe un taux de réponse de 50%. Le nombre de répondants est donc égal à $n_r = 500$. Cette diminution de la taille effective entraînera une augmentation de la variance des estimateurs. Cette variance additionnelle est appelée variance due à la non-réponse.

2.1.2. Causes de la non-réponse et approches pour la réduire

Plusieurs facteurs ont des effets sur le taux de non-réponse, comme le type d'unité échantillonnée. En effet, les enquêtes effectuées dans des établissements ont généralement des taux de réponse plus faibles que les enquêtes individuelles. L'importance du sujet de l'enquête pour l'unité échantillonnée affecte aussi sa probabilité de réponse. De plus, le type d'enquête peut jouer un rôle important. Ainsi, une enquête obligatoire aura un taux de réponse beaucoup plus élevé qu'une enquête optionnelle. Une récompense suite à une réponse incite aussi les unités

à répondre. Enfin, une agence d'enquête réputée aura aussi un taux de réponse plus élevé qu'une agence moins connue.

La conception du questionnaire est une étape ayant un impact important sur le taux de non-réponse. En effet, un questionnaire simple à remplir, court, intéressant et facile à comprendre encourage les personnes à répondre et ainsi diminue la non-réponse. De plus, éviter les répétitions et les questions ouvertes aide à garder l'intérêt du répondant. Une formulation claire des questions est aussi importante. Bref, il est possible d'améliorer la forme d'un questionnaire afin d'augmenter le taux de réponse.

La méthode de collecte de données influence beaucoup le taux de réponse. Dans le cas d'une entrevue en face à face, l'interviewer peut développer des techniques pour mettre le répondant à l'aise et l'aider à remplir correctement le questionnaire de l'enquête. Dans le cas des entrevues téléphoniques, l'interviewer peut, avec une bonne formation et de bonnes techniques, convaincre les personnes de prendre le temps de répondre. Les questionnaires envoyés aux unités échantillonnales, quant à eux, occasionnent plus de non-réponse étant donné l'absence d'interviewer et le fait que le répondant remplisse le questionnaire seul. Il est possible d'utiliser plusieurs méthodes de collecte des données. Par exemple, une enquête pourrait commencer par un questionnaire à remplir soi-même, étant donné que les coûts sont moins élevés. Ensuite, les non-répondants pourraient être contactés personnellement dans le but d'augmenter le taux de réponse. Le temps de la collecte des données est aussi important. Un interviewer devrait avoir le temps de tenter de contacter une unité à plusieurs reprises et de faire varier les moments de la journée auxquels il la contacte, pour ainsi augmenter les chances d'obtenir une réponse.

2.2. Estimateur du total en présence de données imputées

En présence de non-réponse partielle, l'échantillon sélectionné est incomplet. Les valeurs manquantes sont alors imputées et les estimateurs du total Y et de variance présentés au Chapitre 1, dans le contexte d'un jeu de données complet, doivent être adaptés. Pour ce faire, nous commençons par introduire la notation qui sera utilisée dans le reste de ce travail.

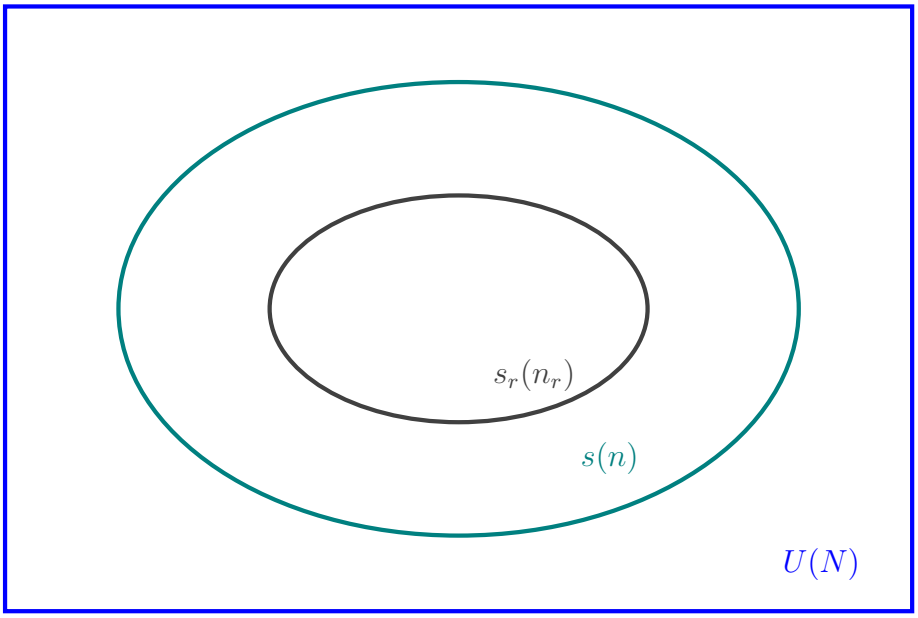


FIGURE 2.1. Représentation de la non-réponse

2.2.1. Contexte de non-réponse

L'ensemble des répondants peut être vu comme un sous-échantillon s_r, de taille aléatoire n_r, de l'échantillon sélectionné s. La population U, l'échantillon s et l'ensemble des répondants s_r sont représentés à la figure 2.1.

À chaque individu de l'échantillon s correspond une variable indicatrice de réponse définie par

$$r_i = \begin{cases} 1 & \text{si } i \text{ est répondant à la variable } y, \\ 0 & \text{sinon.} \end{cases}$$

La probabilité que l'individu i réponde est donnée par

$$p_i = \mathrm{P}(r_i = 1 | i \in s).$$

La probabilité que les individus i et j répondent est donnée par

$$p_{ij} = \mathrm{P}(r_i = 1, r_j = 1 | i \in s, j \in s, i \neq j).$$

Dans ce qui suit, on supposera que les unités répondent indépendamment les unes des autres, i.e., $p_{ij} = p_i p_j$, $i \neq j$. Le mécanisme de réponse peut donc être modélisé par des épreuves de Bernoulli de paramètre (inconnu) p_i.

On parlera d'un mécanisme de réponse uniforme lorsque la probabilité de réponse est constante dans la population, i.e., $p_i = p_0$ pour tout i. On parlera d'un mécanisme de réponse non-uniforme lorsque la probabilité de réponse varie d'un individu à l'autre. La probabilité de réponse peut dépendre de variables

auxiliaires observées pour tous les individus de l'échantillon ou peut dépendre de la variable d'intéret que l'on cherche à imputer.

2.2.2. Estimateur imputé

Soit y_i^* la valeur imputée à la valeur manquante y_i. L'estimateur du total Y en présence de valeurs imputées est

$$\begin{aligned}\widehat{Y}_I &= \sum_{i\in s} r_i \frac{y_i}{\pi_i} + \sum_{i\in s}(1-r_i)\frac{y_i^*}{\pi_i} \\ &= \sum_{i\in s} \frac{\widetilde{y}_i}{\pi_i}, \end{aligned} \tag{2.2.1}$$

où

$$\widetilde{y}_i = r_i y_i + (1-r_i) y_i^*.$$

L'estimateur (2.2.1) est donc un estimateur de la forme (1.1.4), où y_i est remplacée par $\widetilde{y}_i$.

2.3. Méthodes d'imputation

On distingue deux groupes de méthodes d'imputation, soient les méthodes d'imputation déterministe et les méthodes d'imputation aléatoire. Pour les méthodes déterministes, les valeurs imputées restent inchangées si l'on répète le procédé d'imputation sur un échantillon donné. Des méthodes déterministes fréquemment utilisées en pratique sont l'imputation par la régression, l'imputation par le ratio, l'imputation par la moyenne, l'imputation par le plus proche voisin et l'imputation historique. Ces méthodes sont détaillées dans la section 2.3. Dans le cas des méthodes aléatoires, une composante aléatoire est ajoutée à chaque valeur imputée. Dans ce cas, si on répète la méthode d'imputation sur un échantillon donné, on ne retombera pas nécessairement sur les mêmes valeurs imputées à cause de la présence de la composante aléatoire. Un exemple de méthode d'imputation aléatoire est la méthode du hot-deck aléatoire. Notons qu'une méthode d'imputation déterministe peut être transformée en méthode aléatoire en ajoutant simplement une composante aléatoire à chaque valeur imputée.

Les méthodes déterministes sont appropriées lorsque le paramètre que l'on cherche à estimer est le total ou la moyenne de la population, pourvu que le modèle d'imputation soit correctement spécifié. En contrepartie, les méthodes déterministes tendent à distordre la distribution de la variable que l'on cherche à imputer. Par conséquent, de telles méthodes peuvent potentiellement mener à des estimateurs de quantiles (par exemple, la médiane de la population) considérablement biaisés. Les méthodes aléatoires, quant à elles, tendent à préserver

la distribution de la variable que l'on cherche à imputer, et par conséquent, sont appropriées lorsqu'il s'agit d'estimater un quantile. En contrepartie, elles sont potentiellement inefficaces par rapport aux méthodes d'imputation déterministes car elles reposent sur une composante aléatoire additionnelle.

La majorité des méthodes d'imputation sont motivées par le modèle général

$$y_i = f(\boldsymbol{x_i}; \boldsymbol{\beta}) + \epsilon_i, \tag{2.3.1}$$

où $\boldsymbol{x}$ est un vecteur de variables auxiliaires disponible pour toutes les unités de l'échantillon (répondants et non-répondants), $\boldsymbol{\beta}$ est un vecteur de paramètres inconnus, ϵ_i est une variable aléatoire telle que $\mathrm{E}_m(\epsilon_i) = 0$, $\mathrm{E}_m(\epsilon_i \epsilon_j) = 0$ si $i \neq j$, $\mathrm{E}_m(\epsilon_i^2) = \sigma^2 c_i$, où c_i est une constante connue. Notons que $\mathrm{E}_m(.)$ est l'espérance par rapport au modèle d'imputation. Dans le cas des méthodes d'imputation déterministes, la valeur imputée y_i^* est donnée par

$$y_i^* = f(\boldsymbol{x_i}; \widehat{\boldsymbol{\beta}}_{\boldsymbol{r}}),$$

où $\widehat{\boldsymbol{\beta}}_{\boldsymbol{r}}$ est un estimateur de $\boldsymbol{\beta}$ obtenu au moyen des unités répondantes (par exemple l'estimateur du maximum de vraisemblance). Dans ce mémoire, on s'intéresse particulièrement aux méthodes d'imputation déterministes.

2.3.1. Imputation par la régression

L'imputation par la régression est adéquate lorsque la relation entre la variable d'intérêt y et les variables auxiliaires $\boldsymbol{x}$ est linéaire. Le modèle général d'imputation est réduit à

$$f(\boldsymbol{x_i}; \boldsymbol{\beta}) = \boldsymbol{x_i'}\boldsymbol{\beta}$$

et c_i est une constante connue. Le vecteur de paramètres, $\boldsymbol{\beta}$, est estimé, au moyen de l'ensemble des répondants, s_r, par l'estimateur des moindres carrés pondérés, $\widehat{\mathbf{B}}_{\boldsymbol{r}}$. On obtient alors la valeur imputée à l'unité i

$$y_i^* = \boldsymbol{x_i'}\widehat{\mathbf{B}}_{\boldsymbol{r}}, \tag{2.3.2}$$

où

$$\begin{aligned}\widehat{\mathbf{B}}_{\boldsymbol{r}} &= \left(\sum_{i \in s} \frac{r_i}{\pi_i} \boldsymbol{x_i} c_i^{-1} \boldsymbol{x_i'}\right)^{-1} \sum_{i \in s} \frac{r_i}{\pi_i} \boldsymbol{x_i} c_i^{-1} y_i \\ &= \widehat{\boldsymbol{T}}_r^{-1} \widehat{\boldsymbol{t}}_r\end{aligned}$$

avec

$$\widehat{\boldsymbol{T}}_r = \sum_{i \in s} \frac{r_i}{\pi_i} \boldsymbol{x_i} c_i^{-1} \boldsymbol{x_i'}$$

et

$$\widehat{\boldsymbol{t}}_r = \sum_{i \in s} \frac{r_i}{\pi_i} \boldsymbol{x_i} c_i^{-1} y_i.$$

2.3.2. Imputation par la régression linéaire simple

L'imputation par la régression linéaire simple est un cas particulier de l'imputation par la régression, avec $\boldsymbol{x}_i = (1, x_i)'$ et $c_i = 1$. On obtient la valeur imputée à l'unité i suivante :

$$y_i^* = \widehat{\beta}_{0r} + \widehat{\beta}_{1r} x_i, \tag{2.3.3}$$

où

$$\widehat{\beta}_{0r} = \overline{y}_r - \widehat{\beta}_{1r} \overline{x}_r$$

et

$$\widehat{\beta}_{1r} = \frac{\sum\limits_{i \in s} \frac{r_i}{\pi_i} (x_i - \overline{x}_r)(y_i - \overline{y}_r)}{\sum\limits_{i \in s} \frac{r_i}{\pi_i} (x_i - \overline{x}_r)^2},$$

avec $\overline{y}_r = \sum\limits_{i \in s} r_i \frac{y_i}{\pi_i} \Big/ \sum\limits_{i \in s} r_i \frac{1}{\pi_i}$ et $\overline{x}_r = \sum\limits_{i \in s} r_i \frac{x_i}{\pi_i} \Big/ \sum\limits_{i \in s} r_i \frac{1}{\pi_i}$.

2.3.3. Imputation par le ratio

L'imputation par le ratio est un cas particulier de l'imputation par la régression. Cette méthode est adéquate lorsqu'on est en présence d'une seule variable auxiliaire quantitative x et lorsque la relation entre y et x est linéaire et passe par l'origine. Le modèle général d'imputation est réduit à

$$f(\boldsymbol{x_i}; \boldsymbol{\beta}) = \beta x_i$$

avec $c_i = x_i$. On obtient la valeur imputée à l'unité i suivante :

$$y_i^* = \widehat{\mathrm{B}}_r x_i = \frac{\overline{y}_r}{\overline{x}_r} x_i. \tag{2.3.4}$$

2.3.4. Imputation par la moyenne

L'imputation par la moyenne est un autre cas particulier de l'imputation par la régression avec $x_i = 1$ pour tout i et $c_i = 1$. Cette méthode n'utilise pas de variable auxiliaire, sauf la variable auxiliaire valant 1 pour chaque individu. Elle est donc adéquate lorsqu'il n'y a pas de relation entre la variable y et les variables auxiliaires disponibles, ou bien lorsqu'il n'y a simplement pas de variable auxiliaire disponible. Elle consiste à remplacer toutes les valeurs manquantes par la moyenne

des répondants. Le modèle d'imputation général se réduit à

$$f(\boldsymbol{x_i};\boldsymbol{\beta}) = \beta$$

avec $c_i = 1$. On obtient donc comme valeur imputée

$$y_i^* = \widehat{\mathrm{B}}_r = \overline{y}_r. \tag{2.3.5}$$

2.3.5. Imputation par le plus proche voisin

L'imputation par le plus proche voisin est adéquate lorsque la relation entre la variable d'intérêt y et les variables auxiliaires $\boldsymbol{x}$ n'est pas linéaire. C'est une méthode non-paramétrique. Il n'est donc pas nécessaire de formuler des hypothèses à propos de la forme de la fonction $f(\boldsymbol{x}_i;\boldsymbol{\beta})$ et il en est de même pour la structure de variance. Il faut par contre choisir une fonction de distance entre l'unité i et l'unité j, $d(\boldsymbol{x_i},\boldsymbol{x_j})$. Une fois cette fonction choisie, la valeur imputée à l'unité i est

$$y_i^* = y_j,$$

pour l'unité $j \in s_r$ telle que la distance $d(\boldsymbol{x_i},\boldsymbol{x_j})$ est minimale. Cette méthode présente plusieurs avantages. En effet, l'imputation par le plus proche voisin tend à préserver la distribution de la variable d'intérêt. De plus, en remplaçant une valeur manquante par le plus proche voisin en termes des variables $\boldsymbol{x}$, on s'attend à ce que la relation entre y et $\boldsymbol{x}$, si elle existe, soit préservée. Finalement, cette méthode d'imputation utilise des valeurs observées, donc plausibles, ce qui est un avantage en pratique.

2.3.6. Imputation historique

L'imputation historique est adéquate lorsque la variable d'intérêt est connue à un temps antérieur. On s'intéresse à la variable y au temps t pour l'individu i, $y_{i,t}$. On remplace une valeur manquante par la valeur du même individu observée à une occasion précédente, $x_i = y_{i,t-1}$. Cette méthode d'imputation est adéquate seulement si la variable d'intérêt est stable au fil du temps. Autrement dit, il faut que la relation entre $y_{i,t}$ et $y_{i,t-1}$ soit linéaire, passe par l'origine avec une pente proche de 1. C'est une méthode souvent utilisée dans les enquêtes répétées. Le modèle général d'imputation est réduit à

$$f(\boldsymbol{x_i};\boldsymbol{\beta}) = y_{i,t-1}.$$

On obtient alors la valeur imputée

$$y_i^* = y_{i,t-1}.$$

2.3.7. Imputation hot-deck aléatoire

L'imputation hot-deck aléatoire utilise la valeur d'un donneur choisi au hasard dans l'ensemble des répondants afin de remplacer la valeur manquante d'un non-répondant. Cette méthode d'imputation peut être vue comme une imputation par la moyenne avec une composante aléatoire :

$$y_i^* = \overline{y}_r + \epsilon_i^*,$$

où ϵ_i^* est un résidu sélectionné aléatoirement parmi l'ensemble des résidus de l'ensemble des répondants, soit $\widetilde{e}_j = y_j - \overline{y}_r$, pour tout $j \in s_r$. On a donc

$$y_i^* = \overline{y}_r + (y_j - \overline{y}_r) = y_j \text{ pour un certain } j \in s_r.$$

Tout comme la méthode d'imputation par le plus proche voisin, la méthode hot-deck aléatoire préserve la distribution de la variable que l'on impute et conduit à des valeurs imputées observées, donc plausibles.

2.3.8. Erreur totale de l'estimateur imputé

Examinons l'impact du choix du modèle de la variable d'intérêt y. Introduisons d'abord $\mathrm{E}_r(.)$, l'espérance par rapport au mécanisme de réponse et rappelons que les notations $\mathrm{E}_p(.)$ et $\mathrm{E}_m(.)$ désignent l'espérance par rapport au plan de sondage et au modèle d'imputation respectivement. On peut décomposer l'erreur totale de $\widehat{Y}_I$ comme suit :

$$\widehat{Y}_I - Y = \left(\widehat{Y}_{HT} - Y\right) + \left(\widehat{Y}_I - \widehat{Y}_{HT}\right), \tag{2.3.6}$$

où $\widehat{Y}_{HT}$ est l'estimateur d'Horvitz-Thompson en (1.1.4) de l'échantillon complet s. La composante $\widehat{Y}_{HT} - Y$ représente l'erreur due à l'échantillonnage et la composante $\widehat{Y}_I - \widehat{Y}_{HT}$ représente l'erreur due à la non-réponse. Le biais de l'estimateur $\widehat{Y}_I$ est donné par

$$\begin{aligned}
\mathrm{Biais}(\widehat{Y}_I) &= \mathrm{E}(\widehat{Y}_I - Y) \\
&= \mathrm{E}_m\mathrm{E}_p\mathrm{E}_r(\widehat{Y}_I - Y) \\
&= \mathrm{E}_m\mathrm{E}_p\left(\widehat{Y}_{HT} - Y\right) + \mathrm{E}_m\mathrm{E}_p\mathrm{E}_r\left(\widehat{Y}_I - \widehat{Y}_{HT}\right) \\
&= \mathrm{E}_r\mathrm{E}_p\mathrm{E}_m\left(\widehat{Y}_I - \widehat{Y}_{HT}|s,s_r\right) \\
&= \mathrm{E}_r\mathrm{E}_p\left(\mathrm{B}_m\right),
\end{aligned} \tag{2.3.7}$$

où $\mathrm{B}_m = \mathrm{E}_m\left(\widehat{Y}_I - \widehat{Y}_{HT}|s,s_r\right)$ est le biais de non-réponse conditionnel. Notons que la troisième égalité découle du fait que $\widehat{Y}_{HT}$ est sans biais pour Y. L'estimateur

imputé (2.2.1) est sans biais pour Y si $\mathrm{B}_m = 0$. Le biais aura tendance à être petit si le modèle d'imputation (2.3.1) est correctement spécifié.

