Chapter 10

Two Stage Sampling (Subsampling)

In cluster sampling, all the elements in the selected clusters are surveyed. Moreover, the efficiency in cluster sampling depends on the 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.

In subsampling

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

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.**

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.

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 are chosen.

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

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

y_{ij} : Value of the characteristic under study for the jth second stage units of the ith 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 ith 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

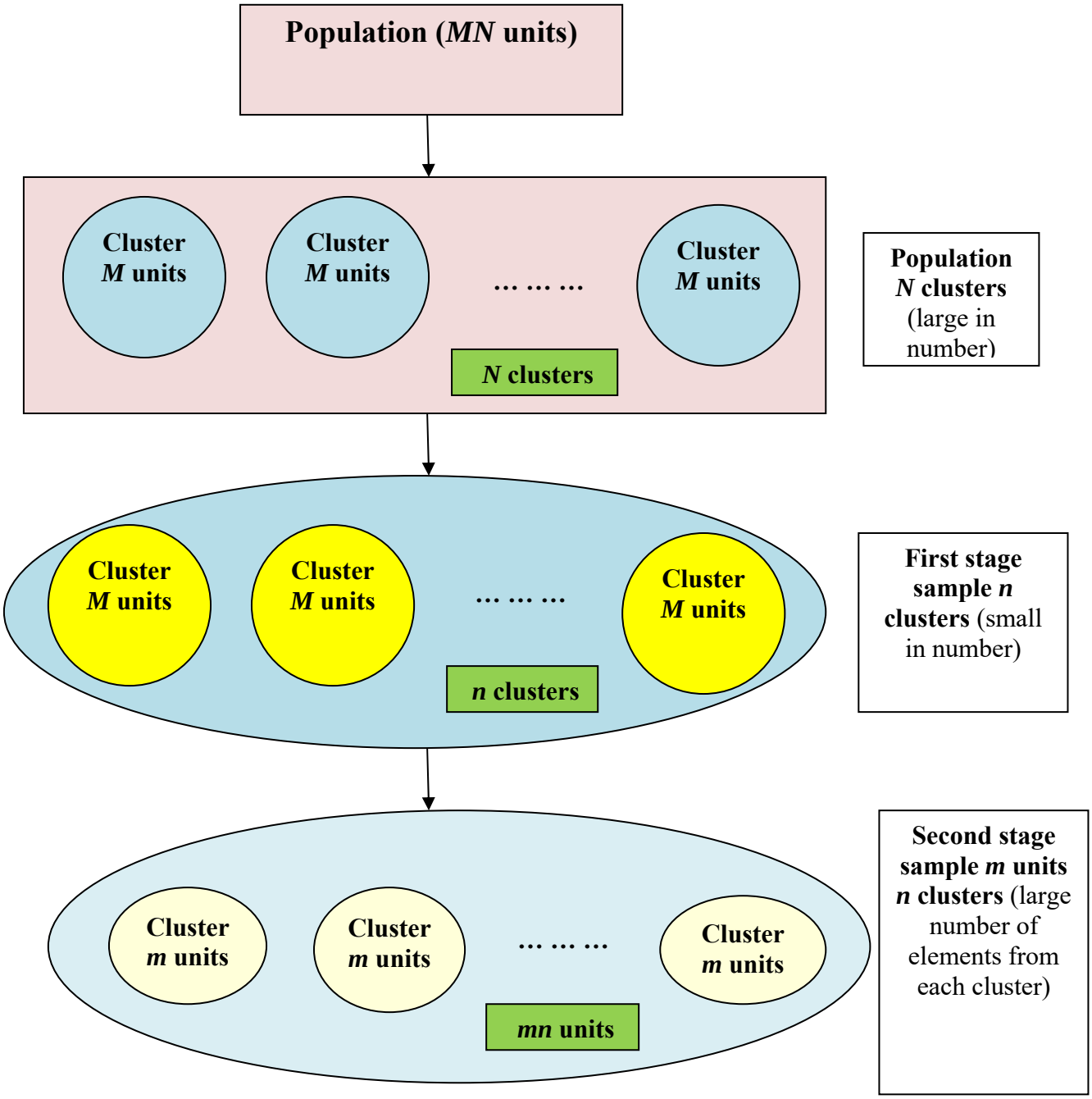
$\bar{y}_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$: mean per second stage unit in the ith first stage unit in the sample.

$\bar{y} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i = \bar{y}_{mn}$: 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. 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.

