Consistency of the Horvitz–Thompson estimator under general sampling and experimental designs

Angèle Delevoye* angele.delevoye@yale.edu

Fredrik Sävje† fredrik.savje@yale.edu

January 7, 2020

Abstract

We extend current concentration results for the Horvitz–Thompson estimator in finite population settings. The estimator is demonstrated to converge in quadratic mean to its target under weaker and more general conditions than previously known. Specifically, we do not require that the variables of interest nor the normalized inclusion probabilities are bounded. Rates of convergence are provided.

*Department of Political Science, Yale University.

†Department of Political Science & Department of Statistics and Data Science, Yale University.

1 Introduction

The quantity of interest is the average of some characteristic y_i in a finite population of N units indexed by $i \in \{1, \ldots, N\}$:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

We observe y_i only for a subset $\mathbf{S}$ of the units, and the task is to estimate μ based on this information.

Narain (1951) and Horvitz & Thompson (1952) provide an estimator of μ when the subset of observed units is random. The estimator is conventionally named after the second set of authors, and we will not depart from that convention here. At the core of the Horvitz–Thompson estimator is the probability distribution of $\mathbf{S}$ over the power set of the unit indices, which is said to be the design of the study. Given a design, the estimator is

$$\hat{\mu} = \frac{1}{N} \sum_{i \in \mathbf{S}} \frac{y_i}{\pi_i},$$

where $\pi_i = \Pr(i \in \mathbf{S})$ is the inclusion probability for unit i. These probabilities are taken to be known in this note, but they may sometimes be estimated. Examples of such settings include estimation of response propensities for unit non-response in survey sampling and estimation of propensity scores for unknown assignment mechanisms in causal inference.

The application of the estimator to questions in survey sampling is immediate. The characteristics are survey responses, and $\mathbf{S}$ collects the sampled units. Its application to causal questions is also straightforward. The characteristics are in this case potential outcomes given by treatments assigned to the units (Neyman, 1923; Holland, 1986). For example, if $y_i(1)$ denotes unit i's outcome when assigned to active treatment and $y_i(0)$ denotes the outcome when assigned to control treatment, the average treatment effect can be written as $\mu_1 - \mu_0$ where μ_1 and μ_0 are the population averages of $y_i(1)$ and $y_i(0)$. The inferential challenge is that no more than one potential outcome is observed for any of the units. The other outcomes are counterfactual, and at least one (but generally both) of μ_1 and μ_0 are unobserved even when the complete population is sampled. In this case,

S collects all sampled units assigned to a certain treatment condition, and the Horvitz–Thompson estimator provides an estimate of the average of the corresponding potential outcome. The contrast between two such estimators is an estimate of the average treatment effect.

The purpose of this note is to investigate some of the asymptotic properties of the Horvitz–Thompson estimator. Our particular focus is to extend current concentration results for general designs. We show that the estimator is consistent under weaker and more general conditions than previously known. The estimator has inspired a large class of estimators providing improvements in various directions. In an appendix, we show that our results extend to some of these improved estimators as well.

2 Consistency of the Horvitz–Thompson estimator

2.1 Sampling inferences

We consider an asymptotic regime in which the number of units in the population grows without limit, $N \to \infty$. All quantities depending on the population, including the design and the quantity of interest, are therefore sequences indexed by N. The index will, however, be suppressed in the following discussion as it eases the exposition without confusion.

A disadvantage of the regime is that the rates of convergence are expressed with respect to the population size rather than the sample size. The sample and the population are, however, connected through the design. If $\bar{\pi} = N^{-1} \sum_{i=1}^{N} \pi_i$ denotes the average inclusion probability, then the sample size $|\mathbf{S}|$ is related to the population as $\mathrm{E}\big[|\mathbf{S}|\big] = \bar{\pi} N$. We require that $\bar{\pi} > 0$, so that $\mathbf{S}$ is not always empty, but we allow $\bar{\pi} \to 0$.

It will prove useful to rewrite the estimator slightly. Let $S_i = \mathbb{1}[i \in \mathbf{S}]$ be an indicator taking the value one when unit i is in the sample, and zero otherwise. The estimator can now be written as a sum over the population:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} S_i w_i y_i,$$

where $w_i = 1/\pi_i$ when $\pi_i > 0$ and $w_i = 0$ otherwise. The definition of w_i ensures that the estimator is well-defined also when some units have no chance of being sampled.

