However, if N is large, then we get

$$n = \left\lceil \frac{S\,Z_{\alpha/2}}{\varepsilon\,\bar{Y}_N} \right\rceil = \left[\text{C.D} > \text{of } Y \text{ in the population} \times \frac{Z_{\alpha/2}}{\varepsilon} \right]^2 \quad \ldots(7\cdot39\,a)$$

The determination of sample size from (7·39) or (7·39a) pre-supposes the knowledge of $(S/\bar{Y}_N)$, the coefficient of dispersion in the population. The C.D. is generally stable over time and provides a reliable estimate with the help of the past data but it is not true for S^2. If S^2 is not known, then it can be estimated by taking a preliminary sample of size m from the given population.

If s^2 in the estimated value of S^2 based on a sample of m units and d is the permissible margin of error, the additional number of units required to estimate $\bar{Y}_N$ with desired precision, assuming N large, is $(n - m)$, where n is given by (7·35) or (7·38 a).

7·10 STRATIFIED RANDOM SAMPLING

In *srswor* we obtained

$$\text{Var}(\bar{y}_n) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

This implies that the variance of the sample estimate of the population mean is :

(*i*) inversely proportional to the sample size, and

(*ii*) directly proportional to the variability of the sampling units in the population.

Since the *precision* of an estimate is defined as the '*reciprocal of its sampling variance*', we see that apart from increasing the sample size n or sampling fraction n/N, the only other way of increasing the precision of $\bar{y}_n$ is to devise a sampling technique which will effectively reduce S^2, the population heterogeneity. One such technique is *Stratified Sampling*.

Stratification means division into layers. *Auxiliary information* (past data or some other information) related to the character under study may be used to divide the population into various groups such that :

(*i*) units *within* each group are as homogeneous as possible, and

(*ii*) the group means are as widely different as possible.

Thus a population consisting of N sampling units is divided into k relatively homogeneous mutually disjoint (non-overlapping) sub-groups, termed as *strata*, of sizes N_1, $N_2, \ldots, N_k$, such that $N = \sum_{i=1}^{k} N_i$. If a simple random sample (generally without replacement) of size n_i, $(i = 1, 2, \ldots, k)$ is drawn from each of the stratum respectively such that $n = \sum_{i=1}^{k} n_i$, the sample is termed as *Stratified Random Sample* of size n and the technique of drawing such a sample is called *Stratified Random Sampling*.

Remarks **1. Main Problems in Stratification.** From the above description of stratified random sampling, the following points need consideration :

(*i*) Principle of stratification, *i.e.*, proper classification of population into various strata.

(*ii*) Given the total sample size n, how to allocate it amongst different strata.

(*iii*) Decision about the value of k, the number of strata.

(*iv*) Having decided about the value of k, how to determine the boundaries of the strata, *i.e.*, *the demarcation of the strata* ?

Scanned with CamScanner

2. The criterion which enables us to classify various sampling units into different strata is termed as '*stratifying factor*' (*s. f.*). Some of the commonly used stratifying factors are age, sex, educational or income level, geographical area, economic status, and so on. An *s.f.* is called effective if it divides the given population into different strata which are homogeneous (or nearly so) within themselves and the units in different strata are as unlike as possible. Such an orgnisation gives estimates with greater precision since a portion of the variability, identifiable as *between strata variance,* is eliminated and the only variance which enters the computation is the *within stratum variance.*

3. If the study relates to a single character, it may be easy to choose a variate *w.r.t.* which the units of the population can be grouped to give homogeneous strata. But if one is dealing with multi-character study one faces the difficulty of choosing an appropriate way of startification. In such a situation, many times, intuition, judgement of subject -matter specialists can all be used effectively in setting up strata. If judgement is exercised in determining the strata, the sample result will be still unbiased provided the sampling within each startum is carried out by a random process. However, if the judgement is good, the sampling variance may be reduced. *In many fields of highly skewed distributions, stratification is an exceedingly valuable tool.*

7·10·1. Principal Advantages of Stratified Random Sampling.

