Chapter 6. Cluster Sampling

1 Definition

Definition: A *cluster sample* is a probability sample in which each sampling unit is a collection, or cluster of elements.

2 Why cluster sampling?

To illustrate, suppose we wish to estimate the average income per household in a large city. How should we choose the sample? If we use the simple random sampling, we will need a frame listing all households in the city, and this frame may be very costly or impossible to obtain. We can not avoid this problem by using stratified random sampling because a frame is still required for each stratum in the population. Rather than draw a simple random sample of elements, we could divide the city into regions such as blocks or clusters of elements and select a simple random sample of blocks form the population. This task is easily accomplished by using a frame that lists all city blocks. Then the income of every household within each sampled block could be measured.

To illustrate the second reason, suppose that a list of households in the city is available. We could select a s.r.s. of households, which probably would be scattered throughout the city. The cost of conduct interviews in the scattered households would be large owing to the interviewer travel time and other related expenses. Stratified random sampling could lower these expenses, but using cluster sampling is a more effective method of reducing travel costs. Elements within a cluster should be close to each other geographically, and hence travel expenses should be reduced. Obviously, travel within a city block would be minimal when the travel associated with s.r.s. of households within the city.

To summarize,

1. A good frame of population elements is not available or costly to obtain; but a list of clusters is available.
2. It saves cost when distance separating the elements increases.

3 How to draw a cluster sample

Suppose that our population consists of N clusters of elements

$$\text{Cluster 1: } u_{11}, \cdots, u_{1M_1} \qquad (\text{with subtotal } u_1 = \textstyle\sum_{j=1}^{M_1} u_{1j})$$

Cluster 2: $u_{21}, \cdots, u_{2M_2}$. (with subtotal $u_2 = \sum_{j=1}^{M_2} u_{2j}$)

.....

Cluster N: $u_{N1}, \cdots, u_{NM_N}$. (with subtotal $u_N = \sum_{j=1}^{M_N} u_{Nj}$)

where

$$\begin{aligned} N &= \text{the number of clusters in the population} \\ M_i &= \text{the number of elements in cluster } i,\ i = 1, 2, \cdots, N \\ M &= \sum_{i=1}^{N} M_i = \text{the number of elements in the population} \\ \overline{M} &= M/N = \text{the average cluster size for the population} \\ u_i &= \sum_{j=1}^{M_i} u_{ij} = \text{the subtotal for the } i\text{th cluster.} \end{aligned}$$

Note that the population mean can be written as

$$(3.1)\quad \mu = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} u_{ij}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N} u_i}{M} = \frac{\sum_{i=1}^{N} u_i/N}{M/N} = \frac{\bar{u}}{\bar{M}} \qquad \text{(a population ratio)}$$

Cluster sampling is simple random sampling from the above N clusters with each element being a cluster. We shall write the cluster sample as

Cluster 1: $y_{11}, \cdots, y_{1m_1}$. (with subtotal $y_1 = \sum_{j=1}^{m_1} y_{1j}$)

Cluster 2: $y_{21}, \cdots, y_{2m_2}$. (with subtotal $y_2 = \sum_{j=1}^{m_2} y_{2j}$)

.....

Cluster n: $y_{n1}, \cdots, y_{nm_n}$. (with subtotal $y_n = \sum_{j=1}^{m_n} y_{nj}$)

where

$$\begin{aligned} n &= \text{the number of clusters selected in a simple random sample} \\ m_i &= \text{the number of elements in cluster } i,\ i = 1, 2, \cdots, n \\ m &= \sum_{i=1}^{n} m_i = \text{the number of elements in the sample} \\ \overline{m} &= \frac{m}{n} = \frac{1}{n}\sum_{i=1}^{n} m_i = \text{the average cluster size for the sample} \\ y_i &= \sum_{j=1}^{m_i} y_{ij} = \text{the total of all observations in cluster } i,\ i = 1, 2, \cdots, n \end{aligned}$$

4 Estimation of population mean

Note that μ in (3.1) is a ratio, which can be estimated by a ratio estimator.

1. The estimator of the population mean μ is

$$\hat{\mu}_c = \frac{\overline{y}}{\overline{m}}.$$

Notice m_i plays the role of x_i in the ratio estimation in the last chapter.