We will compare inclusion probabilities between units, so a normalization will expedite the discussion. Let $\tilde{\pi}_i = \pi_i/\bar{\pi}$ be the inclusion probability of unit i normalized by the average probability. If $\tilde{\pi}_i$ is greater than one, unit i is disproportionately likely to be sampled. If $\tilde{\pi}_i = 1$ for all units, the design is uniform in the sense that all units are equally likely to be sampled. Similarly, let $\tilde{w}_i = \bar{\pi} w_i$ be the normalized version of w_i. That is, $\tilde{w}_i = 1/\tilde{\pi}_i$ when $\pi_i > 0$ and $\tilde{w}_i = 0$ otherwise.

Condition 1. There exist $p > 2$ and $q > 1$ with $pq \geq p + 2q$ such that the pth population moment of the variables of interest and the qth population moment of the sampling weights are bounded:

$$\left[\frac{1}{N}\sum_{i=1}^{N}|y_i|^p\right]^{1/p} \leq k_y, \qquad \text{and} \qquad \left[\frac{1}{N}\sum_{i=1}^{N}\tilde{w}_i^q\right]^{1/q} \leq k_\pi.$$

The constants, p, q, k_y and k_π, are fixed throughout the asymptotic sequence.

The moment conditions provide control over the variable of interest and the design. The first part ensures that units with very large values of y_i are rare in the population. The condition does not rule out that the variable grows indefinitely for some units, but such units must be a diminishing fraction of the population. In particular, the condition will fail if a fraction of the population, diminishing at a sufficiently slow rate, has $y_i \to \infty$ at a sufficiently fast rate.

The second part of the condition ensures that the design does not sample some units disproportionately infrequently. The normalized inclusion probability is small if a unit is sampled rarely compared to the remaining units in the population, making $\tilde{w}_i = 1/\tilde{\pi}_i$ large. The condition thus requires that the inclusion probabilities do not deviate too far from their average. The condition will fail if a sufficiently large fraction of the population has $\pi_i \to 0$ at a sufficiently faster rate than $\bar{\pi} \to 0$.

The two moment conditions are connected because $pq \geq p + 2q$. This means that more control over the variables of interest allow us to have less control over the design. The

appropriate allocation of control will depend on the application at hand. However, it is common to assume that the fourth moment of variables of interest is bounded, $p = 4$, and this setting may provide some intuition. Condition 1 here dictates that $q \geq 2$, which means that the second moment of $\tilde{w}_i$ is bounded. An example of such a design is when $\pi_i = 1/4$ for all units except for $\lfloor N^{0.5} \rfloor$ units with $\pi_i = 1/4N^{0.25}$.

Together with the quantity defined next, Condition 1 provides a bound on the variance of the estimator.

Definition 1. Let $\tilde{\pi}_{i|j} = \Pr(i \in \mathbf{S} \mid j \in \mathbf{S})/\bar{\pi}$ be the normalized inclusion probability of unit i conditional on that unit j is sampled. If $\pi_j = 0$, then set $\tilde{\pi}_{i|j} = \tilde{\pi}_i$ to capture that a unit that is never sampled provides no information about whether another unit is sampled. The *average design dependence* is

$$D(r) = \left[\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i} \left| \tilde{\pi}_{i|j} - \tilde{\pi}_i \right|^{1/r} \right]^r.$$

Lemma 1. *Under Condition 1,* $\operatorname{Var}(\hat{\mu}) \leq k_y^2 k_\pi / \bar{\pi} N + k_y^2 k_\pi D(1 - 1/p - 1/q)$.

Proof. The sampling indicators are the only random variables, so

$$\operatorname{Var}(\hat{\mu}) = \frac{1}{N^2} \sum_{i=1}^{N} w_i^2 y_i^2 \operatorname{Var}(S_i) + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i} w_i w_j y_i y_j \operatorname{Cov}(S_i, S_j).$$

Focusing on the first term, observe that $\operatorname{Var}(S_i) = \pi_i(1 - \pi_i)$. It follows that $w_i^2 \operatorname{Var}(S_i) = w_i(1 - \pi_i) \leq w_i$ when $\pi_i > 0$, and $w_i^2 \operatorname{Var}(S_i) = 0 = w_i$ when $\pi_i = 0$. Use Hölder's inequality to separate the variables of interest from the normalized sampling weights:

$$\frac{1}{N^2} \sum_{i=1}^{N} w_i y_i^2 = \frac{1}{\bar{\pi} N^2} \sum_{i=1}^{N} \tilde{w}_i y_i^2 \leq \frac{1}{\bar{\pi} N^2} \left[\sum_{i=1}^{N} |y_i|^p \right]^{2/p} \left[\sum_{i=1}^{N} \tilde{w}_i^{p/(p-2)} \right]^{(p-2)/p},$$

where $p/2$ and $p/(p-2)$ are Hölder conjugates. Because the reciprocals of the conjugates sum to one, $N = N^{2/p} N^{(p-2)/p}$, and the bound can be written as

$$\frac{1}{\bar{\pi} N} \left[\frac{1}{N} \sum_{i=1}^{N} |y_i|^p \right]^{2/p} \left[\frac{1}{N} \sum_{i=1}^{N} \tilde{w}_i^{p/(p-2)} \right]^{(p-2)/p}.$$