1. *More Representative.* In an unstratified random sample some strata may be over-represented, others may be under-represented while some may be excluded altogether. Stratified sampling ensures any desired representation in the sample of the various strata in the population. It overrules the possibility of any essential group of the population being completely excluded in the sample. Stratified sampling thus provides a more representative cross section of the population and is frequently regarded as the most efficient system of sampling.

2. *Greater Accuracy.* Stratified sampling provides estimates with increased precision. Moreover, stratified sampling enables us to obtain the results of known precision for each of the stratum.

3. *Administrative Convenience.* As compared with simple random sample, the stratified samples would be more concentrated geographically. Accordingly, the time and money involved in collecting the data and interviewing the individuals may be considerably reduced and the supervision of the field work could be allotted with greater ease and convenience.

4. Sometimes the sampling problems may differ markedly in different parts of the population, *e.g.,* a population under study consisting of (*i*) literates and illiterates, or (*ii*) people living in institutions (hostels, prisons, hospitals, etc.) and those living in ordinary homes, or (*iii*) people living in hill areas and plain areas. In such cases, we can deal with the problem through stratified sampling by regarding the different parts of the population as stratum and tackling the problems of the survey within each stratum independently.

7·10·2. Notations and Terminology.

Let k be the number of strata.

N_i = The number of sampling units in the ith stratum ($i = 1, 2, \ldots, k$)

$N = \sum_{i=1}^{k} N_i$, total number of sampling units in the population.

n_i = The number of sampling units selected with *srswor* from the ith stratum.

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

Scanned with CamScanner

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

$\bar{Y}_{Ni}$ = population mean of ith stratum $= \frac{1}{N_i} \sum_{i=1}^{k} Y_{ij}$

$\bar{Y}_N$ = Population mean $= \frac{1}{N} \sum_i \sum_j Y_{ij} = \frac{1}{N} \sum_i^k N_i \bar{Y}_{N_i} = \sum_{i=1}^{k} p_i \bar{Y}_{N_i}$

where $p_i = N_i / N$ is called the *weight* of the ith stratum.

S_i^2 = Population mean square of the ith stratum $= \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_{N_i})^2$,

$(i = 1, 2, \ldots, k)$ ... (7·41)

y_{ij} = Value of jth sampled unit from ith stratum.

$\bar{y}_{n_i}$ = Mean of sample selected from ith stratum.

s_i^2 = Sample mean square of the ith stratum

$= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{y}_{n_i})^2 ; (i = 1, 2, \ldots, k)$ ...(7·42)

We shall consider the following two estimates of the population mean $\bar{Y}_N$:

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{y}_{n_i} \quad \ldots (7·43)$$

$$\bar{y}_{st} = \frac{1}{N} \sum_{i=1}^{k} N_i \bar{y}_{n_i} = \sum_{i=1}^{k} p_i \bar{y}_{n_i}, \quad \ldots (7·44)$$

the latter being weighted mean of the strata sample means, weights being equal to strata sizes.

These two estimates of the population mean are identical

if $\frac{n_i}{n} = \frac{N_i}{N} \Rightarrow \frac{n_i}{N_i} = \frac{n}{N} =$ (constant)$= C$, (say).

i.e., $n_i = CN_i \Rightarrow n_i \propto N_i$ [*c.f.* Proportional Allocation (§ 7·10·3)] ...(7·45)

Estimate of Population Mean and Its Variance.

Theorem 7·6. *$\bar{y}_{st}$ is an unbiased estimate of the population mean $\bar{Y}_N$, i.e.,*

$$E(\bar{y}_{st}) = \bar{Y}_N \quad \ldots (7·46)$$

Proof. Since the sample in each of the stratum is a simple random sample, we have

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

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

Theorem 7·7. $$Var(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} N_i (N_i - n_i) \frac{S_i^2}{n_i} = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) p_i^2 S_i^2 \quad \ldots (7·47)$$

Proof. Since the sample in each stratum is simple random sample without replacement, we have

$$\text{Var}(\bar{y}_{n_i}) = \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_i^2 \quad \ldots (*)$$

where S_i^2 is as defined in (7·41).

