Université de Rennes 1

Mémoire pour l'obtention du diplôme

Habilitation à Diriger des Recherches

Discipline: Mathématiques

Dr. Guillaume CHAUVET

# Some contributions to Sampling and Estimation in Surveys

Soutenue le 28 novembre 2014 devant le jury composé de :

| | | |
|---|---|---|
| Hervé CARDOT | Université de Bourgogne | Rapporteur |
| Maria Giovanna RANALLI | Università degli Studi di Perugia | Rapportrice |
| Chris SKINNER | London School of Economics and Political Science | Rapporteur |
| Jan JOHANNES | Ensai | Examinateur |
| Valérie MONBET | Université de Rennes 1 | Examinatrice |
| Anne RUIZ-GAZEN | Université Toulouse 1 Capitole | Examinatrice |
| Olivier SAUTORY | Insee | Examinateur |

*Je me souviens*

# Remerciements

Je souhaite tout d'abord remercier Hervé Cardot, Gionanna Ranalli et Chris Skinner d'avoir accepté d'être les rapporteurs de ce mémoire, et pour le temps qu'ils y ont consacré. Je remercie également Jan Johannes, Valérie Monbet, Anne Ruiz-Gazen et Olivier Sautory d'avoir accepté de participer à ce jury. Je leur suis sincèrement reconnaissant de leurs efforts pour participer à cette soutenance, malgré leurs emplois du temps chargés. Je remercie tout particulièrement Anne de m'avoir accompagné depuis la rédaction du mémoire jusqu'à la préparation de la soutenance, et pour tous ses encouragements.

Sur le chemin, il y avait beaucoup de collaborateurs qui sont avant tout des amis. J'ai bien sûr une pensée particulière pour David et Camelia, mon grand frère et ma grande soeur de recherche, et pour Jean-Claude qui m'a fait confiance en me recrutant au Laboratoire de Statistique d'Enquête et m'a permis d'en être là aujourd'hui. Je pense bien sûr également à Cyril, Daniel, Eric, et de façon générale aux personnes avec qui j'ai collaboré. Merci aux personnels de l'école qui m'entourent de leur confiance et de leur amitié, et tout particulièrement à Jojo et à l'équipe du département informatique. Merci à tous mes collègues de l'Ensai et d'ailleurs qui ont pris le temps de m'adresser un petit message de soutien.

Dans les bons moments mais surtout dans les moments difficiles, mes amis et ma famille étaient là. Merci à ma bande des anciens de l'Ensai pour leurs encouragements et leur amitié. Merci à ma petite famille de Lyon qui m'accompagne par la pensée. Merci à ma maman et à mon papa qui ont fait une longue route pour venir m'aider et me soutenir dans la mise en place de cette habilitation. Enfin et surtout, merci à Brigitte et Nora pour leur patience et leur amour. Laugh until we think we'll die, barefoot on a summer night, never could be sweeter than with you.

# Présentation générale des travaux

Mon travail de recherche porte sur la théorie de l'échantillonnage et de l'estimation dans le cadre d'une population finie. En dehors de [17], qui constitue une application des méthodes de Bootstrap étudiées dans le cadre de ma thèse de doctorat, et de [1] qui a été écrit en parallèle de ma thèse, tous les travaux de recherche présentés ci-dessous sont postérieurs à ma thèse. La plupart de ces travaux peuvent être répartis en quatre domaines de recherche, qui sont : l'échantillonnage, le traitement de la non-réponse partielle, l'estimation de variance, les méthodes de couplage.

## Echantillonnage

Dans la plupart des enquêtes, on cherche à mobiliser une information auxiliaire afin d'améliorer la précision des estimateurs. A l'étape de l'échantillonnage, cette information est incorporée en imposant que le plan de sondage respecte des contraintes d'équilibrage. Une solution très générale pour sélectionner des échantillons équilibrés a été proposée par Deville et Tillé (2004), sous la forme de la *méthode du Cube* : une part importante de mon travail de recherche concerne l'étude de ses propriétés. Dans [18], nous comparons la méthode du Cube avec une méthode réjective d'échantillonnage équilibrée proposée par Hajek (1981) et Fuller (2009). Dans [6], nous déterminons les probabilités d'inclusion permettant de minimiser la variance de l'estimateur de Horvitz-Thompson pour un tirage équilibré. Dans [12], nous proposons une modification de la méthode du cube dans le cas où la variable d'intérêt peut être représentée par un modèle linéaire mixte. Comme l'algorithme du Cube peut être gourmand en temps de calcul, une procédure plus rapide est proposée dans [1], avec une application de cette méthode dans [2]. Un cas particulier important de cette procédure rapide d'échantillonnage correspond à l'algorithme appelé la *méthode du pivot ordonné* (Deville et Tillé, 1998). J'ai établi dans [11] l'équivalence entre la méthode du pivot ordonné et le *tirage systématique de Deville* (Deville, 1998). Dans [20], nous montrons que toute implémentation de la méthode du pivot est plus efficace que le tirage multinomial.

## Traitement de la non-réponse partielle

A l'étape de l'estimation, l'information auxiliaire peut également être utilisée pour compenser d'une possible non-réponse. Une valeur manquante pour une variable d'intérêt est remplacée par une valeur artificielle, mais plausible: on parle alors d'imputation de la non-réponse partielle (Haziza, 2009). L'imputation se base sur un modèle de comportement pour la variable d'intérêt (modèle d'imputation), que l'on cherche à reproduire lors de la création de la valeur artificielle imputée (mécanisme d'imputation). Dans [7], nous montrons que l'utilisation d'un mécanisme d'imputation aléatoire adapté permet de préserver la distribution de la variable imputée. Nous proposons l'utilisation d'une méthode d'imputation équilibrée adaptée de la méthode du Cube pour préserver la distribution de la variable, tout en réduisant la variance liée à l'imputation ; voir également [22]. L'algorithme d'échantillonnage proposé dans [3] peut être utilisé à cet effet. Une procédure d'imputation permettant d'éliminer complètement la variance liée à l'imputation est proposée dans [26]. Une procédure d'imputation permettant de préserver la distribution de la variable imputée si le modèle d'imputation est adapté ou si la modélisation du mécanisme de non-réponse est adaptée est proposée dans [21]. De façon générale, le mécanisme d'imputation doit refléter les propriétés de la variable d'intérêt. Dans [15], nous étudions des procédures d'imputation permettant de traiter des variables avec sur-représentation de valeurs nulles. Des méthodes d'imputation permettant d'imputer plusieurs variables et de respecter les relations entre elles sont étudiées dans [10] pour des variables quantitatives, et dans [16] pour des variables qualitatives.

## Estimation de variance

Les estimateurs produits dans le cadre d'une enquête sont généralement assortis d'une mesure de précision telle qu'un estimateur de variance, un coefficient de variation ou un intervalle de confiance. L'estimation de variance dans les enquêtes est généralement difficile, car l'ensemble du processus d'échantillonnage et d'estimation doit être pris en compte (Ardilly, 2006). J'ai proposé dans [13] un estimateur de variance pour l'Enquête Logement de 2006, et dans [23] nous considérons des estimateurs de variance pour l'Etude Longitudinale Française depuis l'Enfance (ELFE). Dans [17], nous comparons une méthode de Bootstrap (Efron, 1982; Shao et Tu, 1995) avec la linéarisation (Deville, 1999; Goga et al., 2009), pour l'estimation de l'évolution d'un paramètre complexe. Pour calculer un estimateur sans biais de variance, il est généralement nécessaire de connaître les probabilités d'inclusion d'ordre deux associées au plan de sondage. Ces probabilités sont souvent difficiles à calculer exactement, et peuvent être remplacées par des approximations par simulations (Fattorini, 2006; Thompson et Wu, 2008; Lesage, 2013). Dans [5], nous proposons une méthode d'approximation efficace dans le cas d'un échantillon tiré selon la méthode du Cube. J'utilise également cette méthode dans [8] pour les estimations de variance associées au nouvel Echantillon-Maître. En situation de non-réponse partielle, un mécanisme d'imputation aléatoire engendre une

variance additionnelle qui doit être pris en compte dans les mesures de précision. Dans [10] et [16], nous décrivons une procédure de Bootstrap pour estimer la variance dans le cas de données imputées, et une procédure d'estimation de variance par Jackknife (Shao et Tu, 1995) est proposée dans [15].

## Méthodes de couplage

Dans les enquêtes, la dépendance introduite dans la sélection des unités peut être complexe, ce qui rend des résultats tels qu'un théorème central limite plus difficiles à établir. Les procédures de couplage (Thorisson, 2000) constituent une solution intéressante pour relier un plan de sondage complexe à un plan de sondage à la fois proche et plus simple, et où des propriétés asymptotiques peuvent être plus facilement établies. Cette approche a été utilisée de façon pionnière par Hajek (1960) pour coupler le sondage aléatoire simple avec le tirage de Bernoulli, et par Hajek (1964) pour coupler le sondage réjectif avec le tirage de Poisson. Dans [19], j'utilise les méthodes de couplage pour obtenir des résultats de normalité asymptotique pour des plans à plusieurs degrés, et pour justifier de la consistance d'une méthode de Bootstrap des unités primaires (Rao et Wu, 1988). Dans [25], un couplage entre la méthode du pivot ordonné et une méthode de tirage stratifié est proposé. Il permet d'obtenir un théorème central limite pour la méthode du pivot ordonné, sous des hypothèses standard.

## Autres travaux

Certains de mes travaux de recherche ne rentrent pas dans les quatre grands domaines évoqués ci-dessus: ils sont présentés ici brièvement, mais ils ne seront pas détaillés dans mon document de synthèse. Dans [4], nous avons proposé des méthodes d'échantillonnage et d'estimation dans le cas d'une enquête avec un biais de couverture volontaire. Dans [9], nous avons étudié une allocation optimale d'échantillon pour un plan à plusieurs degrés. Dans [14], nous appliquons des méthodes d'estimation sur bases de sondage multiples (Lohr, 2011) pour mettre en commun des estimations issues de deux vagues d'enquête, dans un plan à plusieurs degrés. Dans [24], nous proposons d'utiliser une méthode de Bootstrap pour sélectionner des variables de calage (Deville et Särndal, 1992) en évitant une inflation de la variance due à un trop grand nombre de régresseurs.

# Contents


**1 Introduction** **1**

1.1 Framework and notation . . . . . 1
1.2 Horvitz-Thompson estimation . . . . . 1
1.3 Weak consistency of the HT-estimator . . . . . 2
1.4 Asymptotic normality . . . . . 3
1.5 Variance estimation . . . . . 4
1.6 Use of models in finite population sampling . . . . . 4

**2 Balanced sampling** **6**

2.1 The Cube method . . . . . 8
2.2 A rejective method for balanced sampling . . . . . 9
2.3 Optimal inclusion probabilities for balanced sampling . . . . . 11
2.4 Penalized balanced sampling . . . . . 12
2.5 A fast method for balanced sampling . . . . . 14
2.6 The pivotal method . . . . . 15
2.7 Future work . . . . . 18

**3 Treatment of item non-response** **19**

3.1 Approaches for inference . . . . . 20
3.2 Imputation methods . . . . . 21
3.3 Balanced random imputation . . . . . 22
3.4 Estimation of the distribution function . . . . . 24
3.5 Tailor-made imputation methods . . . . . 25
3.5.1 Zero-inflated data . . . . . 25
3.5.2 Continuous bivariate data . . . . . 27
3.5.3 Categorical bivariate data . . . . . 29
3.6 Future work . . . . . 30

**4 Variance estimation** **31**
4.1 Simulation-based variance estimation . . . . . 32
4.2 Linearization and replication-based variance estimation . . . . . 33
4.3 Variance estimation for imputed data . . . . . 36
4.4 Future work . . . . . 37

**5 Coupling methods** **38**
5.1 Asymptotic normality for multistage sampling . . . . . 39
5.1.1 Bernoulli sampling of PSUs . . . . . 40
5.1.2 Without replacement simple random sampling of PSUs . . . . . 41
5.2 With-replacement Bootstrap for multistage sampling . . . . . 43
5.2.1 With replacement sampling of PSUs . . . . . 43
5.2.2 A coupling procedure between SIR/SI sampling of PSUs . . . . . 43
5.2.3 With replacement Bootstrap of PSUs . . . . . 44
5.2.4 Bootstrap variance estimation for functions of means . . . . . 46
5.3 Future work . . . . . 47

# Chapter 1

# Introduction

## 1.1 Framework and notation

We consider a finite labeled population $U$ of size $N$, with some variable of interest $y$, and the population vector of values is denoted by $y_U = (y_1, \ldots, y_k, \ldots, y_N)^\top$. We are interested in estimating some parameter $\theta \equiv \theta(y_k,\ k \in U)$, such as the total $t_y = \sum_{k \in U} y_k$ or the mean $\mu_y = N^{-1} \sum_{k \in U} y_k$. A random sample $S$ is selected in $U$ by means of some sampling design $p(\cdot)$, i.e. according to some probability law on the subsets in $U$.

In order to study the asymptotic properties of the sampling designs and estimators that we treat below, we consider the asymptotic framework of Isaki and Fuller (1982). We assume that the population $U$ belongs to a nested sequence $\{U_t\}$ of finite populations with increasing sizes $N_t$, and that the population vector of values $y_{Ut} = (y_{1t}, \ldots, y_{Nt})^\top$ belongs to a sequence $\{y_{Ut}\}$ of $N_t$-vectors. For simplicity, the index $t$ will be suppressed in what follows and all limiting processes will be taken as $t \to \infty$.

Through the paper, we will note $E(\cdot)$ and $V(\cdot)$ for the expectation and the variance of some estimator, and $E_{\{X\}}(\cdot)$ and $V_{\{X\}}(\cdot)$ for the expectation and variance conditionally on some random variable $X$. Also, we will note $x_k$ or $z_k$ for a vector of auxiliary variables for unit $k$ known either on the sample $S$ or on the whole population $U$, and that will be considered as non-random.

## 1.2 Horvitz-Thompson estimation

Let $\pi_k(p)$ denote the first-order inclusion probability of unit $k$ with the sampling design $p(\cdot)$, that is, the probability for unit $k$ to be included in the sample $S$. When there is no risk of confusion, we simply note $\pi_k(p) \equiv \pi_k$. In what follows, we assume that all $\pi_k$'s are positive. We note

$\pi = (\pi_1, \ldots, \pi_N)^\top$, with $\sum_{k \in U} \pi_k = n$ the (integer) average sample size. The Horvitz-Thompson (HT) estimator

$$\hat{t}_{y\pi}(S) \quad = \quad \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k(S) \tag{1.2.1}$$

is then design-unbiased for $t_y$, with $I_k(S)$ the sample membership indicator for unit $k$ (see Horvitz and Thompson, 1952). When there is no risk of confusion, we simply note $I_k(S) \equiv I_k$ and $\hat{t}_{y\pi} \equiv \hat{t}_{y\pi}(S)$. We note $d_k = 1/\pi_k$ the sampling weight. The design variance is

$$V_{\{y_U\}}\left(\hat{t}_{y\pi}\right) = \sum_{k \in U} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \Delta_{kl}(p) \quad \text{with} \quad \Delta_{kl}(p) = \pi_{kl}(p) - \pi_k(p)\pi_l(p), \tag{1.2.2}$$

and with $\pi_{kl}(p)$ the probability that units $k$ and $l$ are selected jointly in the sample $S$ with the sampling design $p(\cdot)$. When there is no risk of confusion, we simply note $\pi_{kl}(p) \equiv \pi_{kl}$ and $\Delta_{kl}(p) \equiv \Delta_{kl}$. The *Poisson sampling design* is a particular important case where each unit $k$ is selected in the sample with probability $\pi_k$, independently on the other units. The components of the vector $I = (I_1, \ldots, I_N)^\top$ are then independent, which leads to

$$V_{\{y_U\}}\left(\hat{t}_{y\pi}\right) = \sum_{k \in U} \left(\frac{y_k}{\pi_k}\right)^2 \pi_k(1 - \pi_k). \tag{1.2.3}$$

The three following properties would be useful for estimation with a sampling design $p(\cdot)$:

P1: The HT-estimator is weakly consistent, i.e.

$$N^{-1}(\hat{t}_{y\pi} - t_y) \quad \underset{Pr}{\longrightarrow} \quad 0,$$

with $\underset{Pr}{\longrightarrow}$ denoting the convergence in probability.

P2: The HT-estimator is asymptotically normal, i.e.

$$\frac{\hat{t}_{y\pi} - t_y}{\sqrt{V_{\{y_U\}}\left(\hat{t}_{y\pi}\right)}} \quad \underset{\mathcal{L}}{\longrightarrow} \quad \mathcal{N}(0,1),$$

with $\underset{\mathcal{L}}{\longrightarrow}$ denoting the convergence in distribution.

P3: There is a weakly consistent variance estimator for $V_{\{y_U\}}\left(\hat{t}_{y\pi}\right)$.

## 1.3 Weak consistency of the HT-estimator

The sampling design $p(\cdot)$ is said to be of fixed-size if only the subsets in $U$ of size $n$ may have a non-zero probability of selection. Many fixed-size sampling designs have been proposed in the literature, such as *conditional Poisson sampling* (Hajek, 1964), *systematic sampling* (e.g. Tillé,

2011), or *pivotal sampling* (Deville and Tillé, 1998) which will be further studied in Section 2.6. In case of a fixed-size sampling design, the variance of the HT-estimator may be rewritten as

$$V_{\{y_U\}}\left(\hat{t}_{y\pi}\right) = -\frac{1}{2}\sum_{k\neq l\in U}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\Delta_{kl}(p), \tag{1.3.1}$$

which is known as the Sen-Yates-Grundy variance formula (see Sen, 1953; Yates and Grundy, 1953).

If the so-called Sen-Yates-Grundy conditions are fulfilled, namely the sampling design $p(\cdot)$ is such that

C1: for any vector $\pi$ of inclusion probabilities, we have $\Delta_{kl}(p)\leq 0$ for any $k\neq l\in U$,

then formula (1.3.1) leads to a variance estimator which is always positive (see Section 4). The Sen-Yates-Grundy conditions have two other interesting consequences. For any variable $y$ with positive values, the variance of the HT-estimator under a sampling design respecting C1 will be systematically lower than the variance of the HT-estimator under Poisson sampling. Also, if the additional moment condition

C2: There exists some constant $C$ such that $N^{-1}\sum_{k\in U}\left(\frac{y_k}{\pi_k}\right)^2 < C$,

is respected, then the property P1 holds true for any variable $y$ with positive values. This property may be extended to the case of any variable of interest (with positive or negative values) if the condition C1 is replaced by

C3: for any vector $\pi$, the variance of the HT-estimator under the sampling design $p(\cdot)$ with inclusion probabilities $\pi$ is lower than the variance of the Hansen-Hurwitz (1953) estimator under $n$ independent draws in $U$, with drawing probabilities $(\pi_1/n,\ldots,\pi_N/n)^\top$.

The property C3 has been shown to hold for the Sampford design (Gabler, 1981 and 1984) and for the conditional Poisson sampling design (Qualité, 2008), for example; see Section 2.6 for pivotal sampling.

## 1.4 Asymptotic normality

Several results of asymptotic normality have been proved for specific sampling designs; see for example Hajek (1960,1961) for simple random sampling without replacement, Hajek (1964) for rejective sampling, Rosen (1972a,1972b), Sen (1979) and Gordon (1983) for successive sampling, and Ohlsson (1986) for the Rao-Hartley-Cochran (1962) procedure. Branden and Jonasson (2012) state a central-limit theorem for the class of sampling algorithms satisfying the strongly Rayleigh property, which includes Sampford sampling, Pareto sampling and the pivotal method. Chen and

Rao (2007) prove asymptotic normality for a class of estimators under two-phase sampling designs, see also Saegusa and Wellner (2013). Asymptotic normality of estimators resulting from multistage samples has been considered by Krewski and Rao (1981) and Ohlsson (1989).