Il est important de noter que l'ordre des espérances a été interchangé dans les égalités menant à (2.3.7). Cette opération est correcte lorsque les données sont *Missing At Random* (*MAR*). Les données sont dites *MAR* lorsque la probabilité de réponse est indépendante du terme d'erreur du modèle d'imputation (2.3.1), après avoir pris en compte le vecteur de variables auxiliaires $\boldsymbol{x}$. Sinon, les données sont *Not Missing At Random* (*NMAR*) ; voir Rubin (1976).

Un exemple de situation *NMAR* survient lorsque la probabilité de réponse dépend d'une variable x, qui est également liée à y, mais qui ne fait pas partie du vecteur $\boldsymbol{x}$ dans le modèle d'imputation. Dans ce cas, il est clair que la probabilité de réponse est liée au terme d'erreur due au modèle d'imputation. Un autre exemple de données *NMAR* survient lorsque la probabilité de réponse dépend directement de la variable y que l'on cherche à imputer.

2.4. Limites de l'imputation

L'imputation présente plusieurs avantages. Comme mentionné précédemment, l'imputation permet d'avoir un jeu de données complet et permet d'utiliser l'information auxiliaire disponible pour les répondants et les non-répondants dans la construction de valeurs imputées, ce qui peut améliorer les estimations à l'étape de l'inférence statistique. Par contre, l'imputation comporte certaines limites. En effet, si la méthode d'imputation ne reflète pas correctement le lien existant entre la variable que l'on impute et l'information auxiliaire, elle peut conduire à des estimateurs considérablement biaisés. De plus, traiter les valeurs imputées comme si elles étaient des valeurs observées va généralement conduire à des inférences invalides due à la sous-estimation de la variance.

2.4.1. Respect des hypothèses

Lors de l'inférence statistique en présence de non-réponse, il faut faire des hypothèses à propos du modèle liant y aux variables auxilaires $\boldsymbol{x}$. Lorsque le modèle d'imputation ne reflète pas adéquatement le vrai lien entre y et $\boldsymbol{x}$, le biais de l'estimateur imputé peut être important. En guise d'illustration, nous examinons le biais pour différentes combinaisons de modèle pour la variable y et de méthode d'imputation utilisée.

Exemple 2.4.1. *Supposons que le vrai modèle soit donné par*

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i, \tag{2.4.1}$$

où $\mathrm{E}_m(\epsilon_i) = 0$, $\mathrm{E}_m(\epsilon_i\epsilon_j) = 0$, $i \neq j$ *et* $\mathrm{V}_m(\epsilon_i) = \sigma^2 c_i$.

Pour remplacer les valeurs manquantes, on utilise l'imputation par la régresion décrite à la section 2.3.1. L'estimateur imputé $\widehat{Y}_I$ peut s'écrire comme

$$\widehat{Y}_I = \sum_{i\in s} r_i \frac{y_i}{\pi_i} + \sum_{i\in s}(1-r_i)\frac{\boldsymbol{x}_i'\widehat{\mathbf{B}}_r}{\pi_i}.$$

L'erreur de non-réponse est donnée par

$$\widehat{Y}_I - \widehat{Y}_{HT} = -\sum_{i\in s}\frac{1}{\pi_i}(1-r_i)\left(y_i - \boldsymbol{x}_i'\widehat{\mathbf{B}}_r\right).$$

Dans ce cas, le biais de non-réponse est égal à

$$\begin{aligned}
\mathrm{B}_m &= \mathrm{E}_m\left(\widehat{Y}_I - \widehat{Y}_{HT}|s,s_r\right)\\
&= -\sum_{i\in s}\frac{1}{\pi_i}(1-r_i)\mathrm{E}_m\left(y_i - \boldsymbol{x}_i'\widehat{\mathbf{B}}_r|s,s_r\right)\\
&= 0,
\end{aligned}$$

en notant que, par rapport au modèle (2.4.1), $\mathrm{E}_m(y_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$ *et* $\mathrm{E}_m(\widehat{\mathbf{B}}_r|s,s_r) = \boldsymbol{\beta}$.

Exemple 2.4.2. *Nous étudions maintenant le biais de non-réponse si la méthode d'imputation est mal spécifiée. Supposons que le vrai modèle est donné par*

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

avec $\mathrm{V}_m(\epsilon_i) = \sigma^2$. *Ce modèle correspond au modèle de régression linéaire simple. Pour remplacer les valeurs manquantes, nous utilisons l'imputation par le ratio décrite à la section 2.3.3. On a*

$$y_i^* = \widehat{\mathrm{B}}_r x_i = \frac{\overline{y}_r}{\overline{x}_r}x_i.$$

Autrement dit, le modèle d'imputation ne contient pas l'ordonnée à l'origine. L'estimateur imputé s'écrit comme

$$\widehat{Y}_I = \frac{\widehat{Y}_r}{\widehat{X}_r}\widehat{X}_{HT},$$

où $\widehat{Y}_r = \sum\limits_{i\in s} r_i\frac{y_i}{\pi_i}$ *et* $\widehat{X}_r = \sum\limits_{i\in s} r_i\frac{x_i}{\pi_i}$. *Le biais de l'estimateur imputé est alors*

$$\begin{aligned}
\mathrm{B}_m &= \mathrm{E}_m\left(\widehat{Y}_I - \widehat{Y}_{HT}|s,s_r\right)\\
&= \mathrm{E}_m\left\{\sum_{i\in s}\frac{1}{\pi_i}\left(\frac{\widehat{X}_{HT}}{\widehat{X}_r}r_i - 1\right)y_i\middle|s,s_r\right\}\\
&= \sum_{i\in s}\frac{1}{\pi_i}\left(\frac{\widehat{X}_{HT}}{\widehat{X}_r}r_i - 1\right)(\beta_0 + \beta_1 x_i)\\
&= \beta_0\sum_{i\in s}\frac{1}{\pi_i}\left(\frac{\widehat{X}_{HT}}{\widehat{X}_r}r_i - 1\right)
\end{aligned}$$

$$\neq 0,$$

en général, en notant que

$$\beta_1 \sum_{i \in s} \frac{1}{\pi_i} \left(\frac{\widehat{X}_{HT}}{\widehat{X}_r} r_i - 1 \right) x_i = 0.$$

Le biais de l'estimateur imputé n'est pas nul en général et dépend, entre autres, de β_0, l'ordonnée à l'origine. Le biais est nul si $\beta_0 = 0$, auquel cas on retrouve le modèle de type ratio décrit dans la section 2.3.3 :

$$y_i = \beta x_i + \epsilon_i.$$

Exemple 2.4.3. *Supposons que le vrai modèle soit donné par*

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,$$

avec $\mathrm{V}_m(\epsilon_i) = \sigma^2 c_i$. Pour remplacer les valeurs manquantes, nous utilisons l'imputation par la régression linéaire simple décrite dans la section 2.3.2. On a

$$y_i^* = \widehat{\beta}_{0r} + \widehat{\beta}_{1r} x_i.$$

Autrement dit, le terme d'imputation n'inclut pas le terme quadratique $\beta_2 x_i^2$. L'estimateur imputé s'écrit comme

$$\begin{aligned}
\widehat{Y}_I &= \sum_{i \in s} r_i \frac{y_i}{\pi_i} + \sum_{i \in s} \frac{(1 - r_i)}{\pi_i} \left(\widehat{\beta}_{0r} + \widehat{\beta}_{1r} x_i \right) \\
&= \sum_{i \in s} r_i \frac{y_i}{\pi_i} + \sum_{i \in s} \frac{(1 - r_i)}{\pi_i} \left(\overline{y}_r - \widehat{\beta}_{1r} \overline{x}_r + \widehat{\beta}_{1r} x_i \right) \\
&= \widehat{N}_{HT} \overline{y}_r + \left(\widehat{N}_{HT} - \widehat{N}_r \right) \overline{y}_r + \widehat{\beta}_{1r} \widehat{N}_{HT} \left(\overline{x}_{HA} - \overline{x}_r \right) \\
&= \widehat{N}_{HT} \left\{ \overline{y}_r + \widehat{\beta}_{1r} \left(\overline{x}_{HA} - \overline{x}_r \right) \right\},
\end{aligned}$$

où $\overline{x}_{HA} = \widehat{X}_{HT} / \widehat{N}_{HT}$ et $\overline{y}_{HA} = \widehat{Y}_{HT} / \widehat{N}_{HT}$. L'erreur de non-réponse est donnée par

$$\begin{aligned}
\widehat{Y}_I - \widehat{Y}_{HT} &= \widehat{N}_{HT} \left\{ \overline{y}_r + \widehat{\beta}_{1r} \left(\overline{x}_{HA} - \overline{x}_r \right) \right\} - \widehat{N} \overline{y}_{HA} \\
&= \widehat{N} \left\{ \left(\overline{y}_r - \overline{y}_{HA} \right) + \widehat{\beta}_{1r} \left(\overline{x}_{HA} - \overline{x}_r \right) \right\}.
\end{aligned}$$

Dans ce cas, le biais de non-réponse est généralement égal à

$$\begin{aligned}
\mathrm{B}_m &= \mathrm{E}_m \left(\widehat{Y}_I - \widehat{Y}_{HT} | s, s_r \right) \\
&= \mathrm{E}_m \left[\widehat{N} \left\{ \left(\overline{y}_r - \overline{y}_{HA} \right) + \widehat{\beta}_{1r} \left(\overline{x}_{HA} - \overline{x}_r \right) \right\} \middle| s, s_r \right] \\
&= \widehat{N} \left\{ \left(\beta_0 + \beta_1 \overline{x}_r + \beta_2 \overline{x^2}_r - \beta_0 - \beta_1 \overline{x}_{HA} + \beta_2 \overline{x^2}_{HA} \right) \right.
\end{aligned}$$

$$+ (\overline{x}_{HA} - \overline{x}_r) \left(\beta_1 + \beta_2 \frac{\sum\limits_{i \in s} \frac{r_i}{\pi_i}(x_i - \overline{x}_r)x_i^2}{\sum\limits_{i \in s} \frac{r_i}{\pi_i}(x_i - \overline{x}_r)^2} \right) \Bigg\}$$

$$= \widehat{N}\beta_2 \left\{ \left(\overline{x^2}_r - \overline{x^2}_{HA} \right) + \frac{\sum\limits_{i \in s} \frac{r_i}{\pi_i}(x_i - \overline{x}_r)x_i^2}{\sum\limits_{i \in s} \frac{r_i}{\pi_i}(x_i - \overline{x}_r)^2} \right\}$$

$$\neq 0,$$

où $\overline{x^2}_r = \sum\limits_{i \in s} \frac{r_i x_i^2}{\pi_i} \Big/ \sum\limits_{i \in s} \frac{r_i}{\pi_i}$ *et* $\overline{x^2}_{HA} = \sum\limits_{i \in s} \frac{x_i^2}{\pi_i} \Big/ \sum\limits_{i \in s} \frac{1}{\pi_i}$. *Le biais de l'estimateur imputé n'est donc pas nul et il dépend, entre autres, du coefficient* β_2. *Le biais est nul si* $\beta_2 = 0$, *auquel cas on retrouve le modèle de type régression linéaire simple décrit dans la section 2.3.2 :*

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Exemple 2.4.4. *Supposons que le vrai modèle soit donné par*

$$y_i = \beta x_i + \epsilon_i.$$

Pour remplacer les valeurs manquantes, nous utilisons l'imputation par la régression linéaire simple décrite à la section 2.3.2. On a

$$y_i^* = \widehat{\beta}_{0r} + \widehat{\beta}_{1r} x_i.$$

Autrement dit, le modèle d'imputation contient l'ordonnée à l'origine qui ne fait pas partie du vrai modèle. Le modèle d'imputation contient donc une variable superflue, la variable qui vaut 1 *pour toutes les unités répondantes. L'estimateur imputé s'écrit comme*

$$\widehat{Y}_I = \sum_{i \in s} r_i \frac{y_i}{\pi_i} + \sum_{i \in s} \frac{(1 - r_i)}{\pi_i} \left\{ \overline{y}_r + \widehat{\beta}_{1r}(x_i - \overline{x}_r) \right\}.$$

L'erreur de non-réponse est donnée par

$$\widehat{Y}_I - \widehat{Y}_{HT} = -\sum_{i \in s} (1 - r_i) \frac{y_i}{\pi_i} + \sum_{i \in s} \frac{(1 - r_i)}{\pi_i} \left\{ \overline{y}_r + \widehat{\beta}_{1r}(x_i - \overline{x}_r) \right\}.$$

Dans ce cas, le biais de non-réponse est égal à

$$\begin{aligned}
\mathrm{B}_m &= \mathrm{E}_m \left(\widehat{Y}_I - \widehat{Y}_{HT} | s, s_r \right) \\
&= \mathrm{E}_m \left(-\sum_{i \in s} (1 - r_i) \frac{y_i}{\pi_i} + \sum_{i \in s} \frac{(1 - r_i)}{\pi_i} \left\{ \overline{y}_r + \widehat{\beta}_{1r}(x_i - \overline{x}_r) \right\} | s, s_r \right) \\
&= -\sum_{i \in s} (1 - r_i) \frac{\beta x_i}{\pi_i} + \sum_{i \in s} \frac{(1 - r_i)}{\pi_i} \left\{ \beta \overline{x}_r + \frac{\sum\limits_{i \in s} \frac{r_i}{\pi_i}(x_i - \overline{x}_r)\beta x_i}{\sum\limits_{i \in s} \frac{r_i}{\pi_i}(x_i - \overline{x}_r)^2} (x_i - \overline{x}_r) \right\}
\end{aligned}$$

$$= -\beta x_i \sum_{i \in s} \frac{(1 - r_i)}{\pi_i} + \beta \sum_{i \in s} \frac{(1 - r_i)}{\pi_i} \overline{x}_r + \beta \sum_{i \in s} \frac{(1 - r_i)}{\pi_i} (x_i - \overline{x}_r)$$

$$= 0.$$

Les exemples 2.4.1 à 2.4.4 montrent que l'omission d'une variable auxiliaire dans le modèle d'imputation mène généralement à un biais. Lorsque le modèle d'imputation inclut des variables auxiliaires superflues, l'estimateur imputé est (approximativement) sans biais, pourvu que les variables importantes soient incluses. Par contre, l'ajout de variables superflues contribuera généralement à hausser la variance des estimateurs imputés.

2.4.2. Relation entre la variable d'intérêt et les variables auxiliaires

L'imputation peut avoir un effet sur la relation entre la variable d'intérêt y et les variables auxiliaires $\boldsymbol{x}$. Par exemple, si on observe une relation linéaire entre y et $\boldsymbol{x}$ et que l'on utilise l'imputation par la moyenne, cela aura pour effet d'atténuer la corrélation entre les variables. En effet, peu importe les valeurs de $\boldsymbol{x}$, la valeur imputée est la moyenne des répondants. Forcément, la corrélation entre la variable d'intérêt et les variables auxiliaires diminuera. Il faut donc faire attention à la méthode d'imputation utilisée.

2.4.3. Traitement des valeurs imputées

Les valeurs imputées ne sont pas des valeurs observées. Il est important de considérer ce fait lors de l'inférence statistique. Traiter les valeurs imputées comme si elles sont des valeurs observées peut occasionner une sous-estimation considérable de la variance des estimateurs. Les intervalles de confiance des estimateurs sont alors moins larges que ceux que l'on aurait obtenus si on avait estimé la variance correctement. L'estimation de la variance en présence de données imputées sera traitée en détail au Chapitre 3.

Chapitre 3

ESTIMATION DE LA VARIANCE EN PRÉSENCE DE DONNÉES IMPUTÉES

Dans ce chapitre, nous traitons le problème de l'estimation de la variance en présence de données imputées. Il est bien connu que le fait de traiter les données imputées comme des valeurs observées mène généralement à une sous-estimation de la variance. La sous-estimation tend à être importante lorsque le taux de non-réponse est élevé. Dans ce cas, l'utilisation d'estimateurs naïfs de la variance dans la construction d'intervalles de confiance conduit à des intervalles trop courts, avec comme conséquense une probabilité de couverture inférieure au taux nominal (par exemple, 95%). Comme nous le verrons, les estimateurs de la variance en présence de données imputées dépendent des probabilités d'inclusion d'ordre deux, π_{ij}. Il s'agira donc d'utiliser les approximations de π_{ij} présentées au chapitre 1 afin d'obtenir des estimateurs de variance approximatifs. Plusieurs approches ont été proposées dans la littérature afin d'estimer la variance d'un estimateur imputé. Särndal (1992) a proposé l'approche à deux-phases détaillée à la section 3.2 et Shao et Steel (1999) ont proposé l'approche renversée présentée à la section 3.3.

3.1. Hypothèses sur le modèle

Dans ce chapitre, nous présentons des estimateurs de variance dans le cas de l'imputation par la régression. Rappelons que le modèle d'imputation sous-jacent est donné par

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i, \tag{3.1.1}$$

avec $\mathrm{V}_m(\epsilon_i) = \sigma^2 c_i$; voir la section 2.3.1. Les valeurs imputées sont données par

$$y_i^* = \boldsymbol{x}_i'\widehat{\mathbf{B}}_r,$$

où

$$\widehat{\mathbf{B}}_r = \left(\sum_{i\in s} \frac{r_i}{\pi_i} \boldsymbol{x}_i c_i^{-1} \boldsymbol{x}_i'\right)^{-1} \sum_{i\in s} \frac{r_i}{\pi_i} \boldsymbol{x}_i c_i^{-1} y_i = \widehat{\boldsymbol{T}}_r^{-1} \widehat{\boldsymbol{t}}_r.$$

L'estimateur imputé (2.2.1) peut s'écrire comme

$$\widehat{Y}_I = \sum_{i\in s} r_i \frac{y_i}{\pi_i} + \sum_{i\in s} \frac{(1-r_i)}{\pi_i} \boldsymbol{x}_i' \widehat{\mathbf{B}}_r. \tag{3.1.2}$$

3.2. Approche à deux-phases

Le processus sur lequel l'approche à deux-phases repose est le suivant :

(i) On tire un échantillon s de la population U selon un plan de sondage $p(s)$;

(ii) l'échantillon de répondants s_r est généré selon un mécanisme de réponse.

Ce processus est représenté à la figure 2.1. Dans ce qui suit, on suppose que l'estimateur $\widehat{Y}_I$ est sans biais pour Y ; c'est-à-dire, $\mathrm{E}_{mpr}\left(\widehat{Y}_I - Y\right) = 0$. Cette hypothèse est appropriée lorsque le modèle (3.1.1) est correctement spécifié.

