Université de Montréal

Estimation simplifiée de la variance pour des plans complexes

par

Isabelle Lefebvre

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

22 décembre 2016

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Estimation simplifiée de la variance pour des plans complexes

présenté par

Isabelle Lefebvre

a été évalué par un jury composé des personnes suivantes :

Pierre Duchesne

(président-rapporteur)

David Haziza

(directeur de recherche)

Jean-François Angers

(membre du jury)

Mémoire accepté le

30 avril 2017

SOMMAIRE

En présence de plans de sondage complexes, les méthodes classiques d'estimation de la variance présentent certains défis. En effet, les estimateurs de variance usuels requièrent les probabilités d'inclusion d'ordre deux qui peuvent être complexes à obtenir pour certains plans de sondage. De plus, pour des raisons de confidentialité, les fichiers externes de microdonnées n'incluent généralement pas les probabilités d'inclusion d'ordre deux (souvent sous la forme de poids bootstrap). En s'inspirant d'une approche développée par Ohlsson (1998) dans le contexte de l'échantillonnage de Poisson séquentiel, nous proposons un estimateur ne requérant que les probabilités d'inclusion d'ordre un. L'idée est d'approximer la stratégie utilisée par l'enquête (consistant du choix d'un plan de sondage et d'un estimateur) par une stratégie équivalente dont le plan de sondage est le plan de Poisson. Nous discuterons des plans proportionnels à la taille avec ou sans grappes. Les résultats d'une étude par simulation seront présentés.

Mots-clés : Estimation de la variance, échantillonnage de Poisson séquentiel, échantillonnage de Poisson, plan proportionnel à la taille, probabilités d'inclusion d'ordre un.

SUMMARY

In a complex design framework, standard variance estimation methods entail substantial challenges. As we know, conventional variance estimators involve second order inclusion probabilities, which can be difficult to compute for some sampling designs. Also, confidentiality standards generally prevent second order inclusion probabilities to be included in external microdata files (often in the form of bootstrap weights). Based on Ohlsson's sequential Poisson sampling method (1998), we suggest a simplified estimator for which we only need first order inclusion probabilities. The idea is to approximate a survey strategy (which consists of a sampling design and an estimator) by an equivalent strategy for which a Poisson sampling design is used. We will discuss proportional to size sampling and proportional to size cluster sampling. Results of a simulation study will be presented.

Keywords : Variance estimation, sequential Poisson sampling, Poisson sampling, proportional to size sampling, first order inclusion probabilities.

TABLE DES MATIÈRES

Sommaire v

Summary vii

Remerciements 1

Introduction 3

Chapitre 1. Introduction à la théorie de l'échantillonnage 5

1.1. Notation et plan de sondage 5
1.1.1. Échantillonnage aléatoire simple sans remise 7
1.1.2. Échantillonnage de Poisson 7
1.1.3. Échantillonnage de Poisson conditionnel 8
1.1.4. Échantillonnage de Poisson séquentiel 8
1.1.5. Échantillonnage de Rao-Sampford 8
1.1.6. Échantillonnage par grappes à un degré 9
1.1.7. Échantillonnage à deux degrés 9

1.2. Estimateur du total et estimation de la variance 10
1.2.1. Cas de l'échantillonnage à un degré 11
1.2.2. Cas de l'échantillonnage par grappes 14
1.2.3. Cas de l'échantillonnage à deux degrés 14
1.2.4. Estimateurs de calage 18

1.3. Entropie d'un plan de sondage 20

1.4. Alternatives à l'estimation classique de la variance 22
1.4.1. Approximation des probabilités d'inclusion d'ordre deux 22
1.4.1.1. Approximation de Hájek 22
1.4.1.2. Approximations de Brewer-Donadio 23
1.4.2. Approche par simulation Monte-Carlo 24
1.4.3. Approche d'Ohlsson 25

1.5. Convergence de l'estimateur de variance de Hájek 26

Chapitre 2. Estimation simplifiée de la variance 29

2.1. Introduction 29

2.2. Étude de l'approche d'Ohlsson 31
2.2.1. Approche d'Ohlsson : une approximation des π_{kl} 31
2.2.2. Forme générale de l'estimateur lié à l'approche d'Ohlsson 32

2.3. Estimation de la variance : cas de l'échantillonnage par grappes à un degré 32
2.3.1. Application de l'approche d'Ohlsson au niveau des grappes 33
2.3.2. Application de l'approche d'Ohlsson au niveau des éléments 33
2.3.2.1. Cas des M_i égaux 35
2.3.2.2. Cas des M_i inégaux 36
2.3.2.3. Variance ajustée pour le biais 37

Chapitre 3. Études par simulation 39

3.1. Étude 1 : Cas d'une population simple 39
3.1.1. Populations et échantillons simulés 39
3.1.2. Critères de comparaison des estimateurs de variance 40
3.1.3. Résultats 41
3.1.4. Discussion 41

3.2. Étude 2 : Cas d'une population de grappes 42
3.2.1. Populations et échantillons simulés 42
3.2.2. Critères de comparaison des estimateurs de variance 43
3.2.3. Résultats 44
3.2.4. Discussion 49

Chapitre 4. Conclusion 51

Bibliographie 53

Annexe A. Approche d'Ohlsson : une approximation des probabilités d'inclusion d'ordre deux ..... A-i

Annexe B. Biais relatif dû à l'utilisation de l'approche d'Ohlsson au niveau des éléments dans le cas des grappes ..... B-i

Annexe C. Ordre de grandeur du terme $\tilde{A}$ ..... C-i

Annexe D. Ordre de grandeur du terme A ..... D-i

Annexe E. Estimation de la variance de l'effet de grappe ..... E-i

REMERCIEMENTS

J'aimerais, en tout premier lieu, remercier mon directeur de recherche, David Haziza, pour son encadrement tout au long de la réalisation de ce mémoire. Je lui suis reconnaissante pour sa patience, sa disponibilité et sa rigueur ainsi que ses grandes qualités de pédagogue. Je tiens aussi à le remercier pour toutes les opportunités de développement professionnel qu'il a mises sur mon chemin.

J'aimerais remercier l'Institut canadien des sciences statistiques (INCASS) de m'avoir offert l'opportunité d'évoluer dans un milieu de recherche gouvernemental, à Statistique Canada, dans le cadre de ce mémoire. Je tiens également à remercier Jean-François Beaumont pour son encadrement lors de cette expérience des plus enrichissantes.

Enfin, je remercie ma famille pour leurs encouragements et leur soutien tout au long de mes études. Un merci particulier à mes parents pour leur support moral et financier, que j'apprécie grandement. Merci également à mon compagnon, Sheldon, et à mes amies Camille, Marie-Ève et Vanessa, pour leur appui constant et leurs conseils.

INTRODUCTION

Dans le secteur publique, les enquêtes constituent un élément essentiel à la prise de décision. Celles-ci sont largement employées par les organismes statistiques (tel que Statistique Canada) dans le but de quantifier plusieurs aspects de la population permettant ainsi d'en dresser un portrait objectif. Une enquête statistique débute toujours avec l'étape importante de la formulation des objectifs, qui permet de définir un cadre de travail bien précis. Une base de sondage est ensuite choisie. Cette dernière contient des informations d'identification et de communication de la population cible. Elle peut également inclure des données de classification qui peuvent être utilisées lors des étapes d'échantillonnage et d'estimation. L'étape suivante consiste à déterminer un plan de sondage. Ainsi, chaque unité de la population cible se voit attribuer une probabilité d'être incluse dans l'échantillon. Cette probabilité est entièrement déterminée par le plan de sondage choisi. On procède ensuite à la rédaction du questionnaire, qui se doit d'être clair pour assurer la qualité de l'information recueillie. Les données sont ensuite collectées avant d'être saisies et codées. Une vérification des données a ensuite lieu dans le but de repérer (et corriger si possible) des erreurs ou des incohérences. En présence de valeurs manquantes, on utilise l'imputation pour corriger la base de données. On procède ensuite à l'estimation des valeurs d'intérêt qui permettront de répondre aux objectifs de l'enquête.

À l'étape d'estimation, il existe plusieurs méthodes permettant d'estimer une quantité d'intérêt. Une approche largement utilisée consiste à faire intervenir les probabilités d'inclusion, qui elles dépendent du plan de sondage employé. Il faut fournir une mesure de volatilité pour accompagner chaque estimation. C'est ce qu'on appelle l'estimation de la variance. Dans ce mémoire, on focalise sur l'estimation de la variance en présence de certains plans de sondage complexes.

Puisqu'ils sont généralement moins coûteux et plus simple à mettre en œuvre, l'emploi de certains plans de sondage complexes peut être favorisé au sein de

grands organismes statistiques. Ces plans de sondage peuvent toutefois entraîner des défis substantiels à l'étape d'estimation, plus particulièrement au niveau de l'estimation de la variance. En effet, les probabilités d'inclusion d'ordre deux (information nécessaire au calcul des estimateurs classiques de variance) sont généralement complexe à calculer. Dans le cas où l'utilisateur des données n'est pas directement employé par l'organisme statistique (collaborateur, chercheur, etc..), les probabilités d'inclusion d'ordre deux peuvent tout simplement avoir été exclues du jeu de données, pour des raisons de confidentialité. En effet, dans certains cas, ces probabilités peuvent permettre de reconstruire de l'information jugée confidentielle, par exemple en permettant d'identifier les éléments qui appartiennent à une même grappe. Dans ce mémoire, nous proposons une alternative à l'estimation classique de la variance qui ne requiert pas les probabilités d'inclusion d'ordre deux. Nous faisons l'hypothèse que les probabilités d'inclusion d'ordre un sont disponibles. Cette approche est étudiée dans le cadre d'un plan de sondage complexe à grande entropie. Une étude par simulation est effectuée pour examiner les propriétés de la méthode suggérée. Cette étude montre que l'estimateur proposé se comporte bien en termes de biais.

Chapitre 1

INTRODUCTION À LA THÉORIE DE L'ÉCHANTILLONNAGE

1.1. Notation et plan de sondage

On considère $U^\star$ une population finie d'éléments de taille $N^\star$. On s'intéresse au total de la variable d'intérêt y au niveau de la population. Il sera noté $Y = \sum_{k \in U^\star} y_k$, où y_k désigne la $k$ième valeur de la variable d'intérêt y. Dans le but d'estimer ce total, un échantillon d'éléments $s^\star$ de taille $n^\star$ est sélectionné de $U^\star$ selon un plan de sondage sans remise. On définit un plan de sondage sans remise $p(\cdot)$ comme une fonction qui assigne à chaque échantillon possible sa probabilité d'être sélectionné et qui satisfait les conditions

(i) $p(s^\star) \geq 0$, $\forall s^\star \subset U^\star$;

(ii) $\sum_{s^\star \subset U^\star} p(s^\star) = 1$.

Un échantillon peut également être caractérisé par le vecteur $\boldsymbol{I} = (I_1, ..., I_{N^\star})^\top$, où I_k prend la valeur 1 si l'élément k est présent dans l'échantillon et la valeur 0 sinon. Un plan de sondage $p(\cdot)$ assigne donc à chaque vecteur $\boldsymbol{I}$ possible sa probabilité de sélection. Pour un plan de sondage donné, l'ensemble de tous les échantillons possibles dont la probabilité de sélection est strictement supérieure à zéro constitue le support $\mathcal{Q}$ de ce plan de sondage. On a $\mathcal{Q} = \{s^\star \subset U^\star : p(s^\star) > 0\}$. La notation $f^\star = n^\star / N^\star$ est utilisée pour désigner la fraction de sondage, soit la proportion d'unités de la population étant incluses dans l'échantillon.

On définira la probabilité que l'unité k soit contenue dans l'échantillon, appelée probabilité d'inclusion d'ordre un, par

$$\pi_k = P(k \in s^\star) = \sum_{\substack{s^\star \in \mathcal{Q} \\ s^\star \ni k}} p(s^\star).$$

On peut interpréter π_k comme la proportion pondérée des échantillons contenant l'unité k. Lorsqu'une variable auxiliaire x, appelée variable de taille, est disponible pour toutes les unités de la population, il est possible d'en tenir compte au niveau du plan de sondage. En effet, on peut définir les probabilités d'inclusion d'ordre un selon

$$\pi_k = n^\star \frac{x_k}{X} \equiv n^\star p_k, \ \forall k \in U^\star, \tag{1.1.1}$$

où $X = \sum_{k \in U^\star} x_k$ est le total de la variable x. Notons qu'à partir de la définition (1.1.1), il est possible d'obtenir des π_k supérieurs à un. Dans ce cas, on sélectionne les unités pour lesquelles $\pi_k > 1$ avec probabilité un, puis on recalcule les π_k pour les unités non-sélectionnées en prenant soin d'ajuster $n^\star$ et X pour tenir compte des unités qui ont déjà été sélectionnées. Si on obtient de nouveau des π_k supérieurs à un, on répète la même démarche de façon itérative, jusqu'à ce que tous les π_k soient inférieurs ou égaux à un. Lorsque les probabilités π_k sont définies selon (1.1.1), on parle de plan de sondage avec probabilités proportionnelles à la taille (EPPT). Pour un plan de taille aléatoire, $n^\star$ désigne la taille espérée de l'échantillon et $n_s^\star$ désigne la taille.

L'inverse de π_k, soit $d_k = 1/\pi_k$, désigne le poids de sondage de l'unité k. On l'interprète comme le nombre d'unités que l'unité k représente dans la population.

De manière similaire, on définit la probabilité d'inclusion d'ordre deux des unités k et l comme

$$\pi_{kl} = P(k \in s^\star, l \in s^\star) = \sum_{\substack{s^\star \in \mathcal{Q} \\ s^\star \ni k,l}} p(s^\star).$$

Il s'agit de la probabilité que l'échantillon contienne à la fois l'unité k et l'unité l. Pour $k = l$, cette probabilité devient $\pi_{kl} = \pi_{kk} = \pi_k$.

On présente quelques propriétés des probabilités d'inclusion. Les notations $E_p(\cdot)$, $V_p(\cdot)$ et $Cov_p(\cdot)$ indiquent que les espérances, variances et covariances sont évaluées par rapport au plan de sondage.

Proposition 1.1.1. *Pour un plan de sondage $p(\cdot)$ donné :*

(i) $E_p(I_k) = \pi_k$;

(ii) $V_p(I_k) = \pi_k(1 - \pi_k)$;

(iii) $E_p(I_k I_l) = \pi_{kl}$;

(iv) $Cov_p(I_k, I_l) = \pi_{kl} - \pi_k \pi_l \equiv \Delta_{kl}$.

Proposition 1.1.2. *Pour un plan de sondage $p(\cdot)$ de taille fixe $n^\star$:*

(i) $\sum_{k\in U^\star} \pi_k = n^\star$;

(ii) $\sum_{\substack{k\in U^\star \\ k\neq l}} \pi_{kl} = (n^\star - 1)\pi_k$;

(iii) $\sum_{\substack{k\in U^\star \\ k\neq l}} \sum_{l\in U^\star} \pi_{kl} = n^\star(n^\star - 1)$;

(iv) $\sum_{k\in U^\star} \Delta_{kl} = 0$, $\forall l \in U^\star$ *et* $\sum_{l\in U^\star} \Delta_{kl} = 0$, $\forall k \in U^\star$.

Les détails des Propositions 1.1.1 et 1.1.2 sont présentés dans Särndal et al. (1992). Les sections 1.1.1 à 1.1.7 présentent quelques plans de sondage fréquemment étudiés dans la littérature.

1.1.1. Échantillonnage aléatoire simple sans remise

L'échantillonnage aléatoire simple sans remise (EASSR) est un plan de sondage de taille fixe qui se caractérise par le fait qu'il accorde à chaque échantillon de $n^\star$ unités distinctes une probabilité égale d'être sélectionné. Ainsi, chacun des $\binom{N^\star}{n^\star}$ échantillons se voit attribué la probabilité de sélection

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

Pour ce plan de sondage, les probabilités d'inclusion d'ordre un et deux sont respectivement égales à

$$\pi_k = \frac{n^\star}{N^\star},\ \forall k \in U^\star;$$
$$\pi_{kl} = \frac{n^\star(n^\star - 1)}{N^\star(N^\star - 1)},\ \forall k, l \in U^\star \text{ tel que } k \neq l.$$

1.1.2. Échantillonnage de Poisson

L'échantillonnage de Poisson est un plan de sondage de taille aléatoire $n_s^\star$ qui consiste à effectuer, pour chaque unité k de la population, une expérience de Bernoulli avec probabilité de succès π_k. Lorsque l'expérience est un succès, l'unité correspondante est incluse dans l'échantillon. La probabilité de tirer un échantillon $s^\star$ donné est

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

Puisque les expériences de Bernoulli sont indépendantes d'une unité à l'autre, on a $\pi_{kl} = \pi_k\pi_l$, $k \neq l$. Le cas particulier pour lequel les probabilités d'inclusion sont constantes ($\pi_k = \pi$) se nomme échantillonnage de Bernoulli. On est dans le cas d'un plan de Poisson avec probabilités proportionnelles à la taille, lorsque π_k est défini par (1.1.1).

1.1.3. Échantillonnage de Poisson conditionnel

L'échantillonnage de Poisson conditionnel, proposé par Hájek (1964), est une variante de taille fixe du plan de Poisson. On assigne d'abord les probabilités d'inclusion $\tilde{\pi}_1, \tilde{\pi}_2, ..., \tilde{\pi}_{N^\star}$ aux unités de la population. On tire ensuite un échantillon au moyen d'un plan de Poisson à l'aide des probabilités $\tilde{\pi}_k$. S'il est de taille $n_s^\star \neq n^\star$, l'échantillon est rejeté. Le processus se poursuit jusqu'à ce qu'un échantillon de taille $n^\star$ soit obtenu. L'échantillon final $s^\star$ a une probabilité de sélection de

$$p(s^\star | n_s^\star = n^\star) = \frac{\prod\limits_{k \in s^\star} \tilde{\pi}_k \prod\limits_{k \in \{U^\star \setminus s^\star\}} (1 - \tilde{\pi}_k)}{\sum\limits_{s \in \mathcal{Q}_{n^\star}} \prod\limits_{k \in s} \tilde{\pi}_k \prod\limits_{k \in \{U^\star \setminus s\}} (1 - \tilde{\pi}_k)},$$

où $\mathcal{Q}_{n^\star}$ est l'ensemble de tous les échantillons possibles de taille $n^\star$. Les probabilités d'inclusion d'ordre un finales, qu'on note π_k, ne sont pas égales aux $\tilde{\pi}_k$. Les π_k ainsi que les π_{kl} se calculent de façon itérative à partir des $\tilde{\pi}_k$. Notons qu'à l'aide d'un algorithme développé par Deville (2000), il est aussi possible de fixer les π_k et de déterminer les $\tilde{\pi}_k$ à utiliser à l'étape d'échantillonnage de Poisson. Lorsque $\tilde{\pi}_k = \tilde{\pi}$, le plan de Poisson conditionnel coïncide avec l'échantillonnage aléatoire simple sans remise. En définissant les π_k par (1.1.1), le plan de Poisson conditionnel permet un échantillonnage PPT sans remise de taille fixe.

1.1.4. Échantillonnage de Poisson séquentiel

Une autre variation à taille fixe du plan de Poisson est le plan de Poisson séquentiel (Ohlsson 1990 ; 1998). On commence par assigner les probabilités $\tilde{\pi}_1, ..., \tilde{\pi}_{N^\star}$ aux $N^\star$ unités de la population. Pour chacune de ces unités, une valeur t_k est générée selon une distribution $\mathcal{U}[0, 1]$. On calcule ensuite

$$\xi_k = t_k / \tilde{\pi}_k.$$

Les $n^\star$ unités associées aux plus petites valeurs de ξ_k sont sélectionnées dans l'échantillon. En général, les probabilités d'inclusion d'ordre un ne correspondent pas aux $\tilde{\pi}_k$ et le calcul des π_{kl} est laborieux. Lorsque $\tilde{\pi}_k = \tilde{\pi}$, on retrouve le plan aléatoire simple sans remise.

1.1.5. Échantillonnage de Rao-Sampford

La procédure de Rao-Sampford permet de tirer un échantillon de taille fixe avec probabilité proportionnelle à la taille. La première unité doit être sélectionnée avec probabilité p_k, alors que les $n^\star - 1$ unités suivantes sont sélectionnées avec

probabilité proportionnelle à

$$\frac{p_k}{(1 - n^\star p_k)}$$

selon un échantillonnage avec remise. Si l'échantillon se compose d'unités distinctes, il est conservé. Sinon, un nouvel échantillon est tiré selon les mêmes indications. Les probabilités d'inclusion d'ordre un sont données par (1.1.1), alors que les π_{kl} peuvent être calculées à l'aide d'un algorithme récursif décrit par Sampford (1967).

