N^o d'ordre :

THÈSE

présentée à

L'UNIVERSITÉ DE RENNES II,
Haute Bretagne

en vue de l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES II
Option : Mathématiques Appliquées
Spécialité : Statistique

par

Camelia GOGA

ESTIMATION DE LA VARIANCE DANS LES SONDAGES À PLUSIEURS ÉCHANTILLONS ET PRISE EN COMPTE DE L'INFORMATION AUXILIAIRE PAR DES MODÈLES NONPARAMÉTRIQUES

Soutenue le 19 Décembre 2003 devant le jury composé de :

Jean-Claude DEVILLE	Inspecteur Général Insee, Crest-ENSAI	Directeur de recherche
Michel CARBON	Prof. Université de Rennes 2	Directeur de recherche
Ray CHAMBERS	Prof. University of Southampton	Rapporteur/Examinateur
Carl-Erik SÄRNDAL	Prof. Université de Montreal	Rapporteur/Examinateur
Gildas BROSSIER	Prof. Université de Rennes 2	Examinateur
Yves TILLÉ	Prof. Université de Neuchâtel	Examinateur

Laboratoire de Statistique d'Enquêtes, CREST-ENSAI

Remerciements

Je tiens tout d'abord à remercier chaleureusement Jean-Claude Deville d'abord pour m'avoir donné l'occasion de faire cette thèse et ensuite pour son soutien dans les moments difficiles, ses idées originales et son experience dans ce domaine. Je le remercie également pour m'avoir fait confiance durant cette thèse.

Je remercie également Michel Carbon qui a bien voulu être l'encadrant officiel de ce travail.

Les travaux de Ray Chambers et Carl-Erik Särndal sont pour moi des références pour ce qui concerne la théorie des sondages. Je les remercie vivement d'avoir accepté d'être les rapporteurs de cette thèse.

Mes remerciements vont aussi à Yves Tillé dont le cours de sondages constitue une base pédagogique indispensable à toute personne voulant s'initier à ce point de vue de la statistique. Je regrette de ne pas avoir pu discuter plus fréquemment avec lui.

Je remercie également Gildas Brossier pour bien avoir voulu participer au jury de ma thèse.

Je tiens à remercier tous les membres de l'ENSAI et plus particulièrement Céline, Hélène, Denys, Pierre et Marian pour les rigolades de la pause café, les cours de roller et tous les petits bonheurs de la vie quotidienne á l'ENSAI.

Mes pensées sont aussi pour ma famille qui malgré la distance a été à mes cotés durant cette période.

Enfin, je remercie mon mari, Hervé Cardot, pour son soutien infatigable de tous les jours et pour l'apport essentiel dans les discussions. Sa confiance m'a aidé à continuer dans les moments de doute sur le chemin choisi.

Table des matières

Présentation

Cette thèse est consacrée à l'estimation de la variance et à l'utilisation de variables auxilaires lorsqu'on dispose de plusieurs échantillons. L'information sur la variance d'un estimateur, indispensable au statisticien, permet d'évaluer sa performance et de déterminer sa précision. Dans le même ordre d'idée l'utilisation de "manière optimale" de l'information auxiliaire permet d'améliorer la qualité d'un estimateur.

Le problème de l'estimation d'une combinaison linéaire de totaux quand l'information provient de plusieurs échantillons apparaît dans beaucoup de situations pratiques.

L'exemple le plus naturel est celui où deux échantillons correspondent à des sondages effectués à des instants différents du temps. Par exemple, l'estimation de l'évolution d'indicateurs socio-économiques nécessite d'utiliser des techniques d'échantillonnage idoines, les méthodes "classiques" n'étant plus satisfaisantes en général puisqu'elles ne tiennent pas compte du caractère temporel du phénomène étudié. Il existe des situations plus générales, comme celle de l'estimation d'une statistique complexe telle que le ratio en présence de nonréponse, qui peuvent être étudiées dans un même cadre formel.

Face à une telle situation, le statisticien se pose naturellement les questions suivantes :

- Comment choisir son échantillon à chaque tirage ?
- Quels sont les effets des taux de recouvrement et des correlations entre chaque échantillon ?
- Comment tenir compte au mieux de la nonréponse et de l'information auxiliaire ?

Ce travail de thèse propose d'apporter des réponses (parfois partielles) à ces questions.

Le premier chapitre, extrait d'un rapport rédigé pour Eurostat, présente une revue bibliographique sur les méthodes d'estimation de la variance en sondage ainsi que leur implémentation dans le logiciel Poulpe lorsque cela est possible.

Nous proposons ensuite dans le chapitre 2 une étude générale du problème "à deux échantillons" qui puisse s'appliquer en particulier aux deux situa-

tions précédentes.
En s'appuyant sur les travaux de Cotton & Hesse (1992) et Salamin (2002), nous étudions l'estimation d'un paramètre d'intérêt qui dépend, via une fonction linéaire où non, de variables d'intérêt qui sont mesurées sur des échantillons différents. Notre analyse porte plus particulièrement sur le cas bidimensionnel. Nous proposons des estimateurs de type Horvitz-Thompson sur deux échantillons.
Un estimateur linéaire pondéré et sans biais est construit pour une combinaison linéaire de totaux. On donne ensuite une formule de type Horvitz-Thompson pour la variance et son estimateur. Une des difficultés provient du fait que dans le cas de deux échantillons il existe une infinité d'estimateurs sans biais, ce qui n'est pas le cas pour un unique échantillon. Nous déterminons celui dont la variance est minimale sous des conditions générales. Le nombre de paramètres qui caractérisent les poids est proportionnel à la taille de la population ce qui rend impossible leur calcul en général. Nous proposons donc des reparamétrisations qui réduisent considérablement ce nombre de paramètres et qui correspondent à des situations pratiques de stratification de la population. Sous ces conditions et sous les plans usuels nous obtenons les meilleurs estimateurs linéaires sans biais.

Nous étudions ensuite le problème de l'estimation de la variance des statistiques complexes, non linéaires, en utilisant la technique de linéarisation par la fonction d'influence. En introduisant des variables artificielles nous sommes en mesure de donner une approximation de la variance d'un estimateur d'une fonction non linéaire de totaux.

Nous développons également un modèle nonparamétrique pour tenir compte de l'information auxiliaire lors de l'estimation d'un total. La relation entre information auxiliaire n'est donc pas nécessairement linéaire ni même paramétrique ce qui permet d'envisager une classe de fonctions de régression beaucoup plus vaste. Il est vrai que les estimateurs, sous un modèle linéaire, sont protégés contre les mauvaises spécifications du modèle au sens où ils sont asymptotiquement sans biais et consistants par rapport au plan de sondage. Neammoins, si la vraie relation n'est pas linéaire, l'efficacité, en terme de variance, de l'estimateur par la régression généralisée sous un modèle linéaire peut se révéler mauvaise, même par rapport à l'estimateur de Horvitz-Thompson qui, pourtant, ne prend pas en compte l'information auxiliaire. Nous proposons un estimateur nonparamétrique de la différence de totaux basé sur les polynômes locaux permettant de tenir compte de l'information auxiliaire. Nous montrons la convergence de tels estimateurs sous des conditions générales. Ces résultats prolongent ceux obtenus par Breidt & Opsomer (2000) dans le cas d'un seul échantillon.

Enfin, nous proposons dans le chapitre 3 une nouvelle approche nonparamétrique basée sur les splines de régression. Il s'agit d'une méthode

alternative aux polynômes locaux qui se montre beaucoup plus simple et rapide à mettre en oeuvre. Notre estimateur de la régression est décomposé sur une base de fonctions B-splines, ce qui revient à construire un modèle linéaire dont le nombre de variables (les B-splines) n'est pas fixé à l'avance. Il est alors facile d'en déduire une prédiction pour un individu qui n'est pas présent dans l'échantillon. L'estimateur obtenu est une somme linéaire en y_k, les valeurs de la variable d'intérêt sur la population, avec des poids qui ne dépendent pas de la variable d'intérêt. De plus, le paramètre de la régression sur la base de B-splines est indépendant de l'unité dans la population contrairement au cas des estimateurs par les polynômes locaux. Cette methode garde alors les propriétés du modèle linéaire classique de régression, qui d'ailleurs est un cas particulier de l'estimation par des B-splines. Par ailleurs, les estimateurs ainsi obtenus peuvent s'étendre facilement au cas de plusieurs variables auxilaires. Nous prouvons la convergence de l'estimateur proposé et validons, sur des simulations, son bon comportement dans la pratique.

Nous avons souhaité que chaque chapitre puisse être lu de manière indépendante, c'est pourquoi certaines notations ou certains arguments pourront paraître redondants lors du parcours de ce manuscript.

Chapitre 1

Variance estimation in survey sampling

Evaluating the accuracy of estimates in survey sampling is of major importance and many papers have been published for forty years on that topic. This report presents a review of the main contributions on variance estimators in the framework of survey sampling. The recent issues of variance estimators in repeated surveys and in the case of nonresponse are also tackled. In each case, we also deal with the practical implementation of these estimators in the `Poulpe` program.

1.1 Introduction

Variance estimation in survey sampling is of major importance. It gives information on the accuracy of the estimators and allows to build confidence intervals. This report intends to make a review of the major techniques used to derive estimators of the variance of an estimated parameter of interest $\hat{t}$ in the framework of survey sampling. Särndal, Swensson & Wretman (1989) state that a variance estimator $\hat{V}(\hat{t})$ should accomplish at the same time all the following requests : it must have good properties with respect to the sampling design, a simple form and applicability in general. If a system of auxiliary information exists, then it would be advisable that the derived estimator would have good properties with respect to a regression model, including applicability to any linear regression model. At the same time, the variance estimator must have such a form that it can be implemented into a computer software. The principal concern is that such a variance estimator would be able to lead to valid confidence interval for the estimated parameter of study :

$$\hat{t} \pm z_{1-\frac{\alpha}{2}} (\hat{V}(\hat{t}))^{\frac{1}{2}}.$$

There are two main directions for deriving a valid quantity for the unknown variance $V(\hat{t})$:
(a) find an unbiased estimator for the variance when we can calculate it,
(b) find a consistent estimator for the approximative variance.

The choice between the two possibilities depends on the particular features of the survey sampling and on the quantity to be estimated.

There are situations when the minimal value of the variance is desired. Godambe (1955) shows that there exists no linear estimator for the population total with uniformly minimum variance but in more restrictive classes of estimators and certain designs or under a model of superpopulation, we can derive such estimators. An example is the optimum regression estimator obtained by Montanari (1987) ; unfortunately, it is more of theoretical interest since the obtained value of the optimal variance depends on all the values of the variable of study which are unknown and thus it can not be calculated explicitly in practice. Problems of this kind are discussed in more details in Cassel *et al.* 1977, ch 3.

Section 2 deals with the estimation of the fundamental quantities in survey sampling, the total and the mean of a finite population. The Horvitz-Thompson (H-T) estimator is introduced and a general unbiased variance estimator is derived. For the most usual survey designs, we give the precise expressions of the H-T estimator as well as of the variance estimator. Except in a few cases, this general variance estimator has a complicate expression and it is hard to calculate. This is mainly due to the evaluation of a double sum and to the difficulty of calculating the probabilities of inclusion of second order.

In section 3 we propose to extend the derivation of a variance estimator when the parameter of interest has a complex form (e.g. non-linear statistics). This will be done with the help of the Taylor expansion. After a presentation of the theoretical results, the technique is used to derive the variance estimator for the ratio estimator, for the mean of the population and for the coefficient of multiple regression.

In sections 2 and 3, no appeal to the auxiliary information was done. Or, it is well-known that the good use of it improves the results. Sections 4 and 5 deal with ways of incorporating auxiliary information in estimating totals and means. Section 4 treats the *calibration technique* proposed by Deville & Särndal (1992) and Deville (2000).

Section 5 deals with the *superpopulation approach*. This approach introduces a new structure for our population. Until now, sections 2-4, we derived results within the context of the *fixed population approach* (Cassel *et al.* 1977) namely each population unit is associated with a fixed and

unknown real number which is the value of the variable of interest. For the *superpopulation approach*, each population unit will be the outcome of a random variable for which a stochastic structure is specified. Section 5 gives a short review on this issue.

Section 6 deals with the situation of survey sampling in the presence of nonresponse. In this case, we are confronted with the unknown response-nonresponse behaviour of the selected individuals. One must derive models for the unknown probabilities of response of the individuals from the population and build appropriate estimators that minimize the error due to non-response. We make a short description of the existing types of non-response, *unit non-response* and *item non-response* and of the proposed methods. For *unit non-response*, the general approach is that of Särndal, Swensson & Wretman (1992) and the general calibration technique proposed by Deville (2000). These methods reduce the bias and variance due to nonresponse. For the *item non-response*, the imputation technique is always used. The end of the chapter describes the way in which the computer software `Poulpe` treats the presence of nonresponse in survey sampling.

When we sample a population during several periods of time, we are confronted with what, in the specific literature, is called *repeated surveys*. Section 7 intends to make a presentation of the different types of repeated surveys, with special attention on rotation sampling. In this case, the approach is that of Patterson (1950) continued by Eckler (1955), Rao & Graham (1964).

We give in the following a description of the computer software `Poulpe` specialized with the calculation of the variance estimator. Several softwares have specially been created to evaluate the variance in survey sampling but none of them treats completely this issue ; they do not provide estimates of the variance for all kinds of sample surveys based on the best existing evaluation formulas of the variance. The difficulties arise from the complicated expressions of the estimates for different sampling designs which entail the use of less exact formulas but which are easier to implement in a computer software.
`Poulpe` is a computer software written in SAS MACRO language and specialized with the evaluation of the variance, being used by INSEE in business and households surveys. In order to be able to use this program, two conditions must be satisfied :

i. a detailed sampling schema must exist ;

ii. we must have enough data in order to calculate the probabilities of inclusion.

`Poulpe` treats the following simple survey designs :

i. simple sample survey without replacement ;

ii. systematic survey ;

iii. balanced survey ;

iv. unequal probabilities survey sampling ;

in the case of one-stage or multi-stage sampling, taking account of the existence of nonresponse. The complex surveys are also treated.
On the contrary, this computer software neglects :

i. the influence of the partial nonresponse ;

ii. the influence of the measurement errors.

A particular feature of `Poulpe` is the way it reduces the more complicated surveys, the complex designs in multi-stages, to a two-stage survey. This is possible because of the following :

i. the tree structure of the survey design made when the units are selected from the population. In this tree, the root is the initial population, the levels are the selected units at every stage and the vertical branches make the link between each unit and the corresponding sampled subunits ; each branch contains information about the used design and the first order probabilities inclusion. The tree ends up with sample elements.

ii. the use of the induction for evaluating the variance in multi-stage sampling. This fact permits to calculate the variance estimate starting with last stage of sampling and climbing up to the root, applying successively the two-stage formula.

These two qualities of `Poulpe` make possible the derivation of the variance estimate, without knowing its final explicit expression. In the next, for each design, the `Poulpe` evaluations of variance are given.

Notations

Let us consider a finite population $\mathcal{U}$ composed of N elements

$$\mathcal{U} = \{u_1, \ldots, u_k, \ldots, u_N\} = \{1, \ldots, k, \ldots, N\}$$

where for simplicity, we identify the k-th element of $\mathcal{U}$ denoted by u_k with its label k. We will consider in the next that our population is such that each unit u_k can be spotted in a unique way by its label, k. This means that the units have the property of identifiability (Cassel *et al* 1977).

Let consider $\mathcal{Y}$, a variable of interest for which the value for the k-th unit, denoted by y_k, is unknown. We designate by $\mathbf{y} = (y_1, \ldots, y_N)$ the *parameter of the finite population* and any real function of it is called a *parametric function*. The goal of a survey sampling is to make inference about a parametric

function such as the total or the mean for example, but more complicated functions may be of interest (the mode, various population quantiles, the population variance).

In the case of a survey sampling, the inference is based on information obtained only from a part of $\mathcal{U}$, called *sample*, obtained from $\mathcal{U}$ by a probabilistic selection scheme. More precisely, let $\mathcal{S}$ be the set of all possible subsets s of $\mathcal{U}$, $s \subset \mathcal{P}(\mathcal{U})$. There are 2^N possible subsets, considering ϕ and $\mathcal{U}$; a sample s is an element of $\mathcal{S}$. Given $\mathcal{U}$, let $p(s)$ be the probability of selecting $s \in \mathcal{S}$. In other words, the function $p(s)$ which is called *the sampling design* satisfies the following conditions :

i. $p(s) \geq 0$ for all $s \in \mathcal{S}$

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

In the present work, we will deal only with *noninformative designs*, namely with designs for which the function $p(\cdot)$ does not depend on the values of $\mathcal{Y}$. For this situation, Basu (1958) shows that it is sufficient to consider only the distinct elements of the sample.

We note by n_s the *sample size*, namely the number of elements of s ; depending on the chosen scheme, n_s may be fixed or not for all the samples $s \in \mathcal{S}$.

We denote by $I_k = \mathbf{1}_{\{k \in s\}}$ for all $k \in \mathcal{U}$ the *sample membership indicator* (Särndal *et al* 1992). The random variable I_k is a Bernoulli variable indicating if the unit k belongs or not to the sample.

Supposing that a sampling design has been fixed, the probabilities of inclusion are defined as follows :

i. π_k, the first order inclusion probability, is the probability that the element k will be included in a sample . For all $k \in \mathcal{U}$, $\pi_k = \sum_{s \ni k} p(s)$.

ii. π_{kl}, the second order inclusion probability, is the probability that the elements k and l will be included in a sample. For all $k, l \in \mathcal{U}$, $\pi_{kl} = \sum_{s \ni k \& l} p(s)$.

Result 1.1.1 *: For a given sampling design $p(\cdot)$, the functions I_k have the following properties :*

i. $E(I_k) = \pi_k$,

ii. $V(I_k) = \pi_k(1 - \pi_k)$,

iii. $Cov\ (I_k, I_l) = \pi_{kl} - \pi_k \pi_l$, $k \neq l$

for all $k, l \in \mathcal{U}$.

Proof : The proof relies on the fact that I_k is a Bernoulli variable of parameter π_k, see for more details Särndal *et al* (1992). □

Consequence 1 *: If the design $p(s)$ has a fixed size, then*

i. $\sum_{\mathcal{U}} \pi_k = n$

ii. $\sum_{k \neq l} \sum_{\mathcal{U}} \pi_{kl} = n(n-1)$

iii. $\sum_{l \in \mathcal{U}, l \neq k} \pi_{kl}$.

For the simplicity of notation, we introduce the Δ-quantities (Särndal *et al* 1992) :

$$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$$

$$\check{\Delta}_{kl} = \Delta_{kl}/\pi_{kl}$$

for all $k, l \in \mathcal{U}$.

We suppose from now on that $\pi_k > 0$ for all $k \in \mathcal{U}$, namely that each unit in the population has a chance to be in the sample.

1.2 Variance estimators for the usual designs

In this section we describe the properties of estimators of finite population totals,

$$t_y = \sum_{\mathcal{U}} y_k.$$

The Horvitz-Thompson estimator for the population total is introduced and unbiased estimators of variance are presented for various classes of sampling designs. In this section as well as in the next two, we are placed in the context of the fixed population approach so the only randomness is the sampling design, $p(\cdot)$. As a result, the definitions of the expectation, variance and mean square error of an estimator Q for t_y can be formulated for a given design $p(s)$. For example,

$$E(Q) = \sum_{s \in \mathcal{S}} p(s)Q(s).$$

We derive formula for the variance estimator specifying the advantages or disadvantages of this estimator. Then, in the next sections, we will propose others estimators and make a comparison of their properties.

1.2.1 The Horvitz-Thompson estimator

We consider the class of linear estimators and among these estimators we take the one proposed by Horvitz-Thompson (1952). This estimator is sometimes called the π *estimator* for the total of $\mathcal{Y}$ because of the probabilities of first degree appear in its formula :

$$\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}. \tag{1.1}$$

We give now the most important result of this section.

Result 1.2.1 *(Horvitz-Thompson 1952). The π estimator for the total of $\mathcal{Y}$, $\hat{t}_\pi$, has the following properties :*

i. $\hat{t}_\pi$ is unbiased for $t = \sum_{\mathcal{U}} y_k$.

ii. The variance of $\hat{t}_\pi$ has the expression :

$$V(\hat{t}_\pi) = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

iii. If $\pi_{kl} > 0$ for all $k\&l \in \mathcal{U}$, an unbiased estimator for $V(\hat{t}_\pi)$ is :

$$\widehat{V}(\hat{t}_\pi) = \sum_s \sum_s \check{\Delta}_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

Proof : The proof relies on the identities :

$$\begin{aligned} \hat{t}_\pi &= \sum_s \frac{y_k}{\pi_k} = \sum_{\mathcal{U}} \frac{y_k}{\pi_k} I_k \\ \widehat{V}(\hat{t}_\pi) &= \sum_{\mathcal{U}} \sum_{\mathcal{U}} \check{\Delta}_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} I_k I_l \end{aligned}$$

and on the properties of the membership indicators I_k, for all $k \in U$ given in the Result 1.1.1. (see for more details Särndal *et al* 1992 and Thompson 1997). □

Remark 1.2.1 *: The H-T estimator is the only unbiased homogeneous linear estimator whose weights do not depend on the sample.*

Remark 1.2.2 *: The variance of $\hat{t}_\pi$ can be written as a quadratic form as follows :*

$$V(\hat{t}_\pi) = \boldsymbol{y} \Delta \boldsymbol{y}'$$

where $\Delta = (\frac{\Delta_{kl}}{\pi_k \pi_l})_{k,l \in \mathcal{U}}$ and $\boldsymbol{y} = (y_1, \ldots, y_N)$ the parameter vector.

For a sampling design of fixed size, $n_s = n$, equivalent formulas can be deduced for the variance and variance estimator of $\hat{t}_\pi$, as obtained by Yates and Grundy (1953) and Sen (1953).

Result 1.2.2 *(Yates-Grundy-Sen 1953). If $p(s) > 0$ is of fixed size, then $V(\hat{t}_\pi)$ and $\widehat{V}(\hat{t}_\pi)$ have the equivalent expressions :*

i. $V(\hat{t}_\pi) = -\frac{1}{2} \sum \sum_{\mathcal{U}} \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$.

ii. If $\pi_{kl} > 0$ for all $k, l \in \mathcal{U}$,

$$\widehat{V}(\hat{t}_\pi) = -\frac{1}{2} \sum \sum_{s} \check{\Delta}_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 .$$

Proof : We use Consequence 1. (Särndal *et al* 1992). □

1.2.2 Sampling without replacement (SI)

We select with equal probability a first element from the population ; this element will be kept away during the following selections. Next, we select with equal probability another element among the $N - 1$ remaining units of the population and we continue the selection in the same way until the sample has n elements. The sampling design has the following expression

$p(s) = \dfrac{1}{\binom{N}{n}}$ and for the probabilities of inclusion of first and second degree, $\pi_k = \dfrac{n}{N}$ and $\pi_{kl} = \dfrac{n(n-1)}{N(N-1)}$. From **Result 1.2.1**, we have :

Result 1.2.3 *: Under a sampling **SI**, the π-estimator for the population total becomes*

i. $\hat{t}_{\pi,SI} = N\overline{y}_s = \dfrac{1}{f}\sum_s y_k$.

ii. The variance has the expression $V_{SI}(\hat{t}_\pi) = N^2\dfrac{1-f}{n}S^2_{y\mathcal{U}}$.

iii. An unbiased estimator of the variance is given by

$$\widehat{V}_{SI}(\hat{t}_\pi) = N^2\frac{1-f}{n}S^2_{ys},$$

where $f = \dfrac{n}{N}$,

$$S^2_{y\mathcal{U}} = \frac{1}{N-1}\sum_{\mathcal{U}}(y_k - \overline{y}_{\mathcal{U}})^2$$

with $\overline{y}_{\mathcal{U}} = N^{-1}\sum_{\mathcal{U}} y_k$ *and*

$$S^2_{ys} = \frac{1}{n-1}\sum_s (y_k - \overline{y}_s)^2,$$

for $\overline{y}_s = n^{-1}\sum_s y_k$.

Remark 1.2.3 *: For **SI** we have* $(\Delta_{k,l})_{k,l\in U} = k(\boldsymbol{I}_N - \boldsymbol{P})$ *with* $k = f(1-f)\dfrac{N}{N-1}$; $\boldsymbol{I}_N$ *is the identity matrix with dimension* N *and* $\boldsymbol{P} = \dfrac{1}{N}\mathbf{1}_N\mathbf{1}'_N$. *The variance has the expression* $V_{SI} = \dfrac{N}{N-1}\dfrac{1-f}{f}\boldsymbol{y}(\boldsymbol{I}_N - \boldsymbol{P})\boldsymbol{y}'$.
As a consequence, det $(\Delta_{k,l}) = 0$ *and the variance will be minimized for all vector* $\boldsymbol{y} = c\ \mathbf{1}'_N$ *with* c *a real constant.*

The evaluation of variance by POULPE

The variance estimation formula given above, at the point (3), is implemented in `Poulpe`. The calculation of S^2_{ys} is made directly from the recorded data for all the sampled individuals.

1.2.3 Sampling with replacement (SIR)

For sampling with replacement, denoted by **SIR**, Hansen and Hurwitz (1943) proposed the *pwr* estimator : p-expanded with replacement, corresponding to a generalization of simple random sampling with replacement.

The method consists in drawing with replacement m different elements with unequal probabilities $p_1, \ldots, p_k, \ldots, p_N$ retaining the independence of the draws.

$$Pr(\text{selecting } k) = p_k \text{ for } k = 1, \ldots, N.$$

Then the sets of $\{p_k\}_{k \in \mathcal{U}}$ satisfy the properties :

i. $p_k > 0$ for all $k \in \mathcal{U}$

ii. $\sum_{\mathcal{U}} p_k = 1$

The proposed *pwr* estimator for the total t_y is :

$$\hat{t}_{pwr} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_{k_i}}{p_{k_i}}$$

We have the following result for the *pwr* estimator :

Result 1.2.4 *:*

i. The variance of $\hat{t}_{pwr}$ *is :*

$$V(\hat{t}_{pwr}) = \frac{V_1}{m} \text{ where } V_1 = \sum_{\mathcal{U}} \left(\frac{y_k}{p_k} - t \right)^2 p_k;$$

ii. An unbiased estimator for $V(\hat{t}_{pwr})$ *is :*

$$\widehat{V}(\hat{t}_{pwr}) = \frac{\hat{V}_1}{m}; \quad \hat{V}_1 = \frac{1}{m-1} \sum_{i=1}^{m} \left(\frac{y_{k_i}}{p_{k_i}} - \hat{t}_{pwr} \right)^2.$$

An alternative estimator is $\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$ where $\pi_k = 1 - \left(1 - \frac{1}{N}\right)^m$ and $\pi_{kl} = 1 - 2\left(1 - \frac{1}{N}\right)^m + \left(1 - \frac{2}{N}\right)^m$. Then the resulting estimator is :

$$\hat{t}_\pi = \frac{1}{1 - \left(1 - \frac{1}{N}\right)^m} \sum_s y_k$$

and the expressions for the variance and variance estimator for $\hat{t}_\pi$ are obtained from the **Result 1.2.1**.

The two sampling designs, **SI** and **SIR**, presented above are sampling designs of fixed size so one can use the Yates-Grundy formula for calculating a variance estimator of $\hat{t}_\pi$. In the case of **SI** sampling, both the variance estimator given by the **Result 1.2.1** and the Yates-Grundy formula give the same result.

1.2.4 Bernoulli sampling (BE)

In this case the sample is composed of all elements k from $\mathcal{U}$ which satisfy $\varepsilon_k < \pi$, where ε_k for all $k \in \mathcal{U}$ are independent realizations of a random variable with uniform distribution in the interval (0, 1) and $0 < \pi < 1$ is a fixed constant. As a result all the units from the population have the same probability of inclusion of first degree, $\pi_k = \pi$ for all $k \in \mathcal{U}$. Besides, the selection of the unit k is made independently from the selection of unit l with $l \neq k$; as a consequence, $\pi_{kl} = \pi^2$ for all $k \neq l \in \mathcal{U}$ and $\Delta_{kl} = 0$ for all $k \neq l \in \mathcal{U}$. Then, the variance-covariance matrix has the expression $\Delta = \text{diag}\pi(1 - \pi)$.

Result 1.2.5 *: Under a **BE** sampling, the π-estimator for the population total t_y can be written*

i. $\hat{t}_{\pi,BE} = \dfrac{1}{\pi} \sum_s y_k$;

ii. The variance has the expression $V_{BE}(\hat{t}_\pi) = \left(\dfrac{1}{\pi} - 1\right) \sum_{\mathcal{U}} y_k^2$.

iii. An unbiased estimator of the variance is given by

$$\hat{V}_{BE}(\hat{t}_\pi) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right) \sum_s y_k^2.$$

The π estimator in the case of Bernoulli sampling is often inefficient because of the variable sample size. Nevertheless, the Bernoulli sampling conditioned to the sample size n_s is a **SI** sample. That's why, once selected the **BE** sample, we can consider the conditional frame.
Although designs with fixed size are desired, there are situations in which variable sample size conducts better. Two examples are relevant. The first one is the selection in a domain of the finite population, situation not treated here and the second one is the selection in the presence of nonresponse. In this case, the response behaviour in different population subgroups is often modeled as a **BE** sample selection, a technique discussed in section 6.

1.2.5 Poisson sampling (PO)

The **BE** sampling is not a fixed size design. Another example is the Poisson sampling (**PO**) when the selection of an element is decided by $\varepsilon_k < \pi_k$ where $\{\pi_1, \ldots, \pi_n\}$ is a set of fixed constants between 0 and 1. We give the first and second order inclusion probabilities. Based on the same arguments as in the case of **BE** sampling, we have that π_k for all $k \in \mathcal{U}$ are the set of first order inclusion probabilities ; as for the second degree inclusion probabilities, we have for $k \neq l \in \mathcal{U}$, $\pi_{kl} = \pi_k \pi_l$. Because of these particular expressions of the inclusion probabilities, the **Result** 1.2.1 will become :

Result 1.2.6 *: Under a* ***PO*** *sampling, the π-estimator for the population total has the following expression :*

i. $\hat{t}_{\pi,PO} = \sum_s \frac{y_k}{\pi_k}$.

ii. The variance is given by $V_{PO}(\hat{t}_\pi) = \sum_{\mathcal{U}} \pi_k(1-\pi_k)\breve{y}_k^2 = \sum_{\mathcal{U}} \frac{(1-\pi_k)}{\pi_k} y_k^2$.

iii. An unbiased variance estimator is $\hat{V}_{PO}(\hat{t}_\pi) = \sum_s (1-\pi_k)\breve{y}_k^2$.

As in the case of **BE** sampling, the **PO** sampling is of variable sample size n_s, fact which entails a great value of $V_{PO}(\hat{t}_\pi)$. By conditioning, we may remove this drawback. Hajek (1981) proves that every conditional **PO** sampling maximizes the entropy in the class of designs having the same first degree inclusion probability and taking part to the same class Ω of samples satisfying $\sum_{s\in\Omega} p(s) = 1$. The entropy is a measure of spread of the sample design defined by

$$e = -\sum_s p(s) \ln p(s).$$

The main role of **PO** sampling is to help to define and analyse other sampling method (Hajek 1981). By conditioning, we can obtain **SI** sampling, two-stage sampling described in sections 2.2 and 2.9, etc.

In the case of a **BE** sampling we have :

$$E_{BE}(n_s) = N\pi \text{ and } \pi_k = \pi \quad \text{for all } k = 1, \ldots, N.$$

Thus, if we fix the expected sample size and we suppose that N is known, then π_k for $k = 1, \ldots, N$ are completely specified. In **PO** sampling, we do not have the same thing :

$$E_{PO}(n_s) = \sum_{k=1}^{N} \pi_k.$$

Thus, for fixed $E_{PO}(n_s)$ we have to make a choice for π_k. We will choose the π_k that minimize the variance. We get the following expression for the first order inclusion probabilities :

$$\pi_k = \frac{n y_k}{\sum_{\mathcal{U}} y_k} \quad \text{for all } k = 1, \ldots, N$$

assuming that $y_k < \frac{\sum_{\mathcal{U}} y_k}{n}$ for all $k = 1, \ldots, N$. Because the expression of π_k requires y_k for all $k = 1, \ldots, N$ which in general are unknown, the inclusion probabilities so obtained can not be used. But, if we have auxiliary information $\mathcal{X}$ which takes the value x_k for the k-th element of the population

and the variable of interest $\mathcal{Y}$ is approximately proportional to $\mathcal{X}$, then we can consider :

$$\pi_k = \frac{nx_k}{\sum_{\mathcal{U}} y_k} \quad \text{for all } k = 1, \ldots, N$$

and

$$x_k < \frac{\sum_{\mathcal{U}} x_k}{n} \quad \text{for all } k = 1, \ldots, N.$$

The inclusion probabilities π_k given by these expressions are called *probability proportional to size*. If $\mathcal{Y}$ is proportional to $\mathcal{X}$, then the associated π-estimator $\hat{t}_{\pi,PO}$ has a small variance.
This discussion states a more general problem : how we can use available auxiliary information at the sampling stage ? One possibility is selecting units with unequal probabilities, such as *probability proportional to size* described above. We will describe briefly in the next subsection the most important methods of units selection with unequal probabilities. Tillé (2001) gives a detailed description of these methods as well as the associated algorithms.

1.2.6 Unequal probabilities sampling designs

We need auxiliary information $\mathcal{X}$ for selecting unequal probabilities sampling designs. The derivation of first order inclusion probabilities of a fixed size design is the first step of all the methods. Conditions on the second order inclusion probabilities are formulated, fact which makes possible the resolution of this problem. Hanif & Brewer (1980) and Brewer & Hanif (1983) present all the existing methods for selecting unequal probabilities sampling designs until that moment.
We can have designs with or without replacement. We start with the study of the πps sampling design when $\pi_k \propto x_k$ for all k in the population, called πps sampling. In this case, several schemes for selecting elements such that $\pi_k \propto x_k$ have been proposed. All these schemes have been devised in order to accomplish the following five requests (Särndal *et al.* 1992) :

i. The selection of the sample is simple ;

ii. the π_k are strictly proportional to x_k for all $k = 1, \ldots, N$;

iii. $\pi_{kl} > 0$ for all $k \neq l$;

iv. the π_{kl} can be computed simply ;

v. $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l < 0$ for all $k \neq l$ to guarantee that the Yates-Grundy variance estimator is nonnegative.

Let n be the sample size. We will study different schemes depending on the values of n.

For $n = 1$, we have the total cumulative method which is strictly a πps design, but $\pi_{kl} = 0$ for all $k \neq l$ so an unbiased variance estimator can not be obtained.

For $n = 2$, we mention only the design proposed by Brewer (1963, 1975) which ensures $\pi_k = \frac{2x_k}{\sum_{\mathcal{U}} x_k}$ for $k = 1, \ldots, N$ and $\pi_{kl} > 0$ for all $k \neq l$. This scheme satisfies also the condition that $\Delta_{kl} < 0$ for all $k \neq l$ that allows the Yates-Grundy variance estimator to be always nonnegative. So, all the above requirements are satisfied.

For $n > 2$, several schemes exit, most of them are complicated because of the requirement that π_k must be proportional to x_k for all $k \in \mathcal{U}$. If we relax this condition, Sunter (1977) proposes a schema which gives π_k strictly proportional to x_k except for a small portion of the population, corresponding to the smallest values of x_k. We have also $\pi_{kl} > 0$ and $\Delta_{kl} < 0$ for $k \neq l$. Sunter (1986) presents a list-sequential scheme which achieves a strictly proportionality between π_k and x_k. Madow (1949) proposes a systematic unequal probabilities sampling design, which is one of the best because of its simplicity. The drawback of this method is that many of the π_{kl} are null.

The method of Rao, Hartley and Cochran (1962) gives an unbiased estimator for the population mean and at the same time a variance estimator.

As for designs with replacement, we present the one proposed by Hansen & Hurwitz (1943). It consists in taking the *pwr*-estimator presented in the case of **SIR** sampling

$$\hat{t}_{pwr} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_{k_i}}{p_{k_i}}$$

when the probabilities used to select units $p_1, \ldots, p_N$ are as follows $p_k = \frac{x_k}{\sum_{\mathcal{U}} x_k}$ for all $k = 1, \ldots, N$.

Then the estimator $\hat{t}_{pwr}$ coupled with a **SIR** sampling design has the properties :

i. $\hat{t}_{pwr} = \left(\sum_{\mathcal{U}} x_k \right) \frac{1}{m} \sum_{i=1}^{m} \frac{y_{k_i}}{x_{k_i}}$

ii. $\hat{V}(\hat{t}_{pwr}) = \frac{\hat{V}_1}{m}, \hat{V}_1 = \frac{(\sum_{\mathcal{U}} x_k)^2}{m-1} \left[\sum_{i=1}^{m} \left(\frac{y_{k_i}}{x_{k_i}} \right)^2 - \frac{1}{m} \left(\sum_{i=1}^{m} \frac{y_{k_i}}{x_{k_i}} \right)^2 \right]$

The same estimator can be obtained if we repeat m times independently, the cumulative method. We start with $p_k = \frac{x_k}{\sum_{\mathcal{U}} x_k}, k \in U$ and let v_k be defined as follows :

$$v_0 = 0 \text{ and } v_k = \sum_{l=1}^{k} p_l \quad \text{for } k = 1, \ldots, N.$$

Then we generate u an Unif(0, 1) variable and the selection or not selection of the element k will be decided as follows : the element k is selected if

$v_{k-1} \leq u < v_k$. We repeat this operation m times until the sample s is obtained.

$$P(\text{ element k is selected }) = P(v_{k-1} \leq u < v_k) = v_k - v_{k-1} = p_k$$

and $\hat{t}_{yHH} = \dfrac{1}{m}\sum_{i=1}^{m} \dfrac{y_{k_i}}{p_{k_i}}$ with the same variance and estimate variance as $\hat{t}_{pwr}$ deduced above. As it can be noticed, the expressions for $V(\hat{t}_{pwr})$ and $\hat{V}(\hat{t}_{pwr})$ are simpler than the corresponding ones if we had used the π-estimator, when the cross-products $\check{\Delta}_{kl}\check{y}_k\check{y}_l$ must be computed. However, $\hat{t}_{pwr}$ is less efficient than the π estimator. The following strategy gives an estimator for the variance that combines the two estimators :

i. a fixed size m πps sampling design is used such that $\pi_k = mp_k = \dfrac{mx_k}{\sum_{\mathcal{U}} x_k}$;

ii. the π-estimator is used to estimate the population total t_y;

iii. the variance of the π-estimator is estimated by the *pps*-sampling formula :

$$\hat{V} = \frac{1}{m(m-1)} \sum_s \left(\frac{y_k}{p_k} - \frac{1}{m} \sum_s \frac{y_k}{p_k} \right)^2$$

The variance estimator $\hat{V}$ will not be unbiased for $V(\hat{t}_\pi)$.

All the methods that have been devised are difficult to apply in practice ; the π_{kl} for $k \neq l$ requires heavy calculations and the calculation of an estimator for the variance estimator given by the Yates-Grundy formula is often impossible to do. A way to avoid this calculation would be to search an estimator for the variance that does not use the quantities π_{kl}. Such an estimator for sampling with replacement exists, it has been proposed by Rao-Hartley-Cochran (1962). It is possible that the variance obtained in a great number of methods would be approximated by an estimator relatively simple. Berger (1998) showed that the variance of Hajek (1981) is a good approximation for the variance, under the conditions that the initial plan does not diverge very much from the plan of maximum entropy. A general approximation for the variance estimation, in the case of sampling without replacement with a large entropy, is given by :

$$V(\hat{t}_{y\pi}) = \sum_{k \in \mathcal{U}} \frac{b_k}{\pi_k^2} (y_k - y_k^*)^2$$

where

$$y_k^* = \pi_k \frac{\sum_{l \in \mathcal{U}} b_l y_l / \pi_l}{\sum_{\mathcal{U}} b_l}.$$

For the quantities b_k we have more approximations :

i. The approximation given by Hajek :

$$b_k = \frac{\pi_k(1-\pi_k)N}{N-1}.$$

This approximation is exact in the case of sampling without replacement with fixed size.

ii. Hajek proposed another approximation. We can write the approximative expression for the variance in the equivalent way :

$$V(\hat{t}_{y\pi}) = \sum_{k\in\mathcal{U}} \frac{y_k^2}{\pi_k^2}\left(b_k - \frac{b_k^2}{\sum_{l\in\mathcal{U}} b_l}\right) - \frac{1}{\sum_{l\in\mathcal{U}} b_l}\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U},k\neq l}\frac{y_k y_l b_k b_l}{\pi_k\pi_l}.$$

The variance proposed by Horvitz-Thompson can be written :

$$V(\hat{t}_{y\pi}) = \sum_{k\in\mathcal{U}} \frac{y_k^2}{\pi_k^2}\pi_k(1-\pi_k) + \sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U},k\neq l}\frac{y_k y_l}{\pi_k\pi_l}(\pi_{kl}-\pi_k\pi_l).$$

We can take b_k such that the two expressions for the variance estimation would be the same. We will take b_k which satisfies the equation :

$$b_k - \frac{b_k^2}{\sum_{l\in\mathcal{U}} b_l} = \pi_k(1-\pi_k).$$

We can use the same technique for obtaining an approximation for the variance estimation. A general formula for an approximation is given by :

$$\widehat{V}(\hat{t}_{y\pi}) = \sum_{k\in s}\frac{c_k}{\pi_k^2}\left(y_k - \widehat{y^*}_k\right)^2$$

where

$$\widehat{y_k^*} = \pi_k\frac{\sum_{l\in s} c_l y_l/\pi_l}{\sum_{l\in s} c_l}.$$

Now, we can choose for c_k :

i. $c_k = (1-\pi_k)\frac{n}{n-1}$.

ii. Deville (1998) proposes for c_k :

$$c_k = (1-\pi_k)\left[1 - \sum_{k\in s}\left\{\frac{1-\pi_k}{\sum_{l\in s}(1-\pi_l)}\right\}^2\right]^{-1}.$$

iii. Berger (1998) proposes :

$$c_k = (1-\pi_k)\frac{n}{n-1}\frac{\sum_{k\in s}(1-\pi_k)}{\sum_{k\in s}\pi_k(1-\pi_k)}$$

iv. We can write the approximative variance as follows :

$$\hat{V}(\hat{t}_{y\pi}) = \sum_{k\in s} \frac{y_k^2}{\pi_k^2}\left(c_k - \frac{c_k^2}{\sum_{l\in s} c_l}\right) - \frac{1}{\sum_{l\in s} c_l}\sum_{k\in s}\sum_{l\in s, k\neq l} \frac{y_k y_l c_k c_l}{\pi_k \pi_l}$$

The variance estimator proposed by Horvitz-Thompson can be written equivalently :

$$\hat{V}(\hat{t}_{y\pi}) = \sum_{k\in s} \frac{y_k^2}{\pi_k^2}(1-\pi_k) + \sum_{k\in s}\sum_{l\in s, k\neq l} \frac{y_k y_l}{\pi_k \pi_l \pi_{kl}}(\pi_{kl} - \pi_k \pi_l)$$

We can take c_k so that the coefficients of the two expressions for the variance estimation would be the same :

$$c_k - \frac{c_k^2}{\sum_{k\in s}} = 1 - \pi_k.$$

The variance evaluation by POULPE

In the case of "unequal probabilities sampling", the difficulty of calculating the double sums and also the π_{ij}, the second-order probabilities of inclusion appears. To overcome this difficulty, the above approximations were introduced. The one which is implemented in `Poulpe` is the following

$$\hat{V}(\hat{t}_\pi) = \frac{n}{n-1}\sum_s (1-\pi_i)\left(\frac{y_i}{\pi_i} - D_2\left(\frac{y}{\pi}\right)\right)^2$$

where

$$D_2\left(\frac{y}{\pi}\right) = \frac{\sum_s (1-\pi_i)\frac{y_i}{\pi_i}}{\sum_s (1-\pi_i)}$$

1.2.7 Stratified sampling

In stratified sampling, the population $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$ is divided into H subpopulations denoted U_h of size N_h for $h = 1, \ldots, H$ such that $U_h \cap U_{h'} = \phi$ for $h \neq h'$.

$$\mathcal{U} = \{1, \ldots, k, \ldots, N\} = \bigcup_{h=1}^{H} U_h$$

$$|U| = N,\ |U_h| = N_h,\ N = \sum_{h=1}^{H} N_h$$

From each subpopulation $\mathcal{U}_h, h = 1, \ldots, H$ we select a sample s_h, $s_h \subset U_h$, of size n_h according to a sample design $p_h(\cdot)$. The selection in each subpopulation is made independently. The resulting sample s is :

$$s = s_1 \cup s_2 \cup \ldots \cup s_H$$

$$|s_h| = n_h \Rightarrow |s| = \sum_{h=1}^{H} n_h.$$

Then the probability of selecting s is :

$$p(s) = p_1(s_1) \dots p_H(s_H).$$

Let π_k^h for $k \in \mathcal{U}$ and π_{kl}^h for $k \neq l \in \mathcal{U}$ be the first and second order inclusion probabilities according to p_h, $h = 1, \dots, H$. Then, π_k, the first order inclusion probability according to $p(\cdot)$, is equal to π_k^h if $k \in s_h$ and π_{kl}, the second order inclusion probability according to $p(\cdot)$, is equal to $\pi_k^h \pi_l^{h'}$ if k, l belong to different strata h, h' and equal to π_{kl}^h if k, l belong to the same stratum h.

We suppose that N_h is known for all $h = 1, \dots, H$. The population total can be written as :

$$t_y = \sum_{\mathcal{U}} y_k = \sum_{h=1}^{H} t_h = \sum_{h=1}^{H} N_h \overline{y}_{U_h}$$

where $t_h = \sum_{\mathcal{U}_h} y_k$ is the total of the stratum h and $\overline{y}_{U_h} = \frac{1}{N_h} t_h$ is the mean of the stratum h.

We can formulate the following result for the π-estimator in the case of stratified sampling :

Result 1.2.7 *: Under a stratified sampling design, the π-estimator for the total of the population is :*

i. $\hat{t}_\pi = \sum_{h=1}^{H} \hat{t}_{h\pi}$ *where* $\hat{t}_{h\pi}$ *is the π-estimator for the stratum h.*

ii. The variance has the expression $V(\hat{t}_\pi) = \sum_{h=1}^{H} V_h(\hat{t}_{h\pi})$*, where* $V_h(\hat{t}_{h\pi})$ *is the variance of* $\hat{t}_{h\pi}$ *for all h.*

iii. An unbiased variance estimator is given by $\hat{V}(\hat{t}_\pi) = \sum_{h=1}^{H} \hat{V}_h(\hat{t}_{h\pi})$*, where* $\hat{V}_h(\hat{t}_{h\pi})$ *is the variance estimator for* V_h *for all h.*

The selection of the sample s_h can be done differently or in the same way in all strata. For example, we can choose all s_h by sampling without replacement or by Bernoulli sampling. In each case, formulas for variance and for an estimation for variance can be obtained using the results derived for the designs **SI**, **BE**. For stratified sampling with sampling without replacement in each stratum, we obtain :

i. $\hat{t}_\pi = \sum_{h=1}^{H} N_h \left(\sum_{s_h} \frac{y_h}{n_h} \right)$.

ii. $V(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} S_{yU_h}^2$; where $S_{yU_h}^2 = \frac{1}{N_h - 1} \sum_{\mathcal{U}_h} (y_k - \overline{y}_{U_h})^2$ is the stratum variance and $f_h = \frac{n_h}{N_h}$ the sampling fraction in stratum h.

iii. $\hat{V}(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{ys_h}^2$ where $S_{ys_h}^2 = \frac{1}{n_h - 1} \sum_{s_h} (y_k - \overline{y}_{s_h})^2$ is the sample variance in stratum h.

The variance evaluation by POULPE

In the case of stratified sampling, calculating the variance estimate is not difficult because of the independence of designs in each of the stratum. For the stratified sampling, with one of the four particular schemas of simple sampling which are implemented in `Poulpe`, the variance estimation can be easily calculated.

1.2.8 Multi-stage sampling

i. The population $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$ is now partitioned into N_I primary sampling units called PSUs, $\mathcal{U}_1, \ldots, \mathcal{U}_i, \ldots, \mathcal{U}_{N_I}$ of size $|\mathcal{U}_i| = N_i$ for $i = 1, \ldots, N_I$ with N_i often unknown. For simplicity, we note :

$$\mathcal{U}_I = \{1, \ldots, i, \ldots, N_I\}.$$

In the first stage a sample $s_I, s_I \subset \mathcal{U}_I$ of PSUs is drawn according to a sampling design $p_I(\cdot)$.

ii. For each $i \in \mathcal{U}_I$, $\mathcal{U}_i$ is partitioned into N_{IIi} secondary sampling units, SSUs, $\mathcal{U}_{i1}, \ldots, \mathcal{U}_{iq}, \ldots, \mathcal{U}_{iN_{IIi}}$ symbolically represented by

$$\mathcal{U}_{IIi} = \{1, \ldots, q, \ldots, N_{IIi}\}.$$

In the second stage, for each $i \in s_I$, a sample s_{IIi} is selected from $\mathcal{U}_{IIi}$ according to a sampling design $p_{IIi}(\cdot)$.

iii. We repeat the two previous steps until the r-th stage when the r-th sampling units are the population elements.

The general procedure will be referred to the $r-1$ subsequent stages. We suppose that we have invariance and independence with respect to the $r-1$ stages of selection.
In multi-stage sampling we have the inclusion probabilities according to the first stage :

$$\pi_{Ii}, \pi_{Iij}, \text{ for } i \neq j \text{ and } i, j \in s_I$$

$$\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}, \quad \check{\Delta}_{Iij} = \frac{\Delta_{Iij}}{\pi_{Iij}}.$$

Let t_y be the population total :

$$t_y = \sum_{\mathcal{U}} y_k = \sum_{\mathcal{U}_\mathcal{I}} t_i;$$

where $t_i = \sum_{\mathcal{U}_i} y_k$ is the total of $\mathcal{U}_i$. We assume that we can build the π-estimator $\hat{t}_{i\pi}$ for t_i with respect to the last $r-1$ stages of selection, $E(\hat{t}_{i\pi}|s_I) = t_i$, and let $V_i = V(\hat{t}_{i\pi}|s_I)$ be the variance of $\hat{t}_{i\pi}$; let $\hat{V}_i$ be an unbiased estimator of V_i, given s_I, namely $E(\hat{V}_i|s_I) = V_i$. With these notations, we can give the result for the π-estimator for the population total and also the expressions for the variance and variance estimator in multi-stage sampling.

Result 1.2.8 *In r-stage sampling, $r \geq 2$, we have :*

i. the estimator $\hat{t}_\pi = \sum_{s_I} \frac{\hat{t}_{i\pi}}{\pi_{Ii}}$ is unbiased for t_y;

ii. the variance of $\hat{t}_\pi$ is :

$$V(\hat{t}_\pi) = \sum_{\mathcal{U}_\mathcal{I}} \sum_{\mathcal{U}_\mathcal{I}} \Delta_{Iij} \check{t}_i \check{t}_j + \sum_{\mathcal{U}_\mathcal{I}} \frac{V_i}{\pi_{Ii}};$$

iii. an unbiased variance estimator is :

$$\hat{V}(\hat{t}_\pi) = \sum_{s_I} \sum_{s_I} \check{\Delta}_{Iij} \frac{\hat{t}_{i\pi}}{\pi_{Ii}} \frac{\hat{t}_{j\pi}}{\pi_{Ij}} + \sum_{s_I} \frac{\hat{V}_i}{\pi_{Ii}}.$$

Proof :

$$1. E(\hat{t}_\pi) = E_I E(\hat{t}_\pi|s_I) = E_I \left(\sum_{s_I} \frac{1}{\pi_{Ii}} E(\hat{t}_{i\pi}|s_I) \right) = E_I \left(\sum_{s_I} \frac{t_i}{\pi_{Ii}} \right) = \sum_{\mathcal{U}_\mathcal{I}} t_i = t.$$

$$\begin{aligned} 2. V(\hat{t}_\pi) &= V_I\{E(\hat{t}_\pi|s_I)\} + E_I\{V(\hat{t}_\pi|s_I)\} = V_I \left\{ \sum_{s_I} \frac{t_i}{\pi_{Ii}} \right\} + E_I \left\{ \sum_{s_I} \frac{V_i}{\pi_{Ii}^2} \right\} \\ &= \sum_{\mathcal{U}_\mathcal{I}} \sum_{\mathcal{U}_\mathcal{I}} \Delta_{Iij} \check{t}_i \check{t}_j + \sum_{\mathcal{U}_\mathcal{I}} \frac{V_i}{\pi_{Ii}}. \end{aligned}$$

$$\begin{aligned} 3. E(\hat{V}(\hat{t}_\pi)) &= E_I E(\hat{V}(\hat{t}_\pi)|s_I) = E_I \left\{ \sum_{s_I} \sum_{s_I} \check{\Delta}_{Iij} \check{t}_i \check{t}_j + \sum_{s_I} (1 - \pi_{Ii}) \frac{V_i}{\pi_{Ii}^2} + \sum_{s_I} \frac{V_i}{\pi_{Ii}} \right\} \\ &= \sum_{\mathcal{U}_\mathcal{I}} \sum_{\mathcal{U}_\mathcal{I}} \Delta_{Iij} \check{t}_i \check{t}_j + \sum_{\mathcal{U}_\mathcal{I}} \frac{V_i}{\pi_{Ii}} = V(\hat{t}). \end{aligned}$$

□

For $r = 2$, we obtain a two-stage sampling and for $r = 3$ a three-stage. From the general result, we can derive formulas for the variance and variance estimator in the particular case of two-stage sampling.

In this case, the second stage will consist in selecting for every $i \in s_I$, a sample s_i of elements from U_i, $s_i \subset U_i$ according to a design $p_i(.|s_I)$. Let $\pi_{k|i}$, $\pi_{k,l|i}$ be the first and second order inclusion probabilities with respect to the second stage sampling, $p_i(\cdot|s_I)$.
The resulting sample is

$$s = \bigcup_{i \in s_I} s_i.$$

We suppose that we have invariance and independence with respect to the second stage of selection. For every $i \in s_I$ let n_i be the size of s_i, then the size of s is :

$$n_s = \sum_{i \in s_I} n_i.$$

We have $\hat{t}_{i\pi}$ an unbiased estimator for t_i with respect to the second stage ; it results then the expression for $\hat{t}_{i\pi}$:

$$\hat{t}_{i\pi} = \sum_{s_i} \frac{y_k}{\pi_{k|i}}$$

From the general result for multi-stage sampling, we have the π-estimator for the total of the population, $t_y = \sum_{\mathcal{U}} y_k$ of the form :

$$\hat{t}_\pi = \sum_{i \in s_I} \frac{\hat{t}_{i\pi}}{\pi_{Ii}} = \sum_{i \in s_I} \frac{1}{\pi_{Ii}} \left(\sum_{k \in s_i} \frac{y_k}{\pi_{k|i}} \right)$$

with the variance :

$$V(\hat{t}_\pi) = V_{PSU} + V_{SSU}$$

where $V_{PSU} = \sum_{\mathcal{U}_\mathcal{I}} \sum_{\mathcal{U}_\mathcal{I}} \Delta_{Iij} \check{t}_i \check{t}_j$ is the variance due to the first stage sampling and $V_{SSU} = \sum_{\mathcal{U}_\mathcal{I}} \frac{V_i}{\pi_{Ii}}$ is the variance due to the second stage sampling. We have $V_i = V(\hat{t}_{i\pi}|s_I) = \sum_{\mathcal{U}_i} \sum_{\mathcal{U}_i} \Delta_{kl|i} \check{y}_{k|i} \check{y}_{l|i}$. An unbiased estimator for V_i is :

$$\hat{V}_i = \sum_{s_I} \sum_{s_I} \check{\Delta}_{kl|i} \check{y}_{k|i} \check{y}_{l|i}.$$

Then, an unbiased estimator for $V(\hat{t}_\pi)$ is :

$$\hat{V}(\hat{t}_\pi) = \sum_{s_I} \sum_{s_I} \check{\Delta}_{Iij} \frac{\hat{t}_{i\pi}}{\pi_{Ii}} \frac{\hat{t}_{j\pi}}{\pi_{Ij}} + \sum_{s_I} \frac{\hat{V}_i}{\pi_{Ii}}.$$

Remark 1.2.4 *: For particular conditions, we can obtain designs that have been already studied.*

i. for $s_I = \mathcal{U}_I$ we obtain stratified sampling and we find the formula for the variance estimator;

ii. for $s_i = \mathcal{U}_i$ for all i, then we have cluster sampling. In this case,

$$\hat{V}(\hat{t}_\pi) = \sum_{s_I} \sum_{s_I} \check{\Delta}_{Iij} \check{t}_i \check{t}_j$$

is an unbiased estimator for the variance of $\hat{t}_\pi$

The evaluation of variance estimation by POULPE

The **Result 1.2.8** is a consequence of a more general theorem of Des Raj (1966) which gives the way of calculating the variance estimation in multi-stage sampling. We suppose that we have invariance and independence with respect to the second stage of selection and we use the same notations as above. Let us consider the general unbiased estimator for the population total

$$\hat{T} = \sum_{i=1}^{N_I} w'_{I,i} \hat{t}_i$$

where $w'_{I,i}$ are weights for each sample s_I of PSUs defined as

$$w'_{I,i} = \begin{cases} w_{I,i} & \text{if } i \in s_I \\ 0 & \text{otherwise} \end{cases}$$

and $\hat{t}_i$ is an unbiased estimator for the total t_i of the PSU $\mathcal{U}_i$, $t_i = \sum_{\mathcal{U}_i} y_k$. We have $E(\hat{t}_i | s_I) = t_i$. We need $E_I(w'_{I,i}) = 1$ for all $i \in \mathcal{U}_I$ in order to ensure the unbiasedness condition for $\hat{T}$; E_I represents the expectation with respect to the first stage.
Let

$$f(T) = \sum_{i=1}^{N_I} a_{i,s} t_i^2 + \sum_{i<j=1}^{N_I} b_{ij,s} t_i t_j$$

be an unbiased estimator for the variance of $\sum_{i=1}^{N_I} w'_{I,i} \hat{t}_i$; $a_{i,s}, b_{ij,s}$ are real numbers, predetermined for every sample s_I. From the condition $E_I(f(T)) = Var_I(\sum_{i=1}^{N_I} w'_{I,i} \hat{t}_i)$ we obtain $E(a_{i,s}) = V(w'_{Ii})$. The theorem of Des Raj states :

Result 1.2.9 *(Des Raj 1966) : An unbiased variance for the population total $t_y = \sum_{\mathcal{U}} y_k$ in two-stage sampling is given by*

$$\hat{V}(\hat{T}) = f(\hat{T}) + \sum_{s_I} w_{I,i} \hat{V}_i$$

where $\hat{V}_i$ is un unbiased estimator for $V_i = Var(\hat{t}_i | s_I)$ and $f(\hat{T})$ is obtained from $f(T)$ after replacing t_i with $\hat{t}_i$ for all $i \in \mathcal{U}_I$.

A variance estimator in two-stage sampling is obtained as follows :

i. We must have an unbiased estimator $f(T)$ for the variance in the case of one-stage sampling and substitute all t_i with $\hat{t}_i$ in the expression of $f(T)$;

ii. We must have an unbiased estimator $\hat{T}$ for the population total in the case of one-stage sampling and substitute all t_i with $\hat{V}_i$.

iii. Finally, we make the sum of the two quantities from above.

So, we only need to know the expressions for the estimator of the total and of the variance, both of them in one-stage sampling. In multi-stage sampling, the mechanism of calculating the variance estimation by `Poulpe` is based on the recurrence induced by this theorem. At the same time with selecting the sample elements, the software creates the tree structure, from top to bottom and containing the elements :

i. the root : the population $\mathcal{U}$ divided in N_I units, called PSU(primary sampling units) ;

ii. the next level contains the selected sample of PSU, s_I, according to a pattern $p_I(.)$;

iii. for each element from s_I,we repeat the second step as many times as degrees we have ; the terminal elements from the created tree will be sample elements.

iv. the links between two successive levels are made with the help of branches which contain the information of the used sampling pattern and of the probabilities of inclusion corresponding to the level.

In order to calculate the variance estimation, we will climb up the tree in the inverse sense, level by level, from the terminal elements to the root and we will apply each time the theorem of Des Raj.
For three-stage sampling, Caron *et al* (1998) give the explicit expressions of the variance estimator corresponding to each stage.

1.2.9 Two-phase sampling

We suppose now that the framework is more general than in two-stage sampling, namely we eliminate the restrictions of independence and invariance in the second stage. The strategy can be described as follows :

i. in the first phase, a sample $s_a \subset \mathcal{U}$ is selected according to a sampling design $p_a(\cdot); |s_a| = n_{s_a}$.

ii. in the second phase, we select a sample $s \subset s_a$, according to $p(\cdot|s_a); |s| = n_s$.

The quantities required for the construction of an unbiased estimator for t, are :

i. the inclusion probabilities with respect to the first phase :

$$\pi_{ak}, \pi_{akl} \quad k, l \in \mathcal{U}$$

$$\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}.$$

ii. given s_a, the inclusion probabilities with respect to the second phase :

$$\pi_{k|s_a}, \pi_{kl|s_a} \text{ for } k, l \in s_a$$

$$\Delta_{kl|s_a} = \pi_{kl|s_a} - \pi_{k|s_a}\pi_{l|s_a}.$$

We introduce the quantities $\pi_k^* = \pi_{ak}\pi_{k|s_a}$ and consider the π^*-estimator for the population total, which attaches the new weights $\dfrac{1}{\pi_k^*}$ to y_k :

$$\hat{t}_{\pi^*} = \sum_s \frac{y_k}{\pi_k^*}.$$

Then, $\hat{t}_{\pi^*}$ has the following properties :

Result 1.2.10 *Consider a two-phase sampling, then the population total $t_y = \sum_{\mathcal{U}} y_k$ is estimated without bias by :*

i. $\hat{t}_{\pi^*} = \sum_s \dfrac{y_k}{\pi_k^*}$

ii. The variance of the π^*-estimator is given by

$$V(\hat{t}_{\pi^*}) = \underbrace{\sum_{\mathcal{U}}\sum_{\mathcal{U}} \Delta_{akl}\check{y}_{ak}\check{y}_{al}}_{\textit{1st phase variance}} + \underbrace{E_{p_a}\left(\sum_{s_a}\sum_{s_a} \Delta_{kl|s_a}\frac{y_k}{\pi_k^*}\frac{y_l}{\pi_l^*}\right)}_{\textit{2nd phase variance}}$$

where $\check{y}_{ak} = \dfrac{y_k}{\pi_{ak}}$

iii. An unbiased variance estimator is given by

$$\hat{V}(\hat{t}_{\pi^*}) = \underbrace{\sum_s\sum_s \frac{\Delta_{akl}}{\pi_{kl}^*}\check{y}_{ak}\check{y}_{al}}_{\textit{1st phase variance estimator}} + \underbrace{\sum_s\sum_s \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}}\frac{y_k}{\pi_k^*}\frac{y_l}{\pi_l^*}}_{\textit{2nd phase variance estimator}}.$$

Proof

i. We have $E(\hat{t}_{\pi^*}) = E_{p_a}E(\hat{t}_{\pi^*}|s_a) = E_{p_a}\left(\sum_{s_a}\dfrac{y_k}{\pi_{ak}}\right) = \sum_{\mathcal{U}} y_k = t_y$ so $\hat{t}_{\pi^*}$ is unbiased for t.

ii. We have the relation

$$V(\hat{t}_{\pi^*}) = V_{p_a} E(\hat{t}_{\pi^*}|s_a) + E_{p_a} V(\hat{t}_{\pi^*}|s_a).$$

We can write $\hat{t}_{\pi^*} - t = \left(\sum_{s_a} \check{y}_{ak} - \sum_{\mathcal{U}} y_k\right) + \left(\sum_s \frac{y_k}{\pi_k^*} - \sum_{s_a} \check{y}_{ak}\right) = Q_{s_a} + R_s$, where Q_{s_a} is the error due to the first phase and R_s the error due to the second phase. It results $V_{p_a} E(\hat{t}_{\pi^*}|s_a) = V_{p_a}(Q_{s_a}) = \sum\sum_{\mathcal{U}} \Delta_{akl} \check{y}_{ak} \check{y}_{al}$ and

$$E_{p_a} V(\hat{t}_{\pi^*}|s_a) = E_{p_a} V(R_s|s_a) = E_{p_a}\left(\sum\sum_{s_a} \Delta_{kl|s_a} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*}\right).$$

iii. $E\left(\sum_s \sum_s \frac{\Delta_{akl}}{\pi_{kl}^*} \check{y}_{ak} \check{y}_{al}\right) = E_{p_a}\left(\sum_{s_a} \sum_{s_a} \frac{\Delta_{akl}}{\pi_{akl}} \check{y}_{ak} \check{y}_{al}\right) = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{akl} \check{y}_{ak} \check{y}_{al}$ and for the second component

$$E\left(\sum_s \sum_s \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*}\right) = E_{p_a}\left(\sum_{s_a} \sum_{s_a} \Delta_{kl|s_a} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*}\right).$$

□

The treatment by POULPE of variance estimation in two-phase sampling

Let us see how this situation is treated by the software program `Poulpe`. We will use the notations and expression for the variance estimator given in **Result 1.2.10**. We remind that for a given first-phase design $p_a(\cdot)$, the software `Poulpe` can calculate double sums of the following form representing the variance estimation for the total of a variable $\mathcal{Z}$:

$$\sum_{k \in s_a} \sum_{l \in s_a} \frac{\Delta_{akl}}{\pi_{akl} \pi_{ak} \pi_{al}} z_k z_l = \sum_{k \in s_a} \sum_{l \in s_a} A_{kl} z_k z_l$$

where $A_{kl} = \dfrac{\pi_{akl} - \pi_{ak}\pi_{al}}{\pi_{akl}\pi_{ak}\pi_{al}}$ for $k \neq l$ and $A_{kk} = \dfrac{1 - \pi_k}{\pi_k^2}$. In the above expression, the quantities A_{kl} are not given explicitly in the formula of variance estimation, but are obtained recursively and implicitly in the software (because the π_{akl} are not known) by means of exact or approximative formulas allowing to estimate the variance at each phase. The variance estimation due to the second phase can be written in the form :

$$\sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}} \frac{y_k}{\pi_{ak}\pi_{k|s_a}} \frac{y_l}{\pi_{al}\pi_{l|s_a}} = \sum_{k \in s} \sum_{l \in s} B_{kl} z_k z_l$$

where $B_{kl} = \dfrac{\pi_{kl|s_a} - \pi_{k|s_a}\pi_{l|s_a}}{\pi_{k|s_a}\pi_{l|s_a}\pi_{kl|s_a}}$ for $k \neq l$ and $B_{kk} = \dfrac{1-\pi_{k|s_a}}{\pi^2_{k|s_a}}$. Because all B_{kl} for $k, l \in s$ can be calculated recursively from the files of data, the second term from the expression of variance estimation can be calculated by `Poulpe`. As for the variance estimator due to the first phase, this can be written in the equivalent form :

$$\sum_{k\in s}\sum_{l\in s}\frac{A_{kl}}{\pi_{kl|s_a}}y_k y_l$$

with A_{kl} given above. It is not possible to calculate this expression because the A_{kl} are not known in general. Nevertheless, it is possible to calculate these quantities for two particular cases of great importance in practice : two-phases sampling design with a Poisson schema in the second phase and stratified two-phase survey with a simple sampling without replacement in each stratum.

Two-phase sample survey with a Poisson design in the second phase

In this case, the inclusion probabilities of second order corresponding to the second phase satisfy the relation, due to the independence of selection the elements in a Poisson schema :

$$\pi_{kl|s_a} = \pi_{k|s_a}\pi_{l|s_a} \text{ if } k \neq l, k, l \in s.$$

It results that $\Delta_{kl|s_a} = 0$ for $k \neq l$ and $k, l \in s$. For $A_{kl} = \dfrac{\pi_{akl} - \pi_{ak}\pi_{al}}{\pi_{akl}\pi_{ak}\pi_{al}}$ the variance estimation can be written as (Caron *et al.* 1998) :

$$\begin{aligned}
\hat{V}(\hat{t}_{\pi^*}) &= \sum_{k\in s}\sum_{l\in s}\frac{A_{kl}}{\pi_{kl|s_a}}y_k y_l + \sum_{k\in s}\frac{\pi_{k|s_a}(1-\pi_{k|s_a})}{\pi_{k|s_a}}\left(\frac{y_k}{\pi_{ak}\pi_{k|s_a}}\right)^2 \\
&= \sum_{k\in s}\sum_{l\in s}\frac{A_{kl}}{\pi_{k|s_a}\pi_{l|s_a}}y_k y_l + \sum_{k\in s}A_{kk}y_k^2\left(\frac{1}{\pi_{k|s_a}} - \frac{1}{\pi^2_{k|s_a}}\right) + \sum_{k\in s}\frac{1-\pi_{k|s_a}}{\pi^2_{k|s_a}}\frac{y_k^2}{\pi^2_{ak}} \\
&= (1) + (2) + (3).
\end{aligned}$$

We can write (1) as follows $(1) = \sum_{k\in s_a}\sum_{l\in s_a}A_{kl}y_k^* y_l^*$ with

$$y_k^* = \begin{cases} \frac{y_k}{\pi_{k|s_a}} & \text{if } k \in s \\ 0 & \text{otherwise} \end{cases}$$

It results that (1) can be calculated by the help of the tree structure of the program, while (2) and (3) are directly computed from the files of the existing data.

