Inference for two-stage sampling designs with application to a panel for urban policy

Guillaume Chauvet, Audrey-Anne Vallée

Inference for two-stage sampling designs with application to a panel for urban policy

Guillaume Chauvet[(1)] and Audrey-Anne Vallée[(2)]

[(1)] Ensai (Irmar), Campus de Ker Lann, Bruz - France

[(2)] Institut de Statistique, Université de Neuchâtel, Switzerland

January 2, 2019

Abstract

Two-stage sampling designs are commonly used for household and health surveys. To produce reliable estimators with assorted confidence intervals, some basic statistical properties like consistency and asymptotic normality of the Horvitz-Thompson estimator are desirable, along with the consistency of assorted variance estimators. These properties have been mainly studied for single-stage sampling designs. In this work, we prove the consistency of the Horvitz-Thompson estimator and of associated variance estimators for a general class of two-stage sampling designs, under mild assumptions. We also study two-stage sampling with a large entropy sampling design at the first stage, and prove that the Horvitz-Thompson estimator is asymptotically normally distributed through a coupling argument. When the

first-stage sampling fraction is negligible, simplified variance estimators which do not require estimating the variance within the Primary Sampling Units are proposed, and shown to be consistent. An application to a panel for urban policy, which is the initial motivation for this work, is also presented.

Keywords: Asymptotic normality, coupling method, rejective sampling, simplified variance estimator.

1 Introduction

In household and health surveys, the population is often sparse over a large territory and there is regularly no sampling frame. Two-stage sampling designs are convenient in such situations. The population are grouped into large blocks (e.g., municipalities or counties), called Primary Sampling Units (PSUs), which are sampled at the first stage. Only a frame of these PSUs is needed at this stage, which is easier to create. At the second stage, a list of population units is obtained inside the selected PSUs, and a sample of these population units is selected. Despite its convenience, multistage sampling has the drawback to lead to estimators with inflated variance, compared to sampling designs where the population units are directly selected. A detailed treatment of multistage sampling may be found in Cochran (1977), Särndal et al. (1992) and Fuller (2011).

To produce reliable estimators with assorted confidence intervals, some statistical properties are needed for a sampling design: (a) the Horvitz-Thompson

estimator should be consistent for the true total; also, (b) this estimator should be asymptotically normally distributed, and (c) consistent variance estimators should be available, to be able to produce normality-based confidence intervals. General conditions for the consistency of the Horvitz-Thompson estimator are given in Isaki and Fuller (1982) and Robinson (1982), see also Prášková and Sen (2009). The asymptotic normality is usually studied design by design, see for example Hájek (1964) for rejective sampling, Rosén (1972) for successive sampling or Ohlsson (1986) for the Rao-Hartley-Cochran (1972) procedure; see also Bickel and Freedman (1984) for stratified simple random sampling and Chen and Rao (2007) for two-phase sampling designs. These properties are also studied in Breidt and Opsomer (2000) for the class of local polynomial regression estimators and in Breidt et al. (2016), but under assumptions that are not generally applicable for multistage sampling designs. More recently, Boistard et al. (2017) and Bertail et al. (2017) established functional central limit theorems for Horvitz-Thompson empirical processes. In summary, these properties have been mainly studied in the literature for one-stage sampling designs.

In two-stage sampling, the asymptotic properties of estimators are more difficult to study, due to the dependence introduced in the selection of the sampling units. Krewski and Rao (1981) studied the case when the primary units are selected with replacement, and Ohlsson (1989) derived a general central limit theorem for such designs. Recently, Chauvet (2015) considered coupling methods to prove the asymptotic normality of the Horvitz-Thompson estimator and the validity of a bootstrap procedure for stratified

simple random sampling at the first stage. However, there is a lack of general conditions ensuring that properties (a)-(c) hold for general two-stage sampling designs, and this is the purpose of the present paper. A notable exception is Breidt and Opsomer (2008), who obtain the consistency of the Horvitz-Thompson estimator under very weak conditions. This is discussed in Section 4.

In this paper, the properties of estimators and variance estimators are studied for a general class of two-stage sampling designs. The framework is introduced in Section 2 and the variance of the Horvitz-Thompson estimator is decomposed in a sum of three components. In Section 3, the assumptions used to establish the asymptotic properties are defined. In Section 4, the Horvitz-Thompson estimator is shown to be consistent under our conditions, and the order of magnitude of the three components of the variance is determined. The consistency of two unbiased variance estimators is established in Section 4.1. A simplified variance estimator which does not require estimating the variance within the PSUs can be produced. We prove in Section 4.2 that this variance estimator is consistent when the total variance within the PSUs is negligible. In Section 5, the specific case of large-entropy sampling designs at the first-stage is considered. When rejective sampling is used at the first-stage, the consistency of a Hájek-type variance estimator is established under reduced assumptions, along with the asymptotic normality of the Horvitz-Thompson estimator. We define a coupling procedure to extend these results to a more general class of large-entropy sampling designs at the first-stage. In Section 6, the properties of the Hájek-type variance estimators

are evaluated in a simulation study. An application to a panel for urban policy, which is the initial motivation for this work, is presented in Section 7.

2 Notation

We are interested in a finite population U of size N, in which a sample is selected by means of a two-stage sampling design. The units in U, called Secondary Sampling Units (SSUs) are partitioned into a population U_I of N_I Primary Sampling Units (PSUs). A sample S_I of n_I PSUs is selected in U_I. We are interested in estimating the population total

$$Y = \sum_{i=1}^{N_I} \sum_{k=1}^{N_i} y_{ik} = \sum_{i=1}^{N_I} Y_i, \tag{1}$$

for some variable of interest y, where $Y_i = \sum_{k=1}^{N_i} y_{ik}$ is the sub-total of the variable y on the PSU i and N_i is the number of SSUs inside the PSU i.

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. The index t is suppressed in what follows but all limiting processes are taken as $t \to \infty$. We assume that $N_I \to \infty$ and $n_I \to \infty$ as $t \to \infty$. We consider a single stratum of PSUs, but our results may be easily generalized to the case of a finite number of strata, see the application to the panel for urban policy in Section 7. An alternative asymptotic set-up is possible, under which the

number of strata tends to infinity while the sample size per stratum remains bounded, see Krewski and Rao (1981) and Breidt et al. (2016).

We note I_{Ii} for the sample membership indicator of the PSU i into S_I, $\pi_{Ii} = E(I_{Ii})$ for the inclusion probability of the PSU i, and $\pi_{Iij} = E(I_{Ii}I_{Ij})$ for the probability that the PSUs i and j are selected jointly in S_I. Inside any PSU $i \in S_I$, a sample S_i of n_i SSUs is selected at the second stage. We note

$$N_0 = \frac{1}{N_I}\sum_{i=1}^{N_I} N_i \quad \text{and} \quad n_0 = \frac{1}{N_I}\sum_{i=1}^{N_I} n_i \tag{2}$$

for the average size of the PSUs and for the average sample size selected inside the PSUs. We do not need particular assumptions on the limit behaviour of n_0 and N_0, and n_0 may be either bounded or unbounded. Our set-up covers in particular the case when the SSUs are comprehensively surveyed inside a selected PSU, which amounts to single-stage sampling on the population of PSUs.

For any SSU k in the PSU i, we note I_k for the sample membership indicator of k in S_i. Also, we note $\pi_{k|i} = E(I_k|i \in S_I)$ for the conditional inclusion probability of k, and $\pi_{kl|i} = E(I_k I_l|i \in S_I)$ for the conditional joint probability that two SSUs $k, l \in i$ are selected together in S_i. We assume invariance of the second-stage designs, as defined by Särndal et al. (1992): the second stage of sampling is independent of S_I. Also, we assume that the second-stage designs are independent from one PSU to another, conditionally on S_I.

