Session 07

Multi Stage Designs

In Section 9.1, we introduce multi-stage design and give a few practical examples. We then provide the notations for two stage design. The unbiased estimators for two stage design with simple random sampling at each stage is discussed. Then we discuss the ratio estimator for two stage design with simple random sampling at each stage.

In Section 9.2, we discuss the two stage design with primary units selected with probability proportional to size and secondary units selected with simple random sampling. The Hansen-Hurwitz estimator is computed for this situation. We also show how to compute the estimated variance for the H-H estimator.

Multi-Stage Sampling: Two Stages with S.R.S at Each Stage

We have learned about cluster sampling where one selects the primary units and then all of the cases from the secondary units. With multi-stage sampling we will only select some of the units from the secondary stages.

For example, in two-stage sampling:

- 1st stage samples *n* primary units
- 2nd stage, for the *i*th primary unit, selects m_i (not all) secondary units

Multistage designs are used in many practical cases. These are just a few:

1. **Large surveys involving the sampling of housing units** - The U.S. Census Bureau selects geographical areas within each state and then select housing units within each selected geographical area.
2. **Practical quality control problems** often involve two (or more) stages of sampling. For example, Ford wants to inspect the quality of a supplier of air filters. They first sample some cartons and then inspect some air filters inside these selected cartons.
3. **Gallop poll samples** approximately 300 election districts. At the second stage, they select 5 households per district.

Notation:

- ***N*** : number of primary units in the population
- $\boldsymbol{M_i}$: number of secondary units in the i^{th} primary unit
- **population total :** $\tau = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$

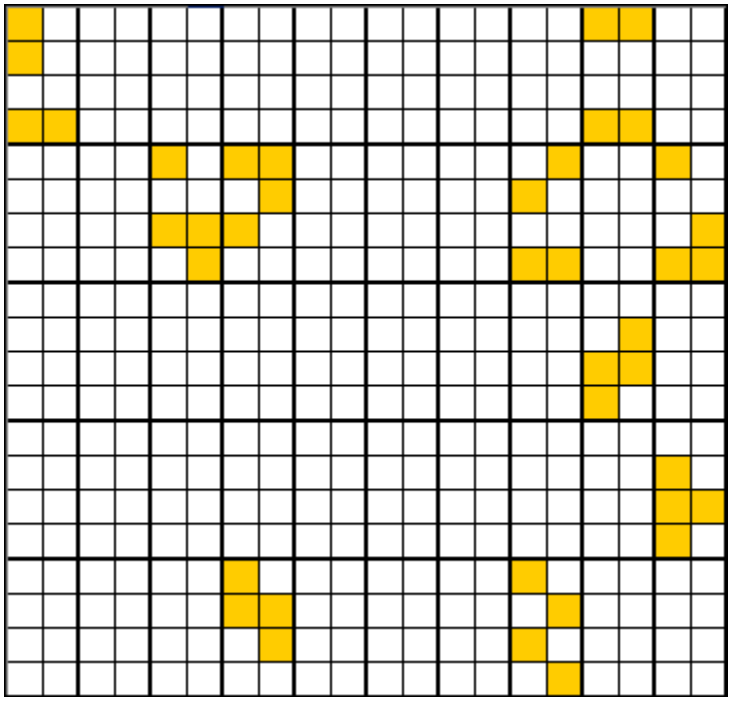
- $\mu = \frac{\tau}{M}$ where $M = \sum_{i=1}^{N} M_i$
- ***n*** : number of primary units selected in the first stage
- $y_i = \sum_{j=1}^{M_i} y_{ij}$
- m_i : number of secondary units selected in the second stage

Multistage Design

This is something that arises in practice quite often. As a result, we need to be able to figure out how this type of sampling design is implemented. Most of the time this deals with two stages of sample with simple random sampling at each stage.

