Monday, Oct 7

Cluster Sampling

A *cluster sampling* design is one where some *sampling units* include *more than one element*.

Steps of a cluster sampling design:

1. Partition the M elements in the population into N clusters.
2. Select n clusters using a probability sampling design (e.g., simple random sampling).
3. Observe the target variable for *all* elements in the sampled clusters.

Note: This is what is called *one-stage* cluster sampling. Later we will discuss *two-stage* and *multi-stage* cluster sampling. But until we discuss these designs, it will be implied that we are referring to a *one-stage* design.

Example: Consider the following sampling design. The population, sampling units, and their respective sizes are as follows.

$$\mathcal{P} = \{\underbrace{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3}_{\mathcal{U}_1}, \underbrace{\mathcal{E}_4, \mathcal{E}_5, \mathcal{E}_6, \mathcal{E}_7}_{\mathcal{U}_2}, \underbrace{\mathcal{E}_8, \mathcal{E}_9, \mathcal{E}_{10}, \mathcal{E}_{11}, \mathcal{E}_{12}}_{\mathcal{U}_3}\},\ N = 3,\ M = 12$$

$$\begin{aligned}
\mathcal{U}_1 &= \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3\},\ m_1 = 3 \\
\mathcal{U}_2 &= \{\mathcal{E}_4, \mathcal{E}_5, \mathcal{E}_6, \mathcal{E}_7\},\ m_2 = 4 \\
\mathcal{U}_3 &= \{\mathcal{E}_8, \mathcal{E}_9, \mathcal{E}_{10}, \mathcal{E}_{11}, \mathcal{E}_{12}\},\ m_3 = 5
\end{aligned}$$

Note that m_i denotes the number of elements in the i-th sampling unit or *cluster*.

If we were to apply simple random sampling to these sampling units to select $n = 2$ clusters, the possible samples and their probabilities are as follows.

$$\begin{aligned}
\mathcal{S}_1 &= \{\mathcal{U}_1, \mathcal{U}_2\} = \{\underbrace{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3}_{\mathcal{U}_1}, \underbrace{\mathcal{E}_4, \mathcal{E}_5, \mathcal{E}_6, \mathcal{E}_7}_{\mathcal{U}_2}\}, P(\mathcal{S}_1) = 1/3 \\
\mathcal{S}_2 &= \{\mathcal{U}_1, \mathcal{U}_3\} = \{\underbrace{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3}_{\mathcal{U}_1}, \underbrace{\mathcal{E}_8, \mathcal{E}_9, \mathcal{E}_{10}, \mathcal{E}_{11}, \mathcal{E}_{12}}_{\mathcal{U}_3}\}, P(\mathcal{S}_2) = 1/3 \\
\mathcal{S}_3 &= \{\mathcal{U}_2, \mathcal{U}_3\} = \{\underbrace{\mathcal{E}_4, \mathcal{E}_5, \mathcal{E}_6, \mathcal{E}_7}_{\mathcal{U}_2}, \underbrace{\mathcal{E}_8, \mathcal{E}_9, \mathcal{E}_{10}, \mathcal{E}_{11}, \mathcal{E}_{12}}_{\mathcal{U}_3}\}, P(\mathcal{S}_3) = 1/3.
\end{aligned}$$

This would be one possible cluster sampling design.

Examples of sampling units, elements, and target variables where cluster sampling might be used.

Sampling Unit	Element	Target Variable
box	widget	weight
block	household	income
county	farm	acres of wheat
classroom	student	test score
hour	minute	number of fish
plot	tree	volume

What are the potential *advantages* of cluster sampling (relative to SRS)?

What are the potential *disadvantages* of cluster sampling (relative to SRS)?

How is cluster sampling different from *stratified* random sampling?[1]

How could we view *simple random sampling* as a special case of cluster sampling?

[1] Note: The symbol $\bar{y}_U$ in that diagram is the population mean (μ).

