Cluster Sampling

Survey Sampling
Statistics 4234/5234
Fall 2018

October 18, 2018

Example: Suppose we want to find out how many bicycles are owned by residents in a community of 10,000 households.

- We could take a SRS of 600 households.
- Or we could divide the community into 500 blocks of 20 households each, and take a random sample of just 30 blocks.

Either way we get information on bicycle ownership in 600 out of 10,000 households.

The latter plan is an example of **cluster sampling**.

- The blocks are the **primary sampling units** (psus), or **clusters**.

- The households are the **secondary sampling units** (ssus).

The cluster sample of 600 households is likely to give less precision than an SRS of 600 households.

Why? Because 20 households in the same block are not as likely to mirror the diversity of the community as well as 20 households chosen at random.

However, it is much cheaper and easier to interview all 20 households in a block than 20 households selected at random from the community.

In cluster sampling, the sampling unit (psu) is not the same as the observation unit (ssu), and the two sizes of experimental units must be considered when calculating standard errors.

Why use cluster sampling?

1. Constructing a sampling frame list of observation units may be difficult, expensive, or impossible.

2. The population may be widely distributed geographically, or may occur in natural clusters such as households or schools, and it is less expensive to take a sample of clusters rather than an SRS of individuals.

Clusters bear a superficial resemblance to strata: A cluster, like a stratum, is a grouping of the members of the population. The selection process, though, is quite different in the two methods.

Whereas stratification generally increases precision when compared with SRS, cluster sampling generally decreases it.

Member of the same cluster tend to be more similar than elements selected at random from the whole population.

Notation

Let $y_{ij} =$ measurement for jth element in ith psu

Let $\mathcal{S}$ denote the sample of psus chosen from the population of psus.

Let $\mathcal{S}_i$ denote the sample of ssus chosen from the ith psu for $i \in \mathcal{S}$.

psu-Level: Population Quantities

$$N = \text{number of psus in the population}$$

$$M_i = \text{number of ssus in psu } i$$

$$M_0 = \sum_{i=1}^{N} M_i = \text{total number of ssus in the population}$$

$$t_i = \sum_{j=1}^{M_i} y_{ij} = \text{total in psu } i$$

$$t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \text{population total}$$

$$S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(t_i - \frac{t}{N} \right)^2 = \text{population variance of psu totals}$$

ssu-Level: Population Quantities

$$\bar{y}_U = \frac{t}{M_0} = \frac{1}{M_0} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \text{population mean}$$

$$\bar{y}_{iU} = \frac{t_i}{M_i} = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} = \text{population mean in psu } i$$

$$S^2 = \frac{1}{M_0 - 1} \sum_{i=1}^{N} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_U)^2 = \text{population variance (per ssu)}$$

$$S_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{iU})^2 = \text{population variance within psu } i$$

Sample Quantities

$$
\begin{aligned}
n &= \text{number of psus in the sample} \\
m_i &= \text{number of ssus in the sample from psu } i \\
\bar{y}_i &= \frac{1}{m_i} \sum_{j \in \mathcal{S}_i} y_{ij} = \text{sample mean (per ssu) for psu } i \\
\hat{t}_i &= \frac{M_i}{m_i} \sum_{j \in \mathcal{S}_i} y_{ij} = \text{estimated total for psu } i \\
\hat{t}_{\text{unb}} &= \frac{N}{n} \sum_{i \in \mathcal{S}} \hat{t}_i = \text{unbiased estimator of population total} \\
s_t^2 &= \frac{1}{n-1} \sum_{i \in \mathcal{S}} \left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N} \right)^2 = \text{sample var of estimated psu totals} \\
s_i^2 &= \frac{1}{m_i - 1} \sum_{j \in \mathcal{S}_i} (y_{ij} - \bar{y}_i)^2 = \text{sample variance within psu } i
\end{aligned}
$$

One-stage cluster sampling for clusters of equal size

In one-stage cluster sampling, we take an SRS of n of the N psus from the population and measure our variable on *every* element in the sample psus; that is, $m_i = M_i$.

In today's discussion we will make the simplifying assumption that each psu has the same number of elements, thus $M_i = M$.

Thus we have an SRS of n data points $\{t_i : i \in \mathcal{S}\}$, then the estimated population total reduces to

$$\hat{t} = \frac{N}{n} \sum_{i \in \mathcal{S}} t_i$$

All the earlier results from our study of SRS apply!