Stratified two-phase sampling with simple sampling without replacement in the second phase

In this case, we have the following sampling design :

i. in the first-phase, we select a sample $s_a, s_a \in \mathcal{U}$ according to $p_a(.)$; s_a is divided into H strata :

$$s_a = \bigcup_{h=1}^{H} s_{ah} \text{ with } |s_{ah}| = N_h \text{ elements ;}$$

ii. in the second phase, for all $h = 1, \ldots, H$ we select a sample $s_h, s_h \in s_{ah}$ of size $n_h = f_h N_h$ by simple sampling without replacement. The resulting sample is $s = \bigcup_{h=1}^{H} s_h$.

The inclusion probabilities corresponding to the second phase have the following expressions

$$\pi_{k|s_a} = f_h \text{ if } k \in s_{ah;}$$

$$\pi_{kl|s_a} = f_h f_{h'} \text{ if } k \in s_{ah}, l \in s_{ah'}, h \neq h'$$

$$\pi_{kl|s_a} = f_h \frac{n_h - 1}{N_h - 1} \text{ if } k, l \in s_{ah}, k \neq l.$$

As a consequence, the π^*-estimator becomes

$$\begin{aligned} \hat{t}_{\pi^*} &= \sum_{h=1}^{H} \sum_{k \in s_h} \frac{y_k}{\pi_{ak}\pi_{k|s_a}} = \sum_{h=1}^{H} N_h \left[\frac{1}{n_h} \sum_{k \in s_h} \underbrace{\frac{y_k}{\pi_{ak}}}_{u_k} \right] \\ &= \sum_{h=1}^{H} N_h \frac{1}{n_h} \sum_{k \in s_h} u_k = \sum_{h=1}^{H} N_h \overline{u}_h \end{aligned}$$

with an estimator for the variance

$$\hat{V}(\hat{t}_{\pi^*}) = \underbrace{\sum_s \sum_s \frac{A_{kl}}{\pi_{kl|s_a}} y_k y_l}_{(1)} + \underbrace{\sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} \left[\frac{1}{n_h - 1} \sum_{s_h} (u_k - \overline{u_h})^2 \right]}_{(2)}.$$

Variance estimation by means of POULPE

The expression (2) from the above formula can be calculated directly from the file of data. We write part (1) as the sum of three terms, as follows :

$$(1) = \sum_s \sum_s A_{kl} \frac{y_k}{\pi_{k|s_a}} \frac{y_l}{\pi_{l|s_a}}$$

$$+ \quad \sum_{h=1}^{H} \sum_{s_h} \sum_{s_h} A_{kl} \frac{y_k}{\pi_{k|s_a}} \frac{y_l}{\pi_{l|s_a}} \frac{1-f_h}{n_h-1}$$

$$+ \quad \sum_{h=1}^{H} \sum_{s_h} A_{kk} y_k^2 \frac{n_h - N_h}{f_h(n_h-1)}$$

The last term of the sum can be calculated directly from the files of recorded dates ; the sum of the first two terms is equal to :

$$\sum_{s_a} \sum_{s_a} A_{kl} z_k^0 z_l^0 + \sum_{h=1}^{H} \sum_{s_a} \sum_{s_a} A_{kl} z_k^h z_l^h$$

where

$$z_k^0 = \begin{cases} \frac{y_k}{\pi_{k|s_a}} & \text{for } k \in s \\ 0 & \text{otherwise} \end{cases}$$

and

$$z_k^h = \begin{cases} \frac{y_k}{\pi_{k|s_a}} \sqrt{\frac{1-f_h}{n_h-1}} & \text{for } k \in s_h \\ 0 & \text{otherwise} \end{cases}$$

$h = 1, \ldots, H$ and we have reduced the calculation to the sum of two terms for which `Poulpe` knows how to calculate the variance estimation.

1.3 Taylor linearization

In the previous section, a π-estimator was presented for the population total, with special attention on the variance estimator and its expression for different sampling designs.

But, in practice, in most of the cases, we are confronted with surveys that involve not only one, but several variables of interest and more generally, the unknown quantity depends on these variables through a general function. An example is the ratio of two unknown population totals

$$\mathcal{R} = \frac{\sum_{\mathcal{U}} y_k}{\sum_{\mathcal{U}} z_k} = \frac{t_y}{t_z}$$

where y and z are two variables of study. First, we will study the linear case, then the general case will be considered by using Taylor series expansions. The use of Taylor series implies the introduction of supplementary conditions on the population and on the sampling design. They will permit the development in Taylor series and also the convergence of it. The subject was treated by Wolter (1985), Särndal *et al.* (1992). We will give a short review of the main results.

1.3.1 π estimators for the linear case

Suppose that there are J variables of study $\mathcal{Y}_1, \ldots, \mathcal{Y}_j, \ldots, \mathcal{Y}_J$, and let y_{jk} be the value of $\mathcal{Y}_j$ for the k-th element of the population, for all $j = 1, \ldots, J$ and all $k = 1, \ldots, N$. Let $t_j = \sum_{\mathcal{U}} y_{jk}$ be the total of the $\mathcal{Y}_j$ variable for all $j = 1, \ldots, J$. The objective is to estimate these quantities, namely the components of the following vector :

$$\mathbf{t} = (t_1, \ldots, t_j, \ldots, t_J)'$$

For each variable of interest the theory of the π estimator, presented in the first section, can be applied. A sample s is drawn from $\mathcal{U}$, according to a sampling design $p(s)$, with the probabilities of inclusion of first and second order, π_k and π_{kl} and for all $k \in s$ we observe the value of the vector :

$$\mathbf{y_k} = (y_{1k}, \ldots, y_{jk}, \ldots, y_{Jk})'$$

and each t_j total is estimated by the π-estimator $\hat{t}_{j\pi} = \sum_s \breve{y}_{jk}$ so that the π-estimator of $\mathbf{t}$ is

$$\hat{\mathbf{t}}_\pi = (\hat{t}_{1\pi}, \ldots, \hat{t}_{j\pi}, \ldots, \hat{t}_{J\pi})'$$

We have the following results :

Result 1.3.1 *(Särndal et al. 1992) : The variance-covariance matrix of $\hat{t}_\pi$ has the following expression :*

$$V(\hat{\mathbf{t}}_\pi) = E\left((\hat{\mathbf{t}}_\pi - \mathbf{t})(\hat{\mathbf{t}}_\pi - \mathbf{t})'\right)$$

and is a symmetric matrix with the j-th diagonal element given by the variance of $\hat{t}_{j\pi}$ and the elements jj' given by the covariance of $\hat{t}_{j\pi}$ and $\hat{t}_{j'\pi}$:

$$V(\hat{t}_{j\pi}) = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \check{y}_{jk} \check{y}_{jl}$$

and

$$C(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \check{y}_{jk} \check{y}_{j'l}$$

The matrix $V(\hat{\mathbf{t}}_\pi)$ is estimated with no bias by the matrix $\hat{V}(\hat{\mathbf{t}}_\pi)$ such that the jth diagonal element is

$$\hat{V}(\hat{t}_{j\pi}) = \sum_{s} \sum_{s} \check{\Delta}_{kl} \check{y}_{jk} \check{y}_{jl}$$

and the jj' element is :

$$\hat{C}(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) = \sum_{s} \sum_{s} \check{\Delta}_{kl} \check{y}_{jk} \check{y}_{j'l}$$

We consider now a more general situation, namely when the requested estimator for a parameter population θ can be written as follows :

$$\theta = f(t_1, \ldots, t_J)$$

and f is a linear function. Then, applying the above result, an estimator for θ is given by :

$$\hat{\theta} = f(\hat{t}_1, \ldots, \hat{t}_J).$$

This is derived from the fact that if f is a linear function then θ can be written as :

$$\theta = a_0 + \sum_{j=1}^{J} a_j t_j = f(t_1, \ldots, t_J).$$

Consequently $\hat{\theta} = a_0 + \sum_{j=1}^{J} a_j \hat{t}_{j\pi} = f(\hat{t}_{1\pi}, \ldots, \hat{t}_{J\pi})$ where $\hat{t}_{j\pi} = \sum_{s} \frac{y_{jk}}{\pi_k}$ is the π-estimator for the total t_j. We can apply now the conclusions from the previous result in order to obtain the expressions for the variance-covariance matrix of θ and also for an estimator of this matrix.

Result 1.3.2 *For a parameter of the population, having the form :*

$$\theta = a_0 + \sum_{j=1}^{J} a_j t_j = f(t_1, \ldots, t_J)$$

an estimator is given by :

$$\hat{\theta} = f(\hat{t}_{1\pi}, \ldots, \hat{t}_{J\pi})$$

with the variance-covariance matrix :

$$V(\hat{\theta}) = \sum_{j=1}^{J} \sum_{j'=1}^{J} a_j a_j' C(\hat{t}_{j\pi}, \hat{t}_{j'\pi})$$

$$\textit{and } \hat{V}(\hat{\theta}) = \sum_{j=1}^{J} \sum_{j'=1}^{J} a_j a_j' \widehat{C}(\hat{t}_{j\pi}, \hat{t}_{j'\pi})$$

where $C(\hat{t}_{j\pi}, \hat{t}_{j'\pi})$ *and* $\widehat{C}(\hat{t}_{j\pi}, \hat{t}_{j'\pi})$ *are given in the result from above.*

As a conclusion, if a parameter of interest is expressed by a linear function of the totals of q variables of study, then the expressions for V and $\hat{V}$ can be deduced in a simple way.

1.3.2 The general case

The case when f is not a linear function will be reduced to the former one. The main idea is to approximate f around the true value by a linear function, for which we know how to derive formulas for the variance and for an estimator of the variance, according to the case one. Under general conditions, we will show how the variance of the estimator can be approximated by the variance of a linear estimator. The approximation will be made by a first-order Taylor series expansion and the method is called *Taylor linearization*. We start with a consistent estimator $\hat{t}$ for t and a function f which satisfies several conditions, the expansion in Taylor series of f around the point $t = (t_1, \ldots, t_J)$ will give :

i. an approximate expression for the design variance of $\hat{\theta} = f(\hat{t})$;

ii. a suitable estimator of the variance of $\hat{\theta}$.

But for a finite population $\mathcal{U}$ we can not define the consistency and asymptotic unbiasedness of an estimator. For achieving these conditions and also for allowing us to develop f in a Taylor series with a remainder of low order, we need supplementary conditions and mathematical results concerning the behaviour of a finite population and of the probabilities π_k, π_{kl} when n and N increase to infinite simultaneously. In order to obtain an infinite population, we will consider the initial population $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$ with the corresponding π_k.
We give here Särndal's way (1980) to obtain an infinite population. Isaki & Fuller (1982), Särndal *et al.* 1992 propose different ones. Särndal (1980) reproduces this population t-1 times. For all t, a sample s is selected from each $\mathcal{U}$ according to $p(s)$, with the same π_k, for all t. The resulting population will have $N_t = Nt$ elements from which we select a sample $s_{(t)}$ consisting of $n_{s_{(t)}} = \sum_{\gamma=1}^{t} n_{s_\gamma} = nt$ elements. Next, we allow $t \to \infty$ and it results that $N_t \to \infty$ and $n_{s(t)} \to \infty$ but n and N remains constant. This framework allows us to define the properties of consistency and asymptotic unbiasedness

for an estimator. Now we can derive the results for Taylor linearization. Let us give two results which will be used in the next.

Result 1.3.3 *(Wolter 1985) : Let $f : \xi_J \to R$ be a real valued function defined on a q dimensional Euclidian space, continuous, differentiable with continuous partial derivatives of order two in an open sphere containing $\mathbf{a} = (a_1, \ldots, a_J)'$. For $\mathbf{X}_n = (X_{1n}, \ldots, X_{Jn})'$ satisfying :*

$$\mathbf{X}_n = \mathbf{a} + O_p(r_n) \text{ where } r_n \to 0,$$

we have :

$$f(\mathbf{X}_n) = f(\mathbf{a}) + \sum_{j=1}^{J} (X_{jn} - a_j) \frac{\partial f}{\partial x_j}(\mathbf{a}) + O_p(r_n^2).$$

Result 1.3.4 *(Wolter 1985) : With the above conditions and*

$$E(\mathbf{X}_n) = \mathbf{a}$$

$$E\left((\mathbf{X}_n - \mathbf{a})(\mathbf{X}_n - \mathbf{a})'\right) = \mathbf{\Sigma}_n < \infty$$

the asymptotic variance of $f(\mathbf{X}_n)$ is :

$$E\left(f(\mathbf{X}_n) - f(\mathbf{a})\right)^2 = \mathbf{d}\mathbf{\Sigma}_n\mathbf{d}' + O_p(r_n^3)$$

where $\mathbf{d} = \left(\frac{\partial f}{\partial x_j}(\mathbf{a})\right)_{j=\overline{1,J}}$.

Let's suppose that :

i. $N^{-1}t_j$ has a finite limit for all $j = 1, \ldots, J$.

ii. $N^{-1}(\hat{t}_{j\pi} - t_j) \to 0$ in probability, for all $j = 1, \ldots, J$.

iii. $n^{\frac{1}{2}}N^{-1}(\hat{t}_{j\pi} - t_j) \to N(0, A)$ in distribution for all $j = 1, \ldots, J$; it results then $N^{-1}(\hat{t}_{j\pi} - t_j) = O_p(n^{-\frac{1}{2}})$.

iv. It exists an estimator for the variance of $\hat{t}_{j\pi}$ for all $j = 1, \ldots, J$.

As mentioned, we want to estimate a parameter expressed by a nonlinear function of totals of J variables of interest :

$$\theta = f(t_1, \ldots, t_J) \text{ where } t_j = \sum_{\mathcal{U}} y_{kj}, j = 1, \ldots, J.$$

with $f : \xi_J \to R$ continuous, differentiable with continuous partial derivatives.

We define an estimator for θ by substitution. It is obtained by replacing each total with the corresponding π estimator :

$$\hat{\theta} = f(\hat{t}_{1\pi}, \ldots, \hat{t}_{J\pi}).$$

We can take as vectors $\mathbf{a}$ and $\mathbf{X}_n$ from **Result 1.3.3** and **Result 1.3.4**, the vectors $\mathbf{t}$ and $\mathbf{t}_\pi$ as defined at the beginning of this section,

$$\mathbf{t} = (t_1, \ldots, t_J)' \text{ and}$$

$$\mathbf{t}_\pi = (\hat{t}_{1\pi}, \ldots, \hat{t}_{J\pi})$$

which have the following properties (as a consequence of the conditions 1-4) :

- $N^{-1}(\mathbf{t}_\pi - \mathbf{t}) \xrightarrow{P} 0$ and
- $N^{-1}(\mathbf{t}_\pi - \mathbf{t}) = O_p(n^{-\frac{1}{2}})$
- $n^{\frac{1}{2}} N^{-1}(\mathbf{t}_\pi - \mathbf{t}) \xrightarrow{\mathcal{L}} N(0, \Sigma)$.

For $N^{-1}\mathbf{t}$, $N^{-1}\mathbf{t}_\pi$ given above and $r_n = n^{-\frac{1}{2}}$ the conditions from **Result 1.3.3** and **Result 1.3.4** are satisfied. We obtain the approximation of the parameter by a first-order Taylor series.

$$\begin{aligned} N^{-1} f(\hat{t}_{1\pi}, \ldots, \hat{t}_{J\pi}) &= N^{-1} f(t_1, \ldots, t_J) \\ &+ N^{-1} \sum_{j=1}^{J} (\hat{t}_{j\pi} - t_j) \left. \frac{\partial f(v_1, \ldots, v_J)}{\partial v_j} \right|_{(v_1,\ldots,v_J)=(t_1,\ldots,t_J)} \\ &+ O_p\left(\frac{1}{n}\right) \end{aligned}$$

and if we note with $\alpha_j = \left. \frac{\partial f(v_1,\ldots,v_J)}{\partial v_j} \right|_{(v_1,\ldots,v_J)=(t_1,\ldots,t_J)}$ for $j = 1, \ldots, J$ we obtain the equivalent formula :

$$N^{-1}\hat{\theta} = N^{-1}\theta + N^{-1} \sum_{j=1}^{J} (\hat{t}_{j\pi} - t_j)\alpha_j + O_p\left(\frac{1}{n}\right)$$

and :

$$N^{-1}\hat{\theta} = N^{-1}\theta + O_p(n^{-\frac{1}{2}}).$$

In particular $\hat{\theta}$ can be approximated by $\hat{\theta} \simeq \theta + \sum_{j=1}^{J} (\hat{t}_{j\pi} - t_j)\alpha_j$ and is approximately unbiased for θ.

We will use **Result 1.3.4** in order to derive the variance of $\hat{\theta}$. Because $B(\hat{\theta}) \simeq 0$, the variance can be approximated by the mean square error :

$$V(\hat{\theta}) \simeq \text{ MSE } (\hat{\theta}) = \boldsymbol{\alpha}\boldsymbol{\Sigma}\boldsymbol{\alpha}' + O_p(n^{-\frac{3}{2}})$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$ is the vector of the partial derivatives of f calculated in $\mathbf{t} = (t_1, \ldots, t_q)$ and $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{t}_\pi - \mathbf{t}$. We obtain then the following approximation for the variance :

$$V(\hat{\theta}) \simeq \text{ MSE } (\hat{\theta}) \simeq \boldsymbol{\alpha}\boldsymbol{\Sigma}\boldsymbol{\alpha}' = V\left(\sum_{i=1}^{J} \alpha_j(\hat{t}_{j\pi} - t_j)\right)$$

and to calculate the last expression, we apply the procedure described in the linear case. We finally obtain the following expression for the approximative variance :

$$V(\hat{\theta}) \simeq \sum_{i=1}^{J}\sum_{j=1}^{J} \alpha_i\alpha_j' \text{ Cov } (\hat{t}_{i\pi}, \hat{t}_{j\pi}) = \boldsymbol{\alpha}\boldsymbol{\Sigma}\boldsymbol{\alpha}'$$

To obtain a variance estimator, we will substitute sample-based estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$. Suppose that the estimator $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\Sigma}}$ are available ; then an estimator for MSE $(\hat{\theta})$ is given by :

$$\widehat{V}(\hat{\theta}) \simeq \widehat{\text{MSE}}\ (\hat{\theta}) = \hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\alpha}}'.$$

It is not unbiased for MSE, but taking $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\Sigma}}$ consistent estimators for $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$, it will be consistent estimator for the MSE.
As it can be observed, the expressions for approximative variance and variance estimator are complicated because they involve the calculation of the variance-covariance matrix $\boldsymbol{\Sigma}$; for the variance estimation, we need to calculate an estimator for each element of $\hat{\boldsymbol{\Sigma}}$. Woodruf (1971) gives a simple method for which the variance estimation is simplified. This method is a generalization of the Keyfitz's (1957) method for obtaining the variances for specific types of estimates derived from specific sample designs. It consists in reordering the components of the sum $\sum_{j=1}^{J} \alpha_j\hat{t}_{j\pi}$. We have :

$$\begin{aligned}\sum_{j=1}^{J} \alpha_j\hat{t}_{j\pi} &= \sum_{j=1}^{J} \alpha_j \left(\sum_s \frac{y_{jk}}{\pi_k}\right) = \sum_s \frac{1}{\pi_k}\left(\sum_{j=1}^{J} \alpha_j y_{jk}\right) \\ &= \sum_s \frac{u_k}{\pi_k} = \sum_s \check{u}_k.\end{aligned}$$

where $u_k = \sum_{j=1}^{J} \alpha_j y_{jk}$. As it can be observed, $\sum_{j=1}^{J} \alpha_j\hat{t}_{j\pi}$ can be written equivalently as the π-estimator for the total of the new introduced quantities u_k. Because the u_k for all $k \in \mathcal{U}$ depend on α_j which on their turn, depend on the $\mathcal{Y}_1, \ldots, \mathcal{Y}_q$ which are unknown we obtain that u_k are unknown and so they can not be used. The quantities α_j are estimated by $\hat{\alpha}_j$ the corresponding π-estimator

$$\hat{\alpha}_j = \left.\frac{\partial f(v_1, \ldots, v_J)}{\partial v_j}\right|_{(v_1,\ldots,v_J)=(\hat{t}_{1\pi},\ldots,\hat{t}_{J\pi})}$$

and then the u_k are estimated by :

$$\hat{u}_k = \sum_{j=1}^{J} \hat{\alpha}_j y_{kj}.$$

It results then $V(\hat{\theta}) \simeq V(\sum_s \check{u}_k)$. Now we can give the result obtained by Woodruff (1971).

Result 1.3.5 *(Woodruff 1971)*
1. An approximatively unbiased estimator for the population parameter

$$\theta = f(t_1, \ldots, t_J)$$

is given by the substitution estimator :

$$\hat{\theta} = f(\hat{t}_{1\pi}, \ldots, \hat{t}_{J\pi})$$

where $\hat{t}_{j\pi}$ is the corresponding π estimator of t_j.
2. Using the Taylor linearization, the approximative variance is :

$$V(\hat{\theta}) \simeq \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}$$

where $u_k = \sum_{j=1}^{J} \alpha_j y_{kj}$ and $\alpha_j = \left. \frac{\partial f(v_1,\ldots,v_J)}{\partial v_j} \right|_{(v_1,\ldots,v_J)=(t_1,\ldots,t_J)}$ and
3. The estimated variance has the expression :

$$\widehat{V}(\hat{\theta}) = \sum_{s} \sum_{s} \check{\Delta}_{kl} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l}$$

where $\hat{u}_k = \sum_{j=1}^{J} \hat{\alpha}_j y_{kj}$ and $\hat{\alpha}_j$ is the π estimator for α_j.

Supposing that the general conditions given by Isaki & Fuller (1982) are fulfilled, $\hat{u}_k$ is a consistent estimator for u_k and as a result $\hat{V}(\hat{\theta})$ is consistent for $V(\hat{\theta})$. For a design of fixed size, we have the alternative formulas for the approximative variance and variance estimation :

$$V(\hat{\theta}) \simeq -\frac{1}{2} \sum \sum_{\mathcal{U}} \Delta_{kl} \left(\check{u}_k - \check{u}_l\right)^2$$

$$\hat{V}(\hat{\theta}) = -\frac{1}{2} \sum \sum_{s} \check{\Delta}_{kl} \left(\check{u}_k - \check{u}_l\right)^2$$

In large sample, we estimate the approximative variance of $\hat{\theta}$ given in Woodruff's result and the found value can be considered as an estimator of the true variance.

From the above result, we can summarize and give the steps when the Taylor technique is applied, namely :

– For the population parameter $\theta = f(t_1, \ldots, t_J)$, expressed as a function of the J totals, we derive the substitution estimator $\hat{\theta}$ which is approximately unbiased for θ. The variance and variance estimator of $\hat{\theta}$ must be calculated.
– The linearized variable, u_k, is derived for all $k \in \mathcal{U}$; these quantities are calculated in the population, u_k being expressed as functions of the J totals $t_1, \ldots, t_J$.
– The unknown quantities u_k are estimated by $\hat{u}_k$.
– The parameter is approximated by $\hat{\theta} \simeq \theta + (\sum_s \check{u}_k - \sum_{\mathcal{U}} u_k)$.
– According to the result of Woodruf we have

$$V(\hat{\theta}) \simeq V\left(\sum_s \frac{u_k}{\pi_k}\right) = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}.$$

– A Horvitz-Thompson variance estimate, based on the unknown u_k can be obtained :

$$\hat{V}(\hat{\theta}) = \sum_s \sum_s \check{\Delta}_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}.$$

– The estimated variance based on the sample estimators $\hat{u}_k$ has the expression :

$$\hat{V}(\hat{\theta}) = \sum_s \sum_s \check{\Delta}_{kl} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l}$$

The evaluation of the variance by means of POULPE

From **Result 1.3.5**, it is possible for `Poulpe` to calculate the variance estimator ; we calculate in fact, the variance estimator for sample estimator of the linearized variable, $\hat{u}_k$ using the Horvitz-Thompson formula and the obtained value is an approximate expression for the variance estimator of $\hat{\theta}$.

In the following, we give several examples in which the Taylor technique is used.

Example 1. Let us consider the ratio between two unknown population totals $t_y = \sum_{\mathcal{U}} y_k$ and $t_x = \sum_{\mathcal{U}} x_k$:

$$R = \frac{t_y}{t_x} = \frac{\sum_{\mathcal{U}} y_k}{\sum_{\mathcal{U}} x_k} = f(t_y, t_x)$$

where $f(v_1, v_2) = \frac{v_1}{v_2}$; we want to estimate R.

1. The substitution estimator of R is $\hat{R} = \dfrac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}}$ where $\hat{t}_{y\pi}, \hat{t}_{x\pi}$ are the π estimators for t_y, t_x.

2. The linearized variable for R is :

$$\begin{aligned} u_k &= y_k \left.\frac{\partial f(v_1,v_2)}{\partial v_1}\right|_{(v_1,v_2)=(t_y,t_x)} + x_k \left.\frac{\partial f(v_1,v_2)}{\partial v_2}\right|_{(v_1,v_2)=(t_y,t_x)} \\ &= y_k \frac{1}{t_x} + x_k \left(\frac{-t_y}{t_x^2}\right) \\ &= \frac{1}{t_x}\left(y_k - R x_k\right). \end{aligned}$$

3. $\hat{R}$ is approximately estimated by

$$\hat{R} \simeq R + \left(\sum_s \frac{u_k}{\pi_k} - \sum_{\mathcal{U}} u_k\right) = R + \frac{1}{t_x}\sum_s \frac{y_k - R x_k}{\pi_k}.$$

because $\sum_{\mathcal{U}} u_k = 0$ in this case.

4. The approximated variance of $\hat{R}$ is :

$$V(\hat{R}) \simeq \left(\frac{1}{t_x}\right)^2 \sum_{\mathcal{U}}\sum_{\mathcal{U}} \Delta_{kl} \frac{y_k - R x_k}{\pi_k}\frac{y_l - R x_l}{\pi_l}.$$

5. The variance estimator is obtained by replacing u_k by $\hat{u}_k = \frac{1}{\hat{t}_{x\pi}}\left(y_k - \hat{R} x_k\right)$ in the expression of the estimated Horvitz-Thompson variance formula :

$$\hat{V}(\hat{R}) = \frac{1}{\hat{t}_{x\pi}^2}\sum_s\sum_s \check{\Delta}_{kl}\frac{y_k - \hat{R} x_k}{\pi_k}\frac{y_l - \hat{R} x_l}{\pi_l}.$$

Example 2. Another example is the derivation of an estimator for the mean of the population.
There are two situations : N is known and not. In the first situation, an unbiased estimator for $\overline{y}_{\mathcal{U}} = \frac{1}{N}\sum_{\mathcal{U}} y_k$ is :

$$\widehat{\overline{y}}_{\mathcal{U}\pi} = \frac{1}{N}\sum_s \frac{y_k}{\pi_k}$$

with the variance and variance estimation, respectively :

$$V(\widehat{\overline{y}}_{\mathcal{U}\pi}) = \frac{1}{N^2}\sum_{\mathcal{U}}\sum_{\mathcal{U}} \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

$$\widehat{V}(\widehat{\overline{y}}_{\mathcal{U}\pi}) = \frac{1}{N^2}\sum_s\sum_s \check{\Delta}_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

derived using the general theory of the π estimator.

For N unknown, we need an estimator for N. The quantity N can be regarded as the total of the variable $x_k = 1$ for all $k \in U$, then $N = t_x$ and the substitution estimator is $\hat{N} = \sum_s \frac{1}{\pi_k}$. Then $\overline{y}_{\mathcal{U}}$ can be regarded as a ratio of two totals variable. Using the first example, an estimator for $\overline{y}_{\mathcal{U}}$ is :

$$\tilde{y}_s = \frac{\sum_s \frac{y_k}{\pi_k}}{\sum_s \frac{1}{\pi_k}}$$

with the approximate variance :

$$V(\tilde{y}_s) \simeq \frac{1}{N^2} \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \frac{y_k - \overline{y}_{\mathcal{U}}}{\pi_k} \frac{y_l - \overline{y}_{\mathcal{U}}}{\pi_l}$$

and the variance estimator :

$$\hat{V}(\tilde{y}_s) = \frac{1}{\hat{N}^2} \sum_s \sum_s \hat{\Delta}_{kl} \frac{y_k - \tilde{y}_s}{\pi_k} \frac{y_l - \tilde{y}_s}{\pi_l}$$

The estimator $\tilde{y}_s$ gives better results even if N is known because of the reduced variability of $\tilde{y}_s$ which lacks to $\overline{y}_{\mathcal{U}}$. Another reason is that $\tilde{y}_s$ works better when we have designs of variable sizes such as Bernoulli or Poisson designs.

Example 3. Suppose we have a study variable, y, and that we have the auxiliary variable x ; x_k is the value of x for the k-th element of the population and x_k is known for all k in $\mathcal{U}$. The objective is to estimate $t_y = \sum_{\mathcal{U}} y_k$; t_y can be written :

$$t_y = t_x \frac{t_y}{t_x}$$

$t_x = \sum_{\mathcal{U}} x_k$ is a known quantity ; then $R = \frac{t_y}{t_x}$ is the ratio of two totals and we can apply the results from the first example. It follows :

i. $\hat{t}_y = t_x \hat{R}$, where $\hat{R} = \dfrac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}}$

ii. $V(\hat{t}_y) \simeq \sum \sum_{\mathcal{U}} \Delta_{kl} \dfrac{y_k - Rx_k}{\pi_k} \dfrac{y_l - Rx_l}{\pi_l}$

iii. $\hat{V}(\hat{t}_y) = \dfrac{t_x^2}{\hat{t}_{x\pi}^2} \sum_s \sum_s \check{\Delta}_{kl} \dfrac{y_k - \hat{R}x_k}{\pi_k} \dfrac{y_l - \hat{R}x_l}{\pi_l}$.

Example 4.

Let us derive an approximately unbiased estimator for the coefficient of multiple regression. Let $\mathcal{X}_1, \ldots, \mathcal{X}_q$ be q auxiliary variables and $B_1, \ldots, B_q$ be the coefficients of regression of $\mathcal{Y}$ through $\mathcal{X}_1, \ldots, \mathcal{X}_q$. We intend to obtain

an approximately unbiased estimator for the vector of regression coefficients $\mathbf{B} = (B_1, \ldots, B_q)'$ whose variance or approximative variance can be calculated. We denote by $\mathbf{x}_k$ the vector containing the values taken by the auxiliary variables for the k-th element in the population,

$$\mathbf{x_k} = (x_{1k}, \ldots, x_{qk})' \text{ for all } k \in \mathcal{U}$$

and $\mathbf{T} = \sum_{\mathcal{U}} \mathbf{x_k x_k}'$

1. The vector of regression coefficients has the expression :

$$\mathbf{B} = (B_1, \ldots, B_q)' = (\sum_{\mathcal{U}} \mathbf{x_k x_k}')^{-1} (\sum_{\mathcal{U}} \mathbf{x_k} y_k).$$

It can be observed that $\mathbf{B}$ can be written as a function of totals. Thus it is possible to apply the method of Taylor linearization. We have $\mathbf{B} = f(t_q, t_z) = \dfrac{t_q}{t_z}$ where :

$$t_q = \sum_{\mathcal{U}} q_k \text{ the total of the new variable } q_k = \mathbf{x_k} y_k, k \in \mathcal{U},$$

and

$$t_z = \sum_{\mathcal{U}} z_k \text{ the total of the new variable } z_k = \mathbf{x_k x_k'}, k \in \mathcal{U}.$$

2. The substitution estimator for $\mathbf{B}$ is :

$$\hat{\mathbf{B}} = \left(\sum_s \frac{\mathbf{x_k x_k'}}{\pi_k} \right)^{-1} \left(\sum_s \frac{\mathbf{x_k} y_k}{\pi_k} \right).$$

3. The linearized variable of $\mathbf{B}$ is :

$$\begin{aligned} u_k &= q_k \left. \frac{\partial f(v_1, v_2)}{\partial v_1} \right|_{(v_1,v_2)=(t_q,t_z)} + z_k \left. \frac{\partial f(v_1, v_2)}{\partial v_2} \right|_{(v_1,v_2)=(t_q,t_z)} \\ &= q_k \frac{1}{t_z} + z_k \left(\frac{-t_q}{t_z^2} \right) \\ &= \mathbf{T}^{-1} \mathbf{x_k} (y_k - \mathbf{x_k'} \mathbf{B}). \end{aligned}$$

4. We can obtain an approximative expression for $\hat{\mathbf{B}}$:

$$\hat{\mathbf{B}} \simeq \mathbf{B} + \left(\sum_s \check{u}_k - \sum_{\mathcal{U}} u_k \right) = \mathbf{B} + \mathbf{T}^{-1} \left(\sum_s \frac{\mathbf{x_k} y_k}{\pi_k} - \hat{\mathbf{T}} \mathbf{B} \right).$$

5. The approximated variance of $\hat{\mathbf{B}}$ has the expression :

$$V(\hat{\mathbf{B}}) \simeq \mathbf{T}^{-1} \mathbf{V} \mathbf{T}^{-1}$$

where $\mathbf{V} = (v_{jj'})_{i,j=1,\ldots,q}$, $v_{ij} = v_{ji}$ and $v_{jj'} = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \frac{x_{jk}(y_k - \mathbf{x_k'B})}{\pi_k} \frac{x_{j'l}(y_l - \mathbf{x_l'B})}{\pi_l}$.

6. The linearized variable u_k is estimated by $\hat{u}_k = \hat{\mathbf{T}}^{-1}\mathbf{x_k}(y_k - \mathbf{x_k'}\hat{\mathbf{B}})$, $\hat{\mathbf{T}} = \sum_s \frac{\mathbf{x_k x_k'}}{\pi_k}$ and the estimator for the variance of $\mathbf{B}$:

$$\hat{V}(\hat{\mathbf{B}}) = \hat{\mathbf{T}}^{-1}\hat{\mathbf{V}}\hat{\mathbf{T}}^{-1}$$

where $\hat{\mathbf{V}} = \left(\hat{v}_{jj'}\right)$ is a $q \times q$ matrix with elements :

$$\hat{v}_{jj'} = \sum_s \sum_s \check{\Delta}_{kl} \frac{\hat{u_k}}{\pi_k} \frac{\hat{u}_l}{\pi_l}.$$

Remark 1.3.1 *:The same result would have been obtained more easily if we had considered* $\mathbf{B}$ *as a ratio of the totals of the variables* q_k *and* z_k *and we had applied directly the linearized variable for a ratio derived at first example.*

1.4 Calibration technique

Until now, no use of auxiliary information was made. Or, in sample survey auxiliary information on the finite population is often used to improve the precision of estimators of the population total. Several approaches are conceivable. We have *the calibration approach* described below, which does not rely explicitly on a model with the more recent *the model-calibration approach*, or *the model-assisted approach* described in the next section and when the inference is based upon a model of superpopulation taking into account at the same time the sampling design. We present in the next the principles of *the calibration approach* developed by Deville & Särndal (1992) and Deville, Särndal & Sautory (1993) and the implementation of this method in the software `Poulpe`. We also give a brief description of the most recent extensions of this approach.

The objective is the estimation of the population total of the variable of interest $\mathcal{Y}$ denoted $t_y = \sum_{\mathcal{U}} y_k$ in the presence of univariate or multivariate auxiliary information for which the only request is that we know its population total, namely we do not need to know the value taken by an auxiliary variable for all the units in the population.

Let $\mathcal{X}_1, \ldots, \mathcal{X}_q$ be q auxiliary variables and for $k \in \mathcal{U}$ and let $\mathbf{x_k} = (x_{k1}, \ldots, x_{kq})'$ be the q-vector with the values of the auxiliary variables for the k-th element in the population. We suppose that the total $t_{\mathbf{x}} = \sum_{U} \mathbf{x_k}$ is known ; the vector $(\mathbf{x_k}, y_k)'$ is observed for all $k \in s$. Let $\hat{t}_{y\pi} = \sum_{s} \frac{y_k}{\pi_k}$ be the π-estimator of t_y and we note with $d_k = \frac{1}{\pi_k}$ the π-weight corresponding to y_k, for all k in s.

The calibration technique consists in finding a new set of weights $\{w_k\}_{k \in s}$ which satisfies the conditions :

i. $\{w_k\}_{k \in s}$ are as close as possible to $\{d_k\}_{k \in s}$ in the sense of a distance between w_k and d_k.

ii. $\{w_k\}_{k \in s}$ satisfy the calibration equations :

$$\sum_{s} w_k \mathbf{x_k} = t_{\mathbf{x}}$$

which means that the new weights must estimate well the auxiliary information.

The calibrated estimator is denoted by $\hat{t}_{yw}$, so that it is related with the w_k weights. We consider a function distance $G_k(w, d)$ such that :

i. for every fixed $d > 0, G_k(w, d) > 0$, differentiable with respect to w , strictly convex, defined on an interval $D_k(d)$ such that $d \in D_k(d)$;

ii. $G_k(d,d) = 0$;

iii. $g_k(w,d) = \dfrac{\partial G_k(w,d)}{\partial w}$ is continuous and the function that transforms the interval $D_k(d)$ in $Im_k(d)$ is one-to-one fashion.

The request that w_k would be as close as possible to d_k is equivalent to minimize the average distance $E_p\{\sum_s G_k(w_k, d_k)\}$. Deville & Särndal (1992) apply the Lagrange multipliers method which leads to the following weights, called *calibration weights* :

$$w_k = d_k F_k(\mathbf{x_k'}\boldsymbol{\lambda})$$

where $F_k(0) = 1, q_k = F_k'(0) > 0$, $d_k F_k$ is the reciprocal mapping of $g_k(\cdot, d_k)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_j, \ldots, \lambda_q)'$ is the vector of Lagrange multipliers. We determine $\boldsymbol{\lambda}$ from the calibration equations :

$$t_{\mathbf{x}} = \sum_s w_k \mathbf{x_k} = \sum_s d_k F_k(\mathbf{x_k'}\boldsymbol{\lambda})\mathbf{x_k}.$$

Deville & Särndal (1992) suppose conditions which ensure that the above equation has a unique solution belonging to $C = \cap_{k\in\mathcal{U}}\{\boldsymbol{\lambda} : \mathbf{x_k'}\boldsymbol{\lambda} \in \mathrm{Im}_k(d_k)\}$ with a probability tending to one. With $\boldsymbol{\lambda}$ determined, we can write the calibration estimator for t_y :

$$\hat{t}_{yw} = \sum_s w_k y_k = \sum_s d_k F_k(\mathbf{x_k'}\boldsymbol{\lambda}) y_k.$$

Several remarks on the derivation of the calibration estimator must be made :

1. The vector $\boldsymbol{\lambda}$ is determined solving the calibration system :

$$\phi_s(\boldsymbol{\lambda}) = t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi} \text{ where}$$

$$\phi_s(\boldsymbol{\lambda}) = \sum_s d_k\{F_k(\mathbf{x_k}'\boldsymbol{\lambda}) - 1\}\mathbf{x_k}.$$

Deville & Särndal (1992) propose Newton's algorithm for obtaining $\boldsymbol{\lambda}$, assuring that this method converge quickly.

2. Different choices of the distance function lead to different estimators. The most important case is $F_k(u) = 1 + q_k u$ when we obtain the generalized regression estimator :

$$\hat{t}_{yreg} = \sum_s w_k y_k = \hat{t}_{y\pi} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi})'\hat{\boldsymbol{\beta}}_s$$

where $\hat{t}_{\mathbf{x}\pi} = \sum_s d_k \mathbf{x_k}$ (respectively $\hat{t}_{y\pi}$) is the π-estimator for the total of $\mathbf{x_k}$ (respectively of $\mathcal{Y}$) and

$$\hat{\boldsymbol{\beta}}_s = \left(\sum_s d_k q_k \mathbf{x_k} \mathbf{x_k}' \right)^{-1} \sum_s d_k q_k \mathbf{x_k} y_k.$$

Focusing on the weights, the calibration technique leads to the same generalized regression estimator obtained by Särndal (1980) when a regression of $\mathcal{Y}$ through the $\mathcal{X}_1, \ldots, \mathcal{X}_q$ was considered. So, we have two different approach, the calibration technique and the regression, which leads to the same estimator.

3. With the assumption that a solution $\boldsymbol{\lambda}$ exists, the different choices for F_k can lead to negative weights, which are not desired for an estimation of the variance. Deville & Särndal (1992) modify properly the function F_k such that the resulting weights are positive.

Suppose now that $n, N \to \infty$ and :
- (C1) $\lim N^{-1} t_{\mathbf{x}} < \infty$;
- (C2) $N^{-1}(\hat{t}_{\mathbf{x}\pi} - t_{\mathbf{x}}) \to 0$ in design probability ;
- (C3) $n^{\frac{1}{2}} N^{-1}(\hat{t}_{\mathbf{x}\pi} - t_{\mathbf{x}})$ converges in distribution to $N(0, A)$;
- (C4) $\max ||\mathbf{x_k}|| = M < \infty$;
- (C5) $\max F_k''(0) = M' < \infty$.

We have the following result :

Result 1.4.1 *(Deville and Särndal 1992) : Under the supposed conditions, $\hat{t}_{yw}$ has the following properties :*

i. $\hat{t}_{yw}$ is design-consistent and at least asymptotically design-unbiased (ADU)

$$N^{-1}(\hat{t}_{yw} - \hat{t}_{y\pi}) = O_p(n^{-\frac{1}{2}})$$

if there exists a solution $\boldsymbol{\lambda}$ of the calibration equations.

ii. $\hat{t}_{yw}$ is asymptotically equivalent to the regression estimator $\hat{t}_{yreg}$ for any F_k that satisfies the condition (C5) from above.

$$N^{-1}(\hat{t}_{yw} - \hat{t}_{yreg}) = O_p(n^{-1})$$

and consequently $V(\hat{t}_{yw}) \simeq V(\hat{t}_{yreg})$.

From the second point, it results that the choice of F_k is not of great importance for the derivation of the variance of $\hat{t}_{yw}$ because all the estimators are asymptotically equivalent with the regression estimator. This result has important consequences on the derivation of the variance and variance estimation of $\hat{t}_{yw}$. We have :

$$V(\hat{t}_{yw}) \simeq V(\hat{t}_{reg}) = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} (d_k E_k)(d_l E_l)$$

where $E_k = y_k - \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}} = (\sum_\mathcal{U} q_k \mathbf{x}_\mathbf{k}\mathbf{x}_\mathbf{k}')^{-1} \sum_\mathcal{U} q_k \mathbf{x}_\mathbf{k} y_k$. To estimate the variance, we need an estimator for $\hat{\boldsymbol{\beta}}$, which is given by :

$$\hat{\boldsymbol{\beta}}_{ws} = \left(\sum_s w_k q_k \mathbf{x}_\mathbf{k}\mathbf{x}_\mathbf{k}'\right)^{-1} \left(\sum_s w_k q_k \mathbf{x}_\mathbf{k} y_k\right)$$

and then

$$\hat{V}(\hat{t}_{ws}) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} (w_k e_k)(w_l e_l)$$

where $e_k = y_k - \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}_{ws}$.

In conclusion, we can summarize the following :

i. For a sample s and for chosen F_k, we solve the calibration equation for obtaining $\boldsymbol{\lambda}$.

ii. When $\boldsymbol{\lambda}$ is determined, we derive the calibration estimator :

$$\hat{t}_{yw} = \sum_s w_k y_k = \sum_s d_k F_k(\mathbf{x}_\mathbf{k}'\boldsymbol{\lambda}) y_k.$$

iii. The variance estimate is equal to the variance estimate for the regression estimator, with the residuals e_k of $\mathcal{Y}$ on the calibrated variables :

$$\hat{V}(\hat{t}_{ws}) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} (w_k e_k)(w_l e_l)$$

1.4.1 The Generalized Calibration

This method (Deville 2000) proposes to consider for all $k \in U$, a function $F_k : R^q \to R$ regular in a domain containing 0 and satisfying $F_k(0) = 1$. The deduction of such functions will be made with the help of auxiliary information. Then, the calibration equations to be solved are

$$t_\mathbf{x} = \sum_s d_k \mathbf{x}_\mathbf{k} F_k(\lambda).$$

The linear case

We have in this case, for all $k \in \mathcal{U}$

$$\begin{aligned} F_k : R^q &\to R \\ F_k(0) &= 1 \\ F_k'(0) &= \mathbf{z}_\mathbf{k} \end{aligned}$$

where $\mathbf{z}_\mathbf{k}$ is the vector of the instrumental variables (Fuller 1996). It results the expression for F_k :

$$F_k(\lambda) = 1 + \mathbf{z}_\mathbf{k}'\lambda.$$

The vector $\mathbf{z_k}$ needs to be known only for the $k \in s$ and it is not an external auxiliary information.

From the above equations of calibration and for the particular expression of F_k we can derive the expression of λ :

$$\boldsymbol{\lambda} = \left(\sum_s d_k \mathbf{z_k} \mathbf{x_k'} \right)^{-1} (t_{\mathbf{x}} - \sum_s d_k \mathbf{x_k}).$$

The calibration estimator has the expression :

$$\begin{aligned} \hat{t}_{yw} &= \sum_s d_k y_k F_k(\boldsymbol{\lambda}) = \hat{t}_{y\pi} + (t_{\mathbf{x}} - \sum_s d_k \mathbf{x_k})' \hat{\boldsymbol{\beta}}_s \\ &= \hat{t}_{y\pi} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi})' \hat{\boldsymbol{\beta}}_s \end{aligned}$$

where $\hat{\boldsymbol{\beta}}_s$ is the solution of the normal equations :

$$\sum_s d_k \mathbf{z_k} (y_k - \mathbf{x_k}' \boldsymbol{\beta}) = 0$$

so $\boldsymbol{\beta}$ is the q-vector of the coefficient of the instrumental regression of $\mathcal{Y}$ through $\mathcal{X}_1, \ldots, \mathcal{X}_q$, with $\mathbf{z_k}$ instruments.
For F_k linear we obtain a result similar as in the case of standard calibration, namely the obtained calibration estimator is equal to a instrumental regression estimator.
As in the case of standard calibration, the variance of $\hat{t}_{yw}$ is asymptotically equivalent to the variance of the regression estimator :

$$V(\hat{t}_{yw}) \simeq V\left(\sum_s d_k E_k \right) \text{ where}$$

$$E_k = y_k - \mathbf{x_k}' \hat{\boldsymbol{\beta}}, k \in \mathcal{U}$$

are the residuals of the instrumental regression and $\hat{\beta}$ satisfies the system of normal equations :

$$\sum_{\mathcal{U}} \mathbf{z_k} (y_k - \mathbf{x_k}' \boldsymbol{\beta}) = 0.$$

With the same principle, the variance estimation of $\hat{t}_{yw}$ has the expression :

$$\hat{V}(\hat{t}_{yw}) = \hat{V}(\sum_s d_k e_k) \text{ where}$$

$$e_k = y_k - \mathbf{x_k}' \hat{\boldsymbol{\beta}}_s, k \in s$$

with $\hat{\boldsymbol{\beta}}_s$ the solution of the normal equations :

$$\sum_s d_k \mathbf{z_k} (y_k - \mathbf{x_k}' \boldsymbol{\beta}) = 0.$$

The general case

With the additional condition $maxF_k''(\lambda) < \infty$, we can suppose that the function F_k can be approximated with a linear function :

$$F_k(\lambda) \simeq 1 + \mathbf{z}_k\lambda, \text{ for all } k \in U$$

where $z_k = gradF_k(0)$.

It results from the calibration equations :

$$\lambda = \left(\sum_s d_k \mathbf{z_k} \mathbf{x_k'}\right)^{-1} (t_x - \hat{t}_{x\pi}) + O(n^{-2}).$$

which implies that the variance of $\hat{t}_{yw}$ in the non-linear case is approximately equal to the variance in the linear case and the variance estimation are the same in the two situations. We can approximate the variance and variance estimate with the same residual technique.

1.4.2 Some other ways to generalise the calibration method

Théberge (1999) extends the calibration technique to estimate population parameters other than totals and means and also extends the technique to the case where there is no solution to the calibration equation. A new method to compute a calibration estimator that uses an arbitrary distance measure is developed and the case of estimating a bilinear parameter is treated.

When the auxiliary information is available for all the units in the population, Wu & Sitter (2001) combine *the model-assisted approach*, discussed in the next section, with *the calibration technique* for estimating the population mean. They call this new approach *model calibration*. More precisely, they consider that the variable of study and the auxiliary information are linked up by a nonlinear model of superpopulation ξ, as follows

$$E_\xi(y_k|\mathbf{x_k}) = \mu(\mathbf{x_k}, \boldsymbol{\theta}), \quad V_\xi(y_k|\mathbf{x_k}) = v_k^2\sigma^2$$

for all $k \in \mathcal{U}$; $\boldsymbol{\theta}$ is a p-dimensional vector, $\boldsymbol{\theta}$ and σ^2 are unknown superpopulation parameters, μ and v_k are known functions. Next, the parameter $\boldsymbol{\theta}$ will be estimated by a design-based method ; $\hat{\boldsymbol{\theta}}$ is obtained.
Then the model calibration estimator of $\overline{y}_\mathcal{U} = \frac{1}{N}\sum_\mathcal{U} y_k$ is denoted by $\hat{t}_{mc} = N^{-1}\sum_s w_k y_k$ where the weights w_k are as close as possible in the sense of a distance function to d_k and verify the calibration equations :

$$N^{-1}\sum_s w_k = 1 \quad \sum_s w_k \mu(\mathbf{x_k}, \hat{\boldsymbol{\theta}}) = \sum_\mathcal{U} \mu(\mathbf{x_k}, \hat{\boldsymbol{\theta}}).$$

We have the following result :

Result 1.4.2 *(Wu & Sitter 2001) :*

i. For a chi-square distance, the calibration estimator has the expression

$$\hat{t}_{mc} = \hat{t}_{y,\pi} + \left\{ N^{-1} \sum_{\mathcal{U}} \hat{\mu}_k - N^{-1} \sum_{s} d_k \hat{\mu}_k \right\} \hat{B}$$

where $\hat{\mu}_k = \mu(\mathbf{x_k}, \hat{\boldsymbol{\theta}})$, $\hat{B} = \dfrac{\sum_s d_k q_k (\hat{\mu}_k - \hat{\overline{\mu}})(y_k - \overline{y})}{\sum_s d_k q_k (\hat{\mu}_k - \hat{\overline{\mu}})^2}$, $\overline{y} = \dfrac{\sum_s d_k q_k y_k}{\sum_s d_k q_k}$ *and* $\hat{\overline{\mu}} = \dfrac{\sum_s d_k q_k \hat{\mu}_k}{\sum_s d_k q_k}$.

ii. Under general conditions, $\hat{t}_{mc} = \hat{t}_{y,\pi} + O_p(n^{-1/2})$ *and thus it is asymptotically design-unbiased for* $\overline{t}_y$ *irrespective of whether the model is correct or not.* $\hat{t}_{mc}$ *is also approximately model-unbiased.*

iii. The variance is approximated by

$$V(\hat{t}_{mc}) \simeq N^{-2} \sum_{\mathcal{U}} \sum_{\mathcal{U}} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \Delta_{kl}$$

where $E_k = y_k - \mu(\mathbf{x_k}, \theta)B$ *are the population fit residuals with* $B = \dfrac{\sum_{\mathcal{U}} q_k (\mu_k - \overline{\mu})(y_k - \overline{t}_y)}{\sum_{\mathcal{U}} q_k (\mu_k - \overline{\mu})^2}$, $\overline{\mu} = N^{-1} \sum_{\mathcal{U}} \mu(x_k, \theta)$.

iv. An estimator for the approximative variance is given by

$$\hat{V}(\hat{t}_{mc}) \simeq N^{-2} \sum_{s} \sum_{s} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}}$$

where $e_k = y_k - \hat{\mu}_k \hat{B}$.

The generalized regression estimator, $\hat{t}_{GREG} = N^{-1} \left(\hat{t}_{y,\pi} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x},\pi})' \hat{\boldsymbol{\beta}} \right)$, $\boldsymbol{\beta}$ the population parameter, (Cassel *et al* 1977, Särndal 1980) is obtained for $\mu(\mathbf{x_k}, \boldsymbol{\theta}) = \mathbf{x_k}' \boldsymbol{\theta}$. Wu & Sitter (2001) propose at the same time a generalisation of GREG by $\hat{t}_{GREG,mc} = N^{-1} \left(\hat{t}_{y,\pi} + \sum_{\mathcal{U}} \hat{\mu}_k - \sum_{s} d_k \hat{\mu}_k \right)$ for which they affirm that is not asymptotically equivalent with $\hat{t}_{mc}$ because $\hat{B}$ does not converge to 1. At the same time, they argue for $\hat{t}_{mc}$ against $\hat{t}_{GREG,mc}$ because the gain of $\hat{t}_{GREG,mc}$ over $\hat{t}_{y,\pi}$ in term of reduction of variance depends on the goodness of the approximation $y_k - \hat{\mu}_k$ while $\hat{t}_{mc}$ uses $\hat{\mu}_k$ as a tool of calibration while keeping as close to $\hat{t}_{y,\pi}$ as possible.

The variance evaluation by POULPE

We consider the case of a simple calibration of the final sample on the known population totals. In this case, CALMAR transforms the initial weights d_k into the final weights w_k, such that they are as close as possible to d_k and satisfy the calibration condition :

$$\sum_{k \in s} w_k \mathbf{x_k} = t_{\mathbf{X}}.$$

According to the general result (Deville & Särndal 1992), the calibration estimator is asymptotically equivalent to the regression estimator with the variance :

$$V(\hat{t}_{wy}) \simeq V\left(\sum_{k \in s} g_k \frac{E_k}{\pi_k}\right)$$

where E_k is the population fit residual of $\mathcal{Y}$ on the calibration variables and $g_k = \frac{w_k}{d_k} = F_k(\mathbf{x}'_{\mathbf{k}}\lambda)$.

For estimating the variance, it is enough to have a formula for the variance estimation of the total of Y, and to replace the quantities y_k with $g_k e_k$, where e_k is the regression residual of Y on the calibration variables.

1.5 Model Approach

The previous sections treated *the parameter vector* $\mathbf{y} = (y_1, \ldots, y_N)$ as a non-random quantity, the only randomness being the sampling design $p(\cdot)$. In the present section, we will consider that $\mathbf{y} = (y_1, \ldots, y_N)$ is the outcome of a vector random variable $\mathbf{Y} = (Y_1, \ldots, Y_N)$ with distribution ξ. We call *superpopulation model* a specified set of conditions for the class of distributions of which ξ belongs to. The main aspect of the statistical analysis in the *superpopulation model* is thus that $\mathbf{y}$ is treated as the outcome of $\mathbf{Y}$ about which certain features are assumed known. The *superpopulation model* ξ can be regarded as a mathematical device used to make explicit the theoretical derivations.

Among the first having used the *superpopulation model*, we mention Cochran (1939, 1946), Deming and Stephan (1941), Madow and Madow (1944).

Although the use of a *superpopulation model* ξ is not accepted by all survey practitioners, there are situations when it is arguable that this approach will perform much better. We mention two such situations. The first one is the inclusion of the treatment of nonsampling errors in survey sampling (Särndal *et al.* 1992, ch. 14). Secondly, it is possible under a *superpopulation model* ξ to make comparison of variances of two p-unbiased strategies, fact which entails the resolution of some of the nonexistence problems in uniformly minimum variance p-unbiased estimation (Cassel *et al.* 1977).

In the case of a *superpopulation model* ξ, we have two kinds of randomness : one already existing, the sampling design $p(\cdot)$ and the new one introduced by the joint distribution ξ of $Y_1, \ldots, Y_N$. We need supplementary notations induced by ξ. Let $Q = Q(Y_1, \ldots, Y_N)$ be a function of $Y_1, \ldots, Y_N$ and we denote by $E_\xi(Q)$ the expectation of Q with respect to ξ defined as follows

$$E_\xi(Q) = \int Q d\xi.$$

In the same manner, we can define other statistical quantities as variance and covariance with respect to ξ. In the next, we will use the index p for all the quantities calculated with respect to the design p; for example, E_p is the p-expectation and ξ for the model.

In the new frame, we call *statistic* a function $T = T(\mathcal{D})$ where $\mathcal{D} = \{(k, Y_k) : k \in S\}$, S being a random variable with values in $\mathcal{S}$ the set of all possible sample s. So, the function T, for any given value s of S, depends on those Y_k for which $k \in s$. The statistic T for $S = s$ used for making inference about the population mean $\overline{Y} = N^{-1} \sum_{\mathcal{U}} Y_k$ is called *predictor* and T for $Y_k = y_k$ is called an *estimate* for $\overline{y} = N^{-1} \sum_{\mathcal{U}} y_k$.

Définition 1.5.1 *:*

i. T is called p-unbiased for $\overline{Y}$ if and only if, for a given design p, $E_p(T(\mathbf{y})) = \overline{y}$ for all $\mathbf{y} = (y_1, \ldots, y_N)$.

ii. T is called ξ-unbiased for $\overline{Y}$ if and only if, for a given design ξ, $E_\xi(T(\overline{Y}) - \overline{Y}) = 0$ for all $s \in \mathcal{S}$.

iii. T is called $p\xi$-unbiased for $\overline{Y}$ if and only if, for given p and ξ $E_\xi E_p(T(\overline{Y}) - \overline{Y}) = 0$.

For a *superpopulation model* approach, the choice of a strategy (p, T) will be dictated by the objective to minimise the ξ-expected p-MSE,

$$E_\xi \mathrm{MSE}_p(p, T) = E_\xi E_p(T - \overline{Y})^2.$$

Although the objective is the same, there are two different ways of obtaining the desired minimum called the *model-based approach* and the *design-based approach*. For the first one (Brewer 1963, Royall 1970, 1971), the sampling design is of minor importance. The objective is to choose T such that for every given sample s, T minimizes $E_\xi(T - \overline{Y})^2$. We will not develop here this approach but we mention its application in repeated surveys (Blight & Scott 1973, Scott & Smith 1974). For the second one (Cassel *et al.* 1976, Särndal 1980, Särndal *et al.* 1989), the support is on the sampling design p. We look for an estimate T of $\overline{y}$ such that T minimizes $E_p(T - \overline{y})^2$. We give a brief presentation of the main results in the case of the *design-based approach*.

Finally, we consider only the *noninformative* designs fact which allows us to interchange the order in calculating the expectations with respect to p and ξ.

Cassel, Särndal & Wretman (1976) introduce the p-unbiased *generalized difference estimator*

$$T_{GD} = \sum_s \frac{Y_k - e_k}{N\pi_k} + \sum_{\mathcal{U}} \frac{e_k}{N}$$

for an arbitrary vector $\mathbf{e} = (e_1, \ldots, e_N)$; T_{GD} is ξ-unbiased if $e_k = \mu_k$, for ξ defined as follows (Cassel, Särndal & Wretman 1976)

$$\xi : \begin{cases} E_\xi(Y_k) = \mu a_k + b_k = \mu_k; \\ E_\xi(Y_k - \mu_k)^2 = a_k^2\sigma^2 = \sigma_k^2; \\ E_\xi\{(Y_k - \mu_k)(Y_l - \mu_l)\} = a_k a_l \rho\sigma^2 = \sigma_{kl} \text{ for } k \neq l; \end{cases}$$

where $a_k > 0, b_k$ for $k = 1, \ldots, N$ are known numbers with $\sum_{k=1}^N a_k = N$ and μ, σ^2 and ρ are unknown and $\frac{-1}{N-1} \leq \rho \leq 1$.

We have the following optimality result :

Result 1.5.1 *: (Cassel, Särndal and Wretman 1976)*
Under the above model, the optimal strategy (p, T) with p a fixed size design of size n and T a p-unbiased linear estimator of $\overline{Y}$ is $(p_0 : T_{GD0})$, where

$$T_{GD0} = \sum_s \frac{Y_k - b_k}{na_k} + \sum_{\mathcal{U}} \frac{b_k}{N}$$

and $p_0 = p_0(s)$ is the sampling design with the inclusion probabilities $\pi_k = fa_k$, $f = \frac{n}{N}$.

The estimator T_{GD0} is also ξ and $p\xi$-unbiased. The HT estimator, $\hat{t}_{y\pi}$ belongs to the class T_{GD} if $\mathbf{e} = 0$ or if $e_k = \pi_k$ for p a fixed size design. From the above result, we have in general $E_\xi V_p(p, \hat{t}_{y\pi}) \geq E_\xi V_p(p_0, T_{GD0})$ with equality for $p = p_0$ and $b_k \propto a_k$.

Särndal (1980) considers a particular case of model ξ. Let us consider $\mathcal{X}_1, \ldots, \mathcal{X}_q$ auxiliary variables with $\mathbf{x}'_\mathbf{k} = (x_{k1}, \ldots, x_{kq})$ and that the variables $Y_1, \ldots, Y_N$ are independent. For the regression model :

$$E_\xi(Y_k) = \mathbf{x}'_\mathbf{k}\boldsymbol{\beta} = \mu_k$$

$$V_\xi(Y_k) = \sigma_k^2 = \sigma^2 v_k$$

where $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_q)$ and σ^2 are unknown, $v_k = v(\mathbf{x_k})$ is a known function for all k in U, T_{GD0} becomes :

$$\begin{aligned} T_{GR} &= \sum_s \frac{Y_k}{N\pi_k} + \sum_{j=1}^q \beta j \left(\frac{1}{N} \sum_U x_{kj} - \sum_s \frac{x_{kj}}{N\pi_k} \right) \\ &= N^{-1} \left[\hat{t}_{y\pi} + (t_\mathbf{x} - \hat{t}_{\mathbf{x}\pi})'\boldsymbol{\beta} \right] \end{aligned}$$

with $\hat{t}_{y\pi} = \sum_s \frac{Y_k}{\pi_k}$ and $\hat{t}_{\mathbf{x}\pi} = \sum_s \frac{\mathbf{x}_k}{\pi_k}$. For the auxiliary information, we need only to know $t_\mathbf{x} = \sum_\mathcal{U} \mathbf{x_k}$. If we suppose that all β_j are known, then the above optimality result give that T_{GR} is optimal for $\pi_k \propto v_k$ for all k.

The estimator T_{GR} can be viewed as a correction of the ξ-model biased but p-unbiased estimator $N^{-1}\hat{t}_{y\pi}$,

$$T_{GR} = N^{-1}\hat{t}_{y\pi} - B_\xi(N^{-1}\hat{t}_{y\pi})$$

where $B_\xi(N^{-1}\hat{t}_{y\pi}) = N^{-1}(t_\mathbf{x} - \hat{t}_{\mathbf{x}\pi})'\boldsymbol{\beta}$ is the ξ-bias of $N^{-1}\hat{t}_{y\pi}$ (Thompson 1997).

More realistic, β_j are unknown, for all $j = 1, \ldots, q$. Särndal (1980) proposes to estimate the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$ by :

$$\hat{\boldsymbol{\beta}}_\boldsymbol{s} = \boldsymbol{G}'_\boldsymbol{s}\boldsymbol{Y}_\boldsymbol{s} = \left(\boldsymbol{W}'_\boldsymbol{s}\boldsymbol{X}_\boldsymbol{s}\right)^{-\mathbf{1}} \boldsymbol{W}'_\boldsymbol{s}\boldsymbol{Y}_\boldsymbol{s}$$

where $\mathbf{W_s} = (w_{kj})_{k=\overline{i,n},j=\overline{i,q}}$ and the w_{kj} may or may not depend on the known quantities $\mathbf{x_k'}$ and v_k, $\mathbf{X}_s = (\mathbf{x_k'})_{k\in s}$ and $\mathbf{Y}_s = (Y_k)_{k\in s}$. It results that the vector $(w_{k1}, \ldots, w_{kq})$ is the vector of weights applied to unit k. The resulting estimator is the so-called *generalized regression estimator* :

$$\hat{T}_{GR} = N^{-1}\left[\hat{t}_{y\pi} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi})'\hat{\boldsymbol{\beta}}_s\right].$$

This estimator can be obtained without implying the variance structure of the model and it has the following properties :

Result 1.5.2 *(Särndal 1980)*
1. The estimator $\hat{\boldsymbol{\beta}}_s$ is ξ-unbiased for $\boldsymbol{\beta}$:

$$E_\xi(\hat{\boldsymbol{\beta}}_s) = \boldsymbol{\beta}.$$

2. $\hat{T}_{GR}$ with $\boldsymbol{\beta}_s$ as above is model unbiased under the ξ model.

Thus we can use the model ξ to obtain ξ-unbiased estimators for $\boldsymbol{\beta}_s$ but the basic properties are not dependent on whether the model ξ holds or not. It implies that we have a model-assisted and not a model-dependent estimator.

If T_{GR} was p-unbiased, $\hat{T}_{GR}$ loses this property because $\hat{\boldsymbol{\beta}}_s$ as a nonlinear function is not p-unbiased for $\boldsymbol{\beta}$. Särndal (1980) states that an exact design-unbiasedness can entail loss in efficiency and he proposes the asymptotic design-unbiasedness (ADU) and consistency of $\hat{T}_{GR}$ as minimum requirements. Robinson & Särndal (1983) gives conditions for which $\hat{T}_{GR}$ is ADU and consistent for $\overline{Y}$. At the same time, they give an approximation for the design mean squared of order $O(n^{-1})$. These conditions do not require the superpopulation model to be true, but optimality is reached if the model is true. Robinson & Särndal (1983) state that in the case of perfect correctness of the model, the mean square error is minimized for $p(s)$ with the inclusion probabilities $\pi_k \propto \sigma_k$.

There is a particular case when $\hat{T}_{GR}$ is exactly p-unbiased, namely for a model ξ for which $E_\xi(Y_k) = \beta x_k$ and a sampling design p with $\pi_k \propto x_k$. In this case, $\hat{T}_{GR} = N^{-1}\hat{t}_{y\pi}$.

