Chapter 9
Cluster Sampling

It is one of the basic assumptions in any sampling procedure that the population can be divided into a finite number of distinct and identifiable units, called **sampling units.** The smallest units into which the population can be divided are called **elements** of the population. The groups of such elements are called **clusters**.

In many practical situations and many types of populations, a list of elements is not available and so the use of an element as a sampling unit is not feasible. The method of cluster sampling or area sampling can be used in such situations.

In cluster sampling

- divide the whole population into clusters according to some well-defined rule.
- Treat the clusters as sampling units.
- Choose a sample of clusters according to some procedure.
- Carry out a complete enumeration of the selected clusters, i.e., collect information on all the sampling units available in selected clusters.

Area sampling

In case, the entire area containing the populations is subdivided into smaller area segments and each element in the population is associated with one and only one such area segment, the procedure is called as area sampling.

Examples:

- In a city, the list of all the individual persons staying in the houses may be difficult to obtain or even maybe not available but a list of all the houses in the city may be available. So every individual person will be treated as sampling unit and every house will be a cluster.
- The list of all the agricultural farms in a village or a district may not be easily available but the list of village or districts are generally available. In this case, every farm in sampling unit and every village or district is the cluster.

Moreover, it is easier, faster, cheaper and convenient to collect information on clusters rather than on sampling units.

In both the examples, draw a sample of clusters from houses/villages and then collect the observations on all the sampling units available in the selected clusters.

Conditions under which the cluster sampling is used:

Cluster sampling is preferred when

(i) No reliable listing of elements is available, and it is expensive to prepare it.

(ii) Even if the list of elements is available, the location or identification of the units may be difficult.

(iii) A necessary condition for the validity of this procedure is that every unit of the population under study must correspond to one and only one unit of the cluster so that the total number of sampling units in the frame may cover all the units of the population under study without any omission or duplication. When this condition is not satisfied, bias is introduced.

Open segment and closed segment:

It is not necessary that all the elements associated with an area segment need be located physically within its boundaries. For example, in the study of farms, the different fields of the same farm need not lie within the same area segment. Such a segment is called an open segment.

In a closed segment, the sum of the characteristic under study, i.e., area, livestock etc. for all the elements associated with the segment will account for all the area, livestock etc. within the segment.

Construction of clusters:

The clusters are constructed such that the sampling units are heterogeneous within the clusters and homogeneous among the clusters. The reason for this will become clear later. This is opposite to the construction of the strata in the stratified sampling.

There are two options to construct the clusters – equal size and unequal size. We discuss the estimation of population means and its variance in both the cases.

Case of equal clusters

- Suppose the population is divided into N clusters and each cluster is of size M.
- Select a sample of n clusters from N clusters by the method of SRS, generally WOR.