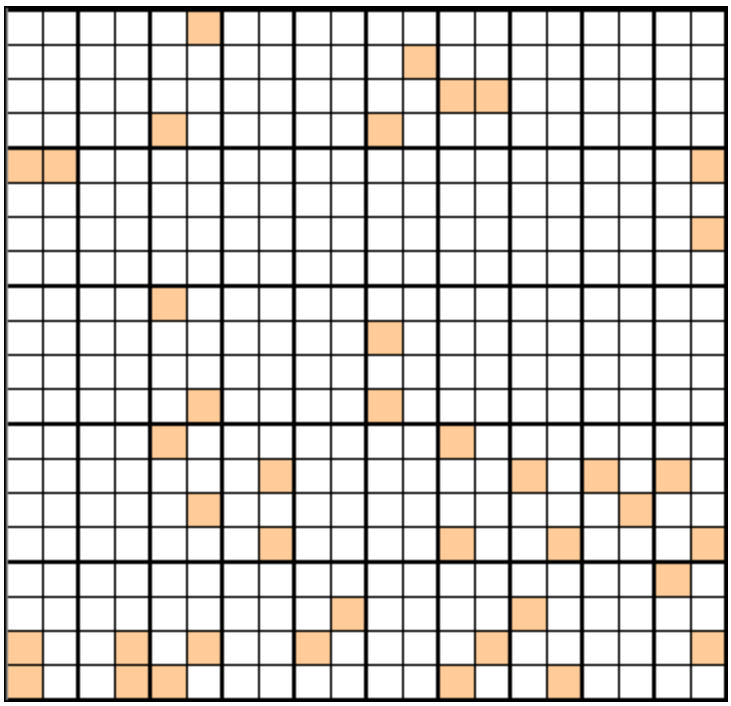
Let's take a look at this graph as a means of understanding how this type of sampling design plays out.

N = 50 for both graphs



(i) . Two-stage sample of 10 primary units
and four secondary units per primary unit.

Here is another graph for another example of two-stage sample



(ii) Two-stage sample of 20 primary units
and two secondary units per primary unit.

Two-stage cluster sampling with simple random sampling at each stage

We will discuss two possible estimators for this sampling design: unbiased estimator and ratio estimator.

A. Unbiased Estimator

Since simple random sampling is used in the second stage, an unbiased estimator of the total *y*-value for the *i*th primary unit is:

$$\hat{y}_i = M_i \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i} = M_i \bar{y}_i \quad \text{where} \quad \bar{y}_i = \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i}$$

The first part of this formula is also known as the expansion estimator.

Also, since simple random sampling is used in the first stage, an unbiased estimator for the population total is:

$$\hat{\tau} = N \cdot \frac{\sum_{i=1}^{n} \hat{y}_i}{n} = N \cdot \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{n}$$

Now we have the expansion estimators from each stage. The next thing we need is the variance.

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

$$\hat{Var}(\hat{\tau}) = N(N-n)\frac{s_u^2}{n} + \frac{N}{n}\sum_{i=1}^{n} M_i(M_i - m_i)\frac{s_i^2}{m_i}$$

s_u^2 is the sample variance among the primary unit totals,
s_i^2 is the sample variance within the *i*th primary unit, here

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{y}_i - \frac{\sum_{i=1}^{n}\hat{y}_i}{n}\right)^2$$, and $$s_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2$$

To estimate the population mean $\mu = \tau / M$, the estimators and the estimated variance are:

$$\hat{\mu} = \frac{N}{M} \cdot \frac{\sum_{i=1}^{n} \hat{y}_i}{n}$$, and $$\hat{Var}(\hat{\mu}) = \frac{1}{M^2}\hat{Var}(\hat{\tau})$$

Let's take a look at an example where we can compute both the estimates and their variances.

Example - Restaurant Employee Satisfaction

A restaurant chain wants to estimate the average employee satisfaction with their job (the scale is from 1 to 7). They have 120 restaurants the total number of employees in the chain is 6860. They use simple random sampling to sample 10 restaurants. They then use simple random sampling to sample and interview about 20% of the employees in those restaurants,. The data are given as follows.

Restaurant	M_i	m_i	**Employee Satisfaction**	$\bar{y}_i$	s_i
1	54	10	5, 7, 6, 5, 4, 7, 6, 6, 4, 5	5.50	1.08

2	48	10	7, 7, 7, 6, 5, 4, 7, 7, 6, 6	6.20	1.03
3	68	14	5, 6, 5, 6, 4, 5, 6, 5, 4, 5, 4, 6, 5, 6	5.14	0.77
4	70	14	6, 5, 7, 6, 7, 6, 5, 7, 5, 7, 6, 5, 7, 6	6.07	0.83
5	52	10	4, 5, 4, 5, 5, 6, 5, 4, 4, 4	4.60	0.70
6	62	12	5, 7, 6, 7, 4, 3, 1, 5, 4, 6, 4, 5	4.75	1.71
7	41	8	7, 6, 7, 7, 6, 6, 5, 7	6.38	0.74
8	53	11	6, 6, 5, 4, 6, 7, 5, 5, 7, 6, 5	5.64	0.92
9	64	12	7, 6, 5, 4, 6, 5, 7, 4, 3, 6, 5, 7	5.42	1.31
10	43	9	7, 6, 6, 5, 7, 3, 5, 4, 5	5.33	1.32

Minitab output:

Mi	mi	yibar	yihead
54	10	5.50	297.00
48	10	6.20	297.60
68	14	5.14	349.52
70	14	6.07	424.90
52	10	4.60	239.20
62	12	4.75	294.50
41	8	6.38	261.58
53	11	5.64	298.92
64	12	5.42	346.88
43	9	5.33	229.19

Descriptive Statistics for yibar and yihead

Variable	N	Mean	StDev
yibar	10	5.503	0.591
yihead	10	303.9	58.1

Here we have output from Minitab that provides the descriptive statistics that you will need to compute the estimators and variance.