On détermine la variance totale de l'estimateur imputé à l'aide de la décomposition de l'erreur totale en (2.3.6) :

$$\begin{aligned}
\mathrm{V}_{tot} &= \mathrm{V}\left(\widehat{Y}_I - Y\right) \\
&= \mathrm{E}\left(\widehat{Y}_I - Y\right)^2 \\
&= \mathrm{E}_m\mathrm{E}_p\mathrm{E}_r\left(\widehat{Y}_I - Y\right)^2 \\
&= \mathrm{E}_m\mathrm{E}_p\mathrm{E}_r\left\{\left(\widehat{Y}_I - \widehat{Y}_{HT}\right) + \left(\widehat{Y}_{HT} - Y\right)\right\}^2 \\
&= \mathrm{E}_m\mathrm{E}_p\mathrm{E}_r\left(\widehat{Y}_{HT} - Y\right)^2 + \mathrm{E}_m\mathrm{E}_p\mathrm{E}_r\left(\widehat{Y}_I - \widehat{Y}_{HT}\right)^2 \\
&\quad + 2\mathrm{E}_m\mathrm{E}_p\mathrm{E}_r\left\{\left(\widehat{Y}_I - \widehat{Y}_{HT}\right)\left(\widehat{Y}_{HT} - Y\right)\right\} \\
&= \mathrm{E}_m\mathrm{V}_p\left(\widehat{Y}_{HT}\right) + \mathrm{E}_p\mathrm{E}_r\mathrm{V}_m\left(\widehat{Y}_I - \widehat{Y}_{HT}|s, s_r\right) \\
&\quad + 2\mathrm{E}_p\mathrm{E}_r\mathrm{Cov}_m\left\{\left(\widehat{Y}_I - \widehat{Y}_{HT}\right)\left(\widehat{Y}_{HT} - Y\right)|s, s_r\right\} \\
&= \mathrm{V}_{sam} + \mathrm{V}_{NR} + 2\mathrm{V}_{mix},
\end{aligned} \tag{3.2.1}$$

où

$$\mathrm{V}_{sam} = \mathrm{E}_m\mathrm{V}_p\left(\widehat{Y}_{HT}\right), \tag{3.2.2}$$

$$\mathrm{V}_{NR} = \mathrm{E}_p\mathrm{E}_r\mathrm{V}_m\left(\widehat{Y}_I - \widehat{Y}_{HT}|s, s_r\right) \tag{3.2.3}$$

et

$$\mathrm{V}_{mix} = \mathrm{E}_p\mathrm{E}_r\mathrm{Cov}_m\left\{\left(\widehat{Y}_I - \widehat{Y}_{HT}\right)\left(\widehat{Y}_{HT} - Y\right)|s, s_r\right\}. \tag{3.2.4}$$

Notons que pour arriver à l'expression (3.2.1), l'ordre des espérances a été interchangé, ce qui est approprié lorsque les données sont *MAR*, voir la section (2.3.8). La variance totale en (3.2.1) s'exprime comme la somme de trois termes : la variance due à l'échantillonnage (3.2.2), la variance due à la non-réponse (3.2.3) et un terme mixte de la covariance entre l'erreur causée par l'échantillonnage et l'erreur causée par la non-réponse donnée en (3.2.4). Un estimateur de la variance totale en (3.2.1) s'obtient en estimant chacun des termes séparément, ce qui conduit à

$$\widehat{\mathrm{V}}_{tot} = \widehat{\mathrm{V}}_{sam} + \widehat{\mathrm{V}}_{NR} + 2\widehat{\mathrm{V}}_{mix}. \tag{3.2.5}$$

Särndal (1992) et Deville et Särndal (1994) ont proposé une manière d'estimer le terme V_{sam} ; la méthode est présentée à la section 3.2.1.1. Beaumont et Bocci (2009) ont proposé une méthode alternative afin d'estimer V_{sam}. Cette méthode est présentée à la section 3.2.1.2. L'estimation de V_{NR} et celle de V_{mix} sont détaillées aux sections 3.2.2 et 3.2.3, respectivement.

3.2.1. Estimation de la variance due à l'échantillonnage

Traiter les valeurs imputées comme des valeurs observées consiste à utiliser l'estimateur de la variance d'Horvitz-Thompson (1.1.6) en remplaçant y_i par $\widetilde{y}_i = r_i y_i + (1 - r_i) y_i^*$, ce qui conduit à l'estimateur de variance dit naïf,

$$\widehat{\mathrm{V}}_{naïf} = \sum_{i \in s} \sum_{j \in s} \frac{\Omega_{ij}}{\pi_{ij}} \widetilde{y}_i \widetilde{y}_j. \tag{3.2.6}$$

En général, l'estimateur naïf (3.2.6) sous-estime la variance due à l'échantillonnage. Nous illustrons cet aspect dans l'exemple suivant.

Exemple 3.2.1. *Considérons le cas de l'EASSR et de l'imputation par la moyenne. Dans ce cas, on a*

$$\pi_i = n/N, \text{ pour tout } i,$$

et

$$\pi_{ij} = n(n-1)/N(N-1), \ i \neq j.$$

De plus, rappelons que le modèle sous-jacent à l'imputation par la moyenne est donné en (3.1.1) avec $x_i = c_i = 1$ pour tout i. On obtient alors

$$\widehat{\mathrm{V}}_{naïf} = \sum_{i \in s} \sum_{j \in s} \frac{\Omega_{ij}}{\pi_{ij}} \widetilde{y}_i \widetilde{y}_j$$

$$
\begin{aligned}
&= \sum_{i\in s}\sum_{\substack{j\in s\\ j\neq i}} \frac{\Omega_{ij}}{\pi_{ij}}\widetilde{y}_i\widetilde{y}_j + \sum_{i\in s}\frac{\Omega_{ii}}{\pi_i}\widetilde{y}_i^2 \\
&= \frac{N^2}{n^2}\left(1-\frac{n}{N}\right)\sum_{i\in s}\widetilde{y}_i^2 - \frac{N^2}{n^2}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{i\in s}\sum_{\substack{j\in s\\ j\neq i}}\widetilde{y}_i\widetilde{y}_j \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\left\{\sum_{i\in s}\widetilde{y}_i^2 - \frac{1}{n}\left(\sum_{i\in s}\widetilde{y}_i\right)^2\right\} \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\left\{\sum_{i\in s} r_i y_i^2 + \overline{y}_r^2\sum_{i\in s}(1-r_i) - \frac{1}{n}\left(\sum_{i\in s} r_i y_i\right)^2\right. \\
&\quad \left. - \frac{\overline{y}_r^2}{n}\left(\sum_{i\in s}(1-r_i)\right)^2 - \frac{2\overline{y}_r}{n}\sum_{i\in s}(1-r_i)\sum_{j\in s} r_i y_i\right\} \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\left\{\sum_{i\in s} r_i y_i^2 - n_r\overline{y}_r^2\right\} \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{n_r-1}{n-1}s_{y_r}^2,
\end{aligned}
\tag{3.2.7}
$$

où

$$
s_{y_r}^2 = \frac{1}{n_r-1}\sum_{i\in s} r_i\left(y_i-\overline{y}_r\right)^2 \tag{3.2.8}
$$

désigne la variance de la variable y observée dans l'ensemble des répondants. Le biais conditionnel relatif de $\mathrm{V}_{naïf}$ *est donné par*

$$
\begin{aligned}
\mathrm{BR}\left(\widehat{\mathrm{V}}_{naïf}\right) &= \frac{\mathrm{E}_m\left(\widehat{\mathrm{V}}_{naïf} - \widehat{\mathrm{V}}_{HT}\middle| s, s_r\right)}{\mathrm{E}_m\left(\widehat{\mathrm{V}}_{HT}\middle| s\right)} \\
&= -\frac{N^2\left(1-\frac{n}{N}\right)\left(1-\frac{n_r}{n}\right)\frac{\sigma^2}{n-1}}{N^2\left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}} \\
&= -\left(\frac{n}{n-1}\right)\left(1-\frac{n_r}{n}\right).
\end{aligned}
\tag{3.2.9}
$$

Dans ce cas, le biais de l'estimateur naïf est négatif, ce qui montre que $\widehat{\mathrm{V}}_{naïf}$ *sous-estime la variance due à l'échantillonnage,* V_{sam}*. La sous-estimation est d'autant plus importante que le taux de réponse* n_r/n *est faible.*

Puisque l'estimateur naïf n'est généralement pas approprié comme estimateur de V_{sam}, on va, dans un premier temps, évaluer son biais. Posons

$$
\mathrm{V}_{diff} = \mathrm{E}_m\left(\mathrm{V}_{sam} - \widehat{\mathrm{V}}_{naïf}|s, s_r\right). \tag{3.2.10}
$$

Dans la situation traitée à l'exemple 3.2.1, on a

$$\mathrm{V}_{diff} = -N^2 \left(1 - \frac{n}{N}\right) \left(1 - \frac{n_r}{n}\right) \frac{\sigma^2}{n-1},$$

qui dépend du paramètre du modèle d'imputation σ^2. On estimera V_{diff} en estimant σ^2 par un estimateur sans biais $\widehat{\sigma}^2$, ce qui conduira à

$$\widehat{\mathrm{V}}_{sam} = \widehat{\mathrm{V}}_{naïf} + \widehat{\mathrm{V}}_{diff}. \tag{3.2.11}$$

L'estimateur $\widehat{\mathrm{V}}_{sam}$ en (3.2.11) peut donc être vu comme un estimateur ajusté pour le biais. En effet, on a

$$\begin{aligned}
\mathrm{E}_{mpr}\left(\widehat{\mathrm{V}}_{sam} - \mathrm{V}_{sam}\right) &= \mathrm{E}_{pr}\mathrm{E}_m\left(\widehat{\mathrm{V}}_{sam} - \mathrm{V}_{sam}\middle| s, s_r\right) \\
&= \mathrm{E}_{pr}\mathrm{E}_m\left(\widehat{\mathrm{V}}_{naïf} + \widehat{\mathrm{V}}_{diff} - \mathrm{V}_{sam}\middle| s, s_r\right) \\
&= -\mathrm{E}_{pr}\mathrm{E}_m\left(\mathrm{V}_{sam} - \widehat{\mathrm{V}}_{naïf}\middle| s, s_r\right) + \mathrm{E}_{pr}\mathrm{E}_m\left(\widehat{\mathrm{V}}_{diff}\middle| s, s_r\right) \\
&= -\mathrm{E}_{pr}\left(\mathrm{V}_{diff}\right) + \mathrm{E}_{pr}\left(\mathrm{V}_{diff}\right) \\
&= 0.
\end{aligned}$$

Il existe plusieurs manières d'obtenir un estimateur de V_{diff} dans la littérature. Ces dernières sont détaillées dans les sections 3.2.1.1 et 3.2.1.2.

3.2.1.1. *Méthode de Särndal*

Särndal (1992) et Deville et Särndal (1994) proposent d'évaluer explicitement le terme V_{diff} en (3.2.10). On peut montrer que (voir Annexe A) l'estimateur de la variance V_{diff} dans le cas de la méthode de Särndal (1992) est

$$\begin{aligned}
\widehat{\mathrm{V}}_{diff} = \widehat{\sigma}^2 \Bigg\{ &\sum_{i\in s}(1-r_i)\frac{\Omega_{ii}}{\pi_{ii}}c_i - 2\sum_{i\in s}\sum_{j\in s} r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}\pi_i}\boldsymbol{x}_i'\widehat{\boldsymbol{T}}_r^{-1}\boldsymbol{x}_j \\
&- \sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}\boldsymbol{x}_i'\widehat{\boldsymbol{T}}_r^{-1}\sum_{k\in s}\frac{r_k}{\pi_k^2}\boldsymbol{x}_k c_k^{-1}\boldsymbol{x}_j'\widehat{\boldsymbol{T}}_r^{-1}\boldsymbol{x}_k \Bigg\},
\end{aligned} \tag{3.2.12}$$

où

$$\widehat{\sigma}^2 = \frac{1}{\sum\limits_{i\in s}\frac{r_i c_i}{\pi_i}} \sum_{i\in s}\frac{r_i}{\pi_i}\left(y_i - \boldsymbol{x}_i'\widehat{\mathbf{B}}_r\right)^2. \tag{3.2.13}$$

Notons que $\widehat{\sigma}^2$ est un estimateur approximativement sans biais pour σ^2 sous le modèle (3.1.1) ; voir Deville et Särndal (1994). En utilisant (3.2.11) et (3.2.12), on obtient (voir Annexe A) un estimateur de V_{sam} :

$$\widehat{\mathrm{V}}_{sam} = \sum_{i\in s}\sum_{j\in s}\frac{\Omega_{ij}}{\pi_{ij}}\widetilde{y}_i\widetilde{y}_j + \widehat{\sigma}^2\left\{\sum_{i\in s}(1-r_i)\frac{\Omega_{ii}}{\pi_{ii}}c_i - 2\sum_{i\in s}\sum_{j\in s} r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}\pi_i}\boldsymbol{x}_i'\widehat{\boldsymbol{T}}^{-1}\boldsymbol{x}_j\right.$$

$$-\sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}\boldsymbol{x}_i'\widehat{\boldsymbol{T}}^{-1}\sum_{k\in s}\frac{r_k}{\pi_k^2}\boldsymbol{x_k}c_k^{-1}\boldsymbol{x}_j'\widehat{\boldsymbol{T}}^{-1}\boldsymbol{x}_k\Bigg\}. \tag{3.2.14}$$

Notons que plusieurs termes dans (3.2.14) dépendent des probabilités d'inclusion d'ordre deux. Il serait possible d'utiliser les approximations des π_{ij} présentées au chapitre 1 afin d'obtenir un estimateur de variance simplifié. Cependant, il en résulterait un estimateur relativement complexe, compte tenu des nombreux termes dans (3.2.14) dépendant des π_{ij}. Dans la prochaine section, nous présentons une méthode d'estimation du terme V_{sam} qui conduira à un estimateur de variance beaucoup plus simple.

3.2.1.2. *Méthode de Beaumont et Bocci*

Beaumont et Bocci (2009) suggèrent d'obtenir V_{diff} conditionnellement à s, à s_r et à $\boldsymbol{y}_r$, qui est la partie observée de $\boldsymbol{y}_s = (y_1,...,y_n)'$. Cette méthode étant beaucoup plus simple que la précédente, nous la privilégierons dans la suite de ce mémoire. On peut montrer que (voir Annexe B) l'estimateur de la variance V_{diff}, dans le cas de la méthode de Beaumont et Bocci, est

$$\widehat{\mathrm{V}}_{diff} = \widehat{\sigma}^2\sum_{i\in s}(1-r_i)\frac{1-\pi_i}{\pi_i^2}c_i.$$

Cela conduit à l'estimateur de V_{sam} suivant :

$$\widehat{\mathrm{V}}_{sam} = \sum_{i\in s}\sum_{j\in s}\frac{\Omega_{ij}}{\pi_{ij}}\widetilde{y}_i\widetilde{y}_j + \widehat{\sigma}^2\sum_{i\in s}(1-r_i)\frac{1-\pi_i}{\pi_i^2}c_i. \tag{3.2.15}$$

3.2.2. Estimation de la variance due à la non-réponse

La composante V_{NR} de la variance totale V_{tot} est la variance qui est due à l'imputation. On peut montrer que (voir Annexe C) l'estimateur de la variance V_{NR} est

$$\widehat{\mathrm{V}}_{NR} = \widehat{\sigma}^2\sum_{i\in s}\frac{1}{\pi_i^2}(r_ig_i-1)^2c_i,$$

où

$$g_i = 1 + \left(\widehat{\boldsymbol{X}}_{HT} - \widehat{\boldsymbol{X}}_r\right)'\widehat{\boldsymbol{T}}_r^{-1}c_i^{-1}\boldsymbol{x_i},$$

avec $\widehat{\boldsymbol{X}}_r = \sum_{i\in s}\frac{r_i}{\pi_i}\boldsymbol{x}_i'$ et $\widehat{\sigma}^2$ donné par (3.2.13).

3.2.3. Estimation du terme mixte

Le terme V_{mix} de la variance totale V_{tot} est un terme de covariance entre l'erreur qui est due à l'échantillonnage et l'erreur qui est due à l'imputation. On

peut montrer que (voir Annexe D) l'estimateur du terme V_{mix} est

$$\widehat{\mathrm{V}}_{mix} = \widehat{\sigma}^2 \sum_{i\in s} \left(\frac{1-\pi_i}{\pi_i^2}\right)(r_i g_i - 1)c_i,$$

où $\widehat{\sigma}^2$ est donné par (3.2.13).

3.2.4. Approximation de la variance totale

L'estimateur de la variance totale V_{tot} sous l'approche à deux-phases et la méthode de Beaumont et Bocci est

$$\widehat{\mathrm{V}}_{D,tot} = \sum_{i\in s}\sum_{j\in s} \frac{\Omega_{ij}}{\pi_{ij}} \widetilde{y}_i \widetilde{y}_j + \widehat{\sigma}^2 \sum_{i\in s} \frac{1}{\pi_i^2} \left\{(r_i g_i - \pi_i)^2 + (1-\pi_i)(\pi_i - r_i)\right\} c_i. \quad (3.2.16)$$

Le premier terme de (3.2.16), soit $\widehat{\mathrm{V}}_{naïf}$, dépend des probabilités d'inclusion d'ordre deux, π_{ij}, et s'exprime comme une double somme. On peut alors utiliser n'importe quelle approximation des π_{ij} présentée au chapitre 1, ce qui conduit à l'estimateur simplifié

$$\widehat{\mathrm{V}}^*_{naïf} = \sum_{i\in s} \varphi_i \left(\frac{\widetilde{y}_i}{\pi_i} - \widehat{B}\right)^2, \quad (3.2.17)$$

où

$$\widehat{B} = \frac{\sum\limits_{i\in s} \alpha_i \frac{\widetilde{y}_i}{\pi_i}}{\sum\limits_{i\in s} \alpha_i}$$

et φ_i et α_i sont des coefficients donnés en fonction de l'approximation utilisée, voir Tableau 1.3. L'estimateur (3.2.17) est de la même forme que (1.4.16), mais avec y_i remplacé par $\widetilde{y}_i$.

L'estimateur simplifié de la variance totale V_{tot} qui en résulte est donc

$$\widehat{\mathrm{V}}^*_{D,tot} = \sum_{i\in s} \varphi_i \left(\frac{\widetilde{y}_i}{\pi_i} - \widehat{B}\right)^2 + \widehat{\sigma}^2 \sum_{i\in s} \frac{1}{\pi_i^2} \left\{(r_i g_i - \pi_i)^2 + (1-\pi_i)(\pi_i - r_i)\right\} c_i. \quad (3.2.18)$$

Remarque 3.2.1. *À l'exception de $\widehat{\mathrm{V}}_{naïf}$ remarquons que tous les termes de l'estimateur de la variance totale, $\widehat{\mathrm{V}}_{tot}$, dépendent du modèle d'imputation. Plus précisément, $\widehat{\mathrm{V}}_{tot}$ s'exprime en fonction des deux premiers moments de la distribution de y, soient $\mathrm{E}_m(y_i)$ et $\mathrm{V}_m(y_i)$. Donc la validité de l'estimateur de la variance totale $\widehat{\mathrm{V}}_{tot}$ dépend de la bonne spécification des deux premiers moments du modèle d'imputation.*

3.3. Approche renversée

Le processus sur lequel l'approche renversée repose est le suivant :

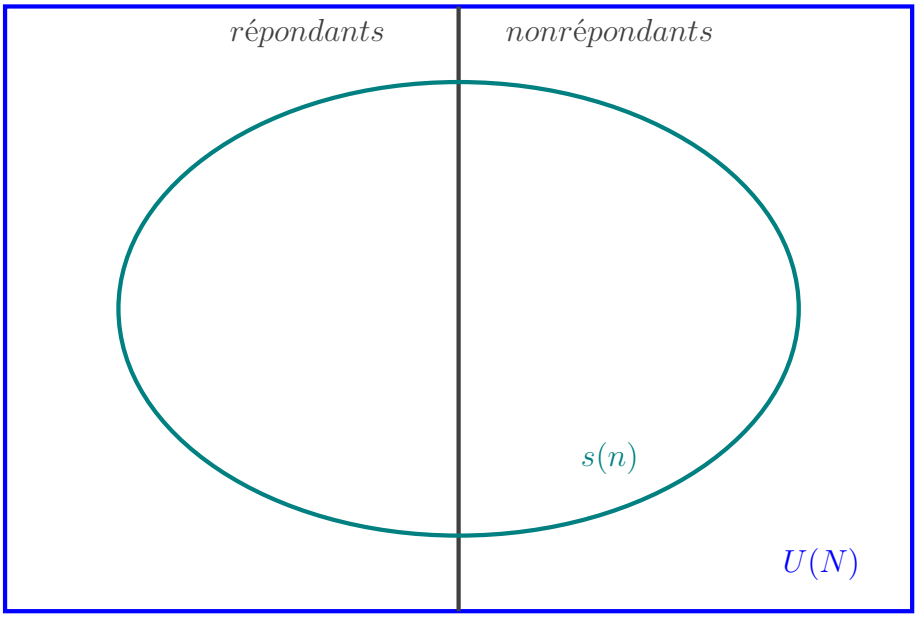


FIGURE 3.1. Représentation de l'approche renversée

(i) La population U est divisée aléatoirement, selon le mécanisme de réponse, en deux sous-populations, soient la population des répondants et celle des non-répondants ;

(ii) Un échantillon s est sélectionné dans la population U et il contient des répondants et des non-répondants.