So

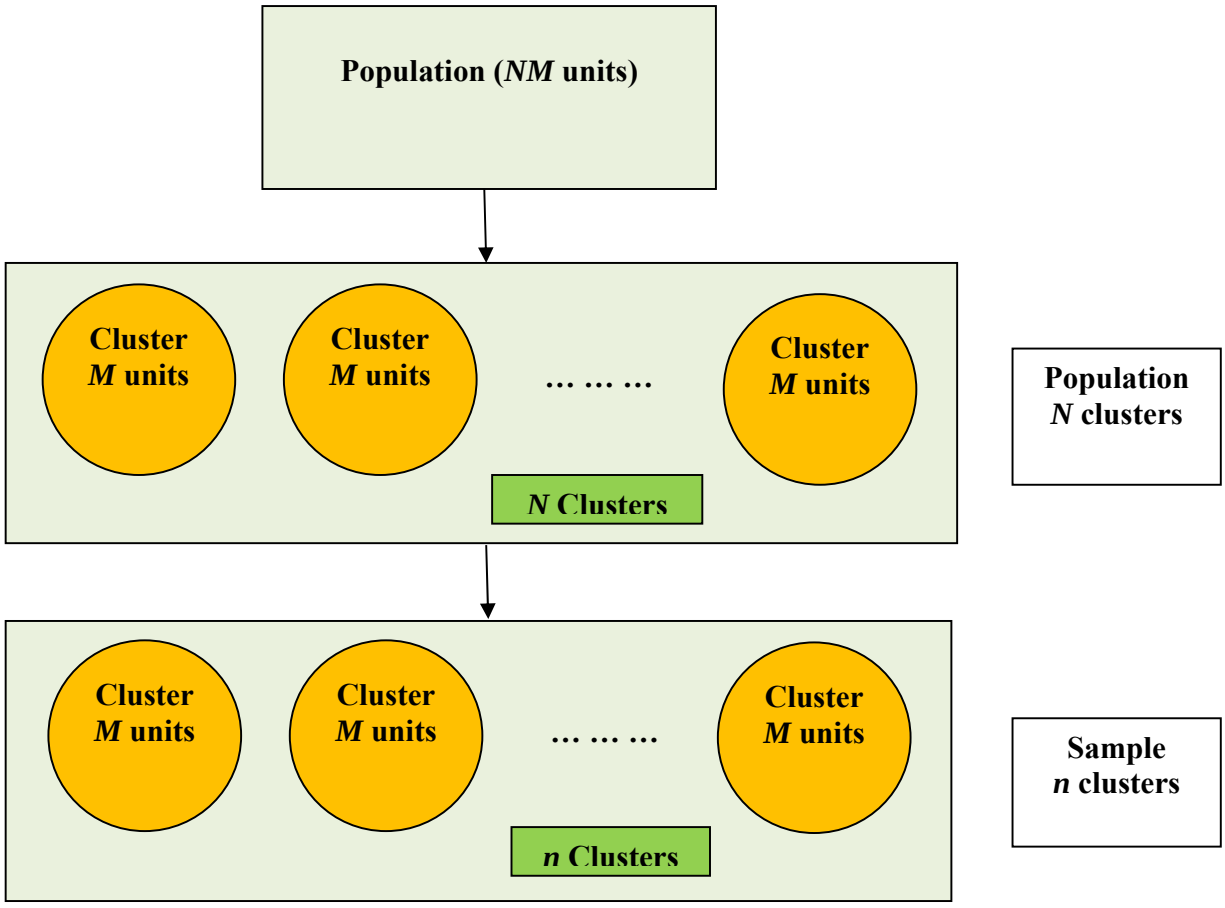
total population size = NM

total sample size = nM.

Let

y_{ij} : Value of the characteristic under study for the value of j^{th} element $(j = 1,2,...,M)$ in the i^{th} cluster $(i = 1,2,...,N)$.

$\bar{y}_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$ mean per element of i^{th} cluster .



Estimation of population mean:

First select n clusters from N clusters by SRSWOR.

Based on n clusters, find the mean of each cluster separately based on all the units in every cluster. So we have the cluster means as $\bar{y}_1, \bar{y}_2, ..., \bar{y}_n$. Consider the mean of all such cluster means as an estimator of population mean as

$$\bar{y}_{cl} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i \ .$$

Bias:

$$\begin{aligned} E(\bar{y}_{cl}) &= \frac{1}{n}\sum_{i=1}^{n} E(\bar{y}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \bar{Y} \qquad \text{(since SRS is used)} \\ &= \bar{Y}. \end{aligned}$$

Thus $\bar{y}_{cl}$ is an unbiased estimator of $\bar{Y}$.

Variance:

The variance of $\bar{y}_{cl}$ can be derived on the same lines as deriving the variance of sample mean in SRSWOR. The only difference is that in SRSWOR, the sampling units are $y_1, y_2, ..., y_n$ whereas in case of $\bar{y}_{cl}$, the sampling units are $\bar{y}_1, \bar{y}_2, ..., \bar{y}_n$.

$\left[\text{Note that is case of SRSWOR, } Var(\bar{y}) = \frac{N-n}{Nn} S^2 \text{ and } \widehat{Var}(\bar{y}) = \frac{N-n}{Nn} s^2\right]$,

$$\begin{aligned} Var(\bar{y}_{cl}) &= E(\bar{y}_{cl} - \bar{Y})^2 \\ &= \frac{N-n}{Nn} S_b^2 \end{aligned}$$

where $S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2$ which is the mean sum of square between the cluster means in the population.

Estimate of variance:

Using the philosophy of estimate of variance in case of SRSWOR again, we can find

$$\widehat{Var}(\bar{y}_{cl}) = \frac{N-n}{Nn} s_b^2$$

where $s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{cl})^2$ is the mean sum of squares between cluster means in the sample .

Comparison with SRS :

If an equivalent sample of nM units were to be selected from the population of NM units by SRSWOR, the variance of the mean per element would be

$$Var(\bar{y}_{nM}) = \frac{NM - nM}{NM} . \frac{S^2}{nM}$$
$$= \frac{f}{n} . \frac{S^2}{M}$$

where $f = \frac{N - n}{N}$ and $S^2 = \frac{1}{NM - 1} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2$.

Also $$Var(\bar{y}_{cl}) = \frac{N - n}{Nn} S_b^2$$
$$= \frac{f}{n} S_b^2 .$$

Consider

$$(NM - 1)S^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2$$
$$= \sum_{i=1}^{N} \sum_{j=1}^{M} \left[(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{Y}) \right]^2$$
$$= \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{N} \sum_{j=1}^{M} (\bar{y}_i - \bar{Y})^2$$
$$= N(M - 1)\bar{S}_w^2 + M(N - 1)S_b^2$$

where

$\bar{S}_w^2 = \frac{1}{N} \sum_{i=1}^{N} S_i^2$ is the mean sum of squares within clusters in the population

$S_i^2 = \frac{1}{M - 1} \sum_{j=1}^{M} (y_{ij} - \bar{y}_i)^2$ is the mean sum of squares for the i^{th} cluster.

The efficiency of cluster sampling over SRSWOR is

$$E = \frac{Var(\bar{y}_{nM})}{Var(\bar{y}_{cl})}$$
$$= \frac{S^2}{MS_b^2}$$
$$= \frac{1}{(NM - 1)} \left[\frac{N(M - 1)}{M} \frac{\bar{S}_w^2}{S_b^2} + (N - 1) \right].$$

Thus the relative efficiency increases when $\bar{S}_w^2$ is large and S_b^2 is small. So cluster sampling will be efficient if clusters are so formed that the variation the between cluster means is as small as possible while variation within the clusters is as large as possible.

