Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities

Peter M. Aronow and Cyrus Samii[1]

August 27, 2012

Abstract

We consider conservative variance estimation for the Horvitz-Thompson estimator of a population total in sampling designs with zero pairwise inclusion probabilities, known as "non-measurable" designs. We decompose the standard Horvitz-Thompson variance estimator under such designs and characterize the bias precisely. We develop a bias correction that is guaranteed to be weakly conservative (nonnegatively biased) regardless of the nature of the non-measurability. The analysis sheds light on conditions under which the standard Horvitz-Thompson variance estimator performs well despite non-measurability and the conservative bias correction may outperform commonly-used approximations.

Keywords: Horvitz-Thompson estimation, non-measurable designs, variance estimation

[1] Peter M. Aronow, Department of Political Science, Yale University, 77 Prospect St., New Haven, CT 06520, Email: peter.aronow@yale.edu. Cyrus Samii, Department of Politics, New York University, 19 West 4th St., New York, NY 10012, Email: cds2083@nyu.edu.

1 Introduction

Sampling designs sometimes result in pairs of units having zero probability of being jointly included in the sample. Horvitz and Thompson (1952)'s statement of the properties of the finite population total makes clear that general, unbiased variance estimation for estimators of population totals is impossible for such *non-measurable* designs (Särndal et al., 1992, 33). Optimal methods for variance estimation in these cases remains an open problem. This paper analyzes the nature of the biases that non-measurability introduces for the standard Horvitz-Thompson estimator and studies an approach to correct for this bias in a conservative manner. While our results cannot offer a solution to the non-measurability problem for all practical applications, we do clarify conditions under which the standard estimator performs well and where the conservative bias correction outperforms commonly-used approximations.

Despite their theoretical drawbacks, sampling designs with zero pairwise inclusion probabilities are quite common. A common non-measurable design is one that draws only single units or clusters from a set of strata. This may occur if the population or a subpopulation of interest is incidentally sparse over stratification cells. Another common non-measurable design is a systematic sample in which unit indices are sampled from a list in multiples from a random starting value. In these designs, units whose indices are multiples from different starting values have zero joint probability of inclusion.

Approximate methods have been proposed for special cases, as discussed in Hansen et al. (1953, Section 9.15), Särndal et al. (1992, Chapter 3), and Wolter (2007, Chapters 2 and 8). In the single-unit per stratum case, a common approach is to collapse strata and assume units were drawn via a simple random sample from the larger, collapsed stratum. For systematic samples, the standard approach is to use an approximation based on an assumption of simple random sampling with replacement. These approximate methods are generally biased to a degree that cannot be determined from the data. In some cases, it can be shown that the bias will tend to be positive, but such is not the case generally, and especially so when the zero pairwise inclusion probabilities occur in a haphazard manner.

This paper begins by decomposing the bias of the Horvitz-Thompson variance estimator under non-measurability. This exposes precisely how conditions on the underlying data result in more or less bias. We also show how a simple application of Young's inequality yields a bias correction and a class of estimators guaranteed to have weakly positive bias as well as no bias under special conditions. We discuss implications for applied work.

2 Variance estimation for the Horvitz-Thompson estimator

Consider a population U indexed by $1,...,k,...N$ and a sampling design such that the probability of inclusion in the sample for unit k is π_k, and the joint inclusion probability for units k and l is π_{kl}. Under a measurable design, two conditions obtain: (1) $\pi_k > 0$ and π_k is known for all $k \in U$ and (2) $\pi_{kl} > 0$ and π_{kl} is known for all $k,l \in U$. Non-measurable designs include those for which either of the two conditions for a measurable design do not hold. Failure to meet the former condition precludes unbiased estimation of totals.

The Horvitz-Thompson estimator of a population total is $\hat{t} = \sum_{k \in s} y_k / \pi_k = \sum_{k \in U} I_k y_k / \pi_k$, where $I_k \in \{0,1\}$ is unit k's inclusion indicator, the only stochastic component of the expression, with $\mathrm{E}(I_k) = \pi_k$, the inclusion probability, and s and U refer to the sample and the population, respectively. Define $\mathrm{E}(I_k I_l) = \pi_{kl}$, which is the probability that both units k and l from U are included in the sample. Since $I_k I_k = I_k$, $\mathrm{E}(I_k I_k) = \pi_{kk} = \pi_k$ by construction. When condition 1 holds, as we assume throughout, the Horvitz-Thompson estimator is unbiased.

