Inference for Two-Stage Sampling

Andrea Boskovic

Department of Statistics, University of Washington Seattle, WA, 98195, USA

Abstract

Two-stage sampling designs are often used in household and health surveys, and the use of the Horvitz-Thompson estimator of the population total is common in such contexts. To produce a statistically valid Horvitz-Thompson estimator with associated confidence intervals, we must prove consistency and asymptotic normality of the estimator along with consistency of its associated variance estimators. These properties have been studied in one-stage sampling designs, but we prove these properties in a general two-stage setting. In addition to generic variance estimators, we also consider simplified variance estimators, which do not require estimating variance within the primary sampling units (PSUs). These are shown to be consistent, although they are only valid when the first-stage sampling fraction is negligible. When large entropy sampling designs are used at the first stage, the Horvitz-Thompson estimator is shown to be asymptotically normal.

1 Introduction

Two-stage sampling designs involve sampling primary units and then sampling from a series of secondary units. These designs are common in health and household surveys, where the population of interest may be limited over a large area and the sampling frame may not exist (Deville and Särndal [1992], Cochran [1977]). In the contexts of these surveys, we are often interested in estimating some population total, and Horvitz-Thompson (HT) estimator is one common approach. This estimator is unbiased and does not rely on model assumptions for the purposes of estimation.

In calculating HT estimators, our goal is to produce a reliable estimate and associated confidence interval. To do so, several properties must hold: the estimator should be consistent for the true total, the estimator should be asymptotically normally distributed, and consistent variance estimators must exist in order to produce valid normality-based confidence intervals.

These properties have been established in the literature for one-stage sampling designs, but there is a dearth of existing literature studying these properties in the context of two-stage sampling. Several papers have established general conditions for the consistency of the HT estimator (Isaki and Fuller [1982]). These properties have also been studied for the class of local polynomial regression estimators as well as under a fine stratification scheme (Breidt and Opsomer [2000], Breidt et al. [2016]). Each of these works focuses on a one-stage sampling design, however.

The asymptotic properties of estimators in two-stage sampling designs are more difficult to study due to the dependence induced by selecting sampling units, but several authors have tackled this problem. In particular, Ohlsson derived a central limit theorem for two-stage sampling procedures as well as a proof of asymptotic normality of estimators for the general case of two-stage sampling (Ohlsson [1989]). More recently, Chauvet proved asymptotic normality of the HT estimator under a multi-stage design using a coupling method, though their results rely on the assumption of simple random sampling in the first stage (Chauvet [2015]). Despite the work done to establish properties of estimators in two-stage sampling design, no work has established that the relevant conditions for reliable estimators hold in this context. This paper provides conditions ensuring that the desired properties outlined above hold in two-stage sampling designs.

In this paper, we study the properties of estimators and their respective variance estimators for the general class of two-stage sampling designs. In Section 2, we discuss the problem setup and decompose the variance of the HT estimator into the sum of three components, which are later used to establish consistency. Next, we lay out the required assumptions for asymptotic properties in Section 3. We then discuss the consistency of estimators in Section 4, namely showing the HT estimator is consistent under various assumptions discussed in Section 3. This section also proves the consistency of two variance estimators. However, these variance estimators require estimating variance withing the primary sampling units (PSUs), so we also define a simplified variance estimator that doesn't require this variance estimation. We prove this simplified variance estimator is also consistent when the total variance within the PSUs is negligible. In Section 5, we study large entropy sampling designs in the first stage. Specifically, we show the consistency of a Hajek-type variance estimator and the HT variance estimator under some assumptions. We also extend these results to a more general class of large entropy sampling designs in the first stage using a coupling argument. Then, we verify the properties from the Hajek-type variance estimator proposed in Section 5 in a simulation study in Section 6. Finally, we apply our methods to data from a panel for urban policy in Section 7, and we summarize our contributions in Section 8.

2 Notation

Define U to be our population of interest, where U is of size N, and suppose we select a sample from U with a two-stage sampling design. The units in U, or secondary sampling units (SSUs), are partitioned into a population U_I of N_I PSUs. Let the sample of n_I PSUs selected in U_I be S_I. Inside each PSU $i \in S_I$, we draw a sample S_i of n_i SSUs in the second stage.

Let y be our variable of interest that we wish to estimate. We are interested in estimating the population total

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

where $Y_i = \sum_{k=1}^{N_i} y_{ik}$ is the subtotal of the variable y over PSU i, and N_i is the number of SSUs within PSU i.

Next, let $\mathbb{I}_{Ii}$ indicate the sample membership in the i-th PSU of sample S_I, let $\pi_{Ii} = \mathbb{P}(i \in S_I) = \mathbb{E}(\mathbb{I}_{Ii})$ be the inclusion probability of PSU i in our sample, and let $\pi_{Iij} = \mathbb{P}(i, j \in S_I) = \mathbb{E}(\mathbb{I}_{Ii}\mathbb{I}_{Ij})$ be the joint inclusion probability of PSUs i and j in our sample.

Define

$$N_0 = \frac{1}{N_I} \sum_{i=1}^{N_I} N_i, \quad n_0 = \frac{1}{N_I} \sum_{i=1}^{N_I} n_i$$

as the the average size of the PSUs and the average sample size inside the PSUs, respectively. This setup covers the case when the SSUs are comprehensively surveyed inside a selected PSU, which is equivalent to single-stage sampling on the population of all PSUs.

For an SSU k in the ith PSU, let $\mathbb{I}_k$ be the indicator that SSU k is in sample S_i, the sample of SSUs. Then, denote $\pi_{k|i} = \mathbb{P}(k \in S_i | i \in S_I) = \mathbb{E}(\mathbb{I}_k | i \in S_I)$, which is the conditional probability that SSU k is chosen, and similarly, define $\pi_{kl|i} = \mathbb{P}(k, l \in S_i | i \in S_I) = \mathbb{E}(\mathbb{I}_k \mathbb{I}_l | i \in S_I)$, which represents the conditional joint probability that SSUs $k, l \in i$ are selected in the sample S_i. Finally, note that we assume invariance of second-stage designs, meaning that the second stage of sampling is independent of sample S_I selected in the first stage (Deville and Särndal [1992]). We also assume that the second-stage designs are independent from one PSU to another conditional on the sample S_I.

Now, define the HT estimator of Y:

$$\hat{Y}_\pi = \sum_{i \in S_I} \frac{\hat{Y}_i}{\pi_{Ii}}, \quad \hat{Y}_i = \sum_{k \in S_i} \frac{y_{ik}}{\pi_{k|i}}. \tag{1}$$

Here, π_{Ii} and $\pi_{k|i}$ are known due and dependent on the survey sampling scheme. Note that we may also be interested in the population mean: $\mu_y = \frac{Y}{N}$. To estimate μ_y, we simply plug in the HT estimator for Y and N, respectively, because N is generally unknown in two-stage surveys. The estimator of the mean is then

$$\hat{\mu}_{y\pi} = \frac{\hat{Y}_\pi}{\hat{N}_\pi},$$

where $\hat{Y}_\pi$ is the HT estimator defined in Equation (1), and $\hat{N}_\pi = \sum_{i \in S_I} \sum_{k \in S_i} \frac{1}{\pi_{Ii}\pi_{k|i}}$.

