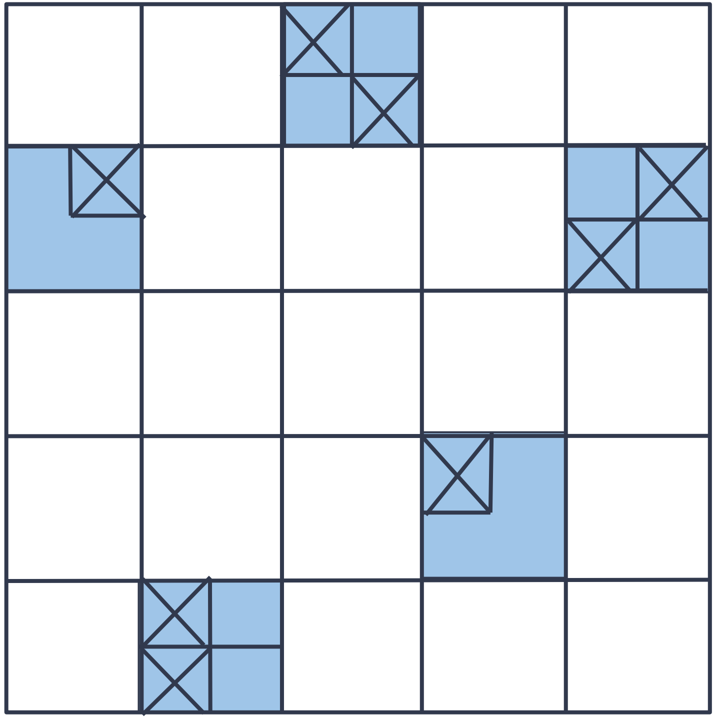
Inference for Two-Stage Sampling

Authors: Guillaume Chauvet and Audrey-Anne Valet

Andrea Boskovic
University of Washington, Department of Statistics

What is two-stage sampling?



PSU

N_I = 25

n_I = 5

SSU

N_i = 4, 2

n_i = 2, 1

Motivation for this work

- We're often interested in estimating a population total Y
- The Horvitz-Thompson (HT) estimator is often used to estimate totals in these surveys

One-Stage HT Estimator

$$\hat{Y}_\pi = \sum_{i \in S_i} \frac{y_i}{\pi_i}$$

Two-Stage HT Estimator

$$\hat{Y}_\pi = \sum_{i \in S_I} \frac{\sum_{k \in S_i} \frac{y_{ik}}{\pi_{k|i}}}{\pi_{Ii}}$$

$$\pi_{Ii} = \mathbb{P}(i \in S_I) \qquad \pi_{k|i} = \mathbb{P}(k \in S_i | i \in S_I)$$

HT variance estimator

We can decompose the HT variance estimator into three components:

$$V(\hat{Y}_\pi) = \boxed{\sum_{i=1}^{N_I}\sum_{j=1}^{N_I} \Delta_{Iij} \frac{\hat{Y}_i}{\pi_{Ii}} \frac{\hat{Y}_j}{\pi_{Ij}}} + \boxed{\sum_{i=1}^{N_I} \frac{1-\pi_{Ii}}{\pi_{Ii}} V_i} + \boxed{\sum_{i=1}^{N_I} V_i}$$

$$= \boxed{V_1(\hat{Y}_\pi)} + \boxed{V_2(\hat{Y}_\pi)} + \boxed{V_3(\hat{Y}_\pi)}$$

Previous work

- Several papers have proven these properties for estimators in one-stage contexts, but there is a dearth of literature pertaining to two-stage designs (Isaki and Fuller 1982)

- Asymptotic normality of the HT estimator under a multi-stage design has been proven, but the results rely on SRS being used in the first stage (Chauvet 2015)

Sampling design conditions

These conditions are needed to produce reliable estimators with associated confidence intervals:

1. The estimator should be consistent for the true total,
2. The estimator should be asymptotically normally distributed, and
3. Consistent variance estimators must exist in order to produce valid normality-based confidence intervals.