The Horvitz-Thompson (HT) estimator of Y is

$$\hat{Y}_\pi = \sum_{i \in S_I} \frac{\hat{Y}_i}{\pi_{Ii}} \quad \text{with} \quad \hat{Y}_i = \sum_{k \in S_i} \frac{y_{ik}}{\pi_{k|i}}. \tag{3}$$

The variance of $\hat{Y}_\pi$ may be written as

$$\begin{aligned} V(\hat{Y}_\pi) &= \sum_{i=1}^{N_I} \sum_{j=1}^{N_I} \Delta_{Iij} \frac{Y_i}{\pi_{Ii}} \frac{Y_j}{\pi_{Ij}} + \sum_{i=1}^{N_I} \left(\frac{1 - \pi_{Ii}}{\pi_{Ii}} \right) V_i + \sum_{i=1}^{N_I} V_i \\ &= V_1(\hat{Y}_\pi) + V_2(\hat{Y}_\pi) + V_3(\hat{Y}_\pi) \end{aligned} \tag{4}$$

with $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$, and

$$V_i \equiv V(\hat{Y}_i) = \sum_{k=1}^{N_i} \sum_{l=1}^{N_i} \Delta_{kl|i} \frac{y_{ik}}{\pi_{k|i}} \frac{y_{il}}{\pi_{l|i}}, \tag{5}$$

with $\Delta_{kl|i} = \pi_{kl|i} - \pi_{k|i}\pi_{l|i}$. The term $V_1(\hat{Y}_\pi)$ is the variance due to the first stage. The sum of the two last terms in (4) may be simplified as

$$V_2(\hat{Y}_\pi) + V_3(\hat{Y}_\pi) = \sum_{i=1}^{N_I} \frac{V_i}{\pi_{Ii}}. \tag{6}$$

This is the variance due to the second stage of sampling.

When estimating the variance, the terms $V_1(\hat{Y}_\pi) + V_2(\hat{Y}_\pi)$ and $V_3(\hat{Y}_\pi)$ are handled separately. Variance estimators for these two terms are considered in Section 4, and proved to be consistent under assumptions which are stated and discussed in Section 3. In case of large entropy sampling designs at the first-stage, consistent variance estimators can be produced under reduced

assumptions, and without using second-order inclusion probabilities. This is studied in Section 5 .

3 Assumptions

To study the asymptotic properties of the estimators and variance estimators that we consider below, a number of assumptions are needed. We present in Section 3.1 the assumptions on the first-stage sampling design, and in Section 3.2 the assumptions on the second-stage sampling designs. The assumptions related to the variable of interest are presented in Section 3.3.

3.1 Assumptions on the first-stage sampling design

FS1: Some constant $f_{I0} < 1$ exists s.t.

$$N_I^{-1} n_I \quad \leq \quad f_{I0}. \tag{7}$$

Some constants $c_{I1}, C_{I1} > 0$ exist s.t. for any PSU i

$$c_{I1} \quad \leq N_I n_I^{-1} \pi_{Ii} \leq \quad C_{I1}. \tag{8}$$

FS2: Some constants $C_{I2}, C_{I3} > 0$ exist s.t. for any PSUs $i \neq j \neq i'$

$$\pi_{Iij} \leq \quad C_{I2} N_I^{-2} n_I^2, \tag{9}$$

$$\pi_{Iiji'} \leq \quad C_{I3} N_I^{-3} n_I^3, \tag{10}$$

with $\pi_{Iiji'}$ the probability that the PSUs i, j, i' are selected together in S_I. Some constants C_{I4}, C_{I5} exist s.t.

$$\Delta_{I1} \equiv \max_{i \neq j = 1,\ldots,N_I} |\pi_{Iij} - \pi_{Ii}\pi_{Ij}| \leq C_{I4} N_I^{-2} n_I, \tag{11}$$
$$\Delta_{I2} \equiv \max_{i \neq j \neq i' \neq j' = 1,\ldots,N_I} |\pi_{Iiji'j'} - \pi_{Ii}\pi_{Ij}\pi_{Ii'}\pi_{Ij'}| \leq C_{I5} N_I^{-4} n_I^3,$$

with $\pi_{Iiji'j'}$ the probability that the PSUs i, j, i', j' are selected together in S_I.

FS3: Some constant $c_{I2} > 0$ exists s.t. for any $i \neq j = 1, \ldots, N_I$

$$c_{I2} N_I^{-2} n_I^2 \leq \pi_{Iij}. \tag{12}$$

The Assumption (FS1) is related to the order of magnitude of the first-stage sample size n_I, and to the first-order inclusion probabilities. Equation (7) ensures that the first-stage sample is not degenerate, in the sense that the PSUs are not comprehensively surveyed. This assumption is compatible with the case $n_I/N_I \to 0$ (negligible first-stage sampling fraction). A similar condition is considered in (Breidt and Opsomer, 2000, assumption A5), and in (Boistard et al., 2017, assumption HT3). Equation (8) states that the first-order inclusion probabilities do not depart much from that obtained under simple random sampling. The same condition is considered in (Boistard et al., 2017, assumption C1). Overall, (FS1) is under the control of the survey sampler.

The Assumption (FS2) is related to the inclusion probabilities of order 2 to 4.

If (FS1) holds, equations (9) and (10) will automatically hold for negatively associated sampling designs (e.g. Brändén and Jonasson, 2012) which includes simple random sampling, rejective sampling (Hájek, 1964), Sampford sampling (Sampford, 1967) and pivotal sampling (Deville and Tillé, 1998; Chauvet, 2012), for example. In equation (11), the quantities Δ_{I1} and Δ_{I2} are two measures of dependency in the selection of units. These quantities will be equal to 0 when the units are selected independently, which is known as Poisson sampling (see for example Fuller, 2011, p. 13). Equation (11) is respected for simple random sampling. If (FS1) holds, it is also respected under rejective sampling (see Boistard et al., 2012, Theorem 1), and it can be proved that it holds for the Rao-Sampford sampling design (see Hajek, 1981, Chapter 8). Similar conditions are considered in (Breidt and Opsomer, 2000, assumption A7), and in (Boistard et al., 2017, conditions C2-C4).

The Assumption (FS3) provides a uniform lower bound for the second-order inclusion probabilities. A similar condition is considered in (Breidt and Opsomer, 2000, assumption A6). This assumption holds for simple random sampling, but is more difficult to prove for unequal probability sampling designs. On the other hand, it is needed to prove the consistency of the Horvitz-Thompson variance estimator and of the Yates-Grundy variance estimator, see our Section 4.1 and Theorem 3 in Breidt and Opsomer (2000). We consider in Section 5 the specific case of large entropy sampling designs at the first-stage, for which alternative consistent variance estimators are possible, and for which the assumption (FS3) can be suppressed.

3.2 Assumptions on the second-stage sampling design

SS0: Some constants $\lambda_1, \Lambda_1 > 0$ and $\phi_1, \Phi_1 > 0$ exist s.t. for any PSU i

$$\lambda_1 n_0 \leq n_i \leq \Lambda_1 n_0, \tag{13}$$

$$\phi_1 N_0 \leq N_i \leq \Phi_1 N_0. \tag{14}$$

SS1: Some constants $c_1, C_1 > 0$ exist s.t. for any PSU i and for any k inside:

$$c_1 \leq N_0 n_0^{-1} \pi_{k|i} \leq C_1. \tag{15}$$

SS2: Some constants $C_2, C_3 > 0$ exists s.t. for any PSU i and any $k \neq l \neq k'$ inside:

$$\pi_{kl|i} \leq C_2 N_0^{-2} n_0^2, \tag{16}$$

$$\pi_{klk'|i} \leq C_3 N_0^{-3} n_0^3, \tag{17}$$

with $\pi_{klk'|i}$ the conditional probability that the SSUs k, l, k' are selected together in S_i. Also, some constants C_4, C_5 exist s.t.

$$\Delta_1 \equiv \max_{i=1,\ldots,N_I} \max_{k \neq l = 1,\ldots,N_i} \left| \pi_{kl|i} - \pi_{k|i}\pi_{l|i} \right| \leq C_4 N_0^{-2} n_0, \tag{18}$$

$$\Delta_2 \equiv \max_{i=1,\ldots,N_I} \max_{k \neq l \neq k' \neq l' = 1,\ldots,N_i} \left| \pi_{klk'l'|i} - \pi_{k|i}\pi_{l|i}\pi_{k'|i}\pi_{l'|i} \right| \leq C_5 N_0^{-4} n_0^3,$$

with $\pi_{klk'l'|i}$ the conditional probability that the SSUs k, l, k', l' are selected together in S_i.