Donc la configuration de l'approche à deux-phases est renversée. Ce processus est représenté à la figure 3.1. L'approche renversée repose sur une hypothèse supplémentaire : le vecteur $\boldsymbol{r} = (r_1, ...,r_N)'$ est indépendant du vecteur $\boldsymbol{\delta} = (\delta_1, ...,\delta_N)'$. Cette hypothèse est similaire à la propriété d'invariance souvent utilisée dans un contexte d'échantillonnage à deux phases ; voir, par exemple, Särndal et coll. (1992). Afin d'exprimer la variance totale, nous exprimons l'erreur totale comme :

$$\widehat{Y}_I - Y = \left(\widehat{Y}_I - \widetilde{Y}_I\right) + \left(\widetilde{Y}_I - Y\right), \tag{3.3.1}$$

où $\widetilde{Y}_I = \mathrm{E}_p\left(\widehat{Y}_I \middle| \boldsymbol{y}_U, \boldsymbol{r}\right)$ et $\boldsymbol{y}_U = (y_1,...,y_N)'$. On supposera que l'estimateur $\widehat{Y}_I$ est sans biais pour Y ; i.e., $\mathrm{E}_{mpr}\left(\widehat{Y}_I - Y\right) = 0$. En utilisant la décomposition (3.3.1), on obtient

$$\begin{aligned}
\mathrm{V}_{tot} &= \mathrm{E}\left(\widehat{Y}_I - Y\right)^2 \\
&= \mathrm{E}_r\mathrm{E}_m\mathrm{E}_p\left\{\left(\widehat{Y}_I - Y\right)^2 \middle| \boldsymbol{y}_U, \boldsymbol{r}\right\} \\
&= \mathrm{E}_r\mathrm{E}_m\mathrm{E}_p\left\{\left(\widehat{Y}_I - \widetilde{Y}_I\right)^2 + \left(\widetilde{Y}_I - Y\right)^2 + 2\left(\widehat{Y}_I - \widetilde{Y}_I\right)\left(\widetilde{Y}_I - Y\right) \middle| \boldsymbol{y}_U, \boldsymbol{r}\right\}
\end{aligned}$$

$$= \mathrm{E}_r \mathrm{E}_m \mathrm{V}_p \left(\widehat{Y}_I | \boldsymbol{y}_U, \boldsymbol{r} \right) + \mathrm{E}_r \mathrm{V}_m \left(\widetilde{Y}_I - Y | \boldsymbol{y}_U, \boldsymbol{r} \right)$$
$$= \mathrm{V}_1 + \mathrm{V}_2, \tag{3.3.2}$$

en notant que

$$\mathrm{E}_p \left\{ \left(\widehat{Y}_I - \widetilde{Y}_I \right) \left(\widetilde{Y}_I - Y \right) \middle| \boldsymbol{y}_U, \boldsymbol{r} \right\} = \left(\widetilde{Y}_I - Y \right) \mathrm{E}_p \left(\widehat{Y}_I - \widetilde{Y}_I \middle| \boldsymbol{y}_U, \boldsymbol{r} \right)$$
$$= 0.$$

Un estimateur de V_{tot} est obtenu en estimant chaque terme séparément dans (3.3.2).

3.3.1. Estimation de V_1

Obtenir un estimateur sans biais de V_1 passe par l'obtention d'un estimateur sans biais de $\mathrm{V}_p \left(\widehat{Y}_I | \boldsymbol{y}_U, \boldsymbol{r} \right)$, qui représente la variance due à l'échantillonnage de l'estimateur $\widehat{Y}_I$ conditionnellement à $\boldsymbol{y}_U$ et à $\boldsymbol{r}$. Or, comme nous le verrons dans le cas de l'imputation par la régression, l'estimateur $\widehat{Y}_I$ s'exprime comme une fonction de totaux estimés conditionnellement à $\boldsymbol{y}_U$ et à $\boldsymbol{r}$. Estimer $\mathrm{V}_p \left(\widehat{Y}_I | \boldsymbol{y}_U, \boldsymbol{r} \right)$ revient donc à estimer la variance due à l'échantillonnage d'une fonction de totaux, ce qui est un problème classique. Pour cela, on peut approximer $\mathrm{V}_p \left(\widehat{Y}_I | \boldsymbol{y}_U, \boldsymbol{r} \right)$ au moyen d'un développement de Taylor du premier ordre, puis on obtient (voir Annexe E) un estimateur de la variance V_1 :

$$\widehat{\mathrm{V}}_1 = \sum_{i \in s} \sum_{j \in s} \frac{\Omega_{ij}}{\pi_{ij}} e_i e_j,$$

où

$$e_i = r_i y_i + (1 - r_i) \boldsymbol{x}_i' \widehat{\mathbf{B}}_r + \left(\sum_{j \in s} \frac{(1 - r_j)}{\pi_i} \boldsymbol{x}_j \right)' \widehat{\boldsymbol{T}}_r^{-1} r_i \boldsymbol{x}_i c_i^{-1} \left(y_i - \boldsymbol{x}_i' \widehat{\mathbf{B}}_r \right). \tag{3.3.3}$$

Remarque 3.3.1. *L'estimateur $\widehat{\mathrm{V}}_1$ étant un estimateur par linéarisation classique, sa validité ne dépend pas de celle du modèle d'imputation. Autrement dit, pourvu que la taille d'échantillon soit suffisament grande et que le taux de réponse soit borné inférieurement, l'estimateur $\widehat{\mathrm{V}}_1$ est approximativement sans biais pour $\mathrm{V}_p \left(\widehat{Y}_I | \boldsymbol{y}_U, \boldsymbol{r} \right)$, peu importe que le modèle d'imputation soit correctement spécifié ou non. Bien sûr, lorsque le modèle d'imputation n'est pas correctement spécifié, l'estimateur imputé, $\widehat{Y}_I$, est potentiellement biaisé. Cependant, l'estimateur de variance $\widehat{\mathrm{V}}_1$ estimera $\mathrm{V}_p \left(\widehat{Y}_I | \boldsymbol{y}_U, \boldsymbol{r} \right)$ correctement. Cette propriété de robustesse est attrayante en pratique. Rappelons que dans le cas de l'approche à deux-phases, tous les termes, à l'exception de $\widehat{\mathrm{V}}_{naïf}$, dépendent de la bonne spécification du premier et du deuxième moment du modèle d'imputation.*

3.3.2. Estimation de V_2

L'estimation du terme V_2 est plus simple que celle du terme V_1. On peut montrer que (voir Annexe F) un estimateur de la variance V_2 est donné par

$$\widehat{\mathrm{V}}_2 = \widehat{\sigma}^2 \sum_{i \in s} \frac{\widehat{\xi}_i^2}{\pi_i} c_i,$$

où

$$\widehat{\xi}_i = \left(\sum_{j \in s} \frac{(1-r_j)}{\pi_j} \boldsymbol{x}_j \right)' \widehat{\boldsymbol{T}}_r^{-1} r_i \boldsymbol{x}_i c_i^{-1} - (1-r_i). \tag{3.3.4}$$

Remarque 3.3.2. *Shao et Steel (1999) montrent que* V_1 *est un terme d'ordre* $O(N^2/n)$ *alors que le terme* V_2 *est un terme d'ordre* $O(N)$*. Ainsi, lorsque la fraction de sondage,* n/N*, est négligeable, la contribution de* V_2 *à la variance totale,* $\frac{\mathrm{V}_2}{\mathrm{V}_1+\mathrm{V}_2}$*, est d'ordre* $O(n/N)$*, qui est négligeable lorsque la fraction de sondage* n/N *est négligeable. Donc, lorsque la fraction de sondage est petite, on peut simplement estimer la variance totale par* $\widehat{\mathrm{V}}_1$*.*

3.3.3. Approximation de la variance totale

Un estimateur de la variance totale V_{tot} sous l'approche renversée est

$$\widehat{\mathrm{V}}_{R,tot} = \sum_{i \in s} \sum_{j \in s} \frac{\Omega_{ij}}{\pi_{ij}} e_i e_j + \widehat{\sigma}^2 \sum_{i \in s} \frac{\widehat{\xi}_i^2}{\pi_i} c_i, \tag{3.3.5}$$

où e_i est donné par (3.3.3) et $\widehat{\xi}_i$ est donné par (3.3.4). Le premier terme de (3.3.5), soit l'estimateur $\widehat{\mathrm{V}}_1$ en (E.0.17), dépend des probabilités d'inclusion du deuxième ordre, π_{ij}, et s'exprime comme une double somme. Pour contrer le problème, on utilisera n'importe quelle approximation des π_{ij} présentée au chapitre 1, ce qui conduit à l'estimateur simplifié

$$\widehat{\mathrm{V}}_1^* = \sum_{i \in s} \varphi_i \left(\frac{e_i}{\pi_i} - \widehat{B} \right)^2, \tag{3.3.6}$$

où

$$\widehat{B} = \frac{\sum_{i \in s} \alpha_i \frac{e_i}{\pi_i}}{\sum_{i \in s} \alpha_i}$$

et φ_i et α_i sont des coefficients donnés en fonction de l'approximation utilisée, voir Tableau 1.3. L'estimateur (3.3.6) est de la même forme que (1.4.16), avec y_i remplacé par e_i, où e_i est donné par (3.3.3).

L'estimateur simplifié de la variance totale V_{tot} sous l'approche renversée est donc

$$\widehat{\mathrm{V}}^{*}_{R,tot} = \sum_{i \in s} \varphi_i \left(\frac{e_i}{\pi_i} - \widehat{B} \right)^2 + \widehat{\sigma}^2 \sum_{i \in s} \frac{\widehat{\xi}_i^2}{\pi_i} c_i. \tag{3.3.7}$$

Chapitre 4

ÉTUDES PAR SIMULATION

Afin de comparer la performance des différentes approximations des probabilités d'inclusion d'ordre deux, π_{ij}, nous avons effectué deux études par simulation. Dans la première étude, nous considérons plusieurs populations dans le cas d'un jeu de données complet et nous étudions l'approximation de la variance de l'estimateur du total d'Horvitz-Thompson (1.1.4) et de l'estimateur du total de Hájek (1.5.1). Dans la seconde étude, nous considérons plusieurs populations dans le cadre d'un jeu de données avec des valeurs imputées. Nous y étudions l'approximation de la variance totale de l'estimateur imputé (2.2.1).

4.1. Étude 1 : jeu de données complet

Une première étude par simulation a été réalisée afin de comparer la performance de différentes approximations de la variance pour l'estimateur du total d'Horvitz-Thompson (1.1.4) et pour l'estimateur du total de Hájek (1.5.1), dans le cadre d'un jeu de données complet. Plusieurs études par simulation ont été réalisées dans la littérature pour comparer différentes approximations dans le cas d'Horvitz-Thompson, voir Haziza et coll. (2008). Par contre, il n'existe pas, à notre connaissance, d'études empiriques étudiant le comportement des estimateurs approximatifs dans le cas de l'estimateur de Hájek (1.5.1).

4.1.1. Populations et échantillons simulés

Un ensemble de seize populations différentes de taille N a été utilisé pour réaliser l'étude. Pour chaque population, nous avons d'abord généré une variable auxiliaire x selon une loi Gamma, de paramètres $\alpha = 5$ et $\beta = 20$, afin d'obtenir x telle que $\mathrm{E}(x) = 100$ et $\mathrm{V}(x) = 2000$. Pour générer les populations, le modèle général suivant a été utilisé :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{4.1.1}$$

où $\mathrm{E}_m(\epsilon_i) = 0$, $\mathrm{E}_m(\epsilon_i\epsilon_j) = 0$ pour $i \neq j$ et $\mathrm{V}_m(\epsilon_i) = \sigma^2 c_i$. Les seize populations ont été générées en faisant varier les coefficients β_0, β_1 et c_i, la taille de la population N, ainsi que la composante σ^2 afin de faire varier la corrélation entre la variable y et la variable x. La variable y a été générée selon 4 choix d'ensembles de coefficients β_0, β_1 et c_i :

$$y_i = 2x_i + \epsilon_i, \text{ avec } c_i = x_i; \tag{4.1.2}$$

$$y_i = 2x_i + \epsilon_i, \text{ avec } c_i = x_i^{3/2}; \tag{4.1.3}$$

$$y_i = 200 + 2x_i + \epsilon_i, \text{ avec } c_i = 1; \tag{4.1.4}$$

$$y_i = 200 + \epsilon_i, \text{ avec } c_i = 1. \tag{4.1.5}$$

Pour les populations sous les modèles (4.1.2)-(4.1.4), nous avons utilisé deux différentes valeurs de σ^2 afin d'obtenir une corrélation entre la variable y et la variable x de 0,6 et de 0,9. Pour les populations sous le modèle (4.1.5), nous avons utilisé deux différentes valeurs de σ^2 afin d'obtenir deux valeurs de $\mathrm{CV}(y) = s_y/\overline{Y}$. Pour chacune des situations, deux différentes tailles N de population ont été utilisées, soient $N = 100$ et $N = 500$.

Pour chacune des seize populations, nous avons tiré $K = 100\,000$ échantillons selon le plan de Poisson conditionnel décrit à la section 1.2.5. Les probabilités d'inclusion d'ordre un, π_i, ont été définies de manière à être proportionnelles à la variable auxiliaire x. Pour les populations de taille $N = 100$, nous avons sélectionné des échantillons de tailles $n = 10, 20, 40$. Pour les populations de taille $N = 500$, nous avons sélectionné des échantillons de tailles $n = 25, 50, 100$.

Les caractéristiques des populations et des échantillons sont exhibées dans le Tableau 4.1.

Tableau 4.1. Populations et échantillons de l'étude 1

β_0	β_1	c_i	Corrélation	N	n
200	0	1	0,6, 0,9	100	10, 20, 40
				500	25, 50, 100
0	2	x	0,6, 0,9	100	10, 20, 40
				500	25, 50, 100
0	2	$x^{3/2}$	0,6, 0,9	100	10, 20, 40
				500	25, 50, 100
200	2	1	0,6, 0,9	100	10, 20, 40
				500	25, 50, 100

4.1.2. Comparaison des approximations

Pour chaque échantillon, l'estimateur du total d'Horvitz-Thompson (1.1.4) a été calculé et sa variance estimée au moyen de la forme Sen-Yates-Grundy (1.1.8). À partir de la forme générale de l'estimateur de variance (1.4.16), différents estimateurs de variance ont été obtenus en y injectant les différentes approximations des π_{ij} (voir le Tableau 1.3 pour les coefficients correspondants). L'estimateur du total de Hájek (1.5.1) a également été calculé et sa variance estimée au moyen de la forme Sen-Yates-Grundy (1.5.3). Encore une fois, différents estimateurs de variance ont été obtenus à partir de la forme générale (1.5.4) au moyen des coefficients présentés au Tableau 1.3. En tout, 11 estimateurs de la variance approximative ont été calculés pour chaque estimateur ponctuel, soient l'estimateur d'Horvitz-Thompson et l'estimateur d'Hájek.

Afin de mesurer le biais des estimateurs de variance, nous avons calculé le biais relatif Monte Carlo (en pourcentage). Nous avons également calculé l'efficacité relative Monte Carlo des estimateurs de variance, avec comme référence l'estimateur Sen-Yates-Grundy, $\widehat{\mathrm{V}}_{SYG}$. Dans ce qui suit, $\widehat{Y}$ représente un estimateur du total de la population ($\widehat{Y}_{HT}$, $\widehat{Y}_{HA}$ ou $\widehat{Y}_c$) et $\widehat{\mathrm{V}}$ est une notation générique représentant n'importe quel estimateur de la variance. Le biais relatif Monte-Carlo de l'approximation de la variance $\widehat{\mathrm{V}}$ du total $\widehat{Y}$ est défini selon

$$\mathrm{RB}_{MC}(\widehat{\mathrm{V}}) = \frac{\mathrm{E}_{MC}(\widehat{\mathrm{V}}) - \mathrm{MSE}_{MC}(\widehat{Y})}{\mathrm{MSE}_{MC}(\widehat{Y})} \times 100, \tag{4.1.6}$$

où

$$\mathrm{MSE}_{MC}(\widehat{Y}) = \mathrm{E}_{MC}(\widehat{Y} - Y)^2,$$

et

$$\mathrm{E}_{MC}(\widehat{\theta}) = \frac{1}{K}\sum_{r=1}^{K} \widehat{\theta}^{(k)}.$$

Notons que $\widehat{\theta}^{(k)}$ est la valeur de l'estimateur $\widehat{\theta}$ à l'itération k, pour $k = 1,...,100\,000$. La stabilité relative est calculée selon

$$\mathrm{RS} = \frac{\mathrm{MSE}_{MC}(\widehat{\mathrm{V}})}{\mathrm{MSE}_{MC}(\widehat{\mathrm{V}}_{SYG})} \times 100$$

avec

$$\mathrm{MSE}_{MC}(\widehat{\mathrm{V}}) = \mathrm{E}_{MC}(\widehat{\mathrm{V}} - \mathrm{MSE}_{MC}(\widehat{Y}))^2.$$

4.1.3. Résultats

Les tableaux 4.2 à 4.17 présentent les résultats des approximations de la variance de l'estimateur du total d'Horvitz-Thompson et de l'estimateur de Hájek.

Pour chaque échantillon, nous présentons le biais relatif des approximations de la variance totale et la stabilité relative correspondante en dessous.