The dependence in the selection of units may be complex, which makes limiting results quite difficult to prove. Coupling methods (see Thorisson, 2000) can be used to link a sampling design under study to a close, simpler sampling design where useful limiting properties may be more easily derived. In a pioneering work, Hajek (1961) introduced a coupling procedure between Bernoulli sampling and simple random sampling without replacement (SI) to obtain a central-limit theorem for the latter. In Hajek (1964), a similar approach was used to link Poisson sampling and rejective sampling, and to derive a central-limit theorem and a variance approximation for rejective sampling. Coupling methods are considered in Section 5 to derive asymptotic normality results for the HT-estimator under without-replacement multistage designs, and to prove the consistency of a Bootstrap procedure. A specific coupling procedure (not presented in this manuscript) has been used in [25] to obtain the asymptotic normality of the HT-estimator under ordered pivotal sampling.

## 1.5 Variance estimation

When the second-order inclusion probabilities may be computed for the sampling design, an unbiased variance estimator is available (see equation 4.0.1). Otherwise, simulation-based approximations for the variance-covariance matrix of the sampling design can be used. In case of variance estimation for complex parameters, we may resort to linearization (Deville, 1999) or Bootstrap (e.g., Shao and Tu, 1995) for variance estimation. These points will be further developed in Section 4.

The consistency of variance estimators typically relies on assumptions on fourth order inclusion probabilities (see Breidt and Opsomer, 2000). These properties are usually difficult to prove for unequal probability sampling algorithms (see Boistard, Lopuhaä and Ruiz-Gazen, 2012, for conditional Poisson sampling). Alternatively, coupling methods could be a promising tool for the consistency of variance estimators, and are a matter of further research for that purpose.

## 1.6 Use of models in finite population sampling

This is often useful to model the variable of interest $y$, making use of some vector $x_k$ of auxiliary variables. For example, we may assume that the variable $y$ follows the linear model

$$y_k = x_k^\top \beta + \epsilon_k, \tag{1.6.1}$$

with $\beta$ a $q$-vector of unknown parameters, and some assumptions on the distribution of the residual terms $\epsilon_k$.

At the sampling stage, the model (1.6.1) is used as a *working model* to define an efficient sampling design. It is not required that underlying model assumptions hold true for the HT-estimator to be unbiased, but the HT-estimator is expected to be more efficient if these assumptions hold true. The model (1.6.1) may also be used as a working model at the estimation stage to define an efficient estimation strategy, e.g. through calibration (see Deville and Särndal, 1992). If the variable $x_k$ is known prior to sampling for any unit $k \in U$, and if the sampling design is such that

$$\hat{t}_{x\pi}(S) \quad = \quad t_x,$$

the sampling design is said to be balanced on $x_k$. It leads to $V_{\{y_U\}}(\hat{t}_{y\pi}) = V_{\{y_U\}}(\hat{t}_{\epsilon\pi})$, so that the variance can be strongly reduced (as compared to an unbalanced sampling design) if the variability of the residuals $\epsilon_k$'s is much lower than that of the $y_k$'s. Balanced sampling strategies will be further studied in Section 2.

At the imputation stage, the model (1.6.1) is used as an *imputation model* so as to justify replacing some missing value with an artificial, imputed value. In such case, explicit assumptions are needed for the imputation procedure to lead to valid (and at least, asymptotically unbiased) estimators. The imputation model needs to be adapted to the type of variable considered (e.g., continuous or categorical), and the imputation process needs to be adapted to the type of estimation needed (e.g., total or distribution function). These considerations will be further studied in Section 3.

# Chapter 2

# Balanced sampling

The accuracy of HT-estimators relies on auxiliary information available for any unit in the population, and which is used to define the sampling design. This auxiliary information is frequently incorporated by using some form of *balanced sampling*, where the sample $S$ is selected so that HT-estimators for the totals of auxiliary variables match (exactly or at least, very closely) the known totals.

Suppose that a $q$-vector $x_k = (x_{1k}, \ldots, x_{qk})^\top$ of auxiliary variables is known at the design stage for any unit $k \in U$. The $N \times q$ matrix

$$A \quad = \quad (x_k/\pi_k)_{k \in U} \tag{2.0.1}$$

is called the matrix of constraints. A sample $s$ selected with inclusion probabilities $\pi = (\pi_1, \ldots, \pi_N)^\top$ is said to be balanced on $x_k$ if the set of balancing equations

$$\hat{t}_{x\pi}(s) \equiv \sum_{k \in s} \frac{x_k}{\pi_k} \quad = \quad t_x \tag{2.0.2}$$

is fulfilled, which means that the HT-estimator exactly match the known vector of totals $t_x$. A sampling design $p(\cdot)$ is said to be balanced on $x_k$ if

$$\forall s \subset U \quad p(s) > 0 \Rightarrow \hat{t}_{x\pi}(s) = t_x, \tag{2.0.3}$$

which means that the support of the sampling design is restricted to balanced samples.

A number of common sampling designs may be seen as balanced for a particular vector $x$. In case of STSI sampling when the population is stratified inside $H$ non-overlapping sub-populations $U_1, \ldots, U_H$ of sizes $N_1, \ldots, N_H$, and the sample $S$ is selected by SI sampling of size $n_h$ inside $U_h$, we have $\pi_k = \frac{n_h}{N_h}$ for any unit $k \in U_h$ and

$$\hat{N}_{h\pi} \equiv \sum_{k \in S} \frac{1(k \in U_h)}{\pi_k} \quad = \quad N_h.$$

Therefore, the sizes of strata are perfectly estimated and the sampling design is balanced on the $H$-vector $x_k = \{1(k \in U_1), \ldots, 1(k \in U_H)\}^\top$. In case of any sampling algorithm with unequal probabilities and fixed size $n = \sum_{k \in U} \pi_k$, we have $n(S) = n$, and the sampling design is balanced on $x_k = \pi_k$.

The variance of the HT-estimator is given by formula (1.2.2) or (1.3.1), but second-order inclusion probabilities are usually difficult to compute for a general balanced sampling design. Deville and Tillé (2005) therefore proposed variance approximations for balanced sampling, under the assumptions that the sampling design is exactly balanced, and performed with maximum entropy among sampling designs balanced on the same variables, with the same inclusion probabilities. Then, under an additional assumption of asymptotic normality of the multivariate HT-estimator under Poisson sampling, they derived the following variance approximation:

$$V_{DT}(\hat{t}_{y\pi}) = \frac{N}{N-q} \sum_{k \in U} b(\pi_k) \left\{ y_k - y_k^*(\pi) \right\}^2, \tag{2.0.4}$$

where

$$b(\pi_k) \quad = \quad 1/\pi_k - 1 \tag{2.0.5}$$

and

$$y_k^*(\pi) = x_k^\top B(\pi) \quad \text{with} \quad B(\pi) = \left\{ \sum_{l \in U} b(\pi_l) x_l x_l^\top \right\}^{-1} \sum_{l \in U} b(\pi_l) x_l y_l. \tag{2.0.6}$$

Other slightly different approximations are proposed in Deville and Tillé (2005), but their simulation results suggest that the approximation (2.0.4) performs well among variance approximations that may be computed in case of any set of inclusion probabilities.

Several partial solutions have been proposed for balanced sampling (see Deville et al., 1988, Ardilly, 1991), before the cube method was introduced (Deville and Tillé, 2004). This method enables to select (approximately) balanced samples for any set of inclusion probabilities $\pi$ and any auxiliary vector $x$, and is described in Section 2.1. An alternative rejective procedure studied by Hajek (1981) and Fuller (2009) is presented in Section 2.2. In Section 2.3, we present a method to compute inclusion probabilities so that an approximation of the variance of the HT estimator is minimized. Adapting the cube method to the case when the variable of interest may be better described by a linear mixed model is the purpose of Section 2.4. The initial implementation of the cube method involved the search for vectors in the kernel of a large matrix, and could thus be time-consuming. A fast implementation proposed by [1] is described in Section 2.5. In case when the vector of auxiliary variables reduces to the inclusion probability, which means fixed-size sampling, this fast implementation leads to the so-called pivotal sampling (Deville and Tillé, 1998). This sampling

algorithm is presented in Section 2.6, and some useful properties are derived.

## 2.1 The Cube method

The cube method proceeds in two steps: a flight phase, at the end of which an exact balancing is maintained, and a landing phase during which the balancing equations may be partly relaxed until the complete sample is obtained, while the inclusion probabilities remain exactly respected.

**Algorithm 1** A general procedure for the cube method

First initialize at $\pi(0) = \pi$. Next, at time $t = 0, \cdots, T$, repeat the following steps:

1. If there exists some vector $u(t) \neq 0$ such that $u(t) \in Ker(A)$ and $u_k(t) = 0$ if $\pi_k(t)$ is an integer, then:

   (a) Take any such vector $u(t)$ (random or not), and compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$, the largest values of $\lambda_1(t)$ and $\lambda_2(t)$ such that

   $$0 \leq \pi(t) + \lambda_1(t)u(t) \leq 1 \quad \text{and} \quad 0 \leq \pi(t) - \lambda_2(t)u(t) \leq 1.$$

   (b) Take $\pi(t+1) = \pi(t) + \delta(t)$, where

   $$\delta(t) = \begin{cases} \lambda_1^*(t)u(t) & \text{with probability } \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}, \\ -\lambda_2^*(t)u(t) & \text{with probability } \lambda_1^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}. \end{cases}$$

2. Otherwise, drop the last column from the matrix $A$ and go back to Step 1.

A general procedure for the cube method (see [1]; Tillé, 2011) which covers both the flight phase and the landing phase is presented in Algorithm 1. It proceeds through a random walk from the vector of inclusion probabilities $\pi$ to the random vector $I \equiv \pi(T)$ of sample membership indicators. During the flight phase, a vector $u(t)$ is chosen in Step 1.a so that the balancing equations remain exactly respected and the units already selected/rejected during the previous steps are not affected. The scalars $\lambda_1^*(t)$ and $\lambda_2^*(t)$ are then determined so that at least one additional unit is definitely selected or rejected. In Step 1.b, the vector $\pi(t+1)$ is randomly updated, in such a way that the inclusion probabilities remain exactly respected.

The flight phase ends at time $T_F$, when it is no more possible to find a suitable vector $u(t)$. We have $\pi_k(T_F) = 0$ if unit $k$ is definitely rejected from the sample, $\pi_k(T_F) = 1$ if unit $k$ is selected. Also, we have $0 < \pi_k(T_F) < 1$ if the decision for unit $k$ remains pending: the associated set of units

is at most of size $q$ (see Deville and Tillé, 2004), and will be denoted as $U(T_F)$. At the end of the flight phase, we have

$$E_{\{y_U\}}\{\pi(T_F)\} = \pi, \tag{2.1.1}$$

$$\sum_{k\in U}\frac{x_k}{\pi_k}\pi_k(T_F) = \sum_{k\in U}x_k. \tag{2.1.2}$$

The landing phase included in Algorithm 1 enables to end the sampling, by successively relaxing in Step 2 one balancing constraint to gain one degree of freedom, and by applying the flight phase to the reduced matrix. Alternatively, the landing phase may be performed by means of an enumerative algorithm on the remaining units in $U(T_F)$ (Tillé, 2011, p. 163). In any case, the landing phase leads to a vector of sample selection indicators $I$ such that

$$E_{\{y_U,\pi(T_F)\}}(I) = \pi(T_F), \tag{2.1.3}$$

$$\hat{t}_{x\pi} \equiv \sum_{k\in U}\frac{x_k}{\pi_k}I_k \simeq \sum_{k\in U}x_k. \tag{2.1.4}$$

The inclusion probabilities are exactly respected, since from (2.1.1) and (2.1.3) we have $E_{\{y_U\}}(I) = \pi$, and the HT-estimator $\hat{t}_{y\pi}$ is exactly design-unbiased for $t_y$. Deville and Tillé (2004) show in their Proposition 4 that, for any variant of the landing phase, the cube method achieves $|\hat{t}_{x\pi} - t_x| \leq q\max_{k\in U}|x_k|\pi_k^{-1}$ element-wise, so that under reasonable hypotheses on $x$ and for standard designs with $(N\min_{k\in U}\pi_k)^{-1} = O(n^{-1})$, we have

$$\hat{t}_{x\pi} = t_x + O\left(q \times Nn^{-1}\right). \tag{2.1.5}$$

In (2.1.5), the remainder term is non-stochastic and can be much smaller for fixed $q$ than the usual $O_p(Nn^{-1/2})$ remainder in the unbalanced case.

## 2.2 A rejective method for balanced sampling

To ensure that the sampling design is balanced, at least approximately, a rejective procedure may alternatively be used (see Hajek, 1981; Fuller, 2009). Under rejective sampling, a basic sampling procedure $p_b(\cdot)$ is repeatedly applied to select a random sample $S_b$ with inclusion probabilities $p_k = Pr(k \in S_b)$, until

$$(\hat{t}_{xp} - t_x)^\top V_{\{y_U\}}(\hat{t}_{xp})^{-1}(\hat{t}_{xp} - t_x) \leq \gamma^2, \tag{2.2.1}$$

where $\hat{t}_{xp} = \sum_{k\in S_b}p_k^{-1}x_k$ and $\gamma > 0$ is a specified balancing tolerance specified by the survey statistician. The resulting rejective sampling procedure $p(\cdot)$ and the associated random sample $S$ are not to be confused with $p_b(\cdot)$ and $S_b$, since in particular $\pi_k \equiv Pr(k \in S) \neq p_k$.

There is an important distinction between the Cube method and the rejective method. In the first, the inclusion probabilities $\pi$ are exactly satisfied but one has no explicit control on the (possible) discrepancy between estimates and the true population totals. In the second, the discrepancy is perfectly controlled through the balancing tolerance $\gamma$ but the exact inclusion probabilities $\pi_k$ are usually unknown. For rejective sampling, Fuller (2009) suggests the use of a GREG type estimator based on the initial inclusion probabilities $p_k$. Suppose that the basic sampling procedure $p_b(\cdot)$ is such that there exists some $\phi_k$ satisfying

$$E_{\{y_U,\ k\in S_b\}}(\hat{t}_{ap} - t_a) \quad = \quad p_k^{-1}\phi_k a_k \tag{2.2.2}$$

for any vector of interest $a_k$. The proposed GREG-type estimator is then

$$\hat{t}_{y,pgreg} = \sum_{k\in U} x_k^\top \hat{B} \text{ where } \hat{B} = \left(\sum_{k\in S} p_k^{-2}\phi_k x_k x_k^\top\right)^{-1} \sum_{k\in S} p_k^{-2}\phi_k x_k y_k, \tag{2.2.3}$$

and is shown to be design-consistent provided that there exists a vector of constants $\lambda$ such that

$$p_k^{-2}\phi_k x_k^\top \lambda = p_k^{-1}, \tag{2.2.4}$$

see Fuller (2009). However, the estimator $\hat{t}_{y,pgreg}$ may suffer from substantial bias for finite sample sizes when the $p_k$'s do not provide a good approximation of the true $\pi_k$.

Two alternative estimation strategies for rejective balanced sampling are proposed in [18]. The first consists in estimating the true $\pi_k$'s through Monte Carlo simulations to obtain a Horvitz-Thompson like estimator (see Fattorini, 2006, and Lesage, 2013). The second consists in approximating the inclusion probabilities through an Edgeworth expansion, when Poisson sampling is used as the basic procedure. In case of a scalar $x$, the final inclusion probability for rejective sampling is

$$\pi_k \quad = \quad p_k \frac{Pr_{\{y_U\}}(-\gamma \le X \le \gamma | I_{bk} = 1)}{Pr_{\{y_U\}}(-\gamma \le X \le \gamma)}, \tag{2.2.5}$$

where $X = d^{-1/2}\sum_{k\in U} p_k^{-1} x_k (I_{bk} - p_k)$ and $d = \sum_{k\in U} x_k^2 p_k^{-1}(1-p_k)$. Using the formal Edgeworth expansion (see Thompson, 1997, equation (3.41)), it is obtained in [18] after some algebra that

$$\pi_k = p_k \left\{ 1 - \frac{1}{d}\frac{\gamma\phi(\gamma)}{2\psi(\gamma)-1}\left(\frac{x_k}{p_k}\right)^2 (1-p_k)(1-2p_k) + \frac{\kappa_3}{\sqrt{d}}\frac{\gamma\phi(\gamma)(3-\gamma^2)}{3(2\psi(\gamma)-1)}\frac{x_k}{p_k}(1-p_k) \right\} + o(d^{-1}), \tag{2.2.6}$$

where $\psi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function and the probability density function of a standard normal distribution, where

$$\begin{aligned}
\kappa_3 \quad &\equiv \quad \mu_3(X) = d^{-3/2}\sum_{k\in U}\left(\frac{x_k}{p_k}\right)^3 p_k(1-p_k)(1-2p_k), \\
\kappa_4 \quad &\equiv \quad \mu_4(X) - 3\{\mu_2(X)\}^2 = d^{-2}\sum_{k\in U}\left(\frac{x_k}{p_k}\right)^4 p_k(1-p_k)\{1-6p_k(1-p_k)\},
\end{aligned}$$

and $\mu_m(X)$ denotes the centered moment of order $m$ of the random variable $X$. When the balancing tolerance may be seen as small, approximation (2.2.6) simplifies as

$$\pi_k \;=\; p_k \left\{ 1 - \frac{1}{2d} \left( \frac{x_k}{p_k} \right)^2 (1-p_k)(1-2p_k) + \frac{\kappa_3}{2\sqrt{d}} \left( \frac{x_k}{p_k} \right) (1-p_k) \right\} + o(d^{-1}). \qquad (2.2.7)$$

## 2.3 Optimal inclusion probabilities for balanced sampling

In many cases, inclusion probabilities are fixed and chosen to be proportional to an auxiliary variable known for any unit in the population. However, if some information on the variable of interest is available at the design stage, it may be of interest to look for inclusion probabilities that minimize, at least approximately, the variance of the HT-estimator. An optimal vector $\pi$ of inclusion probabilities should minimize the variance in (1.2.2), under the constraints that

$$0 \leq \pi_k \leq 1 \text{ for any unit } k \in U \quad \text{ and } \quad \sum_{k \in U} \pi_k = n. \qquad (2.3.1)$$

Since second-order inclusion probabilities are usually untractable, and following Tillé and Favre (2005), it is proposed in [6] to minimize the variance approximation (2.0.4) instead. If the balancing variables $x_k$ do not depend on the inclusion probabilities, they proved that the solution to this problem satisfies

$$\pi_k = n \frac{|y_k - y_k^*(\pi)|}{\sum_{l \in U} \left| y_l - y_l^*(\pi) \right|} \text{ for any } k \in U, \qquad (2.3.2)$$

where $y_l^*(\pi)$ is given in (2.0.6). Formula (2.3.2) states that larger inclusion probabilities should be given to the units for which $y_k$ may not be well predicted by the balancing variables.

