3 Stratified Random Sampling

3.1 Introduction and Description

The objective of any sampling method is usually to estimate the unknown population parameters with the highest precision i.e. the variance of the estimators should be minimized. If the population is heterogeneous as will be in most situations then a sample taken via SRS might yield high levels of variability. As a result in a survey where precision is a main factor to be considered, then a strategy that addresses heterogeneity must be found. One way of achieving higher precision is to divide the population which is originally heterogeneous into sub population which are to a big extent homogeneous with respect to survey characteristics.

In stratified random sampling, the population of N units is first divided into sub-populations $N_1, N_2, ..., N_L$ called strata. The strata are mutually disjoint so that;

$$N_1 + N_2 +, ..., +N_L = \sum_{i=1}^{L} N_i = N \tag{3.1}$$

It is important that the number of units in the stratum denoted by $N_i, i = 1, 2, ..., L$ is known in order to maximize the gain from stratification. After determining the strata, a sample of size $n_i, i = 1, 2, ..., L$ is drawn from each stratum. If simple random sampling procedure is used to obtain the sub-samples in each stratum then the whole procedure is called under **stratified random sampling.**

The basic idea of stratification is that it may be possible to divide heterogeneous population into sub-populations which are internally homogeneous. If each sub-population is homogeneous, a precise estimate of any stratum can be obtained from a small sample of each stratum. This results in an improvement on the precision of the entire estimate.

Example 3.1. 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 sub-populations or strata such as,

Table 4: Average height of students

Students of class	1	2	3	Stratum 1
Students of class	4	5	6	Stratum 2
Students of class	7	8	9	Stratum 3
Students of class	10	11	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: The following is an extension of previous notation used where the suffix i denote the stratum and j denote the j^{th} unit within the stratum.

Let Y_{ij} be the value of the characteristic y on the j^{th} unit in the i^{th} stratum in the population; y_{ij} value in the sample; $j = 1, 2, ..., N_i$ (n_i in the sample), $i = 1, 2, ..., L$

Define:

$N_i =$ Total number of units in the i^{th} stratum

$n_i =$ the number of units in the sample of the i^{th} stratum.

Note: $j = 1, 2, ..., N \rightarrow units$ in a stratum; $i = 1, 2, ..., L \rightarrow strata$

$n = \sum_{i=1}^{L} n_i =$ total sample size from all the strata

$Y_i = \sum_{i=1}^{N_i} Y_{ij} =$ population total for the i^{th} stratum.

$y_i = \sum_{i=1}^{n_i} y_{ij} =$ sample total for the i^{th} stratum.

$\bar{y}_i = \frac{y_i}{n_i} =$ sample mean for the i^{th} stratum.

$\bar{Y} = \sum_{i=1}^{L} \frac{N_i \bar{Y}_i}{N} = \frac{Y}{N} =$ overall population mean.

$S_i^2 = \frac{1}{N_i - 1} \sum_{i=1}^{N_i} \left(Y_{ij} - \bar{Y}\right)^2 =$ population variance for the i^{th} stratum.

$s_i^2 = \frac{1}{n_i - 1} \sum_{i=1}^{n_i} \left(y_{ij} - \bar{y}\right)^2 =$ sample variance for the i^{th} stratum.

$W_i = \frac{N_i}{N} =$ population proportion for the i^{th} stratum or stratum weight and

$f_i = \frac{n_i}{N_i} =$ sampling fraction for the i^{th} stratum.

Note: The divisor of the variance is $(N_i - 1)$

3.1.1 Estimation of Population Mean, Variance and Total

The mean of the target population is given by;

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{L} \sum_{j=1}^{N_i} Y_{ij} = \frac{1}{N} \sum_{i=1}^{L} N_i \bar{Y}_i \tag{3.2}$$

where $N = N_1 + N_2 + ... + N_L$.

For the population mean per unit, the estimate used in stratified sampling is $\bar{y}_{st}$ (st for stratified), where $\bar{y}_{st} = \frac{1}{N} \sum_{i=1}^{L} N_i \bar{y}_i = \sum_{i=1}^{L} W_i \bar{y}_i$. $\left(W_i = \frac{N_i}{N}\right)$

Note: The estimate $\bar{y}_{st}$ is not in general the same as the sample mean. The sample mean $\bar{y}$ can be written as $\bar{y} = \frac{1}{n} \sum n_i \bar{y}_i$. The difference is that in $\bar{y}_{st}$ the estimates from the individual strata receive their correct weights $\frac{N_i}{N}$. It is evident that $\bar{y}$ coincides with $\bar{y}_{st}$ provided that in every stratum, $\frac{n_i}{n} = \frac{N_i}{N}$ or $\frac{n_i}{N_i} = \frac{n}{N} = f_i = f$. This means the sampling fraction is the same in all strata.