Then, we can write the variance of the HT estimator $\hat{Y}_\pi$ as follows:

$$\begin{aligned} V(\hat{Y}_\pi) &= \sum_{i=1}^{N_I}\sum_{j=1}^{N_I}(\pi_{Iij} - \pi_{Ii}\pi_{Ij}) \cdot \frac{Y_i}{\pi_{Ii}}\frac{Y_j}{\pi_{Ij}} + \sum_{i=1}^{N_I}\frac{1-\pi_{Ii}}{\pi_{Ii}}V_i + \sum_{i=1}^{N_I} V_i \\ &= V_1(\hat{Y}_\pi) + V_2(\hat{Y}_\pi) + V_3(\hat{Y}_\pi), \end{aligned} \tag{2}$$

where

$$V_i \equiv V(\hat{Y}_i) = \sum_{k=1}^{N_i}\sum_{l=1}^{N_i}(\pi_{kl|i} - \pi_{k|i}\pi_{l|i}) \cdot \frac{y_{ik}}{\pi_{k|i}}\frac{y_{il}}{\pi_{l|i}}.$$

The variables V_i and $V_i(\hat{Y}_\pi)$ represent different quantities: the former represents the variance within the i-th PSU, and the latter represents a component of the decomposed overall variance. Note that $V_1(\hat{Y}_\pi)$ represents the variance due to the first stage of sampling. On the other hand, we can show that the variance due to the second stage of sampling is equivalent to

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

as shown in equation 4.3.11 of Sarndal (Särndal [1992]). We study the HT variance estimator by individually examining the terms $V_1(\hat{Y}_\pi) + V_2(\hat{Y}_\pi)$ and $V_3(\hat{Y}_\pi)$ in Section 4 given the assumptions laid out in Section 3. Large entropy sampling designs allow us to reduce the amount of assumptions made and provide a more simple variance estimator, which is discussed in Section 5.

3 Assumptions

We must outline assumptions in order to study the asymptotic properties of estimators as well as variance estimators. These assumptions are split into three components: assumptions on the first-stage sampling design in Section 3.1, assumptions on the second-stage sampling design in Section 3.2, and assumptions on the variable of interest in Section 3.3. These assumptions are referenced throughout the paper to establish various results.

3.1 First-stage sampling design assumptions

Assumption 1. Some constant $f_{I0} < 1$ exists such that

$$N_I^{-1} n_I \leq f_{I0}. \tag{3}$$

Also, some constants c_{I1}, C_{I1} exist such that for any PSU i, we have that

$$c_{I1} \leq N_I^{-1} n_I \pi_{Ii} \leq C_{I1}. \tag{4}$$

Overall, assumption 1 is under the control of the survey sampler and is related to the order of magnitude of the first-stage sample size n_I along with the first-order inclusion probabilities π_{Ii}. In particular, Equation (3) ensures that the first-stage sample is non-degenerate by ensuring that the total number of PSUs N_I in the population is greater than the amount of PSUs sampled n_I. Previous work in this field has established this condition as well (Breidt et al. [2016], Boistard et al. [2017]). Equation (4) assures that the first-order inclusion probabilities do not differ much from those obtained under a simple random sample.

Assumption 2. Some constants $C_{I2}, C_{I3} > 0$ exist such that for any PSUs $i \neq j \neq i'$:

$$\pi_{Iij} \leq C_{I2} N_I^{-2} n_I^2, \tag{5}$$

$$\pi_{Iiji'} \leq C_{I3} N_I^{-3} n_I^3, \tag{6}$$

where $\pi_{Iiji'}$ is the probability that PSUs i, j, and i' are selected together in the first-stage sample S_I. Some constants C_{I4}, C_{I5} also exist such that

$$\Delta_{I1} = \max_{i \neq j = 1,\ldots,N_I} |\pi_{Iij} - \pi_{Ii}\pi_{Ij}| \leq C_{I4} N_I^{-2} n_I^2, \tag{7}$$
$$\Delta_{I2} = \max_{i \neq j \neq i' \neq j' = 1,\ldots,N_I} |\pi_{Iiji'j'} - \pi_{Ii}\pi_{Ij}\pi_{Ii'}\pi_{Ij'}| \leq C_{I5} N_I^{-4} n_I^3,$$

where $\pi_{Iiji'j'}$ is the probability that i, j, i', and j' are selected together in the first-stage sample S_I.

This assumption involves the second, third, and fourth order inclusion probabilities. Note that Equation (5) and Equation (6) will automatically hold if assumption 1 holds and the sampling design is negatively associated (Brändén and Jonasson [2012]). Such designs include simple random sampling, rejective sampling (Hájek [1964]), Sampford sampling (Sampford [1967]), and pivotal sampling (Deville and Tille [1998]). Equation (7) checks the dependence in the selection of PSUs. If units in the first-stage are selected independently, i.e., Poisson sampling, both Δ_{I1} and Δ_{I2} will be equal to 0 and the conditions will thus be satisfied. They are also satisfied if satisfied for simple random sampling.

Assumption 3. Some constant $c_{I2} > 0$ exists such that for all $i \neq j = 1, \ldots, N_I$, we have that

$$c_{I2} N_I^{-2} n_I^2 \leq \pi_{Iij}. \tag{8}$$

This assumption gives a uniform lower bound on the second-order inclusion probability (Breidt and Opsomer [2000]). Assumption 3 is satisfied under simple random sampling, but it is not trivial to show it holds for unequal probability sampling designs. This assumption is required to prove consistence of the HT variance estimator and the Yates-Grundy variance estimator, so refer to Section 4 or (Breidt and Opsomer [2000]) for details. Since this assumption is difficult to prove, we consider first-stage large entropy designs to use other consistent variance estimators and allow us to avoid checking this condition.

3.2 Second-stage sampling design assumptions

Assumption 4. Some constants $\lambda_1, \Lambda_1 > 0$ and $\phi_1, \Phi_1 > 0$ exist such that for any PSU i the following inequalities are satisfied:

$$\lambda_1 n_0 \leq n_i \leq \Lambda n_0, \tag{9}$$

$$\phi_1 N_0 \leq N_i \leq \Phi_1 N_0. \tag{10}$$

Here, we assume that the sizes N_i of each PSU is comparable and the number of SSUs n_i within each PSU is also comparable. In practice, this is generally satisfied because PSUs are often grouped into strata such that the PSUs in a stratum are of similar sizes. Also, surveys are usually designed so the number of SSUs within each PSU is equal so that each interviewer has the same workload. Thus, Equation (9) and Equation (10) are reasonable in practice.

Assumption 5. Some constants c_1, $C_1 > 0$ exist such taht for any PSU i and any SSU k within PSU i, we have that

$$c_1 \leq N_0 n_0^{-1} \pi_{k|i} \leq C_1. \tag{11}$$

Assumption 6. There exist constants $C_2, C_3 > 0$ such that for any PSU i and any SSU $k \neq l \neq k'$ within PSU i,

$$\pi_{kl|i} \leq C_2 N_0^{-2} n_0^2, \tag{12}$$

$$\pi_{klk'|i} \leq C_3 N_0^{-3} n_0^3, \tag{13}$$

with $\pi_{klk'|i}$ representing the conditional probability that SSUs k, l, and k' are all selected in the second-stage sample S_i. We also require that constants C_4, $C_5 > 0$ exist such that the following two equations are satisfied:

$$\Delta_1 \equiv \max_{i=1,\dots,N_I} \max_{k \neq l=1,\dots,N_i} |\pi_{kl|i} - \pi_{k|i}\pi_{l|i}| \leq C_4 N_0^{-2} n_0^2, \tag{14}$$

$$\Delta_2 \equiv \max_{i=1,\dots,N_I} \max_{k \neq l \neq k' \neq l'=1,\dots,N_i} |\pi_{klk'l'|i} - \pi_{k|i}\pi_{l|i}\pi_{k'|i}\pi_{l'|i}| \leq C_5 N_0^{-4} n_0^3, \tag{15}$$

where $\pi_{klk'l'|i}$ is the conditional probability that SSUs k, l, k', and l' are selected in the second-stage sample S_i.