In practice, formula (2.3.2) may not be used since the $y$-values are not available on the whole population $U$. An alternative optimization problem is proposed in [6], where $U$ is partitioned into $J$ non-overlapping subsets $U_1, \ldots, U_J$ of sizes $N_1, \ldots, N_J$, and the target inclusion probabilities are required to satisfy

$$\pi_k = \alpha_j \text{ for any unit } k \in U_j, \; j = 1, \ldots, J, \qquad (2.3.3)$$

so that inclusion probabilities are equal inside each subset $U_j$. The variance approximation in (2.0.4) may then be rewritten as

$$V_{DT}(\hat{t}_{y\pi}) \equiv V(\alpha) = \frac{N}{N-q} \sum_{j=1}^{J} b(\alpha_j) \sum_{k \in U_j} \{ y_k - \tilde{y}_k(\alpha) \}^2, \qquad (2.3.4)$$

where $\alpha = (\alpha_1, \ldots, \alpha_J)^\top$, and

$$\tilde{y}_k(\alpha) = x_k^\top \left\{ \sum_{j=1}^{J} b(\alpha_j) G_j \right\}^{-1} \sum_{j=1}^{J} b(\alpha_j) g_{1j}(y)$$

with $G_j = \sum_{k\in U_j} x_k x_k^\top$ and $g_{1j}(y) = \sum_{k\in U_j} x_k y_k$. The vector $\alpha$ that minimizes (2.3.4) under the constraints (2.3.1) and (2.3.3) satisfies

$$\alpha_j = n\frac{\sigma_j(\alpha)}{\sum_{i=1}^J N_i\sigma_i(\alpha)} \quad \text{where} \quad \sigma_j(\alpha) = \frac{1}{N_j}\sum_{k\in U_j}\{y_k - \tilde{y}_k(\alpha)\}^2. \tag{2.3.5}$$

**Algorithm 2** Fixed-point algorithm to compute optimal inclusion probabilities for balanced sampling

First initialize with any vector $\alpha^0 = (\alpha_1^0, \ldots, \alpha_J^0)^\top$. Then:

1. At step $t$, compute $\alpha^t = (\alpha_1^t, \ldots, \alpha_J^t)'$ such that

$$\alpha_j^t = n\frac{\sigma_j(\alpha^{t-1})}{\sum_{i=1}^J N_i\sigma_i(\alpha^{t-1})} \text{ for any } j = 1, \ldots, J.$$

2. The procedure ends at step $T$ when $Max_j\|\alpha_j^t - \alpha_j^{t-1}\|$ is lower than a pre-specified bound $\epsilon$.

The fixed-point Algorithm 2 may be used to determine the inclusion probabilities. From Theorem 1 in [6], it always leads to a reduction in variance, since the sequence $(\alpha^t)_{t\in\mathbb{N}}$ tends to a local minimum. In practice, the needed parameters $G_j$, $g_{1j}(y)$ and $g_{2j}(y) = \sum_{k\in U_j} y_k^2$ are unknown and need to be estimated. In case when another sample $S^p$ (e.g., associated to a former survey) has been selected in $U$ with inclusion probabilities $\pi^p = (\pi_1^p, \ldots, \pi_N^p)^\top$, Algorithm 2 can be used by replacing $G_j$, $g_{1j}(y)$ and $g_{2j}(y)$ with

$$\hat{G}_j^p = \sum_{k\in S_j^p}\frac{x_k x_k^\top}{\pi_k^p}, \quad \hat{g}_{1j}^p(y) = \sum_{k\in S_j^p}\frac{x_k y_k}{\pi_k^p}, \quad \hat{g}_{2j}^p(y) = \sum_{k\in S_j^p}\frac{y_k^2}{\pi_k^p},$$

where $S_j^p = S^p \cap U_j$.

## 2.4 Penalized balanced sampling

Suppose that a particular variable $y$ follows a linear mixed model of the form

$$y_U = X\beta + Z\gamma + \epsilon_U, \tag{2.4.1}$$

where

$$E\begin{pmatrix}\gamma\\ \epsilon_U\end{pmatrix} = 0, \quad V\begin{pmatrix}\gamma\\ \epsilon_U\end{pmatrix} = \sigma^2\begin{pmatrix}\lambda^{-2}Q & 0\\ 0 & I\end{pmatrix},$$

$X$ is a full rank $N\times q$ matrix, $Z$ is a full rank $N\times K$ matrix, and $I$ will denote an identity matrix of appropriate dimension. We suppose that $Q$ is positive definite and known, and it is typically an

identity matrix. The parameter $\sigma^2$ is unknown and the parameter $\lambda^2$ is to be determined.

It is convenient to first orthogonalize the fixed and random effects in model (2.4.1) via

$$C = \{X, (I - P_X)Z\} = (c_1, \ldots, c_{q+K}) \quad \text{where} \quad P_X = X(X^TX)^{-1}X^T.$$

One approach to the problem of using model (2.4.1) in survey design would be to use the cube algorithm to draw samples balanced on $c_j$, so that by (2.1.5),

$$\hat{t}_{c\pi} = t_c + O\{N(q+K)n^{-1}\}. \tag{2.4.2}$$

The flexibility and power of linear mixed models, however, typically come with large $K$, so that (2.4.2) may have unacceptably large errors. Such balance also ignores the mixed effect structure of the linear mixed model, and treats it essentially as an ordinary regression model with $q + K$ fixed effects.

The cube method of Deville and Tillé (2004) is modified in [12] to draw penalized balanced samples. Instead of working directly with the linear mixed model covariates $c_j$, a new set of balancing variables $b_j$ and an ordering of these variables is specified. More precisely, with $C$ as defined above, let

$$M = C^TC + \Lambda = \left\{ \begin{array}{cc} X^TX & 0 \\ 0 & Z^T(I - P_X)Z + \lambda^2 Q^{-1} \end{array} \right\}$$

and compute

$$M^{-1}C^TC = \left( \begin{array}{cc} I & 0 \\ 0 & A_1 D A_2^T \end{array} \right),$$

where $D = \text{diag}(d_1, \ldots, d_K)$, $A_1 D A_2^T$ is the singular value decomposition of

$$\left\{Z^T(I - P_X)Z + \lambda^2 Q^{-1}\right\}^{-1} Z^T(I - P_X)Z,$$

and $1 \geq d_1 \geq \cdots \geq d_K \geq 0$. The $q$ singular values corresponding to the fixed effects are identically equal to one, and $q + \sum_{i=1}^K d_i = \text{tr}(CM^{-1}C^T)$ are the degrees of freedom of the linear mixed model. The factors $d_k$ can be interpreted as fractional degrees of freedom. They decay rapidly to zero for many linear mixed models of interest.

The balancing variables are defined as the columns of the $N \times (q + K)$ matrix

$$B = C \left( \begin{array}{cc} I & 0 \\ 0 & A_1 D \end{array} \right) = (b_1, \ldots, b_{q+K}), \tag{2.4.3}$$

the first $q$ columns of which are $X$. An alternative to the balancing variables (2.4.3) that is useful in practice is to keep only the first $r$ columns of $B$, where $\sum_{i=r+1}^K d_i << 1$; that is, dropping columns

that all together account for much less than one degree of freedom. In Monte Carlo experiments using nonparametric and temporal linear mixed models, the strategy of penalized balanced sampling with Horvitz-Thompson estimation was shown to dominate a variety of standard strategies.

## 2.5 A fast method for balanced sampling

It is noticed in [1] that in Step 1 of Algorithm 1, the search for a vector in the kernel of $A$ may be time-consuming. Finding such vector defines a system of $q$ equations, and therefore $q+1$ degrees of freedom are sufficient. A faster solution is thus as follows: at any step $t$, let $E_t \subset \{1, \ldots, N\}$ denote the set of the $j = 1, \ldots, q+1$ first columns of $A$ such that $u_j(t)$ is not an integer. This set $E_t$ also corresponds to the $q+1$ first units in the population $U$ that are still neither selected nor rejected at step $t$. Also, let $A_t$ denote the sub-matrix of $A$ containing the columns in $E_t$. Then a vector $u(t)$ of $Ker(A)$ is obtained by finding a vector $v(t)$ in $Ker(A_t)$, and by complementing $v(t)$ with zeros for the columns of $A$ that are not in $A_t$. This fast implementation is described in Algorithm 3.

---

**Algorithm 3** A fast procedure for the cube method

---

First initialize at $\pi(0) = \pi$. Next, at time $t = 0, \cdots, T$, repeat the following steps:

1. If there exists some vector $v(t) \neq 0$ such that $v(t) \in Ker(A_t)$, then:

   (a) Take any such vector $v(t)$ (random or not), and take $u(t)$ such that

   $$u_k(t) = \begin{cases} v_k(t) & \text{if } k \in E_t, \\ 0 & \text{otherwise.} \end{cases}$$

   Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$, the largest values of $\lambda_1(t)$ and $\lambda_2(t)$ such that

   $$0 \leq \pi(t) + \lambda_1(t)u(t) \leq 1 \quad \text{and} \quad 0 \leq \pi(t) - \lambda_2(t)u(t) \leq 1.$$

   (b) Take $\pi(t+1) = \pi(t) + \delta(t)$, where

   $$\delta(t) = \begin{cases} \lambda_1^*(t)u(t) & \text{with probability } \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}, \\ -\lambda_2^*(t)u(t) & \text{with probability } \lambda_1^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}. \end{cases}$$

2. Otherwise, drop the last column from the matrix $A_t$ and go back to Step 1.

---

Balanced sampling can be implemented using existing software for the cube method, such as the R function samplecube in the sampling library (Tillé and Matei, 2008; R Development Core Team, 2008), or the SAS Macro Fastcube ([1],[2]). Making use of this fast algorithm, a stratified balanced sampling procedure is proposed in [3]. This procedure is in particular useful for balanced imputation

purpose (see Section 3.3).

## 2.6 The pivotal method

The fast procedure for the cube method in Algorithm 3 leads to a very simple sampling algorithm when $x_k$ is reduced to the inclusion probability $\pi_k$, which means that achieving a fixed sample size is the only balancing constraint. In this case, the matrix of constraints is the $N$-vector $A = (1, \ldots, 1)^\top$ and at any step $t$ of Algorithm 3, $A_t = (1, 1)^\top$. If we note $k_{1t}$ and $k_{2t}$ the two units in $E_t$, we have (up to a scaling factor) $v(t) = (1, -1)^\top$. This leads to

$$\{\lambda_1^*(t), \lambda_2^*(t)\} = \begin{cases} \{\pi_{k_{2t}}(t), \pi_{k_{1t}}(t)\} & \text{if } \pi_{k_{1t}}(t) + \pi_{k_{2t}}(t) \leq 1, \\ \{1 - \pi_{k_{1t}}(t), 1 - \pi_{k_{2t}}(t)\} & \text{if } \pi_{k_{1t}}(t) + \pi_{k_{2t}}(t) > 1. \end{cases}$$

In the case when $\pi_{k_{1t}}(t) + \pi_{k_{2t}}(t) \leq 1$, we obtain

$$\{\pi_{k_{1t}}(t+1), \pi_{k_{2t}}(t+1)\} = \begin{cases} \{\pi_{k_{1t}}(t) + \pi_{k_{2t}}(t), 0\} & \text{with probability } \frac{\pi_{k_{1t}}(t)}{\pi_{k_{1t}}(t)+\pi_{k_{2t}}(t)}, \\ \{0, \pi_{k_{1t}}(t) + \pi_{k_{2t}}(t)\} & \text{with probability } \frac{\pi_{k_{2t}}(t)}{\pi_{k_{1t}}(t)+\pi_{k_{2t}}(t)}. \end{cases}$$

In the case when $\pi_{k_{1t}}(t) + \pi_{k_{2t}}(t) > 1$, we obtain

$$\{\pi_{k_{1t}}(t+1), \pi_{k_{2t}}(t+1)\} = \begin{cases} \{1, \pi_k(t) + \pi_l(t) - 1\} & \text{with probability } \frac{1-\pi_{k_{2t}}(t)}{2-\pi_{k_{1t}}(t)-\pi_{k_{2t}}(t)}, \\ \{\pi_k(t) + \pi_l(t) - 1, 1\} & \text{with probability } \frac{1-\pi_{k_{1t}}(t)}{2-\pi_{k_{1t}}(t)-\pi_{k_{2t}}(t)}. \end{cases}$$

In any case, we have $\pi_m(t+1) = \pi_m(t)$ for $m \notin E_t$.

This method is known as *ordered pivotal sampling*, and may be more simply described as follows. At the first step, only the two first units 1 and 2 are involved. If $\pi_1 + \pi_2 \leq 1$, then with probability $\pi_1/(\pi_1 + \pi_2)$, unit 2 is eliminated while unit 1 gets the cumulated probability $\pi_1 + \pi_2$; with probability $\pi_2/(\pi_1 + \pi_2)$, unit 1 is eliminated while unit 2 gets the cumulated probability $\pi_1 + \pi_2$. If $\pi_1 + \pi_2 > 1$, then with probability $(1 - \pi_2)/(2 - \pi_1 - \pi_2)$, unit 1 is selected while unit 2 gets the residual probability $\pi_1 + \pi_2 - 1$; and with probability $(1 - \pi_1)/(2 - \pi_1 - \pi_2)$, unit 2 is selected while unit 1 gets the residual probability $\pi_1 + \pi_2 - 1$. In other words, units 1 and 2 fight. If $\pi_1 + \pi_2 \leq 1$, the loser is definitely eliminated while the winner gets the cumulated probability. If $\pi_1 + \pi_2 > 1$, the winner is selected in the sample while the loser goes on with the residual probability. In any case, the remaining unit then faces unit 3 in a similar principle. The algorithm stops at step $N - 1$, when the two last units fight.

This sampling algorithm has some appealing properties, but we need further notation for their exposition. We define $V_k = \sum_{l=1}^{k} \pi_l$ for any unit $k \in U$, and $V_0 = 0$. A unit $k \in U$ is said to be *cross-border* if $V_{k-1} < i$ and $V_k \geq i$ for some positive integer $i$. The cross-border units are denoted

as $k_i,\ i = 0, \ldots, n$. We note $a_i = i - V_{k_i - 1}$ and $b_i = V_{k_i} - i$ for $i = 1, \ldots, n-1$. For $k_0$ and $k_n$, we take by convention $a_0 = b_0 = 0$ and $a_n = b_n = 0$. The units $k_0$ and $k_n$ are in fact phantom units with zero associated probabilities.

The $N$ sampling units are grouped to obtain a population $U_c = \{u_0, \ldots, u_{2n}\}$ of clusters. There are the clusters of cross-border units ($n+1$ singletons), denoted as $u_{2i}$ with associated probability $\phi_{2i} = \pi_{k_i} = a_i + b_i$ for $i = 0, \ldots, n$. There are the $n$ clusters of units that are not cross-borders and are between two consecutive integers, denoted as $u_{2i-1}$ with associated probability $\phi_{2i-1} = V_{k_i - 1} - V_{k_{i-1}} = 1 - b_{i-1} - a_i$, for $i = 1, \ldots, n$. We note $\phi = (\phi_0, \ldots, \phi_{2n})^\top$. Let $Y_i = \sum_{k \in u_i} y_k$ denote the subtotal of the variable $y$ on the cluster $u_i$, and $\check{Y}_i = Y_i / \phi_i$, with $\check{Y}_0 = \check{Y}_{2n} = 0$. To fix ideas, useful quantities for population $U_c$ are presented in Figure 2.1.

The microstratum $U_i,\ i = 1, \ldots, n$, is defined as

$$U_i = \{k \in U;\ k_{i-1} \leq k \leq k_i\}. \tag{2.6.1}$$

For any unit $k \in U_i$, we note

$$\alpha_{ik} \quad = \quad \begin{cases} b_{i-1} & \text{if } k = k_{i-1}, \\ \pi_k & \text{if } k_{i-1} < k < k_i, \\ a_i & \text{if } k = k_i. \end{cases} \tag{2.6.2}$$

We have in particular $\sum_{k \in U_i} \alpha_{ik} = 1$. To fix ideas, useful quantities for population $U$ are presented in Figure 2.2. The microstrata are overlapping, since one cross-border unit belongs to two adjacent microstrata: the cross-border unit $k_i$ belongs both to the microstratum $U_i$ (with an associated probability $a_i$) and to the microstratum $U_{i+1}$ (with an associated probability $b_i$).

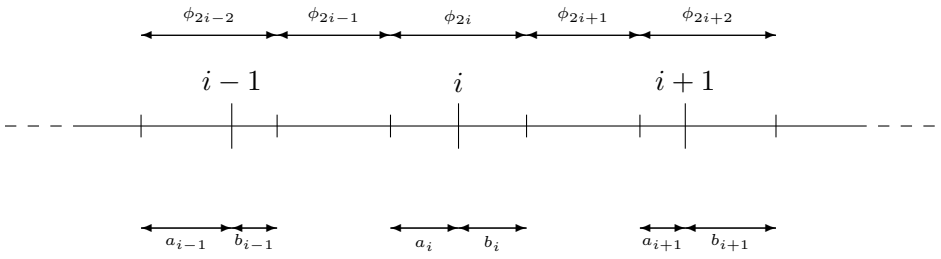


Figure 2.1: Probabilities and cross-border units in the population $U_c$

It is demonstrated in [11] that ordered pivotal sampling is equivalent to the sampling algorithm known as *Deville's systematic sampling* (Deville, 1998), in the sense that both algorithms lead

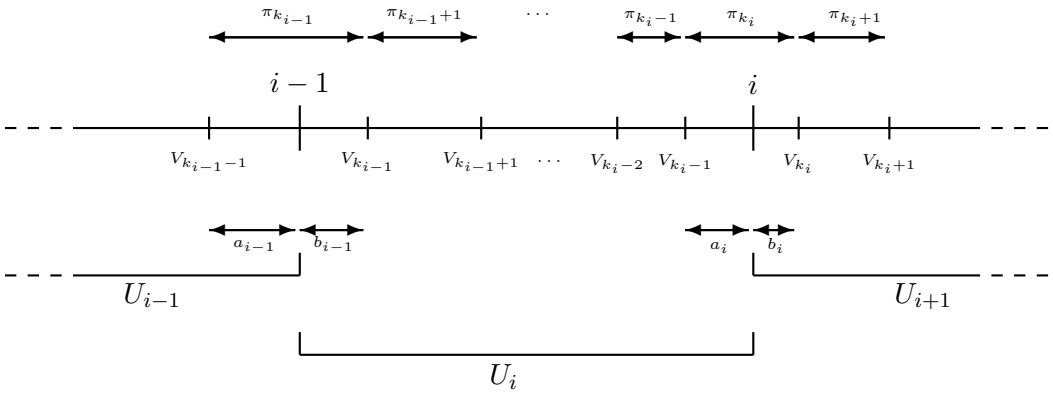


Figure 2.2: Probabilities and cross-border units in microstratum $U_i$, for population $U$

to the same sampling design. This characterization enables in particular the computation of the second-order inclusion probabilities, which are given below:

**Theorem 2.6.1** (Deville, 1998)**.** *Let $k$ and $l$ be two distinct units in $U$. If $k$ and $l$ are two non cross-border units that belong to the same microstratum $U_i$, then $\pi_{kl} = 0$, if $k$ and $l$ are two non cross-border units that belong to distinct microstrata $U_i$ and $U_j$, respectively, where $i < j$, then*

$$\pi_{kl} = \pi_k \pi_l \left\{1 - c(i,j)\right\},$$

*if $k = k_{i-1}$ and $l$ is a non cross-border unit that belongs to the microstratum $U_j$ where $i \leq j$, then*

$$\pi_{kl} = \pi_k \pi_l \left[1 - b_{i-1}(1-\pi_k)\left\{\pi_k(1-b_{i-1})\right\}^{-1} c(i,j)\right],$$

*if $l = k_{j-1}$ and $k$ is a non cross-border unit that belongs to the microstratum $U_i$ where $i < j$, then*

$$\pi_{kl} = \pi_k \pi_l \left\{1 - (1-\pi_l)(1-b_{j-1})(\pi_l b_{j-1})^{-1} c(i,j)\right\},$$

*if $k = p_{i-1}$ and $l = p_{j-1}$, where $i < j$, then*

$$\pi_{kl} = \pi_k \pi_l \left[1 - b_{i-1}(1-b_{j-1})(1-\pi_k)(1-\pi_l)\left\{\pi_k \pi_l b_{j-1}(1-b_{i-1})\right\}^{-1} c(i,j)\right],$$

*where $c(i,j) = \prod_{l=i}^{j-1} c_l$, $c_l = a_l b_l \left\{(1-a_l)(1-b_l)\right\}^{-1}$ and with $c(i,i) = 1$.*

As noticed by Deville (1998), it follows from Theorem 2.6.1 that many of the second-order inclusion probabilities are zero. As a result, no unbiased variance estimator may be found for the HT-estimator. In [20], using alternative representations of ordered pivotal sampling and multinomial sampling, it is proved that any implementation of pivotal sampling is more efficient than multinomial sampling.

