Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 1 Capitole (UT1 Capitole)*

Présentée et soutenue le par :

HÉLÈNE JUILLARD

Méthodes d'estimation et d'estimation de variance pour une enquête longitudinale - Application aux données de l'Etude Longitudinale Française depuis l'Enfance (Elfe)

JURY

École doctorale et spécialité :

MITT : Domaine Mathématiques : Mathématiques appliquées

Unité de Recherche :

T.S.E.

Directeurs de Thèse :

Anne RUIZ-GAZEN et *Guillaume CHAUVET*

Rapporteurs :

Hervé CARDOT et *Chris SKINNER*

Résumé

Dans ce document, on suppose que l'aléa provient du tirage de l'échantillon (inférence basée sur le plan de sondage). Chaque échantillonnage conduit à une variance dite d'échantillonnage. Après déroulement d'une enquête, l'estimation de cette variance va servir de mesure de précision (ou d'incertitude) pour les estimateurs des paramètres étudiés.

La cohorte Elfe, démarrée en 2011, comprend plus de 18 000 enfants dont les parents ont donné leur consentement à l'inclusion. Dans chacune des maternités sélectionnées, les nourrissons de la population d'inférence nés durant quatre périodes spécifiques représentant chacune des quatre saisons de l'année 2011 ont été sélectionnés. Elfe est la première étude longitudinale de ce type en France, suivant les enfants de leur naissance à l'âge adulte. Elle aborde les multiples aspects de la vie de l'enfant sous l'angle des sciences sociales, de la santé et de la santé-environnement. La cohorte Elfe a été sélectionnée selon un plan de sondage non standard appelé échantillonnage produit, avec les sélections indépendantes d'un échantillon de maternités et d'un échantillon de jours. Le suivi de l'enfant commence dès ses premiers jours, à la maternité. Ensuite, lorsque les enfants fêtent leurs deux mois, les parents sont contactés pour un premier entretien téléphonique, puis au premier anniversaire des enfants, à leurs deux ans, 3 ans et demi et cinq ans et demi. L'enquête est longitudinale.

Le premier chapitre de cette thèse introduit des notions relatives à la théorie des sondages et présente l'enquête Elfe (Etude Longitudinale Française depuis l'Enfance) ; ses données serviront d'illustration aux résultats théoriques issus de cette thèse. Le deuxième chapitre porte sur le plan produit et propose dans un cadre théorique général des estimateurs sans biais et des estimateurs simplifiés de variance pour traiter ce plan. Il est aussi montré que ce plan est en général moins efficace que celui classique à deux degrés d'échantillonnage. Le chapitre trois est en continuité avec le précédent : pour le plan produit, cinq estimateurs sans biais de type Yates-Grundy sont proposés à partir de cinq différentes décompositions possibles de la variance. Le chapitre quatre est un article permettant au lecteur de différencier le plan produit du plan à deux degrés, et de mettre en pratique les étapes d'échantillonnage et d'estimation sous les logiciels R, SAS et Stata. Le chapitre cinq est lui, consacré à la variance et l'estimation de la variance pour une enquête de type cohorte avec processus de non-réponse monotone. Le chapitre six est un rapport méthodologique pour les utilisateurs où l'estimation de la variance appropriée au plan Elfe est expliquée et mise en œuvre avec les logiciels R, SAS et Stata.

Tous les résultats des études par simulation présentés dans ce document sont reproductibles, les codes étant proposés en annexe.

Mots-clés : échantillonnage à deux degrés, échantillonnage produit, enquête longitudinale, estimation de variance, logiciels R / SAS / Stata, non-réponse.

Abstract

In this document, we are interested in estimation under a design-based framework, where the randomness arises from the sample selection. Each sampling leads to a sampling variance. After the survey, the estimation of this variance will serve as a measure of precision (or uncertainty) for the estimators of the parameters under study.

The 2011 ELFE cohort comprises more than 18,000 children whose parents consented to their inclusion. In each of the selected maternity units, targeted babies born during four specific periods representing each of the four seasons in 2011 were selected. ELFE is the first longitudinal study of its kind in France, tracking children from birth to adulthood. It will examine every aspect of these children's lives from the perspectives of health, social sciences and environmental health. The ELFE cohort was selected through a non-standard sampling design that is called cross-classified sampling, with independent selections of the sample of maternity units and of the sample of days. In this work, we propose unbiased variance estimators to handle this type of sampling designs, and we derive specific variance estimators adapted to the ELFE case. Tracking of the babies starts when they are just a few days old and still at the maternity unit. When the children reach the age of two months, the parents are contacted for the first telephone interview. When the children are one year old, and again when they reach the ages of two, three and a half years and five and a half years, their parents will once more be contacted by telephone. The survey is longitudinal.

The first chapter of this thesis introduces concepts related to the theory of survey design and presents the survey ELFE (French Longitudinal Study from Childhood) ; its data will be used as illustration for the theoretical results derived in this thesis. The second chapter focuses on the cross-classified design and provides unbiased estimators and simplified variance estimators to treat this design in a general theoretical framework. It is also shown that this design is generally less efficient than the conventional two-stage sampling design. Chapter three is in continuity with the previous one : for the cross-classified sampling design, five unbiased Yates-Grundy like variance estimators are available from five different possible decomposition of the variance. Chapter four is an article allowing the reader to make the difference between the cross-classified sampling design and the two-stage sampling design, and to implement the steps of sampling and estimation under the softwares R, SAS and Stata. Chapter five is devoted to variance computation and variance estimation for a cohort survey with monotone non-response. Chapter six is a methodological report to users in which the appropriate variance estimation for the ELFE design is explained and implemented with softwares R, SAS and Stata.

All the results of simulation studies presented in this document are reproducible, the codes being proposed in the annex.

Keywords : cross-classified sampling, longitudinal survey, non-response, softwares R / SAS / Stata, two-stage sampling, variance estimation.

Contributions

Articles

- Juillard, H., Chauvet, G. et Ruiz-Gazen, A., 2016, "Estimation under cross-classified sampling with application to a childhood French survey". *A paraître dans* Journal of American Statistical Association, *lien*.
- Juillard, H., 2016, "Two-dimensional sampling in practice". *A paraître dans* Case Studies In Business, Industry and Government Statistics.
- Juillard, H. et Chauvet, G., 2016, "Variance estimation under monotone non-response for a panel survey". *Soumis*, *lien*.

Travaux en cours

- Legleye, S., Razakamanana, N., Charrance, G. et Juillard, H., 2016, "A simulation tool for testing the efficiency of using paradata in the weighting process of a real random telephone survey". *En cours.*

Rapport technique

- Juillard, H., 2016, "Estimation de la variance pour l'enquête Elfe - Rapport méthodologique pour l'utilisateur", Document de Travail de l'Ined, n° 226, *lien*

Communications orales et actes de colloque

- Juillard, H., Chauvet, G. et Ruiz-Gazen, A., 2016, "Estimation de la variance pour l'enquête Elfe", Groupes de Projets Scientifiques Elfe (Paris, Elfe).
- Juillard, H., Chauvet, G. et Ruiz-Gazen, A., 2015, "Estimation de la variance pour l'enquête Elfe", Journée doctorale de l'Ined (Paris, Ined), Document de Travail de l'Ined, n° 219, *lien*
- Chauvet, G., Juillard, H. et Ruiz-Gazen, A., 2015, "Estimateurs de variance issue d'un plan produit, pour l'enquête Elfe", XIIème édition des Journées de Méthodologie Statistique (Paris, Insee), *lien*
- Chauvet, G., Juillard, H. and Ruiz-Gazen, A., 2014, "Study of the “Product” Sampling Scheme as Illustrated by the ELFE Survey", 2014 International Methodology Symposium (Canada, Statcan), *lien*
- Juillard, H., Chauvet, G. and Ruiz-Gazen, A., 2014, "Echantillonnage produit : une application à l'Etude Longitudinale Française depuis l'Enfance", 46èmes Journées de Statistiques (Bruz, ENSAI).

Posters

- Legleye, S., Razakamanana, N., Charrance, G. et Juillard, H., 2015, "Is it worth using paradata to correct for total non-response in telephone survey ? A simulation study based on real data", 6th Conference of the European Survey Research Association (Reykjavik, ESRA), page 247.
- Juillard, H., Chauvet, G. et Ruiz-Gazen, A., 2014, "Échantillonnage produit", 8ème Colloque francophone sur les sondages (Dijon, Université de Bourgogne), page 248.

Table des matières

Introduction

Version française

Un institut de recherche comme l'Ined (Institut National d'Etudes Démographiques) met en œuvre ses propres enquêtes et met à disposition de la communauté scientifiques les données récoltées. Pour cela, l'Ined s'appuie sur un service des enquêtes et sondages qui a notamment en charge de créer les plans de sondage, de construire les protocoles de collecte et suivi, puis, après collecte, de mettre à disposition des bases de données apurées et anonymisées et de proposer des jeux de pondérations. Par exemple, au service des enquêtes de l'Ined sont en cours les enquêtes Elap[1], Elfe[2], EMFS[3], l'Enquête modes de vie, santé et sécurité des victimes de violences ou Elipss[4]. Ces enquêtes peuvent être sur un temps comme l'EMFS ou sur plusieurs comme les enquêtes Elipss ou Elfe.

A partir de données issues d'échantillons, l'Ined va chercher à produire des grandeurs descriptives de la population d'intérêt, appelées, en théorie des sondages, paramètres d'intérêt en population finie. Ce peut être par exemple un total, une moyenne, un taux, un coefficient de corrélation. Pour cela il est nécessaire de 1) connaître le plan d'échantillonnage utilisé et avoir accès aux variables du plan de sondage (si le statisticien en charge de l'enquête n'a pas lui-même construit le plan), 2) définir un

[1] Etude Longitudinale sur l'accès à l'Autonomie après le Placement
[2] Etude Longitudinale Française depuis l'Enfance
[3] Enquête Mucoviscidose, Famille et Société
[4] Enquête Longitudinale par Internet Pour les Sciences Sociales

estimateur, comme par exemple l'estimateur ajusté de la non-réponse et/ou calé sur de l'information auxiliaire (qui nécessite l'existence et l'accès à cette dernière). En général, il est livré aux utilisateurs d'une enquête une base de données complétée des variables du plan de sondage et d'un ou plusieurs jeux de pondérations, avec un rapport guidant l'utilisateur vers une "bonne pratique statistique". En effet, l'utilisateur se doit de considérer la précision des chiffres qu'il désire publier en prenant en compte les notions statistiques de base que sont le biais et la variance.

En général, on déterminera des systèmes de pondérations visant à limiter le biais possible des estimateurs. Concernant les variances de ces estimateurs, leurs estimations peuvent être complexes mais restent indispensables : ce sont des quantités qui permettent de mesurer l'incertitude à travailler seulement sur un échantillon et non sur toute la population. L'exemple le plus simple est l'intervalle de confiance demandé dans toute publication d'un paramètre. Il est fonction de la variance estimée. Outre les formules théoriques de variance et celles de ses estimateurs, une difficulté non-négligeable est l'accès aux procédures logicielles : en effet, l'utilisateur demandera des fonctions pré-programmées à usage simple, ceci dans le langage de programmation d'un logiciel qu'il connaît déjà.

L'objectif de cette thèse est de 1) dériver des estimateurs et des estimateurs de variance pour un plan particulier appelé plan produit, résultant du croisement indépendant de deux plans de sondage, 2) permettre au lecteur de différencier le plan produit du plan classique à deux degrés, tant à l'étape de l'échantillonnage qu'à l'étape de l'estimation, 3) développer des estimateurs et des estimateurs de variance pour une enquête longitudinale, aussi communément appelée panel, 4) adapter les résultats du plan produit à l'enquête Elfe et proposer de premières procédures logicielles appropriées à cette enquête.

Cette thèse a été développée sur un plan théorique et simultanément sur un plan pratique en considérant les usages des utilisateurs de données d'enquête : en effet, certaines recherches, comme celles d'estimateurs simplifiés, prennent en compte la possibilité pour les utilisateurs d'utiliser les résultats obtenus (accessibilité, mise en œuvre, compréhension). Les procédures statistiques ont été décrites sur trois logiciels statistiques couramment utilisés en France, R, SAS et Stata, et tout particulièrement par les chercheurs travaillant sur les données issues de l'enquête Elfe.

Ce document est partitionné en six chapitres. Dans quatre de ces chapitres, les études théoriques développées sont validées numériquement sur des populations et exemples simulés : tous les codes utilisés sont reproduits dans ce document, à graines fixées des générateurs aléatoires, rendant ainsi les résultats reproductibles. Dans trois de ces chapitres, les résultats sont illustrés sur des données de l'enquête Elfe non mises à disposition du lecteur.

Le Chapitre 1 contextualise cette thèse. Des notions relatives à la théorie de l'échantillonnage sont rappelées, comme ce qu'est un plan de sondage, une probabilité d'inclusion de premier ordre ou de second ordre. Le plan à deux degrés est présenté et sera rappelé dans les Chapitres 2 et 4, ainsi que celui à deux phases, utile pour modéliser la non-réponse et dont on se servira dans les Chapitres 5 et 6. Ensuite, on listera les grandes familles d'enquêtes dans le temps dont font partie les panels et on définira ce qu'est un processus de non-réponse monotone (utilisé dans le Chapitre 5). L'enquête Elfe, panel commencé en 2011, sera présentée, ainsi que son plan de sondage appelé plan d'échantillonnage produit. Ce plan, traité dans tous les chapitres de cette thèse, dont nous ferons un état de la littérature, sera comparé à d'autres plans de sondage pour bien en comprendre les spécificités.

Le Chapitre 2 **Estimation under cross-classified sampling with application to a childhood survey** est un article co-écrit avec Guillaume Chauvet et Anne Ruiz-Gazen, accepté pour publication dans *Journal of the American Statistical Association* en 2016. Il porte sur le plan produit et propose un cadre théorique général pour des estimations issues de ce plan. Nous montrons notamment que ce plan est en général moins efficace que celui classique à deux degrés d'échantillonnage. Nous obtenons la distribution asymptotique de l'estimateur Horvitz-Thompson et proposons différents estimateurs de variance sans biais. La possibilité de valeurs négatives pour les estimateurs de variance nous amène à proposer des estimateurs simplifiés ainsi que les conditions requises à leur utilisation.

Le Chapitre 3 **Estimation under cross-classified sampling, continuation** est en continuité avec le précédent. Dans ce chapitre, cinq estimateurs sans biais de type Yates-Grundy sont proposés à partir de cinq différentes décompositions possibles de la variance. Si ces cinq estimateurs sont identiques dans le cas particulier de tirages aléatoires simples, nous avons cherché à les comparer à partir de tirages à probabilités inégales. Les plans de Poisson conditionnel, de Sampford et de Midzuno ont servi d'illustration à partir de simulations.

Le Chapitre 4 **Two-dimensional sampling in practice** est un article accepté pour publication dans *Case Studies In Business, Industry and Government Statistics* en 2016. Il permet au lecteur de bien différencier le plan produit du plan à deux degrés, et de mettre en pratique de façon détaillée les étapes d'échantillonnage et d'estimation sous les logiciels R, SAS et Stata pour ces deux plans de sondage. Deux analogies sont proposées : une première entre le plan produit et l'analyse de variance à deux facteurs, et une seconde entre le plan à deux degrés et l'analyse de variance à un facteur.

Le Chapitre 5 **Variance estimation under monotone non-response for a panel survey** est un article co-écrit avec Guillaume Chauvet, soumis en 2016. Il est consacré à la variance et à l'estimation de la variance pour une enquête de type cohorte avec processus de non-réponse monotone. Le cas des estimations avec probabilités de réponse connues est traité ainsi que celui des estimations avec probabilités de réponse estimées. Les formules générales sont illustrées dans le cas particulier d'un modèle de groupes homogènes de réponse.

Le Chapitre 6 **Estimation de la variance pour l'enquête Elfe** est un rapport technique et méthodologique livré en mai 2016 à l'unité Elfe et aux utilisateurs des données Elfe. Sa première version a été publiée mi-2016 dans les Documents de Travail de l'Ined[5]. Ce document explique pas à pas la méthode utilisée pour estimer la variance d'un paramètre total ou ratio, au premier temps de l'enquête Elfe, avec prise en compte de la non-réponse et du calage et propose une mise en œuvre avec les logiciels R, SAS et Stata.

[5] https://www.ined.fr/fr/publications/document-travail/estimation-variance-enquete-elfe/

English version

A research institute such as INED (National Institute of Demographic Studies) implements its own investigations and makes the collected data available to the scientific community. To do so, INED uses its investigations and polls department which is responsible for creating sample designs, building collection and monitoring protocols, and, after collection, providing databases cleared of identifiable data and offering sets of weights. For example, in 2016, the investigation department of INED is working on the surveys ELAP[6], ELFE[7], EMFS[8], the Survey lifestyles, health and safety of victims of violence and ELIPSS[9]. These surveys can be performed once like EMFS, or repeated several times like the ELIPSS or ELFE surveys.

From sample data, INED will try to produce descriptive variables of the population of interest, known as “parameters of interest in finite population” in the sampling theory. It can be a total, an average, a rate, a correlation coefficient. This requires to 1) know the sampling design used and have access to the variables of the survey design (if the statistician in charge of the investigation did not build the design himself), 2) define an estimator such as the estimator adjusted for non-response and / or calibrated on auxiliary information (which requires the existence and the access to such information). Most of the time, a database completed by the survey design variables and one or more weights sets is delivered to users of the investigation, with a report guiding the user for "good statistical practice." Indeed, the user must consider the accuracy of the figures he wants to publish, taking into account basic statistical concepts such as bias and variance.

[6] Etude Longitudinale sur l’accès à l’Autonomie après le Placement (Longitudinal Study on access to Autonomy after Placement)

[7] Etude Longitudinale Française depuis l’Enfance (longitudinal study of children in France)

[8] Enquête Mucoviscidose, Famille et Société (Cystic Fibrosis, Family and Society)

[9] Enquête Longitudinale par Internet Pour les Sciences Sociales (Longitudinal Survey Online For Social Sciences)

Most of the time, weighting systems are produced to limit the possible bias of estimators. Regarding the variances of these estimators, their estimates can be complex but remain essential : they are quantities that measure the uncertainty of working on only a sample rather than on the entire population. The simplest example is the confidence interval required in any publication of a parameter which is a function of the estimated variance. Besides the theoretical formulas of variance and those of its estimators, a non-negligible problem is access to software procedures : the users ask indeed for pre-programmed functions for single use, within a programming language software they already know.

The objectives of this thesis are to 1) derive estimators and variance estimators for a particular design called cross-classified sampling design, resulting from the crossing of two independent sampling designs, 2) allow the reader to differentiate the cross-classified sampling and the classical two-stage sampling, both at the sampling step and at the estimation step, 3) develop estimators and variance estimators for a longitudinal survey - commonly referred to as panel, 4) adapt results of the cross-classified design to the ELFE survey and propose appropriate software procedures for this survey. This thesis was developed simultaneously on a theoretical level and on a practical level by considering the use of survey data users : indeed, some researches, such as simplified estimators, take into account the possibility for users to use the results (accessibility, implementation, understanding). Statistical procedures have been described in three statistical software commonly used in France (R, SAS and Stata), and especially by researchers working on the data from the ELFE survey.

This document is partitioned into six chapters. In four of them, the theoretical studies developed are validated numerically on populations and simulated examples : all codes used are reproduced in this document, with fixed seeds of random generators, making the results easily reproducible. In three of these chapters, the presented results use the ELFE survey data, not available to the reader.

Chapter 1 contextualizes this thesis. Concepts related to sampling theory are recalled (sampling design, first-order or second-order probability of inclusion). The two-stage design is presented and will be reminded in Chapters 2 and 4, as well as the two-phase design which is useful to model non-response and which will be used in Chapters 5 and 6. Next, we will list large families of surveys over time which include the panels and we will define a monotonous process of non-response (used in Chapter 5). The ELFE survey which started in 2011 and its sampling design called cross-classified sampling design will be presented. This design, discussed in all chapters of this thesis, for which we will do a review of the literature, will be compared to other designs to fully understand its specificities.

Chapter 2 **Estimation under cross-classified sampling with application to a childhood survey** is an article co-written with Guillaume Chauvet and Anne Ruiz-Gazen, accepted for publication in *Journal of the American Statistical Association* in 2016. It refers to the cross-classified sampling design and proposes a general theoretical framework for estimates from this design. We demonstrate that this design is generally less effective than the standard two-stage sampling design. We derive the asymptotic distribution of the Horvitz-Thompson estimator and we propose different unbiased variance estimators. The possibility of negative values for the variance estimators leads us to propose simplified estimators and required conditions to their use.

Chapter 3 **Estimation under cross-classified sampling, continuation** is in continuity with the previous chapter. In this chapter, five unbiased estimators of Yates-Grundy type are available from five different possible decomposition of the variance. If these five estimators are identical in the case of simple random sampling without replacement, we sought to compare the five estimators in case of unequal probabilities. Conditional Poisson, Sampford and Midzuno sampling designs served as illustration from simulations.

Chapter 4 **Two-dimensional sampling in practice** is an article accepted for publication in *Case Studies In Business, Industry and Government Statistics* in 2016. It allows the reader to differentiate the cross-classified sampling design and the two-stage sampling design, and to practice in detail the steps of sampling and estimation under the R software, SAS and Stata for these two sampling designs. Two analogies are proposed : the first one between the two-stage sampling design and one-factor analysis of variance ; the second one between the cross-classified sampling design and two-factor analysis of variance.

Chapter 5 **Variance estimation under monotone non-response for a panel survey** is an article co-written with Guillaume Chauvet, submitted in 2016. It is devoted to the variance and the variance estimation for a panel survey with monotone non-response. The case of estimators with known response probabilities is treated as well as the one of estimators with estimated response probabilities. The general formulas are illustrated in the particular case of a model with response homogeneity groups.

Chapter 6 **Estimation de la variance pour l'enquête Elfe** is a technical and methodological report delivered in May 2016 to the ELFE unit and the ELFE data users. Its first version was published in 2016 in the INED Working Papers [10]. This document explains step by step the method used to estimate the variance of a total or a ratio, at the first time of the ELFE survey, taking into account the non-response and the calibration. An implementation with softwares R, SAS and Stata is also described.

[10] https://www.ined.fr/fr/publications/document-travail/estimation-variance-enquete-elfe/

Chapitre 1

Contexte

D'une population, on cherche à connaître certaines caractéristiques à un temps donné ou à en comparer les caractéristiques entre différents temps. A partir d'échantillons tirés aléatoirement, on peut estimer certains paramètres d'intérêt caractérisant cette population en leur associant une marge d'erreur ; on s'intéressera ici à l'inférence basée sur le plan de sondage. Le statisticien d'enquête définit un plan de sondage qualifiant la façon dont sera tiré l'échantillon et anticipant la précision des chiffres qui en découleront. Après ce premier échantillonnage d'unités, apparaît très souvent une seconde phase de non-réponse, correspondant à la non-participation d'unités initialement sollicitées. La taille et la structure de l'échantillon changent, impactant les biais et variances des estimateurs. Lorsque le paramètre d'intérêt mesure des changements dans le temps nécessitant des données récoltées à différentes dates, on s'intéresse à des plans de sondage construits en prenant en compte une dimension supplémentaire, temporelle. Ce peut être une même enquête répétée dans le temps sur des échantillons indépendants, ou les mêmes questions posées aux mêmes personnes à différents temps. Les estimations produites dépendent de la façon dont ont été sélectionnés ces échantillons dans le temps et sont aussi soumises au problème de non-réponse. On s'intéressera à l'enquête Elfe (Etude Longitudinale Française depuis l'Enfance), une enquête de type cohorte démarrée en 2011. Son plan de sondage initial est un plan particulier peu étudié dans la littérature, nommé plan d'échantillonnage produit (cross-classified sampling design).

Sommaire

1.1 Un temps

D'une population, on aimerait connaître certaines caractéristiques à un temps donné au travers de paramètres d'intérêt (par exemple la proportion de nourrissons nés sous césarienne dans la population des enfants nés en France métropolitaine en 2011). N'ayant pas accès à toute la population, ces caractéristiques vont être estimées à partir d'un sous-ensemble, c'est-à-dire qu'on va chercher à estimer ces paramètres sur un échantillon, et ceci avec une certaine marge d'erreur ; on parle d'inférence statistique. Dans ce document, on s'intéresse à l'inférence basée sur le plan de sondage, c'est-à-dire que le seul mécanisme aléatoire considéré est celui qui régit la sélection de l'échantillon, les caractéristiques de la population étant supposées fixes (à l'inverse de l'inférence basée sur le modèle, Särndal *et al.* [1992]).
C'est le statisticien d'enquête, en quête de précision, qui aura pour charge de définir un plan de sondage ainsi que les estimateurs ponctuels et par intervalles utilisés pour les paramètres d'intérêt de l'étude.

1.1.1 Plan de sondage

Soit une population finie U de taille N. Un échantillon S est tiré aléatoirement dans cette population en utilisant une loi de probabilité $p(\cdot)$, appelée plan de sondage, sur $\mathscr{S}$ l'ensemble des parties de U : $\forall s \in \mathscr{S}, p(s) \geq 0$ et $\sum_{\mathscr{S}} p(s) = 1$. Notons que si un plan de sondage est défini de façon unique par sa loi de probabilité, il peut exister plusieurs algorithmes de tirage respectant les propriétés d'un même plan de sondage. Suivant le plan choisi, la taille de l'échantillon $n(\mathrm{S})$ peut être aléatoire ou fixe. On considérera principalement dans ce document des plans de taille fixe, notée n, c'est-à-dire des plans dans lesquels seuls les échantillons de taille n ont une probabilité non nulle d'être tirés.
Pour chaque unité $k \in \mathrm{U}$, on peut définir une indicatrice d'appartenance à l'échan-

tillon, qui est une variable aléatoire :

$$\mathbf{1}_k = \begin{cases} 1 & \text{si } k \in \mathrm{S} \\ 0 & \text{sinon,} \end{cases}$$

ainsi que la probabilité d'inclusion du premier ordre, qui correspond à la probabilité qu'a l'unité k d'être dans l'échantillon S :

$$\pi_k = \mathrm{P}(\mathbf{1}_k = 1) = \mathbf{E}_p(\mathbf{1}_k) = \sum_{s \in \mathscr{S}, s \ni k} p(s)$$

où $\mathbf{E}_p$ est l'espérance par rapport à un plan de sondage p. On appelle probabilité d'inclusion jointe ou probabilité d'inclusion de second ordre, la probabilité pour deux unités distinctes k et l d'appartenir à l'échantillon S :

$$\pi_{kl} = \mathrm{P}(\mathbf{1}_k = 1 \cap \mathbf{1}_l = 1) = \mathbf{E}_p(\mathbf{1}_k \mathbf{1}_l) = \sum_{s \in \mathscr{S}, s \ni k,l} p(s).$$

On note une propriété supplémentaire que vérifie tout plan de sondage, utile au calcul de la variance des estimateurs des paramètres d'intérêt :

$$\Delta_{kl} \equiv \mathbf{Cov}_p(\mathbf{1}_k, \mathbf{1}_l) = \pi_{kl} - \pi_k \pi_l$$

où $\mathbf{Cov}_p$ est la covariance par rapport au plan de sondage p. Notons que deux plans de sondage peuvent présenter les mêmes probabilités d'inclusion de premier ordre et de second ordre sans pour autant être identiques.

Exemple du plan de Poisson

Dans une population U est tiré un échantillon S suivant un plan de Poisson, c'est-à-dire un plan dans lequel chaque unité est sélectionnée indépendamment des autres, selon une probabilité π_k :

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

La taille d'un plan de Poisson est aléatoire et ses probabilités d'inclusion de second ordre sont données par :

$$\pi_{kl} = \begin{cases} \pi_k \pi_l & \forall k \neq l \in \mathrm{U}, \\ \pi_k & \forall k = l. \end{cases}$$

En pratique, on a souvent recours aux plans stratifiés : la population U est partitionnée en H sous-ensembles U_h $(h = 1, \ldots, \mathrm{H})$ appelés strates, de tailles N_h. Dans chaque strate U_h est tiré un échantillon S_h de taille $n(\mathrm{S}_h)$ suivant un plan de sondage $p_h(\cdot)$ défini pour cette strate et indépendant des plans de sondage utilisés dans les autres strates.

Exemple du plan stratifié simple sans remise

Dans chaque strate U_h est tiré un échantillon S_h de taille n_h, fixée, suivant un plan de sondage aléatoire simple, c'est-à-dire un plan pour lequel tous les échantillons possibles de taille n_h sont équiprobables :

$$p_h(s_h) = \begin{cases} \binom{\mathrm{N}_h}{n_h}^{-1} & \text{si } n(s_h) = n_h, \\ 0 & \text{sinon.} \end{cases}$$

Les probabilités d'inclusion de premier et second ordres sont données par :

$$\pi_k^h = \frac{n_h}{\mathrm{N}_h} \ \forall k \in \mathrm{U}_h, \ \forall h = 1, \ldots, \mathrm{H},$$

$$\pi_{kl}^{hg} = \begin{cases} \frac{n_h(n_h-1)}{\mathrm{N}_h(\mathrm{N}_h-1)} & \forall (k,l) \in \mathrm{U}_h^2, \ k \neq l, \\ \frac{n_h}{\mathrm{N}_h} & \forall (k,l) \in \mathrm{U}_h^2, \ k = l, \\ \frac{n_h n_g}{\mathrm{N}_h \mathrm{N}_g} & \forall k \in \mathrm{U}_h, \forall l \in \mathrm{U}_g, \ h \neq g. \end{cases}$$

Les formules pour le plan simple sans remise peuvent être déduites en considérant une seule strate $(\mathrm{H} = 1)$.

Très souvent on a recours à plusieurs étapes d'échantillonnage avant d'obtenir un échantillon final. L'échantillonnage double consiste à tirer un premier échantillon, puis à se servir de cette première information (peu coûteuse) pour en tirer un second. Les deux échantillonnages peuvent être emboités (plan à deux phases) ou ne pas l'être [Hidiroglou, 2001]. Un cas particulier des plans à deux phases est le plan à deux degrés. Dans ce plan, la population $\tilde{\mathrm{U}}$ est partitionnée en N^1 sous-ensembles ($\tilde{\mathrm{U}} = \{u_1, \cdots, u_i, \cdots, u_{\mathrm{N}^1}\}$), appelés Unités Primaires (UP). Un premier plan de sondage est appliqué sur ces UP : on tire un échantillon S^1 de taille $n(\mathrm{S}^1)$. Ensuite à l'intérieur de chaque UP u_i de taille N_i sélectionnée au premier degré, on tire un échantillon S_i d'Unités dites Secondaires (US) de taille $n(\mathrm{S}_i)$ selon un plan de sondage $p_i(\cdot|\mathrm{S}^1)$. L'échantillon final correspond à la réunion de ces US : $\mathrm{S} = \cup_{u_i \in \mathrm{S}^1} \mathrm{S}_i$.

Exemple du plan classique à deux degrés

Pour ce plan, deux hypothèses sont classiquement utilisées :

- l'invariance : le plan de sondage dans une UP ne dépend pas de l'échantillon S^1 sélectionné au premier degré

$$\forall u_i \in \tilde{\mathrm{U}}, \; p_i(\cdot|\mathrm{S}^1) = p_i(\cdot).$$

- l'indépendance : conditionnellement à S^1, les échantillonnages dans chaque UP sont indépendants les uns des autres

$$\mathrm{P}\left(\bigcup_{u_i \in \mathrm{S}^1} \mathrm{S}_i|\mathrm{S}^1\right) = \prod_{u_i \in \mathrm{S}^1} \mathrm{P}(\mathrm{S}_i|\mathrm{S}^1).$$

Au premier degré, notons π_i^1 la probabilité d'inclusion de premier ordre d'une UP u_i de $\tilde{\mathrm{U}}$, π_{ij}^1 la probabilité de second ordre des UP i et j de $\tilde{\mathrm{U}}$ et Δ_{ij}^1 la covariance des indicatrices d'inclusion de i et j de U.
Au deuxième degré, notons $\pi_{k|i}$ la probabilité d'inclusion de premier ordre d'une US k de u_i, $\pi_{kl|i}$ la probabilité de second ordre des US k et l de u_i et $\Delta_{kl|i}$ la covariance des indicatrices d'inclusion de k et l de u_i.

1.1.2 Estimateurs

Soit y une variable d'intérêt prenant la valeur y_k pour l'individu k dans la population ($k = 1,\dots,\mathrm{N}$). On cherche à connaître un paramètre d'intérêt comme le total $t_y = \sum_{k\in \mathrm{U}} y_k$. L'estimateur de Horvitz-Thompson [Horvitz et Thompson, 1952] (aussi appelé π-estimateur ou estimateur par les valeurs dilatées ou estimateur par expansion) du total t_y est calculé à partir de l'échantillon S :

$$\hat{t}_{y\pi} = \sum_{k\in \mathrm{S}} \frac{y_k}{\pi_k}.$$

Cet estimateur est sans biais si tous les π_k sont strictement positifs (pour tout $k \in \mathrm{U}$). Si cet estimateur est espéré proche du total inconnu sur la population, on ne peut le vérifier à partir d'un seul échantillon mais seulement quantifier la précision de cet estimateur. On se tourne alors vers la mesure de dispersion qu'est la variance et qui va dépendre du plan de sondage choisi.

La variance de l'estimateur de Horvitz-Thompson du total t_y, si $\pi_k > 0$ pour tout $k \in \mathrm{U}$, est donnée par :

$$\mathrm{V}_{\mathrm{HT}}\left(\hat{t}_{y\pi}\right) = \sum_{k,l\in \mathrm{U}} \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

et peut se réécrire dans le cas d'un plan de taille fixe [Yates et Grundy, 1953], sous la forme :

$$\mathrm{V}_{\mathrm{YG}}\left(\hat{t}_{y\pi}\right) = -\frac{1}{2} \sum_{k\neq l\in \mathrm{U}} \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2.$$

Cette variance n'est pas calculable à partir d'un seul échantillon mais peut être estimée. L'estimateur de Horvitz-Thompson de la variance de $\hat{t}_{y\pi}$ s'écrit :

$$\hat{\mathrm{V}}_{\mathrm{HT}}\left(\hat{t}_{y\pi}\right) = \sum_{k,l\in \mathrm{S}} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

Il est sans biais si pour toutes les unités k et l de U, $\pi_{kl} > 0$. Sous les mêmes conditions, dans le cas d'un plan de taille fixe, l'estimateur de Sen-Yates-Grundy (Sen [1953], Yates et Grundy [1953])

$$\hat{\mathrm{V}}_{\mathrm{YG}}\left(\hat{t}_{y\pi}\right) = -\frac{1}{2}\sum_{k \neq l \in \mathrm{S}} \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2$$

estime sans biais cette variance. De plus, si le plan de sondage vérifie les conditions de Sen-Yates-Grundy, c'est-à-dire si tous les $\Delta_{kl} \leq 0$ pour $k \neq l \in \mathrm{U}$, alors cet estimateur est positif, propriété intéressante sachant qu'une valeur négative pour cet estimateur serait inutilisable.

Application au plan de Poisson

$$\begin{aligned} \mathrm{V}_{\mathrm{HT}}\left(\hat{t}_{y\pi}\right) &= \sum_{k\in\mathrm{U}} \pi_k\left(1-\pi_k\right)\left(\frac{y_k}{\pi_k}\right)^2, \\ \hat{\mathrm{V}}_{\mathrm{HT}}\left(\hat{t}_{y\pi}\right) &= \sum_{k\in\mathrm{S}} \left(1-\pi_k\right)\left(\frac{y_k}{\pi_k}\right)^2. \end{aligned}$$

Ce plan n'est pas un plan de taille fixe et l'estimateur de variance Yates-Grundy n'est donc pas sans biais.

Dans le cas d'un plan de sondage stratifié en H strates, le total estimé peut se réécrire comme la somme des totaux estimés sur chaque strate h :

$$\hat{t}_{y\pi} = \sum_{h=1}^{\mathrm{H}} \hat{t}_{h\pi} \quad \text{où} \quad \hat{t}_{h\pi} = \sum_{k\in\mathrm{S}_h} \frac{y_k}{\pi_k}.$$

De la même manière, grâce à l'indépendance des tirages entre les différentes strates, la variance (respectivement l'estimateur de variance) peut se réécrire comme la somme des variances (respectivement des estimateurs de variance) au sein de chaque strate :

$$\mathrm{V}_{\mathrm{HT}}\left(\hat{t}_{y\pi}\right)=\sum_{h=1}^{\mathrm{H}}\mathrm{V}_{\mathrm{HT}}\left(\hat{t}_{h\pi}\right)\ \text{ où }\ \mathrm{V}_{\mathrm{HT}}\left(\hat{t}_{h\pi}\right)=\sum_{k,l\in\mathrm{U}_h}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

$$\hat{\mathrm{V}}_{\mathrm{HT}}\left(\hat{t}_{y\pi}\right)=\sum_{h=1}^{\mathrm{H}}\hat{\mathrm{V}}_{\mathrm{HT}}\left(\hat{t}_{h\pi}\right)\ \text{ où }\ \hat{\mathrm{V}}_{\mathrm{HT}}\left(\hat{t}_{h\pi}\right)=\sum_{k,l\in\mathrm{S}_h}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

Application au plan stratifié aléatoire simple sans remise

$$\begin{aligned}
\hat{t}_{h\pi} &= \frac{\mathrm{N}_h}{n_h}\sum_{k\in s_h} y_k,\\
\mathrm{V}_{\mathrm{HT}}\left(\hat{t}_{h\pi}\right) &= \mathrm{N}_h^2\left(\frac{1}{n_h}-\frac{1}{\mathrm{N}_h}\right)\frac{1}{\mathrm{N}_h-1}\sum_{k\in\mathrm{U}_h}\left(y_k-\frac{1}{\mathrm{N}_h}\sum_{l\in\mathrm{U}_h}y_l\right)^2,\\
\hat{\mathrm{V}}_{\mathrm{HT}}\left(\hat{t}_{h\pi}\right) &= \mathrm{N}_h^2\left(\frac{1}{n_h}-\frac{1}{\mathrm{N}_h}\right)\frac{1}{n_h-1}\sum_{k\in\mathrm{S}_h}\left(y_k-\frac{1}{n_h}\sum_{l\in\mathrm{S}_h}y_l\right)^2.
\end{aligned}$$

Ce plan est de taille fixe. On peut montrer pour ce plan particulier que les estimateurs de variance de Horvitz-Thompson et de Yates-Grundy sont égaux.

Application au plan classique à deux degrés

Sous les hypothèses d'indépendance et d'invariance :

$$\begin{aligned}
\hat{t}_{y\pi} &= \sum_{i\in\mathrm{S}^1}\frac{1}{\pi_i^1}\hat{t}_{i\pi}\ \text{ avec }\ \hat{t}_{i\pi}=\sum_{k\in\mathrm{S}_i}\frac{y_k}{\pi_{k|i}},\\
\mathrm{V}_{\mathrm{HT}}\left(\hat{t}_{y\pi}\right) &= \mathrm{V}_{\mathrm{UP}}\left(\hat{t}_{\pi}\right)\ +\ \mathrm{V}_{\mathrm{US}}\left(\hat{t}_{\pi}\right),\\
\text{avec }\ \mathrm{V}_{\mathrm{UP}}\left(\hat{t}_{y\pi}\right) &= \sum_{u_i,u_j\in\tilde{\mathrm{U}}}\Delta_{ij}^1\frac{\hat{t}_{i\pi}}{\pi_i^1}\frac{\hat{t}_{j\pi}}{\pi_j^1}\\
\text{et }\ \mathrm{V}_{\mathrm{US}}\left(\hat{t}_{y\pi}\right) &= \sum_{u_i\in\tilde{\mathrm{U}}}\frac{1}{\pi_i^1}\left(\sum_{k,l\in u_i}\Delta_{kl|i}\frac{y_k}{\pi_{k|i}}\frac{y_l}{\pi_{l|i}}\right),
\end{aligned}$$

où $\mathrm{V}_{\mathrm{UP}}\left(\hat{t}_{\pi}\right)$ correspond à la variance due au premier degré de tirage et $\mathrm{V}_{\mathrm{US}}\left(\hat{t}_{\pi}\right)$ à celle due au second degré de tirage.

$$
\begin{aligned}
\hat{\mathrm{V}}_{\mathrm{HT}}\left(\hat{t}_\pi\right) &= \hat{\mathrm{V}}_{\mathrm{UP}}\left(\hat{t}_\pi\right) + \hat{\mathrm{V}}_{\mathrm{US}}\left(\hat{t}_\pi\right),\\
\text{avec } \hat{\mathrm{V}}_{\mathrm{UP}}\left(\hat{t}_\pi\right) &= \sum_{u_i,u_j\in \mathrm{S}^1} \frac{\Delta^1_{ij}}{\pi^1_{ij}} \frac{\hat{t}_{i\pi}}{\pi^1_i} \frac{\hat{t}_{j\pi}}{\pi^1_j} - \sum_{u_i\in \mathrm{S}^1} \frac{1-\pi^1_i}{\left(\pi^1_i\right)^2} \left(\sum_{k,l\in \mathrm{S}_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}} \right)\\
\text{et } \hat{\mathrm{V}}_{\mathrm{US}}\left(\hat{t}_\pi\right) &= \sum_{u_i\in \mathrm{S}^1} \frac{1}{\left(\pi^1_i\right)^2} \left(\sum_{k,l\in \mathrm{S}_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}} \right),
\end{aligned}
$$

où $\hat{\mathrm{V}}_{\mathrm{UP}}\left(\hat{t}_\pi\right)$ correspond à l'estimateur de la variance due au premier degré de tirage et $\hat{\mathrm{V}}_{\mathrm{US}}\left(\hat{t}_\pi\right)$ à celui de la variance due au second degré de tirage. Les détails de ces calculs sont donnés dans Särndal *et al.* [1992, pages 136-137].

1.2 Quasi deux-phases

La première étape d'échantillonnage est souvent suivie d'une phase de non-réponse, c'est-à-dire qu'une partie de l'échantillon initial refuse de répondre à l'enquête ou est injoignable, changeant taille et structure de l'échantillon (les répondants ont des caractéristiques différentes de celles des non-répondants). Ce processus en deux étapes (plan de sondage, mécanisme de non-réponse) peut être modélisé par un plan de sondage à deux phases, parfois appelé quasi deux-phases. En effet, par rapport à un véritable plan à deux phases [Särndal, Swensson, et Wretman, 1992], dans le cas de la non-réponse, les probabilités de seconde phase (conditionnellement à la première phase) ne sont pas connues mais estimées.

1.2.1 Mécanisme de non-réponse

Une première phase d'échantillonnage est pratiquée : on tire un échantillon S dans la population U selon un plan de sondage $p\,()$. L'échantillon des répondants S_r, un sous-ensemble de S, est généré à partir d'une loi de probabilité $q\left(|\mathrm{S}\right)$ appelée mécanisme de non-réponse.

Pour chacune des unités de l'échantillon initial $k \in \mathrm{S}$, on peut définir une indicatrice de réponse :

$$\mathbf{r}_k = \begin{cases} 1 & \text{si } k \in \mathrm{S}_r, \\ 0 & \text{sinon,} \end{cases}$$

ainsi que la probabilité p_k que l'unité sélectionnée durant la première phase appartienne à l'échantillon des répondant S_r :

$$p_k = \mathrm{P}\left(\mathbf{r}_k = 1|\mathrm{S}\right),$$

aussi appelée probabilité de réponse. En supposant que les unités répondent indépendamment les unes des autres, on peut ainsi écrire p_{kl} la probabilité de réponse jointe pour deux unités distinctes k et l de l'échantillon S :

$$p_{kl} = \mathrm{P}\left(\mathbf{r}_k = 1 \cap \mathbf{r}_l = 1|\mathrm{S}\right) = p_k p_l.$$

Sous cette hypothèse, le mécanisme $q\left(\cdot|\mathrm{S}\right)$ de non-réponse correspond à un plan de Poisson :

$$q\left(\mathrm{S}_r|\mathrm{S}\right) = \prod_{k \in \mathrm{S}} p_k^{\mathbf{r}_k} \left(1 - p_k\right)^{1-\mathbf{r}_k}.$$

Le plan de sondage final (sous les deux phases) $pq\left(\right)$ peut alors s'écrire

$$pq\left(\mathrm{S}_r\right) = \sum_{\mathrm{S} \supset \mathrm{S}_r} p\left(\mathrm{S}\right) q\left(\mathrm{S}_r|\mathrm{S}\right).$$

Ses probabilités d'inclusion de premier ordre pour une unité $k \in \mathrm{U}$, $\mathrm{P}\left(\mathbf{r}_k = 1\right)$ sont difficilement calculables (Särndal *et al.* [1992]). En particulier, $\neq \mathrm{P}\left(\mathbf{1}_k = 1\right) \mathrm{P}\left(\mathbf{r}_k = 1|\mathrm{S}\right)$.

1.2.2 Estimateurs

Si $\pi_k > 0$ pour toute unité k dans U et $p_k > 0$ pour toute unité k dans S, le total t_y peut être estimé sans biais par l'estimateur par expansion :

$$\hat{t}_{ye} = \sum_{k \in S_r} \frac{y_k}{\pi_k p_k}.$$

La variance de cet estimateur se décompose en deux parties :

$$\begin{aligned} V\left(\hat{t}_{ye}\right) &= V_p E_q\left(\hat{t}_{ye}|S\right) + E_p V_q\left(\hat{t}_{ye}|S\right) \\ &= \sum_{k,l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + E_p\left[\sum_{k \in S} \frac{1-p_k}{p_k}\left(\frac{y_k}{\pi_k}\right)^2\right] \end{aligned}$$

où E_q (respectivement V_q) correspond à l'espérance (respectivement la variance) sous le mécanisme de réponse. Cette variance peut être estimée sans biais par :

$$\hat{V}\left(\hat{t}_{ye}\right) = \sum_{k,l \in S_r} \frac{\Delta_{kl}}{\pi_{kl}} \frac{1}{p_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \sum_{k \in S_r} \frac{1-p_k}{\left(p_k\right)^2}\left(\frac{y_k}{\pi_k}\right)^2.$$

En pratique, les probabilités de réponse ne sont pas connues et il existe de nombreuses méthodes pour les estimer. Ces méthodes font partie des méthodes dites de repondération (voir Kalton et Flores-Cervantes [2003] ou Haziza et Lesage [2016]). La méthode la plus utilisée dans les enquêtes est la méthode des Groupes de Réponse Homogènes (GRH). Elle consiste à supposer que l'échantillon peut être partitionné en C groupes S_c $(c = 1, \ldots, C)$ au sein desquels la probabilité de réponse est constante. Ces groupes peuvent être obtenus à l'aide de la méthode des scores. Elle consiste à estimer sur S les probabilités de réponse selon un modèle (logistique en général). Après avoir trié S selon ces probabilités estimées, on constitue les C GRH, dans lesquels les répondants ont pour probabilité de réponse estimée la fréquence empirique de réponse dans le groupe. Ainsi, la probabilité estimée $\hat{p}_k$ pour une unité k appartenant à S_c est égale à

$$\hat{p}_k = \hat{p}_c = \frac{n_{rc}}{n_c}$$

en notant n_c le nombre d'unités dans S_c et n_{rc} le nombre d'unités répondantes dans S_c.

Le total t_y peut être estimé approximativement sans biais par l'estimateur ajusté de la non-réponse :

$$\hat{t}_{y\star} = \sum_{k\in \mathrm{S}_r} \frac{y_k}{\pi_k \hat{p}_k}.$$

En appliquant les résultats de Kim et Kim [2007] au cas de la méthode des scores, la variance de cet estimateur se décompose en deux parties :

$$\begin{aligned} \mathrm{V}\left(\hat{t}_{y\star}\right) &= \mathrm{V}_p\mathrm{E}_q\left(\hat{t}_{y\star}|s\right) + \mathrm{E}_p\mathrm{V}_q\left(\hat{t}_{y\star}|s\right) \\ &= \sum_{k,l\in \mathrm{U}} \varDelta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \mathrm{E}_p\left[\sum_{c=1}^{\mathrm{C}} \frac{1-p_c}{p_c} \sum_{k\in \mathrm{S}_c}\left(\frac{y_k}{\pi_k} - \frac{1}{n_c}\sum_{l\in \mathrm{S}_c}\frac{y_l}{\pi_l}\right)^2\right]. \end{aligned}$$

Cette variance peut être estimée approximativement sans biais par :

$$\hat{\mathrm{V}}\left(\hat{t}_{y\star}\right) = \sum_{k,l\in s_r} \frac{\varDelta_{kl}}{\pi_{kl}}\frac{1}{\hat{p}_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \sum_{c=1}^{\mathrm{C}} \frac{1-\hat{p}_c}{\left(\hat{p}_c\right)^2} \sum_{k\in \mathrm{S}_{rc}}\left(\frac{y_k}{\pi_k} - \frac{1}{n_{rc}}\sum_{l\in \mathrm{S}_{rc}}\frac{y_l}{\pi_l}\right)^2$$

où S_{rc} définit l'échantillon des répondants dans le groupe c. On remarque le terme de centrage qui apparaît dans cette formule, tenant compte du fait que les probabilités ne sont pas connues mais estimées.

1.3 Dans le temps

On peut s'intéresser à une population à un temps donné, mais aussi sur plusieurs temps pour mesurer des changements. La dimension temporelle permet d'observer des changements sur les caractéristiques d'une population et notamment les entrées et sorties d'unités de cette population. Dans un premier temps, on observera les différentes utilisations du mot longitudinal, puis on énumèrera les grandes familles de plans utilisés pour les enquêtes dans le temps. Enfin, on définira les différents processus de non-réponse dans le temps.

1.3.1 Longitudinal

L'adjectif *longitudinal* est beaucoup employé dans la littérature mais pas toujours de la même façon selon le contexte et l'auteur. Il y a la recherche longitudinale, l'approche longitudinale, les plans longitudinaux, les enquêtes longitudinales, les données longitudinales, les analyses longitudinales. On peut commencer par définir le longitudinal en le comparant au transversal. Dans une étude transversale, les données sont collectées à partir d'une ou plusieurs variables à un temps précis, pour une seule période. Pour une étude longitudinale, les données sont aussi collectées à partir d'une ou plusieurs variables, mais à plusieurs temps (deux périodes ou plus), permettant des mesures de changement [Menard, 2008]. Menard [2002] définit aussi la recherche longitudinale en fonction des données et des méthodes d'analyse utilisées :

> "Longitudinal research must be defined in terms of both the data and the methods of analysis used in the research. Longitudinal research is research in which
> (a) data are collected for each item or variable for two or more distinct time periods ;
> (b) the subjects or cases analysed are the same or at least comparable from one period to the next ; and
> (c) the analysis involves some comparison of data between or among periods.

> At a bare minimum, any truly longitudinal design would permit the measurement of differences or change in a variable from one period to another."

L'expression *enquête longitudinale* (*longitudinal survey*) est très souvent employée dans la littérature. Lynn [2009] la définit dans son premier chapitre *Methods for Longitudinal Surveys* de cette façon : "a longitudinal survey is one that collects data from the same sample elements on multiple occasions over time". C'est un terme ainsi souvent utilisé comme synonyme du terme *enquête par panel.*

Toutefois, plusieurs auteurs définissent les *enquêtes dans le temps* (*surveys over time*) sans utiliser le mot longitudinal pour qualifier l'enquête ou le plan de sondage ; le mot est seulement utilisé pour qualifier la mesure, l'analyse ou les données (par exemple, Kalton et Citro [1993]). Ou bien Kalton [2009] qui, dans son chapitre *Designs for surveys over time*, avertit que le terme longitudinal sera utilisé pour décrire les données que le plan génère (et non pas le plan lui-même). Dans la suite de ce document, le terme longitudinal sera peu utilisé.

1.3.2 Enquêtes dans le temps

Les enquêtes dans le temps répondent à deux principaux objectifs : 1) l'étude cherche à estimer un paramètre à un temps donné à un niveau agrégé et à le comparer à un autre temps, à calculer la variation nette entre les deux temps, ou la moyenne sur plusieurs temps, 2) l'étude s'intéresse, à un niveau individuel, à mesurer différentes composantes d'évolution (matrice de transition, flux, trajectoire...). Le plan de sondage de l'enquête se décide selon le ou les objectifs de l'enquête. Kalton [2009] distingue trois grandes familles de plans :

- **Les enquêtes répétées** (*repeated surveys*) : dans une enquête répétée ou enquête transversale répétée (*repeated cross-sectional survey*), on s'intéresse à des paramètres mesurés sur des échantillons différents tirés indépendamment d'une même population (qui peut elle-même avoir légèrement changé de par ses entrées et sorties). Si l'enquête est répétée à intervalles réguliers, on peut l'appeler enquête périodique (*periodic survey*) ou enquête continue (*continuing survey*). Ce type d'enquête satisfait le premier objectif.

- **Les enquêtes par panel** (*panel surveys*) : une enquête par panel ou panel "pur" (*"pure" panel*) ou panel fixe (*fixed panel*) est une enquête dans laquelle les mêmes mesures sont effectuées à différents temps sur le même échantillon. Ce type d'enquête se retrouve souvent dans la littérature sous le terme d'enquête longitudinale et satisfait le second objectif. La **cohorte** est un type particulier de panel : ses unités ont vécu un même événement, comme être né ou s'être marié durant la même année.
- **Les enquêtes par panel rotatif** (*rotating panel surveys*) : c'est un panel avec à chaque temps possibilité qu'un sous-échantillon d'unités sorte du panel et qu'un autre sous-échantillon entre pour le remplacer. Ce type d'enquête a pour avantage de répondre aux deux objectifs.

Afin de statisfaire différents objectifs, on peut trouver plusieurs variantes de ces plans combinés. Duncan et Kalton [1987] distinguent une quatrième famille de plans :

- **Les enquêtes à échantillon partagé** (*split panel surveys*) : il s'agit de la combinaison d'un panel avec une enquête répétée ou un panel rotatif.

Dans Kalton et Citro [1993], deux autres familles apparaissent en complément des quatres précédentes :

- **Les enquêtes chevauchantes** (*overlapping survey*) : ce sont des enquêtes répétées entre lesquelles on va volontairement chercher à avoir un taux de chevauchement important entre les différents échantillons.
- **Les enquêtes par panel répétées** (*repeated panel survey*) : il s'agit d'une série de panels, chacun démarrant et se terminant à des temps choisis (il peut y avoir plusieurs panels en cours à un temps donné).

On pourra aussi trouver dans Caron et Ravalet [2002] une description des quatres grandes premières familles et dans Deville [1998], une étude sur les enquêtes par panel. Dans la suite de ce document, on traitera le cas particulier des enquêtes par panel.

1.3.3 Processus de non-réponse dans le temps

A chaque temps d'enquête, seule une partie de l'échantillon sollicité répond à l'enquête. Dans le cas d'un panel, en observant les différents temps d'enquête (aussi appelés **vagues**), les sous-échantillons de répondants et de non-répondants varient d'un temps à l'autre. On distingue le cas des non-répondants à un temps donné et qui le restent aux temps suivants : ce phénomène par lequel les unités sortent du panel et ne reviennent jamais est appelé **attrition**. Il est difficile à quantifier avant que le panel soit complètement terminé (sauf pour les individus signifiant leur retrait de l'étude). La situation où les unités ne répondant pas à un temps donné sortent définitivement du panel correspond à un **processus de non-réponse monotone**. L'autre cas est la non-réponse **non-monotone** : certains individus cessent de participer à un ou plusieurs temps d'enquête, puis recommencent à répondre aux questionnaires. La possibilité pour ces unités, de sortir puis revenir, permet de récupérer des données et d'agrandir l'échantillon de travail, mais cela complique le traitement statistique des données : il y a des valeurs manquantes dans la base de données, faut-il imputer[1] ? Quelle pondération faut-il utiliser ? En Figure 1.1 sont illustrées toutes les combinaisons possibles de répondants et non-répondants pour trois temps d'enquête successifs.

Dans Kalton [1986] et Kalton, Lepkowski, Heeringa, Ting-Kwong, et Miller [1987] on pourra trouver des descriptions et discussions sur l'imputation et la repondération comme solutions à la non-réponse dans les enquêtes par panel. Dans la suite de ce document, on traitera le cas des enquêtes par panel avec processus de non-réponse

[1] Imputer signifie attribuer une valeur (plausible) lorsqu'une valeur est manquante.

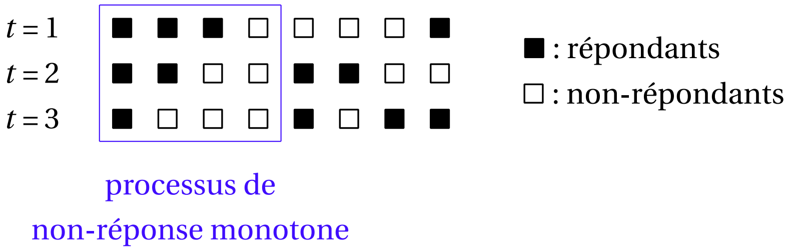


FIGURE 1.1 – Combinaisons de répondants et non-répondants pour trois temps d'enquête successifs

monotone.

1.4 Etude Longitudinale Française depuis l'Enfance

L'enquête Elfe[2] est une cohorte française comprenant plus de 18 000 enfants dont les parents ont consenti à leur inclusion en 2011. Elle est consacrée au suivi des enfants, de la naissance à l'âge adulte, et aborde les multiples aspects de la vie de l'enfant sous l'angle des sciences sociales, de la santé et de la santé-environnement (Pirus *et al.* [2010]). Cette étude est originale de par notamment sa pluridisciplinarité, la participation des deux parents, mais aussi son plan de sondage.

1.4.1 Présentation

Après avoir été testée sur une enquête pilote de 500 familles en 2007, l'enquête Elfe a été généralisée à partir d'avril 2011 en France métropolitaine sur un échantillon de 18 321 enfants. Elfe a reçu le soutien des ministères en charge des affaires sociales et de la santé, de l'enseignement supérieur et de la recherche, celui du ministère en charge de l'écologie et du développement durable ainsi que celui d'un ensemble d'organismes de recherche ou institutions. Elfe est notamment associée à l'Ined[3],

[2] http://www.elfe-france.fr/index.php/fr/
[3] Institut national d'études démographiques

l'Inserm[4], l'Invs[5], l'Insee[6], la DGPR[7], la Drees[8], la DGS[9], la DEPS[10] et la Cnaf[11]. Environ quatre-vingt équipes de recherche travaillent sur le projet Elfe. Les objectifs de l'étude varient selon les chercheurs (Figure 1.2) : par exemple, Wagner *et al.* [2015] traitent de la "Durée de l'allaitement en France selon les caractéristiques des parents et de la naissance", Panico *et al.* [2015] étudient "La fréquence des naissances de petit poids : quelle influence a le niveau d'instruction des mères ?"(Informations extraites du site Elfe [2016]).

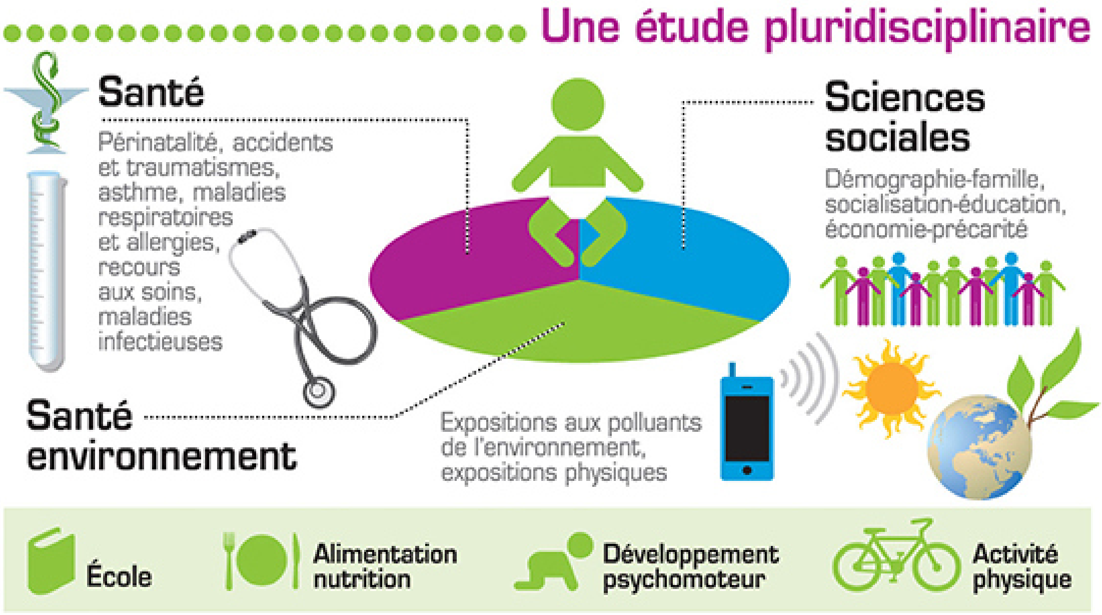

Une étude pluridisciplinaire
Santé
Périnatalité, accidents et traumatismes, asthme, maladies respiratoires et allergies, recours aux soins, maladies infectieuses
Sciences sociales
Démographie-famille, socialisation-éducation, économie-précarité
Santé environnement
Expositions aux polluants de l'environnement, expositions physiques
École
Alimentation nutrition
Développement psychomoteur
Activité physique

FIGURE 1.2 – Elfe, une enquête pluridisciplinaire (infographie Elfe)

Concernant le déroulement de l'enquête, quelques jours après la naissance de l'enfant, les familles ont été contactées à la maternité pour un premier suivi. Sur un sous-

[4] Institut national de la santé et de la recherche médicale
[5] Institut national de veille sanitaire
[6] Institut national de la statistique et des études économiques
[7] Direction générale de la prévention des risques
[8] Direction de la recherche, des études, de l'évaluation et des statistiques
[9] Direction générale de la santé
[10] Département des études de la prospective et des statistiques
[11] Caisse nationale des allocations familiales

échantillon de mères et nourrissons ont été réalisés des prélèvements biologiques (sang veineux, sang de cordon, lait maternel...). Ensuite, à deux mois, les parents ont été contactés pour un entretien téléphonique : conditions de vie, alimentation, environnement et contexte familial ont été évoqués. Entre 2 et 10 mois, un questionnaire concernant l'alimentation du nourrisson a été envoyé aux familles ainsi qu'un piège à poussière sur un sous-échantillon. De nouvelles sollicitations téléphoniques ont eu lieu aux 1 an, 2 ans, 3,5 ans de l'enfant, permettant de noter les changements dans la composition familiale, la profession des parents, le logement, le développement de l'enfant, le suivi médical... (Figure 1.3) (Informations extraites du site Elfe [2016]).

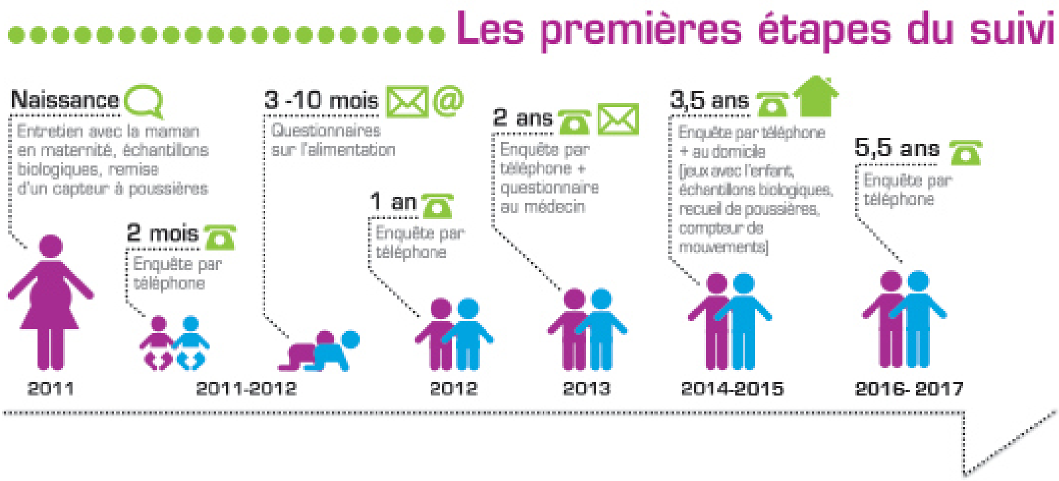

Les premières étapes du suivi
Naissance
Entretien avec la maman en maternité, échantillons biologiques, remise d'un capteur à poussières
2011
2 mois
Enquête par téléphone
3 -10 mois
Questionnaires sur l'alimentation
2011-2012
1 an
Enquête par téléphone
2012
2 ans
Enquête par téléphone + questionnaire au médecin
2013
3,5 ans
Enquête par téléphone + au domicile (jeux avec l'enfant, échantillons biologiques, recueil de poussières, compteur de mouvements)
2014-2015
5,5 ans
Enquête par téléphone
2016- 2017

FIGURE 1.3 – Grandes étapes pour la cohorte Elfe (infographie Elfe)

1.4.2 Plan de sondage

La population d'inférence est fixée au premier temps de l'enquête, c'est celle des nourrissons nés durant l'année 2011 en France métropolitaine, issus d'un accouchement au plus gémellaire, hors grands prématurés, ayant une mère majeure, en mesure de donner un consentement éclairé notamment dans l'une des langues proposées (français, anglais, arabe ou turc), nés dans une maternité métropolitaine et dont les parents ne résidaient pas temporairement en métropole. Toutes les familles sélectionnées ont été enquêtées peu de temps après l'accouchement dans certaines maternités métropolitaines et durant certains jours de l'année.

Les nourrissons inclus dans la cohorte sont issus de deux échantillonnages : leur date de naissance fait partie d'un échantillon de jours de l'année 2011, et leur lieu de naissance appartient à un échantillon de maternités en France métropolitaine. L'échantillon des jours étant le même pour chaque maternité sélectionnée (ou vice versa, l'échantillon des maternités étant le même pour chaque jour sélectionné), on ne peut considérer ce plan comme un tirage à deux degrés classique, c'est-à-dire vérifiant l'hypothèse standard d'indépendance entre les tirages d'unités secondaires relatifs à chaque unité primaire. L'échantillon final se forme au croisement de lieux sélectionnés et de temps choisis : il résulte du produit de deux échantillonnages. Le plan de sondage pour les maternités est un plan probabiliste, 320 maternités ont été sélectionnées. Concernant les jours, 25 ont été choisis durant quatre périodes afin de couvrir les quatre saisons de l'année. Le plan de sondage final utilisé pour l'enquête Elfe résulte du croisement indépendant de ces deux plans de sondage et est appelé *plan d'échantillonnage produit* (ou encore *cross-classified sampling design* dans Ohlsson, 1996).

1.5 Echantillonnage produit

La population dans laquelle est tiré l'échantillon Elfe est obtenue par le croisement de la population des maternités de France métropolitaine et de celle des jours de l'année 2011, les unités de cette population sont des couples (maternité,jour). Plusieurs plans d'échantillonnage sont possibles dans une population produit, dont l'échantillonnage produit pour lequel on fera une revue de la littérature en seconde partie.

1.5.1 Echantillonnage dans une population produit

L'échantillonnage dans une population à deux dimensions ou population produit est discuté dans la littérature dans différents contextes : l'échantillonnage spatial avec longitudes et latitudes, l'échantillonnage dans le temps et l'espace dans Vos [1964], l'échantillonnage de colonnes et lignes dans Bellhouse [1981] ou Ohlsson [1996], l'échantillonnage de points de vente et d'articles pour l'indice des prix à la consommation dans Dalén et Ohlsson [1995]. Dans la Figure 1.4, nous illustrons le croisement produit d'une population de lignes U_L de taille 13 avec une population de colonnes U_C de taille 13 : 169 unités finales sont représentées.

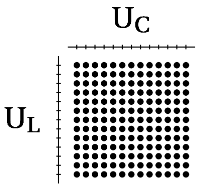


FIGURE 1.4 – Population produit $U_L \times U_C$

Les objectifs de l'enquête peuvent être différents : capter une variabilité en considérant une des populations ou les deux. Les variations périodiques, les mouvements saisonniers, peuvent être importants : par exemple, dans Vos [1964], pour estimer les performances de transport durant une année entière, il est possible d'échantillonner des conducteurs et des semaines. En pratique, le plan de sondage est contraint au niveau logistique et administratif.

Suivant Vos [1964], différentes méthodes d'échantillonnage peuvent être utilisées :

- **Echantillonnage direct** dans la population produit : un échantillon de couples (r, c) est tiré directement dans la population $U_L \times U_C$. Par exemple dans la Figure 1.5, un échantillon de 12 unités est sélectionné.
- **Echantillonnage classique à deux degrés** qui consiste à tirer un échantillon d'uni-

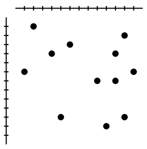


FIGURE 1.5 – Echantillonnage direct dans la population produit

tés primaires, et ensuite un second échantillonnage est effectué indépendamment dans chaque unité primaire. La Figure 1.6 (resp. 1.7) illustre le cas avec les colonnes (resp. lignes) comme unités primaires. Un échantillon final de 12 unités est obtenu en tirant un échantillon d'unités secondaires dans les unités primaires sélectionnées.

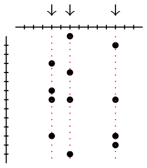


FIGURE 1.6 – Echantillonnage à deux degrés avec les colonnes comme unités primaires

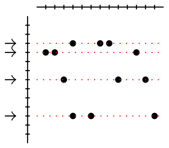


FIGURE 1.7 – Echantillonnage à deux degrés avec les lignes comme unités primaires

- **Echantillonnage produit** : dans ce cas, la précédente propriété d'indépendance entre les tirages est supprimée. Un échantillon de lignes S_L et un échantillon de colonnes S_C sont tirés indépendamment et tous les couples (l,c) in $S_L \times S_C$ sont sélectionnés. Dans la Figure 1.8, un échantillon final de 12 unités est obtenu en utilisant un échantillonnage produit.

- **Sous-échantillonnage dans un échantillon produit** : à la première étape, un échantillon produit est tiré, et à la seconde étape, un échantillon de couples (r, c) est sélectionné dans l'échantillon produit $S_L \times S_C$. Beaucoup de procédures sont possibles

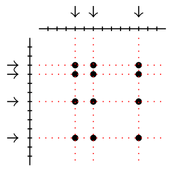


FIGURE 1.8 – Echantillonnage produit

pour cette seconde étape. Dans la Figure 1.9 est illustré un cas particulier appelé *double two stage selection* dans Vos [1964]. C'est un sous-échantillonnage sélectionnant à la seconde étape le même nombre de lignes sur chaque colonne sélectionnée à la première étape, et le même nombre de colonnes pour chaque ligne sélectionnée à la première étape.

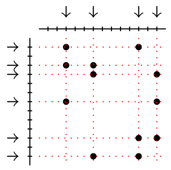


FIGURE 1.9 – Sous-échantillonnage dans un échantillon produit

Parmi tous ces plans, nous allons étudier plus spécifiquement le plan d'échantillonnage produit, qui correspond au plan de sondage utilisé pour l'étude Elfe.

1.5.2 Echantillonnage produit

Dans la littérature, les méthodes présentées dans la section précédente ont été étudiées par Vos [1964], afin de trouver le plan d'échantillonnage optimum dans une population produit, en se basant sur la variance de l'estimateur de la moyenne. Les formules de variance pour un plan d'échantillonnage produit sont proposées pour le croisement de deux plans aléatoires simples sans remise.

