Introduction to Sampling Theory

Lecture 36
Two Stage Sampling (Subsampling)

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Two Stage Sampling (Subsampling):

In cluster sampling, all the elements in the selected clusters are surveyed.

Moreover, the efficiency in cluster sampling depends on size of the cluster.

As the size increases, the efficiency decreases.

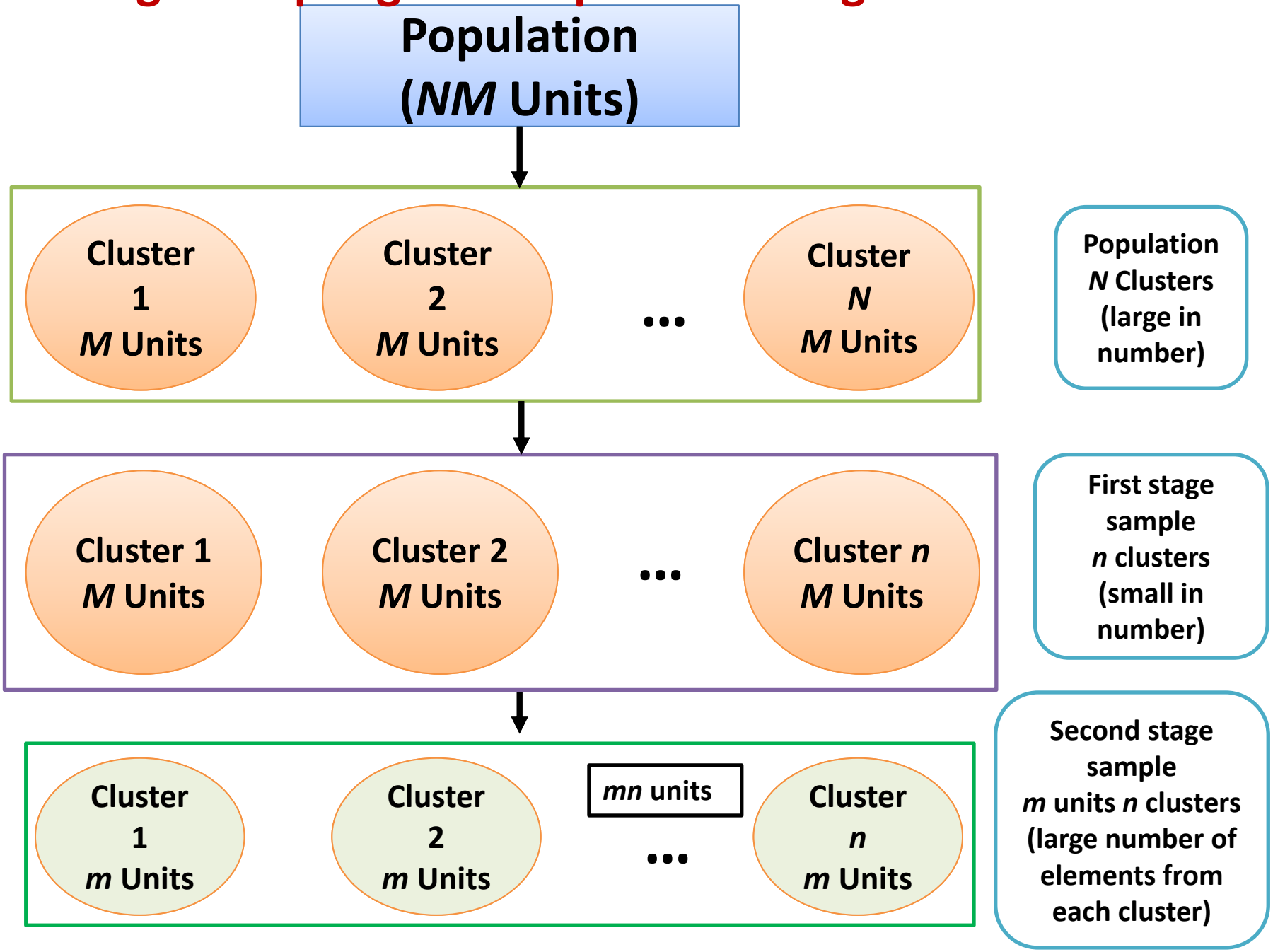
It suggests that higher precision can be attained by distributing a given number of elements over a large number of clusters and then by taking a small number of clusters and enumerating all elements within them. This is achieved in subsampling.

