Chapter 4

Stratified Sampling

An important objective in any estimation problem is to obtain an estimator of a population parameter which can take care of the salient features of the population. If the population is homogeneous with respect to the characteristic under study, then the method of simple random sampling will yield a homogeneous sample, and in turn, the sample mean will serve as a good estimator of the population mean. Thus, if the population is homogeneous with respect to the characteristic under study, then the sample drawn through simple random sampling is expected to provide a representative sample. Moreover, the variance of the sample mean not only depends on the sample size and sampling fraction but also on the population variance. In order to increase the precision of an estimator, we need to use a sampling scheme which can reduce the heterogeneity in the population. If the population is heterogeneous with respect to the characteristic under study, then one such sampling procedure is a stratified sampling.

The basic idea behind the stratified sampling is to

- divide the whole heterogeneous population into smaller groups or subpopulations, such that the sampling units are homogeneous with respect to the characteristic under study within the subpopulation and
- heterogeneous with respect to the characteristic under study between/among the subpopulations. Such subpopulations are termed as **strata.**
- Treat each subpopulation as a separate population and draw a sample by SRS from each stratum.

[Note: 'Stratum' is singular and 'strata' is plural].

Example: In order to find the average height of the students in a school of class 1 to class 12, the height varies a lot as the students in class 1 are of age around 6 years, and students in class 10 are of age around 16 years. So one can divide all the students into different subpopulations or strata such as

Students of class 1, 2 and 3: Stratum 1

Students of class 4, 5 and 6: Stratum 2

Students of class 7, 8 and 9: Stratum 3

Students of class 10, 11 and 12: Stratum 4

Now draw the samples by SRS from each of the strata 1, 2, 3 and 4. All the drawn samples combined together will constitute the final stratified sample for further analysis.

Notations:

We use the following symbols and notations:

N : Population size

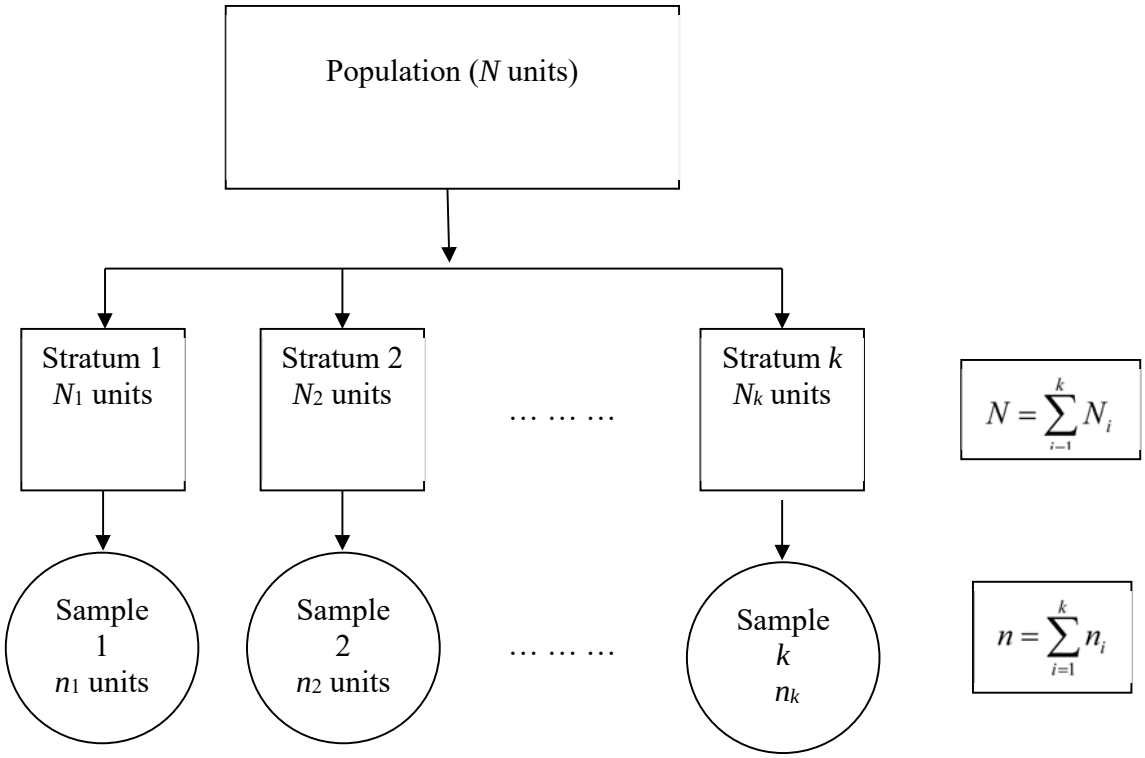
k : Number of strata

N_i : Number of sampling units in i^{th} strata

$$N = \sum_{i=1}^{k} N_i$$

n_i : Number of sampling units to be drawn from i^{th} stratum.

$n = \sum_{i=1}^{k} n_i$: Total sample size



Procedure of stratified sampling

Divide the population of N units into k strata. Let the i^{th} stratum has $N_1, i = 1,2,...,k$ number of units.

- Strata are constructed such that they are non-overlapping and homogeneous with respect to the characteristic under study such that $\sum_{i=1}^{k} N_i = N$.
- Draw a sample of size n_i from i^{th} $(i = 1,2,...,k)$ stratum using SRS (preferably WOR) independently from each stratum.
- All the sampling units drawn from each stratum will constitute a stratified sample of size $n = \sum_{i=1}^{k} n_i$.

Difference between stratified and cluster sampling schemes

In stratified sampling, the strata are constructed such that they are

- within homogeneous and
- among heterogeneous.

In cluster sampling, the clusters are constructed such that they are

- within heterogeneous and
- among homogeneous.

[Note: We discuss the cluster sampling later.]

Issues in the estimation of parameters in stratified sampling

Divide the population of N units in k strata. Let the i^{th} stratum has N_i, $i = 1,2,...,k$ number of units. Note that there are k independent samples drawn through SRS of sizes $n_1, n_2, ..., n_k$ from each of the strata. So, one can have k estimators of a parameter based on the sizes $n_1, n_2, ..., n_k$ respectively. Our interest is not to have k different estimators of the parameters, but the ultimate goal is to have a single estimator. In this case, an important issue is how to combine the different sample information together into one estimator, which is good enough to provide information about the parameter.