Example A sociologist wants to estimate the per-capita income in a certain small city. No list of resident adults is available. How should he design the sample survey?

Solution Cluster sampling seems to be logical for the survey design because no lists of elements are available. The city is marked off into rectangular blocks, except two industrial areas and three parks that contain only a few houses. The sociologist decides that each of the city blocks will be considered one cluster, the two industrial areas will be considered one cluster, and finally, the three parks will be considered one cluster.

The clusters are numbered on a city map, with the numbers from 1 to 415. The experimenter has enough time and money to sample $n = 25$ clusters and to interview every household within each cluster. Hence 25 random numbers between 1 and 415 are selected from random number tables, and the clusters having these numbers are marked on the map. Interviews are then assigned to each of the sampled clusters.

If we use stratified sampling, then we have 415 subpopulation. And from each subpopulation, we get a s.r.s.

2. The variance of $\hat{\mu}_c$ is given by

$$Var(\hat{\mu}_c) \approx \frac{1}{\overline{M}^2} \frac{\sigma^2_{y-\mu m}}{n} \left(\frac{N-n}{N-1} \right),$$

where

$$\begin{aligned}
\sigma^2_{y-\mu m} &\equiv \frac{1}{N} \sum_{i=1}^{N} (u_i - \mu M_i)^2 \\
(\text{since} \quad & P\left((y,m) = (u_i, M_i)\right) = 1/N, \ i = 1, \cdots, N) \\
&= \frac{1}{N} \sum_{i=1}^{N} (u_i - \bar{u} - \mu(M_i - \bar{M}))^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} \left((u_i - \bar{u})^2 - 2\mu(u_i - \bar{u})(M_i - \bar{M}) + \mu^2 (M_i - \bar{M})^2 \right) \\
&= \left(\sigma_y^2 + \mu^2 \sigma_m^2 - 2\mu\rho\sigma_m\sigma_y \right).
\end{aligned}$$

where

$$\rho = corr(y_1, m_1) = \frac{\frac{1}{N}\sum_{i=1}^{N}(u_i - \bar{u})(M_i - \bar{M})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(u_i - \bar{u})^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(M_i - \bar{M})^2}}$$

and $\sigma_m^2 = \frac{1}{N}\sum_{i=1}^{N}(M_i - \bar{M})^2$, is the population variance of m_i, $i = 1, \cdots, N$.

3. An estimate of $\mathrm{Var}(\hat{\mu}_c)$ is given by

$$\begin{aligned} \widehat{Var}(\hat{\mu}_c) &= \frac{1}{\overline{M}^2}\frac{s^2_{y-\hat{\mu}_c m}}{n}(1-f) \qquad \text{if } \overline{M} \text{ is known} \\ &\approx \frac{1}{\overline{m}^2}\frac{s^2_{y-\hat{\mu}_c m}}{n}(1-f) \qquad \text{if } \overline{M} \text{ is unknown} \end{aligned}$$

where

$$\begin{aligned} s^2_{y-\hat{\mu}_c m} &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\mu}_c m_i)^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 + \hat{\mu}_c^2 \sum_{i=1}^{n} m_i^2 - 2\hat{\mu}_c \sum_{i=1}^{n} m_i y_i\right) \\ &= s_y^2 + \hat{\mu}_c^2 s_m^2 - 2\hat{\mu}_c \hat{\rho} s_m s_y, \end{aligned}$$

with

$$s_m^2 = \frac{1}{n-1}\sum_{i=1}^{n}(m_i - \overline{m})^2, \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2, \quad \widehat{\rho} = \frac{1}{n-1}\sum_{i=1}^{n}(m_i - \overline{m})(y_i - \overline{y})/s_m s_y.$$

4. A $(1-\alpha)$ C.I. for μ is

$$\hat{\mu}_c \mp z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\mu}_c)}.$$

5. The sample size n required to estimate μ with error bound B with probability $1-\alpha$ is

$$n \approx \frac{N\sigma^2_{y-\mu m}}{ND + \sigma^2_{y-\mu m}}, \qquad \text{where} \quad D = \frac{B^2\overline{M}^2}{z^2_{\alpha/2}} \quad \text{or} \quad D = \frac{B^2\overline{m}^2}{z^2_{\alpha/2}}.$$

(In practice, $\sigma^2_{y-\mu m}$ may need to be estimated by $s^2_{y-\bar{\mu}_c m}$ from a pilot study.)