TABLEAU 4.2. Modèle (4.1.2) avec corrélation 0,9, pour $\widehat{Y}_{HT}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	0,07	-0,12	-0,31	-2,57	-2,84	-0,12	-0,09	-0,14	0,07	-10,11	-0,14
		99,88	97,36	96,38	93,65	93,27	97,34	97,41	97,29	99,83	82,58	97,30
100	20	0,17	0,00	-0,23	-2,75	-2,89	0,01	0,07	-0,06	0,17	-5,00	-0,05
		99,69	95,00	93,84	91,60	91,46	94,98	95,15	94,83	99,48	87,56	94,87
100	40	-0,24	-0,34	-0,64	-3,97	-4,07	-0,20	-0,02	-0,52	-0,20	-2,83	-0,48
		99,00	89,68	88,24	87,77	87,77	89,75	90,26	89,19	97,78	86,24	89,31
500	25	-0,22	-0,25	-0,28	-0,72	-0,74	-0,25	-0,25	-0,25	-0,22	-4,24	-0,25
		99,99	99,33	99,17	98,55	98,52	99,33	99,33	99,33	99,99	93,07	99,33
500	50	-0,24	-0,26	-0,29	-0,76	-0,77	-0,26	-0,26	-0,27	-0,24	-2,26	-0,27
		99,98	98,65	98,48	97,90	97,89	98,65	98,66	98,64	99,96	95,63	98,65
500	100	-0,27	-0,30	-0,33	-0,86	-0,87	-0,30	-0,29	-0,31	-0,27	-1,30	-0,31
		99,95	96,74	96,55	96,10	96,09	96,75	96,77	96,73	99,82	95,37	96,73

TABLEAU 4.3. Modèle (4.1.2) avec corrélation 0,9, pour $\widehat{Y}_{HA}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	-9,43	-9,28	-9,49	-11,45	-11,69	-9,43	-9,26	-9,46	-9,27	-18,36	-9,42
		99,93	100,32	99,32	98,20	97,99	100,03	100,38	99,97	100,29	90,81	100,05
100	20	-4,76	-3,94	-4,22	-6,47	-6,60	-4,76	-3,87	-4,84	-3,90	-8,74	-4,64
		99,78	100,50	99,25	98,28	98,18	99,16	100,65	99,01	101,23	94,61	99,35
100	40	-2,16	4,68	4,26	1,14	1,05	-2,07	5,01	-2,38	4,82	2,06	-1,28
		99,18	107,14	104,99	100,82	100,69	94,99	108,21	94,45	111,66	100,27	96,09
500	25	-4,80	-4,79	-4,86	-5,19	-5,21	-4,82	-4,79	-4,82	-4,77	-8,60	-4,82
		99,99	99,89	99,59	99,40	99,38	99,83	99,89	99,83	100,05	94,01	99,84
500	50	-2,93	-2,81	-2,89	-3,23	-3,24	-2,95	-2,81	-2,96	-2,79	-4,76	-2,92
		99,98	99,79	99,49	99,32	99,31	99,59	99,80	99,58	100,18	96,81	99,63
500	100	-1,21	-0,54	-0,63	-1,01	-1,01	-1,24	-0,53	-1,25	-0,51	-1,54	-1,09
		99,96	99,99	99,64	99,42	99,41	98,98	100,02	98,95	100,90	98,23	99,19

TABLEAU 4.4. Modèle (4.1.2) avec corrélation 0,6, pour $\widehat{Y}_{HT}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	0,04	0,02	0,00	-2,64	-2,94	0,03	0,05	0,00	0,04	-9,98	0,01
		99,93	99,27	99,10	94,61	94,15	99,29	99,32	99,25	99,93	85,96	99,25
100	20	0,22	0,22	0,20	-2,86	-3,02	0,25	0,29	0,17	0,22	-4,79	0,18
		99,80	97,76	97,46	93,32	93,16	97,83	97,89	97,70	99,78	91,20	97,71
100	40	-0,10	-0,05	-0,08	-4,46	-4,58	0,14	0,26	-0,17	-0,08	-2,55	-0,15
		99,21	90,99	90,28	90,79	90,94	91,39	91,56	90,84	99,00	88,41	90,88
500	25	-0,09	-0,07	-0,06	-0,59	-0,61	-0,07	-0,07	-0,07	-0,09	-4,07	-0,07
		100,00	100,20	100,20	99,24	99,20	100,20	100,21	100,20	100,00	94,27	100,20
500	50	-0,23	-0,22	-0,20	-0,77	-0,78	-0,22	-0,21	-0,22	-0,23	-2,21	-0,22
		100,00	100,28	100,27	99,37	99,35	100,29	100,29	100,28	100,00	97,53	100,28
500	100	-0,28	-0,27	-0,25	-0,91	-0,92	-0,27	-0,25	-0,28	-0,28	-1,26	-0,28
		99,97	99,41	99,35	98,69	98,69	99,43	99,44	99,40	99,99	98,27	99,40

TABLEAU 4.5. Modèle (4.1.2) avec corrélation 0,6, pour $\widehat{Y}_{HA}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	-6,95	-6,83	-6,93	-9,21	-9,48	-6,92	-6,80	-6,95	-6,85	-16,15	-6,92
		99,93	100,14	99,39	97,73	97,51	99,90	100,19	99,84	100,23	91,51	99,91
100	20	-3,25	-2,71	-2,84	-5,51	-5,66	-3,21	-2,63	-3,29	-2,71	-7,57	-3,16
		99,79	100,11	99,08	97,80	97,72	98,92	100,26	98,78	101,08	94,83	99,08
100	40	-1,10	3,38	3,18	-0,69	-0,79	-0,93	3,71	-1,24	3,44	0,80	-0,56
		99,14	104,20	102,40	97,97	97,87	93,93	105,26	93,38	109,88	97,57	94,89
500	25	-3,54	-3,52	-3,56	-3,97	-3,99	-3,54	-3,52	-3,54	-3,52	-7,38	-3,54
		100,00	99,90	99,69	99,36	99,34	99,85	99,90	99,84	100,05	94,71	99,86
500	50	-2,25	-2,19	-2,23	-2,66	-2,67	-2,27	-2,18	-2,27	-2,18	-4,15	-2,25
		99,99	99,77	99,53	99,29	99,28	99,57	99,78	99,56	100,19	97,24	99,61
500	100	-0,97	-0,59	-0,64	-1,13	-1,13	-0,99	-0,58	-1,00	-0,58	-1,59	-0,91
		99,96	99,62	99,34	99,08	99,07	98,68	99,65	98,65	100,87	98,09	98,88

TABLEAU 4.6. Modèle (4.1.3) avec corrélation 0,9, pour $\widehat{Y}_{HT}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	0,04	0,02	0,00	-2,64	-2,94	0,03	0,05	0,00	0,04	-9,98	0,01
		99,93	99,27	99,10	94,61	94,15	99,29	99,32	99,25	99,93	85,96	99,25
100	20	0,22	0,22	0,20	-2,86	-3,02	0,25	0,29	0,17	0,22	-4,79	0,18
		99,80	97,76	97,46	93,32	93,16	97,83	97,89	97,70	99,78	91,20	97,71
100	40	-0,10	-0,05	-0,08	-4,46	-4,58	0,14	0,26	-0,17	-0,08	-2,55	-0,15
		99,21	90,99	90,28	90,79	90,94	91,39	91,56	90,84	99,00	88,41	90,88
500	25	-0,09	-0,07	-0,06	-0,59	-0,61	-0,07	-0,07	-0,07	-0,09	-4,07	-0,07
		100,00	100,20	100,20	99,24	99,20	100,20	100,21	100,20	100,00	94,27	100,20
500	50	-0,23	-0,22	-0,20	-0,77	-0,78	-0,22	-0,21	-0,22	-0,23	-2,21	-0,22
		100,00	100,28	100,27	99,37	99,35	100,29	100,29	100,28	100,00	97,53	100,28
500	100	-0,28	-0,27	-0,25	-0,91	-0,92	-0,27	-0,25	-0,28	-0,28	-1,26	-0,28
		99,97	99,41	99,35	98,69	98,69	99,43	99,44	99,40	99,99	98,27	99,40

TABLEAU 4.7. Modèle (4.1.3) avec corrélation 0,9, pour $\widehat{Y}_{HA}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	-10,52	-10,33	-10,57	-12,43	-12,67	-10,51	-10,30	-10,53	-10,33	-19,30	-10,49
		99,94	100,77	99,77	98,85	98,66	100,45	100,83	100,38	100,35	92,02	100,48
100	20	-5,30	-4,32	-4,64	-6,77	-6,90	-5,29	-4,25	-5,36	-4,29	-9,11	-5,13
		99,80	101,18	99,92	99,10	99,02	99,68	101,35	99,53	101,47	95,52	99,90
100	40	-2,86	5,10	4,62	1,70	1,61	-2,77	5,43	-3,08	5,25	2,47	-1,78
		99,25	109,39	107,07	102,82	102,68	95,30	110,53	94,79	113,74	102,17	96,55
500	25	-5,20	-5,18	-5,25	-5,57	-5,58	-5,21	-5,17	-5,21	-5,16	-8,97	-5,21
		99,99	99,97	99,67	99,50	99,48	99,92	99,98	99,91	100,05	94,04	99,93
500	50	-3,16	-3,01	-3,10	-3,43	-3,44	-3,18	-3,01	-3,19	-3,00	-4,95	-3,14
		99,98	99,93	99,61	99,47	99,46	99,72	99,94	99,71	100,19	96,88	99,76
500	100	-1,27	-0,44	-0,53	-0,90	-0,91	-1,29	-0,43	-1,31	-0,41	-1,43	-1,11
		99,96	100,28	99,92	99,71	99,71	99,24	100,31	99,21	100,96	98,47	99,45

TABLEAU 4.8. Modèle (4.1.3) avec corrélation 0,6, pour $\widehat{Y}_{HT}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	0,04	0,02	0,00	-2,64	-2,94	0,03	0,05	0,00	0,04	-9,98	0,01
		99,93	99,27	99,10	94,61	94,15	99,29	99,32	99,25	99,93	85,96	99,25
100	20	0,22	0,22	0,20	-2,86	-3,02	0,25	0,29	0,17	0,22	-4,79	0,18
		99,80	97,76	97,46	93,32	93,16	97,83	97,89	97,70	99,78	91,20	97,71
100	40	-0,10	-0,05	-0,08	-4,46	-4,58	0,14	0,26	-0,17	-0,08	-2,55	-0,15
		99,21	90,99	90,28	90,79	90,94	91,39	91,56	90,84	99,00	88,41	90,88
500	25	-0,09	-0,07	-0,06	-0,59	-0,61	-0,07	-0,07	-0,07	-0,09	-4,07	-0,07
		100,00	100,20	100,20	99,24	99,20	100,20	100,21	100,20	100,00	94,27	100,20
500	50	-0,23	-0,22	-0,20	-0,77	-0,78	-0,22	-0,21	-0,22	-0,23	-2,21	-0,22
		100,00	100,28	100,27	99,37	99,35	100,29	100,29	100,28	100,00	97,53	100,28
500	100	-0,28	-0,27	-0,25	-0,91	-0,92	-0,27	-0,25	-0,28	-0,28	-1,26	-0,28
		99,97	99,41	99,35	98,69	98,69	99,43	99,44	99,40	99,99	98,27	99,40

TABLEAU 4.9. Modèle (4.1.3) avec corrélation 0,6, pour $\widehat{Y}_{HA}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	-10,44	-10,25	-10,48	-12,37	-12,60	-10,42	-10,22	-10,45	-10,25	-19,22	-10,40
		99,94	100,80	99,81	98,88	98,69	100,47	100,86	100,41	100,36	92,16	100,50
100	20	-5,26	-4,28	-4,59	-6,75	-6,88	-5,24	-4,21	-5,32	-4,25	-9,07	-5,09
		99,81	101,23	99,98	99,16	99,08	99,70	101,39	99,55	101,50	95,64	99,94
100	40	-2,76	5,18	4,72	1,74	1,66	-2,66	5,52	-2,97	5,33	2,55	-1,68
		99,24	109,64	107,32	102,90	102,76	95,21	110,79	94,70	114,10	102,31	96,49
500	25	-5,17	-5,15	-5,22	-5,54	-5,56	-5,19	-5,15	-5,19	-5,14	-8,94	-5,18
		99,99	99,96	99,66	99,49	99,47	99,91	99,97	99,90	100,05	94,04	99,92
500	50	-3,16	-3,02	-3,10	-3,43	-3,44	-3,18	-3,01	-3,18	-3,00	-4,96	-3,14
		99,98	99,92	99,60	99,45	99,44	99,71	99,93	99,70	100,19	96,88	99,75
500	100	-1,27	-0,46	-0,55	-0,92	-0,92	-1,29	-0,44	-1,31	-0,43	-1,45	-1,11
		99,96	100,25	99,89	99,68	99,68	99,21	100,28	99,18	100,95	98,44	99,43

TABLEAU 4.10. Modèle (4.1.4) avec corrélation 0,9, pour $\widehat{Y}_{HT}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	0,29	0,07	-0,28	-2,18	-2,43	-0,08	0,10	-0,11	0,44	-9,94	-0,07
		99,89	98,33	97,02	94,79	94,41	98,04	98,40	97,96	100,15	80,61	98,05
100	20	0,34	0,74	0,34	-1,73	-1,86	-0,04	0,82	-0,12	1,14	-4,30	0,08
		99,73	98,23	96,75	94,61	94,42	96,92	98,40	96,75	100,90	88,98	97,10
100	40	-0,90	4,76	4,23	1,55	1,47	-1,15	5,10	-1,47	5,19	2,14	-0,44
		99,10	102,93	101,01	98,14	98,03	94,48	103,73	93,87	107,36	97,33	95,36
500	25	-0,07	-0,11	-0,18	-0,53	-0,54	-0,14	-0,11	-0,14	-0,04	-4,10	-0,14
		99,99	99,51	99,19	98,88	98,85	99,45	99,52	99,45	100,04	91,99	99,46
500	50	-0,37	-0,30	-0,38	-0,74	-0,75	-0,45	-0,30	-0,46	-0,23	-2,30	-0,42
		99,98	99,33	98,99	98,71	98,70	99,10	99,34	99,09	100,17	95,56	99,15
500	100	0,04	0,71	0,62	0,23	0,22	-0,04	0,72	-0,06	0,79	-0,30	0,12
		99,95	99,49	99,13	98,81	98,80	98,44	99,52	98,41	100,88	97,49	98,66

TABLEAU 4.11. Modèle (4.1.4) avec corrélation 0,9, pour $\widehat{Y}_{HA}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	-9,19	-9,09	-9,33	-11,22	-11,46	-9,23	-9,06	-9,26	-9,03	-18,18	-9,22
		99,91	99,79	98,66	97,32	97,07	99,54	99,85	99,48	100,20	87,82	99,56
100	20	-4,71	-3,97	-4,29	-6,44	-6,57	-4,76	-3,90	-4,83	-3,89	-8,78	-4,64
		99,76	99,84	98,49	97,23	97,11	98,73	100,00	98,58	100,92	92,85	98,87
100	40	-2,09	4,26	3,80	0,86	0,77	-2,04	4,59	-2,36	4,43	1,65	-1,32
		99,12	104,72	102,68	99,41	99,30	95,63	105,64	95,06	108,37	98,62	96,45
500	25	-4,80	-4,80	-4,88	-5,19	-5,21	-4,83	-4,80	-4,83	-4,77	-8,61	-4,82
		99,99	99,83	99,53	99,32	99,30	99,78	99,84	99,78	100,04	93,61	99,79
500	50	-2,84	-2,74	-2,83	-3,15	-3,16	-2,88	-2,74	-2,88	-2,71	-4,69	-2,85
		99,98	99,73	99,42	99,23	99,22	99,54	99,74	99,53	100,16	96,57	99,58
500	100	-1,22	-0,59	-0,69	-1,05	-1,05	-1,26	-0,58	-1,28	-0,55	-1,59	-1,12
		99,95	99,88	99,53	99,29	99,29	98,95	99,90	98,92	100,80	98,09	99,15

TABLEAU 4.12. Modèle (4.1.4) avec corrélation 0,6, pour $\widehat{Y}_{HT}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	0,31	-0,02	-0,40	-2,23	-2,48	-0,09	0,01	-0,12	0,38	-10,02	-0,10
		99,89	98,05	96,64	94,66	94,29	97,86	98,12	97,79	100,01	80,20	97,86
100	20	0,40	0,36	-0,08	-2,04	-2,17	0,00	0,44	-0,07	0,76	-4,66	0,03
		99,72	97,71	96,14	94,29	94,12	96,92	97,88	96,75	100,30	88,50	96,98
100	40	-0,83	1,43	0,87	-1,56	-1,64	-1,08	1,75	-1,40	1,81	-1,11	-0,91
		99,07	99,28	97,43	95,50	95,41	94,87	99,97	94,26	102,91	94,36	95,20
500	25	-0,09	-0,14	-0,20	-0,57	-0,58	-0,15	-0,14	-0,16	-0,07	-4,13	-0,15
		99,99	99,45	99,14	98,81	98,79	99,41	99,46	99,41	100,02	92,02	99,42
500	50	-0,22	-0,22	-0,29	-0,66	-0,67	-0,29	-0,21	-0,29	-0,15	-2,21	-0,27
		99,98	99,19	98,86	98,56	98,55	99,04	99,20	99,02	100,09	95,46	99,06
500	100	-0,06	0,22	0,14	-0,27	-0,28	-0,13	0,24	-0,15	0,29	-0,78	-0,06
		99,95	99,00	98,65	98,34	98,33	98,31	99,02	98,28	100,50	97,07	98,44

TABLEAU 4.13. Modèle (4.1.4) avec corrélation 0,6, pour $\widehat{Y}_{HA}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	-5,80	-5,97	-6,26	-8,14	-8,38	-6,03	-5,95	-6,06	-5,75	-15,38	-6,04
		99,88	98,17	96,89	95,23	94,92	98,06	98,23	97,99	99,95	83,34	98,04
100	20	-3,00	-2,92	-3,28	-5,36	-5,49	-3,22	-2,85	-3,30	-2,69	-7,78	-3,21
		99,72	97,55	96,11	94,60	94,46	97,02	97,70	96,87	100,08	89,63	97,03
100	40	-1,19	0,99	0,51	-2,19	-2,27	-1,29	1,31	-1,60	1,25	-1,53	-1,20
		99,04	97,73	95,95	94,06	93,98	94,35	98,41	93,76	101,87	92,97	94,46
500	25	-3,49	-3,53	-3,61	-3,93	-3,95	-3,55	-3,53	-3,55	-3,48	-7,39	-3,55
		99,99	99,55	99,26	98,99	98,96	99,52	99,55	99,52	100,01	92,99	99,53
500	50	-1,96	-1,96	-2,04	-2,38	-2,39	-2,02	-1,96	-2,02	-1,90	-3,92	-2,01
		99,98	99,29	98,99	98,74	98,73	99,18	99,30	99,17	100,07	95,97	99,20
500	100	-1,00	-0,77	-0,86	-1,23	-1,23	-1,06	-0,76	-1,07	-0,71	-1,76	-1,00
		99,95	98,94	98,61	98,36	98,35	98,40	98,97	98,37	100,35	97,21	98,51

TABLEAU 4.14. Modèle (4.1.5) avec CV(y) = 0,9, pour $\widehat{Y}_{HT}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	0,21	0,05	-0,27	-2,23	-2,48	-0,13	0,08	-0,16	0,41	-9,95	-0,11
		99,90	98,41	97,22	94,75	94,35	98,04	98,48	97,96	100,26	81,11	98,07
100	20	0,21	0,81	0,43	-1,71	-1,84	-0,15	0,89	-0,23	1,19	-4,23	0,01
		99,75	98,26	96,89	94,47	94,28	96,52	98,44	96,35	101,38	89,15	96,78
100	40	-0,76	6,48	5,97	3,13	3,04	-1,00	6,83	-1,31	6,93	3,82	-0,05
		99,13	105,79	103,79	99,78	99,64	93,06	106,72	92,46	111,95	99,39	94,42
500	25	-0,16	-0,20	-0,28	-0,62	-0,63	-0,24	-0,20	-0,24	-0,13	-4,20	-0,23
		99,99	99,50	99,19	98,87	98,84	99,44	99,51	99,43	100,05	92,01	99,45
500	50	-0,45	-0,36	-0,45	-0,79	-0,80	-0,53	-0,36	-0,54	-0,28	-2,36	-0,50
		99,98	99,31	98,99	98,69	98,67	99,05	99,32	99,04	100,20	95,56	99,11
500	100	0,03	0,80	0,71	0,33	0,33	-0,06	0,82	-0,07	0,89	-0,21	0,13
		99,95	99,54	99,18	98,84	98,83	98,30	99,57	98,27	101,06	97,52	98,56

TABLEAU 4.15. Modèle (4.1.5) avec CV(y) = 0,9, pour $\widehat{Y}_{HA}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	-3,52	-3,90	-4,24	-6,05	-6,29	-3,89	-3,87	-3,92	-3,53	-13,51	-3,92
		99,87	96,89	95,51	93,89	93,57	96,90	96,95	96,84	99,73	81,27	96,85
100	20	-1,77	-2,16	-2,57	-4,53	-4,65	-2,14	-2,09	-2,21	-1,79	-7,05	-2,20
		99,70	95,60	94,08	92,70	92,55	95,61	95,75	95,46	99,30	87,51	95,49
100	40	-1,21	-1,57	-2,11	-4,51	-4,58	-1,42	-1,26	-1,73	-1,26	-4,04	-1,69
		99,04	92,73	91,06	90,09	90,03	92,92	93,31	92,35	97,53	88,73	92,45
500	25	-2,10	-2,17	-2,24	-2,59	-2,60	-2,17	-2,17	-2,17	-2,10	-6,08	-2,17
		99,99	99,13	98,88	98,52	98,50	99,13	99,13	99,13	99,97	92,78	99,13
500	50	-1,07	-1,14	-1,22	-1,58	-1,59	-1,14	-1,14	-1,15	-1,08	-3,12	-1,15
		99,98	98,46	98,19	97,87	97,85	98,46	98,47	98,45	99,91	95,26	98,45
500	100	-0,72	-0,79	-0,87	-1,27	-1,27	-0,79	-0,78	-0,80	-0,72	-1,78	-0,80
		99,95	96,99	96,71	96,44	96,44	97,00	97,02	96,98	99,68	95,46	96,98