A pictorial scheme of two stage sampling scheme is as follows:



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

In the case of two stage sampling,

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

$\downarrow$	$\downarrow$	$\searrow$
average over all samples	average over all 1st stage samples	average over all possible 2nd 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,

$$\begin{aligned} Var(\hat{\theta}) &= E(\hat{\theta}-\theta)^2 \\ &= E_1E_2(\hat{\theta}-\theta)^2 \end{aligned}$$

Consider

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

Now average over first stage selection as

$$\begin{aligned} E_1E_2(\hat{\theta}-\theta)^2 &= E_1\left[E_2(\hat{\theta})\right]^2 + E_1\left[V_2(\hat{\theta})\right] - 2\theta E_1E_2(\hat{\theta}) + E_1(\theta^2) \\ &= E_1\left[\left\{E_2(\hat{\theta})\right\}^2 - \theta^2\right] + E_1\left[V_2(\hat{\theta})\right] \\ Var(\hat{\theta}) &= V_1\left[E_2(\hat{\theta})\right] + E_1\left[V_2(\hat{\theta})\right]. \end{aligned}$$

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

$$
\begin{aligned}
E(\bar{y}) &= E_1\left[E_2(\bar{y}_{mn})\right] \\
&= E_1\left[E_2(\bar{y}_{im}|i)\right] \quad \text{(as } 2^{nd} \text{ stage is dependent on } 1^{st} \text{ stage)} \\
&= E_1\left[E_2(\bar{y}_{im}|i)\right] \quad \text{(as } y_i \text{ is unbiased for } \bar{Y}_i \text{ due to SRSWOR)} \\
&= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i \\
&= \bar{Y}.
\end{aligned}
$$

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

Variance

$$
\begin{aligned}
Var(\bar{y}) &= E_1\left[V_2(\bar{y}|i)\right] + V_1\left[E_2(\bar{y}/i)\right] \\
&= E_1\left[V_2\left\{\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i|i\right\}\right] + V_1\left[E_2\left\{\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i/i\right\}\right] \\
&= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}V(\bar{y}_i|i)\right] + V_1\left[\frac{1}{n}\sum_{i=1}^{n}E_2(\bar{y}_i/i)\right] \\
&= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2\right] + V_1\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)E_1(S_i^2) + V_1(\bar{y}_c) \\
&\qquad \text{(where } \bar{y}_c \text{ is based on cluster means as in cluster sampling)} \\
&= \frac{1}{n^2}n\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \frac{N-n}{Nn}S_b^2 \\
&= \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2
\end{aligned}
$$

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 the 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 the variance of $\bar{y}$.

Consider an estimator of

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

where $S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(y_{ij} - \bar{Y}_i\right)^2$

as $\bar{s}_w^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2$

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

So

$$\begin{aligned}
E(\bar{s}_w^2) &= E_1 E_2\left(\bar{s}_w^2 \,|\, i\right) \\
&= E_1 E_2\left[\frac{1}{n}\sum_{i=1}^{n} s_i^2 \,|\, i\right] \\
&= E_1 \frac{1}{n}\sum_{i=1}^{n}\left[E_2(s_i^2 \,|\, i)\right] \\
&= E_1 \frac{1}{n}\sum_{i=1}^{n} S_i^2 \qquad \text{(as SRSWOR is used)} \\
&= \frac{1}{n}\sum_{i=1}^{n} E_1(S_i^2) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{N}\sum_{i=1}^{N} S_i^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} S_i^2 \\
&= \bar{S}_w^2
\end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Thus

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

$$=\frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_w^2+\left(\frac{1}{n}-\frac{1}{N}\right)\left[s_b^2-\left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_w^2\right]$$

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

Allocation of sample to the two stages: Equal first stage units:

The variance of the sample mean in the case of two stage sampling is

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

It depends on $S_b^2, \bar{S}_w^2$, n and m. So the cost of survey of units in the two stage sample depends on n and m.

Case 1. When cost is fixed

We find the values of n and m so that the variance is minimum for given cost.

(I) When cost function is $C = kmn$

Let the cost of the survey be proportional to sample size as

$$C = knm$$

where C is the total cost and k is constant.

When cost is fixed as $C = C_0$. Substituting $m = \frac{C_0}{kn}$ in $Var(\bar{y})$, we get

$$Var(\bar{y}) = \frac{1}{n}\left[S_b^2 - \frac{\bar{S}_w^2}{M}\right] - \frac{S_b^2}{N} + \frac{1}{n}\frac{kn}{C_0}\bar{S}_w^2$$

$$= \frac{1}{n}\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) - \left(\frac{S_b^2}{N} - \frac{k\bar{S}_w^2}{C_0}\right).$$

This variance is a monotonic decreasing function of n if $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) > 0$. The variance is minimum when n assumes maximum value, i.e.,

$$\hat{n} = \frac{C_0}{k} \text{ corresponding to } m = 1.$$

If $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) < 0$ (i.e., intraclass correlation is negative for large N), then the variance is a monotonic increasing function of n, It reaches minimum when n assumes the minimum value, *i.e.*, $\hat{n} = \frac{C_0}{kM}$ (i.e., no subsampling).

(II) **When cost function is** $C = k_1 n + k_2 mn$

Let cost C be fixed as $C_0 = k_1 n + k_2 mn$ where k_1 and k_2 are positive constants. The terms k_1 and k_2 denote the costs of per unit observations in first and second stages respectively. Minimize the variance of sample mean under the two stage with respect to m subject to the restriction $C_0 = k_1 n + k_2 mn$.