The principal properties of the estimate$\bar{y}_{st}$ are outlined in the following theorems. If simple random sample is used in each stratum then, $\bar{y}_{st}$ has the following properties.

Theorem 3.1. In stratified random sampling $\bar{y}_{st} = \sum_{i=1}^{L} \frac{N_i \bar{y}_i}{N} = \sum W_i \bar{y}_i$ is an unbiased estimator of the population mean $\bar{Y}$.

Proof. $E(\bar{y}_{st}) = \sum_{i=1}^{L} \frac{W_i E(\bar{y}_i)}{N} = \bar{Y}$ □

Theorem 3.2. In stratified random sampling using srswor in each stratum $Var(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 Var(\bar{y}_i) = \frac{1}{N^2} \sum_{i=1}^{L} \frac{N_i(N_i - n_i)}{ni} S_i^2$

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

Covariances terms vanish being independent from stratum to stratum

$\frac{1}{N^2} \sum_{i=1}^{L} \frac{N_i(N_i - n_i)}{n_i} S_i^2$ □

Corollary 3.2.1. If sampling fraction $\frac{n_i}{N_i}$ is negligibly small in each stratum, it reduces to

$Var(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{L} \frac{N_i^2 S_i^2}{n_i} = \sum_{i=1}^{L} \frac{W_i S_i^2}{n_i}$

Corollary 3.2.2. If $\hat{Y}_{st} = N\bar{y}_{st}$ is the estimate of the population total Y then $Var\left(\hat{Y}_{st}\right) = \sum_{i=1}^{L} N_i (N_i - n_i) \frac{S_i^2}{n_i}$

Proof. $\hat{Y}_{st} = N\bar{y}_{st}$
$\Rightarrow Var\left(\hat{Y}_{st}\right) = Var(N\bar{y}_{st})$
$= N^2 Var(\bar{y}_{st})$
$= N^2 \left(\frac{1}{N^2} \sum_{i=1}^{L} N_i (N_i - n_i) \frac{S_i^2}{n_i}\right)$

$$= \sum_{i=1}^{L} N_i (N_i - n_i) \frac{S_i^2}{n_i} \tag{3.3}$$

□

3.2 Estimation of Variance

In simple random sampling, the estimate of the variance of each stratum is given by $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ for the i^{th} stratum. We have found that $Var(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{L} N_i (N_i - n_i) \frac{S_i^2}{n_i}$. In stratified random sampling, the unbiased estimate of the variance of $Var(\bar{y}_{st})$ is given by $s_{st}^2 = \frac{1}{N^2} \sum_{i=1}^{L} N_i (N_i - n_i) \frac{s_i^2}{n_i}$. Note if $\bar{y}_{st}$ is normally distributed over $\bar{Y}$ then the confidence interval for $\bar{Y}$ is given by $\left[\bar{y}_{st} - z_{\frac{\alpha}{2}} S_{\bar{y}_{st}}, \bar{y}_{st} + z_{\frac{\alpha}{2}} S_{\bar{y}_{st}}\right]$. Therefore,

$$\bar{y}_{st} \pm z_{\frac{\alpha}{2}} \sqrt{Var(\bar{y}_{st})} \tag{3.4}$$

3.3 Allocation problem and choice of sample sizes is different strata

Question: How to choose the sample sizes $n_1, n_2, ..., n_l$ so that the available resources are used in an effective way? There are two aspects of choosing the sample sizes:

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

(b) 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:

3.3.1 Equal allocation

Choose the sample size n to be the same for all the strata. Draw samples of equal size from each strata. Let n be the sample size and k be the number of strata, then $n_i = \frac{n}{k}$ for $i = 1, 2, ..., L$