## 2.7 Future work

As shown above, ordered pivotal sampling has several interesting features which make it appealing for sample selection. Since it is more efficient than multinomial sampling, the HT-estimator is weakly consistent under a second-order moment condition (see Section 1.3). A result of asymptotic normality for the HT-estimator has been proved in [25], but under the assumption that the inclusion probabilities tend to zero. The use of a coupling algorithm (see Section 5) to weaken this assumption is currently under study. Also, since it may not exist a design unbiased variance estimator for the HT-estimator, this is desirable to exhibit a variance estimator consistent under reasonable assumptions. These aspects are currently under study.

These properties (weak consistency of the HT-estimator, asymptotic normality, weak consistency of a variance estimator) are also of interest for the general Cube algorithm, though they seem much more difficult to prove. The comparison of the penalized balanced sampling strategy with the nonparametric model-assisted estimators considered in Goga and Ruiz-Gazen (2014) is also a matter of further research.

# Chapter 3

# Treatment of item non-response

Imputation is typically used in surveys to compensate for item non-response. It consists of replacing missing values with artificial values in order to reduce the bias and possibly control the variance due to non-response. Imputation methods may be classified into two broad classes: deterministic and random. Unlike random imputation methods, if the imputation process is repeated, deterministic methods yield a fixed imputed value given the sample.

We denote by $S_r$ of size $n_r$ the subset of respondents for the variable $y$ in the sample $S$, and by $r_k$ a response indicator for unit $k$. Let $p_k$ be the response probability of unit $k$. We assume that the units respond independently of one another, so that $p_{kl} \equiv Pr(r_k = 1, r_l = 1) = p_k p_l$ for $k \neq l$. We assume there exists a constant $\kappa > 0$ such that $\kappa < p_k$ for any $k \in s$, so that the response probability is bounded away from 0. An estimated response probability attached to unit $k$ is denoted as $\hat{p}_k$.

In case of imputation, an artificial value $y_k^*$ is used to replace the missing $y_k$ and leads to the imputed estimators. For example, the imputed version of the HT-estimator $\hat{t}_{y\pi}$ is

$$\hat{t}_{yI} = \sum_{k \in S} d_k r_k y_k + \sum_{k \in S} d_k (1 - r_k) y_k^*, \tag{3.0.1}$$

and the imputed version of the estimated distribution function

$$\hat{F}_N(t) = \sum_{k \in S} \tilde{d}_k 1(y_k \leq t) \tag{3.0.2}$$

is

$$\hat{F}_I(t) = \sum_{k \in S} \tilde{d}_k r_k 1(y_k \leq t) + \sum_{k \in S} \tilde{d}_k (1 - r_k) 1(y_k^* \leq t), \tag{3.0.3}$$

where $\tilde{d}_k = (\sum_{l \in s} d_l r_l)^{-1} d_k$. In what follows, we assume that Max $\tilde{d}_k = O(n^{-1})$, so that no extreme sampling weight dominates the others.

Under item non-response, inference may be based on the modeling of the imputed variable or/and of the response mechanism. These two approaches are presented in Section 3.1, and deterministic and random imputation methods are introduced in Section 3.2. The use of random imputation methods results in an additional source of variability, and balanced random imputation methods may be used to reduce or eliminate this imputation variance. In Section 3.3, we explain how the cube method can be adapted for balanced imputation. In Section 3.4, the particular case of estimating the population distribution function is considered. It is shown that an appropriate balanced imputation method leads to a consistent estimation of the distribution function, and a doubly robust imputation procedure is proposed under the common mean model within imputation cells. Tailor-made imputation methods, so as to account for the particular features of imputed data, are presented in Section 3.5.

## 3.1 Approaches for inference

Many imputation methods used in practice can be motivated by the general model

$$m: \; y_k = f(z_k; \beta) + \sigma v_k^{1/2} \epsilon_k, \tag{3.1.1}$$

where $f(\cdot;\cdot)$ is a given function, $z_k$ is a $K$-vector of auxiliary variables available at the imputation stage for all $k \in s$, $\beta$ is a $K$-vector of unknown parameters, $\sigma^2$ is an unknown parameter and $v_k$ is a known constant. We assume that the components of $z_k$ and the number $K$ of components are bounded. The $\epsilon_k$ are independent and identically distributed random variables with mean 0 and variance 1, and their common distribution function is denoted by $F_\epsilon(\cdot)$. The model (3.1.1) is often called an imputation model (e.g., Särndal, 1992). To simplify the presentation, we let $f(z_k; \beta) = z_k^\top \beta$ in (3.1.1), which leads to the imputation regression model

$$m: \; y_k = z_k^\top \; \beta + \sigma v_k^{1/2} \epsilon_k, \tag{3.1.2}$$

We assume that the data are Missing At Random (e.g., Rubin, 1976):

$$E_{\{z_k, r_k = 1\}} (y_k) = E_{\{z_k, r_k = 0\}} (y_k) \, . \tag{3.1.3}$$

The first approach for inference is called the Imputation Model (IM). Inference is made with respect to the joint distribution induced by the imputation model, the sampling design, and the non-response model, but an explicit modeling of the response probabilities is not needed. In such case, we assume that (3.1.1) and (3.1.3) hold.

Alternatively, we may assume that the response probabilities follow some parametric non-response model

$$p_k = p\left(z_k, \alpha\right) \tag{3.1.4}$$

for some vector of unknown parameters $\alpha$. The estimated response probability is then $\hat{p}_k = p\left(z_k, \hat{\alpha}\right)$, where $\hat{\alpha}$ is an estimator of $\alpha$. The second approach for inference is called the Non-response Model (NM). Inference is made with respect to the joint distribution induced by the sampling design and the assumed non-response model in (3.1.4), but an explicit modeling of the variable of interest $y$ is not needed.

Imputation procedures that lead to a consistent estimator if either the imputation model (3.1.1) and/or the non-response model (3.1.4) is correctly specified are often called doubly robust procedures; e.g., Haziza and Rao (2006), Haziza (2009). Doubly robust procedures provide some protection when either the non-response model or the imputation model is misspecified.

## 3.2 Imputation methods

In case of deterministic imputation motivated by the imputation regression model (3.1.2), the imputed value is

$$y_k^* = z_k^\top \hat{B}_r \quad \text{where} \quad \hat{B}_r = \left(\sum_{k\in S} \omega_k r_k v_k^{-1} z_k z_k^\top\right)^{-1} \sum_{k\in S} \omega_k r_k v_k^{-1} z_k y_k, \tag{3.2.1}$$

and $\omega_k$ is an imputation weight attached to unit $k$. Several choices of $\omega_k$ are possible, depending of the approach used for inference. The choice $\omega_k = d_k$ leads to the customary survey weighted imputation, whereas the choice $\omega_k = 1$ leads to unweighted imputation. Both choices lead to an imputed estimator $\hat{t}_{yI}$ approximately unbiased for $t_y$ under the IM approach. The choice $\omega_k = d_k \hat{p}_k^{-1}(1-\hat{p}_k)$ leads to an imputed estimator $\hat{t}_{yI}$ approximately unbiased under both the IM approach and the NM approach; see Haziza and Rao (2006). Hence, this last choice provides a doubly robust estimator for $t_y$. We note $\tilde{\omega}_k = (\sum_{l\in s} \omega_l r_l)^{-1}\omega_k$, and in what follows we assume that Max $\tilde{\omega}_k = O(n^{-1})$.

Random imputation can be seen as a modified deterministic imputation to which a random noise $\epsilon_k^*$ is added. That is, the imputed value is

$$y_k^* = z_k^\top \hat{B}_r + \hat{\sigma} v_k^{1/2} \epsilon_k^*, \tag{3.2.2}$$

where $\hat{\sigma}$ is an estimator of $\sigma$. Although they may be generated from a given parametric distribution, it is natural to select the quantities $\epsilon_k^*$ at random from the empirical distribution function of the respondent residuals. More precisely, denote by $e_k = \hat{\sigma}^{-1} v_k^{-1/2}\left\{y_k - z_k^\top \hat{B}_r\right\}$ the estimated residual with mean $\bar{e}_r = \sum_{k\in s} \tilde{\omega}_k r_k e_k$. We assume that there exists a vector $a$ of known constants such that

$$v_k^{1/2} \quad = \quad a^\top z_k \tag{3.2.3}$$

so that $\bar{e}_r = 0$. As argued by Deville and Särndal (1994), imposing the variance structure in (3.2.3) does not severely restrict the range of imputation models, and many special cases of (3.1.2) satisfy this condition. The random residuals $\epsilon_k^*$ are selected independently and with replacement from the set, $E_r = \{e_l \ ; \ l \in s_r\}$ of standardized residuals observed from the responding units, with probabilities

$$Pr\left(\epsilon_k^* = e_l\right) = \tilde{\omega}_l. \tag{3.2.4}$$

This method for selecting the random residuals $\epsilon_k^*$ is nonparametric in nature since it consists of generating random residuals from the empirical distribution function of the respondent residuals

$$\hat{F}_{\epsilon,r}(t) = \sum_{k \in s} \tilde{\omega}_k r_k 1(e_k \leq t), \tag{3.2.5}$$

where $1(\cdot)$ is the usual indicator function. Random hot-deck imputation is a special case of random regression imputation with $z_k = 1$ and $v_k = 1$ for all $k$.

## 3.3 Balanced random imputation

One drawback of random imputation methods is that they introduce additional variability due to the random selection of residuals. In some cases, the contribution of the imputation variance is appreciable resulting in potentially inefficient estimators. In the literature, three general approaches for reducing the imputation variance have been considered. The fractional imputation approach consists of replacing each missing value with $M \geq 2$ imputed values selected randomly, and assigning a weight to each imputed value (Kalton and Kish, 1981, 1984; Fay, 1996; Kim and Fuller, 2004; Fuller and Kim, 2005). It can be shown that the imputation variance decreases as $M$ increases. The second approach consists of first imputing the missing values using a standard random imputation method, and then adjusting the imputed values in such a way that the imputation variance is eliminated; see Chen, Rao and Sitter (2000). The third approach that we study consists of selecting donors or residuals at random in such a way that the imputation variance is eliminated (Kalton and Kish, 1981, 1984; Deville, 2006).

We consider the case when the total $t_y$ is estimated. Using imputed values given by (3.2.2), the imputed estimator may be written as

$$\hat{t}_{yI} \quad = \quad \sum_{k \in S} d_k r_k y_k + \sum_{k \in S} d_k (1 - r_k)(z_k^\top \hat{B}_r) + \hat{\sigma} \sum_{k \in S} d_k (1 - r_k)(v_k^{1/2} \epsilon_k^*). \tag{3.3.1}$$

In (3.3.1), the imputation variance is only due to the third term on the right-hand side. A balanced random imputation method is proposed in [6]. It consists of selecting the residuals $\epsilon_k^*$ so that

$$\sum_{k \in S} d_k (1 - r_k)(v_k^{1/2} \epsilon_k^*) = 0. \tag{3.3.2}$$

If the equation (3.3.2) is exactly satisfied, then the imputation variance is completely eliminated and the resulting estimator is fully efficient (Kim and Fuller, 2004). In some situations, equation (3.3.2) may only be approximately satisfied and the imputation variance is not completely eliminated but is expected to be significantly reduced. Additional constraints may be added for the selection of the residuals if it is desired to eliminate the imputation variance for other parameters.

Equation (3.3.2) may be seen as a *balancing equation* which is imposed in the with-replacement selection of the random residuals $\epsilon_k^*$ in the set $E_r$ of observed standardized residuals. So as to adapt the Cube method to the with-replacement set-up, we consider the population of cells $U^*$ of size $n_m \times n_r$ given in Table 3.1. Each cell $(k,l)$ is given the value of the standardized residual $e_l$ and the probability of selection $\psi_{kl} = \tilde{\omega}_l$. A random imputation obtained from (3.2.2) may alternatively be seen as selecting a random sample $S^*$ of cells in $U^*$ without replacement, where the non-respondent $k$ is given the residual associated to the respondent $l$ if the cell $(k,l)$ is selected in $S^*$. The sample must be drawn so that each cell has a probability of selection equal to $\psi_{kl}$, and so that exactly one cell per row is selected in $S^*$, since one residual exactly must be selected for each nonrespondent. This is equivalent to select $S^*$ while respecting the system of $n_m$ balancing equations

$$\sum_{(k,l)\in S^*} \frac{x_{kl}}{\psi_{kl}} = \sum_{(k,l)\in U^*} x_{kl} \tag{3.3.3}$$

on a $n_m$ vector of variables $x = (x^1, \ldots, x^{n_m})^\top$, where the variable $x^i$ takes the value $x_{kl}^i = \psi_{kl}\delta_{ik}$ on the cell $(k,l)$, and $\delta_{ik}$ equals 1 if $k = i$ and 0 otherwise. The equation (3.3.2) may be written as the additional balancing equation

$$\sum_{(k,l)\in S^*} \frac{x_{kl}^0}{\psi_{kl}} = \sum_{(k,l)\in U^*} x_{kl}^0, \tag{3.3.4}$$

with $x_{kl}^0 = d_k v_k^{1/2} \psi_{kl} e_l$ for the cell $(k,l)$. Selecting the sample $S^*$ balanced on variables $\tilde{x} = (x^\top, x^0)^\top$ with inclusion probabilities $\psi_{kl}$ ensures that each non-respondent is given a random residual such that the variance imputation is eliminated.

In practice, there may exist no sample $S^*$ such that both equations (3.3.3) and (3.3.4) are exactly satisfied. The Cube method then involves the landing phase in order to end the sampling while exactly respecting the inclusion probabilities. A careful treatment of this landing phase is needed since the balancing equations (3.3.3) must be preserved until the very end of the selection procedure. The stratified balanced sampling algorithm proposed in [3] may be used for this purpose, see also Hasler and Tillé (2014). It is proved in [7] that the landing phase involves no more than two units. Since the balancing constraint $x^0$ is maintained during the whole sampling process, except perhaps for the last step, the imputation variance will be considerably reduced, though perhaps not totally eliminated. An imputation procedure for quantitative variables where the imputation

Table 3.1: Population of cells used for the random selection od residuals

| | 1 | ... | $j$ | ... | $n_r$ |
|---|---|---|---|---|---|
| 1 | $(\psi_{11}, e_1)$ | ... | $(\psi_{1j}, e_j)$ | ... | $(\psi_{1n_r}, e_{n_r})$ |
| ... | ... | | ... | | ... |
| $i$ | $(\psi_{i1}, e_i)$ | ... | $(\psi_{ij}, e_j)$ | ... | $(\psi_{in_r}, e_{n_r})$ |
| ... | ... | | ... | | ... |
| $n_m$ | $(\psi_{n_m1}, e_1)$ | ... | $(\psi_{n_mj}, e_j)$ | ... | $(\psi_{n_mn_r}, e_{n_r})$ |

variance can be fully eliminated is studied in [26].

The proposed imputation method may be readily extended to the case of a categorical variable $y$ with $Q$ possible characteristics. The population $U^*$ is then constituted of $n_m \times Q$ cells, each column being associated to one of the possible characteristics of $y$. The random balanced imputation process then follows the same lines as described above, each non-respondent $i$ being given the $j$-th characteristic of the variable $y$ if the cell $(i, j)$ is selected in $s^*$.

## 3.4 Estimation of the distribution function

While deterministic imputation methods lead to asymptotically unbiased estimators of totals if the underlying imputation or non-response model is correctly specified, they are not appropriate when the objective is to estimate the distribution function because this type of imputation tends to distort the distribution of the variables being imputed. To preserve distributions, it is customary to use some form of random imputation.

The asymptotic properties of the estimated distribution function, under the random regression imputation described in (3.2.2), are studied in [7]. Under the additional assumptions that $(\hat{B}_r, \hat{\sigma})$ is weakly consistent for $(\beta, \sigma)$ and that the distribution function of residuals $F_\epsilon(\cdot)$ is absolutely continuous, they prove that under the IM approach

$$\hat{F}_I(t) - F_N(t) \underset{Pr}{\longrightarrow} 0, \tag{3.4.1}$$

where $\underset{Pr}{\longrightarrow}$ stands for the convergence in probability. It is also proved in [7] that equation (3.4.1) remains true when using the balanced random regression imputation procedure described in Section 3.3, where the random residuals are selected so that (3.3.2) is satisfied.

Doubly robust estimation for the distribution function is considered in [21], assuming a particular form of the imputation model. In this case, the population $U$ is divided into $G$ mutually disjoint

imputation cells $U_1, \ldots, U_G$. The elements in $U_g$ are assumed to be a realization of independently and identically distributed random variables with mean $\mu_g$ and variance $\sigma_g^2$, which we note as

$$m : y_i \sim (\mu_g, \sigma_g^2), \quad i \in U_g, \tag{3.4.2}$$

and which is called the common mean model within imputation cells. This model is frequently used in practice, with the imputation cells being formed on the basis of auxiliary information recorded for both respondents and non-respondents. This is a special case of the imputation regression model (3.1.2), with $z_k = \{1(k \in U_1), \ldots, 1(k \in U_G)\}^\top$, $\beta = (\mu_1, \ldots, \mu_G)^\top$ and $v_k = (\sigma_g/\sigma)^2$ for $k \in U_g$. Under the imputation model (3.4.2), a missing value is treated through random hot-deck imputation within imputation cells where missing $y_k$ in cell $U_g$ is replaced with

$$y_k^* = y_l \quad \text{for} \quad l \in s_{rg}, \tag{3.4.3}$$

with probability

$$Pr(y_k^* = y_l) = \tilde{\omega}_k \quad \text{where} \quad \omega_k = d_k \frac{1 - \hat{p}_k}{\hat{p}_k},$$

and where $s_{rg} = s_r \cap U_g$.

The use of the imputed values in (3.4.3) leads to a doubly robust estimator of a population total. Also, from (3.4.1), $\hat{F}_{I,y}(t) - F_{N,y}(t)$ converges in probability to 0 under the IM approach for any $t \in \mathbb{R}$. Under some additional regularity conditions, it is proved in [21] that (3.4.1) remains valid under the NM approach. Therefore, the imputed distribution function using the imputed values in (3.4.3) is doubly robust.

## 3.5 Tailor-made imputation methods

The imputation regression model in (3.1.2) is not always appropriate. The variable of interest $y$ may be better modeled as a mixture of variables, in which case the model (3.1.2) does not properly reflect the mixture of values. Also, if we are interested in multivariate parameters, it is necessary to model the joint distribution of the variables involved. Such situations are treated below.

### 3.5.1 Zero-inflated data

In some cases, the study variables contain a large number of zeroes. This situation is frequent in business surveys, which collect economic variables (revenue, expenses, etc.). In statistical agencies, it is customary to use some form of regression imputation to fill in the missing values. In the presence of zeroes to item $y$, the finite population $U$ can be viewed as the mixture of a subpopulation $U_0$ of

units for which $y_k = 0$ and of a subpopulation $U_1$ for which $y$ is strictly positive. For this type of populations, it seems natural to postulate the following mixture regression model:

$$m : y_k = \begin{cases} z_k^\top \beta + \epsilon_k & \text{if } \delta_k = 1, \\ 0 & \text{if } \delta_k = 0, \end{cases} \tag{3.5.1}$$

where $\delta_k$ follows a Bernoulli distribution with parameter $\phi_k$. We assume that $\epsilon_k$ and $\delta_k$ are not related after accounting for $z_k$. Also, we assume that conditionally on $\delta_k = 1$ the variable $y_k$ follows the imputation regression model in (3.1.2), so that

$$E(\epsilon_k) = 0, \qquad Cov(\epsilon_k, \epsilon_l) = 0 \text{ for } k \neq l, \qquad V(\epsilon_k) = \sigma^2(a^\top z_k) \tag{3.5.2}$$

where the variance structure $V(\epsilon_k)$ follows from (3.2.3). In practice, the $\phi_k$'s are unknown and need to be estimated. We assume that

$$\phi_k \quad = \quad f(u_k, \gamma), \tag{3.5.3}$$

for some function $f(\cdot, \cdot)$, where $u_k$ is a vector of auxiliary variables attached to unit $k$ and $\gamma$ is a vector of unknown parameters. An estimate of $\phi_k$ is

$$\hat{\phi}_k \quad = \quad f(u_k, \hat{\gamma}), \tag{3.5.4}$$

where $\hat{\gamma}$ is an estimator of $\gamma$. We assume that the data are missing at random. In the context of zero-inflated data, the IM approach assumes that (3.1.3), (3.5.1) and (3.5.3) hold.