We have

$$C_0\left[Var(\bar{y}) + \frac{S_b^2}{N}\right] = k_1\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) + k_2\bar{S}_w^2 + mk_2\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) + \frac{k_1\bar{S}_w^2}{m}.$$

When $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) > 0$, then

$$C_0\left[Var(\bar{y}) + \frac{S_b^2}{N}\right] = \left[\sqrt{k_1\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right)} + \sqrt{k_2\bar{S}_w^2}\right]^2 + \left[\sqrt{mk_2\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right)} - \sqrt{\frac{k_1\bar{S}_w^2}{m}}\right]^2$$

which is minimum when the second term of right-hand side is zero. So we obtain

$$\hat{m} = \sqrt{\frac{k_1}{k_2}\frac{\bar{S}_w^2}{\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right)}}.$$

The optimum n follows from $C_0 = k_1 n + k_2 mn$ as

$$\hat{n} = \frac{C_0}{k_1 + k_2\hat{m}}.$$

When $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) \le 0$ then

$$C_0\left[Var(\bar{y}) + \frac{S_b^2}{N}\right] = k_1\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) + k_2\bar{S}_w^2 + mk_2\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) + \frac{k_1\bar{S}_w^2}{m}$$

is minimum if m is the greatest attainable integer. Hence in this case, when

$$C_0 \ge k_1 + k_2 M;\ \hat{m} = M \text{ and } \hat{n} = \frac{C_0}{k_1 + k_2 M}.$$

If $C_0 \ge k_1 + k_2 M$; then $\hat{m} = \frac{C_0 - k_1}{k_2}$ and $\hat{n} = 1$.

If N is large, then $\bar{S}_w^2 \approx S^2(1-\rho)$

$$\bar{S}_w^2 - \frac{\bar{S}_w^2}{M} \approx \rho S^2$$

$$\hat{m} \approx \sqrt{\frac{k_1}{k_2}\left(\frac{1}{\rho} - 1\right)}.$$

Case 2: When variance is fixed

Now we find the sample sizes when variance is fixed, say as V_0.

$$V_0 = \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \left(\frac{1}{n}-\frac{1}{N}\right)\bar{S}_b^2$$

$$\Rightarrow n = \frac{S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2}{V_0 + \frac{S_b^2}{N}}$$

So

$$C = kmn = km\left(\frac{S_b^2 - \frac{\bar{S}_w^2}{M}}{V_0 + \frac{\bar{S}_b^2}{N}}\right) + \frac{k\bar{S}_w^2}{V_0 + \frac{S_b^2}{N}} .$$

If $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) > 0$, C attains minimum when m assumes the smallest integral value, *i.e.*, 1.

If $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) < 0$, C attains minimum when $\hat{m} = M$.

Comparison of two stage sampling with one stage sampling

One stage sampling procedures are comparable with two stage sampling procedures when either

(i) sampling mn elements in one single stage or

(ii) sampling $\frac{mn}{M}$ first stage units as cluster without sub-sampling.

We consider both the cases.

Case 1: Sampling mn elements in one single stage

The variance of sample mean based on

- mn elements selected by SRSWOR (one stage) is given by

$$V(\bar{y}_{SRS}) = \left(\frac{1}{mn} - \frac{1}{MN}\right)S^2$$

- two stage sampling is given by

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

The intraclass correlation coefficient is

$$\rho = \frac{\left(\frac{N-1}{N}\right)S_b^2 - \frac{\bar{S}_w^2}{M}}{\left(\frac{NM-1}{NM}\right)S^2} = \frac{M(N-1)S_b^2 - N\bar{S}_w^2}{(MN-1)S^2}; -\frac{1}{M-1} \le \rho \le 1 \qquad (1)$$

and using the identity

$$\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{Y})^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{Y}_i)^2 + \sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{Y}_i-\bar{Y})^2$$

$$(NM-1)S^2 = (N-1)MS_b^2 + N(M-1)\bar{S}_w^2 \qquad (2)$$

where $\bar{Y} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}, \bar{Y}_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$.

Now we need to find S_b^2 and $\bar{S}_w^2$ from (1) and (2) in terms of S^2. From (1), we have

$$\bar{S}_w^2 = -\left(\frac{MN-1}{N}\right)MS^2\rho + \left(\frac{N-1}{N}\right)MS_b^2. \qquad (3)$$

Substituting it in (2) gives

$$\begin{aligned}(NM-1)S^2 &= (N-1)MS_b^2 + N(M-1)\left[\left(\frac{N-1}{N}\right)MS_b^2 - \left(\frac{MN-1}{N}\right)MS^2\rho\right] \\ &= (N-1)MS_b^2 + (M-1)(N-1)S_b^2 - \rho M(M-1)(MN-1)S^2 \\ &= (N-1)MS_b^2[1+(M-1)] - \rho M(M-1)(MN-1)S^2 \\ &= (N-1)MS_b^2 - \rho M(M-1)(MN-1)S^2\end{aligned}$$

$$\Rightarrow S_b^2 = \frac{(MN-1)S^2}{M^2(N-1)}\left[1+(M-1)\rho\right]$$

Substituting it in (3) gives

$$\begin{aligned}N(M-1)\bar{S}_w^2 &= (NM-1)S^2 - (N-1)MS_b^2 \\ &= (NM-1)S^2 - (N-1)M\left[\frac{(MN-1)S^2}{M^2(N-1)}\left[1+(M-1)\rho\right]\right] \\ &= (NM-1)S^2\left[\frac{M-1-(M-1)\rho}{M}\right] \\ &= (NM-1)S^2(M-1)(1-\rho)\end{aligned}$$

