Two-Stage Cluster Sampling

Survey Sampling
Statistics 4234/5234
Fall 2018

October 23, 2018

One-stage cluster sampling

Clusters are rarely of equal sizes in social surveys.

An **unbiased** estimator of t is

$$\widehat{t}_{\text{unb}} = \frac{N}{n} \sum_{i \in \mathcal{S}} t_i$$

with standard error

$$\text{SE}(\widehat{t}_{\text{unb}}) = N \frac{s_t}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

The difference between unequal- and equal-sized clusters is that the variation among the individual cluster totals t_i is likely to be large when the clusters have different sizes.

(Expect that t_i is large when the psu size M_i is large, and small when M_i is small.)

Sampling weights

The probabilty that a psu is in the sample is n/N, as an SRS of n of the N psus is taken.

For one-stage cluster sampling, an ssu is included in the sample whenever is psu is included in the sample; thus

$$w_{ij} = \frac{1}{P\{\text{ssu } j \text{ of psu } i \text{ is in sample}\}} = \frac{N}{n}$$

One-stage cluster sampling produces a self-weighting sample when the psus are selected with equal probabilities.

Unbiased estimation

Write

$$\hat{t}_{\text{unb}} = \frac{N}{n}\sum_{i\in\mathcal{S}} t_i = \sum_{i\in\mathcal{S}}\sum_{j=1}^{M_i} w_{ij}y_{ij} = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}$$

And we have an unbiased estimator of $\bar{y}_U$, given by

$$\hat{\bar{y}}_{\text{unb}} = \frac{\hat{t}}{M_0} \quad \text{with} \quad \text{SE}(\hat{\bar{y}}_{\text{unb}}) = \frac{1}{M_0}\text{SE}(\hat{t}_{\text{unb}})$$

In populations with highly variable cluster sizes, the unbiased estimator can be inefficient.

Ratio estimation

Write

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

and this suggests a ratio estimator

$$\hat{\bar{y}}_r = \frac{\hat{t}_{\text{unb}}}{\widehat{M}_0} = \frac{\sum_{i\in\mathcal{S}} t_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{\sum_{i\in\mathcal{S}} M_i \bar{y}_i}{\sum_{i\in\mathcal{S}} M_i}$$

which can also be written

$$\hat{\bar{y}}_r = \frac{\hat{t}_{\text{unb}}}{\hat{M}_0} = \frac{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}}{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}}$$

We need to go back to chapter 4 on ratio estimation to retrieve the standard error formula!

Find

$$\hat{V}(\hat{\bar{y}}_r) = \frac{s_r^2}{n\bar{M}^2}\left(1-\frac{n}{N}\right)$$

where $\bar{M}$ is the sample mean cluster size, and

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

Of course

$$\mathrm{SE}(\hat{\bar{y}}_r) = \sqrt{\hat{V}(\hat{\bar{y}}_r)}$$

Example: Population of 187 high school algebra classes, take an SRS of 12 of those classes, then test every student in a sampled class. This is a one-stage cluster sample.

Two-stage cluster sampling

The stages within a two-stage cluster sample, when we sample the psus and subsample the ssus with equal probabilities, are

1. Select an SRS $\mathcal{S}$ of n psus from the population of N psus.

2. Select an SRS of ssus from each elected psu; the SRS of m_i elements from the ith psu is denoted $\mathcal{S}_i$.

The extra stage complicates the notation and estimators, as we need to consider variability arising from both stages of data collection.

In one-stage cluster sampling we could estimate the population total by $\widehat{t}_{\text{unb}} = (N/n) \sum_{i \in \mathcal{S}} t_i$; the psu totals t_i were known.

In two-stage cluster sampling we need to estimate the individual psu totals, by

$$\widehat{t}_i = \frac{M_i}{m_i} \sum_{j \in \mathcal{S}_i} y_{ij} = M_i \bar{y}_i$$

and thus we estimate the population total by

$$\widehat{t}_{\text{unb}} = \frac{N}{n} \sum_{i \in \mathcal{S}} \widehat{t}_i = \frac{N}{n} \sum_{i \in \mathcal{S}} M_i \bar{y}_i$$

Sampling weights

Let w_{ij} denote the sampling weight for ssu j of psu i.

Then

$$w_{ij} = \frac{1}{P(j\text{th ssu in } i\text{th psu is selected})} = \frac{NM_i}{nm_i}$$

and

$$\widehat{t}_{\text{unb}} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}$$

Standard error

In two-stage sampling, the $\widehat{t}_i$'s are random variables; thus the variance of $\widehat{t}_{\text{unb}}$ has two components: (1) the variability between psus, and (2) the variability of ssus within psus. (In one-stage cluster sampling we had only to worry about the first component.)

For two-stage cluster sampling

$$V(\widehat{t}_{\text{unb}}) = N^2 \frac{S_t^2}{n}\left(1 - \frac{n}{N}\right) + \frac{N}{n}\sum_{i=1}^{N} M_i^2 \frac{S_i^2}{m_i}\left(1 - \frac{m_i}{M_i}\right)$$

where S_t^2 is the population variance of the cluster totals, and S_i^2 is the population variance among the elements within cluster i.

An unbiased estimator of this variance is

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

and the standard error of $\widehat{t}_{\text{unb}}$ is, of course,

$$\text{SE}(\widehat{t}_{\text{unb}}) = \sqrt{\widehat{V}(\widehat{t}_{\text{unb}})}$$

Note the mistake in the text! That N/n term in (5.24) on page 185 should be $(N/n)^2$, as shown above.

Estimating 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

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

and its standard error is

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

Note that (1) if M_0 is unknown you can't use this estimator, and (2) if the M_i are highly variable you probably don't want to use this estimator, since it will have high variance.

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 $\hat{B}$ from ratio estimation, but note that here we have two distinct sources of variability — only a random sample of the psus are observed, and only a random sample of ssus in each psu are observed — and thus two terms in the estimated variance.

We have

$$\text{SE}\left(\hat{\bar{y}}_r\right) = \sqrt{\hat{V}\left(\hat{\bar{y}}_r\right)}$$

where

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

where

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