We return to the expression of $\hat{\boldsymbol{\beta}}_s = \boldsymbol{G_s'Y_s} = \left(\boldsymbol{W_s'X_s}\right)^{-1}\boldsymbol{W_s'Y_s}$. Särndal (1980) discusses two particular choices for the matrix $\mathbf{W}_s$, π-inverse weighting and best linear unbiased weighting. The first one is obtained for that $\mathbf{W}_s$ such that for some vector $\mathbf{c} = (c_1, \ldots, c_q)'$ we have

$$\boldsymbol{1_s'\Pi_s^{-1} = c'W_s'}$$

for $\Pi_s = \mathrm{diag}(\pi_k)_{k\in s}$. This relation is fulfilled for

$$\boldsymbol{W_s = \Pi_s^{-1}X_s}$$

if ξ includes the intercept. In this case, $\hat{\boldsymbol{\beta}}_{s,1} = (\boldsymbol{X}_s'\boldsymbol{\Pi}_s^{-1}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s'\boldsymbol{\Pi}_s^{-1}\boldsymbol{Y}_s$. Another possibility is

$$\boldsymbol{W}_s = \boldsymbol{\Pi}_s^{-1}\boldsymbol{V}_s^{-1}\boldsymbol{X}_s$$

with $\mathbf{V}_s = \text{diag}(v_k)_{k\in s}$ realised if $v_k = \mathbf{c}'\mathbf{x_k}$. In this case

$$\hat{\boldsymbol{\beta}}_{s,2} = (\boldsymbol{X}_s'\boldsymbol{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s'\boldsymbol{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}\boldsymbol{Y}_s.$$

With this expression for $\hat{\boldsymbol{\beta}}_{s,2}$, $\hat{T}_{GR}$ becomes $\hat{T}_{GREG}$ discussed by Särndal, Swensson and Wretman (1989).
The best linear unbiased weighting is obtained for $\boldsymbol{W}_s = \boldsymbol{V}_s^{-1}\boldsymbol{X}_s$ which gives for $\boldsymbol{\beta}$ the estimator

$$\hat{\boldsymbol{\beta}}_{BLU} = (\boldsymbol{X}_s'\boldsymbol{V}_s^{-1}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s'\boldsymbol{V}_s^{-1}\boldsymbol{Y}_s$$

with the corresponding $\hat{T}_{BLU}$. At this point of discussion a question arises : which of the two estimator presented here, $\hat{T}_{GREG}$ and $\hat{T}_{BLU}$, is preferable ? $\hat{T}_{BLU}$ being obtained for the best linear unbiased estimator for $\boldsymbol{\beta}$ is preferred by many statisticians while survey practitioners, such as Särndal *et al.* (1992, pp 519), argue for $\hat{T}_{GREG}$. Their argument is that $\hat{T}_{GREG}$ is obtained for the sample-weighted estimator $\hat{\boldsymbol{\beta}}_{s,2}$ which is more robust than the unweighted $\hat{\boldsymbol{\beta}}_{BLU}$, namely $\hat{\boldsymbol{\beta}}_{s,2}$ remains design-consistent if the model is wrong. Besides, $\hat{\boldsymbol{\beta}}_{s,2}$ is the solution of the optimal sample estimating function (Godambe & Thompson 1986, Godambe 1995) :

$$\sum_s \frac{1}{\sigma_k^2\pi_k}\mathbf{x_k}(Y_k - \mathbf{x_k}'\boldsymbol{\beta}) \quad \text{for}$$

$$\sigma_k^2 = \sigma^2 v_k.$$

We present other properties of $\hat{T}_{GREG}$ in the situation when $v_k = \mathbf{c}'\mathbf{x_k}$, for all $k \in \mathcal{U}$. In the next, we will use for simplicity the notation $\hat{\boldsymbol{\beta}}_s$ for $\hat{\boldsymbol{\beta}}_{s,2}$,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_s &= (\boldsymbol{X}_s'\boldsymbol{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s'\boldsymbol{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}\boldsymbol{Y}_s \\ &= \left(\sum_s \frac{\mathbf{x_k}\mathbf{x_k}'}{\sigma_k^2\pi_k}\right)^{-1}\sum_s \frac{\mathbf{x_k}Y_k}{\sigma_k^2\pi_k}\end{aligned}$$

or equivalently, $\hat{\boldsymbol{\beta}}_s$ is the solution of the normal equation (or the optimal sample estimating function)

$$\sum_s \frac{1}{\sigma_k^2\pi_k}\mathbf{x_k}(Y_k - \mathbf{x_k}'\boldsymbol{\beta}) = 0.$$

The same formula for $\hat{\boldsymbol{\beta}}_s$ would have been obtained if we had used the Taylor linearization method for estimating the coefficient of regression, $\boldsymbol{\beta}$ (Särndal

et al. 1992, pp 193). $\hat{\boldsymbol{\beta}}_s$ is not p-unbiased for $\boldsymbol{\beta}$ but is approximately p-unbiased and ξ-unbiased, $E_\xi(\hat{\boldsymbol{\beta}}_s) = \boldsymbol{\beta}$. Under the conditions of consistency given by Fuller & Isaki (1982), Fuller (2002) $\hat{\boldsymbol{\beta}}_s$ is consistent for $\boldsymbol{\beta}$. More precisely, we have $\hat{\boldsymbol{\beta}}_s = \boldsymbol{\beta} + O_p(n^{-1/2})$.
Särndal, Swensson and Wretman (1989) give equivalent expressions for $\hat{T}_{GREG}$. Let's introduce the following notations :

- the predicted value for the k-th element is denoted by $\hat{Y}_k = \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}_s$,
- $e_{ks} = Y_k - \hat{Y}_k$ is the k-th regression residual,
- $E_k = Y_k - \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}$ is the population fit residual, with $\hat{\boldsymbol{\beta}}$ the solution of the normal equation,

$$\sum_{\mathcal{U}} \frac{1}{\sigma_k^2}\mathbf{x_k}(Y_k - \mathbf{x}_\mathbf{k}'\boldsymbol{\beta}) = 0$$

- $g_{ks} = 1 + \left(t_x - \hat{t}_{\mathbf{x}\pi}\right)' \hat{T}^{-1}\frac{\mathbf{x_k}}{\sigma_k^2}$ where $\hat{T} = \sum_s \frac{\mathbf{x_k}\mathbf{x_k}'}{\sigma_k^2\pi_k}$ and we suppose that $\hat{T}^{-1}$ exists. The quantities g_{ks} are known in the literature as the g-weights and they were introduced by Särndal *et al.* (1989).

In this particular case of model ξ with $v_k = c'\mathbf{x_k}$, for all $k \in \mathcal{U}$, Särndal *et al.* (1992) prove that the regression residuals e_{ks} have the property

$$\sum_s \frac{e_{ks}}{\pi_k} = 0$$

and Thompson (1997) prove that under the same model, the population fit residuals, E_k, satisfy

$$\sum_k E_k = 0.$$

Then, we have the following equivalent expressions for the regression estimator $\hat{T}_{GREG}$:

$$\begin{aligned}
1) \quad \hat{T}_{GREG} &= N^{-1}\sum_{\mathcal{U}} \hat{Y}_k = N^{-1}\sum_{\mathcal{U}} \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}_s \\
2) \quad \hat{T}_{GREG} &= N^{-1}\sum_s g_{ks}\check{Y}_k = \sum_s g_{ks}\frac{Y_k}{\pi_k}; \\
3) \quad \hat{T}_{GREG} &= N^{-1}\sum_{\mathcal{U}} \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}_s + \sum_s g_{ks}\check{E}_k.
\end{aligned}$$

Result 1.5.3 *(Särndal et al. 1992) The generalized regression estimator $\hat{T}_{GREG}$ has the following expression :*

i.

$$\hat{T}_{GREG} = N^{-1}\left\{\hat{t}_{y\pi} + \left(t_\mathbf{x} - \hat{t}_{\mathbf{x}\pi}\right)'\hat{\boldsymbol{\beta}}_s\right\} = \mathbb{N}^{-1}\left\{sum_\mathcal{U}\mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}_\mathbf{s} + \sum_s g_{ks}\check{E}_k\right\}$$

which is approximately unbiased for $\overline{Y}$;

ii. the approximate variance is :

$$V(\hat{T}_{GREG}) \simeq N^{-2} \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \check{E}_k \check{E}_l;$$

iii. the variance estimator is given by :

$$\hat{V}_g(\hat{T}_{GREG}) = N^{-2} \sum_s \sum_s \check{\Delta}_{kl} \left(g_{ks}\check{e}_{ks}\right)\left(g_{ls}\check{e}_{ls}\right).$$

From the expression of the approximate variance of $\hat{T}_{GREG}$, we can give another variance estimator, replacing E_k by its sample-based counterpart e_{ks} :

$$\hat{V}_1 \simeq \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_{ks}}{\pi_k} \frac{e_{ls}}{\pi_l}.$$

When s is of fixed size, the Yates-Grundy variance estimator $\hat{V}_1$ is :

$$\hat{V}_{YG} = -\frac{1}{2} \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} \left[\frac{e_{ks}}{\pi_k} - \frac{e_{ks}}{\pi_k}\right]^2.$$

Kott (1990) gives conditions for the Yates-Grundy estimator to be consistent for MSE $(\hat{T}_{GREG}) \simeq V(\hat{T}_{GREG})$:

i. suppose $\hat{\boldsymbol{\beta}}_s \to \boldsymbol{\beta}$ in probability ;

ii. $E_p\left[(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta})'\right] = O(n^{-2})$;

iii. $N^{-2} E_p\left[(\hat{t}_{\mathbf{x}\pi} - t_{\mathbf{x}})(\hat{t}_{\mathbf{x}\pi} - t_{\mathbf{x}})'\right] = O(n^{-2})$.

Särndal, Swensson & Wretman (1989) proposed three variance estimators for the general regression estimator : $\hat{V}_g, \hat{V}_1$ and $\hat{V}_{YG}$. The question is which one is the best. In the design-approach, $\hat{V}_g, \hat{V}_1$ are approximately equivalent under the supplementary conditions. Since $\hat{V}_1$ is design consistent, $\hat{V}_g$ is design consistent too. For poststratified sampling and ratio estimation $\hat{V}_g$ is preferred. In the model-approach and for the particular case when $\sigma_k^2 = \sigma^2 \lambda' \mathbf{x_k}$, Särndal, Swensson & Wretman (1989) examine the properties of $\hat{V}_g$ under this particular model ξ. For this, they consider :

$$\hat{V}^* = \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_{ks}\epsilon_k}{\pi_k} \frac{g_{ls}\epsilon_l}{\pi_l}$$

where $\epsilon_k = Y_k - \mathbf{x}_\mathbf{k}'\boldsymbol{\beta}, k \in U$ are independents models residuals under the considered model. We can not calculate ϵ_k because of β but we can calculate $\hat{V}_g$ and under the given conditions, we have $\hat{V}^* - \hat{V}_g \to 0$ in probability. Särndal *et al.* (1989) examine if this variance estimator V^* is unbiased for the model mean square error MSE $_\xi(\hat{T}_{BLU}) = E_\xi(\hat{T}_{BLU} - \overline{Y})^2$. We have :

$$\mathrm{MSE}_\xi(\hat{T}_{BLU}) = \sum_s \left(\frac{g_{ks}\sigma_k}{\pi_k}\right)^2 - \sum_{\mathcal{U}} \sigma_k^2$$

$$E_\xi(\hat{V}^*) = \sum_s \left(\frac{g_{ks}\sigma_k}{\pi_k}\right)^2 - \sum_{\mathcal{U}} g_{ks}\sigma_k^2.$$

They introduce the following criterion :

$$\text{RMB}(\hat{V}^*) = \frac{E_\xi(\hat{V}^*) - \text{MSE}_\xi(\hat{T}_{GREG})}{\text{MSE}_\xi(\hat{T}_{GREG})}$$

as a measure of the unbiasedness of $\hat{V}^*$. In general RMB $(\hat{V}^*)$ is small; it is exactly zero in two cases :

(a) for stratification simple random sampling with fixed $\frac{n_h}{N_h}$in stratum h and then $g_{ks} = 1$ for all k and so RMB $(\hat{V}^*) = 0$;

(b) in poststratification where $g_{ks} = \frac{N_h}{\hat{N}_h}$ for all k in the same group.

If we have simple random sampling with poststratification, the ratio will be approximately zero. When RMB $(\hat{V}^*)$ is not zero, it is advisable to remove the model bias. However, in practice this is hard to achieve because :

i. its numerical impact would usually be small;

ii. the resulting expressions for the variance could be complicated;

iii. it appeals to a model, which is an imperfect assumption.

For $\sigma_k^2 = \sigma^2\lambda'\mathbf{x_k}$, Särndal, Sweenson and Wretman (1989) showed that $\hat{V}^*$ is :

i. a design consistent variance estimator;

ii. a nearly model unbiased for MSE $_\xi(\hat{T}_{GREG})$.

The approach of Kott (1990) for estimating the variance is to find a design consistent estimator of the design mean squared error of $\hat{T}_{GREG}$, and if it exists, to multiply it by a factor that removes the model bias of $V_{YG}(\hat{T}_{GREG})$ as an estimator of the conditional variance of $\hat{T}_{GREG}$. As a result, the new variance/mean squared error estimator is simultaneously a design consistent estimator of the design MSE$(\hat{T}_{GREG})$ and a model unbiased estimator of the conditional variance of T_{BLU}. We have a design unbiased estimator for the variance of $\hat{T}_{GREG}$ given by the Yates-Grundy formula :

$$\hat{V}_{YG} = \frac{1}{N^2}\sum_{k<l}\frac{\pi_k\pi_l - \pi_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2$$

Kott (1990) proposes the following estimator :

$$r_V(\hat{V}) = \frac{V_\xi(\hat{T}_{GREG} - \overline{Y})}{E_\xi(V_{YG})}V_{YG}$$

which is difficult to calculate. We have $r_V(\hat{V})$ model unbiased estimator of the conditional variance of $\hat{T}_{GREG}$ and design consistent of MSE$(\hat{T}_{GREG})$ because

$$J_V = \frac{V_\xi(\hat{T}_{GREG} - \overline{Y})}{E_\xi(V_{YG})} \to 1$$

in probability even if the model fails.

Montanari (1987) derives the expression for the coefficient of regression β who minimizes the p-variance of

$$T_{GR} = N^{-1}\left[\hat{t}_{y\pi} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi})'\boldsymbol{\beta}\right]$$

as $\boldsymbol{\beta}_{opt} = \left[\text{Var}\left(\sum_U \frac{\mathbf{x_k}}{\pi_k}\right)\right]^{-1} \text{Cov}\left(\sum_U \frac{\mathbf{x_k}}{\pi_k}, \sum_U \frac{Y_k}{\pi_k}\right)$. Unfortunately, the expression of $\boldsymbol{\beta}_{opt}$ depends on unknown quantities making so impossible its derivation.

If on the contrary, we minimize *the anticipated variance* (Isaki & Fuller 1982),

$$\min[E_\xi E_p(\hat{T}_{GREG} - \overline{Y})^2 - (E_\xi E_p(\hat{T}_{GREG} - \overline{Y}))^2]$$

we will find the optimal sampling design for the regression estimator, namely we find $p(\cdot)$ such that $E(n_s) = n$ and $\pi_k = \frac{n\sigma_k}{\sum_\mathcal{U} \sigma_k}$. For such design,

$$E_\xi E_p(\hat{T}_{GREG} - \overline{Y})^2 \simeq N^{-2}\left\{\frac{1}{n}\left(\sum_\mathcal{U} \sigma_k\right)^2 - \sum_\mathcal{U} \sigma_k^2\right\}$$

which means that $\hat{T}_{GREG}$ reaches the minimum of the Godambi and Joshi (1965) limit.

The emphasis in this section is on improving estimates in the presence of auxiliary information by using regression models. In this case, the only requirement about the auxiliary information is that the population total must be known. When the value of an auxiliary variable for each unit in the population is known, more complex models may be used. A model of regression as studied above will improve our estimate if it reduces its variance. This is achieved if the population fit residuals $E_k = y_k - \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}$ are small, namely that it exists a strong linear relationship between the variable of interest and the auxiliary variable. On the contrary case, the variance could be large. This justifies the use of more general models, as the nonparametric ones. We study in more details this situation in the following.

1.6 Nonresponse

In the previous sections, we have supposed that all the selected individuals give an answer to all the items of the questionnaire. This does not correspond to a real survey, when we are confronted with individuals who do not give responses to some items or to the entire questionnaire. In these situations, we have a survey with *nonresponse* which could happen at a lower or a larger rate (in the case of more than 40% of nonresponse, it is advisable to make the survey again).

We start with a definition and classification of nonresponse. Then, we describe the methods that have been developed to handle such a problem.

1.6.1 Definitions and Notations

We consider the population $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$ from which we select a sample $s_a \subset \mathcal{U}, |s_a| = n_{s_a}$ according to a known sampling design $p_a(.)$ and we want to make inference upon q variables of study :

$$\mathcal{Y}_1, \ldots, \mathcal{Y}_j, \ldots, \mathcal{Y}_q.$$

We denote by y_{jk} the value of the variable $\mathcal{Y}_j$ for the k-th unit in the population. For all $k \in s_a$, let $\mathbf{y}_k$ be the vector of the values of all the variables $\mathcal{Y}_j$, for $j = 1, \ldots, q$ observed for the same individual k :

$$\mathbf{y}_k = (y_{1k}, \ldots, y_{jk}, \ldots, y_{qk}).$$

At the end of the sample survey, the matrix formed with all vectors $\mathbf{y}_k$ for $k \in s_a$ of $n_{s_a} \times q$ dimension, is incomplete. The missing values correspond to the nonresponse of an individual to a certain item. We distinguish two types of nonresponse :

i. *k unit nonresponse* when all the components of $\mathbf{y}_k$ are missing ;

ii. *k item nonresponse* when not all the components of $\mathbf{y}_k$ are missing.

Another useful notation is :

$$r_j = \{k, k \in s_a \text{ for which exists } y_{jk}\} \text{ for all } j = 1, \ldots, q.$$

The quantity r_j, for $j = 1, \ldots, q$, is called *the response set* for the variable $\mathcal{Y}_j$ and it is formed from those selected individuals who have given a valid response to the item j.

We must construct an estimate for the total or mean of all $\mathcal{Y}_j$, having the population $\mathcal{U}$, the sample $s_a \subset \mathcal{U}$ and all the response sets, r_j. As a consequence, we must obtain a good estimate for the total of $\mathcal{Y}_j$ based only on the set of respondents, r_j, and not on the whole sample. In the case of nonresponse, an estimator for the population total is considered only on the set of respondents and this fact introduces a bias in the estimates,

called *nonresponse bias* produces an increase in the variance of the estimates because the effective size of the sample is reduced. In order to decrease it, we want to find a method for regain the loss due to the nonresponding elements. The recovery of the missing values could have been done if we had known the true responding mechanism, namely how r_j is generated. Because the distribution of nonresponse is unknown, assumptions upon it are made. The accuracy of the obtained estimator will depend on the validity of these assumptions, fact that is hard to check because the real distribution is not known. The main means for reducing both the nonresponse bias and the variance is the use of the auxiliary information for which the minimum requirement is that it should be available at least on the sample. Lundström & Särndal (2002) give methods for choosing the most relevant auxiliary information.

1.6.2 Methods of Treatment of Nonresponse

Several techniques have been developed to deal with nonresponse. They consist in a first stage in improving the rate of nonresponse and in a second stage, in improving the estimator in order to obtain a lower bias. We can use methods such as callbacks and follow-ups or a better planning in order to decrease the existing rate of nonresponse. But even after these attempts, we can not reduce to zero the nonresponse. At this point of a sample survey, no efforts will be made to reduce the number of nonrespondents but methods for improving the existing estimates. As we have already mentioned, there are two types of nonresponse : *the unit nonresponse* and *the item nonresponse*. For the first one, we use the method of *reweighting*, namely we modify the weights of the respondents ; *the item nonresponse* will be treated by the *imputation* method by which the missing values will be replaced by proxy values.
In the next, only one variable of interest will be considered for which we want to estimate the population total. In this case, there will be no difference between *the unit nonresponse* and *the item nonresponse*, as a consequence we can act as if we were confronted with *the unit nonresponse*. But, when we have several variables of interest, important decisions must be made about how *the item nonresponse* is to be treated.

1.6.3 Error caused by sampling and nonresponse

We consider the case of a single variable of study, $\mathcal{Y}$ and we intend to estimate its population total : $t_y = \sum_{\mathcal{U}} y_k$. Let $\hat{t}_{nr}$ be an estimator for t_y in the presence of nonresponse constructed by one of the mentioned methods, *reweighting* or *imputation* and let $\hat{t}$ be an estimator for t_y based on full response ; $\hat{t}$ can be the H-T or GREG estimator. We can write

$$\hat{t}_{nr} - t_y = (\hat{t} - t_y) + (\hat{t}_{nr} - \hat{t})$$

where the first term on the right side represents the sampling error and the second the nonresponse error.
A measure of the accuracy of $\hat{t}_{nr}$ is its mean square error MSE($\hat{t}_{nr}$). To evaluate this quantity we need the unknown response mechanism, denoted by $q(r|s_a)$. As we do not know the $q(r|s_a)$, we are not able to tell whether the nonresponse bias $B_{p_aq}(\hat{t}_{nr}) = E_{p_a}(E_q(\hat{t}_{nr}|s_a) - \hat{t})$ is zero; E_{p_a} (respectively E_q) represents the expectation with respect to p_a (respectively the mechanism of response q). In practice we make assumptions about q, assumptions sustained by simulation studies. Lundström & Särndal (2002) discuss in detail these problems. Under the assumption that the nonresponse bias is negligible, the variance has the expression (Lundström & Särndal 2002)

$$V(\hat{t}_{nr}) = V_{sam} + V_{nr}$$

where $V_{sam} = V_{p_a}(\hat{t})$ is *the sampling variance* and $V_{nr} = E_{p_a}V_q(\hat{t}_{nr}|s_a)$ is *the nonresponse variance*. While V_{sam} is estimated by the classical formula, the next section gives an estimator for $V_{nr} = E_{p_a}V_q$ in the case of *reweighting* method. This estimator can be used when the imputed estimator coincides with the reweighting one. As for the other methods of imputation presented here, a general variance estimator does not exist.

1.6.4 Techniques of Treatment the Unit Nonresponse

As we have mentioned, *the unit nonresponse* is treated by the *reweighting* method. This means that new weights are created, weights which are greater than the ones that would have been applied in the case of full response. This is done in order to compensate the lost due to nonresponse. These new weights can be created by *two-phase approach to weighting* or by *calibration approach*. We present below these approaches.

Two-phase Approach to Weighting

We use two-phase sampling as follows :
- in the first phase, a sample $s_a \subset \mathcal{U}$ is selected according to a known sampling design, $p_a(.)$:

$$s_a \subset \mathcal{U}, |s_a| = n_{s_a} \text{ with the inclusion probabilities } \pi_{ak}, \pi_{akl}, k \neq l.$$

- we consider the respondents set r as a second phase of selection, $r \subset s_a$, selected according to an unknown sampling pattern $q(r|s_a)$.

At this stage of selection, we must build up the response model. It consists in a set of assumptions upon the unknown response distribution. We give below the most used response models.

Response Homogeneity Group Model

The method can be described as follows : the first-phase sample s_a is split into H_{s_a} subgroups, denoted s_{ah} of size n_{ah} for $h = 1, \ldots, H_{s_a}$

$$s_a = \bigcup_{h=1}^{H_s} s_{ah}; \text{ for } h = 1, \ldots, H_{s_a}$$

The subsamples s_{ah} for $h = 1, \ldots, H_{s_a}$ are called the response homogeneity groups because each s_{ah} is formed by all the elements having the same response probability and elements with different response probabilities are located in different groups, given s_a. More precisely,

i. $Pr(k \in r|s_a) = \pi_{k|s_a} = \theta_{hs_a}$ for all $k \in s_{ah}$;

ii. $Pr(k\&l \in r|s_a) = \pi_{kl|s_a} = Pr(k \in r|s_a)Pr(l \in r|s_a)$ for $k \neq l \in s_a$

Let r_h be the response subset from s_{ah} of size m_h. Then the total set of respondents r of size m_r has the form

$$r = \bigcup_{h=1}^{H_{s_a}} r_h, r_h \subset s_{ah} \text{ and } m_r = \sum_{h=1}^{H_{s_a}} m_h.$$

Through this model, data are missing at random within each subgroup s_{ah}, conditionally on s_a and not within the entire population, as it is the case of naive model. In the next, let **RHG** be the abbreviation of response homogeneity group model. For the above **RHG**, the second phase corresponds to a stratified **BE** sampling design. If in the second phase, we condition to s_a and to the vector $\mathbf{m} = (m_1, \ldots, m_{H_{s_a}})$ of the sizes of r_h for $h = 1, \ldots, H_{s_a}$ which are supposed known, then the second phase is a stratified simple sampling without replacement. Then the probabilities of inclusion of first and second degree are :

i. $\pi_{k|s_a,\mathbf{m}} = Pr(k \in r|s_a, \mathbf{m}) = f_h = \frac{m_h}{n_h}$ for all $k \in s_{ah}$ and

ii. $\pi_{kl|s_a,\mathbf{m}} = Pr(k\&l \in r|s_a, \mathbf{m}) = f_h \frac{m_h - 1}{n_h - 1}$ for all $k \neq l \in s_{ah}$ and $Pr(k\&l \in r|s_a, \mathbf{m}) = f_h f_{h'}$ for $k \in s_{ah}, l \in s_{ah'}$ and $h \neq h'$.

Under the **RHG** model, we can construct two classes of estimators, whether we consider or not the auxiliary information.

RHG Model without Auxiliary Information

Results from two-phase sampling can be used and in particular for a stratified **SI** in the second phase. We use the same notations for the probability of inclusions, as in the case of two-phase sampling design, namely π_k^* ; these quantities are at the same time the new weights attributed to the values y_k considered only on the set of respondents. We have the following result :

Result 1.6.1 *: Under a **RHG** model,*

i. the weighting unbiased estimator for the population total $t_y = \sum_{\mathcal{U}} y_k$ *is* $\hat{t}_{\pi^*} = \sum_r \frac{y_k}{\pi_k^*}$ *where* $\pi_k^* = \frac{1}{\pi_{ak}} \frac{1}{\pi_{k|s_a,\boldsymbol{m}}}$.

ii. The variance is given by :

$$V(\hat{t}_{\pi^*}) = \sum_U \sum_U \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + E_{p_a} E_{\boldsymbol{m}} \left(\sum_{h=1}^{H_s} n_h^2 \frac{1-f_h}{m_h} S^2_{\check{y}s_{ah}} | s \right)$$

where $S^2_{\check{y}s_{ah}}$ *is the variance in* s_{ah} *of* $\check{y}_k = \frac{y_k}{\pi_{ak}}$ *and* $E_{p_a}(\cdot)$ *(respectively* $E_{\boldsymbol{m}}(\cdot)$*) is the expectation with respect to the sampling design (respectively to the distribution of* $\boldsymbol{m}$*).*

iii. An unbiased estimator for the variance is :

$$\hat{V}(\hat{t}_{\pi^*}) = \sum_r \sum_r \frac{\Delta_{akl}}{\pi_{kl}^*} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \sum_{h=1}^{H_s} n_h^2 \frac{1-f_h}{m_h} S^2_{\check{y}_{r_h}}$$

where $S^2_{\check{y}_{r_h}}$ *is the variance of* $\check{y}_k = \frac{y_k}{\pi_{ak}}$ *in* r_h.

It can be noticed that the expressions for the variance is a sum of two terms. The first corresponds to the variance of the π estimator in the absence of nonresponse and the second one is due to the presence of nonresponse, viewed as a second phase.

RHG with Auxiliary Information

As we have already mentioned, the use of the auxiliary information improves the quality of inference. According to Särndal *et al* (1992), this fact brings two gains :

i. an improved robustness against selection bias when the assumed response model is erroneous.

ii. a reduction of variance.

The approach is a two-phase sampling with the generalized regression estimator as discussed in Särndal *et al* (1992) chapter 9. If in one-phase, auxiliary information is required for the entire population in order to obtain a predictor for y and to compute the regression estimator, in two-phase we will need two sets of auxiliary information, one set for all the elements of the population, called also *population level* and one set for all the elements of the first-phase sample, s_a called *sample level*. We develop these statements, giving the corresponding notations.

Let us consider $\mathbf{X}_k$ a vector of J auxiliary values, available for all $k \in s_a$. These vectors will be used for selection in the second phase. We consider

that $\mathbf{X}_k$ contains at the same time the values available for all $k \in U$ as well as the values available only for $k \in s_a$:

$$\mathbf{X_k} = (\mathbf{X'_{1k}}, \mathbf{X'_{2k}})'$$

where $\mathbf{X}'_{1k}$ is a vector of J_1 auxiliary values available for $k \in U$ and $\mathbf{X}'_{2k}$ of $J - J_1$ values available for $k \in s_a$ only. The vector of auxiliary information, $\mathbf{X}_k$, will be used to get the predicted values $\hat{y}_k$ from r to s_a and $\mathbf{X}_{1k}$ is used to obtain predictions $\hat{y}_{1k}$ from s_a to U.

Let $\mathbf{X}_{1k}$ be a J_1 vector available for all $k \in \mathcal{U}$. We use $\mathbf{X}_{1k}$ to obtain $\hat{y}_{1k}$ from s_a to $\mathcal{U}$.

The vector $\mathbf{X}_2$ is a better predictor than $\mathbf{X}_1$, which is more difficult to obtain. This fact implies a good choice for $\mathbf{X}_2$ at the planning stage, followed by its observation for the elements from s_a.

The Regression Estimator in the Case of Nonresponse

The general theory developed in Särndal, Swensson & Wretman (1992) for two-phase sampling coupled with the regression estimator is used. The objective is to improve the π^* estimator on basis of the auxiliary information at the two levels. We have the following expression for the regression estimator in the presence of nonresponse in the second phase and for general fixed designs in both phases of sampling :

$$\begin{aligned}
\hat{t}_{reg} &= \sum_{\mathcal{U}} \hat{y}_{1k} + \sum_{s_a} \frac{\hat{y}_k - \hat{y}_{1k}}{\pi_{ak}} + \sum_{r} \frac{y_k - \hat{y}_k}{\pi^*_k} \\
&= \sum_{\mathcal{U}} \hat{y}_{1k} + \sum_{h=1}^{H_{s_a}} \left(\sum_{s_{ah}} \frac{\hat{y}_k - \hat{y}_{1k}}{\pi_{ak}} + f_h^{-1} \sum_{r_h} \frac{y_k - \hat{y}_k}{\pi_{ak}} \right)
\end{aligned}$$

where the regression predictors have the expressions $\hat{y}_k = \mathbf{X'_k}\hat{\mathbf{B}}_\mathbf{r}$ for $k \in s_{ah}$ and $\hat{y}_{1k} = \mathbf{X'_{1k}}\hat{\mathbf{B}}_\mathbf{1r}$ for $k \in \mathcal{U}$. For the derivation of $\hat{t}_{reg}$ we need to know $\sum_{s_{ah}} \mathbf{X}_k$ for all h and $\mathbf{X}_k$ for $k \in r$ as well as $\sum_U \mathbf{X}_{1k}$, $\sum_{s_{ah}} \mathbf{X}_{1k}$ for all h and $\mathbf{X}_{1k}$ for all $k \in r$.

The quantity $\hat{\mathbf{B}}_r$ is the π^*-estimator of the coefficient of regression of the linear model explaining the variable of interest $\mathcal{Y}$ based on the predictor vector $\mathbf{X}_k$ available only for $k \in s_a$. The variance of y_k under such a model is supposed to be σ_k^2. It results that $\hat{\mathbf{B}}_\mathbf{r}$ has the following expression (Särndal *et al* 1992) :

$$\hat{\mathbf{B}}_\mathbf{r} = \left(\sum_r \frac{\mathbf{X_k X'_k}}{\sigma_k^2 \pi^*_k} \right)^{-1} \sum_r \frac{\mathbf{X_k} y_k}{\sigma_k^2 \pi^*_k} = \left(\sum_{h=1}^{H_{s_a}} f_h^{-1} \sum_{r_h} \frac{\mathbf{X_k X'_k}}{\sigma_k^2 \pi_{ak}} \right)^{-1} \sum_{h=1}^{H_{s_a}} f_h^{-1} \sum_{r_h} \frac{\mathbf{X_k} y_k}{\sigma_k^2 \pi_{ak}}.$$

The predictions have the expressions $\hat{y}_k = \mathbf{X}'_{\mathbf{k}}\hat{\mathbf{B}}_{\mathbf{r}}$ for $k \in s_a$ with the residuals $e_k = y_k - \hat{y}_k$ for $k \in r$.

The same thing for $\hat{\mathbf{B}}_{1r}$; the model in this case explains $\mathcal{Y}$ based on the predictor vector $\mathbf{X}_{1k}$ available for all $k \in \mathcal{U}$. The variance of y_k is σ^2_{1k}.

$$\hat{\mathbf{B}}_{1r} = \left(\sum_r \frac{\mathbf{X_{1k}X'_{1k}}}{\sigma^2_{1k}\pi^*_k}\right)^{-1} \sum_r \frac{\mathbf{X_{1k}}y_k}{\sigma^2_{1k}\pi^*_k} = \left(\sum_{h=1}^{H_{sa}} f_h^{-1} \sum_{r_h} \frac{\mathbf{X_{1k}X'_{1k}}}{\sigma^2_{1k}\pi_{ak}}\right)^{-1} \sum_{h=1}^{H_{sa}} f_h^{-1} \sum_{r_h} \frac{\mathbf{X_{1k}}y_k}{\sigma^2_{1k}\pi_{ak}}.$$

We will have the predictions $\hat{y}_{1k} = \mathbf{X'_{1k}}\hat{\mathbf{B}}_{\mathbf{1}}$ for $k \in \mathcal{U}$ and the residuals $e_{1k} = y_k - \hat{y}_{1k}$ for $k \in r$.

We give the following notations before deriving a variance and a variance estimate for $\hat{t}_{reg}$:

i. For $k \in s_a$, let $E_k = y_k - \mathbf{X}'_{\mathbf{k}}\mathbf{B}_{\mathbf{s_a}}$ be the residuals with

$$\mathbf{B}_{\mathbf{s_a}} = \left(\sum_{s_a} \frac{\mathbf{X_k X'_k}}{\sigma^2_k \pi_{ak}}\right)^{-1} \sum_{s_a} \frac{\mathbf{X_k}y_k}{\sigma^2_k \pi_{ak}}.$$

ii. $k \in \mathcal{U}$, let $E_{1k} = y_k - \mathbf{X'_{1k}B_1}$ be the residuals with

$$\mathbf{B_1} = \left(\sum_U \frac{\mathbf{X_{1k}X'_{1k}}}{\sigma^2_{1k}}\right)^{-1} \sum_U \frac{\mathbf{X_{1k}}y_k}{\sigma^2_{1k}}.$$

Result 1.6.2 *I. The regression estimator $\hat{t}_{reg}$ is approximately unbiased for the total of $\mathcal{Y}$;*
II. The approximative variance of $\hat{t}_{reg}$ is :

$$V(\hat{t}_{reg}) \simeq \sum_U \sum_U \Delta_{akl} \frac{E_{1k}}{\pi_{ak}} \frac{E_{1l}}{\pi_{al}} + E_{p_a} E_{\boldsymbol{m}} \left(\sum_{h=1}^{H_{sa}} n_h^2 \frac{1-f_h}{m_h} S^2_{\check{E}_{s_{ah}}} | s\right)$$

where $S^2_{\check{E}_{s_{ah}}}$ is the variance of $\frac{E_k}{\pi_{ak}}$ in s_{ah} ;
III. A variance estimate has the expression :

$$\hat{V} = \sum_r \sum_r \frac{\Delta_{akl}}{\pi^*_{kl}} \frac{e_{1k}}{\pi_{ak}} \frac{e_{1l}}{\pi_{al}} + \sum_{h=1}^{H_s} n_h^2 \frac{1-f_h}{m_h} S^2_{\check{e}_{r_h}}$$

where $S^2_{\check{e}_{r_h}}$ is the variance of $\dfrac{e_k}{\pi_{ak}}$ in r_h.

An alternative expression for the variance estimate can be found if we consider the g-weights.

We can summarize the following conclusions : it is advisable to find a good model for the response distribution and moreover, the use of auxiliary information is necessary to improve the results.

Estimation of Response Probabilities

A third approach for treating the nonresponse is the estimation of the response probabilities, θ_k, by means of a model which uses the auxiliary information X_k known for the entire sample s_a. INSEE uses the LOGIT or PROBIT models for estimating θ_k for all $k \in s_a$.

Calibration Technique for Non-Response

As it was mentioned, a smaller nonresponse bias and smaller variance due to nonresponse is obtained by using auxiliary information. An example is the two-phase approach coupled with the regression estimator as proposed by Särndal, Swensson & Wretman (1992) for modelling the nonresponse. Another method for using the auxiliary information for improving the estimator is the calibration technique. We can have the standard calibration technique developed by Deville & Särndal (1992) and modified properly in the case of nonresponse by Lundström (1997) and Lundström & Särndal (2002). We can also use the generalised calibration technique (Deville 1999). We presented these methods in section 4.

The general calibration technique applied to nonresponse

We want to estimate the total of the population, $t_y = \sum_{\mathcal{U}} y_k$, in the presence of auxiliary information, $\mathcal{X}_1, \ldots, \mathcal{X}_q$, searching a set $\{w_k, k \in s_a\}$ that are as close as possible, in a sense of a distance function, to the design weights $d_k = \dfrac{1}{\pi_{ak}}$. In the presence of nonresponse, we have only the respondents set so that all the sums will be considered on r, the set of respondents : we know $\{d_k, k \in r\}$ and we intend to derive $\{w_k, k \in r\}$. The calibration estimator $\hat{t}_{yw}$ is done by the expression :

$$\hat{t}_{yw} = \sum_r w_k y_k$$

where $\{w_k \text{ for } k \in r\}$ is the set of the calibration weights, determined from the calibration equations :

– (C1) $\sum_r w_k \mathbf{x_k} = \sum_{s_a} d_k \mathbf{x_k}$ if $\mathbf{x_k}$ is known for all $k \in s_a$, but we do not know the total $\sum_{\mathcal{U}} \mathbf{x_k}$.

– (C2) $\sum_r w_k \mathbf{x_k} = \sum_{\mathcal{U}} \mathbf{x_k} = t_x$ if we know $t_x = \sum_{\mathcal{U}} \mathbf{x_k}$ and $\mathbf{x_k}$ for all $k \in s_a$.

where $\mathbf{x_k} = (x_{k1}, \ldots, x_{kq})'$ for all $k \in \mathcal{U}$. The distance function to be minimized is

$$\sum_r \frac{(w_k - d_k)^2}{d_k q_k}$$

where q_k are specified positive factors which make the approach more flexible. We have the following result :

Result 1.6.3 *(Lundström & Särndal 1999) : If we know t_x and $\mathbf{x_k}$ is known for all $k \in s_a$, then minimizing the above distance function under the constraint (C2) leads to the following calibration estimator*

$$\hat{t}_{yw} = \sum_r d_k \nu_{Uk} y_k \quad \textit{with}$$

$$\nu_{Uk} = 1 + q_k (\sum_{\mathcal{U}} \mathbf{x_k} - \sum_r d_k \mathbf{x_k})' (\sum_r d_k q_k \mathbf{x_k} \mathbf{x_k}')^{-1} \mathbf{x_k} \quad \textit{for} \quad k \in r.$$

If we are in the second situation, (C1), then $\hat{t}_{yw}$ is obtained by replacing $\sum_{\mathcal{U}} \mathbf{x_k}$ with $\sum_{s_a} d_k \mathbf{x_k}$ in the expression of ν_{Uk}.

They give at the same time the expressions for the variance and for a variance estimator.

The generalised calibration technique applied to nonresponse

The approach by generalized calibration is a little different but leads to the same results in particular situations. As it was mentioned in the first part, in the case of a survey sampling with nonresponse, we have the unknown response mechanism $q(r|s_a)$ which generates the set of respondents r, given the sample s_a. The calibration technique will be used for estimating the unknown parameters of the response-model.

The response mechanism is modelled as a sampling design $q(r, \boldsymbol{\beta}|s_a)$ giving the probability to obtain r given s_a. $\boldsymbol{\beta}$ is an unknown parameter from R^q. We have from the model the first-order inclusion probabilities $\pi_k = F_k^{-1}(\boldsymbol{\beta})$ and eventually the ones of second order. For estimating $\boldsymbol{\beta}$ we use the generalized calibration technique. More precisely, we have the estimating equations :

$$\sum_r F_k(\boldsymbol{\beta}) \mathbf{x}_k = t_{\mathbf{x}}$$

which can be written as equations of calibration in the following way :

$$\sum_r F_k(\boldsymbol{\beta}_0)\mathbf{x}_k G_k(\boldsymbol{\lambda}) = t_x$$

where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$ and $G_k(\boldsymbol{\lambda}) = \dfrac{F_k(\boldsymbol{\beta}_0 + \boldsymbol{\lambda})}{F_k(\boldsymbol{\beta}_0)}$, $G_k(0) = 1$. Moreover, $G_k(\boldsymbol{\lambda})$ are the calibrated functions and because of the fact that $F_k(\boldsymbol{\beta}_0) = d_k$ under the model the above equations are the calibrated equations. From the generalized calibration theory, we can take $G_k(\boldsymbol{\lambda}) = 1 + \mathbf{z}_k'\boldsymbol{\lambda}$ where $\mathbf{z}_k = gradG_k(0)$ are the instrumental variables which explain the nonresponse and known only on the respondents subset. We obtain the calibration equations :

$$\sum_r F_k(\boldsymbol{\beta}_0)x_k(1 + \mathbf{z}_k'\boldsymbol{\lambda}) = t_x$$

which give the expression for $\boldsymbol{\lambda}$ and the calibration estimator

$$\sum_r F_k(\boldsymbol{\beta}_0)y_k(1 + \mathbf{z}_k'\boldsymbol{\lambda}) = \hat{t}_{yw}.$$

The variance and the variance estimation can be calculated with the residual technique. The variance of the calibration estimator is approximately equivalent to the variance of the regression estimator :

$$V(\hat{t}_{yw}) \simeq V\left(\sum_r F_k(\boldsymbol{\beta}_0)E_k\right) = \sum_{\mathcal{U}}\sum_{\mathcal{U}} \Delta_{kl}\frac{E_k}{\pi_k}\frac{E_l}{\pi_l}$$

with the residuals $E_k = y_k - \mathbf{B}'\mathbf{x}_k$ and $\mathbf{B}$ is the solution of the normal equations $\sum_U \mathbf{z}_k(y_k - \mathbf{x}_k'\boldsymbol{\beta}) = 0$.
According to the same reason, the variance estimate has the expression :

$$\hat{V}(\hat{t}_{yw}) \simeq \hat{V}\left(\sum_r F_k(\boldsymbol{\beta}_0)e_k\right) = \sum_r\sum_r \check{\Delta}_{kl}\frac{e_k}{\pi_k}\frac{e_l}{\pi_l}$$

with the residuals $e_k = y_k - \hat{\mathbf{B}}'\mathbf{x}_k$ for $\hat{\mathbf{B}}$ the solution of the normal equations $\sum_r F_k(\boldsymbol{\beta}_0)z_k(y_k - \mathbf{x}_k'\boldsymbol{\beta}) = 0$.

The instrumental variables $\mathbf{z}_k$ are observed only on the set of respondents and they are introduced for reducing the non-response bias, while $\mathbf{x}_k$ reduce the variance. We obtain in this way a reduction of the non-response bias and of the variance at the same time.

If the variables $\mathbf{z}_k$ are of the form $q_k\mathbf{x}_k$, then we obtain the calibration estimator $t_{w\pi}$ derived by Lundstrom (1997) described above. In this case, the variance and the variance estimate can be obtained by replacing in the corresponding formulas from above $\mathbf{z_k}$ with $q_k\mathbf{x}_k$.

1.6.5 Imputation Technique for Treating the Item Non-Response

In this case, proxy values, called *imputed values*, are created for the missing y_k. An *imputed value* is in some sense an estimated value carried out by a given imputation method rather than an observed one. This approach seems to be a better method for the treatment the nonresponse in the case when we have q variables of study and *the weighting technique* becomes more complicated for the following reasons :

i. if we apply for each variable of study the **RHG** model, we will obtain different groups from a variable to another, which entails different weights.

ii. in the case of construction of the cross-tables or the multivariate distributions, the procedure is difficult to apply ;

iii. we can be led to erroneous estimates ;

As a consequence of these drawbacks, the imputation technique is advisable. It is more convenient to fill up the missing values from the matrix $(y_k)_{k\in s_a}$ and after that, to use the new constructed matrix for evaluating the totals of y_j, for $j = \overline{1,q}$ taking as true the imputed values. The accuracy of results will depend on the chosen method of imputation in the sense that the bias and the additional variance would be minimum. The main techniques of imputation are :

i. *Overall Mean Imputation* : the same value, the mean of the respondents values, is imputed for all missing responses.

ii. *Class Mean imputation* : all missing values from the same class are imputed with the mean of the respondents' values found in the corresponding class ; using auxiliary information, each class is built up from elements that can be considered similar ;

iii. *Hold-Deck and Cold-Deck* : in hold-deck imputation, the missing responses are imputed with respondents' values, values recorded during the current survey while the cold-deck procedure supposes to impute with values from earlier surveys ;

iv. *Regression Imputation* : in this case, the respondents are used to fit a regression of a variable, for which we need an imputation, on other variables assumed to have high predicted values. The predictors can be either study variables or auxiliary variables and the fitted regression equation is used to perform imputation.

v. *Multiple Imputations* : this technique was introduced by Rubin (1987) and it consists in replacing each missing response with several values, say m.

Variance estimation in the presence of imputation is a complex statistical problem.

1.6.6 The Evaluation of Variance in the Presence of Nonresponse by POULPE

The first version of `Poulpe` treats only the *unit nonresponse*. The mechanism consists in modelling the response distribution, doubled by a supplementary phase of sampling.

The response probabilities, $\pi_{k|s_a}$ are approximated by a function $G(\mathbf{x}'_{\mathbf{k}}c)$ where $\mathbf{x_k}$ is a vector of auxiliary information. In general a model LOGIT is used; $\pi_{k|s_a}$ are estimated by $G(\mathbf{x}'_{\mathbf{k}}\hat{c})$ where $\hat{c}$ is a convergent estimate for c. The variance is estimated by `Poulpe` applying the mechanism of two-phase sampling with a Poisson schema in the second phase , mechanism implemented in `Poulpe`.

In the case of **RHG** model and the second phase a stratified simple sampling without replacement, `Poulpe` will apply the mechanism of two-phases with in the second phase a stratified simple sampling without replacement where the response probabilities $\pi_{k|s_a}$ are replaced with $\hat{\pi}_{k|s_a}$.

A third possibility is the transformation by CALMAR of the $\frac{1}{\hat{\pi}_k} = \frac{1}{\pi_{ak}\hat{\pi}_{k|s_a}}$ in w_k, the calibration weights, which corresponds to an explicit treatment of nonresponse because CALMAR confers the new weights after the nonresponse was already treated. Improvement of existing nonresponse can be made directly by CALMAR by changing the weights $\frac{1}{\pi_{ak}}$ in w_k. This fact achieves at the same time a correction of nonresponse and a reduction of bias.
The new version, CALMAR 2, treats the generalised calibration method for nonresponse.

In the case of a sampling design made in two-phases, the existence of nonresponse entails the introduction of a supplementary phase of sampling, so it results a three-phase design; in this situation, the formulas from three-phase sampling with stratified simple sampling without replacement in the second phase are applied. These procedures are implemented in `Poulpe` with the difference that the response probabilities are estimated.

1.7 Repeated surveys

In this section, we intend to introduce the notion of time in survey sampling and to give a summary description of the main procedures that are proposed in the literature.

The evolution in time of some social indicators (the number of unemployed persons during one year) needs a special technique of sampling survey. The classical methods used until now are no longer satisfactory because they only give us the state of the population at a specific moment of time and

do not take into account the temporal correlations. Nevertheless, the objective is to make inference on a population which changes in time in order to obtain estimates of parameters of interest and of their evolution in time.

When the statistician has to build a survey sampling that takes place not only at a single moment of time, but during a sequence of successive occasions in time, several questions arise : how the sample must be chosen at every occasion ? Do we have to keep the entire sample at each date or not and if not, what proportion do we replace with the new elements ? At what interval of time do we realise the sample selections ? How could be taken into account the deaths, births during a sample survey as well as the existence of the non-response ? These are the difficulties arisen from the temporal component and which make more difficult the burden of the statistician who must choose the way of realizing the survey as a compromise between all the practical necessities and the existing theoretical frame.

The chapter "Sampling on two occasions" of the reference book Cochran (1977) gives the most important ideas concerning the repeated sampling and points out the contradiction between the estimation of the level and of the evolution, the use of regression technique, the derivation of the optimal overlapping rate. A more thorough discussion is found in Kish (1965) where diverse rotation schemes are studied in details in connection with a large range of estimation problems ; we add Tam (1984), also. Though, all these studies are done for the simple random sampling without replacement. The study of more general sampling designs is made by Särndal *et al.* (1992), Hidiroglu (2001). Duncan & Kalton (1987) make a list of the most important objective to be accomplished :

(a) the estimation of the characteristic of interest at different occasions ; for instance the proportion of unemployed persons from the total population ;

(b) the estimation of the mean of the characteristic at different occasions, such as the mean of the unemployed persons during one year ;

(c) the estimation of the evolution during two occasions, successive or not ;

(d) the measure of certain components of evolution : the transitions between the different activities, the individuals evolutions, the individuals instability ;

(e) the estimation of the mean of the individual indicators ;

(f) the measure of frequencies and duration of certain activities ;

(g) gathering informations upon certain rare groups of population.

As a consequence of the above objectives, Duncan & Kalton (1987) classify the sampling surveys over time as follows :

– the repeating surveys on different samples ;
– the survey by panel ;
– the sampling "partage" ;
– and the rotation sampling at one or several levels.

We present briefly each of them, specifying the particular elements and which of the above objectives are accomplished by each type of these surveys.

Repeating survey sampling on distinct samples

We are confronted with a sequence of sampling surveys realized at the same time, but each selected element is questioned only once. This method is applied in most of the cases on the households surveys realised by Insee. With this method, the objectives a), b), c), g) are fulfilled.

Panel

We deal with the opposite of the above schema ; in this case, a fixed number of individuals are selected and for a period and a frequency established in function of the objectives, we gather information from the same chosen persons ; d) is successfully accomplished, as well as e) ; a) and b) are less satisfied and c) is better accomplished than in the above model ; An example is the european panel realized in all the countries of the European Union since 1994 and until 2002.

Sample "partage"

This method is proposed by Kish and it consists in combining the two above methods, namely the sample is made of two subsamples, one of them is a panel and the other subsample is new at each moment of time. In this way, the advantages of both selections are combined and at the same time, the main drawbacks are overcome.

Rotation Sampling

In this case, an individual belongs to the sample for a limited period of time, period which is determined by the type of the chosen schema of rotation sampling. In fact, there are three types : one-level rotation, half-level and the most general, n-level rotation sampling. The principal feature of such a survey is that at all moments of time, the sample has common elements with the past and the future samples. The problem of matching two samples depends on the kind of rotation.

i. In one-level rotation sampling, at each moment of time t, the sample can be divided in one part common with the sample from the $(t-1)$-th occasion and new one. The common elements are interviewed again at the t-th occasion. The size of the sample as well as the proportion of common individuals, are fixed and independent of time.

ii. The method consists in interviewing a part of the sample for p times, then leaving out for q times and interviewing again for p times. This

schema was treated by Rao & Graham (1964). It is applied in the case of "Current Population Survey" made by the "Bureau of the Census" when a (4, 8) schema of rotation is adopted.

iii. The "n-level" rotation schema consists in considering a new sample at each moment of time t and for all the individuals in the sample, the information relative to the moment $t-1$ and t are recorded. The generalization of this survey comes from the fact that each time not only the information from the present occasion is recorded but also the one relative to the past.

The technique of rotation sampling accomplishes the objectives a), b) and c). Caron & Ravalet (2000) overline the advantages of this kind of selection. As they mentioned, this schema, doubled by the use of the appropriate estimators, can improve the estimate for the level and with a cheaper cost for the sampling survey. The points d) and e) from the list of objectives are better realized than in the case of a panel, even if these are not the principal objectives of the rotation schema.

This particular sampling design is applied in the surveys realized by Statistics Canada for employment ; in this case, each sample element is questioned during six successive periods. In the case of the French survey, a third of the sample is replaced every year.

1.7.1 Notations and assumptions

We consider the same population $\mathcal{U} = \{u_1, \ldots, u_N\}$ consisting of N elements and $\mathcal{Y}$ the variable of interest. We introduce a new notation that specify the fact that the value y_i of the variable $\mathcal{Y}$ for the i-th element of the population was recorded at the t-th occasion. Let $y_{i,t}$ be the value of the variable of interest for the i-th individual, at the moment of time t. An example of such a variable $\mathcal{Y}$ whose evolution in time is requested, could be the inclusion of the i-th individual in one of the groups of activities : employment, unemployment or inactivity. Let θ_t be the mean of the population at the moment t, $\theta_t = \dfrac{1}{N}\sum_{i\in U} y_{i,t}$; for the above example, θ_t represents the proportion of each category in the total population. We will note by $\Delta_s\theta_t$ the evolution of the mean θ_t between the moments $(t-s)$ and t, then $\Delta_s\theta_t = \theta_t - \theta_{t-s}$, and by S_t the empirical variance of the variable $\mathcal{Y}$ within the population :

$$S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_{i,t} - \theta_t)^2.$$

We make the assumptions :

i. The population $\mathcal{U}$ is stationary in time, namely that it is composed of the same N elements at all moments of time t the survey is accom-

plished. Under such assumptions, the births, deaths and even nonresponse are not considered. The case of a non-stationary population was considered by Holt and Skinner (1989).

ii. The quantities S_t , for all t, are assumed to be independent of time, namely :

$$S_t^2 = S^2 \text{ for all } t.$$

The same assumption is made for the covariance :

$$\text{Cov } (\mathcal{Y}_t, \mathcal{Y}_{t-k}) = \rho_k S^2$$

where ρ_k is the correlation, depending only on k, the interval of time between two occasions. For the moment, no assumptions are made on the structure of the correlation. Lee (1990) defines three kinds of correlations in the case of a panel.

iii. The size of population is considered to be large enough so that the correlation between $y_{i,t}$ and $y_{j,t}$, the values of the variable of interest for the i-th and j-th individuals registred at the same occasion of time, t can be zero. It results then :

$$\text{Cov } (y_{i,t}, y_{j,t}) = 0 \text{ and all } t, i, j.$$

Under these general conditions, we have the population $\mathcal{U}$, a sequence of occasions represented as $(0, \ldots, t-1, t)$ where by 0 is meant the first occurrence of the survey and by t the most recent one. Let consider a sample pattern, p, consisting of a set of samples s, where

$s = \{y_{i,t}$ the value of the i-th individual selected in the sample at the moment t$\}$.

Eckler (1955) compares this structure to an incomplete matrix with "m" columns and as many rows as there are distinct elements in U. The sample pattern will be specified later.

At this point of discussion, we need to specify the quantities which must be estimated, the method of estimation adopted and the chosen estimator. In most of the surveys on repeated occasions, it is intended to estimate the sequence $\{\theta_t\}$ of the mean of a certain parameter of the population, or the evolution $\{\Delta\theta_t\}$.

Two methods of estimation exist : *the classical method* and *the temporal one*. The first one considers the sequence $\{\theta_t\}$ unknown and fixed in time, namely non random and without any link between the successive values. In this frame, a structure of correlation unifies the $y_{i,t}$ and allows to build an optimal estimator based on the present and past information ; the first results are obtained by Patterson (1950), Eckler (1955), Rao & Graham (1964) and they are extended in the case of a multi-stage design by Singh (1968).

On the contrary, the second method considers the parameters θ_t to be random and as a result their evolution can be described by means of a stochastic model

$$y_t = \theta_t + e_t$$

where e_t is the error of estimation such that θ_t and e_t are not correlated. Based on the model, it is then possible to derive an estimator of the θ_{t+1} by representing θ_t as a time series, possibly non-stationary. More exactly, the estimation problem is equivalent to a signal extraction in the presence of noise. This frame was considered for the first time by Scott & Smith (1974) and continued by Scott, Smith & Jones (1977). More recent works are Tam (1987), Binder & Dick (1989), Bell & Hillmer (1990).

In the following, we will consider the first approach.

The choice of an estimator is imposed and encumbered by several factors. The method of estimation being chosen, the choice of the estimator will depend on the different types of surveys on repeated occasions and on the quantities which are to be estimated (the mean or the evolution). We must overline the following reality : the quantity to estimate decides in a way the type of sampling to choose ; as an example, if we want to estimate the mean over several periods, it is advisable to select different individuals at successive occasions, but if we want to estimate the evolution, then it is better to preserve the entire sample all the time. It remains for the statistician to choose a sampling design which does a compromise between all these. In comparison with the other designs, the rotation sampling accomplishes the most of the requests. We mention two practical methods for controlling overlaps between different samples : *permanent random numbers* and *collocated random number* (Nordberg 2000 ; Ernst, Valliant & Casady 2000). The techniques can be used for overlap control between samples for the same survey selected at different time periods or between different surveys at the same period of time. Besides, these methods of sampling have the attractive property of covering the deaths and births units.

1.7.2 One-level Rotation Sampling

Under the general framework presented above, we will try to develop the estimators and the corresponding formulas for the variance.

First, we introduce the specific notations : the exponent "nc" is the abbreviation for "non-common" and it is used every time we refer to the new sampled individuals at the t occasion and the symbol "c" for the sample elements present both at the $(t-1)$-th and t-th occasions. From the n individuals sampled at the t-th occasion, a proportion μn will be replaced at the next moment of time. We specify that μ is independent of time.

It is intended to estimate the following parameters of interest :

i. the mean θ_t during the sequence $(0, \ldots, t-1, t)$;

ii. the evolution of θ_t between two periods of time, successive or not.

For the next, it remains to specify the method used for sampling the future interviewed persons for each occasion and also, the estimator which is to be used. Caron & Ravalet (2000), Eckler (1955), Patterson (1950) and others consider the same sampling schema, that of simple sampling without replacement at all occasions in the case of single-stage. The case of two or three stage designs was considered by Singh (1968).
The one-level rotation sampling is performed as follows :

– At the $(t-1)$-th occasion, a sample s of size n is extracted from $\mathcal{U}$,

$$s \subset \mathcal{U}, |s| = n$$

by simple sampling without replacement and information is gathered from all the elements of s.

– A subsample s_c is considered from s of size n_c by the same design,

$$s_c \subset s, |s_c| = n_c.$$

– This sample, s_c will be considered at the next occasion, when the rest of the sample, namely s_{nc} of size $n_{nc} = n - n_c$ elements, will be obtained from the remaining elements $\mathcal{U} - s$ by simple sampling without replacement

$$s_{nc} \subset \mathcal{U} - s; |s_{nc}| = n_{nc}.$$

– At the t-th occasion, both the elements from s_c and the ones from s_{nc} are interviewed.The estimators are constructed upon these elements.

– We note with k the ratio of the not-common elements in the whole sample, namely :

$$k = \frac{n_{nc}}{n};$$

k is called *the changing rate.*

In the literature, several estimators were proposed.

The natural estimate for the mean

Caron & Ravalet (2000) proposed the natural estimator for the mean and for the evolution between $(t-1)$ and t. The approach is based on the elementary estimators proposed by Gurney & Daly, but under supplementary conditions which allow the estimators to be unbiased. The results are below :

Result 1.7.1 *: Suppose that the elementary estimators are unbiased and that* $\overline{y}_t^c$ *and* $\overline{y}_t^{nc}$ *are independent,* $\overline{y}_t^c = \frac{1}{n_c} \sum_{i \in s_c} y_{i,t}$ *and* $\overline{y}_t^{nc} = \frac{1}{n_{nc}} \sum_{i \in s_{nc}} y_{i,t}$.
Then for simple random sampling without replacement :

– *an unbiased estimator for θ_t is the empirical mean $\overline{y}_t$:*

$$\hat{\theta}_t = (1-k)\overline{y}_t^c + k\overline{y}_t^{nc}.$$

The variance is independent of the rotation design and is given by

$$V\left(\hat{\theta}_t\right) = \frac{S^2}{n}.$$

if we neglect the correction term for finite population.

– *An unbiased estimator for the evolution $\Delta\theta_t$ is the difference of the corresponding estimators for $t-1$ and t :*

$$\widehat{\Delta\theta}_{t(1)} = \overline{y}_t - \overline{y}_{t-1}$$

with the corresponding variance formula :

$$V\left(\widehat{\Delta\theta}_{t(1)}\right) = 2\left(1-\rho(1-k)\right)\frac{S^2}{n}.$$

The estimate of the evolution based on the common sample

For the estimation of $\Delta\theta_t$, Caron & Ravalet (2000) give an alternative estimate which only uses the information from the common elements between the $t-1$ and t :

Result 1.7.2 *: An alternative estimate for the evolution is :*

$$\widehat{\Delta\theta}_{t(2)} = \overline{y}_t^c - \overline{y}_{t-1}^c = \frac{1}{n_c}\sum_{s_c}\left(y_{i,t} - y_{i,t-1}\right)$$

with

$$V\left(\widehat{\Delta\theta}_{t(2)}\right) = 2\frac{1-\rho}{1-k}\frac{S^2}{n}.$$

Caron & Ravalet (2000) propose the variance estimate for $\widehat{\Delta\theta}_{t(2)}$:

$$\widehat{Var}\left(\widehat{\Delta\theta}_{t(2)}\right) = \frac{1}{n_c(n_c-1)}\sum_{s_c}(z_i - \overline{z})^2$$

where $z_i = y_{i,t} - y_{i,t-1}$ for all $i \in s_c$.

Comparison between $\widehat{\Delta\theta}_{t(1)}$ and $\widehat{\Delta\theta}_{t(2)}$

Let us consider the ratio of the variances of the two estimators :

$$\frac{\widehat{\Delta\theta}_{t(2)}}{\widehat{\Delta\theta}_{t(1)}} = \frac{1-\rho}{(1-k)\left(1-\rho(1-k)\right)}.$$

We have the following situations :

– for k fixed, then $\widehat{\Delta\theta}_{t(2)}$ is more efficient if $\rho \geq \frac{1}{2-k}$;

– for $\rho < \frac{1}{2}$, $\widehat{\Delta\theta}_{t(1)}$ is better than $\widehat{\Delta\theta}_{t(2)}$.

Best linear unbiased variance estimator

Gurney & Daly (1965) give the best linear unbiased variance estimator on the base of the elementary estimators. Gourieroux & Roy (1978) and Caron & Ravalet (2000) obtain the same result, but through a different way, namely by using the Gauss-Markov theorem.

Suppose we have the following linear model :

$$\overline{Y} = X\Theta + e$$

where $\Theta = (\theta_1, \ldots, \theta_t)'$ is the vector of the mean of the variable of interest $\mathcal{Y}$ considered at the sequence of time $(1, \ldots, t)$; $\overline{Y} = (\overline{y}_1, \ldots, \overline{y}_t)$ is the vector of the empirical mean of $\mathcal{Y}$ considered at the same occasions ; X is a vector of 0 and 1 ; e is the vector of the errors due to the sampling with the properties :

$$E(e) = 0; E(ee') = \Omega$$

where Ω is the variance-covariance matrix. In the case of one-level rotation sampling and simple sampling without replacement for both periods, the model can be written as follows :

$$\begin{pmatrix} \overline{y}_{t-1}^{nc} \\ \overline{y}_{t-1}^{c} \\ \overline{y}_{t}^{nc} \\ \overline{y}_{t}^{c} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + \begin{pmatrix} e_{t-1}^{nc} \\ e_{t-1}^{c} \\ e_{t}^{nc} \\ e_{t}^{nc} \end{pmatrix}.$$

The variance-covariance matrix of the sampling errors has the following form :

$$E(ee') = \Omega = S^2 \begin{pmatrix} \frac{1}{n_c} & 0 & 0 & 0 \\ 0 & \frac{1}{n_c} & 0 & \frac{\rho}{n_c} \\ 0 & 0 & \frac{1}{n_c} & 0 \\ 0 & \frac{\rho}{n_c} & 0 & \frac{1}{n_c} \end{pmatrix}.$$

In these conditions, the Gauss-Markov theorem ensures the existence of the best linear unbiased estimator for Θ and it can be written as :

$$\tilde{\Theta} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y$$

with the variance :

$$V(\tilde{\Theta}) = \left(X'\Omega^{-1}X\right)^{-1}.$$

We obtain in our case, the best linear unbiased estimator for θ_{t-1} and θ_t :

$$\hat{\theta}_{t-1}^{opt} = \overline{y}_{t-1}^{nc} + \frac{1-k}{1-k^2\rho^2}(\overline{y}_{t-1}^{c} - \overline{y}_{t-1}^{nc}) + \frac{\rho k(1-k)}{1-k^2\rho^2}(\overline{y}_{t}^{nc} - \overline{y}_{t}^{c})$$

$$\hat{\theta}_t^{opt} = \frac{\rho k(1-k)}{1-k^2\rho^2}(\overline{y}_{t-1}^c - \overline{y}_{t-1}^{nc}) + \frac{1-k}{1-k^2\rho^2}(\overline{y}_t^c - \overline{y}_t^{nc}) + \overline{y}_t^{nc}$$

with the following expression for the variance :

$$V(\hat{\theta}_{t-1}^{opt}) = V(\hat{\theta}_t^{opt}) = \frac{S^2}{n}\frac{1-k\rho^2}{1-k^2\rho^2}.$$

The Gauss-Markov theorem gives also an estimator for all linear combinations of Θ as a linear combination of the best linear unbiased estimator of Θ, so we can derive the expression of such an estimator for the evolution of θ_t and for its variance :

$$\Delta\hat{\theta}_t^{opt} = \hat{\theta}_t^{opt} - \hat{\theta}_{t-1}^{opt} = \overline{y}_t^{nc} - \overline{y}_{t-1}^{nc} + \frac{1-k}{1-k\rho}\left(\overline{y}_{t-1}^{nc} - \overline{y}_{t-1}^c + \overline{y}_t^c - \overline{y}_t^{nc}\right)$$

$$V(\overline{y}_t^{opt} - \overline{y}_{t-1}^{opt}) = 2\frac{S^2}{n}\frac{1-\rho}{1-k\rho}.$$

A few remarks can be made :

i. The value of k for which the variance is minimum is :

$$k_{opt} = \frac{1-\sqrt{1-\rho^2}}{\rho^2} \text{ for } \rho \neq 0.$$

When $\rho = 0$, we obtain $V(\hat{\theta}_{t-1}^{opt}) = V(\hat{\theta}_t^{opt}) = \frac{S^2}{n}$. This value is independent of the plan of rotation sampling.

ii. For $k = 0$ or $k = 1$, the natural estimator is obtained :

$$\hat{\theta}_{t-1}^{opt} = \overline{y}_{t-1}^c$$
$$\hat{\theta}_t^{opt} = \overline{y}_t^c$$

for $k = 0$ and

$$\hat{\theta}_{t-1}^{opt} = \overline{y}_{t-1}^{nc}$$
$$\hat{\theta}_t^{opt} = \overline{y}_t^{nc}$$

for $k = 1$.

iii. When the aim is to estimate the evolution, it is advisable to keep the whole sample if $\rho > 0$ and to change it entirely, if $\rho < 0$.

The composite estimators

As an alternative to the best linear unbiased estimator, the composite estimator was proposed, applied especially in the case of the multi-stage designs and complex survey sampling. The principle of deriving this estimator consists in improving the estimator which only uses the current information with the help of the correlations between the observations belonging to the common sample. In the case of the estimation of the evolution, we have two possibilities :

– we construct a composite estimator for each period in part and after that we take the difference of them or
– a composite estimator is built up directly for the two levels.

In the next, we consider separately each of the two proposed estimators and then we make a comparison.