$$\Rightarrow \quad \bar{S}_w^2 = \left(\frac{MN-1}{MN}\right)S^2(1-\rho).$$

Substituting S_b^2 and $\bar{S}_w^2$ in $Var(\bar{y}_{TS})$

$$V(\bar{y}_{TS}) = \left(\frac{MN-1}{MN}\right)\frac{S^2}{mn}\left[1 - \frac{m(n-1)}{M(N-1)} + \rho\left\{\frac{N-n}{N-1}\frac{m}{M}(M-1) - \frac{M-m}{M}\right\}\right].$$

When subsampling rate $\frac{m}{M}$ is small, $MN-1 \approx MN$ and $M-1 \approx M$, then

$$V(\bar{y}_{SRS}) = \frac{S^2}{mn}$$

$$V(\bar{y}_{TS}) = \frac{S^2}{mn}\left[1+\rho\left(\frac{N-n}{N-1}m-1\right)\right].$$

The relative efficiency of the two stage in relation to one stage sampling of SRSWOR is

$$RE = \frac{Var(\bar{y}_{TS})}{Var(\bar{y}_{SRS})} = 1+\rho\left(\frac{N-n}{N-1}m-1\right).$$

If $N-1 \approx N$ and finite population correction is ignorable, then $\frac{N-n}{N-1} \approx \frac{N-n}{N} \approx 1$, then

$$RE = 1+\rho(m-1).$$

Case 2: Comparison with cluster sampling

Suppose a random sample of $\frac{mn}{M}$ clusters, without further subsampling is selected.

The variance of the sample mean of equivalent mn/M clusters is

$$Var(\bar{y}_{cl}) = \left(\frac{M}{mn}-\frac{1}{N}\right)S_b^2 .$$

The variance of the sample mean under the two stage sampling is

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

So $Var(\bar{y}_{cl})$ exceedes $Var(\bar{y}_{TS})$ by

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

which is approximately

$$\frac{1}{n}\left(\frac{M}{m}-1\right)\rho S^2 \text{ for large } N \text{ and } \left(S_b^2-\frac{\bar{S}_w^2}{M}\right) > 0.$$

where $S_b^2 = \frac{MN-1}{M(N-1)}\frac{S^2}{M}\left[1+\rho(M-1)\right]$

$$\bar{S}_w^2 = \frac{MN-1}{MN}S^2(1-\rho)$$

So smaller the m/M, larger the reduction in the variance of two stage sample over a cluster sample.

When $\left(S_b^2-\frac{\bar{S}_w^2}{M}\right) < 0$ then the subsampling will lead to loss in precision.

Two stage sampling with unequal first stage units:

Consider two stage sampling when the first stage units are of unequal size and SRSWOR is employed at each stage.

Let

y_{ij} : value of jth second stage unit of the ith first stage unit.

M_i : number of second stage units in i^{th} first stage units $(i = 1, 2, ..., N)$.

$M_0 = \sum_{i=1}^{N} M_i$: total number of second stage units in the population.

m_i : number of second stage units to be selected from i^{th} first stage unit, if it is in the sample.

$m_0 = \sum_{i=1}^{n} m_i$: total number of second stage units in the sample.