We now consider the estimation of population mean and population variance from a stratified sample.

Estimation of population mean and its variance

Let

Y : characteristic under study,

y_{ij} : value of j^{th} unit in i^{th} stratum $j = 1,2,\ldots,n_i$, $i = 1,2,...,k$,

$\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$: population mean of i^{th} stratum

$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$: sample mean from i^{th} stratum

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{k} N_i\bar{Y}_i = \sum_{i=1}^{k} w_i\bar{Y}_i$: population mean where $w_i = \frac{N_i}{N}$.

Estimation of population mean:

First, we discuss the estimation of the population mean.

Note that the population mean is defined as the weighted arithmetic mean of stratum means in the case of stratified sampling where the weights are provided in terms of strata sizes.

Based on the expression $\bar{Y} = \frac{1}{N}\sum_{i=1}^{k} N_i\bar{Y}_i$, one may choose the sample mean

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

as a possible estimator of $\bar{Y}$.

Since the sample in each stratum is drawn by SRS, so

$E(\bar{y}_i) = \bar{Y}_i$,

thus

$$\begin{aligned} E(\bar{y}) &= \frac{1}{n}\sum_{i=1}^{k} n_i E(\bar{y}_i) \\ &= \frac{1}{n}\sum_{i=1}^{k} n_i \bar{Y}_i \\ &\neq \bar{Y} \end{aligned}$$

and $\bar{y}$ turns out to be a biased estimator of $\bar{Y}$. Based on this, one can modify $\bar{y}$ so as to obtain an unbiased estimator of $\bar{Y}$. Consider the stratum mean which is defined as the weighted arithmetic mean of strata sample means with strata sizes as weights given by

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i\bar{y}_i.$$

Now

$$E(\bar{y}_{st}) = \frac{1}{N}\sum_{i=1}^{k} N_i E(\bar{y}_i)$$
$$= \frac{1}{N}\sum_{i=1}^{k} N_i \bar{Y}_i$$
$$= \bar{Y}$$

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

Variance of $\bar{y}_{st}$

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 Var(\bar{y}_i) + \sum_{i(\neq j)=1}^{k}\sum_{j=1}^{n_i} w_i w_j Cov(\bar{y}_i, \bar{y}_j).$$

Since all the samples have been drawn independently from each of the strata by SRSWOR so

$$Cov(\bar{y}_i, \bar{y}_j) = 0, i \neq j$$

$$Var(\bar{y}_i) = \frac{N_i - n_i}{N_i n_i} S_i^2$$

where

$$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2.$$

Thus

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 \frac{N_i - n_i}{N_i n_i} S_i^2$$
$$= \sum_{i=1}^{k} w_i^2 \left(1 - \frac{n_i}{N_i}\right)\frac{S_i^2}{n_i}.$$

Observe that $Var(\bar{y}_{st})$ is small when S_i^2 is small. This observation suggests how to construct the strata. If S_i^2 is small for all i = 1,2,...,k, then $Var(\bar{y}_{st})$ will also be small. That is why it was mentioned earlier that the strata are to be constructed such that they are within homogeneous, i.e., S_i^2 is small and among heterogeneous.

For example, the units in geographical proximity will tend to be more closer. The consumption pattern in the households will be similar within a lower income group housing society and within a higher income group housing society, whereas they will differ a lot between the two housing societies based on income.

Estimate of Variance

Since the samples have been drawn by SRSWOR, so

$$E(s_i^2) = S_i^2$$

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

and $\widehat{Var}(\bar{y}_i) = \frac{N_i - n_i}{N_i n_i} s_i^2$

so $\widehat{Var}(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 \ \widehat{Var}(\bar{y}_i)$

$$= \sum_{i=1}^{k} w_i^2 \left(\frac{N_i - n_i}{N_i n_i}\right) s_i^2$$

Note: If SRSWR is used instead of SRSWOR for drawing the samples from each stratum, then in this case

$$\bar{y}_{st} = \sum_{i=1}^{k} w_i \bar{y}_i$$

$$E(\bar{y}_{st}) = \bar{Y}$$

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 \left(\frac{N_i - 1}{N_i n_i}\right) S_i^2 = \sum_{i=1}^{k} w_i^2 \frac{\sigma_i^2}{n_i}$$

$$\widehat{Var}(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{w_i^2 s_i^2}{n_i}$$

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

Advantages of stratified sampling

1. Data of known precision may be required for certain parts of the population.
   This can be accomplished with a more careful investigation to a few strata.
   Example: In order to know the direct impact of the hike in petrol prices, the population can be divided into strata like lower income group, middle-income group and higher income group. Obviously, the higher income group is more affected than the lower-income group. So more careful investigation can be made in the higher income group strata.
2. Sampling problems may differ in different parts of the population.
   Example: To study the consumption pattern of households, the people living in houses, hotels, hospitals, prison etc. are to be treated differently.

3. Administrative convenience can be exercised in stratified sampling.

   Example: In taking a sample of villages from a big state, it is more administratively convenient to consider the districts as strata so that the administrative set up at district level may be used for this purpose. Such administrative convenience and the convenience in the organization of fieldwork are important aspects in national level surveys.

4. Full cross-section of the population can be obtained through stratified sampling. It may be possible in SRS that some large part of the population may remain unrepresented. Stratified sampling enables one to draw a sample representing different segments of the population to any desired extent. The desired degree of representation of some specified parts of the population is also possible.
5. Substantial gain in efficiency is achieved if the strata are formed intelligently.
6. In the case of skewed population, use of stratification is of importance since larger weight may have to be given for the few extremely large units, which in turn reduces the sampling variability.
7. When estimates are required not only for the population but also for the subpopulations, then the stratified sampling is helpful.
8. When the sampling frame for subpopulations is more easily available than the sampling frame for the whole population, then stratified sampling is helpful.
9. If the population is large, then it is convenient to sample separately from the strata rather than the entire population.
10. The population mean or population total can be estimated with higher precision by suitably providing the weights to the estimates obtained from each stratum.