The composite estimator of the level

For the composite estimator of the level, the natural estimator for the uncommon part is combined with the estimator proposed by Jensen (1942) for the common part. The estimator proposed by Jensen (1942) is obtained as a regression of y_t over y_{t-1}, using as auxiliary information the one gathered from the common elements :

$$\overline{y}_{t,c}^{reg} = \overline{y}_t^c + b(\overline{y}_{t-1} - \overline{y}_{t-1}^c)$$

where b is the linear regression coefficient of y_t over y_{t-1} with the expression :

$$b = \frac{\sum_{\mathcal{U}}(y_{i,t} - \overline{y}_t)(y_{i,t-1} - \overline{y}_{t-1})}{\sum_{\mathcal{U}}(y_{i,t-1} - \overline{y}_{t-1})^2}$$

which is estimated in the common sample s_c by :

$$\hat{b} = \frac{\sum_{s_c}(y_{i,t} - \overline{y}_t^c)(y_{i,t-1} - \overline{y}_{t-1}^c)}{\sum_{s_c}(y_{i,t-1} - \overline{y}_{t-1}^c)^2}.$$

It can be observed that the above expression corresponds to the one of the correlation coefficient, ρ. The variance of $\overline{y}_{t,c}^{reg}$ is given by :

$$V(\overline{y}_{t,c}^{reg}) = \frac{(1-\rho^2)S^2}{n_c} + \rho^2\frac{S^2}{n}$$

We can give now the expression of the composite estimator of θ_t for one level as a combination of the regression estimator on the common part and the empirical mean on the uncommon part ; we use the notation $\overline{y}_t^*$.

$$\overline{y}_t^* = (1-\phi)\overline{y}_t^{nc} + \phi\overline{y}_{t,c}^{reg}$$

where ϕ is a coefficient between 0 and 1. We develop the expression of the variance of $\overline{y}_t^*$; because of the independence of $\overline{y}_t^{nc}$ and $\overline{y}_{t,c}^{reg}$ it results :

$$\begin{aligned} V\left(\overline{y}_t^*\right) &= (1-\phi)^2 V\left(\overline{y}_t^{nc}\right) + \phi^2 V\left(\overline{y}_{t,c}^{reg}\right) \\ &= \frac{S^2}{n}\left(\frac{(1-\phi)^2}{k} + \frac{\phi^2(1-\rho^2)}{1-k} + \phi^2\rho^2\right) \end{aligned}$$

We consider k and ρ fixed and we make the discussion in function of the possible values of ϕ. In order to evaluate the efficiency of this new estimator,

we will compare it with the natural estimator ; we take the ratio of the two variances :

$$\frac{V(\overline{y}_t^*)}{V(\overline{y}_t)} = \frac{(1-\phi)^2}{k} + \frac{\phi^2(1-\rho^2)}{1-k} + \phi^2\rho^2$$

This ratio will be inferior to 1 for $\phi \in [\phi_1, \phi_2]$ where $\phi_1 = \dfrac{1-k}{1+k\rho}$ and $\phi_2 = \dfrac{1-k}{1-k\rho}$. Secondly, we search the value of ϕ which gives the optimal value of the variance. It can be observed that $\overline{y}_t^*$ is expressed as a linear combination of the independent estimators $\overline{y}_{t,c}^{reg}$ and $\overline{y}_t^{nc}$ which are unbiased for θ_t. Then the best combination is obtained for ϕ taking the value

$$\phi_{opt} = \frac{V(\overline{y}_t^{nc})}{V(\overline{y}_t^{nc}) + V(\overline{y}_{t,c}^{reg})}.$$

If we replace the corresponding values for the variances, we get the following expression for ϕ_{opt}

$$\phi_{opt} = \frac{1-k}{1-k^2\rho^2}$$

The obtained value accomplishes the request $\phi \in [\phi_1, \phi_2]$. Now, we can calculate the optimal composite estimator $\overline{y}_t^*$ as well as its variance :

$$\overline{y}_t^* = \left(1 - \frac{1-k}{1-k^2\rho^2}\right)\overline{y}_t^{nc} + \frac{1-k}{1-k^2\rho^2}\left(\overline{y}_t^c + b(\overline{y}_t - \overline{y}_t^c)\right)$$

$$V_{opt}(\overline{y}_t^*) = \frac{S^2}{n}\frac{1-k\rho^2}{1-k^2\rho^2}$$

The optimal number of the common elements, k, can be derived from the condition that

$$\frac{\partial V_{opt}}{\partial k} = 0$$

The optimal value of k is $\frac{1-\sqrt{1-\rho^2}}{\rho^2}$ and then the optimal variance is :

$$\min{}_k V_{opt}(\overline{y}_t^*) = \frac{S^2}{n}\frac{\rho^2}{1-\sqrt{1-\rho^2}}.$$

The composite estimators for the evolution

B.1. The estimator considering the difference between the composite estimators of two successive levels

We are not able to construct a composite estimator for the $(t-1)$-th occasion (unless we take information from the t-th occasion), so we take the estimator $\overline{y}_{t-1}$ and for the t-th one, the corresponding composite estimator, $\overline{y}_t^*$. Then the estimator for the evolution $\Delta\theta = \theta_t - \theta_{t-1}$ is :

$$\widehat{\Delta\theta}_1^* = \overline{y}_t^* - \overline{y}_{t-1} = (1-\phi)\overline{y}_t^{nc} + \phi\left[\overline{y}_t^c + b(\overline{y}_{t-1} - \overline{y}_{t-1}^c)\right] - \overline{y}_{t-1}$$

The variance for $\phi = \phi_{opt} = \frac{1-k}{1-k^2\rho^2}$ and $b = \rho$ has the form :

$$V(\widehat{\Delta\theta}_1^*) = \frac{S^2}{n}\left[1 + \frac{1}{1-k^2\rho^2}(1 - k\rho^2 - 2(1-k)\rho)\right]$$

B.2. The direct composite estimator

Caron & Ravalet (2000) propose as direct composite estimator, the one which combines the estimator for the evolution based on the common part of the two samples with the corresponding one for the uncommon part. It results :

$$\widehat{\Delta\theta}_2^* = \phi\left(\overline{y}_t^c - \overline{y}_{t-1}^c\right) + (1-\phi)\left(\overline{y}_t^{nc} - \overline{y}_{t-1}^{nc}\right)$$

with $\phi \in (0,1)$. As in the previous cases, we are interested in finding the value of ϕ and eventually, of k and ρ, that makes optimal the survey design. Based on the same reason as above, the best value for ϕ is

$$\phi_{opt} = \frac{V_{nc}}{V_c + V_{nc}}$$

where, for simplicity, we have denoted by $V_c = V\left(\overline{y}_t^c - \overline{y}_{t-1}^c\right)$ and similarly, $V_{nc} = V\left(\overline{y}_t^{nc} - \overline{y}_{t-1}^{nc}\right)$. As we deal with simple random sampling without replacement, we obtain :

$$V_{nc} = 2\frac{S^2}{kn}$$

$$V_c = 2\frac{S^2}{n}\frac{1-\rho}{1-k}$$

It results then $\phi_{opt} = \dfrac{1-k}{1-k\rho}$ and

$$\widehat{\Delta\theta}_{2,opt}^* = \frac{1-k}{1-k\rho}\left(\overline{y}_t^c - \overline{y}_{t-1}^c\right) + \left(1 - \frac{1-k}{1-k\rho}\right)\left(\overline{y}_t^{nc} - \overline{y}_{t-1}^{nc}\right) \qquad (1.2)$$

with the corresponding optimal variance :

$$V_{opt}(\widehat{\Delta\theta}_2^*) = 2\frac{S^2}{n}\frac{1-\rho}{1-k\rho}.$$

It is of interest to compare $\widehat{\Delta\theta}_{2,opt}^*$ with the previous estimators ; let us consider the following ratios :

i. between $\widehat{\Delta\theta}_{2,opt}^*$ and the natural estimator of evolution, $\widehat{\Delta\theta}_1 = \overline{y}_t - \overline{y}_{t-1}$:

$$\frac{V_{opt}(\widehat{\Delta\theta}_2^*)}{V(\widehat{\Delta\theta}_1)} = \frac{1-\rho}{1-\rho[1-k\rho(1-k)]} < 1.$$

ii. between $\widehat{\Delta\theta}^*_{2,opt}$ and the estimator of the evolution constructed only on the common part of the sample $\widehat{\Delta\theta}_2 = \overline{y}^c_t - \overline{y}^c_{t-1}$:

$$\frac{V_{opt}(\widehat{\Delta\theta}^*_2)}{V(\widehat{\Delta\theta}_2)} = \frac{1-k}{1-k\rho} < 1.$$

So, the direct composite estimator is more efficient. We can notice that the $\widehat{\Delta\theta}^*_{2,opt}$ is the same $\Delta\hat{\theta}^{opt}_t$.

Patterson's Estimator

Under the general frame used until now, we add the following supplementary conditions imposed by Patterson (1950). We suppose that the empirical variance is stationary and that the covariance between the periods t and $t-s$ is as follows

$$\rho(\mathcal{Y}_{t-s}, \mathcal{Y}_t) = \rho^s.$$

The goal of Patterson's paper is to obtain the estimates of minimum variance of the mean θ_t, assuming that one is restricted to the class of linear unbiased estimates. The method of sampling is one-level rotation sampling with n elements sampled at each occasion, from which λn common elements and μn are replaced each time. Patterson (1950) gives a necessary and sufficient condition for a linear unbiased estimator to have minimum variance.

Patterson (1950) derives the estimator of minimum variance for one level rotation. It is denoted by $\overline{y}^{Pat}_t$ and

$$\overline{y}^{Pat}_t = (1-\varphi_t)\left[\overline{y}^c_t + \rho(\overline{y}^{Pat}_{t-1} - \overline{y}^c_{t-1})\right] + \varphi_t \overline{y}^{nc}_t$$

where φ_t satisfies the relation :

$$(1-\varphi_t)\left[\frac{\sigma^2(1-\rho^2)}{\lambda n} + \rho^2 \text{ Var } (\overline{y}^{Pat}_{t-1})\right] = \frac{\varphi_t \sigma^2}{\mu n}$$

Patterson calculates the form of the variance of $\overline{y}^{Pat}_t$ and he shows that the sequence φ_t converges to a value φ which is reached rapidly and as a result φ_t can be replaced by φ after the second occasion.

Result 1.7.3 *i. Var* $(\overline{y}^{Pat}_t) = \frac{\varphi_t \sigma^2}{\mu n}$.

ii. The sequence φ_t *is convergent.*

The Patterson's estimate for the evolution

i. **The evolution between two successive occasions**

In the case of estimation the evolution of the mean between two successive periods, Patterson (1950) proposes to calculate separately the estimator for each period and then to take the difference of them. Let $\overline{y}_{t-1}^{Pat}$ and $\overline{y}_{t}^{Pat}$ be the corresponding Patterson's estimator for $t-1, t$ levels ; then $\overline{y}_{t}^{Pat} - \overline{y}_{t-1}^{Pat}$ will be the estimator for the evolution.

Evaluation of the variance

We need the formula for the covariance between $\overline{y}_{t-1}^{Pat}$ and $\overline{y}_{t}^{Pat}$:

$$\begin{aligned} \text{Cov } (\overline{y}_{t}^{Pat}; \overline{y}_{t-1}^{Pat}) &= \rho(1-\varphi_t) \text{ Cov } (\overline{y}_{t-1}^{Pat}; \overline{y}_{t-1}^{Pat}) \\ &= \frac{\rho(1-\varphi_t)\varphi_{t-1}\sigma^2}{\mu n} \end{aligned}$$

We have then the variance formula :

$$\text{Var } (\overline{y}_{t}^{Pat} - \overline{y}_{t-1}^{Pat}) = \frac{\sigma^2}{\mu n} [\varphi_t + \varphi_{t-1} - 2\rho(1-\varphi_t)\varphi_{t-1}]$$

and if we replace all φ_t by the common value, φ, we get the simpler value for the variance :

$$\text{Var } (\overline{y}_{t}^{Pat} - \overline{y}_{t-1}^{Pat}) = \frac{2\varphi\sigma^2}{\mu n} [1 - \rho(1-\varphi)]$$

ii. **The evolution between not-successive occasions**

Patterson (1950) gives an estimator in the case of the evolution between two periods, separated by an interval of s periods as the difference between $\overline{y}_{t}^{Pat}$ and $\overline{y}_{t-s}^{Pat}$ the corresponding level estimators at $t-s, t$-th occasions. We are interested in calculating the variance of the estimator for the evolution. As in the case of two successive occasions, we need the covariance.

$$\begin{aligned} \text{Cov } (\overline{y}_{t}^{Pat}; \overline{y}_{t-s}^{Pat}) &= \rho(1-\varphi_t) \text{ Cov } (\overline{y}_{t-s}^{Pat}; \overline{y}_{t-1}^{Pat}) \\ &= \frac{\varphi_{t-s}(1-\varphi_{t-s+1})\ldots(1-\varphi_t)\rho^s\sigma^2}{\mu n} \end{aligned}$$

where $\overline{y}_{t-s}^{Pat}$ is based on information from the first $t-s$ occasions only ; if we suppose that $t-s$ is large enough to put $\varphi_{t-s} = \varphi$, the covariance and variance are given by

$$\text{Cov } (\overline{y}_{t}^{Pat}; \overline{y}_{t-s}^{Pat}) = \frac{\varphi(1-\varphi)^s\rho^s\sigma^2}{\mu n}$$

$$\text{Var } (\overline{y}_{t}^{Pat} - \overline{y}_{t-s}^{Pat}) = \frac{2\varphi\sigma^2}{\mu n} [1 - \rho^s(1-\varphi)^s]$$

The estimator of the mean from a past occasion as a function of the whole sequence of occasions

We can use the t-th occasion to improve the estimate at the $t-1$-th occasion. The proposed estimator has the form

$$\overline{y}_{t-1,t}^{Pat} = \overline{y}_{t-1}^{Pat} - \rho\varphi_{t-1}(\overline{y}_t^{Pat} - \overline{y}_t^{nc})$$

with the variance

$$\text{Var}\ (\overline{y}_{t-1,t}^{Pat}) = \left[1 - \rho^2\varphi_{t-1}\right](1-\varphi_t)\frac{\varphi_{t-1}\sigma^2}{\mu n}$$

It can be observed that this estimator has a smaller variance than $\overline{y}_t^{Pat}$. Also, the estimator for the evolution is constructed as follows

$$\overline{y}_t^{Pat} - \overline{y}_{t-1,t}^{Pat} = (1+\rho\varphi_{t-1})\overline{y}_t^{Pat} - \rho\varphi_{t-1}\overline{y}_t^{nc} - \overline{y}_{t-1}^{Pat}.$$

Patterson generalizes this by obtaining an estimator for the $(t-k)$-th level when there are t levels in total.

1.7.3 The Variance Estimation

In this framework, the calculation of the variance is always almost impossible. Caron & Ravalet (2000) enumerate the existing possibilities for evaluating the variance estimator :

i. It exists a software capable to calculate the variance of the total of a variable. As for the estimator of the evolution, as it has already been specified, the variance estimator is obtained by considering the new variable $z_i = y_{i,t} - y_{i,t-1}$ for all the individuals from the population and by calculating the variance estimation for the total of the new variable z_i.

ii. No software is available. We are confronted with this problem in the case of a complex survey and the solution is the same as in the case of a survey which does not depend on time, namely to reduce to the case of a simple survey in which we know how to calculate the variance estimation. This is achieved by means of the "design effect", noted as deff, when we have its value or its estimation. The general formula is

$$deff = \frac{V(\hat{T})}{V_{swr}(\hat{T}_{swr})}$$

where $V(\hat{T})$ is the variance of the estimator for the total of Y when it was used any survey design and $V_{swr}(\hat{T}_{swr})$ in the case of simple sampling without replacement.

Knowing an estimator for $V_{swr}(\hat{T}_{swr})$ and deff or $\widehat{deff}$, we deduce an estimator for $V(\hat{T})$ as

$$\hat{V}(\hat{T}) = \widehat{deff}\hat{V}_{swr}(\hat{T}_{swr})$$

It can be deduced in this way, the variance estimator for all the survey designs considered until now.

Chapitre 2

Variance et estimation de la variance pour des statistiques complexes dans le cas de deux échantillons

2.1 Introduction

Le problème de l'estimation d'une combinaison linéaire de totaux quand l'information provient de plusieurs échantillons apparaît dans beaucoup de situations pratiques. L'exemple le plus naturel est celui où deux échantillons correspondent à des sondages effectués à des instants différents du temps. On obtient ce que dans la littérature on appelle des *enquêtes répétées*. Il existe des situations plus générales, comme celle de l'estimation d'une statistique complexe telle que le ratio par exemple, en présence de nonréponse. Si pour le premier exemple on trouve une biblioghaphie assez riche (voir chapitre 1) bien qu'en général, les résultats soient obtenus dans des situations plutôt particulières (des plans simples sont utilisés pour les deux époques), peu de travaux sont consacrés à l'estimation d'un ratio où chaque variable est observée sur un échantillon différent.

Nous voulons donner dans la suite une étude générale de ce problème "à deux échantillons" qui puisse s'appliquer en particulier aux deux situations présentées ci-dessus.
L'idée de considérer un plan multidimensionnel apparaît dans la coordination d'échantillons. Pour traiter cela, Cotton & Hesse (1992) et Salamin (2002) définissent un plan multidimensionnel et les probabilités d'inclusion correspondantes. Cotton & Hesse (1992) donnent des exemples de plans mul-

tidimensionnels usuels.

En s'appuyant sur les travaux cités précédemment, nous étudions l'estimation d'un paramètre d'intérêt qui dépend, via une fonction linéaire où non, de variables d'intérêt qui sont mesurées sur des échantillons différents. Notre analyse porte plus particulièrement sur le cas bidimensionnel. Nous proposons des estimateurs de type Horvitz-Thompson sur deux échantillons. Un estimateur linéaire pondéré et sans biais est construit pour une combinaison linéaire de totaux. On donne ensuite une formule de type Horvitz-Thompson pour la variance et son estimateur. Une des difficultés provient du fait que dans le cas de deux échantillons il existe une infinité d'estimateurs sans biais, ce qui n'est pas le cas pour un unique échantillon. Nous déterminons celui dont la variance est minimale sous des conditions générales. Le nombre de paramètres qui caractérisent les poids est proportionnel à la taille de la population ce qui rend impossible leur calcul en général. Nous proposons donc des reparamétrisations qui réduisent considérablement ce nombre de paramètres et qui correspondent à des situations pratiques de stratification de la population. Sous ces conditions et sous les plans usuels nous obtenons les meilleurs estimateurs linéaires sans biais. Enfin, nous étudions le problème des statistiques complexes, non linéaires, en utilisant la technique de linéarisation par la fonction d'influence.

2.2 Problème à deux échantillons

2.2.1 Le plan de sondage multidimendionnel

Nous considérons une population finie $\mathcal{U}_t$ indexée par t, élément de l'ensemble fini $\boldsymbol{\mathcal{T}} = \{1, \ldots, t, \ldots, T\}$. On commence notre étude par la situation la plus simple quand la population $\mathcal{U}_t$ reste inchangée au passage d'un instant à l'autre, c'est à dire $\mathcal{U}_t = \mathcal{U}$ pour tous $t \in \mathcal{T}$ avec $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$ une population finie de taille N. Alors, la population cible est $\otimes_{t\in\mathcal{T}}\mathcal{U}_t = \mathcal{U}^T$.

Définition 2.2.1 *: Tout sous ensemble dans $\mathcal{U}^T$ sera nommé échantillon multidimensionnel où échantillon T-dimensionnel.*

L'ensemble de tous les échantillons T-dimensionnels est noté $\mathcal{S}$ et il est donné par

$$\mathcal{S} \quad = \quad \{\boldsymbol{s} = (s_1, \ldots, s_t, \ldots, s_T) \subset [\mathcal{P}(\mathcal{U})]^T, s_t \subset \mathcal{P}(\mathcal{U}), t \in \boldsymbol{\mathcal{T}}\}. \quad (2.1)$$

Chaque élément de $\mathcal{S}$ peut être considéré comme la réalisation d'une variable aléatoire multidimensionnelle $\mathbf{S} = (S_1, \ldots, S_T)$

$$\begin{aligned} \mathbf{S} \quad &: \quad (\otimes\Omega, \otimes\mathcal{K}, P) \rightarrow (\mathcal{S}, [\mathcal{P}(\mathcal{U})]^T) \\ &\quad\ \mathbf{S}(w) = \boldsymbol{s} \in \mathcal{S} \end{aligned}$$

où P est une probabilité sur un espace mesurable quelconque $(\otimes\Omega, \otimes\mathcal{K})$.

Définition 2.2.2 *: Un plan de sondage T-dimensionnel $p(\boldsymbol{s})$ est la probabilité de sélectionner un échantillon T-dimensionnel $\boldsymbol{s}$.*

Plus précisément, si $\mathbf{S}$ est la variable aléatoire définie ci-dessus alors le plan de sondage T-dimensionnel p est la probabilité sur $(\mathcal{S}, [\mathcal{P}(\mathcal{U})]^T)$ définie par

$$P(\mathbf{S} = \boldsymbol{s}) = p(\boldsymbol{s}) \quad \text{pour tous } \boldsymbol{s} \in \mathcal{S}. \tag{2.2}$$

Comme toute loi de probabilité, $p(\boldsymbol{s})$ vérifie les relations :

$$\begin{cases} p(\boldsymbol{s}) \geq 0 \quad \text{pour tous} \quad \boldsymbol{s} \in \mathcal{S} \\ \sum_{\boldsymbol{s}\in\mathcal{S}} p(\boldsymbol{s}) = 1. \end{cases}$$

Dans un premier temps, des plans de sondage marginaux unidimensionnels et T'-dimensionnels, avec $T' < T$, ou encore conditionnels peuvent être déduits en appliquant les propriétés d'une loi de probabilité multidimensionnelle (Cotton & Hesse 1992). Par exemple, la probabilité de tirer l'échantillon s_t pour $t \in \mathcal{T}$ sachant $p(\boldsymbol{s})$ est $p_t(s_t)$ où

$$p_t(s_t) = \sum_{s_{t'}, t' \neq t \in \mathcal{T}} p(\mathbf{s}).$$

Plus généralement, des probabilités marginales d'ordre plus grand peuvent être obtenues. Par exemple, la probabilité de sélectionner un couple (s_{t_1}, s_{t_2}) sachant le plan $p(\boldsymbol{s})$ est $p_{t_1,t_2}(s_{t_1}, s_{t_2})$ donnée par

$$p_{t_1,t_2}(s_{t_1}, s_{t_2}) = \sum_{s_{t'}, t' \neq t_1, t_2 \in \mathcal{T}} p(\mathbf{s}).$$

Alors des plans conditionnels peuvent être déduits. Par exemple, on obtient l'échantillon s_t sachant la probabilité de sélectionner $(s_1, \ldots, s_{t-1})$ avec $p_{t|1,\ldots,t-1}(s_t|s_1, \ldots, s_{t-1})$ donnée par

$$p_{t|1,\ldots,t-1}(s_t|s_1, \ldots, s_{t-1}) = \frac{p_{1,\ldots,t}(s_1, \ldots, s_t)}{p_{1,\ldots,t-1}(s_1, \ldots, s_{t-1})}$$

Si on a un plan de sondage unidimensionnel comme $p_t(s_t)$ ci-dessus, alors on peut utiliser toutes les notions définies dans le cas classique de la théorie des sondages, comme par exemple les probabilités d'inclusion du premier et du deuxième degré. S'il s'agit par contre d'un plan de sondage multidimensionnel, on a besoin de nouvelles définitions de ces quantités, définitions données dans la suite.

Dans un deuxième temps, on peut calculer la loi de toute opération algébrique (intersection, réunion, différence) entre les s_t pour $t \in \mathcal{T}$. Par exemple, la probabilité de sélectionner l'échantillon intersection $s_{t_1 \cap t_2} = s_{t_1} \cap s_{t_2}$ est $p_{\cap t_1,t_2}$ donnée par

$$p_{\cap t_1,t_2}(\sigma) = \sum_{\sigma \in s_{t_1} \cap s_{t_2}} p_{t_1,t_2}(s_{t_1}, s_{t_2})$$

Enfin, des plans de sondage usuels peuvent être obtenus pour des expressions particulières de $p(\mathbf{s})$:

i. Pour $p(\mathbf{s}) = \Pi_{t \in \mathcal{T}} p_t(s_t)$ on a l'indépendance des échantillons s_t, pour $t \in \mathcal{T}$.

ii. Pour $p(s_1 = \ldots = s_T) = 1$ on obtient le panel.

iii. Pour $T = 2$ et $p(s_2|s_1) = 0$ sauf si $s_2 \subset s_1$, on obtient le plan de sondage à deux-phases.

iv. Pour $T = 2$ et $p(s_2|s_1) = 0$ sauf si $s_1 \cap s_2 = \emptyset$, on sélectionne deux échantillons disjoints dans $\mathcal{U}$ ou le plan à deux-phases dans l'échantillon complémentaire $\mathcal{U} - s_1$.

On donne dans la suite des extensions des notions définies dans le cas classique de la théorie des sondages : les indicatrices d'appartenance dans un échantillon et les probabilités d'inclusion.

Les indicatrices d'appartenance à un échantillon multidimensionnel

Dans le cas d'un seul échantillon s, un individu $k \in \mathcal{U}$ peut appartenir ou non à l'échantillon comme l'indique les valeurs 1 ou 0 d'une variable de Bernoulli $\varepsilon_k = \mathbf{1}_{\{k \in s\}}$ (Särndal, Swensson & Wretman 1992). La situation devient plus compliquée dans le cas T-dimensionnel à cause du nombre beaucoup plus élevé d'échantillons dans lesquels un individu peut se trouver.
Nous allons étendre dans la suite cette définition au cas T-dimensionnel.

Pour chaque $t \in \mathcal{T}$ et $k \in \mathcal{U}$, on note ε_k^t l'indicatrice d'appartenance de l'individu k à l'échantillon s_t, c'est à dire

$$\varepsilon_k^t = \begin{cases} 1 & si \quad k \in s_t \\ 0 & sinon \end{cases}$$

$\varepsilon_k^t = \mathbf{1}_{\{k \in s_t\}}$ est la variable indicatrice dans le cas unidimensionnel (Särndal, Swensson & Wretman 1992).

Soit $\boldsymbol{\varepsilon}_k = (\varepsilon_k^1, \ldots, \varepsilon_k^T)$ la variable aléatoire T-dimensionnelle qui indique l'appartenance de l'individu k à l'échantillon T-dimensionnel $\mathbf{s}$.

$$\begin{aligned} \boldsymbol{\varepsilon}_k &: (\mathcal{S}, [\mathcal{P}(\mathcal{U})]^T) \to \{0,1\}^T \\ \boldsymbol{\varepsilon}_k(\mathbf{s}) &= (\varepsilon_k^1(s_1), \ldots, \varepsilon_k^T(s_T)) \text{ pour tous } k \in \mathcal{U}. \end{aligned} \tag{2.3}$$

Par conséquent, l'appartenance d'un individu $k \in \mathcal{U}$ à l'échantillon joint $\mathbf{s} = (s_1, \ldots, s_T)$ est complètement décrite par l'inclusion de k dans un des $2^{2^T-1}-1$ échantillons de l'algèbre engendrée par le T-uple $(\varepsilon_k^1, \ldots, \varepsilon_k^T)$, notée $\sigma(\boldsymbol{\varepsilon}_k)$. Cette algèbre est de dimension $2^T - 1$. Soit $\mathcal{B}$ une base pour $\sigma(\boldsymbol{\varepsilon}_k)$. On va donner un rôle privillégié à la base formée par les indicatrices qui correspondent aux échantillons disjoints, c'est-à-dire

$$\mathcal{B} = \{\varepsilon_k^t, t \in \sigma(\mathcal{T}) \quad s_t \cap s_{t'} = \emptyset \quad t \neq t' \in \sigma(\mathcal{T})\} \tag{2.4}$$

où $\sigma(\mathcal{T})$ est la tribu engendrée par $\mathcal{T}$. On introduit l'ensemble de tous les échantillons T-dimensionnel avec des composantes disjointes :

$$\mathcal{S}^{disj} = \{\boldsymbol{s}^{disj} = (s_t)_{t \in \sigma(\mathcal{T})} \in [\mathcal{P}(\mathcal{U})]^{2^T-1}, \quad s_t \cap s_{t'} = \emptyset \quad t \neq t' \in \sigma(\mathcal{T})\}$$

Alors, au lieu de considérer comme variable indicatrice la variable $\boldsymbol{\varepsilon}_k$ donnée auparavant, on va prendre plutôt celle qui a comme composantes des indicatrices qui correspondent aux échantillons disjoints. On a la définition suivante :

Définition 2.2.3 *: Pour chaque individu $k \in U$, on appelle variable indicatrice T-dimensionnelle la variable (2^T-1)-dimensionnelle $\boldsymbol{\varepsilon}_k^{\mathcal{B}}$ définie comme suit*

$$\begin{cases} \boldsymbol{\varepsilon}_k^{\mathcal{B}} = (\varepsilon_k^t) \quad avec \quad \varepsilon_k^t \in \mathcal{B} \\ \boldsymbol{\varepsilon}_k^{\mathcal{B}} : (\mathcal{S}^{disj}, [\mathcal{P}(\mathcal{U})]^{2^T-1}) \to \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_j, \ldots, \boldsymbol{e}_{2^T-1}\} \end{cases} \tag{2.5}$$

où $\mathbf{e}_j$ est le vecteur de dimension $2^T - 1$ avec 1 sur la j-ème position et zéro ailleurs.

On voit alors que le problème devient vite assez compliqué à cause du nombre élevé de quantités qui entrent en jeu. Par exemple, pour deux échantillons, l'algèbre va avoir $2^3 - 1$ éléments dont 3 indépendants et si on a trois échantillons, elle a $2^7 - 1$ éléments dont 7 indépendants. Cependant, Salamin (2002) donne une expression pour les probabilités d'inclusion pour le cas général T-dimensionnel.

À cause de ces raisons pratiques, on considère dans la suite le cas bidimensionnel.

2.2.2 Le plan de sondage bidimensionnel

On donne les expressions explicites des notions définies dans la section précédente (indicatrices, base) suivi des définitions des probabilités d'inclusion. À la fin de cette section on donne quelques exemples de plans bidimensionnels.

On a maintenant un échantillon bidimensionnel $\mathbf{s} = (s_1, s_2)$ sélectionné dans $\mathcal{U} \times \mathcal{U}$ selon un plan de sondage bidimensionnel quelconque $p(\mathbf{s})$ donné par (2.2) pour $T = 2$. Pour chaque individu k dans la population, on associe la *variable de Bernoulli deux-dimensionnelle* $\boldsymbol{\varepsilon}_k$ donnée par (2.3)

$$\begin{gathered} \boldsymbol{\varepsilon}_k : (\mathcal{S}, [\mathcal{P}(\mathcal{U})]^2) \to \{0,1\}^2 \\ \boldsymbol{\varepsilon}_k = (\varepsilon_k^1, \varepsilon_k^2) \quad \text{avec} \\ \varepsilon_k^1 = \mathbf{1}_{\{k \in s_1\}}, \quad \varepsilon_k^2 = \mathbf{1}_{\{k \in s_2\}}. \end{gathered} \tag{2.6}$$

Introduisons les notations suivantes :

$$\begin{aligned} s_{12} &= s_1 \cap s_2 \\ s_{1*} &= s_1 - s_{12} \\ s_{2*} &= s_2 - s_{12} \\ s_{**} &= \mathcal{U} - \{s_1 \cup s_2\} \end{aligned}$$

De manière générale, on utilisera les indices 12 pour l'intersection, $1*$ pour le complémentaire de l'intersection dans s_1 et $2*$ pour le complémentaire de l'intersection dans s_2 (voir figure 2.2.2).
Correspondant à (2.4) nous avons besoin des indicatrices d'appartenance aux échantillons disjoints. On note pour cela $\varepsilon_k^{\oplus} = \mathbf{1}_{\{k \in s_{\oplus}\}}$ pour $\oplus \in \{1*, 12, 2*\}$. Par conséquent, les relations suivantes sont vérifiées

$$\varepsilon_k^1 = \varepsilon_k^{1*} + \varepsilon_k^{12}, \quad \varepsilon_k^2 = \varepsilon_k^{2*} + \varepsilon_k^{12}$$

L'algèbre engendrée par $\boldsymbol{\varepsilon}_k$ contient sept éléments dont trois indépendants :

$$\sigma(\boldsymbol{\varepsilon}_k) = \{\varepsilon_k^{1*}, \varepsilon_k^{12}, \varepsilon_k^{2*}, \varepsilon_k^{1*} + \varepsilon_k^{2*}, \varepsilon_k^{1}, \varepsilon_k^{2}, \varepsilon_k^{1} + \varepsilon_k^{2}.\}$$

Plus précisement, on peut voir les variables ε_k^1 et ε_k^2 avec leurs intersection ε_k^{12} comme trois vecteurs partant de la même origine 0. Alors les éléments de $\sigma(\boldsymbol{\varepsilon}_k)$ sont les sommets du cube construit à partir de ε_k^1, ε_k^2 et ε_k^{12} (cf. figure 2.2).

FIG. 2.1 – échantillon bidimensionnel

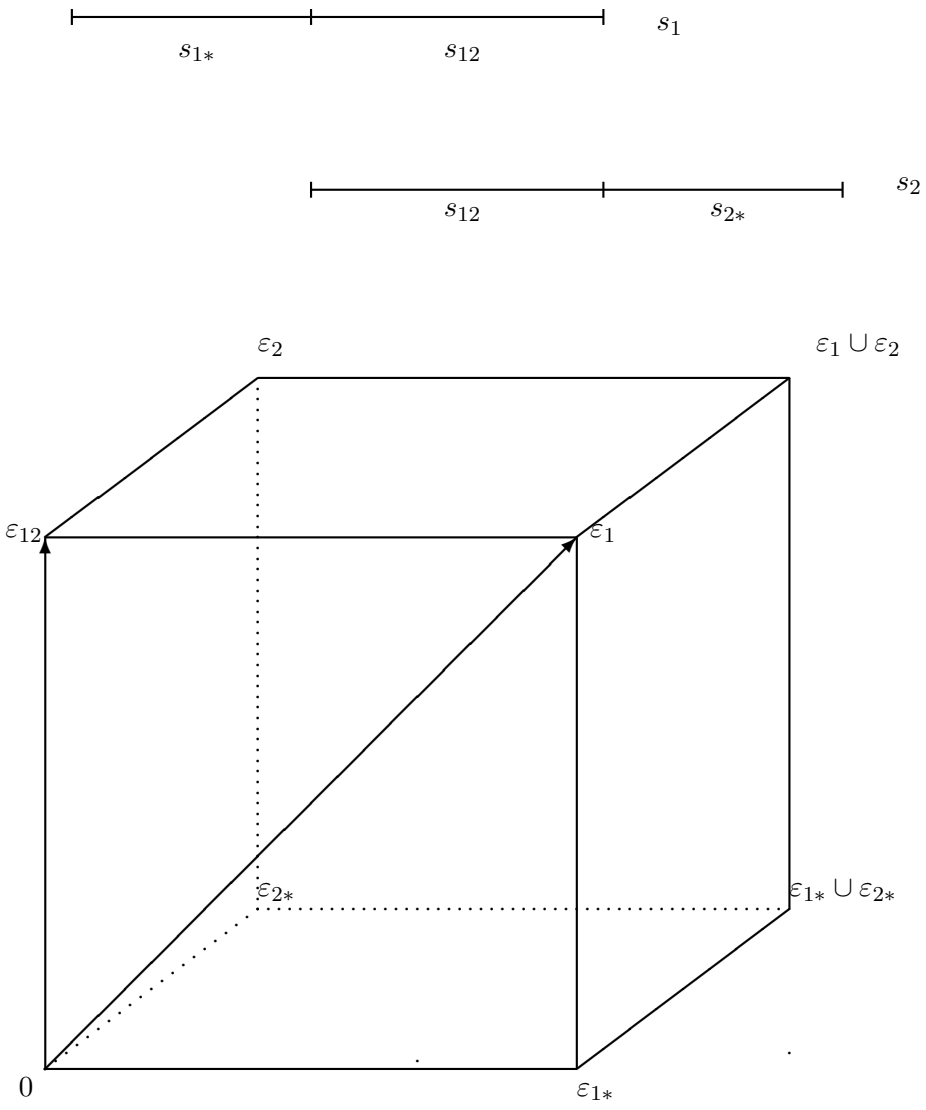

FIG. 2.2 – Plan bidimensionnel

Considérons un triplet de variables dans $\sigma(\boldsymbol{\varepsilon}_k)$ et les vecteurs d'origine 0 associés. Si ces trois vecteurs ainsi obtenus ne se trouvent pas dans le même plan, alors le triplet considéré forme une base pour $\sigma(\boldsymbol{\varepsilon}_k)$. Il y a 6 triplets de variables dans $\sigma(\boldsymbol{\varepsilon}_k)$ qui donnent des vecteurs situés dans le même plan et qui, par conséquent, ne forment pas une base. C'est par exemple le cas des variables $\varepsilon_k^{12}, \varepsilon_k^{2*}, \varepsilon_k^2$. On va donc avoir $C_7^3 - 6 = 29$ bases possibles.

Dans la suite de notre étude nous allons utiliser le triplet des variables donné par (2.4) quand $T = 2$,

$$\mathcal{B} = \{\varepsilon_k^{1*}, \varepsilon_k^{12}, \varepsilon_k^{2*}\} \tag{2.7}$$

comme base pour la tribu $\sigma(\boldsymbol{\varepsilon}_k)$. Remarquons qu'on peut aussi bien développer l'analyse en prenant une autre base $\mathcal{B}'$ parmi les 29 existantes puisqu'elles sont liées par des transformations linéaires $\mathcal{B}' = \mathcal{A}\mathcal{B}$ avec $\mathcal{A}$ matrice ; par exemple on peut determiner la base $\mathcal{B}' = \{\varepsilon_k^1, \varepsilon_k^2, \varepsilon_k^{12}\}$ à partir de $\mathcal{B} = \{\varepsilon_k^{1*}, \varepsilon_k^{2*}, \varepsilon_k^{12}\}$

$$\begin{pmatrix} \varepsilon_k^1 \\ \varepsilon_k^2 \\ \varepsilon_k^{12} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_k^{1*} \\ \varepsilon_k^{2*} \\ \varepsilon_k^{12} \end{pmatrix}$$

La variable indicatrice 3-dimensionnelle est donnée par (2.5) pour le cas particulier $T = 2$:

$$\begin{cases} \boldsymbol{\varepsilon}_k^{\mathcal{B}} = (\varepsilon_k^t) \quad \text{avec} \quad \varepsilon_k^t \in \mathcal{B} \\ \boldsymbol{\varepsilon}_k^{\mathcal{B}} : (\mathcal{S}^{disj}, [\mathcal{P}(\mathcal{U})]^3) \to \{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\} \end{cases} \tag{2.8}$$

avec $\mathcal{B}$ donnée par (2.7) et $\boldsymbol{e}_j$ le vecteur tri-dimensionnel avec 1 sur la j-ème position et zéro ailleurs pour $j = 1, 2, 3$. Il résulte alors que $\boldsymbol{\varepsilon}_k^{\mathcal{B}} = (\varepsilon_k^{1*}, \varepsilon_k^{12}, \varepsilon_k^{2*})$.

On donne maintenant la définition de la taille d'un échantillon bidimensionnel.

Définition 2.2.4 *La taille d'un échantillon bidimensionnel $\boldsymbol{s} = (s_1, s_2)$ est définie par*

$$n_{\boldsymbol{s}} = \sum_{k \in U} \boldsymbol{\varepsilon}_k^{\mathcal{B}} = (n_{1*}, n_{12}, n_{2*}) \tag{2.9}$$

où n_{1*}, n_{2*}, n_{12} sont les tailles des échantillons s_{1*}, s_{2*} et respectivement de l'intersection s_{12}.

La taille $n_{\boldsymbol{s}}$ qui est un vecteur dans le cas multidimensionnel, peut être fixe si toutes ses composantes sont de taille fixe ou variable dans le cas

contraire. Remarquons qu'on peut avoir $\boldsymbol{s}$ de taille $n_{\boldsymbol{s}}$ variable même si s_1 et s_2 sont de taille fixe mais avec une intersection de taille variable.

Nous pouvons maintenant définir les probabilités d'inclusion du premier et du deuxième degré pour $T = 2$.

2.2.3 Les probabilités d'inclusion multiples

On a déjà vu que le tirage d'un échantillon bidimensionnel a compliqué les définitions des indicatrices. Ce fait a comme conséquence directe la définition des probabilités d'inclusion multiples. Dans le cas unidimensionnel, la probabilité d'inclusion de premier degré peut s'écrire

$$\pi_k = P \circ \mathbf{S}^{-1} \circ \boldsymbol{\varepsilon}_k^{-1}(1)$$

pour $\mathbf{S}$ et $\boldsymbol{\varepsilon}_k$ définis en (2.1) et (2.3) pour $T = 1$. Par analogie, on peut donner la définition suivante :

Définition 2.2.5 *: La probabilité que l'individu $k \in U$ se trouve dans un échantillon bidimensionnel sélectionné selon un plan $p(\mathbf{s})$ est noté $\pi_k^{\boldsymbol{e}_l}$*

$$\pi_k^{\boldsymbol{e}_l} = P \circ (\mathbf{S}^{disj})^{-1} \circ (\boldsymbol{\varepsilon}_k^{\mathcal{B}})^{-1}(\boldsymbol{e}_l) \quad pour \quad l = 1, 2, 3$$

avec $\boldsymbol{\varepsilon}_k^{\mathcal{B}}$ et $\boldsymbol{e}_l$ donnés par (2.8) et $\mathbf{S}^{disj} : (\otimes\Omega, \otimes\mathcal{K}, P) \rightarrow (\mathcal{S}^{disj}, [\mathcal{P}(\mathcal{U})]^{2^T-1})$ défini par $\mathbf{S}^{disj}(w) = \mathbf{s}^{disj}$.
Les $\pi_k^{\boldsymbol{e}_l}$ sont appelés probabilités d'inclusion du premier degré bidimensionnelles.

Calculons les expressions explicites de ces probabilités. On a

$$\pi_k^{\boldsymbol{e}_l} = P \circ (\mathbf{S}^{disj})^{-1}(\boldsymbol{\varepsilon}_k^{\mathcal{B}} = \boldsymbol{e}_l) = \sum_{\mathbf{s}^{disj} \in (\boldsymbol{\varepsilon}_k^{\mathcal{B}})^{-1}(\boldsymbol{e}_l)} p(\mathbf{s}^{disj}).$$

Pour $l = 1$, on obtient

$$\pi_k^{\boldsymbol{e}_1} = \sum_{k \in s_{1*}} p(\mathbf{s})$$

probabilité qu'on va noter pour simplifier avec π_k^{1*}. De manière analogue, on obtient les deux autres probabilités notées π_k^{12}, π_k^{2*} pour tous $k \in U$.

Dans la littérature, Cotton & Hesse (1992) et Salamin (2002) définissent de manière différente les probabilités d'inclusion de premier degré bidimensionnelles. On montre qu'elles produisent les mêmes π_k-quantités que celles

construites ci-dessus.

Pour chaque unité k dans U nous pouvons définir la trace de l'échantillon $\mathbf{s}$ sur k, notée $\mathrm{tr}_k(\mathbf{s})$, comme le vecteur d'éléments les traces de s_1 et s_2 sur l'individu k, c'est à dire

$$\mathbf{s_k} = \mathrm{tr}_k(\mathbf{s}) = (s_1 \cap \{k\}, s_2 \cap \{k\}) \in \{\emptyset, \{k\}\} \times \{\emptyset, \{k\}\}.$$

Cotton & Hesse (1992) définissent les *probabilités d'inclusion du premier degré bidimensionnelles* notées $f_k(\mathbf{s_k})$ par

$$f_k(\mathbf{s_k}) = \sum_{\mathbf{s} \in \mathrm{tr}_k^{-1}(\mathbf{s_k})} p(\mathbf{s}) \text{ pour } \mathbf{s_k} \in \{\emptyset, \{k\}\}^2, k \in U.$$

Si on considère l'indicatrice $\boldsymbol{\varepsilon}_k$ donnée par (2.6) et P la probabilité sur $(\otimes\Omega, \otimes\mathcal{K})$ alors $P \circ \mathbf{S}^{-1}$ est une probabilité sur $(\mathcal{S}, [\mathcal{P}(\mathcal{U})]^2)$ et les $f_k(\mathbf{s_k})$ définis ci-dessus peuvent s'écrire sur la forme suivante

$$f_k(\mathbf{s_k}) = P \circ \mathbf{S}^{-1} \circ \boldsymbol{\varepsilon}_k^{-1}(a) \quad \text{avec} \quad a \in \{0,1\}^2$$

et alors on retrouve les probabilités d'inclusion multiples données par Salamin (2002).

On peut calculer les quatre ensembles de probabilités d'inclusion du premier degré $f_k(\{\emptyset, \emptyset\})$, $f_k(\{\emptyset, \{k\}\})$, $f_k(\{\{k\}, \emptyset\})$, $f_k(\{\{k\}, \{k\}\})$. Par exemple,

$$f_k(\{\{k\}, \{k\}\}) = \sum_{\mathbf{s} \in \mathrm{tr}_k^{-1}(\{\{k\},\{k\}\})} p(\mathbf{s}) = \sum_{k \in s_{12}} p(s_1, s_2) = \pi_k^{12}$$

et on obtient la probabilité d'inclusion du premier degré classique par rapport à la loi de l'intersection ; π_k^{12} est la probabilité d'inclusion de l'individu k dans l'intersection s_{12}. On peut obtenir de manière équivalente

$$f_k(\{\{k\}, \emptyset\}) = \pi_k^{1*} = P(k \in s_{1*})$$

$$f_k(\{\emptyset, \{k\}\}) = \pi_k^{2*} = P(k \in s_{2*})$$

$$f_k(\{\emptyset, \emptyset\}) = \pi_k^{**} = P(k \in s_{**}) = 1 - (\pi_k^{1*} + \pi_k^{12} + \pi_k^{2*})$$

pour tous $k \in U$. Contrairement au cas unidimensionnel où on avait un seul ensemble de probabilités de premier degré, pour un tirage joint de deux échantillons on dispose de trois ensembles

$$\pi_k^{1*}, \pi_k^{12}, \pi_k^{2*}$$

qui vérifient la propriété

Propriété 1 *: $E(\varepsilon_k^{\oplus}) = \pi_k^{\oplus}$ pour $\oplus \in \{1*, 12, 2*\}$.*

Nous allons définir dans la suite les probabilités d'inclusion de deuxième ordre calculées par rapport au plan bidimensionnel $\mathbf{s} = (s_1, s_2)$. Soit $\mathbf{s_{k,l}}$ la trace de s sur le couple (k, l) où

$$\mathbf{s_{k,l}} = \text{ tr }_{k,l}(\mathbf{s}) = (s_1 \cap \{k,l\}, s_2 \cap \{k,l\}) \in \{\emptyset, \{k\}, \{l\}, \{k,l\}\}^2.$$

On peut donner la définition suivante :

Définition 2.2.6 *: La probabilité que les individus k, l appartiennent à l'échantillon bidimensionnel* $\mathbf{s}$ *est définie par*

$$f_{k,l}(\mathbf{s_{k,l}}) = \sum_{\mathbf{s}\in\ tr\ _{k,l}^{-1}(\mathbf{s_{k,l}})} p(\mathbf{s}) \ \textit{pour}\ \mathbf{s_{k,l}} \in \{\emptyset, \{k\}, \{l\}, \{k,l\}\}^2.$$

Alors, $f_{k,l}(\mathbf{s_{k,l}})$ est la probabilité d'inclusion de second degré bidimensionnelle.

Quand $\mathbf{s_{k,l}}$ parcourt le domaine de $f_{k,l}$, on obtient seize ensembles de probabilités d'inclusion du second degré bidimensionnelles dépendant de six d'entre eux. On montre d'abord quels sont ces seize ensembles et on donne ensuite les relations de dépendance qui existent entre eux.
Dans le cas particulier $\mathbf{s_{k,l}} = \{\{k,l\}, \{k\}\}$ la probabilité que $k \in s_1 \cap s_2$ et $l \in s_1 - s_2$ est

$$\begin{aligned} f_{k,l}(\{\{k,l\},\{k\}\}) &= \sum_{\mathbf{s}\in\ \text{tr }_{k,l}^{-1}(\{\{k,l\},\{k\}\})} p(\mathbf{s}) = \sum_{k\in s_{12}, l\in s_{1*}} p(s_1,s_2) = \pi_{kl}^{12,1*}, k \neq l \\ f_{k,l}(\{\{k,l\},\{k\}\}) &= 0 \quad \text{si} \quad k = l; \end{aligned}$$

on a utilisé la notation $\pi_{kl}^{12,1*} = P(k \in s_{12}, l \in s_{1*})$ par analogie avec le cas unidimensionnel. En général,

$$\pi_{kl}^{\oplus,\otimes} = 0 \quad \text{pour} \quad k \neq l \quad \text{et} \quad \oplus \neq \otimes \in \{1*, 12, 2*, **\}$$

De manière analogue on obtient toutes les probabilités d'inclusion du deuxième degré bidimensionnelles données dans le tableau ci-dessous.

k,l	1*	12	2*	**
1*	$\pi_{kl}^{1*,1*}$	$\pi_{kl}^{1*,12}$	$\pi_{kl}^{1*,2*}$	$\pi_{kl}^{1*,**}$
12	$\pi_{kl}^{12,1*}$	$\pi_{kl}^{12,12}$	$\pi_{kl}^{12,2*}$	$\pi_{kl}^{12,**}$
2*	$\pi_{kl}^{2*,1*}$	$\pi_{kl}^{2*,12}$	$\pi_{kl}^{2*,2*}$	$\pi_{kl}^{2*,**}$
**	$\pi_{kl}^{**,1*}$	$\pi_{kl}^{**,12}$	$\pi_{kl}^{**,2*}$	$\pi_{kl}^{**,**}$

Propriété 2 *: Les probabilités d'inclusion de deuxième degré bidimensionnelles vérifient*

$$\pi_{kl}^{\oplus,\otimes} = E(\varepsilon_k^{\oplus}\varepsilon_l^{\otimes})$$

pour tous $k, l \in U$ *et* $\oplus, \otimes \in \{1*, 12, 2*, **\}$.

Plusieurs remarques peuvent être faites en analysant le tableau ci-dessus :

i. les éléments qui se trouvent sur la diagonale sont les probabilités d'inclusion d'ordre deux classiques (Särndal, Swensson & Wretman 1992), c'est à dire

$$\pi_{kl}^{\oplus,\oplus} = \pi_{kl}^{\oplus} \text{ pour tous } k, l \in U, \oplus \in \{1*, 12, 2*, **\}.$$

On a alors dans un premier temps trois ensembles de probabilités d'inclusion d'ordre deux bidimensionnelles :

$$\pi_{kl}^{1*,1*}, \quad \pi_{kl}^{12,12}, \quad \pi_{kl}^{2*,2*} \quad \text{pour tous} \quad k, l \in U.$$

ii. la somme des éléments sur chaque ligne donne la probabilité d'inclusion de premier degré, respectivement $\pi_k^{\oplus}$ avec $\oplus \in \{1*, 12, 2*\}$ et par conséquent les éléments de la dernière colonne peuvent être déduits :

$$\pi_{kl}^{\oplus,**} = \pi_k^{\oplus} - (\pi_{kl}^{\oplus,1*} + \pi_{kl}^{\oplus,12} + \pi_{kl}^{\oplus,2*}) \quad \text{pour} \quad \oplus \in \{1*, 12, 2*\}.$$

iii. pour un certain couple k, l avec $k \neq l$ les éléments situés sur les diagonales secondaires sont différents :

$$\pi_{kl}^{\oplus,\otimes} \neq \pi_{kl}^{\otimes,\oplus} \text{ pour } \oplus \neq \otimes \in \{1*, 12, 2*, **\}$$

mais

$$\pi_{kl}^{\oplus,\otimes} = \pi_{lk}^{\otimes,\oplus} \text{ pour } \oplus \neq \otimes \in \{1*, 12, 2*, **\}.$$

La dernière relation implique que les matrices $(\pi_{kl}^{\oplus,\otimes})_{k,l\in U}$ et $(\pi_{kl}^{\otimes,\oplus})_{k,l\in U}$ ont les même éléments, plus precisément l'une est la transposé de l'autre. Par conséquent, on a seulement besoin des ensembles de probabilités qui se trouvent au dessus de la diagonale principale sauf la dernière colonne, c'est à dire

$$\pi_{kl}^{1*,12}, \quad \pi_{kl}^{1*,2*}, \quad \pi_{kl}^{12,2*} \text{ pour tous } k, l \in U.$$

Par analogie avec Särndal *et al.* (1992) dans le cas unidimensionnel, on donne la définition suivante

Définition 2.2.7 *: Soit* $\Delta_{kl}^{\oplus,\otimes} = \pi_{kl}^{\oplus,\otimes} - \pi_k^{\oplus}\pi_l^{\otimes}$ *pour tous* $k, l \in U$ *et pour un certain couple* $\oplus, \otimes \in \{1*, 12, 2*\}$.

Alors pour un certain couple $\oplus, \otimes$ la matrice $(\Delta_{kl}^{\oplus,\otimes})_{k,l\in U}$ a la propriété suivante

Propriété 3 *: La matrice $(\Delta_{kl}^{\oplus,\otimes})_{k,l\in U}$ est la matrice de covariance entre les vecteurs $\boldsymbol{\varepsilon}^{\oplus} = (\varepsilon_k^{\oplus})_{k\in U}$ et $\boldsymbol{\varepsilon}^{\otimes} = (\varepsilon_k^{\otimes})_{k\in U}$ pour $\oplus, \otimes \in \{1*, 12, 2*\}$,*

$$(\Delta_{kl}^{\oplus,\otimes})_{k,l\in U} = Cov(\boldsymbol{\varepsilon}^{\oplus}, \boldsymbol{\varepsilon}^{\otimes})$$

Pour $\oplus = \otimes$ les quantités classiques du cas unidimensionnel sont obtenues et c'est pourquoi on va utiliser la notation avec un seul exposant $\Delta_{kl}^{\oplus} = \pi_{kl}^{\oplus} - \pi_k^{\oplus}\pi_l^{\oplus}$.

Considérons maintenant la matrice $3N \times 3N$ qui contient comme blocs les matrices $(\Delta_{kl}^{\oplus,\otimes})_{k,l\in U}$ pour $\oplus, \otimes \in \{1*, 12, 2*\}$. On note cette matrice $\Delta^{\mathcal{B}}$:

$$\Delta^{\mathcal{B}} = \begin{pmatrix} (\Delta_{kl}^{1*})_{k,l\in U} & (\Delta_{kl}^{1*,2*})_{k,l\in U} & (\Delta_{kl}^{1*,12})_{k,l\in U} \\ (\Delta_{kl}^{2*,1*})_{k,l\in U} & (\Delta_{kl}^{2*})_{k,l\in U} & (\Delta_{kl}^{2*,1*})_{k,l\in U} \\ (\Delta_{kl}^{12,1*})_{k,l\in U} & (\Delta_{kl}^{12,2*})_{k,l\in U} & (\Delta_{kl}^{12})_{k,l\in U} \end{pmatrix}. \tag{2.10}$$

La matrice $\Delta^{\mathcal{B}}$ est la matrice de variance covariace de la variable $\boldsymbol{\varepsilon}^{\mathcal{B}} = (\boldsymbol{\varepsilon}_k^{\mathcal{B}})_{k\in U}$ avec $\boldsymbol{\varepsilon}_k^{\mathcal{B}}$ donné par (2.8) :

$$\Delta^{\mathcal{B}} = \mathrm{Var}(\boldsymbol{\varepsilon}^{\mathcal{B}})$$

et elle contient six blocks différents, les trois sur la diagonale principale et les trois dessous.

Si on utilise une nouvelle base $\mathcal{B}'$, on peut calculer la matrice de variance-covariance par rapport à cette nouvelle base, notée $\Delta^{\mathcal{B}'}$, sachant $\Delta^{\mathcal{B}}$.

Propriété 4 *: Si $\mathcal{B}' = \mathcal{A}\mathcal{B}$ est une autre base alors*

$$\Delta^{\mathcal{B}'} = \mathcal{A}\Delta^{\mathcal{B}}\mathcal{A}'.$$

2.2.4 Exemples de plans bidimensionnels

Nous allons donner quelques exemples de plans bidimensionnels qui sont l'équivalent, en dimension deux, des plans usuels comme par exemple le sondage aléatoire simple sans remise. Les ensembles des probabilités d'inclusion de premier et deuxième degré sont également explicités.
Nous commençons par le cas le plus simple, le tirage bidimensionnel de deux échantillons indépendants.

Exemple 1 *: Le cas de deux échantillons indépendants*

Dans ce cas $p(s_1, s_2) = p_1(s_1)p_2(s_2)$ où p_1 et p_2 sont les plans marginaux de s_1 et respectivement s_2. Pour cette situation, la base $\mathcal{B}' = \{\varepsilon_k^1, \varepsilon_k^2, \varepsilon_k^{12}\}$ facilitera le calcul des probabilités d'inclusion. On donne dans la suite les probabilités d'inclusion de premier et deuxième degré par rapport à $\mathcal{B}'$. Puisque on a l'indépendance de s_1 et s_2,

$$\left\{ \begin{array}{rcl} \pi_k^{12} & = & P(k \in s_{12} = s_1 \cap s_2) = \pi_k^1 \pi_k^2 \\ \pi_{kl}^{1,2} & = & P(k \in s_1 \& l \in s_2) = \pi_k^1 \pi_l^2 \end{array} \right.$$

et alors

$$\left\{ \begin{array}{rcl} \pi_k^{1*} & = & \pi_k^1(1-\pi_k^2), \quad \pi_k^{2*} = \pi_k^2(1-\pi_k^1) \quad \text{pour tous} \quad k \in U \\ \pi_{kl}^{1,12} & = & \pi_{kl}^1 \pi_l^2, \quad \pi_{kl}^{2,12} = \pi_{kl}^2 \pi_l^1 \quad \text{pour tous} \quad k, l \in U. \end{array} \right.$$

On peut calculer les Δ-quantités et la matrice de variance par rapport à la base considérée a la forme suivante :

$$\Delta^{\mathcal{B}'} = \begin{pmatrix} (\Delta_{kl}^1)_{k,l\in U} & 0 & (\Delta_{kl}^1 \pi_l^2)_{k,l\in U} \\ 0 & (\Delta_{kl}^2)_{k,l\in U} & (\Delta_{kl}^2 \pi_l^1)_{k,l\in U} \\ (\Delta_{kl}^1 \pi_l^2)_{k,l\in U} & (\Delta_{kl}^2 \pi_l^1)_{k,l\in U} & (\Delta_{kl}^{12})_{k,l\in U} \end{pmatrix}.$$

Exemple 2 *: Le sondage aléatoire simple sans remise bidimensionnel*

Pour un échantillon bidimensionnel $\boldsymbol{s} = (s_1, s_2)$ sélectionné dans $\mathcal{U} \times \mathcal{U}$ tel que les plan marginaux, s_1 et s_2, sont des échantillons aléatoires simples sans remise dans $\mathcal{U}$, on ne peut rien dire sur l'intersection $s_{12} = s_1 \cap s_2$ qui peut de plus avoir une taille n_{12} variable. On peut coordonner deux échantillons aléatoires simples sans remise dans $\mathcal{U}$ en se fixant la taille de leur intersection comme considéré par Cotton & Hesse (1992) dans le cas T-dimensionnel.
Soit n_{1*}, n_{12} et n_{2*} les tailles de s_{1*}, s_{12} et respectivement la taille de s_{2*}.

Définition 2.2.8 *: On appelle un sondage aléatoire simple sans remise bi-dimensionnel (SAS2) paramétré par n_{1*}, n_{12} et n_{2*} un tirage joint $p(\boldsymbol{s})$ qui charge de façon équiprobable tous les couples $\boldsymbol{s} = (s_1, s_2)$ dont les échantillons s_{1*}, s_{12} et s_{2*} ont les tailles n_{1*}, n_{12} et n_{2*}.*

Cela implique

$$p(\boldsymbol{s}) = \frac{n_{1*}!n_{12}!n_{2*}!(N-(n_{1*}+n_{12}+n_{2*}))!}{N!}. \tag{2.11}$$

Remarque 2.2.1 *: Les plans marginaux et conditionnels sont des plans aléatoires simples sans remise dans $\mathcal{U}$ (Cotton & Hesse 1992). De plus, s_{1*}, s_{12} et s_{2*} sont aussi des SAS dans $\mathcal{U}$ de taille n_{1*}, n_{12} et n_{2*}.*

Suite à cette remarque, les probabilités d'inclusion de premier degré valent :

$$\pi_k^{\oplus} = \frac{n_{\oplus}}{N} = f_{\oplus} \quad \text{pour} \quad \oplus \in \{1, 2, 1*, 12, 2*\}. \tag{2.12}$$

Calculons dans la suite les probabilités d'inclusion de deuxième degré bidimensionnelles $\pi_{kl}^{\oplus,\otimes}$ pour tous $k, l \in U$ et $\oplus, \otimes \in \{1*, 12, 2*\}$.

i. Pour $\oplus = \otimes \in \{1*, 12, 2*\}$ la probabilité d'inclusion d'un couple k, l dans un des échantillons s_{1*}, s_{12} et s_{2*} est obtenue facilement, ce sont des plans de sondages aléatoires simples dans $\mathcal{U}$. Alors,

$$\begin{array}{rcl} \pi_{kl}^{\oplus} & = & \dfrac{n_{\oplus}(n_{\oplus}-1)}{N(N-1)} \quad \text{ce qui nous donne} \\ \Delta_{kl}^{\oplus} & = & \left\{ \begin{array}{cc} -f_{\oplus}\frac{1-f_{\oplus}}{N-1} & \text{pour} \quad k \neq l \\ f_{\oplus}(1-f_{\oplus}) & \text{pour} \quad k = l \end{array} \right. \end{array} \tag{2.13}$$

Alors, pour un certain $\oplus \in \{1*, 12, 2*\}$, la matrice $(\Delta_{kl}^{\oplus})_{k,l\in U}$ a l'expression suivante

$$\begin{array}{rcl} (\Delta_{kl}^{\oplus})_{k,l\in U} & = & f_{\oplus}(1-f_{\oplus})\dfrac{N}{N-1}(\boldsymbol{I}_N - \boldsymbol{P}) \\ & = & k_{\oplus}(\boldsymbol{I}_N - \boldsymbol{P}) \end{array} \tag{2.14}$$

où $\boldsymbol{I}_N$ est la matrice unité d'ordre N et $\boldsymbol{P} = \frac{1}{N}\mathbf{1_N}\mathbf{1'_N}$ et $k_{\oplus} = f_{\oplus}(1-f_{\oplus})\frac{N}{N-1}$.

ii. On considère le cas $\oplus \neq \otimes \in \{1*, 12, 2*\}$.
   1. D'abord, pour $k = l$, on a $\pi_{kk}^{\oplus,\otimes} = 0$.
   2. Pour calculer $\pi_{kl}^{\oplus,\otimes}$ pour $k \neq l$, on utilise la propriété

$$\pi_{kl}^{\oplus,\otimes} = E(\varepsilon_k^{\oplus}\varepsilon_l^{\otimes}) = P(k \in s_{\oplus} \& l \in s_{\otimes}).$$

   Puisque nous sommes dans le cas d'un sondage SAS2, c'est à dire $p(\boldsymbol{s})$ est constant, tous les $E(\varepsilon_k^{\oplus}\varepsilon_l^{\otimes})$ sont égaux. On a de plus

$$\sum_{k\in U}\varepsilon_k^{\oplus}\sum_{l\in U}\varepsilon_l^{\otimes} = n_{\oplus}n_{\otimes}$$

et l'espérance de cette expression vaut

$$\sum_{k\in U}\sum_{l\in U} E\left(\varepsilon_k^{\oplus}\varepsilon_l^{\otimes}\right) = \sum_{k\neq l\in U} E\left(\varepsilon_k^{\oplus}\varepsilon_l^{\otimes}\right) = n_{\oplus}n_{\otimes}$$

ce qui implique

$$\begin{aligned} \pi_{kl}^{\oplus,\otimes} &= \frac{n_{\oplus}n_{\otimes}}{N(N-1)} \\ \Delta_{kl}^{\oplus,\otimes} &= \begin{cases} \frac{f_{\oplus}f_{\otimes}}{N-1} & \text{pour} \quad k \neq l \\ -f_{\oplus}f_{\otimes} & \text{pour} \quad k = l \end{cases} \end{aligned} \tag{2.15}$$

Alors, pour $\oplus \neq \otimes \in \{1*, 12, 2*\}$ la matrice $(\Delta_{kl}^{\oplus,\otimes})_{k,l\in U}$ a l'expression suivante

$$\begin{aligned} (\Delta_{kl}^{\oplus,\otimes})_{k,l\in U} &= -f_{\oplus}f_{\otimes}\frac{N}{N-1}(\boldsymbol{I}_N - \boldsymbol{P}) \\ &= k_{\oplus,\otimes}(\boldsymbol{I}_N - \boldsymbol{P}) \end{aligned} \tag{2.16}$$

où $k_{\oplus,\otimes} = -f_{\oplus}f_{\otimes}\frac{N}{N-1}$.

Exemple 3 *: Le sondage de Poisson bidimensionnel*

On considère un échantillon bidimensionnel $\boldsymbol{s} = (s_1, s_2)$ dans $\mathcal{U} \times \mathcal{U}$. Pour chaque individu k dans la population, on a la trace de $\boldsymbol{s}$ sur k, $\mathbf{s_k} = \text{tr}_{\,k}(\boldsymbol{s}) = (s_1 \cup \{k\}, s_2 \cup \{k\}) \in \{\emptyset, \{k\}\} \times \{\emptyset, \{k\}\}$.
Soit $p_k(\mathbf{s_k})$ une probabilité discrète sur $\{\emptyset, \{k\}\}^2$ pour tous $k \in U$, tel que

$$\begin{cases} p_k(\mathbf{s_k}) \geq 0 \text{ pour tous k} \\ \displaystyle\sum_{\mathbf{s_k}} p_k(\mathbf{s_k}) = 1. \end{cases} \tag{2.17}$$

Cotton & Hesse (1992) donnent la définition suivante :

Définition 2.2.9 *: Un sondage de Poisson bidimensionnel (PO2) est défini par*

$$p(\boldsymbol{s}) = \prod_{k\in U} p_k(\mathbf{s_k}).$$

Cotton & Hesse (1992) montrent que $p_k(\mathbf{s_k})$ sont exactement les probabilités d'inclusion de premier degré bidimensionnelles ; plus précisément, on a

$$p_k(\mathbf{s_k}) = \pi_k^{\oplus} \text{ pour } k \in s_{\oplus} \text{ et } \oplus \in \{1*, 12, 2*, **\}$$

où $\pi_k^{**} = 1 - \left(\pi_k^{1*} + \pi_k^{12} + \pi_k^{2*}\right)$. Par conséquent, un sondage PO2 peut se définir de manière équivalente

Définition 2.2.10 *: Un sondage de Poisson bidimensionnel (PO2) est défini par*

$$p(\boldsymbol{s}) = \prod_{k\in s_{1*}} \pi_k^{1*} \prod_{k\in s_{12}} \pi_k^{12} \prod_{k\in s_{2*}} \pi_k^{2*} \prod_{k\in s_{**}} \pi_k^{**}.$$

Considérons les variables $\boldsymbol{\varepsilon}_k$ données par (2.6).

Définition 2.2.11 *: Alors, un échantillon PO2 est obtenu pour des $\boldsymbol{\varepsilon}_k$ indépendants et*

$$P(\varepsilon_k^{1*} = 1) = \pi_k^{1*}, \quad P(\varepsilon_k^{12} = 1) = \pi_k^{12}, \quad P(\varepsilon_k^{2*} = 1) = \pi_k^{2*}.$$

Cotton & Hesse (1992) montrent que les tirages marginaux et conditionnels sont des sondages de Poisson unidimensionnels.
Calculons les probabilités du second degré bidimensionnelles. Puisque les éléments k et l sont sélectionnés indépendemment l'un de l'autre,

i. les quantités de type $\pi_{kl}^{\oplus}$ pour $\oplus \in \{1*, 12, 2*\}$ deviennent

$$\pi_{kl}^{\oplus} = \begin{cases} \pi_k^{\oplus}\pi_l^{\oplus} \text{ pour } k \neq l \\ \pi_k^{\oplus} \text{ pour } k = l \end{cases} \tag{2.18}$$

ii. les quantités de type $\pi_{kl}^{\oplus,\otimes}$ pour $\oplus \neq \otimes \in \{1*, 12, 2*\}$ deviennent

$$\pi_{kl}^{\oplus,\otimes} = \begin{cases} \pi_k^{\oplus}\pi_l^{\otimes} \text{ pour } k \neq l \\ 0 \text{ pour } k = l. \end{cases} \tag{2.19}$$

Par conséquent, les matrices de variance-covariance ont pour expression :

$$\begin{array}{l} \left(\Delta_{kl}^{\oplus}\right)_{k,l\in U} = \operatorname{diag}\left(\pi_k^{\oplus}(1-\pi_k^{\oplus})\right)_{k\in U} \quad \text{pour } \oplus \in \{1*, 12, 2*\} \\ \left(\Delta_{kl}^{\oplus,\otimes}\right)_{k,l\in U} = \operatorname{diag}\left(-\pi_k^{\oplus}\pi_k^{\otimes}\right)_{k\in U} \quad \text{pour } \oplus \neq \otimes \in \{1*, 12, 2*\}. \end{array} \tag{2.20}$$

Comme dans le cas unidimensionnel, le sondage PO2 est de taille $n_{\boldsymbol{s}} = (n_{1*}, n_{2*}, n_{12})$ variable.

Exemple 4 *: **Le sondage de Bernoulli bidimensionnel***

Considèrons le sondage PO2 dans la situation particulière où chaque individu est tiré indépendamment et avec la même probabilité dans un des quatre échantillons $s_{1*}, s_{12}, s_{2*}, s_{**}$. Nous obtenons ce qu'on va appeler un sondage de Bernoulli bidimensionnel (BE2) $p(\boldsymbol{s} = (s_1, s_2))$ de paramètres $\pi_{1*}, \pi_{12}, \pi_{2*}$. Plus précisément le plan est donné par la définition (2.2.10) pour des probabilités d'inclusion constantes.

Définition 2.2.12 *: Un sondage de Bernoulli bidimensionnel (BE2) est défini par*

$$p(\boldsymbol{s} = (s_1, s_2)) = \pi_{1*}^{n_{1*}} \pi_{12}^{n_{12}} \pi_{2*}^{n_{2*}} \pi_{**}^{N-(n_{1*}+n_{12}+n_{2*})} \tag{2.21}$$

avec $\pi_{1*}, \pi_{12}, \pi_{2*} \in (0,1)$ *et* $\pi_{**} = 1 - (\pi_{1*} + \pi_{12} + \pi_{2*})$.

Les probabilités d'inclusion de second degré bidimensionnelles s'obtiennent à partir de (2.18) et (2.19) pour $\pi_k^{\oplus} = \pi_{\oplus}$ pour tous $k \in U$ et $\oplus \in \{1*, 12, 2*\}$. Alors, les matrices de variance-covariance ont les expressions :

$$\left(\Delta_{kl}^{\oplus}\right)_{k,l\in U} = \operatorname{diag}\left(\pi_{\oplus}(1-\pi_{\oplus})\right) \qquad \text{pour } \oplus \in \{1*, 12, 2*\}$$

$$\left(\Delta_{kl}^{\oplus,\otimes}\right)_{k,l\in U} = \operatorname{diag}\left(-\pi_{\oplus}\pi_{\otimes}\right) \qquad \text{pour } \oplus \neq \otimes \in \{1*, 12, 2*\}$$

La taille $n_{\boldsymbol{s}}$ est une variable aléatoire et la probabilité que l'échantillon $\boldsymbol{s}$ ait exactement la taille $n_{\boldsymbol{s}} = (n_{1*}, n_{12}, n_{2*})$ est

$$\frac{N!}{n_{1*}!n_{12}!n_{2*}!(N-(n_{1*}+n_{12}+n_{2*}))!} \pi_{1*}^{n_{1*}} \pi_{12}^{n_{12}} \pi_{2*}^{n_{2*}} \pi_{**}^{N-(n_{1*}+n_{12}+n_{2*})}. \tag{2.22}$$

Exemple 5 ***: Le sondage de Bernoulli bidimensionnel et conditionnel à la taille***

Comme le plan PO2, le sondage de Bernoulli bidimensionnel a l'inconvénient d'une taille variable. Les relations (4) et (2.22) montrent que conditionnellement à $n_{\boldsymbol{s}} = (n_{1*}, n_{2*}, n_{12})$, le plan $p(\boldsymbol{s} = (s_1, s_2)|n_{1*}, n_{12}, n_{2*})$ est constant et d'expression donnée par (2.11), c'est à dire celle d'un sondage aléatoire simple sans remise et bidimensionnel SAS2. C'est pourquoi, une fois que l'échantillon est sélectionné, on travaille conditionnellement à $n_{\boldsymbol{s}}$.