$$\bar{y}_{i(mi)} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$$

$$\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$$

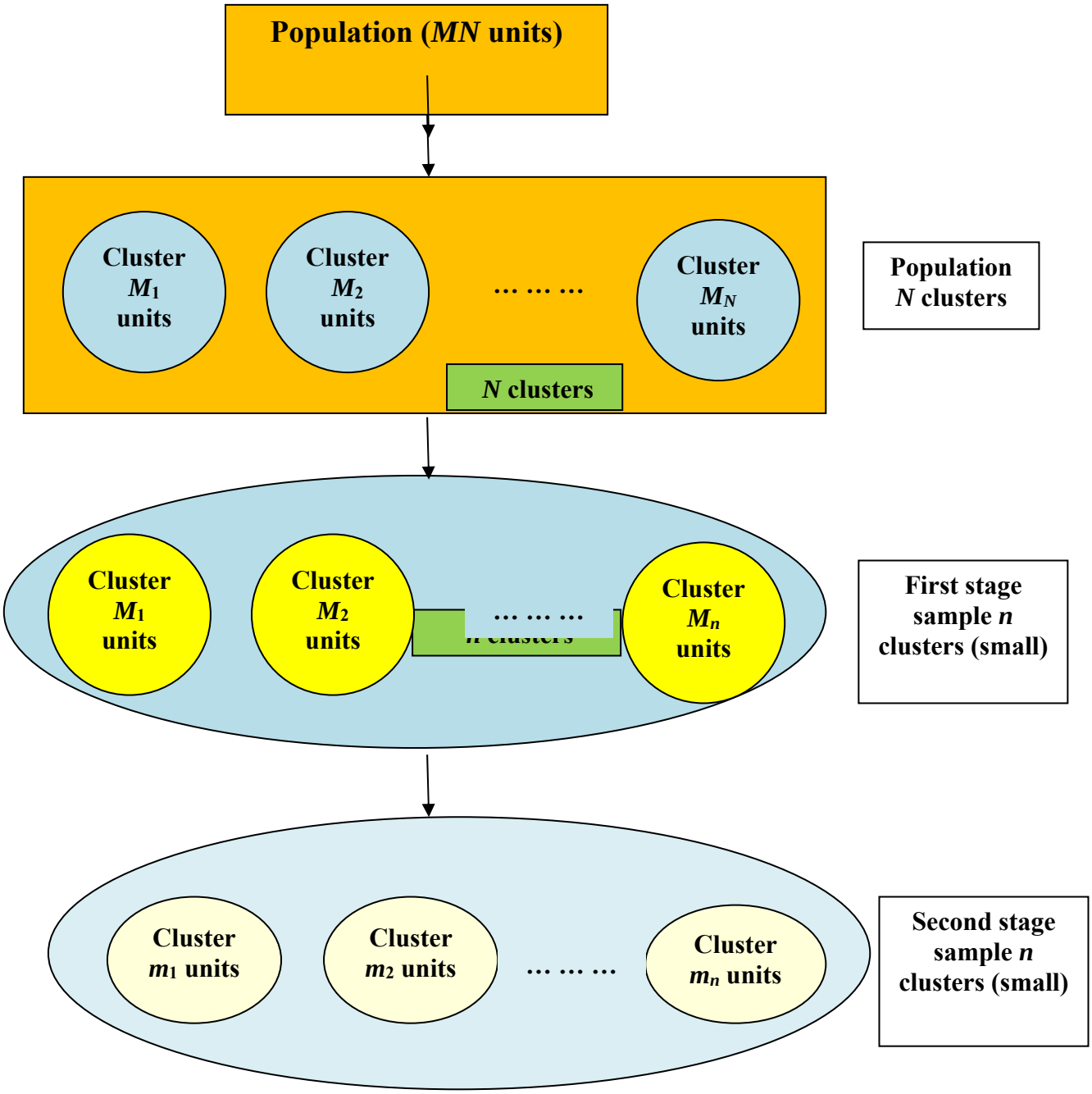
$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} \bar{y}_i = \bar{\bar{Y}}_N$$

$$\bar{Y} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N} M_i \bar{Y}_i}{\bar{M}N} = \frac{1}{N} \sum_{i=1}^{N} u_i \bar{Y}_i$$

$$u_i = \frac{M_i}{\bar{M}}$$

$$\bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i$$

The pictorial scheme of two stage sampling with unequal first stage units case is as follows:



Now we consider different estimators for the estimation of the population mean.

1. Estimator based on the first stage unit means in the sample:

$$\hat{\bar{Y}} = \bar{y}_{S2} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i(mi)}$$

Bias:

$$\begin{aligned}
E(\bar{y}_{S2}) &= E\left[\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i(mi)}\right] \\
&= E_1\left[\frac{1}{n}\sum_{i=1}^{n}E_2(\bar{y}_{i(mi)})\right] \\
&= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i\right] \quad \text{[Since a sample of size } m_i \text{ is selected out of } M_i \text{ units by SRSWOR]} \\
&= \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i \\
&= \bar{\bar{Y}}_N \\
&\neq \bar{Y}.
\end{aligned}$$

So $\bar{y}_{S2}$ is a biased estimator of $\bar{Y}$ and its bias is given by

$$\begin{aligned}
Bias\ (\bar{y}_{S2}) &= E(\bar{y}_{S2}) - \bar{Y} \\
&= \frac{1}{N}\sum_{i=1}^{N}Y_i - \frac{1}{N\bar{M}}\sum_{i=1}^{N}M_i\bar{Y}_i \\
&= -\frac{1}{N\bar{M}}\left[\sum_{i=1}^{N}M_i\bar{Y}_i - \frac{1}{N}\left(\sum_{i=1}^{N}\bar{Y}_i\right)\left(\sum_{i=1}^{N}M_i\right)\right] \\
&= \frac{1}{N\bar{M}}\sum_{i=1}^{N}(M_i - \bar{M})(\bar{Y}_i - \bar{\bar{Y}}_N).
\end{aligned}$$

This bias can be estimated by

$$\widehat{Bias}(\bar{y}_{S2}) = -\frac{N-1}{N\bar{M}(n-1)}\sum_{i=1}^{n}(M_i - \bar{m})(\bar{y}_{i(mi)} - \bar{y}_{S2})$$

which can be seen as follows:

$$\begin{aligned}
E\left[\widehat{Bias}(\bar{y}_{S2})\right] &= -\frac{N-1}{N\bar{M}}E_1\left[\frac{1}{n-1}\sum_{i=1}^{n}E_2\left\{(M_i - \bar{m})(\bar{y}_{i(mi)} - \bar{y}_{S2})/n\right\}\right] \\
&= -\frac{N-1}{N\bar{M}}E\left[\frac{1}{n-1}\sum_{i=1}^{n}(M_i - \bar{m})(\bar{Y}_i - \bar{\bar{y}}_n)\right] \\
&= -\frac{1}{N\bar{M}}\sum_{i=1}^{N}(M_i - \bar{M})(\bar{Y}_i - \bar{\bar{Y}}_N) \\
&= \bar{\bar{Y}}_N - \bar{Y}
\end{aligned}$$

where $\bar{\bar{y}}_n = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i$.