In particular $E(\widehat{t}) = t$ and

$$V(\widehat{t}) = N^2 \frac{S_t^2}{n}\left(1 - \frac{n}{N}\right)$$

and thus the standard error of the estimator is given by

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

where

$$s_t^2 = \frac{1}{n-1}\sum_{i \in \mathcal{S}}\left(t_i - \frac{\widehat{t}}{N}\right)^2$$

To estimate $\bar{y}_U$, divide the estimated total by the number of ssus, obtaining

$$\hat{\bar{y}} = \frac{\hat{t}}{NM}$$

This estimator is also unbiased, of course, with

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

and thus the standard error is

$$\mathrm{SE}(\hat{\bar{y}}) = \frac{1}{M}\frac{s_t}{\sqrt{n}}\left(1 - \frac{n}{N}\right)$$

Example: A student wants to estimate the average GPA in his dormitory. The dorm has 100 suites, each with four students; he chooses 5 of those suites at random, and asks every person in the 5 suites what his or her GPA is.

We have $N = 100$, and $M = 4$, thus 400 students total; take cluster sample of $n = 5$ psus.

Let y_{ij} denote the GPA of the jth student in the ith suite

Suppose the five observed values of $t_i = \sum_{j=1}^{4} y_{ij}$ for $i \in \mathcal{S}$ are 12.16, 11.36, 8.96, 12.96, 11.08.

The estimated population total is

$$\hat{t} = \frac{N}{n} \sum_{i \in \mathcal{S}} t_i = \frac{100}{5} (12.16 + \cdots + 11.08) = 1130.4$$

and the sample variance of the cluster totals is

$$\begin{aligned} s_t^2 &= \frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(t_i - \frac{\widehat{t}}{N}\right)^2 \\ &= \frac{1}{4}\left[(12.16 - 11.304)^2 + \cdots + (11.08 - 11.304)^2\right] \\ &= 2.256 \end{aligned}$$

Thus we estimate the average GPA in this dorm by

$$\widehat{\bar{y}} = \frac{\widehat{t}}{NM} = \frac{1130.4}{400} = 2.826$$

and the standard error of our estimate is

$$SE(\widehat{\bar{y}}) = \sqrt{\frac{s_t^2}{nM^2}\left(1 - \frac{n}{N}\right)} = \sqrt{\frac{2.256}{5\cdot 4^2}\left(1 - \frac{5}{100}\right)} = 0.164$$

Sampling weights

One-stage cluster sampling with an SRS of psus produces a self-weighting sample.

The weight for each observation unit is

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

Then

$$\hat{t} = \frac{N}{n} \sum_{i \in \mathcal{S}} t_i = \frac{N}{n} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} y_{ij} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}$$

and

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

Theory: Relative efficiency of cluster sampling vs. SRS

Cluster sampling almost always provides less precision for the estimators than one would obtain by taking an SRS of the same number of elements.

In stratified sampling, the variance of the estimator of t depended on variability *within* the strata; ideally the S_h^2 are small relative to S^2.

In one-stage cluster sampling where each psu has M ssus, the variability of $\hat{t}$ depends entirely on the *between*-psu part of the variability.

Consider an *Analysis of Variance* of the population of psus and ssus:

$$\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_U)^2=\sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{y}_{iU}-\bar{y}_U)^2+\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{iU})^2$$

or

$$(NM-1)S^2=(N-1)\text{MSB}+N(M-1)\text{MSW}$$

and thus

$$S^2=\frac{(N-1)\text{MSB}+N(M-1)\text{MSW}}{NM-1}$$

It can be shown that

$$\frac{V(\hat{t}_{\text{cluster}})}{V(\hat{t}_{\text{SRS}})} = \frac{\text{MSB}}{S^2} = \frac{NM-1}{M(N-1)}\left[1+(M-1)\text{ICC}\right]$$

where

$$\text{ICC} = 1 - \left(\frac{NM}{NM-1}\right)\left(\frac{\text{MSW}}{S^2}\right)$$

is called the **intraclass correlation coefficient**.

The ICC provides a **measure of homogeneity** within the clusters — it tells us how similar elements within the clusters are — and satisfies

$$-\frac{1}{M-1} \le \text{ICC} \le 1$$

If the clusters occur naturally in the population, the ICC is usually positive — elements within the same cluster tend to be more similar than elements selected at random from the population.

The ICC is negative if elements within a cluster are dispersed *more* than a randomly chosen group would be; in this case, cluster sampling is more efficient than simple random sampling of elements!

Example: $N = 3$ and $M = 3$.

A best case scenario for cluster sampling is:

psu 1 is $\{10, 20, 30\}$, psu 2 is $\{11, 20, 32\}$, psu 3 is $\{9, 17, 31\}$.

A worst case scenario would be:

psu 1 is $\{9, 10, 11\}$, psu 2 is $\{17, 20, 20\}$, psu 3 is $\{31, 32, 30\}$.