Efficiency in terms of intra class correlation

The intra class correlation between the elements within a cluster is given by

$$\rho = \frac{E(y_{ij} - Y)(y_{ik} - \bar{Y})}{E(y_{ij} - \bar{Y})^2}; -\frac{1}{M-1} \le \rho \le 1$$

$$= \frac{\frac{1}{MN(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k(\neq j)=1}^{M} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{\frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2}$$

$$= \frac{\frac{1}{MN(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k(\neq j)=1}^{M} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{\left(\frac{MN-1}{MN}\right) S^2}$$

$$= \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k(\neq j)=1}^{M} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{(MN-1)(M-1)S^2}.$$

Consider

$$\sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2 = \sum_{i=1}^{N} \left[\frac{1}{M} \sum_{j=1}^{M} (y_{ij} - \bar{Y}) \right]^2$$

$$= \sum_{i=1}^{N} \left[\frac{1}{M^2} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2 + \frac{1}{M^2} \sum_{j=1}^{M} \sum_{k(\neq j)=1}^{M} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y}) \right]$$

$$\Rightarrow \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k(\neq j)=1}^{M} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y}) = M^2 \sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2 - \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2$$

or

$$\rho(MN-1)(M-1)S^2 = M^2(N-1)S_b^2 - (NM-1)S^2$$

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

The variance of $\bar{y}_{cl}$ now becomes

$$Var(\bar{y}_{cl}) = \frac{N-n}{Nn} S_b^2$$

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

For large $N, \frac{MN-1}{MN} \approx 1, N-1 \approx N, \frac{N-n}{N} \approx 1$ and so

$$Var(\bar{y}_{cl}) \approx \frac{1}{n} \frac{S^2}{M} \left[1 + (M-1)\rho\right].$$

The variance of the sample mean under SRSWOR for large N is

$$Var(\bar{y}_{nM}) \approx \frac{S^2}{nM}.$$

The relative efficiency for large N is now given by

$$E = \frac{Var(\bar{y}_{nM})}{Var(\bar{y}_{cl})}$$

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

$$= \frac{1}{1+(M-1)\rho}; \quad -\frac{1}{M-1} \le \rho \le 1.$$

- If $M = 1$ then $E = 1,$ i.e., SRS and cluster sampling are equally efficient. Each cluster will consist of one unit, i.e., SRS.
- If $M > 1,$ then cluster sampling is more efficient when

  $$E > 1$$

  or $(M-1)\rho < 0$

  or $\rho < 0.$
- If $\rho = 0,$ then $E = 1,$ i.e., there is no error which means that the units in each cluster are arranged randomly. So sample is heterogeneous.
- In practice, ρ is usually positive and ρ decreases as M increases but the rate of decrease in ρ is much lower in comparison to the rate of increase in $M.$ The situation that $\rho > 0$ is possible when the nearby units are grouped together to form cluster and which are completely enumerated.
- There are situations when $\rho < 0.$

Estimation of relative efficiency:

The relative efficiency of cluster sampling relative to an equivalent SRSWOR is obtained as

$$E = \frac{S^2}{MS_b^2}.$$

An estimator of E can be obtained by substituting the estimates of S^2 and $S_b^2.$

Since $\bar{y}_{cl} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$ is the mean of n means $\bar{y}_i$ from a population of N means $\bar{y}_i,\ i = 1, 2, ..., N$ which are drawn by SRSWOR, so from the theory of SRSWOR,

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

Thus s_b^2 is an unbiased estimator of S_b^2.

Since $s_w^2 = \frac{1}{n}\sum_{i=1}^{n} S_i^2$ is the mean of n mean sum of squares S_i^2 drawn from the population of N mean sums of squares S_i^2, $i = 1,2,...,N$, so it follows from the theory of SRSWOR that

$$E(s_w^2) = E\left[\frac{1}{n}\sum_{i=1}^{n} S_i^2\right] = \frac{1}{n}\sum_{i=1}^{n} E(S_i^2) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{N}\sum_{i=1}^{N} S_i^2\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N} S_i^2$$
$$= \bar{S}_w^2 .$$

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

Consider

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

$$\text{or } (MN-1)S^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}\left[(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{Y})\right]^2$$
$$= \sum_{i=1}^{N}\sum_{j=1}^{M}\left[(y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{Y})^2\right]$$
$$= \sum_{i=1}^{N}(M-1)S_i^2 + M(N-1)S_b^2$$
$$= N(M-1)\bar{S}_w^2 + M(N-1)S_b^2 .$$

An unbiased estimator of S^2 can be obtained as

$$\hat{S}^2 = \frac{1}{MN-1}\left[N(M-1)\bar{s}_w^2 + M(N-1)s_b^2\right].$$

So

$$\widehat{Var}(\bar{y}_{cl}) = \frac{N-n}{Nn}s_b^2$$
$$\widehat{Var}(\bar{y}_{nM}) = \frac{N-n}{Nn}\frac{\hat{S}^2}{M}$$
$$\text{where} \quad s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{y}_{cl})^2 .$$

An estimate of efficiency $E = \frac{S^2}{MS_b^2}$ is

$$\hat{E} = \frac{N(M-1)\bar{s}_w^2 + M(N-1)s_b^2}{M(NM-1)s_b^2}.$$

If N is large so that $M(N-1) \simeq MN$ and $MN - 1 \simeq MN$, then

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

and its estimate is

$$\hat{E} = \frac{1}{M} + \left(\frac{M-1}{M}\right)\frac{\bar{s}_w^2}{Ms_b^2}.$$

Estimation of a proportion in case of equal cluster

Now, we consider the problem of estimation of the proportion of units in the population having a specified attribute on the basis of a sample of clusters. Let this proportion be P.