SS3: Some constant $c_2 > 0$ exists s.t. for any PSU i and for any $k \neq l$ inside:

$$c_2 N_0^{-2} n_0^2 \ \leq \pi_{kl|i}. \tag{19}$$

It is assumed in (SS0) that the sizes N_i of the PSUs are comparable, and that the numbers n_i of SSUs selected inside the PSUs are also comparable. In practice, to reduce the variance associated to the first stage of sampling, the PSUs are usually grouped into strata in such a way that the PSUs inside one stratum are of similar sizes. Also, the number of selected SSUs is commonly the same for any PSU, so that all the interviewers have a comparable workload. Equations (13) and (14) seem therefore reasonable in practice. The assumptions (SS1)-(SS3) are similar to the assumptions (FS1)-(FS3) made for the first-stage sampling design.

As previously mentioned, one-stage sampling designs are a particular case of our set-up. They are obtained when $N_i = 1$ for any PSU i and when $n_i = 1$ for any unit $i \in S_I$. In such case, assumptions (SS0)-(SS1) automatically hold while assumptions (SS2)-(SS3) vanish.

3.3 Assumptions on the variable of interest

VAR1: There exists some constants M_1 and $m_1 > 0$ such that

$$N^{-1} \sum_{i=1}^{N_I} \sum_{k=1}^{N_i} y_{ik}^4 \ \leq M_1, \tag{20}$$

$$m_1 \leq \ N^{-1} \sum_{i=1}^{N_I} \sum_{k=1}^{N_i} y_{ik}. \tag{21}$$

VAR2: There exists some constant $m_2 > 0$ such that

$$m_2 \leq N^{-2} n_I \left\{ V_1(\hat{Y}_\pi) \right\}. \tag{22}$$

It is assumed in (VAR1) that the variable of interest has a bounded moment of order four, and a mean bounded away from 0. It is assumed in (VAR2) that the first-stage sampling variance is non-vanishing. These assumptions are fairly weak, although we may find situations under which they are not respected. The condition (20) is not fulfilled for heavily skewed populations, where a non-negligible part of the individuals exhibit particularly large values for the variable of interest. This may be the case in wealth surveys, for example. Equations (21) and (22) are not fulfilled when we are interested in domain estimation, and when the domain size N_d is negligible as compared to the population size.

4 Consistency of estimators

We begin with determining the orders of magnitude of the components of the variance decomposition in (4). The proof of Proposition 1 follows from some moment inequalities, which are given in Section 1 of the Supplementary Material.

Proposition 1. *Suppose that assumptions (FS1)-(FS2),(SS0)-(SS2) and (VAR1)*

hold. Then

$$\begin{aligned} V_1(\hat{Y}_\pi) &= O\left(N^2 n_I^{-1}\right), \\ V_2(\hat{Y}_\pi) &= O\left(N^2 n_I^{-1} n_0^{-1}\right), \\ V_3(\hat{Y}_\pi) &= O\left(N^2 N_I^{-1} n_0^{-1}\right). \end{aligned} \tag{23}$$

When $n_0 \to \infty$, the first variance component is the leading term and the two last ones are negligible. When n_0 is bounded, the first and second component have the same order of magnitude. The third component is negligible if $N_I^{-1} n_I \to 0$, and has the same order of magnitude otherwise. In practice, the third term is expected to be small as compared to the two first ones.

The consistency of the HT estimator is established in Proposition 2. The proof follows from Proposition 1, and is therefore omitted.

Proposition 2. *Suppose that assumptions (FS1)-(FS2),(SS0)-(SS2) and (VAR1) hold. Then the HT estimator is design-unbiased. Also, we have*

$$E\left[N^{-1}\left\{\hat{Y}_\pi - Y\right\}\right]^2 = O(n_I^{-1}) \quad \textit{and} \quad \frac{\hat{Y}_\pi}{Y} \longrightarrow_{Pr} 1, \tag{24}$$

where $\to_{Pr}$ stands for the convergence in probability.

Proposition 2 implies that the HT estimator is $\sqrt{n}$-consistent for the true total. Also, it is important to note that the consistency of the HT-estimator requires that the sampled number of PSUs n_I tends to infinity, while the consistency is not related to the behaviour of n_0. For example, suppose that a sample of same size $n_i = n_0$ is selected inside any PSU, so that the total num-

ber of SSUs selected is $n = n_I n_0$. Then, even if $n \to \infty$, the HT-estimator may be inconsistent if n_I is bounded. In practice, it is therefore important that a large number of PSUs is selected at the first stage.

The consistency of the HT-estimator is proved in Breidt and Opsomer (2008) under the alternative assumptions:

D4: For any population U, $\min_{k \in U} \pi_k \geq \pi^* > 0$ where $N\pi^* \to \infty$, and there exists $\kappa \geq 0$ such that

$$N^{0.5+\kappa}(\pi^*)^2 \to \infty \quad \text{and} \quad \max_{k \in U} \sum_{l \in U; l \neq k} (\Delta_{kl})^2 = O(N^{-2\kappa}),$$

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

D5: The variable of interest satisfies

$$\limsup \frac{1}{N} \sum_{i=1}^{N_I} \sum_{k=1}^{N_i} y_{ik}^2 \leq \infty.$$

Under (D4) and (D5), we have $E\left[N^{-1}\left\{\hat{Y}_\pi - Y\right\}\right]^2 = o(1)$ (Breidt and Opsomer, 2008, Lemma A.1). Clearly, our condition in (VAR1) on the fourth moment implies (D5). Also, it can be shown that if n_0 is bounded, our assumptions (FS1)-(FS2) and (SS0)-(SS2) imply (D4) with $\kappa = 0$. Our stronger conditions are needed in particular to get the consistency of variance estimators, see Sections 4.1 and 4.2.

4.1 Unbiased variance estimators

We first consider the so-called Horvitz-Thompson variance estimator

$$\begin{aligned}\hat{V}_{HT}(\hat{Y}_\pi) &= \sum_{i,j\in S_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{\hat{Y}_i}{\pi_{Ii}} \frac{\hat{Y}_j}{\pi_{Ij}} + \sum_{i\in S_I} \frac{\hat{V}_{HT,i}}{\pi_{Ii}} \\ &= \hat{V}_{HT,A}(\hat{Y}_\pi) + \hat{V}_{HT,B}(\hat{Y}_\pi), \end{aligned} \tag{25}$$

where

$$\hat{V}_{HT,i} = \sum_{k,l\in S_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} \frac{y_{ik}}{\pi_{k|i}} \frac{y_{il}}{\pi_{l|i}}. \tag{26}$$

Proposition 3. *If assumptions (FS1)-(FS3),(SS0)-(SS2) and (VAR1) hold, we have:*

$$E\left[N^{-2} n_I \left\{\hat{V}_{HT,A}(\hat{Y}_\pi) - V_1(\hat{Y}_\pi) - V_2(\hat{Y}_\pi)\right\}\right]^2 = O(n_I^{-1}). \tag{27}$$

If assumptions (FS1)-(FS2),(SS0)-(SS3) and (VAR1) hold, we have:

$$E\left[N^{-2} N_I n_0 \left\{\hat{V}_{HT,B}(\hat{Y}_\pi) - V_3(\hat{Y}_\pi)\right\}\right]^2 = O(n_I^{-1}). \tag{28}$$

If assumptions (FS1)-(FS3),(SS0)-(SS3),(VAR1)-(VAR2) hold, we have:

$$E\left[N^{-2} n_I \left\{\hat{V}_{HT}(\hat{Y}_\pi) - V(\hat{Y}_\pi)\right\}\right]^2 = O(n_I^{-1}) \text{ and } \frac{\hat{V}_{HT}(\hat{Y}_\pi)}{V(\hat{Y}_\pi)} \to_{Pr} 1. \tag{29}$$

The proof of Proposition 3 is tedious but standard, and is therefore omitted. It implies that $\hat{V}_{HT}(\hat{Y}_\pi)$ is a term by term unbiased and $\sqrt{n}$-consistent variance estimator, in the sense that $\hat{V}_{HT,A}(\hat{Y}_\pi)$ is unbiased and $\sqrt{n}$-consistent

for $V_1(\hat{Y}_\pi)+V_2(\hat{Y}_\pi)$, and $\hat{V}_{HT,B}(\hat{Y}_\pi)$ is unbiased and $\sqrt{n}$-consistent for $V_3(\hat{Y}_\pi)$. In their Theorem 3, Breidt and Opsomer (2000) state a similar result in case of one-stage sampling designs, for a more general class of estimators that they call local polynomial estimators. In the literature, the consistency of the HT-variance estimator is often stated as an assumption; e.g., Kim et al. (2017) for two-stage sampling designs.