2.3 Estimation sans biais linéaire

Après avoir défini un plan de sondage bidimensionnel et les probabilités du premier et du second degré bidimensionnelles, pour avoir une théorie de type Horvitz-Thompson sur deux échantillons il est normal de traiter l'estimation sans biais linéaire. De nouveau, la situation est plus compliquée que dans le cas unidimensionnel. Il faut se poser d'abord la question : que cherche-t-on à estimer ?

Dans le cas d'un seul échantillon, l'estimateur de Horvitz-Thompson pour le total d'une variable est le seul estimateur linéaire sans biais avec des poids qui ne dépendent pas de l'échantillon ni de la variable d'intérêt. Ces poids sont les probabilités d'inclusion dans l'échantillon de premier degré. On peut également considérer plusieurs variables mesurées sur le même échantillon. Alors, quelque soit la combinaison linéaire de leur totaux le meilleur estimateur sans biais est l'estimateur de Horvitz-Thompson.

Une manière logique de généraliser l'estimation sans biais linéaire à plusieurs échantillons est de considérer d'abord une seule variable pour étendre en suite à plusieurs variables. Pour faciliter les calculs, on considère seulement le cas de deux échantillons.

Il s'agit pour nous de trouver dans chaque cas l'estimateur optimal d'un total.

2.3.1 Forme des estimateurs sans biais

Cas d'une seul variable d'intérêt mesurée sur deux échantillons disjoints

On a vu que pour deux échantillons s_1 et s_2,que la dimension de l'algèbre engendrée par $(\varepsilon_k^1, \varepsilon_k^2)$ est de dimension trois dont une base est $\{\varepsilon_k^{1*}, \varepsilon_k^{12}, \varepsilon_k^2\}$. Ce sont les indicatrices associées aux échantillons disjoints. On va donc considérer dans la suite le cas de deux échantillons disjoints. Le cas de deux échantillons non-disjoints qui est équivalent à celui de trois échantillons disjoints se généralise directement.

Soit $\mathcal{X}$ une variable d'intérêt mesurée sur s_{1*} et s_{12} dont nous voulons estimer le total $t_x = \sum_U x_k$. On cherche à estimer t_x par un estimateur pondéré de la forme $\sum_U x_k w_k^{s_1}$ avec les poids $w_k^{s_1}$ qui dépendent de l'échantillon $s_1 = s_{1*} \cup s_{12}$ et qui vont être déterminés dans la suite.

On attribue aux points k qui se trouvent dans s_{1*} un poids w_k^{1*}, pour ceux dans s_{12} un poids w_k^{12} et zéro pour $k \in U - (s_{1*} \cup s_{12})$. Alors, un

estimateur pour t_x est

$$\hat{t}_x = \sum_{k\in U} x_k w_k^{s_1} = \sum_{k\in U} x_k (w_k^{1*}\varepsilon_k^{1*} + w_k^{12}\varepsilon_k^{12}). \tag{2.23}$$

où $w_k^{s_1} = w_k^{1*}\varepsilon_k^{1*} + w_k^{12}\varepsilon_k^{12}$ et $s_1 = s_{1*} \cup s_{12}$.

Résultat 2.3.1 *: L'estimateur $\hat{t}_x$ est sans biais pour t_x si les poids w_k^{1*} et w_k^{12} vérifient*

$$w_k^{1*}\pi_k^{1*} + w_k^{12}\pi_k^{12} = 1$$

Résultat 2.3.2 *: L'estimateur $\hat{t}_x$ optimal est obtenu pour $w_k^{1*}\varepsilon_k^{1*} + w_k^{12}\varepsilon_k^{12} = \frac{1}{\pi_k^1}$ où $\pi_k^1 = Pr(k \in s_1)$, $s_1 = s_{1*} \cup s_{12}$. Autrement dit, l'estimateur optimal est l'estimateur d'H-T considéré sur la réunion de deux échantillons.*

Preuve : On a

$$Var\left(\sum_U x_k w_k^{s_1}\right) = Var \sum_U x_k E(w_k^{s_1}|k \in s_1) + E\sum_U x_k Var(w_k^{s_1}|k \in s_1)$$

qui est minime pour $E(w_k^{s_1}|k \in s_1) = \dfrac{1}{\pi_k^1}$. □

On étend dans la suite ce résultat au cas de deux variables d'intérêt mesurées sur deux échantillons différents avec une intersection que l'on suppose non négligeable.

Cas de deux variables d'intérêt mesurées sur deux échantillons

Soient $\mathcal{X}$ et $\mathcal{Y}$ deux variables d'intérêt avec $\mathcal{X}$ mesurée sur un échantillon s_1 et $\mathcal{Y}$ sur s_2. Notre intention est d'estimer une combinaison linéaire de leurs totaux

$$Z = \phi t_x + \psi t_y$$

par un estimateur pondéré $\sum_U (x_k w_k^{s_1} + y_k w_k^{s_2})$ avec des poids à déterminer. Comme $\mathcal{X}$ est observée sur $s_1 = s_{1*} \cup s_{12}$ on obtiendra alors un estimateur sans biais de ϕt_x si

$$w_k^{1*}\pi_k^{1*} + w_k^{12,x}\pi_k^{12} = \phi \tag{2.24}$$

De même pour $\mathcal{Y}$ qui est estimé sur $s_2 = s_{2*} \cup s_{12}$:

$$w_k^{2*}\pi_k^{2*} + w_k^{12,y}\pi_k^{12} = \psi \tag{2.25}$$

Inversement, pour tout système de variables w_k^{1*}, $w_k^{12,x}$, w_k^{2*} $w_k^{12,y}$ vérifiant (2.24) et (2.25) l'estimateur

$$\hat{Z} = \sum_{s_{1*}} x_k w_k^{1*} + \sum_{s_{12}} x_k w_k^{12,x} + \sum_{s_{2*}} y_k w_k^{2*} + \sum_{s_{12}} y_k w_k^{12,y}$$

sera sans biais pour Z.
C'est un estimateur sans biais dont les poids ne dépendent que des indicatrices de l'échantillon sur lequel est mesurée chaque variable. Ils ne dépendent donc que de deux ensembles de valeurs arbitraires qui vérifient ces deux équations.

On propose une reparamétrisation des $w_k \in \{w_k^{1*}, w_k^{12,x}, w_k^{2*}, w_k^{12,y}\}$ qui respecte une certaine symétrie tout en vérifiant la condition de "sans biais" pour $\hat{Z}$. On prend des poids comme suit

$$\left\{ \begin{array}{lll} w_k^{1*} = \dfrac{a_k}{\pi_k^{1*}} & ; & w_k^{12x} = \dfrac{\phi - a_k}{\pi_k^{12}} \\ w_k^{2*} = \dfrac{b_k}{\pi_k^{2*}} & ; & w_k^{12y} = \dfrac{\psi - b_k}{\pi_k^{12}}, \end{array} \right. \tag{2.26}$$

et cela conduit a l'expression suivante pour $\hat{Z}$:

$$\begin{aligned} \hat{Z} &= \sum_{s_{1*}} x_k \frac{a_k}{\pi_k^{1*}} + \sum_{s_{12}} x_k \frac{\phi - a_k}{\pi_k^{12}} + \sum_{s_{2*}} y_k \frac{b_k}{\pi_k^{2*}} + \sum_{s_{12}} y_k \frac{\psi - b_k}{\pi_k^{12}} \\ &= \sum_U x_k a_k \left(\frac{\varepsilon_k^{1*}}{\pi_k^{1*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}} \right) + \sum_U y_k b_k \left(\frac{\varepsilon_k^{2*}}{\pi_k^{2*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}} \right) + \sum_U (\phi x_k + \psi y_k) \frac{\varepsilon_k^{12}}{\pi_k^{12}} \end{aligned}$$

ce qui peut également s'écrire sous forme matricielle

$$\hat{Z} = \boldsymbol{\theta}' \operatorname{diag} \begin{pmatrix} x \\ \dots \\ y \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} + \hat{t}_{H-T} \tag{2.27}$$

où $\boldsymbol{\theta} = ((a_k)_{k=1}^N, (b_k)_{k=1}^N)' \in \mathbb{R}^{2N}$ est le vecteur de paramètres,

$$\left\{ \begin{array}{l} \boldsymbol{\alpha} = (\alpha_k)_{k=1}^N; \quad \alpha_k = \dfrac{\varepsilon_k^{1*}}{\pi_k^{1*}} - \dfrac{\varepsilon_k^{12}}{\pi_k^{12}} \\ \boldsymbol{\beta} = (\beta_k)_{k=1}^N; \quad \beta_k = \dfrac{\varepsilon_k^{2*}}{\pi_k^{2*}} - \dfrac{\varepsilon_k^{12}}{\pi_k^{12}} \\ \hat{t}_{H-T} = \displaystyle\sum_U \frac{\phi x_k + \psi y_k}{\pi_k^{12}} \end{array} \right.$$

et on a noté

$$\operatorname{diag} \begin{pmatrix} x \\ \dots \\ y \end{pmatrix} = \begin{pmatrix} \operatorname{diag} x & 0 \\ 0 & \operatorname{diag} y \end{pmatrix}$$

où $\operatorname{diag} x$ (resp $\operatorname{diag} y$) est la matrice $N \times N$ avec x_k (resp y_k) sur la diagonale.

Alors, $\hat{Z}$ donné par (2.27) s'écrit comme une somme de trois estimateurs linéaires : l'estimateur d'Hortwitz-Thompson sur s_{12} et qui ne dépend pas des quantités arbitraires a_k et b_k pour tous $k \in U$ et deux estimateurs linéaires en x_k, respectivement en y_k et d'espérances nulles ($E((\boldsymbol{\alpha}, \boldsymbol{\beta})') = 0$).

Remarque 2.3.1 *On peut exprimer le vecteur* $(\boldsymbol{\alpha}, \boldsymbol{\beta})'$ *en fonction de* $\boldsymbol{\varepsilon}^{\mathcal{B}} = (\boldsymbol{\varepsilon}^{1*}, \boldsymbol{\varepsilon}^{2*}, \boldsymbol{\varepsilon}^{12})'$ *comme suit*

$$\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} I_N & 0_N & -I_N \\ 0_N & I_N & -I_N \end{pmatrix} \begin{pmatrix} diag\left(\frac{1}{\pi_k^{1*}}\right)_{k\in U} \\ \cdots \\ diag\left(\frac{1}{\pi_k^{2*}}\right)_{k\in U} \\ \cdots \\ diag\left(\frac{1}{\pi_k^{12}}\right)_{k\in U} \end{pmatrix} \begin{pmatrix} \boldsymbol{\varepsilon}^{1*} \\ \boldsymbol{\varepsilon}^{2*} \\ \boldsymbol{\varepsilon}^{12} \end{pmatrix} \tag{2.28}$$

où I_N*(resp* 0_N*) est la matrice unité (respectivement la matrice nulle) de dimension* N*. Comme on connait l'expression de la variance de* $\boldsymbol{\varepsilon}^{\mathcal{B}}$ *pour les plans usuels tels que PO2 et SAS2, cette relation nous sera utile pour le calcul de la valeur optimale de* $\boldsymbol{\theta}$*.*

On peut obtenir également une formule de type H-T pour la variance de $\hat{Z}$ qui permettra d'en déduire un estimateur pour la variance.

Résultat 2.3.3 *: L'estimateur* $\hat{Z}$ *peut être écrit de la façon suivante :*

$$\hat{Z} = (vect)' \begin{pmatrix} \boldsymbol{\varepsilon}^{1*} \\ \boldsymbol{\varepsilon}^{2*} \\ \boldsymbol{\varepsilon}^{12} \end{pmatrix}$$

avec $(vect)' = \left(\left(\frac{a_k x_k}{\pi_k^{1*}}\right)_{k=1}^{N}, \left(\frac{b_k y_k}{\pi_k^{2*}}\right)_{k=1}^{N}, \left(\frac{x_k(\phi - a_k) + y_k(\psi - b_k)}{\pi_k^{12}}\right)_{k=1}^{N} \right) \in \mathbb{R}^{3N}$ *et alors sa variance a l'expression*

$$V(\hat{Z}) = (vect)' \begin{pmatrix} \Delta_{kl}^{1*,1*} & \Delta_{kl}^{1*,2*} & \Delta_{kl}^{1*,12} \\ \Delta_{kl}^{2*,1*} & \Delta_{kl}^{2*,2*} & \Delta_{kl}^{2*,12} \\ \Delta_{kl}^{12,1*} & \Delta_{kl}^{12,2*} & \Delta_{kl}^{12,12} \end{pmatrix} (vect) \tag{2.29}$$

où chaque $\Delta_{kl}^{\oplus,\otimes}$ *est une matrice de taille* $N \times N$ *pour* $\oplus, \otimes \in \{1*, 12, 2*\}$*.*

2.3.2 Optimisation

Il existe un grand choix d'estimateurs sans biais vérifiant la relation (2.27). On va choisir celui dont la variance est minimale. On a le résultat suivant comme conséquence de (2.27) :

Résultat 2.3.4 *: La variance de $\hat{Z}$ a l'expression*

$$V(\hat{Z}) = \boldsymbol{\theta}'\boldsymbol{\Gamma}\boldsymbol{\theta} + 2\boldsymbol{\theta}'\boldsymbol{\gamma} + c \tag{2.30}$$

où $\boldsymbol{\Gamma}$, $\boldsymbol{\gamma}$ sont des matrices de dimensions $2N * 2N$, $2N * 1$ données par

$$\boldsymbol{\Gamma} \quad = \quad diag\begin{pmatrix} x \\ \dots \\ y \end{pmatrix} \; Var\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} \; diag\begin{pmatrix} x \\ \dots \\ y \end{pmatrix} \tag{2.31}$$

$$\boldsymbol{\gamma} \quad = \quad diag\begin{pmatrix} x \\ \dots \\ y \end{pmatrix} \; Cov\left(\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}, \frac{\boldsymbol{\varepsilon^{12}}}{\boldsymbol{\pi^{12}}}\right)(\phi\boldsymbol{x} + \psi\boldsymbol{y}) \tag{2.32}$$

et $c = \; Var\,(\hat{t}_{H-T}) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl}^{12} \frac{\phi x_k + \psi y_k}{\pi_k^{12}} \frac{\phi x_l + \psi y_l}{\pi_l^{12}}$ *;* $\boldsymbol{x}' = (x_1, \dots, x_N)$, $\boldsymbol{y}' = (y_1, \dots, y_N)$ *et* $\dfrac{\boldsymbol{\varepsilon^{12}}}{\boldsymbol{\pi^{12}}} = \left(\dfrac{\varepsilon_k^{12}}{\pi_k^{12}}\right)_{k=1}^{N}$.

La variance est une forme quadratique en $\boldsymbol{\theta}$ avec $c \in \mathbb{R}$, $\boldsymbol{\gamma} \in \mathbb{R}^{2N}$ et $\boldsymbol{\Gamma}$ une matrice symétrique non négative qui ne dépend pas de $\boldsymbol{\theta}$. On peut déduire la valeur de $\boldsymbol{\theta}$ qui minimise la variance de $\hat{Z}$.

Résultat 2.3.5 *: Si la matrice $\boldsymbol{\Gamma}$ est définie positive alors la variance de $\hat{Z}$ est minimale pour $\boldsymbol{\theta}_{opt} = -\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}$ avec la valeur optimale*

$$V_{opt}(\hat{Z}) = V(\hat{t}_{H-T}) - \boldsymbol{\gamma}`\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma} \tag{2.33}$$

On remarque que l'estimateur ainsi obtenu est toujours supérieur à l'estimateur $\hat{t}_{H-T}$

$$V_{opt}(\hat{Z}) \leq V(\hat{t}_{H-T})$$

Remarque 2.3.2 *: Le resultat obtenu a un intérêt théorique mais en pratique, si les dimensions de $\boldsymbol{\Gamma}$ et $\boldsymbol{\gamma}$ sont grandes, le calcul de $\boldsymbol{\theta}$ est presque impossible. De plus, ces matrices dépendent de toutes les valeurs inconnues y_k, x_k pour toutes les unités dans la population U.*

2.3.3 Estimation de la variance

L'expression (2.29) est l'analogue en deux dimensions de la variance d'un estimateur de Horvitz-Thompson. On a neuf blocs de variance-covariance $(\Delta_{kl}^{\oplus,\otimes})_{k,l\in U}$ au lieu d'un seul pour le cas unidimensionnel classique. On va

utiliser l'expression (2.29) dans le but d'obtenir une estimation de la variance de $\hat{Z}$ similaire au cas unidimensionnel, c'est à dire en fonction des quantités du type $\check{\Delta}_{kl} = \frac{\Delta_{kl}}{\pi_{kl}}$.

La situation est plus compliquée en raison de la présence des termes de covariance et du fait qu'on a plusieurs échantillons. Il existe trois types de termes de variance-covariance qui vont demander un traitement différent. Remarquons que chacun de ces termes est une somme quadratique sur U en $\mathcal{X}$, en $\mathcal{Y}$ ou en $\mathcal{X}$ et $\mathcal{Y}$. On peut alors utiliser la théorie de Horvitz-Thompson unidimensionnelle pour estimer ces quantités. Il reste à établir pour chacun d'entre eux dans quel échantillon nous allons procéder à l'estimation.

Résultat 2.3.6 *: Les produits croisés tels que* $\sum_{k\in U}\sum_{l\in U} \Delta_{kl}^{1*,2*} \frac{x_k}{\pi_k^{1*}} \frac{y_l}{\pi_l^{2*}}$, *peuvent être estimés seulement dans l'espace de l'intersection* s_{12}. *On obtient l'estimateur*

$$\sum_{k\in s_{12}}\sum_{l\in s_{12}} \frac{\Delta_{kl}^{\oplus,\otimes}}{\pi_{kl}^{12}} \frac{x_k}{\pi_k^{\oplus}} \frac{y_l}{\pi_l^{\otimes}}$$

pour $\oplus \in \{1*, 12\}$ *et* $\otimes \in \{12, 2*\}$.

Remarquons qu'on ne peut estimer les termes de ce type-la que sur l'intersection s_{12} car les variables $\mathcal{X}$ et $\mathcal{Y}$ sont observées simultanément uniquement sur s_{12}. Cette formule générale d'estimation de termes de covariance n'est possible que si l'espace intersection n'est pas vide. Dans le cas contraire, aucune conclusion ne peut être déduite.

Les termes restants sont de deux catégories : on a des sommes doubles en $\mathcal{X}$ et des sommes doubles en $\mathcal{Y}$. Concernant leur estimation, on a les deux suivants résultats

Résultat 2.3.7 *: Les sommes doubles en* $\mathcal{X}$ *comme* $\sum_{U}\sum_{U} \Delta_{kl}^{\oplus,\otimes} \frac{x_k}{\pi_k^{\oplus}} \frac{x_l}{\pi_l^{\otimes}}$ *pour* $\oplus, \otimes \in \{1*, 12\}$ *sont estimées dans l'échantillon* s_1 *par*

$$\sum_{k\in s_1}\sum_{l\in s_1} \frac{\Delta_{kl}^{\oplus,\otimes}}{\pi_{kl}^{1}} \frac{x_k}{\pi_k^{\oplus}} \frac{x_l}{\pi_l^{\otimes}}$$

Résultat 2.3.8 *: Les sommes doubles en* $\mathcal{Y}$ *comme* $\sum_{U}\sum_{U} \Delta_{kl}^{\oplus,\otimes} \frac{y_k}{\pi_k^{\oplus}} \frac{y_l}{\pi_l^{\otimes}}$ *pour* $\oplus, \otimes \in \{12, 2*\}$ *sont estimées dans* s_2 *par*

$$\sum_{k\in s_2}\sum_{l\in s_2} \frac{\Delta_{kl}^{\oplus,\otimes}}{\pi_{kl}^{2}} \frac{y_k}{\pi_k^{\oplus}} \frac{y_l}{\pi_l^{\otimes}}.$$

Remarquons que d'autres estimateurs seraient possibles. Par exemple, $\sum_U \sum_U \Delta_{kl}^{\oplus,\otimes} \frac{x_k}{\pi_k^{\oplus}} \frac{x_l}{\pi_l^{\otimes}}$ peut être estimé sur s_{1*} et également sur s_{12} mais comme nous l'avons vu au début de cette section, l'estimateur sur s_1 est optimal. C'est pour cette raison qu'on a opté pour cet estimateur.

Les quantités a_k, b_k, $\phi - a_k$ et $\psi - b_k$ ont été omises dans les expressions de terms de variance-covariance parce que elles n'interviennent pas dans l'estimation.

2.3.4 Réduction du nombre de paramètres

Comme la remarque (2.3.2) du paragraphe précédent le montre, on doit diminuer le nombre de paramètres pour pouvoir les calculer ou au moins les estimer à partir des échantillons. On prend alors $\boldsymbol{a}' = (a_1, \ldots, a_N) = \overline{\boldsymbol{a}}' \times \overline{H}$ et $\boldsymbol{b}' = (b_1, \ldots, b_N) = \overline{\boldsymbol{b}}' \times \overline{J}$ avec $\overline{H}$ (respectivement $\overline{J}$) une matrice connue de taille $H \times N$ (respectivement $J \times N$) et $\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}$ sont deux vecteurs de dimensions H et respectivement J à déterminer. On effectue alors une stratification des vecteurs $\boldsymbol{a}$ et $\boldsymbol{b}$, qui peut être justifiée notamment par l'utilisation de variables catégorielles (géographiques, sociales, ...).
On donne ci-dessous deux exemples particuliers de matrices $\overline{H}$ et $\overline{J}$.

A. Les vecteurs $\boldsymbol{a}$ et $\boldsymbol{b}$ peuvent être de la forme suivante :

$$\boldsymbol{a}' = (a_1, \ldots, a_N) \quad = \quad (\overline{a}_1, \ldots, \overline{a}_H) \times \overline{H} = \overline{\boldsymbol{a}}' \times \overline{H} \tag{2.34}$$

$$\boldsymbol{b}' = (b_1, \ldots, b_N) \quad = \quad (\overline{b}_1, \ldots, \overline{b}_J) \times \overline{J} = \overline{\boldsymbol{b}}' \times \overline{J} \tag{2.35}$$

où les matrices $\overline{H}$ (resp $\overline{J}$) sont de dimensions $H \times N$ (resp $J \times N$) et elles résultent des partitions de la population U réalisées en utilisant l'information auxiliaire comme suit :

$$U = \bigcup_{h=1}^{H} U_h, |U_h| = N_h, \sum_{h=1}^{H} N_h = N$$

$$U = \bigcup_{j=1}^{J} W_j, |W_j| = M_j, \sum_{j=1}^{J} M_j = N.$$

Plus précisément,

$$\overline{H} = \text{ diag } (\mathbf{1}'_{\mathbf{N_h}})_{h=1}^{H} \quad \text{avec} \quad \mathbf{1}'_{\mathbf{N_h}} = (1,\ldots,1) \in R^{N_h}$$

$$\overline{J} = \text{ diag } (\mathbf{1}'_{\mathbf{M_j}})_{j=1}^{J} \quad \text{avec} \quad \mathbf{1}'_{\mathbf{M_j}} = (1,\ldots,1) \in R^{M_j}.$$

Les vecteurs $\boldsymbol{a}, \boldsymbol{b}$ sont de la forme (2.34) et respectivement (2.35) si on prend

$$a_k = \overline{a}_h \quad \forall k \in U_h, \quad h = 1,\ldots,H,$$

$$b_l = \overline{b}_j \quad \forall l \in W_j \quad j = 1,\ldots,J.$$

Dans ce cas, le paramètre $\boldsymbol{\theta}$ s'écrit

$$\begin{aligned} \boldsymbol{\theta}' &= (a_1,\ldots,a_N,b_1,\ldots,b_N) = (\overline{\boldsymbol{a}}' \times \overline{H}, \overline{\boldsymbol{b}}' \times \overline{J}) \\ &= (\overline{\boldsymbol{a}}', \overline{\boldsymbol{b}}') \begin{pmatrix} \overline{H} & 0 \\ 0 & \overline{J} \end{pmatrix}. \end{aligned} \tag{2.36}$$

Alors notre problème d'optimisation de dimension $2N$ est réduit à la dimension $H + J$ puisque on se ramène à trouver

$$\overline{\boldsymbol{\theta}}' = (\overline{\boldsymbol{a}}', \overline{\boldsymbol{b}}') = (a_1,\ldots,a_H,b_1,\ldots,b_J)$$

qui minimise la variance de $\hat{Z}$. On donne le résultat suivant concernant l'expression de $\hat{Z}$

Résultat 2.3.9 *: Pour une stratification de la population U, l'estimateur $\hat{Z}$ devient*

$$\begin{aligned} \hat{Z} &= \overline{\boldsymbol{\theta}}' \begin{pmatrix} \overline{H} & 0 \\ 0 & \overline{J} \end{pmatrix} \begin{pmatrix} diagx & 0 \\ 0 & diagy \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} + \hat{t}_{H-T} \\ &= \overline{\boldsymbol{\theta}}' \begin{pmatrix} \overline{H}\ diagx & 0 \\ 0 & \overline{J}\ diagy \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} + \hat{t}_{H-T} \\ &= \overline{\boldsymbol{\theta}}' \begin{pmatrix} (\boldsymbol{x}'_{\boldsymbol{N_h}} \boldsymbol{\alpha}_{\boldsymbol{N_h}})_{h=1,H} \\ (\boldsymbol{y}'_{\boldsymbol{M_j}} \boldsymbol{\beta}_{\boldsymbol{M_j}})_{j=1,J} \end{pmatrix} + \hat{t}_{H-T} \end{aligned} \tag{2.37}$$

où pour un certain h

$$\begin{cases} \boldsymbol{x}_{\boldsymbol{N_h}} &= (x_k)_{k \in U_h} \in \mathbb{R}^{N_h}, \quad \boldsymbol{\alpha}_{N_h} = (\alpha_k)_{k \in U_h} \\ \boldsymbol{x}'_{\boldsymbol{N_h}} \boldsymbol{\alpha}_{\boldsymbol{N_h}} &= \displaystyle\sum_{k \in U_h} x_k \left(\frac{\varepsilon_k^{1*}}{\pi_k^{1*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}} \right). \end{cases} \tag{2.38}$$

On définit de manière analogue $\boldsymbol{y}'_{\boldsymbol{M_j}}$ et $\boldsymbol{y}'_{\boldsymbol{M_j}} \boldsymbol{\beta}_{\boldsymbol{M_j}}$.

Alors, les valeurs optimales de $\overline{\boldsymbol{\theta}}$ aussi bien que celle de sa variance peuvent être obtenues facilement à partir de (2.38). On a le résultat suivant.

Résultat 2.3.10 *: Le paramètre optimal $\overline{\boldsymbol{\theta}}_{opt}$ a l'expression suivante*

$$\overline{\boldsymbol{\theta}}_{opt} = -\overline{\boldsymbol{\Gamma}}^{-1}\overline{\boldsymbol{\gamma}} \quad avec \tag{2.39}$$

$$\overline{\boldsymbol{\Gamma}} = \begin{pmatrix} \overline{H} & 0 \\ 0 & \overline{J} \end{pmatrix} \boldsymbol{\Gamma} \begin{pmatrix} \overline{H}' & 0 \\ 0 & \overline{J}' \end{pmatrix} \tag{2.40}$$

$$\overline{\boldsymbol{\gamma}} = \begin{pmatrix} \overline{H} & 0 \\ 0 & \overline{J} \end{pmatrix} \boldsymbol{\gamma} \tag{2.41}$$

avec $\boldsymbol{\Gamma}$ et $\boldsymbol{\gamma}$ donnés par (2.31) et (2.32).

Il résulte alors que

$$\overline{\boldsymbol{\Gamma}} = \mathrm{Var} \begin{pmatrix} (\boldsymbol{x'}_{\boldsymbol{N_h}} \boldsymbol{\alpha}_{\boldsymbol{N_h}})_{h=1,H} \\ (\boldsymbol{y'}_{\boldsymbol{M_j}} \boldsymbol{\beta}_{\boldsymbol{M_j}})_{j=1,J} \end{pmatrix}$$

et

$$\overline{\boldsymbol{\gamma}} = \mathrm{Cov} \begin{pmatrix} (\boldsymbol{x'}_{\boldsymbol{N_h}} \boldsymbol{\alpha}_{\boldsymbol{N_h}})_{h=1,H} \\ (\boldsymbol{y'}_{\boldsymbol{M_j}} \boldsymbol{\beta}_{\boldsymbol{M_j}})_{j=1,J} \end{pmatrix} ; \hat{t}_{H-T}$$

Le paramètre optimal $\overline{\boldsymbol{\theta}}$ ainsi obtenu sera toujours inconnu mais dans certains cas il peut être estimé. Les éléments de variance dans $\overline{\boldsymbol{\Gamma}}, \overline{\boldsymbol{\gamma}}$ peuvent être estimés selon la théorie d'estimation de la variance développée auparavant.

B. On considère maintenant le cas extrême où les matrices $\overline{H}$ et $\overline{J}$ sont toutes les deux égales à $\mathbf{1}'_{\mathbf{N}}$, alors $a_k = a$ et $b_k = b$ pour tous $k \in U$; cela conduit à l'expression suivante pour $\hat{Z}$

$$\begin{aligned} \hat{Z} &= (a,b) \begin{pmatrix} \mathbf{1}'_{\mathbf{N}} & 0 \\ 0 & \mathbf{1}'_{\mathbf{N}} \end{pmatrix} \begin{pmatrix} \mathrm{diag}x & 0 \\ 0 & \mathrm{diag}y \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} + \hat{t}_{H-T} \\ &= (a,b) \begin{pmatrix} \mathbf{1}'_{\mathbf{N}} \, \mathrm{diag}\mathbf{x} & 0 \\ 0 & \mathbf{1}'_{\mathbf{N}} \, \mathrm{diag}\mathbf{y} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} + \hat{t}_{H-T} \qquad (2.42) \\ &= (a,b) \begin{pmatrix} \boldsymbol{x'\alpha} \\ \boldsymbol{y'\beta} \end{pmatrix} + \hat{t}_{H-T} \\ &= a(\hat{X}^{1*} - \hat{X}^{12}) + b(\hat{Y}^{2*} - \hat{Y}^{12}) + \hat{t}_{H-T} \qquad (2.43) \end{aligned}$$

où $\boldsymbol{x'} = (x_1, \ldots, x_N)$, $\boldsymbol{y'} = (y_1, \ldots, y_N)$ et $\hat{X}^{\oplus} = \sum_U x_k \frac{\varepsilon_k^{\oplus}}{\pi_k^{\oplus}}$ pour $\oplus \in \{12, 1*\}$ $\boldsymbol{x'\alpha} = \hat{X}^{1*} - \hat{X}^{12}$ et $\hat{Y}^{\otimes} = \sum_U y_k \frac{\varepsilon_k^{\otimes}}{\pi_k^{\otimes}}$ pour $\otimes \in \{12, 2*\}$, $\boldsymbol{y'\beta} =$

$\hat{Y}^{2*} - \hat{Y}^{12}$; $\hat{t}_{H-T} = \phi\hat{X}^{12} + \psi\hat{Y}^{12}$. Dans ce cas, le paramètre est de la forme suivante

$$\boldsymbol{\theta'} = (a, b) \begin{pmatrix} \mathbf{1'_N} & 0 \\ 0 & \mathbf{1'_N} \end{pmatrix}$$

ce que revient à trouver $\overline{\boldsymbol{\theta'}} = (a, b)$. La valeur optimale $\overline{\boldsymbol{\theta}}_{opt} = -\overline{\boldsymbol{\Gamma}}^{-1}\overline{\boldsymbol{\gamma}}$ où

$$\begin{cases} \overline{\boldsymbol{\Gamma}} &= \text{Var} \begin{pmatrix} \boldsymbol{x'\alpha} \\ \boldsymbol{y'\beta} \end{pmatrix} = \begin{pmatrix} V(\hat{X}^{1*} - \hat{X}^{12}) & \text{Cov}(\boldsymbol{x'\alpha}, \boldsymbol{y'\beta}) \\ \text{Cov}(\boldsymbol{x'\alpha}, \boldsymbol{y'\beta}) & V(\hat{Y}^{2*} - \hat{Y}^{12}) \end{pmatrix} \\ \overline{\boldsymbol{\gamma}} &= \text{Cov}\left(\begin{pmatrix} \boldsymbol{x'\alpha} \\ \boldsymbol{y'\beta} \end{pmatrix}, \hat{t}_{H-T} \right) = \begin{pmatrix} \text{Cov}(\boldsymbol{x'\alpha}, \hat{t}_{H-T}) \\ \text{Cov}(\boldsymbol{y'\beta}, \hat{t}_{H-T}) \end{pmatrix} \end{cases} \tag{2.44}$$

Dans cette situation, $\overline{\boldsymbol{\theta}}_{opt}$ dépend seulement de deux termes de variance et de trois de covariance. Tous ces termes ne sont pas connus mais ils peuvent être estimés dans certains cas.

2.3.5 Calcul de paramètre optimal $\boldsymbol{\theta}$ $2N$-dimensionnel pour une variable d'intérêt $Z = \phi t_x + \psi t_y$ et pour des plans de sondage bidimensionnels particuliers

Dans la suite, on va calculer l'expression de $\boldsymbol{\theta'} = ((a_k)_{k=1}^N, (b_k)_{k=1}^N) \in \mathbb{R}^{2N}$ optimal pour les plans de sondages décrits dans la section précédente. On suppose maintenant que les variables $\mathcal{X}$ et $\mathcal{Y}$ sont positives, c'est à dire $x_k > 0$ et $y_k > 0$ pour tous $k \in \mathcal{U}$.
On rappelle l'expression de $\hat{Z}$

$$\hat{Z} = \boldsymbol{\theta'} \, \text{diag} \begin{pmatrix} x \\ \dots \\ y \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} + \hat{t}_{H-T}$$

pour $\hat{t}_{H-T} = \sum_{s_{12}} \frac{\phi x_k + \psi y_k}{\pi_k^{12}}$.

Sondage de Poisson bidimensionnel

On considère le sondage PO2, décrit dans l'exemple (3). La variance de l'estimateur $\hat{Z}^{PO2}$ est donnée par (2.101) avec $c = \sum_{k \in U} \left(\frac{1}{\pi_k^{12}} - 1 \right) (\phi x_k + \psi y_k)^2$. Pour obtenir $\boldsymbol{\theta}_{opt} = -\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}$ avec $\boldsymbol{\Gamma}$ et $\boldsymbol{\gamma}$ donnés par (2.31), (2.32) on

a besoin de la variance de $(\boldsymbol{\alpha}, \boldsymbol{\beta})'$ et de la covariance entre $(\boldsymbol{\alpha}, \boldsymbol{\beta})'$ et $\dfrac{\boldsymbol{\varepsilon}^{12}}{\boldsymbol{\pi}^{12}}$. Avec l'hypothèse que $x_k > 0$ et $y_k > 0$, on obtient l'expression suivante pour $\boldsymbol{\theta}_{opt}$

$$\begin{aligned}
\boldsymbol{\theta}_{opt} &= -\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma} \\
&= -\left[\operatorname{diag} \begin{pmatrix} x \\ \dots \\ y \end{pmatrix} \right]^{-1} \left[\operatorname{Var} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} \right]^{-1} \operatorname{Cov} \left(\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}, \frac{\varepsilon^{12}}{\pi^{12}} \right) (\phi \boldsymbol{x} + \psi \boldsymbol{y}).
\end{aligned}$$

Pour calculer la variance de $(\boldsymbol{\alpha}, \boldsymbol{\beta})'$, on utilise (2.20) et (2.28) et on obtient

$$\begin{aligned}
\operatorname{Var} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} &= \begin{pmatrix} \operatorname{diag} \left(\frac{1}{\pi_k^{1*}} + \frac{1}{\pi_k^{12}} \right)_{k=1,N} & \operatorname{diag} \left(\frac{1}{\pi_k^{12}} \right)_{k=1,N} \\ \operatorname{diag} \left(\frac{1}{\pi_k^{12}} \right)_{k=1,N} & \operatorname{diag} \left(\frac{1}{\pi_k^{2*}} + \frac{1}{\pi_k^{12}} \right)_{k=1,N} \end{pmatrix} \\
&= \left(\begin{array}{c|c} A & B \\ \hline B & D \end{array} \right)
\end{aligned}$$

Ensuite, on obtient l'inverse en utilisant la formule (Rao 1973) :

$$\left[\operatorname{Var} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} \right]^{-1} = \begin{pmatrix} A^{-1} + FE^{-1}F' & -FE^{-1} \\ -E^{-1}F' & E^{-1} \end{pmatrix}$$

où $F = A^{-1}B$ et $E = D - B'A^{-1}B$.
On note $S_k = \pi_k^{1*} + \pi_k^{12} + \pi_k^{2*}$ pour tous $k \in U$, alors l'inverse de la matrice de variance-covariance de $(\boldsymbol{\alpha}, \boldsymbol{\beta})'$ devient :

$$\left[\operatorname{Var} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} \right]^{-1} = \begin{pmatrix} \operatorname{diag} \frac{\pi_k^{1*}}{\pi_k^{1*} + \pi_k^{12}} \left(\pi_k^{12} + \frac{\pi_k^{1*}\pi_k^{2*}}{S_k} \right) & -\operatorname{diag} \frac{\pi_k^{1*}\pi_k^{2*}}{S_k} \\ -\operatorname{diag} \frac{\pi_k^{1*}\pi_k^{2*}}{S_k} & \operatorname{diag} \frac{\pi_k^{2*}(\pi_k^{1*} + \pi_k^{12})}{S_k} \end{pmatrix}.$$

La covariance entre $(\boldsymbol{\alpha}, \boldsymbol{\beta})'$ et $\dfrac{\boldsymbol{\varepsilon}^{12}}{\boldsymbol{\pi}^{12}}$ a l'expression suivante

$$\begin{aligned}
\operatorname{Cov} \left(\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}, \frac{\boldsymbol{\varepsilon}^{12}}{\boldsymbol{\pi}^{12}} \right) &= E \left(\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} \left(\frac{\boldsymbol{\varepsilon}^{12}}{\boldsymbol{\pi}^{12}} \right)' \right) - \underbrace{E \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}}_{0} E \left(\frac{\boldsymbol{\varepsilon}^{12}}{\boldsymbol{\pi}^{12}} \right)' \\
&= \begin{pmatrix} \operatorname{diag} \left(-\frac{1}{\pi_k^{12}} \right)_{k=1,N} \\ \operatorname{diag} \left(-\frac{1}{\pi_k^{12}} \right)_{k=1,N} \end{pmatrix}
\end{aligned}$$

parce que pour tous $k, l \in U$, $E\left(\alpha_k \dfrac{\varepsilon_l^{12}}{\pi_l^{12}} \right) = E\left(\beta_k \dfrac{\varepsilon_l^{12}}{\pi_l^{12}} \right) = -\dfrac{1}{\pi_k^{12}}$ pour $k = l$ et zéro sinon.

Il résulte en particulier les expressions pour $\mathbf{\Gamma}$ et $\boldsymbol{\gamma}$:

$$\mathbf{\Gamma}_{PO2} = \begin{pmatrix} \text{diag}\left(x_k^2\left(\frac{1}{\pi_k^{1*}}+\frac{1}{\pi_k^{12}}\right)\right)_{k\in U} & \text{diag}\left(\frac{x_k y_k}{\pi_k^{12}}\right)_{k\in U} \\ \text{diag}\left(\frac{x_k y_k}{\pi_k^{12}}\right)_{k\in U} & \text{diag}\left(y_k^2\left(\frac{1}{\pi_k^{2*}}+\frac{1}{\pi_k^{12}}\right)\right)_{k\in U} \end{pmatrix} \tag{2.45}$$

$$\boldsymbol{\gamma}_{PO2} = \begin{pmatrix} \left(\frac{-x_k(\phi x_k+\psi y_k)}{\pi_k^{12}}\right)_{k\in U} \\ \left(\frac{-y_k(\phi x_k+\psi y_k)}{\pi_k^{12}}\right)_{k\in U} \end{pmatrix} \tag{2.46}$$

On peut calculer maintenant le paramètre optimal $\boldsymbol{\theta}_{opt}^{PO2}$.

Résultat 2.3.11 *: Pour un sondage PO2 et pour une variable d'intérêt* $Z = \phi t_x + \psi t_y$,

i. le paramètre optimal a l'expression suivante :

$$\boldsymbol{\theta}_{opt}^{PO2} = ((a_{k,opt}^{PO2})_{k=1}^N, (b_{k,opt}^{PO2})_{k=1}^N)' = \begin{pmatrix} \left(\frac{(\phi x_k+\psi y_k)\pi_k^{1*}}{x_k S_k}\right)_{k=1,N} \\ \left(\frac{(\phi x_k+\psi y_k)\pi_k^{2*}}{y_k S_k}\right)_{k=1,N} \end{pmatrix}. \tag{2.47}$$

ii. L'estimateur $\hat{Z}_{opt}^{PO2}$ *devient :*

$$\hat{Z}_{opt}^{PO2} = \sum_{k\in s_1^*\cup s_{12}\cup s_2^*} \frac{\phi x_k+\psi y_k}{\pi_k^{1*}+\pi_k^{12}+\pi_k^{2*}}. \tag{2.48}$$

iii. avec une variance

$$V_{opt}^{PO2} = V(\hat{Z}_{opt}^{PO2}) = \sum_{k\in U}\left(\frac{1}{S_k}-1\right)(\phi x_k+\psi y_k)^2. \tag{2.49}$$

iv. La variance optimale V_{opt}^{PO2} *peut être estimée dans l'échantillon intersection par*

$$\hat{V}_{opt}^{PO2} = \sum_{k\in s_{12}}\left(\frac{1}{S_k}-1\right)\frac{(\phi x_k+\psi y_k)^2}{\pi_k^{12}}$$

Preuve :

i. On remplace les expressions de $\text{Var}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}$ et $\text{Cov}\left(\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}, \frac{\boldsymbol{\varepsilon}^{12}}{\boldsymbol{\pi}^{12}}\right)$ dans celle de $\boldsymbol{\theta}_{opt}$.

ii. On a

$$\begin{aligned}\hat{Z}_{opt}^{PO2} &= \sum_{k\in U} a_{k,opt}^{PO2}\frac{x_k}{\pi_k^{1*}}\varepsilon_k^{1*} + \sum_{k\in U}(\phi - a_{k,opt}^{PO2})\frac{x_k}{\pi_k^{12}}\varepsilon_k^{12} \\ &\quad + \sum_{k\in U} b_{k,opt}^{PO2}\frac{y_k}{\pi_k^{2*}}\varepsilon_k^{2*} + \sum_{k\in U}(\psi - b_{k,opt}^{PO2})\frac{y_k}{\pi_k^{12}}\varepsilon_k^{12} \\ &= \sum_{k\in s_1^*\cup s_{12}\cup s_2^*}\frac{\phi x_k + \psi y_k}{S_k} \\ &= \sum_{k\in s_1^*\cup s_{12}\cup s_2^*}\frac{\phi x_k + \psi y_k}{\pi_k^{1*}+\pi_k^{12}+\pi_k^{2*}}. \qquad (2.50)\end{aligned}$$

iii. La variance se calcule d'après la formule

$$\begin{aligned}V_{opt}^{PO2} &= V(\hat{t}_{H-T}) - \boldsymbol{\gamma}\text{`}\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma} \\ &= \sum_{k\in U}\left(\frac{1}{\pi_k^{12}}-1\right)(\phi x_k+\psi y_k)^2 - \sum_{k\in U}\frac{(\phi x_k+\psi y_k)^2(\pi_k^{1*}+\pi_k^{2*})}{\pi_k^{12}S_k} \\ &= \sum_{k\in U}\left(\frac{1}{S_k}-1\right)(\phi x_k+\psi y_k)^2\end{aligned}$$

iv. L'estimateur de la variance est une conséquence directe du résultat 2.3.6.

□

Les paramètres ainsi calculés dependent des quantités inconnues x_k, y_k pour $k \in U$ où encore de $\frac{x_k}{y_k}$; par conséquent, ils ne peuvent pas être déduits avant une enquête.

Sondage de Bernoulli bidimensionnel

On considère le sondage de Bernoulli bidimensionnel (BE2) décrit dans l'exemple (4) de la section précédente. Alors les paramètres optimaux $a_{k,opt}^{BE2}$ et $b_{k,opt}^{BE2}$ pour tous $k \in U$ s'obtiennent en utilisant (2.51) pour $\pi_k^{\oplus} = \pi^{\oplus}$ pour $\oplus \in \{1*, 12, 2*\}$. On peut obtenir également l'estimateur optimal ; puisque $S_k = S = \pi^{1*} + \pi^{12} + \pi^{2*}$ pour tous $k \in U$, on a

Résultat 2.3.12 *: Pour un sondage BE2 et pour une variable d'intérêt* $Z = \phi t_x + \psi t_y$

i. Le paramètre optimal a l'expression

$$\boldsymbol{\theta}_{opt}^{BE2} = (\boldsymbol{a}_{opt}^{BE2}, \boldsymbol{b}_{opt}^{BE2})' = \left(\begin{array}{c} \left(\dfrac{(\phi x_k + \psi y_k)\pi^{1*}}{x_k S} \right)_{k=1,N} \\ \hline \left(\dfrac{(\phi x_k + \psi y_k)\pi^{2*}}{y_k S} \right)_{k=1,N} \end{array} \right) \quad (2.51)$$

ii. l'estimateur optimal $\hat{Z}_{opt}^{BE2}$ est

$$\hat{Z}_{opt}^{BE2} = \frac{1}{S} \sum_{k \in s_1^* \cup s_{12} \cup s_2^*} (\phi x_k + \psi y_k) \quad (2.52)$$

iii. avec la variance optimale

$$V_{opt}^{BE2} = V(\hat{Z}_{opt}^{BE2}) = \left(\frac{1}{S} - 1 \right) \sum_{k \in U} (\phi x_k + \psi y_k)^2. \quad (2.53)$$

iv. Un estimateur de la variance est

$$\hat{V}_{opt}^{BE2} = \left(\frac{1}{S} - 1 \right) \sum_{k \in s_{12}} \frac{(\phi x_k + \psi y_k)^2}{\pi^{12}}$$

Sondage de Bernoulli bidimensionnel conditionnel à la taille où sondage aléatoire simple sans remise bidimensionnel

Nous reprenons le sondage SAS2 decrit dans l'exemple 2. Nous avons besoin de calculer $\boldsymbol{\theta}_{opt}$ de $V((\boldsymbol{\alpha}, \boldsymbol{\beta})')$ et $Cov((\boldsymbol{\alpha}, \boldsymbol{\beta})', \hat{t}_{H-T})$.

On calcule $V((\boldsymbol{\alpha}, \boldsymbol{\beta})')$. Les relations (2.13), (2.15) et (2.28) nous donnent

$$V((\boldsymbol{\alpha}, \boldsymbol{\beta})') = \begin{pmatrix} k_1(\boldsymbol{I}_N - \boldsymbol{P}) & k_3(\boldsymbol{I}_N - \boldsymbol{P}) \\ k_3(\boldsymbol{I}_N - \boldsymbol{P}) & k_2(\boldsymbol{I}_N - \boldsymbol{P}) \end{pmatrix} \quad (2.54)$$

$$Cov\left((\boldsymbol{\alpha}, \boldsymbol{\beta})', \frac{\boldsymbol{\varepsilon}^{12}}{\boldsymbol{\pi}^{12}} \right) = \begin{pmatrix} -k_3(\boldsymbol{I}_N - \boldsymbol{P}) \\ -k_3(\boldsymbol{I}_N - \boldsymbol{P}) \end{pmatrix}$$

avec les constantes k_1, k_2 et k_3 données par

$$\begin{aligned} k_1 &= \frac{N}{N-1} \left(\frac{1}{f_{1*}} + \frac{1}{f_{12}} \right) \\ k_2 &= \frac{N}{N-1} \left(\frac{1}{f_{2*}} + \frac{1}{f_{12}} \right) \\ k_3 &= \frac{N}{N-1} \frac{1}{f_{12}} \end{aligned}$$

où $\boldsymbol{I}_N$ la matrice unité d'ordre N, et $P = \frac{1}{N}\mathbf{1_N}\mathbf{1'_N}$. Alors (Rao 1973),

$$|V((\boldsymbol{\alpha},\boldsymbol{\beta})')| = |k_1(\boldsymbol{I}_N - \boldsymbol{P})| \times |V(\boldsymbol{\beta}) - Cov(\boldsymbol{\alpha},\boldsymbol{\beta})V^{-1}(\boldsymbol{\alpha})Cov(\boldsymbol{\alpha},\boldsymbol{\beta})|$$

et comme $|k_1(\boldsymbol{I}_N - \boldsymbol{P})| = 0$ on est dans le cas pathologique où la matrice $\boldsymbol{\Gamma}$ est non inversible.

On a vu (chapitre 1, **remark** 1.2.3) dans le cas d'un sondage aléatoire simple sans remise unidimensionnel que tout vecteur $\boldsymbol{y} = (y_1,\ldots,y_N)'$ constant (c'est à dire qui est élément de l'espace $(\boldsymbol{I}_N - \boldsymbol{P})^{\perp}$) minimise la variance de $\sum_s \frac{y_k}{\pi_k}$; les vecteurs dans $(\boldsymbol{I}_N - \boldsymbol{P})^{\perp}$ ont l'espression $\boldsymbol{y}_0 = ct \times \mathbf{1_N}$ où $ct-$constante réelle.

On va utiliser ce résultat pour déterminer un vecteur $\boldsymbol{\theta}$ qui va minimiser la variance de $\hat{Z}$. On écrit $\hat{Z}$ de la manière suivante

$$\hat{Z} = \tilde{\boldsymbol{\theta}}' \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} + \hat{t}_{H-T}$$

avec

$$\begin{aligned} \tilde{\boldsymbol{\theta}}' &= (\tilde{\boldsymbol{\theta}}'_1, \tilde{\boldsymbol{\theta}}'_2) = \boldsymbol{\theta}' \begin{pmatrix} \mathrm{diag}x & 0 \\ 0 & \mathrm{diag}y \end{pmatrix} \\ &= (a_1x_1,\ldots,a_Nx_N, b_1y_1,\ldots,b_Ny_N). \end{aligned}$$

Alors la variance de $\hat{Z}$ a l'expression suivante

$$V(\hat{Z}) = \tilde{\boldsymbol{\theta}}'\mathrm{Var}\left((\boldsymbol{\alpha},\boldsymbol{\beta})'\right)\tilde{\boldsymbol{\theta}} + 2\tilde{\boldsymbol{\theta}}'\mathrm{Cov}\left((\boldsymbol{\alpha},\boldsymbol{\beta})', \frac{\boldsymbol{\varepsilon}^{12}}{\boldsymbol{\pi}^{12}}\right)(\phi\boldsymbol{x} + \psi\boldsymbol{y}) + c$$

où $c = Var(\hat{t}_{H-T})$ et $\boldsymbol{x}' = (x_1,\ldots,x_N)$, $\boldsymbol{y}' = (y_1,\ldots,y_N)$. Pour determiner $\tilde{\boldsymbol{\theta}}_{opt}$ on doit résoudre l'équation

$$\frac{\partial V}{\partial \tilde{\boldsymbol{\theta}}} = 2\mathrm{Var}\left((\boldsymbol{\alpha},\boldsymbol{\beta})'\right)\tilde{\boldsymbol{\theta}} + 2\mathrm{Cov}\left((\boldsymbol{\alpha},\boldsymbol{\beta})', \frac{\boldsymbol{\varepsilon}^{12}}{\boldsymbol{\pi}^{12}}\right)(\phi\boldsymbol{x} + \psi\boldsymbol{y}) = 0 \qquad (2.55)$$

La variance $\mathrm{Var}\left((\boldsymbol{\alpha},\boldsymbol{\beta})'\right)$ est donnée par (2.54). Quant au terme de covariance, on note $\boldsymbol{z} = \phi\boldsymbol{x} + \psi\boldsymbol{y}$ et on a

$$\mathrm{Cov}\left((\boldsymbol{\alpha},\boldsymbol{\beta})', \frac{\boldsymbol{\varepsilon}^{12}}{\boldsymbol{\pi}^{12}}\right)\boldsymbol{z} = \begin{pmatrix} -k_3(\boldsymbol{I}_N - \boldsymbol{P})\boldsymbol{z} \\ -k_3(\boldsymbol{I}_N - \boldsymbol{P})\boldsymbol{z} \end{pmatrix}.$$

Alors l'équation (2.55) devient

$$\begin{cases} k_1(\boldsymbol{I}_N - \boldsymbol{P})\tilde{\boldsymbol{\theta}}_1 & + & k_3(\boldsymbol{I}_N - \boldsymbol{P})(\tilde{\boldsymbol{\theta}}_2 - \boldsymbol{z}) & = & 0 \\ k_3(\boldsymbol{I}_N - \boldsymbol{P})(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{z}) & + & k_2(\boldsymbol{I}_N - \boldsymbol{P})\tilde{\boldsymbol{\theta}}_2 & = & 0 \end{cases} \qquad (2.56)$$

système équivalent à

$$\left\{\begin{array}{ccccc} k_1\tilde{\boldsymbol{\theta}}_1 & + & k_3(\tilde{\boldsymbol{\theta}}_2 - \boldsymbol{z}) & = & q_1\mathbf{1_N} \\ k_3(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{z}) & + & k_2\tilde{\boldsymbol{\theta}}_2 & = & q_2\mathbf{1_N} \end{array}\right. \tag{2.57}$$

avec q_1 et q_2 deux constante réelles. On soustrait la deuxième équation de la première dans (2.56) et on obtient

$$k_1(\boldsymbol{I}_N - \boldsymbol{P})\tilde{\boldsymbol{\theta}}_1 - k_2(\boldsymbol{I}_N - \boldsymbol{P})\tilde{\boldsymbol{\theta}}_2 + k_3(\boldsymbol{I}_N - \boldsymbol{P})(\tilde{\boldsymbol{\theta}}_2 - \tilde{\boldsymbol{\theta}}_1) = 0$$

ou encore

$$\frac{1}{f_{1*}}(\boldsymbol{I}_N - \boldsymbol{P})\tilde{\boldsymbol{\theta}}_1 - \frac{1}{f_{2*}}(\boldsymbol{I}_N - \boldsymbol{P})\tilde{\boldsymbol{\theta}}_2 = 0$$

si on tient compte des expressions de k_1, k_2 et k_3. Cette dernière équation est vérifiée pour $\tilde{\boldsymbol{\theta}}_1$ et $\tilde{\boldsymbol{\theta}}_2$ qui satisfont

$$\frac{1}{f_{1*}}\tilde{\boldsymbol{\theta}}_1 - \frac{1}{f_{2*}}\tilde{\boldsymbol{\theta}}_2 = q_0\mathbf{1_N} \tag{2.58}$$

avec q_0 une constante réelle. Cette relation et la première équation de (2.57) nous donnent

$$\left(\frac{k_1}{f_{2*}} + k_3\right)\tilde{\boldsymbol{\theta}}_2 = k_3\boldsymbol{z} + (k_1 f_{1*} q_0 + q_1)\mathbf{1_N}.$$

Par ailleurs, $\tilde{\boldsymbol{\theta}_2}' = \boldsymbol{b}'$ diag$y$ où $\boldsymbol{b}' = (b_1, \ldots, b_N)$. Alors, pour des $y_k > 0$ on obtient

$$\boldsymbol{b}_{opt}^{SAS2} = \left(\frac{k_1}{f_{2*}} + k_3\right)^{-1} (\text{diag}y)^{-1}\left[k_3\boldsymbol{z} + (k_1 f_{1*} q_0 + q_1)\mathbf{1_N}\right] \tag{2.59}$$

La relation (2.58) nous permet alors d'obtenir

$$\tilde{\boldsymbol{\theta}}_1 = \frac{f_{1*}}{k_1 + k_3 f_{2*}}\left[k_3\boldsymbol{z} + (k_1 f_{1*} q_0 + q_1)\mathbf{1_N}\right] + f_{1*} q_0 \mathbf{1_N}$$

et par conséquent, $\boldsymbol{a}' = (a_1, \ldots, a_N)$ a la valeur optimale suivante

$$\boldsymbol{a}_{opt}^{SAS2} = (\text{diag}x)^{-1}\left\{\frac{f_{1*}}{k_1 + k_3 f_{2*}}\left[k_3\boldsymbol{z} + (k_1 f_{1*} q_0 + q_1)\mathbf{1_N}\right] + f_{1*} q_0 \mathbf{1_N}\right\} \tag{2.60}$$

Finalement, nous obtenons l'expression de $\boldsymbol{\theta}' = (\boldsymbol{a}', \boldsymbol{b}')$ qui minimise la variance de $\hat{Z}$ utilisant les relations (2.59) et (2.60).

La variance minimale a dans ce cas l'expression, suite à (2.55)

$$V_{opt}^{SAS2} = c = Var(\hat{t}_{H-T}).$$

2.3.6 Le calcul du paramètre optimal bidimensionnnel $\overline{\boldsymbol{\theta}} = (a,b)' \in \mathbb{R}^2$ pour $Z = \phi t_x + \psi t_y$

On revient dans cette section au cas extrême où $\overline{\boldsymbol{\theta}} = (a,b)'$.

La relation (2.43) nous donne l'expression de $\hat{Z}$ pour cette situation particulière

$$\hat{Z} = a(\hat{X}^{1*} - \hat{X}^{12}) + b(\hat{Y}^{2*} - \hat{Y}^{12}) + \hat{t}_{H-T}$$

avec $\hat{t}_{H-T} = \phi\hat{X}^{12} + \psi\hat{Y}^{12}$.

Nous étudions dans la suite les paramètres optimaux pour les plans de sondage usuels.

Sondage de Poisson bidimensionnel

On connait les expressions de $\boldsymbol{\Gamma}_{PO2}$ et $\boldsymbol{\gamma}_{PO2}$ pour $\boldsymbol{\theta}$ de dimension $2N$. Alors, grace aux formules (2.40) et (2.41), on obtient

$$\begin{aligned}
\overline{\boldsymbol{\Gamma}}_{PO2} &= \begin{pmatrix} \mathbf{1'_N} & 0 \\ 0 & \mathbf{1'_N} \end{pmatrix} \boldsymbol{\Gamma}_{PO2} \begin{pmatrix} \mathbf{1_N} & 0 \\ 0 & \mathbf{1_N} \end{pmatrix} \\
\overline{\boldsymbol{\gamma}}_{PO2} &= \begin{pmatrix} \mathbf{1'_N} & 0 \\ 0 & \mathbf{1'_N} \end{pmatrix} \boldsymbol{\gamma}_{PO2}
\end{aligned}$$

et $\boldsymbol{\Gamma}_{PO2}$, $\boldsymbol{\gamma}_{PO2}$ donnés par (2.45) et (2.46). Cela implique comme expressions pour $\overline{\boldsymbol{\Gamma}}_{PO2}$ et $\overline{\boldsymbol{\gamma}}_{PO2}$

$$\overline{\boldsymbol{\Gamma}}_{PO2} = \begin{pmatrix} \sum_{k\in U} x_k^2 \left(\frac{1}{\pi_k^{1*}} + \frac{1}{\pi_k^{12}}\right) & \sum\limits_{k\in U} \frac{x_k y_k}{\pi_k^{12}} \\ \sum\limits_{k\in U} \frac{x_k y_k}{\pi_k^{12}} & \sum_{k\in U} y_k^2 \left(\frac{1}{\pi_k^{2*}} + \frac{1}{\pi_k^{12}}\right) \end{pmatrix}$$

$$\overline{\boldsymbol{\gamma}}_{PO2} = \begin{pmatrix} \sum\limits_{k\in U} \frac{-x_k(\phi x_k + \psi y_k)}{\pi_k^{12}} \\ \sum\limits_{k\in U} \frac{-y_k(\phi x_k + \psi y_k)}{\pi_k^{12}} \end{pmatrix}$$

L'inverse de $\overline{\boldsymbol{\Gamma}}_{PO2}$ est donnée par

$$\overline{\boldsymbol{\Gamma}}_{PO2}^{-1} = \frac{1}{\delta} \begin{pmatrix} \sum_{k\in U} y_k^2 \left(\frac{1}{\pi_k^{2*}} + \frac{1}{\pi_k^{12}}\right) & \sum\limits_{k\in U} \frac{-x_k y_k}{\pi_k^{12}} \\ \sum\limits_{k\in U} \frac{-x_k y_k}{\pi_k^{12}} & \sum_{k\in U} x_k^2 \left(\frac{1}{\pi_k^{1*}} + \frac{1}{\pi_k^{12}}\right) \end{pmatrix}$$

$$\text{avec } \delta = \sum_{k\in U} x_k^2 \left(\frac{1}{\pi_k^{1*}} + \frac{1}{\pi_k^{12}}\right) \sum_{k\in U} y_k^2 \left(\frac{1}{\pi_k^{2*}} + \frac{1}{\pi_k^{12}}\right) - \left(\sum_{k\in U} \frac{x_k y_k}{\pi_k^{12}}\right)^2.$$

On peut obtenir alors les quantités $a_{k,opt}^{PO2}$ et $b_{k,opt}^{PO2}$ données par $\overline{\boldsymbol{\theta}}_{opt}^{PO2} = ((a_{k,opt}^{PO2})_{k=1}^N, (b_{k,opt}^{PO2})_{k=1}^N) = -\overline{\boldsymbol{\Gamma}}_{PO2}^{-1}\overline{\boldsymbol{\gamma}}_{PO2}$ et $\overline{\boldsymbol{\Gamma}}_{PO2}$, $\overline{\boldsymbol{\gamma}}_{PO2}$ calculés ci-dessus.

Remarque 2.3.3 *: Les quantités $a_{k,opt}^{PO2}$ et $b_{k,opt}^{PO2}$ sont toujours inconnues pour toutes les unités dans la population. Contrairement au cas $2N$-dimensionnel, elles peuvent être estimées selon le principe presenté dans la section 2.3.3.*

Le sondage BE2 est un cas particulier de PO2. Par conséquent, le paramètre optimal est obtenu directement en appliquant le résultat ci-dessus pour $\pi_k^{1*} = \pi_{1*}$, $\pi_k^{12} = \pi_{12}$ et $\pi_k^{2*} = \pi_{2*}$ pour tous $k \in U$.

Sondage de Bernoulli bidimensionnel conditionnel à la taille où sondage aléatoire simple sans remise bidimensionnel

Nous allons reprendre le sondage SAS2 décrit dans l'exemple (2). Pour calculer $\overline{\boldsymbol{\theta}}_{opt}$, on va utiliser (2.44) quand chaque échantillon s_{1*}, s_{12} ou s_{2*} est un sondage aléatoire simple sans remise dans $\mathcal{U}$. Par conséquent, on peut en déduire facilement les expressions des termes de variance.
On note S_x^2, S_y^2 les variances de $\mathcal{X}, \mathcal{Y}$ dans la population U et $S_{x,y}^2$ la covariance entre $\mathcal{X}$ et $\mathcal{Y}$.

$$S_x^2 = \frac{1}{N-1}\sum_U (x_k - \overline{X})^2, S_y^2 = \frac{1}{N-1}\sum_U (y_k - \overline{Y})^2,$$

$$S_{x,y}^2 = \frac{1}{N-1}\sum_U (x_k - \overline{X})(y_k - \overline{Y})$$

où $\overline{X} = \frac{1}{N}\sum_U x_k, \overline{Y} = \frac{1}{N}\sum_U y_k$. Alors,

$$V(\hat{X}^{\otimes}) = N^2 \frac{1-f_{\otimes}}{n_{\otimes}} S_x^2 \quad \text{pour} \quad \otimes \in \{1*, 12\}$$
$$V(\hat{Y}^{\oplus}) = N^2 \frac{1-f_{\oplus}}{n_{\oplus}} S_y^2 \quad \text{pour} \quad \otimes \in \{12, 2*\}$$

avec $f_{\oplus}$ donné par (2.12).
Quant aux termes de covariance, on utilise les probabilités d'inclusion du second degré données par (2.15). Alors les différents termes de covariance vérifient :

$$\text{Cov } (\hat{X}^{12}, \hat{Y}^{12}) = N\frac{1-f_{12}}{f_{12}} S_{x,y}^2$$
$$\text{Cov } (\hat{X}^{1*}, \hat{X}^{12}) = -NS_x^2, \text{ Cov } (\hat{Y}^{12}, \hat{Y}^{2*}) = -NS_y^2$$
$$\text{Cov } (\hat{X}^{1*}, \hat{Y}^{12}) = \text{ Cov } (\hat{X}^{12}, \hat{Y}^{2*}) = \text{ Cov } (\hat{X}^{1*}, \hat{Y}^{2*}) = -NS_{x,y}^2.$$

L'estimateur $\hat{Z}$ s'écrit

$$\hat{Z} = a(\hat{X}^{1*} - \hat{X}^{12}) + b(\hat{Y}^{2*} - \hat{Y}^{12}) + \phi\hat{X}^{12} + \psi\hat{Y}^{12} \tag{2.61}$$

et remplaçant dans (2.44) les termes de variance et covariance calculés ci-dessus, on obtient la valeur optimale de $\overline{\boldsymbol{\theta}}$.

Résultat 2.3.13 *: Pour un sondage SAS2 et une variable d'intérêt $Z = \phi t_x + \psi t_y$, le paramètre $\overline{\boldsymbol{\theta}} = (a, b)'$ a la valeur optimale*

$$\overline{\boldsymbol{\theta}}_{opt}^{SAS2} = \frac{-h_1 h_2}{1 - \rho^2 h_1 h_2} \begin{pmatrix} \phi\rho^2 + \rho\psi(1 - h_2^{-1})S - \phi h_2^{-1} \\ \psi\rho^2 + \rho\phi(1 - h_1^{-1})S^{-1} - \psi h_1^{-1} \end{pmatrix} \tag{2.62}$$

où $h_1 = \frac{n_1^*}{n_1}$, $h_2 = \frac{n_2^*}{n_2}$ sont les taux de renouvellement, $S = \frac{S_y}{S_x}$ et $\boldsymbol{\rho}$ est le coefficient de corrélation. La variance optimale peut se calculer d'après la formule $Var_{opt} = V(t_{H-T}) - \boldsymbol{\gamma}'\boldsymbol{\Gamma}\boldsymbol{\gamma}$.

Remarque 2.3.4 *: On ne peut pas déduire $\overline{\boldsymbol{\theta}}_{opt}^{SAS2}$ mais on peut l'estimer. Les termes de variance sont estimés comme suit*

$$S_x^2 = \frac{1}{n_1 - 1} \sum_{k \in s_1} (x_k - \overline{x}_{s_1})^2, \quad \overline{x}_{s_1} = \frac{1}{n_1} \sum_{k \in s_1} x_k$$

$$S_y^2 = \frac{1}{n_2 - 1} \sum_{k \in s_2} \left(y_k - \overline{y}_{s_2}\right)^2, \quad \overline{y}_{s_2} = \frac{1}{n_2} \sum_{k \in s_2} y_k$$

et si $n_{12} > 0$, le terme de covariance peut être estimé par

$$S_{xy}^2 = \frac{1}{n_{12} - 1} \sum_{k \in s_{12}} (x_k - \overline{x}_{s_{12}}) \left(y_k - \overline{y}_{s_{12}}\right)$$

avec

$$\overline{x}_{s_{12}} = \frac{1}{n_{12}} \sum_{k \in s_{12}} x_k, \quad \overline{y}_{s_{12}} = \frac{1}{n_{12}} \sum_{k \in s_{12}} y_k.$$

2.3.7 Le calcul des paramètres optimaux $\overline{\boldsymbol{\theta}} = (a, b)' \in \mathbb{R}^2$ pour $Z = t_x - t_y$

Nous nous intéressons maintenant à l'estimation de la différence entre deux totaux. C'est par exemple le cas de l'estimation de l'evolution d'une variable entre deux instants.