Two Stage Sampling (Subsampling):

In subsampling

- **Divide the population into clusters.**
- **Select a sample of clusters [first stage]**
- **From each of the selected cluster, select a sample of specified number of elements [second stage]**

Two Stage Sampling With Equal First Stage Units:



Two Stage Sampling (Subsampling):

The clusters which form the units of sampling at the first stage are called the first stage units and the units or group of units within clusters which form the unit of clusters are called the second stage units or subunits.

Two Stage Sampling (Subsampling):

The procedure is generalized to three or more stages and is then termed as multistage sampling.

For example, in a crop survey

- **villages are the first stage units,**
- **fields within the villages are the second stage units and**
- **plots within the fields are the third stage units.**

In another example, to obtain a sample of fishes from a commercial fishery,

- **first take a sample of boats and**
- **then take a sample of fishes from each selected boat.**

Two Stage Sampling With Equal First Stage Units:

Assume that the population consists of *NM* elements.

- ***NM* elements are grouped into *N* first stage units of *M* second stage units each, (i.e., *N* clusters, each cluster is of size *M*).**
- **Sample of *n* first stage units is selected (i.e., choose *n* clusters)**
- **Sample of *m* second stage units is selected from each selected first stage unit (i.e., choose *m* units from each cluster).**

Units at each stage are selected with SRSWOR.

Two Stage Sampling With Equal First Stage Units:

Cluster sampling is a special case of two stage sampling in the sense that from a population of N clusters of equal size $m = M$, a sample of n clusters chosen.

If further $M = m = 1$, we get SRSWOR.

If $n = N$, we have the case of stratified sampling.

Two Stage Sampling With Equal First Stage Units:

y_{ij} : **Value of the characteristic under study for the *j*th second stage unit of the *i*th first stage unit;** $i = 1, 2, ..., N;\ j = 1, 2, .., m.$

$\bar{Y}_i = \frac{1}{M}\sum_{j=1}^{m} y_{ij}$: **mean per 2nd stage unit of *i*th 1st stage units in the population.**

$\bar{Y} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_i = \bar{Y}_{MN}$: **mean per second stage unit in the population**

$\bar{y}_i = \frac{1}{n}\sum_{j=1}^{m} y_{ij}$: **mean per second stage unit in the *i*th first stage unit in the sample.**

$\bar{y} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i = \bar{y}_{mn}$: **mean per second stage in the sample.**

Advantages

The principle advantage of two stage sampling is that it is more flexible than the one stage sampling.

It reduces to one stage sampling when $m = M$ but unless this is the best choice of m, we have the opportunity of taking some smaller value that appears more efficient.

As usual, this choice reduces to a balance between statistical precision and cost.