Goal of this paper

Problem: These conditions are only established for one-stage sampling designs

Goal of this paper: Establish these conditions for two-stage sampling designs

Main results

1. Consistency of traditional variance estimators under two-stage designs
    - Unbiased variance estimators: term-by-term decomposition
    - Simplified one-term variance estimators

2. Hajek-type variance estimator is consistent under large entropy designs

Required assumptions

Assumptions fall broadly into three categories:

1. Assumptions on the first-stage sampling design
2. Assumptions on the second-stage sampling design
3. Assumptions on the variable of interest

Consistency of the HT estimator

Previous work (Breidt and Opsomer 2008) proved consistency under alternate assumptions, but we show consistency under a set of more flexible assumptions

$$E\{N^{-1}(\hat{Y}_\pi - Y)\}^2 = O(n_I^{-1}),$$

$$\hat{Y}_\pi / Y \xrightarrow{p} 1$$

Consistency of variance estimators

Recall the conditions:

1. ~~The estimator should be consistent for the true total,~~
2. The estimator should be asymptotically normally distributed, and
3. Consistent variance estimators must exist in order to produce valid normality-based confidence intervals.

Unbiased variance estimators

HT variance estimator and the Yates-Grundy (YG) variance estimator are consistent and term-by-term unbiased

- The YG variance estimator is appropriate for sampling designs of fixed size at both stages
- The HT variance estimator is used otherwise

Our results rely on term-by-term decompositions of the variance estimators.

HT and YG variance estimators

Horvitz-Thompson

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

Yates-Grundy

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

Both variance estimators are consistent and term-by-term unbiased

Simplified one-term variance estimators: Approach 1

YG and HT variance estimators are difficult to use in practice because they require unbiased and consistent variance estimators inside any of the selected PSUs

Idea: Consider only the A term components of variance estimators because the last component has a small contribution to the overall variance

Simplified one-term variance estimators: Approach 1

$$V(\hat{Y}_\pi) = \boxed{\sum_{i=1}^{N_I}\sum_{j=1}^{N_I} \Delta_{Iij} \frac{\hat{Y}_i}{\pi_{Ii}} \frac{\hat{Y}_j}{\pi_{Ij}}} + \boxed{\sum_{i=1}^{N_I} \frac{1-\pi_{Ii}}{\pi_{Ii}} V_i} + \boxed{\sum_{i=1}^{N_I} V_i}$$

$$= \boxed{V_1(\hat{Y}_\pi)} + \boxed{V_2(\hat{Y}_\pi)} + \boxed{V_3(\hat{Y}_\pi)}$$

Consistent estimators

$\hat{V}_{HT,A}(\hat{Y}_\pi)$ $\hat{V}_{HT,B}(\hat{Y}_\pi)$

$\hat{V}_{YG,A}(\hat{Y}_\pi)$ $\hat{V}_{YG,B}(\hat{Y}_\pi)$

Simplified one-term variance estimators: Approach 2

Idea: Estimate the variance as if the PSUs were sampled with replacement, i.e., multinomial sampling

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

Large-entropy sampling designs

- Many sampling designs are high entropy, meaning there is a high degree of uncertainty in the sample that will be obtained
- Useful to study entropy in the context of variance estimation (Tille and Haziza 2010)

Entropy of a sampling design

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

Note: U is the support of the sampling design p

Hajek variance estimator

Given Conditional Poisson design in the first stage, we can define the Hajek variance estimator, which only relies on **first order, first-stage inclusion probabilities**:

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

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

Result: Both of these estimators are consistent for the true Horvitz-Thompson variance

Asymptotic Normality of the HT estimator

Recall the conditions:

1. ~~The estimator should be consistent for the true total,~~
2. The estimator should be asymptotically normally distributed, and
3. ~~Consistent variance estimators must exist in order to produce valid normality based confidence intervals.~~

Asymptotic Normality of the HT estimator

Under large entropy sampling in the first-stage and some conditions,

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

We have shown the conditions for two-stage sampling

Recall the conditions:

1. ~~The estimator should be consistent for the true total,~~
2. ~~The estimator should be asymptotically normally distributed, and~~
3. ~~Consistent variance estimators must exist in order to produce valid normality based confidence intervals.~~

Simulation Study: Setup

Goal: Compare the performance of three estimators:

1. With-replacement variance estimator
2. Hajek-type variance estimator
3. Hajek-type simplified (A-term) estimator

Consider 3 populations of 2000 PSUs, and varying number of SSUs.