3.3.2 Proportional Allocations

If the sample sizes in the strata are closer such that $\frac{n_i}{n} = \frac{N_i}{N} = constant$, then the stratification is defined as stratification with proportion al allocation for n_i for $i = 1, 2, ..., L$. Consider, $\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i$, then with proportional allocation, $\bar{y}_{st} = \sum_{i=1}^{L} \frac{Nn_i}{Nn}\bar{y}_i = \frac{1}{n}\sum_{i=1}^{L} n_i \bar{y}_i$ overall mean. In this case, $\bar{y}_{st}$ coincides with $\bar{y}$ (the overall sample mean),

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

Now using proportional allocation the variance is;

$Var\left(\bar{y}_{st}\right)_{prop} = \frac{1}{N^2} N_i\left(N_i - \frac{nN_i}{N}\right)\frac{S_i^2}{\frac{nN_i}{N}}$

$= \frac{1}{N^2}\sum_{i=1}^{L} N_i\left(\frac{NN_i - nN_i}{N}\right)\frac{NS^2}{nN_i}$

$= \frac{1}{N^2}\sum_{i=1}^{L}\left(NN_i - nN_i\right)\frac{S_i^2}{n}$

$$\frac{N-n}{N^2 n}\sum_{i=1}^{L} N_i S_i^2 \tag{3.5}$$

which is the formula for the $Var\left(\bar{y}_{st}\right)$ under proportional allocation.

3.4 Allocation of Sample Sizes(Neymann Allocation)

This allocation considers the size of strata as well as variability; $n_i \propto N_i S_i$, $n_i = C^* N_i S_i$ where C* is the constant of proportionality.

$\sum_{i=1}^{L} n_i = \sum_{i=1}^{L} C^* N_i S_i$

or $n = C^* \sum_{i=1}^{L} N_i S_i$

or $C^* = \frac{n}{\sum_{i=1}^{L} N_i S_i}$, therefore

$n_i = \frac{n N_i S_i}{\sum_{i=1}^{L} N_i S_i}$.

This allocation arises when the $Var\left(\bar{y}_{st}\right)$is minimized subject to the constraint $\sum_{i=1}^{L} n_i$(pre-specified). There are some limitations of the optimum allocation. The knowledge of $S_i, i = 1, 2, ..., L$ 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 cost of survey and variability The cost of survey depends upon the nature of survey. A simple choice of the cost function is,

$$C = C_0 + \sum_{i=1}^{L} C_i n_i \tag{3.6}$$

where

C : total cost

C_0 :overhead cost, e.g., setting up of office, training people e.t.c

C_i cost per unit for the i^{th} stratum.

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

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

$\phi = Var\left(\bar{y}_{st}\right) + \lambda^2 \left(C - C_0\right)$

$= \sum_{i=1}^{L} w_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2 + \lambda^2 \left(\sum_{i=}^{L} C_i n_i\right)$

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

$= \sum_{i=1}^{L} \left[\frac{w_i S_i}{\sqrt{n_i}} - \lambda\sqrt{C_i n_i}\right]^2$ + terms independent of n_i.

Thus ϕ is minimum when; $\frac{w_i S_i}{\sqrt{n_i}} = \lambda\sqrt{C_i n_i}$ for all i or $n_i = \frac{1}{\lambda} \frac{w_i S_i}{\sqrt{C_i}}$.

How to determine λ

There are two ways to determine

(a) Minimize variability for fixed cost.

(b) Minimize cost for given variability.

We consider both the cases.

(i) Minimize variability for fixed cost

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

So,

$\sum_{i=i}^{L} C_i n_i = C_0^*$

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

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

Substituting in the expression for $n_i = \frac{1}{\lambda}\frac{w_i S_i}{\sqrt{C_i}}$ the optimum for n_i is obtained as $n_i^* = \frac{w_i S_i}{\sqrt{C_i}}\left(\frac{C_0^*}{\sum_{i=1}^{L}\sqrt{C_i} w_i S_i}\right)$. The required sample size to estimate $\bar{Y}$ such that the variance is minimum for given cost $C = C_0^*$ is $n = \sum_{i=1}^{L} n_i^*$.

(ii) Minimize cost for 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 $\sum_{i=1}^{L} \frac{w_i S_i^2}{n_i} = V_0 + \sum_{i=1}^{L} \frac{w_i^2 S_i}{N_i}$

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

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

(after substituting $n_i = \frac{1}{\lambda}\frac{w_i S_i}{\sqrt{C_i}}$). Thus the optimum n_i

is $\tilde{n}_i = \frac{w_i S_i}{\sqrt{C_i}}\left(\frac{\sum_{i=1}^{L} w_i S_i \sqrt{C_i}}{V_0 + \sum_{i=1}^{L} \frac{w_i^2 S_i^2}{N_i}}\right)$. So the required sample size to estimate $\bar{Y}$

such that C is minimum for a pre-specified V_0 is $n = \sum_{i=1}^{L} \tilde{n}_i$.

