Unequal Probability Sampling without Replacement

Survey Sampling

Statistics 4234/5234

Fall 2018

November 8, 2018

First we summarize two-stage unequal-probability cluster sampling WITH replacement (assuming SRS at second stage)

1. Determine $\psi_i =$ selection probability for ith psu, for $i = 1, 2, \ldots, N$.

2. Sample n of the N psus, with replacement, using selection probabilities ψ_i; call the sample $\mathcal{R}$, which may contain duplicates.

3. Take separate SRS of M_i from the M_i ssus in the ith psu, for each $i \in \mathcal{R}$.

4. Compute

   $$\frac{\hat{t}_{ij}}{\psi_i} \quad \text{for } j = 1, \ldots, Q_i \text{ for } i \in \mathcal{R}$$

   Here Q_i is the number of times the ith psu is counted in $\mathcal{R}$.

5. Take $\hat{t}_\psi$ to be the average of the n independent estimates in step 4. And that's a straight average, not weighted!

6. The standard error of our estimator is

   $$\text{SE}(\hat{t}_\psi) = \frac{1}{\sqrt{n}}(\text{SD of the } n \text{ independent estimates in step 4})$$

Unequal probability sampling without replacement

(Section 6.4)

Obviously this is more efficient.

Also more complicated to analyze.

Selection probabilities change from draw to draw.

Note

$$P(\text{select unit } i \text{ then unit } k) = \psi_i \times \frac{\psi_k}{1 - \psi_i}$$

and

$$P(\text{select unit } k \text{ then unit } i) = \psi_k \times \frac{\psi_i}{1 - \psi_k}$$

And thus

$$P(\text{select unit } i \text{ then unit } k) \neq P(\text{select unit } k \text{ then unit } i)$$

That complicates things!

One-stage sampling

(6.4.1)

Let $\pi_i = P(\text{unit } i \text{ in sample})$.

Let $\pi_{ik} = P(\text{units } i \text{ and } k \text{ in sample})$.

The **Horvitz-Thompson estimator** of the population total is

$$\hat{t}_{\mathsf{HT}} = \sum_{i \in \mathcal{S}} \frac{t_i}{\pi_i}$$

It is easy to show that $E(\hat{t}_{\mathsf{HT}}) = t$.

It is not at all easy to show that $V(\hat{t}_{\mathsf{HT}})$ is given by (6.20) and (6.21) on page 241.

These two expressions for the true variance are equivalent, of course, but they suggest different estimators that are not equivalent.

Computation of the π_i can be challenging, and the π_{ik} even more so — plus there are $\binom{N}{2}$ of them!

A common simplified approach is:

1. Do unequal probability sampling without replacement, because it's more efficient.

2. Use $\pi_i = n\psi_i$ in the H-T estimator (correct if *with* replacement)

   $$\widehat{t}_{\text{HT,mod}} = \frac{1}{n} \sum_{i \in \mathcal{S}} \frac{t_i}{\psi_i}$$

   which is approximate.

3. Report a standard error based on the with-replacement approximation also,

$$\widehat{V}_{\mathsf{WH}}(\hat{t}_{\mathsf{HT,mod}}) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(\frac{t_i}{\psi_i} - \hat{t}\right)^2$$

which is conservative.

Unequal-prob sampling without replace: two-stage

(6.4.3)

For one-stage we had $\hat{t}_{\mathsf{HT}} = \sum_{i\in\mathcal{S}} \frac{t_i}{\psi_i}$

For two-stage sampling it becomes

$$\hat{t}_{\mathsf{HT}} = \sum_{i \in \mathcal{S}} \frac{\hat{t}_i}{\psi_i}$$

The variance has an additional term, given by

$$\cdots + \sum_{i=1}^{N} \frac{V(\hat{t}_i)}{\pi_i}$$

and estimated by

$$\cdots + \sum_{i \in \mathcal{S}} \frac{\hat{V}(\hat{t}_i)}{\pi_i^2}$$

Typos in text!

Equation (5.24) on page 185, The $\frac{N}{n}$ term should be $\left(\frac{N}{n}\right)^2$.

Equations (6.28) and (6.29) on page 245: The $\frac{\widehat{V}(\widehat{t}_i)}{\pi_i}$ terms should be $\frac{\widehat{V}(\widehat{t}_i)}{\pi_i^2}$.

But never mind anyway, because we're not going to use those estimators.

We will use the with-replacement approximation,

$$\widehat{V}_{\mathsf{WR}}(\widehat{t}_{\mathsf{HT}}) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(\frac{n\widehat{t}_i}{\pi_i} - \widehat{t}\right)^2 = \frac{n}{n-1}\sum_{i\in\mathcal{S}}\left(\frac{\widehat{t}_i}{\pi_i} - \frac{\widehat{t}}{n}\right)^2$$

To sum up.

Unequal-probability sampling without replacement:

One-stage we have

$$\hat{t}_{\mathsf{HT}} = \sum_{i \in \mathcal{S}} \frac{t_i}{\pi_i}$$

Estimated variance is

$$\hat{V}_{\mathsf{WR}}(\hat{t}_{\mathsf{HT}}) = \frac{n}{n-1} \sum_{i \in \mathcal{S}} \left(\frac{t_i}{\pi_i} - \frac{\hat{t}}{n} \right)^2$$

For two-stage we have

$$\widehat{t}_{\mathsf{HT}} = \sum_{i \in \mathcal{S}} \frac{\widehat{t}_i}{\pi_i}$$

Estimated variance is

$$\widehat{V}_{\mathsf{WR}}(\widehat{t}_{\mathsf{HT}}) = \frac{n}{n-1} \left(\frac{\widehat{t}_i}{\pi_i} - \frac{\widehat{t}}{n} \right)^2$$

Weights

(6.4.4)

Next time.