STRATIFIED SAMPLING

Definition:

The type of sampling in which the whole heterogeneous population is divided into smaller groups, known as strata, such that these sampling units are homogeneous with respect to the characteristic under study within the group and heterogeneous with respect to the characteristic under study between the groups , is known as **Stratified Sampling.** Each group is treated as separate population and sampling units are selected from these groups by using **Simple Random Sampling.**

For example: if we want to find the average weight of the students of a school from class I to class X, we know that the weight varies as the student of class I are of age 5/6 years and the students of class X are of age around 15/16 years.

So one can divide all the students into different strata such as:

Students of class 1,2: Stratum 1

Students of class 3,4: Stratum 2

Students of class 5,6: Stratum 3

Students of class 7,8: Stratum 4

Students of class 9,10: Stratum 5

Then the samples are drawn by SRSWOR from each of the strata 1, 2, 3, 4 and 5. The drawn samples are combined together and that sample is termed as the Stratified Sample.

REASONS FOR STRATIFICATION:

(1) To obtain estimates of known precision for certain subdivisions of the population by treating each subdivision as a stratum.
(2) For administrative convenience; for example: stratification can provide survey organization to control the distribution of fieldwork among its regional offices.
(3) Sometimes different parts of the population may call for different sampling procedures. With human populations, people living in institutions(e.g., hotels, hospitals) are often placed in a different stratum from people living in ordinary homes.
(4) Stratification may often produce a gain in precision of the estimates of characteristics of the whole population. The amount in the gain depend on the type of stratification.

ADVANTAGES OF STRATIFIED SAMPLING:

(1) More representatives: in an un-stratified random sampling, units from some groups may be over represented; units from some groups may not be considered at all. Whereas using stratified random sampling, units from all strata are considered.
(2) Greater accuracy: Stratified sampling provides estimates with increased precision.

(3) Administrative convenience: As compared to SRS, information from units within the strata can be collected more quickly, reducing time and money involved.

STEPS OF SELECTION OF STRATIFIED SAMPLE:

Notations:

N : Population size

N : sample size

K : number of strata

N_i : Number of sampling units in i^{th} strata $N = \Sigma N_i$

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

$n = \Sigma n_i$

The steps for selecting stratified sample are :

(1) Divide the population of N units into k strata. Let the i^{th}stratum contains N_i number of units 1, 2, …….,k.

(2) Strata are constructed such that they are non-overlapping and homogeneous with respect to the characteristic under study such that $N = \Sigma N_i$.

(3) Draw a sample of size n_i from the i^{th} stratum by using SRSWOR independently from each stratum.

(4) All the sampling units are drawn from each stratum will constitute a stratified sample of size $n = \Sigma n_i$

Estimation of population mean and its variance:

Let Y : characteristic under study

Let Y_{ij} be the population value of jth unit in ith stratum, $j = 1, 2, \ldots N_i$, $i = 1, 2, \ldots\ldots k$

Population mean of ith stratum = $\bar{Y}_i = \Sigma\, Y_{ij}/N_i$, $j = 1, 2, \ldots, N_i$

Population mean square error for ith stratum = S_i^2

Where $S_i^2 = \Sigma\, (Y_{ij} - \bar{Y}_i)^2/N_i - 1$

Population total = $Y = \Sigma\Sigma\, Y_{ij} = \Sigma\, N_i\bar{Y}_{i.}$

Population mean = $\bar{Y} = \Sigma\Sigma\, Y_{ij}/N = \Sigma\, N_i\bar{Y}_i/N = \Sigma\, W_i\bar{Y}_i$

Where $W_i = N_i/N$

Let y_{ij} be the sample value of jth unit in ith stratum ,

Where $j = 1, 2, \ldots\ldots n_i$, $i = 1, 2, \ldots., k$

Sample mean from ith stratum $= \bar{y}_i = \Sigma y_{ij}/n_i$

Sample mean square error from ith stratum = s_i^2

Where $s_i^2 = \Sigma(y_{ij} - \bar{y}_i)^2/n_i - 1$

Sample mean = $\bar{y} = \Sigma\, n_i\, \bar{y}_i/n = \Sigma\, w_i\, \bar{y}_i$

Where $w_i = n_i/n$

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

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

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

$$= \Sigma\, w_i \bar{Y}_i$$

$$= \Sigma\, (n_i/n)\, \bar{Y}_i \neq \bar{Y}$$

Which shows that sample mean $\bar{y}$ is a **biased estimator** of population mean $\bar{Y}$. Based on this, one can consider an unbiased estimator of $\bar{Y}$.