Scanned with CamScanner

Unit III Stratified Random Sampling

In simple random sampling, it has been seen that the precision of the standard estimators of the population total (mean) depends on two aspects, namely, the sample size (n) and the variability (S^2) of the character under study. ie, In SRSWOR we obtained

$$V(\bar{y}_n) = \frac{N-n}{N}\frac{S^2}{n} = \left(1-\frac{n}{N}\right)\frac{S^2}{n}$$

This implies that $V(\bar{y}_n)$ is

i) inversely proportional to the sample size and

ii) directly proportional to the variability of the sampling units in the population

Therefore in order to get an estimator with increased precision one can increase the sample size. However considerations of cost limits the size of the sample. The other possible way to estimate the population total (mean) with greater precision is to divide the population into several groups each of which is more homogeneous than the entire population and draw a sample of predetermined size from each of these groups. The groups into which the population is divided are called strata and drawing sample from each of the strata is called stratified sampling.

Scanned with CamScanner

Stratification means division into layers. Auxiliary information (past data or some other information) related to the character under study may be used to divide the population into various groups such that

i) units within each group are as homogeneous as possible and ii) the group means are as widely different as possible.

Thus a population consisting of 'N' sampling units is divided into 'k' relatively homogeneous mutually disjoint subgroups, termed as strata, of sizes $N_1, N_2, \ldots, N_k$, such that $N = \sum_{i=1}^{k} N_i$

If a SRS (generally without replacement) of size $n_i, (i = 1, 2, \ldots, k)$ is drawn from each of the stratum respectively such that $n = \sum_{i=1}^{k} n_i$, the sample is termed as Stratified Random Sample of size 'n' and the technique of drawing such a sample is called Stratified Random Sampling.

Notations and Terminology

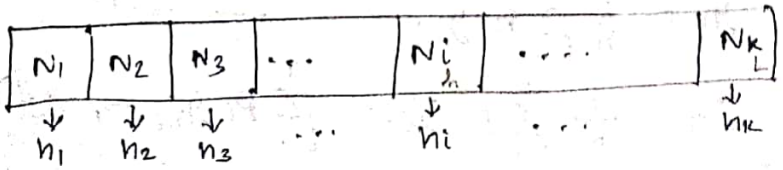
In a population there are 'N' sampling units and they are divided into 'k' number of strata.

N_i = The no. of sampling units in the i^{th} stratum $(i = 1, 2, \ldots k)$

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

n_i = The no. of sampling units selected with SRSWOR from the i^{th} stratum

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



$n_1 + n_2 + n_3 + \cdots + n_i + \cdots + n_k = n$

Scanned with CamScanner

Population

Let y_{ij} $(j=1,2,\ldots,N_i;\ i=1,2,\ldots k)$ be the value of the j^{th} unit in the i^{th} stratum.

$\bar{Y}_{Ni}$ = Population mean of i^{th} stratum

$$= \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$$

$\bar{Y}_N$ = Population mean

$$= \frac{1}{N} \sum_i \sum_j y_{ij}$$

$$= \frac{1}{N} \sum_{i=1}^{k} N_i \bar{Y}_{Ni}$$

$$= \sum_{i=1}^{k} P_i \bar{Y}_{Ni}$$

where $P_i = \frac{N_i}{N}$ is called the weight of the i^{th} stratum.

S_i^2 = Population mean square of the i^{th} stratum

$$= \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_{Ni})^2, \quad (i = 1,2,\ldots,k)$$

example :

if $i = 7^{th}$ stratum and $N_7 = 25$ then

$$S_7^2 = \frac{1}{25-1} \sum_{j=1}^{25} (y_{7j} - \bar{Y}_{25})^2$$

Sample

Let y_{ij} = Value of j^{th} sampled unit from i^{th} stratum

$\bar{y}_{ni}$ = Mean of sample selected from i^{th} stratum

s_i^2 = Sample mean square of the i^{th} stratum

$$= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{ni})^2; \quad (i = 1,2,\ldots,k)$$

We shall consider the following two estimates of the population mean $\bar{Y}_N$, which are

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{y}_{ni} \qquad \text{——— (1)}$$

$$\bar{y}_{st} = \frac{1}{N} \sum_{i=1}^{k} N_i \bar{y}_{ni} = \sum_{i=1}^{k} P_i \bar{y}_{ni} \qquad \text{——— (2)}$$