Assumption 7. Some constant $c_2 > 0$ exists such that for any PSU i and any SSU $k \neq l$ within PSU i,

$$c_2 N_0^{-2} n_0^2 \leq \pi_{kl|i}. \tag{16}$$

Note that assumptions 5-7 are extensions of assumptions 1-3 in Section 3.1 to the second stage of sampling.

3.3 Variable of interest assumptions

Assumption 8. There exist constants M_1, $m_1 > 0$ such that the following holds:

$$N^{-1} \sum_{i=1}^{N_I} \sum_{k=1}^{N_i} y_{ik}^4 \leq M_1, \tag{17}$$

$$m_1 \leq \left| N^{-1} \sum_{i=1}^{N_I} \sum_{k=1}^{N_i} y_{ik} \right|. \tag{18}$$

This assumption assumes a bounded fourth-order moment in the variable of interest and a bounded mean greater than 0.

Assumption 9. Some constant $m_2 > 0$ exists such that

$$m_2 \leq N^{-2} n_I V_1(\hat{Y}_\pi). \tag{19}$$

This assumption states that the first-stage sampling variance $V_1(\hat{Y}_\pi)$ is non-vanishing. Note that these assumptions are all relatively weak, and it's fairly straightforward to find cases where Equation (17), Equation (18), and Equation (19) are not met. In particular, Equation (17) is not satisfied in heavily skewed populations because we have many individuals exhibiting extreme values of our variable of interest. In domain estimation settings where the domain size N_d is negligible compared to the population of interest, Equation (18) and Equation (19) are not satisfied.

4 Consistency of Estimators

First, recall the HT estimator's variance decomposition given in Equation (2). We begin by establishing the order of magnitude of each of these variance components.

Proposition 1. Suppose that first-stage assumptions 1 and 2, second-stage assumptions 4, 5, and 6, and variable assumption 8 hold. Then,

$$\begin{aligned} V_1(\hat{Y}_\pi) &= O(N^2 n_I^{-1}), \\ V_2(\hat{Y}_\pi) &= O(N^2 n_I^{-1} n_0^{-1}), \\ V_3(\hat{Y}_\pi) &= O(N^2 N_I^{-1} n_0^{-1}). \end{aligned} \tag{20}$$

Notice that as $n_0 \to \infty$, $V_1(\hat{Y}_\pi)$ is the leading term, while both $V_2(\hat{Y}_\pi)$ and $V_3(\hat{Y}_\pi)$ go to 0. When n_0 is bounded, we have that $V_1(\hat{Y}_\pi) = V_2(\hat{Y}_\pi) = O(N^2 n_I^{-1})$. In practice, the third variance component is expected to be small.

Next, we establish the consistency of the HT estimator. The proof of the following proposition follows from proposition 1.

Proposition 2. Suppose that first-stage assumptions 1 and 2, second-stage assumptions 4, 5, and 6, and variable assumption 8 hold. Then the HT estimator is design unbiased, i.e., $\mathbb{E}[\hat{Y}_\pi] = Y$. We also have that

$$\begin{aligned} &\mathbb{E}\{N^{-1}(\hat{Y}_\pi - Y)\}^2 = O(n_I^{-1}), \\ &\tfrac{\hat{Y}_\pi}{Y} \xrightarrow{p} 1. \end{aligned} \tag{21}$$

This implies that the HT estimator is $\sqrt{n_I}$ consistent for the true total. This consistency result requires that the sampled number of PSUs, $n_I \to \infty$, but the consistency is not related to the behavior of n_0. That is, even in $n \to \infty$, the HT estimator could be inconsistent if n_I is bounded and thus cannot approach ∞. Our consistency result boils down to the fact that we need a large number of PSUs selected at the first stage in practice.

4.1 Unbiased Variance Estimators

We first consider the canonical two-stage HT variance estimator:

$$\hat{V}_{HT}(\hat{Y}_\pi) = \sum_{i,j \in S_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{\hat{Y}_i}{\pi_{Ii}} \frac{\hat{Y}_j}{\pi_{Ij}} + \sum_{i \in S_i} \frac{\hat{V}_{HT,i}}{\pi_{Ii}} = \hat{V}_{HT,A}(\hat{Y}_\pi) + \hat{V}_{HT,B}(\hat{Y}_\pi), \tag{22}$$

where

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

Proposition 3. Suppose that first-stage assumptions 1-3, second-stage assumptions 4-6, and variable assumption 8 hold. Then,

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

Suppose that first-stage assumptions 1 and 2, second-stage assumptions 4-7, and variable assumption 8 hold. Then,

$$\mathbb{E}\left[N^{-2}N_In_0\left\{\hat{V}_{HT,B}(\hat{Y}_\pi)-V_3(\hat{Y}_\pi)\right\}\right]^2=O(n_I^{-1}). \tag{23}$$

If first-stage assumptions 1-3, second stage assumptions 4-7, and variable assumptions 8 and 9 hold, we have that

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

This proposition shows that $\hat{V}_{HT}(\hat{Y}_\pi)$ is term-by-term unbiased and $\sqrt{n_I}$-consistent because $\hat{V}_{HT,A}(\hat{Y}_\pi)$ is unbiased and $\sqrt{n_I}$-consistent for $V_1(\hat{Y}_\pi)+V_2(\hat{Y}_\pi)$, and $\hat{V}_{HT,B}(\hat{Y}_\pi)$ is unbiased and $\sqrt{n_I}$-consistent for $V_3(\hat{Y}_\pi)$. This result is significant in that consistency of the HT variance estimator is often stated as an assumption in the literature for two-stage sampling designs (Kim et al. [2017]).

Note that if the sampling designs used at both stages are of fixed size, we can may choose to use the Yates-Grundy variance estimator:

$$\hat{V}_{YG}(\hat{Y}_\pi)=-\frac{1}{2}\sum_{i\neq j\in S_I}\frac{\Delta_{Iij}}{\pi_{Iij}}\left(\frac{\hat{Y}_i}{\pi_{Ii}}-\frac{\hat{Y}_j}{\pi_{Ij}}\right)^2+\sum_{i\in S_I}\frac{\hat{V}_{YG,i}}{\pi_{Ii}}=\hat{V}_{YG,A}(\hat{Y}_\pi)+\hat{V}_{YG,B}(\hat{Y}_\pi), \tag{24}$$

where

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

Proposition 4 shows that the Yates-Grundy variance estimator is also term-by-term unbiased and $\sqrt{n_I}$-consistent. The proof of this claim is similar to that of proposition 3.

Proposition 4. If first-stage assumptions 1-3, second-stage assumptions 4-6, and variable assumption 8 hold, we have

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

Suppose that first-stage assumptions 1 and 2, second-stage assumptions 4-7, and variable assumption 8 hold. Then,

$$\mathbb{E}\left[N^{-2}N_In_0\left\{\hat{V}_{YG,B}(\hat{Y}_\pi)-V_3(\hat{Y}_\pi)\right\}\right]^2=O(n_I^{-1}).$$

If first-stage assumptions 1-3, second-stage assumptions 4-7, and variable assumptions 8 and 9 hold, we have that

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

4.2 Simplified One-Term Variance Estimators

Despite our proposed unbiased variance estimators discussed in Section 4.1, these variance estimators are quite difficult to use in practice because they require the consistency of $\hat{V}_{HT,i}$ and $\hat{V}_{YG,i}$, respectively, within any of the randomly selected PSUs. Further, in cases of self-weighted two-stage sampling designs, systematic sampling is often used at the second stage, so assumption 7 is not satisfied, so our variance estimator consistency result does not hold. Self-weighted two-stage sampling is common in practice, and involves of sampling PSUs with π_{Ii} proportional to the size of the PSUs, and a sample of n_0 SSUs inside the PSUs, which results in equal sampling weights for all the SSUs in the population.