TABLEAU 4.16. Modèle (4.1.5) avec CV(y) = 0,6, pour $\widehat{Y}_{HT}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	0,27	0,07	-0,27	-2,19	-2,44	-0,09	0,10	-0,12	0,44	-9,94	-0,08
		99,90	98,36	97,07	94,79	94,41	98,04	98,43	97,97	100,18	80,72	98,06
100	20	0,31	0,78	0,38	-1,71	-1,84	-0,06	0,85	-0,14	1,17	-4,26	0,07
		99,74	98,25	96,80	94,59	94,41	96,84	98,43	96,67	101,01	89,03	97,04
100	40	-0,87	5,29	4,76	2,04	1,95	-1,12	5,63	-1,44	5,72	2,66	-0,34
		99,11	103,62	101,68	98,56	98,44	94,20	104,44	93,59	108,40	97,84	95,19
500	25	-0,09	-0,13	-0,20	-0,54	-0,56	-0,16	-0,13	-0,17	-0,05	-4,12	-0,16
		99,99	99,51	99,19	98,88	98,85	99,45	99,52	99,44	100,04	92,00	99,46
500	50	-0,40	-0,32	-0,40	-0,75	-0,76	-0,48	-0,32	-0,48	-0,24	-2,32	-0,44
		99,98	99,33	98,99	98,71	98,69	99,09	99,34	99,08	100,18	95,57	99,14
500	100	0,04	0,75	0,66	0,27	0,27	-0,04	0,77	-0,06	0,84	-0,26	0,13
		99,95	99,51	99,15	98,83	98,82	98,40	99,54	98,37	100,94	97,50	98,64

TABLEAU 4.17. Modèle (4.1.5) avec CV(y) = 0,6, pour $\widehat{Y}_{HA}$

N	n	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_{PO}$	$\widehat{\mathrm{V}}_R$
100	10	-3,52	-3,90	-4,24	-6,05	-6,29	-3,89	-3,87	-3,92	-3,53	-13,51	-3,92
		99,87	96,89	95,51	93,89	93,57	96,90	96,95	96,84	99,73	81,27	96,85
100	20	-1,77	-2,16	-2,57	-4,53	-4,65	-2,14	-2,09	-2,21	-1,79	-7,05	-2,20
		99,70	95,60	94,08	92,70	92,55	95,61	95,75	95,46	99,30	87,51	95,49
100	40	-1,21	-1,57	-2,11	-4,51	-4,58	-1,42	-1,26	-1,73	-1,26	-4,04	-1,69
		99,04	92,73	91,06	90,09	90,03	92,92	93,31	92,35	97,53	88,73	92,45
500	25	-2,10	-2,17	-2,24	-2,59	-2,60	-2,17	-2,17	-2,17	-2,10	-6,08	-2,17
		99,99	99,13	98,88	98,52	98,50	99,13	99,13	99,13	99,97	92,78	99,13
500	50	-1,07	-1,14	-1,22	-1,58	-1,59	-1,14	-1,14	-1,15	-1,08	-3,12	-1,15
		99,98	98,46	98,19	97,87	97,85	98,46	98,47	98,45	99,91	95,26	98,45
500	100	-0,72	-0,79	-0,87	-1,27	-1,27	-0,79	-0,78	-0,80	-0,72	-1,78	-0,80
		99,95	96,99	96,71	96,44	96,44	97,00	97,02	96,98	99,68	95,46	96,98

4.1.4. Discussion

Nous commençons par discuter du biais des estimateurs relatifs à l'estimateur d'Horvitz-Thompson. Les estimateurs $\widehat{\mathrm{V}}_B$, $\widehat{\mathrm{V}}_{B1}$, $\widehat{\mathrm{V}}_{B2}$, $\widehat{\mathrm{V}}_{D1}$ $\widehat{\mathrm{V}}_{D2}$, $\widehat{\mathrm{V}}_H$, $\widehat{\mathrm{V}}_{HR}$, et $\widehat{\mathrm{V}}_R$ exhibent des biais négligeables (inférieurs à 1% en valeur absolue) dans tous les scénarios. Les estimateurs $\widehat{\mathrm{V}}_{B3}$ et $\widehat{\mathrm{V}}_{B4}$ exhibent un léger biais lorsque $N = 100$ et le biais tend à augmenter lorsque la fraction de sondage n/N augmente. Pour les populations de taille $N = 500$, le biais de ces estimateurs diminue avec des valeurs inférieures à 1% en valeur absolue. Donc, il est peut-être préférable de ne pas utiliser $\widehat{\mathrm{V}}_{B3}$ et $\widehat{\mathrm{V}}_{B4}$ pour des populations de petites tailles telles que $N = 100$. Dans le cas de l'estimateur $\widehat{V}_{PO}$, on note que le biais peut être appréciable pour de petites tailles d'échantillon, mais que ce dernier diminue rapidement au fur et à mesure que la taille de l'échantillon augmente. Ce résultat est attendu puisque l'estimateur $\widehat{V}_{PO}$ repose sur une approximation par série de Taylor, qui requiert une taille d'échantillon suffisamment grande.

Plus généralement, on remarque que le biais diminue au fur et à mesure que la taille de la population N et celle de l'échantillon n augmentent. Ce résultat est cohérent avec les résultats existant dans la litérature qui suggèrent que les approximations des probabilités d'inclusion d'ordre deux, π_{ij}, sont précises pour des plans à grande entropie dans le cas de N et n grands. En fait, pour $N = 500$ tous les estimateurs ont une performance similaire en termes de biais si bien qu'il n'est pas aisé d'identifier la meilleure approximation dans ce cas.

En ce qui concerne l'efficacité relative des estimateurs de variance, on note que, dans la très grande majorité des cas, cette dernière est proche de 100 bien que légérement inférieure à 100. Cela est particulièrement vrai pour $N = 500$, ce qui suggère que les estimateurs approximatifs de la variance exhibent une stabilité très légèrement supérieure à celle de l'estimateur de variance de Sen-Yates-Grundy. La seule exception concerne l'estimateur $\widehat{V}_{PO}$ qui exhibe une stabilité relative significativement inférieure à 100 avec des valeurs avoisinant 80% dans certains cas.

Pour ce qui est du biais des estimateurs relatifs à l'estimateur de Hájek, nous remarquons des biais appréciables pour de petites tailles d'échantillon, ce qui n'est pas surprenant puisque l'estimation ponctuelle ainsi que l'estimation de la variance reposent sur des approximations de Taylor du premier ordre qui requièrent une taille d'échantillon suffisamment grande. Le biais relatif diminue lorsque la taille de l'échantillon n et la taille de la population N augmentent. En termes de stabilité, on remarque que les estimateurs approximatifs exhibent tous des valeurs très proches de 100 dans tous scénarios.

Finalement, on note que le modèle utilisé afin de générer la variable y, le coefficient de corrélation entre x et y et le coefficient de variation de y ne semblent pas affecter les approximations de la variance en termes de biais relatif et de stabilité.

4.2. Étude 2 : en présence de données imputées

Une seconde étude par simulation a été réalisée afin de comparer la performance de différentes approximations de la variance pour l'estimateur du total dans le cadre d'un jeu de données avec des valeurs imputées. Dans cette étude, notre but est d'étudier la performance des approximations de la variance totale pour l'approche à deux-phases et pour l'approche renversée (voir Chapitre 3) et ce, pour différentes populations et mécanismes de réponse.

4.2.1. Populations et échantillons simulés

Un ensemble de différentes populations de taille $N = 500$ a été utilisé pour réaliser l'étude. Pour ce faire, nous avons utilisé le modèle général suivant

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i, \tag{4.2.1}$$

où $\boldsymbol{x}_i = (1, x_{1i}, x_{2i})'$ est un vecteur de variables auxiliaires généré, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ et un vecteur de paramètres, $\mathrm{E}_m(\epsilon_i) = 0$, $\mathrm{E}_m(\epsilon_i\epsilon_j) = 0$ pour $i \neq j$ et $\mathrm{V}_m(\epsilon_i) = \sigma^2 c_i$. Les populations ont été générées en faisant varier le vecteur de paramètres $\boldsymbol{\beta}$, le coefficient c_i ainsi que la composante σ^2 servant à fixer le coefficient de corrélation entre la variable y et la variable x.

Pour chaque population, nous avons d'abord généré la variable auxiliaire x_1 selon une loi Gamma, de paramètres $\alpha_1' = 5$ et $\beta_1' = 20$, de manière à obtenir x_1 telle $\mathrm{E}(x_1) = 100$ et $\mathrm{V}(x) = 2000$. Une deuxième variable auxiliaire x_2 a été générée selon une loi Gamma, de paramètres $\alpha_2' = 10$ et $\beta_2' = 20$, de manière à obtenir x_2 telle $\mathrm{E}(x_2) = 200$ et $\mathrm{V}(x) = 4000$. Ensuite, la variable y a été générée selon 3 choix de paramètres $\boldsymbol{\beta}$ et de coefficients c_i :

$$y_i = 2x_{1i} + \epsilon_i, \text{ avec } c_i = x_i; \tag{4.2.2}$$

$$y_i = 200 + 2x_{1i} + \epsilon_i, \text{ avec } c_i = 1; \tag{4.2.3}$$

$$y_i = 2x_{1i} + 3x_{2i} + \epsilon_i, \text{ avec } c_i = 1. \tag{4.2.4}$$

Pour les populations sous le modèle (4.2.2), nous avons utilisé deux valeurs de σ^2 afin d'obtenir un coefficient de corrélation entre la variable y et la variable x égal à 0,6 et à 0,9.

Pour chacune des populations, nous avons tiré $K = 100\,000$ échantillons selon le plan de Poisson conditionnel décrit à la section 1.2.5 de taille $n = 50, 100$. Les probabilités d'inclusion d'ordre un, π_i, sont définies de manière à être proportionnelles à la variable auxiliaire x_1.

Dans chaque échantillon, la non-réponse à la variable d'intérêt y a été générée selon deux mécanismes de réponse : un mécanisme uniforme et un mécanisme non-uniforme. Pour le mécanisme de réponse uniforme, toutes les unités ont la même probabilité de réponse, i.e., $p_i = p_0$ pour tout i. Nous avons utilisé deux différents taux de réponses, soient $p_0 = 0{,}5$ et $p_0 = 0{,}8$. Pour le mécanisme de réponse non-uniforme, la probabilité de réponse p_i a été générée de manière à ce que la relation entre p_i et la variable auxiliaire x_{1i} soit

$$p_i = \frac{exp(\gamma_1 + x_{1i}\gamma_2)}{1 + exp(\gamma_1 + x_{1i}\gamma_2)},$$

où γ_1 et γ_2 sont des coefficients choisis de manière à obtenir un taux de réponse global moyen de 0,5 et de 0,8.

Pour remplacer les valeurs manquantes dans les populations obtenues au moyen du modèle (4.2.2), nous avons utilisé l'imputation par le ratio décrite à la section 2.3.3. Dans les populations obtenues au moyen du modèle (4.2.3), nous avons utilisé l'imputation par la régression linéaire présentée à la section 2.3.2. Pour les populations obtenues au moyen du modèle (4.2.4), nous avons utilisé l'imputation par la régression décrite à la section 2.3.1.

Les caractéristiques des populations et des échantillons sont exhibées dans le Tableau 4.18.

Tableau 4.18. Populations et échantillons de l'étude 2

β_0	β_1	β_2	c_i	Corrélation	Mécanisme de réponse	Taux de réponse global	N	n
0	2	0	x_i	0,9	Uniforme	0,5, 0,8	500	50, 100
					Logistique	0,5, 0,8	500	50, 100
0	2	0	x_i	0,6	Uniforme	0,5, 0,8	500	50, 100
					Logistique	0,5, 0,8	500	50, 100
200	2	0	1	0,9	Uniforme	0,5, 0,8	500	50, 100
					Logistique	0,5, 0,8	500	50, 100
0	2	3	1	0,9	Uniforme	0,5, 0,8	500	50, 100
					Logistique	0,5, 0,8	500	50, 100

4.2.2. Comparaison des approximations

Pour chaque échantillon, l'estimateur imputé du total (2.2.1) a été calculé. Les approximations de la variance selon l'approche à deux-phases de Beaumont et Bocci (3.2.18) avec les différents coefficients présentés dans le tableau 1.3 ont été calculés, à l'exception de l'estimateur $\widehat{\mathrm{V}}_{PO}$. Les approximations de la variance selon l'approche renversée (3.3.7) avec les différents coefficients présentés dans le tableau 1.3 ont aussi été calculées, à l'exception de l'estimateur $\widehat{\mathrm{V}}_{PO}$. Cela a conduit à un ensemble de 10 estimateurs approximatifs de la variance.

Les approximations de variance sont comparées de deux façons, soient par le biais relatif Monte-Carlo et par le taux de couverture Monte Carlo des intervalles de confiance à 95%. Le biais relatif Monte-Carlo est calculé selon (4.1.6) avec chaque approximation de la variance totale $\widehat{\mathrm{V}}$ et l'estimateur du total imputé $\widehat{Y}_I$. Pour calculer le taux de couverture pour l'approximation de la variance $\widehat{\mathrm{V}}$, nous avons premièrement calculé un intervalle de confiance pour chaque estimateur $\widehat{Y}_I^{(k)}$ comme suit :

$$\widehat{Y}_I^{(k)} \pm 1{,}96\sqrt{\widehat{\mathrm{V}}}.$$

De plus, le taux de couverture pour une approximation de la variance $\widehat{\mathrm{V}}$ est la proportion des intervalles de confiance contenant le vrai total de la population Y. Afin d'évaluer si les taux de couverture sont significativement différents de 95%, on observe si le taux de couverture Monte Carlo est dans l'intervalle

$$0{,}95 \pm 1{,}96\sqrt{\frac{0{,}95 \times 0{,}05}{100\,000}},$$

ce qui correspond à l'intervalle $(0{,}9486; 0{,}9514)$. Dans la présentation des résultats, à la section 4.2.3, on notera un taux de couverture significativement différent de 95% en gras. Notons que nous n'avons pas mesurer la stabilité Monte Carlo des estimateurs de variance, contrairement à l'étude avec données complètes. Rappelons que dans la simulation avec données complètes, la stabilité relative est proche de 100 dans tous les scénarios. C'est pourquoi, il ne nous a pas semblé utile d'inclure cette mesure Monte Carlo dans cette section.

4.2.3. Résultats

Les tableaux 4.19 à 4.34 présentent les résultats des approximations de la variance totale pour les différents modèles de la variable y et les différents mécanismes de réponse (MR). Pour chaque taille d'échantillon, nous présentons le biais relatif des approximations de la variance totale et les taux de couverture correspondants en dessous, et ce, pour les deux approches (Ap.), soient l'approche à

deux-phases (D) et l'approche renversée (R). Rappelons que les taux de couverture en caractères gras sont les taux de couverture significativement différents de 95%.

TABLEAU 4.19. Modèle (4.2.2), corrélation 0,9, $N = 500$, MR uniforme de moyenne 0,5

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-3,20	-3,21	-3,22	-3,32	-3,33	-3,21	-3,21	-3,21	-3,20	-3,21
		93,56	**93,57**	**93,57**	**93,55**	**93,55**	**93,57**	**93,57**	**93,57**	**93,56**	**93,57**
	R	-2,17	-2,21	-2,26	-2,67	-2,68	-2,21	-2,21	-2,22	-2,17	-2,22
		93,52	**93,52**	**93,51**	**93,47**	**93,46**	**93,52**	**93,52**	**93,52**	**93,52**	**93,52**
100	D	-1,08	-1,09	-1,10	-1,21	-1,21	-1,09	-1,08	-1,09	-1,08	-1,09
		94,38	**94,38**	**94,37**	**94,36**	**94,36**	**94,38**	**94,38**	**94,37**	**94,38**	**94,37**
	R	-0,40	-0,44	-0,48	-0,93	-0,93	-0,43	-0,42	-0,45	-0,40	-0,45
		94,32	**94,32**	**94,31**	**94,26**	**94,26**	**94,32**	**94,32**	**94,32**	**94,32**	**94,32**

TABLEAU 4.20. Modèle (4.2.2), corrélation 0,9, $N = 500$, MR uniforme de moyenne 0,8

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-2,77	-2,81	-2,84	-3,10	-3,11	-2,81	-2,80	-2,81	-2,77	-2,81
		94,10	**94,11**	**94,10**	**94,07**	**94,07**	**94,10**	**94,11**	**94,10**	**94,10**	**94,10**
	R	-0,50	-0,55	-0,61	-1,02	-1,03	-0,55	-0,54	-0,56	-0,50	-0,56
		94,25	**94,26**	**94,25**	**94,20**	**94,19**	**94,26**	**94,26**	**94,25**	**94,25**	**94,25**
100	D	-2,34	-2,37	-2,40	-2,69	-2,70	-2,36	-2,35	-2,38	-2,34	-2,37
		94,43	**94,44**	**94,43**	**94,39**	**94,39**	**94,44**	**94,44**	**94,43**	**94,43**	**94,43**
	R	-0,10	-0,14	-0,20	-0,65	-0,66	-0,14	-0,12	-0,16	-0,10	-0,15
		94,63	**94,64**	**94,63**	**94,58**	**94,58**	**94,63**	**94,64**	**94,63**	**94,64**	**94,63**

TABLEAU 4.21. Modèle (4.2.2), corrélation 0,9, $N = 500$, MR logistique de moyenne 0,5

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-1,88	-1,89	-1,90	-2,00	-2,00	-1,89	-1,89	-1,89	-1,88	-1,89
		93,74	**93,74**	**93,74**	**93,72**	**93,72**	**93,74**	**93,74**	**93,74**	**93,74**	**93,74**
	R	-1,77	-1,81	-1,85	-2,27	-2,28	-1,81	-1,81	-1,82	-1,77	-1,82
		93,54	**93,54**	**93,54**	**93,50**	**93,49**	**93,54**	**93,54**	**93,54**	**93,54**	**93,54**
100	D	-0,61	-0,61	-0,62	-0,73	-0,73	-0,61	-0,61	-0,61	-0,60	-0,61
		94,46	**94,46**	**94,46**	**94,45**	**94,45**	**94,46**	**94,46**	**94,46**	**94,46**	**94,46**
	R	-1,00	-1,02	-1,06	-1,51	-1,51	-1,01	-1,00	-1,03	-0,99	-1,03
		94,32	**94,30**	**94,30**	**94,25**	**94,25**	**94,30**	**94,31**	**94,30**	**94,32**	**94,30**

TABLEAU 4.22. Modèle (4.2.2), corrélation 0,9, $N = 500$, MR logistique de moyenne 0,8

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-1,96	-1,98	-2,01	-2,28	-2,29	-1,98	-1,98	-1,99	-1,96	-1,99
		94,18	**94,19**	**94,18**	**94,14**	**94,14**	**94,18**	**94,19**	**94,18**	**94,18**	**94,18**
	R	-0,23	-0,27	-0,31	-0,73	-0,74	-0,27	-0,26	-0,27	-0,23	-0,27
		94,30	**94,29**	**94,28**	**94,22**	**94,22**	**94,29**	**94,29**	**94,29**	**94,30**	**94,29**
100	D	-1,91	-1,94	-1,97	-2,26	-2,27	-1,94	-1,93	-1,95	-1,91	-1,95
		94,43	**94,43**	**94,43**	**94,40**	**94,40**	**94,43**	**94,43**	**94,43**	**94,43**	**94,43**
	R	-0,30	-0,34	-0,39	-0,84	-0,85	-0,33	-0,32	-0,35	-0,29	-0,35
		94,58	**94,57**	**94,56**	**94,51**	**94,51**	**94,57**	**94,57**	**94,56**	**94,58**	**94,57**

TABLEAU 4.23. Modèle (4.2.2), corrélation 0,6, $N = 500$, MR uniforme de moyenne 0,5