2.1 Properties of the Horvitz-Thompson variance estimator under measurability

By Horvitz and Thompson (1952), the variance of the Horvitz-Thompson estimator for the total,

$$\mathrm{Var}(\hat{t}) = \sum_{k\in U}\sum_{l\in U}\mathrm{Cov}(I_k, I_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} = \sum_{k\in U}\mathrm{Var}(I_k)\left(\frac{y_k}{\pi_k}\right)^2 + \sum_{k\in U}\sum_{l\in U\backslash k}\mathrm{Cov}(I_k, I_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

We label a sample from a measurable design, s^M, and an unbiased estimator for $\mathrm{Var}(\hat{t})$ on s^M is

$$\widehat{\mathrm{Var}}(\hat{t}) = \sum_{k\in s^M}\sum_{l\in s^M}\frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} = \sum_{k\in U}\sum_{l\in U} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l},$$

where the only stochastic part of the expression is $I_k I_l$, and unbiasedness is by $\mathrm{E}(I_k I_l) = \pi_{kl}$.

2.2 Properties of the Horvitz-Thompson variance estimator under non-measurability

We now examine the case where condition 2 does not hold: $\pi_{kl} = 0$ for some units $k, l \in U$. Because I_k is a Bernoulli random variable with probability π_k, $\mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$ for $k \neq l$, and $\mathrm{Cov}(I_k, I_k) = \mathrm{Var}(I_k) = \pi_k(1-\pi_k)$. Then, we can re-express the variance above as,

$$\begin{aligned}\mathrm{Var}(\hat{t}) &= \sum_{k\in U}\pi_k(1-\pi_k)\left(\frac{y_k}{\pi_k}\right)^2 + \sum_{k\in U}\sum_{l\in U\backslash k}(\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \\ &= \sum_{k\in U}\pi_k(1-\pi_k)\left(\frac{y_k}{\pi_k}\right)^2 + \sum_{k\in U}\sum_{l\in\{U\backslash k:\pi_{kl}>0\}}(\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} - \underbrace{\sum_{k\in U}\sum_{l\in\{U\backslash k:\pi_{kl}=0\}} y_k y_l}_{A}.\end{aligned}$$

For k and l such that $\pi_{kl} = 0$, the sampling design will never permit unbiased estimation of the component of the variance labeled as A above, since we will never observe y_k and y_l together. We label a sample from a design where condition 2 fails as s^0. When $\widehat{\mathrm{Var}}(\hat{t})$ is applied to s^0, the result is unbiased for $\mathrm{Var}(\hat{t}) + A$. We state this formally as follows:

Proposition 1 *When* s^0 *refers to a sample from a design with some* $\pi_{kl} = 0$,

$$\mathrm{E}\left[\widehat{\mathrm{Var}}(\hat{t})\right] = \mathrm{Var}(\hat{t}) + \sum_{k \in U} \sum_{l \in \{U \backslash k : \pi_{kl} = 0\}} y_k y_l = \mathrm{Var}(\hat{t}) + A.$$

Proof. The result follows from,

$$\mathrm{E}\left[\sum_{k \in s^0} \sum_{l \in s^0} \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}\right] = \mathrm{E}\left[\sum_{k \in U} \sum_{l \in \{U : \pi_{kl} > 0\}} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}\right]$$

$$= \sum_{k \in U} \mathrm{Var}(I_k) \left(\frac{y_k}{\pi_k}\right)^2 + \sum_{k \in U} \sum_{l \in \{U \backslash k : \pi_{kl} > 0\}} \mathrm{Cov}(I_k, I_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = \mathrm{Var}(\hat{t}) + \sum_{k \in U} \sum_{l \in \{U \backslash k : \pi_{kl} = 0\}} y_k y_l = \mathrm{Var}(\hat{t}) + A.$$

□

The standard Horvitz-Thompson variance estimator, if applied to designs with zero pairwise inclusion probabilities, can therefore have a positive or a negative bias. If the y_k, y_l values are always nonnegative (or always nonpositive), then the bias is always nonnegative. When values may be positive or negative, A is the sum of cross-products of outcomes that never appear together under the design sample. If the jointly exclusive outcomes are centered over zero, then no correlation in these outcomes would tend to result in small bias, positive correlation in positive bias, and negative correlation in negative bias.

3 Conservative bias correction for the Horvitz-Thompson variance estimator under non-measurability

The case where A may be less than zero for a non-measurable design suggests the need for some adjustment that will guarantee a bias that is weakly bounded below by zero. We first develop a general bias correction that is guaranteed to be conservative, later providing a special simplified case for practical usage.