Suppose that a sample of n clusters is drawn from N clusters by SRSWOR. Defining $y_{ij} = 1$ if the j^{th} unit in the i^{th} cluster belongs to the specified category (i.e. possessing the given attribute) and $y_{ij} = 0$ otherwise, we find that

$$\bar{y}_i = P_i,$$
$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} P_i = P,$$
$$S_i^2 = \frac{MP_iQ_i}{(M-1)},$$
$$S_w^2 = \frac{M\sum_{i=1}^{N} P_iQ_i}{N(M-1)},$$
$$S^2 = \frac{NMPQ}{NM-1)},$$

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(P_i - P)^2,$$
$$= \frac{1}{N-1}\left[\sum_{i=1}^{N} P_i^2 - NP^2\right]$$
$$= \frac{1}{(N-1)}\left[-\sum_{i=1}^{N} P_i(1-P_i) + \sum_{i=1}^{N} P_i - NP^2\right]$$
$$= \frac{1}{(N-1)}\left[NPQ - \sum_{i=1}^{N} P_i Q_i\right],$$

where P_i is the proportion of elements in the i^{th} cluster, belonging to the specified category and $Q_i = 1 - P_i,\ i = 1, 2, ..., N$ and $Q = 1 - P.$ Then, using the result that $\bar{y}_{cl}$ is an unbiased estimator of $\bar{Y}$, we find that

$$\hat{P}_{cl} = \frac{1}{n}\sum_{i=1}^{n} P_i$$

is an unbiased estimator of P and

$$Var(\hat{P}_{cl}) = \frac{(N-n)}{Nn}\frac{\left[NPQ - \sum_{i=1}^{N} P_i Q_i\right]}{(N-1)}.$$

This variance of $\hat{P}_{cl}$ can be expressed as

$$Var(\hat{P}_{cl}) = \frac{N-n}{N-1}\frac{PQ}{nM}[1 + (M-1)\rho],$$

where the value of ρ can be obtained from

$$\rho = \frac{M(N-1)S_b^2 - N\bar{S}_w^2}{(M-1)(MN-1)S^2} \text{ and } (MN-1)S^2 = N(M-1)\bar{S}_w^2 + M(N-1)S_b^2$$

by substituting $S_b^2, \bar{S}_w^2$ and S^2 in ρ, we obtain

$$\rho = 1 - \frac{M}{(M-1)}\frac{1}{N}\frac{\sum_{i=1}^{N} P_i Q_i}{PQ}\ .$$

The variance of $\hat{P}_{cl}$ can be estimated unbiasedly by

$$\widehat{Var}(\hat{P}_{cl}) = \frac{N-n}{nN}s_b^2$$
$$= \frac{N-n}{nN}\frac{1}{(n-1)}\sum_{i=1}^{n}(P_i - \hat{P}_{cl})^2$$
$$= \frac{N-n}{Nn(n-1)}\left[n\hat{P}_{cl}\hat{Q}_{cl} - \sum_{i=1}^{n} P_i Q_i\right]$$

where $\hat{Q}_{cl} = I - \hat{P}_{cl}$. The efficiency of cluster sampling relative to SRSWOR is given by

$$E = \frac{M(N-1)}{(MN-1)} \frac{1}{[1+(M-1)\rho]}$$
$$= \frac{(N-1)}{NM-1} \frac{NPQ}{\left(NPQ - \sum_{i=1}^{N} P_i Q_i \right)}.$$

If N is large, then $E \cong \frac{1}{M}$.

An estimator of the total number of elements belonging to a specified category is obtained by multiplying $\hat{P}_{cl}$ by NM, *i.e.* by $NM\hat{P}_{cl}$. The expressions of variance and its estimator are obtained by multiplying the corresponding expressions for $\hat{P}_{cl}$ by N^2M^2.

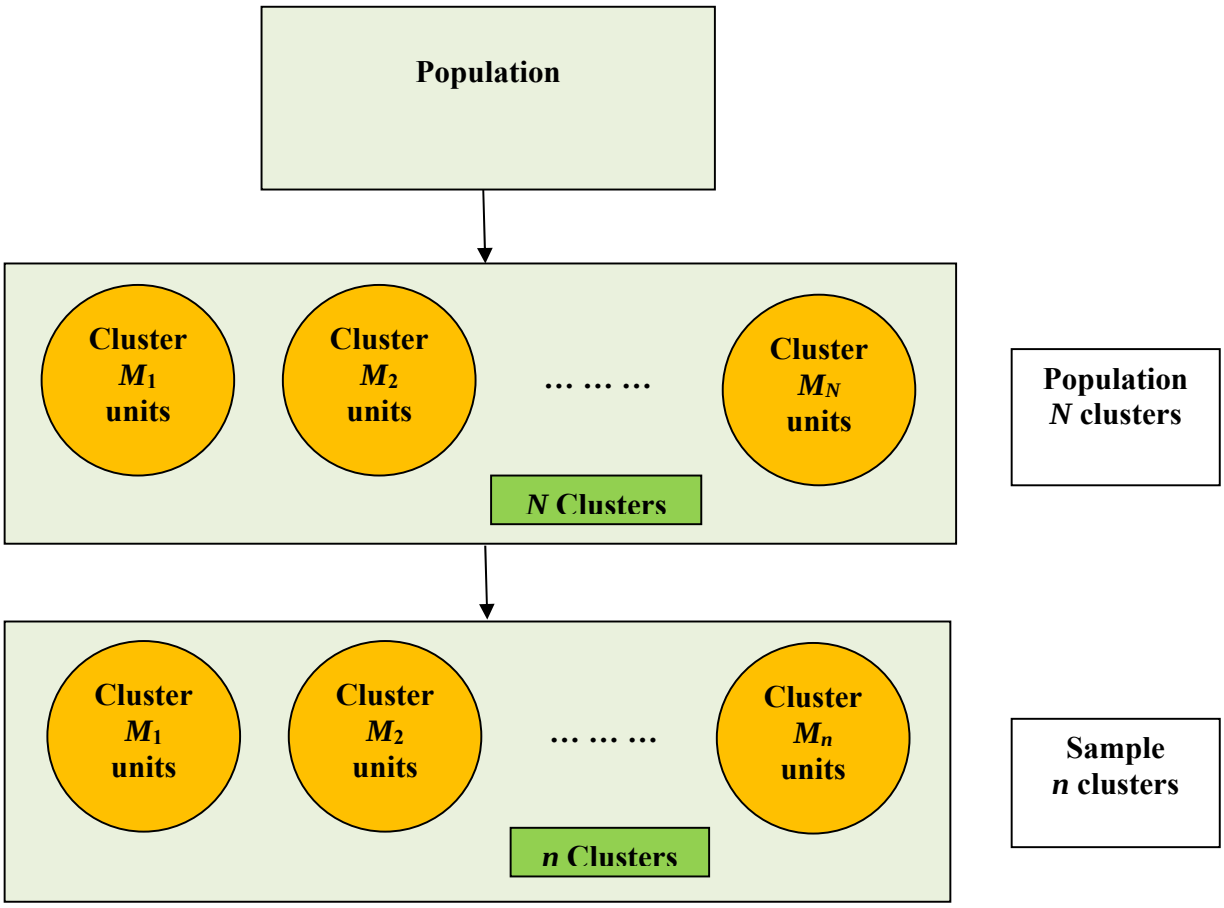
Case of unequal clusters:

In practice, the equal size of clusters are available only when planned. For example, in a screw manufacturing company, the packets of screws can be prepared such that every packet contains same number of screws. In real applications, it is hard to get clusters of equal size. For example, the villages with equal areas are difficult to find, the districts with same number of persons are difficult to find, the number of members in a household may not be same in each household in a given area.

Let there be N clusters and M_i be the size of i^{th} cluster, let

$$M_0 = \sum_{i=1}^{N} M_i$$
$$\bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i$$
$$\bar{y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} \text{ : mean of } i^{th} \text{ cluster}$$
$$\bar{Y} = \frac{1}{M_0} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$$
$$= \sum_{i=1}^{N} \frac{M_i}{M_0} \bar{y}_i$$
$$= \frac{1}{N} \sum_{i=1}^{N} \frac{M_i}{\bar{M}} \bar{y}_i$$

Suppose that n clusters are selected with SRSWOR and all the elements in these selected clusters are surveyed. Assume that M_i's $(i = 1, 2, ..., N)$ are known.



Based on this scheme, several estimators can be obtained to estimate the population mean. We consider four type of such estimators.

1. Mean of cluster means:

Consider the simple arithmetic mean of the cluster means as

$$\bar{\bar{y}}_c = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$$

$$E\left(\bar{\bar{y}}_c\right) = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_i$$

$$\neq \bar{Y} \quad (\text{where } \bar{Y} = \sum_{i=1}^{N}\frac{M_i}{M_0}\bar{y}_i).$$

The bias of $\bar{\bar{y}}_c$ is

$$\begin{aligned}
Bias\left(\bar{\bar{y}}_c\right) &= E\left(\bar{\bar{y}}_c\right) - \bar{Y} \\
&= \frac{1}{N}\sum_{i=1}^{N}\bar{y}_i - \sum_{i=1}^{N}\left(\frac{M_i}{M_0}\right)\bar{y}_i \\
&= -\frac{1}{M_0}\left[\sum_{i=1}^{N} M_i\bar{y}_i - \frac{M_0}{N}\sum_{i=1}^{N}\bar{y}_i\right] \\
&= -\frac{1}{M_0}\left[\sum_{i=1}^{N} M_i\bar{y}_i - \frac{\left(\sum_{i=1}^{N} M_i\right)\left(\sum_{i=1}^{N}\bar{y}_i\right)}{N}\right] \\
&= -\frac{1}{M_0}\sum_{i=1}^{N}(M_i - \bar{M})(\bar{y}_i - \bar{Y}) \\
&= -\left(\frac{N-1}{M_0}\right)S_{m\bar{y}}
\end{aligned}$$

$Bias\left(\bar{\bar{y}}_c\right) = 0$ if M_i and $\bar{y}_i$ are uncorrelated .

The mean squared error is

$$\begin{aligned}
MSE\left(\bar{\bar{y}}_c\right) &= Var\left(\bar{\bar{y}}_c\right) + \left[Bias\left(\bar{\bar{y}}_c\right)\right]^2 \\
&= \frac{N-n}{Nn}S_b^2 + \left(\frac{N-1}{M_0}\right)^2 S_{m\bar{y}}^2
\end{aligned}$$

where

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_i - \bar{Y})^2$$

$$S_{m\bar{y}} = \frac{1}{N-1}\sum_{i=1}^{N}(M_i - \bar{M})(\bar{y}_i - \bar{Y}).$$

An estimate of $Var\left(\bar{\bar{y}}_c\right)$ is

$$\widehat{Var}\left(\bar{\bar{y}}_c\right) = \frac{N-n}{Nn}s_b^2$$

where $s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{y}_i - \bar{\bar{y}}_c\right)^2$.