If the sampling designs used at both stages are of fixed size, we may alternatively use the Yates-Grundy variance estimator

$$\begin{aligned}\hat{V}_{YG}(\hat{Y}_\pi) &= -\frac{1}{2}\sum_{i\neq j\in S_I}\frac{\Delta_{Iij}}{\pi_{Iij}}\left(\frac{\hat{Y}_i}{\pi_{Ii}}-\frac{\hat{Y}_j}{\pi_{Ij}}\right)^2+\sum_{i\in S_I}\frac{\hat{V}_{YG,i}}{\pi_{Ii}}\\ &= \hat{V}_{YG,A}(\hat{Y}_\pi)+\hat{V}_{YG,B}(\hat{Y}_\pi),\end{aligned} \tag{30}$$

with

$$\hat{V}_{YG,i} = -\frac{1}{2}\sum_{k\neq l\in S_i}\frac{\Delta_{kl|i}}{\pi_{kl|i}}\left(\frac{y_{ik}}{\pi_{k|i}}-\frac{y_{il}}{\pi_{l|i}}\right)^2. \tag{31}$$

We prove in Proposition 4 that $\hat{V}_{YG}(\hat{Y}_\pi)$ is also a term by term unbiased and $\sqrt{n}$-consistent variance estimator. The proof is similar to that of Proposition 3.

Proposition 4. *If assumptions (FS1)-(FS3),(SS0)-(SS2) and (VAR1) hold, we have:*

$$E\left[N^{-2}n_I\left\{\hat{V}_{YG,A}(\hat{Y}_\pi)-V_1(\hat{Y}_\pi)-V_2(\hat{Y}_\pi)\right\}\right]^2 = O(n_I^{-1}). \tag{32}$$

If assumptions (FS1)-(FS2),(SS0)-(SS3) and (VAR1) hold, we have:

$$E\left[N^{-2}N_I n_0\left\{\hat{V}_{YG,B}(\hat{Y}_\pi) - V_3(\hat{Y}_\pi)\right\}\right]^2 \;=\; O(n_I^{-1}). \tag{33}$$

If assumptions (FS1)-(FS3),(SS0)-(SS3), (VAR1)-(VAR2) hold, we have:

$$E\left[N^{-2}n_I\left\{\hat{V}_{YG}(\hat{Y}_\pi) - V(\hat{Y}_\pi)\right\}\right]^2 = O(n_I^{-1}) \text{ and } \frac{\hat{V}_{YG}(\hat{Y}_\pi)}{V(\hat{Y}_\pi)} \to_{Pr} 1. \tag{34}$$

4.2 Simplified one-term variance estimators

Both the variance estimators $\hat{V}_{HT}(\hat{Y}_\pi)$ and $\hat{V}_{YG}(\hat{Y}_\pi)$ may be cumbersome in practice, since they require an unbiased and consistent variance estimator $\hat{V}_{HT,i}$ or $\hat{V}_{YG,i}$ inside any of the selected PSUs. Consider the example of self-weighted two-stage sampling designs, which are common in practice. They consist in selecting a sample of PSUs, with probabilities π_{Ii} proportional to the size of the PSUs, and a sample of n_0 SSUs inside any of the selected PSUs. This leads to equal sampling weights for all the SSUs in the population, hence the name. In case of self-weighted two-stage sampling designs, systematic sampling is frequently used at the second stage. In such case, the assumption (SS3) is usually not respected.

A simplified variance estimator can be obtained by using $\hat{V}_{HT,A}(\hat{Y}_\pi)$ only, or for a fixed-size sampling design $\hat{V}_{YG,A}(\hat{Y}_\pi)$ only, see for instance Särndal et al. (1992). Proposition 5 states that these simplified estimators are consistent

when

$$\frac{V_3(\hat{Y}_\pi)}{V_1(\hat{Y}_\pi) + V_2(\hat{Y}_\pi)} \quad \to \quad 0, \tag{35}$$

i.e. when the third component of the variance in the decomposition (4) is negligible. Note that in Proposition 5 we do not need the assumption (SS3) which guarantees a lower bound for the second-order inclusion probabilities at the second stage.

Proposition 5. *Suppose that assumptions (FS1)-(FS3), (SS0)-(SS2), (VAR1)-(VAR2) hold. Suppose that equation (35) holds. Then*

$$E\left[N^{-2} n_I \left\{\hat{V}_{HT,A}(\hat{Y}_\pi) - V(\hat{Y}_\pi)\right\}\right]^2 \quad = \quad o(1), \tag{36}$$

$$\frac{\hat{V}_{HT,A}(\hat{Y}_\pi)}{V(\hat{Y}_\pi)} \quad \longrightarrow_{Pr} \quad 1. \tag{37}$$

If in addition the first-stage sampling design is of fixed-size, we have

$$E\left[N^{-2} n_I \left\{\hat{V}_{YG,A}(\hat{Y}_\pi) - V(\hat{Y}_\pi)\right\}\right]^2 \quad = \quad o(1), \tag{38}$$

$$\frac{\hat{V}_{YG,A}(\hat{Y}_\pi)}{V(\hat{Y}_\pi)} \quad \longrightarrow_{Pr} \quad 1. \tag{39}$$

The proof is immediate from Propositions 3 and 4, and by using equation (35). The simplified variance estimators $\hat{V}_{HT,A}(\hat{Y}_\pi)$ and $\hat{V}_{YG,A}(\hat{Y}_\pi)$ are simpler to compute, since they do not involve variance estimators $\hat{V}_i$ inside PSUs, but only unbiased estimators $\hat{Y}_i$ for the sub-totals over the PSUs.

Under the assumptions (FS1)-(FS3), (SS0)-(SS2) and (VAR1)-(VAR2), a

sufficient condition for equation (35) to hold is that $N_I^{-1} n_I \to 0$ (negligible first-stage sampling rate). In practice, we expect the term $V_3(\hat{Y}_\pi)$ to have a small contribution in the overall variance even if the first-stage sampling rate is not negligible. This is illustrated in Section 6 through a simulation study, and in Section 7 in the application to the panel for urban policy. The two simplified variance estimators $\hat{V}_{HT,A}(\hat{Y}_\pi)$ and $\hat{V}_{YG,A}(\hat{Y}_\pi)$ may therefore be reasonable choices for variance estimation in practice.

5 Case of large entropy sampling designs

In this Section, we focus on the situation when large entropy sampling designs are used at the first stage. We consider a Hájek-type variance estimator, and prove its consistency with limited assumptions, namely by dropping the conditions (FS2) and (FS3). Building on the work of Ohlsson (1989), we also prove that the HT-estimator is asymptotically normally distributed. The rejective sampling design (Hájek, 1964) is first considered in Section 5.1. The results are extended in Section 5.2 to a class of large entropy sampling designs by using a coupling algorithm. The properties of a simplified variance estimator are studied in Section 5.3.

5.1 Rejective sampling

The rejective (or conditional Poisson) sampling design was introduced by Hájek (1964). Rejective sampling in U_I consists in repeatedly selecting samples by means of Poisson sampling, until the sample has the required size n_I. The inclusion probabilities of the Poisson sampling design are chosen so that

the required inclusion probabilities $\pi_{Ii},\ i \in U_I$ are respected; see for example Dupakova (1975). The rejective sampling design has been extensively studied in the literature, see Tillé (2006) for a review. Under a rejective sampling design at the first-stage, the assumption (FS2) is implied by the assumption (FS1), see our discussion in Section 3.1.