1.1.6. Échantillonnage par grappes à un degré

On considère maintenant le cas où la population se compose de N sous-populations distinctes $U_1^\star$, $U_2^\star$, ..., $U_N^\star$, de tailles M_1, M_2, ..., M_N. Ces sous-populations sont nommées grappes ou unités primaires d'échantillonnage (UPE). On a

$$U^\star = \bigcup_{i=1}^{N} U_i^\star \text{ et } N^\star = \sum_{i=1}^{N} M_i.$$

L'échantillonnage par grappes à un degré consiste à tirer un échantillon de n grappes, qui sera noté s, à partir de la population de grappes $U = \{1, 2, ..., N\}$ selon un plan de sondage $p(\cdot)$. L'échantillon final se compose de tous les éléments appartenant aux grappes sélectionnées. Il est défini par

$$s^\star = \bigcup_{i \in s} U_i^\star.$$

Les probabilités d'inclusion d'ordre un et deux des grappes varient selon le plan de sondage utilisé à l'étape de sélection. À l'intérieur des grappes, la probabilité d'inclusion d'un élément est la même que la probabilité d'inclusion de la grappe à laquelle il appartient :

$$\pi_k = \pi_i \text{ pour } k \in U_i^\star.$$

De façon analogue, les probabilités d'inclusion d'ordre deux au niveau des éléments sont

$$\pi_{kl} = \begin{cases} \pi_i & \text{si } k, l \in U_i^\star; \\ \pi_{ij} & \text{si } k \in U_i^\star,\ l \in U_j^\star, \text{ tel que } i \neq j. \end{cases}$$

1.1.7. Échantillonnage à deux degrés

L'échantillonnage à deux degrés comprend deux étapes aléatoires, c'est-à-dire la sélection d'unités primaires d'échantillonnage (UPE) suivie de la sélection d'unités secondaires d'échantillonnage (USE). Considérons une population

du même type que celle présentée à la section 1.1.6. Les UPE correspondent aux grappes, tandis que les éléments qui les composent correspondent aux USE. Au premier degré, un échantillon s de n grappes est sélectionné selon un plan de sondage $p(\cdot)$. À l'intérieur de chaque grappe i sélectionnée au premier degré, on tire un échantillon d'éléments $s_i^\star$ de taille m_i, selon un plan de sondage $p_i(s_i^\star|s)$. L'échantillon final de taille $n^\star = \sum_{i=1}^{n} m_i$ est donné par

$$s^\star = \bigcup_{i \in s} s_i^\star.$$

Dans ce qui suit, on focalise sur l'échantillonnage à deux degrés qui respecte les propriétés d'invariance et d'indépendance. Ainsi, le plan doit respecter la condition $p_i(\cdot|s) = p_i(\cdot)$, $\forall i \in U$, et la sélection d'éléments est effectuée indépendamment d'une grappe à l'autre.

On note π_i et π_{ij} les probabilités d'inclusion d'ordre un et deux correspondant aux grappes. La probabilité d'inclusion de l'élément k dans l'échantillon $s_i^\star$, sachant que s contient i est notée

$$\pi_{k|i} = P(k \in s_i^\star | i \in s).$$

De façon analogue, la probabilité que les éléments k et l soient inclus dans $s_i^\star$, sachant que s contient i est donnée par

$$\pi_{kl|i} = P(k \in s_i^\star, l \in s_i^\star | i \in s).$$

Ainsi, les probabilités d'inclusion d'ordre un et deux correspondant aux USE sont respectivement données par

$$\pi_k = \pi_i \pi_{k|i};$$

$$\pi_{kl} = \begin{cases} \pi_i \pi_{kl|i} & \text{si } k, l \in U_i^\star,\ k \neq l; \\ \pi_{ij} \pi_{k|i} \pi_{l|j} & \text{si } k \in U_i^\star, l \in U_j^\star,\ i \neq j. \end{cases}$$

1.2. Estimateur du total et estimation de la variance

À la section 1.1, quelques techniques permettant de tirer un échantillon ont été introduites. À partir de cet échantillon, on souhaite maintenant estimer le total Y de la variable d'intérêt ainsi que la variance de cette estimation. Les sections suivantes présentent des méthodes pour y arriver. On distingue le cas d'échantillonnage à un degré de celui à deux degrés.

1.2.1. Cas de l'échantillonnage à un degré

Pour estimer le total Y dans un contexte d'échantillonnage à un degré, un estimateur largement utilisé qui présente des propriétés avantageuses est l'estimateur d'Horvitz-Thompson,

$$\widehat{Y}_\pi = \sum_{k\in s^\star} \frac{y_k}{\pi_k} = \sum_{k\in s^\star} d_k y_k = \sum_{k\in U^\star} d_k y_k I_k, \tag{1.2.1}$$

aussi appelé estimateur par dilatation. Notons que, pour un plan à taille fixe, lorsque $y \propto \pi$, on a $\widehat{Y}_\pi = Y$, $\forall s^\star \in \mathcal{Q}$. En effet, lorsque $y_k = a\pi_k$, $\forall k \in U^\star$, où a est une certaine constante, on a

$$\widehat{Y}_\pi = \sum_{k\in s^\star} \frac{y_k}{\pi_k} = \sum_{k\in s^\star} \frac{a\pi_k}{\pi_k} = an^\star = a\sum_{k\in U^\star} \pi_k = \sum_{k\in U^\star} y_k = Y.$$

Dans ce cas, on a $V_p(\widehat{Y}_\pi) = 0$. Les résultats suivants énoncent les propriétés de l'estimateur (1.2.1) en matière de biais et de variance.

Proposition 1.2.1. *Si $\pi_k > 0$, $\forall k \in U^\star$, l'estimateur $\widehat{Y}_\pi$ est sans biais par rapport au plan pour Y.*

Démonstration. L'espérance par rapport au plan de l'estimateur d'Horvitz-Thompson est donnée par

$$E_p(\widehat{Y}_\pi) = E_p\left(\sum_{k\in U^\star} \frac{y_k}{\pi_k} I_k\right) = \sum_{k\in U^\star} \frac{y_k}{\pi_k} E_p(I_k) = \sum_{k\in U^\star} y_k = Y.$$

□

Proposition 1.2.2. *La variance par rapport au plan de l'estimateur d'Horvitz-Thompson est donnée par*

$$V_p(\widehat{Y}_\pi) = \sum_{k\in U^\star}\sum_{l\in U^\star} \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}. \tag{1.2.2}$$

Démonstration.

$$\begin{aligned} V_p(\widehat{Y}_\pi) &= Cov_p\left(\sum_{k\in U^\star} \frac{y_k}{\pi_k} I_k, \sum_{l\in U^\star} \frac{y_l}{\pi_l} I_l\right) \\ &= \sum_{k\in U^\star}\sum_{l\in U^\star} Cov_p(I_k, I_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \\ &= \sum_{k\in U^\star}\sum_{l\in U^\star} \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}. \end{aligned}$$

La dernière égalité découle de la Proposition 1.1.1. □

Proposition 1.2.3. *Pour un plan à taille fixe, la variance par rapport au plan de l'estimateur d'Horvitz-Thompson peut être réécrite comme*

$$V_p(\widehat{Y}_\pi) = -\frac{1}{2}\sum_{k\in U^\star}\sum_{l\in U^\star}\Delta_{kl}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2. \tag{1.2.3}$$

Démonstration. En développant (1.2.3), on obtient

$$\begin{aligned}
V_p(\widehat{Y}_\pi) &= -\frac{1}{2}\sum_{k\in U^\star}\sum_{l\in U^\star}\Delta_{kl}\left(\frac{y_k^2}{\pi_k^2}-2\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}+\frac{y_l^2}{\pi_l^2}\right)\\
&= -\frac{1}{2}\sum_{k\in U^\star}\frac{y_k^2}{\pi_k^2}\sum_{l\in U^\star}\Delta_{kl}+\sum_{k\in U^\star}\sum_{l\in U^\star}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}-\frac{1}{2}\sum_{l\in U^\star}\frac{y_l^2}{\pi_l^2}\sum_{k\in U^\star}\Delta_{kl}\\
&= \sum_{k\in U^\star}\sum_{l\in U^\star}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}
\end{aligned}$$

en notant (voir Proposition 1.1.2) que $\sum_{l\in U^\star}\Delta_{kl} = \sum_{k\in U^\star}\Delta_{kl} = 0$. □

Toutefois, la variance de l'estimateur d'Horvitz-Thompson fait intervenir les y_k de toutes les unités de la population. En pratique, cette information n'est collectée que pour les unités de l'échantillon. La variance de l'estimateur par dilatation demeure donc une quantité inconnue qu'on cherche à estimer.

Dans le but d'estimer $V_p(\widehat{Y}_\pi)$, on considère l'estimateur Horvitz-Thompson de la variance donné par

$$\widehat{V}_{HT}(\widehat{Y}_\pi) = \sum_{k\in s^\star}\sum_{l\in s^\star}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}. \tag{1.2.4}$$

Dans le cas d'un plan de taille fixe, on peut aussi estimer la variance à l'aide de l'estimateur Sen-Yates-Grundy, soit

$$\widehat{V}_{SYG}(\widehat{Y}_\pi) = -\frac{1}{2}\sum_{k\in s^\star}\sum_{l\in s^\star}\frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2. \tag{1.2.5}$$

Proposition 1.2.4. *Si $\pi_{kl} > 0$, $\forall k, l \in U^\star$, l'estimateur Horvitz-Thompson de la variance est sans biais par rapport au plan pour $V_p(\widehat{Y}_\pi)$.*

Démonstration.

$$\begin{aligned}
E_p\{\widehat{V}_{HT}(\widehat{Y}_\pi)\} &= E_p\left(\sum_{k\in s^\star}\sum_{l\in s^\star}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\right)\\
&= \sum_{k\in U^\star}\sum_{l\in U^\star}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}E_p(I_kI_l)\\
&= \sum_{k\in U^\star}\sum_{l\in U^\star}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.
\end{aligned}$$

□

Proposition 1.2.5. *Si $\pi_{kl} > 0$, $\forall k, l \in U^\star$, l'estimateur Sen-Yates-Grundy est sans biais par rapport au plan pour $V_p(\widehat{Y}_\pi)$.*

Démonstration.

$$\begin{aligned} E_p\{\widehat{V}_{SYG}(\widehat{Y}_\pi)\} &= E_p\left\{-\frac{1}{2}\sum_{k\in s^\star}\sum_{l\in s^\star}\frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\right\} \\ &= -\frac{1}{2}\sum_{k\in U^\star}\sum_{l\in U^\star}\frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2 E_p(I_k I_l) \\ &= -\frac{1}{2}\sum_{k\in U^\star}\sum_{l\in U^\star}\Delta_{kl}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2. \end{aligned}$$

□

Ces estimateurs ont donc de bonnes propriétés en termes de biais. On note toutefois qu'ils peuvent tous deux prendre des valeurs négatives. En ce qui concerne l'estimateur (1.2.5), le respect de la condition Sen-Yates-Grundy ($\Delta_{kl} \leq 0, \forall k \neq l$) garantit la non-négativité de l'estimateur. Cependant, cette condition ne garantit pas que l'estimateur (1.2.4) sera positif.

Supposons maintenant qu'un plan à taille fixe est employé et que la condition Sen-Yates-Grundy soit respectée. Ce contexte permet l'utilisation de l'un ou l'autre des estimateurs. Considérons le cas $y \propto \pi$ abordé en début de section. Dans ce cas, la variance est $V_p(\widehat{Y}_\pi) = 0$. Un estimateur de variance adéquat devrait donc, lui aussi, prendre la valeur 0 quelque soit l'échantillon. Dans ce cas, les estimateurs deviennent

$$\widehat{V}_{HT}(\widehat{Y}_\pi) = \sum_{k\in s^\star}\sum_{l\in s^\star}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} = \sum_{k\in s^\star}\sum_{l\in s^\star}\frac{\Delta_{kl}}{\pi_{kl}}\frac{a\pi_k}{\pi_k}\frac{a\pi_l}{\pi_l} = a^2\sum_{k\in s^\star}\sum_{l\in s^\star}\frac{\Delta_{kl}}{\pi_{kl}} \neq 0;$$

$$\widehat{V}_{SYG}(\widehat{Y}_\pi) = -\frac{1}{2}\sum_{k\in s^\star}\sum_{l\in s^\star}\frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2 = -\frac{1}{2}\sum_{k\in s^\star}\sum_{l\in s^\star}\frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{a\pi_k}{\pi_k}-\frac{a\pi_l}{\pi_l}\right)^2 = 0.$$

L'estimateur $\widehat{V}_{SYG}(\widehat{Y}_\pi)$ s'avère donc plus avantageux lorsque $y \propto \pi$. En effet, puisque $\widehat{V}_{HT}(\widehat{Y}_\pi)$ est sans biais pour $V_p(\widehat{Y}_\pi) = 0$, il doit prendre des valeurs positives et des valeurs négatives, alors que $\widehat{V}_{SYG}(\widehat{Y}_\pi)$ estime la variance de façon exacte à tout coup. Rappelons toutefois que ce dernier ne s'applique qu'aux plans à taille fixe.

1.2.2. Cas de l'échantillonnage par grappes

Dans le cas de l'échantillonnage par grappes, l'estimateur du total devient

$$\widehat{Y}_\pi = \sum_{i\in s} \frac{Y_i}{\pi_i} = \sum_{i\in s} d_i Y_i = \sum_{i\in U} d_i Y_i I_i. \tag{1.2.6}$$

Proposition 1.2.6. *Si $\pi_i > 0$, $\forall i \in U$, l'estimateur $\widehat{Y}_\pi$ est sans biais par rapport au plan pour Y.*

La preuve de la Proposition 1.2.6 est identique à celle de la Proposition 1.2.1 en remplaçant y_k par Y_i.

Proposition 1.2.7. *La variance par rapport au plan de l'estimateur $\widehat{Y}_\pi$ est donnée par*

$$V_p(\widehat{Y}_\pi) = \sum_{i\in U}\sum_{j\in U} \Delta_{ij}\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j} \tag{1.2.7}$$

Ce résultat est obtenu à l'aide d'une démarche similaire à celle de la Proposition 1.2.2. De façon analogue au cas de la section 1.2.1, pour un plan de taille fixe, la variance (1.2.7) peut aussi s'écrire comme

$$V_p(\widehat{Y}_\pi) = -\frac{1}{2}\sum_{i\in U}\sum_{j\in U} \Delta_{ij}\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2.$$

Dans le cas d'échantillonnage par grappes, l'estimateur Horvitz-Thompson de la variance devient

$$\widehat{V}_{HT}(\widehat{Y}_\pi) = \sum_{i\in s}\sum_{j\in s} \frac{\Delta_{ij}}{\pi_{ij}}\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j} \tag{1.2.8}$$

et l'estimateur Sen-Yates-Grundy, approprié dans le cas d'un plan de taille fixe, est

$$\widehat{V}_{SYG}(\widehat{Y}_\pi) = -\frac{1}{2}\sum_{i\in s}\sum_{j\in s} \frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2. \tag{1.2.9}$$

Lorsque $\pi_{ij} > 0$, $\forall i, j \in U$, les estimateurs (1.2.8) et (1.2.9) sont tous deux sans biais par rapport au plan pour $V_p(\widehat{Y}_\pi)$. La démonstration de ce résultat est similaire aux démonstrations des Propositions 1.2.4 et 1.2.5.

1.2.3. Cas de l'échantillonnage à deux degrés

Lorsque le plan de sondage employé est un plan à deux degrés, l'estimateur usuel du total est donné par

$$\widehat{Y}_{ts} = \sum_{i\in s}\sum_{k\in s_i^\star} \frac{y_{ik}}{\pi_i \pi_{k|i}} = \sum_{i\in s} \frac{\widehat{Y}_i}{\pi_i} = \sum_{k\in s^\star} \frac{y_k}{\pi_k}, \tag{1.2.10}$$

où $\widehat{Y}_i = \sum_{k \in s_i^\star} y_{ik}/\pi_{k|i}$.

Il s'agit de l'estimateur d'Horvitz-Thompson. Étudions les propriétés de cet estimateur en termes de biais et de variance. Dans un contexte d'échantillonnage à deux degrés, l'espérance par rapport au plan est notée $E_p(\cdot) = E_1 E_2(\cdot|s)$, où $E_1(\cdot)$ désigne l'espérance relative au premier degré et $E_2(\cdot|s)$ désigne l'espérance correspondant au deuxième degré. De façon analogue, la notation $V_p(\cdot)$ désigne $V_p(\cdot) = V_1 E_2(\cdot|s) + E_1 V_2(\cdot|s)$, où $V_1(\cdot)$ et $V_2(\cdot|s)$ correspondent à la variance par rapport au premier degré et deuxième degré, respectivement.

Proposition 1.2.8. *Si $\pi_i > 0$, $\forall i \in U$ et $\pi_{k|i} > 0$, $\forall k \in U_i^\star$, l'estimateur $\widehat{Y}_{ts}$ est sans biais par rapport au plan pour Y.*

Démonstration. Notons d'abord que l'erreur totale de $\widehat{Y}_{ts}$ peut s'écrire comme

$$\widehat{Y}_{ts} - Y = \underbrace{\left(\widehat{Y}_{os} - Y\right)}_{\text{erreur due au } 1^{\text{er}} \text{ degré}} + \underbrace{\left(\widehat{Y}_{ts} - \widehat{Y}_{os}\right)}_{\text{erreur due au } 2^{\text{e}} \text{ degré}}, \tag{1.2.11}$$

où $\widehat{Y}_{os} = \sum_{i \in s} Y_i/\pi_i$ est l'estimateur (1.2.6). Notons aussi que

$$E_2\left(\widehat{Y}_{ts}|s\right) = \sum_{i \in s} \frac{1}{\pi_i} E_2(\widehat{Y}_i|s) = \sum_{i \in s} \frac{Y_i}{\pi_i} = \widehat{Y}_{os}. \tag{1.2.12}$$

Le biais peut donc s'écrire

$$\begin{aligned}
B_p\left(\widehat{Y}_{ts}\right) &= E_p\left(\widehat{Y}_{ts}\right) - Y \\
&= E_1 E_2\left(\widehat{Y}_{ts} - Y|s\right) \\
&= E_1 E_2\left(\widehat{Y}_{os} - Y|s\right) + E_1 E_2\left(\widehat{Y}_{ts} - \widehat{Y}_{os}|s\right) && \text{par (1.2.11)} \\
&= E_1\left(\widehat{Y}_{os} - Y\right) + E_1\left\{E_2\left(\widehat{Y}_{ts}|s\right) - \widehat{Y}_{os}\right\} \\
&= E_1\left(\widehat{Y}_{os}\right) - Y + E_1\left(\widehat{Y}_{os} - \widehat{Y}_{os}\right) && \text{par (1.2.12)} \\
&= 0.
\end{aligned}$$

□

Proposition 1.2.9. *La variance par rapport au plan de l'estimateur $\widehat{Y}_{ts}$ est donnée par*

$$V_p(\widehat{Y}_{ts}) = \sum_{i \in U} \sum_{j \in U} \Delta_{ij} \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} + \sum_{i \in U} \frac{V_i}{\pi_i}, \tag{1.2.13}$$

où $V_i = \sum_{k \in U_i^\star} \sum_{l \in U_i^\star} \Delta_{kl|i} \frac{y_{ik}}{\pi_{k|i}} \frac{y_{il}}{\pi_{l|i}}$.