The estimated variance of the unbiased estimator is then:

$$\hat{Var}(\hat{\tau}) = N(N-n)\frac{s_u^2}{n} + \frac{N}{n}\sum_{i=1}^{n} M_i(M_i - m_i)\frac{s_i^2}{m_i}$$

s_u^2 is the sample variance of $\hat{y}_1$, $\hat{y}_2$, ... , $\hat{y}_{10}$. From the Minitab output, $s_u^2 = (58.1)^2 = 3375.61$

s_i^2 is the sample variance within the primary unit.

$$s_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2$$

s_i has been computed and given in the table.

Remark: If M is unknown, we cannot use the unbiased estimator $\hat{\mu}$.

If the cluster total is proportional to the cluster size, then the ratio estimate is appropriate. We will discuss the ratio estimator in the following:

B. Ratio estimator

For the population total, the ratio estimator and its estimated variance are:

$$\hat{\tau}_r = \frac{\sum_{i=1}^{n}\hat{y}_i}{\sum_{i=1}^{n} M_i} \cdot M = \hat{r}M$$

$$\hat{Var}(\hat{\tau}_r) = \frac{N(N-n)}{n} \cdot \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - M_i\hat{r})^2 + \frac{N}{n}\sum_{i=1}^{n} M_i(M_i - m_i)\frac{s_i^2}{m_i}$$

A similar question can be asked of the population mean. Therefore, for the population mean, the ratio estimator and its estimated variance are:

$$\hat{\mu}_r = \hat{r}$$

$$\hat{V}ar(\hat{\mu}_r) = \frac{1}{M^2}\hat{V}ar(\hat{\tau}_r)$$

Remark: If *M* is unknown, one can use $\hat{\mu}_r$ and estimate *M* by:

$$\frac{\sum_{i=1}^{n} M_i}{n} \times N$$

Recall: $M = \sum_{i=1}^{n} M_i$

Two Stages with Primary Units Selected by Probability Proportional to Size and Secondary Units Selected with S.R.S.

Multi-stage design with primary units selected with p.p.s. and secondary units selected with simple random sampling.

Using the Hansen-Hurwitz estimator, we get the following:

To estimate the population total:

$$\hat{\tau}_p = \frac{M}{n}\sum_{i=1}^{n}\frac{\hat{y}_i}{M_i} = M\frac{\sum \bar{y}_i}{n}$$, where $\bar{y}_i = \frac{\hat{y}_i}{M_i}$

$$\hat{V}ar(\hat{\tau}_p) = \frac{M^2}{n(n-1)}\sum(\bar{y}_i - \hat{\mu}_p)^2$$

To estimate the population mean:

$$\hat{\mu}_p = \frac{\sum \bar{y}_i}{n}$$

[since it is $\left(\frac{\hat{\tau}_p}{M}\right)$ and thus it becomes $\frac{\sum \bar{y}_i}{n}$]

$$\hat{Var}(\hat{\mu}_p) = \frac{1}{n(n-1)} \sum (\bar{y}_i - \hat{\mu}_p)^2$$

Example

There are 36 departments in a small liberal arts college. One wants to estimate the average amount of money the students spent on textbooks last semester. Since the size of each department varies very much, a two-stage cluster sampling using probability proportional to size for the primary unit is carried out. The results are listed in the table below.

Department	M_i	m_i	Textbook expenses in $ for last semester
1	10	4	326, 400, 423, 443
2	20	8	278, 312, 450, 350, 227, 438, 512, 403
3	30	12	512, 256, 332, 402, 512, 309, 411, 610, 422, 630, 550, 470
4	15	6	426, 312, 512, 440, 342, 533

Minitab output:

Descriptive Statistics: dept1, dept2, dept3, dept4

Variable	Mean	SE Mean	StDev	Variance
dept1	398.0	25.6	51.1	2612.7
dept2	371.3	34.1	96.3	9277.4
dept3	451.3	33.9	117.6	13828.8
dept4	427.5	36.1	88.4	7815.9