Allocation problem and choice of sample sizes is different strata

Question: How to choose the sample sizes $n_1, n_2, ..., n_k$ so that the available resources are used in an effective way?

There are two aspects of choosing the sample sizes:

(i) Minimize the cost of survey for a specified precision.

(ii) Maximize the precision for a given cost.

Note: The sample size cannot be determined by minimizing both the cost and variability simultaneously. The cost function is directly proportional to the sample size, whereas variability is inversely proportional to the sample size.

Based on different ideas, some allocation procedures are as follows:

1. Equal allocation

Choose the sample size n_i to be the same for all the strata.

Draw samples of equal size from each stratum.

Let n be the sample size and k be the number of strata, then

$n_i = \frac{n}{k}$ for all $i = 1, 2, ..., k.$

2. Proportional allocation

For fixed k, select n_i such that it is proportional to stratum size N_i, i.e.,

$$n_i \propto N_i$$

or $n_i = \delta N_i$

where δ is the constant of proportionality.

$$\sum_{i=1}^{k} n_i = \delta$$

or $n = \delta N$

$$\Rightarrow \delta = \frac{n}{N}.$$

Thus $n_i = \left(\frac{n}{N}\right) N_i.$

Such allocation arises from considerations like operational convenience.

3. Neyman or optimum allocation

This allocation considers the size of strata as well as variability

$$n_i \propto N_i S_i$$

$$n_i = \delta^* N_i S_i$$

where δ^* is the constant of proportionality.

$$\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \delta^* N_i S_i$$

or $n = \delta^* \sum_{i=1}^{k} N_i S_i$

or $\delta^* = \frac{n}{\sum_{i=1}^{k} N_i S_i}.$

Thus $n_i = \dfrac{nN_iS_i}{\sum_{i=1}^{k} N_iS_i}$.

This allocation arises when the $Var\left(\bar{y}_{st}\right)$ is minimized subject to the constraint $\sum_{i=1}^{k} n_i$ (prespecified).

There are some limitations to the optimum allocation. The knowledge of $S_i (i = 1,2,...,k)$ is needed to know n_i. If there are more than one characteristics, then they may lead to conflicting allocation.

Choice of sample size based on the cost of survey and variability

The cost of the survey depends upon the nature of the survey. A simple choice of the cost function is

$$C = C_0 + \sum_{i=1}^{k} C_i n_i$$

where

C : total cost

C_0 : overhead cost, e.g., setting up the office, training people etc

C_i : cost per unit in the i^{th} stratum

$\sum_{i=1}^{k} C_i n_i$: total cost within the sample.

To find n_i under this cost function, consider the Lagrangian function with a Lagrangian multiplier λ as

$$\begin{aligned}
\phi &= Var(\bar{y}_{st}) + \lambda^2 (C - C_0) \\
&= \sum_{i=1}^{k} w_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_i^2 + \lambda^2 \sum_{i=1}^{k} C_i n_i \\
&= \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{n_i} + \lambda^2 \sum_{i=1}^{k} C_i n_i - \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{N_i} \\
&= \sum_{i=1}^{k} \left[\frac{w_i S_i}{\sqrt{n_i}} - \lambda \sqrt{C_i n_i} \right]^2 + \text{terms independent of } n_i.
\end{aligned}$$

Thus ϕ is minimum when

$$\frac{w_i S_i}{\sqrt{n_i}} = \lambda \sqrt{C_i n_i} \text{ for all } i$$

or $n_i = \dfrac{1}{\lambda} \dfrac{w_i S_i}{\sqrt{C_i}}$.

How to determine λ?

There are two ways to determine λ.

(i) Minimize variability for a fixed cost.

(ii) Minimize cost for given variability.

We consider both cases.

(i) Minimize variability for fixed cost

Let $C = C_0^*$ be the pre-specified cost which is fixed.

So $\sum_{i=1}^{k} C_i n_i = C_0^*$

or $\sum_{i=1}^{k} C_i \frac{w_i S_i}{\lambda\sqrt{C_i}} = C_0^*$

or $\lambda = \frac{\sum_{i=1}^{k} \sqrt{C_i}\, w_i S_i}{C_0^*}$.

Substituting λ in the expression for $n_i = \frac{1}{\lambda}\frac{w_i S_i}{\sqrt{C_i}}$, the optimum n_i is obtained as

$$n_i^* = \frac{w_i S_i}{\sqrt{C_i}}\left(\frac{C_0^*}{\sum_{i=1}^{k}\sqrt{C_i}\, w_i S_i}\right).$$

The required sample size to estimate $\bar{Y}$ such that the variance is minimum for the given cost $C = C_0^*$ is

$$n = \sum_{i=1}^{k} n_i^*.$$

(ii) Minimize cost for a given variability

Let $V = V_0$ be the pre-specified variance. Now determine n_i such that

$$\sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) w_i^2 S_i^2 = V_0$$

or $\sum_{i=1}^{k} \frac{w_i^2 S_i^2}{n_i} = V_0 + \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{N_i}$

or $\sum_{i=1}^{k} \frac{\lambda\sqrt{C_i}}{w_i S_i} w_i^2 S_i^2 = V_0 + \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{N_i}$

or $\lambda = \dfrac{V_0 + \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{N_i}}{\sum_{i=1}^{k} w_i S_i \sqrt{C_i}}$ (after substituting $n_i = \dfrac{1}{\lambda} \dfrac{w_i S_i}{\sqrt{C_i}}$).

Thus the optimum n_i is

$$\tilde{n}_i = \frac{w_i S_i}{\sqrt{C_i}} \left(\frac{\sum_{i=1}^{k} w_i S_i \sqrt{C_i}}{V_0 + \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{N_i}} \right).$$

So the required sample size to estimate $\bar{Y}$ such that cost C is minimum for a prespecified variance V_0 is $n = \sum_{i=1}^{k} \tilde{n}_i$.

Sample size under proportional allocation for fixed cost and for fixed variance

(i) **If cost $C = C_0$ is fixed** then $C_0 = \sum_{i=1}^{k} C_i n_i$.