Let the stratum mean which is defined as the weighted arithmetic mean of strata sample means with strata sizes as weights given by

$\bar{y}_{st} = (1/N)\Sigma N_i\bar{y}_l = \Sigma W_i\bar{y}_i$, where $W_i = \Sigma N_i/N$

Let $E(\bar{y}_{st}) = E(\Sigma W_i\bar{y}_i) = \Sigma W_iE(\bar{y}_i)$

$= \Sigma W_i\bar{Y}_i = \bar{Y}$

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

Variance of an unbiased estimator of $\bar{Y}$:

$V(\bar{y}_{st}) = V(\Sigma W_i\bar{y}_i)$

$= \Sigma W_i^2 V(\bar{y}_i)$

Since samples are drawn independently from each of the strata by SRSWOR,

$V(\bar{y}_{st}) = \Sigma W_i^2(N_i - n_i/N_i)(S_i^2/n_i)$………………(i)

$= \Sigma W_i^2(1 - n_i/N_i)(S_i^2/n_i)$

$= \Sigma W_i^2(1 - f_i)(S_i^2/n_i)$………………..(ii) where $f_i = n_i/N_i$

$= \Sigma W_i^2(S_i^2/n_i)$……………………..(iii) when f_i are small.

We observe that variance of $\bar{y}_{st}$ depends on the values of S_i^2 , variance of $\bar{y}_{st}$ is small when values of S_i^2 are small. And values of S_i^2 are small when strata are homogeneous.

Estimate of Variance:

Since the sample have been drawn by SRSWOR, so

$E(s^2) = S^2$

We have,

$(N\bar{y}_{st}) = \Sigma W_i^2(1 - n_i/N_i)(S_i^2/n_i)$

So an estimate of variance is

$$V(\hat{\bar{y}}_{st}) = \Sigma W_i^2(1 - n_i/N_i)(\hat{S}_i^2/n_i)$$

$$= \Sigma W_i^2(1 - n_i/N_i)(s_i^2/n_i)$$

$$= \Sigma W_i^2\{(N_i - n_i)/N_i\}(s_i^2/n_i)$$

Estimation of population total and its variance:

The population total is Y and population mean is $\bar{Y}$.

So , $Y = N\bar{Y}$ and $E(\bar{y}_{st}) = \bar{Y}$

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

$$= Y$$

So an estimate of population total is $N\bar{y}_{st}$

Variance of an estimate of population total is

$$V(\hat{y}) = V(N\bar{y}_{st}) = N^2\,(N\bar{y}_{st})$$

$$= N^2\,\Sigma W_i^2\{(N_i - n_i)/N_i\}(S_i^2/n_i)$$

$$=N^2*1/\,N^2\,\Sigma N_i^2\{(N_i - n_i)/N_i\}(S_i^2/n_i)$$

$$= \Sigma N_i^2\{(N_i - n_i)/N_i\}(S_i^2/n_i)$$

Estimate of variance of an estimate of population total:

$$\hat{V}(\hat{Y}) = \Sigma N_i^2\{(N_i - n_i)/N_i\}(\hat{S}_i^2/n_i)$$

$$= \Sigma N_i^2\{(N_i - n_i)/N_i\}(s_i^2/n_i)$$

Allocation of sample size:

We know,

$V(\bar{y}_{st}) = \Sigma\ W_i^2(N_i - n_i/N_i)(S_i^2/n_i)$

We observe that the variance depends on n_i sample size from ith stratum. The allocation of sample size to various strata is done in two ways :

- **(A) Proportional Allocation**
- **(B) Optimum Allocation**

(A) Proportional Allocation:

Allocation of n_i sample size to ith stratum is called **Proportional Allocation.** This is a sample fraction which is constant for each stratum.

$$n_1/N_1 = n_2/N_2 = \ldots\ldots = n_i/N_i = \ldots\ldots = n_k/N_k$$

$$= \Sigma\ n_i/N_i = n/N = C$$

Where C is a constant.

Now, $n_i/N_i = C$

⇨ $n_i \alpha N_i$, $i=1, 2, \ldots\ldots., k$

⇨ $n_i=(n/N)N_i$

Here in sample each stratum is represented according to its size.