Démonstration.

$$\begin{aligned}
V_p(\widehat{Y}_{ts}) &= E_p\left(\widehat{Y}_{ts}-Y\right)^2 \\
&= E_1E_2\left\{\left(\widehat{Y}_{ts}-Y\right)^2\Big|s\right\} \\
&= E_1E_2\left[\left\{\left(\widehat{Y}_{os}-Y\right)+\left(\widehat{Y}_{ts}-\widehat{Y}_{os}\right)\right\}^2\Big|s\right] \\
&= E_1E_2\left\{\left(\widehat{Y}_{os}-Y\right)^2\Big|s\right\}+E_1E_2\left\{\left(\widehat{Y}_{ts}-\widehat{Y}_{os}\right)^2\Big|s\right\} \\
&\qquad +2E_1E_2\left\{\left(\widehat{Y}_{os}-Y\right)\left(\widehat{Y}_{ts}-\widehat{Y}_{os}\right)|s\right\} \\
&= E_1\left\{\left(\widehat{Y}_{os}-Y\right)^2\right\}+E_1E_2\left[\left\{\widehat{Y}_{ts}-E_2(\widehat{Y}_{ts}|s)\right\}^2\Big|s\right] \\
&\qquad +2E_1\left[\left(\widehat{Y}_{os}-Y\right)\left\{E_2\left(\widehat{Y}_{ts}|s\right)-\widehat{Y}_{os}\right\}\right] \\
&= E_1\left\{\left(\widehat{Y}_{os}-Y\right)^2\right\}+E_1E_2\left[\left\{\widehat{Y}_{ts}-E_2(\widehat{Y}_{ts}|s)\right\}^2\Big|s\right] \\
&= V_1\left(\widehat{Y}_{os}\right)+E_1\left\{V_2\left(\widehat{Y}_{ts}|s\right)\right\}.
\end{aligned}$$

Puisque $\widehat{Y}_{os}$ a la forme de l'estimateur d'Horvitz-Thompson, l'expression (1.2.2) nous indique que la variance due au premier degré $V_1(\widehat{Y}_{os})$ est donné par

$$V_1(\widehat{Y}_{os})=\sum_{i\in U}\sum_{j\in U}\Delta_{ij}\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}.$$

La variance due au deuxième degré $E_1\left\{V_2\left(\widehat{Y}_{ts}|s\right)\right\}$ se réécrit comme

$$\begin{aligned}
E_1\left\{V_2\left(\widehat{Y}_{ts}|s\right)\right\} &= E_1\left\{V_2\left(\sum_{i\in s}\frac{\widehat{Y}_i}{\pi_i}\Big|s\right)\right\} \\
&= E_1\left\{\sum_{i\in s}\frac{V_2\left(\widehat{Y}_i|s\right)}{\pi_i^2}\right\} \\
&= E_1\left(\sum_{i\in s}\frac{V_i}{\pi_i^2}\right) \\
&= \sum_{i\in U}\frac{V_i}{\pi_i},
\end{aligned}$$

où la notation V_i désigne $V_2\left(\widehat{Y}_i|s\right)$. Puisque $\widehat{Y}_i$ a la forme de l'estimateur Horvitz-Thompson, l'expression (1.2.2) nous permet d'écrire

$$V_i=\sum_{k\in U_i^\star}\sum_{l\in U_i^\star}\Delta_{kl|i}\frac{y_{ik}}{\pi_{k|i}}\frac{y_{il}}{\pi_{l|i}},$$

où $\Delta_{kl|i}=\pi_{kl|i}-\pi_{k|i}\pi_{l|i}$. En combinant les variances dues au premier et au deuxième degré, on trouve (1.2.13). □

On note que cette variance demeure inconnue puisque la variable d'intérêt n'est pas disponible pour toutes les unités de la population. Un estimateur de (1.2.13) est donné par

$$\widehat{V}(\widehat{Y}_{ts}) = \sum_{i\in s}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{\widehat{Y}_i}{\pi_i}\frac{\widehat{Y}_j}{\pi_j} + \sum_{i\in s}\frac{\widehat{V}_i}{\pi_i}, \tag{1.2.14}$$

où $\widehat{V}_i = \sum_{k\in s_i^\star}\sum_{l\in s_i^\star}\frac{\Delta_{kl|i}}{\pi_{kl|i}}\frac{y_{ik}}{\pi_{k|i}}\frac{y_{il}}{\pi_{l|i}}$.

Proposition 1.2.10. *Si $\pi_{ij} > 0$, $\forall i,j \in U$ et si $\pi_{kl|i} > 0$, $\forall k,l \in U_i^\star$, l'estimateur $\widehat{V}(\widehat{Y}_{ts})$ est sans biais par rapport au plan pour $V_p(\widehat{Y}_{ts})$.*

Démonstration. L'espérance de $\widehat{V}(\widehat{Y}_{ts})$ est donnée par $E_1E_2\left\{\widehat{V}(\widehat{Y}_{ts})|s\right\}$. On s'intéresse d'abord à $E_2\left\{\widehat{V}(\widehat{Y}_{ts})|s\right\}$.

$$\begin{aligned}
E_2\left\{\widehat{V}(\widehat{Y}_{ts})|s\right\} &= E_2\left(\sum_{i\in s}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{\widehat{Y}_i}{\pi_i}\frac{\widehat{Y}_j}{\pi_j}\Big|s\right) + E_2\left(\sum_{i\in s}\frac{\widehat{V}_i}{\pi_i}\Big|s\right)\\
&= \sum_{i\in s}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{E_2\left(\widehat{Y}_i\widehat{Y}_j|s\right)}{\pi_i\pi_j} + \sum_{i\in s}\frac{E_2\left(\widehat{V}_i|s\right)}{\pi_i}.
\end{aligned}$$

Or, puisque $E_2\left(\widehat{Y}_i\widehat{Y}_j|s\right) = \begin{cases} Y_iY_j & \text{si } i\neq j \text{ (par indépendance)}\\ V_i + Y_i^2 & \text{si } i = j\end{cases}$

on a

$$\begin{aligned}
E_2\left\{\widehat{V}(\widehat{Y}_{ts})|s\right\} &= \sum_{\substack{i\in s\\ i\neq j}}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j} + \sum_{i\in s}\frac{1-\pi_i}{\pi_i^2}\left(V_i + Y_i^2\right) + \sum_{i\in s}\frac{E_2\left(\widehat{V}_i|s\right)}{\pi_i}\\
&= \sum_{i\in s}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j} + \sum_{i\in s}\frac{1-\pi_i}{\pi_i^2}V_i + \sum_{i\in s}\frac{V_i}{\pi_i}\\
&= \sum_{i\in s}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j} + \sum_{i\in s}\frac{V_i}{\pi_i^2}.
\end{aligned}$$

Finalement,

$$\begin{aligned}
E_p\left\{\widehat{V}(\widehat{Y}_{ts})\right\} &= E_1\left(\sum_{i\in s}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j} + \sum_{i\in s}\frac{V_i}{\pi_i^2}\right)\\
&= \sum_{i\in U}\sum_{j\in U}\frac{\Delta_{ij}}{\pi_{ij}}\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}E_1(I_iI_j) + \sum_{i\in U}\frac{V_i}{\pi_i^2}E_1(I_i)\\
&= \sum_{i\in U}\sum_{j\in U}\Delta_{ij}\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j} + \sum_{i\in U}\frac{V_i}{\pi_i}\\
&= V_p\left(\widehat{Y}_{ts}\right).
\end{aligned}$$

□

1.2.4. Estimateurs de calage

À l'étape de l'estimation, il est possible d'incorporer de l'information auxiliaire afin d'améliorer la qualité des estimateurs. On appellera information auxiliaire toute variable observée dans l'échantillon et dont on connaît le total au niveau de la population. Le calage est une technique de pondération garantissant que les estimations obtenues pour les variables auxiliaires au moyen de cette nouvelle pondération coïncident avec les totaux connus au niveau de la population. Un estimateur qui respecte cette propriété est dit cohérent.

Supposons que l'on dispose de Q variables auxiliaires dont les valeurs observées pour l'unité k sont notées $\boldsymbol{x}_k = (x_{1k}, ..., x_{Qk})^\top$. Le vecteur formé des totaux au niveau de la population de ces variables auxiliaires est $\boldsymbol{X} = (X_1, ..., X_Q)^\top$. On suppose que la relation entre y_k et $\boldsymbol{x}_k$ peut être décrite au moyen du modèle suivant :

$$y_k = \boldsymbol{x_k}^\top \boldsymbol{\beta} + \epsilon_k, \tag{1.2.15}$$

où $\boldsymbol{\beta} = (\beta_1, ..., \beta_Q)^\top$ est un vecteur de paramètres inconnus et ϵ_k est une composante de bruit telle que

$$\begin{aligned} &E_m(\epsilon_k) = 0; \\ &E_m(\epsilon_k \epsilon_l) = 0 \text{ si } k \neq l; \\ &V_m(\epsilon_k) = \sigma^2 \phi_k \text{ (pour } \phi_k \text{ connu)}, \end{aligned}$$

où $\phi_k > 0$ est un coefficient connu associé à l'unité k. Lorsque les ϕ_k sont tous égaux, on se trouve dans un contexte de modèle homoscédastique, sinon on parlera d'un modèle hétéroscédastique.

On note que sous ce modèle, le total Y peut être exprimé comme

$$Y = \sum_{k \in U^\star} \left(\boldsymbol{x_k}^\top \boldsymbol{\beta} + \epsilon_k \right).$$

L'estimateur GREG est obtenu en estimant chacun des deux termes de cette expression séparément, ce qui conduit à

$$\widehat{Y}_{GREG} = \sum_{k \in U^\star} \boldsymbol{x_k}^\top \widehat{\boldsymbol{B}} + \sum_{k \in s^\star} \frac{e_k}{\pi_k},$$

où

$$\widehat{\boldsymbol{B}} = \left(\sum_{k \in s^\star} d_k \boldsymbol{x_k} \phi_k^{-1} \boldsymbol{x_k}^\top \right)^{-1} \sum_{k \in s^\star} d_k \boldsymbol{x_k} \phi_k^{-1} y_k$$

et $e_k = y_k - \boldsymbol{x_k}^\top \widehat{\boldsymbol{B}}$.

L'estimateur GREG peut également s'écrire sous la forme d'une somme pondérée

$$\widehat{Y}_{GREG} = \sum_{k \in s^\star} w_k(s^\star) y_k,$$

où

$$w_k(s^\star) = d_k \left\{ 1 + \phi_k^{-1} \left(\boldsymbol{X} - \widehat{\boldsymbol{X}}_{\boldsymbol{\pi}} \right)^\top \widehat{\boldsymbol{T}}^{-1} \boldsymbol{x_k} \right\},$$

$\widehat{\boldsymbol{X}}_{\boldsymbol{\pi}} = \sum_{k \in s^\star} d_k \boldsymbol{x_k}$ et $\widehat{\boldsymbol{T}} = \sum_{k \in s^\star} d_k \boldsymbol{x_k} \phi_k^{-1} \boldsymbol{x_k}^\top$.
Notons que la cohérence entre $\widehat{\boldsymbol{X}}_{\boldsymbol{GREG}}$ et $\boldsymbol{X}$ est garantie puisque

$$\begin{aligned}
\widehat{\boldsymbol{X}}_{\boldsymbol{GREG}}^\top &= \sum_{k \in s^\star} w_k(s^\star) \boldsymbol{x_k}^\top \\
&= \sum_{k \in s^\star} d_k \boldsymbol{x_k}^\top + \boldsymbol{X}^\top \widehat{\boldsymbol{T}}^{-1} \sum_{k \in s^\star} d_k \boldsymbol{x_k} \phi_k^{-1} \boldsymbol{x_k}^\top - \widehat{\boldsymbol{X}}_{\boldsymbol{\pi}}^\top \widehat{\boldsymbol{T}}^{-1} \sum_{k \in s^\star} d_k \boldsymbol{x_k} \phi_k^{-1} \boldsymbol{x_k}^\top \\
&= \widehat{\boldsymbol{X}}_{\boldsymbol{\pi}}^\top + \boldsymbol{X}^\top - \widehat{\boldsymbol{X}}_{\boldsymbol{\pi}}^\top \\
&= \boldsymbol{X}^\top.
\end{aligned}$$

L'estimateur GREG sera précis dans un contexte où la relation entre $\boldsymbol{x}$ et y est forte et linéaire. Si cette relation est parfaite, c'est-à-dire $y_k = \boldsymbol{x_k}^\top \boldsymbol{\beta}$, $\forall k \in U^\star$, on a $\widehat{Y}_{GREG} = Y$, $\forall s^\star \in \mathcal{Q}$. Il en découle que $EQM_p(\widehat{Y}_{GREG}) = 0$. L'estimateur GREG est biaisé, mais lorsque la taille d'échantillon $n^\star$ est grande, le biais devient négligeable par rapport à la variance. Sa variance approximative, obtenue au moyen d'un développement en série de Taylor de premier ordre, est donnée par

$$AV_p(\widehat{Y}_{GREG}) = \sum_{k \in U^\star} \sum_{l \in U^\star} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l},$$

où $E_k = y_k - \boldsymbol{x_k}^\top \boldsymbol{B}$ et

$$\boldsymbol{B} = \left(\sum_{k \in U^\star} \boldsymbol{x_k} \phi_k^{-1} \boldsymbol{x_k}^\top \right)^{-1} \sum_{k \in U^\star} \boldsymbol{x_k} \phi_k^{-1} y_k.$$

On l'estime par

$$\widehat{AV}_p(\widehat{Y}_{GREG}) = \sum_{k \in s^\star} \sum_{l \in s^\star} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}. \tag{1.2.16}$$

Un cas particulier de l'estimateur GREG consiste à prendre $\boldsymbol{x}_k = 1$ et $\phi_k = 1$, pour tout k. Dans ce cas, on a

$$\begin{aligned}\widehat{Y}_{GREG} &= \sum_{k\in s^\star} w_k(s^\star) y_k \\ &= \sum_{k\in s^\star} d_k \left\{1 + (N^\star - \widehat{N}_\pi)\widehat{N}_\pi^{-1}\right\} y_k \\ &= \sum_{k\in s^\star} d_k \left\{1 + N^\star \widehat{N}_\pi^{-1} - 1\right\} y_k \\ &= N^\star \frac{\widehat{Y}_\pi}{\widehat{N}_\pi} \equiv \widehat{Y}_{HA}, \end{aligned} \tag{1.2.17}$$

où $\widehat{N}_\pi = \sum\limits_{k\in s^\star} d_k$. L'estimateur (1.2.17) est appelé estimateur de Hájek.

Un second cas particulier est celui pour lequel $\boldsymbol{x}_k = x_k$ et $\phi_k = x_k$. Dans ce cas, l'estimateur GREG devient

$$\begin{aligned}\widehat{Y}_{GREG} &= \sum_{k\in s^\star} w_k(s^\star) y_k \\ &= \sum_{k\in s^\star} d_k \left\{1 + x_k^{-1}(X - \widehat{X}_\pi)\widehat{X}_\pi^{-1} x_k\right\} y_k \\ &= \sum_{k\in s^\star} d_k \left\{1 + X\widehat{X}_\pi^{-1} - 1\right\} y_k \\ &= X \frac{\widehat{Y}_\pi}{\widehat{X}_\pi} \equiv \widehat{Y}_{ra}. \end{aligned} \tag{1.2.18}$$

L'estimateur (1.2.18) est l'estimateur par le ratio.

1.3. Entropie d'un plan de sondage

On définit l'entropie d'un plan de sondage $p(\cdot)$ par

$$I(p) = -\sum_{s^\star\in\mathcal{Q}} p(s^\star)\log p(s^\star). \tag{1.3.1}$$

Il s'agit d'une mesure de désordre. Lorsqu'un plan a une grande entropie, il est très difficile de prédire le type d'échantillon qui sera tiré. Comme nous le verrons à la section 1.4, le concept d'entropie est particulièrement utile dans un contexte d'estimation de la variance.

Proposition 1.3.1. *Le plan de sondage qui maximise l'entropie parmi tous les plans à probabilités égales ($\pi_k = \pi$ pour tout $k \ni U^\star$) est le plan de Bernoulli pour lequel $\pi = 1/2$.*

DÉMONSTRATION. Puisque la taille est aléatoire, l'ensemble $\mathcal{Q}$ se compose de $2^{N^\star}$ échantillons possibles, soit $s_1^\star, ..., s_{2^{N^\star}}^\star$. On cherche à maximiser (1.3.1) sous la contrainte $\sum_{s^\star \in \mathcal{Q}} p(s^\star) = 1$. On considère donc la fonction lagrangienne

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

Sa dérivée par rapport à $p(s_t^\star)$ nous donne

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

En la posant égale à zéro, on obtient

$$p(s_t^\star) = e^{\lambda - 1}, \tag{1.3.2}$$

$\forall s_t^\star \in \mathcal{Q}$. En insérant cette expression dans la contrainte $\sum_{s^\star \in \mathcal{Q}} p(s^\star) = 1$, on a

$$\begin{aligned} \sum_{s^\star \in \mathcal{Q}} e^{\lambda - 1} &= 1 \\ \iff e^{\lambda - 1} &= \frac{1}{2^{N^\star}} \\ \iff p(s_t^\star) &= \frac{1}{2^{N^\star}}, \qquad \text{(par 1.3.2)} \end{aligned}$$

$\forall s_t^\star \in \mathcal{Q}$. Ceci revient au plan Bernoulli avec $\pi = 1/2$, puisque pour ce dernier on a

$$p(s^\star) = \pi^{n_s^\star}(1-\pi)^{N^\star - n_s^\star} = \left(\frac{1}{2}\right)^{n_s^\star} \left(1 - \frac{1}{2}\right)^{N^\star - n_s^\star} = \frac{1}{2^{N^\star}}$$

□

Proposition 1.3.2. *Le plan de sondage qui maximise l'entropie parmi les plans à taille fixe $n^\star$ à probabilités égales est le plan aléatoire simple sans remise.*

DÉMONSTRATION. Puisque la taille est fixe, l'ensemble $\mathcal{Q}$ se compose de $\binom{N^\star}{n^\star}$ échantillons possibles, soit $s_1^\star, ..., s_{\binom{N^\star}{n^\star}}^\star$. Une optimisation lagrangienne analogue à celle de la démonstration précédente permet d'obtenir

$$p(s_t^\star) = \frac{1}{\binom{N^\star}{n^\star}},$$

$\forall s_t^\star \in \mathcal{Q}$, ce qui est la définition de l'échantillonnage aléatoire simple sans remise.

□

Notons que parmi les plans à probabilités inégales, c'est le plan de Poisson qui maximise l'entropie, alors que si l'on se restreint aux plans à probabilités inégales

à taille fixe, ce rôle est joué par le plan de Poisson conditionnel (Hájek, 1981). Les plans Rao-Sampford et Poisson séquentiel sont tous deux des plans à grande entropie (Berger, 1998, 2011).

1.4. Alternatives à l'estimation classique de la variance

Les méthodes classiques d'estimation de la variance introduites à la section 1.2 présentent certains défis. En effet, les estimateurs présentés font intervenir les probabilités d'inclusion d'ordre deux qui peuvent être complexes à calculer. Lorsque le plan de sondage comporte plusieurs degrés, cette difficulté est accentuée, sauf dans le cas où la fraction de sondage au premier degré est petite auquel cas il est possible d'utiliser des estimateurs de variance relativement simples. Aussi, l'évaluation d'une double somme peut s'avérer laborieuse lorsque la taille de l'échantillon est grande. On présente donc quelques alternatives visant à simplifier l'estimation de la variance.

1.4.1. Approximation des probabilités d'inclusion d'ordre deux

Une approche consiste à approximer π_{kl} en fonction de π_k et π_l, qui sont toujours disponibles. Les π_{kl} sont ensuite remplacées par leur valeur approximative dans la forme (1.2.3) de la variance. Cette expression permet donc de déduire un estimateur qui ne requiert pas les π_{kl}. Notons que les estimateurs obtenus par cette approche présentent de bonnes propriétés lorsqu'un plan de sondage à grande entropie est utilisé et lorsque les tailles d'échantillon et de population sont grandes ; voir Haziza et al. (2008).

1.4.1.1. *Approximation de Hájek*

Une approximation des π_{kl} proposée par Hájek (1964) est donnée par

$$\pi_{kl} \approx \pi_k \pi_l \left\{ 1 - \frac{(1-\pi_k)(1-\pi_l)}{d} \right\}, \tag{1.4.1}$$

où $d = \sum_{k \in U^\star} \pi_k(1-\pi_k)$. En substituant les π_{kl} de l'équation de variance (1.2.3) par (1.4.1), on obtient

$$V_{HA}(\widehat{Y}_\pi) = \sum_{k \in U^\star} \pi_k(1-\pi_k) \left\{ \frac{y_k}{\pi_k} - \frac{\sum_{l \in U^\star} (1-\pi_l) y_l}{d} \right\}^2.$$

Cette expression permet de déduire l'estimateur de variance

$$\widehat{V}_{HA}(\widehat{Y}_\pi) = \frac{n}{n-1}\sum_{k\in s^\star}(1-\pi_k)\left\{\frac{y_k}{\pi_k} - \frac{\sum_{l\in s^\star}\frac{1-\pi_l}{\pi_l}y_l}{\sum_{l\in s^\star}(1-\pi_l)}\right\}^2. \tag{1.4.2}$$

Notons que ce dernier ne dépend que des probabilités d'inclusion d'ordre un et s'exprime au moyen d'une simple somme.