Under proportional allocation, $n_i = \dfrac{n}{N} N_i = n w_i$

So $C_0 = n \sum_{i=1}^{k} w_i C_i$

or $n = \dfrac{C_0}{\sum_{i=1}^{k} w_i C_i}$.

Thus $n_i = \dfrac{C_o w_i}{\sum w_i C_i}$.

The required sample size to estimate $\bar{Y}$ in this case is $n = \sum_{i=1}^{k} n_i$.

(ii) If variance = V_0 is fixed, then

$$\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)w_i^2S_i^2 = V_0$$

$$\text{or} \quad \sum_{i=1}^{k}\frac{w_i^2S_i^2}{n_i} = V_0 + \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$\text{or} \quad \sum_{i=1}^{k}\frac{w_i^2S_i^2}{nw_i} = V_0 + \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i} \quad (\text{using } n_i = nw_i)$$

$$\text{or} \quad n = \frac{\sum_{i=1}^{k} w_i^2S_i^2}{V_0 + \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}}$$

$$\text{or} \quad n_i = w_i\frac{\sum_{i=1}^{k} w_i^2S_i^2}{V_0 + \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}}.$$

This is known as **Bowley's allocation**.

Variances under different allocations

Now we derive the variance of $\bar{y}_{st}$ under proportional and optimum allocations.

(i) Proportional allocation

Under proportional allocation

$$n_i = \frac{n}{N}N_i$$

and

$$Var(\bar{y})_{st} = \sum_{i=1}^{k}\left(\frac{N_i-n_i}{N_in_i}\right)w_i^2S_i^2$$

$$Var_{prop}(\bar{y})_{st} = \sum_{i=1}^{k}\left(\frac{N_i-\frac{n}{N}N_i}{N_i\frac{n}{N}N_i}\right)\left(\frac{N_i}{N}\right)^2S_i^2$$

$$= \frac{N-n}{Nn}\sum_{i=1}^{k}\frac{N_iS_i^2}{N}$$

$$= \frac{N-n}{Nn}\sum_{i=1}^{k}w_iS_i^2.$$

(ii) Optimum allocation

Under optimum allocation

$$n_i = \frac{nN_iS_i}{\sum_{i=1}^{k} N_iS_i}$$

$$V_{opt}(\bar{y}_{st}) = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)w_i^2S_i^2$$

$$= \sum_{i=1}^{k}\frac{w_i^2S_i^2}{n_i} - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \sum_{i=1}^{k}\left[w_i^2S_i^2\left(\frac{\sum_{i=1}^{k} N_iS_i}{nN_iS_i}\right)\right] - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \sum_{i=1}^{k}\left[\frac{1}{n}.\frac{N_iS_i}{N^2}\left(\sum_{i=1}^{k} N_iS_i\right)\right] - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{k}\frac{N_iS_i}{N}\right)^2 - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i} = \frac{1}{n}\left(\sum_{i=1}^{k} w_iS_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} w_iS_i^2.$$

Comparison of variances of the sample mean under SRS with stratified mean under proportional and optimal allocation:

(a) Proportional allocation:

$$V_{SRS}(\bar{y}) = \frac{N-n}{Nn}S^2$$

$$V_{prop}(\bar{y}_{st}) = \frac{N-n}{Nn}\sum_{i=1}^{k}\frac{N_iS_i^2}{N}.$$

In order to compare $V_{SRS}(\bar{y})$ and $V_{prop}(\bar{y}_{st})$, first we attempt to express S^2 as a function of S_i^2.

Consider

$$(N-1)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij} - \bar{Y})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{N_i}\left[(Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \bar{Y})\right]^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{N_i}(\bar{Y}_i - \bar{Y})^2$$

$$= \sum_{i=1}^{k}(N_i - 1)S_i^2 + \sum_{i=1}^{k} N_i(\bar{Y}_i - \bar{Y})^2$$

$$\frac{N-1}{N}S^2 = \sum_{i=1}^{k}\frac{N_i-1}{N}S_i^2 + \sum_{i=1}^{k}\frac{N_i}{N}(\bar{Y}_i-\bar{Y})^2.$$

For simplification, we assume that N_i is large enough to permit the approximation

$$\frac{N_i-1}{N_i}\approx 1 \text{ and } \frac{N-1}{N}\approx 1.$$

Thus

$$S^2 = \sum_{i=1}^{k}\frac{N_i}{N}S_i^2 + \sum_{i=1}^{k}\frac{N_i}{N}(\bar{Y}_i-\bar{Y})^2$$

or $\frac{N-n}{Nn}S^2 = \frac{N-n}{Nn}\sum_{i=1}^{k}\frac{N_i}{N}S_i^2 + \frac{N-n}{Nn}\sum_{i=1}^{k}\frac{N_i}{N}(\bar{Y}_i-\bar{Y})^2$ (Premultiply by $\frac{N-n}{Nn}$ on both sides)

$$Var_{SRS}(\bar{Y}) = V_{prop}(\bar{y}_{st}) + \frac{N-n}{Nn}\sum_{i=1}^{k}w_i(\bar{Y}_i-\bar{Y})^2$$

Since $\sum_{i=1}^{k}w_i(\bar{Y}_i-\bar{Y})^2 \geq 0,$

$\Rightarrow Var_{prop}(\bar{y}_{st}) \leq Var_{SRS}(\bar{y}).$

A larger gain in the difference is achieved when $\bar{Y}_i$ differs from $\bar{Y}$ more.

(b) Optimum allocation

$$V_{opt}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k}w_iS_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k}w_iS_i^2.$$

Consider

$$V_{prop}(\bar{y}_{st}) - V_{opt}(\bar{y}_{st}) = \left[\left(\frac{N-n}{Nn}\right)\sum_{i=1}^{k}w_iS_i^2\right] - \left[\frac{1}{n}\left(\sum_{i=1}^{k}w_iS_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k}w_iS_i^2\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k}w_iS_i^2 - \left(\sum_{i=1}^{k}w_iS_i\right)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{k}w_iS_i^2 - \frac{1}{n}\bar{S}^2$$

$$= \frac{1}{n}\sum_{i=1}^{k}w_i(S_i-\bar{S})^2$$

where

$$\bar{S} = \sum_{i=1}^{k}w_iS_i$$

$\Rightarrow Var_{prop}(\bar{y}_{st}) - Var_{opt}(\bar{y}_{st}) \geq 0$

or $Var_{opt}(\bar{y}_{st}) \leq Var_{prop}(\bar{y}_{st}).$

The larger gain in efficiency is achieved when S_i differs from $\bar{S}$ more.

