Wednesday, October 29

Two-Stage Cluster Sampling

The type of cluster design we have already discussed is called a *one-stage* cluster sampling design, which can be described as having the following steps.

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

A *two-stage* cluster sampling design can be described as follows.

1. Partition the M elements in a population into N sampling units (i.e., clusters).
2. Select n clusters using a probability sampling design (e.g., SRS or PPS).
3. For each sampled cluster, *sample* elements using a probability sampling design (usually SRS), and observe the target variable for these sampled elements.

Additional notation is necessary for two-stage cluster sampling designs. Let M_i denote the number of elements in the i-th cluster, and let m_i denote the number of *sampled* elements in the i-th cluster. So $m_i \leq M_i$ for all clusters, and $m_i < M_i$ for at least some of the clusters for a two-stage cluster sampling design.

Note: Do not confuse "two-stage" sampling with "two-phase" sampling — the latter is sometimes used describe double sampling designs. While there are some similarities between two-stage and two-phase sampling these are largely superficial.

Example: Consider the following sampling design. The population, clusters, 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 there are actually *two* kinds of sampling units here, the sampling units for the first stage (i.e., clusters), and the sampling units for the second stage (i.e., the elements within the clusters). These are called the *primary* and *secondary* sampling units, respectively. These are denoted above as $\mathcal{U}$ and $\mathcal{E}$, respectively.

Example: What would be some possible samples for a two-stage cluster sampling design with $n = 2$ and $m_1 = 2$, $m_2 = 2$, $m_3 = 3$ (assuming SRS for both stages) using the population shown above?

Note that the table above is not *all possible* samples (there are 108 possible samples for this design).

Clusters	Elements
$\mathcal{U}_1, \mathcal{U}_3$	$\mathcal{E}_2, \mathcal{E}_3; \mathcal{E}_9, \mathcal{E}_{10}, \mathcal{E}_{11}$
$\mathcal{U}_1, \mathcal{U}_2$	$\mathcal{E}_1, \mathcal{E}_2; \mathcal{E}_5, \mathcal{E}_7$
$\mathcal{U}_1, \mathcal{U}_2$	$\mathcal{E}_2, \mathcal{E}_3; \mathcal{E}_4, \mathcal{E}_6$
$\mathcal{U}_1, \mathcal{U}_3$	$\mathcal{E}_1, \mathcal{E}_3; \mathcal{E}_9, \mathcal{E}_{10}, \mathcal{E}_{12}$
$\mathcal{U}_2, \mathcal{U}_3$	$\mathcal{E}_4, \mathcal{E}_5; \mathcal{E}_9, \mathcal{E}_{10}, \mathcal{E}_{11}$
$\mathcal{U}_1, \mathcal{U}_3$	$\mathcal{E}_1, \mathcal{E}_3; \mathcal{E}_9, \mathcal{E}_{10}, \mathcal{E}_{11}$
$\mathcal{U}_1, \mathcal{U}_3$	$\mathcal{E}_1, \mathcal{E}_3; \mathcal{E}_8, \mathcal{E}_9, \mathcal{E}_{12}$
$\mathcal{U}_1, \mathcal{U}_2$	$\mathcal{E}_1, \mathcal{E}_3; \mathcal{E}_4, \mathcal{E}_5$
$\mathcal{U}_1, \mathcal{U}_3$	$\mathcal{E}_1, \mathcal{E}_2; \mathcal{E}_8, \mathcal{E}_{10}, \mathcal{E}_{12}$

Primary and Secondary Sampling Units

Two-stage cluster sampling designs involve two stages/levels of sampling units: **primary sampling units (PSU)** and **secondary sampling units (SSU)**. The primary sampling units are the clusters. The secondary sampling units are the elements within the clusters.

Sampling Unit		
Primary	Secondary	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

Relationship With Other Designs

Consider describing a two-stage cluster sampling design in terms of five quantities:

1. The number of elements in the population (M).
2. The number of clusters/PSUs (N).
3. The number of sampled clusters/PSUs (n).
4. The number of elements/SSUs in each cluster/PSU ($M_1, M_2, \ldots, M_N$).
5. The number of sampled elements/SSUs sampled from each cluster/PSU ($m_1, m_2, \ldots, m_n$).

How is a *one-stage cluster sampling design* a special case?

How is a *stratified sampling design* a special case?