Dans Bellhouse [1981], échantillonnage direct, échantillonnage à deux degrés et échantillonnage produit sont discutés dans un contexte de plans d'échantillonnage spatial avec tendance linéaire, variations périodique et populations spatialement corrélées. Les formules de variance de ces trois méthodes sont proposées pour trois tirages : échantillonnage aléatoire simple, échantillonnage systématique et échantillonnage stratifié avec une seule unité par strate. L'échantillonnage produit apparaît sous le nom “cartesian product".

Dans Ohlsson [1996], l'échantillonnage produit est défini et une formule générale de variance est dérivée, illustrée dans le cas du produit de deux plans simples sans remise et dans le cas du produit de deux plans de Poisson. En utilisant l'indépendance de ces deux plans, le total est décomposé en trois termes indépendants qui conduiront à une décomposition de la variance en trois termes. Une formule de variance est aussi proposée dans le cas du produit de deux plans stratifiés simples. Aucun estimateur de variance n'est proposé.

Dans Dalén et Ohlsson [1995], un échantillonnage produit a été appliqué à l'enquête suédoise concernant l'indice des prix à la consommation : c'est le produit de deux plans stratifié, avec un plan systématique pour les produits et un plan de Poisson séquentiel pour les points de vente. Une formule de variance est donnée et un estimateur de variance simplifié est proposé spécifiquement pour l'enquête concernée.

Dans Skinner [2015], apparaissent les formules d'un estimateur de variance pour un plan produit avec deux plans simples et pour un plan produit avec deux plans stratifiés simples. Ces résultats sont donnés en utilisant l'approche basée sur le plan de sondage et celle basée sur le modèle. Le cas de deux plans à probabilités inégales avec remise est aussi traité. Un estimateur de variance de type bootstrap est proposé dans le cas du plan stratifié simple.

References

D.R. BELLHOUSE : Spatial sampling in the presence of a trend. *Journal of Statistical Planning and Inference*, 5:365–375, 1981. 32, 34

N. CARON et P. RAVALET : Estimation dans les enquêtes répétées : application à l'enquête emploi en continu. *Insee-Méthodes : Actes des Journées de Méthodologie Statistique*, 100:327–392, 2002. 27

J. DALÉN et E. OHLSSON : Variance estimation in the Swedish consumer price index. *Journal of Business & Economic Statistics*, 13(3):347–356, 1995. 32, 35

J.C. DEVILLE : Les enquêtes par panel : en quoi diffèrent-elles des autres enquêtes ? suivi de : comment attraper une population en se servant d'une autre. *Insee-Méthodes : Actes des Journées de Méthodologie Statistique*, 84-85-86:63–82, 1998. 27

G. DUNCAN et G. KALTON : Issues of design and analysis of surveys across time. *International Statistical Review*, 55:97–117, 1987. 26

ELFE : *Grandes étapes [en ligne], rédactrice en chef : L. Gravier*, 2016. URL `http://www.elfe-france.fr/index.php/fr/comment-ca-marche/grandes-etapes`. 29, 30

D. HAZIZA et E. LESAGE : A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32(1):129–145, 2016. 22

M.A. HIDIROGLOU : Double sampling. *Survey Methodology*, 27:143–154, 2001. 16

D. HORVITZ et D. THOMPSON : A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952. 17

G. KALTON : Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2:303–314, 1986. 27

G. KALTON : Designs for survey over time. *Handbook of Statistics*, 29A:89–108, 2009. 25

G. KALTON et C. CITRO : Panel surveys : Adding the fourth dimension. *Survey Methodology*, 19:205–215, 1993. 25, 26

G. KALTON et I. FLORES-CERVANTES : Weighting methods. *Journal of Official Statistics*, 19 (2):81–97, 2003. 22

G. KALTON, J. LEPKOWSKI, S. HEERINGA, L. TING-KWONG et M.E. MILLER : The treatment of person-wave nonresponse in longitudinal surveys. *Survey Research Center, University of Michigan*, 1987. 27

J. K. KIM et J. J. KIM : Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35:501–514, 2007. 23

P. LYNN : *Methodology of Longitudinal Surveys*. John Wiley & Sons, Ltd, 2009. 25

S. MENARD : *Longitudinal Research*, volume 76. SAGE Publications, 2002. 24

S. MENARD : *Handbook of Longitudinal Research*. Elsevier/Academic Pres, 2008. 24

E. OHLSSON : Cross-classified sampling. *Journal of Official Statistics*, 12(3):241–251, 1996. 31, 32, 35

L. PANICO, M. TÔ et O. THÉVENON : La fréquence des naissances de petit poids : quelle influence a le niveau d'instruction des mères ? *Population et Sociétés*, 523, 2015. 29

C. PIRUS, C. BOIS, M.N. DUFOURG, J.L. LANOË, S. VANDENTORREN, H. LERIDON et the ELFE TEAM : Constructing a cohort : Experience with the French Elfe project. *Population*, 65:637–670, 2010. 28

C.-E. SÄRNDAL, B. SWENSSON et J.H. WRETMAN : *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992. 13, 20, 21

A. SEN : On the estimate of the variance in sampling with varying probabilities. *Journal of Indian Society for Agricultural Statistics*, 5:119–127, 1953. 18

C.J. SKINNER : Cross-classified sampling : some estimation theory. *Statistics and Probability Letters*, 104:163–168, 2015. 35

J.W.E. VOS : Sampling in space and time. *Review of the International Statistical Institute*, 32 (3):226–241, 1964. 32, 34

S. WAGNER, C. KERSUZAN, S. GOJARD, C. TICHIT, S. NICKLAUS, B. GEAY, P. HUMEAU, X. THIERRY, M.-A. CHARLES, S. LIORET et B. LAUZON-GUILLAIN : Durée de l'allaitement en France selon les caractéristiques des parents et de la naissance. Résultats de l'étude longitudinale française Elfe, 2011. *BEH*, 29:522–532, 2015. 29

F. YATES et P.M. GRUNDY : Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, 15:235–261, 1953. 17, 18

Chapitre 2

Estimation under cross-classified sampling with application to a childhood survey

This chapter is a reprint of Juillard, H., Chauvet, G. and Ruiz-Gazen, A. (2016). Estimation under cross-classified sampling with application to a childhood survey. *To appear in Journal of American Statistical Association.*

Le plan d'échantillonnage produit consiste à tirer des échantillons, à partir d'une population à deux dimensions, indépendamment dans chaque dimension. Un tel plan est souvent utilisé dans les enquêtes d'indice de prix à la consommation et a été récemment utilisé pour tirer un échantillon de nourrissons dans la cohorte Elfe (Etude Longitudinale Française depuis l'Enfance) en croisant un échantillon de maternités et un échantillon de jours. Nous proposons une théorie générale d'estimation pour un tel plan de sondage. Nous considérons l'estimateur de Horvitz-Thompson pour un total, et montrons que le plan produit, de façon générale, est moins efficace que le plan classique à deux degrés d'échantillonnage. Nous obtenons la distribution asymptotique de l'estimateur de Horvitz-Thompson, ainsi que plusieurs estimateurs de variance sans biais. Face au problème de possibles valeurs négatives, nous proposons des estimateurs de variance simplifiés positifs et étudions leurs biais sous un modèle de super-population. Les estimateurs proposés sont comparés pour des totaux et des ratios sur des données simulées. Une application aux données de Elfe est aussi présentée, et nous proposons certaines recommandations. Des matériels supplémentaires sont disponibles en ligne.

Keywords *: analyse de la variance, échantillonnage à deux degrés, estimateur de Horvitz-Thompson, estimateur de Sen-Yates-Grundy, indépendance, invariance.*

Sommaire

Estimation under cross-classified sampling with application to a childhood survey

The cross-classified sampling design consists in drawing samples from a two-dimension population, independently in each dimension. Such design is commonly used in consumer price index surveys and has been recently applied to draw a sample of babies in the French Longitudinal Survey on Childhood, by crossing a sample of maternity units and a sample of days. We propose to derive a general theory of estimation for this sampling design. We consider the Horvitz-Thompson estimator for a total, and show that the cross-classified design will usually result in a loss of efficiency as compared to the widespread two-stage design. We obtain the asymptotic distribution of the Horvitz-Thompson estimator, and several unbiased variance estimators. Facing the problem of possibly negative values, we propose simplified non-negative variance estimators and study their bias under a super-population model. The proposed estimators are compared for totals and ratios on simulated data. An application on real data from the French Longitudinal Survey on Childhood is also presented, and we make some recommendations. Supplementary materials are available online.

***Mots-clés** : ANOVA, Horvitz-Thompson estimator, Sen-Yates-Grundy estimator, independence, invariance, two-stage sampling.*

2.1 Introduction

The 2011 French Longitudinal Survey on Childhood ELFE (Etude Longitudinale Française depuis l'Enfance) comprises more than 18,000 children selected on the basis of their place and date of birth. On the one hand, a sample of 320 maternity units has been drawn. On the other hand, a sample of 25 days divided in four time periods and spread across the four seasons of 2011 has been selected. The babies born at the sampled locations and on the sampled days have been approached through midwives. Data were collected on babies whose parents consented to their inclusion during their stay at the maternity unit. ELFE is conducted by the National Institute for Demographic Studies, the National Institute for Health and Medical Research and the French Blood Agency. The objective of observing children born within the same year is to analyze their physical and psychological health together with their living and environmental conditions. This large-scale study of children's development and socialization is the first of its kind in France. The collected data are now available to public and private research teams and many projects are underway in areas such as health, health environment and social sciences. In order to derive reliable confidence intervals for finite population parameters such as totals or ratios, the ELFE sampling design has to be taken into account.

The ELFE sample is drawn according to a non-standard sampling design, called Cross-Classified Sampling (CCS), following Ohlsson [1996]. It consists in drawing independently two samples from each component of a two-dimensional population. In the ELFE survey, a sample of maternity units and a sample of days are independently selected. This sampling design appears in other contexts than the ELFE survey. Some examples include consumer price index surveys, as detailed in Dalén and Ohlsson [1995] for the Swedish survey, where outlets and items are sampled, and business surveys (Skinner [2015]), where businesses and products are sampled. Due to its particular properties, CCS deserves a specific attention. However, as noted by Skinner [2015], "the literature on the theory of cross-classified sampling is very limited". In particular, no general theory is derived under the finite population framework. While the papers by Vos [1964] and Ohlsson [1996] focus on simple random sampling without replacement, Skinner [2015] gives some results under stratified without replacement simple random sampling and under with replacement unequal probability sampling. Dalén and Ohlsson [1995] provide some results under probability proportional to size without-replacement sampling.

In the present paper, we develop a general theory for estimation and variance estimation under CCS. The asymptotic normality of the Horvitz-Thompson estimator is derived under some mild conditions. A comparison with a two-stage sampling design is carried out in a general framework. We also raise an issue, not reported before, of possible negative values for Horvitz-Thompson and Yates-Grundy variance estimates. This problem occurs even in the simplest case of simple random sampling without replacement. Non-negative simplified variance estimators are therefore introduced. Conditions for their approximate unbiasedness are given under a design-based and a model-based approach. The properties of our variance estimators are evaluated through a small but realistic simulation study when estimating totals and ratios. Finally, an application to the ELFE data is detailed.

2.2 Cross-classified sampling design

2.2.1 Notations and Horvitz-Thompson estimation

Keeping in mind the ELFE survey, we consider a population U_M of N_M maternity units and a population U_D of N_D days. However, the developments below are completely general and may be applied to any populations U_M and U_D. We will use the indexes i and j for the maternity units, and the indexes k and l for the days. We consider a sampling design $p_M(\cdot)$ on the population U_M, leading to a sample S_M of (average) size n_M, and a sampling design $p_D(\cdot)$ on the population U_D leading to a sample S_D of (average) size n_D. We assume that the two samples are selected independently. The cross-classified sampling design $p(\cdot)$ on the product population $U = U_M \times U_D$ is therefore defined as

$$p(s) = p_M(s_M) \times p_D(s_D) \quad \text{for any} \quad s = s_M \times s_D \subset U_M \times U_D.$$

Let π_i^M denote the probability that i is selected in S_M, π_{ij}^M denote the probability that units i and j are selected jointly in S_M, and let $\Delta_{ij}^M = \pi_{ij}^M - \pi_i^M \pi_j^M$. The quantities π_k^D, π_{kl}^D and Δ_{kl}^D are similarly defined. We assume that the first and second-order inclusion probabilities are non-negative in each population. The probability for the pairs (i,k) to be selected in the product sample $S_M \times S_D$ is $\pi_i^M \pi_k^D$, and the probability for the pairs (i,k) and (j,l) to be selected jointly in the product sample $S_M \times S_D$ is $\pi_{ij}^M \pi_{kl}^D$.

We are interested in some variable of interest with value Y_{ik} for the maternity unit i and the day k. The total $t_{\mathrm{Y}} = \sum_{i \in \mathrm{U_M}} \sum_{k \in \mathrm{U_D}} \mathrm{Y}_{ik}$ is then unbiasedly estimated by the Horvitz-Thompson (HT) estimator

$$\hat{t}_{\mathrm{Y}} = \sum_{i \in \mathrm{S_M}} \sum_{k \in \mathrm{S_D}} \frac{\mathrm{Y}_{ik}}{\pi_i^{\mathrm{M}} \pi_k^{\mathrm{D}}} = \sum_{i \in \mathrm{S_M}} \sum_{k \in \mathrm{S_D}} \check{\mathrm{Y}}_{ik} \text{ where } \check{\mathrm{Y}}_{ik} = \frac{\mathrm{Y}_{ik}}{\pi_i^{\mathrm{M}} \pi_k^{\mathrm{D}}}. \tag{2.2.1}$$

Making use of the independence between $\mathrm{S_M}$ and $\mathrm{S_D}$, the variance of the HT-estimator is

$$\mathrm{V_{CCS}}\left(\hat{t}_{\mathrm{Y}}\right) = \sum_{i,j \in \mathrm{U_M}} \sum_{k,l \in \mathrm{U_D}} \Gamma_{ijkl} \check{\mathrm{Y}}_{ik} \check{\mathrm{Y}}_{jl} \tag{2.2.2}$$

where $\Gamma_{ijkl} = \pi_{ij}^{\mathrm{M}} \pi_{kl}^{\mathrm{D}} - \pi_i^{\mathrm{M}} \pi_j^{\mathrm{M}} \pi_k^{\mathrm{D}} \pi_l^{\mathrm{D}}$. The Sen [1953]-Yates and Grundy [1953] form

$$\mathrm{V_{CCS}}\left(\hat{t}_{\mathrm{Y}}\right) = -\frac{1}{2} \sum_{(i,k) \neq (j,l) \in \mathrm{U_M} \times \mathrm{U_D}} \Gamma_{ijkl} \left(\check{\mathrm{Y}}_{ik} - \check{\mathrm{Y}}_{jl}\right)^2 \tag{2.2.3}$$

can be used alternatively when both sampling designs are of fixed size.

Our set-up can be compared to the usual two-stage framework, by considering $\mathrm{U_M}$ as a population of Primary Sampling Units (PSUs) and $\mathrm{U_D}$ as a population of Secondary Sampling Units (SSUs), each maternity unit i being associated to the same population of days. In case of two-stage sampling, denoted by MD, a first-stage sample $\mathrm{S_M}$ is selected in $\mathrm{U_M}$, and some second-stage samples S_i are selected independently using $p_{\mathrm{D}}(\mathrm{S}_i)$ for each $i \in \mathrm{S_M}$ (see Särndal et al. [1992]). The variance of the HT-estimator is then

$$\mathrm{V_{MD}}\left(\hat{t}_{\mathrm{Y}}\right) = \mathrm{V_{MD}^{PSU}}\left(\hat{t}_{\mathrm{Y}}\right) + \mathrm{V_{MD}^{SSU}}\left(\hat{t}_{\mathrm{Y}}\right) \tag{2.2.4}$$

where

$$\begin{aligned}
\mathrm{V}^{\mathrm{PSU}}_{\mathrm{MD}}\left(\hat{t}_{\mathrm{Y}}\right) &= \sum_{i,j\in\mathrm{U_M}}\sum_{k,l\in\mathrm{U_D}} \Delta^{\mathrm{M}}_{ij}\pi^{\mathrm{D}}_{k}\pi^{\mathrm{D}}_{l}\check{\mathrm{Y}}_{ik}\check{\mathrm{Y}}_{jl}, &(2.2.5)\\
\mathrm{V}^{\mathrm{SSU}}_{\mathrm{MD}}\left(\hat{t}_{\mathrm{Y}}\right) &= \sum_{i\in\mathrm{U_M}}\sum_{k,l\in\mathrm{U_D}} \pi^{\mathrm{M}}_{i}\Delta^{\mathrm{D}}_{kl}\check{\mathrm{Y}}_{ik}\check{\mathrm{Y}}_{il}. &(2.2.6)
\end{aligned}$$

Alternatively, we could consider $\mathrm{U_D}$ as a population of PSUs and $\mathrm{U_M}$ as a population of SSUs, each day k being associated to the same population of maternity units. In this case, the variance of the HT-estimator under two-stage sampling is

$$\mathrm{V}_{\mathrm{DM}}\left(\hat{t}_{\mathrm{Y}}\right) = \mathrm{V}^{\mathrm{PSU}}_{\mathrm{DM}}\left(\hat{t}_{\mathrm{Y}}\right) + \mathrm{V}^{\mathrm{SSU}}_{\mathrm{DM}}\left(\hat{t}_{\mathrm{Y}}\right) \qquad (2.2.7)$$

where

$$\begin{aligned}
\mathrm{V}^{\mathrm{PSU}}_{\mathrm{DM}}\left(\hat{t}_{\mathrm{Y}}\right) &= \sum_{k,l\in\mathrm{U_D}}\sum_{i,j\in\mathrm{U_M}} \Delta^{\mathrm{D}}_{kl}\pi^{\mathrm{M}}_{i}\pi^{\mathrm{M}}_{j}\check{\mathrm{Y}}_{ik}\check{\mathrm{Y}}_{jl}, &(2.2.8)\\
\mathrm{V}^{\mathrm{SSU}}_{\mathrm{DM}}\left(\hat{t}_{\mathrm{Y}}\right) &= \sum_{k\in\mathrm{U_D}}\sum_{i,j\in\mathrm{U_M}} \pi^{\mathrm{D}}_{k}\Delta^{\mathrm{M}}_{ij}\check{\mathrm{Y}}_{ik}\check{\mathrm{Y}}_{il}. &(2.2.9)
\end{aligned}$$

The different features of CCS and two-stage sampling on a two-dimension population are illustrated on Figure 2.1.

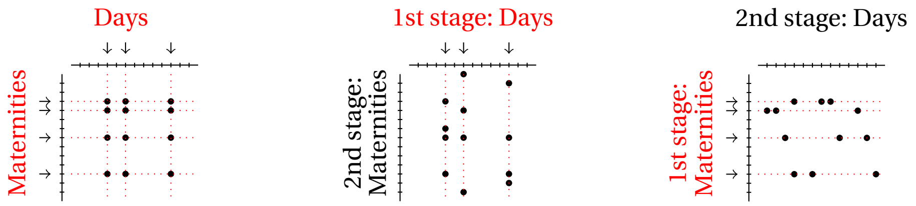


FIGURE 2.1 – Cross-classified sampling (left panel), two-stage sampling DM with primary units in $\mathrm{U_D}$ (central panel), two-stage sampling MD with primary units in $\mathrm{U_M}$ (right panel)

2.2.2 Variance decomposition for cross-classified sampling

The covariance Γ_{ijkl} may be written in several ways, leading to alternative variance decompositions. Plugging $\Gamma_{ijkl} = \pi^{\mathrm{D}}_{kl}\Delta^{\mathrm{M}}_{ij} + \pi^{\mathrm{M}}_{ij}\Delta^{\mathrm{D}}_{kl} - \Delta^{\mathrm{M}}_{ij}\Delta^{\mathrm{D}}_{kl}$ into (2.2.2) gives

$$\mathrm{V}_{\mathrm{CCS}}\left(\hat{t}_{\mathrm{Y}}\right) \quad = \quad \mathrm{V}_1\left(\hat{t}_{\mathrm{Y}}\right) + \mathrm{V}_2\left(\hat{t}_{\mathrm{Y}}\right) - \mathrm{V}_3\left(\hat{t}_{\mathrm{Y}}\right) \tag{2.2.10}$$

where

$$\mathrm{V}_1\left(\hat{t}_{\mathrm{Y}}\right) \quad = \quad \sum_{k,l\in\mathrm{U_D}}\sum_{i,j\in\mathrm{U_M}} \pi^{\mathrm{D}}_{kl}\Delta^{\mathrm{M}}_{ij}\check{\mathrm{Y}}_{ik}\check{\mathrm{Y}}_{jl}, \tag{2.2.11}$$

$$\mathrm{V}_2\left(\hat{t}_{\mathrm{Y}}\right) \quad = \quad \sum_{i,j\in\mathrm{U_M}}\sum_{k,l\in\mathrm{U_D}} \pi^{\mathrm{M}}_{ij}\Delta^{\mathrm{D}}_{kl}\check{\mathrm{Y}}_{ik}\check{\mathrm{Y}}_{jl}, \tag{2.2.12}$$

$$\mathrm{V}_3\left(\hat{t}_{\mathrm{Y}}\right) \quad = \quad \sum_{i,j\in\mathrm{U_M}}\sum_{k,l\in\mathrm{U_D}} \Delta^{\mathrm{M}}_{ij}\Delta^{\mathrm{D}}_{kl}\check{\mathrm{Y}}_{ik}\check{\mathrm{Y}}_{jl}. \tag{2.2.13}$$

Plugging $\Gamma_{ijkl} = \Delta^{\mathrm{M}}_{ij}\pi^{\mathrm{D}}_{k}\pi^{\mathrm{D}}_{l} + \Delta^{\mathrm{D}}_{kl}\pi^{\mathrm{M}}_{i}\pi^{\mathrm{M}}_{j} + \Delta^{\mathrm{M}}_{ij}\Delta^{\mathrm{D}}_{kl}$ into (2.2.2) gives

$$\mathrm{V}_{\mathrm{CCS}}\left(\hat{t}_{\mathrm{Y}}\right) \quad = \quad \mathrm{V}^{\mathrm{PSU}}_{\mathrm{MD}}\left(\hat{t}_{\mathrm{Y}}\right) + \mathrm{V}^{\mathrm{PSU}}_{\mathrm{DM}}\left(\hat{t}_{\mathrm{Y}}\right) + \mathrm{V}_3\left(\hat{t}_{\mathrm{Y}}\right) \tag{2.2.14}$$

and we have $\mathrm{V}_1\left(\hat{t}_{\mathrm{Y}}\right) = \mathrm{V}^{\mathrm{PSU}}_{\mathrm{MD}}\left(\hat{t}_{\mathrm{Y}}\right) + \mathrm{V}_3\left(\hat{t}_{\mathrm{Y}}\right)$ and $\mathrm{V}_2\left(\hat{t}_{\mathrm{Y}}\right) = \mathrm{V}^{\mathrm{PSU}}_{\mathrm{DM}}\left(\hat{t}_{\mathrm{Y}}\right) + \mathrm{V}_3\left(\hat{t}_{\mathrm{Y}}\right)$. This second decomposition was originally derived by Dalén and Ohlsson [1995]. It is also given in Ohlsson [1996], and in equation (3) of Theorem 2.2 of Skinner [2015]. Other decompositions are possible, e.g. through an analysis of variance decomposition as for two-stage sampling.

2.2.3 Comparison with two-stage sampling

From expressions (2.2.7) and (2.2.14), we obtain after some algebra that

$$\mathrm{V}_{\mathrm{CCS}}\left(\hat{t}_{\mathrm{Y}}\right) - \mathrm{V}_{\mathrm{DM}}\left(\hat{t}_{\mathrm{Y}}\right) = \sum_{i,j\in\mathrm{U_M}} \Delta^{\mathrm{M}}_{ij} \sum_{k\neq l\in\mathrm{U_D}} \pi^{\mathrm{D}}_{kl}\check{\mathrm{Y}}_{ik}\check{\mathrm{Y}}_{jl}. \tag{2.2.15}$$

In case of Poisson sampling (PO) inside $\mathrm{U_M}$ and when Y is assumed to be non-negative, the right-hand side in (2.2.15) is non-negative and CCS is thus less efficient than two-stage sampling. In case of fixed-size sampling inside $\mathrm{U_M}$, equation (2.2.15) may be alternatively written as

$$\mathrm{V_{CCS}}\left(\hat{t}_\mathrm{Y}\right)-\mathrm{V_{DM}}\left(\hat{t}_\mathrm{Y}\right)=\sum_{i\neq j\in \mathrm{U_M}}\frac{(-\mathit{\Delta}_{ij}^{\mathrm{M}})}{2}\sum_{k\neq l\in \mathrm{U_D}}\frac{\pi_{kl}^{\mathrm{D}}}{\pi_k^{\mathrm{D}}\pi_l^{\mathrm{D}}}\left(\frac{\mathrm{Y}_{ik}}{\pi_i^{\mathrm{M}}}-\frac{\mathrm{Y}_{jk}}{\pi_j^{\mathrm{M}}}\right)\left(\frac{\mathrm{Y}_{il}}{\pi_i^{\mathrm{M}}}-\frac{\mathrm{Y}_{jl}}{\pi_j^{\mathrm{M}}}\right). \quad (2.2.16)$$

If the so-called Sen-Yates-Grundy conditions are respected for p_M, the quantities $(-\mathit{\Delta}_{ij}^{\mathrm{M}})$ are non-negative. If Y_{ik} is roughly proportional to the size of the maternity unit i, as can be expected for count variables, the quantities

$$\left(\frac{\mathrm{Y}_{ik}}{\pi_i^{\mathrm{M}}}-\frac{\mathrm{Y}_{jk}}{\pi_j^{\mathrm{M}}}\right)\left(\frac{\mathrm{Y}_{il}}{\pi_i^{\mathrm{M}}}-\frac{\mathrm{Y}_{jl}}{\pi_j^{\mathrm{M}}}\right)$$

will tend to be positive unless the inclusion probabilities π_i^{M} are defined proportionally to some measure of size. CCS sampling would then be less efficient than two-stage sampling. This result is illustrated in section 2.4.1 on some simulated populations when both p_M and p_D are simple random sampling without replacement (SI) designs, and for different sample sizes.

2.3 Variance estimation

2.3.1 Design-unbiased variance estimation

The HT variance estimator for $\mathrm{V_{CCS}}\left(\hat{t}_\mathrm{Y}\right)$ is

$$\hat{\mathrm{V}}_{\mathrm{HT}}\left(\hat{t}_\mathrm{Y}\right) = \sum_{i,j\in \mathrm{S_M}}\sum_{k,l\in \mathrm{S_D}}\frac{\mathit{\Gamma}_{ijkl}}{\pi_{ij}^{\mathrm{M}}\pi_{kl}^{\mathrm{D}}}\check{\mathrm{Y}}_{ik}\check{\mathrm{Y}}_{jl}. \quad (2.3.1)$$

It may be also derived from (2.2.10), leading to the alternative writing

$$\hat{\mathrm{V}}_{\mathrm{HT}}\left(\hat{t}_{\mathrm{Y}}\right) = \hat{\mathrm{V}}_{1,\mathrm{HT}}\left(\hat{t}_{\mathrm{Y}}\right)+\hat{\mathrm{V}}_{2,\mathrm{HT}}\left(\hat{t}_{\mathrm{Y}}\right)-\hat{\mathrm{V}}_{3,\mathrm{HT}}\left(\hat{t}_{\mathrm{Y}}\right) \tag{2.3.2}$$

where

$$\begin{aligned}
\hat{\mathrm{V}}_{1,\mathrm{HT}}\left(\hat{t}_{\mathrm{Y}}\right) &= \sum_{i,j\in\mathrm{S_M}}\sum_{k,l\in\mathrm{S_D}}\frac{\mathit{\Delta}^{\mathrm{M}}_{ij}}{\pi^{\mathrm{M}}_{ij}}\check{\mathrm{Y}}_{ik}\check{\mathrm{Y}}_{jl}, && (2.3.3)\\
\hat{\mathrm{V}}_{2,\mathrm{HT}}\left(\hat{t}_{\mathrm{Y}}\right) &= \sum_{i,j\in\mathrm{S_M}}\sum_{k,l\in\mathrm{S_D}}\frac{\mathit{\Delta}^{\mathrm{D}}_{kl}}{\pi^{\mathrm{D}}_{kl}}\check{\mathrm{Y}}_{ik}\check{\mathrm{Y}}_{jl}, && (2.3.4)\\
\hat{\mathrm{V}}_{3,\mathrm{HT}}\left(\hat{t}_{\mathrm{Y}}\right) &= \sum_{i,j\in\mathrm{S_M}}\sum_{k,l\in\mathrm{S_D}}\frac{\mathit{\Delta}^{\mathrm{M}}_{ij}}{\pi^{\mathrm{M}}_{ij}}\frac{\mathit{\Delta}^{\mathrm{D}}_{kl}}{\pi^{\mathrm{D}}_{kl}}\check{\mathrm{Y}}_{ik}\check{\mathrm{Y}}_{jl}. && (2.3.5)
\end{aligned}$$

If p_{M} and p_{D} are both Poisson sampling designs and if Y is assume to be non-negative, this variance estimator is always non-negative. Otherwise, it may take negative values even if p_{M} and p_{D} are both SI designs (denoted by SI^2) as illustrated in section 2.4.2. When p_{M} and p_{D} are both fixed-size sampling designs, we may alternatively consider the Yates-Grundy like variance estimator :

$$\hat{\mathrm{V}}_{\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = \hat{\mathrm{V}}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)+\hat{\mathrm{V}}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)-\hat{\mathrm{V}}_{3,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) \tag{2.3.6}$$

where

$$\begin{aligned}
\hat{\mathrm{V}}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) &= -\frac{1}{2}\sum_{i\neq j\in\mathrm{S_M}}\frac{\mathit{\Delta}^{\mathrm{M}}_{ij}}{\pi^{\mathrm{M}}_{ij}}\left(\frac{\hat{\mathrm{Y}}_{i\bullet}}{\pi^{\mathrm{M}}_{i}}-\frac{\hat{\mathrm{Y}}_{j\bullet}}{\pi^{\mathrm{M}}_{j}}\right)^2 && (2.3.7)\\
\hat{\mathrm{V}}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) &= -\frac{1}{2}\sum_{k\neq l\in\mathrm{S_D}}\frac{\mathit{\Delta}^{\mathrm{D}}_{kl}}{\pi^{\mathrm{D}}_{kl}}\left(\frac{\hat{\mathrm{Y}}_{\bullet k}}{\pi^{\mathrm{D}}_{k}}-\frac{\hat{\mathrm{Y}}_{\bullet l}}{\pi^{\mathrm{D}}_{l}}\right)^2 && (2.3.8)\\
\hat{\mathrm{V}}_{3,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) &= -\frac{1}{2}\sum_{(i,k)\neq(j,l)\in\mathrm{S_M}\times\mathrm{S_D}}\frac{\mathit{\Delta}^{\mathrm{M}}_{ij}\mathit{\Delta}^{\mathrm{D}}_{kl}}{\pi^{\mathrm{M}}_{ij}\pi^{\mathrm{D}}_{kl}}\left(\check{\mathrm{Y}}_{ik}-\check{\mathrm{Y}}_{jl}\right)^2 && (2.3.9)
\end{aligned}$$

with $\hat{\mathrm{Y}}_{\bullet k}=\sum_{i\in\mathrm{S_M}}\mathrm{Y}_{ik}/\pi^{\mathrm{M}}_{i}$ the estimated sub-total for the day k and $\hat{\mathrm{Y}}_{i\bullet}=\sum_{k\in\mathrm{S_D}}\mathrm{Y}_{ik}/\pi^{\mathrm{D}}_{k}$ the estimated sub-total for the maternity unit i.

It can be proved that $\hat{V}_{HT}\left(\hat{t}_Y\right)$ in (2.3.2) and $\hat{V}_{YG}\left(\hat{t}_Y\right)$ in (2.3.6) match term by term, when p_M and p_D are stratified simple random sampling (STSI) designs. In the same STSI context, another variance estimator is given in equation (4) of Theorem 2.2 in Skinner [2015]. This variance estimator does not match $\hat{V}_{HT}\left(\hat{t}_Y\right)$ or $\hat{V}_{YG}\left(\hat{t}_Y\right)$ term by term, since Skinner's variance estimator is based on the variance decomposition in equation (2.2.14), while our variance estimator is based on the variance decomposition in equation (2.2.10). Nevertheless, both variance estimators are globally identical in the STSI case.

Another variance estimator is obtained in Dalén and Ohlsson [1995], in case of a probability proportional to size without-replacement sampling design in both dimensions. Summing the variance component estimators in equations (4.1)-(4.3) of Dalén and Ohlsson [1995] leads to a similar variance estimator than in our equation (2.3.6), except that $-\hat{V}_{3,YG}\left(\hat{t}_Y\right)$ is replaced with $+\hat{V}_{3,YG}\left(\hat{t}_Y\right)$ which results in an overestimation of the variance. This overestimation can be be neglected in cases when $V_3\left(\hat{t}_Y\right)$ is small as compared to the other variance components (see Table 2.1 in Section 2.4.2).

If both sampling designs satisfy the Sen-Yates-Grundy conditions (SYG), the terms $\hat{V}_{1,YG}\left(\hat{t}_Y\right)$ and $\hat{V}_{2,YG}\left(\hat{t}_Y\right)$ are non-negative. However, the term $\hat{V}_{3,YG}\left(\hat{t}_Y\right)$ is usually non-negative, which may lead to negative values for $\hat{V}_{YG}\left(\hat{t}_Y\right)$ as illustrated in the simulations of section 2.4.2. It is thus desirable to have access to non-negative variance estimators with limited bias.

2.3.2 Non-negative variance estimators

We consider the variance decomposition in (2.2.10), and study the relative order of magnitude of the components. We make the following assumptions :

H1 : There exist some constants α_1 and α_2 such that

$$\forall k \in U_D,\ \frac{1}{N_M}\sum_{i\in U_M} Y_{ik}^2 \leq \alpha_1, \quad \text{and} \quad \forall i \in U_M,\ \frac{1}{N_D}\sum_{k\in U_D} Y_{ik}^2 \leq \alpha_2.$$

H2 : There exist some constants $\lambda_1 > 0$ and $\lambda_2 > 0$ such that

$$\forall k \in \mathrm{U_D},\ \pi_k^{\mathrm{D}} \geq \lambda_1 \frac{n_{\mathrm{D}}}{\mathrm{N_D}}, \quad \text{and} \quad \forall i \in \mathrm{U_M},\ \pi_i^{\mathrm{M}} \geq \lambda_2 \frac{n_{\mathrm{M}}}{\mathrm{N_M}}.$$

H3 : There exist some constants γ_1 and γ_2 such that

$$\forall k \neq l \in \mathrm{U_D},\ \frac{\mathrm{N_D^2}}{n_{\mathrm{D}}} \sup_{k \neq l \in \mathrm{U_D}} \left|\Delta_{kl}^{\mathrm{D}}\right| \leq \gamma_1, \quad \text{and} \quad \forall i \neq j \in \mathrm{U_M},\ \frac{\mathrm{N_M^2}}{n_{\mathrm{M}}} \sup_{i \neq j \in \mathrm{U_M}} \left|\Delta_{ij}^{\mathrm{M}}\right| \leq \gamma_2.$$

H4 : There exists some constant $\delta > 0$ such that

$$\mathrm{V_{CCS}}\left(\hat{t}_{\mathrm{Y}}\right) \quad \geq \quad \delta \mathrm{N_M^2 N_D^2} \left(\frac{1}{n_{\mathrm{M}}} + \frac{1}{n_{\mathrm{D}}}\right).$$

It is assumed in (H1) that the variable y has bounded moments of order 2 for each maternity unit i and for each day k. Assumptions (H2) and (H3) are classical in survey sampling and are satisfied for many sampling designs, see for example Cardot et al. [2013]. It is assumed in (H4) that the variance of the HT-estimator under CCS sampling has the order $\mathrm{N_M^2 N_D^2}(n_{\mathrm{M}}^{-1} + n_{\mathrm{D}}^{-1})$. From assumptions (H1-H4), there exist some constants $\mathrm{C_1}$, $\mathrm{C_2}$ and $\mathrm{C_3}$ such that

$$\frac{\mathrm{V_1}\left(\hat{t}_{\mathrm{Y}}\right)}{\mathrm{V_{CCS}}\left(\hat{t}_{\mathrm{Y}}\right)} \quad \leq \quad \mathrm{C_1} \frac{1}{1 + n_{\mathrm{M}} n_{\mathrm{D}}^{-1}}, \tag{2.3.10}$$

$$\frac{\mathrm{V_2}\left(\hat{t}_{\mathrm{Y}}\right)}{\mathrm{V_{CCS}}\left(\hat{t}_{\mathrm{Y}}\right)} \quad \leq \quad \mathrm{C_2} \frac{1}{1 + n_{\mathrm{D}} n_{\mathrm{M}}^{-1}}, \tag{2.3.11}$$

$$\frac{\mathrm{V_3}\left(\hat{t}_{\mathrm{Y}}\right)}{\mathrm{V_{CCS}}\left(\hat{t}_{\mathrm{Y}}\right)} \quad \leq \quad \mathrm{C_3} \frac{1}{n_{\mathrm{D}} n_{\mathrm{M}}^{-1} + n_{\mathrm{M}} n_{\mathrm{D}}^{-1}} \tag{2.3.12}$$

The proof is given in section 2.8. It follows from (2.3.10)-(2.3.12) that if n_{D} is large and n_{M} is bounded, both $\mathrm{V_2}\left(\hat{t}_{\mathrm{Y}}\right)$ and $\mathrm{V_3}\left(\hat{t}_{\mathrm{Y}}\right)$ are negligible and a non-negative simplified variance estimator can be derived by focusing on $\mathrm{V_1}\left(\hat{t}_{\mathrm{Y}}\right)$ only. This leads to

$$\hat{\mathrm{V}}_{\mathrm{SIMP1}}\left(\hat{t}_{\mathrm{Y}}\right) \quad = \quad \hat{\mathrm{V}}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right). \tag{2.3.13}$$

If the sampling design p_{D} satisfies the SYG conditions, this simplified estimator is always non-negative. In the particular SI^2 case, we obtain

$$\hat{\mathrm{V}}_{\mathrm{SIMP1}}\left(\hat{t}_{\mathrm{Y}}\right) \quad = \quad \mathrm{N}_{\mathrm{M}}^2\left(\frac{1}{n_{\mathrm{M}}}-\frac{1}{\mathrm{N}_{\mathrm{M}}}\right)s^2_{\hat{\mathrm{Y}}_{\circ\bullet}} \tag{2.3.14}$$

where

$$s^2_{\hat{\mathrm{Y}}_{\circ\bullet}} \quad = \quad \frac{1}{n_{\mathrm{M}}-1}\sum_{i\in \mathrm{S}_{\mathrm{M}}}\left(\hat{\mathrm{Y}}_{i\bullet}-\frac{1}{n_{\mathrm{M}}}\sum_{j\in \mathrm{S}_{\mathrm{M}}}\hat{\mathrm{Y}}_{j\bullet}\right)^2. \tag{2.3.15}$$

Symmetrically, both $\mathrm{V}_1\left(\hat{t}_{\mathrm{Y}}\right)$ and $\mathrm{V}_3\left(\hat{t}_{\mathrm{Y}}\right)$ may be seen as negligible if n_{M} is large and n_{D} is bounded. Another simplified variance estimator is thus

$$\hat{\mathrm{V}}_{\mathrm{SIMP2}}\left(\hat{t}_{\mathrm{Y}}\right) \quad = \quad \hat{\mathrm{V}}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right). \tag{2.3.16}$$

If the sampling design p_{M} satisfies the SYG conditions, this estimator is non-negative. In the particular SI^2 case, we have

$$\hat{\mathrm{V}}_{\mathrm{SIMP2}}\left(\hat{t}_{\mathrm{Y}}\right) \quad = \quad \mathrm{N}_{\mathrm{D}}^2\left(\frac{1}{n_{\mathrm{D}}}-\frac{1}{\mathrm{N}_{\mathrm{D}}}\right)s^2_{\hat{\mathrm{Y}}_{\bullet\circ}} \tag{2.3.17}$$

where

$$s^2_{\hat{\mathrm{Y}}_{\bullet\circ}} \quad = \quad \frac{1}{n_{\mathrm{D}}-1}\sum_{k\in \mathrm{S}_{\mathrm{D}}}\left(\hat{\mathrm{Y}}_{\bullet k}-\frac{1}{n_{\mathrm{D}}}\sum_{l\in \mathrm{S}_{\mathrm{D}}}\hat{\mathrm{Y}}_{\bullet l}\right)^2. \tag{2.3.18}$$

A third possible simplified variance estimator is

$$\begin{aligned}\hat{\mathrm{V}}_{\mathrm{SIMP3}}\left(\hat{t}_{\mathrm{Y}}\right) \quad &= \quad \hat{\mathrm{V}}_{\mathrm{SIMP1}}+\hat{\mathrm{V}}_{\mathrm{SIMP2}} \\ &= \quad \hat{\mathrm{V}}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)+\hat{\mathrm{V}}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right).\end{aligned} \tag{2.3.19}$$

This estimator is non-negative if both p_{D} and p_{M} satisfy the SYG conditions. It is approximately unbiased for $\mathrm{V}_{\mathrm{CCS}}\left(\hat{t}_{\mathrm{Y}}\right)$ if n_{D} is large and n_{M} is bounded, or if n_{M} is large and n_{D} is bounded. In the particular SI^2 case

$$\hat{\mathrm{V}}_{\mathrm{SIMP3}}\left(\hat{t}_{\mathrm{Y}}\right) = \mathrm{N}_{\mathrm{M}}^2\left(\frac{1}{n_{\mathrm{M}}} - \frac{1}{\mathrm{N}_{\mathrm{M}}}\right) s^2_{\hat{\mathrm{Y}}_{\circ\bullet}} + \mathrm{N}_{\mathrm{D}}^2\left(\frac{1}{n_{\mathrm{D}}} - \frac{1}{\mathrm{N}_{\mathrm{D}}}\right) s^2_{\hat{\mathrm{Y}}_{\bullet\circ}}. \tag{2.3.20}$$

Similar formula can be easily derived in the case of stratified simple random sampling without replacement and will be used in Section 5.

2.3.3 Relative bias under a superpopulation model

We consider the following superpopulation model

$$\mathrm{Y}_{ik} = \mu + \sigma_{\mathrm{M}}\mathrm{U}_i + \sigma_{\mathrm{D}}\mathrm{V}_k + \sigma_{\mathrm{E}}\mathrm{W}_{ik} \tag{2.3.21}$$

where U_i, V_k and W_{ik} are independently generated according to a standard normal distribution. This is a particular case for a single stratum of the stratified cross-classified population model introduced in equation (8) of Skinner [2015], where the fixed and random effects are allowed to depend on the strata. Model (2.3.21) is an analysis of variance model with two crossed random factors and without repetition. Let “E_m" denote the expectation with respect to the model (2.3.21) and “E_p" denote the expectation with respect to the CCS design. For each simplified variance estimator $\hat{\mathrm{V}}_{\mathrm{SIMP}i}$, $i = 1,2,3$, the relative bias RB under the model and under the sampling design is defined by

$$\mathrm{RB}_{m,p}\left[\hat{\mathrm{V}}_{\mathrm{SIMP}i}\left(\hat{t}_{\mathrm{Y}}\right)\right] = \frac{\mathrm{E}_m\left\{\mathrm{E}_p\left[\hat{\mathrm{V}}_{\mathrm{SIMPi}}\left(\hat{t}_{\mathrm{Y}}\right)\right] - \mathrm{V}_{\mathrm{CCS}}\left(\hat{t}_{\mathrm{Y}}\right)\right\}}{\mathrm{E}_m\left[\mathrm{V}_{\mathrm{CCS}}\left(\hat{t}_{\mathrm{Y}}\right)\right]}. \tag{2.3.22}$$

In the SI^2 case, these relative biases are of the form

$$\mathrm{RB}_{m,p}\left[\hat{\mathrm{V}}_{\mathrm{SIMP}i}\left(\hat{t}_{\mathrm{Y}}\right)\right] \quad = \quad -1/(1+\mathrm{A}_i) \tag{2.3.23}$$

for $i = 1$ and 2 and

$$\mathrm{RB}_{m,p}\left[\hat{\mathrm{V}}_{\mathrm{SIMP3}}\left(\hat{t}_{\mathrm{Y}}\right)\right] = 1/(1+\mathrm{A}_3) \tag{2.3.24}$$

for some positive constant A_i, $i = 1,2,3$, depending on σ_{M}, σ_{D}, σ_{E} and n_{M}, N_{M}, n_{D} and N_{D}, see equations (2.3.25)-(2.3.27). Equations (2.3.23) and (2.3.24) imply that the two first simplified variance estimators are negatively biased while the third one is positively biased. Using the notations $r_{\mathrm{M}} = \sigma^2_{\mathrm{M}}/\sigma^2_{\mathrm{E}}$, $r_{\mathrm{D}} = \sigma^2_{\mathrm{D}}/\sigma^2_{\mathrm{E}}$, $f_{\mathrm{M}} = n_{\mathrm{M}}/\mathrm{N}_{\mathrm{M}}$ and $f_{\mathrm{D}} = n_{\mathrm{D}}/\mathrm{N}_{\mathrm{D}}$, we have

$$\begin{aligned}
\mathrm{A}_1 &= \frac{1-f_{\mathrm{M}}}{1-f_{\mathrm{D}}}\frac{n_{\mathrm{D}}r_{\mathrm{M}}+1}{n_{\mathrm{M}}r_{\mathrm{D}}+f_{\mathrm{M}}}, && (2.3.25)\\
\mathrm{A}_2 &= \frac{1-f_{\mathrm{D}}}{1-f_{\mathrm{M}}}\frac{n_{\mathrm{M}}r_{\mathrm{D}}+1}{n_{\mathrm{D}}r_{\mathrm{M}}+f_{\mathrm{D}}}, && (2.3.26)\\
\mathrm{A}_3 &= \frac{n_{\mathrm{D}}r_{\mathrm{M}}+f_{\mathrm{D}}}{1-f_{\mathrm{D}}}+\frac{n_{\mathrm{M}}r_{\mathrm{D}}+f_{\mathrm{M}}}{1-f_{\mathrm{M}}}. && (2.3.27)
\end{aligned}$$

The bias of $\hat{\mathrm{V}}_{\mathrm{SIMP1}}$ increases from -1 to 0 when A_1 increases, which occurs in particular when the ratio r_{M} or the sample size n_{D} increases. In other words, $\hat{\mathrm{V}}_{\mathrm{SIMP1}}$ will have a small bias under model (2.3.21) if the variable of interest contains some maternity unit effect or if the number of sampled days is large enough. Symmetrically, $\hat{\mathrm{V}}_{\mathrm{SIMP2}}$ will have a small bias under model (2.3.21) if the variable of interest contains some day effect or if the number of sampled maternity units is large enough. The bias of $\hat{\mathrm{V}}_{\mathrm{SIMP3}}$ decreases from 1 to 0 when A_3 increases, which occurs in particular when r_{M} or r_{D} increases, or when n_{M} or n_{D} increases. In other words, $\hat{\mathrm{V}}_{\mathrm{SIMP3}}$ will have a small bias under model (2.3.21) if the variable of interest contains some maternity unit or some day effect, or if the number of sampled days or the number of sampled maternity units is large enough. The simulation study in section 2.4 supports these results, and confirms that the variance tends to be underestimated with $\hat{\mathrm{V}}_{\mathrm{SIMP1}}$ or $\hat{\mathrm{V}}_{\mathrm{SIMP2}}$, and overestimated with $\hat{\mathrm{V}}_{\mathrm{SIMP3}}$.

2.3.4 A central-limit theorem

To produce confidence intervals with appropriate asymptotic coverage, it is of interest to state a central-limit theorem (CLT) for CCS. Roughly speaking, Theorem 1 below states that if the HT-estimator follows a CLT under both sampling designs p_D and p_M, then the HT-estimator also follows a CLT under CCS. It is derived almost directly from Theorem 2 in Chen and Rao [2007], and the proof is therefore omitted.

Theorem 1. *Suppose that assumptions (H1)-(H4) hold. Suppose that*

H5 : $\sigma_1^{-1} V_1 \rightarrow_{\mathscr{L}} \mathcal{N}(0,1)$, *where* $\rightarrow_{\mathscr{L}}$ *stands for the convergence in distribution under the sampling-design, with*

$$V_1 = \frac{1}{N}\left(\sum_{i\in S_M} \frac{Y_{i\bullet}}{\pi_i^M} - \sum_{i\in U_M} Y_{i\bullet}\right) \quad \textit{and} \quad \sigma_1^2 = V(V_1) \tag{2.3.28}$$

where $Y_{i\bullet} = \sum_{k\in U_D} Y_{ik}$.

H6 : $\sup_t |P(\sigma_2^{-1} U_1 \leq t | S_M) - \Phi(t)| = o_p(1)$, *where* Φ *is the cumulative distribution function of the standard normal distribution, and where*

$$U_1 = \frac{1}{N}\sum_{i\in S_M} \frac{1}{\pi_i^M}(\hat{Y}_{i\bullet} - Y_{i\bullet}) \quad \textit{and} \quad \sigma_2^2 = V(U_1 | S_M). \tag{2.3.29}$$

H7 : $\sigma_1^2/\sigma_2^2 \rightarrow_P \gamma^2$, *where* $\rightarrow_P$ *stands for the convergence in probability under the sampling-design.*

Then

$$\frac{N^{-1}(\hat{t}_Y - t_Y)}{\sqrt{\sigma_1^2 + \sigma_2^2}} \quad \rightarrow_{\mathscr{L}} \quad \mathcal{N}(0,1). \tag{2.3.30}$$

For illustration, we consider the particular case when p_D and p_M are both SI designs. Suppose that (H2)-(H4) hold, and that (H1) is strengthened to

H1b : There exists $\delta > 0$ and some constants α_1 and α_2 such that

$$\forall k \in U_D \ \frac{1}{N_M} \sum_{i \in U_M} Y_{ik}^{2+\delta} \leq \alpha_1, \quad \text{and} \quad \forall i \in U_M \ \frac{1}{N_D} \sum_{k \in U_D} Y_{ik}^{2+\delta} \leq \alpha_2.$$

Then by using the CLT in Hajek [1961], the assumption (H5) can be shown to hold. By mimicking the proof of Lemma 2 in Chen and Rao [2007], the assumption (H6) can be shown to hold as well.

2.4 Simulations

In this Section, two artificial populations are first generated using the superpopulation model (2.3.21). In Section 2.4.1, CCS is compared with two stage sampling in terms of variance, which illustrates the results in Section 2.2.3. A Monte Carlo experiment is then presented in Section 2.4.2, and the variance estimators introduced in Section 2.3 are compared for the estimation of a total. Some attention is paid to the issue of negative values for the unbiased variance estimator. In Section 2.4.3, two other populations with two variables of interest for each are generated. We focus on variance estimation for a ratio, making use of the variance estimators introduced in Section 2.3 with estimated linearized variables instead of the variable of interest. The results from Tables 2.1 and 2.2 are readily reproducible using the R code provided in the supplementary materials of the present paper.

2.4.1 Comparison with two-stage sampling

Two populations are generated according to model (2.3.21), with $N_M = 1000$ maternity units and $N_D = 1000$ days for each population, and with $\mu = 200$ and $\sigma_E = 5$. Equal random effects standard deviations $\sigma_M = \sigma_D = 5$ are used for population 1, while we use $\sigma_M = 0.5$ and $\sigma_D = 5$ for population 2. For each population, the SI^2 sampling design is used, with sample sizes, n_M and n_D, equal to 5, 10, 100 and 500. The ratios V_{MD}/V_{CCS} between the variance under two-stage sampling and the variance under

CCS are computed, and plotted as a percentage on Figure 2.2. A ratio smaller than 100 indicates that two-stage sampling is more accurate than CCS, which holds true in all cases considered in our experiment.

The ratio increases with n_{D} and decreases when n_{M} increases. Also, it can be observed that the ratio decreases with σ_{M}. This impact of the maternity unit effect is noticeable, and illustrates the substantial loss in accuracy induced by using a CCS instead of a two-stage sampling design if the maternity unit effect is small. Similar conclusions could be derived when computing the ratio $\mathrm{V}_{\mathrm{DM}}/\mathrm{V}_{\mathrm{CCS}}$.

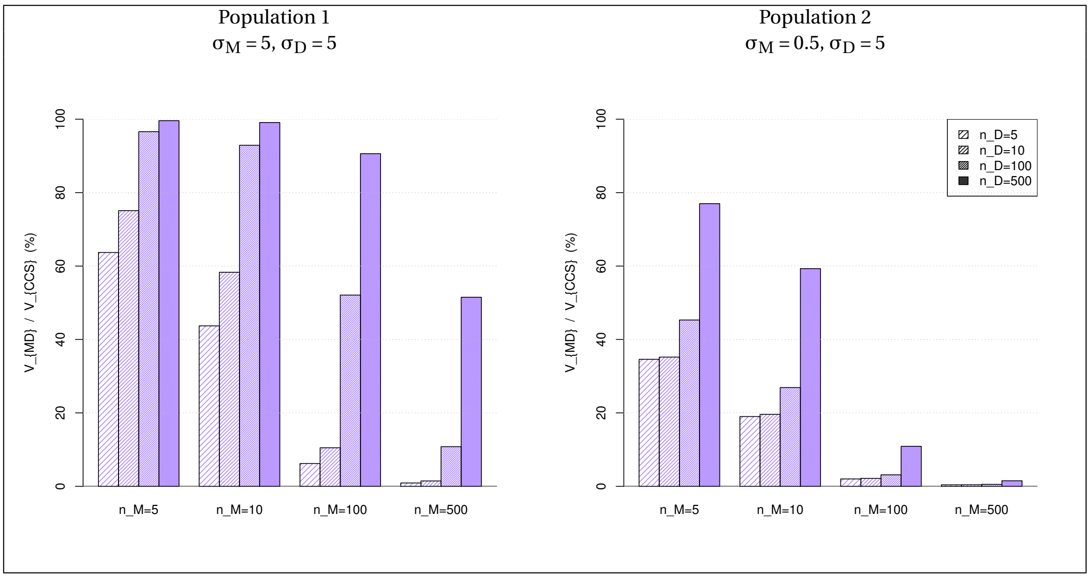


FIGURE 2.2 – $\mathrm{V}_{\mathrm{MD}}/\mathrm{V}_{\mathrm{CCS}}$ (%) for population 1 (left panel) and population 2 (right panel)

2.4.2 Variance estimation for a total

We consider the two artificial populations generated as described in Section 2.4.1. For each population, the SI^2 sampling design is used, with sample sizes equal to 5, 10, 100 and 500, and the sample selection is repeated $\mathrm{B} = 10,000$ times. For each sample $b = 1,\ldots,\mathrm{B}$, we compute the estimate $\hat{t}_{\mathrm{Y}}^{(b)}$ of the total t_{Y}. The unbiased variance estimator $\hat{\mathrm{V}}^{(b)}$ and the simplified variance estimators $\hat{\mathrm{V}}_{\mathrm{SIMP1}}^{(b)}$, $\hat{\mathrm{V}}_{\mathrm{SIMP2}}^{(b)}$, $\hat{\mathrm{V}}_{\mathrm{SIMP3}}^{(b)}$ are also computed for $\hat{t}_{\mathrm{Y}}^{(b)}$.

For each variance estimator $\hat{V}$, we compute the Monte Carlo Percent Relative Bias

$$RB_{MC}(\hat{V}) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{V}^{(b)} - V}{V},$$

where the true variance V was approximated through an independent set of 50,000 simulations. The number (#NEG) of negative variance estimators $\hat{V}^{(b)}$ is also computed.

The results are reported in Table 2.1. The variance estimator $\hat{V}$ is almost unbiased in all situations, as expected. However, this variance estimator is prone to negative values with small sample sizes when the value of σ_M and/or the value of σ_D is small as compared to σ_E. The problem vanishes when the sample sizes increase. We now turn to the simplified variance estimators. The relative bias of $\hat{V}_{SIMP1}$ decreases when n_D increases or when n_M decreases, and when σ_M increases or when σ_D decreases. This supports the findings in Section 2.3.3. Symmetrical conclusions are drawn for the relative bias of $\hat{V}_{SIMP2}$. Turning to $\hat{V}_{SIMP3}$, we note that the relative bias decreases when either σ_M or σ_D increases. This variance estimator is therefore advisable in all cases but those where there is no maternity unit nor day effect.

2.4.3 Variance estimation for a ratio

We now consider variance estimation for a ratio. Two populations are generated with $N_M = 1000$ maternity units and $N_D = 1000$ days. In each population, two count variables are generated so as to mimic the data encountered in the ELFE survey. More precisely, we first generate an auxiliary variable Z_{ik} according to model (2.3.21) with $\mu = 200$, $\sigma_E = \sigma_D = 5$, and $\sigma_M = 5$ or 50. The first variable of interest X_{ik} is generated according to a Poisson distribution with parameter Z_{ik}. The second variable of interest Y_{ik} is generated according to a binomial distribution with parameters X_{ik} and p_{ik}. We consider two cases : (i) equal probabilities with $p_{ik} = 0.3$; (ii) unequal probabilities with $\text{logit}(p_{ik}) = \beta Z_{ik}$, where β was chosen so that the average probability is approximately 0.3. Note that Y_{ik} follows a Poisson distribution with parameter

n_M	5	10	10	100	500	5	10	10	100	500
n_D	5	10	100	100	500	5	10	100	100	500
σ_M	5					50				
σ_D	5					5				
$RB_{MC}(\hat{V})$	1	-1	2	0	-0	1	-1	1	0	0
#NEG	6	0	0	0	0	0	0	0	0	0
$RB_{MC}(\hat{V}_{SIMP1})$	-43	-47	-6	-49	-49	-	-2	1	-1	-1
$RB_{MC}(\hat{V}_{SIMP2})$	-46	-50	-91	-51	-51	-99	-99	-100	-99	-99
$RB_{MC}(\hat{V}_{SIMP3})$	11	3	2	1	-0	1	-0	1	0	0
σ_M	0.5					0.5				
σ_D	5					0.5				
$RB_{MC}(\hat{V})$	1	-1	0	1	-1	1	-1	2	-0	-0
#NEG	91	0	0	0	0	1393	298	0	0	0
$RB_{MC}(\hat{V}_{SIMP1})$	-82	-90	-81	-98	-99	-4	-9	-3	-34	-47
$RB_{MC}(\hat{V}_{SIMP2})$	-1	-2	-10	-0	-2	-5	-10	-52	-36	-49
$RB_{MC}(\hat{V}_{SIMP3})$	18	8	9	2	-0	90	81	45	29	4

TABLE 2.1 – Comparison between variance estimators for a total

$p_{ik}Z_{ik}$.

The reason for this generating process is that some variable of interest X_{ik}, like the number of births in the ELFE survey, may contain some maternity unit and/or day effect which is reflected in the way Z_{ik} is generated. On the other hand, some maternity unit and/or day effect may also be contained in some other variable of interest Y_{ik}, like the number of births per caesarean. Such effects may be either similar to those for X_{ik} like with pattern (i), or may occur differently like with pattern (ii).

For each population, the SI^2 sampling design is used, with sample sizes equal to 5, 10, 100 and 500, and the sample selection is repeated $B = 10,000$ times. For each sample $b = 1,\ldots,B$, we compute the substitution estimator $\hat{R}^{(b)} = \hat{t}_Y^{(b)} / \hat{t}_X^{(b)}$ of the ratio $R = t_Y / t_X$. The variance estimator $\hat{V}^{(b)}$ and the simplified variance estimators $\hat{V}_{SIMP1}^{(b)}$, $\hat{V}_{SIMP2}^{(b)}$, $\hat{V}_{SIMP3}^{(b)}$ are also computed for $\hat{t}_Y^{(b)}$, where the variable of interest Y_{ik} is replaced with the estimated linearized variable of the ratio.

The results are reported in Table 2.2. The variance estimator $\hat{V}$ is almost unbiased in all situations, as expected, but is prone to negative values even when the maternity unit or day effect is small. We now turn to the relative bias for the simplified variance estimators. With pattern (i), the situation is much different from that when a total is estimated, since the relative bias of $\hat{V}_{\mathrm{SIMP3}}$ is much larger than for the other two simplified estimators. This can be explained as follows : when the probabilities p_{ik} are uniform, both Y_{ik} and X_{ik} contain the same maternity unit and day effect, but these effects wear off in the linearized variable. Whatever the values of σ_{M} and σ_{D} are, the situation is therefore comparable to that observed in the bottom right cell of Table 2.1. With pattern (ii), the probabilities p_{ik} depend on i and k, leading potentially to some remaining maternity unit and/or day effect in the linearized variable. In such situation, which seems more realistic in practice, the relative bias of $\hat{V}_{\mathrm{SIMP1}}$ and $\hat{V}_{\mathrm{SIMP2}}$ increase when σ_{M} or σ_{D} increase, while the relative bias of $\hat{V}_{\mathrm{SIMP3}}$ decreases.

	n_{M}	5	10	10	100	500	5	10	10	100	500
	n_{D}	5	10	100	100	500	5	10	100	100	500
	σ_{M}	5					50				
	σ_{D}	5					5				
Case (i)	$\mathrm{RB}_{\mathrm{MC}}(\hat{V})$	-0	-1	-1	0	-0	-2	-1	-1	0	1
	#NEG	1645	484	14	0	0	1656	499	12	0	0
$p_{ik} = 0.3$	$\mathrm{RB}_{\mathrm{MC}}(\hat{V}_{\mathrm{SIMP1}})$	-1	-1	-2	-10	-37	-1	-1	-1	-8	-32
	$\mathrm{RB}_{\mathrm{MC}}(\hat{V}_{\mathrm{SIMP2}})$	0	-2	-10	-8	-30	-2	-1	-9	-8	-31
	$\mathrm{RB}_{\mathrm{MC}}(\hat{V}_{\mathrm{SIMP3}})$	99	96	89	82	33	97	98	90	84	37
Case (ii)	$\mathrm{RB}_{\mathrm{MC}}(\hat{V})$	0	-1	2	0	-0	-4	-3	-1	-0	0
	#NEG	1351	235	0	0	0	67	0	0	0	0
$p_{ik} = \frac{e^{\beta Z_{ik}}}{1+e^{\beta Z_{ik}}}$	$\mathrm{RB}_{\mathrm{MC}}(\hat{V}_{\mathrm{SIMP1}})$	-7	-13	-4	-39	-48	-5	-4	-1	-1	-1
	$\mathrm{RB}_{\mathrm{MC}}(\hat{V}_{\mathrm{SIMP2}})$	-6	-14	-61	-40	-49	-87	-93	-99	-98	-99
	$\mathrm{RB}_{\mathrm{MC}}(\hat{V}_{\mathrm{SIMP3}})$	87	73	35	22	3	8	3	-0	0	0

TABLE 2.2 – Comparison between variance estimators for a ratio

2.5 Application to the ELFE survey

ELFE is the first longitudinal study of its kind in France, tracking children from birth to adulthood (Pirus et al. [2010]). This cohort comprises more than 18,000 children whose parents consented to their inclusion. The population of inference consists of babies born during 2011 in French maternity units, excluding very premature infants.

This is a two-dimensional population with 544 maternity units as spatial units and 365 days as time units. The crossing of one day and one maternity unit represents a cluster of infants.

The sample is obtained by CCS, where days and maternity units are selected independently with selected families surveyed shortly after birth in 320 metropolitan maternity units and during 25 days for one year. The population of maternity units is divided into five strata of equal size. The allocation per stratum is proportional to the number of deliveries recorded in 2008. The sample selection for maternity units is stratified systematic sampling, which can be approximated by stratified simple random sampling (STSI). The sample selection of days is not actually random, due to logistic constraints. A number of $n_{\mathrm{D}} = 25$ days is selected during 4 waves, each wave covering a season. It may be approximated by STSI, with four strata associated to the four seasons of 2011. The sample sizes inside strata are provided in Tables 5.3 and 5.4.

Strata	Strata size	Sample size
g	$\mathrm{N}_{\mathrm{M}g}$	$n_{\mathrm{M}g}$
1	108	21
2	108	41
3	109	55
4	108	80
5	111	90
Total	544	287

TABLE 2.3 – Population and sample strata sizes for the maternity units design p_{M}.

Strata	Strata size	Sample size
h	$\mathrm{N}_{\mathrm{D}h}$	$n_{\mathrm{D}h}$
1	91	4
2	91	6
3	91	7
4	92	8
Total	365	25

TABLE 2.4 – Population and sample strata sizes for the days design p_{D}.

In this Section, we aim at illustrating the results previously obtained on a real data set. Some aspects of the ELFE survey, like the non-response issue or the calibration step, deserve a specific attention but are beyond the scope of the present paper and are therefore not considered. In particular, the ELFE survey is prone to several levels of non-response, since some sampled maternity units and some families refused to participate either for some specific days or for the whole period. In the present study, the sample of respondents is viewed as the original sample and in particular, we consider only the 287 maternity units that participate during the 25 days of survey. The calibration step is not taken into account. The results below are meant to illustrate our theoretical results, but are not intended for use in other contexts.

We consider seven count variables from the ELFE survey. Some of them depend on the characteristics of the maternity units (e.g., the spatial location), like the variable indicating whether the mother is followed by a midwife. Others are related to the days of the survey, like the variable indicating whether the birth occurred by caesarean. For each variable, the estimated total $\hat{t}_{\mathrm{Y}}$ from equation (2.2.1), the estimated variance $\hat{\mathbf{V}}\left(\hat{t}_{\mathrm{Y}}\right)$ from equation (2.3.2) and the three simplified estimators are given in the upper part of Table 2.5. Similar indicators are given in the bottom part of Table 2.5 for ratios, when the totals of the variables of interest are divided by the total number of births.

	Birth	**Born by Caesarean**	**Twins**	**Born within marriage**	**Mother followed by a midwife**	**Mother aged between 18 and 25 years**	**Primiparous mother**	**Immigrant mother**
$\hat{t}_{\mathrm{Y}}$	362924	33873	10187	160283	42337	43238	162316	44169
$\hat{\mathbf{V}}\left(\hat{t}_{\mathrm{Y}}\right)$	7.6E+07	1.5E+07	5.3E+05	2.0E+07	3.9E+06	2.6E+06	1.5E+07	3.6E+06
$\mathrm{RD}\left(\hat{\mathbf{V}}_{\mathrm{SIMP1}}\right)$	-63.7 %	-95.5 %	-63.5 %	-64.6 %	-13.2 %	-49.7 %	-46.5 %	-58.2 %
$\mathrm{RD}\left(\hat{\mathbf{V}}_{\mathrm{SIMP2}}\right)$	-31.1 %	-1.9 %	-13.3 %	-29.7 %	-76.3 %	-35.2 %	-41.4 %	-33.4 %
$\mathrm{RD}\left(\hat{\mathbf{V}}_{\mathrm{SIMP3}}\right)$	5.2 %	2.6 %	23.2 %	5.7 %	10.5 %	15.1 %	12.2 %	8.4 %
$\hat{\mathrm{R}}$	1.00	0.09	0.03	0.44	0.12	0.12	0.45	0.12
$\hat{\mathbf{V}}(\hat{\mathrm{R}})$		7.9E-05	2.8E-06	2.4E-05	2.5E-05	1.2E-05	3.0E-05	1.6E-05
$\mathrm{RD}\left(\hat{\mathbf{V}}_{\mathrm{SIMP1}}\right)$		-96.2 %	-51.0 %	-31.0 %	-7.9 %	-40.2 %	-69.3 %	-49.2 %
$\mathrm{RD}\left(\hat{\mathbf{V}}_{\mathrm{SIMP2}}\right)$		-0.4 %	-17.0 %	-44.7 %	-80.5 %	-35.5 %	-5.0 %	-37.5 %
$\mathrm{RD}\left(\hat{\mathbf{V}}_{\mathrm{SIMP3}}\right)$		3.4 %	31.9 %	24.3 %	11.5 %	24.3 %	25.7 %	13.3 %

TABLE 2.5 – Variance estimates of estimated total and ratio on some ELFE variables

The relative difference RD between $\hat{V}_{SIMP}$ and the unbiased estimator $\hat{V}$ is

$$\text{RD} = \frac{\hat{V}_{SIMP}\left(\hat{t}_Y\right) - \hat{V}\left(\hat{t}_Y\right)}{\hat{V}\left(\hat{t}_Y\right)}.$$

Different behaviours may be observed for the variables of interest, depending on the maternity unit/day effect. For instance, the variable indicating whether the birth occurred by caesarean contains an important day effect, and the RD of $\hat{V}_{SIMP2}$ is therefore small while that of $\hat{V}_{SIMP1}$ is large. Symmetrically, the variable indicating whether the mother is followed by a midwife contains a small day effect as compared to the maternity unit effect, and the RD of $\hat{V}_{SIMP2}$ is therefore large while that of $\hat{V}_{SIMP1}$ is small. Also, we note that the RD of $\hat{V}_{SIMP3}$ is relatively stable for all variables when estimating a total, which is an important feature in favour of this third simplified estimator. We note however that the absolute RD of $\hat{V}_{SIMP3}$ can be large when estimating a ratio, which confirms the simulation results.

2.6 Conclusion

The present paper derives some general estimation theory for the cross-classified sampling design which was used in the recent ELFE survey on childhood. The issue of possibly negative variance estimates may arise even in case of simple random sampling without replacement. Alternative estimators to the usual Horvitz-Thompson and Yates-Grundy variance estimators are thus proposed, and proved to be non-negative under the usual Sen-Yates-Grundy conditions. The relative bias of the proposed variance estimators is derived for a superpopulation model. The behavior of these estimators is also investigated for totals and ratios on simulated data and on data extracted from the ELFE survey. Among the proposals, one variance estimator that leads to a slight overestimation of the variance in many cases, appears to be advisable.

Despite the present results and the recent paper by Skinner [2015], the cross-classified sampling design still deserves some attention. In particular, the treatment of non-response and the calibration should also be taken into account, and is currently under investigation.

2.7 Supplementary Materials

Basic functions required to calculate estimators and commands that display the results in Table 2.1 and Table 2.2 are available in Appendix A.

2.8 Proof of equations (2.3.10)-(2.3.12)

We can rewrite

$$V_1\left(\hat{t}_Y\right) = \sum_{k\in U_D} \frac{V(\hat{Y}_{\bullet k})}{\pi_k^D} + \sum_{k\neq l\in U_D} \frac{\pi_{kl}^D}{\pi_k^D\pi_l^D} Cov(\hat{Y}_{\bullet k}, \hat{Y}_{\bullet l}). \tag{2.8.1}$$

We have

$$V(\hat{Y}_{\bullet k}) = \sum_{i\in U_M} (1-\pi_i^M)\frac{(Y_{ik})^2}{\pi_i^M} + \sum_{i\neq j\in U_M} \frac{\pi_{ij}^M - \pi_i^M\pi_j^M}{\pi_i^M\pi_j^M} Y_{ik}Y_{jk}. \tag{2.8.2}$$

From assumptions (H1), (H2) (H3) and Cauchy-Schwarz inequality, there exists some constant C such that for any $k \in U_D$,

$$V(\hat{Y}_{\bullet k}) \leq C\frac{N_M^2}{n_M}. \tag{2.8.3}$$

Also, from the Cauchy-Schwarz inequality, there exists some constant C such that for any $k \neq l \in U_D$:

$$Cov(\hat{Y}_{\bullet k}, \hat{Y}_{\bullet l}) \leq C\frac{N_M^2}{n_M}. \tag{2.8.4}$$

From equation (2.8.3) and assumption (H2), the first term in the right hand sum of (2.8.1) is $O(N_D^2 N_M^2 n_M^{-1} n_D^{-1})$. From equation (2.8.4) and assumptions (H2) and (H3), the absolute value of the second term in the RHS of (2.8.1) is $O(N_D^2 N_M^2 n_M^{-1})$. Therefore, there exists some constant C such that

$$V_1\left(\hat{t}_Y\right) \leq C\frac{N_D^2 N_M^2}{n_M}. \tag{2.8.5}$$

We can prove similarly that there exists some constant C such that

$$V_2\left(\hat{t}_Y\right) \leq C\frac{N_D^2 N_M^2}{n_D}. \tag{2.8.6}$$

From equation (2.2.13), the term $V_3\left(\hat{t}_Y\right)$ may be split into four terms according to the intersection of $\{i,j\}$ and $\{k,l\}$. From assumptions (H1)-(H3), it is easily shown that the absolute value of each of these four terms is $O(N_D^2 N_M^2 n_M^{-1} n_D^{-1})$. Therefore, there exists some constant C such that

$$V_3\left(\hat{t}_Y\right) \leq C\frac{N_D^2 N_M^2}{n_M n_D}. \tag{2.8.7}$$

Equations (2.3.10)-(2.3.12) follow immediately from equations (2.8.5)-(2.8.7) and assumption (H4).