1.4.1.2. *Approximations de Brewer-Donadio*

Brewer et Donadio (2003) proposent une famille d'approximations des π_{kl} de la forme

$$\pi_{kl} \approx \pi_k\pi_l\frac{c_k + c_l}{2}, \tag{1.4.3}$$

où c_k et c_l sont fixés. Brewer et Donadio (2003) ont considéré les choix de c_k suivants :

(i) $c_k = \dfrac{n^\star - 1}{n^\star - \pi_k}$;

(ii) $c_k = c = \dfrac{n^\star - 1}{n^\star - \frac{1}{n^\star}\sum\limits_{l\in U^\star}\pi_l^2}$;

(iii) $c_k = \dfrac{n^\star - 1}{n^\star - 2\pi_k - \frac{1}{n^\star}\sum\limits_{l\in U^\star}\pi_l^2}$;

(iv) $c_k = \dfrac{n^\star - 1}{n^\star}\left\{1 + \dfrac{2\pi_k}{n^\star} - \dfrac{1}{(n^\star)^2}\sum\limits_{l\in U^\star}\pi_l^2\right\}$.

En remplaçant les probabilités d'inclusion d'ordre deux de la variance (1.2.3) par (1.4.3), on trouve

$$V_{BD}(\widehat{Y}_\pi) = \sum_{k\in U^\star}\pi_k(1-c_k\pi_k)\left(\frac{y_k}{\pi_k} - \frac{Y}{n^\star}\right)^2.$$

Cette expression est estimée par

$$\widehat{V}_{BD}(\widehat{Y}_\pi) = \frac{n^\star}{n^\star - 1}\sum_{k\in s^\star} b_k\left(\frac{y_k}{\pi_k} - \frac{\widehat{Y}_\pi}{n^\star}\right)^2, \tag{1.4.4}$$

où les facteurs b_k correspondant aux approximations (i) à (iv) sont respectivement donnés par

(i) $b_k = 1 - \pi_k$;

(ii) $b_k = 1 - \pi_k + \dfrac{\pi_k}{n^\star} - \dfrac{1}{(n^\star)^2}\sum\limits_{l\in U^\star}\pi_l^2$;

(iii) $b_k = 1 - \pi_k - \dfrac{\pi_k}{n^\star} - \dfrac{1}{(n^\star)^2}\sum\limits_{l\in U^\star}\pi_l^2$;

(iv) $b_k = 1 - \pi_k - \dfrac{\pi_k}{n^\star - 1} + \dfrac{1}{n^\star(n^\star - 1)} \sum\limits_{l \in U^\star} \pi_l^2$.

Comme $\widehat{V}_{HA}(\widehat{Y}_\pi)$, $\widehat{V}_{BD}(\widehat{Y}_\pi)$ s'écrit au moyen d'une simple somme qui ne dépend pas des probabilités d'inclusion d'ordre deux.

Haziza et al. (2008) montrent que les estimateurs (1.4.2) et (1.4.4) peuvent s'écrire sous la forme générale

$$\widehat{V}_{Gen}(\widehat{Y}_\pi) = \sum_{k \in s^\star} c_k e_k^2, \tag{1.4.5}$$

où $e_k = \frac{y_k}{\pi_k} - \widehat{B}$ avec $\widehat{B} = \dfrac{\sum\limits_{k \in s^\star} a_k (y_k/\pi_k)}{\sum\limits_{k \in s^\star} a_k}$ et a_k et c_k sont des constantes spécifiées au tableau 1.1.

TABLEAU 1.1. Constantes a_k et c_k pour la forme générale $\widehat{V}_{Gen}(\widehat{Y}_\pi)$

Estimateur	$\boldsymbol{c_k}$	$\boldsymbol{a_k}$
$\widehat{V}_{HA}(\widehat{Y}_\pi)$	$\frac{n^\star}{n^\star - 1}(1 - \pi_k)$	c_k
$\widehat{V}_{BD}(\widehat{Y}_\pi)$	$\frac{n^\star}{n^\star - 1} b_k$	1

Notons qu'il existe plusieurs autres approximations des π_{kl}, par exemple celles proposées par Hartley et Rao (1962), Berger (1998), Deville (1999) et Rosén (1991).

1.4.2. Approche par simulation Monte-Carlo

Cette approche empirique, introduite par Fattorini (2006), consiste à estimer les π_{kl} au moyen d'une simulation Monte-Carlo. Considérons un échantillon $s^\star$, tiré selon un certain plan de sondage complexe $p(\cdot)$, tel que $\pi_{kl} > 0$, pour $k, l \in U^\star$. Cette méthode par simulation consiste à tirer de façon indépendante un très grand nombre R d'échantillons selon le plan $p(\cdot)$. Ces échantillons sont notés $s_1^\star$,...,$s_R^\star$. La probabilité d'inclusion d'ordre deux des unités k et l est estimée par

$$\hat{\pi}_{kl} = \frac{\sum\limits_{r=1}^{R} \mathbb{1}_{(k \in s_r^\star,\, l \in s_r^\star)}}{R}.$$

En effet, puisque π_{kl} est la probabilité qu'à la fois k et l soient inclus dans l'échantillon, elle est estimée par la proportion d'échantillons (parmi les R tirés) qui contiennent ces deux unités. Ces $\hat{\pi}_{kl}$ remplacent ensuite les π_{kl} à l'intérieur d'un estimateur classique de la variance, tel que l'estimateur (1.2.4) ou (1.2.5). On

choisit une valeur de R suffisamment grande, de manière à ce que $\widehat{V}(\widehat{Y}_\pi)$ présente de bonnes propriétés en matière de biais et de variance ; voir Fattorini (2006) et Thompson et Wu (2008).

1.4.3. Approche d'Ohlsson

Ohlsson (1998) suggère une approche alternative, lorsqu'une variable auxiliaire x, dont le total est connu, est disponible pour les unités de l'échantillon. L'approche d'Ohlsson peut être vue comme une généralisation du résultat bien connu suivant. Dans un contexte de plan à probabilités égales, la stratégie composée du plan aléatoire simple et de l'estimateur d'Horvitz-Thompson est essentiellement équivalente en termes de variance à la stratégie qui consiste à utiliser l'échantillonnage de Bernoulli et l'estimateur de Hájek ; voir Särndal et al. (1992).

Dans le cas de l'approche d'Ohlsson, l'objectif est d'estimer la variance de $\widehat{Y}_\pi$ dans le cadre d'un plan de Poisson séquentiel. Son approche consiste à approximer la stratégie composée d'un plan de Poisson séquentiel et de l'estimateur $\widehat{Y}_\pi$ par une stratégie approximativement équivalente composée d'un plan de Poisson et de l'estimateur par le ratio. Cette stratégie d'approximation revient à utiliser l'estimateur de variance

$$\widehat{V}_{Oh} \equiv \widehat{AV}_p(\widehat{Y}_{ra}) = \sum_{k \in s^\star} \frac{1-\pi_k}{\pi_k^2} \left(y_k - \frac{x_k \widehat{Y}_\pi}{X} \right)^2, \tag{1.4.6}$$

où $s^\star$ est de taille suffisamment grande. Puisqu'on fait intervenir le plan de Poisson, les probabilités d'inclusion d'ordre deux ne sont plus nécessaires au calcul de la variance. On note aussi que l'estimateur résultant se compose d'une simple somme, voir Särndal (1996). On conjecture que cette approche n'est pas seulement appropriée pour le plan de Poisson séquentiel, mais également pour tous les autres plans à grande entropie.

Il est important de noter que l'estimateur de variance simplifié est déduit d'une stratégie qui prévoit un échantillonnage au même niveau que le plan de sondage mis en œuvre. Par exemple, si le plan de Poisson séquentiel a été employé pour sélectionner des grappes, alors la stratégie d'approximation doit faire intervenir un plan de Poisson de grappes. Cette condition d'utilisation de l'approche d'Ohlsson sera analysée dans les prochaines sections. Il serait pratique d'approximer une stratégie faisant intervenir un plan de grappes par une stratégie au niveau des éléments.

1.5. Convergence de l'estimateur de variance de Hájek

Hájek (1981) a montré que l'estimateur de la variance (1.4.2) est un estimateur convergent de la vraie variance sous le plan de Poisson conditionnel. En effet, on peut montrer que, sous certaines conditions de régularité,

$$nE_p\left\{\left|\widehat{V}_{HA}(\widehat{Y}_\pi) - V_p(\widehat{Y}_\pi)\right|\right\} \to 0$$

lorsque $N \to \infty$. La démonstration de ce résultat repose sur un cadre asymptotique particulier que l'on peut décrire comme suit.

Considérons une suite de populations imbriquées $\{U_1^\star \subset U_2^\star \subset ...\}$, où $U_t^\star$ est de taille $N_t^\star$. On note $p_t(\cdot)$ le plan de sondage à grande entropie employé pour tirer, de la population $U_t^\star$, un échantillon $s_t^\star$ de taille $n_t^\star$. On dit, de manière informelle, que l'entropie est grande lorsqu'elle est proche, en un certain sens, de celle du plan de Poisson conditionnel, pour des tailles d'échantillon et de population suffisamment grandes ; voir Berger (2011). Les probabilités d'inclusion d'ordre un relatives au plan de sondage $p_t(s^\star)$ sont notées $\pi_{k,t}$ et on suppose que $\lim_{t\to\infty} \mathcal{D}_t = \infty$, où $\mathcal{D}_t = \sum_{k\in U_t^\star} \pi_{k,t}(1-\pi_{k,t})$. Il s'ensuit donc que $n_t^\star \to \infty$ et $N_t^\star \to \infty$.

Berger (2011) montre que si un estimateur (ponctuel ou de variance) est convergent sous le plan de Poisson conditionnel, alors il est également convergent sous un plan «proche» du plan de Poisson conditionnel. Comme mesure de proximité, Berger (2011) utilise trois critères. On utilise la notation $r_t(\cdot)$ pour désigner le plan de Poisson conditionnel et $p_t(\cdot)$ pour désigner un plan de sondage arbitraire.

Définition 1.5.1. *On dit que $p_t(s^\star)$ converge vers $r_t(s^\star)$ par rapport à leur entropie si*

$$\lim_{t\to\infty}\{I(r_t) - I(p_t)\} = 0. \tag{1.5.1}$$

Définition 1.5.2. *On dit que $p_t(s^\star)$ converge vers $r_t(s^\star)$ par rapport à la norme de variation totale si*

$$\lim_{t\to\infty}\|p_t - r_t\|_1 = 0, \tag{1.5.2}$$

où $\|p_t - r_t\|_1 = \sum_{s^\star\in\mathcal{Q}} |p_t(s^\star) - r_t(s^\star)|$.

Définition 1.5.3. *On dit que $p_t(s^\star)$ converge vers $r_t(s^\star)$ par rapport à la distance du khi-carré si*

$$\lim_{t\to\infty}\|p_t - r_t\|_2 = 0, \tag{1.5.3}$$

où $\|p_t - r_t\|_2 = \sum_{s^\star \in \mathcal{Q}} r_t(s^\star) \left\{ 1 - \frac{p_t(s^\star)}{r_t(s^\star)} \right\}^2$.

Berger (2011) montre que l'estimateur de variance de Hájek est convergent si (1.5.1) ou (1.5.2) ou (1.5.3) est satisfaite.

Chapitre 2

ESTIMATION SIMPLIFIÉE DE LA VARIANCE

2.1. Introduction

Les agences de statistiques telles que Statistique Canada ont pour coutume de produire des fichiers de données rectangulaires, les colonnes de ce fichier correspondant aux variables d'intérêt et les lignes aux unités échantillonnées. La dernière colonne du fichier correspond généralement à un système de pondération. Dans ce mémoire, nous supposons que les poids sont définis comme l'inverse de la probabilité d'inclusion, bien qu'en pratique, la dernière colonne correspond aux poids finaux, ces derniers étant ajustés pour la non-réponse totale et par calage. Une partie des fichiers produits par Statistique Canada est partagée avec des utilisateurs externes (par exemple, des chercheurs en milieu universitaire).

Afin d'estimer la variance des estimateurs ponctuels, on devrait en principe fournir la probabilité d'inclusion d'ordre deux de manière à pouvoir calculer les estimateurs de variance (1.2.4) ou (1.2.5). En pratique, pour des raisons de confidentialité, ces probabilités ne sont jamais fournies aux utilisateurs. Une autre approche consiste à fournir B colonnes de poids bootstrap (par exemple, $B = 500$) grâce auxquelles il est possible d'estimer la variance des estimateurs ponctuels pour certains plans de sondage. Dans certains cas, les poids bootstrap ne sont pas fournis aux utilisateurs externes. De plus, dans un contexte d'enquête par grappes, les identifiants de grappes ne sont pas fournis, encore une fois pour des raisons de confidentialité.

Lorsque la sélection est réalisée directement à partir de la population d'éléments (autrement dit une seule étape aléatoire permet de sélectionner un échantillon d'éléments), il sera possible pour l'utilisateur d'estimer la variance. Dans ce

cas, les valeurs de π_k et y_k des éléments sélectionnés sont partagées à l'utilisateur. Celui-ci détient donc toute l'information nécessaire pour employer la méthode d'approximation des π_{kl} ou l'approche d'Ohlsson. Ainsi, il est possible d'utiliser l'une ou l'autre de ces méthodes, à condition que la sélection d'éléments ait été réalisée à partir d'un plan à grand entropie.

Si l'enquête fait intervenir une sélection de grappes, que ce soit dans le cadre d'un plan à un ou à deux degrés, l'utilisateur ne sera pas en mesure d'estimer la variance. D'abord, si les poids bootstrap ne sont pas fournis, les méthodes usuelles ne peuvent être employées. De plus, puisque l'information permettant d'associer les éléments à leur grappe respective est confidentielle, les méthodes d'Ohlsson et d'approximation des π_{kl} ne conviennent pas. En effet, ces deux approches requièrent les probabilités d'inclusion d'ordre un ainsi que les totaux (ou l'estimation des totaux) au niveau des grappes. Or, l'utilisateur ne connaît π_k et y_k (pour $k \in s^\star$) qu'au niveau des éléments et ne peut en déduire Y_i et π_i ($i \in s$), puisque l'association entre un élément k et sa grappe i est inconnue.

L'utilisateur dispose donc de la variable d'intérêt, de l'information auxiliaire et des poids de sondage ajustés des éléments de l'échantillon, mais ne peut estimer la variance, puisque la littérature ne présente aucune approche convenant à cette situation. Notre recherche sera axée sur l'approche d'Ohlsson et sur la façon dont elle pourrait être adaptée pour répondre aux besoin de ce contexte.

L'approximation d'une stratégie au niveau des grappes par une stratégie au niveau des éléments sera explorée. L'idée est d'amener un estimateur de variance qui remplacerait l'estimateur (1.4.6) de la méthode d'Ohlsson, lorsqu'une stratégie au niveau des grappes est approximée par une stratégie faisant intervenir un plan de Poisson d'éléments. Cet objectif sera poursuivi dans un cadre d'échantillonnage proportionnel à la taille.

Le cas de l'échantillonnage de Poisson comporte plusieurs avantages en termes d'estimation de la variance. Ce plan de sondage faisant intervenir des expériences de Bernoulli indépendantes, on a $\pi_{kl} = \pi_k \pi_l$, pour $k \neq l$. Ceci a pour effet de considérablement simplifier l'estimation de la variance. En effet, l'estimateur de variance (1.2.4) se simplifie pour donner

$$\widehat{V}_{HT}(\widehat{Y}_\pi) = \sum_{k \in s^\star} \frac{1 - \pi_k}{\pi_k^2} y_k^2. \tag{2.1.1}$$

L'estimateur (2.1.1) ne requiert pas les π_{kl} et s'exprime au moyen d'une simple somme.

2.2. Étude de l'approche d'Ohlsson

2.2.1. Approche d'Ohlsson : une approximation des π_{kl}

Considérons le cas d'une enquête réalisée au moyen d'un plan de sondage de taille fixe, qui prévoit une sélection au niveau des éléments. On suppose que l'échantillon d'éléments $s^\star$ a été tiré selon un plan à taille fixe avec probabilités proportionnelles à la taille (voir section 1.1). On suppose également que l'estimateur $\widehat{Y}_\pi = \sum_{k\in s^\star} d_k y_k$ a été utilisé afin d'estimer le total $Y = \sum_{k\in U^\star} y_k$. La variance de $\widehat{Y}_\pi$ est donnée par

$$V_p(\widehat{Y}_\pi) = -\frac{1}{2}\sum_{k\in U^\star}\sum_{l\in U^\star}\Delta_{kl}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2. \tag{2.2.1}$$

L'objectif est d'approximer la stratégie consistant en un plan avec probabilités proportionnelles à la taille et l'estimateur $\widehat{Y}_\pi$ par une stratégie équivalente en termes de variance, consistant en un plan de Poisson et un estimateur par calage (que l'on explicitera ci-dessous). Cette approche a été suggérée par Ohlsson (1998). Considérons un plan de Poisson avec probabilités proportionnelles à la taille données par (1.1.1) et l'estimateur par le ratio suivant :

$$\widehat{Y}_{ra} = X\frac{\widehat{Y}_\pi}{\widehat{X}_\pi}.$$

En utilisant une approximation de Taylor de premier ordre, la variance de $\widehat{Y}_{ra}$ dans le cas d'un plan de Poisson est donnée par

$$V_{Oh} = \sum_{k\in U^\star}(\pi_k^{-1} - 1)\left\{y_k - \frac{Y}{X}x_k\right\}^2. \tag{2.2.2}$$

Pour un plan à grande entropie, la variance (2.2.2) sera proche de la variance (2.2.1). On estimera (2.2.2) par

$$\widehat{V}_{Oh} = \sum_{k\in s^\star}\pi_k^{-1}(\pi_k^{-1} - 1)\left\{y_k - \frac{\widehat{Y}_\pi}{\widehat{X}_\pi}x_k\right\}^2, \tag{2.2.3}$$

en notant que $\widehat{X}_\pi = X$ dans le cas d'un plan proportionnel à la taille.

Il est intéressant de noter que l'approche d'Ohlsson correspond à une approximation des π_{kl}. En effet, on peut montrer (voir Annexe A) que la variance (2.2.2)

s'exprime comme (2.2.1) avec

$$\pi_{kl} = \pi_k \pi_l \left\{ 1 + \frac{1}{n^\star} \left(\pi_k + \pi_l - 2 + \frac{d}{n^\star} \right) \right\} \tag{2.2.4}$$

et $d = \sum_{k \in U^\star} \pi_k (1 - \pi_k)$.

Ce résultat est intéressant car il nous permet de comprendre que l'approche d'Ohlsson peut être vue comme une approche basée sur une approximation des π_{kl}. À notre connaissance, l'approximation des π_{kl} (2.2.4) n'existe pas dans la littérature. Notons qu'elle se rapproche de l'approximation (iv) de la famille des approximations de Brewer-Donadio (voir section 1.4.1.2), soit l'approximation

$$\pi_{kl} \approx \pi_k \pi_l \frac{n^\star - 1}{n^\star} \left\{ 1 + \frac{1}{n^\star} \left(\pi_k + \pi_l - 1 + \frac{d}{n^\star} \right) \right\}.$$

2.2.2. Forme générale de l'estimateur lié à l'approche d'Ohlsson

Comme pour les approximations de Brewer-Donadio et de Hájek, l'estimateur (2.2.3) peut être réécrit sous la forme générale (1.4.5). En effet, pour $a_k = 1$ et $c_k = 1 - \pi_k$, la forme générale devient

$$\begin{aligned}
\widehat{V}_{Gen}(\widehat{Y}_\pi) &= \sum_{k \in s^\star} c_k \left(\frac{y_k}{\pi_k} - \frac{\sum_{k \in s^\star} a_k (y_k / \pi_k)}{\sum_{k \in s^\star} a_k} \right)^2 \\
&= \sum_{k \in s^\star} \pi_k^{-1} (\pi_k^{-1} - 1) \left(y_k - \frac{\pi_k \widehat{Y}_\pi}{n} \right)^2 \\
&= \sum_{k \in s^\star} \pi_k^{-1} (\pi_k^{-1} - 1) \left(y_k - \frac{x_k \widehat{Y}_\pi}{X} \right)^2,
\end{aligned}$$

ce qui correspond à l'estimateur (2.2.3). L'estimateur (2.2.3) appartient donc à la classe des estimateurs de variance décrite à la section 1.4.1.