Pour $\phi = 1$ et $\psi = -1$, les relations (2.61) et (2.62) nous donnent pour $\hat{Z}$

$$\begin{aligned} \hat{Z} &= a\hat{X}^{1*} + (1 - a)\hat{X}^{12} + b\hat{Y}^{2*} + (-1 - b)\hat{Y}^{12} \\ &= a(\hat{X}^{1*} - \hat{X}^{12}) + b(\hat{Y}^{2*} - \hat{Y}^{12}) + \hat{t}_{H-T} \end{aligned}$$

avec $\hat{t}_{H-T} = \hat{X}^{12} - \hat{Y}^{12}$.
Etudions plus en detail le plan de sondage de type SAS2.

Sondage de Bernoulli bidimensionnel conditionnel à la taille où sondage aléatoire simple sans remise bidimensionnel

Résultat 2.3.14 *: Pour un sondage SAS2 et une variable d'intérêt* $Z = t_x - t_y$*, le paramètre* $\overline{\boldsymbol{\theta}} = (a, b)'$ *a pour valeur optimale*

$$\overline{\boldsymbol{\theta}}_{opt}^{SAS2} = \frac{h_1 h_2}{1-\rho^2 h_1 h_2} \begin{pmatrix} -\{\rho^2 - \rho(1-h_2^{-1})S - h_2^{-1}\} \\ \rho^2 - \rho(1-h_1^{-1})S^{-1} - h_1^{-1} \end{pmatrix}. \quad (2.63)$$

avec une variance optimale :

$$V_{opt}(\hat{Z}) = V(t_{H-T}) - \frac{N}{f_{12}} \frac{S_x S_y h_1 h_2}{1-\rho^2 h_1 h_2} \times \quad (2.64)$$

$$\times \quad \left\{ 2\rho(\rho - S)(\rho - \frac{1}{S}) + (\rho^2 - 1)(h_2^{-1} S + h_1^{-1} \frac{1}{S}) + (h_1^{-1} + h_2^{-1})(S - 2\rho + \frac{1}{S}) \right\}.$$

Remarque : *Pour* $a = -b$*, on obtient* $\hat{Z} = a(\hat{X}^{1*} - \hat{Y}^{2*}) + (1-a)(\hat{X}^{12} - \hat{Y}^{12})$*, c'est à dire l'estimateur composite pour l'evolution proposé par Caron & Ravalet (2000).*

Cas particulier 1 : Si on suppose que $S_x^2 = S_y^2$ et $n_1 = n_2 = n$; alors $h_1 = h_2 = h$, $f_1 = f_2 = f$ et

$$\overline{\boldsymbol{\theta}}_{opt}^{SAS2} = \frac{h}{1-\rho h} \begin{pmatrix} 1-\rho \\ \rho - 1 \end{pmatrix}.$$

Alors, $a_{opt} = -b_{opt} = \frac{h(1-\rho)}{1-\rho h}$. Pour les a_{opt}, b_{opt} ainsi obtenus l'estimateur $\hat{Z}_{opt}$ a la même expression que l'estimateur composite optimal pour l'evolution de la variable $\mathcal{X}$ entre deux instants ($\mathcal{Y}$ représente dans ce cas la variable $\mathcal{X}$ mesurée au deuxième instant) obtenu par Gurney & Daly (1965), Gourieroux & Roy (1978), Caron & Ravalet (2000). Par contre la variance de $\hat{Z}_{opt}$ est différente de celle obtenue dans les travaux mentionnés précédemment parce que nous n'avons pas imposé d'indépendance entre les deux instants et le taux de sondage f_{12} n'est plus supposé négligeable :

$$V_{opt}(\hat{Z}) = \frac{2N}{f} S_x^2 (1-\rho)(\frac{1}{1-\rho h} - f)$$

Cas particulier 2 : Si on suppose seulement que $S_x^2 = S_y^2$, alors

$$\overline{\boldsymbol{\theta}}_{opt}^{SAS2} = \frac{h_1 h_2 (1-\rho)}{1-\rho^2 h_1 h_2} \begin{pmatrix} \rho + h_2^{-1} \\ -(\rho + h_1^{-1}) \end{pmatrix}$$

et la variance optimale de $\hat{Z}$ vaut

$$V_{opt}(\hat{Z}) = \frac{N}{f_{12}} S_x^2 (1-\rho) \left(2(1-f_{12}) - \frac{(1-\rho)(2\rho h_1 h_2 + h_1 + h_2)}{1-\rho^2 h_1 h_2} \right)$$

$$
\begin{aligned}
&= -2(1-\rho)NS_x^2 + (1-\rho)NS_x^2(\frac{1}{f_2} - \frac{1}{f_1}) \times \\
&\times \left(\frac{2}{h_1 - h_2} - \frac{(1-\rho)(2\rho h_1 h_2 + h_1 + h_2)}{1-\rho^2 h_1 h_2}\right).
\end{aligned}
$$

On peut remarquer que pour f_1 et f_2 fixés, l'expression de $V_{opt}(\hat{Z})$ ne dépend que de ρ et h_1 car $h_2 = 1 - \frac{f_1}{f_2} + h_1\frac{f_1}{f_2}$.

Nous allons comparer l'estimateur $\hat{Z}_{opt}$ avec deux estimateurs qui sont en quelque sorte des estimateurs naturels de $t_x - t_y$. Le premier est $\hat{t}_{H-T}$, l'estimateur de Horvitz-Thompson pour $Z = t_x - t_y$ sur l'échantillon intersection s_{12}.

Comparaison avec $V(\hat{t}_{H-T})$

Nous avons le rapport

$$
\frac{V_{opt}(\hat{Z})}{V(\hat{t}_{H-T})} = 1 - \frac{(1-\rho)(2\rho h_1 h_2 + h_1 + h_2)}{2\left(1 - \frac{h_1 - h_2}{\frac{1}{f_2} - \frac{1}{f_1}}\right)(1-\rho^2 h_1 h_2)}.
$$

On trace dans Figure 2.3.7 le gain défini comme $1 - V_{opt}(\hat{Z})/V(\hat{t}_{H-T})$, en fonction du coefficient de ρ pour différentes valeurs du taux de renouvellement h_1.

Un gain maximal est obtenu pour un h_1 proche de l'unité, c'est à dire pour un renouvellement total de l'échantillon. Par contre, si les variables $\mathcal{X}$ et $\mathcal{Y}$ sont corrélées positivement, le gain n'est pas important pour un taux de renouvellement h_1 petit. Pour $\rho = 1$ le gain est zéro quel que soit le taux de renouvellement.

Comparaison avec $\hat{X}^1 - \hat{Y}^2$

Un autre estimateur naturel pour $t_x - t_y$ est la différence des estimateurs de type Horvitz-Thompson basés sur les échantillons s_1 et s_2, plus précisement $\hat{X}^1 - \hat{Y}^2$. Comparons la variance de $\hat{X}^1 - \hat{Y}^2$ avec la variance de $\hat{Z}_{opt}$. Pour un sondage SAS2, $\hat{X}^1 - \hat{Y}^2$ devient :

$$
\begin{aligned}
\hat{X}^1 - \hat{Y}^2 &= h_1\hat{X}^{1*} + (1-h_1)\hat{X}^{12} - h_2\hat{Y}^{2*} - (1-h_2)\hat{Y}^{12} \\
&= h_1(\hat{X}^{1*} - \hat{X}^{12}) - h_2(\hat{Y}^{2*} - \hat{Y}^{12}) + \hat{X}^{12} - \hat{Y}^{12} \quad (2.65)
\end{aligned}
$$

c'est à dire, l'estimateur $\hat{Z}$ pour $a = h_1$ et $b = -h_2$. Cela implique $V_{opt}(\hat{Z}) \leq$ Var $(\hat{X}^1 - \hat{Y}^2)$ avec égalité pour $\rho = 0$.

Il est intéressant d'examiner le rapport entre la variance de $\hat{X}^1 - \hat{Y}^2$ et $V(\hat{t}_{H-T})$; on sait que toutes les deux sont inférieures à $V_{opt}(\hat{Z})$. On a

$$
\frac{V(\hat{t}_{H-T})}{\text{Var }(\hat{X}^1 - \hat{Y}^2)} = \frac{2(1-\rho)(1-f_{12})}{-2\rho(h_1h_2 - h_1 - h_2 + 1 - f_{12}) - h_1 - h_2 + 2(1-f_{12})}.
$$

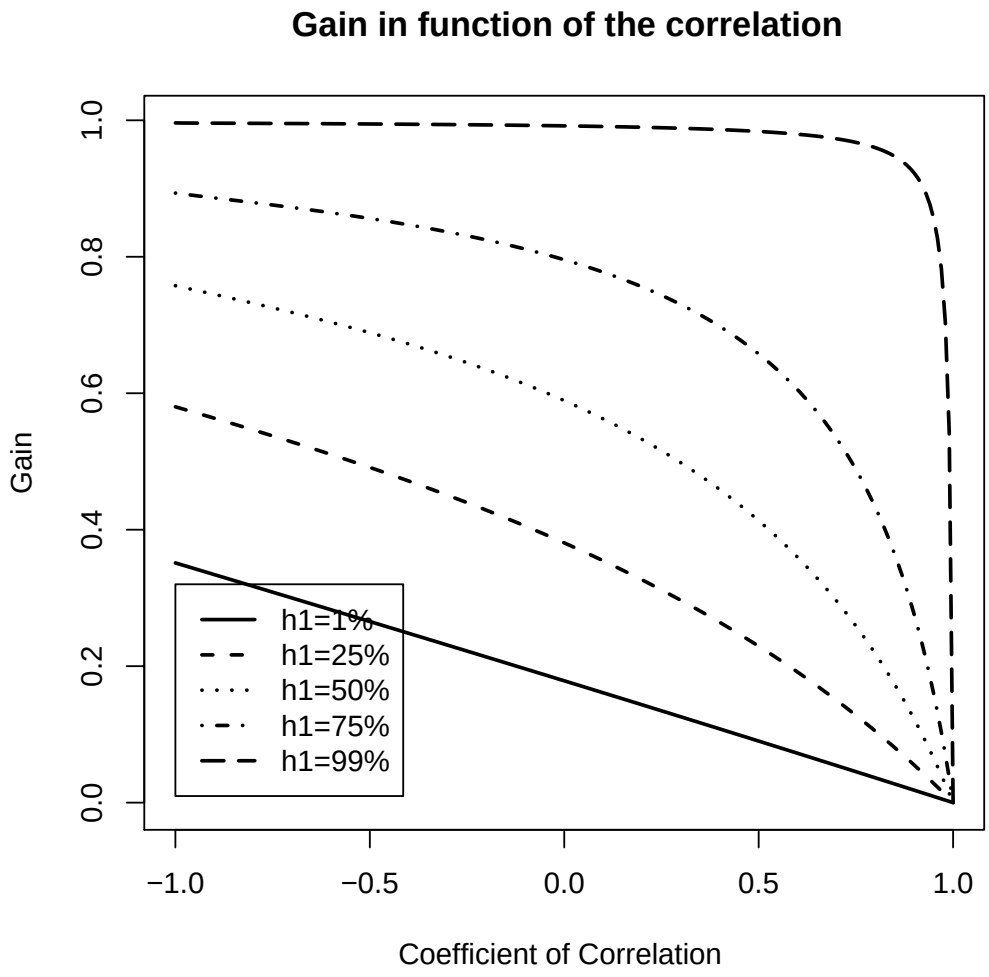

FIG. 2.3 – Gain en fonction du coefficient de corrélation pour différentes valeurs du taux de renouvellement h_1.

Comme fonction de ρ, ce rapport est décroissant avec ρ, il atteint la valeur zéro pour $\rho = 1$ et il est inférieur à l'unité pour $\rho > \frac{h_1+h_2}{2(h_1+h_2-h_1h_2)}$. Pour un coefficient de corrélation supérieur à $\frac{1}{2}$, $\hat{X}^1 - \hat{Y}^2$ est meilleur.

2.3.8 Le calcul de $\overline{\boldsymbol{\theta}} = (a_1, a_2, b_1, b_2)'$ optimal pour sondage aléatoire simple bidimensionnel sans remise conditionnel et $Z = t_x - t_y$

On considère maintenant un autre cas particulier de la situation **A** de la section 2.3.4.

Soient $\mathcal{U}_1$ et $\mathcal{U}_2$ deux partitions de $\mathcal{U}$, $\mathcal{U} = \mathcal{U}_1 \cup \mathcal{U}_2$ et $\mathcal{U}_1 \cap \mathcal{U}_2 = \emptyset$ telles que

$$\begin{cases} a_k = a_1 & k \in \mathcal{U}_1 \\ a_k = a_2 & k \in \mathcal{U}_2 \end{cases} \qquad \begin{cases} b_k = b_1 & k \in \mathcal{U}_1 \\ b_k = b_2 & k \in \mathcal{U}_2 \end{cases}$$

Les matrices $\overline{H}$ et $\overline{J}$ ont les expressions suivantes

$$\overline{H} \;=\; \overline{J} = \begin{pmatrix} \mathbf{1}'_{\boldsymbol{N_1}} & 0 \\ 0 & \mathbf{1}'_{\boldsymbol{N_2}} \end{pmatrix}$$

Alors le paramètre $\boldsymbol{\theta}$ est donné par

$$\boldsymbol{\theta}' \;=\; (a_1, a_2, b_1, b_2) \begin{pmatrix} \mathbf{1}'_{\boldsymbol{N_1}} & 0 & 0 & 0 \\ 0 & \mathbf{1}'_{\boldsymbol{N_2}} & 0 & 0 \\ 0 & 0 & \mathbf{1}'_{\boldsymbol{N_1}} & 0 \\ 0 & 0 & 0 & \mathbf{1}'_{\boldsymbol{N_2}} \end{pmatrix}$$

ce qui revient à chercher $\overline{\boldsymbol{\theta}} = (a_1, a_2, b_1, b_2)'$. L'estimateur $\hat{Z}$ donné par (2.38) devient

$$\hat{Z} = \overline{\boldsymbol{\theta}}' \begin{pmatrix} (\boldsymbol{x}'_{\boldsymbol{N_h}} \boldsymbol{\alpha}_{\boldsymbol{N_h}})_{h=1,2} \\ (\boldsymbol{y}'_{\boldsymbol{N_h}} \boldsymbol{\beta}_{\boldsymbol{N_h}})_{h=1,2} \end{pmatrix} + \hat{t}_{H-T}$$

où pour $h = 1, 2$

$$\begin{cases} \boldsymbol{x}_{\boldsymbol{N_1}} = (x_1, \ldots, x_{N_1})' \in R^{N_1}, \boldsymbol{x}_{\boldsymbol{N_2}} = (x_{N_1+1}, \ldots, x_{N_2})' \in R^{N_2} \\ \boldsymbol{x}'_{\boldsymbol{N_1}} \boldsymbol{\alpha}_{\boldsymbol{N_1}} = \sum_{k \in U_1} x_k \left(\frac{\varepsilon_k^{1*}}{\pi_k^{1*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}} \right), \boldsymbol{x}'_{\boldsymbol{N_2}} \boldsymbol{\alpha}_{\boldsymbol{N_2}} = \sum_{k \in U_2} x_k \left(\frac{\varepsilon_k^{1*}}{\pi_k^{1*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}} \right) \end{cases}$$

Cela implique pour $\hat{Z}$

$$\hat{Z} = (a_1, a_2, b_1, b_2) \begin{pmatrix} \sum_{k \in U_1} x_k \left(\frac{\varepsilon_k^{1*}}{\pi_k^{1*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}} \right) \\ \sum_{k \in U_2} x_k \left(\frac{\varepsilon_k^{1*}}{\pi_k^{1*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}} \right) \\ \sum_{k \in U_1} y_k \left(\frac{\varepsilon_k^{2*}}{\pi_k^{2*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}} \right) \\ \sum_{k \in U_2} y_k \left(\frac{\varepsilon_k^{2*}}{\pi_k^{2*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}} \right) \end{pmatrix} + \hat{t}_{H-T}$$

avec $\hat{t}_{H-T} = \sum_{s_{12}} \frac{x_k - y_k}{\pi_k^{12}}$. Nous allons calculer dans la suite l'expression de $\overline{\boldsymbol{\theta}}_{opt}$ pour le plan de sondage particulier SAS2, $\boldsymbol{s} = (s_1, s_2)$, paramétré par n_{1*}, n_{12} et n_{2*}. Comme déjà mentioné dans l'**exemple** 2, les échantillons unidimensionnels s_1, s_2, s_{12}, s_{1*} et s_{2*} sont des sondages aléatoires simples sans remise dans $\mathcal{U}$. On ne peut pas dire la même chose de leurs intersections avec les deux souspopulations $\mathcal{U}_1$ et $\mathcal{U}_2$ notées

$$s_\oplus^1 = s_\oplus \cap \mathcal{U}_1 \quad s_\oplus^2 = s_\oplus \cap \mathcal{U}_2 \tag{2.66}$$

pour $\oplus \in \{1, 2, 1*, 12, 2*\}$. Par contre, si on considère de nouveau l'aspect conditionnel, la situation devient plus simple.

Soit $n_\oplus^j$ la taille de $s_\oplus^j$ pour $\oplus \in \{1*, 12, 2*\}$ et $j = 1, 2$. Alors, le plan $p(\boldsymbol{s} = (s_1, s_2) | n_\oplus^j, \oplus \in \{1*, 12, 2*\}, j = 1, 2)$ est un tirage aléatoire simple sans remise stratifié sans changement de strate de dimension 2 comme défini par Cotton & Hesse (1992). Plus précisement on note $\boldsymbol{n}^{(1,2)} = (n_\oplus^j)_{\oplus \in \{1*,12,2*\}, j=1,2}$

$$\begin{aligned} p(\boldsymbol{s} = (s_1, s_2) | \boldsymbol{n}^{(1,2)}) &= p^1(\boldsymbol{s}^1 | \boldsymbol{n}^{(1,2)}) p^2(\boldsymbol{s}^2 | \boldsymbol{n}^{(1,2)}) \\ \boldsymbol{s}^1 &= \boldsymbol{s} \cap \mathcal{U}_1 = (s_1 \cap \mathcal{U}_1, s_2 \cap \mathcal{U}_1) \\ \boldsymbol{s}^2 &= \boldsymbol{s} \cap \mathcal{U}_2 = (s_1 \cap \mathcal{U}_2, s_2 \cap \mathcal{U}_2) \end{aligned}$$

avec $p^1(\boldsymbol{s}^1 | \boldsymbol{n}^{(1,2)})$ et $p^2(\boldsymbol{s}^2 | \boldsymbol{n}^{(1,2)})$ des SAS2 dans $\mathcal{U}_1$, respectivement $\mathcal{U}_2$. De plus, les tirages marginaux sont aussi des tirages aléatoires simples sans remise stratifiés (Cotton & Hesse 1992),

$$p_\oplus(s_\oplus | \boldsymbol{n}^{(1,2)}) = p_\oplus^1(s_\oplus^1 | \boldsymbol{n}^{(1,2)}) p_\oplus^2(s_\oplus^2 | \boldsymbol{n}^{(1,2)})$$

avec $s_\oplus^1, s_\oplus^2$ donnés par (2.66) pour $\oplus = 1, 2$. Le fait que $p^1(\boldsymbol{s}^1 | \boldsymbol{n}^{(1,2)})$ est un SAS2 dans $\mathcal{U}_1$ implique que les échantillons unidimensionnels s_{1*}^1, s_{12}^1 et s_{2*}^1 sont des sondages aléatoires simples sans remise de tailles n_{1*}^1, n_{12}^1 et n_{2*}^1. De même, pour s_{1*}^2, s_{12}^2 et s_{2*}^2 dans $\mathcal{U}_2$ avec les tailles n_{1*}^2, n_{12}^2 et n_{2*}^2.

Avec ces considérations, nous allons donner l'expression de $\overline{\boldsymbol{\theta}}_{opt}$. Introduisons les notations suivantes :

$$\hat{X}^{1*,j} = \sum_{k\in U_j} x_k \frac{\varepsilon_k^{1*}}{\pi_k^{1*}}, \quad \hat{X}^{12,j} = \sum_{k\in U_j} x_k \frac{\varepsilon_k^{12}}{\pi_k^{12}} \quad \text{pour} \quad j=1,2$$

$$\hat{Y}^{2*,j} = \sum_{k\in U_j} y_k \frac{\varepsilon_k^{2*}}{\pi_k^{2*}}, \quad \hat{Y}^{12,j} = \sum_{k\in U_j} y_k \frac{\varepsilon_k^{12}}{\pi_k^{12}} \quad \text{pour} \quad j=1,2$$

$$f_{\oplus,j} = \frac{n_{\oplus,j}}{N_j} \quad \text{pour} \quad j=1,2, \oplus \in \{1*, 12, 2*\}$$

Les variances de $\mathcal{X}$, $\mathcal{Y}$ dans $\mathcal{U}_1$ et $\mathcal{U}_2$ sont

$$S^2_{x,\mathcal{U}_j} = \frac{1}{N_j - 1}\sum_{U_j}(x_k - \overline{X}_{U_j})^2 \quad \text{pour} \quad j=1,2$$

$$S^2_{y,\mathcal{U}_j} = \frac{1}{N_j - 1}\sum_{U_j}(y_k - \overline{Y}_{U_j})^2 \quad \text{pour} \quad j=1,2$$

et les covariances entre $\mathcal{X}$, $\mathcal{Y}$ dans $\mathcal{U}_1$ et $\mathcal{U}_2$ sont

$$S^2_{xy,\mathcal{U}_j} = \frac{1}{N_j - 1}\sum_{U_j}(x_k - \overline{X}_{U_j})(y_k - \overline{Y}_{U_j}) \quad \text{pour} \quad j=1,2$$

où $\overline{X}_{U_j} = \frac{1}{N_j}\sum_{U_j} x_k$ et $\overline{Y}_{U_j} = \frac{1}{N_j}\sum_{U_j} y_k$.
Alors,

$$Var(\hat{X}^{\oplus,j}) = N_j^2 \frac{1 - f_{\oplus,j}}{n_{\oplus,j}} S^2_{x,\mathcal{U}_j} \quad \text{pour} \quad j=1,2, \quad \text{et} \quad \oplus \in \{1*, 12\}$$

$$Var(\hat{Y}^{\oplus,j}) = N_j^2 \frac{1 - f_{\oplus,j}}{n_{\oplus,j}} S^2_{y,\mathcal{U}_j} \quad \text{pour} \quad j=1,2, \quad \text{et} \quad \oplus \in \{2*, 12\}$$

$$Cov(\hat{X}^{1*,j}, \hat{X}^{12,j}) = -N_j S^2_{x,\mathcal{U}_j}$$

$$Cov(\hat{Y}^{2*,j}, \hat{Y}^{12,j}) = -N_j S^2_{y,\mathcal{U}_j}$$

$$Cov(\hat{X}^{12,j}, \hat{Y}^{12,j}) = N_j \frac{1 - f_{12,j}}{f_{12,j}} S^2_{xy,\mathcal{U}_j}$$

pour $j=1,2$. Finalement,

$$Cov(\hat{X}^{1*,j}, \hat{Y}^{2*,j}) \;=\; Cov(\hat{X}^{12,j}, \hat{Y}^{2*,j}) = Cov(\hat{X}^{1*,j}, \hat{Y}^{12,j}) = -N_j S^2_{xy,\mathcal{U}_j} \quad j=1,2$$

et en raison de la stratification de la population $\mathcal{U}$ dans $\mathcal{U}_1$ et $\mathcal{U}_2$, on a

$$Cov(\hat{X}^{\oplus,j}, \hat{Y}^{\otimes,k}) \;=\; 0 \quad \text{pour} \quad j \neq k = 1,2 \quad \text{et} \quad \oplus \in \{2*, 12\}, \otimes \in \{12, 2*\}$$

Ainsi, on obtient la matrice de variance-covariance suivante

$$\begin{aligned}\overline{\mathbf{\Gamma}} &= Var\begin{pmatrix}(\hat{X}^{1*,j}-\hat{X}^{12,j})_{j=1,2}\\(\hat{Y}^{2*,j}-\hat{Y}^{12,j})_{j=1,2}\end{pmatrix}\\ &= \begin{pmatrix} N_1 S^2_{x,\mathcal{U}_1}\frac{f_{1,1}}{f_{12,1}f_{1*,1}} & 0 & \frac{N_1}{f_{12,1}}S^2_{xy,U_1} & 0\\ 0 & N_2 S^2_{x,\mathcal{U}_2}\frac{f_{1,2}}{f_{12,2}f_{1*,2}} & 0 & \frac{N_2}{f_{12,2}}S^2_{xy,U_2}\\ \frac{N_1}{f_{12,1}}S^2_{xy,U_1} & 0 & N_1 S^2_{y,\mathcal{U}_1}\frac{f_{2,1}}{f_{12,1}f_{2*,1}} & 0\\ 0 & \frac{N_2}{f_{12,2}}S^2_{xy,U_2} & 0 & N_2 S^2_{y,\mathcal{U}_2}\frac{f_{2,2}}{f_{12,2}f_{2*,2}}\end{pmatrix}\end{aligned}$$

où $f_{1,j} = f_{12,j} + f_{1*,j} = \dfrac{n_{1,j}}{N_j}$ et $f_{2,j} = f_{12,j} + f_{2*,j} = \dfrac{n_{2,j}}{N_j}$ pour $j = 1, 2$.
Par ailleurs,

$$\overline{\mathbf{\Gamma}}^{-1} = \begin{pmatrix}\mathbf{\Gamma}_{11} & \mathbf{\Gamma}_{12}\\ \mathbf{\Gamma}_{21} & \mathbf{\Gamma}_{22}\end{pmatrix}$$

avec

$$\mathbf{\Gamma}_{11} = \begin{pmatrix}\frac{f_{12,1}}{N_1}\frac{h_{1,1}}{S^2_{x,U_1}}\frac{1}{1-h_{1,1}h_{2,1}(\rho_{xy,1})^2} & 0\\ 0 & \frac{f_{12,2}}{N_2}\frac{h_{1,2}}{S^2_{x,U_2}}\frac{1}{1-h_{1,2}h_{2,2}(\rho_{xy,2})^2}\end{pmatrix}$$

$$\mathbf{\Gamma}_{22} = \begin{pmatrix}\frac{f_{12,1}}{N_1}\frac{h_{2,1}}{S^2_{y,U_1}}\frac{1}{1-h_{1,1}h_{2,1}(\rho_{xy,1})^2} & 0\\ 0 & \frac{f_{12,2}}{N_2}\frac{h_{2,2}}{S^2_{y,U_2}}\frac{1}{1-h_{1,2}h_{2,2}(\rho_{xy,2})^2}\end{pmatrix}$$

$$\mathbf{\Gamma}_{12} = \begin{pmatrix}-\frac{f_{12,1}}{N_1}\frac{h_{1,1}h_{2,1}}{S_{x,U_1}S_{y,U_1}}\frac{\rho_{xy,1}}{1-h_{1,1}h_{2,1}(\rho_{xy,1})^2} & 0\\ 0 & -\frac{f_{12,2}}{N_2}\frac{h_{1,2}h_{2,2}}{S_{x,U_2}S_{y,U_2}}\frac{\rho_{xy,2}}{1-h_{1,2}h_{2,2}(\rho_{xy,2})^2}\end{pmatrix}$$

et $\mathbf{\Gamma}_{12} = \mathbf{\Gamma}'_{21}$.
Enfin

$$\overline{\gamma} = Cov\begin{pmatrix}(\hat{X}^{1*,j}-\hat{X}^{12,j})_{j=1,2} & \\ & \hat{t}_{H-T}\\ (\hat{Y}^{2*,j}-\hat{Y}^{12,j})_{j=1,2} & \end{pmatrix}$$

Puisque les estimateurs $\hat{X}^{12}$ et $\hat{Y}^{12}$ peuvent s'écrire de la manière suivante $\hat{X}^{12} = \hat{X}^{12,1} + \hat{X}^{12,2}$, $\hat{Y}^{12} = \hat{Y}^{12,1} + \hat{Y}^{12,2}$ et $\hat{t}_{H-T} = \hat{X}^{12} - \hat{Y}^{12}$, cela

nous donne

$$\overline{\gamma} = \begin{pmatrix} \frac{N_1}{f_{12,1}}(S^2_{xy,U_1} - S^2_{x,U_1}) \\ \frac{N_2}{f_{12,2}}(S^2_{xy,U_2} - S^2_{x,U_2}) \\ \frac{N_1}{f_{12,1}}(S^2_{y,U_1} - S^2_{xy,U_1}) \\ \frac{N_2}{f_{12,2}}(S^2_{y,U_2} - S^2_{xy,U_2}) \end{pmatrix}$$

$\overline{\boldsymbol{\theta}}_{opt} = -\overline{\boldsymbol{\Gamma}}^{-1}\overline{\gamma} = (a_{1,opt}, a_{2,opt}, b_{1,opt}, b_{2,opt})'$. Pour

$$\rho_{xy,1} = \text{coef de corrélation entre} \quad \mathcal{X}, \mathcal{Y} \quad \text{dans} \quad U_1$$

$$\rho_{xy,2} = \text{coef de corrélation entre} \quad \mathcal{X}, \mathcal{Y} \quad \text{dans} \quad U_2.$$

le paramètre $a_{1,opt}$ a l'expression

$$a_{1,opt} = \frac{-h_{1,1}h_{2,1}\left[(\rho_{xy,1})^2 + \rho_{xy,1}((h_{2,1})^{-1} - 1)\frac{S_{y,U_1}}{S_{x,U_1}} - (h_{2,1})^{-1}\right]}{1 - h_{1,1}h_{2,1}(\rho_{xy,1})^2}$$

avec

$$h_{1,1} = \frac{n_{1*,1}}{n_{1,1}}, \quad h_{1,2} = \frac{n_{1*,2}}{n_{1,2}}$$

les taux de renouvellement de s_1 dans $\mathcal{U}_1$, respectivement $\mathcal{U}_2$ et

$$h_{2,1} = \frac{n_{2*,1}}{n_{2,1}}, \quad h_{2,2} = \frac{n_{2*,2}}{n_{2,2}}$$

les taux de renouvellement de s_2 dans $\mathcal{U}_1$, respectivement $\mathcal{U}_2$. On déduit de la même manière les autres composants de $\overline{\boldsymbol{\theta}}_{opt}$. Remarquons que l'expression de $a_{1,opt}$ ressemble à celle de a^{SAS2}_{opt} donnée par 2.63. Le $\overline{\boldsymbol{\theta}}_{opt}$ ainsi obtenu ne peut pas être déduit mais il peut être estimé par la méthode décrite dans la remarque 2.3.4.

2.3.9 Cas général $Z = \phi t_x + \psi t_y + \delta t_v + \chi t_t$

Une situation encore plus générale que celle considérée dans la section précédente est le cas de quatre variables d'intérêt et deux échantillons. On considère $\mathcal{X}$ et $\mathcal{Y}$ définies comme précédemment, la variable $\mathcal{V}$ du total $t_v = \sum_U v_k$ est mesurée sur $s_{12} = s_1 \cap s_2$ et la variable $\mathcal{T}$ du total $t_t = \sum_U t_k$ est mesurée sur $s_1 \cup s_2$.

On veut estimer une combinaison linéaire de leur totaux,

$$Z = \phi t_x + \psi t_y + \delta t_v + \chi t_t$$

avec ϕ, ψ, δ et χ des constantes.

Nous avons déjà mené la discussion sur comment estimer ϕt_x et ψt_y. Ils sont estimés par

$$\phi \hat{t}_x = \sum_{s_{1*}} w_k^{1*,x} x_k + \sum_{s_{12}} w_k^{12,x} x_k$$
$$\psi \hat{t}_y = \sum_{s_{2*}} w_k^{2*,y} y_k + \sum_{s_{12}} w_k^{12,y} y_k$$

où les poids vérifient les équations

$$w_k^{1*,x} \pi_k^{1*} + w_k^{12,x} \pi_k^{12} = \phi$$
$$w_k^{2*,y} \pi_k^{2*} + w_k^{12,y} \pi_k^{12} = \psi.$$

Quant à la variable $\mathcal{V}$, elle est observée sur l'espace intersection ; alors, δt_v est estimé par

$$\delta \hat{t}_v = \sum_{s_{12}} \frac{\delta}{\pi_k^{12}} v_k.$$

La variable $\mathcal{T}$ est mesurée sur $s_1 \cup s_2$, alors elle est estimée par

$$\chi \hat{t}_t = \sum_{s_{1*}} w_k^{1*,t} t_k + \sum_{s_{12}} w_k^{12,t} t_k + \sum_{s_{2*}} w_k^{2*,t} t_k$$

avec les poids $w_k^{1*,t}$, $w_k^{12,t}$ et $w_k^{2*,t}$ qui vérifient

$$w_k^{1*,t} \pi_k^{1*} + w_k^{12,t} \pi_k^{12} + w_k^{2*,t} \pi_k^{2*} = \chi$$

Alors Z est estimé par

$$\hat{Z} = \phi \hat{t}_x + \psi \hat{t}_y + \delta \hat{t}_v + \chi \hat{t}_t.$$

Considérons la paramétrisation suivante pour les poids

$$\begin{cases} w_k^{1*,x} = \dfrac{a_k}{\pi_k^{1*}} \quad ; \quad w_k^{12,x} = \dfrac{\phi - a_k}{\pi_k^{12}} \\ w_k^{2*,y} = \dfrac{b_k}{\pi_k^{2*}} \quad ; \quad w_k^{12,y} = \dfrac{\psi - b_k}{\pi_k^{12}} \\ w_k^{12,v} = \dfrac{\delta}{\pi_k^{12}} \\ w_k^{1*,t} = \dfrac{c_k}{\pi_k^{1*}} \quad ; \quad w_k^{2*,t} = \dfrac{d_k}{\pi_k^{2*}} \quad ; \quad w_k^{12,t} = \dfrac{\chi - c_k - d_k}{\pi_k^{12}}. \end{cases} \tag{2.67}$$

Elle vérifie la condition de sans biais pour $\hat{Z}$. On peut alors écrire $\hat{Z}$ sous la forme matricielle suivante

$$\hat{Z} = \boldsymbol{\theta}' \left(\begin{array}{c|c} \text{diag}x & 0 \\ \hline 0 & \text{diag}y \\ \hline \text{diag}t & 0 \\ \hline 0 & \text{diag}t \end{array} \right) \left(\begin{array}{c} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{array} \right) + \hat{t}_{H-T} \tag{2.68}$$

où $\boldsymbol{\theta} = ((a_k)_{k=1}^N, (b_k)_{k=1}^N, (c_k)_{k=1}^N, (d_k)_{k=1}^N)' \in \mathbb{R}^{4N}$ et $\boldsymbol{\alpha}, \boldsymbol{\beta}$ sont des vecteurs, de dimension N, d'éléments $\alpha_k = \left(\frac{\varepsilon_k^{1*}}{\pi_k^{1*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}} \right), \quad \beta_k = \left(\frac{\varepsilon_k^{2*}}{\pi_k^{2*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}} \right)$ pour tous $k \in U$ et $\hat{t}_{H-T} = \sum_U \frac{\phi x_k + \psi y_k + \delta v_k + \chi t_k}{\pi_k^{12}} \varepsilon_k^{12}$. Si on note

$$\mathbf{H} = \left(\begin{array}{c|c} \text{diag}x & 0 \\ \hline 0 & \text{diag}y \\ \hline \text{diag}t & 0 \\ \hline 0 & \text{diag}t \end{array} \right)$$

alors $\boldsymbol{\theta}_{opt} = -\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}$ avec

$$\left\{ \begin{array}{rcl} \boldsymbol{\Gamma} & = & \mathbf{H} \times \text{Var}\left(\left(\begin{array}{c} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{array} \right) \right) \times \mathbf{H}' \\ \boldsymbol{\gamma} & = & \mathbf{H} \times \text{Cov}\left(\left(\begin{array}{c} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{array} \right), \hat{t}_{H-T} \right). \end{array} \right. \tag{2.69}$$

Cas particuliers

1. Si $\mathcal{V}$ et $\mathcal{T}$ ne sont pas disponibles, on estime $Z = \phi t_x + \psi t_y$ par $\hat{Z}$ obtenu de (2.68) après avoir éliminé les deux dernières lignes dans la matrice $\mathbf{H}$ et les termes qui contiennent t_k et v_k dans $\hat{t}_{H-T}$. L'expression (2.27) de $\hat{Z}$ obtenu dans la section 5 est ainsi retrouvée.

2. Si $\mathcal{V}$ et $\mathcal{T}$ sont disponibles, alors en posant $\delta = \chi = 0$ on estime $Z = \phi t_x + \psi t_y$ en bénéficiant de la connaissance de $\mathcal{V}$ et $\mathcal{T}$ qui peut être considérée comme une information auxiliaire.

$$\hat{Z} = \boldsymbol{\theta}' \times \mathbf{H} + \hat{t}_{H-T}$$

avec le même paramètre $\boldsymbol{\theta}$ de dimension $4N$ et la même matrice $\mathbf{H}$ mais cette fois-ci $\hat{t}_{H-T} = \phi \hat{X}^{12} + \psi \hat{Y}^{12}$.

3. Si on ne dispose que des variables $\mathcal{X}$ et $\mathcal{T}$ alors $Z = \phi t_x + \chi t_t$ et

$$\hat{Z} = \boldsymbol{\theta}' \left(\begin{array}{c|c} \text{diag}x & 0 \\ \hline \text{diag}t & 0 \\ \hline 0 & \text{diag}t \end{array} \right) \left(\begin{array}{c} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{array} \right) + \hat{t}_{H-T} \tag{2.70}$$

où $\boldsymbol{\theta}' = ((a_k)_{k=1}^N, (c_k)_{k=1}^N, (d_k)_{k=1}^N)' \in \mathbb{R}^{3N}$ et $\hat{t}_{H-T} = \phi\hat{X}^{12} + \chi\hat{T}^{12}$.

4. Si on dispose de $\mathcal{X}$, de $\mathcal{Y}$ et de $\mathcal{V}$ où de $\mathcal{X}$, de $\mathcal{Y}$ et de $\mathcal{T}$ alors la matrice $\mathbf{H}$ ne change pas et $\hat{t}_{H-T} = \phi\hat{X}^{12} + \psi\hat{Y}^{12} + \delta\hat{V}^{12}$ ou $\hat{t}_{H-T} = \phi\hat{X}^{12} + \psi\hat{Y}^{12} + \chi\hat{T}^{12}$.

5. Si on dispose que de $\mathcal{X}$ et $\mathcal{Y}$, alors en posant $\phi = \delta = \chi = 0$ et $\psi = 1$ le paramètre d'intérêt Z est exactement le total de la variable $\mathcal{Y}$. On est dans le cas particulier de l'estimation d'un total quand on dispose de l'information auxiliaire.
L'estimateur $\hat{Z}$ est obtenu de (2.68) pour $\hat{t}_{H-T} = \hat{Y}_{12}$ et

$$\mathbf{H} = \left(\begin{array}{c|c} \mathrm{diag}x & 0 \\ \hline 0 & \mathrm{diag}y \end{array}\right).$$

On présente plus en détail maintenant le cas particulier $a_k = a$ et $b_k = b$ pour tous $k \in U$. L'estimateur $\hat{Z}$ a l'expression suivante

$$\hat{Z} = a(\hat{X}^{1*} - \hat{X}^{12}) + b(\hat{Y}^{2*} - \hat{Y}^{12}) + \hat{Y}^{12} \tag{2.71}$$

et pour un sondage BE2 on a,

$$\boldsymbol{\theta}_{opt} \quad = \quad \frac{h_1 h_2}{1 - \rho^2 h_1 h_2} \begin{pmatrix} h_2^{-1}\rho S - \rho S \\ -\rho^2 + h_1^{-1} \end{pmatrix} \tag{2.72}$$

ce qui donne comme variance optimale

$$V_{opt}(\hat{Z}) = \frac{N}{f_{12}} S_y^2 \left((1 - f_{12}) - \frac{h_1 h_2}{1 - \rho^2 h_1 h_2} (h_2^{-1}\rho^2 - 2\rho^2 + h_1^{-1}) \right). \tag{2.73}$$

Remarque 2.3.5 *: $\boldsymbol{\theta}_{opt}$ donné par (2.72) est différent de (2.62) quand on a voulu estimer l'évolution.*

On peut écrire a_{opt} et b_{opt} de la façon suivante

$$a_{opt} = \rho S \frac{h_1(1 - h_2)}{1 - \rho^2 h_1 h_2},$$

$$b_{opt} = 1 - \varphi \quad \text{avec} \quad \varphi = \frac{1 - h_2}{1 - \rho^2 h_1 h_2}.$$

Puisque $\hat{X}^{1*} = h_1^{-1}\hat{X}^1 - (h_1^{-1} - 1)\hat{X}^{12}$, $\hat{Z}$ devient :

$$\hat{Z} = \varphi(\hat{Y}^{12} + B(\hat{X}^1 - \hat{X}^{12})) + (1 - \varphi)\hat{Y}^{2*}$$

où $B = \rho\frac{S_y}{S_x}$ peut être vu comme le coefficient de régression de la variable $\mathcal{Y}$ sur $\mathcal{X}$. Dans le cas particulier $h_1 = h_2 = h$, on obtient $a_{opt} = \frac{\rho S(1-h)h}{1-\rho^2 h^2}$ et $b_{opt} = \frac{h - h^2\rho^2}{1-\rho^2 h^2}$ c'est à dire les mêmes valeurs obtenues par Särndal *et al.* (1992).

2.4 Estimation non-linéaire

Nous allons décrire dans cette section comment un paramètre nonlinéaire Φ qui dépend de variables mesurées sur des échantillons différents, peut être approximé sous des conditiones générales par une statistique linéaire de variables artificielles dont la variance approxime la variance de Φ. La technique de linéarisation utilisée est celle par la fonction d'influence introduite par Deville (1999) qu'on va étendre proprement à deux échantillons.

2.4.1 Cas d'un unique échantillon

Introduisons d'abord les notations puis rappelons les principaux résultats obtenus par Deville (1999) dans le cas d'un seul échantillon.

Soit la population

$$U = \{1, \ldots, k, \ldots, N\}$$

et pour chaque individu $k \in U$ on associe un point $z_k \in \mathbf{R}^p$, caractéristique de l'individu et des valeurs prises par les p variables auxilaires considérées. Attention, il ne s'agit pas en général des valeurs prises par les variables auxiliaires. En effet, ce vecteur permet d'identifier chaque individu dans la population, ainsi deux individus k et l différents ont deux "coordonnées" différentes, $z_k \neq z_l$.

Nous considérons une mesure M définie sur $\mathbf{R}$ et à valeurs dans $\mathbf{R}$ telle que

$$\begin{aligned} M(z) &= 1 \quad \text{si } \exists k \text{ tel que } z = z_k && (2.74) \\ &\quad 0 \quad \text{sinon.} && (2.75) \end{aligned}$$

Cette mesure prend la valeur 1 pour tout élément de la population et vaut zéro ailleurs. La plupart des quantités d'intérêt étudiées dans les sondages peuvent s'écrire comme une fonctionnelle Φ de M.

Par exemple,

– le total de la variable $\mathcal{Y}$ est egal à $\int \mathcal{Y} dM(z)$,

– le ratio du total de deux variables $\mathcal{Y}$ and $\mathcal{X}$ s'écrit

$$\Phi(M) = \frac{\int \mathcal{Y} dM(z)}{\int \mathcal{X} dM(z)}.$$

Des indices plus complexes tels que l'indice de Gini peuvent également s'écrire sous la forme d'une fonctionnelle de la mesure M. On pourra se reporter à Deville (1999, section 12) pour une liste plus complète.

Supposons maintenant que notre paramètre d'intérêt peut s'écrire sous la forme d'une fonctionnelle Φ de la mesure M et définissons la fonction

d'influence de $\Phi(M)$, quand elle existe, par

$$I\Phi(M,x) \quad = \quad \lim_{h\to 0} \frac{1}{h}\left(\Phi(M+h\delta_x) - \Phi(M)\right) \tag{2.76}$$

où δ_x est la fonction Dirac (masse unité) au point $x \in \mathbf{R}$. Les principales propriétés de la fonction d'influence, qui est en fait la différentielle au sens de Gateaux de la masse de Dirac au point x, sont données dans Deville (1999). Elle vérifie, grosso modo, les propriétés de l'opérateur de "dérivation". Si Φ est la somme de deux fonctionnelles, $\Phi(M) = \Phi_1(M) + \Phi_2(M)$ alors

$$I\Phi(M,x) \quad = \quad I\Phi_1 + I\Phi_2.$$

Si Φ se décompose en $\Phi(M) = \Phi_1(M)\Phi_2(M)$ alors elle vérifie

$$I\Phi(M,x) \quad = \quad \Phi_1 I\Phi_2 + \Phi_2 I\Phi_1.$$

Construisons un estimateur $\widehat{M}$ de la mesure M à partir de l'échantillon s sélectionné dans la population U selon un plan de sondage $p(\cdot)$. L'estimateur $\widehat{M}$ est obtenu de la manière suivante :

$$\widehat{M}(z_k) \quad = \quad w_k \quad \text{ si } k \in s \quad \text{ et zéro ailleurs.} \tag{2.77}$$

Cela signifie que $\widehat{M}(z_k) = w_k \varepsilon_k^s$, que nous notons pour simplifier par w_k^s; $\varepsilon_k^s = \mathbf{1}_{\{k\in s\}}$.

Supposons que la fonctionnelle Φ est de degré α, c'est à dire que $N^{-\alpha}\Phi$ existe quand N tend vers l'infini. De manière évidente, si Φ représente un total alors $\alpha = 1$ et si Φ est un ratio alors $\alpha = 0$.

Définissons maintenant les *variables linéarisées*, $u_k = I\Phi(M, x_k)$. Deville (1999) a montré que l'estimateur par substitution, $\Phi(\widehat{M})$, de la fonctionnelle $\Phi(M)$, est linéarisable, c'est à dire qu'il vérifie

$$\sqrt{n}N^{-\alpha}\left(\Phi(M) - \Phi(\widehat{M})\right) = \frac{\sqrt{n}}{N}\sum_U u_k(w_k^s - 1) + o_p(1).$$

En d'autres termes, ceci signifie que la variance de $\Phi(M)$ peut être approximée et estimée à l'aide des variables linéarisées u_k. Ce résultat repose sur les propriétés de la fonction d'influence et sur le fait que

$$d(M/N, \widehat{M}/N) = O_p(n^{-1/2})$$

où la distance $d(M_1, M_2)$ entre deux mesures M_1 et M_2 vérifie :

$$\lim d(M_1, M_2) = 0 \quad \Longleftrightarrow \quad \lim N^{-1}\left(\int \mathcal{Y}\, dM_1(z) - \int \mathcal{Y}\, dM_2(z)\right) = 0,$$

pour toute variable $\mathcal{Y}$.

L'estimation d'un ratio

On a vu que le ratio entre le total de $\mathcal{Y}$ et $\mathcal{X}$ peut s'écrire comme une fonctionnelle de M,

$$R = \frac{\sum_U y_k}{\sum_U x_k} = \frac{\int \mathcal{Y} dM(z)}{\int \mathcal{X} dM(z)}$$

c'est à dire $R = \Phi(Q_y(M), S_x(M))$ avec

$$Q_y(M) = \int \mathcal{Y} dM(z),$$

$$S_x(M) = \int \mathcal{X} dM(z)$$

et

$$\Phi = \frac{Q_y}{S_x}.$$

La fonction d'influence $I\Phi$ est donnée par

$$I\Phi = \Phi'_{Q_y} IQ_y + \Phi'_{S_x} IS_x$$

et $IQ_y = y_k$ et $IS_x = x_k$. Alors, $\hat{R} = \frac{\sum_s y_k}{\sum_s x_k}$, l'estimateur par substitution de R est approximé par

$$\hat{R} - R \simeq \frac{1}{N} \sum_U (u_k w_k^s - u_k)$$

où $u_k = I\Phi$ et $w_k^s = w_k \varepsilon_k^s$.

2.4.2 La mesure M est estimée à partir de plusieurs échantillons

On considère dans la suite les paramètres d'intérêt de la forme suivante

$$\Phi(Q_y(M), S_x(M))$$

pour une fonctionnelle Φ et où $Q_y(M)$ et $S_x(M)$ sont des totaux sur la population U pour les variables $\mathcal{Y}$ et respectivement $\mathcal{X}$. La variable $\mathcal{X}$ est mesurée sur un échantillon s_1 et la variable $\mathcal{Y}$ sur un échantillon s_2, différent de s_1. Notre intention est d'obtenir une variable linéarisée pour

$$\Phi(Q_y(M), S_x(M))$$

avec des poids qui dépendent à la fois de s_1 et de s_2 en utilisant la technique de linéarisation par la fonction d'influence présentée dans le paragraphe

précédent.
La fonction d'influence de $\Phi(Q_y(M), S_x(M))$ est égale à

$$I\Phi = \Phi'_{Q_y} IQ_y + \Phi'_{S_x} IS_x$$

et elle ne dépend pas des échantillons sur lequels les variables $\mathcal{X}$ et $\mathcal{Y}$ sont mesurées. Alors, l'estimateur par substitution $\hat{\Phi}(Q_y(\hat{M}), S_x(\hat{M}))$ est approximé par

$$N^{-\alpha}(\hat{\Phi} - \Phi) \simeq \frac{1}{N}\left(\sum_U I\Phi w_k^s - \sum_U I\Phi\right)$$

avec les poids w_k^s à determiner dans la suite. De plus, la variance de $N^{-\alpha}(\hat{\Phi} - \Phi)$ est approximée par la variance de $\frac{1}{N}\sum_U I\Phi w_k^s$.
Si on tient compte de l'expression de $I\Phi$, on obtient

$$\sum_U I\Phi w_k^s = \sum_U (\Phi'_{Q_y} IQ_y + \Phi'_{S_x} IS_x) w_k^s$$

avec IQ_y connu sur s_2 et IS_x connu sur s_1. De plus, Φ'_{Q_y} et Φ'_{S_x} sont des fonctions des totaux et par conséquent, elles ne dépendent pas de s_1 et de s_2. On peut estimer sans biais $\Phi'_{Q_y} \sum_U IQ_y w_k^s$ avec $IQ_y = y_k$ par

$$\sum_{s_{2*}} y_k w_k^{2*} + \sum_{s_{12}} y_k w_k^{12,y}$$

avec des poids qui vérifient

$$w_k^{2*}\pi_k^{2*} + w_k^{12,y}\pi_k^{12} = \Phi'_{Q_y}.$$

On procède de la même manière avec $\Phi'_{S_x} \sum_U IS_x w_k^s$ pour $IS_x = x_k$. Cette quantité va être estimée sans biais par

$$\sum_{s_{1*}} x_k w_k^{1*} + \sum_{s_{12}} x_k w_k^{12,x}$$

avec des poids qui vérifient

$$w_k^{1*}\pi_k^{1*} + w_k^{12,x}\pi_k^{12} = \Phi'_{S_x}.$$

Alors, pour

$$\begin{aligned} w_k^{1*} &= \frac{a_k}{\pi_k^{1*}}; \quad w_k^{12,x} = \frac{\Phi'_{S_x} - a_k}{\pi_k^{12}} \\ w_k^{2*} &= \frac{b_k}{\pi_k^{2*}}; \quad w_k^{12,y} = \frac{\Phi'_{Q_y} - b_k}{\pi_k^{12}} \end{aligned}$$

l'estimateur par substitution $\hat{\Phi}$ est approximé par

$$N^{-\alpha}(\hat{\Phi}-\Phi) \simeq \frac{1}{N}\left(\sum_{s_{1*}}\frac{x_k a_k}{\pi_k^{1*}}-\sum_{s_{12}}\frac{x_k a_k}{\pi_k^{12}}\right)+\frac{1}{N}\left(\sum_{s_{2*}}\frac{y_k b_k}{\pi_k^{2*}}-\sum_{s_{12}}\frac{y_k b_k}{\pi_k^{12}}\right) +\frac{1}{N}\hat{t}_{H-T}-\frac{1}{N}\sum_U I\Phi \quad (2.78)$$

où

$$\hat{t}_{H-T}=\Phi'_{S_x}\sum_{s_{12}}\frac{x_k}{\pi_k^{12}}+\Phi'_{Q_y}\sum_{s_{12}}\frac{y_k}{\pi_k^{12}}=\Phi'_{S_x}\hat{X}^{12}+\Phi'_{Q_y}\hat{Y}^{12}.$$

On applique alors la théorie développée précédemment pour déterminer les quantités a_k et b_k sous la condition qu'elles minimisent la variance de la quantité (2.78) . Pour a_k et b_k optimaux, la variance de $N^{-\alpha}(\hat{\Phi}-\Phi)$ va être approximée par la variance optimale de

$$\frac{1}{N}\boldsymbol{\theta}'\begin{pmatrix}\text{diag}x & 0\\ 0 & \text{diag}y\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}+\frac{1}{N}\hat{t}_{H-T}$$

où $\boldsymbol{\theta}'=((a_k)_{k=1}^N,(b_k)_{k=1}^N)$.
Des plans de sondage définis dans la première partie peuvent alors être utilisés et on peut déterminer dans chaque cas les paramètres optimaux pour la variable linéarisée.

On donne dans la suite un exemple de statistique complexe estimée sur deux échantillons différents.

Application à l'estimation d'un ratio

On applique les résultats trouvés ci-dessus pour estimer un ratio $R=\frac{t_y}{t_x}$ quand $\mathcal{X}$ est mesurée sur s_1 et $\mathcal{Y}$ sur s_2. Cette situation peut apparaitre par exemple en présence de la nonréponse. Par exemple, un échantillon s est sélectionné dans U et la nonréponse se produit différemment pour les variables $\mathcal{X}$ et $\mathcal{Y}$. Plus précisément, nous avons $s_1 \subset s$ comme ensemble de répondants pour $\mathcal{X}$ et $s_2 \subset s$ pour $\mathcal{Y}$ avec une intersection qui n'est pas négligeable. On suppose comme modèle de nonréponse le modèle de Bernoulli bidimensionnel à la taille largément décrît dans l'**exemple** 2.
Appliquons la technique de linéarisation presentée ci-dessus. La linéarisée de R ne dépend pas de l'échantillon où on connait les variables d'intérêt et elle a l'expression suivante

$$r_k=-\frac{t_y}{t_x^2}x_k+\frac{1}{t_x}y_k$$

quand la fonction d'influence de t_x, $It_x = x_k$ est connue sur s_1 et celle de t_y, $It_y = y_k$ est connue sur s_2. On a donc :

$$\hat{R} - R \;=\; \frac{1}{N}\left\{\sum_U \left(-\frac{t_y}{t_x^2}x_k + \frac{1}{t_x}y_k\right) w_k^s - Z\right\} + o(1)$$

où $Z = \sum_U(\phi x_k + \psi y_k)$ est le total de la variable linéarisée r_k pour $\phi = -\frac{t_y}{t_x^2}$ et $\psi = \frac{1}{t_x}$. Les poids w_k^s sont déterminés selon le procédé décrit auparavant. Par conséquent, l'estimateur par substitution $\hat{R}$ est approximé par

$$\hat{R}_{opt} - R \;\simeq\; \frac{1}{N}\boldsymbol{\theta'}_{opt}\begin{pmatrix} \text{diag}x & 0 \\ 0 & \text{diag}y \end{pmatrix}\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} + \frac{1}{N}\hat{t}_{H-T} - \frac{1}{N}Z$$

avec $\boldsymbol{\theta'}_{opt} = ((a_{k,opt})_{k=1}^N, (b_{k,opt})_{k=1}^N)$. Pour $a_k = a$ et $b_k = b$ pour tous les individus k dans la population U, on obtient

$$\hat{R}_{opt} - R \;\simeq\; a_{opt}(\hat{X}^{1*} - \hat{X}^{12}) + b_{opt}(\hat{Y}^{2*} - \hat{Y}^{12}) + \hat{t}_{H-T} - Z$$

avec

- $\hat{t}_{H-T} = \phi\hat{X}^{12} + \psi\hat{Y}^{12}$;
- a_{opt}, b_{opt} les paramètres optimaux ;
- $Z = \sum_U(\phi x_k + \psi y_k)$ est le total de la variable linéarisée r_k.

Dans ce cas, l'estimateur par substitution de R a l'expression :

$$\hat{R}_{opt} = \frac{b_{opt}\hat{Y}^{2*} + (\psi - b_{opt})\hat{Y}^{12}}{a_{opt}\hat{X}^{1*} + (\phi - a_{opt})\hat{X}^{12}}.$$

Comparaison avec des estimateurs naturels de $\hat{R}$

1. On peut estimer R en ne considérant que l'intersection s_{12}. En utilisant la technique de linéarisation, on déduit que

$$\begin{aligned}\hat{R}^{12} - R \;&\simeq\; \sum_U r_k \frac{\varepsilon_k^{12}}{\pi_k^{12}} - Z = \phi\hat{X}^{12} + \psi\hat{Y}^{12} - Z \\ &=\; \hat{t}_{H-T} - Z,\end{aligned}$$

avec $\hat{R}^{12} = \dfrac{\hat{Y}^{12}}{\hat{X}^{12}}$.

Il résulte qu'on peut approximer la variance de $\hat{R}^{12}$ par la variance de $\hat{t}_{H-T}$ notée V_{H-T}. Alors, on a toujours $\text{AVar}_{opt}(\hat{R}_{opt}) \leq \text{AVar}(\hat{R}^{12})$ indépendemment du plan de sondage utilisé.

2. Une autre possibilité est d'estimer le total de $\mathcal{X}$ sur s_1 et le total de $\mathcal{Y}$ sur s_2. Il résulte que R est estimé par $\widehat{\widehat{R}}^{1,2} = \dfrac{\hat{Y}^2}{\hat{X}^1}$. On peut effectuer l'approximation suivante

$$\widehat{\widehat{R}}^{1,2} - R \simeq \phi\hat{X}^1 + \psi\hat{Y}^2 - Z.$$

Pour un sondage de Bernoulli bidimensionnel conditionnel à la taille, on a

$$\phi\hat{X}^1 + \psi\hat{Y}^2 = \phi h_1(\hat{X}^{1*} - \hat{X}^{12}) + \psi h_2(\hat{Y}^{2*} - \hat{Y}^{12}) + \hat{t}_{H-T}$$

avec $h_1 = \dfrac{n_1^*}{n_1}$ et $h_2 = \dfrac{n_2^*}{n_2}$ les taux de renouvellement. Il résulte alors que $\phi\hat{X}^1 + \psi\hat{Y}^2$ fait partie de la classe des estimateurs considérés (2.43) avec $a = \phi h_1$ et $b = \psi h_2$ et par conséquent la variance approximative de $\widehat{\widehat{R}}^{1,2}$ est supérieure à la variance approximative de $\hat{R}_{opt}$:

$$\mathrm{AVar}_{opt}(\hat{R}_{opt}) \leq \mathrm{AVar}(\widehat{\widehat{R}}^{1,2})$$

où AVar représente la variance approximative.

Exemple du coefficient de régression.

Pour le coefficient de régression $B = \dfrac{\sum_U x_k t_k}{\sum_U t_k^2}$ quand la variable $\mathcal{X}$ est connue sur un échantillon s_1 et la variable $\mathcal{T}$ sur $s_1 \cup s_2$, la variable linéarisée est $\tau^{-1}t_k(x_k - t_k B)$ où $\tau = \sum_U t_k^2$. Donc

$$\hat{B} - B \simeq \tau^{-1}\sum_U (t_k x_k - t_k^2 B)w_k^s - \tau^{-1}\sum_U (t_k x_k - t_k^2 B)$$

ce qui correspond à l'equation (2.70). On peut déduire a_{opt}, c_{opt} et d_{opt} optimaux et

$$\begin{aligned}\hat{B} - B &\simeq \tau^{-1}a_{opt}\left(\widehat{X*T}^{1*} - \widehat{X*T}^{12}\right) + \tau^{-1}B\left[c_{opt}\left(\widehat{\tau}^{1*} - \widehat{\tau}^{12}\right) + d_{opt}\left(\widehat{\tau}^{2*} - \widehat{\tau}^{12}\right)\right] \\ &- \tau^{-1}\sum_U (t_k x_k - t_k^2 B)\end{aligned}$$

où $X * T = \sum_U x_k t_k$, $\tau = \sum_U t_k^2$. Leurs estimateurs de Horvitz-Thompson sur s_{1*}, s_{12} et s_{2*} sont representés par un $\widehat{\ }$.

2.5 Estimation par régression par polynômes locaux dans des enquêtes sur plusieurs échantillons

2.5.1 Introduction

Nous allons traiter dans cette section la prise en compte de l'information auxiliaire dans des enquêtes sur plusieurs échantillons. Nous avons passé en revue dans le chapitre 1 les différentes approches pour tenir compte de l'information auxiliaire pour estimer le total d'une variable d'intérêt. Dans le cas de l'estimation dans des enquêtes répétées, Bell (2001), Fuller & Rao (2001), Hidiroglu (2001) et Singh, Kennedy & Wu (2001) proposent des estimateurs par la régression. Nous allons donner dans la suite une approche différente.

On va se placer dans la suite dans le cadre de l'approche "model-assisted". Pour un modèle de surpopulation linéaire, Särndal *et al* (1992) obtiennent que l'estimateur par la régression généralisée est approximativement sans biais que le modèle soit vrai ou non ; par contre, si la relation entre la variable d'intérêt et la variable auxiliaire n'est pas linéaire, l'estimateur par la regréssion peut avoir une grande variance. Des modèles nonparamétriques ont alors été introduits pour permettre que les modèles soient bien spécifiés pour des classes plus larges des fonctions ; on donne comme références Dorfman & Hall (1993), Chambers (1996) et Chambers, Dorfman & Wehrly (1993) dans le cas de l'approche "prédictive". Pour l'approche "model-assisted", Breidt & Opsomer (2000) proposent d'estimer le total d'une variable à l'aide d'une régression nonparamétrique par des polynômes locaux et ils concluent, à l'aide de simulations, que l'estimateur ainsi obtenu est "*presque aussi bon que l'estimateur par la régression classique quand la fonction de régression est supposée linéaire et il est supérieur pour le cas non-linéaire*".

On propose une extension au cas de deux échantillons s_1 et s_2, quand une variable $\mathcal{X}$ est mesurée sur s_1 et une variable $\mathcal{Y}$ sur s_2.

On commence par donner une description de la théorie asymptotique correspondante aux deux échantillons. On présente ensuite les principes de l'estimation par les polynômes locaux dans le cas d'un seul échantillon comme introduite par Breidt & Opsomer (2000) et par Kim, Breidt & Opsomer (2003) dans le cas de deux-degrés, suivi d'une extension à deux échantillons. On montre que l'estimateur construit vérifie les propriétés d'ADU (asymptotiquement sans biais) et de consistance et nous donnons une approximation de l'erreur quadratique moyenne.

2.5.2 Le cadre général

On s'intéresse aux propriétés d'estimateurs construits dans la suite quand les tailles de l'échantillon et de la population deviennent très larges. Une suite infinie d'éléments $\{u_k\}_k$ est considerée et, pour chaque unité k, une variable auxiliaire z uni-dimensionnelle est observée ; on note cette valeur z_k. On note également x_k et y_k les valeurs des deux variable d'intérêt $\mathcal{X}$ et $\mathcal{Y}$ pour le k-ème individu. On considère une suite de populations finies $\{\mathcal{U}_t\}$ de taille $\{N_t\}$ avec $N_t \to \infty$ comme introduite par Isaki & Fuller (1982) et Robinson & Särndal (1983). Soit $\mathcal{U}_t = \{u_1, \ldots, u_k, \ldots, u_{N_t}\} = \{1, \ldots, k, \ldots, N_t\}$ la population formée par les N_t premiers éléments de la suite $\{u_k\}_k$; on a en particulier, $\mathcal{U}_1 \subset \mathcal{U}_2 \subset \mathcal{U}_3 \subset \ldots$ et $0 < N_1 < N_2 < N_3 < \ldots$. On peut construire alors la suite de populations finies $\mathcal{U}_1 \times \mathcal{U}_1 \subset \mathcal{U}_2 \times \mathcal{U}_2 \subset \mathcal{U}_3 \times \mathcal{U}_3 \subset \ldots$ de tailles $0 < N_1 * N_1 < N_2 * N_2 < N_3 * N_3 < \ldots$.
Dans la suite, nous allons étendre le cadre de Isaki & Fuller (1982) et Robinson & Särndal (1983) au cas où on tire deux échantillons dans la population. On sélectionne dans chaque $\mathcal{U}_t \times \mathcal{U}_t$ un échantillon bidimensionnel $\boldsymbol{s}_t = (s_{1t}, s_{2t})$ selon une loi de probabilité $p_t(\boldsymbol{s}_t = (s_{1t}, s_{2t}))$. L'usage d'un plan bidimensionnel de sondage a été considéré pour la première fois par Cotton & Hesse (1992) et plus récemment par Salamin (2002) et Deville & Goga (2002) dans le but de contrôler la coordination entre échantillons pour les enquêtes répétées dans le temps. On a largement discuté au debut de ce chapitre le plan bidimensionnel.

On note n_{1t}, n_{2t} et n_{12t} les tailles de s_{1t}, s_{2t} et la taille de leur intersection $s_{12t} = s_{1t} \cap s_{2t}$. On note aussi n_{1*t} et n_{2*t} les tailles de s_{1*t} et respectivement s_{2*t} où $s_{1*,t} = s_{1t} - s_{12t}$ et $s_{2*,t} = s_{2t} - s_{12t}$. Nous avons défini la taille d'un échantillon bidimensionnel dans la **définition** 2.2.4. Soit $n_{\boldsymbol{s}_t}$ la taille de $\boldsymbol{s}_t = (s_{1t}, s_{2t})$ comme le triplet $n_{\boldsymbol{s}_t} = (n_{1*t}, n_{2*t}, n_{12t})$. Une relation d'ordre entre deux tailles $n_{\boldsymbol{s}_t}$ et $n_{\boldsymbol{s}_{t'}}$ revient à avoir la même relation d'ordre entre chacune des composantes de $n_{\boldsymbol{s}_t}$ et $n_{\boldsymbol{s}_{t'}}$. On a par exemple $n_{\boldsymbol{s}_t} < n_{\boldsymbol{s}_{t'}}$ si et seulement si $n_{1*t} < n_{1*t'}$, $n_{2*t} < n_{2*t'}$ et $n_{12,t} < n_{12,t'}$. On peut décrire maintenant comment la suite d'échantillons $\{\boldsymbol{s}_t\}$ est obtenue ; un échantillon bidimensionnel $\boldsymbol{s}_1$ de taille $n_{\boldsymbol{s}_1}$ est sélectionné dans $\mathcal{U}_1 \times \mathcal{U}_1$, ensuite un échantillon $\boldsymbol{s}_2$ de taille $n_{\boldsymbol{s}_2}$ est tiré dans $\mathcal{U}_2 \times \mathcal{U}_2$ avec $n_{\boldsymbol{s}_1} < n_{\boldsymbol{s}_2}$ et on poursuit cette opération. On obtient une suite croissante vers l'infini $n_{\boldsymbol{s}_1} < n_{\boldsymbol{s}_2} < \ldots$ et $n_{\boldsymbol{s}_t} < (N_t, N_t, N_t)$. Cela implique en particulier que $\{n_{1t}\}$, $\{n_{2t}\}$ et $\{n_{12t}\}$ sont aussi des suites infinies croissantes et que $\{s_{\oplus,t}\}$ pour $\oplus \in \{1, 2, 12\}$ sont construits comme dans Isaki & Fuller (1982).
Par analogie avec le cas unidimensionnel, les ensembles de probabilités de premier et deuxième ordre par rapport au plan bidimensionnel peuvent être définis (Deville & Goga 2002). Dans le cas de deux échantillons, nous avons sept variables aléatoires attribuées à chaque unité dont trois d'entre elles sont indépendantes ; plus précisement, elles sont $\varepsilon_{kt}^{\oplus} = \mathbf{1}_{\{k \in s_{\oplus,t}\}}$ pour tous $k \in U_t$ et $\oplus \in \{1, 2, 12, 1*, 2*, 1 \cup 2, 1* \cup 2*\}$. Dans la suite de notre étude nous

allons prendre les variables $\varepsilon_{kt}^{1*}, \varepsilon_{kt}^{12}, \varepsilon_{kt}^{2*}$ comme base pour la tribu engendrée par le couple $(\varepsilon_{kt}^{1}, \varepsilon_{kt}^{2})$.