References

Cardot, H., Goga, C., and Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic Journal of Statistics*, 7 :562–596. 50

Chen, J. and Rao, J. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, 17 :1047–1064. 54, 55

Dalén, J. and Ohlsson, E. (1995). Variance estimation in the Swedish consumer price index. *Journal of Business & Economic Statistics*, 13(3) :347–356. 42, 46, 49

Hajek, J. (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. *The Annals of Mathematical Statistics*, 32 :506–523. 55

Ohlsson, E. (1996). Cross-classified sampling. *Journal of Official Statistics*, 12(3) :241–251. 42, 46

Pirus, C., Bois, C., Dufourg, M., Lanoë, J., Vandentorren, S., Leridon, H., and the Elfe team (2010). Constructing a cohort : Experience with the French Elfe project. *Population*, 65 :637–670. 59

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York. 44

Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of Indian Society for Agricultural Statistics*, 5 :119–127. 44

Skinner, C. (2015). Cross-classified sampling : some estimation theory. *Statistics and Probability Letters*, 104 :163–168. 42, 46, 49, 52, 63

Vos, J. (1964). Sampling in space and time. *Review of the International Statistical Institute*, 32(3) :226–241. 42

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, 15 :235–261. 44

Chapitre 3

Estimation under cross-classified sampling, continuation

Le plan d'échantillonnage produit est le produit de deux plans de sondage indépendants. Dans ce chapitre, nous considérons l'estimateur Horvitz-Thompson d'un total et différentes décompositions de sa variance. Dans le cas d'un plan de taille fixe, ces décompositions nous mènent à cinq estimateurs de variance de la forme Yates-Grundy. On propose une étude par simulations pour comparer ces différents estimateurs dans le cas où les deux plans de sondages sont sans remise et avec probabilités proportionnelles à la taille. Les plans considérés sont le plan de Poisson conditionnel, le plan de Sampford et celui de Midzuno.

***Mots-clés** : échantillonnage produit, estimateur Horvitz-Thompson, estimateur Yates-Grundy, variance.*

Sommaire

Estimation under cross-classified sampling, continuation

The cross-classified sampling design is the product of two independent sampling designs. In the present chapter, we consider the Horvitz-Thompson estimator of a total and different decompositions of its variance. In the case of a fixed sample size design, these decompositions lead to five Yates-Grundy forms of variance estimators. We propose a simulation study in order to compare the different estimators when the two sampling designs are without replacement and with inclusion probabilities proportional to size. More precisely, we focus on conditional Poisson sampling, Sampford sampling and Midzuno sampling.

***Keywords** : cross-classified sampling, Horvitz-Thompson estimator, Yates-Grundy estimator, variance.*

3.1 Cross-classified sampling design

We consider a population U_M (of maternities) of size N_M, and a population U_D (of days) of size N_D. In what follows, we will use the indexes i and j for the maternities, and the indexes k and l for the days. We consider a sampling design $p_\text{M}(\cdot)$ on the population U_M, leading to a sample S_M of (average) size n_M, and a sampling design $p_\text{D}(\cdot)$ on the population U_D leading to a sample S_D of (average) size n_D. We assume that the two designs are independent (see Figure 3.1). This enables to define a sampling design $p(\cdot)$ called cross-classified sampling (CCS) design on the product population $\text{U} = \text{U}_\text{M} \times \text{U}_\text{D}$ as

$$p(s) = p_\text{M}(s_\text{M}) \times p_\text{D}(s_\text{D}) \quad \text{for any} \quad s = s_\text{M} \times s_\text{D} \subset \text{U}_\text{M} \times \text{U}_\text{D}.$$

We note π_i^M the probability that the maternity i is selected in S_M, and π_{ij}^M the probability that the maternities i and j are selected jointly in S_M. Similarly, we note π_k^D the

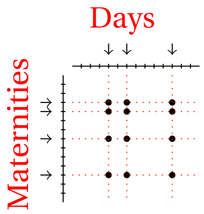


FIGURE 3.1 – Cross-classified sampling

probability that the day k is selected in S_D, and π^D_{kl} the probability that the days k and l are selected jointly in S_D. Because of independence between the two designs, the probability that the maternity i and the day k are selected is :

$$Pr\{(i,k) \in S_M \times S_D\} = \pi^M_i \pi^D_k, \tag{3.1.1}$$

and the second-order inclusion probabilities are :

$$Pr\{(i,k),(j,l) \in S_M \times S_D\} = \pi^M_{ij}\pi^D_{kl}. \tag{3.1.2}$$

In what follows, we assume that the first and second-order inclusion probabilities are strictly positive in each population. The total

$$t_Y = \sum_{i\in U_M}\sum_{k\in U_D} Y_{ik} \tag{3.1.3}$$

is then unbiasedly estimated by the expansion estimator

$$\hat{t}_Y = \sum_{i\in S_M}\sum_{k\in S_D}\frac{Y_{ik}}{\pi^M_i\pi^D_k}. \tag{3.1.4}$$

The estimator may be written as

$$\hat{t}_Y = \sum_{i\in S_M}\frac{\hat{Y}_{i\bullet}}{\pi^M_i} \quad \text{where} \quad \hat{Y}_{i\bullet} = \sum_{k\in S_D}\frac{Y_{ik}}{\pi^D_k}, \tag{3.1.5}$$

or

$$\hat{t}_{\mathrm{Y}} = \sum_{k \in \mathrm{S}_{\mathrm{D}}} \frac{\hat{\mathrm{Y}}_{\bullet k}}{\pi_k^{\mathrm{D}}} \quad \text{where} \quad \hat{\mathrm{Y}}_{\bullet k} = \sum_{i \in \mathrm{S}_{\mathrm{M}}} \frac{\mathrm{Y}_{ik}}{\pi_i^{\mathrm{M}}}. \tag{3.1.6}$$

3.2 Variance estimators for fixed-size CCS

Making use of the independence between the sampling designs $p_{\mathrm{M}}(\cdot)$ and $p_{\mathrm{D}}(\cdot)$, we obtain

$$\mathbf{V}\left(\hat{t}_{\mathrm{Y}}\right) = \sum_{i,j \in \mathrm{U}_{\mathrm{M}}} \sum_{k,l \in \mathrm{U}_{\mathrm{D}}} \mathit{\Gamma}_{ijkl} \frac{\mathrm{Y}_{ik}}{\pi_i^{\mathrm{M}} \pi_k^{\mathrm{D}}} \frac{\mathrm{Y}_{jl}}{\pi_j^{\mathrm{M}} \pi_l^{\mathrm{D}}} \tag{3.2.1}$$

where

$$\begin{aligned} \mathit{\Gamma}_{ijkl} &\equiv \mathrm{Cov}\left(\mathbf{1}_{i \in \mathrm{S}_{\mathrm{M}}} \mathbf{1}_{k \in \mathrm{S}_{\mathrm{D}}}, \mathbf{1}_{j \in \mathrm{S}_{\mathrm{M}}} \mathbf{1}_{l \in \mathrm{S}_{\mathrm{D}}}\right) \\ &= \pi_{ij}^{\mathrm{M}} \pi_{kl}^{\mathrm{D}} - \pi_i^{\mathrm{M}} \pi_j^{\mathrm{M}} \pi_k^{\mathrm{D}} \pi_l^{\mathrm{D}}. \end{aligned} \tag{3.2.2}$$

Other decompositions are possible for the covariances :

$$\begin{aligned} \mathit{\Gamma}_{ijkl} &= \pi_{kl}^{\mathrm{D}} \mathit{\Delta}_{ij}^{\mathrm{M}} + \pi_{ij}^{\mathrm{M}} \mathit{\Delta}_{kl}^{\mathrm{D}} - \mathit{\Delta}_{ij}^{\mathrm{M}} \mathit{\Delta}_{kl}^{\mathrm{D}} && (3.2.3) \\ &= \mathit{\Delta}_{kl}^{\mathrm{D}} \pi_i^{\mathrm{M}} \pi_j^{\mathrm{M}} + \mathit{\Delta}_{ij}^{\mathrm{M}} \pi_k^{\mathrm{D}} \pi_l^{\mathrm{D}} + \mathit{\Delta}_{ij}^{\mathrm{M}} \mathit{\Delta}_{kl}^{\mathrm{D}} && (3.2.4) \\ &= \pi_{kl}^{\mathrm{D}} \mathit{\Delta}_{ij}^{\mathrm{M}} + \pi_i^{\mathrm{M}} \pi_j^{\mathrm{M}} \mathit{\Delta}_{kl}^{\mathrm{D}} && (3.2.5) \\ &= \pi_{ij}^{\mathrm{M}} \mathit{\Delta}_{kl}^{\mathrm{D}} + \pi_k^{\mathrm{D}} \pi_l^{\mathrm{D}} \mathit{\Delta}_{ij}^{\mathrm{M}}. && (3.2.6) \end{aligned}$$

A general unbiased variance estimator for $\mathbf{V}\left(\hat{t}_{\mathrm{Y}}\right)$ is

$$\hat{\mathbf{V}}_{\mathrm{HT}}\left(\hat{t}_{\mathrm{Y}}\right) = \sum_{i,j \in \mathrm{S}_{\mathrm{M}}} \sum_{k,l \in \mathrm{S}_{\mathrm{D}}} \frac{\mathit{\Gamma}_{ijkl}}{\pi_{ij}^{\mathrm{M}} \pi_{kl}^{\mathrm{D}}} \frac{\mathrm{Y}_{ik}}{\pi_i^{\mathrm{M}} \pi_k^{\mathrm{D}}} \frac{\mathrm{Y}_{jl}}{\pi_j^{\mathrm{M}} \pi_l^{\mathrm{D}}}. \tag{3.2.7}$$

We now consider the case when $p_{\mathrm{D}}(\cdot)$ and $p_{\mathrm{M}}(\cdot)$ are both fixed-size sampling designs, and we consider several Yates-Grundy like variance estimators. We make the following assumptions :

H1 : The Sen-Yates-Grundy conditions hold for the sampling design $p_{\mathrm{M}}(\cdot)$, i.e. $\varDelta_{ij}^{\mathrm{M}} \leq 0$ for any $i \neq j \in \mathrm{U_M}$.

H2 : The Sen-Yates-Grundy conditions hold for the sampling design $p_{\mathrm{D}}(\cdot)$, i.e. $\varDelta_{kl}^{\mathrm{D}} \leq 0$ for any $k \neq l \in \mathrm{U_D}$.

3.2.1 Yates-Grundy variance decompositions

We first note that the variance for the expansion estimator given in (3.2.1) may be rewritten as :

$$\mathbf{V}_{\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = -\frac{1}{2}\sum_{(i,k)\neq(j,l)\in\mathrm{U_M}\times\mathrm{U_D}} \varGamma_{ijkl}\left(\frac{\mathrm{Y}_{ik}}{\pi_i^{\mathrm{M}}\pi_k^{\mathrm{D}}} - \frac{\mathrm{Y}_{jl}}{\pi_j^{\mathrm{M}}\pi_l^{\mathrm{D}}}\right)^2. \tag{3.2.8}$$

Using respectively decompositions (3.2.3),(3.2.4),(3.2.5) and (3.2.6) of $\varGamma_{ijkl}$, four other decompositions appear :

$$\begin{aligned}
\mathbf{V}_{\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) &= \mathbf{V}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \mathbf{V}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) - \mathbf{V}_{3,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) && (3.2.9)\\
&= \mathbf{V}_{4,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \mathbf{V}_{5,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \mathbf{V}_{3,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) && (3.2.10)\\
&= \mathbf{V}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \mathbf{V}_{4,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) && (3.2.11)\\
&= \mathbf{V}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \mathbf{V}_{5,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) && (3.2.12)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{V}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) &= -\frac{1}{2}\sum_{(i,k)\neq(j,l)\in\mathrm{U_M}\times\mathrm{U_D}} \varDelta_{ij}^{\mathrm{M}}\pi_{kl}^{\mathrm{D}}\left(\frac{\mathrm{Y}_{ik}}{\pi_i^{\mathrm{M}}\pi_k^{\mathrm{D}}} - \frac{\mathrm{Y}_{jl}}{\pi_j^{\mathrm{M}}\pi_l^{\mathrm{D}}}\right)^2 && (3.2.13)\\
\mathbf{V}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) &= -\frac{1}{2}\sum_{(i,k)\neq(j,l)\in\mathrm{U_M}\times\mathrm{U_D}} \varDelta_{kl}^{\mathrm{D}}\pi_{ij}^{\mathrm{M}}\left(\frac{\mathrm{Y}_{ik}}{\pi_i^{\mathrm{M}}\pi_k^{\mathrm{D}}} - \frac{\mathrm{Y}_{jl}}{\pi_j^{\mathrm{M}}\pi_l^{\mathrm{D}}}\right)^2 && (3.2.14)\\
\mathbf{V}_{3,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) &= -\frac{1}{2}\sum_{(i,k)\neq(j,l)\in\mathrm{U_M}\times\mathrm{U_D}} \varDelta_{kl}^{\mathrm{D}}\varDelta_{ij}^{\mathrm{M}}\left(\frac{\mathrm{Y}_{ik}}{\pi_i^{\mathrm{M}}\pi_k^{\mathrm{D}}} - \frac{\mathrm{Y}_{jl}}{\pi_j^{\mathrm{M}}\pi_l^{\mathrm{D}}}\right)^2 && (3.2.15)
\end{aligned}$$

$$\mathbf{V}_{4,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = -\frac{1}{2}\sum_{(i,k)\neq(j,l)\in \mathrm{U_M}\times\mathrm{U_D}} \mathit{\Delta}^{\mathrm{D}}_{kl}\pi^{\mathrm{M}}_{i}\pi^{\mathrm{M}}_{j}\left(\frac{\mathrm{Y}_{ik}}{\pi^{\mathrm{M}}_{i}\pi^{\mathrm{D}}_{k}} - \frac{\mathrm{Y}_{jl}}{\pi^{\mathrm{M}}_{j}\pi^{\mathrm{D}}_{l}}\right)^2 \quad (3.2.16)$$

$$\mathbf{V}_{5,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = -\frac{1}{2}\sum_{(i,k)\neq(j,l)\in \mathrm{U_M}\times\mathrm{U_D}} \mathit{\Delta}^{\mathrm{M}}_{ij}\pi^{\mathrm{D}}_{k}\pi^{\mathrm{D}}_{l}\left(\frac{\mathrm{Y}_{ik}}{\pi^{\mathrm{M}}_{i}\pi^{\mathrm{D}}_{k}} - \frac{\mathrm{Y}_{jl}}{\pi^{\mathrm{M}}_{j}\pi^{\mathrm{D}}_{l}}\right)^2. \quad (3.2.17)$$

These different expressions of variance are all equal. Decompositions similar to (3.2.9) and (3.2.10) have been proposed with an Horvitz-Thompson form in Chapter 2 (see equations (2.2.10) and (2.2.14)).

3.2.2 Yates-Grundy variance estimators

An unbiased variance estimator for the decomposition in (3.2.8) is

$$\hat{\mathbf{V}}_{a,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = -\frac{1}{2}\sum_{(i,k)\neq(j,l)\in \mathrm{S_M}\times\mathrm{S_D}} \frac{\mathit{\Gamma}_{ijkl}}{\pi^{\mathrm{M}}_{ij}\pi^{\mathrm{D}}_{kl}}\left(\frac{\mathrm{Y}_{ik}}{\pi^{\mathrm{M}}_{i}\pi^{\mathrm{D}}_{k}} - \frac{\mathrm{Y}_{jl}}{\pi^{\mathrm{M}}_{j}\pi^{\mathrm{D}}_{l}}\right)^2. \quad (3.2.18)$$

The estimator $\hat{\mathbf{V}}_{a,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$ is not necessary identical to $\hat{\mathbf{V}}_{\mathrm{HT}}\left(\hat{t}_{\mathrm{Y}}\right)$ but making use of the identity

$$\hat{\mathbf{V}}_{a,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = \hat{\mathbf{V}}_{\mathrm{HT}}\left(\hat{t}_{\mathrm{Y}}\right) - \sum_{i,k\in \mathrm{S_M}\times\mathrm{S_D}}\left(\frac{\mathrm{Y}_{ik}}{\pi^{\mathrm{M}}_{i}\pi^{\mathrm{D}}_{k}}\right)^2 \sum_{j,l\in \mathrm{S_M}\times\mathrm{S_D}} \frac{\mathit{\Gamma}_{ijkl}}{\pi^{\mathrm{M}}_{ij}\pi^{\mathrm{D}}_{kl}}, \quad (3.2.19)$$

it can be shown that both estimators match in the case when $p_{\mathrm{D}}(\cdot)$ and $p_{\mathrm{M}}(\cdot)$ are simple random sampling without replacement (SI) designs or stratified simple random sampling without replacement (STSI) designs.

The variance estimator $\hat{\mathbf{V}}_{a,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$ can take negative values, even if assumptions **H1** and **H2** are satisfied. Indeed, under these two assumptions, we have

$$\mathit{\Gamma}_{ijkl} \leq 0 \quad \text{for any } i \neq j \in \mathrm{U_M} \text{ and } k \neq l \in \mathrm{U_D}, \quad (3.2.20)$$

but in the case when $i = j$ and $k \neq l$ we have $\mathit{\Gamma}_{ijkl} = \pi^{\mathrm{M}}_{i}(\pi^{\mathrm{D}}_{kl} - \pi^{\mathrm{M}}_{i}\pi^{\mathrm{D}}_{k}\pi^{\mathrm{D}}_{l})$, which may take positive values, particularly if the inclusion probability π^{M}_{i} is small.

A term by term unbiased variance estimator for the decomposition in (3.2.9) is

$$\hat{\mathbf{V}}_{b,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = \hat{\mathbf{V}}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \hat{\mathbf{V}}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) - \hat{\mathbf{V}}_{3,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) \tag{3.2.21}$$

where

$$\hat{\mathbf{V}}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = -\frac{1}{2}\sum_{i\neq j\in \mathrm{S_M}} \frac{\Delta^{\mathrm{M}}_{ij}}{\pi^{\mathrm{M}}_{ij}}\left(\frac{\hat{\mathrm{Y}}_{i\bullet}}{\pi^{\mathrm{M}}_{i}} - \frac{\hat{\mathrm{Y}}_{j\bullet}}{\pi^{\mathrm{M}}_{j}}\right)^2 \tag{3.2.22}$$

$$\hat{\mathbf{V}}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = -\frac{1}{2}\sum_{k\neq l\in \mathrm{S_D}} \frac{\Delta^{\mathrm{D}}_{kl}}{\pi^{\mathrm{D}}_{kl}}\left(\frac{\hat{\mathrm{Y}}_{\bullet k}}{\pi^{\mathrm{D}}_{k}} - \frac{\hat{\mathrm{Y}}_{\bullet l}}{\pi^{\mathrm{D}}_{l}}\right)^2 \tag{3.2.23}$$

$$\hat{\mathbf{V}}_{3,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = -\frac{1}{2}\sum_{(i,k)\neq(j,l)\in \mathrm{S_M}\times\mathrm{S_D}} \frac{\Delta^{\mathrm{M}}_{ij}\Delta^{\mathrm{D}}_{kl}}{\pi^{\mathrm{M}}_{ij}\pi^{\mathrm{D}}_{kl}}\left(\frac{\mathrm{Y}_{ik}}{\pi^{\mathrm{M}}_{i}\pi^{\mathrm{D}}_{k}} - \frac{\mathrm{Y}_{jl}}{\pi^{\mathrm{M}}_{j}\pi^{\mathrm{D}}_{l}}\right)^2. \tag{3.2.24}$$

A similar estimator has already been presented in the Chapter 2, equation (2.3.6). Under assumption **H1**, $\hat{\mathbf{V}}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$ is always non-negative and under assumption **H2**, $\hat{\mathbf{V}}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$ is always non-negative, which is usually not true for $\hat{\mathbf{V}}_{3,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$ even if **H1** and **H2** hold.

A term by term unbiased variance estimator for the decomposition in (3.2.10) is

$$\hat{\mathbf{V}}_{c,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = \hat{\mathbf{V}}_{4,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \hat{\mathbf{V}}_{5,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \hat{\mathbf{V}}_{3,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) \tag{3.2.25}$$

where

$$\hat{\mathbf{V}}_{4,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = -\frac{1}{2}\sum_{k,l\in \mathrm{S_D}} \frac{\Delta^{\mathrm{D}}_{kl}}{\pi^{\mathrm{D}}_{kl}}\sum_{i,j\in \mathrm{S_M}}\frac{1}{\pi^{\mathrm{M}}_{ij}}\left(\frac{\mathrm{Y}_{ik}}{\pi^{\mathrm{D}}_{k}} - \frac{\mathrm{Y}_{jl}}{\pi^{\mathrm{D}}_{l}}\right)^2 \tag{3.2.26}$$

$$\hat{\mathbf{V}}_{5,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = -\frac{1}{2}\sum_{i,j\in \mathrm{S_M}} \frac{\Delta^{\mathrm{M}}_{ij}}{\pi^{\mathrm{M}}_{ij}}\sum_{k,l\in \mathrm{S_D}}\frac{1}{\pi^{\mathrm{D}}_{kl}}\left(\frac{\mathrm{Y}_{ik}}{\pi^{\mathrm{M}}_{i}} - \frac{\mathrm{Y}_{jl}}{\pi^{\mathrm{M}}_{j}}\right)^2. \tag{3.2.27}$$

Under assumptions **H1** and **H2**, $\hat{\mathbf{V}}_{4,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$ and $\hat{\mathbf{V}}_{5,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$ are not necessarily non-negative.

Similarly, a term by term unbiased variance estimator for the decomposition in (3.2.11) is

$$\hat{\mathbf{V}}_{d,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = \hat{\mathbf{V}}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \hat{\mathbf{V}}_{4,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right), \tag{3.2.28}$$

and a last term by term unbiased variance estimator for the decomposition in (3.2.12) is

$$\hat{\mathbf{V}}_{e,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = \hat{\mathbf{V}}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \hat{\mathbf{V}}_{5,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right). \tag{3.2.29}$$

All decompositions for $\mathit{\Gamma}_{ijkl}$ and unbiased variance estimators are listed in Table 3.1. None of the five estimators can be assured non-negative, despite the assumptions **H1** and **H2**. In the particular case where $p_{\mathrm{M}}(\cdot)$ and $p_{\mathrm{D}}(\cdot)$ are stratified simple random samplings, each variance estimator matches with the Horvitz-Thompson variance estimator in (3.2.7). In the next section, we will look at sampling designs with inclusion probabilities proportional to size (πps).

$\mathit{\Gamma}_{ijkl} = \pi^{\mathrm{M}}_{ij}\pi^{\mathrm{D}}_{kl} - \pi^{\mathrm{M}}_{i}\pi^{\mathrm{M}}_{j}\pi^{\mathrm{D}}_{k}\pi^{\mathrm{D}}_{l}$ $\hat{\mathrm{V}}_{a,\mathrm{YG}}$	3.2.18
$\mathit{\Gamma}_{ijkl} = \pi^{\mathrm{D}}_{kl}\mathit{\Delta}^{\mathrm{M}}_{ij} + \pi^{\mathrm{M}}_{ij}\mathit{\Delta}^{\mathrm{D}}_{kl} - \mathit{\Delta}^{\mathrm{M}}_{ij}\mathit{\Delta}^{\mathrm{D}}_{kl}$ $\hat{\mathrm{V}}_{b,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = \hat{\mathrm{V}}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \hat{\mathrm{V}}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) - \hat{\mathrm{V}}_{3,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$	3.2.21
$\mathit{\Gamma}_{ijkl} = \mathit{\Delta}^{\mathrm{D}}_{kl}\pi^{\mathrm{M}}_{i}\pi^{\mathrm{M}}_{j} + \mathit{\Delta}^{\mathrm{M}}_{ij}\pi^{\mathrm{D}}_{k}\pi^{\mathrm{D}}_{l} + \mathit{\Delta}^{\mathrm{M}}_{ij}\mathit{\Delta}^{\mathrm{D}}_{kl}$ $\hat{\mathrm{V}}_{c,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = \hat{\mathrm{V}}_{4,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \hat{\mathrm{V}}_{5,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \hat{\mathrm{V}}_{3,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$	3.2.25
$\mathit{\Gamma}_{ijkl} = \pi^{\mathrm{D}}_{kl}\mathit{\Delta}^{\mathrm{M}}_{ij} + \pi^{\mathrm{M}}_{i}\pi^{\mathrm{M}}_{j}\mathit{\Delta}^{\mathrm{D}}_{kl}$ $\hat{\mathrm{V}}_{d,\mathrm{YG}} = \hat{\mathrm{V}}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \hat{\mathrm{V}}_{4,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$	3.2.28
$\mathit{\Gamma}_{ijkl} = \pi^{\mathrm{M}}_{ij}\mathit{\Delta}^{\mathrm{D}}_{kl} + \pi^{\mathrm{D}}_{k}\pi^{\mathrm{D}}_{l}\mathit{\Delta}^{\mathrm{M}}_{ij}$ $\hat{\mathrm{V}}_{e,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) = \hat{\mathrm{V}}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right) + \hat{\mathrm{V}}_{5,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$	3.2.29

TABLE 3.1 – Five different variance estimators

3.3 πps cross-classified sampling

In what follows, we look at sampling methods without replacement, with fixed size and unequal probabilities. We focus for both designs $p_{\mathrm{M}}(.)$ and $p_{\mathrm{D}}(.)$ on probability proportional to size sampling without replacement (πps). More precisely, we consider conditional Poisson, Sampford and Midzuno sampling designs. A sample s_{M} (resp. s_{D}) of fixed size n_{M} (resp. n_{D}) is drawn from population U_{M} (resp. U_{D}) using a πps design.

In the literature on CCS, Dalén and Ohlsson [1995] proposed to estimate the Swedish Consumer Price Index and its variance when both designs p_{M} and p_{D} are random systematic sampling [Tillé, 2006]. For this particular unequal probability sampling design, the authors give an approximate variance formula and derive a variance estimator for the price index. Skinner [2015] considers with replacement sampling with possibly unequal probabilities for p_{M} and p_{D}. He defines a natural unbiased estimator of a total by extending the Hansen-Hurwitz estimator, calculates its variance and derives an unbiased variance estimator.

To define a πps CCS design, we need an auxiliary characteristic denoted by X_{ik} and known for all $i \in \mathrm{U}_{\mathrm{M}}$ and $k \in \mathrm{U}_{\mathrm{D}}$. The probabilities π_i^{M} (resp. π_k^{D}) are taken proportional to $\mathrm{X}_{i\bullet} = \sum_{k \in \mathrm{U}_{\mathrm{D}}} \mathrm{X}_{ik}$ (resp. $\mathrm{X}_{\bullet k} = \sum_{i \in \mathrm{U}_{\mathrm{M}}} \mathrm{X}_{ik}$) and are calculated as follow :

$$\pi_i^{\mathrm{M}} = n_{\mathrm{M}} \times \frac{\mathrm{X}_{i\bullet}}{\mathrm{X}_{\bullet\bullet}} \quad \text{and} \quad \pi_k^{\mathrm{D}} = n_{\mathrm{D}} \times \frac{\mathrm{X}_{\bullet k}}{\mathrm{X}_{\bullet\bullet}}, \tag{3.3.1}$$

where $\mathrm{X}_{\bullet\bullet} = \sum_{i \in \mathrm{U}_{\mathrm{M}}} \sum_{k \in \mathrm{U}_{\mathrm{D}}} \mathrm{X}_{ik}$. Note that if an inclusion probability is larger than 1, then it is taken equal to 1 and other probabilities are recalculated using the remaining population.

In order to compare the different Yates-Grundy variance estimators defined in section 3.2.2, we perform a simulation study using three different πps designs that are implemented in the R package *sampling* [Tillé and Matei, 2015]. More precisely we focus on conditional Poisson, Sampford and Midzuno sampling designs Tillé [2006]. We briefly describe the three designs in the next subsections using generic notations, namely U for a population of size N, s for a sample of size n, $p(.)$ for the sampling design and π_k (resp. π_{kl}) for the associated first (resp. second) order inclusion probabilities.

3.3.1 Conditional Poisson sampling

The conditional Poisson sampling design was introduced by Hájek [1964] and is also called rejective sampling design or maximum entropy sampling design with fixed sample size. It consists in drawing independently n units with given probabilities p_k and rejecting the sample until the sample size $n(s)$ equals n.

For this design, the probability of a sample s is

$$p(s) \quad = \quad \begin{cases} c\prod_{k\in s} p_k \prod_{k\notin s}(1-p_k) & \text{if } \ n(s) = n, \\ 0 & \text{otherwise,} \end{cases} \tag{3.3.2}$$

where c is a constant calculated in order to respect $\sum_{s\in S} p(s) = 1$. The first order probabilities π_k are different from the initial probabilities p_k but can be calculated in terms of the p_k [Hájek, 1964]. Exact second order inclusion probabilities π_{kl} can also be computed by using some recursive formula [Tillé, 2006].

For the simulations in the next section, we use the R package *sampling* [Tillé and Matei, 2015]. For rejective sampling, the functions *UPmaxentropy* and *UPmaxentropypi2* are used. In the function *UPmaxentropy*, a sequential procedure is used to select a conditional Poisson sample [Tillé, 2006]. The function *UPmaxentropypi2* returns exactly the joint inclusion probabilities from the first-order inclusion probability vector.

3.3.2 Sampford sampling

The Sampford sampling design was proposed by Sampford [1967]. The Sampford implementation consists on drawing a unit with replacement with given probabilities $p_k = \pi_k / n$, $k = 1, ..., \mathrm{N}$ and then $n-1$ units with replacement according to the probabilities p_k proportional to $\pi_k / (1-\pi_k)$ under the constraint $\sum_{i=1}^{\mathrm{N}} p_k = 1$. The sample is accepted only if the n units are distinct, otherwise the procedure is repeated.
For this design, the probability of a sample s is

$$p(s) \quad = \quad \begin{cases} c \prod_{k\in\mathrm{U}} p_k^{\mathbf{1}_k} \left(1-p_k\right)^{1-\mathbf{1}_k} \sum_{k\in\mathrm{U}} \left(1-p_k\right) \mathbf{1}_k & \text{if } \ n(s) = n \\ 0 & \text{otherwise} \end{cases} \qquad (3.3.3)$$

where $\mathbf{1}_k$ equals 1 if $k \in s$ and 0 otherwise, and c is a constant calculated in order to respect $\sum_s p(s) = 1$. This drawing process leads to first order inclusion probabilities equal to π_k.

In order to draw a sample with the Sampford design, the functions *UPsampford* and *UPsampfordpi2* from the R package *sampling* [Tillé and Matei, 2015] can be used. In the function *UPsampford*, the multinomial rejective Sampford's procedure is used [Tillé, 2006]. The function *UPsampfordpi2* computes exactly the second order inclusion probabilities for Sampford sampling.

3.3.3 Midzuno sampling

The first algorithm was presented by Midzuno [1952], it consists in drawing one unit with probabilities p_k, $k = 1, ..., \mathrm{N}$ and then using a simple random sampling without replacement of $n-1$ units from the remaining $\mathrm{N}-1$ units. The probability of selection of a sample s is

$$p(s) \quad = \quad \begin{cases} \frac{\sum_{k\in s} p_k}{\binom{\mathrm{N}-1}{n-1}} & \text{if } \ n(s) = n \\ 0 & \text{otherwise.} \end{cases} \qquad (3.3.4)$$

and the first order inclusion probabilities are $\pi_k = \frac{n-1}{N-1} + p_k \frac{N-n}{N-1}$ for $k \in U$. This method implies restrictive constraints ($\pi_k \geq \frac{n-1}{N-1}$, for all $k \in U$). In Deville and Tillé [1998], the method is described as a splitting procedure and is generalized to any set of inclusion probabilities π_k.

In the R package *sampling* [Tillé and Matei, 2015], generalized Midzuno sampling can be used with functions *UPmidzuno* and *UPmidzunopi2*. The algorithm in *UPmidzuno* uses the Tillé's elimination procedure [Tillé, 2006]. The procedure *UPmidzunopi2* returns exactly the joint inclusion probabilities for Midzuno sampling.

3.4 Simulations

A simulation study is conducted in order to compare the performance of the five variance estimators under a πps CCS design. The interest variable Y and the auxiliary variable X are generated using the model :

$$\begin{aligned} X_{ik} &= 200 + \sigma_1 U_i + \sigma_2 V_k + 5W_{ik} \\ Y_{ik} &= \mu + X_{ik} + 5\tilde{W}_{ik} \end{aligned} \tag{3.4.1}$$

where U_i, V_k, W_{ik} and $\tilde{W}_{ik}$ are independently generated using a standard normal distribution, $i = 1, \ldots, 100$ and $k = 1, \ldots, 100$. We use each of the three sampling designs presented previously for p_M and p_D. We recall that for p_M (resp. p_D), inclusion probabilities are taken proportional to the variable $X_{i\bullet} = \sum_{k \in U_D} X_{ik}$ (resp. $X_{\bullet k} = \sum_{i \in U_M} X_{ik}$). We select independently a sample S_D of size n_D, and a sample S_M of size n_M and this sample selection is repeated $B = 10,000$ times. In each of the $\tilde{b} = 1, \ldots, B$, samples, we compute the estimators $\hat{\mathbf{V}}^{(\tilde{b})}_{a,YG}(\hat{t}_Y)$, $\hat{\mathbf{V}}^{(\tilde{b})}_{b,YG}(\hat{t}_Y)$, $\hat{\mathbf{V}}^{(\tilde{b})}_{c,YG}(\hat{t}_Y)$, $\hat{\mathbf{V}}^{(\tilde{b})}_{d,YG}(\hat{t}_Y)$ and $\hat{\mathbf{V}}^{(\tilde{b})}_{e,YG}(\hat{t}_Y)$ and the number of negative values (#NEG) for each estimator. For each variance estimator, the coefficient of variation is calculated in percentage by :

$$\mathrm{CV}\left(\hat{\mathbf{V}}_{q,\mathrm{YG}}\right) = 100 \times \frac{\sqrt{\mathrm{B}^{-1}\sum_{\tilde{b}=1}^{\mathrm{B}}\left(\hat{\mathbf{V}}_{q,\mathrm{YG}}^{(\tilde{b})} - \mathrm{V}\right)^2}}{\mathrm{V}} \tag{3.4.2}$$

for $q = a, b, c, d, e$ and where the true variance V was approximated through an independent set of 50,000 simulations.

We generate different populations using the model (3.4.1) and varying the parameter μ, the maternity unit effect σ_1 and the day effect σ_2. More precisely, we consider two values for the μ parameter : $\mu = 200$ for case 1 in the first subsection below and $\mu = 0$ for case 2 in the second subsection. For each value of μ, we generate 3 populations : POP 1 for $\sigma_1 = 0.5$ and $\sigma_2 = 5$, POP 2 for $\sigma_1 = 5$ and $\sigma_2 = 5$ and POP 3 for $\sigma_1 = 5$ and $\sigma_2 = 50$. The sample sizes n_D and n_M vary from 5 to 50. Only results for the Midzuno design are presented in the main body of the text. Results for the Sampford sampling and the conditional Poisson sampling are very similar and are therefore postponed to the section 3.7. Note that for Sampford sampling, we make the size vary from 5 to 20 only, because the algorithm does not converge for the sample size 50. In terms of computing time, the procedures for the Midzuno sampling and for the Sampford sampling are more efficient than the Poisson conditional sampling which takes 2 to 40 times longer than the other two in our simulation setup.

Under the same simulation framework, we also tested three simplified variance estimators equivalent to those of Chapter 2 :

$$\hat{\mathrm{V}}_{\mathrm{SIMP1}}\left(\hat{t}_\mathrm{Y}\right) = \hat{\mathrm{V}}_{1,\mathrm{YG}}\left(\hat{t}_\mathrm{Y}\right) \tag{3.4.3}$$

$$\hat{\mathrm{V}}_{\mathrm{SIMP2}}\left(\hat{t}_\mathrm{Y}\right) = \hat{\mathrm{V}}_{2,\mathrm{YG}}\left(\hat{t}_\mathrm{Y}\right) \tag{3.4.4}$$

$$\hat{\mathrm{V}}_{\mathrm{SIMP3}}\left(\hat{t}_\mathrm{Y}\right) = \hat{\mathrm{V}}_{1,\mathrm{YG}}\left(\hat{t}_\mathrm{Y}\right) + \hat{\mathrm{V}}_{2,\mathrm{YG}}\left(\hat{t}_\mathrm{Y}\right). \tag{3.4.5}$$

The variance estimator $\hat{\mathrm{V}}_{1,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$ (resp. $\hat{\mathrm{V}}_{2,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$) has the advantage of being always non-negative if the sampling design p_{M} (resp. p_{D}) verifies the Sen-Yates-Grundy conditions. If both p_{M} and p_{D} satisfy the Sen-Yates-Grundy conditions, the third simplified estimator is always non-negative. The Monte Carlo Percent Relative Bias is computed for each of the three simplified variance estimators :

$$\mathrm{RB}_{\mathrm{MC}}\left(\hat{\mathbf{V}}_{\mathrm{SIMP}q}\right) = 100 \times \frac{\mathrm{B}^{-1}\sum_{\tilde{b}=1}^{\mathrm{B}} \hat{\mathbf{V}}_{\mathrm{SIMP}q}^{(b)} - \mathrm{V}}{\mathrm{V}} \qquad (3.4.6)$$

where $q = 1,2,3$ and the true variance V was approximated through an independent set of 50,000 simulations. Note that we do not report the relative bias of the five estimators $\hat{\mathbf{V}}_{a,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$, $\hat{\mathbf{V}}_{b,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$, $\hat{\mathbf{V}}_{c,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$, $\hat{\mathbf{V}}_{d,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$ and $\hat{\mathbf{V}}_{e,\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$ which are all unbiased estimators of $\mathbf{V}_{\mathrm{YG}}\left(\hat{t}_{\mathrm{Y}}\right)$. However we checked that these relative biases are usually less than 2% and not larger than 4%.

We recall that the results are splitted in two cases with the constant μ equal to 200 for case 1, and equal to 0 for case 2. The correlations between the variable of interest and the first order inclusion probabilities are given in Table 3.2 together with the coefficient of determination R^2 of the linear model where Y_{ik} is explained by X_{ik}. They are the same for both cases and depend on the choice of the parameters σ_1 and σ_2. In particular we note that the correlation between Y_{ik} and π_k^{D} and the R^2 are almost equal to 1 for POP 3. Moreover, case 2 is very specific since it implies that the variable of interest is likely to be almost proportional to the auxiliary variable and thus to the first order inclusion probabilities. On the contrary, case 1 seems more realistic. In each case, we observe that results are very similar for the three sampling designs and so we only report results for the Midzuno design while results for conditional Poisson and Sampford samplings are to be found in section 3.7.

POP	σ_1	σ_2	$\mathrm{Cor}(Y_{ik},\pi_k^D)$	$\mathrm{Cor}(Y_{ik},\pi_i^M)$	$R^2(Y_{ik} \sim X_{ik})$
1	0.5	5	0.49	0.09	0.62
2	5	5	0.41	0.53	0.72
3	5	50	0.98	0.13	0.98

TABLE 3.2 – Correlations between Y_{ik} and the inclusion probabilities and the R^2 for both cases

3.4.1 Case 1 : $\mu = 200$

The results for the Midzuno sampling are given in Table 3.3 with n_M and n_D equal to 5, 10, 20 and 50. The coefficient of variation (as a percentage) of each variance estimators decreases when the sample sizes increase. Negative values #NEG appear in the POP 1 and 2 for small sample sizes only (when $n_M = n_D = 5$ and very rarely when $n_M = n_D = 10$) as observed for a similar simulation setup in Chapter 2. Moreover, there is no difference between the five variance estimators.

POP	σ_1	σ_2	n_M	n_D	$\hat{V}_{a,YG}(\hat{t}_Y)$		$\hat{V}_{b,YG}(\hat{t}_Y)$		$\hat{V}_{c,YG}(\hat{t}_Y)$		$\hat{V}_{d,YG}(\hat{t}_Y)$		$\hat{V}_{e,YG}(\hat{t}_Y)$	
					CV	#NEG	CV	#NEG	CV	#NEG	CV	#NEG	CV	#NEG
1	0.5	5	5	5	73	380	73	382	73	380	73	380	73	382
	0.5	5	10	10	43	4	43	4	43	4	43	4	43	4
	0.5	5	20	20	27	0	27	0	27	0	27	0	27	0
	0.5	5	50	50	13	0	13	0	13	0	13	0	13	0
2	5	5	5	5	57	52	57	52	57	51	57	52	57	51
	5	5	10	10	20	0	20	0	20	0	20	0	20	0
	5	5	20	20	19	0	19	0	19	0	19	0	19	0
	5	5	50	50	10	0	10	0	10	0	10	0	10	0
3	5	50	5	5	97	0	97	0	97	0	97	0	97	0
	5	50	10	10	48	0	48	0	48	0	48	0	48	0
	5	50	20	20	49	0	49	0	49	0	49	0	49	0
	5	50	50	50	31	0	31	0	32	0	32	0	31	0

TABLE 3.3 – Five Y-G variance estimators for Midzuno sampling in case 1

For the three simplified estimators, the results are presented in Table 3.4 for the three different populations. For each population, sample sizes n_{M} and n_{D} equal to 5, 20 and 50 are used with Midzuno sampling. The relative bias (RB) and coefficient of variation (CV) of each simplified estimator are given in percent. The RB of $\hat{\mathrm{V}}_{\mathrm{SIMP1}}$ is large in all situations. The RB of $\hat{\mathrm{V}}_{\mathrm{SIMP2}}$ decreases when σ_2 increases or when σ_1 decreases, as observed already in Chapter 2 for an equal probability design. The RB of the estimator $\hat{\mathrm{V}}_{\mathrm{SIMP3}}$ is small and decreases with the sample sizes. As already observed in Chapter 2, its RB also decreases when either σ_1 or σ_2 increases. Concerning the CV, it is comparable between the three estimators when n_{M} and n_{D} equal to 5. But when the sample sizes equal 50, the third simplified estimator outperforms the first simplified estimator and is equivalent to the second one for POP 1 and POP 3 and better for POP 2. The third population presents an important day effect with $\sigma_2 = 50$. In this case, $\hat{\mathrm{V}}_{\mathrm{SIMP1}}$ presents a very large RB, while $\hat{\mathrm{V}}_{\mathrm{SIMP2}}$ becomes an interesting estimator with a null RB. The third simplified estimator can be recommended in all the situations under study.

POP	σ_1	σ_2	n_{M}	n_{D}	$\hat{\mathrm{V}}_{\mathrm{SIMP1}}$		$\hat{\mathrm{V}}_{\mathrm{SIMP2}}$		$\hat{\mathrm{V}}_{\mathrm{SIMP3}}$	
					RB	CV	RB	CV	RB	CV
1	0.5	5	5	5	-60	67	-3	66	37	81
	0.5	5	20	20	-85	86	-3	27	11	29
	0.5	5	50	50	-93	93	-3	13	4	14
2	5	5	5	5	-30	55	-52	61	18	59
	5	5	20	20	-35	39	-61	62	5	20
	5	5	50	50	-36	37	-63	63	1	10
2	5	50	5	5	-98	98	-2	98	0	98
	5	50	20	20	-98	98	-1	48	0	98
	5	50	50	50	-98	98	0	31	1	31

TABLE 3.4 – Simplified variance estimators for Midzuno sampling in case 1

3.4.2 Case 2 : $\mu = 0$

Results are presented for Midzuno sampling in Table 3.5 for the three different populations. Except for the last population (POP 3), the five variance estimators present

POP	σ_1	σ_2	n_M	n_D	$\hat{\mathbf{V}}_{a,\mathrm{YG}}\left(\hat{t}_\mathrm{Y}\right)$		$\hat{\mathbf{V}}_{b,\mathrm{YG}}\left(\hat{t}_\mathrm{Y}\right)$		$\hat{\mathbf{V}}_{c,\mathrm{YG}}\left(\hat{t}_\mathrm{Y}\right)$		$\hat{\mathbf{V}}_{d,\mathrm{YG}}\left(\hat{t}_\mathrm{Y}\right)$		$\hat{\mathbf{V}}_{e,\mathrm{YG}}\left(\hat{t}_\mathrm{Y}\right)$	
					CV	#NEG	CV	#NEG	CV	#NEG	CV	#NEG	CV	#NEG
1	0.5	5	5	5	102	1506	102	1504	102	1506	102	1505	102	1505
	0.5	5	10	10	64	355	64	350	64	356	64	355	64	351
	0.5	5	20	20	41	14	41	16	41	14	41	14	41	15
	0.5	5	50	50	23	0	21	0	23	0	23	0	21	0
2	5	5	5	5	102	1471	102	1470	102	1472	102	1469	102	1470
	5	5	10	10	63	335	63	339	64	338	63	343	63	335
	5	5	20	20	41	19	41	16	41	22	41	19	41	19
	5	5	50	50	27	0	21	0	31	0	26	0	27	0
3	5	50	5	5	111	1604	106	1515	121	1779	117	1714	109	1570
	5	50	10	10	78	669	69	410	99	1200	86	850	85	774
	5	50	20	20	67	217	46	17	136	2021	83	541	117	1279
	5	50	50	50	118	1987	24	0	397	4350	170	3037	360	4151

TABLE 3.5 – Five Y-G variance estimators for Midzuno sampling in case 2

again the same behavior and have approximately the same CVs and #NEGs. When the random effects standard deviations σ_1 and σ_2 are small (POP 1 and POP 2), the number of negative values and the CV decrease when n_M and n_D increase. For the third population however, some differences between the five estimators show up. We observe that the CVs and #NEGs decrease when the sample sizes increase from 5 to 10 but this is not the case anymore for all estimators when the sample sizes grow from 10 to 20 and 20 to 50. Globally, $\hat{\mathbf{V}}_{c,\mathrm{YG}}$, $\hat{\mathbf{V}}_{d,\mathrm{YG}}$ and $\hat{\mathbf{V}}_{e,\mathrm{YG}}$ present the worst CVs and #NEGs. The estimator $\hat{\mathbf{V}}_{a,\mathrm{YG}}$ is slightly better, but only $\hat{\mathbf{V}}_{b,\mathrm{YG}}$ seems acceptable in term of CVs and #NEGs .

To understand better the differences between the five estimators for POP 3 ($\sigma_1 = 5, \sigma_2 = 50$), we construct the boxplots of each estimator as well as the boxplots of the components $\hat{\mathbf{V}}_{1,\mathrm{YG}}$, $\hat{\mathbf{V}}_{2,\mathrm{YG}}$, $\hat{\mathbf{V}}_{3,\mathrm{YG}}$, $\hat{\mathbf{V}}_{4,\mathrm{YG}}$ and $\hat{\mathbf{V}}_{5,\mathrm{YG}}$ under $\mathrm{B} = 10,000$ simulations for Midzuno sampling for POP 2 and POP 3 with $n_\mathrm{M} = n_\mathrm{D} = 20$ in Figure 3.3 and with $n_\mathrm{M} = n_\mathrm{D} = 50$ in Figure 3.4. The results are detailed in Tables 3.6 and 3.7. The contribution CONTR of each component of variance estimators is calculated as a percentage :

$$\mathrm{CONTR}\left(\hat{\mathbf{V}}_{q,\mathrm{YG}}\right) = 100 \times \frac{\hat{\mathbf{V}}_{q,\mathrm{YG}}}{\mathrm{V}} \quad \text{where} \quad q = 1,2,3,4,5 \tag{3.4.7}$$

where the true variance V was approximated through an independent set of 50,000 simulations.

$$\begin{tikzcd}
 & \hat{\mathbf{V}}_{1,\mathrm{YG}} \\
\hat{\mathbf{V}}_{b,\mathrm{YG}} \arrow[ur] \arrow[r] \arrow[dr] & \hat{\mathbf{V}}_{2,\mathrm{YG}} \\
\hat{\mathbf{V}}_{c,\mathrm{YG}} \arrow[r] \arrow[dr] \arrow[ddr] & \hat{\mathbf{V}}_{3,\mathrm{YG}} \\
\hat{\mathbf{V}}_{d,\mathrm{YG}} \arrow[uuur] \arrow[r] & \hat{\mathbf{V}}_{4,\mathrm{YG}} \\
\hat{\mathbf{V}}_{e,\mathrm{YG}} \arrow[uuur] \arrow[r] & \hat{\mathbf{V}}_{5,\mathrm{YG}}
\end{tikzcd}$$

FIGURE 3.2 – The five Y-G estimators and their associated decompositions

In Figure 3.2, we observe that the estimators $\hat{\mathbf{V}}_{c,\mathrm{YG}}$ and $\hat{\mathbf{V}}_{d,\mathrm{YG}}$ have in common the component $\hat{\mathbf{V}}_{4,\mathrm{YG}}$, while $\hat{\mathbf{V}}_{c,\mathrm{YG}}$ and $\hat{\mathbf{V}}_{e,\mathrm{YG}}$ have in common the component $\hat{\mathbf{V}}_{5,\mathrm{YG}}$. In Figures 3.3 and 3.4, we observe that $\hat{\mathbf{V}}_{4,\mathrm{YG}}$ and $\hat{\mathbf{V}}_{5,\mathrm{YG}}$ take values around zero. The variability of $\hat{\mathbf{V}}_{5,\mathrm{YG}}$ increases with σ_2 which contributes to increase the variability of $\hat{\mathbf{V}}_{c,\mathrm{YG}}$ and $\hat{\mathbf{V}}_{e,\mathrm{YG}}$, and the variability of $\hat{\mathbf{V}}_{4,\mathrm{YG}}$ increases also to a lesser extent, contributing to increase the variability of $\hat{\mathbf{V}}_{c,\mathrm{YG}}$ and $\hat{\mathbf{V}}_{d,\mathrm{YG}}$. Note that for reasons of symmetry, the same sort of results could be derived when increasing σ_1.

$\hat{\mathrm{V}}$	$\hat{\mathbf{V}}_{a,\mathrm{YG}}$	$\hat{\mathbf{V}}_{b,\mathrm{YG}}$	$\hat{\mathbf{V}}_{c,\mathrm{YG}}$	$\hat{\mathbf{V}}_{d,\mathrm{YG}}$	$\hat{\mathbf{V}}_{e,\mathrm{YG}}$
#NEG	19	16	22	19	19
CV($\hat{\mathrm{V}}$)	41	41	41	41	41
$\hat{\mathrm{V}}$	$\hat{\mathbf{V}}_{1,\mathrm{YG}}$	$\hat{\mathbf{V}}_{2,\mathrm{YG}}$	$\hat{\mathbf{V}}_{3,\mathrm{YG}}$	$\hat{\mathbf{V}}_{4,\mathrm{YG}}$	$\hat{\mathbf{V}}_{5,\mathrm{YG}}$
CONT($\hat{\mathrm{V}}$)	87	90	78	13	10
#NEG	0	0	0	3645	4015

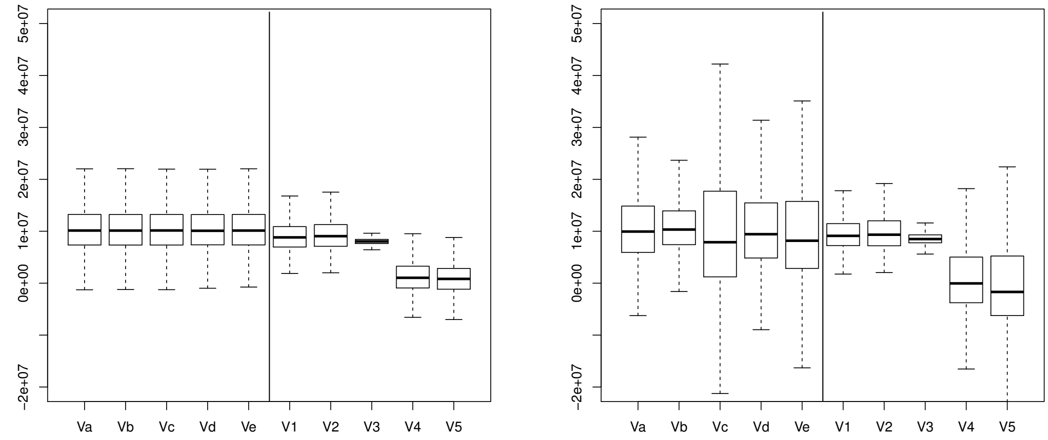
(a) POP 2 : $\sigma_1 = \sigma_2 = 5$

$\hat{\mathbf{V}}_{a,\mathrm{YG}}$	$\hat{\mathbf{V}}_{b,\mathrm{YG}}$	$\hat{\mathbf{V}}_{c,\mathrm{YG}}$	$\hat{\mathbf{V}}_{d,\mathrm{YG}}$	$\hat{\mathbf{V}}_{e,\mathrm{YG}}$
217	17	2021	541	1279
67	46	136	83	117
$\hat{\mathbf{V}}_{1,\mathrm{YG}}$	$\hat{\mathbf{V}}_{2,\mathrm{YG}}$	$\hat{\mathbf{V}}_{3,\mathrm{YG}}$	$\hat{\mathbf{V}}_{4,\mathrm{YG}}$	$\hat{\mathbf{V}}_{5,\mathrm{YG}}$
88	92	79	14	10
0	0	0	5070	5750

(b) POP 3 : $\sigma_1 = 5, \sigma_2 = 50$

TABLE 3.6 – CV and #NEG for the five variance estimators and their different parts, for POP 2 and POP 3 for Midzuno sampling in case 2 for $n_\mathrm{M} = n_\mathrm{D} = 20$

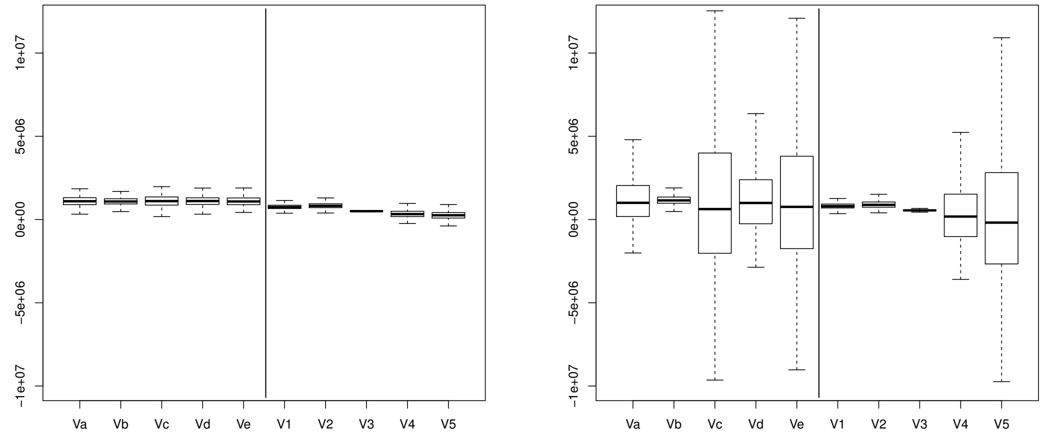
The components $\hat{\mathbf{V}}_{4,\mathrm{YG}}$ and $\hat{\mathbf{V}}_{5,\mathrm{YG}}$ present a small contribution (approximately 10 % in Table 3.6) to the variance estimators which increases with the sample sizes n_M and n_D equal to 50 (approximately 25% in Table 3.7). The number of negative values of



(a) POP 2 : $\sigma_1 = \sigma_2 = 5$

(b) POP 3 : $\sigma_1 = 5, \sigma_2 = 50$

FIGURE 3.3 – Boxplots of the five Y-G variance estimators (on the left of the vertical line) and their different parts (on the right of the vertical line) for POP 2 and POP 3 for Midzuno sampling in case 2 for $n_{\mathrm{M}} = n_{\mathrm{D}} = 20$



(a) POP 2 : $\sigma_1 = \sigma_2 = 5$

(b) POP 3 : $\sigma_1 = 5, \sigma_2 = 50$

FIGURE 3.4 – Boxplots of the five Y-G variance estimators (on the left of the vertical line) and their different parts (on the right of the vertical line) for POP 2 and POP 3 for Midzuno sampling in case 2 for $n_{\mathrm{M}} = n_{\mathrm{D}} = 50$

$\hat{V}$	$\hat{\mathbf{V}}_{a,\text{YG}}$	$\hat{\mathbf{V}}_{b,\text{YG}}$	$\hat{\mathbf{V}}_{c,\text{YG}}$	$\hat{\mathbf{V}}_{d,\text{YG}}$	$\hat{\mathbf{V}}_{e,\text{YG}}$
#NEG	0	0	0	0	0
CV($\hat{V}$)	27	21	31	26	27
$\hat{V}$	$\hat{\mathbf{V}}_{1,\text{YG}}$	$\hat{\mathbf{V}}_{2,\text{YG}}$	$\hat{\mathbf{V}}_{3,\text{YG}}$	$\hat{\mathbf{V}}_{4,\text{YG}}$	$\hat{\mathbf{V}}_{5,\text{YG}}$
CONT($\hat{V}$)	70	77	46	30	24
#NEG	0	0	0	620	1351

(a) POP 2 : $\sigma_1 = \sigma_2 = 5$

$\hat{\mathbf{V}}_{a,\text{YG}}$	$\hat{\mathbf{V}}_{b,\text{YG}}$	$\hat{\mathbf{V}}_{c,\text{YG}}$	$\hat{\mathbf{V}}_{d,\text{YG}}$	$\hat{\mathbf{V}}_{e,\text{YG}}$
1987	0	4350	3037	4151
118	24	397	170	360
$\hat{\mathbf{V}}_{1,\text{YG}}$	$\hat{\mathbf{V}}_{2,\text{YG}}$	$\hat{\mathbf{V}}_{3,\text{YG}}$	$\hat{\mathbf{V}}_{4,\text{YG}}$	$\hat{\mathbf{V}}_{5,\text{YG}}$
69	77	47	32	24
0	0	0	4742	5077

(b) POP 3 : $\sigma_1 = 5, \sigma_2 = 50$

TABLE 3.7 – CV and #NEG for the five variance estimators and their different parts, for POP 2 and POP 3 for Midzuno sampling in case 2 for $n_\text{M} = n_\text{D} = 50$

$\hat{\mathbf{V}}_{4,\text{YG}}$ and $\hat{\mathbf{V}}_{5,\text{YG}}$ also increases and contributes to increase the number of #NEG for the variance estimators $\hat{\mathbf{V}}_{c,\text{YG}}$, $\hat{\mathbf{V}}_{d,\text{YG}}$ and $\hat{\mathbf{V}}_{e,\text{YG}}$. We also observe that the coefficient of variation of each part of the variance estimators increases with the number of negative values. The estimator $\hat{\mathbf{V}}_{b,\text{YG}}$ has the advantage of presenting less negative values than the others : the components $\hat{\mathbf{V}}_{1,\text{YG}}$ and $\hat{\mathbf{V}}_{2,\text{YG}}$ are always non-negative while the non-negative and subtracted term $\hat{\mathbf{V}}_{3,\text{YG}}$ has a contribution which decreases with the sample size.

We also tested the three simplified variance estimators which have the advantage of being always non-negative for a sampling design which verifies the Sen-Yates-Grundy conditions. The results are presented in Table 3.8 for Midzuno sampling. We

POP	σ_1	σ_2	n_M	n_D	$\hat{V}_{\text{SIMP1}}$		$\hat{V}_{\text{SIMP2}}$		$\hat{V}_{\text{SIMP3}}$	
					RB	CV	RB	CV	RB	CV
1	0.5	5	5	5	-3	69	-4	68	93	134
	0.5	5	20	20	-13	31	-10	30	77	87
	0.5	5	50	50	-30	33	-23	28	47	51
2	5	5	5	5	-3	69	-4	68	93	134
	5	5	20	20	-13	31	-10	31	77	87
	5	5	50	50	-30	33	-23	28	47	52
3	5	50	5	5	-4	72	-5	76	90	140
	5	50	20	20	-12	32	-8	36	80	93
	5	50	50	50	-31	34	-23	30	46	52

TABLE 3.8 – Simplified variance estimators for Midzuno sampling in case 2

observe that the absolute value of the relative bias of $\hat{\mathrm{V}}_{\mathrm{SIMP1}}$ and $\hat{\mathrm{V}}_{\mathrm{SIMP2}}$ increases with the sample sizes n_{M} and n_{D}, contrary to the estimator $\hat{\mathrm{V}}_{\mathrm{SIMP3}}$. For the coefficient of variation CV, it decreases when the sample sizes increases for the three simplified estimators. We do not observe significant differences between the three populations. Contrarily to the populations generated in case 1, we do not observe large difference between $\hat{\mathrm{V}}_{\mathrm{SIMP1}}$ and $\hat{\mathrm{V}}_{\mathrm{SIMP2}}$ depending on the values of n_{M} and n_{D}. For small sizes n_{M} and n_{D}, the two first simplified variance estimator can be used, while the third simplified estimator presents a large relative bias and a substantial CV. Note that similar bad performances for $\hat{\mathrm{V}}_{\mathrm{SIMP3}}$ were also pointed out in Chapter 2 when estimating a ratio (see section 2.4.3).

3.5 Conclusion

In this chapter, we presented different variance decompositions and Yates-Grundy estimators. We used Midzuno, Sampford and conditional Poisson sampling in order to observe the differences between these five estimators in a simulation setup. Using a first model which does not seem unrealistic, we do not observe differences between these estimators and the simplified estimators present a behavior similar to the one observed in Chapter 2. The second tested model presents inclusion probabilities almost proportional to the interest variable. Such a situation may be very unlikely in practice but it reveals that differences between the five estimators may appear even for large sample sizes. In this context, only the estimator $\hat{\mathbf{V}}_{b,\mathrm{YG}}$ presents a CV and a number of negative values acceptable and we recommend to use this estimator.

3.6 Supplementary Materials

The basic functions required to calculate the estimators and the commands that display the results in Tables 3.3 to 3.12 or in Figure 3.3 are available in Appendix B at the end of the manuscript.

3.7 Simulations for Sampford and conditional Poisson sampling designs

POP	σ_1	σ_2	n_{M}	n_{D}	$\hat{\mathbf{V}}_{a,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{b,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{c,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{d,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{e,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$	
					CV	#NEG	CV	#NEG	CV	#NEG	CV	#NEG	CV	#NEG
1	0.5	5	5	5	74	410	74	410	74	410	74	410	74	410
	0.5	5	10	10	44	2	44	2	44	2	44	2	44	2
	0.5	5	20	20	27	0	27	0	27	0	27	0	27	0
2	5	5	5	5	57	53	57	53	57	54	57	53	57	53
	5	5	10	10	34	0	34	0	34	0	34	0	34	0
	5	5	20	20	21	0	21	0	21	0	21	0	21	0
3	5	50	5	5	97	0	97	0	97	0	97	0	97	0
	5	50	10	10	70	0	70	0	70	0	70	0	70	0
	5	50	20	20	49	0	49	0	49	0	49	0	49	0

TABLE 3.9 – Five Y-G variance estimators for Sampford sampling in case 1

POP	σ_1	σ_2	n_{M}	n_{D}	$\hat{\mathbf{V}}_{a,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{b,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{c,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{d,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{e,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$	
					CV	#NEG	CV	#NEG	CV	#NEG	CV	#NEG	CV	#NEG
1	0.5	5	5	5	72	360	72	361	72	360	72	360	72	361
	0.5	5	10	10	44	1	44	0	44	1	44	1	44	0
	0.5	5	20	20	27	0	27	0	27	0	27	0	27	0
	0.5	5	50	50	13	0	13	0	13	0	13	0	13	0
2	5	5	5	5	57	52	57	52	57	51	57	52	57	51
	5	5	10	10	33	0	33	0	33	0	33	0	33	0
	5	5	20	20	19	0	19	0	19	0	19	0	19	0
	5	5	50	50	10	0	10	0	10	0	10	0	10	0
3	5	50	5	5	56	41	56	42	56	42	56	41	56	41
	5	50	10	10	34	0	34	0	34	0	34	0	34	0
	5	50	20	20	21	0	21	0	21	0	21	0	21	0
	5	50	50	50	10	0	10	0	10	0	10	0	10	0

TABLE 3.10 – Five Y-G variance estimators for conditional Poisson sampling in case 1

POP	σ_1	σ_2	n_{M}	n_{D}	$\hat{\mathbf{V}}_{a,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{b,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{c,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{d,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{e,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$	
					CV	#NEG	CV	#NEG	CV	#NEG	CV	#NEG	CV	#NEG
1	0.5	5	5	5	103	1586	103	1581	103	1587	103	1586	103	1581
	0.5	5	10	10	65	376	65	378	65	378	65	376	65	378
	0.5	5	20	20	41	19	41	18	41	19	41	19	41	18
2	5	5	5	5	104	1583	104	1581	104	1585	104	1584	104	1582
	5	5	10	10	65	356	65	355	65	358	65	356	65	357
	5	5	20	20	42	18	42	14	42	23	42	14	42	20
3	5	50	5	5	111	1516	110	1508	112	1559	111	1525	111	1536
	5	50	10	10	71	455	69	415	78	664	72	476	75	608
	5	50	20	20	52	60	45	18	101	1378	57	153	94	1190

TABLE 3.11 – Five Y-G variance estimators for Sampford sampling in case 2

POP	σ_1	σ_2	n_{M}	n_{D}	$\hat{\mathbf{V}}_{a,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{b,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{c,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{d,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$		$\hat{\mathbf{V}}_{e,\mathrm{YG}}(\hat{t}_{\mathrm{Y}})$	
					CV	#NEG	CV	#NEG	CV	#NEG	CV	#NEG	CV	#NEG
1	0.5	5	5	5	102	1466	102	1461	102	1466	102	1466	102	1462
	0.5	5	10	10	64	349	64	354	64	348	64	350	64	351
	0.5	5	20	20	41	9	40	9	41	9	40	10	40	8
	0.5	5	50	50	23	0	20	0	23	0	23	0	20	0
2	5	5	5	5	102	1481	102	1479	102	1486	102	1478	102	1481
	5	5	10	10	65	374	65	368	65	376	65	372	65	376
	5	5	20	20	41	8	41	6	41	8	41	9	41	7
	5	5	50	50	26	0	20	0	31	1	25	0	27	0
3	5	50	5	5	108	1511	108	1495	110	1545	108	1522	109	1521
	5	50	10	10	70	402	69	389	77	653	71	432	75	617
	5	50	20	20	52	74	45	15	100	1366	58	154	93	1184
	5	50	50	50	99	1522	24	0	379	4048	141	2501	354	3874

TABLE 3.12 – Five Y-G variance estimators for conditional Poisson sampling in case 2

References

Dalén, J. and Ohlsson, E. (1995). Variance estimation in the swedish consumer price indexy. *Journal of Business & Economic Statistics*, 13(3) :347–356. 76

Deville, J. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85 :89–101. 79

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35 :1491–1523. 77

Midzuno, H. (1952). On the sampling system with probability proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics*, 3 :99–107. 78

Sampford, M. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54 :594–513. 78

Skinner, C. (2015). Cross-classified sampling : some estimation theory. *Statistics and Probability Letters*, 104 :163–168. 76

Tillé, Y. (2006). *Sampling Algorithms*. Springer-Verlag, New York. 76, 77, 78, 79

Tillé, Y. and Matei, A. (2015). *sampling : Survey Sampling*. R package version 2.7. 77, 78, 79

Chapitre 4

Two-dimensional sampling in practice

This chapter is a reprint of Juillard, H. (2016). Two-dimensional sampling in practice. *To appear in Case study In Business, Industry And Government Statistics.*

Cet article explique les principes du plan d'échantillonnage à deux degrés et présente ceux moins connus du plan d'échantillonnage produit. Un des objectifs de l'article est de permettre au lecteur de différencier ces deux plans de sondage et de mettre en pratique les étapes d'échantillonnage et d'estimation. Pour ces deux plans de sondage, des estimateurs respectifs de variance sont calculés pour des cas simples, et des analogies avec l'ANOVA à un facteur et celle à deux facteurs sont proposées. Cette comparaison est motivée par l'enquête française Elfe, et les sélections et les estimations sont illustrées à partir des logiciels R, SAS et Stata.

Mots-clés *: ANOVA, échantillonnage à deux degrés, estimation de variance, population observée bi-dimensionnellement, procédures R / SAS / Stata.*

Sommaire

Two-dimensional sampling in practice

This article explains the principles of the two-stage sampling design and presents the less known cross-classified sampling design. One purpose of the article is to allow the reader to differentiate between the two survey designs and put in practice the sampling and estimation steps. Respective variance estimators for these two designs are calculated in simple cases, and analogies with one-way and two-way ANOVA are proposed. The comparison is motivated by the ELFE French survey, and selections and estimations are illustrated using the softwares R, SAS and Stata.

Keywords *: ANOVA, survey procedures in R / SAS / Stata, population observed bi-dimensionally, two-stage sampling, variance estimation.*

4.1 Introduction

Our population of interest is observed bi-dimensionally and can be represented by a rectangular array. In Figure 4.1, we illustrate the cross product of a population of rows and of a population of columns.

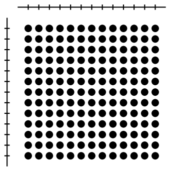


FIGURE 4.1 – Population observed bi-dimensionally

Sampling in a population observed bi-dimensionally is discussed in the literature in different contexts : spatial sampling with the longitude and the latitude as the dimensions, as well as plane sampling or sampling in space and time in Vos [1964]. The use of rows and columns in lattice sampling is presented in Bellhouse [1981] or Ohlsson [1996]. Sampling of outlets and items for the consumer price index is presented in

Dalén and Ohlsson [1995]. A sampling of maternities and days is also used for the ELFE (Etude Longitudinale Française depuis l'Enfance) French cohort of infants.

Various sampling designs are possible in a population observed bi-dimensionally. The sample can be drawn directly with one phase of selection only (as shown in Figure 4.2), or with several steps of selections. For example, a standard two-stage sampling design can be used. This consists in drawing a sample of primary units, and then a second stage sample inside each primary unit independently. Figure 4.3 illustrates a case where rows are used as primary units : 4 rows are selected, and 3 columns are then drawn inside each selected row.