Several imputation procedures motivated by the mixture regression model (3.5.1) are proposed in [15]. The first proposal, called deterministic $\phi$-regression imputation, consists in replacing missing $y_k$ with

$$y_k^* = \hat{\phi}_k z_k \hat{B}_{\phi r} \quad \text{where} \quad \hat{B}_{\phi r} = \left( \sum_{k \in S} \omega_k r_k \hat{\phi}_k^{-1} v_k^{-1} z_k z_k^\top \right)^{-1} \sum_{k \in S} \omega_k r_k v_k^{-1} z_k y_k \tag{3.5.5}$$

and $\omega_k = d_k \hat{p}_k^{-1}(1 - \hat{p}_k)$. Under some regularity conditions, they demonstrate that $\hat{t}_{yI}$ is consistent for $t_y$ under both the IM approach and the NM approach. The estimator $\hat{t}_{yI}$ that uses the imputed values (3.5.6) is thus doubly robust, but it does not respect the mixed structure of the data. Therefore, an alternative procedure is proposed in [15], called random $\phi$-regression imputation, which consists in replacing missing $y_k$ with

$$y_k^* = \begin{cases} z_k^\top \hat{B}_{\phi r} & \text{with probability } \hat{\phi}_k, \\ 0 & \text{with probability } 1 - \hat{\phi}_k. \end{cases} \tag{3.5.6}$$

Under the same regularity conditions, it is demonstrated that $\hat{t}_{yI}$ is also consistent for $t_y$ under both the IM approach and the NM approach. A balanced version of random $\phi$-regression imputation is also proposed in [15].

### 3.5.2 Continuous bivariate data

Sometimes, the interest lies in estimating bivariate parameters such as the correlation coefficients

$$\rho_{xy} = \frac{t_{11} - t_{10}t_{01}/N}{\left(t_{20} - t_{10}^2/N\right)^{1/2}\left(t_{02} - t_{01}^2/N\right)^{1/2}},$$

where $t_{ij} = \sum_{k\in U} x_k^i y_k^j$ for $(i,j) \in \{(1,0),(2,0),(1,1),(0,1),(0,2)\}$. The imputed version of $\rho_{xy}$ is

$$\rho_{xyI} = \frac{t_{11,I} - t_{10,I}t_{01,I}/\hat{N}}{\left(t_{20,I} - t_{10,I}^2/\hat{N}\right)^{1/2}\left(t_{02,I} - t_{01,I}^2/\hat{N}\right)^{1/2}}, \tag{3.5.7}$$

where $\hat{t}_{ij,I} = \sum_{k\in S} d_k\{r_{xk}x_k + (1-r_{xk})x_k^*\}^i\{r_{yk}y_k + (1-r_{yk})y_k^*\}^j$, where $r_{xk}$ is a response indicator for $x_k$, and $r_{yk}$ is a response indicator for $y_k$. While marginal imputation is appropriate for univariate parameters, it may lead to considerably biased estimators of bivariate parameters such as $\rho_{xy}$. Extending the work of Srivastava and Carter (1986), Shao and Wang (2002) considered the bivariate imputation model:

$$m: \begin{array}{rcl} y_k &=& z_k^\top\beta + \sigma_\epsilon\sqrt{v}_k\epsilon_k, \\ x_k &=& z_k^\top\gamma + \sigma_\eta\sqrt{u}_k\eta_k, \end{array} \tag{3.5.8}$$

where $z_k$ is a $K$-vector of auxiliary variables, $\beta$ and $\gamma$ are unknown $K$-vectors of parameters, $v_k$ and $u_k$ are known, and $\epsilon_k$ and $\eta_k$ are random terms independent of $z_k$ with mean 0 and variance 1. Note that $\epsilon_k$ and $\eta_k$ are not independent in general, and their covariance is denoted as $\sigma_{\epsilon\eta}$. In practice, we often have $\sigma_{\epsilon\eta} > 0$.

Shao and Wang (2002) showed that marginal random regression imputation does not preserve the coefficient of correlation between the study variables $x$ and $y$. Motivated by (3.5.8), they proposed a joint random regression imputation procedure, which can be described as follows:

(i) If $y_k$ is observed and $x_k$ is missing, we use the imputed values

$$x_k^* = z_k^\top\hat{\gamma}_r + \frac{\sqrt{u_k}\hat{\sigma}_{\epsilon\eta}}{\sqrt{v_k}\hat{\sigma}_\epsilon^2}\left(y_k - z_k^\top\hat{\beta}_r\right) + \sqrt{u_k}\eta_k^*,$$

where

$$\hat{\gamma}_r = \left(\sum_{k\in S} d_k r_{xk} u_k^{-1} z_k z_k^\top\right)^{-1} \sum_{k\in S} d_k r_{xk} u_k^{-1} z_k x_k, \tag{3.5.9}$$

$$\hat{\beta}_r = \left(\sum_{k\in S} d_k r_{yk} v_k^{-1} z_k z_k^\top\right)^{-1} \sum_{k\in S} d_k r_{yk} v_k^{-1} z_k y_k, \tag{3.5.10}$$

and, given the observed data, the $\eta_k^*$'s are independent random variables with mean 0 and variance $\tilde{\sigma}_\eta^2 = \hat{\sigma}_\eta^2 - \hat{\sigma}_{\epsilon\eta}^2/\hat{\sigma}_\epsilon^2$ with

$$\hat{\sigma}_\epsilon^2 = \frac{1}{\sum_{l\in S} w_l r_{xl} r_{yl}} \sum_{l\in s} w_l r_{xl} r_{yl} v_l^{-1} \left(y_l - z_l^\top \hat{\beta}_r\right)^2, \tag{3.5.11}$$

$$\hat{\sigma}_\eta^2 = \frac{1}{\sum_{l\in S} w_l r_{xl} r_{yl}} \sum_{l\in S} w_l r_{xl} r_{xl} u_l^{-1} \left(x_l - z_l^\top \hat{\gamma}_r\right)^2, \tag{3.5.12}$$

$$\hat{\sigma}_{\epsilon\eta} = \frac{1}{\sum_{l\in S} w_l r_{xl} r_{yl}} \sum_{l\in S} w_l r_{xl} r_{yl} u_l^{-1/2} v_l^{-1/2} \left(x_l - z_l^\top \hat{\gamma}_r\right)\left(y_l - z_l^\top \hat{\beta}_r\right). \tag{3.5.13}$$

(ii) If $x_k$ is observed and $y_k$ is missing, the imputation procedure is similar.

(iii) If both $x_k$ and $y_k$ are missing, we use the imputed values

$$\begin{aligned} x_k^* &= z_k^\top \hat{\gamma}_r + \sqrt{u_k}\eta_k^* \\ y_k^* &= z_k^\top \hat{\beta}_r + \sqrt{v_k}\epsilon_k^*, \end{aligned}$$

where $(\epsilon_k^*, \eta_k^*)$'s are independently distributed with mean 0 and covariance matrix

$$\hat{\Sigma}_1 = \begin{pmatrix} \hat{\sigma}_\epsilon^2 & \hat{\sigma}_{\epsilon\eta} \\ \hat{\sigma}_{\epsilon\eta} & \hat{\sigma}_\eta^2 \end{pmatrix}$$

Under the IM approach, Shao and Wang (2002) showed that this joint regression imputation method leads to asymptotically unbiased estimators of coefficients of correlation, provided the $\tilde{\epsilon}_k^*$'s and $\tilde{\eta}_k^*$'s are independently selected from any distribution with appropriate mean and variance. They argued that the random residuals should be generated from the respondents' residuals if other non-linear parameters such as quantiles are of interest.

The properties of the Shao and Wang procedure under a particular bivariate non-response model are studied in [10]. If $p_{k,\diamond}$ stands for the probability of pattern $\diamond \in \{rr, rm, mr, mm\}$ for unit $k$, this non-response model assumes that a given pattern occurs with the same probability for any unit $k \in S$, so that we may simply note $p_{k,\diamond} = p_\diamond$. Also, it assumed that the sample units respond independently of one another. Then Chauvet and Haziza (2012) proved that under the Shao-Wang procedure, the imputed estimator $\hat{\rho}_{xyI}$ is asymptotically unbiased for $\rho_{xy}$ under the NM approach when the first component of the $z_k$-vector is equal to 1 and $u_k = v_k = 1$.

A balanced version of the Shao and Wang imputation procedure is also proposed in [10]. It succeeds in preserving the coefficient of correlation, while being fully efficient for this parameter.

### 3.5.3 Categorical bivariate data

In household and social surveys, variables are often categorical so that the methods described above are not directly applicable: rather than dealing with means and correlations, we are interested in marginal and joint proportions. Let $x$ denote a study variable with possible characteristics $a = 1, \ldots, A$. Similarly, let $y$ denote a study variable with possible characteristics $b = 1, \ldots, B$. We are interested in estimating $p_{a\bullet} = N^{-1} \sum_{k \in U} 1(x_k = a)$ the marginal proportion of units who possess the characteristic $a$ for $x$; $p_{\bullet b} = N^{-1} \sum_{k \in U} 1(y_k = b)$ the marginal proportion of units who possess the characteristic $b$ for $y$; and $p_{ab} = N^{-1} \sum_{i \in U} 1(x_k = a)1(y_k = b)$ the joint proportion of units who possess both characteristics $a$ for $x$ and $b$ for $y$.

We assume that the finite population $U$ is partitioned into $G$ imputation cells $U_1, \ldots, U_G$ of sizes $N_1, \ldots, N_G$, based on auxiliary variables. Within each class, we assume that the units respond independently of one another. Denote by $s^g = s \cap U^g$ the sample members in class $g$; $s^g_{rr}$ the set of $n^g_{rr}$ respondents to both items in this class; $s^g_{rm}$ the set of $n^g_{rm}$ respondents to just item $x$ in this class; $s^g_{mr}$ the set of $n^g_{mr}$ respondents to just item $y$ in this class; $s^g_{mm}$ the set of $n^g_{mm}$ non-respondents in this class. We note $p^g_{k\diamond} \equiv Pr(k \in s^g_\diamond | k \in s)$ for any pattern $\diamond \in \{rr, rm, mr, mm\}$. We assume that a given pattern occurs with the same probability for any unit $k \in s^g$, so that we simplify the notation as $p^g_{k\diamond} = p^g_\diamond$.

We focus on survey weighted random hot-deck imputation within cells (see Section 3.4), with the choice $\omega_k = d_k$ for the imputation weights. This leads to the following procedure:

(i) for $k \in s^g_{mr}$, missing $x_k$ is imputed by $x^*_k = a$ with probability

$$\hat{p}^g_{a\bullet,ac} \equiv (\hat{N}^g_{r\bullet})^{-1} \sum_{k \in s^g_{r\bullet}} w_k 1(x_k = a) \tag{3.5.14}$$

estimated from the available cases (ac) for item $x$, and $\hat{N}^g_{r\bullet} = \sum_{k \in s^g_{r\bullet}} w_k$;

(ii) for $k \in s^g_{rm}$, missing $y_k$ is imputed by means of an analogous procedure;

(iii) for $k \in s^g_{mm}$, missing $(x_k, y_k)$ is imputed by $(x^*_k, y^*_k) = (a, b)$ with probability

$$\hat{p}^g_{ab,cc} \equiv (\hat{N}^g_{rr})^{-1} \sum_{k \in s^g_{rr}} w_k 1(x_k = a)1(y_k = b) \tag{3.5.15}$$

estimated from the complete cases (cc) for $x$ and $y$, with $\hat{N}^g_{rr} = \sum_{k \in s^g_{rr}} w_k$.

When one variable only is missing, random hot-deck imputation estimates its distribution separately from complete cases for this variable. When both variables are missing, their distribution is estimated jointly from complete cases for both. Random hot-deck imputation succeeds in estimating the marginal distributions of $x$ and $y$, since for any characteristics $a$ and $b$ $B_{qI}(\hat{p}_{a\bullet,I}) \simeq 0$

and $B_{qI}(\hat{p}_{\bullet b,I}) \simeq 0$. Although this imputation procedure generates less bias than marginal random hot-deck imputation, there generally remains some bias when estimating the joint proportions, since:

$$B_{qI}(\hat{p}_{ab,I}) \simeq -\hat{N}^{-1} \sum_{g=1}^{G} (p_{rm}^g + p_{mr}^g) \sum_{k \in s^g} w_k \{1(x_k = a) - \hat{p}_{a\bullet}^g\}\{1(y_k = b) - \hat{p}_{\bullet b}^g\}. \tag{3.5.16}$$

To account for the existing relationship between variables, two imputation procedures are proposed in [16]. The distribution of $x$ is estimated conditionally on $y$ if $x$ only is missing, and e distribution of $y$ is estimated conditionally on $x$ if $y$ only is missing. For any unit $k \in U^g$, we note

$$\hat{p}_{a|b,cc}^g = \frac{\sum_{k \in s_{rr}^g} w_k 1(x_k = a) 1(y_k = b)}{\sum_{k \in s_{rr}^g} w_k 1(y_k = b)}$$

the estimated probability that $x_k = a$ when $y_k = b$, and

$$\hat{p}_{b|a,cc}^g = \frac{\sum_{k \in s_{rr}^g} w_k 1(x_k = a) 1(y_k = b)}{\sum_{k \in s_{rr}^g} w_k 1(x_k = a)}$$

the estimated probability that $y_k = b$ when $x_k = a$. The *joint random hot-deck imputation* procedure is as follows:

(i) for $k \in s_{mr}^g$, missing $x_k$ is imputed by $x_k^* = a$ with probability $\hat{p}_{a|y_k,cc}^g$,

(ii) for $k \in s_{rm}^g$, missing $y_k$ is imputed by $y_k^* = b$ with probability $\hat{p}_{b|x_k,cc}^g$,

(iii) for $k \in s_{mm}^g$, missing $(x_k, y_k)$ is imputed by $(x_k^*, y_k^*) = (a, b)$ with probability $\hat{p}_{ab,cc}^g$.

It can be shown that $B_{qI}(\hat{p}_{\diamond,I}) \simeq 0$ under this imputation procedure, for $\diamond \in \{a\bullet, \bullet b, ab\}$ and any characteristics $a$ and $b$. Guidelines are given in [16] to extend the joint random hot-deck imputation procedure to the case of more than two missing items, and a balanced version of the joint random hot-deck imputation is also described.

## 3.6 Future work

In practice, when using imputed survey data, we may not only be interested in estimating totals or distribution functions, but also complex parameters such as a quantile, a coefficient of regression or a coefficient of logistic regression. It would thus be of interest to develop imputation mechanisms which succeed in obtaining consistent estimators for such parameters, or to prove that existing imputation mechanisms succeed in doing so. Extending the mixture imputation mechanism (see Section 3.5) to a more general mixture imputation model is also currently under investigation.

# Chapter 4

# Variance estimation

Estimators arising from surveys are usually provided with some measure of accuracy, such as a variance estimator, a coefficient of variation or a confidence interval. Variance estimation is usually an intricate task, since it must account for the whole sampling and estimation process, including possible adjustments for non-response. Variance estimation in case of the 2006 French Housing Survey is studied in detail in [13]. The basic national sample was selected by means of stratified multistage sampling with regional extensions, and complementary samples were selected from external databases and local areas, and all samples were joined by composite estimation. Variance estimation for the ELFE sample cohort, with a selection of maternity units, an independent selection of days, and with the survey performed in the maternity units selected for the chosen days, is considered in [23]. The specificity of the ELFE sampling design is that the same sample of days is used for each of the selected maternity units (unlike the usual approach commonly used in two-stage sampling, with independent sub-samplings inside the primary units).

In the situation of full-response, the variance of the HT-estimator may be estimated by

$$\hat{V}_{HT}\left(\hat{t}_{y\pi}\right) = \sum_{k,l\in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{4.0.1}$$

for any sampling design, or by

$$\hat{V}_{YG}\left(\hat{t}_{y\pi}\right) = -\frac{1}{2} \sum_{k\neq l\in S} \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2. \tag{4.0.2}$$

if the sampling design is of fixed-size. These estimators are design-unbiased provided that the second-order inclusion probabilities $\pi_{kl}$ are strictly positive. Otherwise, there exists no design-unbiased variance estimator for the HT-estimator and we may resort to model-assisted variance estimation.

The estimators (4.0.1) and (4.0.2) are computable if so are the second-order inclusion probabilities, which is true for common sampling designs but not for some complex sampling strategies. In Section 4.1, we present simulation-based approximations of the variance-covariance matrix in the context of balanced sampling. Also, if we are interested in estimating non-linear parameters, unbiased variance estimators are usually not available. We may then resort to linearization or Bootstrap variance estimation, which is the purpose of Section 4.2. When the sample suffers from item non-response which is handled by means of random imputation, estimators usually suffer from an additional variance which must be accounted for. We justify in Section 4.3 that the use of balanced imputation makes variance estimation easier, since the imputation variance is usually almost eliminated and need therefore not to be accounted for.

## 4.1 Simulation-based variance estimation

In case of balanced sampling by means of the cube method, second-order inclusion probabilities are usually difficult to compute. We may use the maximum-entropy variance approximation in (2.0.4), and substitute each total by its HT estimator in a plug-in principle to get the variance estimator

$$\hat{V}_{DT}(\hat{t}_{y\pi}) = \frac{n}{n-q}\sum_{k\in S}\frac{1-\pi_k}{\pi_k^2}(y_k-\tilde{y}_k^*)^2 \text{ with } \tilde{y}_k^* = x_k^\top\left(\sum_{l\in S}\frac{1-\pi_l}{\pi_l^2}x_l x_l^\top\right)^{-1}\sum_{l\in S}\frac{1-\pi_l}{\pi_l^2}x_l y_l. \quad (4.1.1)$$

The hypothesis of exact balancing assumed by Deville and Tillé (2005) implies that approximation (2.0.4) accounts for the variance due to the flight phase only. Consequently, the variance estimator given in (4.1.1) may lead to serious bias in variance estimation if the variance due to the landing phase is appreciable.

A second approach consists in using a simulation-based approximation of the design variance-covariance matrix $\Delta = (\Delta_{kl})_{k,l\in U}$, see Fattorini (2006), Thompson and Wu (2008) and Lesage (2013). Since $E_{\{y_U\}}[(I-\pi)(I-\pi)^\top] = \Delta$, a first unbiased simulation-based estimation of $\Delta$ is

$$\Delta_{SIM} = \frac{1}{C}\sum_{c=1}^{C}\{I(S_c)-\pi_{SIM}\}\{I(S_c)-\pi_{SIM}\}^\top, \quad (4.1.2)$$

where $S_1,\ldots,S_C$ are $C$ independent replicates of the sample, and $\pi_{SIM} = C^{-1}\sum_{c=1}^{C} I(S_c)$. A corresponding variance estimator for a given sample $S$ is then obtained by plugging (4.1.2) into (4.0.2), which leads to

$$\hat{V}_{SIM}\left(\hat{t}_{y\pi}\right) = -\frac{1}{2}\sum_{k\neq l\in S}\frac{\Delta_{SIM,kl}}{\Delta_{SIM,kl}+\pi_k\pi_l}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2. \quad (4.1.3)$$

This estimator is not exactly unbiased, and has greater variance than $\hat{V}_{HT}$ due to the simulation, but both the bias and the additional variance can be made arbitrarily small for sufficiently large $C$.

It may be used for any fixed-size sampling design, selected or not by means of the cube method.