The second factor is bounded by k_y^2 and needs no further attention. For the third factor, observe that Condition 1 implies $q \geq p/(p-2)$, and apply Jensen's inequality to get

$$\left[\frac{1}{N}\sum_{i=1}^{N} \tilde{w}_i^{p/(p-2)}\right]^{(p-2)/p} \leq \left[\frac{1}{N}\sum_{i=1}^{N} \tilde{w}_i^{q}\right]^{1/q} \leq k_\pi.$$

The first term of the expression for $\mathrm{Var}(\hat{\mu})$ is thus bounded by $k_y^2 k_\pi/\bar{\pi}N$.

We take care of the second term in a similar way. First, observe that

$$w_i w_j \operatorname{Cov}(S_i, S_j) = \frac{\Pr(i \in \mathbf{S}, j \in \mathbf{S}) - \pi_i\pi_j}{\pi_i\pi_j} = \frac{\tilde{\pi}_{i|j} - \tilde{\pi}_i}{\tilde{\pi}_i},$$

when $\pi_i > 0$ and $\pi_j > 0$. When either probability is zero, $\operatorname{Cov}(S_i, S_j) = 0$. Hence, we have $w_i w_j \operatorname{Cov}(S_i, S_j) = \tilde{w}_i\big(\tilde{\pi}_{i|j} - \tilde{\pi}_i\big)$ for all π_i and π_j. Let $r = 1 - 1/p - 1/q$ and apply Hölder's inequality with conjugates p, q and $1/r$ to get

$$\sum_{i=1}^{N}\sum_{j\neq i} y_i y_j \tilde{w}_i\big(\tilde{\pi}_{i|j} - \tilde{\pi}_i\big) \leq \left[\sum_{i=1}^{N}\sum_{j\neq i} |y_i y_j|^p\right]^{1/p} \left[\sum_{i=1}^{N}\sum_{j\neq i} \tilde{w}_i^q\right]^{1/q} \left[\sum_{i=1}^{N}\sum_{j\neq i} \big|\tilde{\pi}_{i|j} - \tilde{\pi}_i\big|^{1/r}\right]^{r}.$$

Factor N^2 as $N^{2/p}N^{2/q}N^{2r}$, so $1/N^2$ can be distributed into the three factors. Using Condition 1, bound the first factor as

$$\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i} |y_i y_j|^p\right]^{1/p} \leq \left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} |y_i y_j|^p\right]^{1/p} = \left[\frac{1}{N}\sum_{i=1}^{N} |y_i|^p\right]^{2/p} \leq k_y^2,$$

and, in a similar fashion, bound the second factor by k_π. The third factor is equal to the average design dependence, $D(r)$, with $r = 1 - 1/p - 1/q$. □

The first term in the bound in Lemma 1 relates the variance to the expected sample size: $\mathrm{E}\big[|\mathbf{S}|\big] = \bar{\pi}N$. Disregarding the second term, the variance would diminish at the conventional linear rate in the size of the sample. The term makes clear that the estimator may not concentrate unless the design is such that $\bar{\pi}N \to \infty$.

The second term captures sampling dependence between pairs of units. The difference $\Pr(i \in \mathbf{S} \mid j \in \mathbf{S}) - \Pr(i \in \mathbf{S})$ measures the knowledge we gain about whether unit i is sampled when we know that unit j is sampled. The more it deviates from zero, the more knowledge we gain. We get $\tilde{\pi}_{i|j} - \tilde{\pi}_i$ when this difference is normalized to be on the same

scale as the normalized inclusion probabilities. The average design dependence, $D(r)$, is the average of the normalized difference over all pairs of units, providing a measure of the overall dependence introduced by the design. If $r = 1$, the average is the ordinary arithmetic mean. If $r < 1$, the average emphasizes larger dependencies.

Condition 2 (Weak design dependence)**.** $D(1 - 1/p - 1/q) \to 0$.