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-3,20	-3,21	-3,22	-3,32	-3,33	-3,21	-3,21	-3,21	-3,20	-3,21
		93,56	**93,57**	**93,57**	**93,55**	**93,55**	**93,57**	**93,57**	**93,57**	**93,56**	**93,57**
	R	-2,17	-2,21	-2,26	-2,67	-2,68	-2,21	-2,21	-2,22	-2,17	-2,22
		93,52	**93,52**	**93,51**	**93,47**	**93,46**	**93,52**	**93,52**	**93,52**	**93,52**	**93,52**
100	D	-1,08	-1,09	-1,10	-1,21	-1,21	-1,09	-1,08	-1,09	-1,08	-1,09
		94,38	**94,38**	**94,37**	**94,36**	**94,36**	**94,38**	**94,38**	**94,37**	**94,38**	**94,37**
	R	-0,40	-0,44	-0,48	-0,93	-0,93	-0,43	-0,42	-0,45	-0,40	-0,45
		94,32	**94,32**	**94,31**	**94,26**	**94,26**	**94,32**	**94,32**	**94,32**	**94,32**	**94,32**

TABLEAU 4.24. Modèle (4.2.2), corrélation 0,6, $N = 500$, MR uniforme de moyenne 0,8

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-2,77	-2,81	-2,84	-3,10	-3,11	-2,81	-2,80	-2,81	-2,77	-2,81
		94,10	**94,11**	**94,10**	**94,07**	**94,07**	**94,10**	**94,11**	**94,10**	**94,10**	**94,10**
	R	-0,50	-0,55	-0,61	-1,02	-1,03	-0,55	-0,54	-0,56	-0,50	-0,56
		94,25	**94,26**	**94,25**	**94,20**	**94,19**	**94,26**	**94,26**	**94,25**	**94,25**	**94,25**
100	D	-2,34	-2,37	-2,40	-2,69	-2,70	-2,36	-2,35	-2,38	-2,34	-2,37
		94,43	**94,44**	**94,43**	**94,39**	**94,39**	**94,44**	**94,44**	**94,43**	**94,43**	**94,43**
	R	-0,10	-0,14	-0,20	-0,65	-0,66	-0,14	-0,12	-0,16	-0,10	-0,15
		94,63	**94,64**	**94,63**	**94,58**	**94,58**	**94,63**	**94,64**	**94,63**	**94,64**	**94,63**

TABLEAU 4.25. Modèle (4.2.2), corrélation 0,6, $N = 500$, MR logistique de moyenne 0,5

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-1,88	-1,89	-1,90	-2,00	-2,00	-1,89	-1,89	-1,89	-1,88	-1,89
		93,74	**93,74**	**93,74**	**93,72**	**93,72**	**93,74**	**93,74**	**93,74**	**93,74**	**93,74**
	R	-1,77	-1,81	-1,85	-2,27	-2,28	-1,81	-1,81	-1,82	-1,77	-1,82
		93,54	**93,54**	**93,54**	**93,50**	**93,49**	**93,54**	**93,54**	**93,54**	**93,54**	**93,54**
100	D	-0,61	-0,61	-0,62	-0,73	-0,73	-0,61	-0,61	-0,61	-0,60	-0,61
		94,46	**94,46**	**94,46**	**94,45**	**94,45**	**94,46**	**94,46**	**94,46**	**94,46**	**94,46**
	R	-1,00	-1,02	-1,06	-1,51	-1,51	-1,01	-1,00	-1,03	-0,99	-1,03
		94,32	**94,30**	**94,30**	**94,25**	**94,25**	**94,30**	**94,31**	**94,30**	**94,32**	**94,30**

TABLEAU 4.26. Modèle (4.2.2), corrélation 0,6, $N = 500$, MR logistique de moyenne 0,8

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-1,96	-1,98	-2,01	-2,28	-2,29	-1,98	-1,98	-1,99	-1,96	-1,99
		94,18	**94,19**	**94,18**	**94,14**	**94,14**	**94,18**	**94,19**	**94,18**	**94,18**	**94,18**
	R	-0,23	-0,27	-0,31	-0,73	-0,74	-0,27	-0,26	-0,27	-0,23	-0,27
		94,30	**94,29**	**94,28**	**94,22**	**94,22**	**94,29**	**94,29**	**94,29**	**94,30**	**94,29**
100	D	-1,91	-1,94	-1,97	-2,26	-2,27	-1,94	-1,93	-1,95	-1,91	-1,95
		94,43	**94,43**	**94,43**	**94,40**	**94,40**	**94,43**	**94,43**	**94,43**	**94,43**	**94,43**
	R	-0,30	-0,34	-0,39	-0,84	-0,85	-0,33	-0,32	-0,35	-0,29	-0,35
		94,58	**94,57**	**94,56**	**94,51**	**94,51**	**94,57**	**94,57**	**94,56**	**94,58**	**94,57**

TABLEAU 4.27. Modèle (4.2.3), corrélation 0,9, $N = 500$, MR uniforme de moyenne 0,5

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-1,47	-1,38	-1,44	-1,71	-1,72	-1,52	-1,38	-1,53	-1,33	-1,49
		92,92	**92,98**	**92,97**	**92,92**	**92,92**	**92,96**	**92,98**	**92,96**	**92,94**	**92,96**
	R	0,67	0,74	0,67	0,28	0,27	0,60	0,75	0,59	0,81	0,63
		92,96	**92,98**	**92,98**	**92,93**	**92,93**	**92,97**	**92,98**	**92,97**	**92,97**	**92,97**
100	D	-0,42	0,22	0,15	-0,14	-0,14	-0,47	0,23	-0,49	0,27	-0,33
		93,85	**93,95**	**93,94**	**93,91**	**93,91**	**93,87**	**93,95**	**93,87**	**93,93**	**93,88**
	R	-2,41	-1,77	-1,84	-2,26	-2,26	-2,48	-1,76	-2,50	-1,70	-2,34
		93,48	**93,59**	**93,58**	**93,53**	**93,53**	**93,50**	**93,59**	**93,50**	**93,57**	**93,52**

TABLEAU 4.28. Modèle (4.2.3), corrélation 0,9, $N = 500$, MR uniforme de moyenne 0,8

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-0,02	0,08	0,01	-0,33	-0,33	-0,08	0,09	-0,09	0,15	-0,05
		92,99	**93,02**	**93,01**	**92,98**	**92,97**	**93,00**	**93,02**	**93,00**	**93,01**	**93,00**
	R	1,23	1,32	1,24	0,86	0,85	1,15	1,33	1,15	1,39	1,19
		93,07	**93,11**	**93,11**	**93,05**	**93,05**	**93,09**	**93,11**	**93,09**	**93,09**	**93,10**
100	D	0,03	0,77	0,70	0,33	0,32	-0,04	0,79	-0,06	0,84	0,12
		93,66	**93,78**	**93,77**	**93,72**	**93,72**	**93,67**	**93,78**	**93,67**	**93,76**	**93,69**
	R	-0,45	0,30	0,22	-0,20	-0,20	-0,53	0,32	-0,55	0,37	-0,36
		93,53	**93,67**	**93,66**	**93,60**	**93,60**	**93,55**	**93,67**	**93,55**	**93,62**	**93,58**

TABLEAU 4.29. Modèle (4.2.3), corrélation 0,9, $N = 500$, MR logistique de moyenne 0,5

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-1,53	-1,44	-1,50	-1,76	-1,77	-1,58	-1,43	-1,58	-1,39	-1,55
		92,93	**92,98**	**92,98**	**92,93**	**92,93**	**92,97**	**92,98**	**92,97**	**92,95**	**92,97**
	R	0,62	0,69	0,63	0,24	0,23	0,55	0,70	0,55	0,76	0,58
		92,97	**93,01**	**93,01**	**92,95**	**92,95**	**93,00**	**93,01**	**93,00**	**92,99**	**93,00**
100	D	-0,47	0,17	0,10	-0,18	-0,19	-0,52	0,18	-0,53	0,22	-0,38
		93,85	**93,95**	**93,95**	**93,92**	**93,92**	**93,88**	**93,95**	**93,88**	**93,94**	**93,90**
	R	-2,49	-1,85	-1,92	-2,33	-2,34	-2,56	-1,83	-2,57	-1,78	-2,42
		93,48	**93,60**	**93,60**	**93,54**	**93,54**	**93,51**	**93,60**	**93,51**	**93,57**	**93,54**

TABLEAU 4.30. Modèle (4.2.3), corrélation 0,9, $N = 500$, MR logistique de moyenne 0,8

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-0,02	0,08	0,01	-0,33	-0,34	-0,08	0,09	-0,09	0,14	-0,05
		92,99	**93,02**	**93,01**	**92,98**	**92,97**	**93,00**	**93,02**	**93,00**	**93,01**	**93,00**
	R	1,23	1,32	1,24	0,86	0,85	1,15	1,32	1,15	1,39	1,19
		93,07	**93,11**	**93,11**	**93,05**	**93,05**	**93,10**	**93,11**	**93,09**	**93,09**	**93,10**
100	D	0,03	0,78	0,70	0,33	0,33	-0,04	0,80	-0,06	0,85	0,13
		93,66	**93,78**	**93,77**	**93,72**	**93,72**	**93,67**	**93,78**	**93,67**	**93,76**	**93,69**
	R	-0,45	0,30	0,22	-0,19	-0,20	-0,52	0,32	-0,54	0,38	-0,36
		93,53	**93,67**	**93,66**	**93,60**	**93,60**	**93,55**	**93,67**	**93,55**	**93,63**	**93,57**

TABLEAU 4.31. Modèle (4.2.4), corrélation 0,9, $N = 500$, MR uniforme de moyenne 0,5

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-0,38	-0,30	-0,37	-0,75	-0,76	-0,46	-0,29	-0,47	-0,22	-0,43
		92,92	**92,94**	**92,93**	**92,88**	**92,88**	**92,92**	**92,94**	**92,92**	**92,93**	**92,92**
	R	-0,38	-0,29	-0,36	-0,75	-0,75	-0,45	-0,28	-0,46	-0,21	-0,42
		92,92	**92,94**	**92,93**	**92,88**	**92,88**	**92,92**	**92,94**	**92,92**	**92,94**	**92,92**
100	D	0,30	1,08	0,99	0,56	0,56	0,22	1,10	0,20	1,16	0,39
		93,77	**93,92**	**93,90**	**93,83**	**93,83**	**93,79**	**93,92**	**93,79**	**93,87**	**93,81**
	R	0,31	1,09	1,00	0,58	0,57	0,24	1,11	0,22	1,17	0,40
		93,77	**93,92**	**93,91**	**93,83**	**93,83**	**93,79**	**93,92**	**93,79**	**93,87**	**93,82**

TABLEAU 4.32. Modèle (4.2.4), corrélation 0,9, $N = 500$, MR uniforme de moyenne 0,8

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-0,40	-0,27	-0,33	-0,75	-0,76	-0,46	-0,26	-0,47	-0,21	-0,42
		92,87	**92,89**	**92,89**	**92,84**	**92,84**	**92,87**	**92,90**	**92,87**	**92,89**	**92,88**
	R	-0,40	-0,27	-0,34	-0,75	-0,76	-0,46	-0,26	-0,47	-0,21	-0,43
		92,87	**92,90**	**92,89**	**92,84**	**92,83**	**92,87**	**92,90**	**92,87**	**92,90**	**92,88**
100	D	-1,02	-0,10	-0,17	-0,63	-0,64	-1,08	-0,08	-1,11	-0,03	-0,90
		93,65	**93,76**	**93,75**	**93,70**	**93,70**	**93,65**	**93,76**	**93,64**	**93,75**	**93,67**
	R	-1,02	-0,10	-0,17	-0,63	-0,64	-1,09	-0,08	-1,11	-0,03	-0,90
		93,65	**93,76**	**93,75**	**93,70**	**93,69**	**93,65**	**93,76**	**93,65**	**93,75**	**93,66**

TABLEAU 4.33. Modèle (4.2.4), corrélation 0,9, $N = 500$, MR logistique de moyenne 0,5

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-0,52	-0,48	-0,55	-0,92	-0,92	-0,59	-0,47	-0,60	-0,40	-0,57
		92,48	**92,50**	**92,49**	**92,44**	**92,44**	**92,48**	**92,50**	**92,48**	**92,48**	**92,49**
	R	-0,52	-0,48	-0,55	-0,92	-0,92	-0,59	-0,47	-0,60	-0,40	-0,57
		92,48	**92,50**	**92,49**	**92,45**	**92,45**	**92,48**	**92,50**	**92,48**	**92,49**	**92,49**
100	D	-0,62	-0,12	-0,20	-0,60	-0,61	-0,69	-0,11	-0,71	-0,05	-0,58
		93,45	**93,56**	**93,56**	**93,52**	**93,52**	**93,51**	**93,57**	**93,51**	**93,53**	**93,52**
	R	-0,62	-0,12	-0,20	-0,60	-0,61	-0,69	-0,11	-0,71	-0,05	-0,58
		93,45	**93,56**	**93,56**	**93,52**	**93,52**	**93,51**	**93,57**	**93,51**	**93,52**	**93,52**

TABLEAU 4.34. Modèle (4.2.4), corrélation 0,9, $N = 500$, MR logistique de moyenne 0,8

n	Ap,	$\widehat{\mathrm{V}}_B$	$\widehat{\mathrm{V}}_{B1}$	$\widehat{\mathrm{V}}_{B2}$	$\widehat{\mathrm{V}}_{B3}$	$\widehat{\mathrm{V}}_{B4}$	$\widehat{\mathrm{V}}_{D1}$	$\widehat{\mathrm{V}}_{D2}$	$\widehat{\mathrm{V}}_H$	$\widehat{\mathrm{V}}_{HR}$	$\widehat{\mathrm{V}}_R$
50	D	-0,50	-0,45	-0,53	-0,89	-0,90	-0,58	-0,45	-0,59	-0,37	-0,55
		92,65	**92,70**	**92,69**	**92,64**	**92,64**	**92,68**	**92,70**	**92,68**	**92,66**	**92,68**
	R	-0,49	-0,45	-0,53	-0,89	-0,90	-0,58	-0,45	-0,58	-0,37	-0,55
		92,64	**92,70**	**92,70**	**92,64**	**92,64**	**92,69**	**92,70**	**92,69**	**92,66**	**92,69**
100	D	0,17	0,73	0,64	0,25	0,24	0,09	0,75	0,08	0,81	0,22
		93,54	**93,64**	**93,62**	**93,57**	**93,57**	**93,56**	**93,64**	**93,56**	**93,61**	**93,57**
	R	0,18	0,74	0,65	0,25	0,25	0,10	0,75	0,08	0,82	0,23
		93,53	**93,63**	**93,62**	**93,58**	**93,58**	**93,56**	**93,64**	**93,56**	**93,61**	**93,57**

4.2.4. Discussion

Nous remarquons d'emblée que, pour un scénario donné, tous les estimateurs de variance exhibent des biais très similaires. Il en est de même pour la probabilité de couverture qui varie très peu d'un estimateur à l'autre pour un scénario donné. En termes de biais, nous constatons que tous les estimateurs de variance exhibent un biais relatif relativement faible avec des valeurs Monte Carlo inférieures à 4% en valeur absolue. Lorsque la taille de l'échantillon est petite, l'approche à deux-phases est légèrement meilleure en termes de biais relatifs. Nous remarquons aussi que le biais tend à diminuer au fur et à mesure que la taille de l'échantillon augmente. Ce résultat n'est pas surprenant car tous les estimateurs de variance en présence de données imputées reposent sur un développement par séries de Taylor, qui requiert une taille d'échantillon suffisamment grande.

Les taux de couverture des intervalles de confiance, sont tous statistiquement significativement différents de 95,00%. Par contre, des taux de couverture entre 92% et 95% sont généralement considérés comme acceptables. De plus, le fait d'observer des taux de couverture un peu plus faibles que 95,00% peut être expliqué par le fait que les estimateurs de variance exhibent un léger biais négatif dans la très grande majorité des scénarios, conduisant alors à des intervalles un peu trop courts.

Finalement, nous notons que le type de modèle ayant servi à générer la population, le coefficient de corrélation et le mécanisme de réponse ne semblent pas avoir un impact significatif sur le biais et le taux de couverture.

En conclusion, tous les estimateurs de variance semblent avoir des propriétés très similaires pour un scénario donné en termes de biais relatif et de taux de couverture. Choisir l'une des approximations n'est donc pas chose aisée car il est virtuellement impossible de distinguer les approximations dans un contexte de données imputées.

Chapitre 5

CONCLUSION

Dans ce mémoire, nous avons proposé différentes approximations de l'estimation de la variance. Ces approximations sont exprimées en fonction des probabilités d'inclusion d'ordre un et ne dépendent pas des probabilités d'inclusion d'ordre deux. Cette propriété permet d'éviter le calcul des probabilités π_{ij}, qui peut être difficile, voir impossible, dans certains cas. Les approximations de la variance sont aussi exprimées en une simple somme.

Nous avons étudié le comportement des différentes approximations de la variance pour l'estimateur du total d'Horvitz-Thompson et l'estimateur du total de Hájek, selon différentes tailles de population et modèles de la variable y. Nous avons remarqué que les approximations de la variance pour l'estimateur de Hájek sont généralement plus biaisées que celles de l'estimateur d'Horvitz-Thompson. Les approximations se comportent mieux lorsque la taille de la population et la taille de l'échantillon sont grandes.

Nous avons aussi développé des estimateurs de variance en présence de données imputées. Deux approches ont été utilisées pour estimer la variance, soient l'approche à deux-phases et l'approche renversée. Les approximations de la variance sous les deux approches permettent encore une fois de ne pas avoir recours aux probabilités d'inclusion d'ordre deux et de réduire la variance en une simple somme.

Nous avons étudié le comportement des différentes approximations de la variance en présence de données imputées avec l'approche à deux-phases et l'approche renversée, selon plusieurs populations. Les simulations diffèrent en termes de taille de population, de taille d'échantillon, de modèle de la variable y et de mécanisme de réponse. Nous avons remarqué que les approximations se comportent mieux lorsque la taille de l'échantillon et le taux de réponse sont élevés. Les approximations sous les deux approches performent bien en termes de biais relatif

et de taux de couverture. Lorsque la taille de l'échantillon est petite, l'approche à deux-phases est légèrement meilleure en termes de biais relatifs.

Les différentes approximations de la variance se comportent généralement de manière semblable et permettent une estimation de bonne qualité tout en ne requérant pas les probabilités d'ordre deux.

Dans ce mémoire, tous les échantillons ont été tirés selon un plan de sondage de Poisson conditionnel. Il serait intéressant d'étudier la performance des différentes approximations de la variance pour d'autres plans de sondage à grande entropie tels le plan de Rao-Sampford ou le plan systématique randomisé.

Bien que nous ayons développé des estimateurs de variance dans le cas de méthodes d'imputation déterministe, la généralisation au cas de méthodes aléatoires est relativement aisée. En effet, dans ce cas, on compte un terme de variance additionnel, la variance due à l'imputation, qui provient de la sélection aléatoire de valeurs imputées. Ce terme ne dépendant pas des probabilités d'inclusion d'ordre deux, il suffit de l'ajouter aux estimateurs de variance obtenus dans le cas de méthodes déterministes pour obtenir des estimateurs de variance approximativement sans biais dans le cas de méthodes aléatoires.

Bibliographie

Beaumont, J.-F. et Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, **37**, 400–416.

Berger, Y. G. (1998). Rate of convergence for asymptotic variance of the Horvitz–Thompson estimator. *Journal of Statistical Planning and Inference*, **74**, 149–168.

Brewer, K. et Donadio, M. E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, **29**, 189–196.

Chen, X.-H., Dempster, A. P. et Liu, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, **81**, 457–469.

Deville, J. (2000). Note sur l'algorithme de Chen, Dempster et Liu. *Rapport technique,*CREST-ENSAI, Rennes.

Deville, J.-C. et Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, **10**, 381–381.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 1491–1523.

Haziza, D., Mecatti, F. et Rao, J. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron International Journal of Statistics*, **1**, 89–106.

Ohlsson, E. (1998). Sequential Poisson sampling. *Journal of Official Statistics*, **14**, 149–162.

Rao, J. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, **3**, 173–80.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Sampford, M. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, **54**, 499–513.