An alternative simulation-based approximation for $\Delta$, proposed in [5], makes uses of the martingale structure of the cube method. Specifically, note that if the sample $S$ is selected by means of Algorithm 1 or Algorithm 3, we have $I = \pi + \sum_{t=1}^{T} \delta(t)$ where the innovations $\{\delta(t)\}_{t=1,\ldots,T}$ are given in the Algorithm. By construction, $\{\delta(t)\}$ is a martingale difference (MD) sequence with respect to the sequence of sigma-fields $\mathcal{F}_{t-1} = \sigma(\delta(0), \ldots, \delta(t-1))$, and so these random vectors are uncorrelated and have mean zero. This leads to $\Delta = E\left\{\sum_{t=1}^{T} \lambda_1^*(t)\lambda_2^*(t)u(t)u(t)^\top\right\}$. Consequently, the $\Delta$ matrix is unbiasedly estimated by

$$\Delta_{MD} = \frac{1}{C}\sum_{c=1}^{C}\sum_{t=1}^{T} \lambda_1^{*c}(t)\lambda_2^{*c}(t)u^c(t)u^c(t)^\top, \tag{4.1.4}$$

where $S_1, \ldots, S_c, \ldots, S_C$ are $C$ independent replicates of the sample, and the quantities $\lambda_1^{*c}(t)$, $\lambda_2^{*c}(t)$ and $u^c(t)$ are associated to the sample $S_c$. The corresponding variance estimator for a given sample $S$ is then obtained by plugging (4.1.4) into (4.0.2), which leads to

$$\hat{V}_{MD}\left(\hat{t}_{y\pi}\right) = -\frac{1}{2}\sum_{k\neq l\in S} \frac{\Delta_{MD,kl}}{\Delta_{MD,kl} + \pi_k\pi_l}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2. \tag{4.1.5}$$

The simulation results in [5] indicate that the variance estimator $\hat{V}_{MD}\left(\hat{t}_{y\pi}\right)$ is essentially unbiased and tracks well the variance due to the landing phase, but exhibits a large variability as compared to the maximum entropy variance estimator $\hat{V}_{DT}\left(\hat{t}_{y\pi}\right)$. The Martingale Difference variance estimator was also used in [8] for estimators arising from the French Master Sample.

## 4.2 Linearization and replication-based variance estimation

Suppose that we are interested in some parameter $\theta$. Let $M = \sum_{k\in U}\delta_{y_k}$ denote the discrete measure taking unit mass on any point $y_k$ in the population and 0 elsewhere. Most of the parameters of interest $\theta$ studied in surveys can be written as a functional $T$ of $M$, namely $\theta = T(M)$. For instance, the total $t_y = \sum_{k\in U} y_k$ equals $\int y dM$. Let $\hat{M} = \sum_{k\in S} d_k\delta_{y_k}$ denote the discrete measure taking mass $d_k$ on any point in the sample and 0 elsewhere. Substituting $\hat{M}$ into $\theta$ yields the estimator $\hat{\theta} = T(\hat{M})$. In the linear case, the substitution estimator yields the HT estimator for the total $t_y$.

The influence function linearization technique (Deville, 1999) consists in giving a first-order expansion of the substitution estimator $\hat{\theta} = T(\hat{M})$ around the true value $\theta = T(M)$, to approximate the error by a linear estimator of some artificial *linearized variable*. More precisely, the first derivatives of $T$ with respect to $M_1$ are the influence functions

$$IT(M; y) \quad = \quad \lim_{h\to 0}\frac{T(M + h\delta_y) - T(M)}{h},$$

and $u_k = IT(M; y_k)$ is the linearized variable for all $k \in U$. Suppose that $T(\cdot)$ is homogeneous, namely there exists some positive number $\alpha$ dependent on $T$ such that $T(rM) = r^\alpha T(M)$ for any real $r > 0$. Assume also $\lim_{N\to\infty} N^{-\alpha} T(M_1) < \infty$. Under additional regularity assumptions upon $T(\cdot)$ and the sampling design, Deville (1999) establishes that

$$N^{-\alpha}(\hat{\theta} - \theta) \quad = \quad N^{-\alpha} \left( \sum_{k\in s} d_k u_k - \sum_{k\in U} u_k \right) + o_p(n^{-1/2}),$$

so that the error $\hat{\theta} - \theta$ can be approximated by the error of the Horvitz-Thompson estimator for the total of the linearized variable $u_k$. Plugging a sample-based estimator $\hat{u}_k$ of the linearized variable $u_k$ inside (4.0.2) yields the variance estimator of $\hat{\theta}$

$$v_{LIN}(\hat{\theta}) = -\frac{1}{2} \sum_{k\neq l\in S} \frac{\Delta_{kl}}{\pi_{kl}} \left( \frac{\hat{u}_k}{\pi_k} - \frac{\hat{u}_l}{\pi_l} \right)^2. \tag{4.2.1}$$

If in addition the sampling design is such that the HT-estimator satisfies a central-limit theorem, an approximately $(1 - 2\alpha)\%$ confidence interval is $\left[ \hat{\theta}_1 - z_\alpha \sqrt{v_{lin}(\hat{\theta}_1)}, \hat{\theta}_1 + z_\alpha \sqrt{v_{lin}(\hat{\theta}_1)} \right]$ where $z_\alpha$ is the upper $\alpha\%$ cutoff for the standard normal distribution. The linearization technique has been extended to the two-sample case by Goga et al. (2009).

The use of bootstrap techniques in survey sampling has been extensively studied in the literature. The main bootstrap techniques may be thought as particular cases of the weighted bootstrap (Bertail and Combris, 1997; Antal and Tillé, 2011; Beaumont and Patak, 2012); see also Shao and Tu (1995, chap. 6), Davison and Hinkley (1997, section 3.7) and Davison and Sardy (2007) for detailed reviews. Under a weighted bootstrap procedure, the measure $\hat{M} = \sum_s d_k \delta_{y_k}$ is estimated, conditionally on the sample $s$, by the bootstrap measure

$$\hat{M}^* \quad = \quad \sum_{k\in S} d_k D_k \delta_{y_k} \tag{4.2.2}$$

where $D = \{D_k\}_{k\in s}$ denotes a (random) vector of resampling weights. The vector $D$ is usually generated in such a way that the two first moments of the Horvitz-Thompson estimator are matched, at least approximately. That is, we wish to have

$$E_{\{y_U,S\}} \left( \sum_s d_k D_k y_k \right) \simeq \hat{t}_{y1} \quad \text{and} \quad V_{\{y_U,S\}} \left( \sum_s d_k D_k y_k \right) \simeq \hat{V}_{YG}(\hat{t}_{y1}).$$

A bootstrap technique is not suitable for general sampling designs: that is, a particular sampling design usually requires a tailor made resampling scheme.

In case when the sample $S$ is selected by means of SI sampling, we consider the without replacement bootstrap (BWO) introduced by Gross (1980). Suppose that $N/n$ is an integer. Then the vector

$D$ is obtained by, first creating a pseudo-population $U^*$ of size $N$ by duplicating $N/n$ times each unit $k$ in the original sample $s$, and then by selecting a SI resample $S^*$ in $U^*$. The resampling weight $D_k$ is given by the number of times unit $k \in S$ is selected in $s^*$. The building of $U^*$ may be avoided by noting that under the BWO procedure, the vector $D$ follows a multivariate hypergeometric law; therefore, the resampling weights may be directly generated. Several solutions have been proposed to handle the case when $N/n$ is not an integer, see Chao and Lo (1985), Bickel and Freedman (1984), Sitter (1992b), Booth *et al.* (1994), Presnell and Booth (1994), among others. The generalization of BWO variance estimation for unequal probability sampling designs is considered in Särndal *et al.* (2002) and Chauvet (2007, doctoral dissertation). A two-sample BWO technique is developed in [17] for the two-dimensional simple random sampling without replacement.

In case when the sample $S$ is selected by means of multistage sampling, we consider the bootstrap of Primary Sampling Units (PSUs) introduced by Rao and Wu (1988). Assume that the $N$ units are grouped inside $N_I$ non-overlapping PSUs $u_1, \ldots, u_{N_I}$. A with-replacement first-stage sample $S_I$ of size $n_I$ is selected, and a second-stage sample $S_i$ is selected in any $u_i \in S_I$ by means of some sampling design $p_i(\cdot)$. The estimated measure is then $\hat{M} = \sum_{u_i \in S_I} \sum_{k \in S_i} d_k \delta_{y_k}$. Then, a with-replacement resample $S_I^*$ of size $n_I - 1$ is selected in $S_I$, and the bootstrap measure is $\hat{M}^* = \sum_{u_i \in S_I^*} \sum_{k \in S_i} d_k \delta_{y_k}$; the resampling weight $D_k$ is simply the number of times the PSU $u_i \ni k$ is selected in $S_I^*$. The resampling size $n_I - 1$ is used to reproduce the usual unbiased variance estimator in the linear case (see Rao and Wu, 1988). A justification for this technique, making use of coupling arguments, is given in [19] (see Section 5).

Under any weighted bootstrap technique, the plug-in estimator of $\theta = T(M)$ is $\hat{\theta}^* = T(\hat{M}^*)$, and the variance of $\hat{\theta} = T(\hat{M})$ is estimated by

$$V_{\{y_U, S\}}\left(\hat{\theta}^*\right) \quad = \quad E_{\{y_U, S\}}\left\{\hat{\theta}^* - E_{\{y_U, S\}}(\hat{\theta}^*)\right\}^2. \tag{4.2.3}$$

Since the variance estimator (4.2.3) may be difficult to compute exactly, a simulation-based variance estimator may be used instead. More precisely, $C$ independent realizations $D_1, \ldots, D_C$ of the vector $D$ are generated, and we denote $\hat{\theta}_c^* = T(\hat{M}_c^*)$ with $\hat{M}_c^*$ the Bootstrap measure associated to the vector $D_c$. Then $V(\hat{\theta})$ is estimated by

$$\hat{V}_B(\hat{\theta}) \quad = \quad \frac{1}{C-1} \sum_{c=1}^{C} \left\{ \hat{\theta}_c^* - \frac{1}{C} \sum_{c'=1}^{C} \hat{\theta}_{c'}^* \right\}^2. \tag{4.2.4}$$

Two types of confidence intervals are usually computed. The percentile method makes use of the ordered bootstrap estimates $\hat{\theta}_{(c)}^*,\ c = 1, \ldots, C$ to form a $(1-2\alpha)\%$ confidence interval $\left[\hat{\theta}_{(L)}^*, \hat{\theta}_{(U)}^*\right]$ with $L = \alpha C$ and $U = (1-\alpha)C$. The bootstrap-t involves the estimation of the pivotal statistic $t = (\hat{\theta} - \theta)/\sqrt{v_{BWO}(\hat{\theta})}$ by its bootstrap counterpart $t^* = (\hat{\theta}^* - \hat{\theta})/\sqrt{v_{BWO}^*(\hat{\theta}^*)}$, where $v_{BWO}^*(\hat{\theta}^*)$

is obtained by applying the bootstrap procedure to the resample $S^*$. The bootstrap-t is highly computationally intensive since a double bootstrap is required, and is thus less attractive for a data user.

## 4.3 Variance estimation for imputed data

Suppose that the sample $S$ is prone to item non-response corrected by means of simple imputation, and that we consider estimating $\theta = T(M)$ by the imputed estimator $\hat{\theta}_I = T(\hat{M}_I)$ where $\hat{M}_I = \sum_{k\in S} d_k r_k \delta_{y_k} + \sum_{k\in S} d_k(1-r_k)\delta_{y_k^*}$ denotes the discrete measure taking mass $d_k$ on any point in the imputed sample, and 0 elsewhere. To produce a variance estimator for $\hat{\theta}_I$, it is convenient to consider the so-called reverse framework; see Fay (1996), Shao and Steel (1999) and Haziza (2009).

Under the usual framework, a sample $S$ is first selected by means of the sampling design $p(\cdot)$, in which the response mechanism $q(\cdot)$ then leads to the sets of item non-respondents for which imputed values are generated through the imputation mechanism $Imp$. That is, the natural order of the random mechanisms involved is conceptually "p,q,Imp". Under the reverse framework, it is assumed that the sampling design and the non-response mechanism are independent. As a result, we may think of the set of final values as obtained by: first, randomly generating in the whole population $U$ the vector $r$ of response indicators by means of the response mechanism $q(\cdot)$; then, generating by means of the sampling design $p(\cdot)$ the vector $I$ of sample membership indicators for all units, respondents or not; and finally, replacing missing values in the sample through the imputation mechanism $Imp$. That is, the order of the random mechanisms under the reverse framework is conceptually "q,p,Imp".

Under the reverse framework and using an analysis of variance decomposition, we obtain

$$\begin{aligned} V_{\{y_U\}}(\hat{\theta}_I) &= E_{\{y_U\}}V_{\{y_U,r\}}E_{\{y_U,r,I\}}(\hat{\theta}_I) + E_{\{y_U\}}E_{\{y_U,r\}}V_{\{y_U,r,I\}}(\hat{\theta}_I) + V_{\{y_U\}}E_{\{y_U,r\}}E_{\{y_U,r,I\}}(\hat{\theta}_I) \\ &= E_{\{y_U\}}V_{\{y_U,r\}}(\tilde{\theta}_I) + E_{\{y_U\}}E_{\{y_U,r\}}V_{\{y_U,r,I\}}(\hat{\theta}_I) + V_{\{y_U\}}E_{\{y_U,r\}}(\tilde{\theta}_I), \qquad (4.3.1) \end{aligned}$$

where

$$\tilde{\theta}_I = E_{\{y_U,r,I\}}(\hat{\theta}_I).$$

Under mild regularity conditions, the contribution of the third term $V_3 \equiv V_{\{y_U\}}E_{\{y_U,r\}}(\tilde{\theta}_I)$ to the total variance is of order $O(n/N)$, and is therefore negligible when the overall sampling fraction is small. In such case, we only need to account for $V_1 \equiv E_{\{y_U\}}V_{\{y_U,r\}}(\tilde{\theta}_I)$ and $V_2 \equiv E_{\{y_U\}}E_{\{y_U,r\}}V_{\{y_U,r,I\}}(\hat{\theta}_I)$. In case of deterministic imputation, we have $V_2 = 0$. This is also approximately true in case of balanced random imputation for the parameter $\theta$, since then

$$V_{\{y_U,r,I\}}(\hat{\theta}_I) \simeq 0.$$

In case of a small sampling fraction and deterministic/balanced random imputation, a variance estimator of $V\left(\hat{\theta}_I\right)$ is thus simply given by an estimator $\hat{V}_1$ of $V_1 \equiv E_{\{y_U\}}V_{\{y_U,r\}}(\tilde{\theta}_I)$. To estimate $V_1$ consistently, it suffices to estimate $V_{\{y_U,r\}}(\tilde{\theta}_I)$. In the case when the parameter of interest is a smooth function of $K$ totals

$$\theta = f(t_{y_1}, \ldots, t_{y_K}),$$

we have

$$\tilde{\theta}_I \simeq f(\tilde{t}_{y_1 I}, \ldots, \tilde{t}_{y_K I})$$

with $\tilde{t}_{y_j I} = E_{\{y_U,r,I\}}(\hat{t}_{y_j I})$, so that $\tilde{\theta}_I$ may also be (approximately) written as a smooth function of estimated totals. Estimating $V_1$ thus reduces to the classical problem of estimating the sampling variance of a smooth function of estimated totals. Any complete data variance estimation method can thus be used (Taylor linearization, jackknife or bootstrap), and the estimation of $V_1$ can be performed using a complete data variance estimation software, which is attractive for a data user since no specialized variance estimation software is required. Using this approach, a bootstrap procedure for variance estimation with imputed data is proposed in [10] and [16], and a Jackknife procedure is described in [15].

## 4.4 Future work

When the second-order inclusion probabilities are difficult to compute, or when some of them are zero, it is of interest to develop consistent variance estimators for totals, under reasonable model assumptions on the variable of interest. The consistency of such variance estimators for complex parameters (through linearization) is also of interest.

Guidelines for variance estimation with imputed data and a small sampling fraction are described above. In practice, it is of interest to develop variance estimation procedures (e.g., using Bootstrap) suitable in case of non-negligible sampling fractions (e.g., Mashregi et al., 2014).

# Chapter 5

# Coupling methods

So as to improve the accuracy of HT-estimators, many common sampling designs make use of some form of balanced sampling (see Section 2) where a dependence in the selection of units is introduced, and expected to provide a decrease of the variance. For example, when the units are selected independently with equal probabilities $\pi_k = n/N$ (no dependence), which means Bernoulli sampling, the variance of the HT-estimator is

$$V_{\{y_U\}}(\hat{t}_{y\pi}) \quad = \quad N^2 \frac{1-f}{n} \frac{1}{N} \sum_{k \in U} y_k^2.$$

If this sampling design is conditioned to obtain a sample of fixed size equal to $n$ (which corresponds to one balancing variable $x_k = 1$), we obtain SI sampling and the variance of the HT-estimator is

$$V_{\{y_U\}}(\hat{t}_{y\pi}) \quad = \quad N^2 \frac{1-f}{n} \frac{1}{N-1} \sum_{k \in U} (y_k - \mu_y)^2.$$

The variance is thus reduced to the degree of which the variable $y$ may be explained by the balancing variable $x_k = 1$.

The dependence in the selection of units may be complex, which makes limiting results quite difficult to prove. In this chapter, we consider the use of coupling methods (see Thorisson, 2000) to link a sampling design under study to a close, simpler sampling design where useful limiting properties may be more easily derived. The method basically consists in generating a random vector $(X_t, Z_t)^\top$ so that (i) $X_t$ has an appropriate marginal law (e.g., that of the HT estimator for the sampling design under study); (ii) the marginal law of $Z_t$ is simpler to study; and (iii) $X_t$ and $Z_t$ are close, e.g., so that $E(X_t - Z_t)^2$ is smaller than the rate of convergence of $X_t$. In this case, $X_t$ and $Z_t$ share the same limiting variance and the same limiting distribution, as stated in Lemma 5.0.1.

**Lemma 5.0.1.** *Let $X_t$ and $Z_t$ denote two random variables such that $E(X_t) = E(Z_t)$. Assume*

*that* $E(X_t - Z_t)^2 = o(a_t)$ *and that* $V(X_t) = O(a_t)$*, where* $a_t \underset{t\to\infty}{\longrightarrow} 0$*. Then*

$$\frac{V(Z_t)}{V(X_t)} \underset{t\to\infty}{\longrightarrow} 1. \tag{5.0.1}$$

*Also, if* $\sqrt{a_t}\{X_t - E(X_t)\} \underset{\mathcal{L}}{\longrightarrow} X_0$*, then* $\sqrt{a_t}\{Z_t - E(Z_t)\} \underset{\mathcal{L}}{\longrightarrow} X_0$*.*

In a pioneering work, Hajek (1961) introduced a coupling procedure between Bernoulli sampling and SI sampling to obtain a central-limit theorem for the latter. In Hajek (1964), a similar approach was used to link Poisson sampling and rejective sampling, and to derive a central-limit theorem and a variance approximation for rejective sampling. In Section 5.1, we extend the coupling algorithm by Hajek (1961) so as to derive asymptotic normality results for the HT-estimator under without-replacement multistage designs. In Section 5.2, we consider the Bootstrap for multistage sampling. We introduce a new coupling algorithm between SI sampling of PSUs and SIR sampling of PSUs, and we prove that the Hansen-Hurwitz estimator and the HT estimator are close when the first-stage sampling fraction becomes negligible. This coupling algorithm is used to prove a long-standing issue; namely, that the so-called with-replacement Bootstrap of PSUs (see Rao and Wu, 1988) is consistent in case of SI sampling of PSUs with a small first-stage sampling fraction, and yields consistent variance estimators for smooth functions of means.

## 5.1 Asymptotic normality for multistage sampling

In Sections 5.1 and 5.2, we consider multistage sampling and we suppose that the units are grouped inside $N_I$ non-overlapping subpopulations $u_1, \ldots, u_{N_I}$ called primary sampling units (PSUs). We are interested in estimating the population total

$$t_y = \sum_{k\in U} y_k = \sum_{u_i \in U_I} Y_i,$$

where $Y_i = \sum_{k\in u_i} y_k$ is the sub-total of the variable $y$ on the PSU $u_i$. In Sections 5.1 and 5.2, we denote by $\hat{Y}_i$ an unbiased estimator of $Y_i$, and by $V_i \equiv V_{\{y_U\}}(\hat{Y}_i)$ its variance. Also, we denote by $\hat{V}_i$ an unbiased estimator of $V_i$.