To remedy this problem, we can can consider a simplified variance estimator by using $\hat{V}_{HT,A}(\hat{Y}_\pi)$ only, or $\hat{V}_{YG,A}(\hat{Y}_\pi)$ only if we are in a fixed size sampling design setting. Note that the term consistency of these A-term estimators does not rely on assumption 7. These simplified estimators, although they are biased, can be shown to be consistent if $V_3(\hat{Y}_\pi)$ is negligible. Under first-stage assumptions 1-3, second-stage assumptions 4-7, and variable assumptions 8 and 9, a sufficient condition is that the first-stage sampling rate is negligible, i.e., $N_I^{-1}n_I\rightarrow 0$. In practice, however, we expect that even if the first-stage sampling rate is not negligible, we expect that $V_3(\hat{Y}_\pi)$ will still have a small contribution to the overall variance.

Another approach to simplify the variance estimator into one term involves estimating the variance as if the PSUs were sampled with replacement, i.e., via multinomial sampling. This yields the following simplified variance estimator:

$$\hat{V}_{WR}(\hat{Y}_\pi)=\frac{n_I}{n_I-1}\sum_{i\in S_I}\left(\frac{\hat{Y}_i}{\pi_{Ii}}-\frac{\hat{Y}_\pi}{n_I}\right).$$

Note that this estimator is conservative if first-stage sampling is more efficient than multinomial sampling, shown in equation 3.6.15 of Sarndal (Särndal [1992]). We compare this variance estimator in simulations presented in Section 6.

5 Large Entropy Designs

In this section, we narrow our scope to settings where large entropy sampling designs are used in the first stage. In large entropy sampling designs, there is a high amount of variability in the samples that may be chosen. The entropy of a sampling design $p(\cdot)$ is defined as:

$$I(p) = -\sum_{s \in U} p(s) \log p(s),$$

where the sum is taken over the possible samples s on U, the support of $p(\cdot)$.

The authors first consider a Hajek-type variance estimator and show it is consistent with limited assumptions. They also show that the HT estimator is asymptotically normally distributed by building on the work of Ohlsson (Ohlsson [1989]). The large entropy rejective sampling design is first considered in Section 5.1, and then results are extended to a class of large entropy designs in Section 5.2 by using a coupling algorithm. Finally, we study properties of a simplified variance estimator in Section 5.3.

5.1 Rejective Sampling Design

Rejective, or conditional Poisson, sampling was introduced by Hajek (Hájek [1964]). Rejective sampling in U_I consists of using Poisson sampling, where we use independent Bernoulli trials to determine whether or not to include PSU i in the sample S_I, but we condition on the requirement that the sample has fixed size n_I. We can think of this as a rejection procedure where independent samples are drawn sequentially until we achieve a sample of size n_I. Note that the inclusion probabilities in rejective sampling are chosen so that the required inclusion probabilities π_{Ii} are respected. It has been shown that sampling is the largest unequal probability sampling design, while simple random sampling is the largest equal probability sampling design (Särndal [1992]).

We let $p_r(\cdot)$ denote the rejective sampling design with inclusion probabilities π_{Ii} within the population of first-stage units U_I. Let S_{rI} denote a first-stage sample selected via rejective sampling, i.e., with $p_r(\cdot)$, and define

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

as the associated HT estimator in the rejective sampling framework. Hajek proposed a variance estimator that uses a uniform approximation of second-order inclusion probabilities, thereby omitting them from the variance calculation. In the two-stage sampling context, this leads to replacing

$\hat{V}_{HT,A}(\hat{Y}_{\pi})$ in Equation (22) with

$$\hat{V}_{HAJ,A}(\hat{Y}_{r\pi}) = \begin{cases} \sum_{i \in S_{rI}} (1-\pi_{Ii}) \left(\frac{\hat{Y}_i}{\pi_{Ii}} - \hat{\hat{R}}_{r\pi} \right)^2 & \text{if } \hat{d}_{rI} \geq \frac{c_{I0}}{2} n_I, \\ 0 & \text{otherwise}, \end{cases} \tag{25}$$

where

$$\hat{\hat{R}}_{r\pi} = \hat{d}_{rI}^{-1} \sum_{i \in S_{rI}} (1-\pi_{Ii}) \frac{\hat{Y}_i}{\pi_{Ii}} \quad \text{and} \quad \hat{d}_{rI} = \sum_{i \in S_{rI}} (1-\pi_{Ii}), \tag{26}$$

and c_{I0} is defined in Lemma 7 of the Appendix. Our global variance estimator then becomes

$$\hat{V}_{HAJ}(\hat{Y}_{r\pi}) = \hat{V}_{HAJ,A}(\hat{Y}_{r\pi}) + \hat{V}_{HT,B}(\hat{Y}_{r\pi}), \tag{27}$$

where $\hat{V}_{HT,B}(\hat{Y}_{r\pi})$ is defined in Equation (22). Note that this quantity can be replaced with $\hat{V}_{YG,B}(\hat{Y}_{r\pi})$ if the second-stage sampling designs are of fixed size.

Note that the variance estimator $\hat{V}_{HAJ}(\hat{Y}_{r\pi})$ is truncated, i.e., takes value 0 when $\hat{d}_{rI} < \frac{c_{I0}}{2} n_I$, to avoid extreme values for $\hat{\hat{R}}_{r\pi}$. This is needed to establish consistency, which is shown in proposition 5. This estimator does not require first-stage, second-order inclusion probabilities, meaning we do not need first-stage assumption 3 to prove consistency.

Proposition 5. Suppose that a rejective sampling design is used at the first stage. Suppose that first-stage assumption 1, second-stage assumptions 4-6, and assumption 8 hold. Then,

$$E\left[N^{-2} n_I \left\{ \hat{V}_{HAJ,A}(\hat{Y}_{r\pi}) - V_1(\hat{Y}_{r\pi}) - V_2(\hat{Y}_{r\pi}) \right\}\right]^2 = o(1). \tag{28}$$

If variable assumption 9 also holds, we then have that

$$\frac{\hat{Y}_{r\pi} - Y}{\sqrt{V(\hat{Y}_{r\pi})}} \xrightarrow{d} \mathcal{N}(0,1). \tag{29}$$

If we also have that second-stage assumption 7 holds, then

$$E\left[N^{-2} n_I \left\{ \hat{V}_{HAJ}(\hat{Y}_{r\pi}) - V(\hat{Y}_{r\pi}) \right\}\right]^2 = o(1) \text{ and } \frac{\hat{V}_{HAJ}(\hat{Y}_{r\pi})}{V(\hat{Y}_{r\pi})} \rightarrow_{Pr} 1. \tag{30}$$

Asymptotic normality of the HT estimator has been proven by Hajek for a single-stage rejective sampling design, but this results establishes asymptotic normality for a two-stage rejective sampling design (Hájek [1964]). Further, consistency of the Hajek variance estimator has not been established previously, so this proposition is powerful in even one-stage rejective sampling designs. Finally, notice that we can write a two-sided $100(1-2\alpha)\%$ confidence interval for Y in this setting with

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

where $u_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution.