2.3. Estimation de la variance : cas de l'échantillonnage par grappes à un degré

On revient maintenant au cas de l'échantillonnage par grappes à un degré. Ainsi, n grappes sont sélectionnées d'une population U, composée des grappes $U_1^\star$, ..., $U_N^\star$ de tailles M_1,...,M_N. La sélection des grappes est effectuée de manière proportionnelle à la taille. On a donc $\pi_i = nX_i/X$, pour $i \in U$, où X_i est une mesure de taille pour la grappe i et $X = \sum_{i \in U} X_i$. Par exemple, si on fixe les grappes

comme étant les différents ménages de la population, on pourrait choisir la taille d'un ménage M_i comme étant sa mesure de taille. On aurait alors $X_i = M_i$. L'échantillon final $s^\star$ se compose de tous les éléments des grappes sélectionnées. On s'intéresse à l'estimation de la variance de $\widehat{Y}_\pi$.

2.3.1. Application de l'approche d'Ohlsson au niveau des grappes

On estime la variance à l'aide de l'approche d'Ohlsson. Puisque la sélection a été effectuée au niveau des grappes, on doit en tenir compte au moment d'estimer la variance. On approxime donc la variance de l'estimateur $\widehat{Y}_\pi$ (présentée à la section 1.2.2), soit

$$V_p(\widehat{Y}_\pi) = -\frac{1}{2}\sum_{i\in U}\sum_{j\in U}\Delta_{ij}\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 \tag{2.3.1}$$

par la variance correspondant à l'utilisation d'un plan de Poisson au niveau des grappes et de l'estimateur par le ratio, c'est-à-dire

$$V_{Oh} = \sum_{i\in U}(\pi_i^{-1} - 1)\left\{Y_i - \frac{X_i Y}{X}\right\}^2. \tag{2.3.2}$$

La variance (2.3.2) est ensuite estimée par

$$\widehat{V}_{Oh} = \sum_{i\in s}\pi_i^{-1}(\pi_i^{-1} - 1)\left\{Y_i - \frac{X_i \widehat{Y}_\pi}{X}\right\}^2. \tag{2.3.3}$$

Puisque la stratégie d'approximation tient compte de la présence de grappes, l'approche d'Ohlsson demeure appropriée dans ce contexte.

On note toutefois que l'estimateur (2.3.3) fait intervenir les probabilités d'inclusion d'ordre un ainsi que les totaux au niveau des grappes (π_i et Y_i). Tel que mentionné à la section 2.1, ces informations ne sont généralement connues qu'au niveau des éléments dans notre cadre pratique. En d'autres mots, on connaît π_k et y_k, pour $k \in s^\star$ (où $s^\star$ ést l'échantillon d'éléments), mais on ne connaît ni Y_i, ni π_i pour $i \in s$ (où s est l'échantillon de grappes). Dans ce contexte, puisque les variables indicatrices des grappes ne sont généralement pas fournies dans le fichier, il n'est pas possible de calculer l'estimateur (2.3.3).

2.3.2. Application de l'approche d'Ohlsson au niveau des éléments

Comme les valeurs de π_k, y_k et x_k sont connues pour tous les éléments de l'échantillon $s^\star$, on s'interroge maintenant sur le moyen de mettre à profit cette information pour estimer la variance dans un contexte d'échantillonnage par

grappes.

Considérons l'estimateur (2.2.3), soit l'estimateur correspondant à l'approche d'Ohlsson dans le cas d'échantillonnage d'éléments, pour approximer la variance. Si on utilise cet estimateur dans un contexte de grappes, on fait l'hypothèse qu'il est possible d'approximer la variance (2.3.1) par (2.2.2). Cette méthode ignore donc le fait que l'échantillonnage a eu lieu au niveau des grappes et procède comme s'il s'agissait d'un échantillonnage d'éléments. Quel serait l'impact, en termes de biais, de l'utilisation de la variance (2.2.2) pour approximer la variance (2.3.1) dans un contexte d'échantillonnage par grappes ?

Nous proposons d'évaluer cette approche au moyen d'un modèle linéaire mixte. Ainsi, on suppose que la variable d'intérêt peut être modélisée par

$$y_{ik} = \beta x_{ik} + v_i + \epsilon_{ik}, \tag{2.3.4}$$

où y_{ik} est la valeur de la variable d'intérêt pour l'unité k de la grappe i, v_i représente l'effet de la grappe i et ϵ_{ik} désigne le bruit pour l'individu ik. On fera les hypothèses usuelles suivantes : $E_m(v) = E_m(\epsilon) = 0$. Les variances de v et ϵ par rapport au modèle sont respectivement notées $V_m(v) \equiv \sigma_v^2$ et $V_m(\epsilon) \equiv \sigma_\epsilon^2$. La corrélation intra-grappe, notée ρ, est une mesure du lien unissant les éléments à l'intérieur des grappes. Elle est définie par

$$\rho = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\epsilon^2}.$$

La notation $E_m(\cdot)$ désigne l'espérance par rapport au modèle (2.3.4). Dans ce cas, y tire son caractère aléatoire de l'effet de grappe v et du bruit ϵ. On définit le biais relatif de (2.2.2) par

$$BR(V_{Oh}) = \frac{E_m\{V_{Oh} - V_p(\widehat{Y}_\pi)\}}{E_m\{V_p(\widehat{Y}_\pi)\}}, \tag{2.3.5}$$

où V_{Oh} est donné par (2.2.2) et $V_p(\widehat{Y}_\pi)$ correspond à l'expression (2.3.1). On montre (voir Annexe B) que le biais relatif (2.3.5) est donné par

$$BR(V_{Oh}) = A - \frac{\sum\limits_{i \in U} (\pi_i^{-1} - 1) M_i (M_i - 1)}{\sum\limits_{i \in U} (\pi_i^{-1} - 1) M_i^2 + \left(\frac{1}{\rho} - 1\right) \sum\limits_{i \in U} (\pi_i^{-1} - 1) M_i}, \tag{2.3.6}$$

où

$$A = K^{-1}\left\{\frac{\sigma_v^2 \sum\limits_{i\in U} M_i^2}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\left(\sum_{k\in U_i^\star} x_{ik}^2 - 2X\frac{M_i X_i}{\sum\limits_{i\in U} M_i^2}\right)\right.$$
$$\left. + \frac{\sigma_\epsilon^2 \sum\limits_{i\in U} M_i}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\left(\sum_{k\in U_i^\star} x_{ik}^2 - 2X\frac{X_i}{\sum\limits_{i\in U} M_i}\right)\right\} \qquad (2.3.7)$$

et

$$K = \sum_{i\in U}(\pi_i^{-1}-1)M_i(M_i\sigma_v^2+\sigma_\epsilon^2).$$

Par souci de simplicité, nous commençons par étudier le cas des M_i égaux.

2.3.2.1. *Cas des M_i égaux*

Lorsque $M_i = M$, pour tout $i \in U$, le biais relatif (2.3.6) s'écrit comme

$$BR(V_{Oh}) = \tilde{A} - \frac{M-1}{M+\left(\frac{1}{\rho}-1\right)}, \qquad (2.3.8)$$

où

$$\tilde{A} = \frac{\frac{N}{X^2}\sum\limits_{i\in U}(\pi_i^{-1}-1)\left(\sum\limits_{k\in U_i^\star} x_{ik}^2 - 2\frac{X}{MN}X_i\right)}{\sum\limits_{i\in U}(\pi_i^{-1}-1)}.$$

Une étude de l'ordre de grandeur du terme $\tilde{A}$ (voir Annexe C) permet de montrer que $\tilde{A} = O(1/MN)$. Il découle de (2.3.8) que

$$BR(V_{Oh}) = O\left(\frac{1}{MN}\right) - \frac{M-1}{M+\left(\frac{1}{\rho}-1\right)}. \qquad (2.3.9)$$

Le biais relatif s'exprime donc en fonction de la corrélation intra-grappe ρ, de la taille des grappes M et de la taille de la population MN. Considérons d'abord le cas $\rho = 0$, soit l'absence de corrélation intra-grappe. Le biais relatif (2.3.9) devient alors

$$BR(V_{Oh}) = O\left(\frac{1}{MN}\right).$$

Ainsi, pour une population de taille suffisamment grande, le biais relatif est petit et la variance (2.2.2) approxime bien (2.3.1). En présence d'une corrélation intra-grappe parfaite, $\rho = 1$, le biais relatif (2.3.9) s'écrit comme

$$BR(V_{Oh}) = O\left(\frac{1}{MN}\right) - \left(1-\frac{1}{M}\right), \quad \text{pour } M \geq 2.$$

Dans ce cas, le biais relatif dépend directement de la taille des grappes M. Ainsi, lorsque la taille des grappes M augmente, le biais relatif prend rapidement de l'ampleur, et ce même si la taille globale de la population est grande. Dans ce cas, la variance (2.2.2) n'est pas une bonne approximation de (2.3.1).

En résumé, pour $M_i = M$, l'approximation de la variance (2.3.1) par (2.2.2) est appropriée dans le cas où la corrélation intra-grappe est près de zéro et lorsque la taille des grappes est petite. Ce résultat n'est pas surprenant puisque l'approche d'Ohlsson au niveau des éléments consiste à supposer que le plan de sondage utilisé est un plan de Poisson au niveau des éléments. Il s'agit donc d'un plan qui ne tient pas compte de la corrélation intra-grappe.

2.3.2.2. *Cas des M_i inégaux*

Dans le cas où les grappes sont de tailles inégales, une étude de l'ordre de grandeur du terme A de l'expression (2.3.6) du biais relatif (voir Annexe D) indique que $A = O(1/M_0)$, où $M_0 = \sum_{i \in U} M_i$. Ainsi, le biais relatif (2.3.6) se réécrit comme

$$BR(V_{Oh}) = O\left(\frac{1}{M_0}\right) - \frac{\sum_{i \in U} (\pi_i^{-1} - 1) M_i (M_i - 1)}{\sum_{i \in U} (\pi_i^{-1} - 1) M_i^2 + \left(\frac{1}{\rho} - 1\right) \sum_{i \in U} (\pi_i^{-1} - 1) M_i}. \quad (2.3.10)$$

En absence de corrélation intra-grappe, on a que $\rho = 0$ et (2.3.10) devient

$$BR(V_{Oh}) = O\left(\frac{1}{M_0}\right).$$

Le biais relatif sera donc petit pour une population de taille suffisamment grande. Dans le cas d'une corrélation intra-grappe parfaite ($\rho = 1$), l'expression (2.3.10) s'écrit comme

$$\begin{aligned} BR(V_{Oh}) &= O\left(\frac{1}{M_0}\right) - \frac{\sum_{i \in U} (\pi_i^{-1} - 1) M_i (M_i - 1)}{\sum_{i \in U} (\pi_i^{-1} - 1) M_i^2} \\ &= O\left(\frac{1}{M_0}\right) - \left\{1 - \frac{\sum_{i \in U} (\pi_i^{-1} - 1) M_i}{\sum_{i \in U} (\pi_i^{-1} - 1) M_i^2}\right\}, \quad \text{pour } M_i \geq 2,\ \forall i \in U. \end{aligned}$$

L'ampleur du biais relatif dépend donc des valeurs de M_i, pour $i \in U$. De façon similaire au cas des M_i égaux, le biais relatif prendra de l'importance lorsque les tailles de grappes augmentent. Toutefois, dans ce cas, il suffit que certaines

grappes soient de tailles considérablement supérieures aux autres pour que le biais soit important.

Le cas des M_i inégaux présente donc des résultats analogues au cas $M_i = M$, c'est-à-dire que pour qu'il soit approprié d'approximer (2.3.1) par (2.2.2), les tailles de grappes doivent être petites et la corrélation intra-grappe près de zéro.

2.3.2.3. *Variance ajustée pour le biais*

Dans le cas où la corrélation intra-grappe est élevée et/ou la taille des grappes est grande, il est tout de même possible d'obtenir une approximation adéquate de la variance. L'idée est de corriger la variance (2.2.2) pour le biais engendré par de trop grandes valeurs de ρ et/ou M_i.

On présente d'abord une forme alternative (voir Annexe B) du biais relatif (2.3.6). Celui-ci peut s'écrire comme

$$BR(V_{Oh}) = A - \frac{\sigma_v^2 \sum\limits_{i\in U} (\pi_i^{-1} - 1)M_i(M_i - 1)}{\sum\limits_{i\in U} (\pi_i^{-1} - 1)M_i(M_i\sigma_v^2 + \sigma_\epsilon^2)}, \tag{2.3.11}$$

où A est donné par (2.3.7).

En supposant que la taille de la population est suffisamment grande, l'expression (2.3.10) nous permet d'approximer (2.3.11) par

$$BR(V_{Oh}) \approx -\frac{\sigma_v^2 \sum\limits_{i\in U} (\pi_i^{-1} - 1)M_i(M_i - 1)}{\sum\limits_{i\in U} (\pi_i^{-1} - 1)M_i(M_i\sigma_v^2 + \sigma_\epsilon^2)}. \tag{2.3.12}$$

On s'intéresse à ajuster la variance (2.2.2) pour le biais. On cherche donc à déterminer une expression pour le biais (par rapport au modèle) engendré par V_{Oh}, notée $B_m(V_{Oh})$. Ce dernier correspond au numérateur dans la définition de $BR(V_{Oh})$, c'est-à-dire que

$$\begin{aligned} BR(V_{Oh}) &= \frac{E_m\{V_{Oh} - V_p(\widehat{Y}_\pi)\}}{E_m\{V_p(\widehat{Y}_\pi)\}} \\ &= \frac{B_m(V_{Oh})}{E_m\{V_p(\widehat{Y}_\pi)\}}. \end{aligned}$$

Une expression pour $B_m(V_{Oh})$ est donc

$$B_m(V_{Oh}) = BR(V_{Oh})E_m\{V_p(\widehat{Y}_\pi)\}.$$

À partir de l'expression (2.3.12) et en notant que

$$E_m\{V_p(\widehat{Y}_\pi)\} = \sum_{i\in U}(\pi_i^{-1}-1)M_i(M_i\sigma_v^2+\sigma_\epsilon^2),$$

(voir Annexe B), on a donc

$$\begin{aligned} B_m(V_{Oh}) &\approx -\frac{\sigma_v^2\sum\limits_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1)}{\sum\limits_{i\in U}(\pi_i^{-1}-1)M_i(M_i\sigma_v^2+\sigma_\epsilon^2)}E_m\{V_p(\widehat{Y}_\pi)\} \\ &= -\sigma_v^2\sum_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1). \end{aligned}$$

Puisque l'on souhaite corriger V_{Oh} pour le biais, la variance ajustée prend la forme de la variance (2.2.2) à laquelle on soustrait $B_m(V_{Oh})$, soit

$$V_{Oh} - B_m(V_{Oh}) = V_{Oh} + \sigma_v^2\sum_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1).$$

En pratique, la valeur de σ_v^2 n'est toutefois pas connue. Il est possible de l'estimer en considérant un modèle linéaire mixte duquel on dégage des estimateurs par maximum de vraisemblance (voir Annexe E). Cette estimation sera notée $\hat{\sigma}_v^2$. La variance ajustée pour le biais est donnée par

$$V_{Aj} = V_{Oh} + \hat{\sigma}_v^2\sum_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1),$$

où V_{Oh} correspond à la variance (2.2.2). On l'estime par

$$\widehat{V}_{Aj} = \widehat{V}_{Oh} + \hat{\sigma}_v^2\sum_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1), \tag{2.3.13}$$

où $\widehat{V}_{Oh}$ désigne l'estimateur (2.2.3).

Chapitre 3

ÉTUDES PAR SIMULATION

3.1. Étude 1 : Cas d'une population simple

Dans le cadre de cette première étude par simulation, on cherche à vérifier la validité de l'approche d'Ohlsson dans le cas simple où les éléments sont directement tirés de la population. L'estimateur de variance (1.4.6) est donc comparé aux estimateurs classiques de la variance (1.2.4) et (1.2.5). Rappelons que l'estimateur (1.4.6) est avantageux puisque qu'il ne requiert pas les probabilités d'inclusion d'ordre deux et se calcule par une simple somme. L'estimateur correspondant à l'approximation (iv) de Brewer-Donadio est également étudié. En effet, on s'intéresse à la ressemblance qui avait été notée à la section 2.2.1 entre ce dernier et l'estimateur (1.4.6).

3.1.1. Populations et échantillons simulés

Afin de comparer nos différents estimateurs de la variance, nous avons généré quatre populations composées de $N^\star$ éléments. D'abord, une variable auxiliaire x a été générée selon une loi gamma dont le paramètre de forme est $\alpha = 5$ et le paramètre d'échelle est $\theta = 10$. Ensuite, la variable d'intérêt y a été générée selon le modèle

$$y_k = \beta x_k + \epsilon_k,$$

où les erreurs ϵ_k ont été générées à partir d'une loi normale de moyenne égale à 0 et de variance égale à σ_ϵ^2. Nous avons considéré deux valeurs différentes pour le paramètre β, soit $\beta = 1$ et $\beta = 5$. Pour chacun de ces paramètres, nous avons généré deux populations de taille $N^\star = 500, 2000$. La valeur de σ_ϵ^2 a été fixée de manière à ce que le coefficient de corrélation entre les variables x et y soit approximativement égal à 0,3.

Dans chaque population, on a ensuite tiré $R = 100\,000$ échantillons de taille $n^\star$ selon un plan de Poisson conditionnel avec probabilités proportionnelles à la taille. La variable auxiliaire x a été utilisée comme variable de taille. Il s'ensuit que l'élément k est sélectionné avec probabilité $\pi_k = nx_k/X$. Pour les populations de taille $N^\star = 500$ la taille de l'échantillon $n^\star$ a été fixée à 50 et 100, alors que pour celles de taille $N^\star = 2000$ on a utilisé $n^\star = 100, 200$. Les scénarios utilisés sont similaires aux scénarios considérés dans Haziza et al. (2008).

3.1.2. Critères de comparaison des estimateurs de variance

Dans chaque échantillon, on a calculé l'estimateur $\widehat{Y}_\pi$ donné par (1.2.1). La variance de $\widehat{Y}_\pi$ a été estimée selon plusieurs méthodes :

(i) l'estimateur Horvitz-Thompson de la variance (1.2.4) ;

(ii) l'estimateur Sen-Yates-Grundy (1.2.5) ;

(iii) l'estimateur de variance correspondant à l'approche d'Ohlsson (1.4.6) ;

(iv) l'estimateur correspondant à l'approximation (iv) de Brewer-Donadio (voir section 1.4.1.2).

Comme mesure du biais d'un estimateur de variance $\widehat{V}$, nous avons calculé le biais relatif Monte-Carlo (en %) donné par

$$BR_{MC}(\widehat{V}) = \frac{E_{MC}(\widehat{V}) - V}{V} \times 100, \tag{3.1.1}$$

avec V désignant la variance (1.2.2) et

$$E_{MC}(\widehat{V}) = R^{-1} \sum_{i=1}^{R} \widehat{V}^{(i)},$$

où $\widehat{V}^{(i)}$ désigne l'estimateur $\widehat{V}$, calculé à partir du i^e échantillon.

Comme mesure de stabilité, nous avons calculé le ratio des variances Monte-Carlo, qui est évalué à partir des R échantillons tirés. Pour un estimateur $\widehat{V}$, ce ratio (en %) est défini selon

$$RV_{MC}(\widehat{V}) = \frac{V_{MC}(\widehat{V})}{V_{MC}\{\widehat{V}_{SYG}(\widehat{Y}_\pi)\}} \times 100, \tag{3.1.2}$$

où

$$V_{MC}(\widehat{V}) = E_{MC}\left[\{\widehat{V} - E_{MC}(\widehat{V})\}^2\right]$$

et $\widehat{V}_{SYG}(\widehat{Y}_\pi)$ correspond à l'estimateur (1.2.5), qui est ici utilisé comme estimateur de référence.