Comme on a vu au début de ce chapitre, on a trois ensembles de probabilités d'inclusion classiques de premier degré notées π_{kt}^{1*}, π_{kt}^{2*}, π_{kt}^{12} et calculées par rapport aux plans marginaux p_{1*t}, p_{2*t} et au plan de l'intersection p_{12t} respectivement ; on a $\pi_{kt}^{1*} = Pr(k \in s_{1*t}) = E(\varepsilon_{kt}^{1*})$ et également pour π_{kt}^{12}, π_{kt}^{2*}.
Les probabilités d'inclusion multiples du second ordre sont $\pi_{kl,t}^{\oplus,\otimes} = Pr(k \in s_{\oplus,t}, l \in s_{\otimes,t}) = \mathrm{E}(\varepsilon_{kt}^{\oplus}\varepsilon_{lt}^{\otimes})$ pour $\oplus, \otimes \in \{1*, 12, 2*\}$ et $k, l \in U_t$. Pour $\oplus = \otimes$ on obtient les probabilités d'inclusion classiques du second ordre.
On note $\Delta_{klt}^{\oplus,\otimes} = \pi_{klt}^{\oplus,\otimes} - \pi_{kt}^{\oplus}\pi_{lt}^{\otimes}$ pour tous $k, l \in U_t$ et $\oplus, \otimes \in \{1*, 12, 2*\}$. Comme pour $\oplus = \otimes$ on obtient les probabilités d'inclusion classiques du second ordre, la notation $\Delta_{klt}^{\oplus,\oplus} = \Delta_{klt}^{\oplus}$ sera utilisée dans ce cas.
Pour simplifier les notations, l'indice t est suprimé dans la suite. On a alors l'échantillon bidimensionnel $\boldsymbol{s} = (s_1, s_2)$ séléctionné dans $\mathcal{U} \times \mathcal{U}$ selon le plan $p(\boldsymbol{s} = (s_1, s_2))$. L'espérance et la variance par rapport au plan de songage p seront notées E_p et V_p.

2.5.3 L'estimateur proposé

On veut construire un estimateur de $Q = t_x - t_y$ où t_x est le total de la variable $\mathcal{X}$ mesurée sur un échantillon s_1 et t_y est le total de $\mathcal{Y}$ mesurée sur un échantillon s_2 différent de s_1 mais avec une intersection non négligeable ; $t_x = \sum_U x_k$, $t_y = \sum_U y_k$. On considère que notre population peut être vue comme un échantillon dans un surpopulation infinie. Nous supposons le modèle de surpopulation infinie ξ suivant

$$x_k = f(z_k) + \varepsilon_k \tag{2.79}$$

$$y_k = g(z_k) + \eta_k \tag{2.80}$$

où f et g sont deux fonctions supposées continûment dérivables, ε_k et η_k sont des variables aléatoires indépendantes d'espérance zéro et de variance $v(x_k)$ et $r(y_k)$. On note $\mathcal{E}$ et $\mathcal{V}$ l'espérance et la variance sous le modèle ξ.
L'idée est d'utiliser l'estimateur pour le total d'une variable construit par Breidt & Opsomer (2000) et de le modifier proprement dans ce nouveau cadre.

L'estimation par polynômes locaux dans le cas d'un seul échantillon

On rappelle brièvement comment un total est estimé à l'aide de polynômes locaux. Si on veut estimer le total de $\mathcal{X}$ connu seulement sur échantillon s_1 sélectionné dans U selon un plan p_1 (qui selon les notations de la section précédente est le plan marginal de $p(\boldsymbol{s} = (s_1, s_2))$ par rapport à s_1) et si

on suppose que la fonction de régression $f(z_k)$ donnée par (2.79) est connue pour tous $k \in U$ alors la différence entre l'estimateur p_1-sans biais Horvitz-Thompson sur s_1, noté $\hat{X}_{HT}$, et son biais par rapport au modèle ξ vérifie les conditions d'être sans biais par rapport au modèle ξ et au plan p_1.

$$\left\{ \begin{array}{c} \widehat{\mathbf{X}} = \underbrace{\sum_{s_1} \frac{x_k}{\pi_k^1}}_{\hat{X}_{HT}} - \underbrace{\left(\sum_{s_1} \frac{f(z_k)}{\pi_k^1} - \sum_{U} f(z_k) \right)}_{\mathrm{E}_{\ \xi}(\hat{X}_{HT} - X)} \\ \pi_k^1 = Pr(k \in s_1) \end{array} \right. \tag{2.81}$$

On ne connait pas f; pour f paramétrique, on obtient $\hat{X} = \hat{X}_{GREG}$ l'estimateur par la régression généralisée (Särndal, Swensson and Wretman 1992) qui reste ξ-sans biais mais devient asymptotiquement sans biais et consistant quand on remplace le paramètre du modèle par son estimateur sur s_1 (Robinson & Särndal 1982). Cet estimateur est efficace seulement si le modèle supposé est vrai. Breidt & Opsomer (2000) construisent un estimateur plus robuste en supposant un modèle de régression nonparamétrique suivi d'une estimation par des polynômes locaux. En supposant que f est une fonction lisse et v une fonction positive et lisse, les N quantités inconnues $f(z_k)$ sont estimées en deux étapes :

i. **"POPULATION LEVEL" : sous le modèle** ξ, des estimateurs $\widehat{f}_{kU}$ de $f(z_k)$ sont obtenus à l'aide d'une régression nonparamétrique par des polynômes locaux de degré q (Fan 1992) pour tous $k \in U$. Cela revient à faire pour chaque unité $k \in U$ une régression locale pondérée avec des poids qui dépendent, via un noyau, de la taille du voisinage de k. La taille de ce voisinage est contrôlée par un paramètre h, appelé fenêtre dans la littérature. A chaque individu dans la population on associe donc un poids qui dépend de sa distance au point où on estime la régression. Ce poids est d'autant plus faible que le point est éloigné de l'endroit où on fait la régression locale par un polynôme. Les $\widehat{f}_{kU}$ pour tous $k \in U$ sont des quantités basées sur toute la population et par conséquent inconnues.

ii. **"SAMPLE LEVEL" : sous le plan de sondage** $p_1(s_1)$ et pour tous $k \in U$, les quantités $\widehat{f}_{kU}$ sont estimées par des estimateurs de Horvitz-Thompson notés $\widehat{\widehat{f}}_{ks_1}$

$$f(z_k) \xrightarrow{\xi} \widehat{f}_{kU} \xrightarrow{p_1(s_1)} \widehat{\widehat{f}}_{ks_1} \quad k \in U.$$

$t_x = \sum_U x_k$ est alors estimé par l'estimateur obtenu en remplaçant, dans (2.81), les $f(z_k)$ par $\widehat{\widehat{f}}_{ks_1}$:

$$\widehat{X}_{\pi} \ = \ \sum_{s_1} \frac{x_k - \widehat{\widehat{f}}_{ks_1}}{\pi_k^1} + \sum_{U} \widehat{\widehat{f}}_{ks_1}. \tag{2.82}$$

Sous certaines conditions, Breidt & Opsomer (2000) montrent que $\widehat{X}_\pi$ est asymptotiquement sans biais (ADU) et consistant pour $t_x = \sum_U x_k$ avec l'erreur quadratique moyenne approximée par la variance de l'estimateur par la régression généralisée

$$\left\{ \begin{array}{rcl} \mathrm{E}_{p_1}\left(\dfrac{\widehat{X}_\pi - t_x}{N}\right)^2 & = & \dfrac{1}{N^2}\displaystyle\sum_{k,l\in U}(x_k - \widehat{f}_{kU})(x_l - \widehat{f}_{lU})\dfrac{\pi^1_{kl} - \pi^1_k\pi^1_l}{\pi^1_k\pi^1_l} + o\left(\dfrac{1}{n}\right) \\ & & \pi^1_{kl} = Pr(k\&l \in s_1) \end{array} \right. \quad (2.83)$$

La variance ainsi obtenue est inférieure à la variance obtenue en supposant un modèle linéaire, c'est à dire la variance de $\hat{X}_{GREG}$. Par rapport à $\hat{X}_{GREG}$, $\widehat{X}_\pi$ est approximativement ξ sans biais et il est ξ-sans biais si la fonction $f(.)$ est un polynôme de degré inférieur ou égal à q.

L'estimation par polynômes locaux dans le cas de deux échantillons

On propose d'estimer $Q = t_x - t_y$ par $\hat{Q} = \hat{X} - \hat{Y}$ dans le cas où les valeurs $f(z_k)$ et $g(z_k)$ données par (2.79) et (2.80) sont supposées connues pour tous $k \in U$. On estime t_x dans s_{1*} par $\hat{X}_{1*}$ et dans s_{12} par $\hat{X}_{12}$; $\hat{X}_{12}$ et $\hat{X}_{1*}$ sont construits en utilisant (2.81). On considère ensuite, pour estimer t_x, l'estimateur composite de $\hat{X}_{1*}$ et $\hat{X}_{12}$; il résulte que $\hat{X}$ a l'expression suivante

$$\hat{X} = \alpha\left\{\sum_{s_{1*}}\frac{x_k - f(z_k)}{\pi^{1*}_k} + \sum_U f(z_k)\right\} + (1-\alpha)\left\{\sum_{s_{12}}\frac{x_k - f(z_k)}{\pi^{12}_k} + \sum_U f(z_k)\right\}, \quad (2.84)$$

avec α constante. On procède de la même manière avec t_y, pour une constante β :

$$\hat{Y} = \beta\left\{\sum_{s_{2*}}\frac{y_k - g(z_k)}{\pi^{2*}_k} + \sum_U g(z_k)\right\} + (1-\beta)\left\{\sum_{s_{12}}\frac{y_k - g(z_k)}{\pi^{12}_k} + \sum_U g(z_k)\right\}. \quad (2.85)$$

On rappelle que $s_{1*} = s_1 - s_{12}$, $s_{12} = s_1 \cup s_2$ et $s_{2*} = s_2 - s_{12}$.
Les fonctions de régression $f(z_k)$ et $g(z_k)$ sont inconnues ; elles sont estimées en deux étapes comme dans le cas d'un seul échantillon.

1. **POPULATION LEVEL** : Pour tous $k \in U$, on estime $f(z_k)$ et $g(z_k)$ par $\widehat{f}_{kU}$ et $\widehat{g}_{kU}$.
2. **SAMPLE LEVEL** :
• $\widehat{f}_{kU}$ est estimé dans s_{1*} par $\widehat{\widehat{f}}_{1*,k}$ et dans s_{12} par $\widehat{\widehat{f}}_{12,k}$.
• $\widehat{g}_{kU}$ est estimé dans s_{2*} par $\widehat{\widehat{g}}_{2*,k}$ et dans s_{12} par $\widehat{\widehat{g}}_{12,k}$.
On remplace dans (2.84) et (2.85), les f_k et g_k par les estimateurs obtenus

ci-dessus. Il résulte que t_x et t_y sont estimés avec biais par

$$\widehat{X}_\pi = \alpha \sum_{s_{1*}} \frac{x_k - \hat{\hat{f}}_{1*,k}}{\pi_k^{1*}} + (1-\alpha) \sum_{s_{12}} \frac{x_k - \hat{\hat{f}}_{12,k}}{\pi_k^{12}} + \sum_U \left\{ \alpha \hat{\hat{f}}_{1*,k} + (1-\alpha) \hat{\hat{f}}_{12,k} \right\} \tag{2.86}$$

$$\widehat{Y}_\pi = \beta \sum_{s_{2*}} \frac{y_k - \hat{\hat{g}}_{2*k}}{\pi_k^{2*}} + (1-\beta) \sum_{s_{12}} \frac{y_k - \hat{\hat{g}}_{12,k}}{\pi_k^{12}} + \sum_U \left\{ \beta \hat{\hat{g}}_{2*,k} + (1-\beta) \hat{\hat{g}}_{12,k} \right\}. \tag{2.87}$$

On donne les expressions explicites de $\widehat{f}_{kU}$, $\widehat{\widehat{f}}_{\oplus,k}$, $\widehat{g}_{kU}$ et $\widehat{\widehat{g}}_{\otimes,k}$ pour $k \in U$ et $\oplus \in \{1*, 12\}, \otimes \in \{12, 2*\}$. Pour cela, nous considérons les noyaux K, H avec les propriétés classiques (Fan & Gibels 1996) et les fenêtres h_1, h_2. Pour le choix de fenêtres, on se rapporte à Breidt & Opsomer (2000). L'estimateur par polynômes locaux de $f(z_k)$ est $\hat{f}_{kU} = \mathbf{p}'_{Uk} \boldsymbol{x}_U$ où pour tous $k \in U$, le vecteur des poids $\mathbf{p}'_{Uk}$ a l'expression suivante :

$$\left\{ \begin{array}{c} \mathbf{p}'_{Uk} = (1 \quad 0 \ldots 0)(\mathbf{Z}'_{Uk} \mathbf{P}_{Uk} \mathbf{Z}_{Uk})^{-1} \mathbf{Z}'_{Uk} \mathbf{P}_{Uk} \\ \mathbf{Z}_{Uk} = [1 \quad z_l - z_k \ldots (z_l - z_k)^q]_{l \in U} \\ \mathbf{P}_{Uk} = \text{ diag } \left\{ \frac{1}{h_1} K \left(\frac{z_l - z_k}{h_1} \right) \right\}_{l \in U} \end{array} \right. \tag{2.88}$$

L'indice U d'un vecteur indique que celui-ci est défini sur toute la population ; il résulte que $\mathbf{Z}_{Uk}$ est de dimension $N \times (q+1)$ et $\mathbf{P}_{Uk}$ de dimension $N \times N$. Pour $q = 0$, $\hat{f}_{kU}$ est l'estimateur à noyau de $f(z_k)$ ce qui montre qu'en considérant une estimation par des polynômes locaux de degré q on généralise l'estimation par noyau.
D'une manière analogue, on estime $g(z_k)$ par

$$\hat{g}_{kU} = \mathbf{w}'_{Uk} \boldsymbol{y}_U.$$

Le vecteur des poids $\mathbf{w}'_{Uk}$ s'obtiennent comme $\mathbf{p}'_{Uk}$ en remplaçant la matrice $\mathbf{P}_{Uk}$ par

$$\mathbf{W}_{Uk} = \text{ diag } \left\{ \frac{1}{h_2} H \left(\frac{z_l - z_k}{h_2} \right) \right\}_{l \in U}.$$

Les quantités $\hat{f}_{kU}, \hat{g}_{kU}$ sont définies sur toute la population U et par conséquent elles sont inconnues. Nous allons estimer $\hat{f}_{kU}$ par les estimateurs de Horvitz-Thompson basés sur s_{1*}, s_{12}, notés $\hat{\hat{f}}_{1*,k}$ et $\hat{\hat{f}}_{12,k}$; $\hat{g}_{kU}$ est estimé par le même type d'estimateur basé sur s_{12} et s_{2*} et notés $\hat{\hat{g}}_{12,k}$ et $\hat{\hat{g}}_{2*,k}$. On donne ci-dessous l'expression pour $\hat{\hat{f}}_{1*,k}$; les autres estimateurs s'obtiennent de manière analogue en remplaçant les poids π_k^{1*} par les probabilités d'inclusion correspondant à chaque échantillon utilisé. L'indice s_{1*} d'un vecteur

indique que celui-ci est défini sur l'échantillon s_{1*}.

$$\left\{\begin{array}{c} \hat{\hat{f}}_{1*,k} = \mathbf{p}'_{s_{1*,k}} \boldsymbol{x}_{s_{1*}} \\ \mathbf{p}'_{s_{1*,k}} = (1 \quad 0 \ldots 0)(\mathbf{Z}'_{s_{1*,k}} \mathbf{P}_{s_{1*,k}} \mathbf{Z}_{s_{1*,k}})^{-1} \mathbf{Z}'_{s_{1*,k}} \mathbf{P}_{s_{1*,k}} \\ \mathbf{P}_{s_{1*,k}} = \text{ diag } \left\{ \frac{1}{\pi_k^{1*} h_1} K\left(\frac{z_l - z_k}{h_1}\right) \right\}_{l \in s_{1*}} \end{array}\right. \tag{2.89}$$

2.5.4 Quelques résultats

Nous supposons que les conditions générales concernant la distribution des erreurs ε et η sous les modèles ainsi que celles portant sur les noyaux K et H et les fenêtres associées sont vérifiées (ce sont les conditions $A_1 - A_5$ de Breidt & Opsomer 2000). De plus, on suppose que les rapports, deux par deux, entre les tailles n_{1*}, n_{12} et n_{2*} sont toujours finis. Nous supposons également que les probabilités d'inclusion multiples du premier et second ordre vérifient :

$$\min_{k \in U} \pi_k^{\oplus} \geq \lambda_{\oplus} > 0, \quad \min_{k,l \in U} \pi_{k,l}^{\oplus,\otimes} \geq \delta_{\oplus} > 0 \quad \text{et} \quad \limsup_{N \to \infty} n_{\oplus} \max_{k \neq l \in U} |\Delta_{k,l}^{\oplus,\otimes}| < \infty \tag{2.90}$$

pour $\oplus, \otimes \in \{1*, 12, 2*\}$. Quant aux probabilités d'inclusion d'ordre supérieur, on suppose que pour chaque $\oplus, \otimes \in \{1*, 12, 2*\}$ les conditions suivantes sont satisfaites :

$$\lim_{N \to \infty} n_{\oplus}^2 \max_{(i,j,k,l) \in D_{4,N}} \left| \mathrm{E}_p(\varepsilon_i^{\oplus} - \pi_i^{\oplus})(\varepsilon_j^{\oplus} - \pi_j^{\oplus})(\varepsilon_k^{\oplus} - \pi_k^{\oplus})(\varepsilon_l^{\oplus} - \pi_l^{\oplus}) \right| < \infty \tag{2.91}$$

où $D_{4,N}$ est l'ensemble de tous les quadruples differents (i, j, k, l) dans U et

$$\lim_{N \to \infty} \max_{(i,j,k,l) \in D_{4,N}} \left| \mathrm{E}_p(\varepsilon_i^{\oplus} \varepsilon_j^{\oplus} - \pi_i^{\oplus} \pi_j^{\oplus})(\varepsilon_k^{\oplus} \varepsilon_l^{\oplus} - \pi_k^{\oplus} \pi_l^{\oplus}) \right| = 0 \tag{2.92}$$

$$\limsup_{N \to \infty} n_{\oplus} \max_{(i,j,k) \in D_{3,N}} \left| \mathrm{E}_p(\varepsilon_i^{\oplus} - \pi_i^{\oplus})^2(\varepsilon_j^{\oplus} - \pi_j^{\oplus})(\varepsilon_k^{\oplus} - \pi_k^{\oplus}) \right| < \infty \tag{2.93}$$

pour $D_{3,N}$ est l'ensemble de tous les triplets differents (i, j, k) dans U.
La technique de linéarisation par une développement de Taylor est utilisée afin d'étudier les propriétés liées au plan de sondage pour l'estimateur ainsi construit. On décrit en détail dans la suite la linéarisation de $\hat{\hat{f}}_{1*,k}$; toutes les autres quantités $\hat{\hat{f}}_{12,k}$, $\hat{\hat{g}}_{12,k}$, $\hat{\hat{g}}_{2*,k}$ pour tous $k \in U$, s'obtiennent de la même façon. Les $\hat{f}_{kU}$ sont des fonctions des totaux pour tous $k \in U$. Plus précisement, $\hat{f}_{kU} = F(N^{-1}\mathbf{t}_k)$ qui est estimé dans s_{1*} par $\hat{\hat{f}}_{1*,k} = F(N^{-1}\hat{\mathbf{t}}_{1*,k})$; pour chaque $k \in U$, le total $\mathbf{t}_k$ ainsi que son estimateur d'Horvitz-Thompson sur s_{1*}, $\hat{\mathbf{t}}_{1*,k}$ sont des vecteurs de dimension $G = 3q + 2$ quand une estimation par des polynômes locaux de degré q

est employée (Breidt & Opsomer 2000). Leurs expressions explicites sont données ci-dessous :

$$\left\{\begin{array}{rcccl} \mathbf{t}_k & = & (t_{kg})_{g=1}^{G} & = & \left(\sum_{i\in U} u^*_{kgi}\right)_{g=1}^{G} \\ \hat{\mathbf{t}}_{1*,k} & = & (\hat{t}_{1*,kg})_{g=1}^{G} & = & \left(\sum_{i\in U} u^*_{kgi}\frac{\varepsilon_i^{1*}}{\pi_i^{1*}}\right)_{g=1}^{G} \end{array}\right. \tag{2.94}$$

où $u^*_{kgi} = \frac{1}{h_1}K(\frac{z_i-z_k}{h_1})u^+_{kgi}$ avec

$$u^+_{kgi} = \left\{\begin{array}{cl} (z_i-z_k)^{g-1} & , \quad g \le 2q+1 \\ (z_i-z_k)^{g-2q-2}x_i & , \quad g > 2q+1. \end{array}\right. \tag{2.95}$$

Issue d'une développment de Taylor, la quantité $R^x_{k,1*}$ est définie comme suit :

$$\left\{\begin{array}{rcl} R^x_{k,1*} & = & \hat{\hat{f}}_{1*,k} - \hat{f}_{kU} - \frac{1}{N}\sum_{i\in U} u_{ki}\left(\frac{\varepsilon_i^{1*}}{\pi_i^{1*}} - 1\right), \quad k \in U \\ u_{ki} & = & \sum_{g=1}^{G} \frac{\partial \hat{\hat{f}}_{1*,k}}{\partial(N^{-1}\hat{t}_{1*,kg})}\Big|_{\hat{t}_{1*,k}=t_k} u^*_{kgi}, \quad i,k \in U. \end{array}\right. \tag{2.96}$$

Sous les conditions (2.90)-(2.93) quand $\oplus = \otimes = 1*$, on se place dans le cas d'un seul échantillon "s_{1*}" et alors, on peut appliquer les résultats obtenus par Breidt & Opsomer (2000) ; il résulte en particulier que $N^{-1}\mathbf{t}_k$ et $N^{-1}\hat{\mathbf{t}}_{1*,k}$ après $\hat{f}_{kU}$, $\hat{\hat{f}}_{1*,k}$ et les quatre premières dérivées partielles de $\hat{\hat{f}}_{1*,k}$ en $\hat{\mathbf{t}}_{1*,k} = \mathbf{t}_k$, $R^x_{k,1*}$ sont uniformement bornées ; en outre,

$$\frac{n_{1*}}{N}\sum_{k\in U} \mathrm{E}_{p_{1*}}(R^x_{k,1*})^2 = O(\frac{1}{n_{1*}h_1^2}) \quad \text{et alors} \tag{2.97}$$

$$\lim_{N\to\infty} \frac{1}{N}\sum_{k\in U} \mathrm{E}_{p_{1*}}(\hat{\hat{f}}_{1*,k} - \hat{f}_{kU})^2 = 0. \tag{2.98}$$

Par conséquent,

$$\lim_{N\to\infty} \mathrm{E}_{p_{1*}}\left(\left|\frac{\hat{X}_{1*,\pi} - t_x}{N}\right|\right) = 0 \tag{2.99}$$

$$n_{1*}\,\mathrm{E}_{p_{1*}}\left(\frac{\hat{X}_{1*,\pi} - t_x}{N}\right)^2 = \frac{n_{1*}}{N^2}\sum_{k,l\in U}(x_k - \hat{f}_{kU})(x_l - \hat{f}_{lU})\frac{\Delta_{kl}^{1*,1*}}{\pi_k^{1*}\pi_l^{1*}} + o(1) \tag{2.100}$$

où $\hat{X}_{1*,\pi}$ est l'estimateur obtenu en remplaçant $\hat{\hat{f}}_{k,s_1}$ par $\hat{\hat{f}}_{1*,k}$ dans (2.82).

Résultat 2.5.1 *L'estimateur* $\hat{Q}_\pi = \hat{X}_\pi - \hat{Y}_\pi$

i. est asymptotiquement sans biais pour $Q = t_x - t_y$:

$$\lim_{t\to\infty} \; E_{\,p}\left(\frac{\hat{Q}_\pi - Q}{N}\right) = 0$$

avec ξ *probabilité 1.*

ii. est consistant pour $Q = t_x - t_y$:

$$\lim_{t\to\infty} \; P\left(\left|\frac{\hat{Q}_\pi - Q}{N}\right| > \varepsilon\right) = 0$$

avec ξ *probabilité 1 et* $\varepsilon > 0$.

Preuve : On note $\hat{X}_{1*,\pi}$ et $\hat{X}_{12,\pi}$ les estimateurs de $\hat{X}_\pi$ obtenus après avoir remplacé $\hat{\hat{f}}_{k,s_1}$ par $\hat{\hat{f}}_{1*,k}$ et $\hat{\hat{f}}_{12,k}$ dans (2.82). Il résulte que (2.86) devient

$$\hat{X}_\pi = \alpha\hat{X}_{1*,\pi} + (1-\alpha)\hat{X}_{12,\pi}.$$

De manière analogue, $\hat{Y}_\pi = \beta\hat{Y}_{2*,\pi} + (1-\beta)\hat{Y}_{12,\pi}$ et alors

$$\begin{aligned}\mathrm{E}_{\,p}\left(\left|\frac{\hat{Q}_\pi - Q}{N}\right|\right) \;\leq\; & \alpha\,\mathrm{E}_{\,p_{1*}}\left(\left|\frac{\hat{X}_{1*,\pi} - X}{N}\right|\right) + (1-\alpha)\,\mathrm{E}_{\,p_{12}}\left(\left|\frac{\hat{X}_{12,\pi} - X}{N}\right|\right) + \\ & \beta\,\mathrm{E}_{\,p_{2*}}\left(\left|\frac{\hat{Y}_{2*,\pi} - Y}{N}\right|\right) + (1-\beta)\,\mathrm{E}_{\,p_{12}}\left(\left|\frac{\hat{Y}_{12,\pi} - Y}{N}\right|\right)\end{aligned}$$

et tous les termes à droite de l'inégalité ci-dessus tendent vers zéro conformèment à (2.99).
On a donc $\lim_{N\to\infty} \mathrm{E}_{\,p}\left(\left|\frac{\hat{Q}_\pi - Q}{N}\right|\right) = 0$, et en particulier $\hat{Q}_\pi$ est asymptotiquement sans biais et consistant pour Q.

□

Le résultat suivant fournit une approximation de l'erreur quadratique.

Résultat 2.5.2 *Notons* $\hat{\mathbf{f}}_U$ *et* $\hat{\mathbf{g}}_U$ *les vecteurs* N*-dimensionnels d'éléments* $\hat{f}_{kU}$ *et* $\hat{g}_{kU}$ *pour* $k \in U$;
$\Pi_N^{\oplus} =$ *diag* $(\pi_k^{\oplus})_{k=1,\ldots,N}$ *pour* $\oplus \in \{1*, 12, 2*\}$. *Alors*

$$nE_p\left(\frac{(\hat{X}_\pi - \hat{Y}_\pi) - (X - Y)}{N}\right)^2 = \frac{n}{N^2}c'\begin{pmatrix} \Delta_{kl}^{1*} & \Delta_{kl}^{1*,2*} & \Delta_{kl}^{1*,12} \\ \Delta_{kl}^{2*,1*} & \Delta_{kl}^{2*} & \Delta_{kl}^{2*,12} \\ \Delta_{kl}^{12,1*} & \Delta_{kl}^{12,2*} & \Delta_{kl}^{12} \end{pmatrix} c + o(1) \tag{2.101}$$

où

$$c = \begin{pmatrix} (\Pi_N^{1*})^{-1} & 0 & 0 \\ 0 & (\Pi_N^{2*})^{-1} & 0 \\ 0 & 0 & (\Pi_N^{12})^{-1} \end{pmatrix}\begin{pmatrix} \alpha(\boldsymbol{x}_U - \hat{\mathbf{f}}_U) \\ -\beta(\boldsymbol{y}_U - \hat{\mathbf{g}}_U) \\ (1-\alpha)(\boldsymbol{x}_U - \hat{\mathbf{f}}_U) - (1-\beta)(\boldsymbol{y}_U - \hat{\mathbf{g}}_U) \end{pmatrix}$$

Preuve : C'est une application directe du résultat (2.3.3) qui donne l'expression de la variance de type Horvitz-Thompson dans le cas de deux échantillons et de (2.100) considérée à la fois pour $\hat{X}_{1*,\pi}$, $\hat{X}_{12,\pi}$, $\hat{Y}_{2*,\pi}$ et $\hat{Y}_{12,\pi}$. On fait la remarque que la taille n qui intervient dans (2.101) peut être une des valeurs n_{1*}, n_{12} où n_{2*} puisque le rapport de deux parmi les trois est toujours fini.

□

Chapitre 3

Une approche nonparamétrique par splines de régression pour tenir compte de l'information auxiliaire dans les sondages

3.1 Introduction

Ce chapitre est consacré à la prise en compte, par un modèle nonparamétrique, de l'information auxiliaire afin d'améliorer l'estimateur d'un total. L'approche considérée est l'approche dite de superpopulation pour une classe d'estimateurs de type *model-assisted* décrits dans le Chapitre 1, Section 1.5 pour lesquels la relation entre l'information auxiliaire et la variable d'intérêt n'est pas nécessairement linéaire, ni même paramétrique. Il est vrai que les estimateurs, sous un modèle linéaire, sont protégés contre les mauvaises spécifications du modèle au sens où ils sont asymptotiquement sans biais et consistants par rapport au plan de sondage. Néanmoins, si la vraie relation n'est pas linéaire, l'efficacité, en terme de variance, de l'estimateur par la régression généralisée $\hat{T}_{GREG}$ sous un modèle linéaire peut se révéler mauvaise, même par rapport à l'estimateur de H-T qui, pourtant, ne prend pas en compte l'information auxiliaire. La bonne spécification du modèle peut se montrer importante pour les qualités de l'estimateur. Il s'agit alors d'un problème de robustesse (Cassell *et al.*, 1977) du modèle vis-à-vis de la réalité.

Nous proposons ici un modèle, dit nonparamétrique, qui couvre une classe beaucoup plus large de relations entre information auxiliaire et variable d'intérêt, en imposant uniquement des conditions de régularité (dérivabilité) sur la fonction de régression. Contrairement au modèle linéaire où il suffit de connaitre la somme des valeurs de l'information auxilaire sur toute la population, les estimateurs basés sur des modèles nonparamétriques ont besoin que l'information auxiliaire soit connue sur toute la population, ce qui est fréquent de nos jours avec les progrès de l'informatique. Nous laissons ainsi aux données une chance de s'exprimer.

Les premiers auteurs à considérer des modèles nonparamétriques en théorie des sondages sont Kuo (1988), Dorfman (1992), Dorfman & Hall (1993), Chambers (1996) et Chambers, Dorfman & Wehrly (1993) pour la classe des estimateurs *model-based*. Récemment, Breidt & Opsomer (2000) et Kim, Breidt & Opsomer (2003) ont proposé des estimateurs *model-assisted* du total d'une population pour les sondages à une ou deux phases. Ils adoptent pour cela une approche par polynômes locaux (Fan 1992, 1993 ; Fan & Gibels 1996) qui consiste à effectuer en chaque point une régression linéaire pondérée, la pondération étant contrôlée par une fonction appelée noyau et une fenêtre associée. Ces poids donnent moins d'importance aux individus les plus lointains. Cette méthode nécessite donc d'effectuer autant de régressions qu'il y a d'individus dans l'échantillon, ce qui peut se révéler très lourd lorsque la base de sondage est grande.

Nous proposons ici une méthode alternative, beaucoup plus simple et rapide. Notre estimateur de la fonction de régression est décomposé sur une base de fonctions B-splines (Dierckx 1993), ce qui revient à construire un modèle linéaire dont le nombre de variables (les B-splines) n'est pas fixé à l'avance. Dans le cadre classique de la régression non-paramétrique, ces estimateurs possèdent les mêmes qualités que les polynômes locaux (Zhou, Shen & Wolfe 1998). L'estimateur obtenu est une somme linéaire en y_k, les valeurs de la variable d'intérêt sur l'échantillon, avec des poids qui ne dépendent pas de la variable d'intérêt. De plus, le paramètre de la régression sur la base de B-splines est indépendant de l'unité dans la population contrairement au cas des estimateurs par les polynômes locaux. Cette methode garde alors les propriétés du modèle linéaire classique de régression, qui d'ailleurs est un cas particulier de l'estimation par des B-splines, mais en même temps elle donne une généralisation du modèle linéaire classique de régression. Par ailleurs, les estimateurs ainsi obtenus peuvent s'étendre facilement au cas de plusieurs variables auxiliaires.

Le modèle de superpopulation, l'estimateur spline ainsi que l'estimateur du total sont présentés dans la deuxième section de ce chapitre. La section 3 présente les principaux résultats théoriques. L'estimateur obtenu est une somme pondérée des valeurs de la variable d'intérêt sur l'échantillon avec des poids qui vérifient les équations de calage pour les totaux de fonctions B-splines. Sous des conditions générales nous montrons que notre estimateur

est asymptotiquement sans biais et consistant par rapport au plan de sondage. Nous retrouvons en particulier pour les estimateurs de type Horvitz-Thompson des vitesses de convergence de l'ordre de $n^{-1/2}$ ce qui est courant dans la théorie de sondages. En outre, le paramètre de la régression sur la base de B-splines a une norme de l'ordre $K^{1/2}$ ce qui est normal compte tenu des propriétés de B-splines et du fait que son nombre de composantes K tend vers l'infini ce qui n'est pas le cas pour la régression classique où il est fixé. Nous donnons une approximation de la variance sous le plan de sondage de l'estimateur ainsi construit par la variance de type Horvitz-Thompson pour les résidus d'estimation et nous montrons également que la variance anticipée atteint asymptotiquement la borne de Godambe-Joshi (1965). Dans la section 4, des simulations confirment les bonnes propriétés de notre estimateur. Les calculs ont été effectués avec le logiciel R. Nous proposons dans la section 5 une brève discussion sur les extensions possibles. Finalement, la dernière section regroupe les démonstrations.

3.2 Modèle et notations

On considère une population finie $U = \{1, \ldots, k, \ldots, N\}$ et on suppose connues les valeurs d'une variable auxiliaire unidimensionnelle $\mathcal{X}$ pour toutes les unités k dans U; on note ces valeurs x_k pour $k \in \{1, \ldots, N\}$. Un échantillon s de taille fixe n est sélectionné dans U selon un plan de sondage quelconque $p(s)$ et la valeur de la variable d'intérêt $\mathcal{Y}$ est observée pour chaque unité dans l'échantillon ; on obtient y_k pour $k \in s$. Pour chaque individu $k \in U$, la probabilité d'inclusion dans s, $\pi_k = Pr(k \in s)$, est supposée strictement positive ; de même, $\pi_{ik} = Pr(i, k \in s) > 0$ pour tous $i, k \in U$. Pour chaque individu k dans la population, on note $I_k = \mathbf{1}_{\{k \in s\}}$ la variable indicatrice d'appartenance dans l'échantillon s. Nous voulons estimer le total de $\mathcal{Y}$ sur U,

$$t_y = \sum_{k \in U} y_k.$$

On suppose que la population finie U peut être vue comme un échantillon dans une surpopulation infinie ξ, la relation entre les valeurs y_k et x_k est la suivante

$$\xi : \quad y_k = f(x_k) + \varepsilon_k \tag{3.1}$$

où f est une fonction inconnue et les erreurs ε_k sont des variables aléatoires indépendantes, d'espérance nulle et de variance $v(x_k) = v_k$. On suppose que les $x_k \in [0, 1]$ pour tous k.

On construit un estimateur de la fonction f par splines de régression. Définissons pour cela l'espace des fonctions spline d'ordre m ($m \geq 2$) et avec K noeuds intérieurs $0 = \xi_0 < \xi_1 < \ldots < \xi_K < \xi_{K+1} = 1$

$S_{K,m} = \{s \in C^{m-2}[0, 1] : s(x) \quad \text{un polynôme de degré m-1 sur} \quad (\xi_j, \xi_{j+1})\}.$

On considère dans la suite la situation où les noeuds sont équidistants. Pour $m = 1$, $S_{K,1}$ est l'ensemble des fonctions en escalier sur les sous intervalles de $[0, 1]$ définis par les noeuds et pour $m = 2$, $S_{K,2}$ est l'ensemble de fonctions continues sur $[0, 1]$ et linéaires par morceaux. L'espace $S_{K,m}$ est un espace linéaire de dimension $q = K + m$ dont une base est constituée des fonctions B-splines $(B_j(.))_{j=1}^q$ définies de la manière suivante (Schumaker 1981, Dieckx 1993) :

$$B_j \text{ a les noeuds } \xi_{j-m}, \ldots, \xi_j$$
$$B_j(x) = (\xi_j - \xi_{j-m}) \sum_{l=0}^{m} \frac{(\xi_{j-l} - x)_+^{m-1}}{\Pi_{r=0, r\neq l}^{m} (\xi_{j-l} - \xi_{j-r})}$$

où $(\xi_{j-l} - x)_+^{m-1} = (\xi_{j-l} - x)^{m-1}$ si $\xi_{j-l} \geq x$ et zéro sinon. Les fonctions B_j sont représentées dans la figure 3.1, elles ont les propriétés suivantes :

i. $B_j(x) \geq 0$ pour tout $x \in [0, 1]$.

ii. $B_j(x) = 0$ pour $x \notin [\xi_{j-m}, \xi_j]$.

iii. $\sum_{j=1}^q B_j(x) = 1$ pour tout $x \in [0, 1]$.

On estime f par l'estimateur des moindres carrés $\hat{f}(x) \in S_{K,m}$

$$\begin{cases} \hat{f}(x) = \sum_{j=1}^q \hat{\theta}_j B_j(t) \quad \text{où} \\ \hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_q)' = \min_{\theta_j} \sum_{k=1}^N \left(y_k - \sum_{j=1}^q \theta_j B_j(x_k) \right)^2 . \end{cases} \tag{3.2}$$

Introduisons quelques notations utiles pour la suite.

(N1) : Soit $\boldsymbol{B}_U$ la matrice de taille $N \times q$ d'éléments $B_j(x_k)$ pour $k \in U$, $j = 1, \ldots, q$ et $\boldsymbol{B}_s$ sa restriction à l'échantillon s, c'est-à-dire

$$\boldsymbol{B}_U = (B_j(x_k))_{k\in U, j=1,\ldots,q}, \quad \boldsymbol{B}_s = (B_j(x_k))_{k\in s, j=1,\ldots,q}.$$

(N2) : $\boldsymbol{b}'(x_k)$ pour $k \in U$ sont les lignes de $\boldsymbol{B}_U$; $\boldsymbol{b}'(x_k) = (B_1(x_k), \ldots, B_q(x_k))$.

$$\boldsymbol{B}_U = (\mathbf{b}'(x_k))_{k\in U}$$

(N3) : On note $\boldsymbol{y}_U$, (respectivement $\boldsymbol{f}_U$) le vecteur d'éléments y_k (respectivement $f(x_k)$) pour $k \in U$. Leur restriction sur s sont notées $\boldsymbol{y}_s$ et $\boldsymbol{f}_s$.

$$\boldsymbol{y}_U = (y_1, \ldots, y_N)', \quad \boldsymbol{y}_s = (y_k)'_{k\in s}$$
$$\boldsymbol{f}_U = (f(x_1), \ldots, f(x_N))', \quad \boldsymbol{f}_s = (f(x_k))'_{k\in s}$$

(N4) : Pour simplifier les écritures, on pose $f(x_k) = f_k$, $k \in U$.

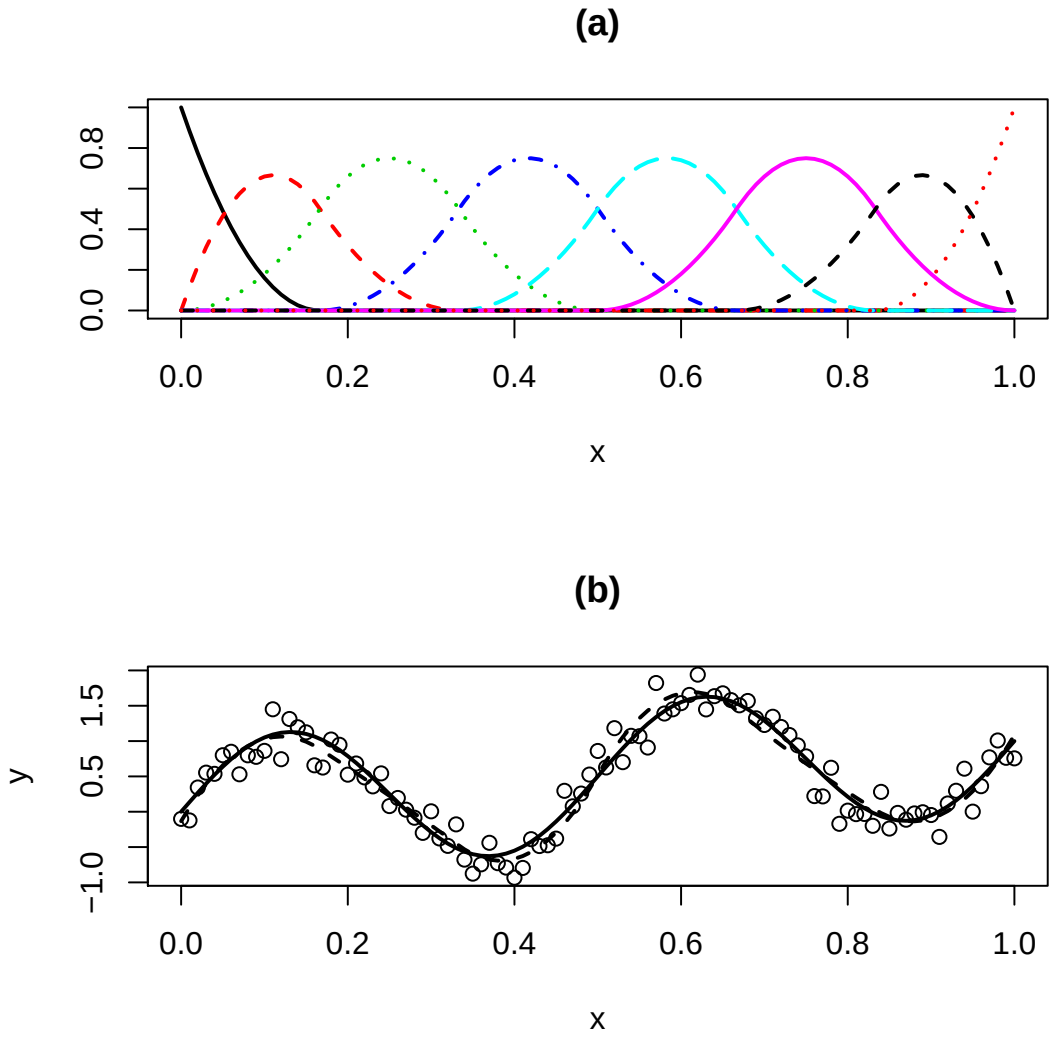

FIG. 3.1 – (a) Description de la base des fonctions B-splines pour 5 noeuds intérieurs et $m = 3$. (b) Exemple d'approximation spline (en pointillés) de la fonction $f(x) = x + \sin(4\pi x)$ observée de manière bruitée en 100 points de mesure.

Avec ces nouvelles notations, l'estimateur par B-splines de f_k défini en (3.2) s'écrit pour chaque x_k, $k \in U$:

$$\left\{ \begin{array}{l} \hat{f}_k = \mathbf{b}'(x_k)\hat{\boldsymbol{\theta}} \quad k \in U \\ \hat{\boldsymbol{\theta}} = (\boldsymbol{B'}_U \boldsymbol{B}_U)^{-1} \boldsymbol{B'}_U \boldsymbol{y}_U \end{array} \right. \tag{3.3}$$

en supposant que la matrice $\boldsymbol{B'}_U \boldsymbol{B}_U$ est inversible.

On note

(N5) : $\hat{\boldsymbol{f}}_U = (\hat{f}_k)_{k \in U}$ et $\hat{\boldsymbol{f}}_s = (\hat{f}_k)_{k \in s}$ sa restriction sur l'échantillon s.
On a

$$\hat{\boldsymbol{f}}_U = \boldsymbol{B}_U \hat{\boldsymbol{\theta}} \quad \text{et} \quad \hat{\boldsymbol{f}}_s = \boldsymbol{B}_s \hat{\boldsymbol{\theta}}.$$

Si on connait f_k pour tous $k \in U$, on estime sans-biais par rapport au plan p et au modèle ξ le total t_y par l'estimateur par la différence généralisée (Cassel *et al.* 1976)

$$\hat{t}_y^0 = \sum_{k \in s} \frac{y_k - f_k}{\pi_k} + \sum_{k \in U} f_k. \tag{3.4}$$

Malheureusement, dans la pratique on ne connait pas les f_k. Nous proposons d'expliciter la construction de notre estimateur en deux étapes.

On considère d'abord, pour chaque f_k, l'estimateur par B-splines $\hat{f}_k$ (3.3). Si on remplace dans (3.4) les f_k par $\hat{f}_k$, on obtient un estimateur pour le total t_y qui est toujours p-sans biais mais approximativement ξ-sans biais :

$$\hat{t}_y \quad = \quad \sum_{k \in s} \frac{y_k - \hat{f}_k}{\pi_k} + \sum_{k \in U} \hat{f}_k. \tag{3.5}$$

On observe qu'on ne peut pas calculer les quantités $\hat{f}_k$ pour $k \in U$ puisqu'elles sont définies en fonction du vecteur inconnu $\boldsymbol{y}_U$. On considère donc le π-estimateur de $\hat{f}_k$ dans l'échantillon s, noté $\hat{\hat{f}}_k$ pour tous $k \in U$ ce qui revient à considérer le π-estimateur de $\hat{\boldsymbol{\theta}}$ dans l'échantillon s, noté $\hat{\hat{\boldsymbol{\theta}}}$

$$\left\{ \begin{array}{l} \hat{\hat{f}}_k = \boldsymbol{b'}(x_k)\hat{\hat{\boldsymbol{\theta}}} \quad k \in U \\ \hat{\hat{\boldsymbol{\theta}}} = (\boldsymbol{B'}_s \boldsymbol{\Pi}_s^{-1} \boldsymbol{B}_s)^{-1} \boldsymbol{B'}_s \boldsymbol{\Pi}_s^{-1} \boldsymbol{y}_s \end{array} \right. \tag{3.6}$$

avec $\boldsymbol{\Pi}_s$ donné ci-dessous dans (N6). Alors,

$$\hat{\hat{\boldsymbol{f}}}_U = (\hat{\hat{f}}_k)_{k \in U} = \boldsymbol{B}_U \hat{\hat{\boldsymbol{\theta}}} \quad \text{et} \quad \hat{\hat{\boldsymbol{f}}}_s = (\hat{\hat{f}}_k)_{k \in s} = \boldsymbol{B}_s \hat{\hat{\boldsymbol{\theta}}}.$$

Finalement, l'estimateur par B-splines du total t_y est obtenu en remplaçant dans (3.5) les $\hat{f}_k$ par $\hat{\hat{f}}_k$ comme suit :

$$\hat{t}_{y,\pi} \quad = \quad \sum_{k \in s} \frac{y_k - \hat{\hat{f}}_k}{\pi_k} + \sum_{k \in U} \hat{\hat{f}}_k. \tag{3.7}$$

On a utilisé la notation

(N6) : $\mathbf{\Pi}_U = \text{diag}(\pi_k)_{k\in U}$ avec la restriction sur s : $\mathbf{\Pi}_s = \text{diag}(\pi_k)_{k\in s}$

et on introduit les notations suivantes fréquentes dans la régression linèaire

(N7) : $\boldsymbol{T}_U = \boldsymbol{B'}_U \boldsymbol{B}_U$ avec $\hat{\boldsymbol{T}}_s = \boldsymbol{B'}_s \mathbf{\Pi}_s^{-1} \boldsymbol{B}_s$ son estimateur sur s.

On a $\boldsymbol{T}_U = \sum_U \boldsymbol{b}(x_k)\boldsymbol{b'}(x_k)$ et $\hat{\boldsymbol{T}}_s = \sum_s \frac{1}{\pi_k}\boldsymbol{b}(x_k)\boldsymbol{b'}(x_k)$.

(N8) : On note $\boldsymbol{I}_N$ la matrice unité d'ordre N et $\boldsymbol{I}_s$ sa restriction sur s.

3.3 Les principaux résultats

Nous présentons dans cette section les propriétés asymptotiques de l'estimateur $\hat{t}_{y\pi}$.

Pour cela nous devons supposer que les tailles de la population et de l'échantillon deviennent de plus en plus grandes. Isaki & Fuller (1982) donnent une méthode pour obtenir une population et un échantillon avec des tailles qui tendent vers l'infini. Ce thème a déjà été abordé dans chapitre 2, section 6. Une suite infinie d'éléments $\{u_k\}_k$ est considerée et, pour chaque unité k, on observe la valeur x_k de la variable auxiliaire $\mathcal{X}$. On considère une suite de populations finies $\{\mathcal{U}_t\}$ de taille $\{N_t\}$ avec $N_t \to \infty$. Soit $\mathcal{U}_t = \{u_1, \ldots, u_k, \ldots, u_{N_t}\} = \{1, \ldots, k, \ldots, N_t\}$ la population formée par les N_t premiers éléments de la suite $\{u_k\}_k$; on a en particulier, $\mathcal{U}_1 \subset \mathcal{U}_2 \subset \mathcal{U}_3 \subset \ldots$ et $0 < N_1 < N_2 < N_3 < \ldots$. On sélectionne dans chaque $\mathcal{U}_t$ un échantillon $s_t \subset \mathcal{U}_t$ de taille n_t selon une loi de probabilité p_t. La suite des échantillons n'est plus emboitée, contrairement à celle des populations, mais avec des tailles qui tendent vers l'infini : $n_1 < n_2 < \ldots < n_t < \ldots$ et $n_t < N_t$ pour tous t. Pour chaque plan p_t on peut calculer les probabilités d'inclusion du premier et du deuxième degré, notées $\pi_{k,t} = Pr(k \in s_t)$ et $\pi_{kl,t} = Pr(k, l \in s_t)$. Les conditions posées ci-dessous concernent plus precisément les suites $(N_t)_t$, $(n_t)_t$ et $(\pi_{k,t})_t$, $(\pi_{kl,t})_t$ où pour alléger les notations on a supprimé l'indice t.

(C1). $\lim_{N\to\infty} \frac{n}{N} = \pi \in (0, 1)$.

(C2). La fonction f est m fois continuement dérivable sur $[0, 1]$.

(C3). $\frac{1}{N}\sum_{k\in U} y_k^2 < \infty$ conditionnellement aux valeurs réalisées $y_1, \ldots, y_N$ sous le modèle ξ.

(C4). $\min_{k\in U} \pi_k \geq \lambda > 0, \quad \min_{i,k\in U} \pi_{ik} \geq \lambda* > 0,$

$\overline{\lim}_{N\to\infty} n \max_{i\neq k\in U} |\pi_{ik} - \pi_i\pi_k| < \infty.$

(C5). $K = o(N)$

(C6). Il existe une fonction de répartition $Q(x)$ admettant une densité

strictement positive sur $[0,1]$ telle que

$$\sup_{x\in[0,1]} |Q_N(x) - Q(x)| = o(K^{-1})$$

où $Q_N(x)$ est la distribution empirique de $(x_i)_{i=1}^N$.

(C7). $\frac{1}{N}\sum_{k\in U} \varepsilon_k^2 < \infty$ conditionnellement aux valeurs réalisées $y_1, \ldots, y_N$ sous le modèle ξ.

(C8). $\sup_{k\in U} v_k < \infty$.

La condition (C1), classique pour les modèles de superpopulation, indique qu'asymptotiquement la taille de la population ne doit pas être trop grande par rapport à la taille de l'échantillon et inversement que la taille de l'échantillon ne doit pas être trop grande par rapport à la taille de la population. Les conditions (C2) et (C5), classiques en estimation nonparamétrique (cf. par exemple Zhou, Shen, & Wolfe 1998, Burman 1981), signifient que la fonction à estimer f doit être suffisamment régulière et que la dimension de la base des B-splines ne doit pas croitre trop rapidement avec la taille de l'échantillon. La condition (C3) signifie que la variable $\mathcal{Y}$ admet un moment d'ordre deux (la même chose pour C7) et la condition (C4), que l'on retrouve par exemple dans Isaki & Fuller (1982), Robinson & Särndal (1983), Breidt & Opsomer (2000) est une condition technique sur les probabilités d'inclusion du premier et deuxième degré qui assure que tous les éléments dans la population ont une probabilité strictement positive d'être sélectionnés et qu'il est possible de construire un estimateur de la variance. La condition (C6) classique en estimation nonparamétrique assure qu'asymptotiquement il n'existe pas de sous intervalle $[0,1]$ qui ne contienne aucun point x. Cette hypothèse assure l'inversibilité de la matrice $\boldsymbol{B}'_U\boldsymbol{B}_U$. Finalement, la condition (C8) suppose que la variance du bruit est une fonction bornée. Les preuves des principales propriétés se trouvent dans la dernière section.

3.3.1 Calage et estimation par splines de régression

L'estimateur $\hat{t}_{y,\pi}$ défini en (3.7) est une somme pondérée des y_k, $k \in s$ avec des poids qui ne dépendent pas de $\mathcal{Y}$ et qui indiquent à la fois l'appartenance de l'unité k dans l'échantillon et l'information auxiliaire. On donne ci-dessous leur expression :

$$\hat{\hat{f}}_k = \underbrace{\boldsymbol{b}'(x_k)(\boldsymbol{B}'_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s)^{-1}\boldsymbol{B}'_s\boldsymbol{\Pi}_s^{-1}}_{w_{ks}}\boldsymbol{y}_s = \boldsymbol{w}_{ks}\boldsymbol{y}_s.$$

Alors,

$$\hat{t}_{y,\pi} = \sum_{k\in s}\frac{y_k}{\pi_k} + \sum_{k\in U}\left(1-\frac{\varepsilon_k}{\pi_k}\right)\boldsymbol{w}_{ks}\boldsymbol{y}_s$$

$$
\begin{aligned}
&= \sum_{l\in s}\frac{y_l}{\pi_l} + \sum_{l\in s}\sum_{k\in U}\left(1-\frac{\varepsilon_k}{\pi_k}\right)\boldsymbol{w}_{ks}\boldsymbol{e}^l_{n_s}y_l \\
&= \sum_{l\in s}\left\{\frac{1}{\pi_l} + \sum_{k\in U}\left(1-\frac{\varepsilon_k}{\pi_k}\right)\boldsymbol{w}_{ks}\boldsymbol{e}^l_{n_s}\right\}y_l \\
&= \sum_{l\in s} w^c_{ls}y_l
\end{aligned}
$$

où $\boldsymbol{e}^l_{n_s}$ est le vecteur de taille n_s avec la valeur 1 sur la $l$ème position et zéro ailleurs. Les quantités w^c_{ls} sont appelées dans la suite les poids de calage :

$$
w^c_{ls} = \frac{1}{\pi_l} + \sum_{k\in U}\left(1-\frac{\varepsilon_k}{\pi_k}\right)\boldsymbol{w}_{ks}\boldsymbol{e}^l_{n_s}. \tag{3.8}
$$

On peut appliquer alors les poids ainsi obtenus, w^c_{ls}, aux variables $B_j(x_k)$ des totaux $t_{B_j} = \sum_{k\in U} B_j(x_k)$ connus pour tous $j = 1, \ldots, q$.

Proposition 3.3.1 *: Les poids w^c_{ls} vérifient les équations de calage suivantes (d'où le nom de poids de calage) :*

$$
\sum_{l\in s} w^c_{ls}B_j(x_l) = \sum_{l\in U} B_j(x_l) \text{ pour tous } j = 1, \ldots, q. \tag{3.9}
$$

3.3.2 Quelques propriétés de l'estimateur

Proposition 3.3.2 *: Les vecteurs $\boldsymbol{b}(x_k) = (B_1(x_k), \ldots, B_q(x_k))'$ satisfont également l'équation de calage (3.9) :*

$$
\sum_{k\in s} w^c_{ks}\boldsymbol{b}(x_k) = \sum_{k\in U}\boldsymbol{b}(x_k).
$$

Proposition 3.3.3 *: Les vecteurs $\boldsymbol{b}(x_k) = (B_1(x_k), \ldots, B_q(x_k))'$, pour tout $k \in U$, satisfont la relation*

$$
\mathbf{1}'_q \times \boldsymbol{b}(x_k) = 1.
$$

où $\mathbf{1}_q$ est le vecteur unité de dimension q, $\mathbf{1}'_q = (1, \ldots, 1)$.

Cette propriété résulte du fait que les quantités $B_j(x_k)$ pour $j = 1, \ldots, q$ satisfont $\sum_{j=1}^q B_j(x_k) = 1$ pour tous x_k.
Alors, les vecteurs $\boldsymbol{b}(x_k)$ satisfont $\boldsymbol{c}' \times \boldsymbol{z}_k = 1$ (Särndal 1980) pour $\boldsymbol{c} = \mathbf{1}_q$ et le vecteur d'information auxiliaire $\boldsymbol{z}_k = \boldsymbol{b}(x_k)$ ce qui correspond au système des π-poids appliqués à $\boldsymbol{\theta}$ et discutés par Särndal (1980) comme alternative au meilleur estimateur linéaire sans biais dans le cas d'un modèle linéaire.

Proposition 3.3.4 *: On note $E_k = y_k - \hat{f}_k$ pour tous $k \in U$ les résidus dans la population. Alors,*

$$\sum_U E_k = 0.$$

Cette relation est obtenue en considérant les équations normales pour $\hat{\theta}$

$$\sum_U \boldsymbol{b}(x_k)[y_k - \boldsymbol{b'}(x_k)\hat{\boldsymbol{\theta}}] = 0$$

et en faisant le produit scalaire avec le vecteur $\mathbf{1}_q$:

$$\sum_U \mathbf{1}_q' \boldsymbol{b}(x_k) E_k = 0.$$

En appliquant la propriété (3.3.3), on obtient $\sum_U E_k = 0$.

Proposition 3.3.5 *: On note $e_k = y_k - \hat{\hat{f}}_k$ pour tous $k \in U$. Alors*

$$\sum_s \frac{e_k}{\pi_k} = 0.$$

Pour montrer cette relation, on prend les équations normales pour $\hat{\hat{\boldsymbol{\theta}}}$

$$\sum_s \frac{\boldsymbol{b}(x_k)}{\pi_k}[y_k - \boldsymbol{b'}(x_k)\hat{\hat{\boldsymbol{\theta}}}] = 0$$

et on procède de la même manière que pour montrer la propriété (3.3.4).

Proposition 3.3.6 *: L'estimateur $\hat{t}_{y\pi}$ peut s'écrire comme le total des quantités $\hat{\hat{f}}_k$,*

$$\hat{t}_{y\pi} = \sum_{k \in U} \hat{\hat{f}}_k.$$

C'est une conséquence directe de la propriété 3.3.5.

Proposition 3.3.7 *: On note $\boldsymbol{G}_s = \hat{\boldsymbol{T}}_s^{-1} \boldsymbol{B'}_s \boldsymbol{\Pi}_s^{-1}$ alors*

$$\mathbf{1'}_s \boldsymbol{\Pi}_s^{-1} (\boldsymbol{B}_s \boldsymbol{G}_s - \boldsymbol{I_s}) = 0.$$

On a $\boldsymbol{B}_s \hat{\boldsymbol{T}}_s^{-1} \boldsymbol{B'}_s = \left(\boldsymbol{b'}(x_i) \hat{\boldsymbol{T}}_s^{-1} \boldsymbol{b}(x_k) \right)_{i,k \in s}$ et alors

$$\begin{aligned} \mathbf{1'}_s \boldsymbol{\Pi}_s^{-1} \boldsymbol{B}_s \boldsymbol{G}_s &= \mathbf{1'}_s \boldsymbol{\Pi}_s^{-1} \boldsymbol{B}_s \hat{\boldsymbol{T}}_s^{-1} \boldsymbol{B'}_s \boldsymbol{\Pi}_s^{-1} \\ &= \left(\frac{1}{\pi_1} \left(\sum_s \frac{1}{\pi_k} \boldsymbol{b'}(x_k) \hat{\boldsymbol{T}}_s^{-1} \right) \boldsymbol{b}(x_1), \ldots, \frac{1}{\pi_n} \left(\sum_s \frac{1}{\pi_k} \boldsymbol{b'}(x_k) \hat{\boldsymbol{T}}_s^{-1} \right) \boldsymbol{b}(x_n) \right) \end{aligned}$$

et comme $\mathbf{1'}_s\mathbf{\Pi}_s^{-1}\boldsymbol{I_s} = \left(\frac{1}{\pi_k}\right)_{k\in s}$ il suffit de montrer que $\sum_s \frac{1}{\pi_k}\boldsymbol{b'}(x_k)\hat{\boldsymbol{T}}_s^{-1}\boldsymbol{b}(x_i) = 1$ pour tous $i \in s$. On a

$$\mathbf{1'}_q\hat{\boldsymbol{T}}_s = \sum_s \frac{1}{\pi_k}\mathbf{1'}_q\boldsymbol{b}(x_k)\boldsymbol{b'}(x_k) = \sum_s \frac{1}{\pi_k}\boldsymbol{b'}(x_k)$$

par la propriété 3.3.3. Avec cette identité, la relation à prouver devient

$$\left[\sum_s \frac{1}{\pi_k}\boldsymbol{b'}(x_k)\right]\hat{\boldsymbol{T}}_s^{-1}\boldsymbol{b}(x_i) = \mathbf{1'}_q\hat{\boldsymbol{T}}_s\hat{\boldsymbol{T}}_s^{-1}\boldsymbol{b}(x_i) = \mathbf{1'}_q\boldsymbol{b}(x_i) = 1$$

toujours par la propriété 3.3.3.

On peut montrer de la même façon la propriété suivante :

Proposition 3.3.8 *: On note* $\boldsymbol{G}_U = \boldsymbol{T}_U^{-1}\boldsymbol{B'}_U$ *alors*

$$(\boldsymbol{I_N} - \boldsymbol{G'}_U\boldsymbol{B'}_U)\mathbf{1}_N = 0. \tag{3.10}$$

3.3.3 Propriétés de $\hat{t}_{y\pi}$ par rapport au plan de sondage

Nous allons montrer que l'estimateur ainsi construit $\hat{t}_{y\pi}$ est asymptotiquement équivalent à l'estimateur d'Horvitz-Thompson $\hat{t}_{HT}$ et à l'estimateur par la régression généralisée $\hat{t}_y$. Nous donnons les preuves dans la section 6.

Proposition 3.3.9 *: Sous les conditions précédentes C1-C7,*

$$\frac{1}{N}\left(\hat{t}_{y\pi} - \hat{t}_{HT}\right) = O_p((K/n)^{1/2})$$

où $\hat{t}_{HT} = \sum_s \frac{y_k}{\pi_k}$ *est l'estimateur d'Horvitz-Thompson.*
L'estimateur $\hat{t}_{y\pi}$ *est donc asymptotiquement sans biais et consistant pour* t_y.

Proposition 3.3.10 *: Sous les conditions C1-C7*

$$\frac{1}{N}\left(\hat{t}_{y\pi} - \hat{t}_y\right) = O_p\left(K^{3/2}/n\right)$$

où $\hat{t}_y = \sum_s \frac{y_k - \hat{f}_k}{\pi_k} + \sum_U \hat{f}_k$.

Il résulte de la propriété ci-dessus que lorsque $K = o(n^{1/3})$, on a

$$N^{-1}(\hat{t}_{y\pi} - t_y) = N^{-1}(\hat{t}_y - t_y) + o_p(n^{-1/2}). \tag{3.11}$$

Proposition 3.3.11 *:La variance de $\hat{t}_{y\pi}$ peut être approximée par la variance de l'estimateur $\hat{t}_y$, c'est à dire la variance de type Horvitz-Thompson pour l'erreur de prévision $E_k = y_k - \hat{f}_k$ pour tous $k \in U$,*

$$Var_p(\hat{t}_{y\pi}) \simeq \frac{1}{N^2} \sum_{k\in U} \sum_{i\in U} \Delta_{ki} \frac{y_k - \hat{f}_k}{\pi_k} \frac{y_i - \hat{f}_i}{\pi_i}. \tag{3.12}$$

On obtient ainsi que la variance approximative de $\hat{t}_{y\pi}$ est inférieure à celle donnée par un modèle paramétrique.

3.3.4 Variance sous le modèle et sous le plan

Isaki & Fuller (1982) introduisent le critère de la variance anticipée ("anticipated variance") pour évaluer un prédicteur d du total t_y. La variance anticipée est la variance de $\frac{1}{N}(d-t_y)$ calculée sous le plan p et sous le modèle ξ,

$$Var\left(\frac{1}{N}(d-t_y)\right) = E_\xi E_p \left(\frac{1}{N}(d-t_y)\right)^2 - \left(E_\xi E_p \left(\frac{1}{N}(d-t_y)\right)\right)^2.$$

Pour des prédicteurs d qui sont ξ-sans biais, cette variance devient l'espérance sous le modèle de la "design mean square error", $E_\xi E_p(\frac{1}{N}(d-t_y))^2$, considérée aussi par Särndal (1980) comme un indicateur de la qualité du prédicteur utilisé. Dans notre cas, $\hat{t}_{y\pi}$ est asymptotiquement $p\xi$-sans biais ce qui implique que la variance anticipée $Var(\frac{1}{N}(\hat{t}_{y\pi} - t_y))$ et $E_\xi E_p \frac{1}{N}(\hat{t}_{y\pi} - t_y)^2$ sont asymptotiquement équivalentes ; on donne la justification ci-dessous.
On utilise la propriété (3.3.6) ; alors $\frac{1}{N}(\hat{t}_{y\pi} - t_y)$ peut s'écrire de la manière équivalente :

$$\begin{aligned}
\frac{1}{N}(\hat{t}_{y\pi} - t_y) &= \frac{1}{N}\left(\sum_{k\in U} \hat{\hat{f}}_k - \sum_{k\in U} \hat{f}_k\right) = \frac{1}{N}(\mathbf{1'}_N \hat{\hat{\boldsymbol{f}}}_U - \mathbf{1'}_N \hat{\boldsymbol{f}}_U) \\
&= \frac{\mathbf{1'}_N}{N} \boldsymbol{B}_U \left(\boldsymbol{G}_s \boldsymbol{y}_s - \boldsymbol{G}_U \boldsymbol{y}_U\right) \qquad (3.13) \\
&= \frac{\mathbf{1'}_N}{N} \boldsymbol{B}_U (\boldsymbol{G}_s \boldsymbol{f}_s - \boldsymbol{G}_U \boldsymbol{f}_U + \boldsymbol{G}_s \boldsymbol{\varepsilon}_s - \boldsymbol{G}_U \boldsymbol{\varepsilon}_U).
\end{aligned}$$

On peut changer l'ordre dans le calcul des espérances, cela nous donne

$$E_\xi E_p \frac{1}{N}(\hat{t}_{y\pi} - t_y) = E_p E_\xi \frac{1}{N}(\hat{t}_{y\pi} - t_y) = \frac{\mathbf{1'}_N}{N} \boldsymbol{B}_U E_p(\boldsymbol{G}_s \boldsymbol{f}_s - \boldsymbol{G}_U \boldsymbol{f}_U)$$

et par (3.24) et le lemme 3.6.1 point 2, cette quantité tend vers zéro.

D'après Godambe (1982), une stratégie d'échantillonnage (p, d) est robuste sous le modèle ξ si $E_\xi E_p(\frac{1}{N}(d - t_y))^2$ atteint la borne inférieure de

Godambe-Joshi (1965), $\frac{1}{N^2}\sum_{k\in U} v(x_k)\frac{1-\pi_k}{\pi_k}$ qui est aussi la variance anticipée de $\hat{t}_y$. On sait que l'estimateur par la régression généralisée $\hat{t}_{GREG}$ atteint cette borne dans le cas d'un modèle linéaire (Särndal, Swensson & Wretman, p. 453). Plus récemment, Breidt & Opsomer (2000) ont montré cette propriété pour des estimateurs assistés par un modèle nonparamétrique en utilisant une approche par des polynômes locaux.

Nous allons montrer dans la suite que $\hat{t}_{y\pi}$ est robuste dans le sens mentionné ci-dessus.

Proposition 3.3.12 *: Sous les conditions C1-C7 et pour $K \approx N^{\frac{1}{2m+1}}$, $\hat{t}_{y\pi}$ atteint asymptotiquement la borne inférieure de Godambe-Joshi (1965) :*

$$E_\xi E_p\left(\frac{1}{N}(\hat{t}_{y\pi}-t_y)\right)^2 = \frac{1}{N^2}\sum_{k\in U} v(x_k)\frac{1-\pi_k}{\pi_k} + o(1)$$

Preuve : On donne la preuve de cette proposition dans la section 6. □

3.4 Une étude par simulations

Cette section est consacrée à une étude par simulations du comportement de notre estimateur du total d'une variable d'intérêt $\mathcal{Y}$. Nous effectuons une comparaison avec l'estimateur de Horvitz-Thompson, $\hat{t}_{HT}$, ainsi que l'estimateur par la régression, $\hat{t}_{GREG}$. Les calculs ont été effectués avec le logiciel R.

Nous considérons, comme Breidt et Opsomer (2000), une population $\mathcal{U}$ constituée de $N = 1000$ individus. Nous associons les deux variables y_k et x_k à chaque individu k de $\mathcal{U}$ et nous supposons de plus qu'il existe une fonction f telle que

$$y_k \quad = \quad f(x_k) + \epsilon_k, \tag{3.14}$$

où les ϵ_k sont des erreurs de mesure indépendantes, de loi $N(0, \sigma^2)$. Les x_k sont tirés de manière indépendantes selon une loi uniforme sur l'intervalle $[0, 1]$.