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

Simulation Study: Performance evaluation

Goal: Compare the performance of three estimators of interest via three metrics

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

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

95% Confidence Interval Coverage

Simulation Study: Results Comparison

Original Paper

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

My Results

			RB_{MC}		RS_{MC}		CI_{MC}	
ICC	n_I	n_i	$\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ,A}(\widehat{Y}_\pi)$	$\widehat{V}_{HAJ}(\widehat{Y}_\pi)$
0.1	20	5	0.09	0.68	33.20	33.18	0.94	0.94
		10	-0.95	-0.56	31.32	31.33	0.93	0.93
	40	5	-1.02	0.23	21.60	21.57	0.94	0.94
		10	-2.67	-1.82	21.88	21.75	0.93	0.93
	100	5	-3.24	-0.09	14.04	13.63	0.94	0.94
		10	-2.36	-0.28	14.32	14.16	0.95	0.95
	200	5	-6.63	-0.16	11.17	9.05	0.94	0.94
		10	-4.13	0.18	10.44	9.58	0.94	0.95
0.2	20	5	-0.39	0.03	33.21	33.19	0.93	0.93
		10	-0.82	-0.57	32.02	32.01	0.93	0.93
	40	5	-0.81	0.03	22.21	22.19	0.94	0.94
		10	-2.18	-1.73	22.00	21.95	0.93	0.93
	100	5	-2.21	-0.14	14.09	13.86	0.95	0.95
		10	-1.72	-0.58	14.33	14.24	0.94	0.95
	200	5	-4.57	-0.14	10.21	9.19	0.94	0.94
		10	-2.27	0.27	9.92	9.76	0.94	0.94
0.3	20	5	-0.76	-0.42	32.91	32.83	0.94	0.94
		10	-0.71	-0.53	32.36	32.35	0.93	0.93
	40	5	-0.75	-0.13	22.57	22.54	0.94	0.94
		10	-1.84	-1.53	22.00	21.98	0.93	0.93
	100	5	-1.62	-0.17	14.10	14.04	0.95	0.95
		10	-1.43	-0.69	14.27	14.26	0.95	0.95
	200	5	-3.25	-0.14	9.81	9.23	0.95	0.95
		10	-1.29	0.33	9.81	9.74	0.95	0.95

Application to Urban Policy Data: Setup

- Survey on security, employment, housing conditions, and schooling for those in urban areas
- Initial panel is selected through two-stage sampling with districts as PSUs and households as SSUs
- Total of n_i = 1065 households selected overall

Sampling scheme:

- n_I = 40 districts selected with probability proportional to the number of main dwellings
- n_i households selected within each PSU with equal probability

Application to Urban Policy Data: Results

	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]

- CI for simplified Hajek is always narrower
- Nearly identical performance of Hajek and simplified Hajek variance estimators

Conclusions

- Provided an asymptotic set-up for studying two-stage designs
- Given conditions for consistency of HT estimator and various variance estimators
- When first-stage sampling fraction is negligible, simplified variance estimators are also consistent and perform well according to our simulation and real-world application

References

1. Isaki, Cary T., and Wayne A. Fuller. "Survey design under the regression superpopulation model." *Journal of the American Statistical Association* 77.377 (1982): 89-96.
2. Chauvet, Guillaume. "Coupling methods for multistage sampling." (2015): 2484-2506.
3. Tillé, Yves, and David Haziza. "An interesting property of the entropy of some sampling designs." *Survey Methodology* 36.2 (2010): 229-31.
4. Breidt, F. Jay, and Jean D. Opsomer. "Endogenous post-stratification in surveys: Classifying with a sample-fitted model." (2008): 403-427.