3.1.3. Résultats

Les tableaux 3.1 et 3.2 présentent les biais relatifs Monte-Carlo (en %) ainsi que les ratios des variances Monte-Carlo (en %) correspondant aux différents estimateurs de variance étudiés. Le tableau 3.1 comprend les résultats relatifs aux populations pour lesquelles $\beta = 1$, alors que le tableau 3.2 présente les résultats des populations pour lesquelles $\beta = 5$.

TABLEAU 3.1. Résultats relatifs aux estimateurs étudiés dans le cadre du modèle $\beta = 1$

$N^\star$	$n^\star$	$\widehat{V}_{HT}$	$\widehat{V}_{SYG}$	$\widehat{V}_{Oh}$	$\widehat{V}_{BD(iv)}$
500	50	-0,04	-0,04	-2,12	-0,04
		98,75	100,00	94,84	99,08
	100	-0,16	-0,17	-1,24	-0,16
		97,53	100,00	95,84	98,12
2000	100	-0,07	-0,07	-1,09	-0,07
		99,32	100,00	97,46	99,52
	200	0,04	0,03	-0,49	0,03
		98,70	100,00	97,99	99,06

TABLEAU 3.2. Résultats relatifs aux estimateurs étudiés dans le cadre du modèle $\beta = 5$

$N^\star$	$n^\star$	$\widehat{V}_{HT}$	$\widehat{V}_{SYG}$	$\widehat{V}_{Oh}$	$\widehat{V}_{BD(iv)}$
500	50	-0,04	-0,04	-2,12	-0,05
		95,33	100,00	95,11	99,40
	100	-0,11	-0,12	-1,20	-0,12
		94,68	100,00	96,60	98,94
2000	100	0,07	0,06	-0,97	0,06
		97,46	100,00	97,64	99,72
	200	0,08	0,09	-0,55	0,09
		95,69	100,00	98,31	99,40

3.1.4. Discussion

À la lumière des résultats obtenus, les estimateurs classiques de la variance $\widehat{V}_{HT}$ et $\widehat{V}_{SYG}$, qui sont des estimateurs sans biais, présentent, tel qu'attendu, des biais relatifs très faibles. Pour ce qui est de l'estimateur $\widehat{V}_{Oh}$, on note un léger biais relatif pour les populations de 500 éléments, qui diminue au fur et à mesure

que $n^\star$ et $N^\star$ augmentent. Par exemple, pour $n^\star = 50$, le biais relatif est égal à -2,12%, alors que pour $n^\star = 100$ il est de -1,24% si $\beta = 1$ et de -1,20% si $\beta = 5$. En ce qui concerne l'estimateur correspondant à l'approximation (iv) de Brewer-Donadio, on obtient des résultats proches de ceux obtenus pour $\widehat{V}_{HT}$ et $\widehat{V}_{SYG}$. Les résultats de simulation ne reflètent donc pas la ressemblance théorique qui avait été établie entre $\widehat{V}_{Oh}$ et$\widehat{V}_{BD(iv)}$ à la section 2.2.1. Le biais relatif ne semble pas être affecté lorsque le paramètre β passe de un à cinq. De manière plus générale, on note que l'estimateur $\widehat{V}_{Oh}$ sous estime légèrement la variance, dans tous les cas.

Au niveau de la stabilité des estimateurs, on remarque que dans le cas $\beta = 1$, l'estimateur $\widehat{V}_{Oh}$ est légèrement plus stable que les estimateurs classiques de la variance avec des valeurs de RV se situant entre 94,84% et 97,99%.

3.2. Étude 2 : Cas d'une population de grappes

Dans le cadre de cette seconde étude par simulation, on s'intéresse au cas où la population est divisée en grappes. L'étude vise à comparer l'application de l'approche d'Ohlsson au niveau des éléments (voir section 2.3.2) ainsi que l'estimateur ajusté pour le biais (2.3.13) aux estimateurs classiques (1.2.8) et (1.2.9). L'estimateur correspondant à l'application de l'approche d'Ohlsson au niveau des grappes est aussi étudié (voir section 2.3.1).

3.2.1. Populations et échantillons simulés

Dans le but de comparer les différents estimateurs de variance, 36 populations constituées de N grappes, chacune composée de M éléments, ont été générées. D'abord, une variable x a été générée pour chaque élément de la population selon une loi gamma de paramètres $\alpha = 5$ et $\theta = 10$. Puis, un effet de grappe, noté v, a été généré pour chaque grappe de la population selon une loi normale de moyenne nulle et de variance σ_v^2. La variable d'intérêt y a été obtenue au moyen du modèle

$$y_{ik} = \beta x_{ik} + v_i + \epsilon_{ik},$$

où les termes d'erreur ϵ_{ik} ont été générés à partir d'une loi normale de moyenne 0 et de variance égale à σ_ϵ^2. On s'intéresse à ce modèle pour les valeurs $\beta = 1$ et 5 ainsi que pour différentes valeurs de σ_v^2 qui correspondent aux corrélations intra-grappe $\rho = \{0,025\}, \{0,05\}, \{0,10\}$. On forme ainsi 6 modèles caractérisés par les différentes combinaisons de β et ρ. Pour chacun de ces modèles, 6 populations ont été générées à partir des différentes combinaisons formées par les tailles $N = 2000, 5000$ et $M = 2, 5, 10$. Le paramètre σ_ϵ^2 a été fixé de manière à ce que

la corrélation entre x et y soit approximativement égale à 0,3.

À l'intérieur de chacune des populations décrites précédemment, on a tiré $R = 100\,000$ échantillons de n grappes et tous les éléments de ces grappes ont été inclus dans l'échantillon final. Les grappes ont été sélectionnées selon un plan de Poisson conditionnel avec probabilités proportionnelles à la taille. La variable x a été utilisée comme mesure de taille. Une grappe i a donc une probabilité d'inclusion égale à $\pi_i = nX_i/X$, où X_i représente la somme de tous les x_{ik} des éléments de la grappe i. Pour les populations de 2000 grappes, on a tiré des échantillons de $n = 60$ et $n = 100$ grappes, tandis que pour celles de 5000 grappes, les échantillons tirés sont de taille $n = 50, 150$ grappes. Le tableau 3.3 présente un sommaire des populations et des échantillons de cette étude.

TABLEAU 3.3. Sommaire des populations et des échantillons étudiés

β	ρ	M	N	n
1	0,025, 0,05, 0,10	2, 5, 10	2000	60, 100
			5000	50, 150
5	0,025, 0,05, 0,10	2, 5, 10	2000	60, 100
			5000	50, 150

Les scénarios utilisés sont similaires aux scénarios considérés dans Haziza et al. (2008).

3.2.2. Critères de comparaison des estimateurs de variance

Pour chaque échantillon sélectionné, on a d'abord calculé l'estimateur $\widehat{Y}_\pi$, donné par (1.2.6). Sa variance est ensuite estimée à partir de différents estimateurs :

(i) l'estimateur de variance d'Horvitz-Thompson (1.2.8) ;

(ii) l'estimateur Sen-Yates-Grundy (1.2.9) ;

(iii) l'estimateur (2.3.3) qui correspond à l'application de la méthode d'Ohlsson au niveau des grappes (noté $\widehat{V}_{OhG}$) ;

(iv) l'estimateur (1.4.6) qui est associé à l'application de l'approche d'Ohlsson au niveau des éléments (noté $\widehat{V}_{OhE}$) ;

(v) l'estimateur ajusté pour le biais (2.3.13) qui est calculé en procédant tel que décrit à la section 2.3.2.3 (noté $\widehat{V}_{Aj}$).

On a ensuite calculé, pour chacun de ces estimateurs, le biais relatif Monte-Carlo (en %) ainsi que le ratio des variances Monte-Carlo (en %) qui sont respectivement donnés par les expressions (3.1.1) et (3.1.2).

3.2.3. Résultats

Les tableaux 3.4 à 3.7 présentent les biais relatif Monte-Carlo (en %) ainsi que les ratios des variances Monte-Carlo (en %) des estimateurs étudiés. Le tableau 3.4 comprend les résultats correspondant aux modèles pour lesquels $\beta = 1$ pour les populations de 2000 grappes. Ces mêmes modèles sont étudiés au tableau 3.5, pour les populations de 5000 grappes. De façon analogue, les tableaux 3.6 et 3.7 présentent les résultats des modèles pour lesquels $\beta = 5$ pour les populations respectives de 2000 et 5000 grappes.

TABLEAU 3.4. Résultats relatifs aux modèles pour lesquels $\beta = 1$ pour une population de 2000 grappes

		n									
		60					100				
M	ρ	$\widehat{V}_{HT}$	$\widehat{V}_{SYG}$	$\widehat{V}_{OhG}$	$\widehat{V}_{OhE}$	$\widehat{V}_{Aj}$	$\widehat{V}_{HT}$	$\widehat{V}_{SYG}$	$\widehat{V}_{OhG}$	$\widehat{V}_{OhE}$	$\widehat{V}_{Aj}$
2	0,025	-0,04	-0,04	-1,72	-5,39	0,29	0,11	0,11	-0,90	-5,14	-0,34
		99,39	100,00	96,33	50,29	63,56	98,96	100,00	97,41	50,76	64,06
	0,050	0,06	0,06	-1,62	-6,96	1,28	0,02	0,02	-0,99	-6,70	0,76
		99,41	100,00	96,35	44,24	64,81	99,01	100,00	97,44	45,01	66,88
	0,100	-0,09	-0,09	-1,76	-10,90	0,44	0,06	0,05	-0,96	-10,55	0,44
		99,36	100,00	96,32	46,75	72,68	98,94	100,00	97,41	47,10	75,13
5	0,025	0,02	0,02	-1,65	-12,33	0,05	-0,04	-0,04	-1,04	-12,20	-0,60
		99,44	100,00	96,50	17,09	60,89	99,07	100,00	97,69	17,10	65,24
	0,050	0,03	0,03	-1,64	-14,99	0,00	-0,02	-0,02	-1,03	-14,87	-0,33
		99,42	100,00	96,50	18,14	66,93	98,98	100,00	97,67	18,42	72,08
	0,100	0,01	0,01	-1,66	-28,67	-1,42	0,02	0,02	-0,99	-28,55	-1,03
		99,45	100,00	96,49	13,97	80,60	99,07	100,00	97,66	13,95	81,82
10	0,025	0,04	0,04	-1,63	-17,35	0,06	-0,13	-0,13	-1,13	-17,30	-0,38
		99,46	100,00	96,58	8,92	73,26	99,09	100,00	97,82	9,07	80,07
	0,050	-0,01	-0,01	-1,68	-32,46	-1,08	0,07	0,07	-0,93	-32,43	-0,60
		99,56	100,00	96,58	5,87	86,11	99,25	100,00	97,82	5,90	87,86
	0,100	0,08	0,08	-1,59	-50,84	-1,96	0,00	0,00	-1,00	-50,80	-1,56
		99,66	100,00	96,58	3,24	89,77	99,41	100,00	97,81	3,32	90,18

TABLEAU 3.5. Résultats relatifs aux modèles pour lesquels $\beta = 1$ pour une population de 5000 grappes

		n									
		50					150				
M	ρ	$\widehat{V}_{HT}$	$\widehat{V}_{SYG}$	$\widehat{V}_{OhG}$	$\widehat{V}_{OhE}$	$\widehat{V}_{Aj}$	$\widehat{V}_{HT}$	$\widehat{V}_{SYG}$	$\widehat{V}_{OhG}$	$\widehat{V}_{OhE}$	$\widehat{V}_{Aj}$
2	0,025	0,11	0,11	-1,90	-4,87	3,09	-0,03	-0,03	-0,70	-4,34	1,85
		99,82	100,00	95,93	55,62	72,89	99,43	100,00	98,35	56,35	76,08
	0,050	0,03	0,03	-1,98	-7,44	-0,05	-0,04	-0,04	-0,71	-6,87	-1,22
		99,82	100,00	95,93	46,13	61,84	99,44	100,00	98,35	46,47	64,36
	0,100	0,06	0,06	-1,95	-9,98	0,58	-0,09	-0,09	-0,76	-9,44	0,08
		99,81	100,00	95,92	48,74	69,64	99,41	100,00	98,33	48,98	73,60
5	0,025	0,03	0,03	-1,97	-9,19	2,01	0,08	0,08	-0,59	-8,96	0,48
		99,82	100,00	95,98	20,46	59,84	99,43	100,00	98,48	21,16	69,04
	0,050	-0,08	-0,08	-2,08	-17,53	-0,43	-0,02	-0,02	-0,69	-17,27	-0,51
		99,81	100,00	95,97	18,34	71,54	99,41	100,00	98,47	18,48	81,77
	0,100	-0,05	-0,05	-2,05	-27,16	-1,56	-0,07	-0,07	-0,74	-26,94	-0,81
		99,83	100,00	95,97	14,12	74,55	99,47	100,00	98,47	14,40	78,43
10	0,025	-0,07	-0,07	-2,06	-19,06	0,04	0,00	0,00	-0,67	-18,93	-0,06
		99,84	100,00	96,00	9,12	75,52	99,49	100,00	98,56	9,35	89,13
	0,050	0,05	0,05	-1,96	-29,65	-1,46	-0,04	-0,04	-0,70	-29,55	-0,77
		99,85	100,00	96,00	6,60	85,40	99,52	100,00	98,56	6,66	90,86
	0,100	0,02	0,02	-1,98	-47,63	-1,20	0,01	0,01	-0,65	-47,52	-0,24
		99,88	100,00	96,01	3,84	90,24	99,62	100,00	98,56	3,92	91,29

TABLEAU 3.6. Résultats relatifs aux modèles pour lesquels $\beta = 5$ pour une population de 2000 grappes

		n									
		60					100				
M	ρ	$\widehat{V}_{HT}$	$\widehat{V}_{SYG}$	$\widehat{V}_{OhG}$	$\widehat{V}_{OhE}$	$\widehat{V}_{Aj}$	$\widehat{V}_{HT}$	$\widehat{V}_{SYG}$	$\widehat{V}_{OhG}$	$\widehat{V}_{OhE}$	$\widehat{V}_{Aj}$
2	0,025	0,06	0,07	-1,61	-2,70	2,35	0,02	-0,01	-1,02	-2,47	1,47
		94,21	100,00	96,35	50,33	61,97	91,62	100,00	97,46	50,50	61,60
	0,050	-0,07	-0,08	-1,75	-5,13	0,99	-0,05	-0,04	-1,05	-4,86	0,36
		94,00	100,00	96,34	53,49	66,99	90,98	100,00	97,43	53,84	67,71
	0,100	0,00	-0,01	-1,68	-9,26	-0,49	0,02	0,03	-0,98	-8,96	-0,78
		93,21	100,00	96,29	48,40	69,12	89,94	100,00	97,34	48,86	71,09
5	0,025	-0,05	-0,03	-1,70	-13,68	0,49	-0,05	-0,04	-1,04	-13,60	-0,02
		93,70	100,00	96,49	18,60	68,16	94,45	100,00	97,66	18,73	72,92
	0,050	0,05	0,05	-1,62	-17,96	-0,23	-0,02	-0,03	-1,04	-17,85	-0,31
		93,58	100,00	96,48	17,29	70,78	93,44	100,00	97,64	17,60	75,91
	0,100	0,03	0,03	-1,65	-29,64	-0,71	-0,02	-0,04	-1,04	-29,53	-0,31
		94,56	100,00	96,50	13,77	83,89	93,77	100,00	97,68	14,05	85,52
10	0,025	-0,06	-0,06	-1,73	-16,92	-0,01	0,02	-0,01	-1,01	-16,89	-0,24
		97,14	100,00	96,57	9,82	73,68	106,33	100,00	97,80	9,76	80,74
	0,050	0,02	0,01	-1,66	-28,72	-1,38	-0,03	-0,02	-1,02	-28,64	-1,10
		96,87	100,00	96,59	6,32	84,37	100,69	100,00	97,83	6,47	87,72
	0,100	-0,03	-0,04	-1,71	-47,05	-0,86	0,04	0,06	-0,95	-46,98	-0,29
		96,43	100,00	96,58	4,34	88,60	98,30	100,00	97,80	4,37	88,69

TABLEAU 3.7. Résultats relatifs aux modèles pour lesquels $\beta = 5$ pour une population de 5000 grappes

		n									
		50					150				
M	ρ	$\widehat{V}_{HT}$	$\widehat{V}_{SYG}$	$\widehat{V}_{OhG}$	$\widehat{V}_{OhE}$	$\widehat{V}_{Aj}$	$\widehat{V}_{HT}$	$\widehat{V}_{SYG}$	$\widehat{V}_{OhG}$	$\widehat{V}_{OhE}$	$\widehat{V}_{Aj}$
2	0,025	-0,02	-0,02	-2,03	-2,69	2,72	-0,02	-0,02	-0,69	-2,11	1,29
		97,89	100,00	95,93	51,37	62,78	94,42	100,00	98,36	51,47	62,94
	0,050	-0,11	-0,12	-2,12	-7,24	1,49	-0,02	-0,02	-0,69	-6,61	0,78
		97,93	100,00	95,93	49,99	68,38	94,40	100,00	98,35	50,65	72,54
	0,100	-0,03	-0,03	-2,03	-11,33	-0,56	0,03	0,02	-0,65	-10,67	-0,83
		97,81	100,00	95,92	43,96	68,10	94,10	100,00	98,31	44,57	73,52
5	0,025	0,05	0,04	-1,96	-8,92	2,08	0,04	0,05	-0,62	-8,67	0,40
		96,87	100,00	95,97	21,91	60,66	93,55	100,00	98,47	22,57	69,88
	0,050	0,02	0,02	-1,98	-15,25	1,06	-0,06	-0,06	-0,73	-15,05	0,59
		97,31	100,00	95,98	18,85	71,67	94,56	100,00	98,48	19,35	83,04
	0,100	0,07	0,07	-1,93	-29,56	-2,08	0,07	0,08	-0,59	-29,36	-1,25
		97,76	100,00	95,98	12,24	75,28	94,67	100,00	98,48	12,50	78,17
10	0,025	0,03	0,02	-1,98	-22,25	-1,79	0,05	0,03	-0,64	-22,16	-1,74
		96,95	100,00	96,00	8,01	72,15	95,13	100,00	98,54	8,03	82,94
	0,050	0,07	0,06	-1,94	-30,82	-1,71	-0,02	-0,02	-0,69	-30,71	-1,01
		97,61	100,00	96,00	6,08	81,94	96,20	100,00	98,56	6,22	86,55
	0,100	0,10	0,10	-1,90	-48,14	-1,16	-0,02	-0,02	-0,68	-48,04	-0,32
		98,07	100,00	96,00	3,93	88,01	96,31	100,00	98,56	4,03	89,19

3.2.4. Discussion

En comparant les biais relatifs des estimateurs étudiés, on note, sans surprise, que les estimateurs classiques de variance $\widehat{V}_{HT}$ et $\widehat{V}_{SYG}$ présentent une fois de plus des biais relatifs négligeables. L'estimateur lié à l'approche d'Ohlsson appliquée au niveau des grappes ($\widehat{V}_{OhG}$) se comporte bien, avec un biais relatif qui ne dépasse jamais 2,2%, en valeur absolue. On note aussi que le biais relatif de l'estimateur $\widehat{V}_{OhG}$ diminue lorsque la taille de l'échantillon augmente. Le biais relatif de $\widehat{V}_{OhG}$ ne semble pas affecté par une variation de ρ ou M. En ce qui concerne l'approche d'Ohlsson appliquée au niveau des éléments, l'estimateur $\widehat{V}_{OhE}$ présente des valeurs de biais relatif qui prennent rapidement de l'ampleur lorsque la corrélation intra-grappe ou la taille des grappes augmentent, ce qui est cohérent avec l'expression (2.3.9). Pour une valeur de ρ et M fixée, on note toutefois que le biais relatif de $\widehat{V}_{OhE}$ diminue légèrement lorsque n augmente. Pour ce qui est de l'estimateur ajusté $\widehat{V}_{Aj}$, son biais relatif ne dépasse pas 3,1%, en valeur absolue. Il ne semble pas affecté par une variation de la taille de la population, de l'échantillon, des grappes, ni par la corrélation intra-grappe.

On note finalement que les estimateurs $\widehat{V}_{OhG}$ et $\widehat{V}_{OhE}$ sous estiment la variance de $\widehat{Y}_{\pi}$ dans tous les cas. On n'observe aucun effet dû à la variation du paramètre β, au niveau du biais relatif.