Scanned with CamScanner

The estimates ① & ② are identical if $\frac{n_i}{n} = \frac{N_i}{N}$

$$\Rightarrow \frac{n_i}{N_i} = \frac{n}{N} = C \text{ (constant)}$$

$$n_i = C N_i$$

$$\Rightarrow n_i \propto N_i$$

Theorem: 1

P.T. in Stratified random sampling, the sample mean is an unbiased estimate of the population mean

ie., $E(\bar{y}_{st}) = \bar{Y}_N$

Proof

Let $\bar{y}_{ni}$ is the mean of sample selected from ith stratum. Since the sample in each of the stratum is a simple random sample, we have

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

We know that

$$\bar{y}_{st} = \frac{1}{N} \sum_{i=1}^{K} N_i \bar{y}_{ni}$$

Taking Expectation on both sides, we get

$$E(\bar{y}_{st}) = E\left[\frac{1}{N} \sum_{i=1}^{K} N_i \bar{y}_{ni}\right]$$

$$= \frac{1}{N} \sum_{i=1}^{K} N_i E(\bar{y}_{ni})$$

$$= \frac{1}{N} \sum_{i=1}^{K} N_i \bar{Y}_{Ni}$$

$$= \bar{Y}_N$$

and hence the proof.

Scanned with CamScanner

Theorem 2:

Show that in Stratified random sampling

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

Proof:

Since the sample in each stratum is simple random sample without replacement, we have

$$V(\bar{y}_{ni}) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2 \quad \text{---------- (1)}$$

where S_i^2 is defined in SRSWOR

We know that

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

$$= \sum_{i=1}^{k} P_i \bar{y}_{ni} \qquad \because \frac{N_i}{N} = P_i$$

$$\therefore V(\bar{y}_{st}) = V\left[\sum_{i=1}^{k} P_i (\bar{y}_{ni})\right]$$

$$= \sum_{i=1}^{k} P_i^2 V(\bar{y}_{ni})$$

the Covariance terms vanish since the samples from different strata are independent

$$V(\bar{y}_{st}) = \sum_{i=1}^{k} P_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2 \quad \text{from (1)}$$

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

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

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

and hence the proof.

Note: From the above, we see that $V(\bar{y}_{st})$ depends on S_i^2, the heterogenity within the strata. Thus if S_i^2 are small ie, strata are homogeneous then Stratified sampling provides estimates with greater precision. (ie., when S_i^2 increases $V(\bar{y}_{st})$ also increases)

Scanned with CamScanner

Remark :

1. In general S_i^2 are not known. Since a simple random sample is drawn from each stratum, $E(s_i^2) = S_i^2$ $i=1,2,\ldots,k$

Accordingly an unbiased estimate of the $V(\bar{y}_{st})$ is given by

$$\text{Estimate of } V(\bar{y}_{st}) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) P_i^2 s_i^2$$

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

2. With stratified random sampling there is in general no single finite population correction factor (f.p.c) since $\frac{N_i - n_i}{n_i}$ may be different for different values of $i = 1, 2, \ldots, k$. When N_i is very large and n_i is very small, we take $(N_i - n_i)$ as N_i. Hence we get

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} N_i^2 \frac{s_i^2}{n_i}$$

ALLOCATION OF SAMPLE SIZE

We know that

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} N_i (N_i - n_i) \frac{s_i^2}{n_i}$$

$$= \sum_{i=1}^{k} P_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) s_i^2 \quad \text{where } P_i = \frac{N_i}{N}$$

From the above expression we observe that $V(\bar{y}_{st})$ depends on 'n_i' (the sample size for the i^{th} stratum) which can be fixed at will. The allocation of the sample sizes for different strata is done in the following ways :

(a) Proportional allocation, and

(b) Optimum allocation.

Scanned with CamScanner

a) Proportional allocation

Allocation of n_i's to various strata is proportional if the sample fraction is constant for each stratum, i.e.

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_k}{N_k} = \frac{\Sigma n_i}{\Sigma N_i} = \frac{n}{N} = c \text{ (Constant)}$$

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

$$\Rightarrow n_i \propto N_i, \quad i = 1, 2, \ldots k$$

Thus in proportional allocation each stratum is represented according to its size.