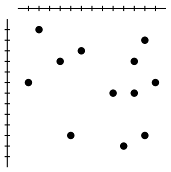


FIGURE 4.2 – Direct sampling in a population observed bi-dimensionally

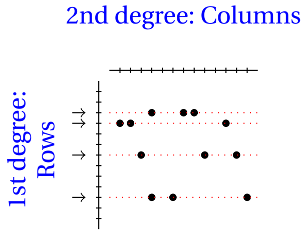


FIGURE 4.3 – Two-stage sampling in a population observed bi-dimensionally with rows as primary units

A cross-classified sampling design (CCS) can also be used, which proceeds as follows : two samples are drawn independently, and then crossed. In Figure 4.4, a sample of 4 rows and a sample of 3 columns are selected, which results in a final sample of 12 units row × column.

For the two-stage and the cross-classified designs, we distinguish two steps of sampling : one on rows and one on columns. Nevertheless, the CCS design can not be

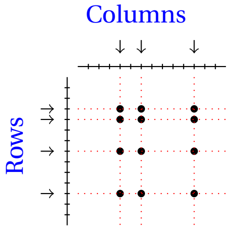


FIGURE 4.4 – Cross-classified sampling

regarded as a classic two-stage design. A classic two-stage design requires two assumptions : independence between the drawings made at each stage, also called the invariance property [Särndal et al., 1992] ; independence between the various drawings at the second stage, conditionally on the first stage sample. For a CCS design, the invariance property is verified (independence between the sample of rows and the sample of columns), but the independence property is not (a same sample of columns is used for each row).

If the two-stage sampling design is well known, the CCS design presents a limited literature, recently completed by Skinner [2015] and Juillard et al. [2016]. In practice, it is specifically used in the Consumer Price Index designs in different countries like the United States [Wilkerson, 1957] and Sweden [Dalén and Ohlsson, 1995]. One purpose of this article is to allow the reader to differentiate between these two sampling designs, and to put in practice the sampling and estimation steps. In practice, softwares like R, SAS or Stata propose sampling and estimation procedures for two-stage sampling, but to the best of our knowledge there is no such offer for the CCS design. This case study aims at illustrating the error committed by users, when treating the CCS design as a two-stage sampling design for variance computation and variance estimation. An R program which enables to perform variance estimation for a CCS design is available as supplementary material.

The comparison between two-stage sampling and CCS is motivated by the ELFE survey presented in Section 4.2 together with the data used for this case study. For these two designs, the total and ratio parameters are studied and corresponding variances as well as variance estimators are computed in a simple case. Analogies with one-way and two-way ANOVA are proposed, which enables to interpret the variance formulas in terms of column effect or row effect. In Section 4.3, we focus on the two-stage sampling design and in Section 4.4, we focus on the CCS design. We will compare the softwares advantages (R, SAS, Stata) in terms of selection procedures and variance estimation when estimating totals and ratios. The various estimators will be progressively illustrated in this article. A comparison between the different methods of estimation for the two designs through simulations is proposed in Section 4.5.

4.2 ELFE survey, data and softwares

The ELFE [1] French cohort consists of more than 18,000 children whose parents consented to their inclusion. In each of the 320 selected maternity units, targeted babies born during 25 days (during four specific periods representing each of the four seasons) in 2011 were selected.

In the ELFE survey, spatial (metropolitan France) and temporal (year 2011) variabilities was sought. In practice, logistical and administrative reasons oriented the sample design : a direct sampling (as illustrated in Figure 4.2) or a two-stage sampling design (as illustrated in Figure 4.3) could not be used. A CCS was implemented, crossing independently a sample of maternities and a sample of days. Stratified simple random sampling was used for the two populations, but in our study, we will consider a simple random sampling for the two designs. Owing to its two selection steps, the CCS design may be considered by data users as a two-stage sampling design, leading to erroneous variance estimation. This article aims at differentiating these two sampling designs, and at quantifying the bias in variance induced by such approximation of the CCS design.

[1] http://www.elfe-france.fr/index.php/en/

The dataset delivered with this article represents the ELFE population with $N_M = 544$ maternities in the population U_M and $N_D = 365$ days in the population U_D in 2011. Given the confidentiality issues, the interest variables in the dataset are count variables simulated taking into account different maternity and day effects. So as to mimic the variables in the ELFE survey, we consider the *Number of infants with a mother followed by a midwife* for the variable Y_{ik} and the *Number of infants born by caesarean* for Z_{ik} where i denotes the index for the maternity and k the index for the day. In this article, we will focus on the estimation of total and ratio parameters and the variable X_{ik} in the dataset, that will be used as the denominator for the ratio, can be considered as the *Number of births*. The construction of this count variables is detailed in section 4.6.1.

The code is provided in order to replicate all results obtained in this article. Three softwares are used and compared : R 3.2.2 [R Core Team, 2015], SAS 9.4 [SAS Institute Inc., 2015], Stata 13.1 [StataCorp., 2013]. R is available from Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/.

4.3 Two-stage sampling : selection and estimation

We begin by describing the basic principles of two-stage sampling. Assume that we are interested in some population $U_M = \{u_1, \ldots, u_i, \ldots, u_{N_M}\}$ of non-overlapping Primary Sampling Units (PSUs), where each PSU u_i is itself a population of Secondary Sampling Units (SSUs) of size N_i. A sample S_M of size n_M is selected in U_M by means of some sampling design $p_M(\cdot)$. Inside each $u_i \in S_M$, a second stage sample S_i of size n_i is then selected according to some sampling design $p_{iD}(\cdot|S_M)$. The final sample of SSUs is $S = \bigcup_{u_i \in S_M} S_i$.

A two-stage sampling design is usually required to match the following assumptions :

H1 Invariance : the design $p_{iD}(\cdot|S_M)$ used in the second stage for a PSU u_i does not depend on the first-stage sample $\mathrm{S_M}$ selected, that is

$$\forall u_i \in \mathrm{U_M}, \ \ p_{i\mathrm{D}}\left(.|\mathrm{S_M}\right) = p_{i\mathrm{D}}\left(.\right).$$

H2 Independence : conditionally on $\mathrm{S_M}$, the sub-sampling inside the selected PSUs is independent from one PSU to another. That is,

$$Pr\left(\bigcup_{u_i \in \mathrm{S_M}} \mathrm{S}_i|\mathrm{S_M}\right) = \prod_{u_i \in \mathrm{S_M}} Pr\left(\mathrm{S}_i|\mathrm{S_M}\right).$$

4.3.1 Selecting a two-stage sample

In this part, the possibilities to draw two-stage samples using the softwares R, SAS and Stata are scanned. In our case study, a SI (simple random) sampling is drawn in $\mathrm{U_M}$ and the SI sampling is also used in each $u_i \in \mathrm{U_M}$ (which we denote {SI,SI}) ; in order to mimic the ELFE sample size, the same number $n_\mathrm{D} = 25$ of SSUs is drawn inside each of the $n_\mathrm{M} = 320$ selected PSUs.

R implementation The function *mstage* of the sampling package [Tillé and Matei, 2015] in R can be used to select a two-stage sample in a single step (see the frame Code 4.1). With the argument *stage*, four methods of selection can be used but it has to be the same for the two stages : simple random sampling without replacement or with replacement, Poisson sampling or systematic sampling. The option *pik* has to be applied in the case of unequal probabilities of selection. The argument *size* used indicates the sample size of PSUs, and the vector of sample sizes of SSUs.

```
library(sampling)
tableR=read.csv2("..../Data2stCCS.csv")
n_m=320; n_d=25; N_m=544; N_d=365; N=N_m*N_d

m=mstage(tableR,stage=list("cluster","cluster"),varnames=list("ID_i","ID_k"),size=list(n_m,c(rep(n_
    d,n_m))), method=c("srswor","srswor"))

ech=getdata(tableR,m)[[2]]
```

Code 4.1 – An R code to select a two-stage sample in a population observed bi-dimensionally

SAS implementation The SAS software proposes to call two procedures *SURVEYSELECT* as proposed in the frame Code 4.2. In order to identify the PSUs, the first procedure uses the *cluster* statement and the second the *strata* statement. The *strata* statement can also be applied at both stages and a lot of different methods of selection are available (simple random sampling with or without replacement, Bernoulli sampling, sampling with probabilities proportional to size, with sequential or systematic selection, . . .).

```
proc import datafile = ".../Data2stCCS.csv"
out=pop dbms=csv replace;DELIMITER=";" ; run;

proc SURVEYSELECT data=pop method=srs n=320 seed=1357 out=ech1;
   cluster ID_i;
run;

proc SURVEYSELECT data=ech1 method=srs n=25 seed=7548 out=ech;
   strata ID_i;
run;
```

Code 4.2 – A SAS code to select a two-stage sample in a population observed bi-dimensionally

Stata implementation The software Stata proposes the command *sample* (*bsample*, respectively) to draw a random sample without replacement (with replacement, respectively). The command *sample* can be used with the option *by* followed by the name of the stratum. In this case the same number or the same percentage of units is drawn inside each stratum. In the frame Code 4.3, a table 'ech1' containing only one row by PSU is created and a first SI sample of size 320 is selected. The command *merge* enables to create the sampling base for the second step of selection. A SI sample of 25 units is drawn in each selected PSU using *by id_i : sample 25, count.*

```
. clear
. insheet using /.../Data2stCCS.csv,delimiter(;)
. save POP, replace
. contract id_i
. sample 320, count

. sort id_i
. keep id_i
. save /.../ech1.dta, replace
. clear
. use POP
. sort id_i
. merge m:1 id_i using /.../ech1.dta
. drop if _merge != 3

. sort id_i
. by id_i: sample 25, count
. count
```

Code 4.3 – A Stata code to select a two-stage sample in a population observed bi-dimensionally

The same two steps as in Stata could also be used with R and SAS. This would enable to make use of the one-stage sampling procedures available in each software.

4.3.2 Estimating a total

We consider a study variable Y taking the value Y_{ik} for the PSU u_i and the SSU k. We are interested in estimating the total

$$t_Y = \sum_{u_i \in U_M} \sum_{k \in u_i} Y_{ik}.$$

In the particular case of SI sampling in U_M and SI sampling inside each $u_i \in S_M$, the expansion estimator

$$\hat{t}_Y = \frac{N_M}{n_M} \sum_{u_i \in S_M} \frac{N_i}{n_i} \sum_{k \in S_i} Y_{ik}$$

is unbiased for t_Y [Särndal et al., 1992].

4.3.3 Calculating the variance

Under the invariance and the independence assumptions, the variance of $\hat{t}_Y$ is obtained by conditioning on the first stage sample S_M. This leads to

$$\mathbf{V}_{2d}\left(\hat{t}_Y\right) = \mathbf{V}_{PSU}\left(\hat{t}_Y\right) + \mathbf{V}_{SSU}\left(\hat{t}_Y\right).$$

In case of {SI,SI}, we obtain

$$\mathbf{V}_{PSU}\left(\hat{t}_Y\right) = N_M^2\left(\frac{1}{n_M} - \frac{1}{N_M}\right) S^2_{Y_{\circ\bullet}}, \tag{4.3.1}$$

$$\mathbf{V}_{SSU}\left(\hat{t}_Y\right) = \frac{N_M}{n_M}\sum_{u_i \in U_M} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right) S^2_{Y_{i\circ}}, \tag{4.3.2}$$

with

$$S^2_{Y_{\circ\bullet}} = \frac{1}{N_M - 1}\sum_{u_i \in U_M}\left(Y_{i\bullet} - \frac{1}{N_M}\sum_{u_j \in U_M} Y_{j\bullet}\right)^2,$$

$$S^2_{Y_{i\circ}} = \frac{1}{N_i - 1}\sum_{k \in u_i}\left(Y_{ik} - \frac{1}{N_i}\sum_{l \in u_i} Y_{il}\right)^2$$

where $Y_{i\bullet} = \sum_{k \in u_i} Y_{ik}$.

In the particular case where two-stage sampling is used inside a product population $U_M \times U_D$ (as illustrated in Figure 4.3), all the PSUs u_i (with associated size N_i) in the above formulas can be replaced by a same notation U_D (with associated size N_D) for all $i \in U_M$. In this case, if the same number of SSUs is drawn inside each selected PSU, we may note $n_i = n_D$ for any $i \in S_M$.

An analogy can be made between the two-stage variance decomposition and the analysis of variance (ANOVA) which uses the partitioning of sums of squared deviations. For one-way ANOVA, the total sum of squares $SS_T = \sum_{u_i \in U_M}\sum_{k \in u_i}\left(Y_{ik} - \bar{Y}_{\bullet\bullet}\right)^2$ (see section 4.6.2 for the notations) may be written as

$$\mathrm{SS_T} \quad = \quad \mathrm{SS_M} + \mathrm{SS_E}$$

where $\mathrm{SS_M}$ is the explained sum of squares (a.k.a. the sum of squares between classes) and $\mathrm{SS_E}$ denotes the residual sum of squares (a.k.a. sum of squares within classes), see section 4.6.3 for details. For example, in our case study, the variable *Number of infants born by caesarean* (Z_{ik}) presents a smaller $\mathrm{SS_M}$ than the variable *Number of births* (X_{ik}).

We consider the {SI,SI} sampling case, and assume for simplicity that all the PSUs are of the same size $\mathrm{N}_i = \mathrm{N_D}$, and that the same sample size $n_i = n_{\mathrm{D}}$ is used inside each selected PSU. In this case, we have

$$\begin{aligned} \mathrm{SS_M} &= \frac{\mathrm{N_M}-1}{\mathrm{N_D}} \, \mathrm{S}^2_{\mathrm{Y}_{\circ\bullet}}, \\ \mathrm{SS_E} &= (\mathrm{N_D}-1) \sum_{u_i \in \mathrm{U_M}} \mathrm{S}^2_{\mathrm{Y}_{i\circ}}. \end{aligned}$$

The variance in (4.3.1) due to the selection of PSUs may be rewritten as

$$\mathbf{V}_{\mathrm{PSU}}\left(\hat{t}_{\mathrm{Y}}\right) = \mathrm{N_M^2}\left(\frac{1}{n_{\mathrm{M}}} - \frac{1}{\mathrm{N_M}}\right)\frac{\mathrm{N_D}}{\mathrm{N_M}-1} \, \mathrm{SS_M},$$

and depends on the explained sum of squares $\mathrm{SS_M}$. The variable X_{ik} will present a more important part of first-stage variance than the variable Z_{ik}. The variance in (4.3.2) due to the selection of SSUs may be rewritten as

$$\mathbf{V}_{\mathrm{SSU}}\left(\hat{t}_{\mathrm{Y}}\right) = \frac{\mathrm{N_M}}{n_{\mathrm{M}}}\mathrm{N_D^2}\left(\frac{1}{n_{\mathrm{D}}} - \frac{1}{\mathrm{N_D}}\right)\frac{1}{\mathrm{N_D}-1} \, \mathrm{SS_E}$$

and depends on the residual sum of squares $\mathrm{SS_E}$.

4.3.4 Estimating the variance

An unbiased variance estimator of $\hat{t}_{\mathrm{Y}}$ can be written as

$$\hat{\mathbf{V}}_{2d}\left(\hat{t}_{\mathrm{Y}}\right) = \hat{\mathbf{V}}_{2d,a}\left(\hat{t}_{\mathrm{Y}}\right) + \hat{\mathbf{V}}_{2d,b}\left(\hat{t}_{\mathrm{Y}}\right) \tag{4.3.3}$$

where

$$\hat{\mathbf{V}}_{2d,a}\left(\hat{t}_{\mathrm{Y}}\right) = \mathrm{N}_{\mathrm{M}}^2\left(\frac{1}{n_{\mathrm{M}}} - \frac{1}{\mathrm{N}_{\mathrm{M}}}\right)s^2_{\hat{\mathrm{Y}}_{\circ\bullet}}, \tag{4.3.4}$$

$$\hat{\mathbf{V}}_{2d,b}\left(\hat{t}_{\mathrm{Y}}\right) = \frac{\mathrm{N}_{\mathrm{M}}}{n_{\mathrm{M}}}\sum_{u_i\in\mathrm{S}_{\mathrm{M}}}\mathrm{N}_i^2\left(\frac{1}{n_i} - \frac{1}{\mathrm{N}_i}\right)s^2_{\mathrm{Y}_{i\circ}}, \tag{4.3.5}$$

with

$$s^2_{\hat{\mathrm{Y}}_{\circ\bullet}} = \frac{1}{n_{\mathrm{M}}-1}\sum_{u_i\in\mathrm{S}_{\mathrm{M}}}\left(\hat{\mathrm{Y}}_{i\bullet} - \frac{1}{n_{\mathrm{M}}}\sum_{u_j\in\mathrm{S}_{\mathrm{M}}}\hat{\mathrm{Y}}_{j\bullet}\right)^2,$$

$$s^2_{\mathrm{Y}_{i\circ}} = \frac{1}{n_i-1}\sum_{k\in\mathrm{S}_i}\left(\mathrm{Y}_{ik} - \frac{1}{n_i}\sum_{l\in\mathrm{S}_i}\mathrm{Y}_{il}\right)^2,$$

and where

$$\hat{\mathrm{Y}}_{i\bullet} = \sum_{k\in\mathrm{S}_i}\frac{\mathrm{N}_i}{n_i}\mathrm{Y}_{ik}$$

denotes the Horvitz-Thompson estimator of the sub-total Y_i. For an estimation term by term of the variance in formula (4.3.1), see the section 4.6.2.

Using the same one-way ANOVA as in the previous section but calculated on the sample $s_{\mathrm{M}} \times s_{\mathrm{D}}$, the total sum of squares ss_{T} may be written as

$$ss_{\mathrm{T}} = ss_{\mathrm{M}} + ss_{\mathrm{E}}$$

where each term is defined in section 4.6.3. The first part of the variance estimator in (4.3.4) can be rewritten as

$$\hat{\mathbf{V}}_{2d,a}\left(\hat{t}_{\mathrm{Y}}\right) = \mathrm{N}_{\mathrm{M}}^{2}\left(\frac{1}{n_{\mathrm{M}}} - \frac{1}{\mathrm{N}_{\mathrm{M}}}\right)\frac{n_{\mathrm{D}}}{n_{\mathrm{M}}-1}\quad ss_{\mathrm{M}},$$

and depends on the explained sum of squares ss_{M}. The second part in (4.3.5) can be rewritten as

$$\hat{\mathbf{V}}_{2d,b}\left(\hat{t}_{\mathrm{Y}}\right) = \frac{\mathrm{N}_{\mathrm{M}}}{n_{\mathrm{M}}}\mathrm{N}_{\mathrm{D}}^{2}\left(\frac{1}{n_{\mathrm{D}}} - \frac{1}{\mathrm{N}_{\mathrm{D}}}\right)\frac{1}{n_{\mathrm{D}}-1}\quad ss_{\mathrm{E}}$$

and depends on the explained sum of squares ss_{E}. Note that the term $\hat{\mathbf{V}}_{2d,a}\left(\hat{t}_{\mathrm{Y}}\right)$ is occasionally considered as a simplified variance estimator of $\hat{\mathbf{V}}_{2d}\left(\hat{t}_{\mathrm{Y}}\right)$. The underestimation is seen as negligible when the first-stage inclusion probabilities $n_{\mathrm{M}}/\mathrm{N}_{\mathrm{M}}$ are small [Särndal et al., 1992].

4.3.5 Estimation in practice

In this Section, we propose to study the estimation of a more complex parameter than a total using different procedures from the R, SAS or Stata softwares. A ratio $\mathrm{R} = t_{\mathrm{Y}}/t_{\mathrm{X}}$ can be easily estimated by $\hat{\mathrm{R}} = \hat{t}_{\mathrm{Y}}/\hat{t}_{\mathrm{X}}$ using a plug-in principle. To estimate the variance, a linearization method can be used [Deville, 1999]. The estimated linearized variable is then plugged into the formula (4.3.3).

From a particular selected sample (using variable *Dummy_2d* in the dataset, which takes the value 1 if the unit is selected in the {SI,SI} sample and 0 otherwise), the estimated ratios $\hat{t}_{\mathrm{Y}}/\hat{t}_{\mathrm{X}}$ and $\hat{t}_{\mathrm{Z}}/\hat{t}_{\mathrm{X}}$ and their estimated variance $\hat{\mathbf{V}}_{2d}$ can be calculated together with the approximation $\hat{\mathbf{V}}_{2d,a}$.

R implementation The functions *svydesign* and *twophase* of the R package survey [Lumley, 2014] can be used to describe the two-stage sample. Only the first one is illustrated in the frame Code 4.4. Note that other packages are available to estimate the sampling variance. In the argument *id*, the vector of PSU IDs has to be entered, followed by the vector of SSU IDs. The argument *fpc* can be specified as the PSU population size N_M in the form of a vector, followed by the vector of the SSU populations size N_D. The appropriate set of weights can be set using the argument *weights*. Note that it is possible to take into account the stratified sampling by using the argument *strata*. The function *svyratio* estimates the ratio and its associated standard error. The command *SE(yxratio)ˆ2* displays the estimated variance $\hat{\mathbf{V}}_{2d}(\hat{R})$.

```
> tableR=read.csv2(".../Data2stCCS.csv")
> ech=tableR[tableR$Dummy_2d==1,]
> attach(ech)
> library(survey)
> n_m=320; n_d=25; N_m=544; N_d=365
> infoplan<-svydesign(id=~ID_i+ID_k,fpc=~N_M+N_D, weights=(N_m*N_d)/(n_m*n_d), data=ech)
> (yxratio <- svyratio(~Yik+Zik ,~Xik,infoplan))

Ratio estimator: svyratio.survey.design2(~Yik + Zik, ~Xik, infoplan)
Ratios=
          Xik
Yik 0.1507968
Zik 0.1510090

SEs=
             Xik
Yik 0.0008873011
Zik 0.0009162289

> SE(yxratio)^2

     Yik/Xik      Zik/Xik
7.873033e-07 8.394754e-07

> confint(yxratio)
#confint(yxratio, level=0.90)
             2.5 %    97.5 %
Yik/Xik 0.1490577 0.1525359
Zik/Xik 0.1492132 0.1528047
```

Code 4.4 – R code and results when estimating the ratio and its variance $\hat{\mathbf{V}}_{2d}(\hat{R})$

The command *vcov(yxratio)* permits also to display the estimated variance. The function *svytotal(~ Xik + Yik + Zik, infoplan)* can be used to estimate the totals $\hat{t}_X$, $\hat{t}_Y$ and

$\hat{t}_Z$ while the function *svymean(~ Xik + Yik + Zik , infoplan)* can be used to estimate the respective means of X_{ik}, Y_{ik} and Z_{ik}. By default the function *confint* produces a confidence interval of level 0.95 and it can be changed using the option *level*.

Note that $\hat{\mathbf{V}}_{2d,a}\left(\hat{R}\right)$ can also be calculated with R, with a simple modification of the previous procedures (see frame Code 4.5).

```
> infoplan<-svydesign(id=~ID_i,fpc=~N_M, weights=(N_m*N_d)/(n_m*n_d), data=ech)
> yxratio <- svyratio(~Yik+Zik ,~Xik, infoplan)
> SE(yxratio)^2

     Yik/Xik      Zik/Xik
3.377375e-07 3.732472e-07

> confint(yxratio)

               2.5 %    97.5 %
Yik/Xik 0.1496578 0.1519359
Zik/Xik 0.1498116 0.1522064
```

Code 4.5 – R code and results when estimating the ratio and its part of variance $\hat{\mathbf{V}}_{2d,a}\left(\hat{R}\right)$

SAS implementation The procedure *SURVEYMEANS* is used in the frame Code 4.6 with the argument *cluster* to indicate the PSU IDs, and *weight* for the set of weights wik. The option *strata* is available. This procedure calculates $\hat{R}$ and only the first part $\hat{\mathbf{V}}_{2d,a}\left(\hat{R}\right)$ of the estimated variance $\hat{\mathbf{V}}_{2d}\left(\hat{R}_Y\right)$.

```
proc IMPORT datafile = ".../Data2stCCS.csv"
out  = ech (where= (Dummy_2d=1))
dbms = csv
replace;
DELIMITER=";" ;
run;

data ech; set ech; wik=(544*365)/(320*25); run;

proc SURVEYMEANS data=ech total=544 mean sum var varsum missing clm /* alpha=0.10 */;
CLUSTER ID_i ;
/* VAR Xik Yik Zik ; */
RATIO Yik Zik / Xik ;
WEIGHT wik ;
```

```
run ;

                        Ratio Analysis
Numerator Denominator Ratio Std Err Var 95% CL for Ratio
Yik Xik 0.150797 0.000581 0.000000338 0.149653 0.151940
Zik Xik 0.151009 0.000611 0.000000373 0.149807 0.152211
```

Code 4.6 – SAS code and results when estimating the ratio and its part of variance $\hat{\mathbf{V}}_{2d,a}(\hat{R})$

The default *alpha* option is 0.05. The line of code *VAR Xik Yik Zik;* can be used to estimate the totals $\hat{t}_X$, $\hat{t}_Y$ and $\hat{t}_Z$ using results of options *sum* and *varsum*. The same command is used to estimate the means with options *mean* and *var*. Note that the second term of $\hat{\mathbf{V}}_{2d}(\hat{R})$ can be calculated using a supplementary step [Aragon and Ruiz-Gazen, 2004].

Stata implementation The command *svyset* of Stata in the frame Code 4.7 is used to describe the two-stage sample. In the first place id_i stands for the PSU IDs, followed by the vector of weights wik which in this application equals $(N_m N_D)/(n_M n_D)$. The argument *fcp* takes into account the PSU population size. After the two vertical bars, the second stage is defined in the same way. The command *svy : ratio* calculates the estimated ratio $\hat{R}$ and its associated standard error which corresponds to the square root of $\hat{\mathbf{V}}_{2d}(\hat{R})$.

```
. clear
. insheet using /.../Data2stCCS.csv,delimiter(;)
(14 vars, 198560 obs)
. save POP, replace
. keep if dummy_2d==1
. gen wik=(544*365)/(25*320)
. svyset id_i [pweight=wik], fpc(n_m) || id_k, fpc(n_d)

      pweight: wik
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: id_i
        FPC 1: n_m
     Strata 2: <one>
         SU 2: id_k
         FPC 2: n_d

. svy : ratio (yik/xik) (zik/xik)
* svy : ratio (yik/xik) (zik/xik), level(90) ;
(running ratio on estimation sample)

Survey: Ratio estimation

Number of strata =    1  Number of obs    =    8000
Number of PSUs   = 320   Population size = 198560
                         Design df        =     319

     _ratio_1: yik/xik
     _ratio_2: zik/xik

--------------------------------------------------------------
        |           Linearized
        |   Ratio   Std.Err.  [97.5% Conf.Interval]
--------+-----------------------------------------------------
_ratio_1| .1507968 .0008873 .1490511 .1525425
_ratio_2|  .151009 .0009162 .1492064 .1528116
--------------------------------------------------------------
```

Code 4.7 – Stata code and results when estimating the ratio and its variance $\hat{\mathbf{V}}_{2d}\left(\hat{\mathrm{R}}\right)$

To estimate $\hat{t}_{\mathrm{X}}$, $\hat{t}_{\mathrm{Y}}$ and $\hat{t}_{\mathrm{Z}}$, the command *svy : total xik yik zik* can be used. So as to estimate means, we may use the command *svy : mean xik yik zik*. The default *level* option for the confidence interval is 95 %.

Note that the variance estimator $\hat{\mathbf{V}}_{2d,a}\left(\hat{\mathrm{R}}\right)$ can also be obtained with Stata in the frame Code 4.8.

```
. svyset id_i [pweight=wik], fpc(n_m)
. svy : ratio (yik/xik) (zik/xik)


--------------------------------------------------------------
        |           Linearized
        |   Ratio   Std.Err.  [97.5% Conf.Interval]
--------+-----------------------------------------------------
_ratio_1| .1507968 .0005812 .1496534 .1519402
_ratio_2|  .151009 .0006109 .149807   .152211
--------------------------------------------------------------
```

Code 4.8 – Stata code and results when estimating the ratio and its part of variance $\hat{\mathbf{V}}_{2d,a}\left(\hat{\mathrm{R}}\right)$

4.4 Cross-classified sampling : selection and estimation

We now consider the cross-classified sampling design. We consider a sampling design p_{M} in $\mathrm{U_M}$, leading to a sample $\mathrm{S_M}$ of size n_{M}. We consider a sampling design p_{D} in $\mathrm{U_D}$, leading to a sample $\mathrm{S_D}$ of size n_{D}. We assume that the two designs $p_{\mathrm{M}}(\cdot)$ and $p_{\mathrm{D}}(\cdot)$ are independent. This enables to define a sampling design $p(\cdot)$ on the product population $\mathrm{U} = \mathrm{U_M} \times \mathrm{U_D}$ as

$$\begin{aligned} p(s) &= p_{\mathrm{M}}(s_{\mathrm{M}}) \times p_{\mathrm{D}}(s_{\mathrm{D}}) \\ \text{for any } s &= s_{\mathrm{M}} \times s_{\mathrm{D}} \subset \mathrm{U_M} \times \mathrm{U_D}. \end{aligned}$$

The assumption of independence for a cross-classified sampling design is equivalent to the standard assumption **H1** of invariance between two successive drawings in a two-stage sampling design.

4.4.1 Selecting a cross-classified sample

There is no standard procedure to perform CCS in one step, but all possible one-stage sampling procedures can be used to select $\mathrm{S_M}$ and $\mathrm{S_D}$ independently. The samples are then crossed to obtain the final sample $\mathrm{S_M} \times \mathrm{S_D}$. In our case study, we are interesting in the crossing of a SI sample of size $n_{\mathrm{M}} = 320$ drawn in $\mathrm{U_M}$, and of a SI sample of size n_{D} drawn in $\mathrm{U_D}$. Such design will be denoted as SI × SI.

R implementation A selection of a SI × SI sample with the software R is presented in the frame Code 4.9.

```
> tableR=read.csv2(".../Data2stCCS.csv")
> n_m=320; n_d=25; N_m=544; N_d=365
>
> s_m=sample(1:N_m,n_m) ; s_d=sample(1:N_d,n_d)

> Dummy_CCS2 <- rep(0,N); Dummy_CCS2[which(tableR$ID_i %in% s_m & tableR$ID_k %in% s_d)] <-1
> echCCS=tableR[Dummy_CCS2==1, ]
```

Code 4.9 – An R code to select the CCS sample

SAS implementation With SAS, the procedures *SURVEYSELECT* and *merge* can be used to select and cross the two samples (frame Code 4.10).

```
proc IMPORT datafile = ".../Data2stCCS.csv"
out=pop dbms=csv replace;DELIMITER=";" ; run;

proc freq data=pop;tables ID_i/out=popM;run;
proc SURVEYSELECT data=popM method=srs n=320 seed=2289 stats out=echM ;
run;
proc freq data=pop;tables ID_k/out=popD;run;
proc SURVEYSELECT data=popD method=srs n=25 seed=2368 stats out=echD ;
run;

proc sort data=pop; by ID_i; run;
proc sort data=echM; by ID_i; run;
data echA; merge echM (in=A) pop; by ID_i; if A;
run;
proc sort data=echA; by ID_k; run;
proc sort data=echD; by ID_k; run;
data ech; merge echD (in=A) echA; by ID_k; if A;
Dummy_CCS2=1;
run;
```

Code 4.10 – A SAS code to select the CCS sample

Stata implementation Following the same logic, the commands *sample* and *merge* may be used with the Stata software, as illustrated in the frame Code 4.11.

```
. clear
. insheet using /.../Data2stCCS.csv,delimiter(;)
. save POP, replace
. contract id_i
. sample 320, count
. sort id_i
. keep id_i
. save echM, replace

. clear
. use POP
. contract id_k
. sample 25, count
. sort id_k
. keep id_k
. save echD, replace

. clear
. use POP
. sort id_i
. merge m:1 id_i using echM.dta
. drop if _merge != 3
. drop _merge
. sort id_k
. merge m:1 id_k using echD.dta
. drop if _merge != 3

. gen Dummy_CCS2=1;
```

Code 4.11 – A Stata code to select the CCS sample

4.4.2 Estimating a total

In the particular case of SI × SI, the total

$$t_{\mathrm{Y}} = \sum_{i \in \mathrm{U_M}} \sum_{k \in \mathrm{U_D}} \mathrm{Y}_{ik}$$

is unbiasedly estimated by the expansion estimator

$$\hat{t}_{\mathrm{Y}} = \sum_{i \in \mathrm{S_M}} \sum_{k \in \mathrm{S_D}} \frac{\mathrm{N_M N_D}}{n_{\mathrm{M}} n_{\mathrm{D}}} \mathrm{Y}_{ik},$$

see Juillard et al. [2016] for details.

4.4.3 Calculating the variance

In this Section, the variance of $\hat{t}_{\mathrm{Y}}$ is calculated in the SI × SI case. The analogy between the decomposition of the SI × SI variance and the decomposition of a two-way

ANOVA was noted in Ohlsson [1996], and is described here. For a two-way ANOVA without replication, the total sum of squares may be written as

$$\mathrm{SS_T} = \mathrm{SS_M} + \mathrm{SS_D} + \mathrm{SS_E} \tag{4.4.1}$$

where the terms $\mathrm{SS_D}$, $\mathrm{SS_M}$ and $\mathrm{SS_E}$ represent respectively the sum of squares explained by the factor D, the one explained by the factor M and the residual sum of squares. The details are given in section 4.6.4. In our case study, the variable *Number of infants born by caesarean* presents a large $\mathrm{SS_D}$, since caesarean sections are operations which are rarely scheduled during a week-end. On the other hand, the $\mathrm{SS_D}$ is small for the variable *Number of infants with a mother followed by a midwife*. Using the different terms of this ANOVA, the variance of $\hat{t}_\mathrm{Y}$ can be rewritten as

$$\mathrm{V_{CCS}}\left(\hat{t}_\mathrm{Y}\right) = \mathrm{V_1}\left(\hat{t}_\mathrm{Y}\right) + \mathrm{V_2}\left(\hat{t}_\mathrm{Y}\right) + \mathrm{V_3}\left(\hat{t}_\mathrm{Y}\right) \tag{4.4.2}$$

where

$$\begin{aligned}
\mathrm{V_1}\left(\hat{t}_\mathrm{Y}\right) &= \left(\frac{1}{n_\mathrm{D}} - \frac{1}{\mathrm{N_D}}\right) \frac{\mathrm{N_D^2 N_M}}{\mathrm{N_D} - 1} \; \mathrm{SS_D} \\
\mathrm{V_2}\left(\hat{t}_\mathrm{Y}\right) &= \left(\frac{1}{n_\mathrm{M}} - \frac{1}{\mathrm{N_M}}\right) \frac{\mathrm{N_M^2 N_D}}{\mathrm{N_M} - 1} \; \mathrm{SS_M} \\
\mathrm{V_3}\left(\hat{t}_\mathrm{Y}\right) &= \left(\frac{1}{n_\mathrm{D}} - \frac{1}{\mathrm{N_D}}\right)\left(\frac{1}{n_\mathrm{M}} - \frac{1}{\mathrm{N_M}}\right) \frac{\mathrm{N_D^2}}{\mathrm{N_D} - 1} \frac{\mathrm{N_M^2}}{\mathrm{N_M} - 1} \; \mathrm{SS_E}.
\end{aligned}$$

We note that the CCS variance is divided into three terms associated respectively to a maternity effect, a day effect and a residual effect. On the other hand, the two-stage variance was divided into two terms associated to a maternity effect and to a residual effect. The term $\mathrm{SS_M}$ is the same in both decompositions, but the term $\mathrm{SS_E}$ is obviously different.

4.4.4 Estimating the variance

A term by term unbiased estimator of the variance of $\hat{t}_{\mathrm{Y}}$ in formula (4.4.2) is presented in section 4.6.5. This variance estimator simplifies as

$$\hat{\mathbf{V}}_{\mathrm{CCS}}\left(\hat{t}_{\mathrm{Y}}\right)=\hat{\mathbf{V}}_{\mathrm{D}}\left(\hat{t}_{\mathrm{Y}}\right)+\hat{\mathbf{V}}_{\mathrm{M}}\left(\hat{t}_{\mathrm{Y}}\right)-\hat{\mathbf{V}}_{\mathrm{E}}\left(\hat{t}_{\mathrm{Y}}\right) \tag{4.4.3}$$

where

$$\begin{aligned}
\hat{\mathbf{V}}_{\mathrm{D}}\left(\hat{t}_{\mathrm{Y}}\right) &= \left(\frac{1}{n_{\mathrm{D}}}-\frac{1}{\mathrm{N}_{\mathrm{D}}}\right)\frac{\mathrm{N}_{\mathrm{D}}^{2}}{n_{\mathrm{D}}-1}\frac{\mathrm{N}_{\mathrm{M}}^{2}}{n_{\mathrm{M}}}\; ss_{\mathrm{D}},\\
\hat{\mathbf{V}}_{\mathrm{M}}\left(\hat{t}_{\mathrm{Y}}\right) &= \left(\frac{1}{n_{\mathrm{M}}}-\frac{1}{\mathrm{N}_{\mathrm{M}}}\right)\frac{\mathrm{N}_{\mathrm{M}}^{2}}{n_{\mathrm{M}}-1}\frac{\mathrm{N}_{\mathrm{D}}^{2}}{n_{\mathrm{D}}}\; ss_{\mathrm{M}},\\
\hat{\mathbf{V}}_{\mathrm{E}}\left(\hat{t}_{\mathrm{Y}}\right) &= \left(\frac{1}{n_{\mathrm{M}}}-\frac{1}{\mathrm{N}_{\mathrm{M}}}\right)\left(\frac{1}{n_{\mathrm{D}}}-\frac{1}{\mathrm{N}_{\mathrm{D}}}\right)\frac{\mathrm{N}_{\mathrm{M}}^{2}\mathrm{N}_{\mathrm{D}}^{2}}{(n_{\mathrm{M}}-1)(n_{\mathrm{D}}-1)}\; ss_{\mathrm{E}},
\end{aligned}$$

where the terms come from an ANOVA decomposition on the sample $s_{\mathrm{M}} \times s_{\mathrm{D}}$ as detailed in section 4.6.4. The variance estimator is divided into three terms : $\hat{\mathbf{V}}_{\mathrm{D}}\left(\hat{t}_{\mathrm{Y}}\right)$ which represents an inter-day effect, $\hat{\mathbf{V}}_{\mathrm{M}}\left(\hat{t}_{\mathrm{Y}}\right)$ which represents an inter-maternity effect, and $\hat{\mathbf{V}}_{\mathrm{E}}\left(\hat{t}_{\mathrm{Y}}\right)$ which represents a residual effect.

4.4.5 Estimation in practice

To the best of our knowledge, there are no direct procedures in the softwares R, SAS and Stata to calculate CCS variance estimates. In this paper, we develop R functions to estimate a total and a ratio along with variance estimators. More precisely, from a selected sample (using variable *Dummy_CCS* in the dataset, which takes the value 1 if the unit is selected in the SI × SI sample and 0 otherwise), the estimated total $\hat{t}_{\mathrm{X}}$ and its estimated variance can be calculated using the R functions *EstTccsSISI* and *EstVARTccsSISI* proposed in the supplementary material. In the frame Code 4.12, these functions require that you enter the cross-classified sample (matrix of size $n_{\mathrm{D}} \times n_{\mathrm{M}}$), the sample sizes n_{M} and n_{D} and the population sizes N_{M} and N_{D}.

```
> echCCS=tableR[tableR$Dummy_CCS==1,];attach(echCCS)
> n_m=320; n_d=25; N_m=544; N_d=365
> echXCCS=matrix(Xik,nrow=n_d)
> EstTccsSISI(ECH=echXCCS,n_m,n_d,N_m,N_d)

[1] 3981426

> EstVARTccsSISI(ECH=echXCCS,n_m,n_d,N_m,N_d)

[1] 307219631
```

Code 4.12 – R code and results when estimating the total and its variance $\hat{\mathbf{V}}_{\mathrm{CCS}}\left(\hat{t}_{\mathrm{Y}}\right)$

To estimate the ratio $t_{\mathrm{Y}}/t_{\mathrm{X}}$, the function *EstRccsSISI* can be used. The linearized variable for the ratio estimator is then calculated by *LinearizedR*, and is plugged in the function *EstVARTccsSISI* as illustrated in the frame Code 4.13.

```
> echYCCS=matrix(Yik,nrow=n_d)
> EstRccsSISI(ECHY=echYCCS,ECHX=echXCCS,n_m,n_d,N_m,N_d)

[1] 0.1495898

> LinR=LinearizedR(ECHY=echYCCS,ECHX=echXCCS,n_m,n_d,N_m,N_d)
> EstVARTccsSISI(ECH=LinR,n_m,n_d,N_m,N_d)

[1] 1.006684e-06
```

Code 4.13 – R code and results when estimating a ratio and its variance $\hat{\mathbf{V}}_{\mathrm{CCS}}\left(\hat{\mathrm{R}}_{\mathrm{Y}}\right)$

4.5 Illustration

A small simulation study is conducted to compare the performance of several variance estimators under a two-stage sampling design and under a CCS design. We also evaluate the performance of various variance estimators. For a two-stage sampling design where the primary units are the maternities and where the number of secondary units n_{D} is the same inside all the primary units, we calculated the unbiased variance estimator $\hat{\mathbf{V}}_{2d}$ as well as its first part $\hat{\mathbf{V}}_{2d,a}$. For the CCS design, the unbiased variance estimator $\hat{\mathbf{V}}_{\mathrm{CCS}}$ is calculated as well as $\hat{\mathbf{V}}_{2d}$ and we also calculate the first part $\hat{\mathbf{V}}_{2d,a}$ of $\hat{\mathbf{V}}_{2d}$ in order to examine the error due to using the two-stage variance estimator instead of the cross-classified variance estimator. The two sampling designs and the various variance estimators are summarized in Table 4.1.

SAMPLING DESIGN	
two-stage	cross-classified
UNBIASED VARIANCE ESTIMATOR	
$\hat{\mathbf{V}}_{2d}$ in (4.3.3)	$\hat{\mathbf{V}}_{\mathrm{CCS}}$ in (4.4.3)
APPROXIMATION	
$\hat{\mathbf{V}}_{2d,a}$ in (4.3.4)	$\hat{\mathbf{V}}_{2d}$ in (4.3.3) $\hat{\mathbf{V}}_{2d,a}$ in (4.3.4)

TABLE 4.1 – Variance estimators of two-stage sampling and CCS

For the two-stage sampling design, the {SI,SI} sampling is used : a sample $\mathrm{S_M}$ of n_{M} maternities is selected and in each selected maternity, a sample s_{D} of size n_{D} is selected. For the CCS design, the SI × SI sampling is used : a sample $\mathrm{S_D}$ of n_{D} days, and a sample $\mathrm{S_M}$ of n_{M} maternities are selected. We used various sample sizes are used, namely n_{M} or n_{D} equal to 5, 25 and 320 (the two last sizes corresponding to the true ELFE sample sizes). These two sample selections were respectively repeated $\mathrm{B} = 10,000$ times. For CCS and for two-stage sampling, and in each of the $b = 1, \ldots, \mathrm{B}$ samples, the estimator $\hat{\mathrm{R}}^{(b)}$ of the ratio $\mathrm{R} = t_{\mathrm{Y}} / t_{\mathrm{X}}$ is computed. Also, for each cross-classified sample, the unbiased variance estimator $\hat{\mathrm{V}}^{(b)}_{\mathrm{CCS}}$ and the simplified variance estimators $\hat{\mathrm{V}}^{(b)}_{2d}$, $\hat{\mathrm{V}}^{(b)}_{2d,a}$ are computed, and for each two-stage sample, the unbiased variance estimator $\hat{\mathrm{V}}^{(b)}_{2d}$ and the simplified variance estimator $\hat{\mathrm{V}}^{(b)}_{2d,a}$ are computed. For each variance estimator $\hat{\mathrm{V}}$, the Monte Carlo Percent Relative Bias (RB), given by

$$\mathrm{RB_{MC}}(\hat{\mathrm{V}}) = 100 \times \frac{\mathrm{B}^{-1} \sum_{b=1}^{\mathrm{B}} \hat{\mathrm{V}}^{(b)} - \mathrm{V}}{\mathrm{V}}$$

is computed, where the true variance V was approximated through an independent set of 50,000 simulations.

Results for two ratios are reported in Table 4.2. In the top part of the table (case 1), we consider the plug-in estimator $\hat{t}_Y/\hat{t}_X$ of the proportion of infants with a mother followed by a midwife. In the bottom part of the table (case 2), we consider the plug-in estimator $\hat{t}_Z/\hat{t}_X$ of the proportion of infants born by caesarean. As expected, the variance estimator $\hat{V}_{CCS}$ is unbiased for the CCS variance, and the variance estimator $\hat{V}_{2d}$ is unbiased for the two-stage sampling variance. For the two-stage sampling, the estimator $\hat{V}_{2d,a}$ gives a good approximation of $\hat{V}_{2d}$ when the sample size n_M is small (5 or 25). But it presents an important underestimation when n_M increases (320), especially when n_D is small (25) : -57 % for both cases. For the CCS, in all cases, the relative biases of $\hat{V}_{2d}$ and $\hat{V}_{2d,a}$ increase when n_M increases or when n_D decreases. The relative bias of $\hat{V}_{2d,a}$ is always greater than the relative bias of $\hat{V}_{2d}$. In case 2, for all samples sizes, the relative biases are larger than in case 1. In this case, the variable *Number of infants born by caesarean* that we use presents an important day variability. The approximation of $\hat{V}_{CCS}$ by $\hat{V}_{2d}$ or $\hat{V}_{2d,a}$, which captures principally maternity effect, is therefore not appropriate. In case 1, the day effect (of Y_{ik}) is not as strong as for case 2, and the relative biases are therefore smaller.

	n_M	5	25	320	25	320
	n_D	5	25	25	320	320
Case 1 : $\hat{t}_Y/\hat{t}_X$						
CCS	$RB_{MC}\left(\hat{V}_{CCS}\right)$	0	0	-1	-1	-1
	$RB_{MC}\left(\hat{V}_{2d}\right)$	1	-1	-16	-1	-5
	$RB_{MC}\left(\hat{V}_{2d,a}\right)$	-0	-5	-64	-1	-18
2d	$RB_{MC}\left(\hat{V}_{2d}\right)$	-0	0	-1	-1	0
	$RB_{MC}\left(\hat{V}_{2d,a}\right)$	-1	-4	-57	-2	-13
Case 2 : $\hat{t}_Z/\hat{t}_X$						
CCS	$RB_{MC}\left(\hat{V}_{CCS}\right)$	-2	1	-0	-1	1
	$RB_{MC}\left(\hat{V}_{2d}\right)$	-28	-63	-96	-16	-83
	$RB_{MC}\left(\hat{V}_{2d,a}\right)$	-29	-65	-98	-17	-85
2d	$RB_{MC}\left(\hat{V}_{2d}\right)$	-0	-1	-1	0	-0
	$RB_{MC}\left(\hat{V}_{2d,a}\right)$	-1	-5	-57	-0	-13

TABLE 4.2 – Comparison between variance estimators of the estimated ratio for CCS and two-stage sampling (2d)

In the two-stage sampling design case, this simulation study recalls that the variance estimator $\hat{V}_{2d,a}$ is a fair approximation for $\hat{V}_{2d}$ only if the first stage sampling rate is small. The results also indicate that it seems hazardous to approximate a CCS variance estimator by a two-stage sampling variance estimator with a first stage on the maternity population. The behaviour of this simplified estimator depends on the importance of the day effect contained in the interest variable, and also depends on the sample sizes. In the ELFE case ($n_M = 320$ and $n_D = 25$), the underestimation is very high and its use is therefore not recommended. In Juillard et al. [2016], some alternative variance estimators are studied and proposed for a CCS design.

All the results of this paper are reproducible using the supplementary files (in Appendix C) which contain programming codes.

Acknowledgements

I thank Guillaume Chauvet and Anne Ruiz-Gazen for their help.

4.6 Supplementary Materials

4.6.1 Models used to generate the variables

In the dataset delivered with this article, the count variable X_{ik} is randomly generated by a Poisson distribution with parameter P_{ik}, generated according to the model

$$200 + \sigma_1 U_i + \sigma_2 V_k + \sigma_3 W_{ik} \tag{4.6.1}$$

where U_i, V_k and W_{ik} are independently generated with a distribution $N(0,1)$ and with $\sigma_1 = 2$ and $\sigma_2 = \sigma_3 = 0.2$.
Conditionally to the value of x_{ik}, the variable Y_{ik} (respectively Z_{ik}) is a binomial variable of parameters x_{ik} and p^Y_{ik} (respectively p^Z_{ik}). The probabilities p^Y_{ik} and respec-

tively p_{ik}^{Z} are dependent on i and k :

$$\begin{aligned} p_{ik}^{Y} &= \frac{e^{\beta A_{ik}}}{1+e^{\beta A_{ik}}} \\ p_{ik}^{Z} &= \frac{e^{\beta B_{ik}}}{1+e^{\beta B_{ik}}} \end{aligned}$$

where the variable A_{ik} (respectively B_{ik}) is generated according to the model (4.6.1) with $\sigma_1 = \sigma_2 = \sigma_3 = 0.2$ (respectively $\sigma_2 = 2$, $\sigma_1 = \sigma_3 = 0.2$) and β is chosen in order to the average probability is 0.3.

4.6.2 Term by term variance estimation for two-stage sampling

The variance in (4.3.1) may be unbiasedly estimated term by term by

$$\hat{\mathbf{V}}_{2d}\left(\hat{t}_Y\right) = \hat{\mathbf{V}}_{PSU}\left(\hat{t}_Y\right) + \hat{\mathbf{V}}_{SSU}\left(\hat{t}_Y\right)$$

where

$$\begin{aligned} \hat{\mathbf{V}}_{PSU}\left(\hat{t}_Y\right) &= \hat{\mathbf{V}}^{1}_{PSU}\left(\hat{t}_Y\right) - \hat{\mathbf{V}}^{2}_{PSU}\left(\hat{t}_Y\right), \\ \hat{\mathbf{V}}_{SSU}\left(\hat{t}_Y\right) &= \left(\frac{N_M}{n_M}\right)^2 \sum_{u_i\in S_M} N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right) s^2_{Y_{i\circ}} \\ \hat{\mathbf{V}}^{1}_{PSU}\left(\hat{t}_Y\right) &= N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right) s^2_{\hat{Y}_{\circ\bullet}}, \\ \hat{\mathbf{V}}^{2}_{PSU}\left(\hat{t}_Y\right) &= \frac{N_M^2}{n_M}\left(\frac{1}{n_M}-\frac{1}{N_M}\right) \sum_{u_i\in S_M} N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right) s^2_{Y_{i\circ}}. \end{aligned}$$

4.6.3 Analogy between two-stage sampling and one-way ANOVA : formula details

Analysis of variance (ANOVA) uses the partitioning of sums of squared deviations. For one-way ANOVA, the total sum of squares $SS_T = \sum_{u_i\in U_M}\sum_{k\in u_i}\left(Y_{ik}-\bar{Y}_{\bullet\bullet}\right)^2$ may be written as

$$\mathrm{SS_T} \quad = \quad \mathrm{SS_M} + \mathrm{SS_E}.$$

We have

$$\mathrm{SS_M} \quad = \quad \sum_{u_i \in \mathrm{U_M}} \sum_{k \in u_i} \left(\bar{\mathrm{Y}}_{i\bullet} - \bar{\mathrm{Y}}_{\bullet\bullet}\right)^2$$

the explained sum of squares (a.k.a. the sum of squares between classes), where $\bar{\mathrm{Y}}_{\bullet\bullet} = \mathrm{N}^{-1} \sum_{u_i \in \mathrm{U_M}} \sum_{k \in u_i} \mathrm{Y}_{ik}$ is the population mean and $\bar{\mathrm{Y}}_{i\bullet} = \mathrm{N}_i^{-1} \sum_{k \in u_i} \mathrm{Y}_{ik}$ is the mean inside the Primary Sampling Unit u_i. Also,

$$\mathrm{SS_E} \quad = \quad \sum_{u_i \in \mathrm{U_M}} \sum_{k \in u_i} \left(\mathrm{Y}_{ik} - \bar{\mathrm{Y}}_{i\bullet}\right)^2$$

denotes the residual sum of squares (a.k.a. sum of squares within classes).

In what follows, the factor, which is the categorical variable used to explain Y, is the belonging to one particular PSU u_i ($\mathrm{N_M}$ modalities). The total number of cases is $\mathrm{N} = \sum_{u_i \in \mathrm{U_M}} \mathrm{N}_i$. We consider the {SI,SI} sampling case, and assume for simplicity that all the PSUs are of the same size $\mathrm{N}_i = \mathrm{N_D}$, and that the same sample size $n_i = n_\mathrm{D}$ is used inside any selected PSU. In this case, we have

$$\begin{aligned} \mathrm{SS_M} &= \frac{\mathrm{N_M} - 1}{\mathrm{N_D}} \mathrm{S}^2_{\mathrm{Y}_{\circ\bullet}} \\ \mathrm{SS_E} &= (\mathrm{N_D} - 1) \sum_{u_i \in \mathrm{U_M}} \mathrm{S}^2_{\mathrm{Y}_{i\bullet}}. \end{aligned}$$

Now, we use ANOVA on the sample $\mathrm{S_M} \times \mathrm{S_D}$. The total number of cases is $n = n_\mathrm{M} \times n_\mathrm{D}$, and we denote

$$\begin{aligned} ss_{\mathrm{T}} &= \sum_{u_i \in \mathrm{S_M}} \sum_{k \in \mathrm{S}_i} \left(\mathrm{Y}_{ik} - \hat{\hat{\mathrm{Y}}}_{\bullet\bullet}\right)^2 \\ &= ss_{\mathrm{M}} + ss_{\mathrm{E}}, \end{aligned}$$

where

$$\begin{aligned} ss_{\mathrm{M}} &= \sum_{u_i \in \mathrm{S_M}} \sum_{k \in \mathrm{S}_i} \left(\hat{\hat{\mathrm{Y}}}_{i\bullet} - \hat{\hat{\mathrm{Y}}}_{\bullet\bullet}\right)^2 \\ ss_{\mathrm{E}} &= \sum_{u_i \in \mathrm{S_M}} \sum_{k \in \mathrm{S}_i} \left(\mathrm{Y}_{ik} - \hat{\hat{\mathrm{Y}}}_{i\bullet}\right)^2 \end{aligned}$$

with $\hat{\hat{\mathrm{Y}}}_{i\bullet} = \frac{1}{n_{\mathrm{D}}} \sum_{u_i \in \mathrm{S_D}} \mathrm{Y}_{ik}$ the estimated population mean inside u_i and $\hat{\hat{\mathrm{Y}}}_{\bullet\bullet} = \frac{1}{n} \sum_{u_i \in \mathrm{S_M}} \sum_{k \in \mathrm{S}_i} \mathrm{Y}_{ik}$ the estimated population mean.

4.6.4 Analogy between CCS and two-way ANOVA : formula details

For a two-way ANOVA without replication, the total sum of squares $\mathrm{SS_T} = \sum_{u_i \in \mathrm{U_M}} \sum_{k \in u_i} \left(\mathrm{Y}_{ik} - \bar{\mathrm{Y}}_{\bullet\bullet}\right)^2$ may be written as

$$\mathrm{SS_T} = \mathrm{SS_M} + \mathrm{SS_D} + \mathrm{SS_E}.$$

The total number of cases is $\mathrm{N} = \mathrm{N_M} \times \mathrm{N_D}$. We have

$$\mathrm{SS_M} = \mathrm{N_D} \sum_{i \in \mathrm{U_M}} \left(\bar{\mathrm{Y}}_{i\bullet} - \bar{\mathrm{Y}}_{\bullet\bullet}\right)^2$$

the sum of squares explained by the belonging to one particular unit i ($\mathrm{N_M}$ modalities), where $\bar{\mathrm{Y}}_{\bullet\bullet} = \frac{1}{\mathrm{N}} \sum_{i \in \mathrm{U_M}} \sum_{k \in \mathrm{U_D}} \mathrm{Y}_{ik}$ is the population mean and $\bar{\mathrm{Y}}_{i\bullet} = \frac{1}{\mathrm{N_D}} \sum_{k \in \mathrm{U_D}} \mathrm{Y}_{ik}$ is the mean inside the unit i. Then, we have

$$\mathrm{SS_D} = \mathrm{N_M} \sum_{k \in \mathrm{U_D}} \left(\bar{\mathrm{Y}}_{\bullet k} - \bar{\mathrm{Y}}_{\bullet\bullet}\right)^2$$

the sum of squares explained by the belonging to one particular unit k ($\mathrm{N_D}$ modalities), where $\bar{\mathrm{Y}}_{\bullet k} = \frac{1}{\mathrm{N_M}} \sum_{i \in \mathrm{U_M}} \mathrm{Y}_{ik}$ is the mean inside the unit k. Also,

$$\mathrm{SS_E} \quad = \quad \sum_{i \in \mathrm{U_M}} \sum_{k \in \mathrm{U_D}} \left(\mathrm{Y}_{ik} - \bar{\mathrm{Y}}_{i\bullet} - \bar{\mathrm{Y}}_{\bullet k} + \bar{\mathrm{Y}}_{\bullet\bullet}\right)^2$$

denotes the residual sum of squares.

Now, we use ANOVA on the sample $\mathrm{S_M} \times \mathrm{S_D}$. The total number of cases is $n = n_\mathrm{M} \times n_\mathrm{D}$, and we denote

$$\begin{aligned} ss_\mathrm{T} &= \sum_{i \in \mathrm{S_M}} \sum_{k \in \mathrm{S_D}} \left(\mathrm{Y}_{ik} - \hat{\bar{\mathrm{Y}}}_{\bullet\bullet}\right)^2 \\ &= ss_\mathrm{M} + ss_\mathrm{D} + ss_\mathrm{E}, \end{aligned}$$

where

$$\begin{aligned} ss_\mathrm{M} &= \sum_{i \in \mathrm{S_M}} \sum_{k \in \mathrm{S_D}} \left(\hat{\bar{\mathrm{Y}}}_{i\bullet} - \hat{\bar{\mathrm{Y}}}_{\bullet\bullet}\right)^2 \\ ss_\mathrm{D} &= \sum_{i \in \mathrm{S_M}} \sum_{k \in \mathrm{S_D}} \left(\hat{\bar{\mathrm{Y}}}_{\bullet k} - \hat{\bar{\mathrm{Y}}}_{\bullet\bullet}\right)^2 \\ ss_\mathrm{E} &= \sum_{i \in \mathrm{S_M}} \sum_{k \in \mathrm{S_D}} \left(\mathrm{Y}_{ik} - \hat{\bar{\mathrm{Y}}}_{i\bullet} - \hat{\bar{\mathrm{Y}}}_{\bullet k} + \hat{\bar{\mathrm{Y}}}_{\bullet\bullet}\right)^2 \end{aligned}$$

with $\hat{\bar{\mathrm{Y}}}_{\bullet k} = \frac{1}{n_\mathrm{M}} \sum_{i \in \mathrm{S_M}} \mathrm{Y}_{ik}$ the estimated population mean inside the unit k, $\hat{\bar{\mathrm{Y}}}_{i\bullet} = \frac{1}{n_\mathrm{D}} \sum_{i \in \mathrm{S_D}} \mathrm{Y}_{ik}$ the estimated population mean inside the unit i and $\hat{\bar{\mathrm{Y}}}_{\bullet\bullet} = \frac{1}{n} \sum_{i \in \mathrm{S_M}} \sum_{k \in \mathrm{S_D}} \mathrm{Y}_{ik}$ the estimated population mean.

4.6.5 Term by term variance estimation for CCS

A term by term unbiased estimator of the variance of $\hat{t}_\mathrm{Y}$ in formula (4.4.2) is

$$\hat{\mathbf{V}}_\mathrm{CCS}\left(\hat{t}_\mathrm{Y}\right) \quad = \quad \hat{\mathbf{V}}_1\left(\hat{t}_\mathrm{Y}\right) + \hat{\mathbf{V}}_2\left(\hat{t}_\mathrm{Y}\right) + \hat{\mathbf{V}}_3\left(\hat{t}_\mathrm{Y}\right)$$

with

$$
\begin{aligned}
\hat{\mathbf{V}}_1\left(\hat{t}_{\mathrm{Y}}\right) &= \hat{\mathbf{V}}_{\mathrm{D}}\left(\hat{t}_{\mathrm{Y}}\right)-\hat{\mathbf{V}}_{\mathrm{E}}\left(\hat{t}_{\mathrm{Y}}\right)\\
\hat{\mathbf{V}}_2\left(\hat{t}_{\mathrm{Y}}\right) &= \hat{\mathbf{V}}_{\mathrm{M}}\left(\hat{t}_{\mathrm{Y}}\right)-\hat{\mathbf{V}}_{\mathrm{E}}\left(\hat{t}_{\mathrm{Y}}\right)\\
\hat{\mathbf{V}}_3\left(\hat{t}_{\mathrm{Y}}\right) &= \hat{\mathbf{V}}_{\mathrm{E}}\left(\hat{t}_{\mathrm{Y}}\right)
\end{aligned}
$$

where $\hat{\mathbf{V}}_{\mathrm{D}}\left(\hat{t}_{\mathrm{Y}}\right)$, $\hat{\mathbf{V}}_{\mathrm{M}}\left(\hat{t}_{\mathrm{Y}}\right)$ and $\hat{\mathbf{V}}_{\mathrm{E}}\left(\hat{t}_{\mathrm{Y}}\right)$ are done in formula (4.4.3).

References

Aragon, Y. and Ruiz-Gazen, A. (2004). Utilisation des procédures sas dans l'enseignement des sondages. In Dunod, editor, *Echantillonnage et méthodes d'enquêtes*. 109

Bellhouse, D. (1981). Spatial sampling in the presence of a trend. *Journal of Statistical Planning and Inference*, 5 :365–375. 95

Dalén, J. and Ohlsson, E. (1995). Variance estimation in the swedish consumer price indexy. *Journal of Business & Economic Statistics*, 13(3) :347–356. 96, 97

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators : Linearization and residual techniques. *Survey Methodology*, 25(2) :193–203. 106

Juillard, H., Chauvet, G., and Ruiz-Gazen, A. (2016). Estimation under cross-classified sampling with application to a chilhood survey. *to appear in Journal of the American Statistical Association*. 97, 113, 119

Lumley, T. (2014). survey : analysis of complex survey samples. R package version 3.30. 106

Ohlsson, E. (1996). Cross-classified sampling. *Journal of Official Statistics*, 12(3) :241–251. 95, 114

R Core Team (2015). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 99

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York. 97, 102, 106

SAS Institute Inc. (2015). *SAS/STAT® 14.1 User's Guide*. Cary, NC. 99

Skinner, C. (2015). Cross-classified sampling : some estimation theory. *Statistics and Probability Letters*, 104 :163–168. 97

StataCorp. (2013). *Stata Statistical Software : Release 13*. StataCorp LP, College Station, TX. 99

Tillé, Y. and Matei, A. (2015). *sampling : Survey Sampling*. R package version 2.7. 100

Vos, J. (1964). Sampling in space and time. *Review of the International Statistical Institute*, 32(3) :226–241. 95

Wilkerson, M. (1957). Sampling error in the consumer price index. *Journal of the American Statistical Association*, 62(319) :899–914. 97

Chapitre 5

Variance estimation under monotone non-response for a panel survey

This chapter is a reprint of Juillard, H. and Chauvet, G. (2016). Variance estimation under monotone non-response for a panel survey. *Submitted*

Les enquêtes par panel sont souvent utilisées pour mesurer l'évolution de paramètres dans le temps. Les échantillons du panel peuvent souffrir de non-réponse totale, souvent corrigée en estimant les probabilités de réponse et en augmentant le poids des répondants. Dans ce travail, nous considérons des estimations et des estimations de variance pour une enquête par panel avec non-réponse d'unités dans le temps. Par extension du travail de Kim and Kim (2007), nous considérons un estimateur par expansion prenant en compte la non-réponse initiale et l'attrition, et proposons un estimateur de variance adapté. Ces résultats sont ensuite étendus afin de couvrir la plupart des estimateurs rencontrés dans les enquêtes, y compris les estimateurs calés, les paramètres complexes et les estimateurs longitudinaux. Les propriétés de l'estimateur de variance proposé et celles d'un estimateur de variance simplifié sont estimées à travers une étude par simulations. Une illustration des méthodes proposées est aussi présentée à partir des données de l'enquête Elfe.

Sommaire

Variance estimation under monotone non-response for a panel survey

Panel surveys are frequently used to measure the evolution of parameters over time. Panel samples may suffer from different types of unit non-response, which is currently handled by estimating the response probabilites and by reweighting respondents. In this work, we consider estimation and variance estimation under unit non-response for panel surveys. Extending the work by Kim and Kim [2007], we consider an expansion estimator accounting for initial non-response and attrition, and propose a suitable variance estimator. It is then extended to cover most estimators encountered in surveys, including calibrated estimators, complex parameters and longitudinal estimators. The properties of the proposed variance estimator and of a simplified variance estimator are estimated through a simulation study. An illustration of the proposed methods on data from the ELFE survey is also presented.

5.1 Introduction

Surveys are not only used to produce estimators for one point in time (cross-sectional estimations), but also to measure the evolution of parameters (longitudinal estimations), and are thus repeated over time. Kalton [2009] distinguishes three broad families of sampling designs for such surveys : the repeated cross-sectional surveys, in which estimations are produced through samples selected independently at each time ; the panel surveys, in which measures are repeated over time for units in a same sample ; the rotating panel surveys, which correspond to panel surveys with a sub-sample of units being replaced at each time by another incoming sub-sample. In this paper, we are interested in estimation and variance estimation for panel surveys.

Among the panel surveys (a.k.a. longitudinal surveys, see Lynn, 2009), cohort surveys are particular cases where the units in the sample are linked by a common original event, such as being born on the same year for children in the ELFE survey (Enquête Longitudinale Française depuis l'Enfance), which is the motivating example for this work. ELFE is the first longitudinal study of its kind in France, tracking children from birth to adulthood [Pirus et al., 2010]. Covering the whole metropolitan France, it was launched in 2011 and consists of more than 18,000 children whose parents consented to their inclusion. It will examine every aspect of these children's lives from the perspectives of health, social sciences and environmental health.

Surveys suffer from unit non-response, which needs to be accounted for by using available auxiliary information, so as to limit the bias of estimators. Non-response is currently handled by modeling the response probabilities [Kim and Kim, 2007] and by reweighting respondents with the inverse of these estimated probabilities. A panel sample may suffer from three types of unit non-response [Hawkes and Plewis, 2009] : initial non-response refers to the original absence of selected units ; wave non-response occurs when some units in the panel sample temporarily do not answer at some point in time, while attrition occurs when some units in the panel sample permanently do not answer from some point in time. Wave non-response was fairly uncommon in the first waves of the ELFE survey which were at our disposal. We therefore simplify this set-up by assuming monotone non-response, where only initial non-response and attrition occur.

There is a vast literature on the treatment of unit non-response for surveys over time, see Ekholm and Laaksonen [1991], Fuller et al. [1994], Rizzo et al. [1996], Clarke and Tate [2002], Laaksonen and Chambers [2006], Hawkes and Plewis [2009], Rendtel and Harms [2009], Laaksonen [2007], Slud and Bailey [2010], Zhou and Kim [2012]. Variance estimation for longitudinal estimators is considered in Tam [1984], Laniel [1988], Nordberg [2000], Berger [2004], Skinner and Vieira [2005], Qualité and Tillé [2008] and Chauvet and Goga [2016], but with focus on the sampling variance only. Variance estimation in case of non-response weighting adjustments on cross-sectional surveys is considered in Kim and Kim [2007]. To the best of our knowledge, and despite the interest for applications, variance estimation accounting for non-response for panel surveys has not been treated in the literature, with the exception of Zhou and Kim [2012].

The paper is organized as follows. Basic notations are given in Section 5.2. In Section 5.3, we define the expansion estimator for a total and a corresponding variance estimator when the response probabilities are assumed to be known at each time. Though this case appears unrealistic in most applications, it is common practice in some surveys that the response probabilities are assumed to be known without error to simplify variance estimation. Consequently, the simplified set-up in Section 5.3 enables to propose a first simplified variance estimator. In Section 5.4, we consider the usual case when the response probabilities are unknown. A parametric model is postulated leading to estimated response probabilities and to a reweighted estimator, and a variance estimator is derived by following the approach in Kim and Kim [2007]. Some illustrations on the particular important case of the response homogeneity groups are also given. The proposed variance estimator is extended to cover calibrated estimators and complex parameters in Section 5.5. Longitudinal estimation is discussed in Section 5.6, and the proposed variance estimator is used to cover such cases. The variance estimators are compared in Section 5.7 through a simulation study, and an illustration on the ELFE data is proposed in Section 5.8. We draw some conclusions in Section 5.9.

5.2 Notation

We are interested in a finite population U. A sample s_0 is first selected according to some sampling design $p(\cdot)$, and we assume that the first-order inclusion probabilities π_i are strictly positive for any $i \in U$. This first sampling phase corresponds to the original inclusion of units in the sample. For example, the 2011 ELFE survey involved the selection of a sample of babies according to a cross-classified sampling design (CCS), where a sample of maternity units and a sample of days were selected independently, the survey being performed in the maternity units selected on the days selected [Juillard et al., 2016]. Also, we note π_{ij} for the probability that units i and j are selected jointly in the sample, and $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$.

We consider the case of a panel survey in which the sole units in the original sample s_0 are followed over time, without reentry or late entry units at subsequent times to represent possible newborns. We are therefore interested in estimating some parameter defined over the population U, for some study variable y taking the value y_i for the unit i. The units in the sample s_0 are then followed at subsequent times $\delta = 1, \ldots, t$, and the sample is prone to unit non-response at each time. We note r_i^δ for the response indicator for unit i at time δ, and s_δ for the subset of respondents at time δ.

We assume monotone non-response resulting in the nested sequence

$$s_0 \supset s_1 \supset \ldots \supset s_t. \tag{5.2.1}$$

For $\delta = 1, \ldots, t$, we note

$$p_i^\delta \quad = \quad Pr\left(i \in s_\delta | s_{\delta-1}\right) \tag{5.2.2}$$

for the response probability of some unit i to be a respondent at time δ. We assume that the non-response mechanisms are ignorable, in the sense that the response probability p_i^δ at time δ can be explained by the variables observed at times $0, \ldots, \delta - 1$. Also, we assume that at any time δ the units answer independently of one another, and we note

$$p_{ij}^\delta \quad = \quad Pr\left(i, j \in s_\delta | s_{\delta-1}\right) = p_i^\delta p_j^\delta \tag{5.2.3}$$

for the probability that two distinct units i and j answer jointly at time δ.

5.3 Estimation with known response probabilities

5.3.1 Expansion estimator

We are interested in estimating the total $Y = \sum_{i \in U} y_i$. In a situation of full response, the Horvitz-Thompson estimator

$$\tilde{Y}_0 = \sum_{i \in s_0} \frac{y_i}{\pi_i} \tag{5.3.1}$$

is design-unbiased for Y. In the situation of unit non-response, the sub-sample s_t only is observed at time t. Assuming that the response probabilities at each time are known, the expansion estimator at time t is

$$\tilde{Y}_t = \sum_{i \in s_t} \frac{y_i}{\pi_i p_i^{1\to t}} \quad \text{with} \quad p_i^{1\to t} = \prod_{\delta=1}^{t} p_i^{\delta}. \tag{5.3.2}$$

The expansion estimator $\tilde{Y}_t$ is design unbiased for Y, provided that $p_i^{\delta} > 0$ for any unit $i \in s_{\delta-1}$ and at any time $\delta = 1, \ldots, t$, which we assume to hold in the rest of the paper. Here and elsewhere, the subscript δ will be used when the sample observed at time δ is used for estimation. The superscript δ will be used when we account for non-response at time δ, like for the probability p_i^{δ} of unit i to be a respondent at time δ.