Notation

Let y_{ij} be the value of the target variable for the j-th element in the i-th cluster, and let y_i be the *sum* of the m_i values of the target variable for all the elements in the i-th cluster so that

$$y_i = \sum_{j=1}^{m_i} y_{ij}.$$

Example: Suppose a population of $M = 12$ elements are partitioned into $N = 3$ clusters as follows.

i	m_i	y_{ij}	y_i
1	3	y_{11}, y_{12}, y_{13}	$y_1 = y_{11} + y_{12} + y_{13}$
2	4	$y_{21}, y_{22}, y_{23}, y_{24}$	$y_2 = y_{21} + y_{22} + y_{23} + y_{24}$
3	5	$y_{31}, y_{32}, y_{33}, y_{34}, y_{35}$	$y_3 = y_{31} + y_{32} + y_{33} + y_{34} + y_{35}$

Note that the three clusters have *sizes* of $m_1 = 3$, $m_2 = 4$, and $m_3 = 5$.

The mean and total of a target variable for all elements in the population can be computed as

$$\mu = \frac{1}{M}\sum_{i=1}^{N} y_i \quad \text{and} \quad \tau = \sum_{i=1}^{N} y_i,$$

respectively, noting that $M = \sum_{i=1}^{N} m_i$ is the number of elements in the population.

Note: Much of the estimation theory of cluster sampling (assuming simple random sampling of clusters) is essentially treating the *clusters* as *elements* where the cluster total y_i is the target variable. But one key difference is that $\mu = \tau/M$ and not $\mu = \tau/N$.

Estimation of μ

Note that we can write μ as

$$\mu = \frac{1}{M}\sum_{i=1}^{N} y_i = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} m_i}.$$

We can also write this as

$$\mu = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i}{\frac{1}{N}\sum_{i=1}^{N} m_i} = \frac{\mu_y}{\mu_m},$$

where μ_y is the *mean cluster total across all clusters* (which is *not* necessarily the same as μ) and μ_m is the *mean cluster size across all clusters* (which is also called $\bar{M}$ and equals M/N). Assuming simple random sampling of clusters, this suggests that we might estimate μ with

$$\hat{\mu} = \frac{\bar{y}}{\bar{m}} = \frac{\frac{1}{n}\sum_{i\in\mathcal{S}} y_i}{\frac{1}{n}\sum_{i\in\mathcal{S}} m_i} = \frac{\sum_{i\in\mathcal{S}} y_i}{\sum_{i\in\mathcal{S}} m_i},$$

(i.e., the ratio of the totals of the clusters totals and the cluster sizes *for the sampled clusters*). Where have we seen this kind of estimator before?

Example: A cluster sampling design selects $n = 3$ boxes using simple random sampling of the boxes. The number of widgets in these boxes are $m_1 = 3$, $m_2 = 4$, and $m_3 = 5$. The total weight of the widgets in these boxes are $y_1 = 6.2$, $y_2 = 7.5$, and $y_3 = 10.3$. What is the estimate of μ?

Variance of the Estimator of μ

The estimated variance of the estimator

$$\hat{\mu} = \frac{\sum_{i \in \mathcal{S}} y_i}{\sum_{i \in \mathcal{S}} m_i},$$

assuming simple random sampling of clusters, is

$$\hat{V}(\hat{\mu}) = \frac{1}{\bar{M}^2}\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n} \quad \text{where} \quad s_r^2 = \frac{\sum_{i \in \mathcal{S}}(y_i - \hat{\mu} m_i)^2}{n-1},$$

where $\bar{m}$ can be used in place of $\bar{M}$ if it is unknown.

Example: Assume that in the previous example that there are a total of 100 boxes, and that the total number of widgets in all those boxes is 425. What is the variance and bound on the error of estimation for $\hat{\mu}$?

An Alternative Estimator of μ?

Recall that

$$\mu = \frac{\mu_y}{\mu_m},$$

where μ_y is the *mean cluster total across all clusters* and μ_m is the *mean cluster size across all clusters*. If we *know* μ_m we might use it instead of $\bar{m}$ and therefore use the estimator

$$\hat{\mu} = \frac{\bar{y}}{\mu_m}$$

instead of the estimator introduced earlier which can be written as $\hat{\mu} = \bar{y}/\bar{m}$. These estimators are not equivalent *unless all clusters are of the same size* (in which case $\mu_m = \bar{m}$). Should we use this alternative estimator? Probably not. Why? Consider our discussion of two estimators of a ratio of totals.