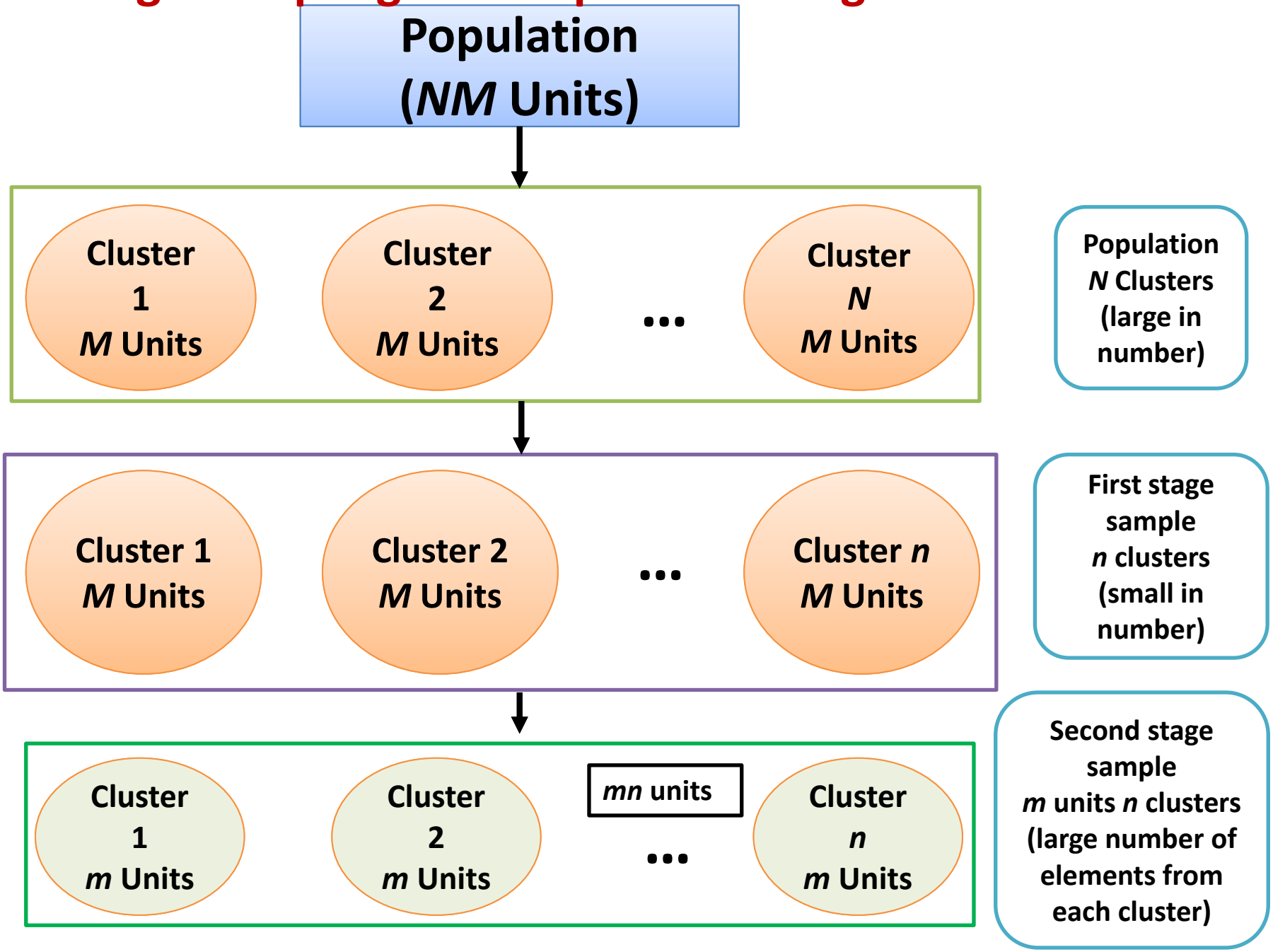
Advantages

When units of the first stage agree very closely, then consideration of precision suggests a small value of *m*.

On the other hand, it is sometimes as cheap to measure the whole of a unit as to a sample.

For example, when the unit is a household and a single respondent can give as accurate data as all the members of the household.

Two Stage Sampling With Equal First Stage Units:



Note

The expectations under two stage sampling scheme depend on the stages. For example, the expectation at second stage unit will be dependent on first stage unit in the sense that second stage unit will be in the sample provided it was selected in the first stage.

To calculate the average:

- First average the estimator over all the second stage selections that can be drawn from a fixed set of units that the plan selects.

- Then average over all the possible selections of units by the plan.