3.1 General formulation

Consider the following variance estimator:

$$\widehat{\mathrm{Var}}_C(\hat{t}) = \sum_{k\in U}\sum_{l\in\{U:\pi_{kl}>0\}} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \sum_{k\in U}\sum_{l\in\{U\backslash k:\pi_{kl}=0\}} \left(I_k \frac{|y_k|^{a_{kl}}}{a_{kl}\pi_k} + I_l \frac{|y_l|^{b_{kl}}}{b_{kl}\pi_l} \right),$$

where a_{kl}, b_{kl} are positive real numbers such that $1/a_{kl} + 1/b_{kl} = 1$ for all pairs k, l with $\pi_{kl} = 0$. The estimator is guaranteed to produce an expected value greater than or equal to the true variance for all designs, and is thus conservative. We state this property formally:

Proposition 2 *The expected value of* $\widehat{\mathrm{Var}}_C(\hat{t})$,

$$\mathrm{E}\left[\widehat{\mathrm{Var}}_C(\hat{t})\right] \geq \mathrm{Var}(\hat{t}).$$

Proof. By Young's inequality,

$$\frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} \geq |y_k||y_l|,$$

if $1/a_{kl} + 1/b_{kl} = 1$. Define A^* such that,

$$A^* = \sum_{k \in U} \sum_{l \in \{U \backslash k : \pi_{kl}=0\}} \frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} \geq \sum_{k \in U} \sum_{l \in \{U \backslash k : \pi_{kl}=0\}} |y_k||y_l| \geq \sum_{k \in U} \sum_{l \in \{U \backslash k : \pi_{kl}=0\}} y_k y_l = A$$

and

$$A^* \geq \sum_{k \in U} \sum_{l \in \{U \backslash k : \pi_{kl}=0\}} |y_k||y_l| \geq \sum_{k \in U} \sum_{l \in \{U \backslash k : \pi_{kl}=0\}} -y_k y_l = -A.$$

Therefore

$$\text{Var}(\hat{t}) + A + A^* \geq \text{Var}(\hat{t}).$$

The associated Horvitz-Thompson estimator of A^* would be

$$\widehat{A^*} = \sum_{k \in U} \sum_{l \in \{U \backslash k : \pi_{kl}=0\}} \left(I_k \frac{|y_k|^{a_{kl}}}{a_{kl}\pi_k} + I_l \frac{|y_l|^{b_{kl}}}{b_{kl}\pi_l} \right),$$

which is unbiased by $E(I_k) = \pi_k$ and $E(I_l) = \pi_l$.

Since $E(\hat{A}^*) = A^*$, by Proposition 1,

$$\text{E}\left[\sum_{k \in s^0} \sum_{l \in s^0} \frac{\text{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \widehat{A^*} \right] = \text{Var}(\hat{t}) + A + A^*$$

$$\text{E}\left[\widehat{\text{Var}}_C(\hat{t}) \right] \geq \text{Var}(\hat{t}).$$

Substituting terms,

$$\text{E}\left[\sum_{k \in U} \sum_{l \in \{U : \pi_{kl} > 0\}} I_k I_l \frac{\text{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \sum_{k \in U} \sum_{l \in \{U \backslash k : \pi_{kl}=0\}} \left(I_k \frac{|y_k|^{a_{kl}}}{a_{kl}\pi_k} + I_l \frac{|y_l|^{b_{kl}}}{b_{kl}\pi_l} \right) \right] \geq \text{Var}(\hat{t}).$$

□

$\widehat{\mathrm{Var}}_C(\hat{t})$ is justified as a conservative estimator for the case when A is not known to be positive. This estimator is unbiased under a special condition:

Corollary 1 *If, for all pairs* k,l *such that* $\pi_{kl}=0$, *(i)* $|y_k|^{a_{kl}}=|y_l|^{b_{kl}}$ *and (ii)* $-y_k y_l=|y_k||y_l|$,

$$\mathrm{E}\left[\widehat{\mathrm{Var}}_C(\hat{t})\right]=\mathrm{Var}(\hat{t}).$$

Proof. By (i), (ii) and Young's inequality,

$$\frac{|y_k|^{a_{kl}}}{a_{kl}}+\frac{|y_l|^{b_{kl}}}{b_{kl}}=|y_k||y_l|=-y_k y_l.$$

Therefore,

$$A^*=\sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}}\frac{|y_k|^{a_{kl}}}{a_{kl}}+\frac{|y_l|^{b_{kl}}}{b_{kl}}=\sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}}|y_k||y_l|=\sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}}-y_k y_l=-A.$$

It follows that $\mathrm{Var}(\hat{t})+A+A^*=\mathrm{Var}(\hat{t})$ and $\mathrm{E}\left[\widehat{\mathrm{Var}}_C(\hat{t})\right]=\mathrm{Var}(\hat{t})$. □

If any units k,l are in clusters (i.e., $\Pr\left(I_k\neq I_l\right)=0$), these units should be totaled into one larger unit before estimation. Combining units will tend to reduce the bias of the variance estimator because only pairs of cluster-level totals will be included in A^*, as opposed to all constituent pairs.