2. Weighted mean of cluster means

Consider the arithmetic mean based on cluster total as

$$\bar{y}_c^* = \frac{1}{n\bar{M}} \sum_{i=1}^{n} M_i \bar{y}_i$$

$$\begin{aligned} E(\bar{y}_c^*) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\bar{M}} E(\bar{y}_i M_i) \\ &= \frac{n}{n} \frac{1}{M_0} \sum_{i=1}^{N} M_i \bar{y}_i \\ &= \frac{1}{M_0} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} \\ &= \bar{Y}. \end{aligned}$$

Thus $\bar{y}_c^*$ is an unbiased estimator of $\bar{Y}$. The variance of $\bar{y}_c^*$ and its estimate are given by

$$\begin{aligned} Var(\bar{y}_c^*) &= Var\left(\frac{1}{n} \sum_{i=1}^{n} \frac{M_i}{\bar{M}} \bar{y}_i \right) \\ &= \frac{N-n}{Nn} S_b^{*2} \\ \widehat{Var}(\bar{y}_c^*) &= \frac{N-n}{Nn} s_b^{*2} \end{aligned}$$

where

$$S_b^{*2} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{M_i}{\bar{M}} \bar{y}_i - \bar{Y} \right)^2$$

$$s_b^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{M_i}{\bar{M}} \bar{y}_i - \bar{y}_c^* \right)^2$$

$$E(s_b^{*2}) = S_b^{*2}.$$

Note that the expressions of variance of $\bar{y}_c^*$ and its estimate can be derived using directly the theory of SRSWOR as follows:

Let $z_i = \frac{M_i}{\bar{M}} \bar{y}_i$, then $\bar{y}_c^* = \frac{1}{n} \sum_{i=1}^{n} z_i = \bar{z}$.

Since SRSWOR is followed, so

$$Var(\bar{y}_c^*) = Var(\bar{z}) = \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{n}(z_i - \bar{Y})^2$$
$$= \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}\bar{y}_i - \bar{Y}\right)^2$$
$$= \frac{N-n}{Nn}S_b^{*2}.$$

Since

$$E(s_b^{*2}) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2\right]$$
$$= E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i}{\bar{M}}\bar{y}_i - \bar{y}_c^*\right)^2\right]$$
$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}\bar{y}_i - \bar{Y}\right)^2$$
$$= S_b^{*2}$$

So an unbiased estimator of variance can be easily derived.

3. Estimator based on ratio method of estimation

Consider the weighted mean of the cluster means as

$$\bar{y}_c^{**} = \frac{\sum_{i=1}^{n} M_i\bar{y}_i}{\sum_{i=1}^{n} M_i}$$

It is easy to see that this estimator is a biased estimator of the population mean. Before deriving its bias and mean squared error, we note that this estimator can be derived using the philosophy of ratio method of estimation. To see this, consider the study variable U_i and auxiliary variable V_i as

$$U_i = \frac{M_i\bar{y}_i}{\bar{M}}$$
$$V_i = \frac{M_i}{\bar{M}}\ i = 1, 2, ..., N$$

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

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$$

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i.$$

The ratio estimator based on U and V is

$$\hat{\bar{Y}}_R = \frac{\bar{u}}{\bar{v}}\bar{V}$$

$$= \frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} v_i}$$

$$= \frac{\sum_{i=1}^{n} \frac{M_i \bar{y}_i}{\bar{M}}}{\sum_{i=1}^{n} \frac{M_i}{\bar{M}}}$$

$$= \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}.$$

Since the ratio estimator is biased, so $\bar{y}_c^{**}$ is also a biased estimator. The approximate bias and mean squared errors of $\bar{y}_c^{**}$ can be derived directly by using the bias and *MSE* of ratio estimator. So using the results from the ratio method of estimation, the bias up to second order of approximation is given as follows

$$Bias(\bar{y}_c^{**}) = \frac{N-n}{Nn}\left(\frac{S_v^2}{\bar{V}^2} - \frac{S_{uv}}{\bar{U}\bar{V}}\right)\bar{U}$$

$$= \frac{N-n}{Nn}\left(S_v^2 - \frac{S_{uv}}{\bar{U}}\right)\bar{U}$$

where $\bar{U} = \frac{1}{N}\sum_{i=1}^{N} U_i = \frac{1}{N\bar{M}}\sum_{i=1}^{N} M_i \bar{y}_i$

$$S_v^2 = \frac{1}{N-1}\sum_{i=1}^{N}(V_i - \bar{V})^2$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}} - 1\right)^2$$

$$S_{uv} = \frac{1}{N-1}\sum_{i=1}^{N}(U_i - \bar{U})(V_i - \bar{V})$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i\bar{y}_i}{\bar{M}} - \frac{1}{N\bar{M}}\sum_{i=1}^{N}M_i\bar{y}_i\right)\left(\frac{M_i}{\bar{M}} - 1\right)$$

$$R_{uv} = \frac{\bar{U}}{\bar{V}} = \bar{U} = \frac{1}{N\bar{M}}\sum_{i=1}^{N}M_i\bar{y}_i.$$

The *MSE* of $\bar{y}_c^{**}$ up to second order of approximation can be obtained as follows:

$$MSE(\bar{y}_c^{**}) = \frac{N-n}{Nn}\left(S_u^2 + R^2S_v^2 - 2RS_{uv}\right)$$

where $S_u^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i\bar{y}_i}{\bar{M}} - \frac{1}{N\bar{M}}\sum_{i=1}^{N}M_i\bar{y}_i\right)^2$

Alternatively,

$$MSE(\bar{y}_c^{**}) = \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}\left(U_i - R_{uv}V_i\right)^2$$

$$= \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}\left[\frac{M_i\bar{y}_i}{\bar{M}} - \left(\frac{1}{N\bar{M}}\sum_{i=1}^{N}M_i\bar{y}_i\right)\frac{M_i}{\bar{M}}\right]^2$$

$$= \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}\right)^2\left[\bar{y}_i - \frac{\sum_{i=1}^{N}M_i\bar{y}_i}{N\bar{M}}\right]^2.$$

An estimator of *MSE* can be obtained as

$$\widehat{MSE}(\bar{y}_c^{**}) = \frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i}{\bar{M}}\right)^2(\bar{y}_i - \bar{y}_c^{**})^2.$$

The estimator $\bar{y}_c^{**}$ is biased but consistent.