La population est simulée pour les fonctions f suivantes :
- $f_{lin}(x) = 1 + 2(x - 0.5)$, qui correspond à la situation où l'estimateur par la régression linéaire devrait se montrer le plus performant.
- $f_{\exp}(x) = \exp(-8x)$,
- $f_{\sin}(x) = 2 + \sin(2\pi x)$.

MSE	f	$\hat{t}_{HT}$	$\hat{t}_{GREG}$	$\hat{t}_{y\pi}$
$\sigma = 0.1$	f_{lin}	2980	**94**	99
	$f_{\exp}$	513	281	**100**
	$f_{\sin}$	4706	1835	**102**
$\sigma = 0.4$	f_{lin}	4504	**1515**	1633
	$f_{\exp}$	1788	1638	**1552**
	$f_{\sin}$	5476	3103	**1565**

TAB. 3.1 – Valeurs de l'erreur quadratique moyenne des estimateurs pour différentes fonctions f et différents niveaux de bruit.

Nous effectuons des sondages simples sans remise de taille $n = 100$ dans la population, ainsi $\pi_k = n/N = 0.1$. Ensuite nous comparons les différents estimateurs du total selon le critère de l'erreur quadratique, définie pour tout estimateur $\hat{t}$ par

$$MSE(\hat{t}) \quad = \quad (\hat{t} - t_y)^2, \tag{3.15}$$

où $t_y = \sum_{k \in \mathcal{U}} y_k$.

Afin de mieux évaluer la précision de chaque estimateur, mille réplications de l'expérience sont effectuées, c'est à dire qu'on réalise 1000 sondages et qu'on évalue l'erreur quadratique moyenne à partir de ces 1000 expériences.

L'estimateur spline est construit pour $m = 3$ et $K = 5$ noeuds intérieurs positionnés aux quantiles de la variable $\mathcal{X}$.

Les résultats sont rassemblés dans la table 3.1 pour les différentes fonctions f considérées et deux valeurs de σ, correspondant à un bruit faible ($\sigma = 0.1$) et un bruit important ($\sigma = 0.4$).

On peut remarquer tout d'abord que l'estimateur spline fournit généralement les meilleurs résultats. En effet il se montre presque équivalent à $\hat{t}_{GREG}$ lorsque le vrai modèle est linéaire et nettement supérieur aux autres estimateurs lorsque la relation entre $\mathcal{X}$ et $\mathcal{Y}$ est non linéaire. Remarquons également que l'estimateur de Horvitz-Thompson, qui ne tient pas compte de l'information auxiliaire se montre le moins performant. Enfin, lorsque le niveau de bruit est plus important, $\sigma = 0.4$, le gain apporté par le modèle diminue.

Nous avons considéré dans cette simulation une base de fonctions B-splines avec un nombre de noeuds fixe pour chaque sondage, égal à $K = 5$. Si la base est trop grande, certains problèmes d'identifiabilité peuvent apparaitre lors du sondage en raison de l'absence d'observations entre deux noeuds. D'une manière générale le choix du nombre de noeuds et de leur position semble être un problème délicat qui peut avoir une influence importante sur la qualité de l'estimation de la fonction f et par conséquent de

l'estimateur du total. Il semble qu'un petit nombre de noeuds puisse quand même autoriser notre estimateur à approximer de manière correcte une fonction "régulière". C'est pourquoi nous préconisons d'utiliser, en fonction de la taille de l'échantillon, le moins de noeuds possible. Ce conseil peut d'ailleurs être relier aux résultats asymptotiques de la section précédente qui montrent que la variance augmente avec le nombre de noeuds.

3.5 Discussions et perspectives

Plusieurs prolongements de ce travail peuvent être envisagés.

Il s'agit tout d'abord de poursuivre les calculs afin d'exhiber un estimateur de la variance sous le plan et d'en déduire sa normalité asymptotique.

On pourra également appliquer cette technique par splines de régression à l'estimation en présence de deux échantillons (cf. chapitre précédent).

On peut aussi considérer le choix pratique des paramètres de lissage dans la régression nonparamétrique. La régularité de l'estimateur est contrôlée ici par le nombre et la position des noeuds. Dans la pratique il semble difficile (et long) d'optimiser ces paramètres de manière automatique. C'est pourquoi O'Sullivan (1986) dans le cadre de la régression nonparamétrique et très récemment Zheng et Little (2003) dans le cadre d'échantillonnage "PPS" ont proposé des estimateurs splines pénalisés. Il s'agit d'une fonction spline de régression dont la régularité est contrôlée par une pénalisation via un paramètre de lissage. Ce terme de pénalisation, qui s'ajoute au critère des moindres carrés, est généralement de la forme suivante

$$J(\hat{f},\rho) \quad = \quad \rho \int_0^1 \left(\hat{f}^{(2)}(t)\right)^2 \, dt \tag{3.16}$$

où le paramètre ρ permet de donner plus ou moins d'importance au caractère régulier de l'estimateur qui est mesuré par la norme de sa dérivée seconde.

Cette approche présente le double avantage de pouvoir à la fois assurer l'existence de l'estimateur de manière quasi systématique en éliminant les problèmes d'identifiabilité, même pour un nombre de noeuds important, mais aussi de contrôler la régularité de l'estimateur à l'aide d'un unique paramètre (dont en plus une valeur peut être choisie de manière automatique par validation croisée).

Un prolongement de ce travail est d'inclure un terme de pénalisation et de déterminer un critère automatique de choix du paramètre de lissage afin de s'affranchir des inconvénients des splines de régression.

Enfin, cette approche par splines de régression peut se généraliser assez rapidement au cas multivarié qui est fréquent dans la pratique, c'est à dire losqu'on dispose de plusieurs variables auxiliaires. Il est possible d'estimer directement des modèles nonparamétriques "purs",

$$y_k \quad = \quad f(\mathbf{x}_k) + \epsilon_k \tag{3.17}$$

où l'information auxiliaire pour l'individu k est contenue dans le vecteur $\mathbf{x}_k$. Malheureusement, en raison du "fleau" de la dimension, il est connu que l'estimation peut être de très mauvaise qualité si la taille du vecteur $\mathbf{x}$ est supérieure à 3 et si l'échantillon n'est pas de taille suffisante. Cependant on peut considérer des sous modèles tels que les modèles additifs qui s'écrivent

$$y_k = \mu + f_1(x_k^1) + \cdots + f_p(x_k^p) + \epsilon_k \tag{3.18}$$

et ne souffrent pas du "fléau" de la dimension. Par ailleurs, les estimateurs par splines de régression sont relativement faciles à mettre en oeuvre pour ces modèles additifs (Hastie et Tibshirani 1990).

3.6 Preuves

***Preuve* de la proposition** 3.3.1 : On a

$$\begin{aligned}\sum_{l\in s} w_{ls}^c B_j(x_l) &= \sum_{l\in s}\left\{\frac{1}{\pi_l}+\sum_{k\in U}\left(1-\frac{\varepsilon_k}{\pi_k}\right)\boldsymbol{w}_{ks}\boldsymbol{e}_{n_s}^l\right\}B_j(x_l)\\ &= \sum_{l\in s}\frac{1}{\pi_l}B_j(x_l)+\sum_{k\in U}\left(1-\frac{\varepsilon_k}{\pi_k}\right)\boldsymbol{w}_{ks}\sum_{l\in s}\boldsymbol{e}_{n_s}^l B_j(x_l)\end{aligned}$$

et $\sum_{l\in s}\boldsymbol{e}_{n_s}^l B_j(x_l) = (B_j(x_l))_{l\in s} = \boldsymbol{B}_s\boldsymbol{e}_q^j$ où $\boldsymbol{e}_{n_s}^l$, respectivement $\boldsymbol{e}_q^j$, est le vecteur de taille n_s, respectivement q, prenant la valeur 1 sur la l, respectivment j-eme position et zéro ailleurs. Il résulte en particulier que

$$\begin{aligned}\boldsymbol{w}_{ks}\sum_{l\in s}\boldsymbol{e}_{n_s}^l B_j(x_l) &= \boldsymbol{w}_{ks}\boldsymbol{B}_s\boldsymbol{e}_q^j = \boldsymbol{b'}(x_k)(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s)^{-1}\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s\boldsymbol{e}_q^j\\ &= \boldsymbol{b'}(x_k)\boldsymbol{e}_q^j = B_j(x_k).\end{aligned}$$

Alors,

$$\begin{aligned}\sum_{l\in s} w_{ls}^c B_j(x_l) &= \sum_{l\in s}\frac{1}{\pi_l}B_j(x_l)+\sum_{k\in U}\left(1-\frac{\varepsilon_k}{\pi_k}\right)B_j(x_k)\\ &= \sum_U B_j(x_k).\end{aligned}$$

□

On donne dans la suite quelques lemmes et propriétés utiles dans les preuves des propositions 3.3.9, 3.3.10 et 3.3.12.

Lemmes et propriétés

Lemme 3.6.1 *: Sous les conditions C1-C7, on a :*

i. $\left\|\frac{1}{N}\boldsymbol{B'}_U\boldsymbol{B}_U\right\| = O(K^{-1})$. *(Agarwal & Studden 1980, lemme 6.3)*

ii. $\left\|\frac{1}{N}\sum_{k\in U}\boldsymbol{b}(x_k)\right\| = O\left(K^{-1/2}\right)$.

iii. $\|\hat{\boldsymbol{\theta}}\| = O(K^{1/2})$.

Preuve :
2. On a

$$\left\|\frac{1}{N}\sum_{k\in U}\boldsymbol{b}(x_k)\right\|^2 = \frac{1}{N^2}\sum_{j=1}^{q}\sum_{k\in U}\sum_{i\in U}B_j(x_k)B_j(x_i)$$

$$
\begin{aligned}
&= \frac{1}{N^2}\sum_{j=1}^{q}\sum_{k\in U}\sum_{i\in U,|k-i|<m} B_j(x_k)B_j(x_i)\\
&\leq \frac{1}{N^2}\sum_{j=1}^{q}\sum_{k\in U}\frac{N}{K}B_j(x_k) = \frac{1}{NK}\sum_{k\in U}\sum_{j=1}^{q}B_j(x_k) = \frac{1}{K}
\end{aligned}
$$

parce que pour $k, i \in U$ tels que $|k-i| > m$ on a $B_j(x_k)B_j(x_i) = 0$ et dans la somme $\sum_{i\in U,|k-i|<m} B_j(x_k)B_j(x_i)$, seuls $\frac{N}{K}$ termes sont non nuls et ils sont tous inférieurs à 1. On a utilisé aussi le fait que $\sum_{j=1}^{q} B_j(x_k) = 1$.
3. On a

$$
\begin{aligned}
\|\hat{\boldsymbol{\theta}}\| &= \left\|\left(\frac{1}{N}\boldsymbol{B'}_U\boldsymbol{B}_U\right)^{-1}\frac{1}{N}\boldsymbol{B'}_U\boldsymbol{y}_U\right\| \leq \left\|\left(\frac{1}{N}\boldsymbol{B'}_U\boldsymbol{B}_U\right)^{-1}\right\|\left\|\frac{1}{N}\boldsymbol{B'}_U\boldsymbol{y}_U\right\|\\
&= O(K)O(K^{-1/2}) = O(K^{1/2})
\end{aligned}
$$

par le premier point et par le lemme 3.6.3, point 1. □

Lemme 3.6.2 *: On note* $\alpha_i = \frac{I_i}{\pi_i} - 1$ *pour tous* $i \in U$ *et soit* $(\beta_i)_{i\in U}$ *un ensemble de réels. On suppose une population telle que la condition C1 soit satisfaite et les probabilités d'inclusion de premier et deuxième degré* π_i, π_{ik} *satisfont la condition C4. Si* $\frac{1}{N}\sum_{i\in U}\beta_i^2 < \infty$ *et* $\frac{1}{N^2}\sum_{i\neq k\in U}|\beta_i||\beta_k| < \infty$ *, alors*

$$
E\left(\frac{1}{N^2}\sum_{i\in U}\sum_{k\in U}\beta_i\beta_k\alpha_i\alpha_k\right) = O(n^{-1}).
$$

Preuve : On a $E(\alpha_i^2) = \frac{1-\pi_i}{\pi_i}$ pour tous $i \in U$ et $E(\alpha_i\alpha_k) = \frac{\Delta_{ik}}{\pi_i\pi_k}$ pour tous $i \neq k \in U$ ce qui implique

$$
\begin{aligned}
E\left(\frac{1}{N^2}\sum_{i\in U}\sum_{k\in U}\beta_i\beta_k\alpha_i\alpha_k\right) &= \frac{1}{N^2}\sum_{i\in U}\beta_i^2\frac{1-\pi_i}{\pi_i} + \frac{1}{N^2}\sum_{i\neq k\in U}\beta_i\beta_k\frac{\Delta_{ik}}{\pi_i\pi_k}\\
&\leq \frac{1-\lambda}{\lambda}\frac{1}{N^2}\sum_{i\in U}\beta_i^2 + \frac{1}{N^2}\sum_{i\neq k\in U}|\beta_i||\beta_k|\frac{|\Delta_{ik}|}{\lambda^2}\\
&\leq \frac{1}{n}\frac{1-\lambda}{\lambda}\frac{n}{N}\left(\frac{1}{N}\sum_{i\in U}\beta_i^2\right)\\
&\quad + \frac{1}{n}\frac{n\max_{i\neq k}|\Delta_{ik}|}{\lambda^2}\left(\frac{1}{N^2}\sum_{i\neq k\in U}|\beta_i||\beta_k|\right).
\end{aligned}
$$

On a utilisé la condition C4 pour obtenir ces majorations. Si de plus on tient compte de la condition C1 et de celles de l'hypothèse, alors la conclusion est obtenue. □

Ce lemme va nous être utile pour les trois lemmes suivants.

Lemme 3.6.3 *: Sous les conditions de régularité C1-C7, on a*

i. $\frac{1}{N}\boldsymbol{B'}_U\boldsymbol{y}_U = O\left(K^{-1/2}\right).$

ii. $\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{y}_s - \boldsymbol{B'}_U\boldsymbol{y}_U\right) = O_p\left(n^{-1/2}\right).$

iii. $\frac{1}{N}\boldsymbol{B}'_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{y}_s = O_p(K^{-1/2}).$

Preuve :

i. On a

$$\begin{aligned}\left\|\frac{1}{N}\boldsymbol{B'}_U\boldsymbol{y}_U\right\|^2 &= \frac{1}{N^2}\sum_{j=1}^{q}\left[\sum_{i\in U}B_j(x_i)y_i\right]^2\\ &\leq \frac{c^2}{N^2}\sum_{j=1}^{q}\left[\sum_{i\in U}B_j(x_i)\right]^2\\ &\leq \frac{c^2}{K}\end{aligned}$$

ce qui implique $\frac{1}{N}\boldsymbol{B'}_U\boldsymbol{y}_U = O\left(\frac{1}{\sqrt{K}}\right).$

ii. On calcule $\mathrm{E}_p(\left\|\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{y}_s - \boldsymbol{B'}_U\boldsymbol{y}_U\right)\right\|^2)$.

$$\begin{aligned}\left\|\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{y}_s - \boldsymbol{B'}_U\boldsymbol{y}_U\right)\right\|^2 &= \left\|\frac{1}{N}\sum_{i\in U}\boldsymbol{b}(x_i)y_i\left(\frac{I_i}{\pi_i}-1\right)\right\|^2\\ &= \frac{1}{N^2}\left[\sum_{i,k\in U}\left(\sum_{j=1}^{q}B_j(x_i)y_iB_j(x_k)y_k\right)\alpha_i\alpha_k\right].\end{aligned}$$

On applique le lemme précédent pour $\beta_i\beta_k = \sum_{j=1}^{q}B_j(x_i)y_iB_j(x_k)y_k$ pour tous $i,k\in U$. On vérifie d'abord que les conditions du lemme (3.6.2) sont satisfaites :

$$\frac{1}{N}\sum_{i\in U}\beta_i^2 = \frac{1}{N}\sum_{i\in U}\sum_{j=1}^{q}B_j^2(x_i)y_i^2 < \frac{1}{N}\sum_{i\in U}y_i^2 < \infty$$

ce qui prouve que la première condition du lemme (3.6.2) est vérifiée. On a utilisé le fait que $\sum_{j=1}^{q}B_j^2(x_i) = \|b(x_i)\|^2 \leq 1$ (Burman, 1991

pp. 270) et la condition C3. Pour le deuxième point, on doit vérifier que

$$\frac{1}{N^2}\sum_{i\neq k\in U}(|\beta_i\beta_k|)<\infty.$$

Comme $\sum_{j=1}^{q}B_j(x_i)B_j(x_k)=<b(x_i),b(x_k)>\leq\|b(x_i)\|\times\|b(x_k)\|\leq 1$, (Burman 1991) on obtient

$$\begin{aligned}\frac{1}{N^2}\sum_{i\neq k\in U}(|\beta_i\beta_k|) &\leq \frac{1}{N^2}\sum_{i\neq k\in U}\left(\sum_{j=1}^{q}B_j(x_i)B_j(x_k)\right)|y_iy_k|\\ &\leq \frac{1}{N^2}\sum_{i\neq k\in U}|y_iy_k|\leq\frac{1}{N}\sum_{i\in U}y_i^2<\infty\end{aligned}$$

toujours selon la condition C3. On peut déduire alors que

$$\mathrm{E}_p\left(\left\|\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{y}_s-\boldsymbol{B'}_U\boldsymbol{y}_U\right)\right\|^2\right)=O(n^{-1})$$

et par conséquent, $\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{y}_s-\boldsymbol{B'}_U\boldsymbol{y}_U\right)=O_p(n^{-1/2})$.

iii. On a $\frac{1}{N}\boldsymbol{B}'_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{y}_s=\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{y}_s-\boldsymbol{B'}_U\boldsymbol{y}_U\right)+\frac{1}{N}\boldsymbol{B'}_U\boldsymbol{y}_U$.

□

Lemme 3.6.4 *: Sous les conditions C1-C7, on a*

i. $\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s-\boldsymbol{B'}_U\boldsymbol{B}_U\right)=O_p(n^{-1/2})$.

ii. $\left(\frac{1}{N}\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s\right)^{-1}=O_p(K)$

Preuve :

i. On a $\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s-\boldsymbol{B'}_U\boldsymbol{B}_U\right)=\frac{1}{N}\sum_{i\in U}\boldsymbol{b}(x_i)\boldsymbol{b'}(x_i)\alpha_i$ ou $\alpha_i=\frac{I_i}{\pi_i}-1$.
Alors,

$$\left\|\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s-\boldsymbol{B'}_U\boldsymbol{B}_U\right)\right\|^2=\frac{1}{N^2}\sum_{i,k\in U}\mathrm{tr}[\boldsymbol{b'}(x_i)\boldsymbol{b}(x_i)\boldsymbol{b}(x_k)\boldsymbol{b'}(x_k)]\alpha_i\alpha_k.$$

Pour calculer l'espérance de la quantité ci-dessus, on va utiliser de nouveau le lemme (3.6.2) pour $\beta_i\beta_k=\mathrm{tr}[\boldsymbol{b'}(x_i)\boldsymbol{b}(x_i)\boldsymbol{b}(x_k)\boldsymbol{b'}(x_k)]$. Puisque $\mathrm{tr}[\boldsymbol{b'}(x_i)\boldsymbol{b}(x_i)\boldsymbol{b}(x_k)\boldsymbol{b'}(x_k)]=\sum_{j=1}^{q}B_j^2(x_i)\sum_{j=1}^{q}B_j^2(x_k)\leq 1$, (Burman 1991)

alors

$$\begin{aligned}\frac{1}{N}\sum_{i\in U}\beta_i^2 &= \frac{1}{N}\sum_{i\in U}\left(\sum_{j=1}^{q}B_j^2(x_i)\right)^2 < 1\\ \frac{1}{N^2}\sum_{i\neq k\in U}\beta_i\beta_k &= \frac{1}{N^2}\sum_{i\neq k\in U}\sum_{j=1}^{q}B_j^2(x_i)\sum_{j=1}^{q}B_j^2(x_k) < \frac{N(N-1)}{N^2} < 1\end{aligned}$$

On obtient par conséquent $\mathrm{E}_p\left\|\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s - \boldsymbol{B'}_U\boldsymbol{B}_U\right)\right\|^2 = O\left(n^{-1}\right)$ et $\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s - \boldsymbol{B'}_U\boldsymbol{B}_U\right) = O_p\left(n^{-1/2}\right)$.

ii. On a $\frac{1}{N}\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s = \frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s - \boldsymbol{B'}_U\boldsymbol{B}_U\right) + \frac{1}{N}\boldsymbol{B'}_U\boldsymbol{B}_U$ et en utilisant le lemme (3.6.1) et le premier point de ce lemme, on obtient que $\frac{1}{N}\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s = O_p(K^{-1})$. Par conséquent, $\left(\frac{1}{N}\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s\right)^{-1} = O_p(K)$.

□

Lemme 3.6.5 *:* $\frac{1}{N}\sum_{k\in U}\boldsymbol{b'}(x_k)\alpha_k = O_p(n^{-1/2})$.

Preuve : On a

$$\left\|\frac{1}{N}\sum_{k\in U}\boldsymbol{b'}(x_k)\alpha_k\right\|^2 = \frac{1}{N^2}\sum_{i,k\in U}\left(\sum_{j=1}^{q}B_j(x_i)B_j(x_k)\right)\alpha_i\alpha_k$$

et pour calculer son espérance, on applique de nouveau le lemme (3.6.2) avec $\beta_i\beta_k$ dans ce cas égal à $\sum_{j=1}^{q}B_j(x_i)B_j(x_k)$.

$$\begin{aligned}\frac{1}{N}\sum_{i\in U}\beta_i^2 &= \frac{1}{N}\sum_{i\in U}\sum_{j=1}^{q}B_j^2(x_i) < 1\\ \frac{1}{N^2}\sum_{i\neq k\in U}|\beta_i\beta_k| &< \frac{1}{N^2}\sum_{j=1}^{q}\sum_{k,i\in U}B_j(x_k)B_j(x_i) \leq \frac{1}{K} < \infty\end{aligned}$$

où la dernière inégalité résulte du lemme (3.6.1), point 2.

On obtient alors que $\mathrm{E}_p\left(\left\|\frac{1}{N}\sum_{k\in U}\boldsymbol{b'}(x_k)\alpha_k\right\|^2\right) = O(1/n)$ et

$$\frac{1}{N}\sum_{k\in U}\boldsymbol{b'}(x_k)\alpha_k = O_p(n^{-1/2}).$$

□

On donne dans la suite quelques propriétés concernant le paramètre de la régression par des fonctions B-splines.

Proposition 3.6.1 *: Sous les conditions de régularité C1-C7, on a*

$$\hat{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{\theta}} = O_p(K^{3/2} n^{-1/2}).$$

Preuve : On peut écrire $\hat{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{\theta}}$ de la manière suivante :

$$\begin{aligned}
\hat{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{\theta}} &= \left(\frac{1}{N}\boldsymbol{B'}_U\boldsymbol{B}_U\right)^{-1}\left[\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{y}_s - \boldsymbol{B'}_U\boldsymbol{y}_U\right)\right] \\
&- \left(\frac{1}{N}\boldsymbol{B'}_U\boldsymbol{B}_U\right)^{-1}\left[\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s - \boldsymbol{B'}_U\boldsymbol{B}_U\right)\right]\left(\frac{1}{N}\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s\right)^{-1}\left(\frac{1}{N}\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{y}_s\right).
\end{aligned}$$

On utilise alors les lemmes (3.6.1), (3.6.3) et (3.6.4) pour conclure. □

La propriété suivante donne l'approximation de $\hat{\boldsymbol{\theta}}$ par un developpement de Taylor.

Proposition 3.6.2 *: Sous les conditions C1-C7, on a*

$$\hat{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{\theta}} = \left(\frac{1}{N}\boldsymbol{T}_U\right)^{-1}\left(\frac{1}{N}\sum_{k\in s}\frac{\boldsymbol{b}(x_k)E_k}{\pi_k}\right) + O_p(K^{5/2}n^{-1})$$

où $\boldsymbol{T}_U = \boldsymbol{B'}_U\boldsymbol{B}_U$ *et* $E_k = y_k - \boldsymbol{b'}(x_k)\hat{\boldsymbol{\theta}}$ *pour tous* $k \in U$.

Preuve : On a

$$\begin{aligned}
\hat{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{\theta}} &= \hat{\boldsymbol{T}}_s^{-1}(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{y}_s - \hat{\boldsymbol{T}}_s\hat{\boldsymbol{\theta}}) \\
&= \hat{\boldsymbol{T}}_s^{-1}\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}(\boldsymbol{y}_s - \boldsymbol{B}_s\hat{\boldsymbol{\theta}}) \\
&= \left(\frac{1}{N}\hat{\boldsymbol{T}}_s\right)^{-1}\frac{1}{N}\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{E}_s \qquad (3.19)
\end{aligned}$$

avec $\hat{\boldsymbol{T}}_s = \boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s$ et $\boldsymbol{E}_s = (E_k)_{k\in s}$. De plus,

$$\frac{1}{N}\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{E}_s = \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}\boldsymbol{b}(x_k)E_k = O_p((K/n)^{1/2}) \qquad (3.20)$$

parce que

$$\frac{1}{N}\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{E}_s = \underbrace{\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{y}_s - \boldsymbol{B'}_U\boldsymbol{y}_U\right)}_{O_p(n^{-1/2})} - \underbrace{\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s - \boldsymbol{B'}_U\boldsymbol{B}_U\right)}_{O_p(n^{-1/2})}\underbrace{\hat{\boldsymbol{\theta}}}_{O(\sqrt{K})}$$

d'après les lemmes (3.6.1), (3.6.3) et (3.6.4).
Le lemme (3.6.4) nous donne $\frac{1}{N}(\hat{\boldsymbol{T}}_s - \boldsymbol{T}_U) = O_p(n^{-1/2})$ et $\left(\frac{1}{N}\hat{\boldsymbol{T}}_s\right)^{-1} = O_p(K)$ et le lemme (3.6.1) $\left(\frac{1}{N}\boldsymbol{T}_U\right)^{-1} = O(K)$; alors,

$$\left(\frac{1}{N}\hat{\boldsymbol{T}}_s\right)^{-1} - \left(\frac{1}{N}\boldsymbol{T}_U\right)^{-1} = O_p(K^2 n^{-1/2}). \tag{3.21}$$

Enfin, on combine (3.20) et (3.21) dans (3.19) et on obtient la conclusion. □

Cette propriété a comme conséquence immédiate :

Proposition 3.6.3 *: Sous les conditions C1-C7, on a*

$$\begin{aligned} \boldsymbol{G}_s\boldsymbol{f}_s - \boldsymbol{G}_U\boldsymbol{f}_U &= \boldsymbol{T}_U^{-1}\boldsymbol{B}'_s\boldsymbol{\Pi}_s^{-1}\tilde{\boldsymbol{E}}_s + O_p(K^{5/2}n^{-1}) \\ &= \boldsymbol{T}_U^{-1}\sum_{k\in s}\frac{\boldsymbol{b}(x_k)\tilde{E}_k}{\pi_k} + O_p(K^{5/2}n^{-1}) \end{aligned} \tag{3.22}$$

avec $\tilde{E}_k = f_k - \boldsymbol{b}'(x_k)\boldsymbol{G}_U\boldsymbol{f}_U$ *pour* $k \in U$.

Preuve : Procédant de la même manière que dans la proposition 3.6.2, on obtient

$$\begin{cases} \boldsymbol{G}_s\boldsymbol{f}_s - \boldsymbol{G}_U\boldsymbol{f}_U = \hat{\boldsymbol{T}}_s^{-1}\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\tilde{\boldsymbol{E}}_s \quad \text{avec} \\ \tilde{\boldsymbol{E}}_s = (\tilde{E}_k)_{k\in s}, \quad \tilde{E}_k = f_k - \boldsymbol{b'}(x_k)\boldsymbol{G}_U\boldsymbol{f}_U, \quad k \in U \end{cases} \tag{3.23}$$

Pour $\frac{1}{N}\sum_{k\in U} f_k^2 < \infty$, il résulte de la démonstration du lemme 3.6.3, deuxième point que $\frac{1}{N}\left(\boldsymbol{B'}_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{f}_s - \boldsymbol{B'}_U\boldsymbol{f}_U\right) = O_p(n^{-1/2})$. Cette relation plus $(\boldsymbol{B'}_U\boldsymbol{B}_U)^{-1}\boldsymbol{B'}_U\boldsymbol{f}_U = O(K^{1/2})$ donne la conclusion. □
Cela nous donne :

$$\begin{cases} \mathrm{E}_p(\boldsymbol{G}_s\boldsymbol{f}_s - \boldsymbol{G}_U\boldsymbol{f}_U) \simeq 0 \\ \mathrm{E}_p[(\boldsymbol{G}_s\boldsymbol{f}_s - \boldsymbol{G}_U\boldsymbol{f}_U)(\boldsymbol{G}_s\boldsymbol{f}_s - \boldsymbol{G}_U\boldsymbol{f}_U)'] \simeq \mathrm{Var}_p\left[\boldsymbol{T}_U^{-1}\sum_{k\in s}\frac{\boldsymbol{b}(x_k)\widetilde{E}_k}{\pi_k}\right] = \boldsymbol{T}_U^{-1}\widetilde{\boldsymbol{V}}\boldsymbol{T}_U^{-1} \\ \widetilde{\boldsymbol{V}} = (v_{jl})_{j,l=\overline{1,q}}, \quad v_{jl} = \sum_{i=1}^N\sum_{k=1}^N \frac{B_j(x_i)\widetilde{E}_i}{\pi_i}\frac{B_l(x_k)\widetilde{E}_k}{\pi_k}\Delta_{ik} \end{cases} \tag{3.24}$$

On peut donner maintenant les preuves des propositions 3.3.9, 3.3.10 et 3.3.12.

Preuve **de la proposition** 3.3.9 :

$$\begin{aligned} \frac{1}{N}\left(\hat{t}_{y\pi} - \hat{t}_{HT}\right) &= \frac{1}{N}\left\{\sum_{k\in s}\frac{y_k - \hat{\hat{f}}_k}{\pi_k} + \sum_{k\in U}\hat{\hat{f}}_k - \sum_s\frac{y_k}{\pi_k}\right\} \\ &= \left[\frac{1}{N}\sum_{k\in U}\boldsymbol{b}'(x_k)\alpha_k\right](\hat{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{\theta}}) + \left[\frac{1}{N}\sum_{k\in U}\boldsymbol{b}'(x_k)\alpha_k\right]\hat{\boldsymbol{\theta}} \end{aligned}$$

et on en déduit le résultat en utilisant la proposition 3.6.1 ci-dessus et les lemmes (3.6.1) et (3.6.5). □

***Preuve* de la proposition** 3.3.10.
On a

$$\frac{1}{N}\left(\hat{t}_{y\pi}-\hat{t}_y\right)=\left[\frac{1}{N}\sum_{k\in U}\boldsymbol{b}'(x_k)\alpha_k\right](\hat{\hat{\boldsymbol{\theta}}}-\hat{\boldsymbol{\theta}})$$

La proposition 3.6.1 montre que $\hat{\hat{\boldsymbol{\theta}}}-\hat{\boldsymbol{\theta}}=O_p(K^{3/2}n^{-1/2})$ et $\frac{1}{N}\sum_{k\in U}\boldsymbol{b}'(x_k)\alpha_k = O_p(n^{-1/2})$ d'après le lemme (3.6.5), d'où le résultat annoncé. □

***Preuve* de la proposition** 3.3.12 :
On change de nouveau l'ordre dans le calcul des espérances et on va appliquer la **proposition** 3.6.3 pour les vecteurs $\boldsymbol{G}_s\boldsymbol{f}_s$, $\boldsymbol{G}_s\boldsymbol{v}_s$ avec $\boldsymbol{v}_s=(v_k)_{k\in s}$ et $\boldsymbol{G}_s\boldsymbol{\varepsilon}_s$.
On note $\mathbf{P}=\boldsymbol{G}_s\boldsymbol{f}_s-\boldsymbol{G}_U\boldsymbol{f}_U$. Alors,

$$\begin{aligned}
&\frac{1}{N^2}E_\xi(\hat{t}_{y\pi}-t_y)^2=\\
=\;&\frac{1}{N^2}1'_N\boldsymbol{B}_U\left(\mathbf{P}\mathbf{P}'+\boldsymbol{G}_s\boldsymbol{V}_s\boldsymbol{G}'_s-\boldsymbol{G}_s(\boldsymbol{V}_s|0)\boldsymbol{G}'_U-\boldsymbol{G}_U(\boldsymbol{V}_s|0)'\boldsymbol{G}'_s+\boldsymbol{G}_U\boldsymbol{V}_U\boldsymbol{G}'_U\right)\boldsymbol{B}'_U\mathbf{1}_N\\
=\;&\frac{1}{N^2}(\mathbf{1}'_N\boldsymbol{B}_U\mathbf{P}\mathbf{P}'\boldsymbol{B}'_U\mathbf{1}_N+\mathbf{1}'_N\boldsymbol{B}_U\boldsymbol{G}_s\boldsymbol{V}_s\boldsymbol{G}'_s\boldsymbol{B}'_U\mathbf{1}_N-2\mathbf{1}'_N\boldsymbol{B}_U\boldsymbol{G}_s\boldsymbol{V}_s\mathbf{1}_s+\sum_{k\in U}v_k)
\end{aligned}$$

où $\boldsymbol{V}_U=Var_\xi(\boldsymbol{\varepsilon}_U)=\mathrm{diag}(v_k)_{k\in U}$ avec $\boldsymbol{V}_s$ sa restriction sur s.

$$\begin{aligned}
&\frac{1}{N^2}E_pE_\xi(\hat{t}_{y\pi}-t_y)^2=\\
=\;&\frac{1}{N^2}(\mathbf{1}'_N\boldsymbol{B}_UE_p(\mathbf{P}\mathbf{P}')\boldsymbol{B}'_U\mathbf{1}_N+\mathbf{1}'_N\boldsymbol{B}_UE_p(\boldsymbol{G}_s\boldsymbol{V}_s\boldsymbol{G}'_s)\boldsymbol{B}'_U\mathbf{1}_N-2\mathbf{1}'_N\boldsymbol{B}_UE_p(\boldsymbol{G}_s\boldsymbol{V}_s\mathbf{1}_s))+\\
&+\frac{1}{N^2}\sum_{k\in U}v_k.
\end{aligned}$$

On a besoin de calculer $E_p(\mathbf{P}\mathbf{P}')$, $E_p(\boldsymbol{G}_s\boldsymbol{V}_s\boldsymbol{G}'_s)$ et $E_p(\boldsymbol{G}_s\boldsymbol{V}_s\mathbf{1}_s)$.

A : Pour le premier terme, on utilise (3.24),

$$\mathrm{E}_p[\mathbf{P}\mathbf{P}']\simeq\mathrm{Var}_p\left[\boldsymbol{T}_U^{-1}\sum_{k\in s}\frac{\boldsymbol{b}(x_k)\widetilde{E}_k}{\pi_k}\right]=\boldsymbol{T}_U^{-1}\widetilde{\boldsymbol{V}}\boldsymbol{T}_U^{-1}$$

avec

$$\widetilde{\boldsymbol{V}}=(v_{jl})_{j,l=\overline{1,q}},\quad v_{jl}=\sum_{i=1}^N\sum_{k=1}^N\frac{B_j(x_i)\widetilde{E}_i}{\pi_i}\frac{B_l(x_k)\widetilde{E}_k}{\pi_k}\Delta_{ik}.$$

Par ailleurs,

$$\begin{aligned}
\frac{1}{N^2}v_{jl} &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{B_j(x_i)\widetilde{E}_i}{\pi_i}\frac{B_l(x_k)\widetilde{E}_k}{\pi_k}\Delta_{ik} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}B_j(x_i)B_l(x_i)\widetilde{E}_i^2\frac{1-\pi_i}{\pi_i} + \frac{1}{N^2}\sum_{i\neq k=1}^{N}B_j(x_i)B_l(x_k)\widetilde{E}_i\widetilde{E}_k\frac{\Delta_{ik}}{\pi_i\pi_k} \\
&\leq \frac{1}{N^2}\frac{1-\lambda}{\lambda}\sum_{i=1}^{N}\widetilde{E}_i^2 + \frac{1}{N^2}\frac{\max_{i\neq k}|\Delta_{ik}|}{\lambda^2}\sum_{i\neq k=1}^{N}|\widetilde{E}_i\widetilde{E}_k| \\
&\leq \frac{1}{N^2}\frac{1-\lambda}{\lambda}\sum_{i=1}^{N}\widetilde{E}_i^2 + \frac{1}{N^2}\frac{\max_{i\neq k}|\Delta_{ik}|}{\lambda^2}\left(\sum_{i=1}^{N}|\widetilde{E}_i|\right)^2 \\
&\leq \frac{1}{N}\frac{1-\lambda}{\lambda}\left(\frac{1}{N}\sum_{i=1}^{N}\widetilde{E}_i^2\right) + \frac{1}{n}\frac{n\max_{i\neq k}|\Delta_{ik}|}{\lambda^2}\left(\frac{1}{N}\sum_{i=1}^{N}\widetilde{E}_i^2\right).
\end{aligned}$$

On a, sous la condition de régularité C.2 et pour $K \approx N^{\frac{1}{2m+1}}$ (Zhou, Shen & Wolfe 1998), que $\tilde{E}_k = f_k - \boldsymbol{b}'(x_k)\boldsymbol{G}_U\boldsymbol{f}_U = f_k - E_\xi(\hat{f}_k) = O(K^{-m})$ uniformément en $k \in U$. Cela implique avec les conditions C.1 et C.4 pour $\frac{1}{N^2}v_{jl}$,

$$\frac{1}{N^2}v_{jl} = O\left(\frac{1}{nK^{2m}}\right).$$

Alors,

$$\begin{aligned}
\frac{1}{N^2}\mathbf{1}_N'\boldsymbol{B}_U\mathrm{E}_p(PP')\boldsymbol{B}_U'\mathbf{1}_N &\simeq \frac{1}{N^2}\mathbf{1}_N'\boldsymbol{B}_U\boldsymbol{T}_U^{-1}\widetilde{\boldsymbol{V}}\boldsymbol{T}_U^{-1}\boldsymbol{B}'\mathbf{1}_N \\
&= \sum_{j=1}^{q}\sum_{l=1}^{q}c_jc_l\left(\frac{1}{N^2}v_{jl}\right) \qquad (3.25)
\end{aligned}$$

où les c_j pour $j = 1, \ldots, q$ sont les éléments du vecteur $\mathbf{1}_N'\boldsymbol{B}_U\boldsymbol{T}_U^{-1}$. Soient $\boldsymbol{T}_U^{-1} = (a_{lj})_{l,j=1}^q$ et $\boldsymbol{a}_j = (a_{1j}, \ldots, a_{qj})'$ les q colonnes de $\boldsymbol{T}_U^{-1}$. Burman (1991) donne une majoration des éléments de la matrice $\boldsymbol{T}_U^{-1}$:

$$|a_{lj}| \leq \tilde{c}_1\frac{K}{N}\rho^{|l-j|} \quad \text{pour } 0 < \rho < 1$$

où $\tilde{c}_1$ est une constante. Le dénominateur N apparait parce que dans notre matrice $\boldsymbol{T}_U$ les éléments ne sont pas divisés par N comme dans Burman (1991). Alors $c_j = \sum_{i=1}^N \boldsymbol{b}'(x_i)\boldsymbol{a}_j$ et

$$||\boldsymbol{a}_j|| \leq \tilde{c}_1\frac{K}{N}\left(\sum_{l=1}^{q}\rho^{2|l-j|}\right)^{1/2} \qquad (3.26)$$

Cela implique pour c_j :

$$\begin{aligned} |c_j| &\leq \left\| \sum_{i\in U} \boldsymbol{b}'(x_i) \right\| \|\boldsymbol{a}_j\| \leq \left\| \frac{1}{N} \sum_{i\in U} \boldsymbol{b}'(x_i) \right\| \tilde{c}_1 K \left(\sum_{l=1}^{q} \rho^{2|l-j|} \right)^{1/2} \\ &\leq ct\tilde{c}_1 K^{1/2} \left(\sum_{l=1}^{q} \rho^{2|l-j|} \right)^{1/2} \end{aligned} \tag{3.27}$$

d'après le lemme (3.6.1), deuxième point. Alors,

$$\begin{aligned} \left| \sum_{j=1}^{q} \sum_{l=1}^{q} c_j c_l \left(\frac{1}{N^2} v_{jl} \right) \right| &\leq ct \frac{K}{nK^{2m}} \left(\sum_{j=1}^{q} \sum_{l=1}^{q} \left(\sum_{l_1=1}^{q} \rho^{2|l_1-j|} \right)^{1/2} \left(\sum_{l_2=1}^{q} \rho^{2|l_2-l|} \right)^{1/2} \right) \\ &= ct \frac{K}{nK^{2m}} \left(\sum_{l=1}^{q} \left(\sum_{j=1}^{q} \rho^{2|j-l|} \right)^{1/2} \right)^2 \end{aligned}$$

et après avoir appliqué l'inégalité de Cauchy-Shwartz, on obtient

$$\left| \sum_{j=1}^{q} \sum_{l=1}^{q} c_j c_l \left(\frac{1}{N^2} v_{jl} \right) \right| \leq ct \frac{Kq}{nK^{2m}} \sum_{j=1}^{q} \sum_{l=1}^{q} \rho^{2|j-l|}.$$

On a

$$\begin{aligned} \sum_{j=1}^{q} \sum_{l=1}^{q} \rho^{2|j-l|} &= \frac{2}{\rho^2-1} \left[(\rho^2)^q + (\rho^2)^{q-1} + (\rho^2)^{q-2} + \ldots + 1 - q - 1 \right] \\ &= \frac{2}{\rho^2-1} \left[\frac{(\rho^2)^{q+1}-1}{\rho^2-1} - q - 1 \right] \end{aligned}$$

et comme $q = K + m$, le terme prédominant dans $\left| \sum_{j=1}^{q} \sum_{l=1}^{q} c_j c_l \left(\frac{1}{N^2} v_{jl} \right) \right|$ est $\frac{K^3}{nK^{2m}}$ qui tend vers zéro pour $K \approx N^{\frac{1}{2m+1}}$.
On a obtenu alors que

$$\frac{1}{N^2} \mathbf{1}_N' \boldsymbol{B}_U \mathrm{E}_p(PP') \boldsymbol{B}_U' \mathbf{1}_N \to 0. \tag{3.28}$$

B : Calculons maintenant $E_p(\mathbf{1}_s' \boldsymbol{V}_s \boldsymbol{G}_s')$. On a $\mathbf{1}_s' \boldsymbol{V}_s = (v_1, \ldots, v_n) = \boldsymbol{v}_s'$ et alors $\boldsymbol{v}_U' = (v_1, \ldots, v_N)$. On utilise la proposition 3.6.3 pour $\frac{1}{N} \sum_{k\in U} v_k^2 < \infty$, alors

$$\left\{ \begin{array}{rcl} \boldsymbol{G}_s \boldsymbol{v}_s - \boldsymbol{G}_U \boldsymbol{v}_U &=& \boldsymbol{T}_U^{-1} \sum_{k\in s} \frac{\boldsymbol{b}(x_k) \bar{E}_k}{\pi_k} + O_p(K^{5/2} n^{-1}) \\ \bar{E}_k &=& v_k - \boldsymbol{b}'(x_k) \boldsymbol{G}_U \boldsymbol{v}_U, \quad k = 1, \ldots, N \end{array} \right. \tag{3.29}$$

et $E_p(\boldsymbol{G}_s\boldsymbol{v}_s)' \simeq (\boldsymbol{G}_U\boldsymbol{v}_U)'$. Cela nous donne

$$\mathbf{1}'_N\boldsymbol{B}_U E_p(\boldsymbol{G}_s\boldsymbol{V}_s\mathbf{1}_s) \simeq \mathbf{1}'_N\boldsymbol{B}_U\boldsymbol{G}_U\boldsymbol{v}_U = \mathbf{1}'_N\boldsymbol{v}_U = \sum_{k\in U} v_k \tag{3.30}$$

d'après (3.10).

C : Il reste à calculer $E_p(\boldsymbol{G}_s\boldsymbol{V}_s\boldsymbol{G}'_s)$. Ce terme peut être écrit différement si on tient compte du fait que $\boldsymbol{V}_s = E_\xi(\boldsymbol{\varepsilon}_s\boldsymbol{\varepsilon}'_s)$ et que le plan de sondage p est noninformatif, c'est à dire qu'on peut changer l'ordre entre E_p et E_ξ

$$E_p(\boldsymbol{G}_s\boldsymbol{V}_s\boldsymbol{G}'_s) = E_pE_\xi(\boldsymbol{G}_s\boldsymbol{\varepsilon}_s\boldsymbol{\varepsilon}'_s\boldsymbol{G}'_s) = E_\xi E_p(\boldsymbol{G}_s\boldsymbol{\varepsilon}_s(\boldsymbol{G}_s\boldsymbol{\varepsilon}_s)').$$

Pour calculer $E_p(\boldsymbol{G}_s\boldsymbol{\varepsilon}_s(\boldsymbol{G}_s\boldsymbol{\varepsilon}_s)')$ on va utiliser de nouveau la proposition 3.6.3 pour $\dfrac{1}{N}\sum_{k\in U}\varepsilon_k^2 < \infty$. On obtient

$$\left\{\begin{array}{rcl} \boldsymbol{G}_s\boldsymbol{\varepsilon}_s - \boldsymbol{G}_U\boldsymbol{\varepsilon}_U &=& \boldsymbol{T}_U^{-1}\displaystyle\sum_{k\in s}\frac{\boldsymbol{b}(x_k)\bar{F}_k}{\pi_k} + O_p(K^{5/2}n^{-1}) \\ \bar{F}_k &=& \varepsilon_k - \boldsymbol{b}'(x_k)\boldsymbol{G}_U\boldsymbol{\varepsilon}_U, \quad k=1,\ldots,N \end{array}\right. \tag{3.31}$$

avec

$$\left\{\begin{array}{l} \mathrm{Var}(\boldsymbol{G}_s\boldsymbol{\varepsilon}_s) \simeq \boldsymbol{T}_U^{-1}\bar{\boldsymbol{F}}\boldsymbol{T}_U^{-1}, \quad \bar{\boldsymbol{F}} = (\bar{f}_{jl})_{j,l=1}^q \\ \bar{f}_{jl} = \displaystyle\sum_{i\in U}\sum_{k\in U}\frac{B_j(x_i)\bar{F}_i}{\pi_i}\frac{B_l(x_k)\bar{F}_k}{\pi_k}\Delta_{ik}. \end{array}\right. \tag{3.32}$$

On note $\check{\boldsymbol{\Delta}}$ la matrice de variance-covariance de $\boldsymbol{\alpha} = (\alpha_i)_{i\in U}$, $\alpha_i = \frac{I_i}{\pi_i} - 1$; $\check{\boldsymbol{\Delta}} = \left(\dfrac{\Delta_{ki}}{\pi_k\pi_i}\right)_{k,i\in U}$. Alors, la matrice de variance-covariance $\bar{\boldsymbol{F}}$ des éléments $\bar{f}_{jl}$ peut s'écrire de manière équivalente

$$\bar{\boldsymbol{F}} = \boldsymbol{B}'_U \times \boldsymbol{H} \times \boldsymbol{B}_U$$

avec $\boldsymbol{H} = \mathrm{diag}(\bar{F}_k)_{k\in U} \times \check{\boldsymbol{\Delta}} \times \mathrm{diag}(\bar{F}_k)_{k\in U}$.

$$\begin{aligned} & \frac{1}{N^2}\mathbf{1}'_N\boldsymbol{B}_U E_p(\boldsymbol{G}_s\boldsymbol{V}_s\boldsymbol{G}'_s)\boldsymbol{B}'_U\mathbf{1}_N = \frac{1}{N^2}E_\xi\left\{\mathbf{1}'_N\boldsymbol{B}_U E_p[(\boldsymbol{G}_s\boldsymbol{\varepsilon}_s)(\boldsymbol{G}_s\boldsymbol{\varepsilon}_s)']\boldsymbol{B}'_U\mathbf{1}_N\right\} \\ = & \frac{1}{N^2}E_\xi\left\{\mathbf{1}'_N\boldsymbol{B}_U\mathrm{Var}(\boldsymbol{G}_s\boldsymbol{\varepsilon}_s)\boldsymbol{B}'_U\mathbf{1}_N + \underbrace{\mathbf{1}'_N\boldsymbol{B}_U\boldsymbol{G}_U}_{\mathbf{1}'_N}\boldsymbol{\varepsilon}_U\boldsymbol{\varepsilon}'_U\underbrace{\boldsymbol{G}'_U\boldsymbol{B}'_U\mathbf{1}_N}_{\mathbf{1}_N}\right\} \\ \simeq & \frac{1}{N^2}E_\xi\left\{\mathbf{1}'_N\boldsymbol{B}_U\boldsymbol{T}_U^{-1}\bar{\boldsymbol{F}}\boldsymbol{T}_U^{-1}\boldsymbol{B}'_U\mathbf{1}_N + \mathbf{1}'_N\boldsymbol{\varepsilon}_U\boldsymbol{\varepsilon}'_U\mathbf{1}_N\right\} \\ = & \frac{1}{N^2}E_\xi\left\{\underbrace{\mathbf{1}'_N\boldsymbol{B}_U\boldsymbol{T}_U^{-1}\boldsymbol{B}'_U}_{\mathbf{1}'_N}\boldsymbol{H}\boldsymbol{B}_U\boldsymbol{T}_U^{-1}\boldsymbol{B}'_U\mathbf{1}_N\right\} + \frac{1}{N^2}\sum_{k\in U}v_k \\ = & \frac{1}{N^2}E_\xi(\mathbf{1}'_N\boldsymbol{H}\mathbf{1}_N) + \frac{1}{N^2}\sum_{k\in U}v_k \end{aligned}$$

$$\mathbf{1}'_N \boldsymbol{H} \mathbf{1}_N = (\bar{F}_1, \ldots, \bar{F}_N)\check{\boldsymbol{\Delta}}(\bar{F}_1, \ldots, \bar{F}_N)' = \bar{\boldsymbol{F}}'_U \check{\boldsymbol{\Delta}} \bar{\boldsymbol{F}}_U$$

pour $\boldsymbol{F}_U = (\bar{F}_1, \ldots, \bar{F}_N)'$. De plus, $\boldsymbol{F}_U = (\boldsymbol{I}_N - \boldsymbol{B}_U \boldsymbol{G}_U)\boldsymbol{\varepsilon}_U$ où $\boldsymbol{I}_N$ est la matrice unité d'ordre N. Cela implique que $\boldsymbol{F}_U = \boldsymbol{R}_B \boldsymbol{\varepsilon}_U$ où $\boldsymbol{R}_B$ est le projecteur sur l'espace orthogonal à l'espace engendré par les fonctions des B-splines. Alors,

$$\begin{aligned} E_\xi(\mathbf{1}'_N \boldsymbol{H} \mathbf{1}_N) &= E_\xi(\boldsymbol{\varepsilon}'_U \boldsymbol{R}'_B \check{\boldsymbol{\Delta}} \boldsymbol{R}_B \boldsymbol{\varepsilon}_U) = tr(\boldsymbol{R}'_B \check{\boldsymbol{\Delta}} \boldsymbol{R}_B \boldsymbol{V}_U) \\ &= tr(\boldsymbol{V}_U \check{\boldsymbol{\Delta}}) - tr(\boldsymbol{V}_U \check{\boldsymbol{\Delta}} \boldsymbol{B}_U \boldsymbol{G}_U) \\ &- tr(\boldsymbol{B}_U \boldsymbol{G}_U \check{\boldsymbol{\Delta}} \boldsymbol{V}_U) + tr(\boldsymbol{B}_U \boldsymbol{G}_U \check{\boldsymbol{\Delta}} \boldsymbol{B}_U \boldsymbol{G}_U \boldsymbol{V}_U) \end{aligned}$$

On a $0 \leq tr(\boldsymbol{V}_U \check{\boldsymbol{\Delta}} \boldsymbol{B}_U \boldsymbol{G}_U) \leq tr(\boldsymbol{V}_U \check{\boldsymbol{\Delta}})$, $0 \leq tr(\boldsymbol{B}_U \boldsymbol{G}_U \check{\boldsymbol{\Delta}} \boldsymbol{V}_U) \leq tr(\check{\boldsymbol{\Delta}} \boldsymbol{V}_U) = tr(\boldsymbol{V}_U \check{\boldsymbol{\Delta}})$ et $0 \leq tr(\boldsymbol{B}_U \boldsymbol{G}_U \check{\boldsymbol{\Delta}} \boldsymbol{B}_U \boldsymbol{G}_U \boldsymbol{V}_U) \leq tr(\boldsymbol{V}_U \check{\boldsymbol{\Delta}})$ parce que $\boldsymbol{V}_U \check{\boldsymbol{\Delta}}$ et $\boldsymbol{B}_U \boldsymbol{G}_U = \boldsymbol{B}_U (\boldsymbol{B}'_U \boldsymbol{B}_U)^{-1} \boldsymbol{B}'_U$ sont deux matrices non négatives (respectivement matrice de variance-covariance et projecteur) et la plus petite valeur propre de la matrice $\boldsymbol{B}_U \boldsymbol{G}_U$ est 0 et la plus grande valeur propre est 1, c'est un projecteur. On a utilisé le résultat suivant $\lambda^A_{min} tr(\boldsymbol{B}) \leq tr(\boldsymbol{AB}) \leq \lambda^A_{max} tr(\boldsymbol{B})$ pour $\boldsymbol{A}, \boldsymbol{B}$ matrices non-négatives et $\lambda^A_{min}, \lambda^A_{max}$ est la plus petite, respectivement la plus grande valeur propre de $\boldsymbol{A}$ (Zhou, Shen & Wolfe 1998). Ensuite,

$$tr(\boldsymbol{V}_U \check{\boldsymbol{\Delta}}) = \sum_{k \in U} v_k \check{\Delta}_{kk} = \sum_{k \in U} v_k \frac{1 - \pi_k}{\pi_k} = \sum_{k \in U} \frac{v_k}{\pi_k} - \sum_{k \in U} v_k$$

ce qui donne comme expression pour $\frac{1}{N^2} \mathbf{1}'_N \boldsymbol{B}_U E_p(\boldsymbol{G}_s \boldsymbol{V}_s \boldsymbol{G}'_s) \boldsymbol{B}'_U \mathbf{1}_N$,

$$\frac{1}{N^2} \mathbf{1}'_N \boldsymbol{B}_U E_p(\boldsymbol{G}_s \boldsymbol{V}_s \boldsymbol{G}'_s) \boldsymbol{B}'_U \mathbf{1}_N \simeq \frac{1}{N^2} \left(\sum_{k \in U} \frac{v_k}{\pi_k} - \sum_{k \in U} v_k \right) + \frac{1}{N^2} \sum_{k \in U} v_k$$

parce que $0 \leq \frac{1}{N^2} tr(\boldsymbol{V}_U \check{\boldsymbol{\Delta}} \boldsymbol{B}_U \boldsymbol{G}_U) \leq \frac{1}{N^2} \sum_{k \in U} v_k \frac{1 - \pi_k}{\pi_k} \leq \frac{1 - \lambda}{\lambda} \frac{1}{N^2} \sum_{k \in U} v_k \to 0$; la même chose pour $tr(\boldsymbol{B}_U \boldsymbol{G}_U \check{\boldsymbol{\Delta}} \boldsymbol{V}_U)$ et $tr(\boldsymbol{B}_U \boldsymbol{G}_U \check{\boldsymbol{\Delta}} \boldsymbol{B}_U \boldsymbol{G}_U \boldsymbol{V}_U)$. Alors,

$$\frac{1}{N^2} \mathbf{1}'_N \boldsymbol{B}_U E_p(\boldsymbol{G}_s \boldsymbol{V}_s \boldsymbol{G}'_s) \boldsymbol{B}'_U \mathbf{1}_N \simeq \frac{1}{N^2} \sum_{k \in U} \frac{v_k}{\pi_k}. \tag{3.33}$$

On peut donner l'expression finale de $\frac{1}{N^2} E_\xi E_p(\hat{t}_{y\pi} - t_y)^2$ en utilisant les relations (3.28), (3.30) et (3.33) :

$$\begin{aligned} \frac{1}{N^2} E_\xi E_p(\hat{t}_{y\pi} - t_y)^2 &\simeq \frac{1}{N^2} \left\{ \sum_{k \in U} \frac{v_k}{\pi_k} - 2 \sum_{k \in U} v_k + \sum_{k \in U} v_k \right\} \\ &= \frac{1}{N^2} \left\{ \sum_{k \in U} \frac{v_k}{\pi_k} - \sum_{k \in U} v_k \right\}. \end{aligned}$$

□

Bibliographie

Agarwal, G. and Studden, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *The Annals of Statistics*, vol. 8, no. 6, 1307-1325.

Basu, D. (1958). On sampling with and without replacement. *Sankhya* 20, 287-294.

Bell, P. (2001). Comparaison d'autres estimateurs pour l'Enquête sur la population active. *Technique d'enquête*, 27, 1, pp. 57-68.

Blight, B. J. N. and Scott, A. J. (1973). A stochastic model for repeated surveys. *J. R. Statist. Soc., B*, 35, 61-66.

Breidt, F. J. , Opsomer, J. D. (2000). Local Polynomial Survey Regression Estimators in Survey Sampling. *The Annals of Statistics* 28, No. 4, 1026-1053.

Brewer, K. R. W. (1963). Ratio estimation and finite populations : Some results deducible from the assumption of an underlying stochastic process. *Aust. J. Statist.* 5, 93-105.

Brewer, K. R. W. (1975). A simple procedure for $\pi pswor$. *Aust. J. Statist.* 17, 166-172.

Brewer, K. R. W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*. New-York, Springer-Verlag.

Burman, P. (1981). Regression Function Estimation from Dependent Observations. *Journal of Multivariate Analysis* 36, 263-279.

Cochran, W. G. (1977). *Sampling Techniques*. 3ème édition, New-Yok, Wiley.

Caron, N., (1996). Les principales techniques de la correction de la non-réponse et les modèles associées. *Document de travail INSEE de la Direction des Statistiques Démographiques et Sociales N°9604.*

Caron, N., Deville, J.C., Sautory, O. (1998). Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE. *Document de travail INSEE de la Direction des Statistiques Démographiques et Sociales N°9806.*

Caron N., and Ravalet, P. (2000) Estimation dans les enquêtes répétées : Application à l'Enquête Emploi en continu. *Document de travail INSEE de la Direction des Statistiques Démographiques et Sociales N°0005.*

Cassel, C.M., Särndal, C.E., and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 3, pp. 615-20.

Chambers, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *J. Official Statist.* 12, 3-32.

Chambers, R. L. , Dorfman, A. H. and Wherly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association* 88, 268-277.

Cochran, W. G. (1939). The use of analysis of variance in enumeration by sampling. *J. Amer. Statist. Ass.* 34, 492-510.

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Statist.*17, 164-177.

Cotton, F. , Hesse, C. (1992). Tirages coordonnes d'échantillons. *Document de travail INSEE de la Direction des Statistiques Economiques E9206.*

Deming, W.E. and Stephan, F. (1941). On the interpretation of census as samples. *J. Amer. Statist. Ass.* 36, 45-49.

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

Deville, J.C. (2000). Generalized calibration and application to weighting for non-response. *Document Compstat, Utrecht ; Physica-Verlag*, pp. 65-76.

Deville, J. C., Goga, C. (2002). The Horvitz-Thompson theory for two samples. preprint.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of American Statistical Association*, 87, 418, pp. 376-382.

Dierckx, P. (1993). *Curves and surface fitting with splines.* Oxford, Clarendon Press.

Dorfman, A. H. (1992). Nonparametric regression for estimating totals in finite populations. *Proceedings of the Section on Survey Research Methods* 622-625. Amer. Statist. Assoc., Alexandria, VA.

Dorfman, A. H. and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Ann. Statist.* 21, 1452-1475.

Eckler, A. R. (1955). Rotation sampling. *Annals of Mathematical Statistics*, 26, 664-85.

Fan, J. (1992). Design-adaptive nonparametric regression. J. Amer. Statist. Assoc. 87,998-1004.

Fan, J. and Gibels, I.(1996) *Local Polynomial Modeling and Its Applications.* Chapman and Hall, London.

Fuller, W. (2002). Regression estimation for survey samples. *Survey Methodolgy*, 28, 5-23.

Fuller, W. (1996). *An introduction to statistical time series.* 2nd, Willey Ney-York.

Fuller, W., Rao, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquêtes*, 27, 1, pp. 49-56.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *J. R. Statist. Soc. B* 17, 269-278.

Godambe, V. P. (1995). Estimation of parameters in survey sampling : optimality. *Canadian Journal of Statistics* 23, 227-243.

Godambe, V. P. and Joshi, V. M. (1965). Admissibility and Bayes estimation in sampling finite populations. 1. *Annals of Mathematical Statistics*, 36, 1707-1722.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population : their relationships and estimation. *International Statistical Review* 54, 127-138.

Gourieroux, C. and Roy, G. (1978). Enquêtes en deux vagues : renouvellement de l'échantillon. *Annales de l'INSEE*, 29, 115-144.

Gurney, M. and Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 247-257.

Hajek, J. (1981). *Sampling from a Finite Population.* Marcel Dekker, New York.

Hanif, M. and Brewer, K. R. W. (1980). Sampling with unequal probabilities without replacement :a review. *International Statistical Review*, 48, 317-335.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.* 14, 333-362.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hidiroglu, M. A. (2001). Double sampling. *Survey methodology*, to appear.

Horvitz, D. G. and Thompson, D. J. (1952). A generalisation of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, 663-685.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 377, pp. 89-96.

Keyfitz, N. (1957). Estimates of Sampling Variance Where Two Units Are Selected From Each Stratum. *Journal of the American Statistical Association*, 52, (December 1957), 503-10.

Kim, J.Y., Breidt, F.J. and Opsomer, J.D. (2003). Nonparametric regression estimation of finite population totals under two-stage sampling. *preprint.*

Kish, L. (1965). *Survey Sampling.* New-York, Wiley.

Kott, P.S. (1990). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference*, 24, pp. 287-96.

Kuo, L. (1998). Classical and prediction approaches to estimating distribution functions from survey data. *Proceedings of the Section on Survey Research Methods* 280-285. Amer. Statist. Assoc., Alexandria, VA.

Lundström, S. (1997). Calibration as a standard method for treatement of non-response. Doctoral dissertation, Stockholm University.

Lundström, S. & Särndal, C. E. (1999). Calibration as Standard Method for Treatment of nonresponse. *Journal of Official Statistics*, 15,(2), 305-327.

Lundström, S. & Särndal, C. E (2002). *Estimation in the Presence of Nonresponse and Frame imperfections.* Statistics Sweeden, 172 p.

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-sample surveys. *International Statistical Review*, 55, 2, pp. 191-202.

Madow, W. G. and Madow, L. H. (1944). On the theory of systematic sampling. *Ann. Math. Statist.* 15, 1-24.

Madow, W. G. (1949). On the theory of systematic sampling, II. *Ann. Math. Statist.*, 20, 333-354.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, 4, 502-527.

Patterson, H.D. (1950). Sampling on Sucessive Occasions with Partial Replacement of Units *Journal of the Royal Statistical Society*, B, 12, 2, 241-255.

Raj, Des (1966). Some Remarks on a Simple Procedure of Sampling Without Replacement *Journal of the American Statistical Association*, 61, 314, 391-396.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 2, pp. 153-165.

Rao, J.N.K. and Graham, J.E. (1964). Rotation Designs for Sampling on Repeated Occasions *Journal of the American Statistical Association*, 59, 306, 492-509.

Rao, J.N.K., Hartley, H. O. and Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, B, 24, 482-491.

Robinson, P.M. and Särndal, C.E. (1983). Asymptotic property of the generalized regression estimator in probability sampling. *Sankhya : The Indian Journal of Statistics*, 45, B, 240-248.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Royall, R. M. (1971). Linear regression models in finite population sampling theory. In V.P. Godambe and D.A. Sprott, Eds., *Foundations of statistical inference.* Torento : Holt, Rinehart & Winston, 259-274.

Royall, R. M. ,Cumberland, W. G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association* 73, 351-358.

Salamin, P. A. (2002) Introduction to the problem of sampling co-ordination. *notes de cours.*

Särndal, C.E. (1980). On the π-inverse weighting best linear unbiased weighting in probability sampling. *Biometrika* , 67, 3 pp. 639-50.

Särndal, C.E. , Swensson, B. and Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 3, pp. 527-37.

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

Schumaker, L. L. (1981). *Spline Functions : Basic Theory.* Wiley, New-York.

Scott, A. J. and Smith, T. M. F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *J. Indian Soc. Agric. Statist.* 5, 119-127.

Singh, A. C., Kennedy, B. and Wu, S. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel. *Techniques d'enquêtes.* 27,1, pp. 35-48.

Sunter, A. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 3, 261-268.

Sunter, A. (1986). Solutions to the problem of unequal probabilities without replacement. *International Statistical Review*, 54, 1, 33-50.

Tam, S.M. (1984). On Covariances From Overlapping Samples. *The American Statistician*, 38, 4, pp.288-289.

Tillé, Y. (2000). *Théorie des sondages.* Notes de cours, ENSAI.

Tillé, Y. (2001). *Théorie des sondages : échantillonnage et estimation en populations finies.* Dunod, Paris.

Théberge, A. (1999). Extensions of Calibration Estimators in Survey Sampling ; *Journal of the American Statistical Association*, vol. 94, no. 446.

Thompson, M. E. (1997). *Theory of Sample Surveys.* Chapman & Hall.

Wolter, K.M. (1985). *Introduction to variance estimation.* Springer-Verlag.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 334, pp. 411-414.

Wu, C. and Sitter, R.R. (2001). A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data. *Journal of the American Statistical Association*, vol. 96, no. 453, 2001.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *J. R. Statist. Soc. B* 15, 253-261.

Zheng, H., Little, R.J.A (2003). Penalized Spline Model-Based Estimation of the Finite Populations Total from Probability-Proportional-to-Size Samples. *J. of Official Statistics*, to appear.

Zhou, S., Shen, X. and Wolfe, D. A. (1998). Local Asymptotics For Regression Splines And Confidence Regions. *The Annals of Statistics*, vol. 26, no. 5, 1760-1782.

Résumé

Cette thèse est consacrée à l'estimation de la variance et à l'utilisation de variables auxilaires lorsqu'on dispose de plusieurs échantillons. Le premier chapitre, extrait d'un rapport rédigé pour Eurostat, présente une revue bibliographique sur les méthodes d'estimation de la variance en sondage ainsi que leur implémentation dans le logiciel Poulpe lorsque cela est possible. Nous définissons dans le chapitre 2 les plans de sondage bidimensionnels et nous donnons une formule de la variance de type Horvitz-Thompson pour des estimateurs linéaires qui dépendent des deux échantillons. Nous exhibons le meilleur estimateur linéaire sans biais et proposons un estimateur de sa variance. Des formules explicites sont données pour certains plans bidimensionnels. Une technique de linéarisation sur deux échantillons est proposée afin de pouvoir estimer la variance de fonctions non linéaires de totaux. Ensuite, nous développons un modèle nonparamétrique basé sur les polynômes locaux permettant de tenir compte de l'information auxiliaire. Enfin, nous proposons dans le chapitre 3 une nouvelle approche nonparamétrique basée sur les splines de régression. Nous prouvons la convergence de l'estimateur proposé et validons, sur des simulations, son bon comportement dans la pratique.

English Title : Variance estimation in multi-occasions sampling and nonparametric regression estimation in order to take auxiliary information into account

Abstract

This Phd deals with the variance estimation and the use of auxiliary information when we have more than one sample. The first chapter, which is a part of a report made for Eurostat, gives a review of the main techniques for estimating the variance and their implementation in the software POULPE, when it is possible. We define, in chapter 2, a bidimensional sampling design and we give a variance Horvitz-Thompson type formula for linear estimators that depend on these two samples. We derive the best linear unbiased estimator and a variance estimator is also proposed. Explicit formulas are given for particular bidimensional sampling designs. A technique of linearization on two samples is developped in order to deal with non linear statistics of totals. Then, we use auxiliary information for improving estimates by means of nonparametric regression estimators based on local polynomials. Finally, we study in chapter 3 a new approach based on regression splines for taking into account the auxiliary information for estimating the poulation total. Consistency results are proved and simulations confirm the good behaviour of this estimator in practice.

Discipline : Mathématiques Appliquées, Statistique.

Mots-clés : échantillons multiples, enquêtes répétées, estimateurs de Horvitz-Thompson, fonctions splines, fonction d'influence, linéarisation, polynômes locaux, convergence.

Ecole doctorale des Sciences Humaines et Sociales. Université Rennes 2, Haute Bretagne.