The formula $n_i/N_i=n/N$ is known as Bowley's formula for proportional allocation. Here variance of $\bar{y}_{st}$ is given by

$$V(\bar{y}_{st})= \Sigma\ W_i^2(N_i - n_i/N_i)(S_i^2/n_i)$$

$$= \Sigma N_i^2/N^2(1 - n_i/N_i)(S_i^2/n_i)$$

$$=1/N^2\ \Sigma N_i/(1 - n_i/N_i)(S_i^2/n_i/N_i)$$

$V(\bar{y}_{st})_{prop} = 1/N^2\ \Sigma N_i/(1 - n_i/N_i)(S_i^2/n_i/N_i)$ where $n_i/N_i = n/N$

Multiplying by n/N, we get,

$$V(\bar{y}_{st})_{prop} = 1/N^2\ \Sigma\ N_i(1/n/N - n/N*n/N)S_i^2$$
$$=1/N^2\ \Sigma\ N_i\ (N/n)(1 - n/N)S_i^2$$
$$=1/N\ \Sigma\ N_i(1/n - 1/N)S_i^2$$
$$=(1/n - 1/N)\ \Sigma\ (N_i/N)\ S_i^2$$
$$=(1/n - 1/N)\ \Sigma\ W_i\ S_i^2$$

Hence, $V(\bar{y}_{st})_{prop} = (1/n - 1/N)\ \Sigma\ W_i\ S_i^2$

(B) Optimum Allocation:

Here we use two methods:

(a) Determine the value of n_i's which minimizes the variance of $\bar{y}_{st}$ for (i) fixed sample size (n)
(ii) for fixed cost.

(b) Determine the values of n_i's which minimizes the total cost for fixed precision.

The allocation of n_i's to various strata in accordance with the above objective is known as **Optimum Allocation**.

Thus, in optimum allocation the values of n_i's are obtained such that

(i)Variance of $\bar{y}_{st}$ is minimum for fixed value of n (Neyman's allocation)

(ii) Variance of $\bar{y}_{st}$ is minimum for fixed value of total cost C.

(iii)Total cost C is minimum for fixed value of Variance of $\bar{y}_{st}$.

Cost function:

In any sample survey the cost associated with must be balanced with the value of the information. For stratified random sample, the cost function can be written as

$C = a + \Sigma c_i n_i$

Where a= overhead cost, e.g.: setting up of office, training people etc.

And c_i = cost per unit in the ith stratum

Theorem 1:

$Var(\bar{y}_{st})$ is minimum for fixed value of sample size n if $n_i \alpha N_i$.

We have $n_i = nN_iS_i/\Sigma N_iS_i$

Neyman's formula for optimum allocation is given by $n_i \alpha N_i$

This allocation arises when the $Var(\bar{y}_{st})$ is minimised subject to the constraint $\Sigma n_i = n$ (where sample size n is pre-specified). By substituting value of n_i in expression for $V(\bar{y}_{st})$ we get

$$V(\bar{y}_{st}) = 1/N^2 \Sigma\, N_i(N_i/n_i - 1)S_i^2$$

$$= 1/N^2 \Sigma\, N_i(\Sigma N_iS_i/nS_i - 1)\, S_i^2$$

$$V(\bar{y}_{st}) = 1/n(\Sigma W_iS_i)^2 - 1/N\, \Sigma W_iS_i^2$$

Theorem 2: $V(\bar{y}_{st})$ is minimum for a given cost function of the form $C = a + \Sigma c_i n_i$ if $n_i \alpha N_iS_i/\sqrt{C_i}$

We have $n_i = N_iS_i/N\sqrt{\lambda}\sqrt{C_i}$

$\emptyset = V(\bar{y}_{st}) + \lambda\, [\, \Sigma c_i n_i - C + a\,]$

$= (1/N^2)\Sigma\, N_i(N_i/n_i - 1)\, S_i^2 + \lambda\, [\, \Sigma c_i n_i - C + a\,]$

$\emptyset$ is minimum when $n_i = N_iS_i/N\sqrt{\lambda}\sqrt{C_i}$

Where $\sqrt{\lambda} = (\Sigma N_iS_i/\sqrt{C_i})/Nn$

Putting the value of $\sqrt{\lambda}$ in expression for n_i,

We get $n_i = N_iS_i/\sqrt{C_i}$

Thus $n_i \propto N_iS_i/\sqrt{C_i}$

This indicates that a larger sample would be required from each stratum if

(i)Stratum size N_i is large

(ii)Stratum variability S_i is large

(iii)Sampling cost per unit is low in the stratum .

Hence, the value of n_i depends on the value of n-sample size.