4. Estimator based on unbiased ratio type estimation

Since $\bar{\bar{y}}_c = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$ (where $\bar{y}_i = \frac{1}{M_i}\sum_{i=1}^{M_i} y_{ij}$) is a biased estimator of population mean and

$$Bias(\bar{\bar{y}}_c) = -\left(\frac{N-1}{M_0}\right) S_{m\bar{y}}$$
$$= -\left(\frac{N-1}{N\bar{M}}\right) S_{m\bar{y}}$$

Since SRSWOR is used, so

$$s_{m\bar{y}} = \frac{1}{n-1}\sum_{i=1}^{n}(M_i - \bar{m})(\bar{y}_i - \bar{\bar{y}}_c), \qquad \bar{m} = \frac{1}{n}\sum_{i=1}^{n} M_i$$

is an unbiased estimator of

$$S_{m\bar{y}} = \frac{1}{N-1}\sum_{i=1}^{N}(M_i - \bar{M})(\bar{y}_i - \bar{Y}),$$

i.e., $E(s_{m\bar{y}}) = S_{m\bar{y}}$.

So it follow that

$$E(\bar{\bar{y}}_c) - \bar{Y} = -\left(\frac{N-1}{N\bar{M}}\right) E(s_{m\bar{y}})$$

or $E\left[\bar{\bar{y}}_c + \left(\frac{N-1}{N\bar{M}}\right) s_{m\bar{y}}\right] = \bar{Y}$.

So

$$\bar{\bar{y}}_c^{**} = \bar{\bar{y}}_c + \left(\frac{N-1}{N\bar{M}}\right) s_{m\bar{y}}$$

is an unbiased estimator of the population mean $\bar{Y}$.

This estimator is based on unbiased ratio type estimator. This can be obtained by replacing the study variable (earlier y_i) by $\frac{M_i}{\bar{M}}\bar{y}_i$ and auxiliary variable (earlier x_i) by $\frac{M_i}{\bar{M}}$. The exact variance of this estimate is complicated and does not reduces to a simple form. The approximate variance upto first order of approximation is

$$Var\left(\bar{\bar{y}}_c^{**}\right) = \frac{1}{n(N-1)}\sum_{i=1}^{N}\left[\left(\frac{M_i}{\bar{M}}\bar{y}_i - \bar{Y}\right) - \left(\frac{1}{N\bar{M}}\sum_{i=1}^{N}\bar{y}_i\right)(M_i - \bar{M})\right]^2.$$

A consistent estimate of this variance is

$$\widehat{Var}\left(\bar{\bar{y}}_{\boldsymbol{c}}^{**}\right)=\frac{1}{n(n-1)}\sum_{i=1}^{n}\left[\left(\frac{M_i}{\bar{M}}\bar{y}_i-\bar{y}_c\right)-\left(\frac{1}{n\bar{M}}\sum_{i=1}^{n}\bar{y}_i\right)\left(M_i-\frac{\sum_{i=1}^{n}M_i}{n}\right)\right]^2.$$

The variance of $\bar{\bar{y}}_{\boldsymbol{c}}^{**}$ will be smaller than that of $\bar{y}_c^{**}$ (based on the ratio method of estimation) provided the regression coefficient of $\frac{M_i\bar{y}_i}{\bar{M}}$ on $\frac{M_i}{\bar{M}}$ is nearer to $\frac{1}{N}\sum_{i=1}^{N}\bar{y}_i$ than to $\frac{1}{M_0}\sum_{i=1}^{N}M_i\bar{y}_i$.

Comparison between SRS and cluster sampling:

In case of unequal clusters, $\sum_{i=1}^{n}M_i$ is a random variable such that

$$E\left(\sum_{i=1}^{n}M_i\right)=n\bar{M}.$$

Now if a sample of size $n\bar{M}$ is drawn from a population of size $N\bar{M}$, then the variance of corresponding sample mean based on SRSWOR is

$$\begin{aligned}Var(\bar{y}_{SRS})&=\frac{N\bar{M}-n\bar{M}}{N\bar{M}}\frac{S^2}{n\bar{M}}\\&=\frac{N-n}{Nn}\frac{S^2}{\bar{M}}.\end{aligned}$$

This variance can be compared with any of the four proposed estimators.