We want the average design dependence to be low because this provides more control over the variance. In particular, if the quantity diminishes, as captured in Condition 2, then the dependencies are sufficiently weak to ensure that the effective sample size grows with the nominal one. It may be instructive to consider the rates at which the dependence diminishes for conventional designs. First, consider a Bernoulli or Poisson design, in which case the units are sampled independently. Here, $\tilde{\pi}_{i|j} = \tilde{\pi}_i$ and $D(r) = 0$. Next, consider when a simple random sample is taken, so $|\mathbf{S}| = \bar{\pi}N$ is fixed and the design is otherwise uniform. In this case, $\tilde{\pi}_{i|j} - \tilde{\pi}_i = (1 - \bar{\pi})/\bar{\pi}(N - 1)$ and $D(r) = \mathcal{O}(1/\bar{\pi}N)$. Finally, consider when the units are sampled in clusters. Partition $\{1, \ldots, N\}$ into $N_c = \lfloor N/k \rfloor$ disjoint groups of size k, where we set $N = kN_c$ throughout the asymptotic sequence for convenience. Sample the groups independently, each with probability $\bar{\pi}$. Here, $\tilde{\pi}_{i|j} = 1/\bar{\pi}$ if i and j are in the same group, and $\tilde{\pi}_{i|j} = \tilde{\pi}_i$ otherwise. It follows that $D(r) = \mathcal{O}(1/\bar{\pi}N_c^r)$. For designs with strong dependencies, such as cluster sampling with a fixed number of clusters and various systematic sampling design, the average design dependence may be fixed, indicating that the effective sample size does not grow with N.

Lemma 1 together with Condition 2 ensure that the sampling distribution of the estimator concentrates, but they do not control where it does so. The concern is that we may never observe some units. These units would affect the population mean but not the estimator. In the worst case, when $\pi_i = 0$ for all units, $\mathrm{Var}(\hat{\mu}) = 0$, but the estimator is constant at zero no matter the value of μ. Unless one makes structural assumptions allowing for extrapolation, the only way to control the point of convergence is to ensure that a sufficiently small fraction of the population is excluded by the design so to not noticeably affect μ.

Definition 2. Let $e_i = \mathbb{1}[\pi_i = 0]$, so that $e_i = 1$ if unit i is excluded by design and $e_i = 0$ otherwise. Let $\bar{e} = N^{-1} \sum_{i=1}^{N} e_i$ denote the share of excluded units in the population.

Lemma 2. *Under Condition 1,* $\big|\mu - \mathrm{E}[\hat{\mu}]\big| \leq 2k_y \bar{e}^{1-1/p}$.

Proof. It is only units with $e_i = 1$ that could shift the location of the sampling distribution away from the population mean:

$$\mu - \mathrm{E}[\hat{\mu}] = \frac{1}{N} \sum_{i=1}^{N} \big[y_i - \pi_i w_i y_i\big] = \frac{1}{N} \sum_{i=1}^{N} e_i y_i,$$

where the last equality follows from $\pi_i w_i = 1 - e_i$. To gain control over this quantity, let $a_i = \mathbb{1}\big[|y_i| > c\big]$ where $c = k_y / \bar{e}^{1/p}$, except $a_i = 0$ when $\bar{e} = 0$. Write the bias as

$$\big|\mu - \mathrm{E}[\hat{\mu}]\big| \leq \frac{1}{N} \sum_{i=1}^{N} e_i |y_i| = \frac{1}{N} \sum_{i=1}^{N} a_i e_i |y_i| + \frac{1}{N} \sum_{i=1}^{N} (1 - a_i) e_i |y_i|.$$

For the first term, observe that $c^{p-1} a_i |y_i| \leq |y_i|^p$ because Condition 1 ensures that $p > 2$. It follows that

$$\frac{1}{N} \sum_{i=1}^{N} a_i e_i |y_i| \leq \frac{c^{1-p}}{N} \sum_{i=1}^{N} |y_i|^p \leq c^{1-p} k_y^p.$$

For the second term, observe that $(1 - a_i)|y_i| \leq c$, so

$$\frac{1}{N} \sum_{i=1}^{N} (1 - a_i) e_i |y_i| \leq \frac{c}{N} \sum_{i=1}^{N} e_i = c\bar{e}.$$

Recall that $c = k_y / \bar{e}^{1/p}$, so $c^{1-p} k_y^p$ and $c\bar{e}$ are both equal to $k_y \bar{e}^{1-1/p}$. □

One can prove the lemma using Hölder's inequality in a way similar to the proof of Lemma 1, but the more elementary proof presented here demonstrates the underlying idea better.

The two lemmas provide control over the location and width of the sampling distribution. Together, they provide control over the deviation from the population mean, and we are ready for the main result.