Expression for variance:

$V(\bar{y}_{st}) = 1/N^2 \Sigma N_i(N_i/n_i - 1)S_i^2$

Since $n_i = (nN_iS_i/\sqrt{C_i})/\Sigma(N_iS_i/\sqrt{C_i})$

- ⇨ $N_i/n_i = \Sigma(N_iS_i/\sqrt{C_i})/ (N_iS_i/\sqrt{C_i})$
- ⇨ $V(\bar{y}_{st}) = 1/N^2\Sigma N_i[\Sigma(N_iS_i/\sqrt{C_i})/ (nS_i/\sqrt{C_i}) - 1]S_i^2$
- ⇨ $V(\bar{y}_{st}) = 1/n\Sigma[W_iS_i/\sqrt{C_i}]\ \Sigma[W_iS_i/\sqrt{C_i}] - 1/N\ \Sigma W_iS_i^2$

The variance depends on the value of n and the value of n depends on whether the sample is selected to meet the specified cost or a given precision.

Case I: Fixed cost is given:

Let specified cost is C_0.

$V(\bar{y}_{st}) = 1/(C - a)[\ \Sigma[W_iS_i/\sqrt{C_i}]^2 - \Sigma W_iS_i^2$

(i) Particular case 1: C_i=c(constant)for i=1, 2,……..,k

Then $c - a = \Sigma c_in_i = c\Sigma n_i = cn$

⇨ $n=(C-a)/c$

and $n_i=(nN_iS_i/\sqrt{C_i})/\Sigma(N_iS_i/\sqrt{C_i})= nN_iS_i/\Sigma N_iS_i$

$= nW_iS_i/ \Sigma W_iS_i$

Which is the formula for Neyman's Allocation.

Comparison of Simple Random Sampling, Stratified Random Sampling using Proportional Allocation and Neyman's Allocation:

Comparison of SRS, Stratified Random Sampling – Proportional Allocation:

Consider a sample of size n is selected from a population of size N. If the sampling technique used is SRSWOR, then the sample mean is an unbiased estimator of population mean and its variance is given by $V(\bar{y})_{SRSWOR}= (1/n - 1/N)S^2$

Where $S^2 = (1/N-1)\ \Sigma\ \Sigma(Y_{ij}-\bar{Y})^2$

$V(\bar{y}_{st})_{prop} = (1/n - 1/N)\Sigma W_iS_i^2$

Where $W_i= N_i/N$ and

$S_i^2 = (1/N_i -1)\ \Sigma(Y_{ij} -\bar{Y})^2$ where j=1, 2, …..N_i

Let $S^2 = (1/N-1)\ \Sigma\ \Sigma(Y_{ij}-\bar{Y})^2$

⇨ $(N-1)\ S^2= \Sigma\ \Sigma(Y_{ij}-\bar{Y})^2$

⇨ $(N-1)\ S^2= \Sigma(N_i-1)S_i^2 +\Sigma N_i(Y_i- \bar{Y})^2$

If N_i is large so N is also large , then

$N \approx N-1$ and $N_i=N_i-1$

$S^2= \Sigma\ W_iS_i^2 + \Sigma\ W_i(\bar{Y}_i-\bar{Y})^2$

Hence, $V(\bar{y})_{SRSWOR}=(1/n - 1/N)S^2$

$=(1/n - 1/N)[\Sigma W_iS_i^2 - \Sigma\ W_i(\bar{Y}_i-\bar{Y})^2$

$=V(\bar{y}_{st})_{prop} + (1/n - 1/N)\ \Sigma\ W_i(\bar{Y}_i - \bar{Y})^2$

⇨ $V(\bar{y})_{SRSWOR} \geq V(\bar{y}_{st})_{prop}$

Comparison between Stratified Random Sampling- Proportional Allocation and Neyman Allocation:

We know,

$V(\bar{y}_{st})_{prop} = (1/n - 1/N)\Sigma W_i S_i^2$

And $V(\bar{y}_{st})_{Ney} = 1/n(\Sigma W_i S_i)^2 - 1/N\ \Sigma W_i S_i^2$

$V(\bar{y}_{st})_{prop} - V(\bar{y}_{st})_{Ney} \geq 0$

⇨ $V(\bar{y}_{st})_{prop} \geq V(\bar{y}_{st})_{Ney}$

Hence we get,

⇨ $V(\bar{y})_{SRSWOR} \geq V(\bar{y}_{st})_{prop} \geq V(\bar{y}_{st})_{Ney}$