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

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, **18**, 241–252.

Shao, J. et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, **94**, 254–265.

Tillé, Y. (2006). *Sampling Algorithms*. New York : Springer.

Annexe A

ESTIMATEUR DE V_{SAM} DANS LE CAS DE LA MÉTHODE DE SÄRNDAL

En examinant la variance (3.2.10) conditionnellement à s et à s_r, on a

$$\begin{aligned}\mathrm{V}_{diff} &= \mathrm{E}_m\left(\widehat{\mathrm{V}}_{HT} - \widehat{\mathrm{V}}_{naïf}\Big|s,s_r\right)\\ &= \mathrm{E}_m\left(\sum_{i\in s}\sum_{j\in s}\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j - \sum_{i\in s}\sum_{j\in s}\frac{\Omega_{ij}}{\pi_{ij}}\widetilde{y}_i\widetilde{y}_j\Big|s,s_r\right). \qquad \text{(A.0.5)}\end{aligned}$$

En détaillant (A.0.5) en termes de répondants et de non-répondants, on obtient

$$\begin{aligned}\mathrm{V}_{diff} = \mathrm{E}_m\Bigg(&\sum_{i\in s}\sum_{j\in s}r_ir_j\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j + \sum_{i\in s}\sum_{j\in s}r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j\\ &+\sum_{i\in s}\sum_{j\in s}(1-r_i)r_j\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j + \sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j\\ &-\sum_{i\in s}\sum_{j\in s}r_ir_j\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j \ -\sum_{i\in s}\sum_{j\in s}r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j^*\\ &-\sum_{i\in s}\sum_{j\in s}(1-r_i)r_j\frac{\Omega_{ij}}{\pi_{ij}}y_i^*y_j - \sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_i^*y_j^*\Big|s,s_r\Bigg)\\ = \mathrm{E}_m\Bigg(&2\sum_{i\in s}\sum_{j\in s}r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j + \sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j\\ &-2\sum_{i\in s}\sum_{j\in s}r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j^*\\ &-\sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_i^*y_j^*\Big|s,s_r\Bigg). \qquad \text{(A.0.6)}\end{aligned}$$

Détaillons les termes de (A.0.6) un à la fois. On a d'abord

$$\mathrm{E}_m\left(2\sum_{i\in s}\sum_{j\in s}r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j\Big|s,s_r\right) = 2\sum_{i\in s}\sum_{j\in s}r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}\mathrm{E}_m(y_iy_j)$$

$$= 2\sum_{i\in s}\sum_{j\in s} r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}\left\{\mathrm{Cov}_m(y_i,y_j) + \mathrm{E}_m(y_i)\mathrm{E}_m(y_j)\right\}$$

$$= 2\sum_{i\in s}\sum_{j\in s} r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}\boldsymbol{x}_i'\boldsymbol{\beta}\boldsymbol{x}_j'\boldsymbol{\beta}. \tag{A.0.7}$$

Puis, on a que

$$\mathrm{E}_m\left(\sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j\middle| s,s_r\right)$$

$$= \mathrm{E}_m\left(\sum_{i\in s}\sum_{\substack{j\in s\\ j\neq i}}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j + \sum_{i\in s}(1-r_i)\frac{\Omega_{ii}}{\pi_{ii}}y_i^2\middle| s,s_r\right)$$

$$= \sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}\boldsymbol{x}_i'\boldsymbol{\beta}\boldsymbol{x}_j'\boldsymbol{\beta} + \sigma^2\sum_{i\in s}(1-r_i)\frac{\Omega_{ii}}{\pi_{ii}}c_i. \tag{A.0.8}$$

On a aussi que

$$\mathrm{E}_m\left(2\sum_{i\in s}\sum_{j\in s} r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_iy_j^*\middle| s,s_r\right)$$

$$= \mathrm{E}_m\left(2\sum_{i\in s}\sum_{j\in s} r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_i\boldsymbol{x}_j'\widehat{\boldsymbol{T}}_r^{-1}\sum_{k\in s}\frac{r_k}{\pi_k}\boldsymbol{x}_kc_k^{-1}y_k\middle| s,s_r\right)$$

$$= 2\sum_{j\in s}(1-r_j)\left\{\sum_{i\in s} r_i\frac{\Omega_{ij}}{\pi_{ij}\pi_i}\boldsymbol{x}_j'\widehat{\boldsymbol{T}}_r^{-1}\boldsymbol{x}_ic_i^{-1}\mathrm{E}_m(y_i^2)\right.$$

$$\left. + \sum_{i\in s}\sum_{\substack{k\in s\\ k\neq i}} r_ir_k\frac{\Omega_{ij}}{\pi_{ij}\pi_k}\boldsymbol{x}_j'\widehat{\boldsymbol{T}}_r^{-1}\boldsymbol{x}_kc_k^{-1}\mathrm{E}_m(y_iy_k)\right\}$$

$$= 2\left\{\sum_{i\in s}\sum_{j\in s} r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}\boldsymbol{x}_i'\boldsymbol{\beta}\boldsymbol{x}_j'\boldsymbol{\beta}\right.$$

$$\left. +\sigma^2\sum_{i\in s}\sum_{j\in s} r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}\pi_i}\boldsymbol{x}_i'\widehat{\boldsymbol{T}}_r^{-1}\boldsymbol{x}_j\right\}. \tag{A.0.9}$$

Finalement, on a que

$$\mathrm{E}_m\left(\sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}y_i^*y_j^*\middle| s,s_r\right)$$

$$= \mathrm{E}_m\left(\sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}\boldsymbol{x}_i'\widehat{\mathbf{B}}_r\boldsymbol{x}_j'\widehat{\mathbf{B}}_r\middle| s,s_r\right)$$

$$= \sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}\boldsymbol{x}_i'\boldsymbol{\beta}\boldsymbol{x}_j'\boldsymbol{\beta}$$

$$+ \sigma^2 \sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}\boldsymbol{x}_i'\widehat{\boldsymbol{T}}_r^{-1}\sum_{k\in s}\frac{r_k}{\pi_k^2}\boldsymbol{x}_{\boldsymbol{k}}c_k^{-1}\boldsymbol{x}_j'\widehat{\boldsymbol{T}}_r^{-1}\boldsymbol{x}_k. \quad \text{(A.0.10)}$$

En additionnant les termes (A.0.7), (A.0.8), (A.0.9) et (A.0.10), on obtient

$$\mathrm{V}_{diff} = \sigma^2 \left\{ \sum_{i\in s}(1-r_i)\frac{\Omega_{ii}}{\pi_{ii}}c_i - 2\sum_{i\in s}\sum_{j\in s} r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}\pi_i}\boldsymbol{x}_{\boldsymbol{i}}'\widehat{\boldsymbol{T}}_r^{-1}\boldsymbol{x}_{\boldsymbol{j}} \right.$$
$$\left. - \sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}\boldsymbol{x}_i'\widehat{\boldsymbol{T}}_r^{-1}\sum_{k\in s}\frac{r_k}{\pi_k^2}\boldsymbol{x}_{\boldsymbol{k}}c_k^{-1}\boldsymbol{x}_j'\widehat{\boldsymbol{T}}_r^{-1}\boldsymbol{x}_k \right\}. \quad \text{(A.0.11)}$$

Un estimateur sans biais de V_{diff} est obtenu en estimant σ^2 par $\widehat{\sigma}^2$ en (3.2.13), dans (A.0.11). On obtient alors l'estimateur (3.2.12). Finalement, un estimateur de V_{sam} est donné par (3.2.14).

Annexe B

ESTIMATEUR DE V_{SAM} DANS LE CAS DE LA MÉTHODE DE BEAUMONT ET BOCCI

On commence par écrire la variance (3.2.10) conditionnellement à s, à s_r et à $\boldsymbol{y}_r$. On a alors

$$\begin{aligned}\mathrm{V}_{diff} &= \mathrm{E}_m\left(\widehat{\mathrm{V}}_{HT} - \widehat{\mathrm{V}}_{naïf}|s,s_r,\boldsymbol{y}_r\right)\\ &= \mathrm{E}_m\left(\widehat{\mathrm{V}}_{HT}|s,s_r,\boldsymbol{y}_r\right) - \widehat{\mathrm{V}}_{naïf},\end{aligned}$$

étant donné que $\widehat{\mathrm{V}}_{naïf}$ dépend uniquement des valeurs de y observées chez les répondants. On a alors

$$\begin{aligned}\mathrm{V}_{diff} &= \mathrm{E}_m\left(\widehat{\mathrm{V}}_{HT}|s,s_r,\boldsymbol{y}_r\right) - \widehat{\mathrm{V}}_{naïf}\\ &= \mathrm{E}_m\left(\sum_{i\in s}\sum_{j\in s} r_i r_j \frac{\Omega_{ij}}{\pi_{ij}} y_i y_j + 2\sum_{i\in s}\sum_{j\in s} r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}} y_i y_j\right.\\ &\quad \left. + \sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}} y_i y_j|s,s_r,\boldsymbol{y}_r\right) - \widehat{\mathrm{V}}_{naïf}\\ &= \sum_{i\in s}\sum_{j\in s} r_i r_j \frac{\Omega_{ij}}{\pi_{ij}} y_i y_j + 2\sum_{i\in s}\sum_{j\in s} r_i(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}} y_i \mathrm{E}_m(y_j|s,s_r,\boldsymbol{y}_r)\\ &\quad + \sum_{i\in s}\sum_{j\in s}(1-r_i)(1-r_j)\frac{\Omega_{ij}}{\pi_{ij}}\mathrm{E}_m(y_i y_j|s,s_r,\boldsymbol{y}_r) - \widehat{\mathrm{V}}_{naïf}\\ &= \sum_{i\in s}\sum_{j\in s}\frac{\Omega_{ij}}{\pi_{ij}}\widetilde{y}_i\widetilde{y}_j + \sigma^2\sum_{i\in s}(1-r_i)\frac{\Omega_{ii}}{\pi_i}c_i - \widehat{\mathrm{V}}_{naïf}\\ &= \sigma^2\sum_{i\in s}(1-r_i)\frac{\Omega_{ii}}{\pi_i}c_i.\end{aligned}$$

En estimant σ^2 par $\widehat{\sigma}^2$ en (3.2.13), on obtient l'estimateur

$$\widehat{\mathrm{V}}_{diff} = \widehat{\sigma}^2\sum_{i\in s}(1-r_i)\frac{1-\pi_i}{\pi_i^2}c_i.$$

B-ii

L'estimateur de V_{sam} est donc

$$\widehat{\mathrm{V}}_{sam} = \sum_{i\in s}\sum_{j\in s}\frac{\Omega_{ij}}{\pi_{ij}}\widetilde{y}_i\widetilde{y}_j + \widehat{\sigma}^2\sum_{i\in s}(1-r_i)\frac{1-\pi_i}{\pi_i^2}c_i.$$

Annexe C

ESTIMATEUR DE V_{NR}

Afin de calculer V_{NR}, examinons d'abord l'erreur due à la non-réponse. On a

$$\begin{aligned}
\widehat{Y}_I - \widehat{Y}_{HT} &= \sum_{i\in s} r_i \frac{y_i}{\pi_i} + \sum_{i\in s}(1-r_i)\frac{y_i^*}{\pi_i} - \sum_{i\in s}\frac{y_i}{\pi_i} \\
&= \sum_{i\in s} r_i \frac{y_i}{\pi_i} + \sum_{i\in s}\frac{(1-r_i)}{\pi_i}\boldsymbol{x}_i'\widehat{\boldsymbol{T}}_r^{-1}\sum_{i\in s}\frac{r_i}{\pi_i}\boldsymbol{x}_i c_i^{-1} y_i - \sum_{i\in s}\frac{y_i}{\pi_i} \\
&= \sum_{i\in s} r_i \frac{y_i}{\pi_i} + \left(\widehat{\boldsymbol{X}}_{HT} - \widehat{\boldsymbol{X}}_r\right)'\widehat{\boldsymbol{T}}_r^{-1}\sum_{i\in s}\frac{r_i}{\pi_i}\boldsymbol{x}_i c_i^{-1} y_i - \sum_{i\in s}\frac{y_i}{\pi_i} \\
&= \sum_{i\in s}\frac{r_i}{\pi_i}\left\{\left(\widehat{\boldsymbol{X}}_{HT} - \widehat{\boldsymbol{X}}_r\right)'\widehat{\boldsymbol{T}}_r^{-1} c_i^{-1}\boldsymbol{x_i} + 1\right\} y_i - \sum_{i\in s}\frac{y_i}{\pi_i} \\
&= \sum_{i\in s}\frac{r_i}{\pi_i} g_i y_i - \sum_{i\in s}\frac{y_i}{\pi_i} \\
&= \sum_{i\in s}\frac{1}{\pi_i}(r_i g_i - 1) y_i. \qquad \text{(C.0.12)}
\end{aligned}$$

En insérant (C.0.12) dans la variance (3.2.3), on obtient

$$\begin{aligned}
\mathrm{V}_{NR} &= \mathrm{E}_p\mathrm{E}_r\mathrm{V}_m\left(\widehat{Y}_I - \widehat{Y}_{HT}|s, s_r\right) \\
&= \sigma^2\mathrm{E}_p\mathrm{E}_r\left\{\sum_{i\in s}\frac{1}{\pi_i^2}(r_i g_i - 1)^2 c_i\right\}. \qquad \text{(C.0.13)}
\end{aligned}$$

L'estimateur de V_{NR} est obtenu en estimant σ^2 par $\widehat{\sigma}^2$ en (3.2.13) et on obtient

$$\widehat{\mathrm{V}}_{NR} = \widehat{\sigma}^2\sum_{i\in s}\frac{1}{\pi_i^2}(r_i g_i - 1)^2 c_i. \qquad \text{(C.0.14)}$$

Annexe D

ESTIMATEUR DE V_{MIX}

En insérant la décomposition de l'erreur qui est due à la non-réponse (C.0.12) dans la variance mixte, on obtient

$$\begin{aligned}
\mathrm{V}_{mix} &= \mathrm{E}_p\mathrm{E}_r\mathrm{Cov}_m\left(\widehat{Y}_I - \widehat{Y}_{HT}, \widehat{Y}_{HT} - Y | s, s_r\right) \\
&= \mathrm{E}_p\mathrm{E}_r\mathrm{Cov}_m\left(\sum_{i\in s}\frac{1}{\pi_i}(r_i g_i - 1)y_i,\ \sum_{i\in s}\frac{y_i}{\pi_i} - \sum_{i\in U} y_i \middle| s, s_r\right) \\
&= \mathrm{E}_p\mathrm{E}_r\left\{\sum_{j\in s}\frac{1}{\pi_j}\sum_{i\in s}\frac{1}{\pi_i}(r_i g_i - 1)\mathrm{Cov}_m\left(y_j, y_i | s, s_r\right)\right. \\
&\left. - \sum_{j\in U}\sum_{i\in s}\frac{1}{\pi_i}(r_i g_i - 1)\mathrm{Cov}_m\left(y_j, y_i | s, s_r\right)\right\} \\
&= \sigma^2\mathrm{E}_p\mathrm{E}_r\left\{\sum_{i\in s}\left(\frac{1-\pi_i}{\pi_i^2}\right)(r_i g_i - 1)c_i\right\}. \qquad \text{(D.0.15)}
\end{aligned}$$

L'estimateur de V_{mix} est obtenu en estimant σ^2 par $\widehat{\sigma}^2$ en (3.2.13) et on obtient

$$\widehat{\mathrm{V}}_{mix} = \widehat{\sigma}^2\sum_{i\in s}\left(\frac{1-\pi_i}{\pi_i^2}\right)(r_i g_i - 1)c_i. \qquad \text{(D.0.16)}$$

Annexe E

ESTIMATEUR DE V_1

Remarquons d'abord que l'estimateur $\widehat{Y}_I$ dans le cas de l'imputation par la régression peut s'écrire comme (3.1.2), qui est bien une fonction de totaux estimés. Un développement de Taylor du premier ordre conduit à

$$\widehat{Y}_I - \widetilde{Y}_I = \sum_{i \in s} \frac{E_i}{\pi_i} + O_p\left(\frac{1}{n}\right),$$

où

$$E_i = r_i y_i + (1 - r_i)\boldsymbol{x}_i' \mathbf{B}_r + \left(\sum_{j \in U} (1 - r_j)\boldsymbol{x}_j\right)' \boldsymbol{T}_r^{-1} r_i \boldsymbol{x}_i c_i^{-1} \left(y_i - \boldsymbol{x}_i' \mathbf{B}_r\right)$$

et

$$\begin{aligned} \mathbf{B}_r &= \left(\sum_{i \in U} r_i \boldsymbol{x}_i c_i^{-1} \boldsymbol{x}_i'\right)^{-1} \sum_{i \in U} r_i \boldsymbol{x}_i c_i^{-1} y_i \\ &= \boldsymbol{T}_r^{-1} \boldsymbol{t}_r. \end{aligned}$$

En ignorant les termes d'ordre supérieur, on peut approximer $\mathrm{V}_p\left(\widehat{Y}_I \middle| \boldsymbol{y}_U, \boldsymbol{r}\right)$ par

$$\mathrm{V}_p\left(\widehat{Y}_I \middle| \boldsymbol{y}_U, \boldsymbol{r}\right) \approx \sum_{i \in U} \sum_{j \in U} \Omega_{ij} E_i E_j,$$

que l'on estimera par

$$\widehat{\mathrm{V}}_1 = \sum_{i \in s} \sum_{j \in s} \frac{\Omega_{ij}}{\pi_{ij}} e_i e_j, \tag{E.0.17}$$

où e_i est donné par (3.3.3).

Annexe F

ESTIMATEUR DE V_2

On commence par écrire

$$\begin{aligned}
\mathrm{V}_2 &= \mathrm{E}_r \mathrm{V}_m \mathrm{E}_p (\widehat{Y}_I - Y | \boldsymbol{y}_U, \boldsymbol{r}) \\
&= \mathrm{E}_r \mathrm{V}_m \mathrm{E}_p \left(\sum_{i \in s} r_i \frac{y_i}{\pi_i} + \sum_{i \in s} \frac{(1-r_i)}{\pi_i} \boldsymbol{x}_{\boldsymbol{i}}' \widehat{\mathbf{B}}_r - \sum_{i \in U} y_i \middle| \boldsymbol{y}_U, \boldsymbol{r} \right) \\
&\approx \mathrm{E}_r \mathrm{V}_m \left(\sum_{i \in U} r_i y_i + \sum_{i \in U} (1-r_i) \boldsymbol{x}_i' \mathbf{B}_r - \sum_{i \in U} y_i \middle| \boldsymbol{r} \right). \qquad \text{(F.0.18)}
\end{aligned}$$

L'approximation (F.0.18) provient d'un développement de Taylor du premier ordre en ignorant les termes d'ordre supérieur. La variance (F.0.18) peut donc s'écrire comme

$$\mathrm{V}_2 \approx \mathrm{E}_r \mathrm{V}_m \left(\sum_{i \in U} \xi_i y_i \right), \qquad \text{(F.0.19)}$$

où

$$\xi_i = \left(\sum_{j \in U} (1-r_j) \boldsymbol{x}_j \right)' \boldsymbol{T}_r^{-1} r_i \boldsymbol{x}_i c_i^{-1} - (1-r_i).$$

On trouve finalement

$$\begin{aligned}
\mathrm{V}_2 &= \mathrm{E}_r \left\{ \sum_{i \in U} \xi_i^2 \mathrm{V}_m(y_i) \right\} \\
&= \sigma^2 \mathrm{E}_r \left\{ \sum_{i \in U} \xi_i^2 c_i \right\}. \qquad \text{(F.0.20)}
\end{aligned}$$

Un estimateur de la variance (F.0.20) est donné par

$$\widehat{\mathrm{V}}_2 = \widehat{\sigma}^2 \sum_{i \in s} \frac{\widehat{\xi}_i^2}{\pi_i} c_i,$$

où $\widehat{\xi}_i$ est donné par (3.3.4) et $\widehat{\sigma}^2$ est donné par (3.2.13).