Proposition 1. *Under Condition 1, the mean square error is bounded as*

$$\mathrm{E}\big[(\hat{\mu} - \mu)^2\big] \leq 4k_y^2 \bar{e}^{2-2/p} + k_y^2 k_\pi / \bar{\pi} N + k_y^2 k_\pi D(1 - 1/p - 1/q).$$

Proof. Decompose the mean square error into the squared bias and variance, and apply the bounds from Lemmas 1 and 2. □

Corollary 1. *The Horvitz–Thompson estimator converges in quadratic mean to the population mean if* $\bar{e} \to 0$ *and* $\bar{\pi} N \to \infty$ *in addition to Conditions 1 and 2.*

2.2 Causal inferences

The literature on causal inference has a long-standing interest in the Horvitz–Thompson estimator, where it sometimes is called the inverse probability weighted estimator. This literature tends to focus on an asymptotic regime in which the sample is assumed to consist of independently and identically distributed observations from an infinite super-population. Some of the nuances of the design-based perspective are lost with this regime. Focus has instead been on estimated inclusion probabilities and the estimator's efficiency (see, for example, Robins & Ritov, 1997; Hahn, 1998; Hirano et al., 2003). There are, however, exceptions. A contribution of particular note is Aronow & Middleton (2013). The authors study finite sample properties of the Horvitz–Thompson estimator for causal quantities in a design-based framework. The discussion in this section complements their analysis with large sample results.

Let $\mathbf{Z}$ be a set of treatment conditions. In the standard setting, as described in the introduction, $\mathbf{Z} = \{0, 1\}$, but $\mathbf{Z}$ may be any countable set. A potential outcome $y_i(a)$ is the response of unit i when assigned to treatment $a \in \mathbf{Z}$. The notation presumes that the responses are unambiguous for each treatment condition, ruling out, for example, that the treatment assigned to one unit affects the response of another unit. The quantity of interest is the contrast between the average of the potential outcomes for two treatment conditions a and b:

$$\tau(a, b) = \frac{1}{N} \sum_{i=1}^{N} y_i(a) - \frac{1}{N} \sum_{i=1}^{N} y_i(b).$$

We can approach the estimation of $\tau(a, b)$ as two separate estimation exercises of the type considered in the previous section. In the first exercise, the population characteristic

is $y_i(a)$, and the sample, denoted by $\mathbf{S}_a$, collects all units assigned to condition a, for which we observe $y_i(a)$. The design is the probability distribution of $\mathbf{S}_a$. The second sampling exercise is the analog with b substituted for a.

The estimator of $\tau(a, b)$ is the difference between the two Horvitz–Thompson estimators for the corresponding two sampling exercises:

$$\hat{\tau}(a,b) = \frac{1}{N}\sum_{i\in\mathbf{S}_a}\frac{y_i(a)}{\Pr(i\in\mathbf{S}_a)} - \frac{1}{N}\sum_{i\in\mathbf{S}_b}\frac{y_i(b)}{\Pr(i\in\mathbf{S}_b)}.$$

Corollary 2. *If Condition 1 holds with respect to the design and potential outcomes for both a and b, the mean square error is bounded as*

$$\mathrm{E}\Big[\big(\hat{\tau}(a,b)-\tau(a,b)\big)^2\Big] \leq 16k_y^2\bar{e}_{ab}^{2-2/p} + 4k_y^2k_\pi/\bar{\pi}_{ab}N + 4k_y^2k_\pi D_{ab}(1-1/p-1/q),$$

where $\bar{\pi}_{ab}$ is the minimum of the corresponding quantity for the designs of $\mathbf{S}_a$ and $\mathbf{S}_b$, and $\bar{e}_{ab}$ and $D_{ab}(1-1/p-1/q)$ are the corresponding maximums.

The corollary follows directly from Proposition 1 because it provides control over the constitute terms. In particular, use Young's inequality for products to write the square estimation error as

$$\mathrm{E}\Big[\big(\hat{\tau}(a,b)-\tau(a,b)\big)^2\Big] = \mathrm{E}\Big[\big((\hat{\mu}_a-\mu_a)-(\hat{\mu}_b-\mu_b)\big)^2\Big] \leq 2\,\mathrm{E}\big[(\hat{\mu}_a-\mu_a)^2\big] + 2\,\mathrm{E}\big[(\hat{\mu}_b-\mu_b)^2\big],$$

where μ_z and $\hat{\mu}_z$ are the terms of the estimand and estimator, respectively. The difference compared to the sampling setting is that there are two designs here, and the mean squared error is governed by the design of the least favorable treatment group. The intuition is otherwise unchanged, and the analog of Corollary 1 applies.