Example :

From a population of 8000 units, a sample of 30 units is to be taken. The population is divided into three strata with the units 4000, 2400 and 1600 units. How would you draw the sample using proportional allocation technique?

Soln : It is given that $N = 8000$, $n = 30$

$N_1 = 4000$, $N_2 = 2400$, $N_3 = 1600$.

In proportional allocation, we have

$$n_i = \frac{n}{N} \cdot N_i = \frac{30}{8000} \times N_i$$

$$\Rightarrow n_1 = \frac{30}{8000} \times N_1 = \frac{30}{8000} \times 4000 = 15$$

$$n_2 = \frac{30}{8000} \times N_2 = \frac{30}{8000} \times 2400 = 9$$

$$n_3 = \frac{30}{8000} \times N_3 = \frac{30}{8000} \times 1600 = \frac{6}{30}$$

Scanned with CamScanner

Theorem:

In proportional allocation, Prove that

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

Proof:

In ~~proportional allocation, the~~ Stratified r.s $V(\bar{y}_{st})$ is given by

$$V(\bar{y}_{st})_{prop} = \frac{1}{N^2} \sum_{i=1}^{k} N_i [N_i - n_i] \frac{S_i^2}{n_i}$$

$$= \sum_{i=1}^{k} \frac{N_i}{N^2} \left[\frac{N_i}{n_i} - 1\right] S_i^2$$

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

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

$$= \frac{1}{N} \left(\frac{N-n}{n}\right) \sum_{i=1}^{k} P_i S_i^2$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} P_i S_i^2$$

and hence the proof.

b) Optimum allocation

Another guiding principle in the determination of the n_i's is to choose them so as to:

a) Minimise the Variance (ie., maximise the precision) of the estimate for
   (i) fixed sample size n, (ii) fixed Cost

b) Minimise the total Cost for fixed desired precision.

The allocation of n_i's to various strata in accordance with the above principles is known as optimum allocation.

Scanned with CamScanner

In optimum allocation n_i's are to be obtained such that

1. $V(\bar{y}_{st})$ is minimum for fixed n.
2. $V(\bar{y}_{st})$ is minimum for fixed total cost C (say)
3. Total cost C is minimum for fixed value of $Var(\bar{y}_{st})$

Cost function

The simplest form of the cost function in stratified sampling may be given by the linear model

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

where a = overhead (or) fixed cost

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

Theorem : 1

$V(\bar{y}_{st})$ is minimum for fixed total size of the sample (n) if $n_i \propto N_i S_i$

Proof :

Here the problem is to minimize

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

subject to the condition $\sum_{i=1}^{k} n_i = n$ (fixed)

This is equivalent to minimising

$$\phi = V(\bar{y}_{st}) + \lambda \left(\sum_{i=1}^{k} n_i - n \right)$$

$$= \frac{1}{N^2} \sum_{i=1}^{k} N_i (N_i - n_i) \frac{S_i^2}{n_i} + \lambda \left(\sum_{i=1}^{k} n_i - n \right) \quad \text{---} (1)$$

unconditionally for variations in n_i and where λ is the unknown Lagrange multiplier.

Scanned with CamScanner

The function ϕ attains its minimum if

i) $\frac{\partial \phi}{\partial n_i} = 0$ and solve for n_i

ii) $\frac{\partial^2 \phi}{\partial n_i^2} > 0$ at n_i

Hence $\frac{\partial \phi}{\partial n_i} = 0$

$$\Rightarrow \frac{\partial}{\partial n_i}\left[\frac{1}{N^2}\sum_{i=1}^{k}\frac{N_i^2 S_i^2}{n_i} - \frac{1}{N^2}\sum_{i=1}^{k} N_i n_i \frac{S_i^2}{n_i}\right] + \frac{\partial}{\partial n_i}\left[\lambda \sum_{i=1}^{k} n_i - \lambda n\right] = 0$$

$$\Rightarrow \frac{\partial}{\partial n_i}\left[\frac{1}{N^2}\left\{\frac{N_1^2 S_1^2}{n_1} + \frac{N_2^2 S_2^2}{n_2} + \cdots + \frac{N_i^2 S_i^2}{n_i} + \cdots + \frac{N_k^2 S_k^2}{n_i}\right\} - \frac{1}{N^2}\sum_{i=1}^{k} N_i S_i^2\right] +$$

$$\frac{\partial}{\partial n_i}\left[\lambda\{n_1 + n_2 + \cdots + n_i + \cdots + n_k\} - \lambda n\right] = 0$$