5.2 Other Sampling Designs

Now, we consider a more general class of high entropy sampling designs at the first-stage, which are close to rejective sampling with respect to the χ^2-distance. Other distance functions have been considred in the literature, such as the Hellinger distance and the total variation distance. Let $p(\cdot)$ be a fixed size sampling design with inclusion probabilities π_{Ii} in the population U_I. We have that $p(\cdot)$ is close to the rejective sampling design $p_r(\cdot)$ with respect to the χ^2-distance if

$$d_2(p, p_r) \to 0 \quad \text{where} \quad d_2(p, p_r) = \sum_{s_I \subset U_I;\ p_r(s_I)>0} \frac{\{p(s_I) - p_r(s_I)\}^2}{p_r(s_I)}. \tag{31}$$

Section 5.2 holds for many designs, such as the Rao-Sampford sampling design. Let S_{pI} be a first-stage sample selected by means of $p(\cdot)$. The associated HT estimator is then

$$\hat{Y}_{p\pi} \quad = \quad \sum_{i \in S_{pI}} \frac{\hat{Y}_i}{\pi_{Ii}}. \tag{32}$$

We introduce a coupling procedure in Algorithm 1 to obtain the estimators $\hat{Y}_{p\pi}$ and $\hat{Y}_{r\pi}$ jointly in order to be able to extend the results from proposition 5 to our more general large entropy design estimate $\hat{Y}_{p\pi}$. Define

$$\alpha = 1 - d_{TV}(p, p_r) \quad \text{where} \quad d_{TV}(p, p_r) = \frac{1}{2} \sum_{s_I \in U_I} |p(s_I) - p_r(s_I)|, \tag{33}$$

where $d_{TV}(p, p_r)$ is the total variation distance between $p(\cdot)$ and $p_r(\cdot)$. We can then show that algorithm 1 leads to estimators $\hat{Y}_{p\pi}$ and $\hat{Y}_{r\pi}$ associated with the required two-stage sampling designs.

Proposition 6. Suppose that S_{rI} and S_{pI} are selected via the coupling procedure described in Algorithm 1. Then,

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

Proposition 7. Suppose that the samples S_{rI} and S_{pI} are selected via the coupling procedure described in Algorithm 1. Suppose that first-stage assumption 1, second-stage assumptions 4-6, and variable assumption 8 hold. Further assume that $d_2(p, p_r) \to 0$. Then,

$$E\left(\hat{Y}_{p\pi} - \hat{Y}_{r\pi}\right)^2 \quad = \quad o\left(N^2 n_I^{-1}\right), \tag{34}$$

and if in addition variable assumption 9 holds, we have that

$$\frac{V\left(\hat{Y}_{p\pi}\right)}{V\left(\hat{Y}_{r\pi}\right)} \quad \to \quad 1. \tag{35}$$