We note $p_r(\cdot)$ the rejective sampling design with inclusion probabilities π_{Ii} in the population U_I. Also, we note S_{rI} a first-stage sample selected by means of p_r, and

$$\hat{Y}_{r\pi} \ = \ \sum_{i \in S_{rI}} \frac{\hat{Y}_i}{\pi_{Ii}} \tag{40}$$

the associated HT-estimator. Making use of a uniform approximation of the second-order inclusion probabilities, Hájek (1964) proposed a very simple variance estimator for which these second-order inclusion probabilities are not needed. In our two-stage sampling context, this leads to replacing in (25) the term $\hat{V}_{HT,A}(\hat{Y}_{r\pi})$ with

$$\hat{V}_{HAJ,A}(\hat{Y}_{r\pi}) \ = \ \begin{cases} \displaystyle\sum_{i \in S_{rI}} (1-\pi_{Ii}) \left(\frac{\hat{Y}_i}{\pi_{Ii}} - \hat{\hat{R}}_{r\pi} \right)^2 & \text{if } \hat{d}_{rI} \geq \frac{c_{I0}}{2} n_I, \\ 0 & \text{otherwise,} \end{cases} \tag{41}$$

with

$$\hat{\hat{R}}_{r\pi} = \hat{d}_{rI}^{-1} \sum_{i \in S_{rI}} (1-\pi_{Ii}) \frac{\hat{Y}_i}{\pi_{Ii}} \quad \text{and} \quad \hat{d}_{rI} = \sum_{i \in S_{rI}} (1-\pi_{Ii}), \tag{42}$$

and where c_{I0} is defined in Lemma **??** (see the Supplementary Material). This leads to the global variance estimator

$$\hat{V}_{HAJ}(\hat{Y}_{r\pi}) \;=\; \hat{V}_{HAJ,A}(\hat{Y}_{r\pi}) + \hat{V}_{HT,B}(\hat{Y}_{r\pi}), \tag{43}$$

where $\hat{V}_{HT,B}(\hat{Y}_{r\pi})$ is defined in equation (25). If the second-stage sampling designs are all of fixed-size, we could alternatively replace $\hat{V}_{HT,B}(\hat{Y}_{r\pi})$ with $\hat{V}_{YG,B}(\hat{Y}_{r\pi})$ given in equation (30).

Note that the variance estimator $\hat{V}_{HAJ}(\hat{Y}_{r\pi})$ is truncated to avoid extreme values for $\hat{\hat{R}}_{r\pi}$. This is needed to establish its consistency, which is done in Proposition 6. An advantage of this variance estimator is that the first-stage second-order inclusion probabilities are not required. In particular, the condition (FS3) is not needed to prove the consistency. We also prove in Proposition 6 that the HT-estimator is asymptotically normally distributed, by using Theorem 2.1 in Ohlsson (1989).

Proposition 6. *Suppose that a rejective sampling design is used at the first stage. Suppose that assumptions (FS1), (SS0)-(SS2) and (VAR1) hold. Then*

$$E\left[N^{-2}n_I\left\{\hat{V}_{HAJ,A}(\hat{Y}_{r\pi}) - V_1(\hat{Y}_{r\pi}) - V_2(\hat{Y}_{r\pi})\right\}\right]^2 \;=\; o(1). \tag{44}$$

If in addition the assumption (VAR2) holds, then

$$\frac{\hat{Y}_{r\pi} - Y}{\sqrt{V(\hat{Y}_{r\pi})}} \;\longrightarrow_{\mathcal{L}}\; \mathcal{N}(0,1), \tag{45}$$

where $\rightarrow_{\mathcal{L}}$ *stands for the convergence in distribution. If in addition the assumption (SS3) holds, then*

$$E\left[N^{-2}n_I\left\{\hat{V}_{HAJ}(\hat{Y}_{r\pi})-V(\hat{Y}_{r\pi})\right\}\right]^2=o(1)\ \textit{and}\ \frac{\hat{V}_{HAJ}(\hat{Y}_{r\pi})}{V(\hat{Y}_{r\pi})}\rightarrow_{Pr}1. \quad (46)$$

The proof is given in Section 2 of the Supplementary Material. The asymptotic normality of the HT-estimator has been proved by Hájek (1964) for a single stage rejective sampling design, but the consistency of the Hájek-type variance estimator has not been rigorously established previously. Proposition 6 has therefore its own interest, even for one-stage sampling designs. It follows that under rejective sampling at the first-stage, an approximate two-sided $100(1-2\alpha)\%$ confidence interval for Y is obtained as

$$\left[\hat{Y}_{r\pi}\pm u_{1-\alpha}\{\hat{V}_{HAJ}(\hat{Y}_{r\pi})\}^{0.5}\right], \quad (47)$$

with $u_{1-\alpha}$ the quantile of order $1-\alpha$ of the standard normal distribution.

5.2 Other sampling designs

We consider a more general class of sampling designs at the first-stage, which are close to the rejective sampling design with respect to the Chi-square distance. Other distance functions have been considered in the literature, such as the Hellinger distance (Conti, 2014) or the total variation distance (Bertail et al., 2017). We note $p(\cdot)$ for a fixed-size sampling design with inclusion probabilities π_{Ii} in the population U_I. It is said to be close to the

rejective sampling design $p_r(\cdot)$ with respect to the Chi-square distance if

$$d_2(p, p_r) \to 0 \quad \text{where} \quad d_2(p, p_r) = \sum_{s_I \subset U_I;\ p_r(s_I)>0} \frac{\{p(s_I) - p_r(s_I)\}^2}{p_r(s_I)} \quad (48)$$

Equation (48) holds for the Rao-Sampford (Sampford, 1967) sampling design, for example. We note S_{pI} a first-stage sample selected by means of $p(\cdot)$, and the associated HT-estimator is

$$\hat{Y}_{p\pi} = \sum_{i \in S_{pI}} \frac{\hat{Y}_i}{\pi_{Ii}}. \quad (49)$$

We introduce in Algorithm 1 a coupling procedure to obtain the estimators $\hat{Y}_{p\pi}$ and $\hat{Y}_{r\pi}$ jointly, which is the main tool in extending the results in Proposition 6 to $\hat{Y}_{p\pi}$. We note

$$\alpha = 1 - d_{TV}(p, p_r) \quad \text{where} \quad d_{TV}(p, p_r) = \frac{1}{2} \sum_{s_I \in U_I} |p(s_I) - p_r(s_I)| \quad (50)$$

is the total variation distance between $p(\cdot)$ and $p_r(\cdot)$. By using Lemma 11 in Section 3 of the Supplementary Material, it can be proved that the coupling procedure in Algorithm 1 leads to estimators $\hat{Y}_{r\pi}$ and $\hat{Y}_{p\pi}$ associated to the required two-stage sampling designs; see also van Der Hofstad (2016), Theorem 2.9.

Proposition 7. *Suppose that the samples S_{rI} and S_{pI} are selected by means of the coupling procedure in Algorithm 1. Then:*

$$E\left(\hat{Y}_{p\pi} - \hat{Y}_{r\pi}\right)^2 \leq \sum_{s_I \in U_I} |p(s_I) - p_r(s_I)| \left\{ \left(\sum_{i \in s_I} \frac{Y_i}{\pi_{Ii}} - Y \right)^2 + \sum_{i \in s_I} \frac{V_i}{\pi_{Ii}^2} \right\}.$$

Algorithm 1 A coupling procedure between two-stage sampling designs

1. Draw u from a uniform distribution $U[0,1]$.
2. If $u \leq \alpha$, then:
   (a) Select a sample s_I with probabilities $\dfrac{p(s_I) \wedge p_r(s_I)}{\alpha}$, and take $S_{rI} = S_{pI} = s_I$.
   (b) For any $i \in S_{rI} = S_{pI}$, select the same second-stage sample S_i for both $\hat{Y}_{r\pi}$ and $\hat{Y}_{p\pi}$.
3. If $u > \alpha$, then:
   (a) Select the sample S_{pI} with probabilities $\dfrac{p(s_I) - p_r(s_I)}{1-\alpha}$ in the set $\{s_I \in U_I;\ p(s_I) > p_r(s_I)\}$. For any $i \in S_{pI}$, select a second-stage sample S_i for $\hat{Y}_{p\pi}$.
   (b) Independently of S_{pI} and of the associated second-stage samples S_i's, select the sample S_{rI} with probabilities $\dfrac{p_r(s_I) - p(s_I)}{1-\alpha}$ in the set $\{s_I;\ p(s_I) \leq p_r(s_I)\}$. For any $i \in S_{rI}$, select a second-stage sample S_i for $\hat{Y}_{r\pi}$.