3 Concluding remarks

3.1 Previous results

To the best of our knowledge, Robinson (1982) provides the most comprehensive concentration results for the Horvitz–Thompson estimator. He proves consistency under two different sets of conditions. The first proof requires that the variables of interest are bounded

throughout the imagined asymptotic sequence. The rate of convergence is governed by the average of the reciprocal of the units' first order inclusion probabilities and the average of the difference between second and first order probabilities. The second proof requires only that the second population moment of the variables of interest is bounded, but the rate of convergence is now governed by the minimum and maximum of the quantities that were averaged in the first proof. It is in this case necessary for the minimum first order inclusion probability to be of the same order as the average inclusion probability. In both proofs, the sample size is assumed to be of fixed, so $|\mathbf{S}| = \bar{\pi}N$ with probability one.

The two settings studied by Robinson are important, but they may be too restrictive for some investigations. On the one hand, the requirement of bounded variables of interest in the first proof is a poor approximation in studies of characteristics with heavy-tailed distributions, such as wealth or income. On the other hand, the restriction on the inclusion probabilities in the second proof is a poor approximation in studies with skewed designs. As we discuss in the following section, this is a particular concern when drawing causal inferences because naturally occurring designs tend to disproportionately favor some units for certain treatments.

The concentration result presented here addresses these concerns by considering consistency of the Horvitz–Thompson estimator under more general conditions. Our result can be seen as a generalization of the two results in Robinson (1982). In particular, Robinson's results can be reproduced, in spirit, as special cases of Proposition 1. The condition in Robinson's first proof corresponds to Condition 1 when we let $p \to \infty$, so the moment is the uniform norm of the variables of interest. This allows $q \to 1$ as in Robinson's first proof. Robinson's second proof corresponds to Condition 1 when one lets $q \to \infty$, which allows $p \to 2$. Proposition 1 demonstrates that the estimator is consistent also in the intermediate cases, where neither quantity is bounded, as long as the trade-off described in Condition 1 is respected.

The results in Robinson (1982) have been extended in various other directions before us. Robinson & Särndal (1983) use techniques similar to Robinson's second proof to prove

consistency for a larger class of design-based estimators. Opsomer & Breidt (2000), Breidt & Opsomer (2008) and Cardot et al. (2010) further build on this work.

In this strand of the literature, Chauvet (2014) provides the most recent concentration results for the Horvitz–Thompson estimator. His initial conditions are similar to the ones in the papers just mentioned, requiring that $\tilde{\pi}_i$ is uniformly bounded away from zero and that $|\tilde{\pi}_{i|j} - \tilde{\pi}_i|$ uniformly diminishes at a faster rate than $1/N$. Chauvet continues by showing that the second condition can be relaxed if the variables of interest are known to be non-negative.

The approach used by Isaki & Fuller (1982) differs from both approaches in Robinson (1982). Similar to this note, they use Hölder's inequality, which avoids the two extremes Robinson considers. However, unlike Robinson and this note, Isaki & Fuller impose a composite condition on the design and variables of interest together, which may be hard to reason about. From this perspective, our contribution can be seen as merging these different approaches, taking advantage of the benefits each has to offer.

3.2 Practical implications[1]

In light of the previous literature, the results in this note are relevant primarily in settings that are ill-behaved in some way. As mentioned above, this could be when the distribution of the variable of interest is heavy-tailed or when the design is skewed. Practitioners should, if possible, avoid such extreme settings. This is because the estimator may be inefficient compared to a more well-behaved setting, even if it is still consistent. Practitioners will therefore find our results most helpful when an ill-behaved setting cannot be avoided, and in particular when they have limited control over the design.

One such setting is when the design includes trade-offs between statistical and practical concerns. For example, the United States Census Bureau conducts the American Community Survey on a monthly basis to complement its decennial census. Unlike the census, the survey excludes some rural areas in Alaska from the sampling frame because they are too

[1]We thank an anonymous reviewer for suggesting the inclusion of this section.

difficult to access (Torrieri, 2014, Section 4.8). Other areas are excluded for only part of the year. These communities are in the population of interest, but they have disproportionately low, or no, probability of being sampled.

Another example is a study by Niccolai et al. (2010) who sought to estimate the HIV incidence among injection drug users in St. Petersburg, Russia. The authors could not freely select the inclusion probabilities in this setting. One reason, among others, is that a non-negligible portion of the population was homeless, without a reliable way to contact them. To reach as many people as possible, the authors used a respondent driven sampling design where drug users already sampled were asked to recruit other drug users to the study. A consequence of this procedure was that well-connected drug users were more likely to be sampled than isolated users. Indeed, a drug user who did not know any other drug users was only observed when included in the initial sample, which was very small because of the cost of its construction.