An unbiased estimator of the population mean $\bar{Y}$ is thus obtained as

$$\bar{y}_{S2} + \frac{N-1}{N\bar{M}} \frac{1}{n-1} \sum_{i=1}^{n} (M_i - \bar{m})(\bar{y}_{i(mi)} - \bar{y}_{S2}).$$

Note that the bias arises due to the inequality of sizes of the first stage units and probability of selection of second stage units varies from one first stage to another.

Variance:

$$\begin{aligned}
Var(\bar{y}_{S2}) &= Var\left[E(\bar{y}_{S2}|n)\right] + E\left[Var(\bar{y}_{S2}|n)\right] \\
&= Var\left[\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i\right] + E\left[\frac{1}{n^2}\sum_{i=1}^{n} Var(\bar{y}_{i(mi)}|i)\right] \\
&= \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2\right] \\
&= \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{Nn}\sum_{i=1}^{N}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2
\end{aligned}$$

where $S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i - \bar{\bar{Y}}_N\right)^2$

$$S_i^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}\left(y_{ij} - \bar{Y}_i\right)^2.$$

The *MSE* can be obtained as

$$MSE(\bar{y}_{S2}) = Var(\bar{y}_{S2}) + \left[Bias(\bar{y}_{S2})\right]^2.$$

Estimation of variance:

Consider mean square between cluster means in the sample

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{y}_{i(m_i)} - \bar{y}_{S2}\right)^2.$$

It can be shown that

$$E(s_b^2) = S_b^2 + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2.$$

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

$$E(s_i^2) = S_i^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i)^2$$

So $E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2\right] = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2.$

Thus

$$E(s_b^2) = S_b^2 + E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_i^2\right]$$

and an unbiased estimator of S_b^2 is

$$\hat{S}_b^2 = s_b^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_i^2.$$

So an estimator of the variance can be obtained by replacing S_b^2 and S_i^2 by their unbiased estimators as

$$\widehat{Var}(\bar{y}_{S2}) = \left(\frac{1}{n} - \frac{1}{N}\right)\hat{S}_b^2 + \frac{1}{Nn}\sum_{i=1}^{N}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)\hat{S}_i^2.$$

2. Estimation based on first stage unit totals:

$$\hat{\bar{Y}} = \bar{y}_{S2}^* = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{y}_{i(mi)}}{\bar{M}}$$

$$= \frac{1}{n}\sum_{i=1}^{n}u_i\bar{y}_{i(mi)}$$

where $u_i = \frac{M_i}{\bar{M}}$.

Bias

$$E(y_{S2}^*) = E\left[\frac{1}{n}\sum_{i=1}^{n}u_i\bar{y}_{i(mi)}\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}u_iE_2(\bar{y}_{i(mi)}|i)\right]$$

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

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

$$= \bar{Y}.$$

Thus $\bar{y}_{S2}^*$ is an unbiased estimator of $\bar{Y}$.

Variance:

$$Var(\bar{y}_{S2}^*) = Var\left[E(\bar{y}_{S2}^*|n)\right] + E\left[Var(\bar{y}_{S2}^*|n)\right]$$

$$= Var\left[\frac{1}{n}\sum_{i=1}^{n}u_i\bar{Y}_i\right] + E\left[\frac{1}{n^2}\sum_{i=1}^{n}u_i^2Var(\bar{y}_{i(mi)}|i)\right]$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right)S_b^{*2} + \frac{1}{nN}\sum_{i=1}^{N}u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_i^2$$

where $S_i^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i)^2$

$$S_b^{*2} = \frac{1}{N-1}\sum_{j=1}^{N}(u_i\bar{Y}_i - \bar{Y})^2.$$

3. Estimator based on ratio estimator:

$$\hat{\bar{Y}} = \bar{y}_{S2}^{**} = \frac{\sum_{i=1}^{n} M_i \bar{y}_{i(mi)}}{\sum_{i=1}^{n} M_i}$$

$$= \frac{\sum_{i=1}^{n} u_i \bar{y}_{i(mi)}}{\sum_{i=1}^{n} u_i}$$

$$= \frac{\bar{y}_{S2}^{*}}{\bar{u}_n}$$

where $u_i = \frac{M_i}{\bar{M}}$, $\bar{u}_n = \frac{1}{n}\sum_{i=1}^{n} u_i$.