3.2 Simplified special case

In general, it would be difficult to assign optimal values of a_{kl} and b_{kl} for all pairs k,l such that $\pi_{kl}=0$. Instead, we examine one intuitive case, assigning all $a_{kl}=b_{kl}=2$:

$$\widehat{\mathrm{Var}}_{C2}(\hat{t})=\sum_{k\in U}\sum_{l\in\{U:\pi_{kl}>0\}}I_k I_l\frac{\mathrm{Cov}(I_k,I_l)}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}+\sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}}\left(I_k\frac{y_k^2}{2\pi_k}+I_l\frac{y_l^2}{2\pi_l}\right).$$

As a special case of $\widehat{\text{Var}}_C(\hat{t})$, $\widehat{\text{Var}}_{C2}(\hat{t})$ is also conservative:

Corollary 2 *The expected value of* $\widehat{\text{Var}}_{C2}(\hat{t})$,

$$\text{E}\left[\widehat{\text{Var}}_{C2}(\hat{t})\right] \geq \text{Var}(\hat{t}).$$

Proof. For all pairs k,l such that $\pi_{kl} = 0$, $1/a_{kl} + 1/b_{kl} = 1$. Proposition 1 therefore holds. □

The choice to set all $a_{kl} = b_{kl} = 2$ is justified by the fact that it will yield the lowest value of the estimator $\widehat{\text{Var}}_C(\hat{t})$ subject to the constraint that a_{kl} and b_{kl} are fixed as constants a and b over all k,l.

Corollary 3 *Among the class of estimators* $\widehat{\text{Var}}_{Cab}(\hat{t})$, *defined as the set of estimators* $\widehat{\text{Var}}_C(\hat{t})$ *such that all* $a_{kl} = a$ *and all* $b_{kl} = b$, $\widehat{\text{Var}}_{C2}(\hat{t}) = \min_{a,b}\left[\widehat{\text{Var}}_{Cab}(\hat{t})\right]$.

Proof. By simple algebra,

$$\widehat{\text{Var}}_{Cab}(\hat{t}) = \sum_{k\in U}\sum_{l\in\{U:\pi_{kl}>0\}} I_k I_l \frac{\text{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \sum_{k\in U}\sum_{l\in\{U\backslash k:\pi_{kl}=0\}} \left(I_k \frac{|y_k|^a}{a\pi_k} + I_l \frac{|y_l|^b}{b\pi_l} \right)$$

$$= \widehat{\text{Var}}(\hat{t}) + \sum_{k\in U}\sum_{l\in\{U\backslash k:\pi_{kl}=0\}} \left(I_k \frac{|y_k|^a}{a\pi_k} + I_l \frac{|y_l|^b}{b\pi_l} \right)$$

$$= \widehat{\text{Var}}(\hat{t}) + \frac{1}{2}\sum_{k\in U}\sum_{l\in\{U\backslash k:\pi_{kl}=0\}} \left(I_k \frac{|y_k|^a}{a\pi_k} + I_l \frac{|y_l|^b}{b\pi_l} + I_k \frac{|y_k|^b}{b\pi_k} + I_l \frac{|y_l|^a}{a\pi_l} \right)$$

$$= \widehat{\text{Var}}(\hat{t}) + \frac{1}{2}\sum_{k\in U}\sum_{l\in\{U\backslash k:\pi_{kl}=0\}} \left(\frac{I_k}{\pi_k}\left(\frac{|y_k|^a}{a} + \frac{|y_k|^b}{b} \right) + \frac{I_l}{\pi_l}\left(\frac{|y_l|^a}{a} + \frac{|y_l|^b}{b} \right) \right).$$

By Young's inequality, given $1/a+1/b=1$, $|y_k|^a/a+|y_k|^b/b \geq y_k^2$, but equality must hold if $a=b=2$. Similarly, $|y_l|^a/a+|y_l|^b/b \geq y_l^2$, but equality must hold if $a=b=2$. Since $I_k/\pi_k \geq 0$ and $I_l/\pi_l \geq 0$, any choice $a \neq b$ can only yield $\widehat{\mathrm{Var}}_{Cab}(\hat{t}) \geq \widehat{\mathrm{Var}}_{C2}(\hat{t})$. □

Given all values of y_k and y_l, it is possible to derive an optimal vector of a_{kl} and b_{kl} values that varies over k,l, but such a derivation may not be of practical value.