For example, in case of

$$\bar{y}_c^{*}=\frac{1}{n\bar{M}}\sum_{i=1}^{n}M_i\bar{y}_i$$

$$\begin{aligned}Var(\bar{y}_c^{*})&=\frac{N-n}{Nn}S_b^{*2}\\&=\frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}\bar{y}_i-\bar{Y}\right)^2.\end{aligned}$$

The relative efficiency of $\bar{y}_c^{**}$ relative to SRS based sample mean

$$\begin{aligned}E&=\frac{Var(\bar{y}_{SRS})}{Var(\bar{y}_c^{*})}\\&=\frac{S^2}{\bar{M}S_b^{*2}}.\end{aligned}$$

For $Var(\bar{y}_c^{*}) < Var(\bar{y}_{SRS})$, the variance between the clusters (S_b^{*2}) should be less. So the clusters should be formed in such a way that the variation between them is as small as possible.

Sampling with replacement and unequal probabilities (PPSWR)

In many practical situations, the cluster total for the study variable is likely to be positively correlated with the number of units in the cluster. In this situation, it is advantageous to select the clusters with probability proportional to the number of units in the cluster instead of with equal probability, or to stratify the clusters according to their sizes and then to draw a SRSWOR of clusters from each of the stratum. We consider here the case where clusters are selected with probability proportional to the number of units in the cluster and with replacement.

Suppose that n clusters are selected with ppswr, the size being the number of units in the cluster. Here P_i is the probability of selection assigned to the i^{th} cluster which is given by

$$P_i = \frac{M_i}{M_0} = \frac{M_i}{N\bar{M}}, \quad i = 1, 2, ..., N.$$

Consider the following estimator of the population mean:

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

Then this estimator can be expressed as

$$\hat{\bar{Y}}_c = \frac{1}{n}\sum_{i=1}^{N} \alpha_i \bar{y}_i$$

where α_i denotes the number of times the i^{th} cluster occurs in the sample. The random variables $\alpha_1, \alpha_2, ..., \alpha_N$ follow a multinomial probability distribution with

$$E(\alpha_i) = nP_i, \; Var(\alpha_i) = nP_i(1 - P_i)$$
$$Cov(\alpha_i, \alpha_j) = -nP_iP_j, \quad i \neq j.$$

Hence,

$$\begin{aligned} E(\hat{\bar{Y}}_c) &= \frac{1}{n}\sum_{i=1}^{N} E(\alpha_i)\bar{y}_i \\ &= \frac{1}{n}\sum_{i=1}^{N} nP_i\bar{y}_i \\ &= \sum_{i=1}^{N} \frac{M_i}{N\bar{M}}\bar{y}_i \\ &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}}{N\bar{M}} = \bar{Y}. \end{aligned}$$

Thus $\hat{\bar{Y}}_c$ is an unbiased estimator of $\bar{Y}$.

We now derive the variance of $\hat{\bar{Y}}_c$.

From $\hat{\bar{Y}}_c = \frac{1}{n}\sum_{i=1}^{N}\alpha_i \bar{y}_i$,

$$
\begin{aligned}
Var(\hat{\bar{Y}}_c) &= \frac{1}{n^2}\left[\sum_{i=1}^{N} Var(\alpha_i)\bar{y}_i^2 + \sum_{i\neq j}^{N} Cov(\alpha_i,\alpha_j)\bar{y}_i\bar{y}_j\right] \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} P_i(1-P_i)\bar{y}_i^2 - \sum_{i\neq j}^{N} P_iP_j\bar{y}_i\bar{y}_j\right] \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} P_i\bar{y}_i^2 - \left(\sum_{i\neq j}^{N} P_i\bar{y}_i\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{N} P_i\left(\bar{y}_i - \bar{Y}\right)^2 \\
&= \frac{1}{nN\bar{M}}\sum_{i=1}^{N} M_i(\bar{y}_i - \bar{Y})^2.
\end{aligned}
$$

An unbiased estimator of the variance of $\hat{\bar{Y}}_c$ is

$$\widehat{Var}(\hat{\bar{Y}}_c) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\bar{Y}}_c)^2$$

which can be seen to satisfy the unbiasedness property as follows:

Consider

$$
\begin{aligned}
&E\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\bar{Y}}_c)^2\right] \\
&= E\left[\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}(\bar{y}_i^2 - n\hat{\bar{Y}}_c)^2\right)\right] \\
&= \frac{1}{n(n-1)}\left[E\left(\sum_{i=1}^{n}\alpha_i\bar{y}_i^2\right) - nVar(\hat{\bar{Y}}_c) - n\bar{Y}^2\right]
\end{aligned}
$$

where $E(\alpha_i) = nP_i, Var(\alpha_i) = nP_i(1-P_i), Cov(\alpha_i,\alpha_j) = -nP_iP_j, i \neq j$

$$
\begin{aligned}
E\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\bar{Y}}_c)^2\right] &= \frac{1}{n(n-1)}\left[\sum_{i=1}^{N} n_iP_i\bar{y}_i^2 - n\frac{1}{n}\sum_{i=1}^{N} P_i(\bar{y}_i - \bar{Y})^2 - n\bar{Y}^2\right] \\
&= \frac{1}{(n-1)}\left[\sum_{i=1}^{N} P_i(\bar{y}_i^2 - \bar{Y}^2) - \frac{1}{n}\sum_{i=1}^{N} P_i(\bar{y}_i - \bar{Y})^2\right] \\
&= \frac{1}{(n-1)}\left[\sum_{i=1}^{N} P_i(\bar{y}_i - \bar{Y})^2 - \frac{1}{n}\sum_{i=1}^{N} P_i(\bar{y}_i - \bar{Y})^2\right] \\
&= \frac{1}{(n-1)}\sum_{i=1}^{N} P_i(\bar{y}_i - \bar{Y})^2 \\
&= Var(\hat{\bar{Y}}_c).
\end{aligned}
$$