5.3.2 Variance computation

At time t, we have

$$
\begin{aligned}
\mathrm{V}(\tilde{\mathrm{Y}}_t) &= \mathrm{VE}(\tilde{\mathrm{Y}}_t|s_{t-1}) + \mathrm{EV}(\tilde{\mathrm{Y}}_t|s_{t-1}) && (5.3.3)\\
&= \mathrm{V}(\tilde{\mathrm{Y}}_{t-1}) + \mathrm{EV}(\tilde{\mathrm{Y}}_t|s_{t-1}). && (5.3.4)
\end{aligned}
$$

Using a proof by induction, we obtain

$$\mathrm{V}(\tilde{\mathrm{Y}}_t) = \mathrm{V}(\tilde{\mathrm{Y}}_0) + \mathrm{E}\left\{\sum_{\delta=1}^{t} \mathrm{V}(\tilde{\mathrm{Y}}_\delta|s_{\delta-1})\right\}. \tag{5.3.5}$$

The first term in the right-hand side of (5.3.5) is the variance due to the sampling design, that we note as $\mathrm{V}^p(\tilde{\mathrm{Y}}_t)$, and that may be rewritten as

$$\mathrm{V}^p(\tilde{\mathrm{Y}}_t) = \sum_{i,j\in\mathrm{U}} \Delta_{ij}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}. \tag{5.3.6}$$

The second term in the right-hand side of (5.3.5) is the variance due to non-response, that we note as $\mathrm{V}^{nr}(\tilde{\mathrm{Y}}_t)$ and that may be rewritten as

$$\mathrm{V}^{nr}(\tilde{\mathrm{Y}}_t) = \mathrm{E}\left\{\sum_{\delta=1}^{t} \mathrm{V}^{nr\delta}(\tilde{\mathrm{Y}}_t)\right\} \tag{5.3.7}$$

with

$$\mathrm{V}^{nr\delta}(\tilde{\mathrm{Y}}_t) = \sum_{i\in s_{\delta-1}} p_i^\delta(1-p_i^\delta)\left(\frac{y_i}{\pi_i p_i^{1\to\delta}}\right)^2. \tag{5.3.8}$$

5.3.3 Variance estimation

At time t, an estimator for the variance due to the sampling design $\mathrm{V}^p(\tilde{\mathrm{Y}}_t)$ is

$$\hat{\mathrm{V}}_t^p(\tilde{\mathrm{Y}}_t) = \sum_{i,j\in s_t} \frac{\Delta_{ij}}{\pi_{ij}}\frac{1}{p_{ij}^{1\to t}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}, \tag{5.3.9}$$

where $p_{ij}^{1\to t} \equiv \prod_{\delta=1}^{t} p_{ij}^{\delta}$. This estimator is unbiased for $\mathrm{V}^p(\tilde{\mathrm{Y}}_t)$, provided that $\pi_{ij} > 0$ for any units $i \neq j \in \mathrm{U}$. An unbiased estimator for the variance due to non-response $\mathrm{V}^{nr}(\tilde{\mathrm{Y}}_t)$ is

$$\hat{\mathrm{V}}_t^{nr}(\tilde{\mathrm{Y}}_t) = \sum_{\delta=1}^{t} \hat{\mathrm{V}}_t^{nr\delta}(\tilde{\mathrm{Y}}_t) \tag{5.3.10}$$

with

$$\hat{\mathrm{V}}_t^{nr\delta}(\tilde{\mathrm{Y}}_t) = \sum_{i\in s_t} \frac{p_i^{\delta}(1-p_i^{\delta})}{p_i^{\delta\to t}} \left(\frac{y_i}{\pi_i p_i^{1\to\delta}}\right)^2. \tag{5.3.11}$$

By using the writing

$$\hat{\mathrm{V}}_t^{\mathrm{nr}\delta}(\tilde{\mathrm{Y}}_t) = \sum_{i\in s_t} \left(\frac{y_i}{\pi_i}\right)^2 \times \frac{1}{p_i^{1\to t}} \times \left(\frac{1}{p_i^{1\to\delta}} - \frac{1}{p_i^{1\to\delta-1}}\right), \tag{5.3.12}$$

and by summing for $\delta = 1, \ldots, t$, the estimator for the variance due to non-response simplifies as

$$\hat{\mathrm{V}}_t^{nr}(\tilde{\mathrm{Y}}_t) = \sum_{i\in s_t} \frac{1-p_i^{1\to t}}{\left(p_i^{1\to t}\right)^2} \left(\frac{y_i}{\pi_i}\right)^2. \tag{5.3.13}$$

This leads to the global variance estimator at time t

$$\hat{\mathrm{V}}_t(\tilde{\mathrm{Y}}_t) = \hat{\mathrm{V}}_t^{p}(\tilde{\mathrm{Y}}_t) + \hat{\mathrm{V}}_t^{nr}(\tilde{\mathrm{Y}}_t). \tag{5.3.14}$$

5.3.4 Application to Response Homogeneity Groups

For the purpose of illustration, we consider the model of Response Homogeneity Groups (RHG) which is often used in practice. More precisely, we assume that at each time $\delta = 1, \ldots, t$, the sub-sample $s_{\delta-1}$ may be partitioned into $\mathrm{C}(\delta-1)$ groups $s_{\delta-1}^{c}$, $c = 1, \ldots, \mathrm{C}(\delta-1)$, such that the response probability p_i^{δ} is constant inside a group. In such case, we simplify the notation as

$$p_i^{\delta} = p_c^{\delta} \quad \text{for any} \quad i \in s_{\delta-1}^c. \tag{5.3.15}$$

Note that the number of groups, and the groups themselves, may vary over time.

If the expansion estimator is computed at time $t = 1$, the estimator in (5.3.13) for the variance due to non-response may be rewritten as

$$\hat{\mathrm{V}}_1^{nr}(\tilde{\mathrm{Y}}_1) = \sum_{c=1}^{\mathrm{C}(0)} \frac{1-p_c^1}{(p_c^1)^2} \sum_{i \in s_1 \cap s_0^c} \left(\frac{y_i}{\pi_i}\right)^2, \tag{5.3.16}$$

and the global variance estimator at time 1 is

$$\hat{\mathrm{V}}_1(\tilde{\mathrm{Y}}_1) = \sum_{i,j \in s_1} \frac{\mathit{\Delta}_{ij}}{\pi_{ij}} \frac{1}{p_{ij}^1} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} + \sum_{c=1}^{\mathrm{C}(0)} \frac{1-p_c^1}{(p_c^1)^2} \sum_{i \in s_1 \cap s_0^c} \left(\frac{y_i}{\pi_i}\right)^2. \tag{5.3.17}$$

If the expansion estimator is computed at time $t = 2$, the estimator in (5.3.13) for the variance due to non-response may be rewritten as

$$\hat{\mathrm{V}}_2^{nr}(\tilde{\mathrm{Y}}_2) = \sum_{c=1}^{\mathrm{C}(0)} \sum_{d=1}^{\mathrm{C}(1)} \frac{1-p_c^1 p_d^2}{(p_c^1 p_d^2)^2} \sum_{i \in s_2 \cap s_1^d \cap s_0^c} \left(\frac{y_i}{\pi_i}\right)^2. \tag{5.3.18}$$

A simple case occurs when the same system of RHGs is kept over time. In this case, the number of groups at each time is equal to C(0), and we obtain a nested sequence of sub-samples

$$s_0^c \supset s_1^c \supset \ldots \supset s_t^c \quad \text{for any} \quad c = 1, \ldots, \mathrm{C}(0). \tag{5.3.19}$$

The variance estimator in (5.3.18) simplifies as

$$\hat{\mathrm{V}}_2^{nr}(\tilde{\mathrm{Y}}_2) = \sum_{c=1}^{\mathrm{C}(0)} \frac{1-p_c^{1\to 2}}{(p_c^{1\to 2})^2} \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_i}{\pi_i}\right)^2, \tag{5.3.20}$$

with $p_c^{1\to 2} = \prod_{\delta=1}^{2} p_c$ for $c = 1, \dots, \mathrm{C}(0)$.

5.4 Estimation with unknown response probabilities

5.4.1 Reweighted estimator

In practice, the response probabilities at each time are unknown and need to be estimated. We assume that at each time δ the probability of response is parametrically modeled as

$$p_i^\delta = p^\delta(z_i^\delta, \alpha^\delta) \tag{5.4.1}$$

for some known function $p^\delta(\cdot,\cdot)$, where z_i^δ is a vector of auxiliary variables observed for all the units in the subsample $s_{\delta-1}$, and α^δ denotes some unknown parameter. Following the approach in Kim and Kim [2007], we assume that the true parameter is estimated by $\hat{\alpha}^\delta$, which is the solution of the estimating equation

$$\frac{\partial}{\partial \alpha} \sum_{i \in s_{\delta-1}} k_i^\delta \left\{ r_i^\delta \ln(p_i^\delta) + (1 - r_i^\delta) \ln(1 - p_i^\delta) \right\} = 0, \tag{5.4.2}$$

with k_i^δ some weight of unit i in the estimating equation. Customary choices for these weights include $k_i^\delta = 1$ and $k_i^\delta = \pi_i^{-1}$, see Fuller and An [1998], Beaumont [2005] and Kim and Kim [2007].

The estimated response probability at time δ is

$$\hat{p}_i^\delta = p^\delta(z_i^\delta, \hat{\alpha}^\delta). \tag{5.4.3}$$

The reweighted estimator at time t is

$$\hat{Y}_t = \sum_{i \in s_\delta} \frac{y_i}{\pi_i \hat{p}_i^{1\to t}} \quad \text{with} \quad \hat{p}_i^{1\to t} = \prod_{\delta=1}^{t} \hat{p}_i^\delta. \tag{5.4.4}$$

It is obtained by substituting in (5.3.2) each unknown response probability p_i^δ with its estimator in (5.4.3). Under some mild assumptions on the response mechanisms and some regularity conditions on $p^\delta(\cdot,\cdot)$ (see conditions R.1 and R.2 in Kim and Kim, 2007), the reweighted estimator $\hat{Y}_t$ is approximately unbiased for Y.

5.4.2 Variance computation

The variance of $\hat{Y}_t$ is approximately given by

$$V(\hat{Y}_t) \simeq V(\tilde{Y}_0) + E\left\{\sum_{\delta=1}^{t} V(\hat{Y}_\delta | s_{\delta-1})\right\}. \tag{5.4.5}$$

The proof is similar to that of equation (5.3.5), and is thus omitted. The first term in the right-hand side of (5.4.5) is the variance due to the sampling design, that we note as $V^p(\hat{Y}_t)$. It is identical to the variance due to the sampling design for the expansion estimator. The second term in the right-hand side of (5.4.5) is the variance due to non-response, that we note as $V^{nr}(\hat{Y}_t)$. Adapting to the panel context equation (14) in Kim and Kim [2007], this variance is approximately given by

$$V^{nr}(\hat{Y}_t) \simeq E\left\{\sum_{\delta=1}^{t} V^{nr\delta}(\hat{Y}_t)\right\}, \tag{5.4.6}$$

where

$$V^{nr\delta}(\hat{Y}_t) = \sum_{i\in s_{\delta-1}} p_i^\delta(1-p_i^\delta)\left(\frac{y_i}{\pi_i \hat{p}_i^{1\to\delta-1} p_i^\delta} - k_i^\delta (h_i^\delta)^\top \gamma^\delta\right)^2, \tag{5.4.7}$$

where h_i^δ is the value of $h_i^\delta(\alpha) = \partial \text{logit}(p_i^\delta)/\partial\alpha$ evaluated at $\alpha = \alpha^\delta$, and where

$$\gamma^\delta = \left\{\sum_{i\in s_{\delta-1}} k_i^\delta p_i^\delta (1-p_i^\delta) h_i^\delta (h_i^\delta)^\top\right\}^{-1} \sum_{i\in s_{\delta-1}} \frac{1-p_i^\delta}{\hat{p}_i^{1\to\delta-1}} h_i^\delta \frac{y_i}{\pi_i}. \tag{5.4.8}$$

We now compare the variance due to non-response for the reweighted estimator with estimated response probabilities $\hat{Y}_t$, given in equation (5.4.6), and the variance due to non-response for the expansion estimator with known response probabilities $\tilde{Y}_t$, given in equation (5.3.7). For each of its component $\delta = 1,\ldots,t$, the term $V^{nr\delta}(\hat{Y}_t)$ in (5.4.7) includes a centering term $k_i^{\delta}(h_i^{\delta})^{\top}\gamma^{\delta}$, which is essentially a prediction of $(\pi_i \hat{p}_i^{1\to\delta-1} p_i^{\delta})^{-1} y_i$ by means of regressors h_i^{δ}. This centering is due to the estimation of the response probabilities, and therefore does not appear in equation (5.3.7). It usually leads to a smaller variance than that of $\tilde{Y}_t$; see also Beaumont [2005], equation (5.7) and Kim and Kim [2007], equation (17), for the case $t = 1$.

5.4.3 Variance estimation

At time t, an approximately unbiased estimator for the variance due to the sampling design $V^p(\hat{Y}_t)$ is

$$\hat{V}_t^p(\hat{Y}_t) = \sum_{i,j\in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}, \tag{5.4.9}$$

where $\hat{p}_{ij}^{1\to t} \equiv \prod_{\delta=1}^{t} \hat{p}_{ij}^{\delta}$, and where

$$\hat{p}_{ij}^{\delta} = \begin{cases} \hat{p}_i^{\delta} & \text{if } i = j, \\ \hat{p}_i^{\delta}\hat{p}_j^{\delta} & \text{otherwise.} \end{cases} \tag{5.4.10}$$

Following equation (25) in Kim and Kim [2007], $V^{nr}(\hat{Y}_t)$ may be approximately unbiasedly estimated at time t by

$$\hat{V}_t^{nr}(\hat{Y}_t) = \sum_{\delta=1}^{t} \hat{V}_t^{nr\delta}(\hat{Y}_t) \tag{5.4.11}$$

where

$$\begin{aligned}\hat{V}_t^{nr\delta}(\hat{Y}_t) &= \sum_{i\in s_t} \frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} \left(\frac{y_i}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^\delta (\hat{h}_i^\delta)^\top \hat{\gamma}_t^\delta\right)^2, & (5.4.12)\\ \hat{h}_i^\delta &= h(z_i, \hat{\alpha}^\delta), & (5.4.13)\\ \hat{\gamma}_t^\delta &= \left\{\sum_{i\in s_t} k_i^\delta \frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} \hat{h}_i^\delta (\hat{h}_i^\delta)^\top\right\}^{-1} \sum_{i\in s_t} \frac{1-\hat{p}_i^\delta}{\hat{p}_i^{1\to t}} \hat{h}_i^\delta \frac{y_i}{\pi_i}. & (5.4.14)\end{aligned}$$

This leads to the global variance estimator at time t

$$\hat{V}_t(\hat{Y}_t) = \hat{V}_t^p(\hat{Y}_t) + \hat{V}_t^{nr}(\hat{Y}_t). \qquad (5.4.15)$$

A simplified estimator of the variance due to non-response is obtained by ignoring the prediction terms $k_i^\delta (\hat{h}_i^\delta)^\top \hat{\gamma}_t^\delta$ for each of the $\delta = 1, \ldots, t$ variance components. Mimicking the reasoning in Section 5.3.3, this leads to the simplified variance estimator

$$\begin{aligned}\hat{V}_{t,simp}^{nr}(\hat{Y}_t) &= \sum_{\delta=1}^{t} \sum_{i\in s_t} \frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} \left(\frac{y_i}{\pi_i \hat{p}_i^{1\to\delta}}\right)^2\\ &= \sum_{i\in s_t} \frac{1-\hat{p}_i^{1\to t}}{\left(\hat{p}_i^{1\to t}\right)^2} \left(\frac{y_i}{\pi_i}\right)^2. \qquad (5.4.16)\end{aligned}$$

This simplified variance estimator is computed as if in the reweighted estimator $\hat{Y}_t$, the response probabilities were known. It will tend to overestimate the variance due to non-response of $\hat{Y}_t$ if the prediction term $k_i^\delta (h_i^\delta)^\top \gamma^\delta$ partly explains $(\pi_i \hat{p}_i^{1\to\delta-1} p_i^\delta)^{-1} y_i$.

5.4.4 Application to the logistic regression model

In the particular case when a logistic regression model is used at each time δ, the model (5.4.1) may be rewritten as

$$\text{logit}(p_i^\delta) = (z_i^\delta)^\top \alpha^\delta. \qquad (5.4.17)$$

We obtain $\hat{h}_i^\delta = z_i^\delta$, and the estimator for the variance due to non-response is given by (5.4.11), with

$$\hat{V}_t^{nr\delta}(\hat{Y}_t) = \sum_{i\in s_t} \frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} \left(\frac{y_i}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^\delta (z_i^\delta)^\top \hat{\gamma}_t^\delta \right)^2, \quad (5.4.18)$$

$$\hat{\gamma}_t^\delta = \left\{ \sum_{i\in s_t} k_i^\delta \frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} z_i^\delta (z_i^\delta)^\top \right\}^{-1} \sum_{i\in s_t} \frac{1-\hat{p}_i^\delta}{\hat{p}_i^{1\to t}} z_i^\delta \frac{y_i}{\pi_i}. \quad (5.4.19)$$

If the reweighted estimator is computed at time $t=1$, the estimator in (5.4.11) for the variance due to non-response may be rewritten as

$$\begin{aligned} \hat{V}_1^{nr}(\hat{Y}_1) &= \hat{V}_1^{nr,1}(\hat{Y}_1) \\ &= \sum_{i\in s_1} (1-\hat{p}_i^1) \left(\frac{y_i}{\pi_i \hat{p}_i^1} - k_i^1 (z_i^1)^\top \hat{\gamma}_1^1 \right)^2. \quad (5.4.20) \end{aligned}$$

If the reweighted estimator is computed at time $t=2$, the estimator in (5.4.11) for the variance due to non-response may be rewritten as

$$\begin{aligned} \hat{V}_2^{nr}(\hat{Y}_2) &= \hat{V}_2^{nr,1}(\hat{Y}_2) + \hat{V}_2^{nr,2}(\hat{Y}_2) \\ &= \sum_{i\in s_2} \frac{(1-\hat{p}_i^1)}{\hat{p}_i^2} \left(\frac{y_i}{\pi_i \hat{p}_i^1} - k_i^1 (z_i^1)^\top \hat{\gamma}_2^1 \right)^2 \\ &+ \sum_{i\in s_2} (1-\hat{p}_i^2) \left(\frac{y_i}{\pi_i \hat{p}_i^1 \hat{p}_i^2} - k_i^2 (z_i^2)^\top \hat{\gamma}_2^2 \right)^2. \quad (5.4.21) \end{aligned}$$

5.4.5 Application to Response Homogeneity Groups

We consider the model of Response Homogeneity Groups introduced in Section 5.3.4. At each time $\delta = 1, \ldots, t$, the sub-sample $s_{\delta-1}$ is partitioned into $C(\delta-1)$ groups $s_{\delta-1}^c$, $c = 1, \ldots, C(\delta-1)$. The response probabilities are assumed to be constant within the groups.

This model is equivalent to the logistic regression model in (5.4.17), with

$$z_i^{\delta} \;=\; \left[1\left\{i \in s_{\delta-1}^{1}\right\},\ldots,1\left\{i \in s_{\delta-1}^{C(\delta-1)}\right\}\right]^{\top}. \tag{5.4.22}$$

Solving the estimating equation (5.4.2) leads to the estimated response probabilities

$$\hat{p}_i^{\delta} = \frac{\sum_{i\in s_{\delta-1}^{c}} k_i^{\delta} r_i^{\delta}}{\sum_{i\in s_{\delta-1}^{c}} k_i^{\delta}} \quad \text{for} \quad i \in s_{\delta-1}^{c}. \tag{5.4.23}$$

That is, the response probability is estimated by the weighted response rate inside the RHG.

We first consider the case when the reweighted estimator is computed at time $t = 1$. In the estimator of the variance due to non-response given in (5.4.20), the vector $\hat{\gamma}_1^1$ simplifies as

$$\hat{\gamma}_1^1 \;=\; \left(\frac{\sum_{i\in s_1\cap s_0^1} \frac{y_i}{\pi_i}}{\hat{p}_1^1 \sum_{i\in s_1\cap s_0^1} k_i^1},\ldots,\frac{\sum_{i\in s_1\cap s_0^{C(0)}} \frac{y_i}{\pi_i}}{\hat{p}_{C(0)}^1 \sum_{i\in s_1\cap s_0^{C(0)}} k_i^1}\right)^{\top}. \tag{5.4.24}$$

After some algebra, the variance estimator in (5.4.20) may be rewritten as

$$\hat{V}_1^{nr}(\hat{Y}_1) \;=\; \sum_{c=1}^{C(0)} \frac{(1-\hat{p}_c^1)}{(\hat{p}_c^1)^2} \sum_{i\in s_1\cap s_0^c} \left(\frac{y_i}{\pi_i} - k_i^1 \frac{\sum_{j\in s_1\cap s_0^c} \frac{y_j}{\pi_j}}{\sum_{j\in s_1\cap s_0^c} k_j^1}\right)^2. \tag{5.4.25}$$

We now consider the case when the reweighted estimator is computed at time $t = 2$. We focus on the simpler case when the same system of RHGs is kept over time. In the estimator of the variance due to non-response given in (5.4.21), the vectors $\hat{\gamma}_2^1$ and $\hat{\gamma}_2^2$

simplify as

$$
\begin{aligned}
\hat{\mathrm{Y}}_2^1 &= \left(\frac{\sum_{i \in s_2 \cap s_1^1} \frac{y_i}{\pi_i}}{\hat{p}_1^1 \sum_{i \in s_2 \cap s_1^1} k_i^1}, \ldots, \frac{\sum_{i \in s_2 \cap s_1^{\mathrm{C}(0)}} \frac{y_i}{\pi_i}}{\hat{p}_{\mathrm{C}(0)}^1 \sum_{i \in s_2 \cap s_1^{\mathrm{C}(0)}} k_i^1} \right)^\top, \qquad (5.4.26) \\
\hat{\mathrm{Y}}_2^2 &= \left(\frac{\sum_{i \in s_2 \cap s_1^1} \frac{y_i}{\pi_i}}{\hat{p}_1^1 \hat{p}_1^2 \sum_{i \in s_2 \cap s_1^1} k_i^2}, \ldots, \frac{\sum_{i \in s_2 \cap s_1^{\mathrm{C}(0)}} \frac{y_i}{\pi_i}}{\hat{p}_{\mathrm{C}(0)}^1 \hat{p}_{\mathrm{C}(0)}^2 \sum_{i \in s_2 \cap s_1^{\mathrm{C}(0)}} k_i^2} \right)^\top. \qquad (5.4.27)
\end{aligned}
$$

After some algebra, the variance estimator in (5.4.21) may be rewritten as

$$
\begin{aligned}
\hat{\mathrm{V}}_2^{nr}(\hat{\mathrm{Y}}_2) &= \sum_{c=1}^{\mathrm{C}(0)} \frac{(1-\hat{p}_c^1)}{\hat{p}_c^2} \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_i}{\pi_i \hat{p}_c^1} - k_i^1 \frac{\sum_{j \in s_2 \cap s_1^c} \frac{y_j}{\pi_j}}{\sum_{j \in s_2 \cap s_1^c} k_j^1} \right)^2 \\
&+ \sum_{c=1}^{\mathrm{C}(0)} (1-\hat{p}_c^2) \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_i}{\pi_i \hat{p}_c^1 \hat{p}_c^2} - k_i^2 \frac{\sum_{j \in s_2 \cap s_1^c} \frac{y_j}{\pi_j}}{\sum_{j \in s_2 \cap s_1^c} k_j^2} \right)^2. \qquad (5.4.28)
\end{aligned}
$$

If we further assume that k_i^δ is constant over times $\delta = 1, 2$, and may thus be rewritten as k_i, the expression in (5.4.28) simplifies as

$$
\hat{\mathrm{V}}_2^{nr}(\hat{\mathrm{Y}}_2) = \sum_{c=1}^{\mathrm{C}(0)} \frac{(1-\hat{p}_c^{1\to 2})}{(\hat{p}_c^{1\to 2})^2} \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_i}{\pi_i} - k_i \frac{\sum_{j \in s_2 \cap s_1^c} \frac{y_j}{\pi_j}}{\sum_{j \in s_2 \cap s_1^c} k_j} \right)^2. \qquad (5.4.29)
$$

with $\hat{p}_c^{1\to 2} = \prod_{\delta=1}^2 \hat{p}_c^\delta$ for $c = 1, \ldots, \mathrm{C}(0)$. This simplification of the variance estimator can be extended to the reweighted estimator computed at time t. Assuming that the RHGs are kept over time, and that $k_i^\delta = k_i$ for any $\delta = 1, \ldots, t$, the variance estimator in (5.4.11) may be written after some algebra as

$$
\hat{\mathrm{V}}_t^{nr}(\hat{\mathrm{Y}}_t) = \sum_{c=1}^{\mathrm{C}(0)} \frac{(1-\hat{p}_c^{1\to t})}{(\hat{p}_c^{1\to t})^2} \sum_{i \in s_t \cap s_{t-1}^c} \left(\frac{y_i}{\pi_i} - k_i \frac{\sum_{j \in s_t \cap s_{t-1}^c} \frac{y_j}{\pi_j}}{\sum_{j \in s_t \cap s_{t-1}^c} k_j} \right)^2 \qquad (5.4.30)
$$

with $\hat{p}_c^{1\to t} = \prod_{\delta=1}^t \hat{p}_c^\delta$ for $c = 1, \ldots, \mathrm{C}(0)$.

5.5 Calibrated estimators and complex parameters

In most surveys, a calibration step is used to obtain adjusted weights which enable to improve the accuracy of total estimates. Such calibrated estimators are considered in Section 5.5.1. Also, more complex parameters than totals are frequently of interest, and a linearization step can be used for variance estimation. This is the purpose of Section 5.5.2. The estimation of complex parameters with calibrated weights is treated in Section 5.5.3. In each case, explicit formulas for variance estimation and simplified variance estimation are derived, and the bias of the simplified variance estimator is discussed.

5.5.1 Variance estimation for calibrated total estimators

Assume that a vector x_i of auxiliary variables is available for any unit $i \in s_t$, and that the vector of totals X on the population U is known. Then an additional calibration step [Deville and Särndal, 1992] is usually applied to $\hat{Y}_t$. It consists in modifying the weights $d_{ti} = \pi_i^{-1}(\hat{p}_i^{1\to t})^{-1}$ to obtain calibrated weights w_{ti} which enable to match the real total X, in the sense that

$$\sum_{i \in s_t} w_{ti} x_i = X. \tag{5.5.1}$$

The new calibrated weights are chosen so as to minimize a distance function with the original weights, while satisfying (5.5.1). This leads to the calibrated estimator

$$\hat{Y}_{wt} = \sum_{i \in s_t} w_{ti} y_i. \tag{5.5.2}$$

Under some mild conditions on the chosen distance function, on the sampling design and on the response mechanisms, it can be shown that the calibrated estimator $\hat{Y}_{wt}$ is approximately unbiased for Y.

The estimated residual for the weighted regression of y_i on x_i is denoted by

$$\begin{aligned} e_i &= y_i - \hat{b}_t x_i & (5.5.3)\\ \text{with } \hat{b}_t &= \left(\sum_{i\in s_t} \frac{1}{\pi_i \hat{p}_i^{1\to t}} x_i x_i^\top\right)^{-1} \sum_{i\in s_t} \frac{1}{\pi_i \hat{p}_i^{1\to t}} x_i y_i. & (5.5.4)\end{aligned}$$

Replacing in (5.4.9) the variable y_i with e_i yields the estimator of the variance due to the sampling design

$$\hat{\mathrm{V}}_t^p(\hat{\mathrm{Y}}_{wt}) = \sum_{i,j\in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{e_i}{\pi_i} \frac{e_j}{\pi_j}. \quad (5.5.5)$$

Similarly, replacing in (5.4.11) the variable y_i with e_i yields the estimator of the variance due to the non-response

$$\begin{aligned} \hat{\mathrm{V}}_t^{nr}(\hat{\mathrm{Y}}_{wt}) &= \sum_{\delta=1}^{t} \sum_{i\in s_t} \frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} \left(\frac{e_i}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^\delta (\hat{h}_i^\delta)^\top \hat{\gamma}_{te}^\delta\right)^2 & (5.5.6)\\ \hat{\gamma}_{te}^\delta &= \left\{\sum_{i\in s_t} k_i^\delta \frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} \hat{h}_i^\delta (\hat{h}_i^\delta)^\top\right\}^{-1} \sum_{i\in s_t} \frac{1-\hat{p}_i^\delta}{\hat{p}_i^{1\to t}} \hat{h}_i^\delta \frac{e_i}{\pi_i}. & (5.5.7)\end{aligned}$$

The global variance estimator for $\hat{\mathrm{Y}}_{wt}$ is

$$\hat{\mathrm{V}}_t(\hat{\mathrm{Y}}_{wt}) = \hat{\mathrm{V}}_t^p(\hat{\mathrm{Y}}_{wt}) + \hat{\mathrm{V}}_t^{nr}(\hat{\mathrm{Y}}_{wt}). \quad (5.5.8)$$

The simplified estimator of the variance due to non-response is

$$\begin{aligned} \hat{\mathrm{V}}_{t,simp}^{nr}(\hat{\mathrm{Y}}_{wt}) &= \sum_{\delta=1}^{t} \sum_{i\in s_t} \frac{\hat{p}_i^\delta(1-\hat{p}_i^\delta)}{\hat{p}_i^{\delta\to t}} \left(\frac{e_i}{\pi_i \hat{p}_i^{1\to\delta}}\right)^2 \\ &= \sum_{i\in s_t} \frac{1-\hat{p}_i^{1\to t}}{(\hat{p}_i^{1\to t})^2} \left(\frac{e_i}{\pi_i}\right)^2. & (5.5.9)\end{aligned}$$

Here again, this simplified variance estimator ignores the prediction terms $k_i^\delta (\hat{h}_i^\delta)^\top \hat{\gamma}_{te}^\delta$. If all the auxiliary variables that are explanatory for y_i are included in the calibration, which means that the underlying calibration model is appropriate, then e_i is essentially a white noise. The explanatory power of $\hat{h}_i^\delta$ for e_i is then expected to be small, as well as the prediction term $k_i^\delta(\hat{h}_i^\delta)^\top \hat{\gamma}_{te}^\delta$. In such case, we expect the bias of the simplified variance estimator to be small. If it is believed that some important auxiliary variables are not included in the calibration, then there may remain in e_i some significant part of y_i that may not been explained by the sole x_i. In such case, there may remain some explanatory power for $\hat{h}_i^\delta$ on e_i, and the bias of the simplified variance estimator may be non-negligible.

5.5.2 Variance estimation for complex parameters

We may be interested in estimating more complex parameters than totals. Suppose that the variable of interest y is q-multivariate, and that the parameter of interest is $\theta = f(Y)$ with $f(\cdot)$ a known function. At time t, substituting $\hat{Y}_t$ into θ yields the plug-in estimator $\hat{\theta}_t = f(\hat{Y}_t)$. Under some mild regularity conditions on the function f, on the sampling design and on the response mechanisms (see Deville, 1999 ; Goga et al., 2009), the plug-in estimator $\hat{\theta}_t$ is approximately unbiased for θ.

The estimated linearized variable of θ is

$$u_i = \{f'(\hat{Y}_t)\}^\top y_i, \qquad (5.5.10)$$

with $f'(\hat{Y}_t)$ the q-vector of first derivatives of f at point $\hat{Y}_t$. Replacing in (5.4.9) the variable y_i with u_i yields the estimator of the variance due to the sampling design

$$\hat{V}_t^p(\hat{\theta}_t) = \sum_{i,j\in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{u_i}{\pi_i} \frac{u_j}{\pi_j}. \qquad (5.5.11)$$

Similarly, replacing in (5.4.11) the variable y_i with u_i yields the estimator of the variance due to the non-response

$$\begin{aligned}
\hat{V}_t^{nr}(\hat{\theta}_t) &= \sum_{\delta=1}^{t}\sum_{i\in s_t}\frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}}\left(\frac{u_i}{\pi_i\hat{p}_i^{1\to\delta}}-k_i^{\delta}(\hat{h}_i^{\delta})^{\top}\hat{\gamma}_{t\theta}^{\delta}\right)^2 && (5.5.12)\\
\hat{\gamma}_{t\theta}^{\delta} &= \left\{\sum_{i\in s_t}k_i^{\delta}\frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}}\hat{h}_i^{\delta}(\hat{h}_i^{\delta})^{\top}\right\}^{-1}\sum_{i\in s_t}\frac{1-\hat{p}_i^{\delta}}{\hat{p}_i^{1\to t}}\hat{h}_i^{\delta}\frac{u_i}{\pi_i}. && (5.5.13)
\end{aligned}$$

The global variance estimator for $\hat{\theta}_t$ is

$$\hat{V}_t(\hat{\theta}_t) = \hat{V}_t^{p}(\hat{\theta}_t)+\hat{V}_t^{nr}(\hat{\theta}_t). \quad (5.5.14)$$

The simplified estimator of the variance due to non-response is

$$\begin{aligned}
\hat{V}_{t,simp}^{nr}(\hat{\theta}_t) &= \sum_{\delta=1}^{t}\sum_{i\in s_t}\frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}}\left(\frac{u_i}{\pi_i\hat{p}_i^{1\to\delta}}\right)^2\\
&= \sum_{i\in s_t}\frac{1-\hat{p}_i^{1\to t}}{\left(\hat{p}_i^{1\to t}\right)^2}\left(\frac{u_i}{\pi_i}\right)^2. && (5.5.15)
\end{aligned}$$

The bias of this simplified variance estimator will depend on the explanatory power for $\hat{h}_i^{\delta}$ on the linearized variable u_i.

5.5.3 Variance estimation for complex parameters under calibration

The calibrated weights w_{ti} may also be used to obtain an estimator of the parameter θ at time t. Substituting $\hat{Y}_{wt}$ into $\theta = f(Y)$ yields the calibrated plug-in estimator $\hat{\theta}_{wt} = f(\hat{Y}_{wt})$. So as to obtain a variance estimator for $\hat{\theta}_{wt}$, we first compute the estimated linearized variable $u_i = \{f'(\hat{Y}_t)\}^{\top} y_i$. Then, we compute

$$\begin{aligned}
e_{\theta i} &= u_i-\hat{b}_{\theta t}x_i && (5.5.16)\\
\text{with } \hat{b}_{\theta t} &= \left(\sum_{i\in s_t}\frac{1}{\pi_i\hat{p}_i^{1\to t}}x_i x_i^{\top}\right)^{-1}\sum_{i\in s_t}\frac{1}{\pi_i\hat{p}_i^{1\to t}}x_i u_i. && (5.5.17)
\end{aligned}$$

Replacing in (5.4.9) the variable y_i with $e_{\theta i}$ yields the estimator of the variance due to the sampling design

$$\hat{\mathrm{V}}_t^p(\hat{\theta}_{wt}) \quad = \quad \sum_{i,j \in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{e_{\theta i}}{\pi_i} \frac{e_{\theta j}}{\pi_j}. \tag{5.5.18}$$

Similarly, replacing in (5.4.11) the variable y_i with $e_{\theta i}$ yields the estimator of the variance due to the non-response

$$\hat{\mathrm{V}}_t^{nr}(\hat{\theta}_{wt}) \quad = \quad \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}} \left(\frac{e_{\theta i}}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^{\delta} (\hat{h}_i^{\delta})^\top \hat{\gamma}_{te\theta}^{\delta} \right)^2 \tag{5.5.19}$$

$$\hat{\gamma}_{te\theta}^{\delta} \quad = \quad \left\{ \sum_{i \in s_t} k_i^{\delta} \frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}} \hat{h}_i^{\delta} (\hat{h}_i^{\delta})^\top \right\}^{-1} \sum_{i \in s_t} \frac{1-\hat{p}_i^{\delta}}{\hat{p}_i^{1\to t}} \hat{h}_i^{\delta} \frac{e_{\theta i}}{\pi_i}. \tag{5.5.20}$$

The global variance estimator for $\hat{\theta}_{wt}$ is

$$\hat{\mathrm{V}}_t(\hat{\theta}_{wt}) \quad = \quad \hat{\mathrm{V}}_t^p(\hat{\theta}_{wt}) + \hat{\mathrm{V}}_t^{nr}(\hat{\theta}_{wt}). \tag{5.5.21}$$

The simplified estimator of the variance due to non-response is

$$\begin{aligned} \hat{\mathrm{V}}_{t,simp}^{nr}(\hat{\theta}_{wt}) \quad &= \quad \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}} \left(\frac{e_{\theta i}}{\pi_i \hat{p}_i^{1\to\delta}} \right)^2 \\ &= \quad \sum_{i \in s_t} \frac{1-\hat{p}_i^{1\to t}}{(\hat{p}_i^{1\to t})^2} \left(\frac{e_{\theta i}}{\pi_i} \right)^2 . \end{aligned} \tag{5.5.22}$$

The bias of this simplified variance estimator will depend on the explanatory power for $\hat{h}_i^{\delta}$ on $e_{\theta i}$. Since the variable $e_{\theta i}$ is obtained as the residual in the regression of the linearized variable u_i on the calibration variables x_i, the explanatory power for $\hat{h}_i^{\delta}$ on $e_{\theta i}$ is expected to be small in practice, and the bias of the simplified variance estimator is expected to be small as well.

As an illustration, we consider the model of Response Homogeneity Groups, and the simple case when RHGs are kept over time and when $k_i^\delta = k_i$ for any $\delta = 1, \ldots, t$. In such case, the estimator of the variance due to the non-response in (5.5.19) may be rewritten as

$$\hat{\mathrm{V}}_t^{nr}(\hat{\theta}_{wt}) = \sum_{c=1}^{\mathrm{C}(0)} \frac{(1-\hat{p}_c^{1\to t})}{(\hat{p}_c^{1\to t})^2} \sum_{i\in s_t\cap s_{t-1}^c} \left(\frac{e_{\theta i}}{\pi_i} - k_i \frac{\sum_{j\in s_t\cap s_{t-1}^c} \frac{e_{\theta j}}{\pi_j}}{\sum_{j\in s_t\cap s_{t-1}^c} k_j} \right)^2 . \tag{5.5.23}$$

5.6 Longitudinal estimators

We may be interested in a change in parameters, such as the difference between the totals of a variable of interest measured at two different times $u < t$. Denoting by y_{ui} and y_{ti} the value of this variable of interest for unit i at times u and t, respectively, and denoting by $\mathrm{Y}(u) = \sum_{i\in \mathrm{U}} y_{ui}$ and $\mathrm{Y}(t) = \sum_{i\in \mathrm{U}} y_{ti}$ their totals, the parameter of interest is

$$\Delta(u \to t) \quad = \quad \mathrm{Y}(t) - \mathrm{Y}(u). \tag{5.6.1}$$

Since the variable y_{ui} is measured on all sub-samples $s_{u'}$ for $u' = u, \ldots, t$, there are several possible estimators for $\Delta(u \to t)$. For $u' = u, \ldots, t$, we denote by

$$\hat{\Delta}_{u't}(u \to t) \quad = \quad \sum_{i\in s_t} \frac{y_{ti}}{\pi_i \hat{p}_i^{1\to t}} - \sum_{i\in s_{u'}} \frac{y_{ui}}{\pi_i \hat{p}_i^{1\to u'}} \tag{5.6.2}$$

the estimator which makes use of the sample s_t for the estimation of $\mathrm{Y}(t)$, and of the sample $s_{u'}$ for the estimation of $\mathrm{Y}(u)$. The case $u' = u$ corresponds to the estimation of $\mathrm{Y}(u)$ on the largest available sub-sample, s_u. The case $u' = t$ corresponds to the estimation of $\mathrm{Y}(u)$ and $\mathrm{Y}(t)$ on the common sub-sample, s_t.

In the context of full response, several authors have recommended the estimator $\hat{\Delta}_{tt}(u \to t)$ which makes use of the common sample only, if the variables y_{ui} and y_{ti} are strongly positively correlated ; see Caron and Ravalet [2000], Qualité and Tillé

[2008], Goga et al. [2009], Chauvet and Goga [2016]. In our context, this choice may be heuristically justified as follows. For $u' < t$, and by conditioning on the sub-sample $s_{u'}$, we obtain

$$\mathrm{V}\left\{\hat{\Delta}_{u't}(u \to t)\right\} \simeq \mathrm{V}\left\{\sum_{i \in s_{u'}} \frac{y_{ti} - y_{ui}}{\pi_i \hat{p}_i^{1\to u'}}\right\} + \mathrm{EV}\left\{\sum_{i \in s_t} \frac{y_{ti}}{\pi_i \hat{p}_i^{1\to t}} \middle| s_{u'}\right\}, \quad (5.6.3)$$

$$\mathrm{V}\left\{\hat{\Delta}_{tt}(u \to t)\right\} \simeq \mathrm{V}\left\{\sum_{i \in s_{u'}} \frac{y_{ti} - y_{ui}}{\pi_i \hat{p}_i^{1\to u'}}\right\} + \mathrm{EV}\left\{\sum_{i \in s_t} \frac{y_{ti} - y_{ui}}{\pi_i \hat{p}_i^{1\to t}} \middle| s_{u'}\right\}. \quad (5.6.4)$$

In equations (5.6.3) and (5.6.4), the first term in the right-hand side is identical. If the variables y_{ui} and y_{ti} are positively correlated, then the difference $y_{ti} - y_{ui}$ is expected to be smaller than y_{ti}, so that the second term in the right-hand side of (5.6.4) is expected to be smaller than the second term in the right-hand side of (5.6.3). Therefore, the estimator $\hat{\Delta}_{tt}(u \to t)$ based on the common sample is expected to be more efficient in terms of variance.

The results of a small simulation study in Section 5.7.2 support this heuristic reasoning. Therefore, we focus only in this Section on the estimator $\hat{\Delta}_{tt}(u \to t)$ for the estimation of $\Delta(u \to t)$. Replacing in (5.4.9) the variable y_i with $y_{ti} - y_{ui}$ yields the estimator of the variance due to the sampling design

$$\hat{\mathrm{V}}_t^p\left\{\hat{\Delta}_{tt}(u \to t)\right\} = \sum_{i,j \in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{(y_{ti} - y_{ui})}{\pi_i} \frac{(y_{tj} - y_{uj})}{\pi_j}. \quad (5.6.5)$$

Similarly, replacing in (5.4.11) the variable y_i with $y_{ti} - y_{ui}$ yields the estimator of the variance due to the non-response

$$\hat{\mathrm{V}}_t^{nr}\left\{\hat{\Delta}_{tt}(u \to t)\right\} = \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^{\delta}(1 - \hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}} \left(\frac{y_{ti} - y_{ui}}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^{\delta}(\hat{h}_i^{\delta})^\top \hat{\gamma}_{t\Delta}^{\delta}\right)^2 \quad (5.6.6)$$

with

$$\hat{\gamma}_{t\Delta}^{\delta} = \left\{\sum_{i\in s_t} k_i^{\delta}\frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}}\hat{h}_i^{\delta}(\hat{h}_i^{\delta})^{\top}\right\}^{-1}\sum_{i\in s_t}\frac{1-\hat{p}_i^{\delta}}{\hat{p}_i^{1\to t}}\hat{h}_i^{\delta}\frac{y_{ti}-y_{ui}}{\pi_i}. \tag{5.6.7}$$

The global variance estimator for $\hat{\Delta}_{tt}(u\to t)$ is

$$\hat{V}_t\left\{\hat{\Delta}_{tt}(u\to t)\right\} = \hat{V}_t^p\left\{\hat{\Delta}_{tt}(u\to t)\right\}+\hat{V}_t^{nr}\left\{\hat{\Delta}_{tt}(u\to t)\right\}. \tag{5.6.8}$$

Variance estimation for measures of change is also considered in Berger [2004], Qualité and Tillé [2008], Goga et al. [2009], Chauvet and Goga [2016], among others.

The simplified estimator of the variance due to non-response is

$$\begin{aligned}\hat{V}_{t,simp}^{nr}(\hat{\Delta}_{tt}(u\to t)) &= \sum_{\delta=1}^{t}\sum_{i\in s_t}\frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}}\left(\frac{y_{ti}-y_{ui}}{\pi_i\hat{p}_i^{1\to\delta}}\right)^2\\ &= \sum_{i\in s_t}\frac{1-\hat{p}_i^{1\to t}}{(\hat{p}_i^{1\to t})^2}\left(\frac{y_{ti}-y_{ui}}{\pi_i}\right)^2.\end{aligned} \tag{5.6.9}$$

If the variables y_{ti} and y_{ui} are strongly positively correlated, the explanatory power for $\hat{h}_i^{\delta}$ on $y_{ti}-y_{ui}$ is expected to be small in practice. In such case, the bias of the simplified variance estimator is also expected to be small.

5.7 A simulation study

In this Section, several artificial populations are generated according to some superpopulation model described in Section 5.7.1. In Section 5.7.2, we consider several estimators for a change between totals, which illustrates the heuristic reasoning in Section 5.6. A Monte Carlo experiment is then presented in Section 5.7.3, and several variance estimators for estimating a total, a ratio or a parameter change are compared. The results from Tables 5.1 and 5.2 are readily reproducible using the R code provided in the supplementary material of the present paper.

5.7.1 Simulation set-up

We consider seven populations of size 10,000, each containing three variables of interest y_{1i}, y_{2i} and y_{3i} observed at times $t = 1, 2$ and 3, respectively. The variables of interest are generated according to the superpopulation model

$$y_{1i} = \alpha^0 + \alpha^a x_{ai} + \alpha^b x_{bi} + \sigma u_{1i}, \tag{5.7.1}$$

$$y_{2i} = \rho y_{1i} + \sigma u_{2i}, \tag{5.7.2}$$

$$y_{3i} = \rho y_{2i} + \sigma u_{3i}. \tag{5.7.3}$$

The auxiliary variables x_{ai} and x_{bi} are independently generated from a Gamma distribution with shape and scale parameters 2 and 1. Two other auxiliary variables x_{ci} and x_{di} are also independently generated from a Gamma distribution with shape and scale parameters 2 and 1. These two last variables are not related to the variables of interest. The variables u_{1i}, u_{2i} and u_{3i} are independently generated according to a standard normal distribution. We use $\alpha^0 = 10$, $\alpha^a = \alpha^b = 5$ and $\sigma = 10$, which leads to a coefficient of determination (R^2) in model (5.7.1) approximately equal to 0.50. The parameter ρ is set to 0 for population 1, 0.2 for population 2, 0.4 for population 3, 0.6 for population 4, 0.8 for population 5, 1.0 for population 6 and 1.2 for population 7.

For each population, a simple random sample s_0 of size $n = 1,000$ is selected. Three non-response phases are then successively simulated. At each phase $\delta = 1, 2, 3$, the sub-sample of respondents s_δ is obtained by Poisson sampling with a response probability p_i^δ for unit i, defined as

$$\text{logit}(p_i^\delta) = \beta^{\delta 0} + \beta^{\delta a} x_{ai} + \beta^{\delta b} x_{bi}. \tag{5.7.4}$$

We use $\beta^{\delta 0} = -1$ at each phase $\delta = 1, 2, 3$. For $\delta = 1$, we use $\beta^{1a} = \beta^{1b} = 0.60$, which corresponds to an average response rate of 0.75. For $\delta = 2, 3$, we use $\beta^{\delta a} = \beta^{\delta b} = 0.75$, which corresponds to an average response rate of 0.81. Inside each sub-sample s_δ, the estimated response probabilities $\hat{p}_i^\delta$ are obtained by means of an unweighted logistic regression.

5.7.2 Comparison of estimators for a difference of totals

In this Section, we are interested in comparing the accuracy of two estimators for a difference of totals $\Delta(u \rightarrow t)$ for $u = 1$ and $t = 2$, for $u = 1$ and $t = 3$, and for $u = 2$ and $t = 3$. We consider the estimator $\hat{\Delta}_{ut}(u \rightarrow t)$, which makes use of the whole appropriate sub-samples for variables y_{ui} and y_{ti}, and the estimator $\hat{\Delta}_{tt}(u \rightarrow t)$, which makes use of the common sub-sample only. These two estimators are compared through the relative difference (RD) of their variances, which are defined as follows :

$$\begin{aligned}
\mathrm{RD}(1 \rightarrow 2) &= 100 \times \frac{\mathrm{V}\{\hat{\Delta}_{12}(1 \rightarrow 2)\} - \mathrm{V}\{\hat{\Delta}_{22}(1 \rightarrow 2)\}}{\mathrm{V}\{\hat{\Delta}_{22}(1 \rightarrow 2)\}}, && (5.7.5)\\
\mathrm{RD}(1 \rightarrow 3) &= 100 \times \frac{\mathrm{V}\{\hat{\Delta}_{13}(1 \rightarrow 3)\} - \mathrm{V}\{\hat{\Delta}_{33}(1 \rightarrow 3)\}}{\mathrm{V}\{\hat{\Delta}_{33}(1 \rightarrow 3)\}}, && (5.7.6)\\
\mathrm{RD}(2 \rightarrow 3) &= 100 \times \frac{\mathrm{V}\{\hat{\Delta}_{23}(2 \rightarrow 3)\} - \mathrm{V}\{\hat{\Delta}_{33}(2 \rightarrow 3)\}}{\mathrm{V}\{\hat{\Delta}_{33}(2 \rightarrow 3)\}}. && (5.7.7)
\end{aligned}$$

The true variances are replaced by their Monte Carlo approximation, obtained by repeating $\mathrm{B} = 100,000$ times the sample selection and the non-response phases.

The results are presented in Table 5.1. A positive RD indicates that the use of the common sample only leads to a more accurate estimator. As could be expected, the RD increases in all cases with ρ, that is, when the correlation between y_{ti} and y_{ui} increases. For $u = 1$ and $t = 2$, and for $u = 2$ and $t = 3$, the estimator $\hat{\Delta}_{tt}(u \rightarrow t)$ is more accurate for ρ greater than 0.6. For $u = 1$ and $t = 3$, $\hat{\Delta}_{tt}(u \rightarrow t)$ is more accurate for ρ greater than 0.8.

ρ	RD$(1 \rightarrow 2)$	RD$(1 \rightarrow 3)$	RD$(2 \rightarrow 3)$
0.0	-12	-27	-13
0.2	-09	-25	-11
0.4	-04	-20	-03
0.6	05	-09	11
0.8	17	11	39
1.0	30	33	83
1.2	40	46	127

TABLE 5.1 – Relative Difference (RD) between two estimators for a difference of totals

5.7.3 Performances of the variance estimators

In this Section, we consider the artificial population 5 ($\rho = 0.8$) generated as described in Section 5.7.1. The sample selection by means of simple random sampling of size $n = 1,000$ and the three non-response phases are applied $B = 5,000$ times. We are interested in evaluating the variance estimators and the simplified variance estimators, in case of estimating a total, a ratio or a change in totals.

As for the total Y, we consider at each time $t = 1,2,3$, three estimators. The estimator $\hat{Y}_t$ makes use of the weights $d_{ti} = \pi_i^{-1}(\hat{p}_i^{1\rightarrow t})^{-1}$. The estimator $\hat{Y}_{wt}$ makes use of the weights w_i, obtained by calibrating the weights d_{ti} on the population size and on the totals of the auxiliary variables x_{ai} and x_{bi}. In view of model (5.7.1), the working model underlying this calibration is well-specified. Finally, the estimator $\hat{Y}_{\tilde{w}t}$ makes use of the weights $\tilde{w}_i$, obtained by calibrating the weights d_{ti} on the population size and on the totals of the auxiliary variables x_{ci} and x_{di}. In view of model (5.7.1), the working model underlying this calibration is therefore not correctly specified. The proposed variance estimator for $\hat{Y}_t$ is obtained from equation (5.4.15), and the simplified variance estimator is obtained by plugging in (5.4.15) the simplified variance estimator for non-response given in (5.4.16). The proposed variance estimators for $\hat{Y}_{wt}$ and $\hat{Y}_{\tilde{w}t}$ are obtained from equation (5.5.8), and the simplified variance estimators are obtained by plugging in (5.5.8) the simplified variance estimator for non-response given in (5.5.9).

We are also interested in estimating the ratio $R_t = Y_t/Y_1$ for $t = 2,3$. At each time t, we consider three estimators. The estimator $\hat{R}_t$ makes use of the weights d_i. The proposed variance estimator is obtained from equation (5.5.14), by using the estimated linearized variable $u_i = (\hat{Y}_1)^{-1}(y_{ti} - \hat{R}_t y_{1i})$. The simplified variance estimator is obtained by plugging in (5.5.14) the simplified variance estimator for non-response given in (5.5.15). The estimators $\hat{R}_{wt}$ and $\hat{R}_{\tilde{w}t}$ make use of the calibrated weights w_i and $\tilde{w}_i$. The proposed variance estimators are obtained from equation (5.5.21). The simplified variance estimators are obtained by plugging in (5.5.21) the simplified variance estimator for non-response given in (5.5.22).

Finally, we are interested in estimating the change in totals $\Delta(1 \to t)$ for $t = 2,3$. At each time t, we consider three estimators. The estimator $\hat{\Delta}_{tt}(1 \to t)$ makes use of the weights d_i. The proposed variance estimator is obtained from equation (5.6.8), and the simplified variance estimator is obtained by plugging in (5.6.8) the simplified variance estimator for non-response given in (5.6.9). The estimators $\hat{\Delta}_{tt,w}(1 \to t)$ and $\hat{\Delta}_{tt,\tilde{w}}(1 \to t)$ make use of the calibrated weights w_i and $\tilde{w}_i$. The proposed variance estimators are obtained from equation (5.6.8), by replacing $y_{ti} - y_{ui}$ by the estimated residual for the weighted regression of $y_{ti} - y_{ui}$ on the calibration variables. The simplified variance estimators are obtained by plugging in (5.6.8) the simplified variance estimator for non-response given in (5.6.9).

For a proposed variance estimator $\hat{V}$, we computed the Monte Carlo Percent Relative Bias

$$RB_{MC}(\hat{V}) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{V}^{(b)} - V}{V}$$

where the global variance V was approximated through an independent set of 100,000 simulations. So as to evaluate the contribution of some component $\hat{V}_a$ into the proposed variance estimator $\hat{V}$, we also computed the contribution (in percent)

$$\mathrm{CONTR}_{\mathrm{MC}}(\hat{V}_a) = 100 \times \frac{\frac{1}{B}\sum_{b=1}^{B} \hat{V}_a^{(b)}}{\frac{1}{B}\sum_{b=1}^{B} \hat{V}^{(b)}}.$$

So as to evaluate the simplified variance estimator for the non-response $\hat{V}_{simp}^{nr}$, we also computed the Monte Carlo Percent Relative Bias

$$\mathrm{RB}_{\mathrm{MC}}(\hat{V}_{simp}^{nr}) = 100 \times \frac{B^{-1}\sum_{b=1}^{B} \hat{V}_{simp}^{(b)} - V^{nr}}{V^{nr}},$$

where the variance V^{nr} due to non-response was approximated through an independent set of $100,000$ simulations.

The simulation results are presented in Table 5.2. The proposed variance estimator is almost unbiased in all cases. As could be expected, the contribution of the variance due to the sampling design decreases with time, as the number of respondents decreases and as the variance due to non-response becomes larger. The simplified variance estimator is highly biased for the variance due to non-response in case of $\hat{Y}_t$. The bias decreases quickly with time, but remains large at time $t = 3$. The simplified variance estimator is almost unbiased for a calibrated estimator when the working model is adequately specified, but is severely biased otherwise. This is consistent with our reasoning in Section 5.5.1. The simplified variance estimator is almost unbiased for the three estimators of the ratio, and for the calibrated estimators of the change in totals. In case of the non-calibrated estimator for the change in totals, the bias can be as high as 30 % .

	$t=1$	$t=2$	$t=3$	$t=1$	$t=2$	$t=3$	$t=1$	$t=2$	$t=3$
	$\hat{Y}_t$			$\hat{Y}_{wt}$			$\hat{Y}_{\tilde{w}t}$		
$RB_{MC}(\hat{V})$	-0	-1	-2	-1	-1	-2	-1	-1	-3
$CONTR_{MC}(\hat{V}_t^p)$	81	57	35	69	49	32	80	56	35
$CONTR_{MC}(\hat{V}_t^{nr1})$	19	19	13	31	22	15	20	18	13
$CONTR_{MC}(\hat{V}_t^{nr2})$	-	25	18	-	28	19	-	25	17
$CONTR_{MC}(\hat{V}_t^{nr3})$	-	-	34	-	-	34	-	-	34
$RB_{MC}(\hat{V}_{simp}^{nr})$	559	188	80	0	-1	-2	83	34	15
	$\hat{R}_t$			$\hat{R}_{wt}$			$\hat{R}_{\tilde{w}t}$		
$RB_{MC}(\hat{V})$	-	-0	-2	-	-1	-2	-	-1	-2
$CONTR_{MC}(\hat{V}_t^p)$	-	49	32	-	49	32	-	50	33
$CONTR_{MC}(\hat{V}_t^{nr1})$	-	22	15	-	22	15	-	22	15
$CONTR_{MC}(\hat{V}_t^{nr2})$	-	28	19	-	28	19	-	28	19
$CONTR_{MC}(\hat{V}_t^{nr3})$	-	-	34	-	-	34	-	-	34
$RB_{MC}(\hat{V}_{simp}^{nr})$	-	0	0	-	-1	-2	-	-1	-1
	$\hat{\Delta}_{tt}(1 \rightarrow t)$			$\hat{\Delta}_{tt,w}(1 \rightarrow t)$			$\hat{\Delta}_{tt,\tilde{w}}(1 \rightarrow t)$		
$RB_{MC}(\hat{V})$	-	-0	-2	-	-0	-2	-	-1	-3
$CONTR_{MC}(\hat{V}_t^p)$	-	50	33	-	49	32	-	50	33
$CONTR_{MC}(\hat{V}_t^{nr1})$	-	22	14	-	22	15	-	22	14
$CONTR_{MC}(\hat{V}_t^{nr2})$	-	28	18	-	28	19	-	28	18
$CONTR_{MC}(\hat{V}_t^{nr3})$	-	-	34	-	-	34	-	-	34
$RB_{MC}(\hat{V}_{simp}^{nr})$	-	19	30	-	-1	-2	-	3	5

TABLE 5.2 – Relative bias of a global variance estimator, relative contribution to the estimators of variance components and relative bias of a simplified variance estimator for the variance due to non-response for the estimation of a total, a ratio or a change in totals with three sets of weights

5.8 Illustration

In this Section, we aim at illustrating the results previously obtained on a real data set from the ELFE survey. Covering the whole metropolitan France, it was launched in 2011 and consists of more than 18,000 children whose parents consented to their inclusion. The population of inference consists of infants born in one of the 544 French maternity units during 2011, except very premature infants.

An original sample s_0 of about $35,600$ infants was originally selected when the babies were just a few days old and were still at the maternity unit. The sample was selected using a cross-classified sampling design (Skinner, 2015 ; Juillard et al., 2016). A sample of days and a sample of maternity units were independently selected, and both sample selections may be approximated by stratified simple random sampling (STSI). The sample sizes inside strata are provided in Tables 5.3 and 5.4. The sample consisted in all the infants born during one of the 25 selected days in one of the 320 selected maternity units.

Strata g	Strata size N_{Mg}	Sample size n_{Mg}
1	108	21
2	108	41
3	109	55
4	108	80
5	111	90
Total	544	287

TABLE 5.3 – Population and sample strata sizes for the maternity units design.

Among the $35,600$ infants originally selected, a total of $18,329$ face-to-face interviews were completed with their families, which represents a response rate of 51 % . This led to the subsample s_1 after accounting for non-response. The weights at time $t=1$ were computed on the basis of the original sampling weights, adjusted in two steps. First, response probabilities were estimated by means of a model of Response Homo-

Strata h	Strata size N_{Dh}	Sample size n_{Dh}
1	91	4
2	91	6
3	91	7
4	92	8
Total	365	25

TABLE 5.4 – Population and sample strata sizes for the days design.

geneity Groups (RHGs), with 20 RHGs defined by using a logistic regression model with explanatory variables *Age of the mother*, *Gemellary identity* and *Season of birth*. Then, a calibration by means of the raking ratio method was performed on the binary variables *Born within marriage*, *Immigrant mother* and *Gemellary identity*.

When the children reached the age of two months, the parents had the first telephone interview with a response rate of 87 % . This leads to the subsample s_2. The weights at time $t = 2$ were computed on the basis on the weight obtained at time $t = 1$, with a two-step adjustment. First, response probabilities were estimated by means of 20 RHGs, defined by using a logistic regression with explanatory variables *Age of the mother*, *Mother nationality* and *Father present at childbirth*. Then, a calibration by means of the raking ratio method was performed on the same calibration variables as at time $t = 1$.

When the children were one year old, the parents were contacted by phone with a response rate of 77 % . This led to the subsample s_3. The weights at time $t = 3$ were computed on the basis on the weights obtained at time $t = 2$, with a two-step adjustment similar to that realized at time $t = 2$. The parents were expected to be also interviewed when the infants would reach the age of two, three and half years and five and half years, but at the time when the paper was written, the three first waves only were available.

We considered three variables of interest : *Breastfeeding exclusivity at the childbirth, at two month, at one year.* For each of these variables, we computed the estimator $\hat{Y}_t$ from equation (5.4.4) and the estimated variance $\hat{V}_t(\hat{Y}_t)$ from the equation (5.4.15). We also computed the estimated coefficient of variation (in percent), defined as

$$\widehat{CV}_t\left(\hat{Y}_t\right) \quad = \quad 100 \times \frac{\sqrt{\hat{V}_t\left(\hat{Y}_t\right)}}{\hat{Y}_t}. \tag{5.8.1}$$

For each component $\hat{V}_{ta}$ in the estimated variance $\hat{V}_t$, we computed its contribution (in percent) defined as

$$\mathrm{CONTR}(\hat{V}_{ta}) = 100 \times \frac{\hat{V}_{ta} - \hat{V}_t}{\hat{V}_t}. \tag{5.8.2}$$

We also computed the simplified variance estimator for non-response $\hat{V}^{nr}_{t,simp}$ given in (5.4.16), and the relative difference (in percent) with the approximately unbiased variance estimator $\hat{V}^{nr}$ defined as

$$\mathrm{RD}(\hat{V}^{nr}_{simp}) \quad = \quad 100 \times \frac{\hat{V}^{nr}_{simp} - \hat{V}^{nr}_t}{\hat{V}^{nr}_t}. \tag{5.8.3}$$

The results are given in the upper left of Table 5.5. For each of the three variables of interest, we also computed the calibrated estimator $\hat{Y}_{wt}$, and the same indicators. They are given in the upper right of Table 5.5. Finally, for each variable interest, we computed the estimator $\hat{R}_t$ and the calibrated estimator $\hat{R}_{wt}$ for the percentage of breastfeeding among all the children. The same indicators were computed. They are presented in the lower part of Table 5.5. As observed in the simulation study, the RD of the simplified variance estimator for non-response can be large in case of the estimator of the total without calibration, but the bias decreases with time. The bias appears as negligible for the calibrated estimator of the total, and for both estimators of the ratio.

Breastfeeding	$t=1$	$t=2$	$t=3$	$t=1$	$t=2$	$t=3$
exclusivity	maternity	2 months	1 year	maternity	2 months	1 year
	without calibration			with calibration		
$\hat{Y}_t$	402409	209009	22658	415272	214262	23276
$\hat{V}_t(\cdot)$	8.51E+07	2.32E+07	1.58E+06	5.95E+06	6.93E+06	1.21E+06
$\hat{CV}_t(\cdot)$ (%)	2.3	2.3	5.6	0.6	1.2	4.7
CONTR($\hat{V}_t^p$)	94	78	42	28	34	25
CONTR($\hat{V}_t^{nr1}$)	6	17	32	72	51	42
CONTR($\hat{V}_t^{nr2}$)	-	5	10	-	15	13
CONTR($\hat{V}_t^{nr3}$)	-	-	16	-	-	21
RD($\hat{V}^{nr}_{simp}$)	91	31	3	1	2	0
$\hat{R}_t$ (%)	59.0	30.6	3.3	59.4	31.0	3.4
$\hat{V}(\hat{R}_t)$	1.34E-05	1.50E-05	2.58E-06	1.28E-05	1.48E-05	2.60E-06
$\hat{CV}(\hat{Y}_t)$ (%)	0.6	1.3	4.8	0.6	1.2	4.7
CONTR($\hat{V}_t^p$)	31	34	24	28	34	25
CONTR($\hat{V}_t^{nr1}$)	69	51	42	72	51	41
CONTR($\hat{V}_t^{nr2}$)	-	15	13	-	15	13
CONTR($\hat{V}_t^{nr3}$)	-	-	21	-	-	21
RD($\hat{V}^{nr}_{simp}$)	2	2	0	1	2	0

TABLE 5.5 – Estimates for a total and a ratio, variance estimates, estimated coefficient of variation, relative contributions of variance components and relative difference of a simplified variance estimator for some variables in the ELFE survey

5.9 Conclusion

In this paper, we considered variance estimation accounting for weighting adjustments in panel surveys. We proposed both an approximately unbiased variance estimator and a simplified variance estimator for estimators of totals, complex parameters and measures of change, which covers most cases that may be encountered in practice. Our simulation results indicate that the proposed variance estimator performs well in all cases considered. The simplified variance estimator tends to overestimate the variance of the expansion estimator for totals, and to overestimate the variance for calibrated estimators when the calibration variables lack of explanatory power for the variable of interest.

The assumption of independent response behaviour is usually not tenable for multistage surveys, since units within clusters tend to be correlated with respect to the response behaviour. In this context, estimation of response probabilities based upon conditional logistic regression in the context of correlated responses has been studied by Skinner and D'Arrigo [2011], see also Kim et al. [2016]. Extending the present work in the context of correlated response behaviour is a challenging problem for further research.

5.10 Supplementary Materials

Basic functions required to calculate estimators and commands that display the results in Table 5.1 and Table 5.2 are available in Appendix D.

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society B*, 67 :445–458. 137, 139

Berger, Y. (2004). Variance estimation for measures of change in probability sampling. *Canadian Journal of Statistics*, 32(4) :451–467. 130, 151

Caron, N. and Ravalet, P. (2000). Estimation dans les enquêtes répétées : application à l'enquête emploi en continu. *Technical report INSEE, Paris.* 149

Chauvet, G. and Goga, C. (2016). Linearization versus bootstrap for variance estimation of the change between Gini indexes. *In revision for Survey Methodology.* 130, 150, 151

Clarke, P. and Tate, P. (2002). An application of non-ignorable non-response models for gross flows estimation in the British labour force survey. *Australian and New Zealand Journal of Statistics*, 4 :413–425. 130

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators : linearization and residual techniques. *Survey Methodology*, 25 :193–203. 146

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87 :376–382. 144

Ekholm, A. and Laaksonen, S. (1991). Weighting via response modeling in the Finnish household budget survey. *Journal of Official Statistics*, 7 :325–327. 130

Fuller, W. and An, A. (1998). Regression adjustment for non-response. *Journal of the Indian Society of Agricultural Statistics*, 51 :331–342. 137

Fuller, W. A., Loughin, M. M., and Baker, H. D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 national food consumption survey. *Survey Methodology*, 20 :75–85. 130

Goga, C., Deville, J.-C., and Ruiz-Gazen, A. (2009). Composite estimation and linearization method for two-sample survey data. *Biometrika*, 96 :691–709. 146, 150, 151

Hawkes, D. and Plewis, I. (2009). Modelling nonresponse in the national child development study. *Journal of the royal Statistical Society Series A*, 169 :479–491. 130

Juillard, H., Chauvet, G., and Ruiz-Gazen, A. (2016). Estimation under cross-classified sampling with application to a childhood survey. *To appear in Journal of the American Statistical Association*. 132, 158

Kalton, G. (2009). Design for surveys over time. *Handbook of Statistics*, 29 :89–108. 129

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35 :501–514. 129, 130, 131, 137, 138, 139

Kim, J. K., Kwon, Y., and Park, M. (2016). Calibrated propensity score method for survey nonresponse in cluster sampling. *Biometrika*, 103 :461–473. 162

Laaksonen, S. (2007). Weighting for two-phase surveyed data. *Survey Methodology*, 33 :121–130. 130

Laaksonen, S. and Chambers, R. L. (2006). Survey estimation under informative nonresponse with follow-up. *Journal of Official Statistics*, 22 :81–95. 130

Laniel, N. (1988). Variances for a rotating sample from a changing population. *Proceedings of the Business and Economics Statistics Section, American Statistical Association*, pages 246–250. 130

Lynn, P. (2009). Methods for longitudinal surveys. *Methodology of Longitudinal Surveys*, pages 1–19. 129

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16 :363–378. 130

Pirus, C., Bois, C., Dufourg, M., Lanoë, J., Vandentorren, S., Leridon, H., and the Elfe team (2010). Constructing a cohort : Experience with the French Elfe project. *Population*, 65 :637–670. 130

Qualité, L. and Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34 :173–181. 130, 149, 151

Rendtel, U. and Harms, T. (2009). Weighting and calibration for household panels. *Methodology of Longitudinal Surveys*, pages 265–286. 130

Rizzo, L., Kalton, G., and Brick, J. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22 :43–53. 130

Skinner, C. (2015). Cross-classified sampling : some estimation theory. *Statistics and Probability Letters*, 104 :163–168. 158

Skinner, C. and D'Arrigo, J. (2011). Inverse probability weighting for clustered non-response. *Biometrika*, 98 :953–966. 162

Skinner, C. and Vieira, M. (2005). Design effects in the analysis of longitudinal survey data. *S3RI Methdology Working Papers, M05/13. Southampton, UK : Southampton Statistical Sciences Research Institute*. 130

Slud, E. V. and Bailey, L. (2010). Evaluation and selection of models for attrition nonresponse adjustment. *Journal of Official Statistics*, 26 :1–18. 130

Tam, S. (1984). On covariance from overlapping samples. *The American Statistician*, 38 :1–18. 130

Zhou, M. and Kim, J. (2012). An efficient method of estimation for longitudinal surveys with monotone missing data. *Biometrika*, 99 :631–648. 130, 131

Chapitre 6

Estimation de la variance pour l’enquête Elfe - Rapport méthodologique pour l’utilisateur

This chapter is a reprint of Juillard, H. (2016). Estimation de la variance pour l’enquête Elfe - Rapport méthodologique pour l’utilisateur. *Document de Travail de l’Ined, n°226.*

Chaque échantillonnage conduit à une variance dite d’échantillonnage. Cette variance est une mesure d’incertitude (ou de précision), relative au fait de sélectionner un échantillon et reflète la façon dont l’échantillon a été tiré. La cohorte Elfe (2011) comprend plus de 18 000 enfants dont les parents ont donné leur consentement à la maternité. Dans chacune des maternités sélectionnées, les nourrissons (de la population d’inférence), nés durant quatre périodes spécifiques représentant chacune des quatre saisons de l’année 2011, ont été sélectionnés. Le plan d’échantillonnage utilisé pour l’étude Elfe n’est pas standard, il s’agit d’un plan d’échantillonnage produit (cross-classified sampling design), avec les sélections indépendantes d’un échantillon de maternités et d’un échantillon de jours. Dans ce travail, des estimateurs sans biais sont dérivés, ainsi que des estimateurs spécifiques adaptés au plan Elfe. Les résultats sont illustrés et les procédures dans les logiciels R, SAS et Stata sont décrites.