En étudiant le ratio des variances, on remarque que l'estimateur $\widehat{V}_{OhG}$ se comporte de façon légèrement plus stable que les estimateurs classiques de variance, lorsque $\beta = 1$. Dans le cas où $\beta = 5$, cette tendance est maintenue pour une population de 5000 grappes de taille quelconque, ainsi que pour une population de 2000 grappes de taille $M = 10$. Toutefois, lorsqu'on a 2000 grappes de taille $M = 2$ ou $M = 5$, l'estimateur $\widehat{V}_{HT}$ présente une meilleur stabilité que $\widehat{V}_{OhG}$. En ce qui concerne l'estimateur $\widehat{V}_{OhE}$, il présente une stabilité largement supérieure aux estimateurs classiques de variance, qui se reflète par un ratio des variances qui ne dépasse pas 56,35%. On note aussi que $\widehat{V}_{OhE}$ gagne en stabilité lorsque ρ ou M augmente et perd légèrement en stabilité lorsque n augmente. Finalement, l'estimateur $\widehat{V}_{Aj}$ est plus stable que $\widehat{V}_{HT}$, $\widehat{V}_{SYG}$ et $\widehat{V}_{OhG}$, mais moins stable que $\widehat{V}_{OhE}$. Sa stabilité semble diminuer lorsque ρ ou n augmente, ainsi que lorsque M augmente pour des valeurs de $\rho > 0,025$. Le paramètre β n'influence pas la stabilité de $\widehat{V}_{OhE}$ ou $\widehat{V}_{Aj}$.

Chapitre 4

CONCLUSION

Dans ce mémoire, nous avons proposé une approche alternative à l'estimation classique de la variance pour les plans de sondage à grande entropie. La méthode présentée a pour avantages de ne faire intervenir que les probabilités d'inclusion d'ordre un et de s'écrire comme une simple somme. Ainsi, les probabilités d'inclusion d'ordre deux, qui peuvent être complexes, voire impossibles à déterminer, ne sont pas nécessaires au calcul de l'estimateur de variance simplifié.

Nous avons d'abord montré que l'approche proposée par Ohlsson (1998) pouvait être vue comme une méthode basée sur une approximation des π_{kl}. Nous avons ensuite étudié le cas des plans de grappes à un degré, pour lequel nous avons conclu que la simple application de l'approche d'Ohlsson au niveau des grappes, bien que facile à mettre en place, n'est pas adaptée au cas de fichiers de données produits par Statistique Canada. Une solution alternative, qui consiste à appliquer l'approche d'Ohlsson au niveau des éléments, a donc été étudiée.

Les conditions sous laquelle cette approche s'avère adéquate ont été étudiées sous le modèle linéaire mixte. Nous avons montré que la méthode proposée offre de bonnes propriétés en termes de biais lorsque la taille des grappes est petite et la corrélation intra-grappe est près de zéro. Nous avons ensuite proposé un estimateur ajusté pour le biais, qu'il est préférable d'employer lorsque l'une et/ou l'autre de ces conditions ne sont pas respectées.

Une première étude par simulation nous a permis de confirmer la validité de l'approche d'Ohlsson dans le cas d'une population simple. Dans une seconde étude par simulation, nous avons comparé l'approche que nous suggérons ainsi que l'estimateur ajusté, aux méthodes classiques d'estimation de la variance. Les résultats obtenus ont confirmé les conditions de validité qui avaient été établies

dans un cadre théorique. De plus, nous avons noté que l'estimateur ajusté est adéquat lorsque l'une et/ou l'autre des conditions de validité ne sont pas respectées.

Nous avons montré que la méthode simplifiée d'estimation de la variance présentée dans ce mémoire est appropriée lorsque certaines conditions sont respectées. Dans le cas contraire, un estimateur ajusté pour le biais a été proposé. Il serait toutefois intéressant de développer une stratégie qui, plutôt que de recourir à un ajustement, permet directement d'estimer la variance, pour toute taille de grappes et corrélation intra-grappe. Nous croyons que l'estimateur de variance adéquat pourrait être développé en étudiant la combinaison d'un plan de Poisson d'éléments et d'un estimateur de calage.

Bibliographie

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

Berger, Y. G. (2011). Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan Journal of Statistics*, **27**, 407–426.

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

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators : Linearization and residual techniques. *Survey Methodology*, **25**, 193–203.

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

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs : a computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, **93**, 269–278.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 1491–1523.

Hájek, J. (1981). Sampling from a finite population. New York, Marcel Dekker.

Hartley, H. O. et Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *The Annals of Mathematical Statistics*, **33**, 350–374.

Haziza, D., Mecatti, F. et Rao, J. N. K. (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. (1990). Sequential Poisson sampling from a business register and its application to the swedish consumer price index. *Statistics Sweden Research and Development Report*, **1990 :6**.

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

Rosén, B. (1991). Variance estimation for systematic PPS-sampling. *Statistics Sweden Technical Report*, **1991 :15**.

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

Särndal, C.E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, **91**, 1289–1300.

Särndal, C.E., Swensson B. et Wretman, J.H. (1992). *Model assisted survey sampling.* New York : Springer–Verlag.

Thompson, M. E. et Wu, C. (2008). Simulation-based randomized systematic PPS sampling under substitution of units. *Survey Methodology*, **34**, 3–10.

Annexe A

APPROCHE D'OHLSSON : UNE APPROXIMATION DES PROBABILITÉS D'INCLUSION D'ORDRE DEUX

Nous commençons par écrire d comme

$$d = \sum_{k \in U^\star} \pi_k (1 - \pi_k) = n^\star - \sum_{k \in U^\star} \pi_k^2. \tag{A.0.1}$$

En insérant l'approximation (2.2.4) dans l'expression de variance (1.2.3), on obtient

$$\begin{aligned} V_p(\widehat{Y}_\pi) &\approx -\frac{1}{2} \sum_{k \in U^\star} \sum_{l \in U^\star} \frac{\pi_k \pi_l}{n^\star} \left(\pi_k + \pi_l - 2 + \frac{d}{n^\star} \right) \left\{ \frac{y_k^2}{\pi_k^2} + \frac{y_l^2}{\pi_l^2} - 2 \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \right\} \\ &= - \sum_{k \in U^\star} \sum_{l \in U^\star} \frac{\pi_k \pi_l}{n^\star} \left(\pi_k + \pi_l - 2 + \frac{d}{n^\star} \right) \frac{y_k^2}{\pi_k^2} \\ &\qquad + \sum_{k \in U^\star} \sum_{l \in U^\star} \frac{1}{n^\star} \left(\pi_k + \pi_l - 2 + \frac{d}{n^\star} \right) y_k y_l \end{aligned} \tag{A.0.2}$$

Le premier terme dans (A.0.2) se réduit à

$$\begin{aligned} &- \sum_{k \in U^\star} \sum_{l \in U^\star} \frac{\pi_k \pi_l}{n^\star} \left(\pi_k + \pi_l - 2 + \frac{d}{n^\star} \right) \frac{y_k^2}{\pi_k^2} \\ &\quad = - \sum_{k \in U^\star} \frac{y_k^2}{n^\star} \sum_{l \in U^\star} \pi_l - \frac{1}{n^\star} \sum_{k \in U^\star} \frac{y_k^2}{\pi_k} \sum_{l \in U^\star} \pi_l^2 - \frac{1}{n^\star} \left(\frac{d}{n^\star} - 2 \right) \sum_{k \in U^\star} \frac{y_k^2}{\pi_k} \sum_{l \in U^\star} \pi_l \\ &\quad = - \sum_{k \in U^\star} y_k^2 - \frac{1}{n^\star} (n^\star - d) \sum_{k \in U^\star} \frac{y_k^2}{\pi_k} - \left(\frac{d}{n^\star} - 2 \right) \sum_{k \in U^\star} \frac{y_k^2}{\pi_k} \qquad \text{(par A.0.1)} \\ &\quad = - \sum_{k \in U^\star} y_k^2 + \left(\frac{d}{n^\star} - 1 - \frac{d}{n^\star} + 2 \right) \sum_{k \in U^\star} \frac{y_k^2}{\pi_k} \\ &\quad = \sum_{k \in U^\star} y_k^2 (\pi_k^{-1} - 1). \end{aligned}$$

A-ii

Le deuxième terme devient

$$\begin{aligned}
&\sum_{k\in U^\star}\sum_{l\in U^\star}\frac{1}{n^\star}\left(\pi_k+\pi_l-2+\frac{d}{n^\star}\right)y_k y_l\\
&=\frac{2Y}{n^\star}\sum_{k\in U^\star}\pi_k y_k+\frac{1}{n^\star}\left(\frac{d}{n^\star}-2\right)Y^2\\
&=\frac{2Y}{n^\star}\left(\sum_{k\in U^\star}\pi_k y_k-Y\right)+\frac{Y^2}{(n^\star)^2}d\\
&=\frac{2Y}{n^\star}\sum_{k\in U^\star}y_k(\pi_k-1)+\frac{Y^2}{(n^\star)^2}\sum_{k\in U^\star}\pi_k(1-\pi_k).
\end{aligned}$$

On a donc

$$\begin{aligned}
V_p(\widehat{Y}_\pi)&\approx\sum_{k\in U^\star}y_k^2(\pi_k^{-1}-1)+\frac{2Y}{n^\star}\sum_{k\in U^\star}y_k(\pi_k-1)+\frac{Y^2}{(n^\star)^2}\sum_{k\in U^\star}\pi_k(1-\pi_k)\\
&=\sum_{k\in U^\star}(\pi_k^{-1}-1)\left\{y_k^2-2Y\frac{y_k\pi_k}{n^\star}+\frac{Y^2\pi_k^2}{(n^\star)^2}\right\}\\
&=\sum_{k\in U^\star}(\pi_k^{-1}-1)\left\{y_k^2-2Y\frac{y_k x_k}{X}+\frac{Y^2x_k^2}{X^2}\right\}\\
&=\sum_{k\in U^\star}(\pi_k^{-1}-1)\left\{y_k-\frac{x_kY}{X}\right\}^2.
\end{aligned}$$

De cette expression, on déduit l'estimateur

$$\sum_{k\in s^\star}\pi_k^{-1}(\pi_k^{-1}-1)\left\{y_k-\frac{x_k\widehat{Y}_\pi}{X}\right\}^2,$$

qui correspond à l'estimateur (2.2.3).

Annexe B

BIAIS RELATIF DÛ À L'UTILISATION DE L'APPROCHE D'OHLSSON AU NIVEAU DES ÉLÉMENTS DANS LE CAS DES GRAPPES

Posons $p_k = x_k/X$ et $B = \sum_{t\in U^\star} (\pi_t^{-1} - 1)p_t^2$. Réécrivons d'abord la variance (2.2.2) à l'aide de la notation de grappes.

$$
\begin{aligned}
V_{Oh} &= \sum_{k\in U^\star} (\pi_k^{-1} - 1)\left(y_k - \frac{x_k Y}{X}\right)^2 \\
&= \sum_{k\in U^\star} (\pi_k^{-1} - 1)\left(y_k^2 - y_k p_k Y - y_k p_k Y + p_k^2 Y^2\right) \\
&= \sum_{k\in U^\star} (\pi_k^{-1} - 1)y_k^2 - \sum_{k\in U^\star} (\pi_k^{-1} - 1)y_k p_k \sum_{l\in U^\star} y_l - \sum_{l\in U^\star} (\pi_l^{-1} - 1)y_l p_l \sum_{k\in U^\star} y_k \\
&\qquad + \sum_{k\in U^\star} (\pi_k^{-1} - 1)p_k^2 \sum_{l\in U^\star}\sum_{s\in U^\star} y_l y_s \\
&= \sum_{k\in U^\star} (\pi_k^{-1} - 1)y_k^2 - \sum_{k\in U^\star}\sum_{l\in U^\star} (\pi_k^{-1} - 1)p_k y_k y_l - \sum_{k\in U^\star}\sum_{l\in U^\star} (\pi_l^{-1} - 1)p_l y_k y_l \\
&\qquad + \sum_{k\in U^\star}\sum_{l\in U^\star} B y_k y_l \\
&= \sum_{k\in U^\star} (\pi_k^{-1} - 1)y_k^2 + \sum_{k\in U^\star}\sum_{l\in U^\star} y_k y_l \left\{B - (\pi_k^{-1} - 1)p_k - (\pi_l^{-1} - 1)p_l\right\} \\
&= \sum_{k\in U^\star} y_k^2 \left\{(\pi_k^{-1} - 1) + B - 2(\pi_k^{-1} - 1)p_k\right\} \\
&\qquad + \sum_{\substack{k\in U^\star \\ k\neq l}}\sum_{l\in U^\star} y_k y_l \left\{B - (\pi_k^{-1} - 1)p_k - (\pi_l^{-1} - 1)p_l\right\} \\
&= \sum_{i\in U}\sum_{k\in U_i} \left\{(\pi_k^{-1} - 1) + B - 2(\pi_k^{-1} - 1)p_{ik}\right\} y_{ik}^2 \\
&\qquad + \sum_{i\in U}\sum_{j\in U}\sum_{\substack{k\in U_i \\ k\neq l}}\sum_{l\in U_j} y_{ik} y_{jl} \left\{B - (\pi_k^{-1} - 1)p_{ik} - (\pi_l^{-1} - 1)p_{jl}\right\}
\end{aligned}
$$

B-ii

$$= \sum_{i\in U}\sum_{k\in U_i}\left\{(\pi_i^{-1}-1)+B-2(\pi_i^{-1}-1)p_{ik}\right\}y_{ik}^2$$
$$+\sum_{i\in U}\sum_{\substack{k\in U_i\\k\neq l}}\sum_{l\in U_i} y_{ik}y_{il}\left\{B-(\pi_i^{-1}-1)(p_{ik}+p_{il})\right\}$$
$$+\sum_{\substack{i\in U\\i\neq j}}\sum_{j\in U}\sum_{k\in U_i}\sum_{l\in U_j} y_{ik}y_{jl}\left\{B-(\pi_i^{-1}-1)p_{ik}-(\pi_j^{-1}-1)p_{jl}\right\},$$

où $B=\sum\limits_{t\in U^\star}(\pi_t^{-1}-1)p_t^2=\sum\limits_{i\in U}(\pi_i^{-1}-1)\sum\limits_{k\in U_i}p_{ik}^2$.
Puisqu'on s'intéresse au biais relatif de (2.2.2), on cherche à calculer

$$E_m\left\{V_{Oh}-V_p(\widehat{Y}_\pi)\right\}/E_m\left\{V_p(\widehat{Y}_\pi)\right\}, \tag{B.0.1}$$

où $V_p(\widehat{Y}_\pi)$ représente la vraie variance de l'estimateur Horvitz-Thompson, dans le cas d'échantillonnage par grappes, soit

$$V_p(\widehat{Y}_\pi)=\sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j)\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}.$$

À partir de V_{Oh}, écrit selon la notation des grappes, déterminons $E_m(V_{Oh})$ sous le modèle (2.3.4). Notons d'abord que ce modèle entraîne les résultats suivants, pour $k\neq l$ et $i\neq j$:

(i) $E_m(y_{ik}^2)=V_m(y_{ik})+E_m(y_{ik})^2=\sigma_v^2+\sigma_\epsilon^2+\beta^2x_{ik}^2$;

(ii) $E_m(y_{ik}y_{il})=E\{(\beta x_{ik}+v_i+\epsilon_{ik})(\beta x_{il}+v_i+\epsilon_{il})\}=\beta^2x_{ik}x_{il}+\sigma_v^2$;

(iii) $E_m(y_{ik}y_{jl})=E\{(\beta x_{ik}+v_i+\epsilon_{ik})(\beta x_{jl}+v_j+\epsilon_{jl})\}=\beta^2x_{ik}x_{jl}$.

On a donc

$$E_m(V_{Oh})=E_m\Bigg[\sum_{i\in U}\sum_{k\in U_i}y_{ik}^2\left\{(\pi_i^{-1}-1)+B-2(\pi_i^{-1}-1)p_{ik}\right\}$$
$$+\sum_{i\in U}\sum_{\substack{k\in U_i\\k\neq l}}\sum_{l\in U_i}y_{ik}y_{il}\left\{B-(\pi_i^{-1}-1)(p_{ik}+p_{il})\right\}$$
$$+\sum_{\substack{i\in U\\i\neq j}}\sum_{j\in U}\sum_{k\in U_i}\sum_{l\in U_j}y_{ik}y_{jl}\left\{B-(\pi_i^{-1}-1)p_{ik}-(\pi_j^{-1}-1)p_{jl}\right\}\Bigg]$$
$$=\sum_{i\in U}\sum_{k\in U_i}(\beta^2x_{ik}^2+\sigma_v^2+\sigma_\epsilon^2)\left\{(\pi_i^{-1}-1)+B-2(\pi_i^{-1}-1)p_{ik}\right\}$$
$$+\sum_{i\in U}\sum_{\substack{k\in U_i\\k\neq l}}\sum_{l\in U_i}(\beta^2x_{ik}x_{il}+\sigma_v^2)\left\{B-(\pi_i^{-1}-1)(p_{ik}+p_{il})\right\}$$
$$+\sum_{\substack{i\in U\\i\neq j}}\sum_{j\in U}\sum_{k\in U_i}\sum_{l\in U_j}\beta^2x_{ik}x_{jl}\left\{B-(\pi_i^{-1}-1)p_{ik}-(\pi_j^{-1}-1)p_{jl}\right\}$$
$$=\sum_{i\in U}\sum_{j\in U}\sum_{k\in U_i}\sum_{l\in U_j}\beta^2x_{ik}x_{jl}\left\{B-(\pi_i^{-1}-1)p_{ik}-(\pi_j^{-1}-1)p_{jl}\right\}$$

$$
\begin{aligned}
&+ \sum_{i\in U}\sum_{k\in U_i} \beta^2 x_{ik}^2(\pi_i^{-1}-1) \\
&\quad + \sum_{i\in U}\sum_{k\in U_i} (\sigma_v^2+\sigma_\epsilon^2)\left\{(\pi_i^{-1}-1)+B-2(\pi_i^{-1}-1)p_{ik}\right\} \\
&\quad\quad + \sum_{i\in U}\sum_{\substack{k\in U_i\\ k\neq l}}\sum_{l\in U_i} \sigma_v^2\left\{B-(\pi_i^{-1}-1)(p_{ik}+p_{il})\right\} \\
&= \sum_{i\in U}\sum_{j\in U}\sum_{k\in U_i}\sum_{l\in U_j} \beta^2 x_{ik}x_{jl}\left\{B-(\pi_i^{-1}-1)p_{ik}-(\pi_j^{-1}-1)p_{jl}\right\} \\
&\quad + \sum_{i\in U}\sum_{k\in U_i} (\pi_i^{-1}-1)(\beta^2 x_{ik}^2+\sigma_v^2+\sigma_\epsilon^2)+\sigma_\epsilon^2\left\{B-2(\pi_i^{-1}-1)p_{ik}\right\} \\
&\quad\quad + \sum_{i\in U}\sum_{k\in U_i}\sum_{l\in U_i} \sigma_v^2\left\{B-(\pi_i^{-1}-1)(p_{ik}+p_{il})\right\}.
\end{aligned}
$$

En développant, le premier terme devient

$$
\begin{aligned}
&\sum_{i\in U}\sum_{j\in U}\sum_{k\in U_i}\sum_{l\in U_j} \beta^2 x_{ik}x_{jl}\left\{B-(\pi_i^{-1}-1)p_{ik}-(\pi_j^{-1}-1)p_{jl}\right\} \\
&\quad = \beta^2 BX^2 - \beta^2\sum_{i\in U}(\pi_i^{-1}-1)\sum_{k\in U_i} p_{ik}x_{ik}\sum_{j\in U}\sum_{l\in U_j} x_{jl} \\
&\quad\quad - \beta^2\sum_{j\in U}(\pi_j^{-1}-1)\sum_{l\in U_j} p_{jl}x_{jl}\sum_{i\in U}\sum_{k\in U_i} x_{ik} \\
&\quad = \beta^2 BX^2 - \beta^2\sum_{i\in U}(\pi_i^{-1}-1)\sum_{k\in U_i} x_{ik}^2 - \beta^2\sum_{j\in U}(\pi_j^{-1}-1)\sum_{l\in U_j} x_{jl}x_{jl} \\
&\quad = \beta^2\sum_{i\in U}(\pi_i^{-1}-1)\sum_{k\in U_i} x_{ik}^2 - 2\beta^2\sum_{i\in U}(\pi_i^{-1}-1)\sum_{k\in U_i} x_{ik}^2 \\
&\quad = -\beta^2\sum_{i\in U}(\pi_i^{-1}-1)\sum_{k\in U_i} x_{ik}^2.
\end{aligned}
$$