Combining the results in (a) and (b), we have

$Var_{opt}(\bar{y}_{st}) \le Var_{prop}(\bar{y}_{st}) \le Var_{SRS}(\bar{y})$

Estimate of variance and confidence intervals

Under SRSWOR, an unbiased estimate of S_i^2 for the i^{th} stratum ($i = 1,2,...,k$) is

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

In stratified sampling,

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 \frac{N_i - n_i}{N_i n_i} S_i^2.$$

So, an unbiased estimate of $Var(\bar{y}_{st})$ is

$$\begin{aligned}\widehat{Var}(\bar{y}_{st}) &= \sum_{i=1}^{k} w_i^2 \frac{N_i - n_i}{N_i n_i} s_i^2 \\ &= \sum_{i=1}^{k} \frac{w_i^2 s_i^2}{n_i} - \sum_{i=1}^{k} \frac{w_i^2 s_i^2}{N_i} \\ &= \sum_{i=1}^{k} \frac{w_i^2 s_i^2}{n_i} - \frac{1}{N}\sum_{i=1}^{k} w_i s_i^2\end{aligned}$$

The second term in this expression represents the reduction due to finite population correction.

The confidence limits of $\bar{Y}$ can be obtained as

$$\bar{y}_{st} \pm t\sqrt{\widehat{Var}(\bar{y}_{st})}$$

assuming $\bar{y}_{st}$ is normally distributed and $\sqrt{\widehat{Var}(\bar{y}_{st})}$ is well determined so that t can be read from normal distribution tables. If only few degrees of freedom are provided by each stratum, then t values are obtained from the table of student's t-distribution.

The distribution of $\sqrt{\widehat{Var}(\bar{y}_{st})}$ is generally complex. An approximate method of assigning an effective number of degrees of freedom (n_e) to $\sqrt{\widehat{Var}(\bar{y}_{st})}$ is $n_e = \dfrac{\left(\sum_{i=1}^{k} g_i s_i^2\right)^2}{\sum_{i=1}^{k} \frac{g_i^2 s_i^4}{n_i - 1}}$

where $g_i = \dfrac{N_i(N_i - n_i)}{n_i}$ and $Min(n_i - 1) \le n_e \le \sum_{i=1}^{k}(n_i - 1)$ assuming y_{ij} are normally distributed.

Modification of optimal allocation

Sometimes in the optimal allocation, the size of subsample exceeds the stratum size. In such a case,

replace n_i by N_i

and recompute the rest of $n_i's$ by the revised allocation.

For example, if $n_1 > N_1$, then take the revised $n_i's$ as

$$\tilde{n}_1 = N_1$$

and

$$\tilde{n}_i = \frac{(n - N_1)w_i S_i}{\sum_{i=2}^{k} w_i S_i} \; ; \quad i = 2,3,...,k$$

provided $\tilde{n}_i \le N_i$ for all $i = 2,3,\ldots,k$.

Suppose in revised allocation, we find that $\tilde{n}_2 > N_2$ then the revised allocation would be

$$\tilde{n}_1 = N_1$$
$$\tilde{n}_2 = N_2$$
$$\tilde{n}_i = \frac{(n - N_1 - N_2)w_i S_i}{\sum_{i=3}^{k} w_i S_i}; i = 3,4,...,k.$$

provided $\tilde{n}_i < N_i$ for all $i = 3,4,...,k$.

We continue this process until every $\tilde{n}_i < N_i$.

In such cases, the formula for the minimum variance of $\bar{y}_{st}$ need to be modified as

$$Min\ Var(\bar{y}_{st}) = \frac{(\sum^{*} w_i S_i)^2}{n^*} - \frac{\sum^{*} w_i S_i^2}{N}$$

where $\sum^{*}$ denotes the summation over the strata in which $\tilde{n}_i \le N_i$ and n^* is the revised total sample size in the strata.

Stratified sampling for proportions

If the characteristic under study is qualitative in nature, then its values will fall into one of the two mutually exclusive complimentary classes *C* and *C'*. Ideally, only two strata are needed in which all the units can be divided depending on whether they belong to *C* or its complement *C'*. Thus is difficult to achieve in practice. So the strata are constructed such that the proportion in *C* varies as much as possible among strata.

Let

$P_i = \frac{A_i}{N_i}$: Proportion of units in *C* in the i^{th} stratum

$p_i = \frac{a_i}{n_i}$: Proportion of units in *C* in the sample from the i^{th} stratum

An estimate of population proportion based on the stratified sampling is

$$p_{st} = \sum_{i=1}^{k} \frac{N_i p_i}{N}.$$

which is based on the indicator variable

$$Y_{ij} = \begin{cases} 1 & \text{when } j^{th} \text{ unit belongs to the } i^{th} \text{ stratum is in } C \\ 0 & \text{otherwise} \end{cases}$$

and $\bar{y}_{st} = p_{st}$.

Here $S_i^2 = \frac{N_i}{N_i - 1} P_i Q_i$

where $Q_i = 1 - P_i$.

Also $Var(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{N_i - n_i}{N_i n_i} w_i^2 S_i^2$.

So $Var(p_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} \frac{N_i^2 (N_i - n_i)}{N_i - 1} \frac{P_i Q_i}{n_i}$.