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

(i) If cost $C = C_0$ is fixed then $C_0 = \sum_{i=1}^{L} C_i n_i$, under the proportional allocation, $n_i = \frac{n}{N} N_i = nw_i$.

So,

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

or $n = \frac{C_o}{\sum_{i=1}^{L} w_i C_i}$,

therefore $n_i = \frac{C_0 w_i}{\sum w_i C_i}$.

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

(ii) If variance = V_0 is fixed then,

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

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

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

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

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

3.4.1 Variances under different allocations

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

(i) under Proportional allocation

Under proportional allocation; $n_i = \frac{n}{N} N_i$

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

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

$$= \frac{N-n}{Nn} \sum_{i=1}^{L} w_i S_i^2 \tag{3.7}$$

(ii) under Optimum allocation

Under optimum allocation;

$n_i = \frac{n N_i S_i}{\sum_{i=1}^{L} N_i S_i}$.

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

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

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

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

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

Example 3.2. A population of size 800 is divided into three strata. Their sizes and deviations are as given below.

Table 5: Population

Stata	1	2	3
Size of N_i	200	300	300
Standard deviation S_i	6	8	12

A sample of 120 is to be drawn from the population. Determine the sample size based on;

(a) Proportional allocation

(b) Optimum allocation

(c) Obtain the variance of the estimates of the population mean i.e. $Var_{prop}(\bar{y}_{st})$ and $Var_{opt}(\bar{y}_{st})$

Solution 4. $\frac{n_i}{n} = \frac{N_i}{N} \Rightarrow n_i = \frac{nN_i}{N}$, $n = 120 = \sum_{i=1}^{L} n_i$, $N = N_1 + N_2 + N_3 = 200 + 300 + 300 = 8000$

Therefore under proportional allocation, $n_1 = \frac{nN_1}{N} = \frac{120\times200}{800} = 30$,

$n_2 = \frac{nN_2}{N} = \frac{120\times300}{800} = 45$, $n_3 = \frac{nN_3}{N} = \frac{120\times300}{800} = 45$.

Under optimal allocation, $n_i = \frac{nN_iS_i}{\sum_{i=1}^{L} N_iS_i}$, $\sum_{i=1}^{L} N_iS_i = 200(6) + 300(8) + 300(12) = 72,000$,

$n_1 = \frac{nN_1S_1}{72,00} = \frac{120\times200\times6}{7200} = 20$, $n_2 = \frac{120\times300\times8}{7200} = 40$, $n_3 = \frac{120\times300\times12}{7200} = 60$.

$Var(\bar{y}_{st})_{prop} = \frac{N-n}{N^2n}\sum_{i=1}^{L} N_iS_i^2$, $\sum_{i=1}^{L} N_iS_i^2 = 200\left(6^2\right) + 300\left(8^2\right) + 300\left(12^2\right) = 69,600$,

$\Rightarrow \frac{800-120}{(800)^2(120)}(69,600) = \frac{680}{64,000(120)}(69,600) = 0.61625$,

$Var(\bar{y}_{st})_{opt} = \frac{1}{N^2}\left[\frac{1}{n}\left(\sum_{i=1}^{L} N_iS_i\right)^2 - \sum_{i=1}^{L} N_iS_i^2\right] = \frac{1}{800^2}\left[\frac{1}{120}(7200)^2 - 69,600\right] = 0.56626$.

Note: $Var(\bar{y}_{st})_{opt} < Var(\bar{y}_{st})_{prop}$.