In case of Two Stage Sampling:

$$E(\hat{\theta}) = E_1[E_2(\hat{\theta})]$$

$E(\hat{\theta})$	E_1	$E_2(\hat{\theta})$
average over all samples	average over all 1^{st} stage samples	average over all possible 2^{nd} stage selections from a fixed set of units

In case of three stage sampling,

$$E(\hat{\theta}) = E_1\left[E_2\left\{E_3(\hat{\theta})\right\}\right].$$

To calculate the variance, we proceed as follows:

In case of two stage sampling, $Var(\hat{\theta}) = E(\hat{\theta} - \theta)^2$

$$= E_1E_2(\hat{\theta} - \theta)^2.$$

In case of Two Stage Sampling:

Consider $E_2(\hat{\theta}-\theta)^2 = E_2(\hat{\theta}^2) - 2\theta E_2(\hat{\theta}) + \theta^2$

$$= \left[\left\{E_2(\hat{\theta})\right\}^2 + V_2(\hat{\theta})\right] - 2\theta E_2(\hat{\theta}) + \theta^2.$$

Now average over first stage selection as

$$E_1E_2(\hat{\theta}-\theta)^2 = E_1\left[E_2(\hat{\theta})\right]^2 + E_1\left[V_2(\hat{\theta})\right] - 2\theta E_1E_2(\hat{\theta}) + E_1(\theta^2)$$

$$= E_1\left[E_1\left\{E_2(\hat{\theta})\right\}^2 - \theta^2\right] + E_1\left[V_2(\hat{\theta})\right]$$

$$Var(\hat{\theta}) = V_1\left[E_2(\hat{\theta})\right] + E_1\left[V_2(\hat{\theta})\right].$$

In case of Two Stage Sampling:

In case of three stage sampling,

$$Var(\hat{\theta}) = V_1\left[E_{2}\left\{E_3(\hat{\theta})\right\}\right] + E_1\left[V_2\left\{E_3(\hat{\theta})\right\}\right] + E_1\left[E_2\left\{V_3(\hat{\theta})\right\}\right].$$

Estimation of Population Mean:

Consider $\bar{y} = \bar{y}_{mn}$ as an estimator of the population mean $\bar{Y}$.

Bias

Consider

$$E(\bar{y}) = E_1\left[E_2(\bar{y}_{mn})\right]$$

$$= E_1\left[E_2(\bar{y}_{im} \mid i)\right] \quad \text{(as } 2^{nd} \text{ stage is dependent on } 1^{st} \text{ stage)}$$

$$= E_1\left[E_2(\bar{y}_{im} \mid i)\right] \quad \text{(as } y_i \text{ is unbiased for } \bar{Y}_i \text{ due to SRSWOR)}$$

$$= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i = \bar{Y}.$$

Thus $\bar{y}_{mn}$ is an unbiased estimator of the population mean.

Estimation of Population Mean: Variance

$$Var(\bar{y}) = E_1\left[V_2(\bar{y}\,|\,i)\right] + V_1\left[E_2(\bar{y}\,|\,i)\right]$$

$$= E_1\left[V_2\left\{\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i\,|\,i\right\}\right] + V_1\left[E_2\left\{\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i\,|\,i\right\}\right]$$

$$= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}V(\bar{y}_i\,|\,i)\right] + V_1\left[\frac{1}{n}\sum_{i=1}^{n}E_2(\bar{y}_i\,|\,i)\right]$$

$$= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2\right] + V_1\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)E(S_i^2)\,|\,i + V_1(\bar{y}_c)$$

(**where $\bar{y}_c$ is based on cluster means as in cluster sampling**)

$$= \frac{1}{n^2}n\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \frac{N-n}{Nn}S_b^2, = \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2$$

where $\bar{S}_w^2 = \frac{1}{N}\sum_{i=1}^{N}S_i^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{ij}-\bar{Y}_i\right)^2$, $\bar{S}_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i-\bar{Y})^2$.

Estimate of Variance:

An unbiased estimator of variance of $\bar{y}$ can be obtained by replacing S_b^2 and $\bar{S}_w^2$ by their unbiased estimators in the expression of variance of $\bar{y}$.