4 Applications

Proposition 1 shows that the bias of the Horvitz-Thompson variance estimator under non-measurability is

$$A=\sum_{k \in U} \sum_{l \in \{U \backslash k: \pi_{kl}=0\}} y_k y_l .$$

This expression, along with the fact that $A^* \geq A$, makes it evident that the degree of bias in $\widehat{\mathrm{Var}}(\hat{t})$ and $\widehat{\mathrm{Var}}_C(\hat{t})$ depends a great deal on the number of pairs with zero pairwise inclusion probabilities. For designs where this number is small, $\widehat{\mathrm{Var}}(\hat{t})$ may provide a reasonable and conservative estimator for cases where y_k takes the same sign for all k, and $\widehat{\mathrm{Var}}_C(\hat{t})$ may provide a reasonable and conservative estimator for cases where y_k may take different signs for some k. An example that arises frequently is stratified sampling where for a relatively small proportion of cases, we have small strata from which we draw only one unit.

For designs that result in many pairs having zero inclusion probabilities, $\widehat{\mathrm{Var}}(\hat{t})$ and $\widehat{\mathrm{Var}}_C(\hat{t})$ could be wildly over-conservative and other estimators may be preferred in terms of criteria such as mean square error. A prominent example is systematic sampling. Indeed, Särndal et al. (1992, 76) propose that under systematic sampling, the Horvitz-Thompson variance estimator, $\widehat{\mathrm{Var}}(\hat{t})$, can give a "non-sensical result." The expression for A makes it clear why this would be the case. Wolter (2007, Ch. 8) shows that simpler biased estimators, such as the with-replacement (Hansen-Hurwitz) variance estimator, can be reliable, if slightly conservative, in a broad range of data scenarios under equal probability and probability proportional to size (PPS) systematic sampling. Nonetheless, the with-replacement estimator fails to account adequately for sampling variance when outcome variance within systematic sample clusters is smaller than the between cluster variance. In such cases, $\widehat{\text{Var}}(\hat{t})$ would bound this variance in expectation when outcomes are all of the same sign, and $\widehat{\text{Var}}_C(\hat{t})$ would always bound this variance in expectation. Of course, it may still be the case that the bias is too large to be of much use, and so we would not suggest that $\widehat{\text{Var}}(\hat{t})$and $\widehat{\text{Var}}_C(\hat{t})$ provides a full solution to the variance estimation problem for systematic sampling under high intra-cluster correlation.

Results from simulation studies are available in a supplement (at https://files.nyu.edu/cds2083/public/docs/smj_suppl.pdf). They illustrate how $\widehat{\text{Var}}(\hat{t})$ and $\widehat{\text{Var}}_C(\hat{t})$ perform relative to commonly-used alternatives in applied scenarios. The simulations demonstrate situations when these estimators are preferable to the alternatives. For one-unit-per-stratum sampling, we show that these estimators are less biased than the "collapsed stratum" estimator in a range of scenarios. For PPS systematic sampling, these estimators perform favorably when the population exhibits substantial periodicity, a case when the commonly-used with-replacement estimator may be grossly negatively biased.

5 Conclusion

We have characterized precisely the bias of the Horvitz-Thompson variance estimator under non-measurability and used this characterization to develop a conservative bias correction. These estimators reflect the fundamental uncertainty inherent to non-measurable designs. Compared to available approximate methods, these estimators may sometimes perform better and sometimes worse from a practical perspective. But available approximate methods may be biased in ways that cannot always be evaluated in terms of either magnitude or sign. The estimators developed in this paper may therefore provide an informative measure of sampling variability with which analysts can agree without invoking additional assumptions or resorting to methods that carry the potential for negative bias. The bias term, A, has a simple form that suggests the possibility of refinements to the estimators developed here, something that we leave open for future research.

Acknowledgements

The authors wish to thank Joel Middleton, Allison Carnegie, an anonymous reviewer and the associate editor for helpful comments.

References

Hansen, M. H., W. N. Hurwitz, and W. G. Madow (1953). *Sample Survey Methods and Theory: Volume 1 Methods and Theory*. New York: Wiley.

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

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer.

Wolter, K. (2007). *Introduction to Variance Estimation, Second Edition*. New York: Springer.