If the finite population correction can be ignored, then

$$Var(p_{st}) = \sum_{i=1}^{k} w_i^2 \frac{P_i Q_i}{n_i}.$$

If the proportional allocation is used for n_i , then the variance of p_{st} is

$$Var_{prop}(p_{st}) = \frac{N-n}{N} \frac{1}{Nn} \sum_{i=1}^{k} \frac{N_i^2 P_i Q_i}{N_i - 1}$$

$$= \frac{N-n}{Nn} \sum_{i=1}^{k} w_i P_i Q_i$$

and its estimate is

$$\widehat{Var}_{prop}(p_{st}) = \frac{N-n}{Nn}\sum_{i=1}^{k} w_i \frac{p_i q_i}{n_i - 1}.$$

The best choice of n_i such that it minimizes the variance for fixed total sample size is

$$\begin{aligned} n_i &\propto N_i\sqrt{\frac{N_i P_i Q_i}{N_i - 1}} \\ &= N_i\sqrt{P_i Q_i} \end{aligned}$$

Thus $n_i = n \dfrac{N_i\sqrt{P_i Q_i}}{\sum_{i=1}^{k} N_i\sqrt{P_i Q_i}}$.

Similarly, the best choice of n_i such that the variance is minimum for fixed cost $C = C_0 + \sum_{i=1}^{k} C_i n_i$ is

$$n_i = \frac{nN_i\sqrt{\dfrac{P_i Q_i}{C_i}}}{\sum_{i=1}^{k} N_i\sqrt{\dfrac{P_i Q_i}{C_i}}}.$$

Estimation of the gain in precision due to stratification

An obvious question crops up that what is the advantage of stratifying a population in the sense that instead of using SRS, the population is divided into various strata? This is answered by estimating the variance of estimators of population mean under SRS (without stratification) and stratified sampling by evaluating

$$\frac{\widehat{Var}_{SRS}(\bar{y}) - \widehat{Var}(\bar{y}_{st})}{\widehat{Var}(\bar{y}_{st})}.$$

This gives an idea about the gain in efficiency due to stratification.

Since $Var_{SRS}(\bar{y}) = \dfrac{N-n}{Nn}S^2,$ so there is a need to express S^2 in terms of S_i^2. How to estimate S^2 based on a stratified sample?

Consider

$$
\begin{aligned}
(N-1)S^2 &= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{N_i}\left[(Y_{ij}-\bar{Y}_i) + (\bar{Y}_i-\bar{Y})\right]^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y})^2 + \sum_{i=1}^{k}N_i(\bar{Y}_i-\bar{Y})^2 \\
&= \sum_{i=1}^{k}(N_i-1)S_i^2 + \sum_{i=1}^{k}N_i(\bar{Y}_i-\bar{Y})^2 \\
&= \sum_{i=1}^{k}(N_i-1)S_i^2 + N\left[\sum_{i=1}^{k}w_i\bar{Y}_i^2 - \bar{Y}^2\right].
\end{aligned}
$$

In order to estimate S^2, we need to estimates of S_i^2, $\bar{Y}_i^2$ and $\bar{Y}^2$. We consider their estimation one by one.

(I) For an estimate of S_i^2, we have

$$E(s_i^2) = S_i^2$$

So $\hat{S}_i^2 = s_i^2$.

(II) For estimate of $\bar{Y}_i^2$, we know

$$
\begin{aligned}
Var(\bar{y}_i) &= E(\bar{y}_i^2) - [E(\bar{y}_i)]^2 \\
&= E(\bar{y}_i^2) - \bar{Y}_i^2 \\
\text{or } \bar{Y}_i^2 &= E(\bar{y}_i^2) - Var(\bar{y}_i).
\end{aligned}
$$

An unbiased estimate of $\bar{Y}_i^2$ is

$$
\begin{aligned}
\hat{\bar{Y}}_i^2 &= \bar{y}_i^2 - \widehat{Var}(\bar{y}_i) \\
&= \bar{y}_i^2 - \left(\frac{N_i-n_i}{N_i n_i}\right)s_i^2.
\end{aligned}
$$

(III) For the estimation of $\bar{Y}^2$, we know

$$
\begin{aligned}
Var(\bar{y}_{st}) &= E(\bar{y}_{st}^2) - [E(\bar{y}_{st})]^2 \\
&= E(\bar{y}_{st}^2) - \bar{Y}^2 \\
\Rightarrow \bar{Y}^2 &= E(\bar{y}_{st}^2) - Var(\bar{y}_{st})
\end{aligned}
$$

So, an estimate of $\overline{Y}^2$ is

$$\hat{\overline{Y}}^2 = \overline{y}_{st}^2 - \widehat{Var}(\overline{y}_{st})$$
$$= \overline{y}_{st}^2 - \sum_{i=1}^{k}\left(\frac{N_i - n_i}{N_i n_i}\right) w_i^2 s_i^2 .$$

Substituting these estimates in the expression $(n-1)S^2$ as follows, the estimate of S^2 is obtained as

$$(N-1)S^2 = \sum_{i=1}^{k}(N_i - 1)S_i^2 + N\left[\sum_{i=1}^{k} w_i \overline{Y}_i^{\,2} - \overline{Y}^2\right]$$

as
$$\hat{S}^2 = \frac{1}{N-1}\sum_{i=1}^{k}(N_i - 1)\hat{S}_i^2 + \frac{N}{N-1}\left[\sum_{i=1}^{k} w_i \hat{\overline{Y}}_i^{\,2} - \hat{\overline{Y}}^2\right]$$
$$= \frac{1}{N-1}\left[\sum_{i=1}^{k}(N_i - 1)s_i^2\right] + \frac{N}{N-1}\left[\left(\sum_{i=1}^{k} w_i\left(\overline{y}_i^2 - \left(\frac{N_i - n_i}{N_i n_i}\right)s_i^2\right)\right) - \left(\overline{y}_{st}^2 - \sum_{i=1}^{k}\frac{N_i - n_i}{N_i n_i} w_i^2 s_i^2\right)\right]$$
$$= \frac{1}{N-1}\left[\sum_{i=1}^{k}(N_i - 1)s_i^2\right] + \frac{N}{N-1}\left[\sum_{i=1}^{k} w_i(\overline{y}_i - \overline{y}_{st})^2 - \sum_{i=1}^{k} w_i(1-w_i)\frac{N_i - n_i}{N_i n_i}s_i^2\right].$$

