Designing a Cluster Sample

Survey Sampling
Statistics 4234/5234
Fall 2018

October 25, 2018

Two-stage cluster sampling

Take an SRS of n of the N psus, denote the sample $\mathcal{S}$.

For each $i \in \mathcal{S}$, take an SRS of m_i of the M_i ssus, denote it $\mathcal{S}_i$.

Estimate

$$t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{if}$$

by

$$\hat{t}_{\text{unb}} = \frac{N}{n} \sum_{i \in \mathcal{S}} \hat{t}_i = \frac{N}{n} \sum_{i \in \mathcal{S}} M_i \bar{y}_i = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} \frac{N M_i}{n m_i} y_{ij}$$

The standard error is $\mathrm{SE}(\hat{t}_{\text{unb}}) = \sqrt{\hat{V}(\hat{t}_{\text{unb}})}$ of course.

We have

$$\hat{V}(\hat{t}_{\text{unb}}) = N^2 \frac{s_t^2}{n}\left(1 - \frac{n}{N}\right) + \left(\frac{N}{n}\right)^2 \sum_{i \in \mathcal{S}} M_i^2 \frac{s_i^2}{m_i}\left(1 - \frac{m_i}{M_i}\right)$$

where

$$s_t^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} \left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N}\right)^2$$

and

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j \in \mathcal{S}_i} \left(y_{ij} - \bar{y}_i\right)^2$$

Estimating the population mean

Wish to estimate the population mean,

$$\bar{y}_U = \frac{t}{M_0} = \frac{\sum_{i=1}^{N} t_i}{\sum_{i=1}^{N} M_i}$$

Method 1: **Unbiased estimation**

An unbiased estimator of $\bar{y}_U$ is

$$\widehat{\bar{y}}_{\text{unb}} = \frac{\hat{t}_{\text{unb}}}{M_0}$$

Its standard error is

$$\mathsf{SE}(\widehat{\bar{y}}_{\mathsf{unb}}) = \frac{\mathsf{SE}(\hat{t}_{\mathsf{unb}})}{M_0}$$

Notes:

1. If M_0 is unknown, we can't do unbiased estimation.

2. If the M_i vary a lot, we don't want to do unbiased estimation.

Method 2: **Ratio estimation**

The ratio estimator of $\bar{y}_U$ is

$$\hat{\bar{y}}_r = \frac{\sum_{i \in \mathcal{S}} \hat{t}_i}{\sum_{i \in \mathcal{S}} M_i} = \frac{\sum_{i \in \mathcal{S}} M_i \bar{y}_i}{\sum_{i \in \mathcal{S}} M_i} = \frac{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij}}$$

Recall the standard error of $\hat{B} = \frac{\bar{y}}{\bar{x}}$ from ratio estimation.

But there are two components now, since there are two sources of variability in the estimate, corresponding to the two stages of cluster sampling.

The standard error is $\mathrm{SE}(\hat{\bar{y}}_r) = \sqrt{\hat{V}}$ as usual.

Here we have

$$\hat{V}(\hat{\bar{y}}_r) = \frac{s_r^2}{n\bar{M}^2}\left(1 - \frac{n}{N}\right) + \frac{1}{nN\bar{M}^2}\sum_{i\in\mathcal{S}} M_i^2 \frac{s_i^2}{m_i}\left(1 - \frac{m_i}{M_i}\right)$$

where

$$s_r^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}} M_i^2 \left(\bar{y}_i - \hat{\bar{y}}_r\right)^2$$

the sample variance of the $M_i(\bar{y}_i - \hat{\bar{y}}_r)$.

Designing a cluster sample

(Section 5.4)

Four issues:

1. What overall precision is needed?
2. How to define the psus?
3. How many ssus should be sampled in each sampled psus?
4. How many psus should be sampled?