Here we are differentiating only for i^{th} value, we get

$$\Rightarrow -\frac{N_i^2 S_i^2}{N^2 n_i^2} + \lambda = 0 \quad \text{---------- (2)}$$

$$\Rightarrow \lambda = \frac{N_i^2 S_i^2}{N^2 n_i^2}$$

$$\Rightarrow n_i^2 = \frac{N_i^2 S_i^2}{N^2 \lambda}$$

$$\Rightarrow n_i = \frac{N_i S_i}{N\sqrt{\lambda}} \quad \text{---------- (3)}$$

$$\frac{\partial^2 \phi}{\partial n_i^2} = \frac{\partial}{\partial n_i}\left[-\frac{N_i^2 S_i^2}{N^2 n_i^2} + \lambda\right]$$

$$= \frac{2 N_i^2 S_i^2}{N^2 n_i^3} \quad \text{---------- (4)}$$

Put (3) in (4), we get

$$\left.\frac{\partial^2 \phi}{\partial n_i^2}\right]_{n_i} = \frac{2 N_i^2 S_i^2}{N^2 \left(\frac{N_i S_i}{N\sqrt{\lambda}}\right)^3} > 0$$

Hence the function ϕ attains its minimum at $n_i = \frac{N_i S_i}{N\sqrt{\lambda}}$

Scanned with CamScanner

To find the value of λ, Consider eqn. (3)

Consider $n_i = \frac{N_i S_i}{N\sqrt{\lambda}}$

Taking summation over $i = 1, 2, \ldots k$ on both sides, we get

$$\sum_{i=1}^{k} n_i = \frac{\sum_{i=1}^{k} N_i S_i}{N\sqrt{\lambda}}$$

$$\Rightarrow n = \frac{\sum_{i=1}^{k} N_i S_i}{N\sqrt{\lambda}} \quad \text{since } \sum_{i=1}^{k} n_i = n$$

$$\Rightarrow \sqrt{\lambda} = \frac{\sum_{i=1}^{k} N_i S_i}{Nn} \quad \text{---} (5)$$

Substitute the value of λ obtained from (5) in (3), we get

$$n_i = \frac{N_i S_i}{N\sqrt{\lambda}}$$

$$= \frac{N_i S_i}{N \sum_{i=1}^{k} N_i S_i} \times Nn$$

$$\therefore n_i = \frac{N_i S_i}{\sum_{i=1}^{k} N_i S_i} \times n$$

Thus in optimum allocation for a fixed total sample size, we have

$n_i \propto N_i S_i$

This is known as Neyman's formula for optimum allocation.

Hence in stratified random sampling $V(\bar{y}_{st})$ is minimum for fixed sample size n if $n_i \propto N_i S_i$

This suggests that a larger sample size is to be selected from the stratum if

i) Stratum size (N_i) is large

ii) Stratum variability (S_i) is large

in order to obtain the most precise of the population mean.

Scanned with CamScanner

Theorem 2 :

In stratified random sampling with given cost function of the form $C = a + \sum_{i=1}^{k} c_i n_i$, $V(\bar{y}_{st})$ is minimum if $n_i \propto \frac{N_i S_i}{\sqrt{c_i}}$

Proof :

Here we have to minimise $V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} N_i (N_i - n_i) \frac{S_i^2}{n_i}$

subject to the condition $C = a + \sum_{i=1}^{k} c_i n_i$

This is equivalent to minimise

$$\phi = V(\bar{y}_{st}) + \lambda \left(\sum_{i=1}^{k} c_i n_i - C + a \right) \quad \text{---- (1)}$$

unconditionally for variations in n_i and λ is the unknown Lagrange multiplier.