Proposition 8. *Suppose that the samples S_{rI} and S_{pI} are selected by means of the coupling procedure in Algorithm 1. Suppose that assumptions (FS1), (SS0)-(SS2) and (VAR1) hold. Suppose that $d_2(p, p_r) \to 0$. Then*

$$E\left(\hat{Y}_{p\pi} - \hat{Y}_{r\pi}\right)^2 \;=\; o\left(N^2 n_I^{-1}\right). \tag{51}$$

If in addition the assumption (VAR2) holds, then

$$\frac{V\left(\hat{Y}_{p\pi}\right)}{V\left(\hat{Y}_{r\pi}\right)} \;\to\; 1. \tag{52}$$

The proofs of Propositions 7 and 8 are given in Sections 3.2 and 3.3 of

the Supplementary Material. These propositions state that if the sampling designs $p(\cdot)$ and $p_r(\cdot)$ are close with respect to the Chi-square distance, then $E\left(\hat{Y}_{p\pi} - \hat{Y}_{r\pi}\right)^2$ is smaller than the rate of convergence of $\hat{Y}_{r\pi}$. Consequently, the results in Proposition 6 can be extended to the sampling design $p(\cdot)$, as stated in Proposition 9. Similar coupling arguments are used by Chauvet (2015) to obtain asymptotic results for multistage sampling designs with stratified simple random without replacement sampling at the first stage.

Proposition 9. *Suppose that assumptions (FS1), (SS0)-(SS2), (VAR1)-(VAR2) hold, and that $d_2(p, p_r) \to 0$. Then*

$$\frac{\hat{Y}_{p\pi} - Y}{\sqrt{V(\hat{Y}_{p\pi})}} \quad \longrightarrow_{\mathcal{L}} \quad \mathcal{N}(0,1). \tag{53}$$

If in addition the assumption (SS3) holds, we have

$$E\left[N^{-2} n_I \left|\hat{V}_{HAJ}(\hat{Y}_{p\pi}) - V(\hat{Y}_{p\pi})\right|\right] = o(1) \quad \textit{and} \quad \frac{\hat{V}_{HAJ}(\hat{Y}_{p\pi})}{V(\hat{Y}_{p\pi})} \to_{Pr} 1 \tag{54}$$

The proof is given in Section 3.4 of the Supplementary Material. From Proposition 9, the two-sided $100(1-2\alpha)\%$ confidence interval given in (47) is also asymptotically valid for $\hat{Y}_{p\pi}$.

We now turn back to the choice of the distance function. Let $X(s_I)$ denote some function of a sample s_I. Equation (51) in Proposition 8 is based on the

inequality

$$\begin{aligned}\sum_{s_I \subset U_I} |p(s_I) - p_r(s_I)| X(s_I) &\leq \sqrt{d_2(p, p_r)} \times \sqrt{\sum_{s_I \subset U_I} p_r(s_I) X(s_I)^2} \\ &\leq \sqrt{d_2(p, p_r)} \times \sqrt{E\{X(S_{rI})^2\}}. \qquad (55)\end{aligned}$$

From equation (55) and Proposition 7, $X(S_{rI})$ and $X(S_{pI})$ are asymptotically equivalent if (a) $d_2(p, p_r) \to 0$, and if (b) we can control the second moment of $X(S_{rI})$. This last point may be obtained through standard algebra for rejective sampling, see Lemma 8 for example.

If we rather resort to the Kullback-Leibler divergence

$$d_{KL}(p, p_r) = \sum_{s_I \subset U_I;\ p_r(s_I) > 0} p(s_I) \log \left\{ \frac{p(s_I)}{p_r(s_I)} \right\}, \qquad (56)$$

we can obtain the similar inequality

$$\sum_{s_I \subset U_I} |p(s_I) - p_r(s_I)| X(s_I) \leq \sqrt{d_{KL}(p, p_r)} \times \sqrt{\frac{4}{3} E\{X(S_{rI})^2\} + \frac{2}{3} E\{X(S_{pI})^2\}}.$$

Consequently, we may alternatively demonstrate that $X(S_{rI})$ and $X(S_{pI})$ are asymptotically equivalent if (a') $d_{KL}(p, p_r) \to 0$, if (b) we can control the second moment of $X(S_{rI})$, and if (c) we can control the second moment of $X(S_{pI})$. This last point is difficult to prove for a general sampling design.

5.3 A simplified variance estimator

The variance estimator $\hat{V}_{HAJ}(\hat{Y}_{r\pi})$ proposed in (43) has been proved to be consistent for large entropy sampling designs, with limited assumptions on the first-stage sampling design. However, unbiased and consistent variance estimators are required inside the PSUs, which can be cumbersome for a data user. It is stated in Proposition 10 that the simplified one-term variance estimator $\hat{V}_{HAJ,A}(\hat{Y}_{r\pi})$ is consistent, provided that the third component of the variance in the decomposition (4) is negligible. The proof readily follows from Propositions 6 and 9, and is therefore omitted. Note that the assumption (SS3) providing a lower bound for the second order inclusion probabilities at the second stage is not needed any more.

Proposition 10. *Suppose that assumptions (FS1), (SS0)-(SS2), (VAR1)-(VAR2) hold. Suppose that equation (35) holds. If a rejective sampling design p_r is used at the first-stage, we have*

$$E\left[N^{-2}n_I\left\{\hat{V}_{HAJ,A}(\hat{Y}_{r\pi})-V(\hat{Y}_{r\pi})\right\}\right]^2 \quad = \quad o(1), \tag{57}$$

$$\frac{\hat{V}_{HAJ,A}(\hat{Y}_{r\pi})}{V(\hat{Y}_{r\pi})} \quad \longrightarrow_{Pr} \quad 1. \tag{58}$$

If the first-stage sampling design p is such that $d_2(p,p_r)\to 0$, then

$$E\left[N^{-2}n_I\left|\hat{V}_{HAJ,A}(\hat{Y}_{p\pi})-V(\hat{Y}_{p\pi})\right|\right]=o(1) \quad \textit{and} \quad \frac{\hat{V}_{HAJ,A}(\hat{Y}_{p\pi})}{V(\hat{Y}_{p\pi})}\to_{Pr} 1 \tag{59}$$

6 Simulation study

A simulation study was conducted to evaluate the asymptotic properties of the Hájek-type variance estimators $\hat{V}_{HAJ}(\hat{Y}_\pi)$ and $\hat{V}_{HAJ,A}(\hat{Y}_\pi)$. Three populations U_1, U_2, U_3 of $N_I = 2,000$ PSUs were generated. The number of SSUs per PSU were randomly generated, with mean $N_0 = 40$ and with a coefficient of variation equal to 0, 0.03 and 0.06 for population 1, 2, and 3 respectively. The PSUs are therefore of equal size in the first population.

In each population, a value ν_i was generated for any PSU i from a standard normal distribution. Three variables were generated, for any SSU k inside PSU i, in each population according to the model

$$y_{ikh} = \lambda + \sigma\nu_i + [\rho_h^{-1}(1-\rho_h)]^{0.5}\sigma\varepsilon_k,$$

where $\lambda = 20$, $\sigma = 2$, where ε_k was generated from a standard normal distribution, and ρ_h was such that the intra-cluster correlation coefficient (ICC) was approximately 0.1, 0.2 and 0.3 for $h = 1, 2$ and 3 respectively.