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 $\frac{p(s_I) \wedge p_r(s_I)}{\alpha}$, 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 $\frac{p(s_I) - p_r(s_I)}{1-\alpha}$ 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 $\frac{p_r(s_I) - p(s_I)}{1-\alpha}$ 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 $\hat{Y}_{r\pi}$.
```

Propositions 6 and 7 state that, assuming $p(\cdot)$ and $p_r(\cdot)$ are close with respect to the χ^2-distance, then $E\left(\hat{Y}_{p\pi} - \hat{Y}_{r\pi}\right)^2$ is smaller than the rate of convergence of $\hat{Y}_{r\pi}$.

The results from proposition 5 can be extended to the sampling design $p(\cdot)$, as stated in proposition 8. The coupling method is a powerful tool to study the asymptotic behavior of estimators under complex sampling designs, and this tool is used to obtain asymptotic results for multistage sampling designs with stratified simple random sampling without replacement at the first stage (Chauvet [2015]).

Proposition 8. Suppose that first-stage assumption 1, second-stage assumptions 4-6, and variable assumptions 8 and 9 hold. Also suppose that $d_2(p, p_r) \to 0$. Then,

$$\frac{\hat{Y}_{p\pi} - Y}{\sqrt{V(\hat{Y}_{p\pi})}} \quad \overset{d}{\to} \quad \mathcal{N}(0,1). \tag{36}$$

If in addition second-stage assumption 7 holds, we have that

$$E\left[N^{-2} n_I \left|\hat{V}_{HAJ}(\hat{Y}_{p\pi}) - V(\hat{Y}_{p\pi})\right|\right] = o(1) \quad \text{and} \quad \frac{\hat{V}_{HAJ}(\hat{Y}_{p\pi})}{V(\hat{Y}_{p\pi})} \overset{p}{\to} 1. \tag{37}$$

The two-sided $100(1-2\alpha)\%$ confidence interval defined previously is also asymptotically valid for $\hat{Y}_{p\pi}$, so we have

$$\left[\hat{Y}_{p\pi} \pm u_{1-\alpha}\{\hat{V}_{HAJ}(\hat{Y}_{p\pi})\}^{0.5}\right].$$

We now return to our choice of distance function. Let $X(s_I)$ denote some function of a sample s_I. Section 5.2 in proposition 7 is based on the inequality

$$\begin{aligned} \sum_{s_I \subset U_I} |p(s_I) - p_r(s_I)| X(s_I) &\leq \sqrt{d_2(p, p_r)} \times \sqrt{\sum_{s_I \subset U_I} p_r(s_I) X(s_I)^2} \\ &\leq \sqrt{d_2(p, p_r)} \times \sqrt{E\{X(S_{rI})^2\}}. \end{aligned} \quad (38)$$

From Section 5.2 and proposition 6, we have the $X(S_{rI})$ and $X(S_{pI})$ are asymptotically equivalent if $d_2(p, p_r) \to 0$ and if we can control the second moment of $X(S_{rI})$.

If we would rather use the Kullback-Leibler divergence,

$$d_{KL}(p, p_r) = \sum_{s_I \subset U_I;\ p_r(s_I)>0} p(s_I) \log\left\{\frac{p(s_I)}{p_r(s_I)}\right\}, \quad (39)$$

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 can demonstrate that $X(S_{rI})$ and $X(S_{pI})$ are asymptotically equivalent if $d_{KL}(p, p_r) \to 0$, if we can control the second moment of $X(S_{rI})$, and if we can control the second moment of $X(S_{pI})$, although the third condition is difficult to prove for a general sampling design.

5.3 A Simplified Variance Estimator

The variance estimator $\hat{V}_{HAJ}(\hat{Y}_{r\pi})$ that was proposed in Equation (22) has been proved to be consistent for large entropy sampling designs with limited assumptions on the first-stage sampling design. However, the estimator requires unbiased and consistent variance estimators within the PSUs, which are difficult to obtain in practice. Recall that in proposition 9, we state 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 term in the decomposition is negligible. The proof follows from propositions 5 and 8. Note that the second-stage assumption of 7 providing a lower bound for the second-order inclusion probabilities in the second stage is no longer needed.

Proposition 9. Suppose that first-stage assumption 1, second-stage assumptions 4-5, and variable assumptions 8 and 9 hold. Further suppose that $\frac{V_3(\hat{Y}_\pi)}{V_1(\hat{Y}_\pi)+V_2(\hat{Y}_\pi)} \to 0$. If a rejective sampling design $p_r(\cdot)$ is used at the first stage, we have

$$E\left[N^{-2} n_I \left\{\hat{V}_{HAJ,A}(\hat{Y}_{r\pi}) - V(\hat{Y}_{r\pi})\right\}\right]^2 = o(1), \quad (40)$$

$$\frac{\hat{V}_{HAJ,A}(\hat{Y}_{r\pi})}{V(\hat{Y}_{r\pi})} \longrightarrow_{Pr} 1. \quad (41)$$

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

$$E\left[N^{-2} n_I \left|\hat{V}_{HAJ,A}(\hat{Y}_{p\pi}) - V(\hat{Y}_{p\pi})\right|\right] = o(1) \quad \text{and} \quad \frac{\hat{V}_{HAJ,A}(\hat{Y}_{p\pi})}{V(\hat{Y}_{p\pi})} \to_{Pr} 1. \tag{42}$$

6 Simulation Study

We conduct a simulation study to evaluate the asymptotic properties of the Hajek-type variance estimators $\hat{Y}_{HAJ}(\hat{Y}_\pi)$, $\hat{Y}_{HAJ,A}(\hat{Y}_\pi)$, and the with-replacement variance estimator $\hat{Y}_{WR}(\hat{Y}_\pi)$. We generate three populations U_1, U_2, and U_3 of $N_I = 2000$ PSUs. 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.

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 $\lambda = 20$ and $\sigma = 2$, ε_k follows a standard normal distribution, and ρ_h is such that the intracluster correlation coefficient ICC was approximately 0.1, 0.2, and 0.3 for $h = 1$, 2, and 3 respectively.

From each population ,we repeated the following two stage sampling design $R = 1000$ times. A first-stage sample S_I of $n_I = 20$, 40, 100, and 200 PSUs was selected via 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 $n_i = n_0 = 10$ 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 Hajek-type variance estimators $\hat{V}_{HAJ,A}(\hat{Y}_\pi)$, $\hat{V}_{HAJ}(\hat{Y}_\pi)$, and $\hat{V}_{WR}(\hat{Y}_\pi)$.

We use the Monte Carlo percentage relative bias as a measure of bias of a variance estimator $\hat{V}$:

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

where $\hat{V}^{(r)}$ is the value of the estimator in the r^{th} sample, and $V(\hat{Y}_\pi)$ is the true, exact variance of the estimator $\hat{Y}_\pi$. The Monte Carlo percentage relative stability was calculated as a measure of variability of the variance estimator $\hat{V}$, and is defined as follows:

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

We also calculate the error rates of the normality-based confidence interval with a nominal one-tailed error rate of 2.5% in each tail.

The results displayed in Table 1 match those of the original paper very closely, and we see that all estimators perform similarly in terms of 95% confidence interval coverage. However, we see the relative bias of the first-term estimator and with-replacement variance estimator are much higher than that of the overall Hajek variance estimator, which aligns with our expectations because the former estimators are biased by construction. In particular, we see that the relative bias of the with-replacement variance estimator grows with the sample size. We also see that the relative stability of the three estimators are very similar, and in particular, we see that the relative stability for each estimator decreases with n_I but not with n_i, as expected.

7 Application to Urban Policy

We consider an application of this work to the Panel for Urban Policy. This is a panel survey performed by the French General Secretariat of the Inter-ministerial Committee for Cities conducted in four waves between 2011 and 2014. The goal of the survey is to collect information about security, employment, housing conditions, schooling, and health for people living in urban zones. The initial panel S_I was selected through two-stage sampling, where districts are PSUs and households are SSUs. The individuals in the selected households are comprehensively surveyed.

In the first stage, the population of districts U_I 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. Within each stratum, the sample is selected via the Hanurav-Vijayan sampling procedure.