These concerns are common also in causal inference. Unlike the sampling setting, the assignment of treatments can have life-changing consequences for the participants. Ethical considerations are therefore a common component in experimental design. For example, Harvey et al. (2005) studies the effect of pulmonary artery catheters on critically ill patients. This type of catheter is a monitoring device of heart function that gets physically inserted into the pulmonary artery. It provides valuable information to the patient's physicians, but the insertion can lead to serious complications, including the death of the patient. Statistical efficiency may suggest to uniformly randomize among the patients, but such a design could have grave consequences. Instead, to minimize the number of complications, the design used in the study conditioned the randomization on physicians' assessments of the risk for the patient. As a consequence, patients with a high risk of complications had a low, or no, chance of being assigned to get a catheter inserted. The design is, thus, skewed with respect to the full population, and the authors chose to instead focus on a subpopulation for which the design was uniform.

The problem of skewed designs is particularly common in field experiments. Rather

than being confined to a lab, field experiments are conducted in real-world settings, allowing practitioners to study the effect of various treatment in their natural environment. The field setting, however, makes the implementation strenuous and costly, and it is common to embed these experiments in the normal operations of companies, governments or other organizations. This means that the partner organization often has the final say about the design, and these organizations tend to give little weight to statistical efficiency. An example is the study by Karlan & Zinman (2009) of the effect of access to consumer credit. The authors partnered with a company offering micro-credit loans in South Africa. They convinced the company to introduce a random component in their credit scoring model, making a customer applying for a loan more or less likely to be approved. The company was, however, not prepared to forgo good customers, nor did they want to give loans to people who were likely to default on their payments. The randomness thus mainly affected the chance of approval for "marginal" customers: those that was just below or above the required credit score. As a consequence, the only design the company would accept was one that gave a very high probability of approval to good customers and very low probability to poor customers. As in the previous example, the design is skewed with respect to the full population, and as before, the authors chose to restrict their focus to the subpopulation of marginal customers.

The problem is taken to its extreme in natural experiments. For this type of experiment, practitioners take advantage of some naturally occurring phenomena which assign treatments at random, so they have no influence over the design. One such example is a study by Shayo & Zussman (2011). The authors investigate ingroup bias in Israeli small claims courts. They take advantage of the fact that judges are assigned at random to cases in these courts. This introduces random variation in whether a case is assigned an Arab or Jewish judge. Unintentionally, the arrangement creates a randomized experiment which can be used to estimate the causal effect of a defendant being assigned to an ingroup versus outgroup judge, effectively measuring the ingroup bias. While the assignment is random, the design is skewed because some courts are ethically homogeneous.

The results in this note are of interest to studies of the types discussed in this section. Of particular interest is the close connection between the main conditions for consistency as captured in Conditions 1 and 2. This connection shows that if the variables of interest is known to be well-behaved, perhaps uniformly bounded, then practitioners have more leeway in the design of the survey or experiment, allowing them to give more attention to other important considerations. When practitioners have no control over the design, the results illuminate which naturally occurring designs can be used for inference.

References

Aronow, P. M. & Middleton, J. A. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, 1(1).

Breidt, F. J. & Opsomer, J. D. (2008). Endogenous post-stratification in surveys: Classifying with a sample-fitted model. *Annals of Statistics*, 36(1), 403–427.

Cardot, H., Chaouch, M., Goga, C., & Labruère, C. (2010). Properties of design-based functional principal components analysis. *Journal of Statistical Planning and Inference*, 140(1), 75–91.

Cassel, C. M., Särndal, C. E., & Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3), 615–620.

Chauvet, G. (2014). A note on the consistency of the Narain–Horvitz–Thompson estimator. arXiv:1412.2887v1.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2), 315–331.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling,

part one". In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.

Harvey, S., Harrison, D. A., Singer, M., Ashcroft, J., Jones, C. M., Elbourne, D., Brampton, W., Williams, D., Young, D., & Rowan, K. (2005). Assessment of the clinical effectiveness of pulmonary artery catheters in management of patients in intensive care (pac-man): a randomised controlled trial. *The Lancet*, 366(9484), 472–477.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161–1189.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.

Isaki, C. T. & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377), 89–96.

Karlan, D. & Zinman, J. (2009). Expanding credit access: Using randomized supply decisions to estimate the impacts. *The Review of Financial Studies*, 23(1), 433–464.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169—-175.

Neyman, J. (1990/1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4), 465–472. (Original work published 1923).

Niccolai, L. M., Verevochkin, S. V., Toussova, O. V., White, E., Barbour, R., Kozlov, A. P., & Heimer, R. (2010). Estimates of hiv incidence among drug users in st. petersburg, russia: continued growth of a rapidly expanding epidemic. *European Journal of Public Health*, 21(5), 613–619.

