Unequal Probability Sampling: Two-Stage with Replacement

Survey Sampling

Statistics 4234/5234

Fall 2018

November 1, 2018

For now we are still working in one-stage sampling, unequal probabilities, with replacement.

Designing selection probabilities

(6.2.3)

Example: Supermarkets

The estimator has lower variance if we assign higher selection probabilities to bigger stores.

The *ideal* is $\psi_i = t_i/t$ for $i = 1, 2, \ldots, N$.

In that case

$$\widehat{t}_\psi = \frac{1}{n} \sum_{i \in \mathcal{R}} \frac{t_i}{\psi_i} = t$$

for *any* possible sample $\mathcal{R}$.

The variance is zero!

Of course, in practice, this can't be done because of course the t_i are unknown.

One common approach is to use the cluster sizes M_i, that is, $\psi_i = M_i/M_0$, and get

$$\widehat{t}_\psi = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{t_i}{\psi_i} = \frac{M_0}{n}\sum_{i\in\mathcal{R}}\bar{y}_i$$

and

$$\widehat{\bar{y}}_\psi = \frac{\widehat{t}_\psi}{\widehat{M}_{0\psi}} = \frac{\widehat{t}_\psi}{M_0} = \frac{1}{n}\sum_{i\in\mathcal{R}}\bar{y}_i \qquad (1)$$

Estimate the population mean by the sample average of the cluster means.

Wait a second. The *straight* average? Not weighted by cluster sizes?

Yes, the straight average! The appropriate weighting was taken care of by the selection probabilities!

The Chapter 5 way of doing things: Sample clusters with equal probability, then take a weighted average of the sampled cluster means, weighted by cluster sizes, to estimate the population mean

$$\hat{\bar{y}} = \frac{\sum_{i \in \mathcal{S}} M_i \bar{y}_i}{\sum_{i \in \mathcal{S}} M_i}$$

The new way (Chapter 6) is to sample with probabilities $\propto M_i$, and take the straight average

$$\hat{\bar{y}}_\psi = \frac{1}{n} \sum_{i \in \mathcal{R}} \bar{y}_i$$

Consider (1), where $\psi_i = M_i/M_0$, why does it follow that $\hat{M}_{0\psi} = M_0$ for any possible sample?

Because

$$\hat{M}_{0\psi} = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{M_i}{\psi_i} = M_0$$

For the standard error of our estimator note that

$$\hat{V}(\hat{\bar{y}}_\psi) = \frac{1}{n\hat{M}_{0\psi}^2}\frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(\frac{t_i}{\psi_i} - \hat{\bar{y}}_\psi\frac{M_i}{\psi_i}\right)^2$$

which for $\psi_i = M_i/M_0$ gives

$$\hat{V}(\hat{\bar{y}}_\psi) = \frac{1}{n(n-1)}\sum_{i\in\mathcal{R}}\left(\bar{y}_i - \hat{\bar{y}}_\psi\right)^2 = \frac{s^2}{n}$$

where s^2 is the sample variance of the $\bar{y}_i$'s.

Sampling Weights

Write

$$\widehat{t}_\psi = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{t_i}{\psi_i} = \sum_{i\in\mathcal{R}} w_i t_i$$

The sampling weight for unit i is

$$w_i = \frac{1}{n\psi_i} = \frac{1}{E(Q_i)}$$

where Q_i is the number of times unit i is counted in the sample — remember we're sampling with replacement!

Here w_i represents the sampling weight of the ith psu, $i = 1, 2, \ldots, N$.

What about the ssu wieghts w_{ij}?

For one-stage sampling we have

$$\hat{t}_\psi = \sum_{i \in \mathcal{R}} w_i t_i = \sum_{i \in \mathcal{R}} \sum_{j=1}^{M_i} w_{ij} y_{ij}$$

where $w_{ij} = w_i$, and

$$\hat{\bar{y}}_\psi = \frac{\sum_{i \in \mathcal{R}} \sum_{j=1}^{M_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{R}} \sum_{j=1}^{M_i} w_{ij}}$$

PPS sampling is not self-weighting.

Elements in larger psus are given smaller weight (but are more likely to get counted).

Elements in smaller psus are less likely to get counted, but are given greater weight if they do.

Two-Stage Sampling

(Section 6.3)

Still **with** replacement

The story here is basically the same as one-stage, just replace t_i everywhere by $\widehat{t}_i$!

Notation

Let Q_i = the number of times psu i is counted in the sample.

We have Q_i independent estimates of t_i, denote them

$$\hat{t}_{i1}, \ldots, \hat{t}_{iQ_i}$$

Then

$$\hat{t}_\psi = \frac{1}{n} \sum_{i=1}^{N} \sum_{j=1}^{Q_i} \frac{\hat{t}_{ij}}{\psi_i}$$

The standard error is $\text{SE}(\hat{t}_\psi) = \sqrt{\hat{V}}$ where

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{N} \sum_{j=1}^{Q_i} \left(\frac{\hat{t}_{ij}}{\psi_i} - \hat{t}_\psi \right)^2$$

Also

$$\hat{\bar{y}}_\psi = \frac{\hat{t}_\psi}{\hat{M}_{0\psi}}$$

where

$$\hat{M}_{0\psi} = \frac{1}{n}\sum_{i=1}^{N} Q_i \frac{M_i}{\psi_i}$$

Based on the theory for ratio estimation we get

$$\hat{V}(\hat{\bar{y}}_\psi) = \frac{1}{n\hat{M}_{0\psi}^2}\frac{1}{n-1}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\left(\frac{\hat{t_{ij}}}{\psi_i} - \hat{\bar{y}}_\psi\frac{M_i}{\psi_i}\right)^2$$

Weights

The expected number of times that the jth ssu of the ith psu gets counted in the sample is, assuming SRS at the second state,

$$E(Q_{ij}) = n\psi_i \left(\frac{m_i}{M_i}\right)$$

Thus the sampling weight for this unit is

$$w_{ij} = \frac{1}{n\psi_i}\frac{M_i}{m_i}$$

Under PPS sampling, $\psi_i = M_i/M_0$, and we get

$$w_{ij} = \frac{M_0}{nm_i}$$

Summary

Two-stage, unequal-probability cluster sampling with replacement (using SRS at second stage)

1. Determine the selection probabilities ψ_i.

2. Take a sample of size n from the N psus using selection probabilities ψ_i.

3. Take Q_i separate SRSs of size m_i from M_i ssus in psu i, for each $i =\in \mathcal{R}$.

4. Compute

$$\frac{\widehat{t}_{ij}}{\psi_i} \quad \text{for} \quad j = 1, \ldots, Q_i \quad \text{for} \quad i = 1, \ldots, N$$

   We thus have n independent estimates of t.

5. Take $\widehat{t}_\psi$ to be the average of the estimates in step 4.

6. The standard error of our estimator is

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