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

(a) Proportional allocation:

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

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

In order to compare $V_{srs}(\bar{y})$

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

Consider, $S^2 = \frac{1}{N-1}\sum_{i=1}^{L}\sum_{j=1}^{N_i}\left(Y_{ij} - \bar{Y}\right)^2$

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

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

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

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

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

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

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

Therefore; $S^2 = \sum_{i=1}^{L}\frac{N_i}{N}S_i^2 + \sum_{i=1}^{L}\frac{N_i}{N}\left(\bar{Y}_i - \bar{Y}\right)^2$

or $\frac{N-n}{Nn}S^2 = \frac{N-n}{Nn}\sum_{i=1}^{L}\frac{N_i}{N}S_i^2 + \frac{N-n}{Nn}\sum_{i=1}^{L}\frac{N_i}{N}\left(\bar{Y}_i - \bar{Y}\right)^2$

(pre-multiply by $\frac{N-n}{Nn}$ on both sides)

$Var_{srs}\left(\bar{Y}\right) = V_{prop}\left(\bar{y}s_{st}\right) + \frac{N-n}{Nn}\sum_{i=1}^{L} w_i\left(\bar{Y}_i - \bar{Y}\right)^2.$

Since $\sum_{i=1}^{L}\left(\bar{Y}_i - \bar{Y}\right)^2 \geq 0 \Rightarrow Var_{prop}\left(\bar{y}_{st}\right) \leq Var_{srs}\left(\bar{y}\right)$

Larger gain in the difference is achieved when $\bar{Y}_i$

differs from $\bar{Y}$ more.

(b) Optimum allocation

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

Consider

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

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

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

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

where $\bar{S} = \sum_{i=1}^{L} w_iS_i \Rightarrow Var_{prop}\left(\bar{y}_{st}\right) - Var_{opt}\left(\bar{y}_{st}\right) \geq 0$

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

Larger gain in efficiency is achieved when S_i differ from $\bar{S}$

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

$$Var_{opt}\left(\bar{y}_{st}\right) \leq Var_{prop}\left(\bar{y}_{st}\right) \leq Var_{srs}\left(\bar{y}\right) \tag{3.8}$$

3.5 Estimate of variance and confidence intervals

Under SRSWOR, an unbiased estimate of S_i^2 for the i^{th} stratum $(i = 1, 2, ..., L)$ is $s_i^2\frac{1}{n_i - 1}\sum_{i=1}^{L}\left(y_{ij} - \bar{y}_i\right)$

In stratified sampling, $Var\left(\bar{y}_{st}\right) = \sum_{i=1}^{L} w_i^2\frac{N_i - n_i}{N_in_i}S_i^2$. So an unbiased estimate of $Var\left(\bar{y}_{st}\right)$ is $\widehat{Var}\left(\bar{y}_{st}\right) = \sum_{i=1}^{L} w_i^2\frac{N_i - n_i}{N_in_i}s_i^2$

or $\sum_{i=1}^{L}\frac{w_i^2s_i^2}{n_i} - \sum_{i=}^{L}\frac{w_i^2s_i^2}{N_i}$

or $\sum_{i=1}^{L}\frac{w_i^2s_i^2}{n_i} - \frac{1}{N}\sum_{i=1}^{L} w_is_i^2$

The second term in this expression represents the reduction due to finite population correction. The confidence limits of $\bar{Y}$ can be obtained as

$$\bar{y}_{st} \pm t\sqrt{\widehat{Var}\left(\bar{y}_{st}\right)} \tag{3.9}$$

assuming $\bar{y}_{st}$ is normally distributed and $\widehat{Var}(\bar{y}_{st})$ 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 t-distribution. The distribution of $\widehat{Var}(\bar{y}_{st})$ 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 $n_e = \frac{\left(\sum_{i=1}^{L} g_i s_i^2\right)^2}{\sum_{i=1}^{L} \frac{g_i^2 s_i^4}{n_i - 1}}$ where $g_i = \frac{N_i(N_i - n_i)}{n_i}$ and $min(n_i - 1) \leq n_e \leq \sum_{i=1}^{L}(n_i - 1)$ assuming y_{ij} are normally distributed.

Example 3.3. (a) A market researcher is allocated Ksh. 20,000 to conduct a survey by means of stratified random sampling. The population consists of stratum A of size 40,000, B of size 20,000 and C of size 10,000. The set cost of administering the survey is 200 and the cost of sampling one unit are 2.25, 4.00 and 1.00 for stratum A, B and C respectively. The deviations of observations in stratum A is thought to be twice that of stratum B and C. Find the optimum and proportional allocations, assuming that all the money is to be spent on the survey.

Solution $N_1 = 40,000$, $N_2 = 20,000$, $N_3 = 10,000$, $c_0 = 200$ (fixed cost), $c = 20,000$, $c_1 = 2.25$, $c_1 = 4.0$,$c_1 = 1.0$, $A = 2S_3, B = S_3, C = S_3$.