This estimator can be seen as if arising by the ratio method of estimation as follows:

Let $y_i^* = u_i \bar{y}_{i(mi)}$

$$x_i^* = \frac{M_i}{\bar{M}}, \; i = 1, 2, ..., N$$

be the values of study variable and auxiliary variable in reference to the ratio method of estimation. Then

$$\bar{y}^* = \frac{1}{n}\sum_{i=1}^{n} y_i^* = \bar{y}_{S2}^*$$

$$\bar{x}^* = \frac{1}{n}\sum_{i=1}^{n} x_i^* = \bar{u}_n$$

$$\bar{X}^* = \frac{1}{N}\sum_{i=1}^{N} X_i^* = 1.$$

The corresponding ratio estimator of $\bar{Y}$ is

$$\hat{\bar{Y}}_R = \frac{\bar{y}^*}{\bar{x}^*}\bar{X}^* = \frac{\bar{y}_{S2}^*}{\bar{u}_n} 1 = \bar{y}_{S2}^{**}.$$

So the bias and mean squared error of $\bar{y}_{S2}^{**}$ can be obtained directly from the results of ratio estimator.

Recall that in ratio method of estimation, the bias and MSE of the ratio estimator upto second order of approximation is

$$Bias(\hat{\bar{y}}_R) \approx \frac{N-n}{Nn}\bar{Y}(C_x^2 - 2\rho C_x C_y)$$

$$= \bar{Y}\left[\frac{Var(\bar{x})}{\bar{X}^2} - \frac{Cov(\bar{x}, \bar{y})}{\bar{X}\bar{Y}}\right]$$

$$MSE(\hat{\bar{Y}}_R) \approx \left[Var(\bar{y}) + R^2 Var(\bar{x}) - 2RCov(\bar{x}, \bar{y})\right]$$

where $R = \frac{\bar{Y}}{\bar{X}}$.

Bias:

The bias of $\bar{y}_{S2}^{**}$ up to second order of approximation is

$$Bias(\bar{y}_{S2}^{**}) = \bar{Y}\left[\frac{Var(\bar{x}_{S2}^{*})}{\bar{X}^2} - \frac{Cov(\bar{x}_{S2}^{*}, \bar{y}_{S2}^{*})}{\bar{X}\bar{Y}}\right]$$

where $\bar{x}_{S2}^{*}$ is the mean of auxiliary variable similar to $\bar{y}_{S2}^{*}$ as $\bar{x}_{S2}^{*} = \frac{1}{n}\sum_{i=1}^{n}\bar{x}_{i(mi)}$.

Now we find $Cov(\bar{x}_{S2}^{*}, \bar{y}_{S2}^{*})$.

$$\begin{aligned}
Cov(\bar{x}_{S2}^{*}, \bar{y}_{S2}^{*}) &= Cov\left[E\left(\frac{1}{n}\sum_{i=1}^{n}u_i\bar{x}_{i(mi)}, \frac{1}{n}\sum_{i=1}^{n}u_i\bar{y}_{i(mi)}\right)\right] + E\left[Cov\left(\frac{1}{n}\sum_{i=1}^{n}u_i\bar{x}_{i(mi)}, \frac{1}{n}\sum_{i=1}^{n}u_i\bar{y}_{i(mi)}\right)\right] \\
&= Cov\left[\frac{1}{n}\sum_{i=1}^{n}u_iE(\bar{x}_{i(mi)}), \frac{1}{n}\sum_{i=1}^{n}u_iE(\bar{y}_{i(mi)})\right] + E\left[\frac{1}{n^2}\sum_{i=1}^{n}u_i^2Cov(\bar{x}_{i(mi)}, \bar{y}_{i(mi)})\middle| i\right] \\
&= Cov\left[\frac{1}{n}\sum_{i=1}^{n}u_i\bar{X}_i, \frac{1}{n}\sum_{i=1}^{n}u_i\bar{Y}_i\right] + E\left[\frac{1}{n^2}\sum_{i=1}^{n}u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{ixy}\right] \\
&= \left(\frac{1}{n} - \frac{1}{N}\right)S_{bxy}^{*} + \frac{1}{nN}\sum_{i=1}^{N}u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{ixy}
\end{aligned}$$

where

$$S_{bxy}^{*} = \frac{1}{N-1}\sum_{i=1}^{N}(u_i\bar{X}_i - \bar{X})(u_i\bar{Y}_i - \bar{Y})$$

$$S_{ixy} = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(x_{ij} - \bar{X}_i)(y_{ij} - \bar{Y}_i).$$

Similarly, $Var(\bar{x}_{S2}^{*})$ can be obtained by replacing x in place of y in $Cov(\bar{x}_{S2}^{*}, \bar{y}_{S2}^{*})$ as

$$Var(\bar{x}_{S2}^{*}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_{bx}^{*2} + \frac{1}{nN}\sum_{i=1}^{N}u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{ix}^2$$

where $S_{bx}^{*2} = \frac{1}{N-1}\sum_{i=1}^{N}(u_i\bar{X}_i - \bar{X})^2$

$$S_{ix}^{*2} = \frac{1}{M_i - 1}\sum_{i=1}^{M_i}(x_{ij} - \bar{X}_i)^2.$$

Substituting $Cov(\bar{x}_{S2}^{*}, \bar{y}_{S2}^{*})$ and $Var(\bar{x}_{S2}^{*})$ in $Bias(\bar{y}_{S2}^{**})$, we obtain the approximate bias as

$$Bias(\bar{y}_{S2}^{**}) \approx \bar{Y}\left[\left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{S_{bx}^{*2}}{\bar{X}^2} - \frac{S_{bxy}^{*}}{\bar{X}\bar{Y}}\right) + \frac{1}{nN}\sum_{i=1}^{N}\left\{u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)\left(\frac{S_{ix}^2}{\bar{X}^2} - \frac{S_{ixy}}{\bar{X}\bar{Y}}\right)\right\}\right].$$