The derivation is sketched as follows,

From

$$B = z_{\alpha/2}\sqrt{Var(\hat{\mu}_c)} = z_{\alpha/2}\sqrt{\frac{1}{\overline{M}^2}\frac{\sigma^2_{y-\mu m}}{n}\left(\frac{N-n}{N-1}\right)}$$

we have

$$D = \frac{\sigma^2_{y-\mu m}}{N-1}\left(\frac{N-n}{n}\right),$$

from which we have

$$n \approx \frac{N\sigma^2_{y-\mu m}}{(N-1)D + \sigma^2_{y-\mu m}} \approx \frac{N\sigma^2_{y-\mu m}}{ND + \sigma^2_{y-\mu m}}.$$

You may also use $\widehat{Var}(\hat{\mu}_c)$.

Example A sociologist wants to estimate the per-capita income in a certain small city. Interviews are conducted in each of the 25 blocks. The data on incomes are presented in the table. Use the data to estimate the per-capita income in the city and place a bound on the error of estimation.

Cluster	Number of residents	Total income per cluster
1	8	$ 96,000
2	12	121,000
3	4	42,000
4	5	65,000
5	6	52,000
6	6	40,000
7	7	75,000
8	5	65,000
9	8	45,000
10	3	50,000
11	2	85,000
12	6	43,000
13	5	54,000
14	10	49,000
15	9	53,000
16	3	50,000
17	6	32,000
18	5	22,000
19	5	45,000
20	4	37,000
21	6	51,000
22	8	30,000
23	7	39,000
24	3	47,000
25	8	41,000

Solution

$$\hat{\mu}_c = \frac{\sum_{i=1}^{25} y_i}{\sum_{i=1}^{25} m_i} = \frac{1,329,000}{151} = 8801$$

$$\begin{aligned} \widehat{Var}(\hat{\mu}_c) &= \frac{N-n}{Nn\overline{m}^2} \frac{\sum_{i=1}^{25}(y_i - \hat{\mu}_c m_i)^2}{n-1} \\ &= \frac{415-25}{415 \times 25 \times 6.04^2} \times 25189^2 = 653785 \end{aligned}$$

The error bound $B = 2\sqrt{\widehat{Var}(\hat{\mu}_c)} = 1617$, which is rather large. It could be reduced by sampling more clusters and consequently, increasing the sample size.

5 Estimation of population total

1. The estimator of the population total τ is

$$\hat{\tau}_c = M \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} = M\hat{\mu}_c.$$

2. The variance of $\hat{\tau}_c$ is given by

$$Var(\hat{\tau}_c) = M^2 Var(\hat{\mu}_c) \approx M^2 \frac{1}{\overline{M}^2} \frac{\sigma^2_{y-\mu m}}{n} \left(\frac{N-n}{N-1}\right) = N^2 \frac{\sigma^2_{y-\mu m}}{n} \left(\frac{N-n}{N-1}\right).$$

3. An estimate of $\text{Var}(\hat{\tau}_c)$ is given by

$$\widehat{Var}(\hat{\tau}_c) = N^2 \frac{s^2_{y-\hat{\mu}_c m}}{n}(1-f).$$

4. A $(1-\alpha)$ C.I. for τ is

$$\hat{\tau}_c \mp z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\tau}_c)}.$$

5. The sample size n required to estimate τ with error bound B with probability $1-\alpha$ is

$$n \approx \frac{N\sigma^2_{y-\mu m}}{ND + \sigma^2_{y-\mu m}}, \qquad \text{where} \quad D = \frac{B^2}{N^2 z^2_{\alpha/2}}.$$

Example Use the data in the last example to estimate the total income of all residents of the city and place a bound on the error of estimation. There are 2500 residents in the city.