Inside any selected district i, a sample S_i of n_i households is selected with equal probabilities from a sampling frame built by gathering five rotation groups of the French census. This sample of households is prone to unit non-response, but we ignore this issue in this application for simplicity. In other words, we treat the sample of responding households as the true sample, and we treat it as it were selected by means of simple random sampling without replacement. We obtain a data set containing a sample of 1,065 households obtained by stratified two-stage sampling.

We are interested in four categorical variables related to security, town planning, and residential mobility. The variable y_1 gives the perceived reputation of the district (good, fair, poor, or no opinion), y_2 indicates whether a member of the household has witnessed trafficking (never, rarely, sometimes, or no opinion), y_3 indicates whether significant road work has been done in the neighborhood in the last twelve months (yes, no, or no opinion), and y_4 represents whether the household intends to leave the district during the next 12 months (certainly or probably, certainly not, probably not, or no opinion). For each category c of variable y, we are interested in the proportion:

$$p_c = \frac{\sum_{h=1}^{H} \sum_{i=1}^{N_{Ih}} Y_i}{\sum_{h=1}^{H} \sum_{i=1}^{N_{Ih}} N_i} \quad \text{with} \quad Y_i = \sum_{k=1}^{N_i} 1(y_{ik} = c), \tag{43}$$

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

$$\hat{p}_c = \frac{\sum_{h=1}^{H} \sum_{i \in S_{Ih}} \frac{\hat{Y}_i}{\pi_{Ii}}}{\hat{N}_\pi} \quad \text{with} \quad \hat{N}_\pi \equiv \sum_{h=1}^{H} \sum_{i \in S_{Ih}} \sum_{k \in S_i} \frac{1}{\pi_{Ii}\pi_{k|i}}, \tag{44}$$

where S_{Ih} is the sample of PSUs in stratum h. For each proportion, we consider the two variance estimators presented in Section 5, namely V_{HAJ} and $V_{HAJ,A}$. For each proportion $\hat{p}_c$, we compute the linearized variable of p_c, which is

$$e_{ik} = \frac{1}{\hat{N}_k}\{\mathbb{I}(y_{ik} = c) - \hat{p}_c\}.$$

We then compute the variance estimator by replacing the variable y_{ik} with e_{ik}, and without truncating the first term of the variance for simplicity. With stratified sampling at the first stage, and since the second stage samples are selected with equal probabilities, we get the following variance estimator:

$$\hat{V}_{HAJ}(\hat{p}_c) = \hat{V}_{HAJ,A}(\hat{p}_c) + \hat{V}_{HT,B}(\hat{p}_c), \tag{45}$$

with

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

and where

$$\begin{aligned} \hat{\hat{R}}_{eh\pi} &= \frac{\sum_{i \in S_{Ih}} (1 - \pi_{Ii}) \frac{\hat{E}_i}{\pi_{Ii}}}{\sum_{i \in S_{Ih}} (1 - \pi_{Ii})} \quad \text{with} \quad \hat{E}_i = \sum_{k \in S_i} \frac{e_{ik}}{\pi_{k|i}}, \\ s_{ei}^2 &= \frac{1}{n_i - 1} \sum_{k \in S_i} (e_{ik} - \bar{e}_i)^2 \quad \text{with} \quad \bar{e}_i = \frac{1}{n_i} \sum_{k \in S_i} e_{ik}. \end{aligned} \tag{46}$$

Note that the simplified variance estimator $V_{HAJ,A}(\hat{p}_c)$ can be obtained from Section 7 by dropping the second term, $\hat{V}_{HT,B}(\hat{p}_c)$. Each of 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 shown in Table 4 show nearly identical performance of both variance estimators.

8 Discussion

In this paper, we propose an asymptotic set-up to study two-stage sampling designs. We outline general conditions for which the HT estimator and its usual variance estimators are consistent. Under large entropy sampling designs at the first stage, we also proved that the HT estimator is asymptotically normally distributed and that a truncated Hajek-like variance estimator is consistent. When the first-stage sampling fraction is negligible, simplified variance estimators are also shown to be consisted under limited assumptions.

Multistage sampling designs are often used in longitudinal household surveys, but such estimation in such a setting poses difficulty in terms of variance estimation. Even in the simplest case when estimates are produced at baseline with a single sample, variance estimation is challenging due to the different sources of randomness that we must account for, including that of the sampling design, unit non-response, item non-response, and the corresponding statistical treatments. In these more realistic contexts, variance estimation is an important topic for further study.

References

Hélène Boistard, Hendrik P Lopuhaä, and Anne Ruiz-Gazen. Functional central limit theorems for single-stage sampling designs. 2017.

Petter Brändén and Johan Jonasson. Negative dependence in sampling. *Scandinavian Journal of Statistics*, 39(4):830–838, 2012.

F Jay Breidt and Jean D Opsomer. Local polynomial regression estimators in survey sampling. *Annals of statistics*, pages 1026–1053, 2000.

F Jay Breidt, Jean D Opsomer, and Ismael Sanchez-Borrego. Nonparametric variance estimation under fine stratification: an alternative to collapsed strata. *Journal of the American Statistical Association*, 111(514):822–833, 2016.

Guillaume Chauvet. Coupling methods for multistage sampling. 2015.

William G Cochran. *Sampling techniques*. John Wiley & Sons, 1977.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

Jean-Claude Deville and Yves Tille. Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85(1):89–101, 1998.

Jaroslav Hájek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.

Cary T Isaki and Wayne A Fuller. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96, 1982.

Jae Kwang Kim, Seunghwan Park, and Youngjo Lee. Statistical inference using generalized linear mixed models under informative cluster sampling. *Canadian Journal of Statistics*, 45(4):479–497, 2017.

Esbjörn Ohlsson. Asymptotic normality for two-stage sampling from a finite population. *Probability theory and related fields*, 81(3):341–352, 1989.

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

Carl-Erik Särndal. Methods for estimating the precision of survey estimates when imputation has been used. *Survey methodology*, 18(2):241–252, 1992.

A Moment Inequalities

Lemma 1. Suppose that first-stage assumption 1 and 2 hold. Then, there exists some constant C'_{I5} such that

$$\max_{i\neq j\neq i'\neq j'=1,\ldots,N_I} |\Delta_{Iij,i'j'}| \leq C'_{I5} N_I^{-4} n_I^3,$$

where $\Delta_{Iij,i',j'} = Cov(I_{Ii}I_{Ij}, I_{Ii'}I_{Ij'})$.

Lemma 2. Suppose that second stage assumption 5 and 6 hold. There exists some constant C_5' such that

$$\max_{i\in U_I}\max_{k\neq l\neq k'\neq l'=1,\ldots,N_i}|\Delta_{kl,k'l'|i}|\leq C_5' N_0^{-4}n_0^3,$$

where $\Delta_{kl,k'l'|i}=Cov(I_kI_l,I_{k'}I_{l'}|i\in S_I)$.

Lemma 3. Under second stage assumption 4 and and variable assumption 8, there exists M such that

$$N_I^{-1}\sum_{i=1}^{N_I}Y_i^4\leq MN_0^4.$$

Lemma 4. Under second-stage assumptions 4-6 and variable assumption 8, there exists some constant M_2 such that

$$N_I^{-1}\sum_{i=1}^{N_I}V_i^2\leq M_2N_0^4n_0^{-2}.$$

Lemma 5. Under second-stage assumptions 4-6 and variable assumption 8, there exists some constant M_3 such that

$$N_I^{-1}\sum_{i=1}^{N_I}V(\hat{Y}_i^2)\leq M_3N_0^4n_0^{-1}.$$

Lemma 6. Under second-stage assumptions 4-7 and variable assumption 8, there exists some constant M_4 such that

$$N_I^{-1}\sum_{i=1}^{N_I}V(\hat{V}_i^2)\leq M_4N_0^4n_0^{-3}.$$

B Proof of Proposition 5

B.1 Preliminary Lemmas

We state the following Lemmas needed to prove proposition 5.

Lemma 7. Suppose that first stage assumption 1 holds. Then there exists c_{I0} such that

$$c_{I0}n_I\leq\sum_{i=1}^{N_I}\pi_{Ii}(1-\pi_{Ii})\equiv d_I.$$

Lemma 8. Suppose that first stage assumption 1, second stage assumptions 4-6, and variable assumption 8 hold. Then,