Advantages and Disadvantages Relative to Other Designs

What are the advantages and disadvantages of two-stage cluster sampling relative to *simple random sampling*?

What advantage does two-stage cluster sampling having relative to *one-stage cluster sampling*?

Estimators of τ and μ for Two-Stage Cluster Sampling

Assuming we are using simple random sampling at *both* stages, how might we estimate τ and μ?

Unbiased Estimators

Recall that with a *one-stage* cluster sampling design, the "unbiased estimator" of τ is

$$\hat{\tau} = \frac{N}{n} \sum_{i \in \mathcal{S}} y_i,$$

where y_i is the *total* of the target variable for all the elements in the i-th cluster. In a two-stage cluster sampling design we do not know y_i if we do not sample all elements in the cluster, but we can *estimate* it with the "expansion estimator" $M_i \bar{y}$, where $\bar{y}_i$ is the *mean of the target variable for the sampled elements from the i-th cluster*. So the "unbiased" estimator of τ for a two-stage cluster sampling design is

$$\hat{\tau} = \frac{N}{n} \sum_{i \in \mathcal{S}} M_i \bar{y}_i.$$

The "unbiased" estimator of μ is then obtained by dividing this estimator by M to get

$$\hat{\mu} = \frac{N}{Mn} \sum_{i \in \mathcal{S}} M_i \bar{y}_i.$$

Note that the set $\mathcal{S}$ denotes the indices of the set of sampled clusters.

Example: A two-stage cluster sampling design selects $n = 3$ boxes using simple random sampling. These boxes each contain 10, 12, and 5 widgets, but we select 2, 3, and 1 widget(s) from these boxes, respectively, using simple random sampling. The mean weight of the sampled widgets from each box are 2.2, 1.9, and 2.1, respectively. Assume that there are $N = 100$ boxes of widgets in the warehouse, containing a total of $M = 425$ widgets. What are the estimates of τ and μ using the unbiased estimators?

The variance of $\hat{\tau}$ is

$$V(\hat{\tau}) = \underbrace{N^2 \left(1 - \frac{n}{N}\right) \frac{\sigma_c^2}{n}}_{\text{first stage of sampling}} + \underbrace{\frac{N}{n} \sum_{i=1}^{N} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{\sigma_i^2}{m_i}}_{\text{second stage of sampling}}$$

where σ_c^2 and σ_i^2 are the variances of the cluster totals and within the i-th cluster, respectively, defined as

$$\sigma_c^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \tau/N)^2 \quad \text{and} \quad \sigma_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \mu_i)^2,$$

where μ_i is the mean of all the elements in the i-th cluster. The variance of $\hat{\mu}$ is $V(\hat{\mu}) = V(\hat{\tau})/M^2$. Modified formulas (not shown) can be used to *estimate* these variances.

Ratio Estimators

Recall that for a *one-stage* cluster sampling design, the "ratio estimator" of τ is

$$\hat{\tau} = M \frac{\sum_{i \in \mathcal{S}} y_i}{\sum_{i \in \mathcal{S}} m_i}$$

where y_i is again the *total* of the target variable of the elements in the i-th cluster, and m_i is the number of elements in the i-th cluster (which we would call M_i in the context of two-stage cluster sampling). For a *two-stage cluster sampling design* we estimate y_i with $M_i \bar{y}_i$ and replace m_i with M_i to get

$$\hat{\tau} = M \frac{\sum_{i \in \mathcal{S}} M_i \bar{y}_i}{\sum_{i \in \mathcal{S}} M_i}.$$

The corresponding estimator of μ is then

$$\hat{\mu} = \frac{\sum_{i \in \mathcal{S}} M_i \bar{y}_i}{\sum_{i \in \mathcal{S}} M_i}.$$

Example: A two-stage cluster sampling design selects $n = 3$ boxes using simple random sampling. These boxes each contain 10, 12, and 5 widgets, but we select 2, 3, and 1 widget(s) from these boxes, respectively, using simple random sampling. The mean weight of the sampled widgets from each box are 2.2, 1.9, and 2.1, respectively. Assume that there are $N = 100$ boxes of widgets in the warehouse, containing a total of $M = 425$ widgets. What are the estimates of τ and μ using the ratio estimators?