Mots-clés *: estimation de variance, échantillonnage produit, procédures R / SAS / Stata.*

Sommaire

ESTIMATION DE LA VARIANCE POUR L'ENQUETE ELFE
Rapport méthodologique pour l'utilisateur

Hélène Juillard$^{(\star)}$ [1]

$^{(\star)}$ dans le cadre de ses travaux de doctorat financés par l'Ined
sous la direction d'Anne Ruiz-Gazen et de Guillaume Chauvet

Travail en collaboration avec Marie Cheminat, administrateur bases de données Elfe.
Ce document est destiné aux utilisateurs des données issues de la cohorte Elfe et propose un estimateur de variance prenant en compte le plan de sondage utilisé pour l'étude Elfe.

Chaque échantillonnage conduit à une variance dite d'échantillonnage. Cette variance est une mesure d'incertitude (ou de précision), relative au fait de sélectionner un échantillon et reflète la façon dont l'échantillon a été tiré. Dans le cas d'un recensement (tirage exhaustif), cette variance est nulle. Après déroulement d'une enquête, les informations relatives à un seul échantillon sont connues et les calculs du paramètre estimé $\hat{\theta}$ et de sa variance estimée $\hat{\mathbf{V}}(\hat{\theta})$ sont possibles. De ces calculs dépendront les intervalles de confiance associés à chaque paramètre estimé. Dans ce document, nous considérons uniquement des paramètres θ en population finie (totaux, ratios, coefficients de corrélation...) et nous supposons que l'aléa provient du tirage de l'échantillon (inférence basée sur le plan, voir Särndal, Swensson, et Wretman, 1992).

Le plan utilisé pour l'enquête Elfe n'est pas standard. Il s'agit du produit de deux échantillonnages indépendants suivi de plusieurs phases de non-réponse. Le calcul de l'estimateur de variance n'est pas directement disponible dans la littérature et fait l'objet d'un travail de recherche de la part de l'auteur de cette note. Ce document propose de détailler l'application au cas spécifique de l'enquête Elfe.

Une modélisation du plan de sondage est proposée, avec prise en compte de la non-réponse et du calage et les estimateurs de variance associés sont dérivés. L'estimateur de variance sans biais n'étant a priori programmé dans aucun logiciel, plusieurs estimateurs simplifiés prenant en compte les procédures logicielles déjà existantes (R / SAS / Stata) sont décrits et illustrés sur données Elfe. Après comparaison, un unique estimateur simplifié est recommandé aux utilisateurs des données Elfe. Les détails des calculs présentés dans ce document sont donnés dans Juillard, Chauvet, et Ruiz-Gazen [2016].

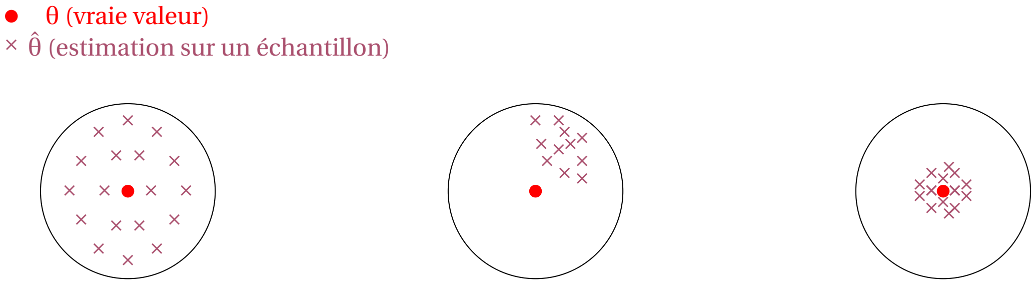
Il est conseillé à l'utilisateur de bien lire la première section. Un résumé de la suite du document est proposé en page 200.

[1] L'auteur remercie pour leurs conseils Emmanuel Gros et Stéphane Legleye.

6.1 Echantillonnage et variance

Dans les enquêtes, on s'intéresse à des populations de tailles finies, dans lesquelles on choisit parfois de sélectionner un échantillon : on parle alors d'enquête par sondage. On s'intéresse à un paramètre θ inconnu (calculable seulement sur toute la population) que l'on estime à partir d'un échantillon par $\hat{\theta}$ (l'accent circonflexe symbolise l'estimateur). On veut inférer les résultats de l'échantillon à la population. Par exemple, on veut estimer le nombre total de naissances sous césarienne qui ont eu lieu durant l'année 2011 en enquêtant seulement quelques maternités durant quelques jours.

Ce qui nous intéresse c'est de savoir si le $\hat{\theta}$ obtenu à partir de notre échantillon sélectionné est proche de θ. Si l'on avait sélectionné un autre échantillon, aurait-on obtenu le même $\hat{\theta}$? Et si l'on en avait sélectionné un autre ? Ou encore un autre ? C'est en imaginant toutes ces différentes valeurs de $\hat{\theta}$ que l'on peut se représenter la **variance dite d'échantillonnage** $\mathbf{V}(\hat{\theta})$ (voir la Figure 6.1).
En pratique, la variance $\mathbf{V}(\hat{\theta})$ est inconnue mais estimée sur l'échantillon par $\hat{\mathbf{V}}(\hat{\theta})$.



Les estimations sont dispersées
(grande variance)

Les estimations sont rapprochées
(petite variance)
mais éloignées de la cible (biais)

Les estimations sont rapprochées
(petite variance)
et centrées vers la cible :
situation recherchée

FIGURE 6.1 – Biais et variances

En théorie des sondages, l'aléa provient de la sélection de l'échantillon. Pour comparaison, en statistique classique, la variable d'intérêt y est une variable aléatoire, alors qu'en statistique d'enquête elle est fixée : c'est l'indicatrice d'appartenance à l'échantillon qui est aléatoire.

La méthode de tirage de l'échantillon est importante : les calculs de $\hat{\theta}$ et $\hat{\mathbf{V}}(\hat{\theta})$ vont en dépendre. Par exemple, si pour connaître le nombre de césariennes en 2011, on sélectionnait certaines régions de France, puis à l'intérieur de ces régions, certaines maternités, les formules associées à $\hat{\theta}$ et $\hat{\mathbf{V}}(\hat{\theta})$ ne seraient pas les mêmes que si l'on sélectionnait directement certaines maternités parmi toutes les maternités de France. Autrement dit, une formule de variance d'échantillonnage doit refléter la structure du plan de sondage.

La variance est donc une **mesure d'incertitude dépendant du plan de sondage** : rien ne nous dit que la valeur issue de notre échantillon est exacte, on l'espère seulement rapprochée de θ. Elle va par exemple permettre à notre estimateur $\hat{\theta}$ d'être associé à un intervalle de confiance, c'est-à-dire deux valeurs entre lesquelles θ aura 95 % (ou 90 %, ou 99 %...) de chance d'être compris. Si à partir d'une enquête, on estime le taux de césariennes à 20 %, ce chiffre est-il sûr, précis ? L'intervalle de confiance estimé construit autour de cette valeur est-il [19 %, 21 %] ? Ou plutôt [15 %, 25 %] ? **L'enquête effectuée, la précision issue du plan de sondage ne se choisit plus, c'est avant** que les choix sont faits concernant le plan d'échantillonnage et la précision qui en découle. Toutefois, il existe différentes méthodes post-échantillonnage usant d'informations auxiliaires, permettant d'améliorer la précision des estimateurs, comme le calage que nous verrons dans ce document.

6.2 L'enquête Elfe : contexte et modélisation du plan de sondage

La population d'inférence est celle des nourrissons nés durant l'année 2011 en France métropolitaine, issus d'un accouchement au plus gémellaire, hors grands prématurés, ayant une mère majeure, en mesure de donner un consentement éclairé notamment dans l'une des langues proposées (français, anglais, arabe ou turc), nés dans une maternité métropolitaine et dont les parents ne résidaient pas temporairement en métropole. Toutes les familles sélectionnées ont été enquêtées peu de temps après l'accouchement dans certaines maternités métropolitaines et durant certains jours de l'année (voir Figure 6.2).

Le plan de sondage pour les maternités est un plan probabiliste. Concernant les jours, 25 ont été choisis durant quatre périodes (appelées vagues) couvrant les quatre saisons de l'année[2] (dont la moitié devait coïncider avec l'échantillon démographique permanent E.D.P.). Notons que les deux échantillons (maternités et jours) ont été sélectionnés indépendamment.

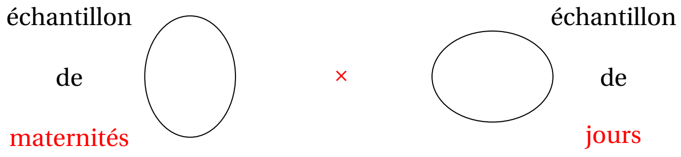


FIGURE 6.2 – Représentation schématique du plan de sondage utilisé pour l'enquête Elfe

L'échantillonnage probabiliste des maternités correspond à un plan stratifié : cinq strates à effectifs égaux avec tirages à allocation proportionnelle au nombre d'accouchements recensés en 2008. Il s'agissait d'un tirage systématique avec pour variables de stratification implicite le statut juridique de la maternité, le niveau de médicalisa-

[2] Du 1er au 4 avril, les 27 et 28 juin, du 1er au 4 juillet, du 27 au 29 septembre, du 1er au 4 octobre, du 28 au 30 novembre et du 1er au 5 décembre.

tion et la région en cinq postes. Par la suite, on supposera être dans le cas d'un plan stratifié avec plan SI (tirage aléatoire simple sans remise) dans chaque strate : plan STSI.

L'échantillonnage des jours n'est pas aléatoire, d'où la nécessité de le modéliser. Une modélisation est proposée dans la suite de ce document et deux autres modélisations possibles sont développées dans Juillard, Chauvet, et Ruiz-Gazen [2015a]. La modélisation proposée consiste en un plan STSI avec quatre strates (voir Figure 6.3) que nous nommerons pour simplifier “saisons" dans la suite de ce document et tirage SI à l'intérieur de chaque strate de respectivement 4, 6, 7 et 8 jours. Cette modélisation permet de représenter l'effet saisonnier du plan mais néglige l'effet grappe (jours presque consécutifs sélectionnés durant chaque saison).

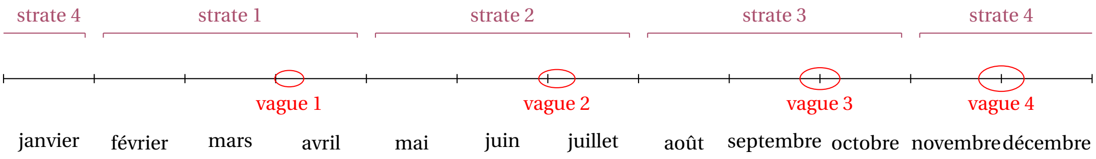


FIGURE 6.3 – Exemple de découpage de l'année 2011 en 4 strates et sélection de jours

Le plan de sondage final utilisé pour l'enquête Elfe résulte du croisement indépendant de ces deux plans de sondage (l'un dans la population des maternités, l'autre dans celle des jours) et est appelé *plan produit* (ou encore *cross-classified sampling*, voir Ohlsson, 1996 ou Skinner, 2015). L'échantillonnage bien particulier utilisé pour l'enquête Elfe est comparé dans la section suivante à d'autres plans de sondage, afin d'en comprendre les spécificités.

6.3 Plan produit et autres plans (que l'on ne peut confondre)

Notons $\mathrm{U_M}$ la population des maternités de taille $\mathrm{N_M}$ et $\mathrm{U_D}$ la population des jours de taille $\mathrm{N_D}$. Les indices i et j sont utilisés pour les maternités et les indices k et l pour les jours. On considère un plan de sondage p_{M} dans la population $\mathrm{U_M}$ menant à un échantillon S_M de taille n_M et un plan de sondage p_D dans la population U_D menant à un échantillon S_D de taille n_D. Pour un plan produit (le cas de l'enquête Elfe), ces deux plans sont indépendants (voir Figure 6.4). L'unité finale d'échantillonnage qui nous intéresse est caractérisée par un couple maternité × jour (i, k), avec $i \in U_M$ et $k \in U_D$.

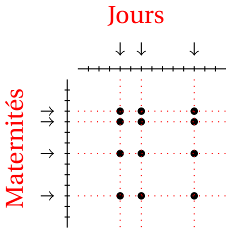


FIGURE 6.4 – Echantillonnage de maternités et de jours pour un plan produit

Si le plan produit est bien un plan dans la population produit $U_M \times U_D$, il est caractérisé par deux plans sources (tirage de i dans U_M, tirage de k dans U_D), et diffère d'un plan de sondage direct dans cette population, c'est-à-dire qui tirerait directement des unités (i, k) dans $U_M \times U_D$ comme illustré dans la Figure 6.5.

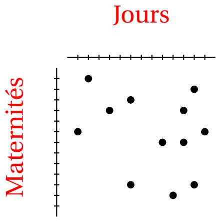


FIGURE 6.5 – Echantillonnage de maternités et de jours pour un tirage direct dans la population produit

Pour l'enquête Elfe, on distingue deux phases d'échantillonnage : celle sur les jours et celle sur les maternités. Néanmoins le plan Elfe ne peut être considéré comme un plan classique à deux degrés avec au premier degré un échantillonnage de maternités et au second degré un échantillonnage de jours (Figure 6.6). Il ne peut non plus symétriquement être considéré comme un plan classique à deux degrés avec au premier degré un échantillonnage des jours et au second degré un échantillonnage des maternités.

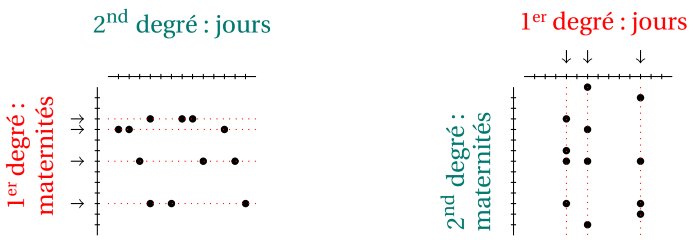


FIGURE 6.6 – Echantillonnage de maternités et de jours pour un plan à deux degrés, avec tirage de maternités au premier degré (à gauche) ou tirage de jours au premier degré (à droite)

Un plan classique à deux degrés requiert deux hypothèses : l'indépendance entre les tirages effectués à chaque degré, encore appelée propriété d'invariance ; l'indépendance entre les différents tirages effectués au second degré, conditionnellement au premier degré de tirage. Pour un plan produit, la première hypothèse est vérifiée (indépendance entre l'échantillon de maternités et l'échantillon de jours) mais la seconde ne l'est pas (le même échantillon de jours est utilisé pour chaque maternité).

Des estimateurs de variance pour un plan produit quelconque sont dérivés dans Juillard, Chauvet, et Ruiz-Gazen [2016]. Une comparaison a été effectuée entre la variance issue d'un plan produit et celle issue d'un plan à deux degrés. La comparaison entre ces deux plans est aussi détaillée dans Juillard [2016] avec possibilité de mettre en pratique sur les logiciels R, SAS et Stata, les étapes d'échantillonnage et d'estimation. Dans la suite de ce document, nous proposons des estimateurs pour le cas particulier de l'enquête Elfe.

6.4 Estimation de la variance issue du plan Elfe

On considère le plan de sondage pour lequel p_{M} est un plan aléatoire simple stratifié de taille $n_{\mathrm{M}g}$ à l'intérieur de chaque strate $\mathrm{U}_{\mathrm{M}g}$ de taille $\mathrm{N}_{\mathrm{M}g}$ avec $g = 1, \ldots, \mathrm{G}$ (voir le Tableau 6.1), et où p_{D} est un plan aléatoire simple stratifié de taille $n_{\mathrm{D}h}$ à l'intérieur de chaque strate $\mathrm{U}_{\mathrm{D}h}$ de taille $\mathrm{N}_{\mathrm{D}h}$ avec $h = 1, \ldots, \mathrm{H}$ (voir le Tableau 6.2). Notre variable d'intérêt Y prend la valeur Y_{ik} pour la maternité i et le jour k. On s'intéresse au total $t_{\mathrm{Y}} = \sum_{i \in \mathrm{U}_{\mathrm{M}}} \sum_{k \in \mathrm{U}_{\mathrm{D}}} \mathrm{Y}_{ik}$ estimé sans biais par

$$\hat{t}_{\mathrm{Y}} = \sum_{g=1}^{\mathrm{G}} \sum_{h=1}^{\mathrm{H}} \sum_{i \in \mathrm{S}_{\mathrm{M}g}} \sum_{k \in \mathrm{S}_{\mathrm{D}h}} \frac{\mathrm{N}_{\mathrm{D}h}}{n_{\mathrm{D}h}} \frac{\mathrm{N}_{\mathrm{M}g}}{n_{\mathrm{M}g}} \mathrm{Y}_{ik} = \sum_{g=1}^{\mathrm{G}} \sum_{i \in \mathrm{S}_{\mathrm{M}g}} \frac{\mathrm{N}_{\mathrm{M}g}}{n_{\mathrm{M}g}} \hat{\mathrm{Y}}_{i\bullet} = \sum_{h=1}^{\mathrm{H}} \sum_{k \in \mathrm{S}_{\mathrm{D}h}} \frac{\mathrm{N}_{\mathrm{D}h}}{n_{\mathrm{D}h}} \hat{\mathrm{Y}}_{\bullet k} \tag{6.4.1}$$

avec $\hat{\mathrm{Y}}_{i\bullet}$, l'estimateur d'Horvitz-Thompson du total sur la maternité i et $\hat{\mathrm{Y}}_{\bullet k}$, l'estimateur d'Horvitz-Thompson du total sur le jour k. Un estimateur sans biais de la variance de $\hat{t}_{\mathrm{Y}}$ est donné par :

$$\hat{\mathbf{V}}_{prod}\left(\hat{t}_{\mathrm{Y}}\right) = \hat{\mathbf{V}}_{\mathrm{D}} + \hat{\mathbf{V}}_{\mathrm{M}} - \hat{\mathbf{V}}_{\mathrm{E}} \tag{6.4.2}$$

avec

$$\hat{\mathbf{V}}_{\mathrm{D}}\left(\hat{t}_{\mathrm{Y}}\right) = \sum_{h=1}^{\mathrm{H}} \mathrm{N}_{\mathrm{D}h}^2 \left(\frac{1}{n_{\mathrm{D}h}} - \frac{1}{\mathrm{N}_{\mathrm{D}h}}\right) s^2_{\hat{\mathrm{Y}}_{\bullet\circ,h}}, \tag{6.4.3}$$

$$\hat{\mathbf{V}}_{\mathrm{M}}\left(\hat{t}_{\mathrm{Y}}\right) = \sum_{g=1}^{\mathrm{G}} \mathrm{N}_{\mathrm{M}g}^2 \left(\frac{1}{n_{\mathrm{M}g}} - \frac{1}{\mathrm{N}_{\mathrm{M}g}}\right) s^2_{\hat{\mathrm{Y}}_{\circ\bullet,g}}, \tag{6.4.4}$$

$$\hat{\mathbf{V}}_{\mathrm{E}}\left(\hat{t}_{\mathrm{Y}}\right) = \sum_{g=1}^{\mathrm{G}} \mathrm{N}_{\mathrm{M}g}^2 \left(\frac{1}{n_{\mathrm{M}g}} - \frac{1}{\mathrm{N}_{\mathrm{M}g}}\right) \sum_{h=1}^{\mathrm{H}} \mathrm{N}_{\mathrm{D}h}^2 \left(\frac{1}{n_{\mathrm{D}h}} - \frac{1}{\mathrm{N}_{\mathrm{D}h}}\right) \frac{1}{(n_{\mathrm{M}g}-1)(n_{\mathrm{D}h}-1)} s_{\mathrm{E},hg}, \tag{6.4.5}$$

où

$$s^2_{\hat{\mathrm{Y}}_{\bullet\circ,h}} = \frac{1}{n_{\mathrm{D}h}-1}\sum_{k\in\mathrm{S}_{\mathrm{D}h}}\left(\hat{\mathrm{Y}}_{\bullet k}-\frac{1}{n_{\mathrm{D}h}}\sum_{l\in\mathrm{S}_{\mathrm{D}h}}\hat{\mathrm{Y}}_{\bullet l}\right)^2, \quad (6.4.6)$$

$$s^2_{\hat{\mathrm{Y}}_{\circ\bullet,g}} = \frac{1}{n_{\mathrm{M}g}-1}\sum_{i\in\mathrm{S}_{\mathrm{M}g}}\left(\hat{\mathrm{Y}}_{i\bullet}-\frac{1}{n_{\mathrm{M}g}}\sum_{j\in\mathrm{S}_{\mathrm{M}g}}\hat{\mathrm{Y}}_{j\bullet}\right)^2, \quad (6.4.7)$$

$$s_{\mathrm{E},hg} = \sum_{i\in\mathrm{S}_{\mathrm{M}g}}\sum_{k\in\mathrm{S}_{\mathrm{D}h}}\left[\mathrm{Y}_{ik}-\frac{1}{n_{\mathrm{M}g}}\sum_{j\in\mathrm{S}_{\mathrm{M}g}}\mathrm{Y}_{jk}-\frac{1}{n_{\mathrm{D}h}}\sum_{l\in\mathrm{S}_{\mathrm{D}h}}\mathrm{Y}_{il}+\frac{1}{n_{\mathrm{M}g}}\frac{1}{n_{\mathrm{D}h}}\sum_{j\in\mathrm{S}_{\mathrm{M}g}}\sum_{l\in\mathrm{S}_{\mathrm{D}h}}\mathrm{Y}_{jl}\right]^2 (6.4.8)$$

L'estimateur de variance se décompose en trois termes : $\hat{\mathbf{V}}_{\mathrm{D}}\left(\hat{t}_{\mathrm{Y}}\right)$ qui représente un effet inter-jours, $\hat{\mathbf{V}}_{\mathrm{M}}\left(\hat{t}_{\mathrm{Y}}\right)$ qui représente un effet inter-maternités, $\hat{\mathbf{V}}_{\mathrm{E}}\left(\hat{t}_{\mathrm{Y}}\right)$ qui représente un effet résiduel.

Strate	Taille de la strate	Taille de l'échantillon	Critère de stratification
g	$\mathrm{N}_{\mathrm{M}g}$	$n_{\mathrm{M}g}$	Nombre d'accouchements en 2008
1	108	28	[145;699]
2	108	47	[700;1009]
3	109	66	[1010;1418]
4	108	97	[1422;2187]
5	111	111	[2197;5215]

TABLEAU 6.1 – Tailles des strates et des échantillons dans chaque strate pour le plan de sondage p_{M}

Strate	Taille de la strate	Taille de l'échantillon	Critère de stratification
h	$\mathrm{N}_{\mathrm{D}h}$	$n_{\mathrm{D}h}$	Saison
1	91	4	Printemps
2	91	6	Ete
3	91	7	Automne
4	91	8	Hiver (fin automne)

TABLEAU 6.2 – Tailles des strates et des échantillons dans chaque strate pour une modélisation du plan de sondage p_{D} sous la forme d'un plan STSI

Dans le Tableau 6.2 sont présentées les tailles de strates et des échantillons dans chaque strate pour une modélisation STSI du plan de sondage sur les jours.

6.5 Estimation de la variance issue du plan Elfe avec prise en compte de la non-réponse

Le traitement de la non-réponse au niveau du biais de l'estimateur est présenté ici succinctement (voir Juillard *et al.*, 2015b). L'estimateur de variance prenant en compte l'échantillonnage produit mais aussi la non-réponse est calculé.

6.5.1 Phase de non-réponse

Durant l'enquête Elfe, 29 maternités parmi les 349 sélectionnées n'ont pas participé à l'enquête. Cette première étape de non-réponse a été traitée par la méthode des Groupes de Réponses Homogènes (G.R.H.). Ensuite, parmi ces 320 maternités, certaines n'ont pas participé à toutes les vagues d'enquête : 15 maternités n'ont pas participé au trimestre 1, 8 au trimestre 2, 9 au trimestre 3 et 11 au trimestre 4. Cette non-réponse a été traitée dans chaque strate de maternités en ajustant les probabilités d'inclusion par un quotient représentant le nombre de maternités participant au trimestre sur le nombre de maternités attendues. Avec des taux de non-réponse relativement faibles pour les maternités (7 %) et pour les jours (3 % en moyenne), ces deux premières phases de non-réponse ne sont pas prises en compte dans le calcul de la variance de non-réponse mais traitées en ajustant simplement les probabilités d'inclusion.

Ensuite, il y a une phase de non-réponse au niveau nourrisson (voir la Figure 6.7) : 49 % des 36 000 familles approchées n'ont pas souhaité participer. La méthode des G.R.H. a de nouveau été utilisée pour traiter cette phase, puis, pour finir, un calage a été réalisé sur des variables socio-démographiques (âge de la mère, groupe de région

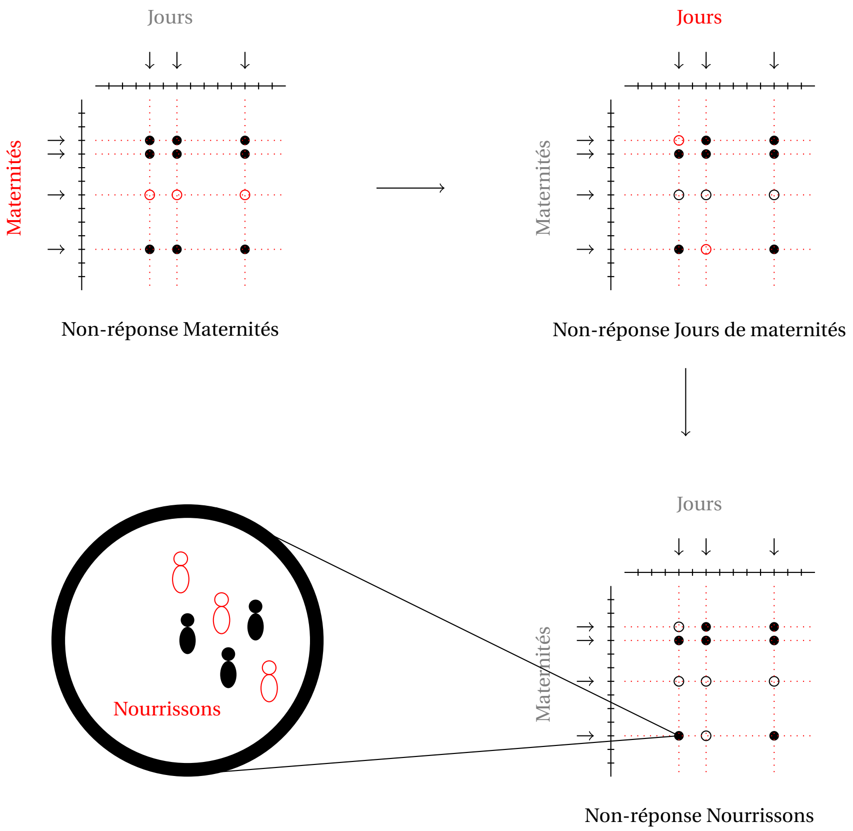


FIGURE 6.7 – Non-réponse au niveau maternité, puis jour de maternité, puis nourrisson

d'habitation, statut immigré de la mère, état matrimonial de la mère, primiparité et niveau d'étude de la mère). Cette dernière phase de non-réponse est considérée dans le calcul de l'estimateur de variance qui suit mais l'étape de calage ne l'est pas pour l'instant.

Notre variable d'intérêt prend la valeur y_a pour le nourrisson a de la maternité i du jour k. Le total t_Y peut alors s'écrire

$$t_Y = \sum_{i \in U_M} \sum_{k \in U_D} Y_{ik} \quad \text{avec} \quad Y_{ik} = \sum_{a \in U_{ik}} y_a, \tag{6.5.1}$$

où U_{ik} représente la sous-population des nourrissons de la maternité i le jour k. On

note $\mathrm{S}_{\mathrm{R}_{ik}}$ l'échantillon des répondants de la sous-population U_{ik}. La non-réponse est modélisée par une seconde phase de tirage au sein de l'échantillon complet des nourrissons. Pour cela, on fait l'hypothèse qu'il existe des groupes homogènes de réponse, avec comportements de réponse indépendants dans ces G.R.H.. En se basant sur la méthode des scores [Eltinge et Yansaneh, 1997] afin d'estimer les probabilités de réponse, F groupes de réponses homogènes sont créés. On notera $\hat{p}_f$ la probabilité de réponse estimée pour le G.R.H. f, et $\mathrm{S}_{\mathrm{R}_f}$ l'échantillon des n_{R_f} répondants du G.R.H. f. On a donc $\hat{p}_a = \hat{p}_f$ pour tout $a \in \mathrm{S}_{\mathrm{R}_f}$.
Dans ce cas, le total t_Y est estimé approximativement sans biais par

$$\begin{aligned} \hat{t}_{\mathrm{Y}\star} &= \sum_{g=1}^{\mathrm{G}} \sum_{h=1}^{\mathrm{H}} \sum_{i \in \mathrm{S}_{\mathrm{M}g}} \sum_{k \in \mathrm{S}_{\mathrm{D}h}} \frac{\mathrm{N}_{\mathrm{D}h}}{n_{\mathrm{D}h}} \frac{\mathrm{N}_{\mathrm{M}g}}{n_{\mathrm{M}g}} \hat{\mathrm{Y}}_{ik} \quad \text{avec} \quad \hat{\mathrm{Y}}_{ik} = \sum_{a \in \mathrm{S}_{\mathrm{R}_{ik}}} \frac{y_a}{\hat{p}_a}, \qquad (6.5.2) \\ &= \sum_{g=1}^{\mathrm{G}} \sum_{i \in \mathrm{S}_{\mathrm{M}g}} \frac{\mathrm{N}_{\mathrm{M}g}}{n_{\mathrm{M}g}} \hat{\hat{\mathrm{Y}}}_{i\bullet} \quad \text{avec} \quad \hat{\hat{\mathrm{Y}}}_{i\bullet} = \sum_{h=1}^{\mathrm{H}} \sum_{k \in \mathrm{S}_{\mathrm{D}h}} \frac{\mathrm{N}_{\mathrm{D}h}}{n_{\mathrm{D}h}} \hat{\mathrm{Y}}_{ik}, \\ &= \sum_{h=1}^{\mathrm{H}} \sum_{k \in \mathrm{S}_{\mathrm{D}h}} \frac{\mathrm{N}_{\mathrm{D}h}}{n_{\mathrm{D}h}} \hat{\hat{\mathrm{Y}}}_{\bullet k} \quad \text{avec} \quad \hat{\hat{\mathrm{Y}}}_{\bullet k} = \sum_{g=1}^{\mathrm{G}} \sum_{i \in \mathrm{S}_{\mathrm{M}g}} \frac{\mathrm{N}_{\mathrm{M}g}}{n_{\mathrm{M}g}} \hat{\mathrm{Y}}_{ik}. \end{aligned}$$

6.5.2 Estimation de la variance avec phase de non-réponse

Pour un plan produit et la phase de non-réponse présentée dans le paragraphe précédent, lorsqu'on utilise les estimations des probabilités de réponse issues de la méthode des scores, un estimateur approximativement sans biais de la variance peut être obtenu en adaptant le travail de Kim et Kim [2007]. Dans le cas particulier de l'enquête Elfe cela conduit à :

$$\hat{\mathbf{V}}\left(\hat{t}_{\mathrm{Y}\star}\right) = \hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{\mathrm{Y}\star}\right) + \hat{\mathbf{V}}_{\mathrm{NR}}\left(\hat{t}_{\mathrm{Y}\star}\right) \qquad (6.5.3)$$

où

$$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{\mathrm{Y}\star}\right) = \hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech1}}\left(\hat{t}_{\mathrm{Y}\star}\right) - \hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech2}}\left(\hat{t}_{\mathrm{Y}\star}\right) \tag{6.5.4}$$

$$\tag{6.5.5}$$

$$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech1}}\left(\hat{t}_{\mathrm{Y}\star}\right) = \hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{D}}\left(\hat{t}_{\mathrm{Y}\star}\right) + \hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{M}}\left(\hat{t}_{\mathrm{Y}\star}\right) - \hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{E}}\left(\hat{t}_{\mathrm{Y}\star}\right) \tag{6.5.6}$$

$$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech2}}\left(\hat{t}_{\mathrm{Y}\star}\right) = \sum_{g=1}^{\mathrm{G}}\left(\frac{\mathrm{N}_{\mathrm{M}g}}{n_{\mathrm{M}g}}\right)^2 \sum_{h=1}^{\mathrm{H}}\left(\frac{\mathrm{N}_{\mathrm{D}h}}{n_{\mathrm{D}h}}\right)^2 \sum_{f=1}^{\mathrm{F}}\sum_{a\in\mathrm{S}_{\mathrm{R}_f}}\left(1-\frac{n_{\mathrm{M}g}n_{\mathrm{D}h}}{\mathrm{N}_{\mathrm{M}g}\mathrm{N}_{\mathrm{D}h}}\right)\frac{1-\hat{p}_f}{\hat{p}_f^2}{y_a}^2 \tag{6.5.7}$$

$$\hat{\mathbf{V}}_{\mathrm{NR}}\left(\hat{t}_{\mathrm{Y}\star}\right) = \sum_{f=1}^{\mathrm{F}}\sum_{a\in\mathrm{S}_{\mathrm{R}_f}}\frac{1-\hat{p}_f}{\hat{p}_f^2}\left(\check{y}_a-\frac{1}{n_{\mathrm{R}_f}}\sum_{b\in\mathrm{S}_{\mathrm{R}_f}}\check{y}_b\right)^2 \text{ avec } \check{y}_a=\frac{y_a}{\pi^{\mathrm{M}}_i\pi^{\mathrm{D}}_k}, \tag{6.5.8}$$

avec

$$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{D}}\left(\hat{t}_{\mathrm{Y}\star}\right) = \sum_{h=1}^{\mathrm{H}}(\mathrm{N}_{\mathrm{D}h})^2\left(\frac{1}{n_{\mathrm{D}h}}-\frac{1}{\mathrm{N}_{\mathrm{D}h}}\right)s^2_{\hat{\hat{\mathrm{Y}}}_{\bullet\circ,h}} \tag{6.5.9}$$

$$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{M}}\left(\hat{t}_{\mathrm{Y}\star}\right) = \sum_{g=1}^{\mathrm{G}}(\mathrm{N}_{\mathrm{M}g})^2\left(\frac{1}{n_{\mathrm{M}g}}-\frac{1}{\mathrm{N}_{\mathrm{M}g}}\right)s^2_{\hat{\hat{\mathrm{Y}}}_{\circ\bullet,g}} \tag{6.5.10}$$

$$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{E}}\left(\hat{t}_{\mathrm{Y}\star}\right) = \sum_{g=1}^{\mathrm{G}}(\mathrm{N}_{\mathrm{M}g})^2\left(\frac{1}{n_{\mathrm{M}g}}-\frac{1}{\mathrm{N}_{\mathrm{M}g}}\right)\sum_{h=1}^{\mathrm{H}}(\mathrm{N}_{\mathrm{D}h})^2\left(\frac{1}{n_{\mathrm{D}h}}-\frac{1}{\mathrm{N}_{\mathrm{D}h}}\right)\frac{1}{(n_{\mathrm{M}g}-1)(n_{\mathrm{D}h}-1)}s_{\hat{\mathrm{E}},hg} \tag{6.5.11}$$

et

$$s^2_{\hat{\hat{\mathrm{Y}}}_{\bullet\circ,h}} = \frac{1}{n_{\mathrm{D}h}-1}\sum_{k\in\mathrm{S}_{\mathrm{D}h}}\left(\hat{\hat{\mathrm{Y}}}_{\bullet k}-\frac{1}{n_{\mathrm{D}h}}\sum_{l\in\mathrm{S}_{\mathrm{D}h}}\hat{\hat{\mathrm{Y}}}_{\bullet l}\right)^2, \tag{6.5.12}$$

$$s^2_{\hat{\hat{\mathrm{Y}}}_{\circ\bullet,g}} = \frac{1}{n_{\mathrm{M}g}-1}\sum_{i\in\mathrm{S}_{\mathrm{M}g}}\left(\hat{\hat{\mathrm{Y}}}_{i\bullet}-\frac{1}{n_{\mathrm{M}g}}\sum_{j\in\mathrm{S}_{\mathrm{M}g}}\hat{\hat{\mathrm{Y}}}_{j\bullet}\right)^2, \tag{6.5.13}$$

$$s_{\hat{\mathrm{E}},hg} = \sum_{i\in\mathrm{S}_{\mathrm{M}g}}\sum_{k\in\mathrm{S}_{\mathrm{D}h}}\left[\hat{\mathrm{Y}}_{ik}-\frac{1}{n_{\mathrm{M}g}}\sum_{j\in\mathrm{S}_{\mathrm{M}g}}\hat{\mathrm{Y}}_{jk}-\frac{1}{n_{\mathrm{D}h}}\sum_{l\in\mathrm{S}_{\mathrm{D}h}}\hat{\mathrm{Y}}_{il}+\frac{1}{n_{\mathrm{M}g}}\frac{1}{n_{\mathrm{D}h}}\sum_{j\in\mathrm{S}_{\mathrm{M}g}}\sum_{l\in\mathrm{S}_{\mathrm{D}h}}\hat{\mathrm{Y}}_{jl}\right]^2. \tag{6.5.14}$$

La partie $\hat{\mathbf{V}}_{\mathrm{NR}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ correspond à l’estimateur de la variance due à la non-réponse avec probabilités de réponse estimées et $\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ correspond à l’estimateur de la variance due à l'échantillonnage. On retrouve dans $\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech1}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ les trois termes qui composaient la variance présentée en formule (6.4.2) (à la différence que les sous-totaux Y_{ik} sont ici estimés, prenant en compte l'ajustement de la non-réponse), auxquels on soustrait le terme $\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech2}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ afin d'obtenir un estimateur sans biais de la variance d'échantillonnage.

6.6 A la recherche d'estimateurs simplifiés

Précédemment, un estimateur de la variance issu du plan de sondage Elfe a été présenté, avec prise en compte de la non-réponse. Dans cette section, plusieurs estimateurs simplifiés sont étudiés pour différentes raisons :

- l'estimateur sans biais n'est programmé dans aucun logiciel à notre connaissance ;
- l'estimateur sans biais peut théoriquement prendre des valeurs négatives, d'où la recherche d'estimateurs simplifiés, potentiellement biaisés mais positifs.

6.6.1 Estimateurs simplifiés

En prenant en compte les procédures logicielles existantes dans R, SAS et Stata, cinq estimateurs simplifiés ont été retenus :

- le premier estimateur correspond à une partie de l'estimateur sans biais, représentant la variance estimée inter-maternités en formule (6.5.10),

$$\hat{\mathbf{V}}_{\mathrm{SIMP1}} \equiv \hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{M}}\left(\hat{t}_{\mathrm{Y}\star}\right) = \sum_{g=1}^{\mathrm{G}} (\mathrm{N}_{\mathrm{M}g})^2 \left(\frac{1}{n_{\mathrm{M}g}} - \frac{1}{\mathrm{N}_{\mathrm{M}g}}\right) s^2_{\hat{\hat{\mathrm{Y}}}_{\circ\bullet,g}}, \qquad (6.6.1)$$

- le deuxième estimateur correspond à une partie de l'estimateur sans biais, représentant la variance estimée inter-jours en formule (6.5.9),

$$\hat{\mathbf{V}}_{\mathrm{SIMP2}} \equiv \hat{\mathbf{V}}_{\mathrm{D}}^{\mathrm{NR}}\left(\hat{t}_{\mathrm{Y}\star}\right) = \sum_{h=1}^{\mathrm{H}} (\mathrm{N}_{\mathrm{D}h})^2 \left(\frac{1}{n_{\mathrm{D}h}} - \frac{1}{\mathrm{N}_{\mathrm{D}h}}\right) s^2_{\hat{\hat{\mathrm{Y}}}_{\bullet\circ,h}}, \tag{6.6.2}$$

- le troisième estimateur correspond à la somme des deux précédents estimateurs simplifiés,

$$\begin{aligned} \hat{\mathbf{V}}_{\mathrm{SIMP3}} &\equiv \hat{\mathbf{V}}_{\mathrm{SIMP1}} + \hat{\mathbf{V}}_{\mathrm{SIMP2}} \\ &= \hat{\mathbf{V}}_{\mathrm{D}}^{\mathrm{NR}}\left(\hat{t}_{\mathrm{Y}\star}\right) + \hat{\mathbf{V}}_{\mathrm{M}}^{\mathrm{NR}}\left(\hat{t}_{\mathrm{Y}\star}\right), \end{aligned} \tag{6.6.3}$$

- le quatrième estimateur correspond à l'estimateur de variance adapté à un plan classique à deux degrés, dans lequel les maternités constituent les Unités Primaires (UP) et les jours les Unités Secondaires (US),

$$\begin{aligned} \hat{\mathbf{V}}_{\mathrm{SIMP4}} &\equiv \hat{\mathbf{V}}_{\mathrm{M}}^{\mathrm{NR}}\left(\hat{t}_{\mathrm{Y}\star}\right) + \sum_{g=1}^{\mathrm{G}} \frac{\mathrm{N}_{\mathrm{M}g}}{n_{\mathrm{M}g}} \sum_{i\in \mathrm{S}_{\mathrm{M}g}} \sum_{h=1}^{\mathrm{H}} \mathrm{N}^2_{\mathrm{D}h} \left(\frac{1}{n_{\mathrm{D}h}} - \frac{1}{\mathrm{N}_{\mathrm{D}h}}\right) s^2_{\hat{\mathrm{Y}}_{i\circ,h}} \\ \text{avec} \quad s^2_{\mathrm{Y}_{i\circ,h}} &= \frac{1}{n_{\mathrm{D}h}-1} \sum_{k\in \mathrm{S}_{\mathrm{D}h}} (\hat{\mathrm{Y}}_{ik} - \frac{1}{n_{\mathrm{D}h}} \sum_{l\in \mathrm{S}_{\mathrm{D}h}} \hat{\mathrm{Y}}_{il})^2, \end{aligned} \tag{6.6.4}$$

- le cinquième estimateur correspond à l'estimateur de variance adapté à un plan classique à deux degrés, dans lequel les jours constituent les UP et les maternités les US,

$$\begin{aligned} \hat{\mathbf{V}}_{\mathrm{SIMP5}} &\equiv \hat{\mathbf{V}}_{\mathrm{D}}^{\mathrm{NR}}\left(\hat{t}_{\mathrm{Y}\star}\right) + \sum_{h=1}^{\mathrm{H}} \frac{\mathrm{N}_{\mathrm{D}h}}{n_{\mathrm{D}h}} \sum_{k\in \mathrm{S}_{\mathrm{D}h}} \sum_{g=1}^{\mathrm{G}} \mathrm{N}^2_{\mathrm{M}g} \left(\frac{1}{n_{\mathrm{M}g}} - \frac{1}{\mathrm{N}_{\mathrm{M}g}}\right) s^2_{\hat{\mathrm{Y}}_{\circ k,g}} \\ \text{avec} \quad s^2_{\mathrm{Y}_{\circ k,g}} &= \frac{1}{n_{\mathrm{M}g}-1} \sum_{i\in \mathrm{S}_{\mathrm{M}g}} (\hat{\mathrm{Y}}_{ik} - \frac{1}{n_{\mathrm{M}g}} \sum_{j\in \mathrm{S}_{\mathrm{M}g}} \hat{\mathrm{Y}}_{jk})^2. \end{aligned} \tag{6.6.5}$$

Ces estimateurs simplifiés sont positifs et calculables à partir de procédures déjà programmées (voir Tableau 6.3) mais ne sont **pas sans biais**. Dans le paragraphe suivant, ces cinq estimateurs sont comparés.

Estimateur simplifié	Logiciels
$\hat{\mathbf{V}}_{\text{SIMP1}}$	R/SAS/Stata
$\hat{\mathbf{V}}_{\text{SIMP2}}$	R/SAS/Stata
$\hat{\mathbf{V}}_{\text{SIMP3}}$	R/SAS/Stata
$\hat{\mathbf{V}}_{\text{SIMP4}}$	R/Stata
$\hat{\mathbf{V}}_{\text{SIMP5}}$	R/Stata

TABLEAU 6.3 – Procédures logicielles R/SAS/Stata et estimateurs simplifiés

6.6.2 Comparaisons entre l'estimateur sans biais et les estimateurs simplifiés sur données Elfe

Dans cette partie, les résultats associés à l'estimateur $\hat{\mathbf{V}}$ (sans biais) ainsi qu'aux cinq estimateurs simplifiés présentés dans la section précédente sont illustrés sur données Elfe.

Dans le Tableau 6.4, pour chacune des variables Elfe choisie, on calcule le total $\hat{t}_{Y\star}$ donné en formule (6.5.2), sa variance estimée $\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)$ donnée en formule (6.5.3), ainsi que chaque partie qui la compose : $\hat{\mathbf{V}}^{\text{NR}}_{\text{ech1}}\left(\hat{t}_{Y\star}\right)$ en (6.5.6), $\hat{\mathbf{V}}^{\text{NR}}_{\text{ech2}}\left(\hat{t}_{Y\star}\right)$ en (6.5.7) et $\hat{\mathbf{V}}_{\text{NR}}\left(\hat{t}_{Y\star}\right)$ en (6.5.8). Il est à noter que les résultats affichés ne tiennent pas compte de l'étape de calage qui est prise en compte dans les poids livrés aux utilisateurs. L'utilisateur ne peut donc retomber (exactement) sur les mêmes résultats à partir des variables de sa base de données. Le calage sera pris en compte dans la prochaine section.

On calcule l'écart relatif entre $\hat{\mathbf{V}}_{\text{SIMP}}$ et l'estimateur sans biais $\hat{\mathbf{V}}$ défini par :

$$\text{ER} = \frac{\hat{\mathbf{V}}_{\text{SIMP}}\left(\hat{t}_{Y\star}\right) - \hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)}{\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)}.$$

On constate dans le Tableau 6.4 que la part de variance estimée due à la non-réponse $\hat{\mathbf{V}}_{\text{NR}}$ est faible comparée à celle d'échantillonnage $\hat{\mathbf{V}}^{\text{NR}}_{\text{ech}}\left(\hat{t}_{Y\star}\right)$. On rappelle qu'il s'agit d'une non-réponse non pas sur la première phase d'échantillonnage des unités groupées (i,k), mais à la seconde phase sur l'unité nourrisson.

Mis à part $\hat{\mathbf{V}}_{\text{SIMP3}}$, tous les estimateurs simplifiés présentent des valeurs inférieures à l'estimateur sans biais. Rappelons que l'estimateur $\hat{\mathbf{V}}$ a déjà lui-même subi des sim-

plifications (non prise en compte de la non-réponse au niveau maternité, ni celle au niveau jour) et présente des valeurs certainement plus petites qu'elles ne l'auraient été sans ces simplifications.
L'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ présente des ER relativement faibles et peu variables (entre 0 et 20 %, sauf pour la variable *Nombre de nourrissons ayant une mère âgée entre 18 et 25 ans* qui atteint 29 %). Tous les autres estimateurs présentent au moins un cas avec un ER supérieur à 45 % en valeur absolue. On observe que l'estimateur $\hat{\mathbf{V}}_{\text{SIMP5}}$ s'avère intéressant dans plusieurs cas (parmi les dix variables étudiées, huit présentent un ER inférieur à 20 % en valeur absolue) mais extrêmement mauvais pour des variables présentant une variabilité inter-maternités importante (-47 % pour la variable *Nombre de nourrissons ayant une mère suivie par sage-femme*). Les estimateurs $\hat{\mathbf{V}}_{\text{SIMP1}}$ et $\hat{\mathbf{V}}_{\text{SIMP4}}$ s'avèrent inacceptables avec jusqu'à -95 % d'erreur relative.

L'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ reste le seul estimateur simplifié acceptable quelle que soit la variable d'intérêt et est recommandé aux utilisateurs des données Elfe.

Lecture du Tableau 6.4 : on estime à 7.5×10^5 le nombre total de naissance dans la population (définie) avec une variance estimée à 2.9×10^8. L'estimateur de variance simplifié préconisé vaut 3.3×10^8, soit une sur-estimation de 14 % par rapport à l'estimateur sans biais.

Modélisation STSI × STSI, NR	Nombre de naissances	Nombre de naissances sous césarienne en début du travail	Nombre de nourrissons ayant une mère suivie par sage-femme	Nombre de nourrissons ayant une mère primipare	Nombre de nourrissons ayant une mère mariée ou remariée
$\hat{t}_{\mathrm{Y}\star}$	7.5×10^{5}	7.4×10^{4}	9.8×10^{4}	3.3×10^{5}	3.3×10^{5}
$\hat{\mathbf{V}}\left(\hat{t}_{\mathrm{Y}\star}\right) = \hat{\mathbf{v}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{\mathrm{Y}\star}\right) + \hat{\mathbf{v}}_{\mathrm{NR}}\left(\hat{t}_{\mathrm{Y}\star}\right)$	2.9×10^{8}	7.1×10^{7}	1.9×10^{7}	6.2×10^{7}	8.9×10^{7}
$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{\mathrm{Y}\star}\right)$	2.8×10^{8}	6.7×10^{7}	1.3×10^{7}	5.2×10^{7}	7.8×10^{7}
$\hat{\mathbf{V}}_{\mathrm{NR}}\left(\hat{t}_{\mathrm{Y}\star}\right)$	8.1×10^{6}	4.2×10^{6}	6.1×10^{6}	1.0×10^{7}	1.1×10^{7}
$\hat{\mathbf{V}}_{\mathrm{SIMP1}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ (ER)	8.9×10^{7} (-69 %)	3.7×10^{6} (-95 %)	1.4×10^{7} (-29 %)	2.3×10^{7} (-63 %)	2.4×10^{7} (-73 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP2}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ (ER)	2.4×10^{8} (-16 %)	7.0×10^{7} (-02 %)	8.7×10^{6} (-55 %)	5.0×10^{7} (-20 %)	7.7×10^{7} (-14 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP3}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ (ER)	3.3×10^{8} (14 %)	7.3×10^{7} (3.1 %)	2.2×10^{7} (15 %)	7.3×10^{7} (17 %)	1.0×10^{8} (13 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP4}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ (ER)	1.2×10^{8} (-58 %)	8.3×10^{6} (-88 %)	1.9×10^{7} (-01 %)	3.8×10^{7} (-40 %)	4.0×10^{7} (-55 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP5}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ (ER)	2.5×10^{8} (-13 %)	7.1×10^{7} (-0.3 %)	1.0×10^{7} (-47 %)	5.4×10^{7} (-13 %)	8.1×10^{7} (-10 %)
	Nombre de nourrissons ayant une mère âgée entre 18 à 25 ans	Nombre de nourrissons ayant une mère avec un IMC supérieur à 30	Nombre de nourrissons ayant une mère ayant suivi des séances préparation	Nombre de nourrissons ayant une mère étrangère ou apatride	Nombre de jumeaux
$\hat{t}_{\mathrm{Y}\star}$	1.2×10^{5}	8.1×10^{4}	3.6×10^{5}	9.2×10^{4}	2.4×10^{4}
$\hat{\mathbf{V}}\left(\hat{t}_{\mathrm{Y}\star}\right) \hat{\mathbf{v}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{\mathrm{Y}\star}\right) + \hat{\mathbf{v}}_{\mathrm{NR}}\left(\hat{t}_{\mathrm{Y}\star}\right)$	2.1×10^{7}	1.6×10^{7}	6.4×10^{7}	2.5×10^{7}	4.5×10^{6}
$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{\mathrm{Y}\star}\right)$	1.4×10^{7}	1.1×10^{7}	5.6×10^{7}	1.8×10^{7}	3.4×10^{6}
$\hat{\mathbf{V}}_{\mathrm{NR}}\left(\hat{t}_{\mathrm{Y}\star}\right)$	7.0×10^{6}	5.0×10^{6}	7.9×10^{6}	6.6×10^{6}	1.1×10^{6}
$\hat{\mathbf{V}}_{\mathrm{SIMP1}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ (ER)	1.0×10^{7} (-50 %)	5.2×10^{6} (-67 %)	2.8×10^{7} (-57 %)	7.0×10^{6} (-72 %)	1.4×10^{6} (-70 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP2}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ (ER)	1.6×10^{7} (-21 %)	1.3×10^{7} (-16 %)	4.5×10^{7} (-30 %)	2.3×10^{7} (-08 %)	4.0×10^{6} (-12 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP3}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ (ER)	2.7×10^{7} (29 %)	1.9×10^{7} (17 %)	7.3×10^{7} (14 %)	2.9×10^{7} (20 %)	5.3×10^{6} (18 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP4}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ (ER)	1.9×10^{7} (-09 %)	9.7×10^{6} (-38 %)	4.0×10^{7} (-37 %)	1.6×10^{7} (-36 %)	3.9×10^{6} (-14 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP5}}\left(\hat{t}_{\mathrm{Y}\star}\right)$ (ER)	1.9×10^{7} (-08 %)	1.5×10^{7} (-06 %)	4.9×10^{7} (-24 %)	2.4×10^{7} (-01 %)	4.4×10^{6} (-02 %)

TABLEAU 6.4 – Comparaison entre différents estimateurs simplifiés et l'estimateur sans biais.

6.7 Estimation de la variance avec prise en compte du calage

Le calage est une méthode permettant d'intégrer des informations connues sur l'ensemble de la population, après que l'enquête ait eu lieu. La méthode consiste à modifier les poids de sondage en utilisant certaines équations respectant certaines contraintes (méthode détaillée dans Deville et Särndal, 1992). Cette modification impacte les estimateurs (paramètre **et** variance), cela peut diminuer le biais et améliorer la précision si la variable d'intérêt est liée aux variables de calage.

6.7.1 Calage dans Elfe

Le choix du vecteur de variables de calage s'est porté sur *Classe d'âge, Groupe de régions, Etat matrimonial, Statut immigré, Niveau d'étude* et *Primiparité*, permettant un calage caractérisant la situation familiale, géographique et socio-demographique. Pour plus de détails concernant le découpage de ces variables catégorielles, voir Juillard *et al.*, 2015b. Notons que le calage ne prenant pas en compte les données manquantes, les variables de calage avaient été imputées (à petits taux) et que les poids calés ont subi une phase de troncature afin de limiter leur dispersion. Les sources de calage pour l'enquête Elfe sont l'état civil et l'enquête nationale périnatale (ENP) 2010.
On note w_a le poids calé associé à l'individu a de l'échantillon des répondants $\mathrm{S_R}$. Le total t_Y est estimé approximativement sans biais par

$$\hat{t}_{\mathrm{Y}_c} = \sum_{a \in \mathrm{S_R}} w_a y_a, \tag{6.7.1}$$

L'idée du calage est de réduire la variance associée aux estimateurs calés : plus la variable d'intérêt y sera corrélée aux variables de calage, meilleure sera la précision.

6.7.2 Estimation de la variance après calage

Pour calculer l'estimateur de variance, on s'intéresse aux résidus estimés e de la régression (pondérée) de notre variable d'intérêt y sur les variables de calage x :

$$e_a = y_a - \hat{b} x_a \tag{6.7.2}$$

où $\hat{b} = \left(\sum_{a \in \mathrm{S_R}} d_a x_a x_a^\mathrm{T}\right)^{-1} \sum_{a \in \mathrm{S_R}} d_a x_a y_a$ avec d_a, les poids de sondage corrigés de la non-réponse avant calage. L'estimateur de la variance après calage est obtenu en remplaçant chaque y_a par son résidu associé e_a dans la formule (6.5.3). On comprend alors pourquoi la variance estimée sera d'autant plus faible que les variables y et x sont liées.

	Nombre de naissances	Nombre de naissances sous césarienne en début de travail	Nombre de nourrissons ayant une mère suivie par sage-femme	Nombre de nourrissons ayant une mère primipare	Nombre de nourrissons ayant une mère mariée ou remariée
$\hat{t}_{Y\star}$	7.6×10^5	7.5×10^4	9.8×10^4	3.4×10^5	3.4×10^5
$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)$	$3,1 \times 10^8$	$7,4 \times 10^7$	$2,2 \times 10^7$	$6,8E \times 10^7$	$9,7 \times 10^7$
$\hat{t}_{Y_c}$	7.6×10^5	7.5×10^4	9.6×10^4	3.2×10^5	3.3×10^5
$\hat{\mathbf{V}}\left(\hat{t}_{Y_c}\right)$		$4,7 \times 10^7$	$1,2 \times 10^7$	$4,2 \times 10^5$	$6,6 \times 10^5$
$\hat{\mathbf{V}}_{\text{SIMP1}}\left(\hat{t}_{Y_c}\right)$ (ER)		$1,7 \times 10^6$ (-96 %)	$9,4 \times 10^6$ (-21 %)	$1,4 \times 10^5$ (-66 %)	$2,4 \times 10^5$ (-64 %)
$\hat{\mathbf{V}}_{\text{SIMP2}}\left(\hat{t}_{Y_c}\right)$ (ER)		$4,6 \times 10^7$ (-01 %)	$3,8 \times 10^6$ (-68 %)	$3,7 \times 10^5$ (-11 %)	$5,6 \times 10^5$ (-16 %)
$\hat{\mathbf{V}}_{\text{SIMP3}}\left(\hat{t}_{Y_c}\right)$ (ER)		$4,8 \times 10^7$ (02 %)	$1,3 \times 10^7$ (11 %)	$5,2 \times 10^5$ (22 %)	$8,0 \times 10^5$ (20 %)
$\hat{\mathbf{V}}_{\text{SIMP4}}\left(\hat{t}_{Y_c}\right)$ (ER)		$5,7 \times 10^6$ (-88 %)	$1,4 \times 10^7$ (17 %)	$6,2 \times 10^5$ (47 %)	$7,1 \times 10^5$ (07 %)
$\hat{\mathbf{V}}_{\text{SIMP5}}\left(\hat{t}_{Y_c}\right)$ (ER)		$4,7 \times 10^7$ (01 %)	$4,9 \times 10^6$ (-59 %)	$4,3 \times 10^5$ (03 %)	$6,9 \times 10^5$ (04 %)

	Nombre de nourrissons ayant une mère âgée entre 18 et 25 ans	Nombre de nourrissons ayant une mère avec un IMC supérieur à 30	Nombre de nourrissons ayant mère ayant suivi des séances de préparation	Nombre de nourrissons ayant une mère étrangère ou apatride	Nombre de jumeaux
$\hat{t}_{Y\star}$	1.2×10^5	8.2×10^4	3.7×10^5	9.6×10^4	2.5×10^4
$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)$	$2,1 \times 10^7$	$1,8 \times 10^7$	$6,8 \times 10^7$	$2,7 \times 10^7$	$4,6 \times 10^6$
$\hat{t}_{Y_c}$	1.1×10^5	8.2×10^4	3.7×10^5	9.8×10^4	2.5×10^4
$\hat{\mathbf{V}}\left(\hat{t}_{Y_c}\right)$	$2,4 \times 10^4$	$5,7 \times 10^6$	$1,4 \times 10^7$	$3,7 \times 10^6$	$3,3 \times 10^6$
$\hat{\mathbf{V}}_{\text{SIMP1}}\left(\hat{t}_{Y_c}\right)$ (ER)	$9,3 \times 10^3$ (-62 %)	$2,0 \times 10^6$ (-65 %)	$5,6 \times 10^6$ (-59 %)	$7,5 \times 10^5$ (-80 %)	$1,3 \times 10^6$ (-61 %)
$\hat{\mathbf{V}}_{\text{SIMP2}}\left(\hat{t}_{Y_c}\right)$ (ER)	$2,3 \times 10^4$ (-05 %)	$5,0 \times 10^6$ (-12 %)	$1,1 \times 10^7$ (-24 %)	$3,5 \times 10^6$ (-05 %)	$2,8 \times 10^6$ (-16 %)
$\hat{\mathbf{V}}_{\text{SIMP3}}\left(\hat{t}_{Y_c}\right)$ (ER)	$3,3 \times 10^4$ (33 %)	$7,0 \times 10^6$ (23 %)	$1,6 \times 10^7$ (17 %)	$4,3 \times 10^6$ (15 %)	$4,1 \times 10^6$ (23 %)
$\hat{\mathbf{V}}_{\text{SIMP4}}\left(\hat{t}_{Y_c}\right)$ (ER)	$3,5 \times 10^4$ (43 %)	$5,9 \times 10^6$ (02 %)	$1,2 \times 10^7$ (-13 %)	$3,0 \times 10^6$ (-18 %)	$3,7 \times 10^6$ (12 %)
$\hat{\mathbf{V}}_{\text{SIMP5}}\left(\hat{t}_{Y_c}\right)$ (ER)	$2,7 \times 10^4$ (10 %)	$6,0 \times 10^6$ (06 %)	$1,2 \times 10^7$ (-12 %)	$3,9 \times 10^6$ (05 %)	$3,3 \times 10^6$ (-02 %)

TABLEAU 6.5 – Comparaison entre différents estimateurs simplifiés et l'estimateur sans biais avant et après calage.

Dans le Tableau 6.5, afin de comparer les résultats avant et après calage, l'estimateur avant calage a été ajusté sur le nombre total de naissance (764000). Le calage permet bien de diminuer l'estimation de la variance et ceci d'autant plus que les variables d'intérêt sont corrélées aux variables de calage : ceci est flagrant pour les variables *Nombre de nourrissons ayant une mère primipare*, *Nombre de nourrissons ayant une mère mariée ou remariée* et *Nombre de nourrissons ayant une mère âgée entre 18 et 25 ans* qui correspondent respectivement à des modalités des variables de calage *Primiparité*, *Etat matrimonial* et *Classe d'âge*. Notons cependant qu'il ne s'agit pas exactement des variables de calage, qui avaient été imputées, mais des variables brutes. Les véritables variables de calage donnent une estimation de la variance approximativement nulle.

Lecture du Tableau 6.5 : sans calage, on estime à 9.8×10^4 le nombre total de nourrissons ayant une mère suivie par une sage-femme dans la population (définie) avec une variance estimée à 2.2×10^7. Après calage, on estime cette variance à 1.2×10^7 et l'estimateur de variance simplifié préconisé vaut 1.3×10^7, soit une sur-estimation de 11 % par rapport à l'estimateur sans biais calé.

6.8 Estimation de la variance pour une statistique plus complexe

Les parties précédentes concernent l'estimation de la variance d'un total estimé. Pour d'autres paramètres, tels qu'un ratio ou un coefficient de corrélation, la méthode de linéarisation [Deville, 1999] peut être utilisée afin de pouvoir estimer leurs variances.

Prenons l'exemple d'un ratio $R = t_Y / t_X$, il peut simplement s'estimer par $\hat{R}_c = \hat{t}_{Y_c} / \hat{t}_{X_c}$. Pour l'estimation de sa variance, il est nécessaire d'estimer la linéarisée l_a de notre paramètre ratio (les calculs pour plusieurs paramètres sont expliqués dans Dell *et al.*, 2002) :

$$\hat{l}_a = \frac{1}{\hat{t}_{X_c}} \left(y_a - \hat{R}_c x_a \right). \tag{6.8.1}$$

Ensuite, pour prendre en compte l'étape de calage, il faut (tout comme dans la section précédente) estimer les résidus e_a de la régression des $\hat{l}_a$ sur les variables de calage et de nouveau remplacer les y_a de la formule (6.5.3) par ces résidus.

Le Tableau 6.6 permet de comparer les résultats avant et après calage pour différents ratios estimés. Tout comme observé pour le paramètre total dans la section précédente, on remarque que la variance estimée du ratio diminue en prenant en compte le calage, et ceci d'autant plus que la variable d'intérêt est liée aux variables de calage. Concernant les différents estimateurs simplifiés, les mêmes commentaires faits

	% de naissances	% de naissances sous césarienne en début de travail	% de nourrissons ayant une mère suivie par sage-femme	% de nourrissons ayant une mère primipare	% de nourrissons ayant une mère mariée ou remariée
$\hat{\mathrm{R}}_\star$	1,000	0,098	0,128	0,440	0,444
$\hat{\mathbf{V}}(\hat{\mathrm{R}}_\star)$		$8,8 \times 10^{-5}$	$2,9 \times 10^{-5}$	$3,3 \times 10^{-5}$	$4,4 \times 10^{-5}$
$\hat{\mathrm{R}}_c$		0,098	0,126	0,427	0,439
$\hat{\mathbf{V}}(\hat{\mathrm{R}}_c)$		$8,0 \times 10^{-5}$	$2,1 \times 10^{-5}$	$7,2 \times 10^{-7}$	$1,1 \times 10^{-6}$
$\hat{\mathbf{V}}_{\mathrm{SIMP1}}(\hat{\mathrm{R}}_c)$ (ER)		$3,0 \times 10^{-6}$ (-96 %)	$1, \times 10^{-5}$ (-21 %)	$2,4 \times 10^{-7}$ (-66 %)	$4,1 \times 10^{-7}$ (-64 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP2}}(\hat{\mathrm{R}}_c)$ (ER)		$7,9 \times 10^{-5}$ (-01 %)	$6,6 \times 10^{-6}$ (-68 %)	$6,4 \times 10^{-7}$ (-11 %)	$9,5 \times 10^{-7}$ (-16 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP3}}(\hat{\mathrm{R}}_c)$ (ER)		$8,2 \times 10^{-5}$ (02 %)	$2,3 \times 10^{-5}$ (11 %)	$8,8 \times 10^{-7}$ (22 %)	$1,4 \times 10^{-6}$ (20 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP4}}(\hat{\mathrm{R}}_c)$ (ER)		$9,7 \times 10^{-6}$ (-88 %)	$2,4 \times 10^{-5}$ (17 %)	$1,1 \times 10^{-6}$ (47 %)	$1,2 \times 10^{-6}$ (07 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP5}}(\hat{\mathrm{R}}_c)$ (ER)		$8,0 \times 10^{-5}$ (01 %)	$8,5 \times 10^{-6}$ (-59 %)	$7,4 \times 10^{-7}$ (03 %)	$1,2 \times 10^{-6}$ (04 %)

	% de nourrissons ayant une mère âgée entre 18 à 25 ans	% de nourrissons ayant une mère avec un IMC supérieur à 30	% de nourrissons ayant mère ayant suivi des séances de préparation	% de nourrissons ayant une mère étrangère ou apatride	% de jumeaux
$\hat{\mathrm{R}}_\star$	0,153	0,107	0,481	0,125	0,032
$\hat{\mathbf{V}}(\hat{\mathrm{R}}_\star)$	$2,1 \times 10^{-5}$	$1,7 \times 10^{-5}$	$4,8 \times 10^{-5}$	$3,5 \times 10^{-5}$	$6,3 \times 10^{-6}$
$\hat{\mathrm{R}}_c$	0,139	0,108	0,480	0,128	0,033
$\hat{\mathbf{V}}(\hat{\mathrm{R}}_c)$	$4,2 \times 10^{-8}$	$9,8 \times 10^{-6}$	$2,4 \times 10^{-5}$	$6,4 \times 10^{-6}$	$5,7 \times 10^{-6}$
$\hat{\mathbf{V}}_{\mathrm{SIMP1}}(\hat{\mathrm{R}}_c)$ (ER)	$1,6 \times 10^{-8}$ (-62 %)	$3,4 \times 10^{-6}$ (-65 %)	$9,6 \times 10^{-6}$ (-59 %)	$1,3 \times 10^{-6}$ (-80 %)	$2,2 \times 10^{-6}$ (-61 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP2}}(\hat{\mathrm{R}}_c)$ (ER)	$4,0 \times 10^{-8}$ (-05 %)	$8,6 \times 10^{-6}$ (-12 %)	$1,8 \times 10^{-5}$ (-24 %)	$6,1 \times 10^{-6}$ (-05 %)	$4,8 \times 10^{-6}$ (-16 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP3}}(\hat{\mathrm{R}}_c)$ (ER)	$5,6 \times 10^{-8}$ (33 %)	$1,2 \times 10^{-5}$ (23 %)	$2,8 \times 10^{-5}$ (17 %)	$7,4 \times 10^{-6}$ (15 %)	$7,0 \times 10^{-6}$ (23 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP4}}(\hat{\mathrm{R}}_c)$ (ER)	$6,0 \times 10^{-8}$ (43 %)	$1,0 \times 10^{-5}$ (02 %)	$2,1 \times 10^{-5}$ (-13 %)	$5,2 \times 10^{-6}$ (-18 %)	$6,4 \times 10^{-6}$ (12 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP5}}(\hat{\mathrm{R}}_c)$ (ER)	$4,6 \times 10^{-8}$ (10 %)	$1,0 \times 10^{-5}$ (06 %)	$2,1 \times 10^{-5}$ (-12 %)	$6,7 \times 10^{-6}$ (05 %)	$5,6 \times 10^{-6}$ (-02 %)

TABLEAU 6.6 – Comparaison entre différents estimateurs simplifiés de variance d'un ratio estimé et l'estimateur sans biais avant et après calage.

pour le cas d'un total en sous-section 6.6.2 peuvent s'appliquer au ratio. Seul $\hat{\mathbf{V}}_{\mathrm{SIMP3}}$ s'avère intéressant avec des ER relativement faibles et peu variables.

L'estimateur $\hat{\mathbf{V}}_{\mathrm{SIMP3}}$ reste le seul estimateur simplifié acceptable quelle que soit la variable d'intérêt et est recommandé aux utilisateurs des données Elfe.

Lecture du Tableau 6.6 : sans calage, on estime à 0.44 le pourcentage de nourrissons ayant une mère suivie par une sage-femme dans la population (définie) avec une variance estimée à 3.3×10^{-5}. Après calage, on estime cette variance à 7.2×10^{-7} et l'estimateur de variance simplifié préconisé vaut 8.8×10^{-7}, soit une sur-estimation de 22 % par rapport à l'estimateur sans biais calé.

6.9 Procédures logicielles (SAS/R/Stata)

Les codes proposés concernent trois logiciels couramment utilisés : R 3.2.2 [R Core Team, 2015], SAS 9.4 [SAS Institute Inc., 2013], Stata 13.1 [StataCorp., 2013]. Le logiciel R est disponible à partir du CRAN (Comprehensive R Archive Network, `http://CRAN.R-project.org/`).

L'estimation se déroule en trois étapes. En premier lieu, il est nécessaire d'estimer la linéarisée de votre paramètre. Si votre paramètre est un total, vous pouvez passer directement à l'étape 2. Sinon, il vous faut coder la formule de la linéarisée, vous trouverez par exemple dans Dell *et al.* [2002] la formule pour l'indice de Gini et dans la suite du document sera traité le cas du ratio (moyenne, proportion...). Ensuite, il vous faudra effectuer une régression de cette linéarisée sur les variables de calage et récupérer les résidus : ceci permet de prendre en compte l'étape de calage censée diminuer la variance si votre variable est corrélée aux variables de calage. En dernier lieu, insérer ces résidus dans les procédures logicielles proposées pour estimer la variance (il s'agit de l'estimateur simplifié $\hat{\mathbf{V}}_{\text{SIMP3}}$ décrit précédemment).

Notons que pour des calculs corrects, il faut prendre en compte tous les croisements jour × maternité existants, c'est-à-dire 25 jours × 320 maternités = 8000 croisements. Puisque certains croisements n'apparaissent pas dans la base de données (puisqu'il n'y a pas eu de naissances Elfe ce jour-là dans cette maternité-là), il faut ajouter une ligne pour chacun de ces croisements avec un 0 pour la variable d'intérêt. En effet, ces 0 modifient les estimations de variance.