Thus

$$\widehat{Var}_{SRS}(\overline{y}) = \frac{N-n}{Nn}\hat{S}^2$$
$$= \frac{N-n}{N(N-1)n}\left[\sum_{i=1}^{k}(N_i - 1)s_i^2\right] + \frac{N(N-n)}{nN(N-1)}\left[\sum_{i=1}^{k} w_i(\overline{y}_i - \overline{y}_{st})^2 - \sum_{i=1}^{k} w_i(1-w_i)\frac{N_i - n_i}{N_i n_i}s_i^2\right]$$

and

$$\widehat{Var}(\overline{y}_{st}) = \sum_{i=1}^{k}\frac{N_i - n_i}{N_i n_i} w_i^2 s_i^2 .$$

Substituting these expressions in

$$\frac{\widehat{Var}_{SRS}(\overline{y}) - \widehat{Var}(\overline{y}_{st})}{\widehat{Var}(\overline{y}_{st})},$$

the gain in efficiency due to stratification can be obtained.

If any other particular allocation is used, then substituting the appropriate n_i under that allocation, such gain can be estimated.

Interpenetrating subsampling

Suppose a sample consists of two or more subsamples which are drawn according to the same sampling scheme. The samples are such that each subsample yields an estimate of the parameter. Such subsamples are called interpenetrating subsamples.

The subsamples need not necessarily be independent. The assumption of independent subsamples helps in obtaining an unbiased estimate of the variance of the composite estimator. This is even helpful if the sample design is complicated and the expression for variance of the composite estimator is complex.

Let there be g independent interpenetrating subsamples and $t_1, t_2, ..., t_g$ be g unbiased estimators of parameter θ where $t_j (j = 1, 2, ..., g)$ is based on j^{th} interpenetrating subsample.

Then an unbiased estimator of θ is given by

$$\hat{\theta} = \frac{1}{g}\sum_{j=1}^{g} t_j = \bar{t}, \text{ say.}$$

Then

$$E(\hat{\theta}) = E(\bar{t}) = \theta$$

and

$$\widehat{Var}(\hat{\theta}) = \widehat{Var}(\bar{t}) = \frac{1}{g(g-1)}\sum_{j=1}^{g}(t_j - \bar{t})^2 .$$

Note that

$$E\left[\widehat{Var}(\bar{t})\right] = \frac{1}{g(g-1)} E\left[\sum_{j=1}^{g}(t_j - \theta)^2 - g(\bar{t} - \theta)^2\right]$$

$$= \frac{1}{g(g-1)}\left[\sum_{j=1}^{g} Var(t_j) - g\,Var(\bar{t})\right]$$

$$= \frac{1}{g(g-1)}(g^2 - g)Var(\bar{t}) = Var(\bar{t})$$

If the distribution of each estimator t_j is symmetric about θ, then the confidence interval of θ can be obtained by

$$P\left[Min(t_1, t_2, ..., t_g) < \theta < Max(t_1, t_2, ..., t_g)\right] = 1 - \left(\frac{1}{2}\right)^{g-1} .$$

Implementation of interpenetrating subsamples in stratified sampling

Consider the set up of stratified sampling. Suppose that each stratum provides an independent interpenetrating subsample. So based on each stratum, there are L independent interpenetrating subsamples drawn according to the same sampling scheme.

Let $\hat{Y}_{ij(tot)}$ be an unbiased estimator of the total of j^{th} stratum based on the i^{th} subsample ,

$i = 1,2,...,L$; $j = 1,2,...,k$.

An unbiased estimator of the j^{th} stratum total is given by

$$\hat{Y}_{j(tot)} = \frac{1}{L}\sum_{i=1}^{J}\hat{Y}_{ij(tot)}$$

and an unbiased estimator of the variance of $\hat{Y}_{j(tot)}$ is given by

$$\widehat{Var}(\hat{Y}_{j(tot)}) = \frac{1}{L(L-1)}\sum_{i=1}^{L}(\hat{Y}_{ij(tot)} - \hat{Y}_{j(tot)})^2 .$$

Thus an unbiased estimator of population total Y_{tot} is

$$\hat{Y}_{tot} = \sum_{j=1}^{k}\hat{Y}_{j(tot)} = \frac{1}{k}\sum_{i=1}^{L}\sum_{j=1}^{k}\hat{Y}_{ij(tot)}$$

And an unbiased estimator of its variance is given by

$$\begin{aligned}\widehat{Var}(\hat{Y}_{tot}) &= \sum_{j=1}^{k}\widehat{Var}(\hat{Y}_{j(tot)}) \\ &= \frac{1}{L(L-1)}\sum_{i=1}^{L}\sum_{j=1}^{k}(\hat{Y}_{ij(tot)} - \hat{Y}_{j(tot)})^2 .\end{aligned}$$

Post Stratifications

Sometimes the stratum to which a unit belongs may be known after the field survey only. For example, the age of persons, their educational qualifications etc. can not be known in advance. In such cases, we adopt the post-stratification procedure to increase the precision of the estimates.

Note**:** *This topic is to be read after the next module on ratio method of estimation. Since it is related to the stratification, so it is given here.*

In post-stratification,

- draw a sample by simple random sampling from the population and carry out the survey.
- After the completion of the survey, stratify the sampling units to increase the precision of the estimates.

Assume that the stratum size N_i is fairly accurately known. Let

m_i : number of sampling units from i^{th} stratum, $i = 1,2,...,k$.

$$\sum_{i=1}^{k} m_i = n.$$

Note that m_i is a random variable (and that is why we are not using the symbol n_i as earlier).

Assume n is large enough or the stratification is such that the probability that some $m_i = 0$ is negligibly small. In case, $m_i = 0$ for some strata, two or more strata can be combined to make the sample size non-zero before evaluating the final estimates.