For optimum allocation, we need to find the size of the sample in each of the stratum i.e.
$n_i = \frac{\frac{(c-c_0)N_i S_i}{\sqrt{c_i}}}{\sum_{i=1}^{3} N_i S_i \sqrt{c_i}}$.

Now, $\sum_{i=1}^{3} N_i S_i \sqrt{c_i} = 40,000\,(2S_3)\left(\sqrt{2.25}\right) + 20,000\,(S_3)\left(\sqrt{4}\right) + 10,000\,(S_3)\sqrt{1} = 170,000S_3$.

$n_1 = \frac{\frac{(20,000-200)(40,000)2S_3}{1.5}}{170,000S_3} \simeq 6211.$

$n_2 = \frac{\frac{(20,000-200)(20,000)S_3}{2}}{170,000S_3} \simeq 1164.7.$

$n_3 = \frac{\frac{(20,000-200)(10,000)S_3}{2}}{170,000S_3} \simeq 1164.7.$

Under proportional allocation,

$\frac{n_i}{n} = \frac{N_i}{N} \Rightarrow n_i = \frac{nN_i}{N}$,

$c = c_0 + \sum_{i=1}^{L} c_i n_i = c_0 + \sum_{i=1}^{L} ci\frac{nN_i}{N}$,

$c_0 + \frac{n}{N}\sum_{i=1}^{L} C_i N_i \Rightarrow c - c_0 = \frac{n}{N}\sum_{i=1}^{L} C_i N_i \Rightarrow n = \frac{N(c-c_0)}{\sum_{i=1}^{L} C_i N_i}$, $N = N_1 + N_2 + N_3 = 70,000.$

$\Rightarrow n = \frac{70,000(20,000-2,000)}{(2.25)(40,000)+4(20,000)+1(10,000)} = 77,000.$

Therefore,

$n_i = \frac{nN_i}{N} \Rightarrow n_1 = \frac{7700(40,000)}{70,000} = 4,400,$

$n_2 = \frac{7700(20,000)}{70,000} = 2200,$

$n_3 = \frac{7700(10,000)}{70,000} = 1,100.$

3.6 Exercises

(a) Given a population $U = 1, 2, 3, 4$ and $y_1 = y_2 = 0, y_3 = 1, y_4 = -1$, the values taken by the characteristic y.

i. Calculate the variance of the mean estimator for a simple random design without replacement of size $n = 2$.

ii. Calculate the variance of the mean estimator for a stratified random design for which only one unit is selected per stratum and the strata are given by $U_1 = 1, 2$ and$U_2 = 3, 4$.

(b) A sample of 30 students is to be drawn from a population of 300 students belonging to two colleges A and B. The means and deviations of their marks are given below. Use the information to confirm that Neyman allocation scheme is a more efficient scheme when compared to proportional allocation.

Table 6: A sample of 30 students

	Number of students	Mean	SD
College A	200	30	10
College B	100	60	40

(c) A stratified population has 5 strata. The stratum sizes N_i and means $\bar{Y}_i$ and S_i^2 of some variable Y are as follows.

Table 7: Stratified population

$Stratum$	N_i	$\bar{Y}_i$	S_i^2
1	117	7.3	1.31
2	98	6.9	2.03
3	74	11.2	1.13
4	41	9.1	1.96
5	45	9.6	1.74

i. Calculate the overall population mean and variance.

ii. For a stratified simple random sample of size 80, determine the appropriate stratum sample sizes under proportional allocation and Neyman allocation.

(d) Among the 7500 employees of a company, we wish to know the proportion P of them that owns at least one vehicle. For each individual in the sampling frame, we have the value of his income. We then decide to construct three strata in the population: individuals with low income (stratum 1), with medium income (stratum 2), and with high income (stratum 3). We denote:

N_h = the stratum size h,

n_h = the sample size in stratum h (simple random sampling),

p_h = the estimator of the proportion of individuals in stratum h owning at least one vehicle.

The results are given in Table 8

Table 8: Employees according to income

	h=1	h=2	h=3
Nh	3500	2000	2000
nh	500	300	200
ph	0.13	0.45	0.5

i. What estimator $\hat{P}$ of P do you propose? What can we say about its bias?

ii. Calculate the accuracy of $\hat{P}$, and give a 95% confidence interval for P.

iii. Do you consider the stratification criteria to be adequate?