Pour chacun des codes proposés, en Tableau 6.7 sont listés les noms des variables de la base de données Elfe :

VARIABLES A NE PAS OUBLIER LORS DE LA DEMANDE D’ACCES	
Pondération en maternité	M00E_PONDVALC2 ou M00F_PONDVALC2
Strates pour le plan sur les jours	M00M1_VAGUE
Identifiant du jour	M00M2_JNAISSEALEA
Strates pour le plan sur les maternités	M00M1_MATSTRATEC1
Identifiant de la maternité	M00M1_IDGROUPNAMEALEAC1 ou M00M1_IDGROUPNAMEALEAC1B
Variables de calage	CAL1 CAL2 CAL3 CAL4 CAL5 CAL6

TABLEAU 6.7 – Liste des variables du plan de sondage Elfe nécessaires pour estimer la variance

6.9.1 Logiciel R

1$^{\text{ère}}$ étape : linéarisation du paramètre

Pour le cas d’un ratio t_{num}/t_{den} (moyenne, proportion...), le code permet de récupérer la linéarisée notée *lin* qui sera utilisée à la deuxième étape :

```
########################################
#DESCRIPION: estimates the linearized variable of a ratio Y2/Y1
#USAGE: LINratio(w,Y1,Y2)
#ARGUMENTS:
#              Y2      vector of the numerator variable
#              Y1      vector of the denominator variable
#              w       vector of the weights
#VALUE: the function returns a vector
LINratio <- function(w,Y1,Y2)
        {rpi=EstTOTAL(w,Y2)/EstTOTAL(w,Y1)
        txpi=EstTOTAL(w,Y1)
        upi = (Y2 - Y1*rpi) / txpi ; return(upi)
        }

#########################################
#DESCRIPION: estimates a population total
#USAGE: EstTOTAL(w,Y1)
#ARGUMENTS:
#              Y1      vector of the interest variable
#              w       vector of the weights
#VALUE: the function returns a numeric
EstTOTAL <- function(w,Y1)
               {t=sum(Y1*w) ; return(t)}
```

```
lin=LINratio(w=M00E_PONDVALC2,Y1=denominator,Y2=numerator)
```

2$^{\text{ème}}$ étape : régression sur les variables de calage

Pour cette étape, il suffit de prendre la linéarisée *lin* issue de la première étape et de l'injecter dans la régression (pondérée) sur les variables de calage. On récupère les résidus *res* de cette régression pour la prochaine étape.

```
########################################
#DESCRIPION: estimates a population total
#USAGE: REScalib(w,Y,X1,X2,X3,X4,X5,X6)
#ARGUMENTS:
#               Y1      vectors of the auxiliary variables
#               w       vector of the weights before calibration
#VALUE: the function returns a vector
REScalib <- function(w,Y,X1,X2,X3,X4,X5,X6)
        {modele=lm(Y ~ X1+X2+X3+X4+X5+X6,weights=w,na.action=na.exclude)
        e=residuals(modele)
        return(e)
        }

#pour le ratio
res=REScalib(w=M00E_PONDVALC2,Y=lin,X1=CAL1,X2=CAL2,X3=CAL3,X4=CAL4,X5=CAL5,X6=CAL6)

#Remplacer lin par votre variable d'interet si vous voulez estimer un total.
```

3$^{\text{ème}}$ étape : estimation de la variance

Dans cette section sont décrites les procédures logicielles à sommer pour pouvoir calculer $\hat{\mathbf{V}}_{\text{SIMP3}}$. L'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ en formule (6.6.3), recommandé aux utilisateurs, est la somme de deux termes calculables séparément. Les résidus *res* calculés dans l'étape précédente sont insérés dans la formule de l'estimateur de variance d'un total. Attention, si la somme des deux termes correspond bien à l'estimation de la variance, le total estimé, lui, correspond au total estimé des résidus. Il faut donc calculer séparément votre paramètre (total, ratio...). Pour le logiciel R, différents packages sont possibles, voici les procédures utilisant le package *survey* [Lumley, 2014] :

```
library(survey)

fpcm=c()
fpcm[M00M1_MATSTRATEC1==1] <- 108 ; fpcm[M00M1_MATSTRATEC1==2] <- 108
fpcm[M00M1_MATSTRATEC1==3] <- 109 ; fpcm[M00M1_MATSTRATEC1==4] <- 108
fpcm[M00M1_MATSTRATEC1==5] <- 111

infoplan<-svydesign(id=~M00M1_IDGROUPNAMEALEAC1, strata =~M00M1_MATSTRATEC1, fpc=~fpcm, weights=
    M00E_PONDVALC2)
infoplan

(Result <- svytotal(~res , infoplan, na.rm=TRUE))
vcov(Result)
SE(Result)^2

#Si vous utilisez la fonction confint(), l'intervalle de confiance n'est pas correct.
#Cette premiere fonction calcule le premier terme de l'estimateur de variance.

M00M2_JNAISSEALEA[M00M1_VAGUE==2 & M00M2_JNAISSEALEA==3] <- 7 ; M00M2_JNAISSEALEA[M00M1_VAGUE==2 &
    M00M2_JNAISSEALEA==4] <- 8 ; M00M2_JNAISSEALEA[M00M1_VAGUE==2 & M00M2_JNAISSEALEA==5] <- 9 ;
    M00M2_JNAISSEALEA[M00M1_VAGUE==2 & M00M2_JNAISSEALEA==6] <- 10 ; M00M2_JNAISSEALEA[M00M1_VAGUE
    ==2 & M00M2_JNAISSEALEA==1] <- 5 ; M00M2_JNAISSEALEA[M00M1_VAGUE==2 & M00M2_JNAISSEALEA==2] <-
     6 ;
M00M2_JNAISSEALEA[M00M1_VAGUE==3 & M00M2_JNAISSEALEA==1] <- 11 ; M00M2_JNAISSEALEA[M00M1_VAGUE==3 &
     M00M2_JNAISSEALEA==2] <- 12 ; M00M2_JNAISSEALEA[M00M1_VAGUE==3 & M00M2_JNAISSEALEA==3] <- 13
    ; M00M2_JNAISSEALEA[M00M1_VAGUE==3 & M00M2_JNAISSEALEA==4] <- 14 ; M00M2_JNAISSEALEA[M00M1_
    VAGUE==3 & M00M2_JNAISSEALEA==5] <- 15 ; M00M2_JNAISSEALEA[M00M1_VAGUE==3 & M00M2_JNAISSEALEA
    ==6] <- 16 ; M00M2_JNAISSEALEA[M00M1_VAGUE==3 & M00M2_JNAISSEALEA==7] <- 17 ;
M00M2_JNAISSEALEA[M00M1_VAGUE==4 & M00M2_JNAISSEALEA==1] <- 18 ; M00M2_JNAISSEALEA[M00M1_VAGUE==4 &
     M00M2_JNAISSEALEA==2] <- 19 ; M00M2_JNAISSEALEA[M00M1_VAGUE==4 & M00M2_JNAISSEALEA==3] <- 20
    ; M00M2_JNAISSEALEA[M00M1_VAGUE==4 & M00M2_JNAISSEALEA==4] <- 21 ; M00M2_JNAISSEALEA[M00M1_
    VAGUE==4 & M00M2_JNAISSEALEA==5] <- 22 ; M00M2_JNAISSEALEA[M00M1_VAGUE==4 & M00M2_JNAISSEALEA
    ==6] <- 23 ; M00M2_JNAISSEALEA[M00M1_VAGUE==4 & M00M2_JNAISSEALEA==7] <- 24 ; M00M2_
    JNAISSEALEA[M00M1_VAGUE==4 & M00M2_JNAISSEALEA==8] <- 25 ;

fpcd=c()
fpcd[M00M1_VAGUE==1] <- 91 ; fpcd[M00M1_VAGUE==2] <- 91
fpcd[M00M1_VAGUE==3] <- 91 ; fpcd[M00M1_VAGUE==4] <- 91

infoplan2<-svydesign(id=~M00M2_JNAISSEALEA, strata =~M00M1_VAGUE, fpc=~fpcd, weights=M00E_PONDVALC2
    )
infoplan2
```

```
(Result <- svytotal(~res , infoplan2, na.rm=TRUE))
vcov(Result)
SE(Result)^2

#Si vous utilisez la fonction confint(), l'intervalle de confiance n'est pas correct.
#Cette seconde fonction calcule le second terme de l'estimateur de variance.

#!!Les deux termes doivent etre sommes afin d'obtenir l'estimateur recommande V_SIMP3.
```

6.9.2 Logiciel SAS

1$^{\text{ère}}$ étape : linéarisation du paramètre

Pour le cas d'un ratio t_{num}/t_{den} (moyenne, proportion...), le code permet de récupérer la linéarisée notée *lin* qui sera utilisée à la deuxième étape :

```
PROC MEANS DATA=table sum ;
VAR numerator denominator ;
WEIGHT M00E_PONDVALC2;
OUTPUT OUT = stat sum= totnum totden ;  /* recupere les totaux estimes du numerateur et du
     denominateur */
RUN ;

DATA _null_; SET stat;
call symput('tnum', totnum);
call symput('tden', totden);
RUN;

DATA table ; SET table ;
lin = (numerator - (&tnum/&tden)*denominator) * (1/&tden) ;    /* formule de la linearisee */
RUN;
```

2$^{\text{ème}}$ étape : régression sur les variables de calage

Pour cette étape, il suffit de prendre la linéarisée *lin* issue de la première étape et de l'injecter dans la régression (pondérée) sur les variables de calage. On récupère les résidus *res* de cette régression pour la prochaine étape.

```
PROC GLM data=table noprint ;
CLASS CAL1 CAL2 CAL3 CAL4 CAL5 CAL6 ;
```

```
MODEL lin = CAL1 CAL2 CAL3 CAL4 CAL5 CAL6 ;
WEIGHT M00E_PONDVALC2 ;
OUTPUT OUT = table r=res ;
quit ;

/* Remplacer lin par votre variable d'interet si vous souhaitez estimer un total */
```

$3^{\text{ème}}$ étape : estimation de la variance

Dans cette section sont décrites les procédures logicielles à sommer pour pouvoir calculer $\hat{\mathbf{V}}_{\text{SIMP3}}$. L'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ en formule (6.6.3), recommandé aux utilisateurs, est la somme de deux termes calculables séparément. Les résidus *res* calculés dans l'étape précédente sont insérés dans la formule de l'estimateur de variance d'un total. Attention, si la somme des deux termes correspond bien à l'estimation de la variance, le total estimé, lui, correspond au total estimé des résidus. Il faut donc calculer séparément votre paramètre (total, ratio...). Les deux morceaux de l'estimateur peuvent se calculer à partir de la procédure *surveymeans* du logiciel SAS :

```
DATA NBDegreM ;                           /* Tailles des strates de maternites */
input M00M1_MATSTRATEC1 _TOTAL_;
datalines ;
1 108
2 108
3 109
4 108
5 111
;

proc SURVEYMEANS data=table TOTAL=NBDegreM sum mean var varsum missing;
CLUSTER M00M1_IDGROUPNAMEALEAC1;
STRATA M00M1_MATSTRATEC1 ;
VAR res ;
WEIGHT M00E_PONDVALC2 ;
run;

/* L'intervalle de confiance que peut produire cette procedure n'est pas correct.
/* Cette 1ere procedure calcule le 1er terme de l'estimateur de variance dans la colonne "Var of
    Sum". */

DATA NBDegreD ;                           /* Tailles des strates de jours */
```

```
input M00M1_VAGUE _TOTAL_;
datalines ;
1 91
2 91
3 91
4 91
;

proc SURVEYMEANS data=table TOTAL=NBDegreD sum mean var varsum missing;
CLUSTER M00M2_JNAISSEALEA;
STRATA M00M1_VAGUE ;
VAR  res ;
WEIGHT M00E_PONDVALC2 ;
run;

/* L'intervalle de confiance que peut produire cette procedure n'est pas correct.
/* Cette 2nde procedure calcule le 2nd terme de l'estimateur de variance dans la colonne "Var of
    Sum". */


/* !! Les deux termes doivent etre sommes afin d'obtenir l'estimateur recommande V_SIMP3. */
```

6.9.3 Logiciel Stata

1$^{\text{ère}}$ étape : linéarisation du paramètre

Pour le cas d'un ratio t_{num}/t_{den} (moyenne, proportion...), le code permet de récupérer la linéarisée notée *lin* qui sera utilisée à la deuxième étape :

```
. clear
* total estime du numerateur
. egen tnum= sum(numerator*M00E_PONDVALC2)
* total estime du denominateur
. egen tden= sum(denominator*M00E_PONDVALC2)
* formule de la linearisee lin
. gen lin = (numerator - (tnum/tden)*denominator) / tden
```

2ème étape : régression sur les variables de calage

Pour cette étape, il suffit de prendre la linéarisée *lin* issue de la première étape et de l'injecter dans la régression (pondérée) sur les variables de calage. On récupère les résidus *res* de cette régression pour la prochaine étape.

```
. reg lin CAL1 CAL2 CAL3 CAL4 CAL5 CAL6 [weight=M00E_PONDVALC2]
* Recuperer les residus res de la regression
. predict res, residuals

* Remplacer lin par votre variable d'interet si vous voulez estimer un total.
```

3ème étape : estimation de la variance

Dans cette section sont décrites les procédures logicielles à sommer pour pouvoir calculer $\hat{\mathbf{V}}_{\text{SIMP3}}$. L'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ en formule (6.6.3), recommandé aux utilisateurs, est la somme de deux termes calculables séparément. Les résidus *res* calculés dans l'étape précédente sont insérés dans la formule de l'estimateur de variance d'un total. Attention, si la somme des deux termes correspond bien à l'estimation de la variance, le total estimé, lui, correspond au total estimé des résidus. Il faut donc calculer séparément votre paramètre (total, ratio...).

```
* Tailles des strates de maternites
. gen fpcm=108 if M00M1_MATSTRATEC1==1
. replace fpcm=108 if M00M1_MATSTRATEC1==2
. replace fpcm=109 if M00M1_MATSTRATEC1==3
. replace fpcm=108 if M00M1_MATSTRATEC1==4
. replace fpcm=111 if M00M1_MATSTRATEC1==5

. svyset M00M1_IDGROUPNAMEALEAC1 [pweight=M00E_PONDVALC2], strata(M00M1_MATSTRATEC1) fpc(fpcm)

. svy: total res

* L'intervalle de confiance n'est pas correct.
* Il faut prendre la valeur du "Std. Err." et la mettre au carre.

* Tailles des strates de jours
. gen fpcd=91 if M00M1_VAGUE==1
. replace fpcd=91 if M00M1_VAGUE==2
. replace fpcd=91 if M00M1_VAGUE==3
. replace fpcd=91 if M00M1_VAGUE==4

. svyset M00M2_JNAISSEALEA [pweight=M00E_PONDVALC2], strata(M00M1_VAGUE) fpc(fpcd)

. svy: total res

* L'intervalle de confiance n'est pas correct.
* Il faut prendre la valeur du "Std. Err." et la mettre au carre.

* !! Les deux termes doivent etre sommes afin d'obtenir l'estimateur recommande V_SIMP3.
```

A retenir :

La **population d'inférence** est celle des nourrissons nés durant l'année 2011 en maternité de France métropolitaine, issus d'un accouchement au plus gémellaire, hors grands prématurés, ayant une mère majeure, en mesure de donner un consentement éclairé notamment dans l'une des langues proposées et dont les parents ne résidaient pas temporairement en métropole.

Le plan de sondage utilisé par l'enquête Elfe est appelé **plan d'échantillonnage produit** (cross-classified sampling design) et résulte du croisement indépendant d'un plan de sondage sur la population des maternités et d'un plan de sondage sur la population des jours. Le plan sur les maternités est un plan stratifié (cinq strates relatives à la taille des maternités) et le plan sur les jours peut être modélisé par un plan stratifié (sur les saisons).

Un **estimateur sans biais** de la variance pour l'enquête Elfe a été présenté en formule (6.5.3). Comme sa forme (complexe) demande une programmation spécifique, plusieurs **estimateurs simplifiés** et calculables avec des procédures logicielles déjà existantes ont été proposés et comparés : les résultats montrent que l'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ peut être recommandé aux utilisateurs pour une estimation de la variance simple, et peu biaisée. Les **codes R, SAS, Stata** pour cet estimateur sont proposés en section 6.9 ainsi que la démarche à suivre en trois étapes. Les **variables des bases de données Elfe** nécessaires pour estimer la variance sont listées dans le Tableau 6.7.

Un travail est en cours concernant l'estimation de la variance dans le longitudinal (pour un panel avec processus de non-réponse monotone dans le temps).

References

F. Dell, X. D'Haultfoeuille, P. Février et E. Massé : Mise en œuvre du calcul de variance par linéarisation. *Insee-Méthodes : Actes des Journées de Méthodologie Statistique*, 2002. 189, 191

J.-C. Deville : Variance estimation for complex statistics and estimators : Linearization and residual techniques. *Survey Methodology*, 25(2):193–203, 1999. 189

J.-C. Deville et C.-E Särndal : Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382, 1992. 186

J. L. Eltinge et I. S. Yansaneh : Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. *Survey Methodology*, 23:33–40, 1997. 180

H. Juillard : Two-dimensional sampling in practice. *En révision*, 2016. 175

H. Juillard, G. Chauvet et A. Ruiz-Gazen : Estimateurs de variance issus d'un plan produit pour l'enquête Elfe. *XIIème édition des Journées de Méthodologie Statistique*, 2015a. 173

H. Juillard, G. Chauvet et A. Ruiz-Gazen : Estimation under cross-classified sampling with application to a chilhood survey. *A paraître dans Journal of the American Statistical Association*, 2016. 169, 175

H. Juillard, X. Thierry, N. Razafindratsima, A. Bringé et J.L. Lanoë : Pondérations de l'enquête Elfe en maternité. Rapport technique, 2015b. URL https://pandora.vjf.inserm.fr/public/docs/ELFE_NoteDet0.pdf. 178, 187

J. K. Kim et J. J. Kim : Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35:501–514, 2007. 180

T. Lumley : survey : analysis of complex survey samples, 2014. R package version 3.30. 193

E. OHLSSON : Cross-classified sampling. *Journal of Official Statistics*, 12(3):241–251, 1996. 173

R CORE TEAM : *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/. 191

C.-E. SÄRNDAL, B. SWENSSON et J.H. WRETMAN : *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992. 169

SAS INSTITUTE INC. : *SAS/STAT® 14.1 User's Guide*. Cary, NC, 2013. URL http://www.sas.com/. 191

C. J. SKINNER : Cross-classified sampling : some estimation theory. *Statistics and Probability Letters*, 104:163–168, 2015. 173

STATACORP. : *Stata Statistical Software : Release 13*. StataCorp LP, College Station, TX, 2013. URL http://www.stata.com/. 191

Conclusion

Version française

Dans ce travail, nous avons présenté un cadre théorique général pour des estimateurs issus d'un plan spécifique, appelé plan produit et résultant du croisement indépendant de deux plans de sondage. Ce plan, qui paraît assez naturel, reste un plan très particulier, différent du plan classique à deux degrés et est encore un sujet pour des recherches ultérieures. Des estimateurs de variance de type Horvitz-Thompson et Yates-Grundy ont été décrits et à partir d'une des décompositions possibles de la variance, trois estimateurs simplifiés ont été proposés et étudiés afin de contrer le problème de valeurs négatives possibles prises par ces estimateurs. Pour la suite, les codes logiciels d'un estimateur sans biais de variance et des estimateurs simplifiés pour le plan produit pourraient être insérés dans des packages existants ou dans un nouveau package R, permettant un outil plus accessible.

Pour ce plan produit, dans un cadre général, cinq décompositions possibles de la variance apparaissent, entrainant cinq différents estimateurs de variance de la forme Yates-Grundy. Ces cinq estimateurs peuvent prendre des valeurs différentes selon le plan de sondage utilisé. Les tirages de Poisson conditionnel, de Sampford et de Midzuno illustrent ces différences, sur simulations, et permettent de conclure à un estimateur préférable au moins dans les conditions de tailles d'échantillons et de paramètres considérés pour générer la population. Ces conditions peuvent paraître spécifiques, il serait intéressant de travailler sur d'autres modèles générant les populations pour confirmer les résultats obtenus.

Nous avons comparé le plan classique à deux degrés et le plan produit afin de bien les différencier : les hypothèses utilisées au stade de l'échantillonnage diffèrent, ainsi que les variances et estimateurs de variance. Les procédures logicielles respectives pour les étapes d'échantillonnage et d'estimation ont été présentées sous R, SAS et Stata. Nous avons aussi montré que de façon générale, le plan produit est moins efficace que le plan classique à deux degrés d'échantillonnage. En partant de l'échantillonnage aléatoire simple, nous avons rappelé les correspondances entre un plan à deux degrés et une ANOVA à un facteur et détaillé les correspondances entre un plan produit et une ANOVA à deux facteurs.

Un cadre théorique général a été proposé pour des estimations issues d'une enquête par panel avec processus de non-réponse monotone. Variances et estimateurs de variance ont été calculés dans le cas de probabilités de réponse connues ou estimées. Les paramètres total, ratio et la différence entre deux totaux ont été présentés et leurs estimateurs de variance calculés. Nous avons pris en compte une étape de calage dans l'estimation et effectué plusieurs simulations à partir de trois temps d'enquête successifs. Dans le cas pratique où les probabilités sont estimées, nous avons testé l'estimateur de variance considérant ces probabilités connues et avons pu distinguer les cas où il peut être utilisé de ceux où son biais n'est pas négligeable. Pour la suite, le modèle utilisé pour les simulations pourrait être retravaillé afin de prendre en compte la structure des plans à deux degrés, souvent utilisés en pratique.

Un rapport technique a été remis à l'unité Elfe pour les utilisateurs. Ce document concerne l'estimation de la variance du premier temps de l'enquête, il détaille la prise en compte de l'échantillonnage produit, de la première phase de non-réponse et du calage dans l'estimation de la variance. En considérant les procédures logicielles existant dans R, SAS et Stata, l'utilisation d'un estimateur simplifié a été préconisée aux utilisateurs. En partant de ce rapport, le service des méthodes statistiques de l'Ined, dirigé par A. Bringé, travaille actuellement à des procédures logicielles plus accessibles (de type macro) pour les utilisateurs. Une suite intéressante à ce travail serait l'adaptation, à l'enquête Elfe, des résultats précédents concernant les estimations issues d'une enquête par panel, c'est-à-dire sur plusieurs temps.

En pratique, les poids de sondage ajustés pour le non-réponse ne sont pas nécessairement pris en compte par les utilisateurs de données, une perspective à notre travail est l'étude des biais et des variances des estimateurs naïfs tels que l'estimateur sans aucun poids et l'estimateur ne considérant que les poids de sondage mais omettant l'ajustement de la non-réponse. Une autre perspective à court terme serait l'étude de paramètres plus complexes que ceux présentés dans cette note ; la prise en compte de la variance issue d'un plan de sondage comme celui de l'enquête Elfe dans le calcul d'un test du chi-deux ou dans celui d'une régression logistique fait partie des questionnements actuels des utilisateurs de données.

English version

In this work, we presented a general theoretical framework for estimators from a particular design, called cross-classified sampling design and resulting from the crossing of two independent sampling designs. This design, which seems quite natural, remains a very specific design, different from the standard two-stage sampling design and is a matter for further research. Horvitz-Thompson and Yates-Grundy variance estimators have been described and from a possible decomposition of variance, three simplified estimators have been proposed and studied to counteract the possible negative values taken by these estimators. In the future, the software code of an unbiased estimator and simplified variance estimators could be included in existing packages or in a new R package enabling a more accessible tool.

For this design, in a general framework, five possible decompositions of the variance appear, leading to five different variance estimators of Yates-Grundy form. These five estimators can have different values depending on the used sampling design. The conditional Poisson drawing, the Sampford drawing and the Midzuno drawing illustrate these differences under simulations, and support the conclusion that one estimator is preferable at least in terms of sample sizes and parameters used to generate the population. These conditions may seem specific ; it would be interesting to work on other models generating populations to confirm the results.

We compared the conventional two-stage design and the cross-classified design in order to differentiate them : the underlying assumptions differ, and so do the variance and the variance estimators. The respective software procedures for sampling and estimation steps were presented in R, SAS and Stata. We also showed that generally the cross-classified sampling design is less efficient than the standard conventional two-stage sampling design. Starting from simple random sampling, we recalled the connection between a two-stage design and a one-way ANOVA and detailed the connection between a cross-classified design and a two-factor ANOVA.

A general theoretical framework was proposed for estimators from a panel survey with monotone non-response process. Variance and variance estimators have been proposed in the case of known or estimated response probabilities. We consider the estimation of a total, a ratio and a difference between two totals, and their variance estimators. We took into account a calibration step in the estimator and conducted several simulations from three successive times of survey. In the practical case where the probabilities are estimated, we compared the (approximately) unbiased variance estimator with a simplified variance estimator, computed as if the response probabilities were known. We were able to distinguish cases where the simplified variance estimator can be used, and cases where its bias is not negligible. For the continuation, the model used for the simulations could be reworked to take into account the structure of the designs with two stages, often used in practice for household ans social surveys.

A technical report was delivered to the ELFE unit and to the ELFE data users. This document is related to the estimated variance at the first time of the survey. It details taking into account the cross-classified sampling design, the first phase of non-response and the calibration step in the estimation of variance. Considering the existing software procedures in R, SAS and Stata, the use of a simplified estimator has been advocated to users. Based on this report, the service of statistical methods at INED, led by A. Bringe, currently works at more affordable software procedures (macro) for users. An interesting continuation of this work would be the adaptation, for the ELFE survey, of previous results to estimators from a panel survey.

In practice, the survey weights adjusted for non-response are not necessarily taken into account by the data users. A perspective of future work is the study of bias and variance estimators for "naive" unweighted estimators, and for the estimator using the sampling weights, but omitting the adjustment for non-response. Another short-term perspective is the study of more complex parameters than those presented in this note ; taking into account the sampling variance of the ELFE survey in the calculation of a chi-squared test or in the logistic regression is part of the current questions of data users.

Annexe A

Matériel supplémentaire relatif au Chapitre 2

Dans la première annexe se trouvent les fonctions nécessaires au calcul des estimateurs. Dans la seconde annexe, ce sont les commandes pour faire tourner ces fonctions et afficher les résultats des Tableaux 2.1 et 2.2.

A.1 Fonctions

```
#R functions documented:
#DRAWsi
#DRAWsisi
#ESTratio
#ESTtotal
#ESTVARlinratio
#ESTVARtotal
#LINratio
#MEASURESratio
#MEASUREStotal
#POPgenerationRATIO
#POPgenerationTOTAL
#POPratio
#POPtotal
#SAMPLEsisi
#SIMULATIONtotal
#SIMULATIONratio
#VARIANCE_MCratio
#VARIANCE_MCtotal

#Note replicability: the seed was fixed using the command set.seed() in the
#functions SIMULATIONtotal, SIMULATIONratio, VARIANCE_MCratio and VARIANCE_MCtotal.
```

```
#################################################################################################
#DESCRIPION: draws (and fixes) simple sample (equal probabilities) without replacement
#USAGE: DRAWsi(n,N)
#ARGUMENTS:
#               n        size of the sample
#               N        size of the population
#VALUE: the function returns a vector of lenght n

DRAWsi <- function(n,N)
        { s=c(); s=sample(1:N,n); return(s) }


#################################################################################################
#DESCRIPION: draws (and fixes) two simple samples (equal probabilities) without replacement in two
    populations
#USAGE: DRAWsisi(n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               n_m      size of the sample drawn in the population U_m
#               n_d      size of the sample drawn in the population U_d
#               N_m      size of the population U_m
#               N_d      size of the population U_d
#VALUE: the function returns a list of two vectors of respective lenghts n_m and n_d

DRAWsisi <- function(n_m,n_d,N_m,N_d)
        {
        sd=DRAWsi(n_d,N_d); sm=DRAWsi(n_m,N_m)
        S=list(sm,sd) ; return(S)
        }


#################################################################################################
#DESCRIPION: estimates a population total
#USAGE: ESTtotal(ECH,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               ECH      matrix of the interest variable, of size [n_d,n_m]
#               n_m      size of the sample drawn in the population U_m
#               n_d      size of the sample drawn in the population U_d
#               N_m      size of the population U_m
#               N_d      size of the population U_d
#VALUE: the function returns a value for the estimated total

ESTtotal <- function(ECH,n_m,n_d,N_m,N_d)
        {w=(N_m/n_m)*(N_d/n_d); Tpi=sum(w*c(ECH)) ; return(Tpi) }


#################################################################################################
#DESCRIPION: estimates a population ratio Y/X
#USAGE: ESTratio(ECHY,ECHX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               ECHY     matrix of the numerator variable, of size [n_d,n_m]
#               ECHX     matrix of the denominator variable, of size [n_d,n_m]
#               n_m      size of the sample drawn in the population U_m
#               n_d      size of the sample drawn in the population U_d
#               N_m      size of the population U_m
#               N_d      size of the population U_d
#VALUE: the function returns a value for the estimated ratio

ESTratio <- function(ECHY,ECHX,n_m,n_d,N_m,N_d)
        {Rpi=ESTtotal(ECHY,n_m,n_d,N_m,N_d)/ESTtotal(ECHX,n_m,n_d,N_m,N_d) ; return(Rpi) }


#################################################################################################
```

```
#DESCRIPION: estimates the variance of the estimated total of the estimated linearized variable of
    a ratio Y/X
#USAGE: ESTVARlinratio(ECHY,ECHX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#              ECHY    matrix of the numerator variable, of size [n_d,n_m]
#              ECHX    matrix of the denominator variable, of size [n_d,n_m]
#              n_m     size of the sample drawn in the population U_m
#              n_d     size of the sample drawn in the population U_d
#              N_m     size of the population U_m
#              N_d     size of the population U_d
#VALUE: the function returns a list of four values (the unbiased estimator and 3 simplified
    estimators: Vsimp1, Vsimp2, Vsimp3)

ESTVARlinratio <- function(ECHY,ECHX,n_m,n_d,N_m,N_d)
       {
       Rlin=LINratio(ECHY,ECHX,n_m,n_d,N_m,N_d)
       vl=ESTVARtotal(Rlin,n_m,n_d,N_m,N_d) ; return(vl)
       }


####################################################################################################
#DESCRIPION: estimates the variance of an estimated total
#USAGE: ESTVARtotal(ECH,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#              ECH     matrix of the interest variable, of size [n_d,n_m]
#              n_m     size of the sample drawn in the population U_m
#              n_d     size of the sample drawn in the population U_d
#              N_m     size of the population U_m
#              N_d     size of the population U_d
#VALUE: the function returns a list of four values (the unbiased estimator and 3 simplified
    estimators: Vsimp1, Vsimp2, Vsimp3)

ESTVARtotal <- function(ECH,n_m,n_d,N_m,N_d)
       {
       w_m=N_m/n_m; w_d=N_d/n_d
       VarM=var(apply(ECH*w_d,2,sum)); Vsimp1=(N_m^2)*(1/n_m - 1/N_m)*VarM
       VarD=var(apply(ECH*w_m,1,sum)); Vsimp2=(N_d^2)*(1/n_d - 1/N_d)*VarD
       d=apply(ECH,1,mean); e=apply(ECH,2,mean); f=mean(ECH)
       X = t(t(ECH - d) - e) + f; X2=X*X
       Vco=(1/(n_d - 1))*(1/(n_m - 1))*sum(X2)
       VCORR=(N_d^2)*(1/n_d - 1/N_d)*(N_m^2)*(1/n_m - 1/N_m)*Vco
       VCCS=Vsimp1+Vsimp2-VCORR ; Vsimp3=Vsimp1+Vsimp2
       v=list(VCCS,Vsimp1,Vsimp2,Vsimp3) ; return(v)
       }


####################################################################################################
#DESCRIPION: estimates the linearized variable of a ratio Y/X
#USAGE: LINratio(ECHY,ECHX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#              ECHY    matrix of the numerator variable, of size [n_d,n_m]
#              ECHX    matrix of the denominator variable, of size [n_d,n_m]
#              n_m     size of the sample drawn in the population U_m
#              n_d     size of the sample drawn in the population U_d
#              N_m     size of the population U_m
#              N_d     size of the population U_d
#VALUE: the function returns a matrix of length [n_d,n_m]

LINratio <- function(ECHY,ECHX,n_m,n_d,N_m,N_d)
       {
       rpi=ESTratio(ECHY,ECHX,n_m,n_d,N_m,N_d)
       txpi=ESTtotal(ECHX,n_m,n_d,N_m,N_d)
```

```
        upi = (ECHY - ECHX*rpi) / txpi ; return(upi)
        }


##############################################################################################
#DESCRIPION: measures (with simulations in Population) the relative biases of an estimated ratio Y/
     X and of its variance estimators
#USAGE: MEASURESratio(NBech,NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               NBech               number of simulations to calculate the estimated ratio and its
     variance estimators
#               NBechMC number of simulations to calculate the 'true' variance
#               POPY     matrix of the numerator variable, of size [N_d,N_m]
#               POPX     matrix of the denominator variable, of size [N_d,N_m]
#               n_m      size of the sample drawn in the population U_m
#               n_d      size of the sample drawn in the population U_d
#               N_m      size of the population U_m
#               N_d      size of the population U_d
#VALUE: the function returns a list of 17 values: the sizes n_m, N_m, n_d, N_d,
#the ratio population, its estimate under NBech simulations, its relative bias,
#the variance monteCarlo under NBechMC simulation, the unbiased estimated variance under NBech
     simulations, its relative bias,
#the number of negative value of the unbiased estimated variance under NBech simulations,
#the simplified variance estimator Vsimp1 under NBech simulations, its relative bias,
#the simplified variance estimator Vsimp2 under NBech simulations, its relative bias,
#the simplified variance estimator Vsimp3 under NBech simulations, its relative bias

MEASURESratio <- function(NBech,NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
        {
        esp=SIMULATIONratio(NBech,POPY,POPX,n_m,n_d,N_m,N_d)
        varmc=VARIANCE_MCratio(NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
        t=POPratio(POPY,POPX)
        RBratio=((esp[[1]]-t)/t) *100
        RBvar=((esp[[2]]-varmc)/varmc) *100
        RBvar1=((esp[[4]]-varmc)/varmc) *100;RBvar2=((esp[[5]]-varmc)/varmc) *100;RBvar3=((esp[[6]]
            -varmc)/varmc) *100
        m=list(n_m,N_m,n_d,N_d,t,esp[[1]],RBratio,varmc,esp[[2]],RBvar,esp[[3]],esp[[4]],RBvar1,esp
             [[5]],RBvar2,esp[[6]],RBvar3)
        return(m)
        }


##############################################################################################
#DESCRIPION: measures (with simulations in Population) the relative biases of an estimated total
     and of its variance estimators
#USAGE: MEASUREStotal(NBech,NBechMC,POP,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               NBech               number of simulations to calculate the estimated total and its
     variance estimators
#               NBechMC number of simulations to calculate the 'true' variance
#               POP      matrix of the interest variable, of size [N_d,N_m]
#               n_m      size of the sample drawn in the population U_m
#               n_d      size of the sample drawn in the population U_d
#               N_m      size of the population U_m
#               N_d      size of the population U_d
#VALUE: the function returns a list of 17 values: the sizes n_m, N_m, n_d, N_d,
#the total population, its estimate under NBech simulations, its relative bias,
#the variance monteCarlo under NBechMC simulation, the unbiased estimated variance under NBech
     simulations, its relative bias,
#the number of negative value of the unbiased estimated variance under NBech simulations,
#the simplified variance estimator Vsimp1 under NBech simulations, its relative bias,
#the simplified variance estimator Vsimp2 under NBech simulations, its relative bias,
#the simplified variance estimator Vsimp3 under NBech simulations, its relative bias

MEASUREStotal <- function(NBech,NBechMC,POP,n_m,n_d,N_m,N_d)
          {
          esp=SIMULATIONtotal(NBech,POP,n_m,n_d,N_m,N_d)
          varmc=VARIANCE_MCtotal(NBechMC,POP,n_m,n_d,N_m,N_d)
          t=POPtotal(POP)
          RBtotal=((esp[[1]]-t)/t) *100
          RBvar=((esp[[2]]-varmc)/varmc) *100
          RBvar1=((esp[[4]]-varmc)/varmc) *100;RBvar2=((esp[[5]]-varmc)/varmc) *100;RBvar3=((esp[[6]]
                -varmc)/varmc) *100
          m=list(n_m,N_m,n_d,N_d,t,esp[[1]],RBtotal,varmc,esp[[2]],RBvar,esp[[3]],esp[[4]],RBvar1,esp
                [[5]],RBvar2,esp[[6]],RBvar3)
          return(m)
          }


##########################################################################################
#DESCRIPION: generates the interest variables presented in Table 2 (count variables)
#USAGE: POPgenerationRATIO(case,mu,sigma1,sigma2,sigma3,N_m,N_d)
#ARGUMENTS:
#                    case      1 corresponds to equal probabilities / 2 corresponds to inequal
      probabilities
#                    mu        constant
#                    sigma1    maternity effect
#                    sigma2    day effect
#                    sigma3    residual effect
#                    N_m       size of the population U_m
#                    N_d       size of the population U_d
#VALUE: the function returns a list of two matrices of length [N_m,N_d]

POPgenerationRATIO <- function(case,mu,sigma1,sigma2,sigma3,N_m,N_d)
          {
          model=POPgenerationTOTAL(mu,sigma1,sigma2,sigma3,N_m,N_d)
          set.seed(54321) ; lambd=exp(log(model)) ; N=N_m*N_d
          Xik=rpois(N, lambda=lambd) ; POP1x=matrix(Xik,nrow=N_d)
          if (case==1)
                    {p=0.3}
          if (case==2)
                    {beta=log(0.3/(1-0.3))/mean(POP1x) ; betak=matrix(rep(beta,N),nrow=N_d) ; p=(exp(
                          betak*POP1x))/(1+exp(betak*POP1x)) }
          Yik=rbinom(N,Xik,p) ; POP1y=matrix(Yik,nrow=N_d)
          a=list(POP1x,POP1y) ; return(a)
          }


##########################################################################################
#DESCRIPION: generates the interest variables presented in Table 1 (model of two-way ANOVA)
#USAGE: POPgenerationTOTAL(mu,sigma1,sigma2,sigma3,N_m,N_d)
#ARGUMENTS:
#                    mu        constant
#                    sigma1    maternity effect
#                    sigma2    day effect
#                    sigma3    residual effect
#                    N_m       size of the population U_m
#                    N_d       size of the population U_d
#VALUE: the function returns a matrix of length [N_m,N_d]

POPgenerationTOTAL <- function(mu,sigma1,sigma2,sigma3,N_m,N_d)
          {
          set.seed(4321) ; N=N_m*N_d
          Yi=rnorm(N_m,0,1)*sigma1; YI=rep(Yi,N_d); POP1yi=matrix(YI,nrow=N_m); POP1yi=t(POP1yi)
```

```
        Yk=rnorm(N_d,0,1)*sigma2; YK=rep(Yk,N_m); POP1yk=matrix(YK,nrow=N_d)
        Y=rnorm(N,0,1)*sigma3; POP1yy=matrix(Y,nrow=N_d)
        POP1=mu+POP1yi+POP1yk+POP1yy ; return(POP1)
        }


##################################################################################################
#DESCRIPION: calculates the population ratio Y/X
#USAGE: POPratio(POPY,POPX)
#ARGUMENTS:
#               POPY    matrix of the numerator variable
#               POPX    matrix of the denominator variable
#VALUE: the function returns a value for the population ratio

POPratio <- function(POPY,POPX)
        { return(sum(POPY)/sum(POPX)) }


##################################################################################################
#DESCRIPION: calculates the population total
#USAGE: POPtotal(POP)
#ARGUMENTS:
#               POP     matrix of the interest variable
#VALUE: the function returns a value for the population total

POPtotal <- function(POP)
        { return(sum(POP)) }


##################################################################################################
#DESCRIPION: calculates the population ratio Y/X
#USAGE: SAMPLEsisi(POP, s_m, s_d)
#ARGUMENTS:
#               POP     matrix of the interest variable
#               s_m     vector of lenght<dim(POP)[2]
#               s_d     vector of lenght<dim(POP)[1]
#VALUE: the function returns a matrix of length [length(s_m),length(s_d)]

SAMPLEsisi <- function(POP, s_m, s_d)
        {s=POP[s_d,s_m] ; return(s) }


##################################################################################################
#DESCRIPION: simulations in Population for an estimated ratio Y/X and its variance estimators
#USAGE: SIMULATIONratio(NBech,POPY,POPX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               NBech           number of simulations to calculate the estimated ratio and its
    variance estimators
#               POPY    matrix of the numerator variable, of size [N_d,N_m]
#               POPX    matrix of the denominator variable, of size [N_d,N_m]
#               n_m     size of the sample drawn in the population U_m
#               n_d     size of the sample drawn in the population U_d
#               N_m     size of the population U_m
#               N_d     size of the population U_d
#VALUE: the function returns a list of 6 values:
#the ratio estimator under NBech simulations, its unbiased variance estimator under NBech
    simulations,
#the number of negative value of the unbiased variance estimator under NBech simulations
#the simplified variance estimator Vsimp1 under NBech simulations
#the simplified variance estimator Vsimp2 under NBech simulations
#the simplified variance estimator Vsimp3 under NBech simulations
```

```
SIMULATIONratio <- function(NBech,POPY,POPX,n_m,n_d,N_m,N_d)
        {
        Est<-c(); EstVar<-c(); EstVar1<-c(); EstVar2<-c(); EstVar3<-c() ; set.seed(12345)
        for (i in 1:NBech)
                {
                s=DRAWsisi(n_m,n_d,N_m,N_d)
                echY=SAMPLEsisi(POPY,s[[1]],s[[2]])
                echX=SAMPLEsisi(POPX,s[[1]],s[[2]])
                Est<-c(Est, ESTratio(echY,echX,n_m,n_d,N_m,N_d))
                v=ESTVARlinratio (echY,echX,n_m,n_d,N_m,N_d)
                EstVar<-c(EstVar,v[[1]])
                EstVar1<-c(EstVar1,v[[2]]); EstVar2<-c(EstVar2,v[[3]]); EstVar3<-c(EstVar3,v[[4]])
                }
        ESPest=mean(Est) ; compteur=sum(EstVar<0)
        ESPvar=mean(EstVar);ESPvar1=mean(EstVar1);ESPvar2=mean(EstVar2);ESPvar3=mean(EstVar3)
        E=list(ESPest,ESPvar,compteur,ESPvar1,ESPvar2,ESPvar3) ; return(E)
        }


##################################################################################################
#DESCRIPION: simulations in Population for an estimated total and its variance estimators
#USAGE: SIMULATIONtotal(NBech,POP,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               NBech           number of simulations to calculate the estimated ratio and its
     variance estimators
#               POP     matrix of the interest variable, of size [N_d,N_m]
#               n_m     size of the sample drawn in the population U_m
#               n_d     size of the sample drawn in the population U_d
#               N_m     size of the population U_m
#               N_d     size of the population U_d
#VALUE: the function returns a list of 6 values:
#the ratio estimator under NBech simulations, its unbiased variance estimator under NBech
     simulations,
#the number of negative value of the unbiased variance estimator under NBech simulations
#the simplified variance estimator Vsimp1 under NBech simulations
#the simplified variance estimator Vsimp2 under NBech simulations
#the simplified variance estimator Vsimp3 under NBech simulations

SIMULATIONtotal <- function(NBech,POP,n_m,n_d,N_m,N_d)
        {
        Est<-c(); EstVar<-c(); EstVar1<-c(); EstVar2<-c(); EstVar3<-c() ; set.seed(12345)
        for (i in 1:NBech)
                {
                s=DRAWsisi(n_m,n_d,N_m,N_d)
                ech=SAMPLEsisi(POP,s[[1]],s[[2]])
                Est<-c(Est, ESTtotal(ech,n_m,n_d,N_m,N_d))
                v=ESTVARtotal(ech,n_m,n_d,N_m,N_d)
                EstVar<-c(EstVar,v[[1]])
                EstVar1<-c(EstVar1,v[[2]]); EstVar2<-c(EstVar2,v[[3]]); EstVar3<-c(EstVar3,v[[4]])
                }
        ESPest=mean(Est) ; compteur=sum(EstVar<0)
        ESPvar=mean(EstVar);ESPvar1=mean(EstVar1);ESPvar2=mean(EstVar2);ESPvar3=mean(EstVar3)
        E=list(ESPest,ESPvar,compteur,ESPvar1,ESPvar2,ESPvar3) ; return(E)
        }


##################################################################################################
#DESCRIPION: calculates independently the MonteCarlo variance of an estimated ratio Y/X
#USAGE: VARIANCE_MCratio(NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               NBechMC         number of simulations to calculate the MC variance
#               POPY    matrix of the numerator variable, of size [N_d,N_m]
#          POPX     matrix of the denominator variable, of size [N_d,N_m]
#          n_m      size of the sample drawn in the population U_m
#          n_d      size of the sample drawn in the population U_d
#          N_m      size of the population U_m
#          N_d      size of the population U_d
#VALUE: the function returns a value for the MC variance

VARIANCE_MCratio <- function(NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
     {
     Est<-c() ; set.seed(123456)
     for (i in 1:NBechMC)
          {
          s=DRAWsisi(n_m,n_d,N_m,N_d)
          echY=SAMPLEsisi(POPY,s[[1]],s[[2]])
          echX=SAMPLEsisi(POPX,s[[1]],s[[2]])
          Est<-c(Est, ESTratio(echY,echX,n_m,n_d,N_m,N_d))
          }
     return(var(Est))
     }


###############################################################################################
#DESCRIPION: calculates independently the MonteCarlo variance of an estimated ratio Y/X
#USAGE: VARIANCE_MCtotal(NBechMC,POP,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#          NBechMC       number of simulations to calculate the MC variance
#          POP      matrix of the interest variable, of size [N_d,N_m]
#          n_m      size of the sample drawn in the population U_m
#          n_d      size of the sample drawn in the population U_d
#          N_m      size of the population U_m
#          N_d      size of the population U_d
#VALUE: the function returns a value for the MC variance

VARIANCE_MCtotal <- function(NBechMC,POP,n_m,n_d,N_m,N_d)
     {
     Est<-c() ; set.seed(123456)
     for (i in 1:NBechMC)
          {
          s=DRAWsisi(n_m,n_d,N_m,N_d)
          ech=SAMPLEsisi(POP,s[[1]],s[[2]])
          Est<-c(Est, ESTtotal(ech,n_m,n_d,N_m,N_d) )
          }
     return(var(Est))
     }
```

Code A.1 – CodeR_functions.r : Basic functions required to calculate estimators.

A.2 Simulations

```


############## STEP 1
#Compile all the code in "CodeR_functions.r"
#(this is the required functions for the simulations in Tables 1 and 2)
#For this, you can choose one of the two following commands:
source("C:/.../CodeR_functions.r")
source(file.choose())
```

```
############## STEP 2: GENERATION VARIABLES
# Variables used in Table 1 using the function POPgenerationTOTAL
# POP1y : top left (5,5)          # POP2y : bottom left (0.5,5)
# POP3y : top right (50,5)        # POP4y : bottom right (0.5,0.5)
# Variables used in Table 2 using the function POPgenerationRATIO
# POP5a : numerator top left      # POP5b : numerator bottom left
# POP6a : numerator top right # POP6b : numerator bottom right
# POP5x : denominator left        # POP6x : denominator right

N_m=1000 ; N_d=1000 #Sizes of POPulations

POP1y=POPgenerationTOTAL(mu=200,sigma1=5,sigma2=5,sigma3=5,N_m,N_d)
POP2y=POPgenerationTOTAL(mu=200,sigma1=0.5,sigma2=5,sigma3=5,N_m,N_d)
POP3y=POPgenerationTOTAL(mu=200,sigma1=50,sigma2=5,sigma3=5,N_m,N_d)
POP4y=POPgenerationTOTAL(mu=200,sigma1=0.5,sigma2=0.5,sigma3=5,N_m,N_d)

POP5=POPgenerationRATIO(case=1,mu=200,sigma1=5,sigma2=5,sigma3=5,N_m=1000,N_d=1000)
POP5x=POP5[[1]] ; POP5ay=POP5[[2]]
POP5=POPgenerationRATIO(case=2,mu=200,sigma1=5,sigma2=5,sigma3=5,N_m=1000,N_d=1000)
POP5by=POP5[[2]]
POP6=POPgenerationRATIO(case=1,mu=200,sigma1=50,sigma2=5,sigma3=5,N_m=1000,N_d=1000)
POP6x=POP6[[1]] ; POP6ay=POP6[[2]]
POP6=POPgenerationRATIO(case=2,mu=200,sigma1=50,sigma2=5,sigma3=5,N_m=1000,N_d=1000)
POP6by=POP6[[2]]

############## STEP 3: SIMULATIONS


############## STEP 3 a) Table 1
#Top left cell of Table 1
a=MEASUREStotal(10000,50000,POP1y,5,5,N_m,N_d)
b=MEASUREStotal(10000,50000,POP1y,10,10,N_m,N_d)
c=MEASUREStotal(10000,50000,POP1y,10,100,N_m,N_d)
d=MEASUREStotal(10000,50000,POP1y,100,100,N_m,N_d)
e=MEASUREStotal(10000,50000,POP1y,500,500,N_m,N_d)
#Bottom left cell of Table 1
a=MEASUREStotal(10000,50000,POP2y,5,5,N_m,N_d)
b=MEASUREStotal(10000,50000,POP2y,10,10,N_m,N_d)
c=MEASUREStotal(10000,50000,POP2y,10,100,N_m,N_d)
d=MEASUREStotal(10000,50000,POP2y,100,100,N_m,N_d)
e=MEASUREStotal(10000,50000,POP2y,500,500,N_m,N_d)
#Top right cell of Table 1
a=MEASUREStotal(10000,50000,POP3y,5,5,N_m,N_d)
b=MEASUREStotal(10000,50000,POP3y,10,10,N_m,N_d)
c=MEASUREStotal(10000,50000,POP3y,10,100,N_m,N_d)
d=MEASUREStotal(10000,50000,POP3y,100,100,N_m,N_d)
e=MEASUREStotal(10000,50000,POP3y,500,500,N_m,N_d)
#Bottom right cell of Table 1
a=MEASUREStotal(10000,50000,POP4y,5,5,N_m,N_d)
b=MEASUREStotal(10000,50000,POP4y,10,10,N_m,N_d)
c=MEASUREStotal(10000,50000,POP4y,10,100,N_m,N_d)
d=MEASUREStotal(10000,50000,POP4y,100,100,N_m,N_d)
e=MEASUREStotal(10000,50000,POP4y,500,500,N_m,N_d)

resultA=cbind(a,b,c,d,e)
rownames(resultA)=c("$n_M$","$N_M$","$n_D$","$N_D$"
,"$ParameterPOP$","$E(Parameter)$","$mbox{RB}_{MC} left(hat{Parameter} right)$"
,"$Var_{MC}$","$E(hat{V})$","$mbox{RB}_{MC} left(hat{V} right)$","NEG"
,"$E(hat{V}_{SIMP1})$","$mbox{RB}_{MC} left(hat{V}_{SIMP1} right)$"
,"$E(hat{V}_{SIMP2})$","$mbox{RB}_{MC} left(hat{V}_{SIMP2} right)$"
,"$E(hat{V}_{SIMP3})$","$mbox{RB}_{MC} left(hat{V}_{SIMP3} right)$")
```

```
resultA


############## STEP 3 b) Table 2
#Top left cell of Table 2
a=MEASURESratio(10000,50000,POP5ay,POP5x,5,5,N_m,N_d)
b=MEASURESratio(10000,50000,POP5ay,POP5x,10,10,N_m,N_d)
c=MEASURESratio(10000,50000,POP5ay,POP5x,10,100,N_m,N_d)
d=MEASURESratio(10000,50000,POP5ay,POP5x,100,100,N_m,N_d)
e=MEASURESratio(10000,50000,POP5ay,POP5x,500,500,N_m,N_d)
#Bottom left cell of Table 2
a=MEASURESratio(10000,50000,POP5by,POP5x,5,5,N_m,N_d)
b=MEASURESratio(10000,50000,POP5by,POP5x,10,10,N_m,N_d)
c=MEASURESratio(10000,50000,POP5by,POP5x,10,100,N_m,N_d)
d=MEASURESratio(10000,50000,POP5by,POP5x,100,100,N_m,N_d)
e=MEASURESratio(10000,50000,POP5by,POP5x,500,500,N_m,N_d)
#Top right cell of Table 2
a=MEASURESratio(10000,50000,POP6ay,POP6x,5,5,N_m,N_d)
b=MEASURESratio(10000,50000,POP6ay,POP6x,10,10,N_m,N_d)
c=MEASURESratio(10000,50000,POP6ay,POP6x,10,100,N_m,N_d)
d=MEASURESratio(10000,50000,POP6ay,POP6x,100,100,N_m,N_d)
e=MEASURESratio(10000,50000,POP6ay,POP6x,500,500,N_m,N_d)
#Bottom right cell of Table 2
a=MEASURESratio(10000,50000,POP6by,POP6x,5,5,N_m,N_d)
b=MEASURESratio(10000,50000,POP6by,POP6x,10,10,N_m,N_d)
c=MEASURESratio(10000,50000,POP6by,POP6x,10,100,N_m,N_d)
d=MEASURESratio(10000,50000,POP6by,POP6x,100,100,N_m,N_d)
e=MEASURESratio(10000,50000,POP6by,POP6x,500,500,N_m,N_d)

resultB=cbind(a,b,c,d,e)
rownames(resultB)=c("$n_M$","$N_M$","$n_D$","$N_D$"
,"$ParameterPOP$","$E(Parameter)$","$mbox{RB}_{MC} left(hat{Parameter} right)$"
,"$Var_{MC}$","$E(hat{V})$","$mbox{RB}_{MC} left(hat{V} right)$","NEG"
,"$E(hat{V}_{SIMP1})$","$mbox{RB}_{MC} left(hat{V}_{SIMP1} right)$"
,"$E(hat{V}_{SIMP2})$","$mbox{RB}_{MC} left(hat{V}_{SIMP2} right)$"
,"$E(hat{V}_{SIMP3})$","$mbox{RB}_{MC} left(hat{V}_{SIMP3} right)$")
resultB
```

Code A.2 – CodeR_Tables.r : Generation of the interest variables and results of simulations.

Annexe B

Matériel supplémentaire relatif au Chapitre 3

Dans la première annexe se trouvent les fonctions nécessaires au calcul des estimateurs. Dans la seconde annexe, ce sont les commandes pour faire tourner ces fonctions et afficher les résultats des Tableaux 3.3 à 3.12.

B.1 Fonctions

```
#R FUNCTIONS REQUIRED FOR tables 1 and 2

#R functions documented:
#EstTOTAL
#PIKetSIMU
#RECpps
#Sampling
#SIMU
#TOTALpopulation
#varestYG
#VARIANCE_MC

#Note replicability: the seed was fixed using the command set.seed() in the functions SIMU,
     VARIANCE_MC.


library(sampling)

####################################################################################################
#DESCRIPION: estimates the variance of an estimated total
#USAGE: varestYG(cc, pikl, piij)
#ARGUMENTS:
#               cc      matrix of the interest variable, of size [n_d,n_m]
#               pikl    matrix of variance-covariance of pikl (population U_m)
#               piij    matrix of variance-covariance of piij (population U_d)
#VALUE: the function returns a list of 10 values (the 5 variance estimators and their parts)

varestYG <-function(cc, pikl, piij)
```

```
         {n_m=dim(piij)[1] ; n_d=dim(pikl)[1]
         dk=diag(pikl) ; PIK=t(matrix(rep(dk,n_m),ncol=n_m))
         di=diag(piij) ; PII=(matrix(rep(di,n_d),ncol=n_d))

         dd=cc/(PIK*PII) ; k=kronecker(dd,dd,"-") ; k2=k^2

         r=t(rep(as.vector(t(pikl)),n_m^2))
         Pikl=matrix(r,ncol=n_m^2) ; Pikl=t(Pikl)
         s=(rep(as.vector((piij)),n_d^2))
         Piij=matrix(s,ncol=n_d^2)

         mdk=matrix(dk) %*% t(matrix(dk))
         rmdk=rep(as.vector(mdk),n_m^2)
         PikPil=matrix(rmdk,ncol=n_m^2) ; PikPil=t(PikPil) ; PikPil

         mdi=matrix(di) %*% t(matrix(di))
         rmdi=rep(as.vector(mdi),n_d^2)
         PiiPij=matrix(rmdi,ncol=n_d^2) ; PiiPij=(PiiPij) ; PiiPij

         coeffijkl= 1 - ( (PikPil*PiiPij)/(Pikl*Piij) )
         vYGa=-0.5*sum(coeffijkl*k2)

         Yk=apply(cc*(1/PII),2,sum)
         A <- ((pikl-mdk)/pikl)*(kronecker(Yk/dk,matrix(1,1,n_d)) - kronecker(matrix(1,n_d,1),t(Yk/
               dk)))^2
         vYGb1=-0.5*sum(A)

         Yi=apply(cc*(1/PIK),1,sum)
         B <- ((piij-mdi)/piij)*(kronecker(Yi/di,matrix(1,1,n_m)) - kronecker(matrix(1,n_m,1),t(Yi/
               di)))^2
         vYGc1=-0.5*sum(B)

         ee=cc/PII
         e=kronecker(ee,ee,"-") ; e2=e^2
         coeffijkl2= (Piij - PiiPij)/(Pikl*Piij)
         vYGb2=-0.5*sum(coeffijkl2*e2)

         ff=cc/PIK
         f=kronecker(ff,ff,"-") ; f2=f^2
         coeffijkl3= (Pikl - PikPil)/(Pikl*Piij)
         vYGc2=-0.5*sum(coeffijkl3*f2)

         coeffijkl4= ((Piij-PiiPij)*(Pikl-PikPil))/(Pikl*Piij)
         vYGd3=-0.5*sum(coeffijkl4*k2)

         #Vsimp4
         Ykpond = (sum((Yk/dk)*(1-dk)))/sum(1-dk)
         V4a = (n_d/(n_d-1)) * sum( (1-dk) * (Yk/dk - Ykpond)^2 )
         Yipond = (sum((Yi/di)*(1-di)))/sum(1-di)
         V4b = (n_m/(n_m-1)) * sum( (1-di) * (Yi/di - Yipond)^2 )
         Vsimp4=V4a+V4b

         #Vsimp5
         Ytot= sum(Yk/dk)
         V5a = (n_d/(n_d-1)) * sum( (Yk/dk - Ytot/n_d)^2 )
         V5b = (n_m/(n_m-1)) * sum( (Yi/di - Ytot/n_m)^2 )
         Vsimp5=V5a+V5b

         vb=vYGb1+vYGb2; vc=vYGc1+vYGc2; vd=vYGb1+vYGc1-vYGd3; ve=vYGb2+vYGc2+vYGd3 ; Vsimp3=vYGb1+
               vYGc1
         return(c(vYGa,vd,ve,vb,vc,vYGb1,vYGc1,vYGd3,vYGb2,vYGc2,Vsimp3,Vsimp4,Vsimp5))
}
```

```
###############################################################################################
#DESCRIPION: draws (and fixes) two simple samples (equal probabilities) without replacement in two
      populations
#USAGE: Sampling(tirage,PIk,PIi)
#ARGUMENTS:
#                 tirage   MIDZUNO, SAMPFORD or REJECTIF
#                 PIk      probabilities for the sample drawn in the population U_m
#                 PIi      probabilities for the sample drawn in the population U_d
#VALUE: the function returns a list of two vectors of respective lenghts n_m and n_d

Sampling <- function(tirage,PIk,PIi)
        {
        if (tirage=="MIDZUNO") {s_d=UPmidzuno(PIk) ; s_m=UPmidzuno(PIi)}
        if (tirage=="REJECTIF") {s_d=UPmaxentropy(PIk) ; s_m=UPmaxentropy(PIi)}
        if (tirage=="SAMPFORD") {s_d=UPsampford(PIk) ; s_m=UPsampford(PIi)} #, max_iter=50000000
        S=list(s_m,s_d)
        return(S)
        }

###############################################################################################
#DESCRIPION: calculates the population total
#USAGE: TOTALpopulation(POP)
#ARGUMENTS:
#                 POP      matrix of the interest variable
#VALUE: the function returns a value for the population total

TOTALpopulation <- function(POP)
        { return(sum(POP)) }


###############################################################################################
#DESCRIPION: estimates a population total
#USAGE: EstTOTAL(y,PIkl,PIij)
#ARGUMENTS:
#                 y        matrix of the interest variable, of size [n_d,n_m]
#                 PIkl     matrix of variance-covariance of pikl (population U_m)
#                 PIij     matrix of variance-covariance of pikl (population U_d)
#VALUE: the function returns a value for the estimated total

EstTOTAL <- function(y,PIkl,PIij)
        {n_m=dim(PIij)[1] ; n_d=dim(PIkl)[1]
        pii=diag(PIij) ; PIi=matrix(rep(pii,n_d),nrow=n_m,ncol=n_d)
        pik=diag(PIkl) ; PIk=t(matrix(rep(pik,n_m),nrow=n_d,ncol=n_m))
        Y=y*(1/PIi)*(1/PIk)
        return(sum(Y))
        }


###############################################################################################
#DESCRIPION: calculates independently the MonteCarlo variance of an estimated total
#USAGE: VARIANCE_MC(tirage,NBechMC,Y,PIk,PIi)
#ARGUMENTS:
#                 tirage   MIDZUNO, SAMPFORD or REJECTIF
#                 NBechMC  number of simulations to calculate the MC variance
#                 Y        matrix of the interest variable, of size [N_d,N_m]
#                 PIk      probabilities for the sample drawn in the population U_m
#                 PIi      probabilities for the sample drawn in the population U_d
#VALUE: the function returns a value for the MC variance

VARIANCE_MC <- function(tirage,NBechMC,Y,PIk,PIi)
```

```
        {Est<-c()  ; set.seed(4391)
        if (tirage=="MIDZUNO") {POPpiij=UPmidzunopi2(PIi) ; POPpikl=UPmidzunopi2(PIk)}
        if (tirage=="REJECTIF") {POPpiij=UPmaxentropypi2(PIi) ; POPpikl=UPmaxentropypi2(PIk)}
        if (tirage=="SAMPFORD") {POPpiij=UPsampfordpi2(PIi) ; POPpikl=UPsampfordpi2(PIk)}
        for (i in 1:NBechMC)
                {sss=Sampling(tirage,PIk,PIi)
                ECHpikl=POPpikl[sss[[2]]==1,sss[[2]]==1]
                ECHpiij=POPpiij[sss[[1]]==1,sss[[1]]==1]
                Yik=Y[sss[[1]]==1,sss[[2]]==1]
                Est<-c(Est, EstTOTAL(Yik,ECHpikl,ECHpiij) )
                }
        return(var(Est))
        }


##############################################################################################
#DESCRIPION: simulations in Population for an estimated total and its variance estimators
#USAGE: SIMU(tirage,NBMC,NB,Y,PIk,PIi)
#ARGUMENTS:
#               tirage  MIDZUNO, SAMPFORD or REJECTIF
#               NBMC            number of simulations to calculate the MC variance
#               NB              number of simulations to calculate the estimated total and its
     variance estimators
#               Y       matrix of the interest variable, of size [N_d,N_m]
#               PIk     probabilities for the sample drawn in the population U_m
#               PIi     probabilities for the sample drawn in the population U_d
#VALUE: the function returns a list of values:
#the total population, the total estimator under NB simulations, the MC variance,
#the number of negative value of the 5 unbiased variance estimator under NB simulations, and of its
      parts
#the 5 unbiased variance estimator under NB simulations, its parts
#the variances of each variance estimators under NB estimators and of each part,
#the coefficient of variation of each variance estimator, and of each part,
#the boxplot of each variance estimator, and of each part.

SIMU <- function(tirage,NBMC,NB,Y,PIk,PIi)
        {x<-c() ; Tpi<-c() ; T0<-Sys.time() ; set.seed(43111)
        T=TOTALpopulation(Y)
        if (tirage=="MIDZUNO") {POPpiij=UPmidzunopi2(PIi) ; POPpikl=UPmidzunopi2(PIk)}
        if (tirage=="REJECTIF") {POPpiij=UPmaxentropypi2(PIi) ; POPpikl=UPmaxentropypi2(PIk)}
        if (tirage=="SAMPFORD") {POPpiij=UPsampfordpi2(PIi) ; POPpikl=UPsampfordpi2(PIk)}

        VMC=VARIANCE_MC(tirage,NBMC,Y,PIk,PIi)[1]
        for(t in 1:NB)
                {sss=Sampling(tirage,PIk,PIi)
                 ECHpikl=POPpikl[sss[[2]]==1,sss[[2]]==1]
                 ECHpiij=POPpiij[sss[[1]]==1,sss[[1]]==1]
                 Yik=Y[sss[[1]]==1,sss[[2]]==1]
                 a=varestYG(Yik,ECHpikl,ECHpiij)
                 x <- rbind(x,a)
                 tpi=EstTOTAL(Yik,ECHpikl,ECHpiij)
                 Tpi<- rbind(Tpi,tpi)
                 }
        meanvar=apply(x,2,mean)
        compteur=apply(x,2,function(y) sum(y<0))
        varvar=apply(x,2,var)

        CVvar=sqrt(varvar)/VMC #medianvar=apply(x,2,median)
        t=mean(Tpi)
        T1<-Sys.time() ; Tdiff1= difftime(T1, T0, units ="secs"); Tdiff1
        v=c(T,t,VMC,compteur,meanvar,varvar,Tdiff1,CVvar)
        windows() ; boxplot(x,ylim=range(boxplot(x,plot=FALSE)$stats),outline =FALSE,names=c("Va","
            Vb","Vc","Vd","Ve","V1","V2","V3","V4","V5","Vsp3","Vsp4","Vsp5"))
        return(v)
}


###############################################################################################
#DESCRIPION: fonction recursive qui dit pik=1 si pik>1 sinon pik=n (xk/sum(xk))
#USAGE: RECpps (n,M)
#ARGUMENTS:
#               n       sample size
#               M       matrix which contains the auxiliary variable X

#VALUE: the function returns a matrix which contains a column with the pik proportional to X

RECpps <- function(n,M)
        {M[,3]=0
        M[,3]=n*(M[,2]/sum(M[,2]))
        a=length(which(M[,3]>=1))
        if ((a==0)| (a==n) ) {M[ ,4] = M[,3] ; resultat=M}
        else {M[ which(M[,3]>=1),4 ]=1
                resultat <- rbind(M[ which(M[,4]==1), ], RECpps (n-a,M[ which(M[,4]!=1), ]))
                }
        return(resultat)
        }




###############################################################################################
#DESCRIPION: create pik proportional to X and simulations in Population
#USAGE: PIKetSIMU (tirage,n_m,n_d,NBMC,NB,Y)
#ARGUMENTS:
#               tirage  MIDZUNO, SAMPFORD or REJECTIF
#               n_m             sample size for the sample drawn in the population U_m
#               n_d             sample size for the sample drawn in the population U_d
#               NBMC            number of simulations to calculate the MC variance
#               NB              number of simulations to calculate the estimated total and its
    variance estimators
#               Y       matrix of the interest variable, of size [N_m,N_d]

#VALUE: the function returns a list of values obtained by the function SIMU

PIKetSIMU <- function(tirage,n_m,n_d,NBMC,NB,Y)
        {
        recA=RECpps(n_d,dataA) ; A=recA[order(recA[,1]),] ; pii=A[,4]
        recB=RECpps(n_m,dataB) ; B=recB[order(recB[,1]),] ; pik=B[,4]
        SIMU(tirage,NBMC,NB,Y,pik,pii)
        }
```

Code B.1 – CodeR_functions.r : Basic functions required to calculate estimators.

B.2 Simulations

```


###################
```

```
# Model for Y and X
####################

# Choose the maternity unit effect and the day effect:

sigma1=50 ; sigma2=50

N_m=100; N_j=100        ; N=N_m*N_j

mu=200 ; sigma3=5 ; sigma4=5 ; alpha=1 ; nu=200

set.seed(438)
POP1y=0 ; Yi=rnorm(N_m,0,1)*sigma1; Yil=rep(Yi,N_j); POP1yi=matrix(Yil,nrow=N_m); POP1yi=t(POP1yi)
Yk=rnorm(N_j,0,1)*sigma2; Yjk=rep(Yk,N_m); POP1yk=matrix(Yjk,nrow=N_j)
eps=rnorm(N,0,1); POP1eps=matrix(eps,nrow=N_j)  ; uik=rnorm(N,0,1); POP1uik=matrix(uik,nrow=N_j)
    ; vik=rnorm(N,0,1); POP1vik=matrix(vik,nrow=N_j)

POP1x = mu+POP1yi+POP1yk + sigma3*POP1uik
POP1y = nu+POP1x + sigma4*POP1eps


###############################
#                    Xi et Xk
###############################

piix=apply(POP1x,1,mean) ; pikx=apply(POP1x,2,mean)
dataA <-as.data.frame(matrix(ncol=4,nrow=N_j))  ;  dataA[,1]=1:N_j      ; dataA [,2]=piix        ;
    dataA [,4]=0
dataB <-as.data.frame(matrix(ncol=4,nrow=N_m))  ;  dataB[,1]=1:N_m      ; dataB [,2]=pikx        ;
    dataB [,4]=0


##################################
#Simulations (choose sample sizes)
##################################

res1=PIKetSIMU(tirage="MIDZUNO",n_m=5,n_d=5,NBMC=50000,NB=10000,Y=POP1y) ; res1
res2=PIKetSIMU(tirage="SAMPFORD",n_m=10,n_d=10,NBMC=50000,NB=10000,Y=POP1y) ; res2
res3=PIKetSIMU(tirage="REJECTIF",n_m=20,n_d=20,NBMC=50000,NB=10000,Y=POP1y) ; res3
aa=rbind(res1,res2,res3)
colnames(aa)=c("Total","EstT","V_{MC}","#NEGa","#NEGb","#NEGc","#NEGd","#NEGe","#NEG1","#NEG2","#
    NEG3","#NEG4","#NEG5","#NEGsp3","#NEGsp4","#NEGsp5",
"E_{MC}(Va)","E_{MC}(Vb)","E_{MC}(Vc)","E_{MC}(Vd)","E_{MC}(Ve)","E_{MC}(V1)","E_{MC}(V2)","E_{MC}(
    V3)","E_{MC}(V4)","E_{MC}(V5)","E_{MC}(Vsp3)","E_{MC}(Vsp4)","E_{MC}(Vsp5)",
"V_{MC}(Va)","V_{MC}(Vb)","V_{MC}(Vc)","V_{MC}(Vd)","V_{MC}(Ve)","V_{MC}(V1)","V_{MC}(V2)","V_{MC}(
    V3)","V_{MC}(V4)","V_{MC}(V5)","V_{MC}(Vsp3)","V_{MC}(Vsp4)","V_{MC}(Vsp5)" ,
"temps","CVa","CVb","CVc","CVd","CVe","CV1","CV2","CV3","CV4","CV5","CVsp3","CVsp4","CVsp5")
aa
```

Code B.2 – Generation of the interest and auxiliary variables and results of simulations.

Annexe C

Matériel supplémentaire relatif au Chapitre 4

Dans la première annexe se trouvent les fonctions nécessaires au calcul des estimateurs. Dans la seconde annexe, ce sont les commandes pour faire tourner ces fonctions et afficher les résultats du Tableau 4.2 (pour cela il faut avoir accès à un fichier au format csv contenant les données, qui sera disponible sur le site CSBIGS [1]).