Opsomer, J. D. & Breidt, F. J. (2000). Local polynomial regresssion estimators in survey sampling. *The Annals of Statistics*, 28(4), 1026–1053.

Robins, J. M. & Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3), 285–319.

Robinson, P. M. (1982). On the convergence of the Horvitz-Thompson estimator. *Australian Journal of Statistics*, 24(2), 234–238.

Robinson, P. M. & Särndal, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā: The Indian Journal of Statistics, Series B*, 45(2), 240–248.

Shayo, M. & Zussman, A. (2011). Judicial ingroup bias in the shadow of terrorism. *Quarterly Journal of Economics*, 126(3), 1447–1484.

Torrieri, N. (2014). *American Community Survey: Design and Methodology.* Technical report, U.S. Census Bureau. Version 2.0, January 30, 2014.

A Consistency of other design-based estimators

The Horvitz–Thompson estimator tends to perform poorly in small samples, and practitioners generally benefit from using one of its many improvements. We focused on the original estimator in this note because the improved estimators often inherit its asymptotic properties. We here illustrate this by demonstrating that Proposition 1 implies consistency of two of these improvements.

The first is the estimator introduced by Hájek (1971):

$$\hat{\mu}_{\text{HÁ}} = \sum_{i \in \mathbf{S}} \frac{y_i}{\pi_i} \Big/ \sum_{i \in \mathbf{S}} \frac{1}{\pi_i},$$

and $\hat{\mu}_{\text{HÁ}} = 0$ if $\mathbf{S}$ is empty. This estimator adjusts for the average reciprocal of the inclusion probabilities. If the probabilities vary between the units or the sample size is random, this adjustment tends to stabilize the estimator.

Proposition A1. *The Hájek estimator is consistent for the population mean if Conditions 1 and 2 hold and* $\bar{e} \to 0$ *and* $\bar{\pi} N \to \infty$.

Proof. Let $h = N^{-1} \sum_{i \in \mathbf{S}} \pi_i^{-1}$, and note that $\hat{\mu}_{\text{HÁ}} = \hat{\mu}/h$. Corollary 1 provides convergence of $\hat{\mu}$ to μ under the stated conditions. Observe that h is the Horvitz–Thompson estimator for a population in which $y_i = 1$ for all units. Corollary 1 thus provides convergence of h to 1. The continuous mapping theorem completes the proof. □

The second estimator is the difference estimator by Cassel et al. (1976):

$$\hat{\mu}_{\text{DE}} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i + \frac{1}{N} \sum_{i \in \mathbf{S}} \frac{y_i - \hat{y}_i}{\pi_i},$$

where $\hat{y}_i$ is a non-random prediction of y_i.

The difference estimator takes advantage of auxiliary information about the responses in the population. In particular, if the auxiliary information provides predictions of the responses, our inferences could be improved because we can use the predictions to impute responses we do not observe. In fact, if the predictions are sufficiently good, their average alone will be a reasonable estimate of the population mean. This is the first term of the difference estimator.

The concern with this estimate is that it will be inaccurate if the predictions are poor. We can assess the quality of the predictions by comparing them to the responses in the sample. In fact, we can estimate the systematic prediction error using the Horvitz–Thompson estimator. This is the second term. If we detect a systematic prediction error, we want to adjust the first term by subtracting the estimated error, and this yields the difference estimator.

The estimator allows us to change Condition 1 to one that generally is weaker.

Condition A1. There exist $p > 2$ and $q > 1$ with $pq \geq p + 2q$ such that

$$\left[\frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|^p \right]^{1/p} \leq k_y, \qquad \text{and} \qquad \left[\frac{1}{N} \sum_{i=1}^{N} \tilde{w}_i^q \right]^{1/q} \leq k_\pi.$$

Proposition A2. *The difference estimator is consistent for the population mean if Conditions A1 and 2 hold and* $\bar{e} \to 0$ *and* $\bar{\pi} N \to \infty$.

Proof. Let $m_i = \hat{y}_i - y_i$, and note that m_i is observed when i is in $\mathbf{S}$. Let

$$m = \frac{1}{N} \sum_{i=1}^{N} m_i \qquad \text{and} \qquad \hat{m} = \frac{1}{N} \sum_{i \in \mathbf{S}} \frac{m_i}{\pi_i},$$

so $\hat{\mu}_{\text{DE}} = \mu + m - \hat{m}$. Note that $\hat{m}$ is the Horvitz–Thompson estimator of m, so $\hat{m} \to m$ under the stated conditions. □