Mean squared error

$$MSE(\bar{y}_{S2}^{**}) \approx Var(\bar{y}_{S2}^{*}) - 2R^{*}Cov(\bar{x}_{S2}^{*}, \bar{y}_{S2}^{*}) + R^{*2}Var(\bar{x}_{S2}^{*})$$

$$Var(\bar{y}_{S2}^{**}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_{by}^{*2} + \frac{1}{nN}\sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{iy}^2$$

$$Var(\bar{x}_{S2}^{**}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_{bx}^{*2} + \frac{1}{nN}\sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{ix}^2$$

$$Cov(\bar{x}_{S2}^{*}, \bar{y}_{S2}^{**}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_{bxy}^{*} + \frac{1}{nN}\sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{ixy}$$

where

$$S_{by}^{*2} = \frac{1}{N-1}\sum_{i=1}^{N}(u_i\bar{Y}_i - \bar{Y})^2$$

$$S_{iy}^{*2} = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i)^2$$

$$R^{*} = \frac{\bar{Y}}{\bar{X}} = \bar{Y}.$$

Thus

$$MSE(\bar{y}_{S2}^{**}) \approx \left(\frac{1}{n} - \frac{1}{N}\right)\left(S_{by}^{*2} - 2R^{*}S_{bxy}^{*} + R^{*2}S_{bx}^{*2}\right) + \frac{1}{nN}\sum_{i=1}^{N}\left[u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)\left(S_{iy}^2 - 2R^{*}S_{ixy} + R^{*2}S_{ix}^2\right)\right].$$

Also

$$MSE(\bar{y}_{S2}^{**}) \approx \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}u_i^2\left(\bar{Y}_i - R^{*}\bar{X}_i\right)^2 + \frac{1}{nN}\sum_{i=1}^{N}\left[u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)\left(S_{iy}^2 - 2R^{*}S_{ixy} + R^{*2}S_{ix}^2\right)\right].$$

Estimate of variance

Consider

$$s_{bxy}^{*} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(u_i\bar{y}_{i(mi)} - \bar{y}_{S2}^{*}\right)\left(u_i\bar{x}_{i(mi)} - \bar{x}_{S2}^{*}\right)\right]$$

$$s_{ixy} = \frac{1}{m_i - 1}\sum_{j=1}^{n}\left[\left(x_{ij} - \bar{x}_{i(mi)}\right)\left(y_{ij} - \bar{y}_{i(mi)}\right)\right].$$

It can be shown that

$$E\left(s_{bxy}^{*}\right) = S_{bxy}^{*} + \frac{1}{N}\sum_{i=1}^{N}u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{ixy}$$

$$E(s_{ixy}) = S_{ixy}.$$

So

$$E\left[\frac{1}{n}\sum_{i=1}^{n}u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_{ixy}\right] = \frac{1}{N}\sum_{i=1}^{N}\left[u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{ixy}\right].$$

Thus

$$\hat{S}_{bxy}^{*} = s_{bxy}^{*} - \frac{1}{n}\sum_{i=1}^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{ixy}$$

$$\hat{S}_{bx}^{*2} = s_{bx}^{*2} - \frac{1}{n}\sum_{i=1}^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{ix}^2$$

$$\hat{S}_{by}^{*2} = s_{by}^{*2} - \frac{1}{n}\sum_{i=1}^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{iy}^2.$$

Also

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left\{u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_{ix}^2\right\}\right] = \frac{1}{N}\sum_{i=1}^{N}\left[u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{ix}^2\right]$$

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left\{u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_{iy}^2\right\}\right] = \frac{1}{N}\sum_{i=1}^{N}\left[u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{iy}^2\right].$$

A consistent estimator of MSE of $\bar{y}_{S2}^{**}$ can be obtained by substituting the unbiased estimators of respective statistics in $MSE(\bar{y}_{S2}^{**})$ as

$$\begin{aligned}
\widehat{MSE}(\bar{y}_{S2}^{**}) &\approx \left(\frac{1}{n} - \frac{1}{N}\right)\left(s_{by}^{*2} - 2r^{*}s_{bxy}^{*} + r^{*2}s_{bx}^{*2}\right) \\
&\quad + \frac{1}{nN}\sum_{i=1}^{n} u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)\left(s_{iy}^2 - 2r^{*}s_{ixy} + r^{*2}s_{ix}^2\right) \\
&\approx \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{y}_{i(mi)} - r^{*}\bar{x}_{i(mi)}\right)^2 \\
&\quad + \frac{1}{nN}\sum_{i=1}^{n}\left[u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)\left(s_{iy}^2 - 2r^{*}s_{ixy} + r^{*2}s_{ix}^2\right)\right] \qquad \text{where } r^{*} = \frac{\bar{y}_{S2}^{*}}{\bar{x}_{S2}^{*}}.
\end{aligned}$$