In the population $U_I = \{u_1, \ldots, u_{N_I}\}$ of PSUs, a first-stage sample $S_I$ is selected according to some sampling design $p_I(\cdot)$. For clarity of exposition, we consider non-stratified sampling designs for $p_I(\cdot)$, but the results may be easily extended to the case of stratified first-stage sampling designs with a finite number of strata. If the PSU $u_i$ is selected in $S_I$, a second-stage sample $S_i$ is selected in $u_i$ by means of some sampling design $p_i(\cdot|S_I)$. We assume invariance of the second-stage designs: that is, the second stage of sampling is independent of $S_I$ and we may simply write $p_i(\cdot|S_I) = p_i(\cdot)$. Also, we assume that the second-stage designs are independent from one PSU to another, conditionally

on $S_I$. This implies that

$$Pr\left(\bigcup_{u_i\in S_I}\{S_i=s_i\}\middle|S_I\right) = \prod_{u_i\in S_I} p_i(s_i|S_I)$$
$$= \prod_{u_i\in S_I} p_i(s_i) \quad (5.1.1)$$

for any set of samples $s_i \subset u_i,\ i=1,\ldots,N_I$, where the second line in (5.1.1) follows from the invariance assumption; see Särndal et al (1992, chapter 4) for further details. The second-stage sampling designs $p_i(\cdot)$ are left arbitrary. For example, they may involve censuses inside some PSUs (which means cluster sampling), or additional stages of sampling.

We will make use of the following assumptions:

H1: $N_I \underset{t\to\infty}{\longrightarrow} \infty$ and $n_I \underset{t\to\infty}{\longrightarrow} \infty$. Also, $f_I = n_I/N_I \underset{t\to\infty}{\longrightarrow} f \in [0,1[$.

H2: There exists $\delta > 0$ and constants $C_1, C_2$ such that $C_1 < N_I^{-1}\sum_{u_i\in U_I} E_{\{y_U\}}|\hat{Y}_i|^{2+\delta} < C_2$.

It is assumed in (H1) that a large number $n_I$ of PSUs is selected. The assumption (H2) implies that the sequence of $\{Y_i\}_{u_i\in U_I}$ has bounded $2+\delta$ moments and that the sequence of $\{V_{\{y_U\}}(\hat{Y}_i)\}_{u_i\in U_I}$ has a bounded first moment; this assumption requires in particular that the total number of SSUs within PSUs remains bounded.

### 5.1.1 Bernoulli sampling of PSUs

We first consider the case when a first-stage sample $S_I^B$ is selected in $U_I$ by means of Bernoulli sampling (BE) with expected size $n_I$ (Särndal et al., 1992, p. 62; Fuller, 2009, p. 16), which we note as $S_I^B \sim BE(U_I; n_I)$. That is, the PSUs are independently selected in $S_I^B$ with inclusion probabilities $f_I = n_I/N_I$, and the size $n_I^B$ of $S_I^B$ is random. The Horvitz-Thompson estimator

$$\hat{Y}_B = \frac{N_I}{n_I}\sum_{u_i\in U_I} I_i^B \hat{Y}_i = \frac{N_I}{n_I}\sum_{u_i\in S_I^B}\hat{Y}_i \quad (5.1.2)$$

is unbiased for $Y$, with $I_i^B$ the sample membership indicator for the PSU $u_i$ in the sample $S_I^B$. The variance of $\hat{Y}_B$ is

$$V_{\{y_U\}}(\hat{Y}_B) = \frac{N_I^2}{n_I}\left\{(1-f_I)\frac{1}{N_I}\sum_{u_i\in U_I} Y_i^2 + \frac{1}{N_I}\sum_{u_i\in U_I} V_i\right\} \quad (5.1.3)$$

where $V_i = V_{\{y_U\}}(\hat{Y}_i)$. If (H1) and (H2) hold, we have

$$\frac{\hat{Y}_B - Y}{\sqrt{V_{\{y_U\}}(\hat{Y}_B)}} \underset{\mathcal{L}}{\longrightarrow} \mathcal{N}(0,1). \quad (5.1.4)$$

A coupling procedure between BE sampling and SI sampling will be used in Section 5.1.2 to extend (5.1.4) to the context of multistage sampling with SI sampling at the first stage. Anyway, as noted by Hajek (1964, p. 1499), the Narain-Horvitz-Thompson estimator under Bernoulli sampling of PSUs presents a variability due to the random sample size which is unneeded for comparison with simple random sampling. Therefore, we need to consider a modified version of $\hat{Y}_B$, defined as

$$\tilde{Y}_B = \frac{N_I}{n_I} \sum_{u_i \in U_I} I_i^B (\hat{Y}_i - \mu_Y) = \frac{N_I}{n_I} \sum_{u_i \in S_I^B} (\hat{Y}_i - \mu_Y) \tag{5.1.5}$$

where $\mu_Y = N_I^{-1} \sum_{u_i \in U_I} Y_i$; $\tilde{Y}_B$ is not an estimator per se, since its definition involves the unknown quantity $\mu_Y$. We have

$$V_{\{y_U\}}(\tilde{Y}_B) = \frac{N_I^2}{n_I} \left\{ (1 - f_I) \frac{1}{N_I} \sum_{u_i \in U_I} (Y_i - \mu_Y)^2 + \frac{1}{N_I} \sum_{u_i \in U_I} V_i \right\}, \tag{5.1.6}$$

and if (H1) and (H2) hold, we have

$$\frac{\tilde{Y}_B}{\sqrt{V_{\{y_U\}}(\tilde{Y}_B)}} \underset{\mathcal{L}}{\longrightarrow} \mathcal{N}(0,1). \tag{5.1.7}$$

### 5.1.2 Without replacement simple random sampling of PSUs

We now consider the case when a first-stage sample $S_I$ is selected in $U_I$ by means of simple random sampling without replacement (SI) of size $n_I$, which we note as $S_I \sim SI(U_I; n_I)$. The Narain-Horvitz-Thompson estimator is

$$\hat{Y} = \frac{N_I}{n_I} \sum_{u_i \in U_I} I_i \hat{Y}_i = \frac{N_I}{n_I} \sum_{u_i \in S_I} \hat{Y}_i, \tag{5.1.8}$$

with $I_i$ the sample membership indicator for the PSU $u_i$ in the sample $S_I$. We may alternatively rewrite the Narain-Horvitz-Thompson estimator as

$$\hat{Y} = N_I \bar{Z} \quad \text{with} \quad \bar{Z} = \frac{1}{n_I} \sum_{j=1}^{n_I} Z_j, \tag{5.1.9}$$

where the sample $S_I$ of PSUs is obtained by drawing $n_I$ times without replacement one PSU in $U_I$, and where $Z_j$ stands for the estimator of the total for the PSU selected at the $j$-th draw. The variance of $\hat{Y}$ is

$$V_{\{y_U\}}(\hat{Y}) = \frac{N_I^2}{n_I} \left\{ (1 - f_I) S_{Y,U_I}^2 + \frac{1}{N_I} \sum_{u_i \in U_I} V_i \right\}. \tag{5.1.10}$$

with $S_{Y,U_I}^2 = (N_I - 1)^{-1} \sum_{u_i \in U_I} (Y_i - \mu_Y)^2$ the population dispersion of the sub-totals $Y_i$.

Like we did for Bernoulli sampling, we also define

$$\tilde{Y} \quad = \quad \frac{N_I}{n_I} \sum_{u_i \in S_I} (\hat{Y}_i - \mu_Y).$$

Hajek (1961) proposed a coupling procedure to draw simultaneously a BE sample and a SI sample. This procedure is adapted in Algorithm 4 to the context of multistage sampling, and Proposition 1 below generalizes the Lemma 2.1 in Hajek (1961).

---

**Algorithm 4** A coupling procedure for Bernoulli sampling of PSUs and simple random sampling without replacement of PSUs

---

1. Draw the sample $S_I^B \sim BE(U_I; n_I)$. Denote by $n_I^B$ the (random) size of $S_I^B$.

2. Draw the sample $S_I$ as follows:

   - if $n_I^B = n_I$, take $S_I = S_I^B$;
   - if $n_I^B < n_I$, draw $S_I^+ \sim SI(U_I \setminus S_I^B; n_I - n_I^B)$ and take $S_I = S_I^B \cup S_I^+$;
   - if $n_I^B > n_I$, draw $S_I^+ \sim SI(S_I^B; n_I^B - n_I)$ and take $S_I = S_I^B \setminus S_I^+$.

3. For any PSU $u_i$:

   - if $u_i \in S_I^B \cap S_I$, select the same second-stage sample $S_i$ for both $\hat{Y}_B$ and $\hat{Y}$;
   - if $u_i \in S_I^B \setminus S_I$, select a second-stage sample $S_i$ for $\hat{Y}_B$;
   - if $u_i \in S_I \setminus S_I^B$, select a second-stage sample $S_i$ for $\hat{Y}$.

---

*Proposition* 1. Assume that the samples $S_I^B$ and $S_I$ are selected according to Algorithm 1. Then

$$\frac{E_{\{y_U\}}(\tilde{Y} - \tilde{Y}_B)^2}{V_{\{y_U\}}(\tilde{Y}_B)} \quad \leq \quad \sqrt{\frac{1}{n_I} + \frac{1}{N_I - n_I}}. \tag{5.1.11}$$

The result in Proposition 1 can be easily generalized to the multivariate case: if $y_k = (y_{1k}, \ldots, y_{qk})^\top$ denotes the value taken for unit $k$ by some $q$-vector of interest $y$, we have

$$V_{\{y_U\}}(\tilde{Y} - \tilde{Y}_B) \quad \leq \quad \sqrt{\frac{1}{n_I} + \frac{1}{N_I - n_I}} V_{\{y_U\}}(\tilde{Y}_B),$$

where for symmetric matrices $A$ and $B$ of size $q$, $A \leq B$ means that $B - A$ is nonnegative definite.

Under (H1) and (H2), we have $V_{\{y_U\}}(\tilde{Y}_B) = O(N_I^2 \; n_I^{-1})$ and from Proposition 1 $E_{\{y_U\}}(\tilde{Y} - \tilde{Y}_B)^2 = o(N_I^2 \; n_I^{-1})$. It then follows from from Lemma 5.0.1 that

$$\frac{\hat{Y} - Y}{\sqrt{V_{\{y_U\}}(\hat{Y})}} \quad \xrightarrow[\mathcal{L}]{} \quad \mathcal{N}(0, 1) \tag{5.1.12}$$

## 5.2 With-replacement Bootstrap for multistage sampling

Bootstrap for multistage sampling under without-replacement sampling of PSUs has been considered for example in Rao and Wu (1988), Rao, Wu and Yue (1992), Nigam and Rao (1996), Funaoka et al. (2006), Preston (2009) and Lin et al. (2013), among others. Testing the validity of a bootstrap procedure has primarily consisted in showing that it led to the correct variance estimator in the linear case, and then in evaluating empirically the behavior of the method for complex parameters through simulations. In this section, we consider the so-called with-replacement Bootstrap of PSUs (see Rao and Wu, 1988). We prove that this Bootstrap method is suitable for multistage sampling with SI sampling of PSUs and a small first-stage sampling fraction. More precisely, we prove that the Bootstrap pivotal statistic (see equation 5.2.9) is asymptotically normally distributed, and that the Bootstrap yields consistent variance estimators for smooth functions of means.

### 5.2.1 With replacement sampling of PSUs

We consider the case when a first-stage sample $S_I^{WR}$ is selected in $U_I$ according to simple random sample with replacement (SIR) of size $n_I$ inside $U_I$, which we note as $S_I^{WR} \sim SIR(U_I; n_I)$. Denote by $W_i$ the number of selections of the PSU $u_i$ in $S_I^{WR}$, and by $S_I^d$ of size $n_I^d$ the set of distinct PSUs associated to $S_I^{WR}$. Each time $j = 1, \ldots, W_i$ that unit $u_i$ is drawn in $S_I^{WR}$, a second-stage sample $S_{i[j]}$ is selected in $u_i$. The total $Y$ is unbiasedly estimated by the Hansen-Hurwitz (1942) estimator

$$\hat{Y}_{WR} = \sum_{u_i \in S_I^d} \frac{1}{E(W_i)} \sum_{j=1}^{W_i} \hat{Y}_{i[j]} = \frac{N_I}{n_I} \sum_{u_i \in S_I^d} \sum_{j=1}^{W_i} \hat{Y}_{i[j]} \tag{5.2.1}$$

where $\hat{Y}_{i[j]}$ stands for an unbiased estimator of $Y_i$ computed on $S_{i[j]}$. We may alternatively rewrite the Hansen-Hurwitz estimator as

$$\hat{Y}_{WR} = N_I \bar{X} \quad \text{with} \quad \bar{X} = \frac{1}{n_I} \sum_{j=1}^{n_I} X_j, \tag{5.2.2}$$

where the sample $S_I^{WR}$ of PSUs is obtained by drawing $n_I$ times with replacement one PSU in $U_I$ and where $X_j$ stands for the estimator of the total for the PSU selected at the $j$-th draw. Then if (H1) and (H2) hold, we have

$$\frac{\hat{Y}_{WR} - Y}{\sqrt{V_{\{y_U\}}(\hat{Y}_{WR})}} \xrightarrow[\mathcal{L}]{} \mathcal{N}(0,1). \tag{5.2.3}$$

### 5.2.2 A coupling procedure between SIR/SI sampling of PSUs

The procedure is described in Algorithm 5. Conditionally on $n_I^d$, the sample $S_I^d$ obtained in Step 1 is by symmetry a SI sample of size $n_I^d$ from $U_I$, which implies that $S_I^d \cup S_I^c$ is a SI sample of size

$n_I$ from $U_I$. Consequently, this procedure leads to a sample $S_I$ drawn by means of SI sampling of PSUs.

**Algorithm 5** A coupling procedure for simple random sampling with-replacement of PSUs and simple random sampling without replacement of PSUs for multistage sampling

1. Draw the sample $S_I^{WR} \sim SIR(U_I; n_I)$. Denote by $S_I^d$ of (random) size $n_I^d$ the set of distinct PSUs in $S_I^{WR}$.
2. Draw a complementary sample $S_I^c \sim SI(U_I \setminus S_I^d; n_I - n_I^d)$ and take $S_I = S_I^d \cup S_I^c$.
3. For any $u_i \in S_I^d$:
   - Each time $j = 1, \ldots, W_i$ that unit $u_i$ is drawn in $S_I^{WR}$, select a second-stage sample $S_{i[j]}$ with associated estimator $\hat{Y}_{i[j]}$ for $\hat{Y}_{WR}$.
   - Take $S_i = S_{i[1]}$ and $\hat{Y}_i = \hat{Y}_{i[1]}$ for $\hat{Y}$.
4. For any $u_i \in S_I^c$, select a second-stage sample $S_i$ with associated estimator $\hat{Y}_i$ for $\hat{Y}$.

*Proposition* 2. Assume that the samples $S_I^{WR}$ and $S_I$ are selected according to Algorithm 5. Then

$$\frac{E_{\{y_U\}}(\hat{Y}_{WR} - \hat{Y})^2}{V_{\{y_U\}}(\hat{Y}_{WR})} \quad \leq \quad \frac{n_I - 1}{N_I - 1}. \tag{5.2.4}$$

The right bound in (5.2.4) is mainly of interest when $f_I \underset{t\to\infty}{\longrightarrow} 0$. In this case, from the trivial inequality $\frac{n_I - 1}{N_I - 1} \leq \frac{n_I}{N_I}$, Algorithm 3 may be used to select the samples $S_I^{WR}$ and $S_I$ so that the difference between $\hat{Y}_{WR}$ and $\hat{Y}$ is asymptotically negligible. A similar result holds for the dispersions between the estimated totals inside PSUs, as stated in Proposition 3 below.

*Proposition* 3. Assume that the samples $S_I^{WR}$ and $S_I$ are selected according to Algorithm 5. Assume that (H1) and (H2) hold, and that $f_I \underset{t\to\infty}{\longrightarrow} 0$. Then

$$E_{\{y_U\}}(\bar{Z} - \bar{X})^2 = o(n_I^{-1}), \tag{5.2.5}$$

$$E_{\{y_U\}}|s_Z^2 - s_X^2| \underset{t\to\infty}{\longrightarrow} 0. \tag{5.2.6}$$

where $\bar{X}$ and $\bar{Z}$ are defined in equations (5.2.2) and (5.1.9), and with

$$s_X^2 = \frac{1}{n_I - 1} \sum_{u_i \in S_I} (X_i - \bar{X})^2 \quad \text{and} \quad s_Z^2 = \frac{1}{n_I - 1} \sum_{u_i \in S_I} (Z_i - \bar{Z})^2.$$

### 5.2.3 With replacement Bootstrap of PSUs

We consider the with-replacement Bootstrap of PSUs described for example in Rao and Wu (1988). Using the notation introduced in equation (5.1.9), let $(Z_1, \ldots, Z_{n_I})^\top$ denote the sample of estimators

under SI sampling of PSUs. Also, let $(Z_1^*, \ldots, Z_m^*)^\top$ be obtained by sampling $m$ times independently in $(Z_1, \ldots, Z_{n_I})^\top$. Similarly, using the notation introduced in equation (5.2.2), let $(X_1, \ldots, X_{n_I})^\top$ denote the sample of estimators under SIR sampling of PSUs. Also, let $(X_1^*, \ldots, X_m^*)^\top$ be obtained by sampling $m$ times independently in $(X_1, \ldots, X_{n_I})^\top$.
We first consider the Bootstrap consistency. We note

$$\bar{Z}_m^* = \frac{1}{m}\sum_{j=1}^{m} Z_j^*, \qquad s_Z^{*2} = \frac{1}{m-1}\sum_{j=1}^{m}\left(Z_j^* - \bar{Z}_m^*\right)^2, \tag{5.2.7}$$

and

$$\bar{X}_m^* = \frac{1}{m}\sum_{j=1}^{m} X_j^*, \qquad s_X^{*2} = \frac{1}{m-1}\sum_{j=1}^{m}\left(X_j^* - \bar{X}_m^*\right)^2. \tag{5.2.8}$$

The proof proceeds by showing that, using Algorithm 5, the samples $S_I$ and $S_I^{WR}$ can be drawn so that the pivotal statistics

$$\frac{\sqrt{m}(\bar{Z}_m^* - \bar{Z})}{s_Z^*} \quad \text{and} \quad \frac{\sqrt{m}(\bar{X}_m^* - \bar{X})}{s_X^*} \tag{5.2.9}$$

are close. More precisely, we make use of the Mallows metric (Mallows, 1972; Bickel and Freedman, 1981), also known as the Wasserstein metric. Let $1 \leq q < \infty$, and let $\alpha$ and $\beta$ denote two distributions on $\mathbb{R}^s$ with finite moments of order $q$. Then $d_q(\alpha, \beta) = \inf\{E\|X - Z\|^q\}^{1/q}$, where the infimum is taken over all couples $(X, Z)$ with marginal distributions $\alpha$ and $\beta$. For two random vectors $X$ and $Z$, we note $d_q(\alpha, \beta)$ for the $d_q$-distance between the distributions of $X$ and $Z$. In what follows, we consider $q = 1$ or $q = 2$.

*Proposition* 4. Assume that (H1) and (H2) hold, that $f_I \underset{t\to\infty}{\longrightarrow} 0$ and that $m \underset{t\to\infty}{\longrightarrow} \infty$. Then :

$$d_2\left[\sqrt{m}(\bar{Z}_m^* - \bar{Z}), \sqrt{m}(\bar{X}_m^* - \bar{X})\right] \underset{t\to\infty}{\longrightarrow} 0, \tag{5.2.10}$$

$$d_1\left[s_Z^{*2}, s_X^{*2}\right] \underset{t\to\infty}{\longrightarrow} 0. \tag{5.2.11}$$

From Proposition 4, the pivotal statistics in (5.2.9) share the same limiting distribution. Theorem 1 below follows from Theorem 2.1 of Bickel and Freedman (1981).