The function ϕ attains its minimum if i) $\frac{\partial \phi}{\partial n_i} = 0$ and ii) $\frac{\partial^2 \phi}{\partial n_i^2} > 0$ at n_i

Hence $\frac{\partial \phi}{\partial n_i} = 0$

$$\Rightarrow \frac{\partial}{\partial n_i} \left[\frac{1}{N^2} \sum_{i=1}^{k} N_i (N_i - n_i) \frac{S_i^2}{n_i} + \lambda \left(\sum_{i=1}^{k} c_i n_i - C + a \right) \right] = 0$$

$$\Rightarrow -\frac{N_i^2 S_i^2}{N^2 n_i^2} + \lambda c_i = 0 \quad \text{---- (2)}$$

$$\Rightarrow \lambda c_i = \frac{N_i^2 S_i^2}{N^2 n_i^2}$$

$$\Rightarrow n_i^2 = \frac{N_i^2 S_i^2}{N^2 \lambda c_i}$$

$$\Rightarrow n_i = \frac{N_i S_i}{N \sqrt{\lambda} \sqrt{c_i}} \quad \text{---- (3)}$$

Scanned with CamScanner

$$\frac{\partial^2 \phi}{\partial n_i^2} = \frac{\partial}{\partial n_i}\left[-\frac{N_i^2 S_i^2}{N^2 n_i^2} + \lambda C_i\right]$$

$$= \frac{2 N_i^2 S_i^2}{N^2 n_i^3} > 0 \quad \text{at} \quad n_i = \frac{N_i S_i}{N\sqrt{\lambda}\sqrt{C_i}}$$

Hence the function ϕ attains its minimum at n_i

To find the value of λ, Consider (3)

$$n_i = \frac{N_i S_i}{N\sqrt{\lambda}\sqrt{C_i}}$$

$$\Rightarrow \sqrt{\lambda} = \frac{N_i S_i}{N n_i \sqrt{C_i}}$$

Summing over i from 1 to k, we get

$$\sum_{i=1}^{k} n_i = \frac{\sum_{i=1}^{k}[N_i S_i/\sqrt{C_i}]}{N\sqrt{\lambda}}$$

$$\Rightarrow n = \frac{\sum_{i=1}^{k}[N_i S_i/\sqrt{C_i}]}{N\sqrt{\lambda}}$$

$$\Rightarrow \sqrt{\lambda} = \frac{\sum_{i=1}^{k}[N_i S_i/\sqrt{C_i}]}{Nn} \quad \text{---} \quad (4)$$

Substitute (4) in (3), we get

$$n_i = \frac{N_i S_i}{N \times \frac{\sum_{i=1}^{k} N_i S_i/\sqrt{C_i}}{Nn} \times \sqrt{C_i}}$$

$$= \frac{n N_i S_i/\sqrt{C_i}}{\sum_{i=1}^{k}[N_i S_i/\sqrt{C_i}]}$$

Thus, in optimum allocation for a fixed cost

$$n_i \propto \frac{N_i S_i}{\sqrt{C_i}}$$

Scanned with CamScanner

This leads to the following important Conclusion:

A larger sample would be required from a 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.

Theorem:

In Stratified random sampling, Prove that

$$V(\bar{y})_R > V(\bar{y}_{st})_{prop} > V(\bar{y}_{st})_{opt}.$$

Proof:

We know that, in SRSWOR

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

$$= \left(1-\frac{n}{N}\right)\frac{S^2}{n}$$

If finite population correction (f.p.c) is ignored, then

$$V(\bar{y})_R = \frac{S^2}{n} \quad \text{——— (1)}$$

We know that in Stratified random sampling.