The variance of $\hat{\tau}$ is

$$V(\hat{\tau}) \approx \underbrace{N^2 \left(1 - \frac{n}{N}\right) \frac{\sigma_r^2}{n}}_{\text{first stage of sampling}} + \underbrace{\frac{N}{n} \sum_{i=1}^{N} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{\sigma_i^2}{m_i}}_{\text{second stage of sampling}},$$

where

$$\sigma_r^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - M_i \mu)^2.$$

The variance of $\hat{\mu}$ is $V(\hat{\mu}) = V(\hat{\tau})/M^2$. Modified formulas (not shown) can be used to *estimate* these variances.

Comparison of Estimators

We have two estimators of τ, the "unbiased" and "ratio" estimators which are

$$\hat{\tau}_u = \frac{N}{n}\sum_{i\in\mathcal{S}} M_i\bar{y}_i \quad \text{and} \quad \hat{\tau}_r = M\frac{\sum_{i\in\mathcal{S}} M_i\bar{y}_i}{\sum_{i\in\mathcal{S}} M_i},$$

respectively. Why might we choose one over the other?

We have two estimators of μ, the "unbiased" and "ratio" estimators which are

$$\hat{\mu}_u = \frac{N}{Mn}\sum_{i\in\mathcal{S}} M_i\bar{y}_i \quad \text{and} \quad \hat{\mu}_r = \frac{\sum_{i\in\mathcal{S}} M_i\bar{y}_i}{\sum_{i\in\mathcal{S}} M_i},$$

respectively. How do these compare?

Two-Stage Cluster Sampling With PPS Sampling of Clusters

If the clusters are sampled *with replacement* with selection probabilities proportional to cluster size (i.e., PPS sampling with $\delta_i = M_i/M$), then we can use a variant of the Hansen-Hurwitz estimator. Recall that for one-stage cluster sampling this estimator can be written as

$$\hat{\tau} = \frac{M}{n} \sum_{i \in \mathcal{S}} \frac{y_i}{m_i},$$

where y_i is the total of the target variable for the elements in the i-th cluster. To derive the estimator for a two-stage cluster sampling design, replace y_i with the estimator $M_i \bar{y}_i$ and m_i with M_i to get

$$\hat{\tau} = \frac{M}{n} \sum_{i \in \mathcal{S}} \bar{y}_i,$$

where $\bar{y}_i$ is the mean of the target variable for the sampled elements from the i-th cluster. The estimator of μ is then

$$\hat{\mu} = \frac{1}{n} \sum_{i \in \mathcal{S}} \bar{y}_i.$$

Example: A two-stage cluster sampling design selects $n = 3$ boxes using simple random sampling. These boxes each contain 10, 12, and 5 widgets, but we select 2, 3, and 1 widget(s) from these boxes, respectively, using PPS sampling. The mean weight of the sampled widgets from each box are 2.2, 1.9, and 2.1, respectively. Assume that there are $N = 100$ boxes of widgets in the warehouse, containing a total of $M = 425$ widgets. What are the estimates of τ and μ using Hansen-Hurwitz estimator?

The variance of $\hat{\tau}$ can be written as

$$V(\hat{\tau}) = \underbrace{\frac{M}{n} \sum_{i \in \mathcal{S}} M_i (\mu_i - \mu)^2}_{\text{first stage of sampling}} + \underbrace{\frac{M}{n} \sum_{i \in \mathcal{S}} M_i \left(1 - \frac{m_i}{M_i}\right) \frac{\sigma_i^2}{m_i}}_{\text{second stage of sampling}},$$

where σ_i^2 is as defined earlier, and μ_i is the mean of the target variable for all elements in the i-th cluster. The variance of $\hat{\mu}$ is then $V(\hat{\mu}) = V(\hat{\tau})/M^2$. Interestingly the *estimated* variance of $\hat{\tau}$ is very simple. It can be written as

$$\hat{V}(\hat{\tau}) = \frac{M^2}{n(n-1)} \sum_{i \in \mathcal{S}} (\bar{y}_i - \hat{\mu})^2,$$

where $\hat{\mu}$ is the Hansen-Hurwitz estimator of μ. The variance of $\hat{\mu}$ is then

$$\hat{V}(\hat{\mu}) = \frac{1}{n(n-1)} \sum_{i \in \mathcal{S}} (\bar{y}_i - \hat{\mu})^2.$$