C.1 Fonctions

```
#R FUNCTIONS REQUIRED FOR the simulation table
#R functions documented:
#DRAWsi
#DRAWsisi2d
#DRAWsisiccs
#ESTratio
#ESTtotal
#ESTVARlinratio2d
#ESTVARlinratioccs
#ESTVARtotal2d
#ESTVARtotalccs
#LINratio
#MEASURESratio2d
#MEASURESratioccs
#POPratio
#POPtotal
#SAMPLEsisiccs
#SIMULATIONratio2d
#SIMULATIONratioccs
#VARIANCE_MCratio2d
#VARIANCE_MCratioccs


#Note replicability: the seed was fixed using the command set.seed() in the
```

[1] http://publications-sfds.fr/index.php/csbigs

```
#functions SIMULATIONratio2d, SIMULATIONratioccs, VARIANCE_MCratio2d and VARIANCE_MCratioccs.


###############################################################################################
#DESCRIPION: draws (and fixes) simple sample (equal probabilities) without replacement
#USAGE: DRAWsi(n,N)
#ARGUMENTS:
#               n        size of the sample
#               N        size of the population
#VALUE: the function returns a vector of lenght n

DRAWsi <- function(n,N)
        { s=c(); s=sample(1:N,n); return(s) }


###############################################################################################
#DESCRIPION: draws (and fixes) two-stage sample (equal probabilities at the two degrees) without
    replacement and with same sample size in each selected PSU
#USAGE: DRAWsisi2d(n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               n_m      size of the sample drawn in the population U_m
#               n_d      size of the sample drawn in the population U_d
#               N_m      size of the population U_m
#               N_d      size of the population U_d
#VALUE: the function returns a vector of lenght n_m*n_d

DRAWsisi2d <- function(n_m,n_d,N_m,N_d)
        {
        Vs_d <- c()
        s_m=sort(DRAWsi(n_m,N_m))
        for (i in 1:n_m)
                {
                s_d=sort(DRAWsi(n_d,N_d)) + N_d*(s_m[i]-1)
                Vs_d=c(Vs_d,s_d)
                }
        Dummy=rep(0,N_m*N_d) ; Dummy[Vs_d] <-1 ; return(Dummy)
        }


###############################################################################################
#DESCRIPION: draws (and fixes) two simple samples (equal probabilities) without replacement in two
    populations
#USAGE: DRAWsisiccs(n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               n_m      size of the sample drawn in the population U_m
#               n_d      size of the sample drawn in the population U_d
#               N_m      size of the population U_m
#               N_d      size of the population U_d
#VALUE: the function returns a list of two vectors of respective lenghts n_m and n_d

DRAWsisiccs <- function(n_m,n_d,N_m,N_d)
        {
        sd=DRAWsi(n_d,N_d); sm=DRAWsi(n_m,N_m)
        S=list(sm,sd) ; return(S)
        }


###############################################################################################
#DESCRIPION: estimates a population total
#USAGE: ESTtotal(ECH,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               ECH      matrix of the interest variable, of size [n_d,n_m]
#                n_m     size of the sample drawn in the population U_m
#                n_d     size of the sample drawn in the population U_d
#                N_m     size of the population U_m
#                N_d     size of the population U_d
#VALUE: the function returns a value for the estimated total

ESTtotal <- function(ECH,n_m,n_d,N_m,N_d)
        {w=(N_m/n_m)*(N_d/n_d); Tpi=sum(w*c(ECH)) ; return(Tpi) }


#################################################################################################
#DESCRIPION: estimates a population ratio Y/X
#USAGE: ESTratio(ECHY,ECHX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#                ECHY    matrix of the numerator variable, of size [n_d,n_m]
#                ECHX    matrix of the denominator variable, of size [n_d,n_m]
#                n_m     size of the sample drawn in the population U_m
#                n_d     size of the sample drawn in the population U_d
#                N_m     size of the population U_m
#                N_d     size of the population U_d
#VALUE: the function returns a value for the estimated ratio

ESTratio <- function(ECHY,ECHX,n_m,n_d,N_m,N_d)
        {Rpi=ESTtotal(ECHY,n_m,n_d,N_m,N_d)/ESTtotal(ECHX,n_m,n_d,N_m,N_d) ; return(Rpi) }


#################################################################################################
#DESCRIPION: estimates the variance of the estimated total of the estimated linearized variable of
     a ratio Y/X
#USAGE: ESTVARlinratio2d(ECHY,ECHX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#                ECHY    matrix of the numerator variable, of size [n_d,n_m]
#                ECHX    matrix of the denominator variable, of size [n_d,n_m]
#                n_m     size of the sample drawn in the population U_m
#                n_d     size of the sample drawn in the population U_d
#                N_m     size of the population U_m
#                N_d     size of the population U_d
#VALUE: the function returns a list of two values (the unbiased estimator V2d and its
     approximations V2d,a)

ESTVARlinratio2d <- function(ECHY,ECHX,n_m,n_d,N_m,N_d)
        {
        Rlin=LINratio(ECHY,ECHX,n_m,n_d,N_m,N_d)
        vl=ESTVARtotal2d(Rlin,n_m,n_d,N_m,N_d) ; return(vl)
        }


#################################################################################################
#DESCRIPION: estimates the variance of the estimated total of the estimated linearized variable of
     a ratio Y/X
#USAGE: ESTVARlinratioccs(ECHY,ECHX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#                ECHY    matrix of the numerator variable, of size [n_d,n_m]
#                ECHX    matrix of the denominator variable, of size [n_d,n_m]
#                n_m     size of the sample drawn in the population U_m
#                n_d     size of the sample drawn in the population U_d
#                N_m     size of the population U_m
#                N_d     size of the population U_d
#VALUE: the function returns a list of three values (the unbiased estimator Vccs and its
     approximations V2d and V2d,a)

ESTVARlinratioccs <- function(ECHY,ECHX,n_m,n_d,N_m,N_d)
```

```
        {
        Rlin=LINratio(ECHY,ECHX,n_m,n_d,N_m,N_d)
        vl=ESTVARtotalccs(Rlin,n_m,n_d,N_m,N_d) ; return(vl)
        }


####################################################################################################
#DESCRIPION: estimates the variance of an estimated total for a two-stage sampling {SI,SI}
#USAGE: ESTVARtotal2d(ECH,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               ECH     matrix of the interest variable, of size [n_d,n_m]
#               n_m     size of the sample drawn in the population U_m
#               n_d     size of the sample drawn in the population U_d
#               N_m     size of the population U_m
#               N_d     size of the population U_d
#VALUE: the function returns a list of two values (the unbiased estimator V2d and its approximation
      V2d,a)

ESTVARtotal2d <- function(ECH,n_m,n_d,N_m,N_d)
        {
        w_m=N_m/n_m;  w_d=N_d/n_d
        VarM=var(apply(ECH*w_d,2,sum));    V2da=(N_m^2)*(1/n_m - 1/N_m)*VarM
        Vard=(N_d^2)*(1/n_d - 1/N_d)*apply(ECH,2,var);  V2db=w_m*sum(Vard)
        V2d=V2da+V2db
        v=list(V2d,V2da) ; return(v)
        }


####################################################################################################
#DESCRIPION: estimates the variance of an estimated total for a CCS SIxSI
#USAGE: ESTVARtotalccs(ECH,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               ECH     matrix of the interest variable, of size [n_d,n_m]
#               n_m     size of the sample drawn in the population U_m
#               n_d     size of the sample drawn in the population U_d
#               N_m     size of the population U_m
#               N_d     size of the population U_d
#VALUE: the function returns a list of three values (the unbiased estimator Vccs and its
     approximations V2d and V2d,a)

ESTVARtotalccs <- function(ECH,n_m,n_d,N_m,N_d)
        {
        w_m=N_m/n_m;  w_d=N_d/n_d
        VarM=var(apply(ECH*w_d,2,sum));    V2da=(N_m^2)*(1/n_m - 1/N_m)*VarM
        VarD=var(apply(ECH*w_m,1,sum));    Vsimp=(N_d^2)*(1/n_d - 1/N_d)*VarD
        d=apply(ECH,1,mean);  e=apply(ECH,2,mean);  f=mean(ECH)
        X = t(t(ECH - d) - e) + f;  X2=X*X
        Vco=(1/(n_d - 1))*(1/(n_m - 1))*sum(X2)
        VCORR=(N_d^2)*(1/n_d - 1/N_d)*(N_m^2)*(1/n_m - 1/N_m)*Vco
        VCCS=V2da+Vsimp-VCORR
        Vard=(N_d^2)*(1/n_d - 1/N_d)*apply(ECH,2,var);  V2db=w_m*sum(Vard)
        V2d=V2da+V2db
        v=list(VCCS,V2d,V2da) ; return(v)
        }


####################################################################################################
#DESCRIPION: estimates the linearized variable of a ratio Y/X
#USAGE: LINratio(ECHY,ECHX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               ECHY    matrix of the numerator variable, of size [n_d,n_m]
#               ECHX    matrix of the denominator variable, of size [n_d,n_m]
#               n_m     size of the sample drawn in the population U_m
#               n_d     size of the sample drawn in the population U_d
#               N_m     size of the population U_m
#               N_d     size of the population U_d
#VALUE: the function returns a matrix of length [n_d,n_m]

LINratio <- function(ECHY,ECHX,n_m,n_d,N_m,N_d)
        {
        rpi=ESTratio(ECHY,ECHX,n_m,n_d,N_m,N_d)
        txpi=ESTtotal(ECHX,n_m,n_d,N_m,N_d)
        upi = (ECHY - ECHX*rpi) / txpi ; return(upi)
        }


##############################################################################################
#DESCRIPION: measures (with simulations in Population) the relative biases of an estimated ratio Y/
    X and of its variance estimators for two-stage sampling
#USAGE: MEASURESratio2d(NBech,NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               NBech           number of simulations to calculate the estimated ratio and its
    variance estimators
#               NBechMC number of simulations to calculate the 'true' variance
#               POPY    matrix of the numerator variable, of size [N_d,N_m]
#               POPX    matrix of the denominator variable, of size [N_d,N_m]
#               n_m     size of the sample drawn in the population U_m
#               n_d     size of the sample drawn in the population U_d
#               N_m     size of the population U_m
#               N_d     size of the population U_d
#VALUE: the function returns a list of 12 values: the sizes n_m, N_m, n_d, N_d,
#the ratio population, its estimate under NBech simulations, its relative bias,
#the variance monteCarlo under NBechMC simulation, the unbiased estimated variance under NBech
    simulations, its relative bias,
#the simplified variance estimator V2d,a under NBech simulations, its relative bias

MEASURESratio2d <- function(NBech,NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
        {
        esp=SIMULATIONratio2d(NBech,POPY,POPX,n_m,n_d,N_m,N_d)
        varmc=VARIANCE_MCratio2d(NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
        t=POPratio(POPY,POPX)
        RBratio=((esp[[1]]-t)/t) *100
        RBvar=((esp[[2]]-varmc)/varmc) *100
        RBvar1=((esp[[3]]-varmc)/varmc) *100
        m=list(n_m,N_m,n_d,N_d,t,esp[[1]],RBratio,varmc,esp[[2]],RBvar,esp[[3]],RBvar1)
        return(m)
        }


##############################################################################################
#DESCRIPION: measures (with simulations in Population) the relative biases of an estimated ratio Y/
    X and of its variance estimators for CCS
#USAGE: MEASURESratioccs(NBech,NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               NBech           number of simulations to calculate the estimated ratio and its
    variance estimators
#               NBechMC number of simulations to calculate the 'true' variance
#               POPY    matrix of the numerator variable, of size [N_d,N_m]
#               POPX    matrix of the denominator variable, of size [N_d,N_m]
#               n_m     size of the sample drawn in the population U_m
#               n_d     size of the sample drawn in the population U_d
#               N_m     size of the population U_m
#               N_d     size of the population U_d
#VALUE: the function returns a list of 12 values: the sizes n_m, N_m, n_d, N_d,
#the ratio population, its estimate under NBech simulations, its relative bias,
#the variance monteCarlo under NBechMC simulation, the unbiased estimated variance under NBech
    simulations, its relative bias,
#the simplified variance estimator V2d under NBech simulations, its relative bias,
#the simplified variance estimator V2d,a under NBech simulations, its relative bias

MEASURESratioccs <- function(NBech,NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
       {
       esp=SIMULATIONratioccs(NBech,POPY,POPX,n_m,n_d,N_m,N_d)
       varmc=VARIANCE_MCratioccs(NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
       t=POPratio(POPY,POPX)
       RBratio=((esp[[1]]-t)/t) *100
       RBvar=((esp[[2]]-varmc)/varmc) *100
       RBvar1=((esp[[3]]-varmc)/varmc) *100;RBvar2=((esp[[4]]-varmc)/varmc) *100
       m=list(n_m,N_m,n_d,N_d,t,esp[[1]],RBratio,varmc,esp[[2]],RBvar,esp[[3]],RBvar1,esp[[4]],
           RBvar2)
       return(m)
       }


###############################################################################################
#DESCRIPION: calculates the population ratio Y/X
#USAGE: POPratio(POPY,POPX)
#ARGUMENTS:
#             POPY     matrix of the numerator variable
#             POPX     matrix of the denominator variable
#VALUE: the function returns a value for the population ratio

POPratio <- function(POPY,POPX)
       { return(sum(POPY)/sum(POPX)) }


###############################################################################################
#DESCRIPION: calculates the population total
#USAGE: POPtotal(POP)
#ARGUMENTS:
#             POP      matrix of the interest variable
#VALUE: the function returns a value for the population total

POPtotal <- function(POP)
       { return(sum(POP)) }


###############################################################################################
#DESCRIPION: calculates the population ratio Y/X
#USAGE: SAMPLEsisiccs(POP, s_m, s_d)
#ARGUMENTS:
#             POP      matrix of the interest variable
#             s_m      vector of lenght<dim(POP)[2]
#             s_d      vector of lenght<dim(POP)[1]
#VALUE: the function returns a matrix of length [length(s_m),length(s_d)]

SAMPLEsisiccs <- function(POP, s_m, s_d)
       {s=POP[s_d,s_m] ; return(s) }


###############################################################################################
#DESCRIPION: simulations in Population for an estimated ratio Y/X and its variance estimators for a
     two-stage sampling {SI,SI}
#USAGE: SIMULATIONratio2d(NBech,POPY,POPX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
```

```
#               NBech           number of simulations to calculate the estimated ratio and its
     variance estimators
#               POPY     matrix of the numerator variable, of size [N_d,N_m]
#               POPX     matrix of the denominator variable, of size [N_d,N_m]
#               n_m      size of the sample drawn in the population U_m
#               n_d      size of the sample drawn in the population U_d
#               N_m      size of the population U_m
#               N_d      size of the population U_d
#VALUE: the function returns a list of 3 values:
#the ratio estimator under NBech simulations, its unbiased variance estimator under NBech
     simulations,
#the simplified variance estimator V2d,a under NBech simulations

SIMULATIONratio2d <- function(NBech,POPY,POPX,n_m,n_d,N_m,N_d)
        {
        Est<-c(); EstVar<-c(); EstVar1<-c(); set.seed(12345)
        for (i in 1:NBech)
                {
                echY<-c(); echX<-c(); Dummy_2=DRAWsisi2d(n_m,n_d,N_m,N_d)
                echY=matrix(POPY[Dummy_2==1],nrow=n_d)
                echX=matrix(POPX[Dummy_2==1],nrow=n_d)
                Est<-c(Est, ESTratio(echY,echX,n_m,n_d,N_m,N_d))
                v=ESTVARlinratio2d (echY,echX,n_m,n_d,N_m,N_d)
                EstVar<-c(EstVar,v[[1]])
                EstVar1<-c(EstVar1,v[[2]])
                }
        ESPest=mean(Est);ESPvar=mean(EstVar);ESPvar1=mean(EstVar1)
        E=list(ESPest,ESPvar,ESPvar1) ; return(E)
        }


##################################################################################################
#DESCRIPION: simulations in Population for an estimated ratio Y/X and its variance estimators for a
      CCS SIxSI
#USAGE: SIMULATIONratioccs(NBech,POPY,POPX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               NBech           number of simulations to calculate the estimated ratio and its
     variance estimators
#               POPY     matrix of the numerator variable, of size [N_d,N_m]
#               POPX     matrix of the denominator variable, of size [N_d,N_m]
#               n_m      size of the sample drawn in the population U_m
#               n_d      size of the sample drawn in the population U_d
#               N_m      size of the population U_m
#               N_d      size of the population U_d
#VALUE: the function returns a list of 4 values:
#the ratio estimator under NBech simulations, its unbiased variance estimator under NBech
     simulations,
#the simplified variance estimator V2d under NBech simulations
#the simplified variance estimator V2d,a under NBech simulations

SIMULATIONratioccs <- function(NBech,POPY,POPX,n_m,n_d,N_m,N_d)
        {
        Est<-c(); EstVar<-c(); EstVar1<-c(); EstVar2<-c() ; set.seed(1235)
        for (i in 1:NBech)
                {
                s=DRAWsisiccs(n_m,n_d,N_m,N_d)
                POPYmat=matrix(POPY,nrow=N_d)
                POPXmat=matrix(POPX,nrow=N_d)
                echY=SAMPLEsisiccs(POPYmat,s[[1]],s[[2]])
                echX=SAMPLEsisiccs(POPXmat,s[[1]],s[[2]])
                Est<-c(Est, ESTratio(echY,echX,n_m,n_d,N_m,N_d))
                v=ESTVARlinratioccs (echY,echX,n_m,n_d,N_m,N_d)
                EstVar<-c(EstVar,v[[1]])
                EstVar1<-c(EstVar1,v[[2]]); EstVar2<-c(EstVar2,v[[3]])
                }
        ESPest=mean(Est);ESPvar=mean(EstVar);ESPvar1=mean(EstVar1);ESPvar2=mean(EstVar2)
        E=list(ESPest,ESPvar,ESPvar1,ESPvar2) ; return(E)
        }


##########################################################################################
#DESCRIPION: calculates independently the MonteCarlo variance of an estimated ratio Y/X for a two-
    stage sampling {SI,SI}
#USAGE: VARIANCE_MCratio2d(NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               NBechMC         number of simulations to calculate the MC variance
#               POPY    matrix of the numerator variable, of size [N_d,N_m]
#               POPX    matrix of the denominator variable, of size [N_d,N_m]
#               n_m     size of the sample drawn in the population U_m
#               n_d     size of the sample drawn in the population U_d
#               N_m     size of the population U_m
#               N_d     size of the population U_d
#VALUE: the function returns a value for the MC variance

VARIANCE_MCratio2d <- function(NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
        {
        Est<-c() ; set.seed(13456)
        for (i in 1:NBechMC)
                {
                echY<-c(); echX<-c() ; Dummy_2=DRAWsisi2d(n_m,n_d,N_m,N_d)
                echY=matrix(POPY[Dummy_2==1],nrow=n_d)
                echX=matrix(POPX[Dummy_2==1],nrow=n_d)
                Est<-c(Est, ESTratio(echY,echX,n_m,n_d,N_m,N_d))
                }
        return(var(Est))
        }


##########################################################################################
#DESCRIPION: calculates independently the MonteCarlo variance of an estimated ratio Y/X for CCS
    SIxSI
#USAGE: VARIANCE_MCratioccs(NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
#ARGUMENTS:
#               NBechMC         number of simulations to calculate the MC variance
#               POPY    matrix of the numerator variable, of size [N_d,N_m]
#               POPX    matrix of the denominator variable, of size [N_d,N_m]
#               n_m     size of the sample drawn in the population U_m
#               n_d     size of the sample drawn in the population U_d
#               N_m     size of the population U_m
#               N_d     size of the population U_d
#VALUE: the function returns a value for the MC variance

VARIANCE_MCratioccs <- function(NBechMC,POPY,POPX,n_m,n_d,N_m,N_d)
        {
        Est<-c() ; set.seed(1256)
        for (i in 1:NBechMC)
                {
                s=DRAWsisiccs(n_m,n_d,N_m,N_d)
                POPYmat=matrix(POPY,nrow=N_d)
                POPXmat=matrix(POPX,nrow=N_d)
                echY=SAMPLEsisiccs(POPYmat,s[[1]],s[[2]])
                echX=SAMPLEsisiccs(POPXmat,s[[1]],s[[2]])
                Est<-c(Est, ESTratio(echY,echX,n_m,n_d,N_m,N_d))
                }
        return(var(Est))
        }
```

Code C.1 – CodeR_simu_functions.r : Basic functions required to calculate estimators.

C.2 Simulations

```
############## STEP 1
#Compile all the code in "CodeR_simu_functions.r"
#(this is the required functions for the simulations in the article)
#For this, you can choose one of the two following commands:
source(".../CodeR_simu_functions.r")
#or
source(file.choose())

############## STEP 2: DATA
#Read the table "Data2stCCS.csv"
#For this, you can choose one of the two following commands:
tableR=read.csv2(".../Data2stCCS.csv")
#or
tableR=read.csv2(file.choose())

attach(tableR)
N_m=544 ; N_d=365

############## STEP 3: SIMULATIONS

B=10000 ; C=50000

################ STEP 3 case 1: tY/tX

################## STEP 3 case 1: tY/tX 2d

a1=MEASURESratio2d(B,C,POPY=Yik,POPX=Xik,n_m=5,n_d=5,N_m,N_d)
b1=MEASURESratio2d(B,C,POPY=Yik,POPX=Xik,n_m=25,n_d=25,N_m,N_d)
c1=MEASURESratio2d(B,C,POPY=Yik,POPX=Xik,n_m=320,n_d=25,N_m,N_d)
d1=MEASURESratio2d(B,C,POPY=Yik,POPX=Xik,n_m=25,n_d=320,N_m,N_d)
e1=MEASURESratio2d(B,C,POPY=Yik,POPX=Xik,n_m=320,n_d=320,N_m,N_d)

result2d_1=cbind(a1,b1,c1,d1,e1)
rownames(result2d_1)=c("$n_M$","$N_M$","$n_D$","$N_D$"
,"$ParameterPOP$","$E(Parameter)$","$mbox{RB}_{MC}left(hat{Parameter}right)$"
,"$Var_{MC}$","$E(hat{V}_{2d})$","$mbox{RB}_{MC}left(hat{V}_{2d}right)$"
,"$E(hat{V}_{2d,a})$","$mbox{RB}_{MC}left(hat{V}_{2d,a}right)$")
result2d_1

################## STEP 3 case 1: tY/tX ccs

f1=MEASURESratioccs(B,C,POPY=Yik,POPX=Xik,n_m=5,n_d=5,N_m,N_d)
g1=MEASURESratioccs(B,C,POPY=Yik,POPX=Xik,n_m=25,n_d=25,N_m,N_d)
h1=MEASURESratioccs(B,C,POPY=Yik,POPX=Xik,n_m=320,n_d=25,N_m,N_d)
i1=MEASURESratioccs(B,C,POPY=Yik,POPX=Xik,n_m=25,n_d=320,N_m,N_d)
j1=MEASURESratioccs(B,C,POPY=Yik,POPX=Xik,n_m=320,n_d=320,N_m,N_d)

resultccs_1=cbind(f1,g1,h1,i1,j1)
```

```
rownames(resultccs_1)=c("$n_M$","$N_M$","$n_D$","$N_D$"
,"$ParameterPOP$","$E(Parameter)$","$mbox{RB}_{MC} left(hat{Parameter} right)$"
,"$Var_{MC}$","$E(hat{V}_{CCS})$","$mbox{RB}_{MC} left(hat{V}_{CCS} right)$"
,"$E(hat{V}_{2d})$","$mbox{RB}_{MC} left(hat{V}_{2d} right)$"
,"$E(hat{V}_{2d,a})$","$mbox{RB}_{MC} left(hat{V}_{2d,a} right)$")
resultccs_1

################ STEP 3 case 2: tZ/tX

################## STEP 3 case 2: tZ/tX 2d

a2=MEASURESratio2d(B,C,POPY=Zik,POPX=Xik,n_m=5,n_d=5,N_m,N_d)
b2=MEASURESratio2d(B,C,POPY=Zik,POPX=Xik,n_m=25,n_d=25,N_m,N_d)
c2=MEASURESratio2d(B,C,POPY=Zik,POPX=Xik,n_m=320,n_d=25,N_m,N_d)
d2=MEASURESratio2d(B,C,POPY=Zik,POPX=Xik,n_m=25,n_d=320,N_m,N_d)
e2=MEASURESratio2d(B,C,POPY=Zik,POPX=Xik,n_m=320,n_d=320,N_m,N_d)

result2d_2=cbind(a2,b2,c2,d2,e2)
rownames(result2d_2)=c("$n_M$","$N_M$","$n_D$","$N_D$"
,"$ParameterPOP$","$E(Parameter)$","$mbox{RB}_{MC} left(hat{Parameter} right)$"
,"$Var_{MC}$","$E(hat{V}_{2d})$","$mbox{RB}_{MC} left(hat{V}_{2d} right)$"
,"$E(hat{V}_{2d,a})$","$mbox{RB}_{MC} left(hat{V}_{2d,a} right)$")
result2d_2

################## STEP 3 case 2: tZ/tX ccs

f2=MEASURESratioccs(B,C,POPY=Zik,POPX=Xik,n_m=5,n_d=5,N_m,N_d)
g2=MEASURESratioccs(B,C,POPY=Zik,POPX=Xik,n_m=25,n_d=25,N_m,N_d)
h2=MEASURESratioccs(B,C,POPY=Zik,POPX=Xik,n_m=320,n_d=25,N_m,N_d)
i2=MEASURESratioccs(B,C,POPY=Zik,POPX=Xik,n_m=25,n_d=320,N_m,N_d)
j2=MEASURESratioccs(B,C,POPY=Zik,POPX=Xik,n_m=320,n_d=320,N_m,N_d)

resultccs_2=cbind(f2,g2,h2,i2,j2)
rownames(resultccs_2)=c("$n_M$","$N_M$","$n_D$","$N_D$"
,"$ParameterPOP$","$E(Parameter)$","$mbox{RB}_{MC} left(hat{Parameter} right)$"
,"$Var_{MC}$","$E(hat{V}_{CCS})$","$mbox{RB}_{MC} left(hat{V}_{CCS} right)$"
,"$E(hat{V}_{2d})$","$mbox{RB}_{MC} left(hat{V}_{2d} right)$"
,"$E(hat{V}_{2d,a})$","$mbox{RB}_{MC} left(hat{V}_{2d,a} right)$")
resultccs_2
```

Code C.2 – CodeR : Results of simulations.

Annexe D

Matériel supplémentaire relatif au Chapitre 5

Dans la première annexe se trouvent les fonctions nécessaires au calcul des estimateurs. Dans la seconde annexe, ce sont les commandes pour faire tourner ces fonctions et afficher les résultats des Tableaux 5.1 et 5.2.

D.1 Fonctions

```
#R FUNCTIONS REQUIRED FOR tables 1 and 2

#R functions documented:

#CALAGE
#CalculSAMPLE
#DUMMYnonreponse
#DUMMYsample
#EstProbareponseLOGIT
#EstTOTAL
#EstTOTALcal
#EstVAR
#FinalSAMPLE
#FinalSAMPLEdiff
#LINratio
#POP
#SIMULATIONS
#VARIANCE_MC
#VARIANCE_MCnr
#VARIANCE_MC_VtVu


#Note replicability: the seed was fixed using the command set.seed() in the
#functions POP, SIMULATIONS and VARIANCE_MC.

# To change calibration variables, change a row in Finalsample and a row in CALAGE
```

```
library(MASS)
library(sampling)

##########################################################################################
#DESCRIPION: generates the interest variables presented in Table 2
#USAGE: POP(N,rho,sigma,alpha,alphaA,alphaB,beta1,beta2,beta3)
#ARGUMENTS:
#           N       population size
#           rho,sigma,alpha,alphaA,alphaB,beta1,beta2,beta3        parameter of model

#VALUE: the function returns a data.frame [N,12]

POP <- function(N,rho,sigma,alpha,alphaA,alphaB,beta1,beta2,beta3)
      {
      set.seed(124)
      x1=rgamma(N, shape= 2) ; x2=rgamma(N, shape= 2)
      x3=rgamma(N, shape= 2) ; x4=rgamma(N, shape= 2)
      u1=rnorm(N,0,1) ; u2=rnorm(N,0,1); u3=rnorm(N,0,1)
      y1 = alpha + alphaA*x1 + alphaB*x2 + sigma*u1
      y2 = rho*y1 + sigma*u2
      y3 = rho*y2 + sigma*u3
      p1=exp(-1+beta1*x1+beta1*x2)/(1+exp(-1+beta1*x1+beta1*x2))
      p2=exp(-1+beta2*x1+beta2*x2)/(1+exp(-1+beta2*x1+beta2*x2))
      p3=exp(-1+beta3*x1+beta3*x2)/(1+exp(-1+beta3*x1+beta3*x2))
      id=c(1:N); cste=rep(-1,N)
      dataf=data.frame(id,x1,x2,y1,y2,y3,p1,p2,p3,cste,x3,x4)
      names(dataf) = c("ID","X1","X2","Y1","Y2","Y3","prob1","prob2","prob3","Cte","X3","X4")
      return(dataf)
      }


##########################################################################################
#DESCRIPION: draws SI simple sample (equal probabilities) without replacement
#USAGE: DUMMYsample (POP,n)
#ARGUMENTS:
#           POP     data frame
#           n       size of the sample drawn in the population

#VALUE: the function returns a list: the dataframe with the dummy sample, the value of the second
    order probabilities (SI) and the covariance

DUMMYsample <- function(POP,n)
            {N=nrow(POP) ; piij=(n*(n-1))/(N*(N-1)) ; pii=n/N ; Is<-c()
            deltaij=piij-pii^2
            Is=srswor(n,N) ; POPech = cbind.data.frame(POP,Is)
            l=list(POPech,piij,deltaij)
            return(l)
            }

##########################################################################################
#DESCRIPION: draws Poisson sampling PO(p)
#USAGE: DUMMYnonreponse (dds,p)
#ARGUMENTS:
#           dfs     data frame
#           p       response probabilities

#VALUE: the function returns the dataframe with the dummy response sample

DUMMYnonreponse <- function(dfs,p)
            {IdR = runif(nrow(dfs))
            for (i in 1:length(IdR)) {if (IdR[i] > p[i]) IdR[i] = 0 else IdR[i] = 1}
            dfsIr = cbind.data.frame(dfs, IdR)
```

```
                names(dfsIr)[names(dfsIr )=="IdR"] <- "Id"
                return(dfsIr)
                }

################################################################################################
#DESCRIPION: estimates response probabilities using logistic regression
#USAGE: EstProbareponseLOGIT (tab)
#ARGUMENTS:
#               tab     data frame

#VALUE: the function returns the dataframe with the estimate response probabilities

EstProbareponseLOGIT <- function(tab)
                {modelogit=glm(Id ~ X1+X2,family=binomial(link = "logit"),data=tab)
                ee=modelogit$fitted.value ; tabP = cbind.data.frame(tab, ee)
                return(tabP)
                }

################################################################################################
#DESCRIPION: draw SI sample, respondent sample, calculate weights with and without calibration,
    linearized variable for ratio and estimated residual for the regression (calibration)
#USAGE: FinalSAMPLE (POP,n,temps,ratio)
#ARGUMENTS:
#               POP     data frame
#               n       sample size to draw a SI sample
#               temps   number of non-response phases (1, 2 or 3)
#               ratio   1 for a ratio parameter, else it is a total
#               param           1 only to calculate MC variance nonresponse

#VALUE: the function returns a list: a vector of probabilities (sampling and non-response, PIij the
      second order probabilities, DELTAij the covariance,
#Ys the variable which will be used in the variance function, g the calibration factor, e the
     estimated residual for the weighted regression of y on x (calibration variables),
#w the final weights before calibration, XX the variables used for non-response, Y2 the numerator
     for the ratio, Y1 the denominator for the ratio.

FinalSAMPLE <- function(POP,n,temps,ratio,param)
                {
                s0<-c() ; prob<-c()
                D=DUMMYsample(POP,n)
                step0=D[[1]]
                s0=step0[step0$Is==1,] ; pi0=rep(n/nrow(POP),n) ; PIij=D[[2]] ; DELTAij=D[[3]]
                s=cbind.data.frame(s0,pi0)
                w=1/(s$pi0) ; prob<-cbind(prob,pi0)

        if (param==1)
                {
                s=POP
                }

        if (temps==1 | temps==2 | temps==3)
                {
                step1=DUMMYnonreponse(s,s$prob1)
                step1P=EstProbareponseLOGIT(step1)
                names(step1P)[names(step1P)=="Id"] <- "I1" ; names(step1P)[names(step1P)=="ee"] <-
                     "Estp1"
                s=step1P[step1P$I1==1,]
                w=1/(s$pi0*s$Estp1) ; prob<-cbind(s$pi0,s$Estp1)
                }

        if (temps==2 | temps==3)
                {
```

```
                step2=DUMMYnonreponse(s, s$prob2)
                step2P=EstProbareponseLOGIT(step2)
                names(step2P)[names(step2P)=="Id"] <- "I2" ; names(step2P)[names(step2P)=="ee"] <-
                    "Estp2"
                s=step2P[step2P$I2==1,]
                w=1/(s$pi0*s$Estp1*s$Estp2) ; prob<-cbind(s$pi0, s$Estp1, s$Estp2)
                }

        if (temps==3)
                {
                step3=DUMMYnonreponse(s, s$prob3)
                step3P=EstProbareponseLOGIT(step3)
                names(step3P)[names(step3P)=="Id"] <- "I3" ; names(step3P)[names(step3P)=="ee"] <-
                    "Estp3"
                s=step3P[step3P$I3==1,]
                w=1/(s$pi0*s$Estp1*s$Estp2*s$Estp3) ; prob<-cbind(s$pi0, s$Estp1, s$Estp2, s$Estp3)
                }

        #Xs=cbind(s$Cte, s$X3, s$X4)        #use this row to change variables for calibration
        Xs=cbind(s$Cte, s$X1, s$X2)
        g=CALAGE(Xs, temps=temps,w)
        Ys=s$Y1 ; Y1=s$Y1 ; Y2=s$Y2
        if (temps==3) {Y2=s$Y3} #pour le ratio
        if (ratio==1) {Ys=LINratio(w,Y1,Y2)     } #CONDITION RATIO

        q=rep(1,length(w)) ; ww=(g*w)*q ; b=t(Xs*ww) ; beta=ginv(b%*%Xs)%*%b%*%Ys
        e=Ys-Xs%*%beta #regression de Y1 sur X pondere par dk
        XX=cbind(s$Cte, s$X1, s$X2)        #for the non-response variance

                l=list(prob, PIij, DELTAij, Ys, g, e,w,XX, Y2, Y1)
                return(l)
                }

###################################################################################################
#DESCRIPION: calculate estimated linearized variable of ratio Y2/Y1
#USAGE: LINratio (w,Y1,Y2)
#ARGUMENTS:
#               w       final weight
#               Y1      denominator
#               Y2      numerator

#VALUE: the function returns a vector

LINratio <- function(w,Y1,Y2)
        {
        rpi=EstTOTAL(w,Y2)/EstTOTAL(w,Y1)
        txpi=EstTOTAL(w,Y1) ; upi = (Y2 - Y1*rpi) / txpi ; return(upi)
        }


###################################################################################################
#DESCRIPION: calculate estimated total
#USAGE: EstTOTAL (w,Y1)
#ARGUMENTS:
#               w       final weight without calibration
#               Y1      interest variable

#VALUE: the function returns a numeric

EstTOTAL <- function(w,Y1)
                {t=sum(Y1*w) ; return(t)
                }
```

```
##################################################################################################
#DESCRIPION: calculate estimated calibrated total
#USAGE: EstTOTALcal (w,g,Y1)
#ARGUMENTS:
#               w        final weight without calibration
#               g        g-weights of the calibration estimator
#               Y1       interest variable

#VALUE: the function returns a numeric

EstTOTALcal <- function(w,g,Y1)
               {t=sum(Y1*g*w)  ; return(t)
               }

##################################################################################################
#DESCRIPION: calculate estimated variance of a total
#USAGE: EstVAR (ECH,piij,deltaij,temps,simp,Y,Xs)
#ARGUMENTS:
#               ECH      matrix with (temps+1) columns: sampling probabilities, non-response
     probabilities of each phase
#               piij             second order probabilities
#               deltaij covariance
#               temps            number of non-response phases (1, 2 or 3)
#               simp             1 for calculate the estimated variance with known probabilities, 0
     else
#               Y                the variable for the total
#               Xs               the non-response variables

#VALUE: the function returns a list: V the estimated variance,Vp the part du to the sampling, Vnr
     the part due to the nonresponse,
#Vnrd a list of each part du to the nonresponse phase t, Vsimp the estimated variance like response
      probabilities are known,VsimpNR its part due to te nonresponse

EstVAR <- function(ECH,piij,deltaij,temps,simp,Y,Xs)
               {n=nrow(ECH)  ;  pi0=ECH[,1]  ;  p1_t=rep(1,n)
               for (i in 1:temps) {p1_t=ECH[,i+1]*p1_t}
               #sampling variance SI
               Vp1=sum( ((1-pi0)/p1_t) * (Y/pi0)^2 )
               Vp2=0
               for (i in 1:n)
                       {
                       for (j in 1:n)
                               {if (i != j)
                                       {a= deltaij/piij * (Y[i]/(pi0[i]*p1_t[i])*Y[j]/(pi0[j]*p1_t
                                            [j])) ; Vp2=Vp2+a}
                               }
                       }
               Vp=Vp1+Vp2
               Z=as.matrix(Xs) #if NR with logit
               Vnrd <- c()
               for (delta in 1:temps)
                       {pdelta=ECH[,delta+1]
                                       pdelta_t=rep(1,n)
                                       for (i in delta:temps)
                                               {pdelta_t=ECH[,i+1]*pdelta_t
                                               }
                                       p1_delta=(p1_t/pdelta_t)*pdelta
                                       gammalnum=  t((1-pdelta)/p1_t * (Z)) %*% (Y/pi0)
                                       gammalden=  solve ( t( (pdelta*(1-pdelta))/pdelta_t* (Z) )
                                            %*% Z  )
                                       gammal=gammalden %*% gammalnum
```

```
                                    Vnrdelta = sum ( (pdelta*(1-pdelta))/pdelta_t * ( Y/(pi0*p1
                                        _delta) - Z %*% gamma1 )^2  )
                                    Vnrd <- c(Vnrd,Vnrdelta)
                  }
              Vnr=sum(Vnrd)
              V=Vp+Vnr
              if (simp==1){ VsimpNR=sum((Y/pi0)^2 * (1-p1_t)/(p1_t^2)) ; Vsimp=Vp + VsimpNR}
              if (simp==0){Vsimp=0 }
              l=c(V,Vp,Vnr,Vp1,Vp2,Vnrd,Vsimp,VsimpNR) ; return(l)
              }


##########################################################################################
#DESCRIPION: calculate estimated total or ratio and its variance (with and without calibration)
#USAGE: CalculSAMPLE (POP,n,temps,simp,ratio)
#ARGUMENTS:
#             ECH          (temps+1) columns: sampling probabilities, non-response
    probabilities of each phase
#             piij         second order probabilities
#             deltaij covariance
#             temps        number of non-response phases (1, 2 or 3)
#             simp         1 for calculate the estimated variance with known probabilities, 0
    else
#             Y            the variable for the total
#             Xs           the non-response variables

#VALUE: the function returns a list: V the estimated variance,Vp the part du to the sampling, Vnr
    the part due to the nonresponse,
#Vnrd a list of each part du to the nonresponse phase t, Vsimp the estimated variance like response
     probabilities are known,VsimpNR its part due to te nonresponse

CalculSAMPLE <- function(POP,n,temps,simp,ratio)
              {f<-c();t<-c();v<-c();tcal<-c();vcal<-c()
              f=FinalSAMPLE(POP,n,temps,ratio)
              t=EstTOTAL(f[[7]],f[[10]])
              tcal=EstTOTALcal(f[[7]],f[[5]],f[[10]])
              if (ratio==1){t2=EstTOTAL(f[[7]],f[[9]]) ; tcal2=EstTOTALcal(f[[7]],f[[5]],f[[9]])
                  ; t=t2/t ; tcal=tcal2/tcal }
              v=EstVAR(f[[1]],f[[2]],f[[3]],temps,simp,f[[4]],f[[8]])
              vcal=EstVAR(f[[1]],f[[2]],f[[3]],temps,simp,f[[6]],f[[8]])
              l=c(t,v,tcal,vcal) ; return(l)
              }


##########################################################################################
#DESCRIPION: calculate the g-weights (calibration step)
#USAGE: CALAGE (Xs,temps,w)
#ARGUMENTS:
#             Xs           matrix of sample calibration variables
#             temps        number of non-response phases (1, 2 or 3)
#             w            final weight without calibration

#VALUE: the function returns a vector of the g-weights

CALAGE <- function(Xs,temps,w)
              {piks=1/w      # inclusion probabilities
              total=c(sum(tableR$Cte),sum(tableR$X1),sum(tableR$X2)) # population totals
              g=calib(Xs,d=1/piks,total,method="raking")
              return(g)
              }
```

```
##########################################################################################
#DESCRIPION: simulations in Population for an estimated total or ratio and its variance estimators
#USAGE: SIMULATIONS (NBsimu,POPu,nn,temps,simp,ratio)
#ARGUMENTS:
#            NBech          number of simulations to calculate the estimated total ratio and
    its variance estimators
#            POPu           dataframe ("ID","X1","X2","Y1","Y2","Y3","prob1","prob2","prob3","
    Cte","X3","X4")
#            nn             size of the SI sample
#            temps          number of non-response phases (1, 2 or 3)
#            simp           1 for calculate the estimated variance with known probabilities, 0
    else
#            ratio   1 for a ratio parameter, else it is a total

#VALUE: the function returns a list of numerics

SIMULATIONS <- function(NBsimu,POPu,nn,temps,simp,ratio)
              {tot<-c() ; var<-c() ; o<-c() ; set.seed(12345)
              for (i in 1:NBsimu)
                     {o<-c() ; o=CalculSAMPLE(POP=POPu,n=nn,temps,simp,ratio) ; tot <- rbind(tot
                          ,o)}
              EspT=mean(tot[,1]) ; EspV=mean(tot[,2]) ; EspVp=mean(tot[,3]); EspVnr=mean(tot[,4])
                    ; Vnrdelta<-c() ; Vnrdeltacal <- c()
              for (i in 1:temps)
                     {Vnrdelta <- c(Vnrdelta, mean(tot[,6+i]))}
              EspVsimp=mean(tot[,7+temps]) ; EspVsimpNR=mean(tot[,8+temps]) ; EspTcal=mean(tot[
                   ,9+temps]) ; EspVcal=mean(tot[,10+temps]) ; EspVpcal=mean(tot[,11+temps]);
                   EspVnrcal=mean(tot[,12+temps]);
              for (j in 1:temps)
                     {Vnrdeltacal <- c(Vnrdeltacal, mean(tot[,14+temps+j]))}
              EspVsimpcal=mean(tot[,15+2*temps]) ; EspVsimpcalNR=mean(tot[,16+2*temps])
              l=c(EspT,EspV,EspVp,EspVnr,Vnrdelta,EspVsimp,EspVsimpNR,EspTcal,EspVcal,EspVpcal,
                   EspVnrcal,Vnrdeltacal,EspVsimpcal,EspVsimpcalNR)
              return(l)
              }


##########################################################################################
#DESCRIPION: calculates independently the MonteCarlo variance of an estimated ratio or total
#USAGE: VARIANCE_MC (NBsimuMC,POP,n,temps,ratio)
#ARGUMENTS:
#            NBsimuMC       number of simulations to calculate the MC variance
#            POP            dataframe ("ID","X1","X2","Y1","Y2","Y3","prob1","prob2","prob3","
    Cte","X3","X4")
#            n              size of the SI sample
#            temps          number of non-response phases (1, 2 or 3)
#            ratio   1 for a ratio parameter, else it is a total

#VALUE: the function returns a list of 2 numerics

VARIANCE_MC <- function(NBsimuMC,POP,n,temps,ratio)
              {tot<-c() ; totcal<-c() ; set.seed(1234)
              for (i in 1:NBsimuMC)
                     {f=FinalSAMPLE(POP,n,temps,ratio)
                     TOT=EstTOTAL(f[[7]],f[[10]])
                     TOTCAL=EstTOTALcal(f[[7]],f[[5]],f[[10]])
                     if (ratio==1)
                            {TOT2=EstTOTAL(f[[7]],f[[9]]) ; TOTCAL2=EstTOTALcal(f[[7]],f[[5]],f
                                 [[9]]) ; TOT=TOT2/TOT ; TOTCAL=TOTCAL2/TOTCAL}
                     tot <- c(tot, TOT) ; totcal <- c(totcal, TOTCAL)
                     }
              VARMC=var(tot);VARMCcal=var(totcal) ; l=c(VARMC,VARMCcal) ; return(l)
```

```
                }



################################################################################################
#DESCRIPION: draw SI sample, respondent sample, calculate weights with and without calibration
#USAGE: FinalSAMPLEdiff (POP,n,temps1,temps2)
#ARGUMENTS:
#               POP             dataframe ("ID","X1","X2","Y1","Y2","Y3","prob1","prob2","prob3","
     Cte","X3","X4")
#               n               size of the SI sample
#               temps1  number of non-response phases (1 or 2)
#               temps2  number of non-response phases (2 or 3)

#VALUE: the function returns a list of 4 numerics: the estimated total of temps1 without
     calibration, with calibration, the estimated total of temps2 whitout calibration, with
     calibration

FinalSAMPLEdiff <- function(POP,n,temps1,temps2)
                {s0<-c()
                D=DUMMYsample(POP,n)
                step0=D[[1]]
                s0=step0[step0$Is==1,] ; pi0=rep(n/nrow(POP),n) ; PIij=D[[2]] ; DELTAij=D[[3]]
                s=cbind.data.frame(s0,pi0)

        if (temps2>=1)
                {
                step1=DUMMYnonreponse(s,s$prob1)
                step1P=EstProbareponseLOGIT(step1)
                names(step1P)[names(step1P)=="Id"] <- "I1" ; names(step1P)[names(step1P)=="ee"] <-
                     "Estp1"
                s=step1P[step1P$I1==1,]
                }

        if (temps1==1)
                {Y1=s$YY1 ; w=1/(s$pi0*s$Estp1) ; tot1=EstTOTAL(w,Y1)
                Xs=cbind(s$Cte,s$X1,s$X2) ; g=CALAGE(Xs,temps=temps1,w)
                Y1=s$YY1*g ; tot1cal=EstTOTAL(w,Y1)
                }

        if (temps2>=2)
                {
                step2=DUMMYnonreponse(s,s$prob2)
                step2P=EstProbareponseLOGIT(step2)
                names(step2P)[names(step2P)=="Id"] <- "I2" ; names(step2P)[names(step2P)=="ee"] <-
                     "Estp2"
                s=step2P[step2P$I2==1,]
                }
        if (temps1==2)
                {Y1=s$YY2 ; w=1/(s$pi0*s$Estp1*s$Estp2) ; tot1=EstTOTAL(w,Y1)
                Xs=cbind(s$Cte,s$X1,s$X2) ; g=CALAGE(Xs,temps=temps1,w)
                Y1=s$YY2*g ; tot1cal=EstTOTAL(w,Y1)}
        if (temps2==2)
                {Y1=s$YY2 ; w=1/(s$pi0*s$Estp1*s$Estp2) ; tot2=EstTOTAL(w,Y1)
                Xs=cbind(s$Cte,s$X1,s$X2) ; g=CALAGE(Xs,temps=temps2,w)
                Y1=s$YY2*g ; tot2cal=EstTOTAL(w,Y1)}

        if (temps2==3)
                {
                step3=DUMMYnonreponse(s,s$prob3)
                step3P=EstProbareponseLOGIT(step3)
                names(step3P)[names(step3P)=="Id"] <- "I3" ; names(step3P)[names(step3P)=="ee"] <-
                    "Estp3"
                s=step3P[step3P$I3==1,]
                }
        if (temps2==3)
                {Y1=s$YY3 ; w=1/(s$pi0*s$Estp1*s$Estp2*s$Estp3) ; tot2=EstTOTAL(w,Y1)
                Xs=cbind(s$Cte,s$X1,s$X2) ; g=CALAGE(Xs,temps=temps2,w)
                Y1=s$YY3*g ; tot2cal=EstTOTAL(w,Y1)}

                l=list(tot1,tot2,tot1cal,tot2cal)
                return(l)
                }


##################################################################################################
#DESCRIPION: calculates independently the MonteCarlo variance of an estimated ratio or total
#USAGE: VARIANCE_MC_VtVu (NBsimuMC,POP,n,temps1,temps2)
#ARGUMENTS:
#               NBsimuMC        number of simulations to calculate the MC variance
#               POP             dataframe ("ID","X1","X2","Y1","Y2","Y3","prob1","prob2","prob3","
     Cte","X3","X4")
#               n               size of the SI sample
#               temps1  number of non-response phases (1 or 2)
#               temps2  number of non-response phases (2 or 3)

#VALUE: the function returns a list of 2 numerics

VARIANCE_MC_VtVu <- function(NBsimuMC,POP,n,temps1,temps2)
                {Tot1<-c() ; Tot2<-c() ; Tot3<-c() ; Tot4<-c() ; ; set.seed(123)
                for (i in 1:NBsimuMC)
                        {
                f=FinalSAMPLEdiff(POP,n,temps1,temps2)
                        Tot1 <- c(Tot1,f[[1]])
                        Tot2 <- c(Tot2,f[[2]])
                        Tot3 <- c(Tot1,f[[3]])
                        Tot4 <- c(Tot2,f[[4]])
                        }
                difftot=Tot2-Tot1 ; difftotCAL=Tot4-Tot3
                VARMC=var(difftot); VARMCcal=var(difftotCAL)
                l=c(VARMC,VARMCcal) ; return(l)
                }


##################################################################################################
#DESCRIPION: calculates independently the MonteCarlo nonresponse variance of an estimated ratio or
     total
#USAGE: VARIANCE_MCnr (NBsimuMC,NBsimuMCb,POP,n,temps,ratio)
#ARGUMENTS:
#               NBsimuMC        number of simulations to calculate the MC variance
#               NBsimuMCb       number of simulations to calculate the MC esperance
#               POP             dataframe ("ID","X1","X2","Y1","Y2","Y3","prob1","prob2","prob3","
     Cte","X3","X4")
#               n               size of the SI sample
#               temps           number of non-response phases (1, 2 or 3)
#               ratio   1 for a ratio parameter, else it is a total

#VALUE: the function returns a list of 2 numerics

VARIANCE_MCnr <- function(NBsimuMC,NBsimuMCb,POP,n,temps,ratio)
                {varmc<-c(); varmccal<-c(); set.seed(114)
        for (i in 1:NBsimuMCb)
                {
```

```
          tot<-c() ; totcal<-c()
          s0<-c() ; D=DUMMYsample(POP,n)
          step0=D[[1]] ; s0=step0[step0$Is==1,] ; pi0=rep(n/nrow(POP),n) ; s=cbind.data.frame
              (s0,pi0)

          for (i in 1:NBsimuMC)
                {f=FinalSAMPLE(s,n,temps,ratio,param=1)
                TOT=EstTOTAL(f[[7]],f[[10]])
                TOTCAL=EstTOTALcal(f[[7]],f[[5]],f[[10]])
                if (ratio==1)
                      {TOT2=EstTOTAL(f[[7]],f[[9]]) ; TOTCAL2=EstTOTALcal(f[[7]],f[[5]],f
                          [[9]]) ; TOT=TOT2/TOT ; TOTCAL=TOTCAL2/TOTCAL}
                tot <- c(tot, TOT) ; totcal <- c(totcal, TOTCAL)
                }
          VARMC=var(tot) ; VARMCcal=var(totcal)
          varmc<-c(varmc,VARMC); varmccal<-c(varmccal,VARMCcal)
          }
  EVARMC=mean(varmc) ; EVARMCcal=mean(varmccal)
  l=c(EVARMC,EVARMCcal) ; return(l)
          }
```

Code D.1 – CodeR_Functions.r : Basic functions required to calculate estimators.

D.2 Simulations

```
############## STEP 1
#Compile all the code in "CodeR_Functions.r"
#(this is the required functions for the simulations in Tables 1 and 2)
#For this, you can choose one of the two following commands:
source(".../CodeR_Functions.r")
#source(file.choose())

############## STEP 2: GENERATION VARIABLES
#POPULATION Variables used in Table1 and Table 2 (with rho=0.8) using the function POP
tableR=POP(N=10000,rho=0.8,sigma=10,alpha=10,alphaA=5,alphaB=5,beta1=0.6,beta2=0.75,beta3=0.75)
tableR$YY1=tableR$Y1; tableR$YY2=tableR$Y2 ; tableR$YY3=tableR$Y3


############## STEP 3: SIMULATIONS

############# TABLE 1

tableR$Y1=tableR$YY2-tableR$YY1
Dif21=VARIANCE_MC(NBsimuMC=100000,POP=tableR,n=1000,temps=2,ratio=0)
names(Dif21)=c("V_MC","CAL␣V_MC") ; Dif21 #First value
dif21=VARIANCE_MC_VtVu(NBsimuMC=100000,POP=tableR,n=1000,temps1=1,temps2=2)
dif21

tableR$Y1=tableR$YY3-tableR$YY1
Dif31=VARIANCE_MC(NBsimuMC=100000,POP=tableR,n=1000,temps=3,ratio=0)
names(Dif31)=c("V_MC","CAL␣V_MC") ; Dif31 #First value
dif31=VARIANCE_MC_VtVu(NBsimuMC=100000,POP=tableR,n=1000,temps1=1,temps2=3)
dif31

tableR$Y1=tableR$YY3-tableR$YY2
Dif32=VARIANCE_MC(NBsimuMC=100000,POP=tableR,n=1000,temps=3,ratio=0)
```

```
names(Dif32)=c("V_MC","CAL␣V_MC") ; Dif32 #First value
dif32=VARIANCE_MC_VtVu(NBsimuMC=100000,POP=tableR,n=1000,temps1=2,temps2=3)
dif32




############# TABLE 2

Sys.time()
t1=SIMULATIONS(NBsimu=5000,POPu=tableR,nn=1000,temps=1,simp=1,ratio=0)
names(t1)=c("hat{Y}","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}_{known}","hat{V}^{nr}
       _{known}",
"hat{Y}CAL","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}_{known}","hat{V}^{nr}_{known}"
       )
t1
tt1=VARIANCE_MC(NBsimuMC=100000,POP=tableR,n=1000,temps=1,ratio=0)
names(tt1)=c("V_MC","CAL␣V_MC") ; tt1
ttt1=VARIANCE_MCnr(NBsimuMC=100,NBsimuMCb=1000,POP=tableR,n=1000,temps=1,ratio=0)
names(ttt1)=c("V_MCnr","CAL␣V_MCnr") ; ttt1

tableR$Y1=tableR$YY2
t2=SIMULATIONS(NBsimu=5000,POPu=tableR,nn=1000,temps=2,simp=1,ratio=0)
names(t2)=c("hat{Y}","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}^{nr2}","hat{V}_{known
       }","hat{V}^{nr}_{known}",
"hat{Y}CAL","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}^{nr2}","hat{V}_{known}","hat{V
       }^{nr}_{known}")
t2
tt2=VARIANCE_MC(NBsimuMC=100000,POP=tableR,n=1000,temps=2,ratio=0)
names(tt2)=c("V_MC","CAL␣V_MC") ; tt2
ttt2=VARIANCE_MCnr(NBsimuMC=100,NBsimuMCb=1000,POP=tableR,n=1000,temps=2,ratio=0)
names(ttt2)=c("V_MCnr","CAL␣V_MCnr") ; ttt2

tableR$Y1=tableR$YY3
t3=SIMULATIONS(NBsimu=5000,POPu=tableR,nn=1000,temps=3,simp=1,ratio=0)
names(t3)=c("hat{Y}","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}^{nr2}","hat{V}^{nr3}"
       ,"hat{V}_{known}","hat{V}^{nr}_{known}",
"hat{Y}CAL","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}^{nr2}","hat{V}^{nr3}","hat{V}_
       {known}","hat{V}^{nr}_{known}")
t3
tt3=VARIANCE_MC(NBsimuMC=100000,POP=tableR,n=1000,temps=3,ratio=0)
names(tt3)=c("V_MC","CAL␣V_MC") ; tt3
ttt3=VARIANCE_MCnr(NBsimuMC=100,NBsimuMCb=1000,POP=tableR,n=1000,temps=3,ratio=0)
names(ttt3)=c("V_MCnr","CAL␣V_MCnr") ; ttt3


tableR$Y1=tableR$YY1 ; tableR$Y2=tableR$YY2
r2=SIMULATIONS(NBsimu=5000,POPu=tableR,nn=1000,temps=2,simp=1,ratio=1)
names(r2)=c("hat{R}","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}^{nr2}","hat{V}_{known
       }","hat{V}^{nr}_{known}",
"hat{R}CAL","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}^{nr2}","hat{V}_{known}","hat{V
       }^{nr}_{known}")
r2
rr2=VARIANCE_MC(NBsimuMC=100000,POP=tableR,n=1000,temps=2,ratio=1)
names(rr2)=c("V_MC","CAL␣V_MC") ; rr2
rrr2=VARIANCE_MCnr(NBsimuMC=100,NBsimuMCb=1000,POP=tableR,n=1000,temps=2,ratio=1)
names(rrr2)=c("V_MCnr","CAL␣V_MCnr") ; rrr2

tableR$Y1=tableR$YY1 ; tableR$Y2=tableR$YY3
r3=SIMULATIONS(NBsimu=5000,POPu=tableR,nn=1000,temps=3,simp=1,ratio=1)
names(r3)=c("hat{R}","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}^{nr2}","hat{V}^{nr3}"
       ,"hat{V}_{known}","hat{V}^{nr}_{known}",
"hat{R}CAL","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}^{nr2}","hat{V}^{nr3}","hat{V}_
    {known}","hat{V}^{nr}_{known}")
r3
rr3=VARIANCE_MC(NBsimuMC=100000,POP=tableR,n=1000,temps=3,ratio=1)
names(rr3)=c("V_MC","CAL␣V_MC")  ;  rr3
rrr3=VARIANCE_MCnr(NBsimuMC=100,NBsimuMCb=1000,POP=tableR,n=1000,temps=3,ratio=1)
names(rrr3)=c("V_MCnr","CAL␣V_MCnr")  ;  rrr3



tableR$Y1=tableR$YY2-tableR$YY1
Diff21=SIMULATIONS(NBsimu=5000,POPu=tableR,nn=1000,temps=2,simp=1,ratio=0)
names(Diff21)=c("hat{Y}","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}^{nr2}","hat{V}_{
    known}","hat{V}^{nr}_{known}",
"hat{Y}CAL","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}^{nr2}","hat{V}_{known}","hat{V
    }^{nr}_{known}")
Diff21
Dif21=VARIANCE_MC(NBsimuMC=100000,POP=tableR,n=1000,temps=2,ratio=0)
names(Dif21)=c("V_MC","CAL␣V_MC")  ;  Dif21
Dif21d=VARIANCE_MCnr(NBsimuMC=100,NBsimuMCb=1000,POP=tableR,n=1000,temps=2,ratio=0)
names(Dif21d)=c("V_MCnr","CAL␣V_MCnr")  ;  Dif21d


tableR$Y1=tableR$YY3-tableR$YY1
Diff31=SIMULATIONS(NBsimu=5000,POPu=tableR,nn=1000,temps=3,simp=1,ratio=0)
names(Diff31)=c("hat{Y}","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}^{nr2}","hat{V}^{
    nr3}","hat{V}_{known}","hat{V}^{nr}_{known}",
"hat{Y}CAL","hat{V}","hat{V}^p","hat{V}^{nr}","hat{V}^{nr1}","hat{V}^{nr2}","hat{V}^{nr3}","hat{V}_
    {known}","hat{V}^{nr}_{known}")
Diff31
Dif31=VARIANCE_MC(NBsimuMC=100000,POP=tableR,n=1000,temps=3,ratio=0)
names(Dif31)=c("V_MC","CAL␣V_MC")  ;  Dif31
Dif31d=VARIANCE_MCnr(NBsimuMC=100,NBsimuMCb=1000,POP=tableR,n=1000,temps=3,ratio=0)
names(Dif31d)=c("V_MCnr","CAL␣V_MCnr")  ;  Dif31d
```

Code D.2 – Generation of the interest and auxiliary variables and results of simulations.

Echantillonnage Produit

Hélène Juillard avec Guillaume Chauvet et Anne Ruiz-Gazen

Enquête Elfe

Etude Longitudinale
Francaise depuis l'Enfance

Ined
Inserm

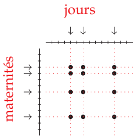
POPULATION : nourrissons nés en 2011 dans l'une des 544 maternités métropolitaines.

THÈMES : leur **santé**, leur **alimentation**, leur **lieu** d'habitation, leur **scolarité** ainsi que leur **vie familiale** et **sociale**.

ECHANTILLON : constitué des 18 300 nourrissons nés dans 320 maternités sélectionnées durant 25 jours sélectionnés. Résultat du **croisement de deux échantillons** de natures différentes, tirés indépendamment : celui des jours et celui des maternités.

PARAMÈTRE D'INTERÊT : par exemple, t_Y est le nombre total de nourrissons nés sous césarienne.

La variable d'étude Y prend la valeur Y_{ik} pour $i \in U_M$ et $k \in U_D$. Y_{ik} correspond au sous-total pour les naissances dans la maternité i, le jour k.



$$t_Y = \sum_{i\in U_M}\sum_{k\in U_D} Y_{ik}$$
$$= \sum_{i\in U_M} Y_{i\bullet} = \sum_{k\in U_D} Y_{\bullet k}$$

$$\hat{t}_{\text{HT}^*} = \sum_{i\in S_M}\sum_{k\in S_D} \frac{Y_{ik}}{\pi_i^M \pi_k^D}$$

$$\text{V}_{\text{prod}}\left(\hat{t}_{\text{HT}}\right) = \sum_{i,j\in U_M}\sum_{k,l\in U_D} \Gamma_{ijkl}\frac{Y_{ik}}{\pi_i^M\pi_k^D}\frac{Y_{jl}}{\pi_j^M\pi_l^D}$$

$$\hat{\text{V}}_{\text{HT}}\left(\hat{t}_{\text{HT}}\right) = \sum_{i,j\in S_M}\sum_{k,l\in S_D} \frac{\Gamma_{ijkl}}{\pi_{ij}^M\pi_{kl}^D}\frac{Y_{ik}}{\pi_i^M\pi_k^D}\frac{Y_{jl}}{\pi_j^M\pi_l^D}$$

*Horvitz-Thompson

Deux plans sources

- MATERNITES $p_M(\cdot)$ $\quad\forall\ i,j \in U_M$:

$$\begin{aligned}\mathbf{E}\left(\mathbf{1}_{\{i\in s_M\}}\right) &= \pi_i^M\\ \mathbf{E}\left(\mathbf{1}_{\{i\in s_M\}}\mathbf{1}_{\{j\in s_M\}}\right) &= \pi_{ij}^M\\ \Delta_{ij}^M \equiv \textbf{Cov}\left(\mathbf{1}_{\{i\in s_M\}},\mathbf{1}_{\{j\in s_M\}}\right) &= \pi_{ij}^M - \pi_i^M\pi_j^M\end{aligned}$$

- JOURS $p_D(\cdot)$ $\quad\forall\ k,l \in U_D$:

$$\begin{aligned}\mathbf{E}\left(\mathbf{1}_{\{k\in s_D\}}\right) &= \pi_k^D\\ \mathbf{E}\left(\mathbf{1}_{\{k\in s_D\}}\mathbf{1}_{\{l\in s_D\}}\right) &= \pi_{kl}^D\\ \Delta_{kl}^D \equiv \textbf{Cov}\left(\mathbf{1}_{\{k\in s_D\}},\mathbf{1}_{\{l\in s_D\}}\right) &= \pi_{kl}^D - \pi_k^D\pi_l^D\end{aligned}$$

Produit des deux plans sources

Hypothèse : **les deux plans sont indépendants.**

L'échantillonnage produit $p(\cdot)$ est défini sur la population produit $U = U_M \times U_D$ par

$p(s) = p_M(s_M) \times p_D(s_D)$ **pour tout** $s = s_M \times s_D \subset U_M \times U_D$**.**

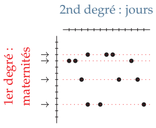
$\forall\ (i,k),(j,l) \in U_M \times U_D$:

$$\begin{aligned}\mathbf{E}\left(\mathbf{1}_{\{(i,k)\in s\}}\right) &= \pi_i^M\pi_k^D\\ \mathbf{E}\left(\mathbf{1}_{\{(i,k)\in s\}}\mathbf{1}_{\{(j,l)\in s\}}\right) &= \pi_{ij}^M\pi_{kl}^D\\ \Gamma_{ijkl} \equiv \textbf{Cov}\left(\mathbf{1}_{\{(i,k)\in s\}},\mathbf{1}_{\{(j,l)\in s\}}\right) &= \pi_i^M\pi_j^M\Delta_{kl}^D + \pi_k^D\pi_l^D\Delta_{ij}^M + \Delta_{ij}^M\Delta_{kl}^D\end{aligned}$$

Quelle différence entre un plan produit SI* × SI et un plan à 2 degrés {SI,SI} ?

$$\text{V}_{\text{prod}}\left(\hat{t}_{\text{HT}}\right) = N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)S^2_{Y_{\circ\bullet}} + N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)S^2_{Y_{\bullet\circ}} + N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)S^2$$

Un plan classique à deux degrés requiert deux hypothèses : l'**indépendance** entre les différents tirages effectués au 2nd degré et l'**invariance** (indépendance entre les tirages effectués à chaque degré). La seconde hypothèse est vérifiée (indépendance entre l'échantillon de maternités et l'échantillon de jours). La première hypothèse n'est pas vérifiée (le même échantillon de jours est utilisé pour chaque maternité).



$$\text{V}_{\mathbf{MD}}\left(\hat{t}_{\text{HT}}\right) = N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)S^2_{Y_{\bullet\circ}} + \frac{N_M}{n_M}N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)\sum_{i\in U_M}S^2_{Y_{i\circ}}$$

$$S^2_{Y_{\circ\bullet}} = \frac{1}{N_M-1}\sum_{i\in U_M}\left(Y_{i\bullet} - \frac{1}{N_M}\sum_{j\in U_M}Y_{j\bullet}\right)^2 \quad \text{inter maternités}$$

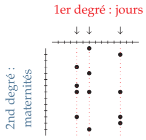
$$S^2_{Y_{\bullet\circ}} = \frac{1}{N_D-1}\sum_{k\in U_D}\left(Y_{\bullet k} - \frac{1}{N_D}\sum_{l\in U_D}Y_{\bullet l}\right)^2 \quad \text{inter jours}$$

$$S^2 = \frac{1}{N_D-1}\frac{1}{N_M-1}\sum_{i\in U_M}\sum_{k\in U_D}\left(Y_{ik} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet k} + \bar{Y}_{\bullet\bullet}\right)^2 \quad \text{résiduel}$$

$$S^2_{Y_{i\circ}} = \frac{1}{N_D-1}\sum_{k\in U_D}\left(Y_{ik} - \frac{1}{N_D}\sum_{l\in U_D}Y_{il}\right)^2 \quad \text{intra maternités}$$

$$S^2_{Y_{\circ k}} = \frac{1}{N_M-1}\sum_{i\in U_M}\left(Y_{ik} - \frac{1}{N_M}\sum_{j\in U_M}Y_{il}\right)^2 \quad \text{intra jours}$$

$$\text{V}_{\mathbf{DM}}\left(\hat{t}_{\text{HT}}\right) = N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)S^2_{Y_{\circ\bullet}} + \frac{N_D}{n_D}N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)\sum_{k\in U_D}S^2_{Y_{\circ k}}$$



*échantillon aléatoire SImple sans remise

Comparaison sous un modèle

modèle m : $Y_{ik} = \mu + \sigma_1 U_i + \sigma_2 V_k + \sigma_3 W_{ik}$ $\quad U_i, V_k, W_{ik} \sim \mathcal{N}(0,1)$

Sous ce modèle, la variance anticipée issue d'un plan produit SI × SI est toujours **plus grande** que celle issue d'un plan à deux degrés {SI,SI}. Elle est d'autant plus grande que la variabilité entre les unités secondaires est grande :

$$E_m\left[\text{V}_{\text{prod}}\left(\hat{t}_{\text{HT}}\right) - \text{V}_{\mathbf{MD}}\left(\hat{t}_{\text{HT}}\right)\right] = N_M^2 N_D^2\frac{n_M-1}{n_M}\left(\frac{1}{n_D}-\frac{1}{N_D}\right){\sigma_2}^2$$

$$E_m\left[\text{V}_{\text{prod}}\left(\hat{t}_{\text{HT}}\right) - \text{V}_{\mathbf{DM}}\left(\hat{t}_{\text{HT}}\right)\right] = N_M^2 N_D^2\frac{n_D-1}{n_D}\left(\frac{1}{n_M}-\frac{1}{N_M}\right){\sigma_1}^2$$

$N_M = N_D = 1000,\ \sigma_1 = \sigma_2 = \sigma_3 = 5$

n_M	10	10	300	300
n_D	10	300	10	300
$\text{V}_{\mathbf{MD}}/\text{V}_{\text{prod}}(\%)$	58.3	98.0	3.07	51.6
$\text{V}_{\mathbf{DM}}/\text{V}_{\text{prod}}(\%)$	55.9	2.81	97.8	48.9

Estimation

Pour un plan produit de taille fixe, nous remarquons la possibilité d'obtenir des estimations négatives de la variance malgré le respect des conditions de **Sen-Yates-Grundy** pour chacun des plans $p_M(\cdot)$ et $p_D(\cdot)$. Afin d'y remédier, plusieurs **estimateurs simplifiés** ont été proposés avec leurs conditions d'utilisation et tenant compte des procédures logicielles existantes (R, SAS, Stata).

Références

- Pirus, C., Bois, C., Dufourg, M.-N., Lanoë, J.-L., Vandentorren, S., Leridon, H. et l'équipe Elfe (2010), La construction d'une cohorte : l'expérience du projet français Elfe, Population 65(4) : 637-670.
- Särndal, C.-E., Swensson, B. et Wretman, J.H. (1992), Model Assisted Survey Sampling, Springer-Verlag.

Novembre 2014
Université de Bourgogne
8ème Colloque Francophone sur les Sondages

Perspectives

En jouant avec les différentes décompositions de Γ_{ijkl}, 5 différents estimateurs de variance à formes Yates-Grundy apparaissent (étude en cours à partir du tirage réjectif avec probabilités inégales).

$$\begin{aligned}\Gamma_{ijkl} &= \pi_{ij}^M\pi_{kl}^D - \pi_i^M\pi_j^M\pi_k^D\pi_l^D\\ &= \pi_{ij}^M\Delta_{kl}^D + \pi_k^D\pi_l^D\Delta_{ij}^M\\ &= \pi_i^M\pi_j^M\Delta_{kl}^D + \pi_{kl}^D\Delta_{ij}^M\\ &= \pi_{ij}^M\Delta_{kl}^D + \pi_{kl}^D\Delta_{ij}^M - \Delta_{ij}^M\Delta_{kl}^D\\ &= \pi_i^M\pi_j^M\Delta_{kl}^D + \pi_k^D\pi_l^D\Delta_{ij}^M + \Delta_{ij}^M\Delta_{kl}^D\end{aligned}$$

A cette première phase d'échantillonnage, vient s'ajouter une phase de non-réponse. La variance associée sera prise en compte dans l'estimateur qui sera proposé aux utilisateurs des données Elfe. De plus, nous envisageons une estimation de variance par linéarisation ou par bootstrap pour des paramètres plus complexes qu'un total.