Solution The sample mean $\hat{\mu}_c = 8801$ from the last example. Thus

$$\hat{\tau} = M\hat{\mu}_c = 2500 \times 8801 = 22,002,500$$

The error bound is

$$B = 2\sqrt{M^2 Var(\hat{\mu}_c)} = 2\sqrt{2500^2 \times 653,785} = 4042848$$

So the error bound is large, it could be reduced by increasing the sample size.

6 Comparison to simple random sampling under equal cluster sizes

THEOREM **6.1** *If all cluster sizes are equal, then*

$$\widehat{RE}\left(\frac{\widehat{\mu}_c}{\widehat{\mu}_{srs}}\right) > 1 \qquad \textit{iff} \qquad MSW > MSB.$$

That is, cluster sampling is more efficient than s.r.s. if the clusters are similar and the variations within each cluster is big. (MSW=mean square error within the cluster, MSB=mean square error between the clusters)

Proof. Since all cluster sizes are equal, we have

$$M_1 = \cdots = M_N = \overline{M}, \qquad m_1 = \cdots = m_n = \overline{m}, \qquad \overline{M} = \overline{m}.$$

Our cluster sample is

Cluster 1: $y_{11}, \cdots, y_{1m_1}$, $\quad \overline{y}_1 = \sum_{j=1}^{\overline{m}} y_{1j}/m_1 = y_1/\overline{m}$

Cluster 2: $y_{21}, \cdots, y_{2m_2}$ $\quad \overline{y}_2 = \sum_{j=1}^{\overline{m}} y_{2j}/m_2 = y_2/\overline{m}$

.....

Cluster n: $y_{n1}, \cdots, y_{nm_n}$ $\quad \overline{y}_n = \sum_{j=1}^{m} y_{nj}/m_n = y_n/\overline{m}$

Define

$$SST = \sum_{i=1}^{n}\sum_{j=1}^{\overline{m}}(y_{ij} - \widehat{\mu}_c)^2, \qquad MST = SST/(\overline{m}n - 1),$$

(one constraint $\sum_{i=1}^{n}\sum_{j=1}^{\overline{m}} y_{ij}/(n\overline{m}) = \widehat{\mu}_c$)

$$SSW = \sum_{i=1}^{n}\sum_{j=1}^{\overline{m}}(y_{ij} - \overline{y}_i)^2, \qquad MSW = SSW/(\overline{m}n - n),$$

(n constraints $\sum_{j=1}^{\overline{m}} y_{ij}/\overline{m} = \overline{y}_i$)

$$SSB = \sum_{i=1}^{n}\sum_{j=1}^{\overline{m}}(\overline{y}_i - \widehat{\mu}_c)^2, \qquad MSB = SSB/(n-1).$$

$$= \sum_{i=1}^{n}\overline{m}(\overline{y}_i - \widehat{\mu}_c)^2$$

By the ANOVA decomposition (which we met in systematic sampling), we have

$$SST = SSW + SSB.$$

For cluster sampling, we have

$$\begin{aligned}
\widehat{\mu}_c &= \frac{\sum_{i=1}^n y_i}{\overline{m}n} = \frac{\overline{y}}{\overline{m}}, \\
s^2_{y-\hat{\mu}_c m} &= \frac{1}{n-1}\sum_{i=1}^n (y_i - \hat{\mu}_c m_i)^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \overline{y})^2, \\
\widehat{Var}(\widehat{\mu}_c) &= \frac{1}{\overline{m}^2}\frac{1}{n} s^2_{y-\hat{\mu}_c m}(1-f) \\
&= \frac{1}{\overline{m}^2 n}\Big[\frac{1}{n-1}\sum_{i=1}^n (y_i - \overline{y})^2\Big](1-f) \\
&= \frac{1}{\overline{m}^2 n}\Big[\frac{1}{n-1}\sum_{i=1}^n (\overline{m}\ \overline{y}_i - \overline{m}\ \widehat{\mu}_c)^2\Big](1-f) \\
&= \frac{1}{\overline{m}n(n-1)} SSB(1-f) \\
&= \frac{1}{\overline{m}n} MSB(1-f).
\end{aligned}$$