From each population, we repeated $R = 1,000$ times the following two-stage sampling design. A first-stage sample S_I of $n_I = 20, 40, 100$ or 200 PSUs was selected by means of a rejective sampling design, with inclusion probabilities π_{Ii} proportional to the size N_i. A second-stage sample S_i of $n_i = n_0 = 5$ or 10 was selected inside any $i \in S_I$ by simple random sampling without replacement. In each sample, we computed the HT-estimator $\hat{Y}_\pi$ and the Hájek-type variance estimators $\hat{V}_{HAJ,A}(\hat{Y}_\pi)$ and $\hat{V}_{HAJ}(\hat{Y}_\pi)$.

As a measure of bias of a variance estimator $\hat{V}$, we computed the Monte Carlo percent relative bias

$$\mathrm{RB}_{MC}(\hat{V}) = \frac{\frac{1}{R}\sum_{r=1}^{R} \hat{V}^{(r)} - V(\hat{Y}_\pi)}{V(\hat{Y}_\pi)} \times 100,$$

with $\hat{V}^{(r)}$ the value of the estimator in the rth sample, and $\mathrm{V}(\hat{Y}_\pi)$ the exact variance. The Monte Carlo percent relative stability,

$$\mathrm{RS}_{MC}(\hat{V}) = \frac{\left\{\frac{1}{R}\sum_{r=1}^{R} \left[\hat{V}^{(r)} - \mathrm{V}(\hat{Y}_\pi)\right]^2\right\}^{1/2}}{\mathrm{V}(\hat{Y}_\pi)} \times 100,$$

was calculated as a measure of variability of $\hat{V}$. We also calculated the error rates of the normality-based confidence interval given in (47), with nominal one-tailed error rate of 2.5 % in each tail.

The results are presented in Table 1 for the population 3. We observed no qualitative difference with populations 1 and 2, and the results are therefore omitted for conciseness. As expected, the variance estimator $\hat{V}_{HAJ}(\hat{Y}_\pi)$ is almost unbiased in any case, with RB_{MC} lower than 2% in absolute value. The stability RS_{MC} decreases with n_I but not with n_i, as expected. The bias of the simplified variance estimator $\hat{V}_{HAJ,A}(\hat{Y}_\pi)$ is comparable with a small first-stage sampling fraction, but increases with n_I/N_I. Even with the largest sampling fraction, the bias of $\hat{V}_{HAJ,A}(\hat{Y}_\pi)$ is limited and no greater than 7%.

This supports the fact that the term of variance $V_3(\hat{Y}_\pi)$ in the decomposition (23) has a small contribution to the global variance. Both variance estimators perform similarly in terms of stability, with RS_{MC} being slightly larger for $\hat{V}_{HAJ,A}(\hat{Y}_\pi)$ with the largest sampling fraction. The coverage probabilities are well respected in any case, lying between 93% and 95%.

Table 1: Percent relative biases, percent relative stabilities and coverage probabilities of $\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$ and $\widehat{V}_{HAJ}(\widehat{Y}_\pi)$ in population 3

			RB_{MC}		RS_{MC}		CI_{MC}	
ICC	n_I	n_i	$\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ}(\widehat{Y}_\pi)$
0.1	20	5	0.08	0.70	33.58	33.59	0.94	0.94
		10	-0.98	-0.57	31.30	31.30	0.93	0.93
	40	5	-1.00	0.24	21.59	21.56	0.94	0.94
		10	-2.66	-1.84	21.85	21.77	0.93	0.93
	100	5	-3.23	-0.08	14.02	13.64	0.94	0.94
		10	-2.36	-0.27	14.34	14.15	0.95	0.95
	200	5	-6.59	-0.19	11.17	9.03	0.94	0.94
		10	-4.15	0.17	10.42	9.57	0.94	0.95
0.2	20	5	-0.37	0.05	33.13	33.13	0.93	0.93
		10	-0.80	-0.57	32.03	32.02	0.93	0.93
	40	5	-0.82	0.01	22.20	22.18	0.94	0.94
		10	-2.17	-1.71	21.99	21.94	0.93	0.93
	100	5	-2.25	-0.13	14.07	13.89	0.95	0.95
		10	-1.75	-0.56	14.34	14.25	0.94	0.95
	200	5	-4.54	-0.17	10.20	9.14	0.94	0.94
		10	-2.22	0.28	9.96	9.72	0.94	0.94
0.3	20	5	-0.72	-0.43	32.89	32.88	0.94	0.94
		10	-0.69	-0.54	32.39	32.39	0.93	0.93
	40	5	-0.77	-0.19	22.58	22.56	0.94	0.94
		10	-1.85	-1.55	22.02	21.99	0.93	0.93
	100	5	-1.63	-0.14	14.09	14.00	0.95	0.95
		10	-1.44	-0.67	14.29	14.24	0.95	0.95
	200	5	-3.26	-0.16	9.80	9.25	0.95	0.95
		10	-1.29	0.32	9.83	9.75	0.95	0.95

7 Illustration on the panel for urban policy

We consider an application to the Panel for Urban Policy (PUP), which is the original motivation for this work. This is a panel survey in four waves, performed by the French General Secretariat of the Inter-ministerial Committee for Cities (SGCIV) and conducted between 2011 and 2014. The scope of the survey is the collection of various information about security, employment, precariousness, schooling and health, for people living in the Sensitive Urban Zones (ZUS). The initial panel S_I is selected through two-stage sampling, with districts as PSUs and households as SSUs. The individuals in the selected households are comprehensively surveyed.

At the first stage, the population U_I of districts is partitioned into $H = 4$ strata defined according to the progress of the urban renewal program. A stratified sample S_I of $n_I = 40$ districts is selected, with probabilities proportional to the number of main dwellings. The first-stage inclusion probabilities range from 0.04 to 0.67, for a first-stage sampling rate of approximately 0.09. Inside any selected district i, a sample S_i of n_i households is selected with equal probabilities. The sample of households is prone to unit non-response, but this issue is not considered here for the sake of simplicity. In this illustration, the sample of responding households is viewed as the true sample. In summary, the data set is a sample of $1,065$ households obtained by stratified two-stage sampling.

We are interested in four variables related to security, town planning and

residential mobility. The variable y_1 gives the perceived reputation of the district (good, fair, poor, no opinion). The variable y_2 indicates if a member of the household has witnessed trafficking (never, rarely, sometimes, no opinion). The variable y_3 indicates if some significant roadworks have been done in the neighborhood in the twelve last months (yes, no, no opinion). The variable y_4 indicates if the households intends to leave the district during the next twelve next months (certainly/probably, certainly not, probably not, no opinion). For any possible characteristic c of some variable y, we are interested in the proportion

$$p_c = \frac{\sum_{h=1}^{H} \sum_{i=1}^{N_{Ih}} Y_i}{\sum_{h=1}^{H} \sum_{i=1}^{N_{Ih}} N_i} \quad \text{with} \quad Y_i = \sum_{k=1}^{N_i} 1(y_{ik} = c), \tag{60}$$

and where N_{Ih} is the number of PSUs in the stratum h. The proportion p_c is estimated by its substitution estimator

$$\hat{p}_c = \frac{\sum_{h=1}^{H} \sum_{i \in S_{Ih}} \frac{\hat{Y}_i}{\pi_{Ii}}}{\hat{N}_\pi} \quad \text{with} \quad \hat{N}_\pi \equiv \sum_{h=1}^{H} \sum_{i \in S_{Ih}} \sum_{k \in S_i} \frac{1}{\pi_{Ii}\pi_{k|i}}, \tag{61}$$

and where S_{Ih} is the sample of PSUs in the stratum h.

For each proportion, we consider the two variance estimators presented in Section 5. We first compute the linearized variable of p_c, which is

$$e_{ik} \quad = \quad \frac{1}{\hat{N}_\pi}\{1(y_{ik} = c) - \hat{p}_c\}. \tag{62}$$

We then compute the variance estimator in (43) by replacing the variable y_{ik} with e_{ik}, and without truncating the first term of variance for simplicity.