Consider an estimator of

$$\bar{S}_w^2 = \frac{1}{N}\sum_{i=1}^{N} S_i^2$$

where $S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(y_{ij} - \bar{Y}_i\right)^2, \bar{s}_w^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2, s_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_{ij} - \bar{y}_i)^2.$

Estimate of Variance:

So $E(\bar{s}_w^2) = E_1 E_2 \left(\bar{s}_w^2 \,|\, i\right)$

$$= E_1 E_2 \left[\frac{1}{n}\sum_{i=1}^{n} s_i^2 \,|\, i\right]$$

$$= E_1 \frac{1}{n}\sum_{i=1}^{n}\left[E_2(s_i^2 \,|\, i)\right]$$

$$= E_1 \frac{1}{n}\sum_{i=1}^{n} S_i^2 \qquad \text{(as SRSWOR is used)}$$

$$= \frac{1}{n}\sum_{i=1}^{n} E_1(S_i^2)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{N}\sum_{i=1}^{N} S_i^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} S_i^2 = \bar{S}_w^2$$

so $\bar{s}_w^2$ **is an unbiased estimator of** $\bar{S}_w^2$ **.**

Estimate of Variance:

Consider $s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{y})^2$

as an estimator of $S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i - \bar{Y})^2.$

So

$$E(s_b^2) = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(\bar{y}_i - \bar{y})^2\right]$$

$$(n-1)E(s_b^2) = E\left[\sum_{i=1}^{n}\bar{y}_i^2 - n\bar{y}^2\right]$$

$$= E\left[\sum_{i=1}^{n}\bar{y}_i^2\right] - nE(\bar{y}^2)$$

$$= E_1\left[E_2\left(\sum_{i=1}^{n}\bar{y}_i^2\right)\right] - n\left[Var(\bar{y}) + \left\{E(\bar{y})\right\}^2\right]$$

Estimate of Variance:

$$= E_1\left[\sum_{i=1}^{n} E_2(\bar{y}_i^2)\,|\,i)\right] - n\left[\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{n}\bar{S}_w^2 + \bar{Y}^2\right]$$

$$= E_1\left[\sum_{i=1}^{n}\left\{Var(\bar{y}_i) + \left(E(\bar{y}_i)\right)^2\right\}\right] - n\left[\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{n}\bar{S}_w^2 + \bar{Y}^2\right]$$

$$= E_1\left[\sum_{i=1}^{n}\left\{\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2 + \bar{Y}_i^2\right\}\right] - n\left[\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{n}\bar{S}_w^2 + \bar{Y}^2\right]$$

$$= nE_1\left[\frac{1}{n}\left\{\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2 + \bar{Y}_i^2\right\}\right] - n\left[\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{n}\bar{S}_w^2 + \bar{Y}^2\right]$$

$$= n\left[\left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{N}\sum_{i=1}^{N} S_i^2 + \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i^2\right] - n\left[\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{n}\bar{S}_w^2 + \bar{Y}^2\right]$$

$$= n\left[\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i^2\right] - n\left[\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{n}\bar{S}_w^2 + \bar{Y}^2\right]$$

Estimate of Variance:

$$= (n-1)\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \frac{n}{N}\sum_{i=1}^{N}\bar{Y}_i^2 - n\bar{Y}^2 - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2$$

$$= (n-1)\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \frac{n}{N}\left[\sum_{i=1}^{N}\bar{Y}_i^2 - N\bar{Y}^2\right] - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2$$

$$= (n-1)\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \frac{n}{N}(N-1)S_b^2 - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2$$

$$= (n-1)\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + (n-1)S_b^2.$$

$$\Rightarrow E(s_b^2) = \left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + S_b^2$$

or $$E\left[s_b^2 - \left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_w^2\right] = S_b^2.$$

Estimate of Variance:

Thus

$$\widehat{Var}(\bar{y}) = \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\hat{\bar{S}}_{\omega}^{2} + \left(\frac{1}{n}-\frac{1}{N}\right)\hat{S}_{b}^{2}$$

$$= \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_{w}^{2} + \left(\frac{1}{n}-\frac{1}{N}\right)\left[s_{b}^{2} - \left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_{w}^{2}\right]$$

$$= \frac{1}{N}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_{w}^{2} + \left(\frac{1}{n}-\frac{1}{N}\right)s_{b}^{2}.$$