If we regard the cluster sample as a simple random sample, i.e. $y_{11}, \cdots, y_{1m_1}$, $y_{21}, \cdots, y_{2m_2}$, $\cdots$, $y_{n1}, \cdots, y_{nm_n}$ is a s.r.s., we have $\hat{\mu}_c = \widehat{\mu}_{srs}$, and

$$\widehat{Var}(\widehat{\mu}_{srs}) = \frac{1}{\overline{m}n}\Big[\frac{1}{\overline{m}n-1}\sum_{i=1}^n\sum_{j=1}^{\bar{m}} (y_{ij} - \hat{\mu}_c)^2\Big](1-f) = \frac{1}{\overline{m}n} MST(1-f).$$

Therefore,

$$\begin{aligned}
\widehat{RE}\left(\frac{\widehat{\mu}_c}{\widehat{\mu}_{srs}}\right) &= \frac{\widehat{Var}(\widehat{\mu}_{srs})}{\widehat{Var}(\widehat{\mu}_c)} = \frac{MST}{MSB} \\
&= \frac{(SSW + SSB)/(\overline{m}n - 1)}{MSB} \\
&= \frac{[n(\overline{m}-1)MSW + (n-1)MSB]}{(\overline{m}n-1)MSB} \\
&\approx \frac{[n(\overline{m}-1)MSW + nMSB]}{\overline{m}nMSB} \\
&= \frac{(\overline{m}-1)MSW + MSB}{\overline{m}MSB}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\widehat{RE}\left(\frac{\widehat{\mu}_c}{\widehat{\mu}_{srs}}\right) > 1 \qquad & \text{iff} \quad (\overline{m}-1)MSW + MSB > \overline{m}MSB \\
& \text{iff} \quad (\overline{m}-1)MSW > (\overline{m}-1)MSB \\
& \text{iff} \quad MSW > MSB.
\end{aligned}$$

That is, cluster sampling is more efficient than s.r.s. if the clusters are similar and the variations within each cluster is big.

Remark: From the above theorem, we see that *cluster sampling* and *the stratified random sampling* are quite opposite sampling schemes. Which one we pick in practice depends on the variations amongst subgroups (either clusters or strata) and the variations between subgroups. The following is a guideline.

(a). Use the stratified random sampling if the subgroups are quite different but within each subgroup, the elments are similar (or homogeneous). (i.e. $MSB > MSW$).

(b). Use the cluster sampling if the subgroups are similar but within each subgroup, the elments are quite different. (i.e. $MSB < MSW$).

7 A special case

Suppose $M_1 = M_2 = \cdots = M_N = \bar{M}$. Then $m_1 = m_2 = \cdots = m_n = \bar{M}$, and

$$\hat{\mu}_c = n^{-1} \sum_{i=1}^{n} \bar{y}_i.$$

Let us calculate $E\hat{\mu}_c$ and $V(\hat{\mu}_c)$.

Since $\{\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_n\}$ is a s.r.s. from $\{\bar{u}_1, \cdots, \bar{u}_N\}$, where $\bar{u}_i = u_i/\bar{M}$,

$$E\hat{\mu}_c = \mu, \quad V(\hat{\mu}_c) = \frac{\sigma_b^2}{n} \frac{N-n}{N-1},$$

where

$$\sigma_b^2 = N^{-1} \sum_{i=1}^{N} (\bar{u}_i - \mu)^2.$$

From the last section,

$$\widehat{V}(\widehat{\mu}_c) = \frac{1}{\bar{m}n} MSB(1-f) = n^{-1}(n-1)^{-1} \sum_{i=1}^{n} (\bar{y}_i - \hat{\mu}_c)^2 (1-f).$$

So

$$E\widehat{V}(\widehat{\mu}_c) = n^{-1} E(n-1)^{-1} \sum_{i=1}^{n} (\bar{y}_i - \hat{\mu}_c)^2 (1-f) = n^{-1} \frac{N}{N-1} \sigma_b^2 (1-f) = \frac{\sigma_b^2}{n} \frac{N-n}{N-1}.$$

In other words, $\widehat{V}(\widehat{\mu}_c)$ is an unbiased estimator of $V(\hat{\mu}_c)$ in this special case.