With stratified sampling at the first stage, and since the second-stage samples are selected with equal probabilities, this leads to the variance estimator

$$\begin{aligned} \hat{V}_{HAJ}(\hat{p}_c) &= \hat{V}_{HAJ,A}(\hat{p}_c) + \hat{V}_{HT,B}(\hat{p}_c), \\ \text{with } \hat{V}_{HAJ,A}(\hat{p}_c) &= \sum_{h=1}^{4} \sum_{i \in S_{Ih}} (1 - \pi_{Ii}) \left(\frac{\hat{E}_i}{\pi_{Ii}} - \hat{\hat{R}}_{eh\pi} \right)^2, \\ \text{with } \hat{V}_{HT,B}(\hat{p}_c) &= \sum_{h=1}^{4} \sum_{i \in S_{Ih}} \frac{N_i^2}{\pi_{Ii}} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) s_{ei}^2, \end{aligned} \tag{63}$$

and where

$$\begin{aligned} \hat{\hat{R}}_{eh\pi} &= \frac{\sum_{i \in S_{Ih}} (1 - \pi_{Ii}) \frac{\hat{E}_i}{\pi_{Ii}}}{\sum_{i \in S_{Ih}} (1 - \pi_{Ii})} \quad \text{with} \quad \hat{E}_i = \sum_{k \in S_i} \frac{e_{ik}}{\pi_{k|i}}, \\ s_{ei}^2 &= \frac{1}{n_i - 1} \sum_{k \in S_i} (e_{ik} - \bar{e}_i)^2 \quad \text{with} \quad \bar{e}_i = \frac{1}{n_i} \sum_{k \in S_i} e_{ik}. \end{aligned} \tag{64}$$

The second, simplified variance estimator is $\hat{V}_{HAJ,A}(\hat{p}_c)$, obtained from equation (63) by dropping the second component.

The two variance estimators are then plugged into a normality-based confidence interval, with a nominal one-tailed error rate of 2.5 % . The results are presented in Table 2, and show almost identical performance of both variance estimators.

8 Discussion

In this article, we proposed an asymptotic set-up for the study of two-stage sampling designs. We gave general conditions under which the Horvitz-

Table 2: Substitution estimator of the marginal proportions and normality-based Confidence Intervals (CI) for four variables

	Perceived Reputation of District Status			
	Good	Fair	Poor	No opinion
Estimator $\hat{p}_c$	0.218	0.227	0.527	0.028
CI with $\hat{V}_{HAJ}$	[0.182,0.253]	[0.205,0.250]	[0.485,0.569]	[0.018,0.038]
CI with $\hat{V}_{HAJ,A}$	[0.183,0.252]	[0.206,0.248]	[0.486,0.568]	[0.019,0.038]
	Witnessed trafficking			
	Never	Rarely	Sometimes	No opinion
Estimator $\hat{p}_c$	0.582	0.053	0.163	0.049
CI with $\hat{V}_{HAJ}$	[0.537,0.628]	[0.037,0.068]	[0.135,0.192]	[0.036,0.063]
CI with $\hat{V}_{HAJ,A}$	[0.538,0.627]	[0.038,0.068]	[0.136,0.191]	[0.037,0.062]
	Roadworks in neighborhood			
	Yes	No	No opinion	
Estimator $\hat{p}_c$	0.463	0.503	0.034	
CI with $\hat{V}_{HAJ}$	[0.398,0.528]	[0.434,0.572]	[0.022,0.045]	
CI with $\hat{V}_{HAJ,A}$	[0.399,0.527]	[0.435,0.572]	[0.023,0.044]	
	Intention to leave the district			
	Certainly/Probably	Probably not	Certainly not	No opinion
Estimator $\hat{p}_c$	0.275	0.129	0.562	0.034
CI with $\hat{V}_{HAJ}$	[0.255,0.295]	[0.098,0.159]	[0.531,0.594]	[0.025,0.043]
CI with $\hat{V}_{HAJ,A}$	[0.257,0.292]	[0.099,0.158]	[0.532,0.593]	[0.036,0.042]

Thompson estimator is consistent, and under which usual variance estimators are consistent. In case of large entropy sampling designs at the first stage, we also proved that the Horvitz-Thompson estimator is asymptotically normally distributed and that a truncated Hájek-like variance estimator is consistent. When the first-stage sampling fraction is negligible, simplified variance estimators are also shown to be consistent, under limited assumptions.

Multistage sampling designs are often used at baseline for longitudinal household surveys. If we wish to perform longitudinal estimations, individuals from the initial sample are followed over time. If we also wish to perform cross-sectional estimations at several times, additional samples are selected at further waves and mixed with the individuals originally selected. Even in the simplest case when estimations are produced at baseline with a single sample, variance estimation is challenging due to the different sources of randomness which need to be accounted for: this includes not only the sampling design, but also unit non-response, item non-response and the corresponding statistical treatments. Variance estimation in such more realistic context is an important matter for further investigation.

References

Bertail, P., Chautru, E., and Clémençon, S. (2017). Empirical processes in survey sampling with (conditional) poisson designs. *Scand. J. Stat.*, 44(1):97–111.

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the

bootstrap in stratified sampling. *Ann. Stat.*, pages 470–482.

Boistard, H., Lopuhaä, H. P., Ruiz-Gazen, A., et al. (2012). Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Electron. J. Stat.*, 6:1967–1983.

Boistard, H., Lopuhaä, H. P., Ruiz-Gazen, A., et al. (2017). Functional central limit theorems for single-stage sampling designs. *Ann. Stat.*, 45(4):1728–1758.

Brändén, P. and Jonasson, J. (2012). Negative dependence in sampling. *Scand. J. Stat.*, 39(4):830–838.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regresssion estimators in survey sampling. *Ann. Stat.*, 28(4):1026–1053.

Breidt, F. J. and Opsomer, J. D. (2008). Endogenous post-stratification in surveys: classifying with a sample-fitted model. *Ann. Stat.*, 36(1):403–427.

Breidt, F. J., Opsomer, J. D., and Sanchez-Borrego, I. (2016). Nonparametric variance estimation under fine stratification: An alternative to collapsed strata. *J. Am. Stat. Assoc.*, 111(514):822–833.

Chauvet, G. (2012). On a characterization of ordered pivotal sampling. *Bernoulli*, 18(4):1320–1340.

Chauvet, G. (2015). Coupling methods for multistage sampling. *Ann. Stat.*, 43(6):2484–2506.

Chen, J. and Rao, J. N. K. Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, 17(3):1047–1064.

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons, New York-London-Sydney, third edition. Wiley Series in Probability and Mathematical Statistics.

Conti, P. L. (2014). On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. *Sankhya B*, 76(2):234–259.

Deville, J.-C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85(1):89–101.

Dupakova, J. (1975). A note on rejective sampling. *Contribution to Statistics (J. Hajek memorial volume) Academia Prague*.

Fuller, W. A. (2011). *Sampling statistics*, volume 560. John Wiley & Sons.

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

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *J. Am. Stat. Assoc.*, 77(377):89–96.

Kim, J. K., Park, S., and Lee, Y. (2017). Statistical inference using generalized linear mixed models under informative cluster sampling. *Canadian Journal of Statistics*, 45(4):479–497.

Krewski, D. and Rao, J. N. K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Ann. Stat.*, 9(5):1010–1019.

Ohlsson, E. (1986). Asymptotic normality of the Rao-Hartley-Cochran estimator: an application of the martingale CLT. *Scand. J. Stat.*, 13(1):17–28.

Ohlsson, E. (1989). Asymptotic normality for two-stage sampling from a finite population. *Probab. Theory Relat. Fields*, 81(3):341–352.

Prášková, Z. and Sen, P. K. (2009). Asymptotics in finite population sampling. In *Handbook of Statistics*, volume 29, pages 489–522. Elsevier.

Qualité, L. (2008). A comparison of conditional poisson sampling versus unequal probability sampling with replacement. *J. Stat. Plan. Infer.*, 138(5):1428–1432.

Rao, J. N. K., Hartley, H. O., and Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *J. Royal Stat. Soc. B*, 24:482–491.

Robinson, P. (1982). On the convergence of the horvitz-thompson estimator. *Australian Journal of Statistics*, 24(2):234–238.

Rosén, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement. I, II. *Ann. Stat.*, 43:373–397; ibid. 43 (1972), 748–776.

Sampford, M. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54(3-4):499–513.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics.

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

van Der Hofstad, R. (2016). *Random graphs and complex networks*. Cambridge series in statistical and probabilistic mathematics.