$$\sum_{i=1}^{N_I}\frac{\mathbb{E}[\hat{Y}_i-Y_i]^4}{\pi_{Ii}^3}=O(\frac{N^4}{n_I^3}).$$

Lemma 9. Suppose that a rejective sampling design is used at the first stage. Suppose that first stage assumption 1, second stage assumption 4-6, and variable assumption 8 hold. Then,

$$\mathbb{E}[N^{-2}n_I\{\hat{V}_{HAJ,A}(\hat{Y}_\pi) - \tilde{V}_{HAJ,A}(\hat{Y}_\pi)\}1(\Omega_r)]^2 = O(n_I^{-1}),$$

where

$$\tilde{V}_{HAJ,A}(\hat{Y}_\pi) = \sum_{i=1}^{N_I} \pi_{Ii}(1-\pi_{Ii})\left(\frac{\hat{Y}_i}{\pi_{Ii}} - \hat{R}\right)^2,$$

$$\hat{R} = d_I^{-1}\sum_{i=1}^{N_I} \pi_{Ii}(1-\pi_{Ii})\frac{\hat{Y}_i}{\pi_{Ii}},$$

and with $\Omega_r = \{\hat{d}_{rI} \geq \frac{c_{I0}}{2}n_I\}$.

Lemma 10. Suppose that a rejective sampling design is used at the first stage. Further suppose that first stage assumption 1, second stage assumption 4-6, and variable assumption 8 hold. Then we have that

$$\mathbb{E}[N^{-2}n_I\{\tilde{V}_{HAJ,A}(\hat{Y}_\pi)1(\Omega_r) - V_1(\hat{Y}_\pi) - V_2(\hat{Y}_\pi)\}]^2 = o(1).$$

B.2 Proof of Proposition 5

The proof of follows from Lemmas 9 and 10, and the the proof of follows from and , so to complete the proof of Proposition 5, we must simply prove the asymptotic normality property. Based on Theorem 2.1 in Ohlsson, it suffices to show conditions C1, C2, and 2.8 hold (Ohlsson [1989]). In the context of our problem, this amounts to showing that

$$\frac{\tilde{Y}_\pi - Y}{\sqrt{V(\tilde{Y}_\pi)}} \xrightarrow{d} \mathcal{N}(0,1), \tag{47}$$

$$\frac{\sum_{i=1}^{N_I} \frac{E(\hat{Y}_i - Y_i)^4}{\pi_{Ii}^3}}{\{V(\hat{Y}_\pi)\}^2} \rightarrow 0, \tag{48}$$

$$\pi_{Iij} - \pi_{Ii}\pi_{Ij} \leq 0, \tag{49}$$

where $\tilde{Y_\pi} = \sum_{i \in S_I} \frac{Y_i}{\pi_{Ii}}$. Equation 48 represents the Yates-Grundy conditions, which is a property of rejective sampling designs, so it is satisfied in the context of our problem. Further, Equation 49 follows from Lemma 8 and variable assumption 9. Thus, it remains to show that $\tilde{Y_\pi}$ is asymptotically normally distributed. Note that we can show that

$$\sum_{i=1}^{N_I} \frac{(Y_i - R\pi_{Ii})^4}{\pi_{Ii}^3} = O\left(\frac{N^4}{n_I^3}\right).$$

Also, a fundamental theorem by Hajek gives us that

$$\sum_{i=1}^{N_I} (Y_i - R\pi_{Ii})^2 (\frac{1}{\pi_{Ii}} - 1) = V(\tilde{Y}_\pi)\{1 + o(1)\}.$$

Substituting these results into the Lyapunov condition,

$$\frac{\sum_{i=1}^{N_I} \frac{(Y_i - R\pi_{Ii})^4}{\pi_{Ii}^3}}{\left\{\sum_{i=1}^{N_I} (Y_i - R\pi_{Ii})^2 (\frac{1}{\pi_{Ii}} - 1)\right\}^2} \to 0,$$

we get the desired result. Therefore, $\tilde{Y_\pi}$ is asymptotically normal.

C Results from Original Paper

In this appendix, we present the results from the original paper.

C.1 Simulation Study

The original simulation study results are shown in Table 3.

C.2 Real Data Application

The results originally produced by Chauvet and Vallee in the data application are given in Table 4.

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

			RB_{MC}			RS_{MC}			CI_{MC}		
ICC	n_I	n_i	$\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ}(\widehat{Y}_\pi)$	$\widehat{V}_{WR}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ}(\widehat{Y}_\pi)$	$\widehat{V}_{WR}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ}(\widehat{Y}_\pi)$	$\widehat{V}_{WR}(\widehat{Y}_\pi)$
0.1	20	5	-0.85	0.03	0.27	32.86	33.01	32.85	0.93	0.93	0.93
		10	-0.97	-1.04	-0.58	32.97	33.12	33.17	0.94	0.95	0.94
	40	5	-2.45	-0.87	-0.49	22.59	22.28	22.37	0.95	0.95	0.95
		10	-1.28	0.53	1.83	23.14	23.02	23.85	0.94	0.94	0.95
	100	5	-3.63	-0.50	1.47	13.88	13.39	14.18	0.95	0.95	0.95
		10	-1.67	0.42	3.90	14.23	14.51	15.73	0.95	0.95	0.95
	200	5	-5.92	0.48	4.88	11.36	9.74	11.10	0.95	0.95	0.95
		10	-4.74	-0.12	6.11	10.09	9.49	12.57	0.94	0.95	0.95
0.2	20	5	-0.32	-0.13	0.63	32.49	32.40	32.58	0.94	0.94	0.94
		10	-1.03	-0.84	-0.23	33.64	32.81	33.05	0.95	0.94	0.94
	40	5	-1.68	-0.99	0.44	22.29	22.65	22.71	0.95	0.94	0.95
		10	0.10	0.74	2.31	22.84	22.82	23.34	0.95	0.95	0.95
	100	5	-2.78	-0.82	2.52	13.43	13.85	14.73	0.95	0.95	0.95
		10	-1.05	0.35	4.24	13.90	14.13	15.82	0.95	0.95	0.95
	200	5	-4.25	0.28	7.12	10.34	9.89	12.70	0.95	0.95	0.95
		10	-2.63	0.12	8.42	9.92	9.63	13.23	0.94	0.94	0.95
0.3	20	5	0.26	0.18	0.67	31.83	32.10	32.44	0.94	0.95	0.94
		10	-1.05	-1.17	-0.52	32.53	32.72	33.11	0.93	0.93	0.94
	40	5	-1.52	-0.84	0.56	22.28	22.43	22.57	0.95	0.95	0.95
		10	0.05	0.41	1.89	22.93	22.67	22.72	0.95	0.95	0.95
	100	5	-2.01	-0.38	3.12	14.05	13.57	14.49	0.95	0.95	0.96
		10	-0.12	0.64	5.38	14.52	14.61	16.92	0.95	0.95	0.96
	200	5	-2.52	0.26	7.53	10.12	9.96	13.36	0.95	0.95	0.96
		10	-1.97	0.09	9.25	9.54	9.48	14.54	0.95	0.95	0.95

Table 2: Substitution estimator of the marginal proportions and normality-based confidence intervals (CI) for four variables of interest.

	Perceived Reputation of District Status			
	Good	Fair	Poor	No opinion
Estimator $\hat{p}_c$	0.218	0.227	0.527	0.028
CI with $\hat{V}_{HAJ}$	[0.179,0.256]	[0.201,0.254]	[0.478,0.575]	[0.015,0.041]
CI with $\hat{V}_{HAJ,A}$	[0.180,0.255]	[0.202,0.254]	[0.479,0.574]	[0.016,0.041]

	Witnessed trafficking			
	Never	Rarely	Sometimes	No opinion
Estimator $\hat{p}_c$	0.582	0.053	0.163	0.049
CI with $\hat{V}_{HAJ}$	[0.525,0.640]	[0.036,0.069]	[0.132,0.195]	[0.028,0.071]
CI with $\hat{V}_{HAJ,A}$	[0.525,0.640]	[0.037,0.068]	[0.133,0.194]	[0.029,0.070]

	Roadworks in neighborhood			
	Yes	No	No opinion	
Estimator $\hat{p}_c$	0.463	0.503	0.034	
CI with $\hat{V}_{HAJ}$	[0.394,0.532]	[0.430,0.577]	[0.021,0.046]	
CI with $\hat{V}_{HAJ,A}$	[0.394,0.532]	[0.431,0.576]	[0.022,0.045]	

	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 $\hat{V}_{HAJ}$	[0.252,0.298]	[0.095,0.163]	[0.524,0.601]	[0.022,0.046]
CI with $\hat{V}_{HAJ,A}$	[0.254,0.296]	[0.095,0.162]	[0.525,0.600]	[0.023,0.045]

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

			RB_{MC}			RS_{MC}			CI_{MC}		
ICC	n_I	n_i	$\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ}(\widehat{Y}_\pi)$	$\widehat{V}_{WR}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ}(\widehat{Y}_\pi)$	$\widehat{V}_{WR}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ}(\widehat{Y}_\pi)$	$\widehat{V}_{WR}(\widehat{Y}_\pi)$
0.1	20	5	-0.66	-0.05	0.34	32.64	32.63	32.96	0.93	0.93	0.93
		10	-1.30	-0.89	-0.30	33.01	33.00	33.33	0.94	0.94	0.94
	40	5	-2.36	-1.14	-0.36	22.42	22.33	22.76	0.95	0.95	0.95
		10	-0.06	0.75	1.99	23.07	23.09	23.63	0.94	0.94	0.95
	100	5	-3.63	-0.50	1.47	13.88	13.39	14.18	0.94	0.95	0.95
		10	-1.49	0.60	3.73	14.68	14.62	15.83	0.94	0.95	0.95
	200	5	-6.03	0.37	4.48	11.23	9.48	11.45	0.95	0.95	0.96
		10	-4.34	-0.05	6.35	10.50	9.55	12.38	0.94	0.94	0.95
0.2	20	5	-0.44	-0.02	0.57	32.43	32.42	32.76	0.94	0.94	0.94
		10	-1.18	-0.95	-0.18	32.70	32.70	33.02	0.94	0.94	0.94
	40	5	-1.73	-0.90	0.29	22.47	22.42	22.87	0.95	0.95	0.95
		10	0.06	0.53	2.11	22.93	22.93	23.50	0.95	0.95	0.95
	100	5	-2.62	-0.49	2.54	13.79	13.54	14.48	0.95	0.95	0.95
		10	-0.50	0.71	4.78	14.53	14.54	16.03	0.94	0.95	0.95
	200	5	-4.10	0.30	6.63	10.56	9.74	12.69	0.95	0.95	0.95
		10	-2.55	-0.04	8.36	9.79	9.45	13.43	0.94	0.94	0.94
0.3	20	5	-0.34	-0.05	0.67	32.36	32.36	32.70	0.94	0.94	0.94
		10	-1.18	-1.03	-0.17	32.47	32.47	32.78	0.93	0.93	0.94
	40	5	-1.30	-0.71	0.73	22.47	22.45	22.91	0.95	0.95	0.95
		10	0.06	0.37	2.12	22.81	22.82	22.38	0.95	0.95	0.95
	100	5	-1.97	-0.46	3.23	13.81	13.67	14.75	0.95	0.95	0.96
		10	-0.05	0.73	5.25	14.48	14.50	16.13	0.95	0.95	0.96
	200	5	-2.89	0.26	7.99	10.25	9.84	13.55	0.95	0.95	0.96
		10	-1.67	-0.03	9.35	9.55	9.40	14.03	0.94	0.94	0.95

Table 4: Substitution estimator of the marginal proportions and normality-based Confidence Intervals (CI) for four variables of interest.

	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]