A post stratified estimator of the population mean $\bar{Y}$ is

$$\bar{y}_{post} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i \ .$$

Now

$$\begin{aligned} E(\bar{y}_{post}) &= \frac{1}{N} E\left[\sum_{i=1}^{k} N_i E(\bar{y}_i \,|\, m_1, m_2, ..., m_k)\right] \\ &= \frac{1}{N} E\left[\sum_{i=1}^{k} N_i \bar{Y}_i\right] \\ &= \bar{Y} \end{aligned}$$

$$Var(\bar{y}_{post}) = E\left[Var(\bar{y}_{post}\,|m_1, m_2, ..., m_k)\right] + Var\left[E(\bar{y}_{post}\,|m_1, m_2, ..., m_k)\right]$$

$$= E\left[\sum_{i=1}^{k} w_i^2 \left(\frac{1}{m_i} - \frac{1}{N_i}\right) S_i^2\right] + Var(\bar{Y})$$

$$= \sum_{i=1}^{k} w_i^2 \left[E\left(\frac{1}{m_i}\right) - \left(\frac{1}{N_i}\right)\right] S_i^2 \quad (\text{Since } Var(\bar{Y}) = 0).$$

To find $E\left(\frac{1}{m_i}\right) - \frac{1}{N_i}$, proceed as follows:

Consider the estimate of ratio based on ratio method of estimation as

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{j=1}^{n} y_j}{\sum_{j=1}^{n} x_j}, \qquad R = \frac{\bar{Y}}{\bar{X}} = \frac{\sum_{j=1}^{N} Y_j}{\sum_{j=1}^{N} X_j}.$$

We know that

$$E(\hat{R}) - R = \frac{N-n}{Nn} \cdot \frac{RS_X^2 - S_{XY}}{\bar{X}^2}.$$

Let $x_j = \begin{cases} 1 \text{ if } j^{th} \text{ unit belongs to } i^{th} \text{ stratum} \\ 0 \text{ otherwise} \end{cases}$

and

$y_j = 1$ for all $j = 1,2,...,N$.

Then $R, \hat{R}$ and S_x^2 reduces to

$$\hat{R} = \frac{\sum_{j=1}^{n} y_j}{\sum_{j=1}^{n} x_j} = \frac{n}{n_i}$$

$$R = \frac{\sum_{j=1}^{N} Y_j}{\sum_{j=1}^{N} X_j} = \frac{N}{N_i}$$

$$S_x^2 = \frac{1}{N-1}\left[\sum_{j=1}^{N} X_j^2 - N\bar{X}^2\right] = \frac{1}{N-1}\left[N_i - N\frac{N_i^2}{N^2}\right] = \frac{1}{N-1}\left(N_i - \frac{N_i^2}{N}\right)$$

$$S_{xy} = \frac{1}{N-1}\left[\sum_{j=1}^{N} X_j Y_j - N\bar{X}\bar{Y}\right] = \frac{1}{N-1}\left[N_i - \frac{N_i N}{N^2}\right] = 0.$$

Using these values in $E(\hat{R})-R,$ we have

$$E(\hat{R})-R=E\left(\frac{n}{n_i}\right)-\frac{N}{N_i}=\frac{N(N-n)(N-N_i)}{nN_i^2(N-1)}.$$

Thus

$$E\left(\frac{1}{n_i}\right)-\frac{1}{N_i}=\frac{N}{nN_i}+\frac{N(N-n)(N-N_i)}{n^2N_i^2(N-1)}-\frac{1}{N_i}$$
$$=\frac{(N-n)N}{n(N-1)N_i}\left(1+\frac{N}{N_in}-\frac{1}{n}\right).$$

Replacing m_i in place of $n_i,$ we obtain

$$E\left(\frac{1}{m_i}\right)-\frac{1}{N_i}=\frac{(N-n)N}{n(N-1)N_i}\left(1+\frac{N}{N_in}-\frac{1}{n}\right)$$

Now substitute this in the expression of $Var(\bar{y}_{post})$ as

$$Var(\bar{y}_{post})=\sum_{i=1}^{k}w_i^2\left[E\left(\frac{1}{m_i}\right)-\frac{1}{N_i}\right]S_i^2$$
$$=\sum_{i=1}^{k}w_i^2S_i^2\left[\frac{N-n}{(N-1)n}.\frac{N}{N_i}\left(1+\frac{N}{nN_i}-\frac{1}{n}\right)\right]$$
$$=\frac{N-n}{n(N-1)}\sum_{i=1}^{k}w_i^2S_i^2\left[\frac{1}{w_i}\left(1+\frac{1}{nw_i}-\frac{1}{n}\right)\right]$$
$$=\frac{N-n}{n^2(N-1)}\sum_{i=1}^{k}w_i\,S_i^2\left[n-1+\frac{1}{w_i}\right]$$
$$=\frac{N-n}{n^2(N-1)}\sum_{i=1}^{k}(nw_i+1-w_i)S_i^2$$
$$=\frac{N-n}{n(N-1)}\sum_{i=1}^{k}w_iS_i^2+\frac{N-n}{n^2(N-1)}\sum_{i=1}^{k}(1-w_i)S_i^2.$$

Assuming $N-1\approx N.$

$$V(\bar{y}_{post})=\frac{N-n}{Nn}\sum_{i=1}^{n}w_iS_i^2+\frac{N-n}{n^2N}\sum_{i=1}^{n}(1-w_i)S_i^2$$
$$=V_{prop}(\bar{y}_{st})+\frac{N-n}{Nn^2}\sum_{i=1}^{n}(1-w_i)S_i^2.$$

The second term is the contribution to the variance of $\bar{y}_{post}$ due to $m_i's$ not being proportionately distributed.

If $S_i^2 \approx S_w^2$, say for all i, then the last term in the expression is

$$\frac{N-n}{Nn^2}\sum_{i=1}^{k}(1-w_i)S_w^2 = \frac{N-n}{Nn^2}S_w^2(k-1) \quad (\text{Since } \sum_{i=1}^{k} w_i = 1)$$
$$= \left(\frac{k-1}{n}\right)\left(\frac{N-n}{Nn}\right)S_w^2$$
$$= \frac{k-1}{n}Var(\bar{y}_{st}).$$

The increase in the variance over $Var_{prop}(\bar{y}_{st})$ is small if the average sample size $\bar{n} = \frac{n}{2}$ per stratum is reasonably large.

Thus a post-stratification with a large sample produces an estimator which is almost as precise as an estimator in the stratified sampling with proportional allocation.