Le deuxième terme devient

$$
\begin{aligned}
&\sum_{i\in U}\sum_{k\in U_i}\sum_{l\in U_i} \sigma_v^2\left\{B-(\pi_i^{-1}-1)(p_{ik}+p_{il})\right\} \\
&\quad = \sigma_v^2 B\sum_{i\in U} M_i^2 - \sigma_v^2\sum_{i\in U}(\pi_i^{-1}-1)M_i\sum_{k\in U_i^\star}\frac{x_{ik}}{X} - \sigma_v^2\sum_{i\in U}(\pi_i^{-1}-1)M_i\sum_{l\in U_i^\star}\frac{x_{il}}{X} \\
&\quad = \frac{\sigma_v^2\sum\limits_{i\in U} M_i^2}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\sum_{k\in U_i^\star} x_{ik}^2 - \frac{2\sigma_v^2}{X}\sum_{i\in U}(\pi_i^{-1}-1)M_iX_i \\
&\quad = \frac{\sigma_v^2\sum\limits_{i\in U} M_i^2}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\left(\sum_{k\in U_i^\star} x_{ik}^2 - 2X\frac{M_iX_i}{\sum\limits_{i\in U} M_i^2}\right)
\end{aligned}
$$

et le troisième terme devient

$$\begin{aligned}
&\sum_{i\in U}\sum_{k\in U_i}(\pi_i^{-1}-1)(\beta^2x_{ik}^2+\sigma_v^2+\sigma_\epsilon^2)+\sigma_\epsilon^2\left\{B-2(\pi_i^{-1}-1)p_{ik}\right\}\\
&\qquad=\beta^2\sum_{i\in U}(\pi_i^{-1}-1)\sum_{k\in U_i^\star}x_{ik}^2+(\sigma_v^2+\sigma_\epsilon^2)\sum_{i\in U}(\pi_i^{-1}-1)M_i+\sigma_\epsilon^2B\sum_{i\in U}M_i\\
&\qquad\qquad-\frac{2\sigma_\epsilon^2}{X}\sum_{i\in U}(\pi_i^{-1}-1)X_i\\
&\qquad=\beta^2\sum_{i\in U}(\pi_i^{-1}-1)\sum_{k\in U_i^\star}x_{ik}^2+(\sigma_v^2+\sigma_\epsilon^2)\sum_{i\in U}(\pi_i^{-1}-1)M_i\\
&\qquad\qquad+\frac{\sigma_\epsilon^2\sum\limits_{i\in U}M_i}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\sum_{k\in U_i^\star}x_{ik}^2-\frac{2\sigma_\epsilon^2}{X}\sum_{i\in U}(\pi_i^{-1}-1)X_i\\
&\qquad=\beta^2\sum_{i\in U}(\pi_i^{-1}-1)\sum_{k\in U_i^\star}x_{ik}^2+(\sigma_v^2+\sigma_\epsilon^2)\sum_{i\in U}(\pi_i^{-1}-1)M_i\\
&\qquad\qquad+\frac{\sigma_\epsilon^2\sum\limits_{i\in U}M_i}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\left(\sum_{k\in U_i^\star}x_{ik}^2-2X\frac{X_i}{\sum\limits_{i\in U}M_i}\right).
\end{aligned}$$

Ainsi, on a que

$$\begin{aligned}
E_m(V_{Oh})=&\frac{\sigma_v^2\sum\limits_{i\in U}M_i^2}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\left(\sum_{k\in U_i^\star}x_{ik}^2-2X\frac{M_iX_i}{\sum\limits_{i\in U}M_i^2}\right)\\
&+(\sigma_v^2+\sigma_\epsilon^2)\sum_{i\in U}(\pi_i^{-1}-1)M_i\\
&\qquad+\frac{\sigma_\epsilon^2\sum\limits_{i\in U}M_i}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\left(\sum_{k\in U_i^\star}x_{ik}^2-2X\frac{X_i}{\sum\limits_{i\in U}M_i}\right).
\end{aligned}$$

Calculons maintenant $E_m\left\{V_p(\widehat{Y}_\pi)\right\}$. Dans le contexte d'un plan de taille fixe, la Proposition 1.2.3 nous permet d'écrire

$$\begin{aligned}
V_p(\widehat{Y}_\pi)&=-\frac{1}{2}\sum_{i\in U}\sum_{j\in U}\Delta_{ij}\left(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right)^2\\
&=\sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j)\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}. \qquad \text{(B.0.2)}
\end{aligned}$$

Pour des raisons de simplicité, on calcule donc l'espérance par rapport au modèle à partir de la forme B.0.2 de la variance.

$$
\begin{aligned}
E_m\left\{V_p(\widehat{Y}_\pi)\right\} &= E_m\left\{\sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j)\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}\right\} \\
&= E_m\Bigg\{\sum_{i\in U}\sum_{k\in U_i^\star} y_{ik}^2(\pi_i^{-1}-1)+\sum_{i\in U}\sum_{\substack{k\in U_i^\star\\ k\neq l}}\sum_{l\in U_i^\star} y_{ik}y_{il}(\pi_i^{-1}-1) \\
&\qquad +\sum_{\substack{i\in U\\ i\neq j}}\sum_{j\in U}\sum_{k\in U_i^\star}\sum_{l\in U_j^\star} y_{ik}y_{jl}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\Bigg\} \\
&= \sum_{i\in U}\sum_{k\in U_i^\star}(\beta^2x_{ik}^2+\sigma_v^2+\sigma_\epsilon^2)(\pi_i^{-1}-1) \\
&\qquad +\sum_{i\in U}\sum_{\substack{k\in U_i^\star\\ k\neq l}}\sum_{l\in U_i^\star}(\beta^2x_{ik}x_{il}+\sigma_v^2)(\pi_i^{-1}-1) \\
&\qquad\qquad +\sum_{\substack{i\in U\\ i\neq j}}\sum_{j\in U}\sum_{k\in U_i^\star}\sum_{l\in U_j^\star}\beta^2x_{ik}x_{jl}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right) \\
&= \sum_{i\in U}\sum_{j\in U}\sum_{k\in U_i^\star}\sum_{l\in U_j^\star}\beta^2x_{ik}x_{jl}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right) \\
&\qquad +\sum_{i\in U}\sum_{k\in U_i^\star}\sum_{l\in U_i^\star}\sigma_v^2(\pi_i^{-1}-1)+\sum_{i\in U}\sum_{k\in U_i^\star}\sigma_\epsilon^2(\pi_i^{-1}-1) \\
&= \beta^2\sum_{i\in U}\sum_{j\in U}\pi_{ij}\frac{X_i}{\pi_i}\frac{X_j}{\pi_j}-\beta^2\sum_{i\in U}\sum_{j\in U}X_iX_j \\
&\qquad +\sigma_v^2\sum_{i\in U}(\pi_i^{-1}-1)M_i^2+\sigma_\epsilon^2\sum_{i\in U}(\pi_i^{-1}-1)M_i \\
&= \beta^2\sum_{i\in U}\sum_{j\in U}\pi_{ij}\frac{X^2}{n^2}-\beta^2X^2+\sum_{i\in U}(\pi_i^{-1}-1)M_i(M_i\sigma_v^2+\sigma_\epsilon^2) \\
&= \sum_{i\in U}(\pi_i^{-1}-1)M_i(M_i\sigma_v^2+\sigma_\epsilon^2)
\end{aligned}
$$

Le numérateur de (B.0.1) est donc

$$
\begin{aligned}
E_m\left\{V_{Oh}-V_p(\widehat{Y}_\pi)\right\} &= E_m(V_{Oh})-E_m\left\{V_p(\widehat{Y}_\pi)\right\} \\
&= \frac{\sigma_v^2\sum\limits_{i\in U}M_i^2}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\left(\sum_{k\in U_i^\star}x_{ik}^2-2X\frac{M_iX_i}{\sum\limits_{i\in U}M_i^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{\sigma_\epsilon^2\sum\limits_{i\in U}M_i}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\left(\sum_{k\in U_i^\star}x_{ik}^2-2X\frac{X_i}{\sum\limits_{i\in U}M_i}\right)\\
&\quad+(\sigma_v^2+\sigma_\epsilon^2)\sum_{i\in U}(\pi_i^{-1}-1)M_i\\
&\quad-\sigma_v^2\sum_{i\in U}(\pi_i^{-1}-1)M_i^2-\sigma_\epsilon^2\sum_{i\in U}(\pi_i^{-1}-1)M_i\\
=&\frac{\sigma_v^2\sum\limits_{i\in U}M_i^2}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\left(\sum_{k\in U_i^\star}x_{ik}^2-2X\frac{M_iX_i}{\sum\limits_{i\in U}M_i^2}\right)\\
&+\frac{\sigma_\epsilon^2\sum\limits_{i\in U}M_i}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\left(\sum_{k\in U_i^\star}x_{ik}^2-2X\frac{X_i}{\sum\limits_{i\in U}M_i}\right)\\
&\quad+\sigma_v^2\sum_{i\in U}(\pi_i^{-1}-1)M_i(1-M_i).
\end{aligned}
$$

On peut maintenant calculer le biais relatif.

$$
BR(V_{Oh})=\frac{E_m\left\{V_{Oh}-V_p(\widehat{Y}_\pi)\right\}}{E_m\left\{V_p(\widehat{Y}_\pi)\right\}}=A-\frac{\sigma_v^2\sum\limits_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1)}{\sum\limits_{i\in U}(\pi_i^{-1}-1)M_i(M_i\sigma_v^2+\sigma_\epsilon^2)},\quad\text{(B.0.3)}
$$

où

$$
\begin{aligned}
A=K^{-1}\Bigg\{&\frac{\sigma_v^2\sum\limits_{i\in U}M_i^2}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\left(\sum_{k\in U_i^\star}x_{ik}^2-2X\frac{M_iX_i}{\sum\limits_{i\in U}M_i^2}\right)\\
&+\frac{\sigma_\epsilon^2\sum\limits_{i\in U}M_i}{X^2}\sum_{i\in U}(\pi_i^{-1}-1)\left(\sum_{k\in U_i^\star}x_{ik}^2-2X\frac{X_i}{\sum\limits_{i\in U}M_i}\right)\Bigg\}
\end{aligned}
$$

pour $K=\sum\limits_{i\in U}(\pi_i^{-1}-1)M_i(M_i\sigma_v^2+\sigma_\epsilon^2)$.

Le second terme de (B.0.3) se réécrit

$$
\begin{aligned}
&-\frac{\sigma_v^2\sum\limits_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1)}{\sum\limits_{i\in U}(\pi_i^{-1}-1)M_i(M_i\sigma_v^2+\sigma_\epsilon^2)}\\
=&-\frac{\sum\limits_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1)}{\sum\limits_{i\in U}M_i(\pi_i^{-1}-1)\left(\dfrac{M_i\sigma_v^2+\sigma_\epsilon^2}{\sigma_v^2}\right)}\\
=&-\frac{\sum\limits_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1)}{\sum\limits_{i\in U}M_i(\pi_i^{-1}-1)\left\{\dfrac{(M_i-1)\sigma_v^2+\sigma_\epsilon^2+\sigma_v^2}{\sigma_v^2}\right\}}
\end{aligned}
$$

$$= -\frac{\sum_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1)}{\sum_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1)+\sum_{i\in U}M_i(\pi_i^{-1}-1)\left(\frac{1}{\rho}\right)}$$

$$= -\frac{\sum_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1)}{\sum_{i\in U}(\pi_i^{-1}-1)M_i^2+\sum_{i\in U}(\pi_i^{-1}-1)M_i\left(\frac{1}{\rho}-1\right)},$$

où $\rho = \dfrac{\sigma_v^2}{\sigma_\epsilon^2+\sigma_v^2}$. Le biais relatif de V_{Oh} est donc

$$BR(V_{Oh}) = A - \frac{\sum_{i\in U}(\pi_i^{-1}-1)M_i(M_i-1)}{\sum_{i\in U}(\pi_i^{-1}-1)M_i^2+\sum_{i\in U}(\pi_i^{-1}-1)M_i\left(\frac{1}{\rho}-1\right)}.$$

Annexe C

ORDRE DE GRANDEUR DU TERME $\tilde{A}$

Nous faisons les hypothèses suivantes :

(i) $\max_{i \in U} d_i = O(N/n)$;

(ii) $\max_{i \in U} \sum_{k \in U_i^\star} x_{ik}^2 / M_i = O(1)$;

(iii) $\overline{X} = \sum_{i \in U} \sum_{k \in U_i^\star} x_{ik}/(NM) = O(1)$.

De plus, nous supposons que $x_{ik} > 0$, $\forall (ik)$. Cette hypothèse est crédible puisque x est une variable de taille.

Le numérateur de $\tilde{A}$ s'écrit comme

$$\begin{aligned} A_{num} &= \frac{N}{X^2} \sum_{i \in U} (\pi_i^{-1} - 1) \left(\sum_{k \in U_i^\star} x_{ik}^2 - 2 \frac{X}{MN} X_i \right) \\ &= \frac{N}{(MN)^2 \overline{X}^2} \sum_{i \in U} (d_i - 1) \sum_{k \in U_i^\star} x_{ik}^2 - 2 \frac{1}{(M^2 N) \overline{X}} \sum_{i \in U} (d_i - 1) X_i. \end{aligned}$$

On a que

$$\begin{aligned} A_{num} &\leq \frac{N}{(MN)^2 \overline{X}^2} \sum_{i \in U} (d_i - 1) \sum_{k \in U_i^\star} x_{ik}^2 \\ &\leq \frac{N}{(MN)^2 \overline{X}^2} \max_{i \in U} (d_i - 1) \sum_{i \in U} M_i \left(\frac{\sum_{k \in U_i^\star} x_{ik}^2}{M_i} \right) \\ &\leq \frac{N}{(MN) \overline{X}^2} \max_{i \in U} (d_i - 1) \max_{i \in U} \left(\frac{\sum_{k \in U_i^\star} x_{ik}^2}{M_i} \right) \\ &= O\left(\frac{N}{Mn} \right). \end{aligned}$$

C-ii

Le dénominateur s'écrit

$$\sum_{i\in U}(d_i-1)\leq \max_{i\in U}(d_i-1)N = O\left(\frac{N^2}{n}\right).$$

On a donc

$$\tilde{A} = O\left(\frac{N}{Mn}\frac{n}{N^2}\right) = O\left(\frac{1}{MN}\right).$$

Annexe D

ORDRE DE GRANDEUR DU TERME A

Posons $M_0 = \sum_{i\in U} M_i$. Nous faisons les hypothèses suivantes :

(i) $\max_{i\in U} d_i = O(N/n)$;

(ii) $\max_{i\in U} \sum_{k\in U_i^\star} x_{ik}^2/M_i = O(1)$;

(iii) $\sum_{i\in U} M_i^2/M_0 = O(1)$;

(iv) $\overline{X} = \sum_{i\in U} X_i/M_0 = O(1)$;

(v) $\sum_{i\in U} M_i X_i/M_0 = O(1)$;

(vi) $\sigma_\epsilon^2 = O(1)$ et $\sigma_v^2 = O(1)$.

Le terme K de (B.0.3) s'écrit comme

$$\begin{aligned} K &= \sigma_v^2 \sum_{i\in U} (d_i - 1)M_i^2 + \sigma_\epsilon^2 \sum_{i\in U} (d_i - 1)M_i \\ &\leq \sigma_v^2 \max_{i\in U} (d_i - 1)M_0 \left(\sum_{i\in U} M_i^2/M_0\right) + \sigma_\epsilon^2 \max_{i\in U} (d_i - 1)M_0 \\ &= O\left(M_0 \frac{N}{n}\right) \end{aligned}$$

Ensuite, pour le premier terme de A, on a que

$$\begin{aligned} A_1 &= \frac{\sigma_v^2 \left(\sum_{i\in U} M_i^2\right)}{X^2} \sum_{i\in U} (d_i - 1) \left(\sum_{k\in U_i^\star} x_{ik}^2 - 2X \frac{M_i X_i}{\sum_{i\in U} M_i^2}\right) \\ &= \frac{\sigma_v^2 M_0 \left(\sum_{i\in U} M_i^2/M_0\right)}{M_0^2 \overline{X}^2} \sum_{i\in U} (d_i - 1)M_i \left(\sum_{k\in U_i^\star} x_{ik}^2/M_i\right) \\ &\qquad - 2\frac{\sigma_v^2 M_0 \left(\sum_{i\in U} M_i^2/M_0\right)}{M_0 \overline{X}} \frac{1}{M_0 \sum_{i\in U} M_i^2/M_0} \sum_{i\in U} (d_i - 1)M_i X_i \end{aligned}$$

D-ii

$$
\begin{aligned}
&\leq \frac{\sigma_v^2 \left(\sum_{i \in U} M_i^2 / M_0 \right)}{M_0 \overline{X}^2} \max_{i \in U} (d_i - 1) \max_{i \in U} \left(\sum_{k \in U_i^\star} x_{ik}^2 / M_i \right) M_0 \\
&= O\left(\frac{N}{n} \right).
\end{aligned}
$$

Pour le second terme de A, on a que

$$
\begin{aligned}
A_2 &= \frac{\sigma_\epsilon^2 M_0}{X^2} \sum_{i \in U} (d_i - 1) \left(\sum_{k \in U_i^\star} x_{ik}^2 - 2X \frac{X_i}{M_0} \right) \\
&= \frac{\sigma_\epsilon^2}{M_0 \overline{X}^2} \sum_{i \in U} (d_i - 1) M_i \left(\sum_{k \in U_i^\star} x_{ik}^2 / M_i \right) - 2 \frac{\sigma_\epsilon^2}{M_0 \overline{X}} \sum_{i \in U} (d_i - 1) X_i \\
&\leq \frac{\sigma_\epsilon^2}{M_0 \overline{X}^2} \max_{i \in U} (d_i - 1) \max_{i \in U} \left(\sum_{k \in U_i^\star} x_{ik}^2 / M_i \right) M_0 \\
&= O\left(\frac{N}{n} \right).
\end{aligned}
$$

En combinant les résultats pour K, A_1 et A_2, on a que

$$
\begin{aligned}
A &= K^{-1}(A_1 + A_2) \\
&= \frac{O\left(N/n\right)}{O\left(M_0 N/n\right)} \\
&= O\left(\frac{1}{M_0} \right).
\end{aligned}
$$

Annexe E

ESTIMATION DE LA VARIANCE DE L'EFFET DE GRAPPE

Pour estimer σ_v^2, on ajuste d'abord un modèle linéaire mixte aux données de l'échantillon. Pour la grappe i, ce modèle correspond à

$$\boldsymbol{y_i} = \beta \boldsymbol{x_i} + \mathbf{1} v_i + \boldsymbol{\epsilon_i},$$

qu'on peut aussi écrire

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{iM_i} \end{pmatrix} = \beta \begin{pmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{iM_i} \end{pmatrix} + v_i \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \dots \\ \epsilon_{iM_i}, \end{pmatrix}.$$

On fait l'hypothèse que $\epsilon_{ik} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ et $v_i \sim \mathcal{N}(0, D\sigma_\epsilon^2)$ avec $D = (1-\rho)^{-1}\rho$. Aussi, on suppose que la matrice $\sum_{i \in U} \boldsymbol{x_i}^\top \boldsymbol{x_i}$ est non-singulière et que $N^\star > N > 1$.

À partir du modèle linéaire mixte, la composante $\sigma_v^2 = D\sigma_\epsilon^2$ est estimée par maximisation de la log-vraisemblance. Cette dernière est donnée par

$$l(\beta, \sigma_\epsilon^2, D) = -\frac{N^\star}{2} \ln(2\pi) - \frac{1}{2} \left\{ N^\star \ln \sigma_\epsilon^2 + \sum_{i \in U} \left(\ln |V_i| + \frac{1}{\sigma_\epsilon^2} \boldsymbol{e_i}^\top V_i^{-1} \boldsymbol{e_i} \right) \right\},$$

où $\boldsymbol{e_i} = \boldsymbol{y_i} - \beta \boldsymbol{x_i}$ et $V_i = I + \mathbf{1} D \mathbf{1}^\top$, avec I correspondant à la matrice identité de dimension $M_i \times M_i$.