*Theorem* 1. Assume that $m \underset{t\to\infty}{\longrightarrow} \infty$. Then the Bootstrap pivotal quantity

$$\frac{\sqrt{m}(\bar{Z}_m^* - \bar{Z})}{s_Z^*}$$

converges in distribution to the standard normal distribution.

### 5.2.4 Bootstrap variance estimation for functions of means

We now consider the case when $y_k = (y_{1k}, \ldots, y_{qk})^\top$ is multivariate, and denotes the value taken for unit $k$ by some $q$-vector of interest $y$. We are interested in a parameter $\theta = f(\mu_Y)$ for some function $f : \mathbb{R}^q \longrightarrow \mathbb{R}$. We consider the additional regularity assumption:

H4: $f(\cdot)$ is a differentiable function on $\mathbb{R}^q$ with bounded partial derivatives, and $f'(\mu_Y) \neq 0$.

Under SI sampling of PSUs, the plug-in estimator of $\theta$ is denoted by $\hat{\theta} = f(\bar{Z})$. Under SIR sampling of PSUs, the plug-in estimator of $\theta$ is denoted by $\hat{\theta}_{WR} = f(\bar{X})$.

*Proposition* 5. Assume that the samples $S_I^{WR}$ and $S_I$ are selected according to Algorithm 3. Assume that assumptions (H1), (H2) and (H4) hold. Assume that $f_I \underset{t\to\infty}{\longrightarrow} 0$. Then :

$$E_{\{y_U\}}(\|\bar{Z} - \bar{X}\|^2) = o(n_I^{-1}), \tag{5.2.12}$$

$$E_{\{y_U\}}(\hat{\theta} - \hat{\theta}_{WR})^2 = o(n_I^{-1}). \tag{5.2.13}$$

with $\|\cdot\|$ the Euclidean norm.

In proving Proposition 5, equation (5.2.5) easily generalizes as (5.2.12). Also, equation (5.2.13) follows directly from (5.2.12) and the regularity assumptions on $f(\cdot)$. We can prove a similar result for the Bootstrap estimators, which we note as $\hat{\theta}^* = f(\bar{Z}_m^*)$ and $\hat{\theta}_{WR}^* = f(\bar{X}_m^*)$, where $\bar{Z}_m^*$ and $\bar{X}_m^*$ are defined in (5.2.7) and (5.2.8).

*Proposition* 6. Assume that the samples $S_I^{WR}$ and $S_I$ are selected according to Algorithm 3. Assume that assumptions (H1), (H2) and (H4) hold. Assume that $f_I \underset{t\to\infty}{\longrightarrow} 0$ and $m \underset{t\to\infty}{\longrightarrow} \infty$. Then :

$$E_{\{y_U\}}(\|\bar{Z}^* - \bar{X}^*\|^2) = o(m^{-1}) + o(n_I^{-1}), \tag{5.2.14}$$

$$E_{\{y_U\}}(\hat{\theta}^* - \hat{\theta}_{WR}^*)^2 = o(m^{-1}) + o(n_I^{-1}). \tag{5.2.15}$$

Since $f'(\mu_Y) \neq 0$, we have $V(\hat{\theta}_{WR}) = O(n_I^{-1})$ under (H2). Making use of Lemma 5.0.1, Proposition 5 implies that $V_{\{y_U\}}(\hat{\theta}_{WR})$ and $V_{\{y_U\}}(\hat{\theta})$ are asymptotically equivalent, i.e.

$$\frac{V_{\{y_U\}}(\hat{\theta})}{V_{\{y_U\}}(\hat{\theta}_{WR})} \underset{t\to\infty}{\longrightarrow} 1. \tag{5.2.16}$$

Similarly, Proposition 6 implies that

$$\frac{V_{\{y_U,Z\}}(\hat{\theta}^*)}{V_{\{y_U,X\}}(\hat{\theta}_{WR}^*)} \underset{Pr}{\longrightarrow} 1. \tag{5.2.17}$$

If the with-replacement Bootstrap provides consistent variance estimation for $\hat{\theta}_{WR}$ in case of SIR sampling of PSUs, we have $\dfrac{V_{\{y_U,X\}}(\hat{\theta}_{WR}^*)}{V_{\{y_U\}}(\hat{\theta}_{WR})} \underset{Pr}{\longrightarrow} 1$. Then, from (5.2.16) and (**??**) follows that the with-replacement Bootstrap also provides consistent variance estimation for $\hat{\theta}$ in case of SI sampling of

PSUs. The regularity assumption (H4) is somewhat strong, and may be weakened to differentiability of $f(\cdot)$ on a compact set, under additional assumptions on the vector of interest $y$ and on the second-stage sampling weights.

## 5.3 Future work

The coupling method is a promising tool to obtain asymptotic properties for estimators and sampling designs. We intend to develop specific coupling procedures for unequal probability sampling algorithms, to obtain asymptotic normality results under some mild assumptions. Also, the extension of Algorithm 5 to sampling with unequal probabilities is currently under investigation, in order to prove consistency of the with-replacement Bootstrap of PSUs in this context. Looking for Bootstrap techniques which are consistent under a non-negligible sampling fraction, or proving that some existing Bootstrap methods succeed in so doing, is also a topic of practical and theoretical interest.

# List of papers

## Published or accepted for publication


[1] G. Chauvet, Y. Tillé (2006). *A fast algorithm of Balanced Sampling.* Computational Statistics, 21, 53 - 61.

[2] G. Chauvet, Y. Tillé (2007). *Application of Fast SAS Macros for Balancing Samples to the Selection of Addresses.* Case Studies in Business, Industry, and Government Statistics, 2.

[3] G. Chauvet (2009). *Stratified Balanced Sampling.* Survey Methodology, 35, 115 - 119.

[4] D. Haziza, G. Chauvet, J.C. Deville (2010). *A note on sampling and estimation in the presence of cut-off sampling.* Australian and New Zealand Journal of Statistics, 52, 303 - 319.

[5] F.J. Breidt, G. Chauvet (2011). *Improved variance estimation for balanced samples drawn via the Cube method.* Journal of Statistical Planning and Inference, 141, 479 - 487.

[6] G. Chauvet, D. Bonnery, J.C. Deville (2011). *Optimal inclusion probabilities for balanced sampling.* Journal of Statistical Planning and Inference, 141, 984 - 994.

[7] G. Chauvet, J.C. Deville, D. Haziza (2011). *On balanced random imputation in surveys.* Biometrika, 98, 459-471.

[8] G. Chauvet (2011). *On variance estimation for the French Master Sample.* Journal of Official Statistics, 27, n° 4, 651–668.

[9] M. Chandesris, G. Chauvet, J.C. Deville (2011). *Allocation optimale pour un plan à plusieurs degrés. Application à l'estimation de la fraude tarifaire grandes lignes à la SNCF.* Journal de la SFdS, 152, n°4, 47-59.

[10] G. Chauvet, D. Haziza (2012). *Fully efficient estimation of coefficients of correlation in the presence of imputed data.* Canadian Journal of Statistics, 40, n° 1, 124-149.

[11] G. Chauvet (2012). *On a characterization of ordered pivotal sampling.* Bernoulli, 18, n°4, 1320-1340.

[12] F.J. Breidt, G. Chauvet (2012). *Penalized Balanced Sampling.* Biometrika, 99, n° 4, 945-958.

[13] G. Chauvet (201X). *Variance Estimation for the 2006 French Housing Survey.* To appear in Mathematical Population Studies (invited submission for a special issue).

[14] G. Chauvet, G. Tandeau de Marsac (201X). *Méthodes d'estimation sur bases de sondage multiples dans le cadre de plans de sondage à deux degrés.* To appear in Survey Methodology.

[15] D. Haziza, C-O. Nambeu, G. Chauvet (201X). *Doubly robust imputation procedures for populations containing a large amount of zeroes in surveys.* Canadian Journal of Statistics, 42, n° 4, 650-669.

## Submitted or work in progress

[16] H. Chaput, G. Chauvet, D. Haziza, L. Salembier, J. Solard (201X). *Joint imputation procedures for categorical variables with application to the French Wealth Survey.* Second revision for Journal of the Royal Statistical Society C.

[17] G. Chauvet, C. Goga (201X). *Gini coefficient and Gini coefficient change: linearization versus Bootstrap to estimate the variance.* In revision for Survey Methodology.

[18] G. Chauvet, D. Haziza et E. Lesage (201X). *Examining some aspects of balanced sampling in surveys.* In revision for Statistica Sinica.

[19] H. Boistard, G. Chauvet, D. Haziza (201X). *Consistency of the estimated distribution function with missing data under a non-response model.* In revision for Scandinavian Journal of Statistics.

[20] G. Chauvet (201X). *Coupling Methods for multistage sampling.* Submitted.

[21] G. Chauvet, A. Ruiz-Gazen (201X). *A comparison of pivotal sampling and unequal probability sampling with replacement.* Submitted.

[22] G. Chauvet, J.C. Deville, D. Haziza (201X). *Adapting the Cube algorithm for balanced random imputation in surveys.* Submitted.

[23] G. Chauvet, H. Juillard, A. Ruiz-Gazen (201X). *Variance estimation for product sampling: an application to the ELFE survey.*

[24] G. Chauvet, C. Goga (201X). *Sélection de variables de calage par une méthode de Bootstrap.*

[25] G. Chauvet, J.C. Deville (201X). *Asymptotic Results for Deville's Systematic Sampling.*

[26] G. Chauvet, Do Paco, W., Haziza, D. (201X). *Exact balanced imputation for sample survey data.*

# References


- Antal, E., and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. Journal of the American Statistical Association, 106(494), 534-543.
- Ardilly, P. (1991). Echantillonnage représentatif optimum à probabilités inégales. Annales d'Economie et de Statistique, 91-113.
- Ardilly, P. (2006). Les techniques de sondage. Editions Technip.
- Beaumont, J. F., and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. International Statistical Review, 80(1), 127-148.
- Bertail, P. and Combris, P. (1997). Bootstrap généralisé d'un sondage. Annales d'Economie et de Statistique, 46, 49-83.
- Bickel, P. J., and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. The Annals of Statistics, 1196-1217.
- Bickel, P. J., and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. The annals of statistics, 470-482.
- Boistard, H., Lopuhaä, H. P., and Ruiz-Gazen, A. (2012). Approximation of rejective sampling inclusion probabilities and application to high order correlations. Electronic Journal of Statistics, 6, 1967-1983.
- Booth, J. G., and Butler, R. W., and Hall, P. (1994). Bootstrap methods for finite populations. Journal of the American Statistical Association, 89(428), 1282-1289.
- Branden, P., and Jonasson, J. (2012). Negative dependence in sampling. Scandinavian Journal of Statistics, 39, 830-838.
- Breidt, F. J., and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. Annals of Statistics, 1026-1053.
- Chao, M. T., and Lo, S. H. (1985). A bootstrap method for finite population. Sankhya: The Indian Journal of Statistics, Series A, 399-405.

- Chen, J., and Rao, J.N.K. (2007). Asymptotic normality under two-phase sampling designs. Statistica Sinica, 17, 1047-1064.
- Chen, J., and Rao, J. N. K., and Sitter, R. R. (2000). Efficient random imputation for missing data in complex surveys. Statistica Sinica, 10(4), 1153-1170.
- Christine, M., and Faivre, S. (2009). Le projet OCTOPUSSE de nouvel Echantillon-Maître de l'INSEE. Journées de Méthodologie Statistique, 2009.
- Davison, A. C., and Hinkley, D. V. (1997). Bootstrap methods and their application. Cambridge University Press. Cambridge, UK, 193.
- Davison, A. C., and Sardy, S. (2007). Resampling variance estimation in surveys with missing data. Journal of Official Statistics, 23(3), 371.
- Deville, J. C. (1998). Une nouvelle méthode de tirage à probabilités inégales. Technical Report 9804, Ensai, France.
- Deville, J. C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. Survey methodology, 25(2), 193-204.
- Deville (2006). Random imputation using balanced sampling. Presentation to the Joint Statistical Meeting of the American Statistical Association, Seattle, USA.
- Deville, J. C., Grosbras, J. M., and Roth, N. (1988, January). Efficient sampling algorithms and balanced samples. In Compstat (pp. 255-266). Physica-Verlag HD.
- Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418), 376-382.
- Deville, J. C., and Särndal, C. E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. Journal of Official Statistics, 10, 381-381.
- Deville, J. C., and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. Biometrika, 85(1), 89-101.
- Deville, J. C., and Tillé, Y. (2004). Efficient balanced sampling: the cube method. Biometrika, 91(4), 893-912.
- Deville, J. C., and Tillé, Y. (2005). Variance approximation under balanced sampling. Journal of Statistical Planning and Inference, 128(2), 569-591.
- Efron, B. (1982). The jackknife, the bootstrap and other resampling plans (Vol. 38). Philadelphia: Society for industrial and applied mathematics.

- Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. Biometrika, 93(2), 269-278.
- Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. Journal of the American Statistical Association, 91(434), 490-498.
- Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. Biometrika, 96(4), 933-944.
- Fuller, W. A. (2011). Sampling statistics (Vol. 560). John Wiley and Sons.
- Fuller, W. A., and Kim, J. K. (2005). Hot deck imputation for the response model. Survey Methodology, 31(2), 139.
- Funaoka, F., and Saigo, H., and Sitter, R. R., and Toida, T. (2006). Bernoulli bootstrap for stratified multistage sampling. Survey Methodology, 32(2), 151.
- Gabler, S. (1981), A comparison of Sampford's sampling procedure versus unequal probability sampling with replacement, Biometrika, 68, 725–727.
- Gabler, S. (1984), On unequal probability sampling: sufficient conditions for the superiority of sampling without replacement, Biometrika, 71, 171–175.
- Goga, C., and Deville, J. C., and Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample survey data. Biometrika, 96(3), 691-709.
- Goga, C., and Ruiz-Gazen, A. (2014). Efficient estimation of non-linear finite population parameters by using non-parametrics. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1), 113-140.
- Gordon, L. (1983). Successive sampling in large finite populations. The Annals of Statistics, 11, 702-706.
- Gross, S. (1980). Median estimation in sample surveys. In Proceedings of the Section on Survey Research Methods (Vol. 1814184). American Statistical Association.
- Hajek, J. (1960). Limiting distributions in simple random sampling from a finite population. Publications of the Mathematics Institute of the Hungarian Academy of Science, 5, 361-74.
- Hajek, J. (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. The Annals of Mathematical Statistics, 506-523.
- Hajek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics, 1491-1523.

- Hajek, J. (1981). Sampling from a finite population (Vol. 37). V. Dupac (Ed.). M. Dekker.
- Hansen, M. H., and Hurwitz, W. N. (1953). Sample survey methods and theory. Vol. I et II, Wiley, New-York.
- Hasler, C., and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. Computational Statistics and Data Analysis.
- Haziza, D. (2009). Imputation and inference in the presence of missing data. Handbook of Statistics, 29, 215-246.
- Haziza, D., and Rao, J. N. (2006). A nonresponse model approach to inference under imputation for missing survey data. Survey Methodology, 32(1), 53.
- Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260), 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), 89-96.
- Kalton, G., and Kish, L. (1981). Two efficient random imputation procedures. Proceedings of the Survey Research Methods, American Statistical Association, 146-151.
- Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. Communications in Statistics-Theory and Methods, 13(16), 1919-1939.
- Kim, J. K., and Fuller, W. (2004). Fractional hot deck imputation. Biometrika, 91(3), 559-578.
- Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, Jackknife and balanced repeated replication methods. The Annals of Statistics, 9, 1010-1019.
- Lesage, E. (2013). Utilisation d'information auxiliaire en théorie des sondages à l'étape de l'échantillonnage et à l'étape de l'estimation (Doctoral dissertation, University of Rennes 1).
- Lin, C. D., and Lu, W. W., and Rust, K., and Sitter, R. R. (2013). Replication variance estimation in unequal probability sampling without replacement: One stage and two stage. Canadian Journal of Statistics, 41(4), 696-716.
- Lohr, S. L. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. Survey Methodology, 37(2), 197-213.
- Mallows, C. L. (1972). A note on asymptotic joint normality. The Annals of Mathematical Statistics, 508-515.

- Mashregi, Z., and Léger, C. and Haziza, D. (2014). Bootstrap Methods for Imputed Data from Regression, Ratio and Hot Deck Imputation. Canadian Journal of Statistics, 42, 142-167
- Nigam, A. K., and Rao, J. N. K. (1996). On balanced bootstrap for stratified multistage samples. Statistica sinica, 6(1), 199-214.
- Ohlsson, E. (1986). Asymptotic normality of the Rao, Hartley, Cochran Estimator: An Application of the Martingale CLT. Scandinavian Journal of Statistics, 13, 17-28.
- Ohlsson, E. (1989). Asymptotic normality for two-stage sampling from a finite population. Probability Theory and Related Fields, 81, 341-352.
- Presnell, B., and Booth, J. G. (1994). Resampling methods for sample surveys. Presnell, B. and Booth, JG (1994) Resampling methods for sample surveys. Technical Report 470, Department of Statistics, University of Florida, Gainesville, FL.
- Preston, J. (2009). Rescaled bootstrap for stratified multistage sampling. Survey Methodology, 35(2), 227-234.
- Qualité, L. (2008), A comparison of conditional Poisson sampling versus unequal probability sampling with replacement, Journal of Statistical Planning and Inference, 138, 1428–1432.
- R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rao, J.N.K., and 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.
- Rao, J. N.K., and Wu, C. F. J. (1988). Resampling inference with complex survey data. Journal of the american statistical association, 83(401), 231-241.
- Rao, J. N. K., and Wu, C. F. J., and Yue, K. (1992). Some recent work on resampling methods for complex surveys. Survey methodology, 18(2), 209-217.
- Rosen (1972a). Asymptotic theory for successive sampling with varying probabilities without replacement, I. The Annals of Mathematical Statistics, 43, 373-397.
- Rosen (1972b). Asymptotic theory for successive sampling with varying probabilities without replacement, II. The Annals of Mathematical Statistics, 43, 748-776.
- Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.
- Saegusa, T., and Wellner, J.A. (2013). Weighted likelihood estimation under two-phase sampling. The Annals of Statistics, 41, 269-295.

- Särndal, C. E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. Survey Methodology, 18(2), 241-252.
- Särndal, C. E., and Swensson, B., and Wretman, J. (2003). Model assisted survey sampling. Springer.
- Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. Journal of the Indian Society of Agricultural Statistics, 5(1194), 127.
- Sen, P. K. (1979). Invariance principles for the coupon collector's problem: a martingale approach. The Annals of Statistics, 7(2), 372-380.
- Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. Journal of the American Statistical Association, 94(445), 254-265.
- Shao, J., and Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation. Journal of the American Statistical Association, 97(458), 544-552.
- Shao, J., and Tu, D. (1995). The jackknife and bootstrap. Springer.
- Sitter, R. R. (1992). A resampling procedure for complex survey data. Journal of the American Statistical Association, 87(419), 755-765.
- Srivastava, M., and Carter, E. (1986). The maximum likelihood method for non-response in sample surveys. Statistics Canada, 12, 61-72.
- Thompson, M. (1997). Theory of sample surveys (Vol. 74). CRC Press.
- Thompson, M. E., and Wu, C. (2008). Simulation-based randomized systematic PPS sampling under substitution of units. Survey Methodology, 34(1), 3.
- Thorisson, H. (2000). Coupling, stationarity, and regeneration (pp. 90095-1555). New York: Springer.
- Tillé, Y. (2011). Sampling algorithms. Springer Berlin Heidelberg.
- Tillé, Y., and Favre, A.C. (2005). Optimal allocation in balanced sampling. Statistics and Probability Letters, 74(1), 31-37.
- Tillé, Y., and Matei, A. (2008). Sampling: Survey Sampling. `R` package version 2.0.
- Yates, F., and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. Journal of the Royal Statistical Society. Series B (Methodological), 253-261.