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

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

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

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

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

Scanned with CamScanner

If f.p.c is ignored, then

$$V(\bar{y}_{st})_{prop} = \frac{1}{Nn}\sum_{i=1}^{k} N_i S_i^2 \quad \text{——— (2)}$$

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

$$\therefore V(\bar{y}_{st})_{opt} = \frac{1}{N^2}\sum_{i=1}^{k} N_i\left[\frac{N_i}{\frac{N_i S_i \cdot n}{\sum N_i S_i}} - 1\right] S_i^2$$

$$= \frac{1}{N^2}\sum_{i=1}^{k} N_i \cdot \left[\frac{N_i \sum_{i=1}^{k} N_i S_i}{N_i S_i \cdot n} - 1\right] S_i^2$$

$$= \frac{1}{N^2}\sum_{i=1}^{k} N_i S_i^2 \left[\frac{N_i \sum_{i=1}^{k} N_i S_i - N_i S_i \cdot n}{N_i S_i \cdot n}\right]$$

$$= \frac{1}{N^2}\left[\frac{\sum_{i=1}^{k} N_i S_i \cdot \sum_{i=1}^{k} N_i S_i \cdot N_i}{N_i S_i \cdot n}\right] - \frac{1}{N^2}\left[\frac{\sum N_i S_i^2 \cdot N_i S_i \cdot n}{N_i S_i \cdot n}\right]$$

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

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

if f.p.c is ignored, then we get

$$V(\bar{y}_{st})_{opt} = \frac{1}{N^2 n}\left(\sum_{i=1}^{k} N_i S_i\right)^2 \quad \text{——— (3)}$$

Let us consider $$S^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{N_i}(y_{ij} - \bar{Y})^2$$

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

Scanned with CamScanner

Add and Subtract $\bar{y}_n$ within the bracket, we get

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

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

$$2\sum_{i=1}^{k}\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_n)(\bar{y}_n - \bar{Y})$$

$$= \sum_{i=1}^{k}(N_i - 1)S_i^2 + \sum_{i=1}^{k} N_i(\bar{y}_n - \bar{Y})^2 + 0$$

Since sum of the product of deviations from mean is zero

and $S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_n)^2$

$$(N_i - 1)S_i^2 = \sum_{j=1}^{N_i}(y_{ij} - \bar{y}_n)^2, \quad i = 1, 2, \ldots k$$

$$\therefore (N-1)S^2 = \sum_{i=1}^{k} N_i\left(1 - \frac{1}{N_i}\right)S_i^2 + \sum_{i=1}^{k} N_i(\bar{y}_n - \bar{Y})^2$$

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

if f.p.c. is ignored, we get.

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

Divide N.n on both sides, we get

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

$$\Rightarrow V(\bar{y})_R = V(\bar{y}_{st})_{prop} + \text{Some +ve quantity.}$$

$$\Rightarrow V(\bar{y})_R > V(\bar{y}_{st})_{prop} \quad \text{——— (A)}$$

Scanned with CamScanner

Gain in efficiency in St.R.S over SRSWOR:

a) due to proportional allocation is

$$\frac{V(\bar{y}_n)_R - V(\bar{y}_{st})_P}{V(\bar{y}_{st})_P}$$

b) due to Neyman's optimum allocation is

$$\frac{V(\bar{y}_n)_R - V(\bar{y}_{st})_N}{V(\bar{y}_{st})_N}$$

Consider the difference

$$V(\bar{y}_{st})_{prop} - V(\bar{y}_{st})_{opt}$$

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

$\because P_i = \frac{N_i}{N}$

$$= \frac{1}{n}\left[\sum_{i=1}^{k} P_i (S_i - \bar{S})^2\right] \quad \text{where } \bar{S} = \sum P_i S_i$$

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

$$\Rightarrow V(\bar{y}_{st})_{prop} \geq V(\bar{y}_{st})_{opt} \quad \text{——— (B)}$$

Comparing (A) & (B), we get

$$V(\bar{y})_R > V(\bar{y}_{st})_{prop} > V(\bar{y}_{st})_{opt}$$

Theory Questions

1. Explain Stratified random sampling. Also state its merits and demerits (Apr' 2008)
2. Write down the principles of stratification? (Nov' 2008)
3. What is stratification? When is it useful? (Nov' 2008)
4. What is Stratified random sampling? State its principal advantages (Nov' 2006)
5. State the advantages and Disadvantages of Stratified random sampling. (Nov' 2007)

$$\sum P_i (S_i - \bar{S})^2 = \sum_{i=1}^{k} P_i (S_i^2 - 2S_i\bar{S} + \bar{S}^2)$$

$$= \sum P_i S_i^2 - 2\bar{S}\sum P_i S_i + \bar{S}^2 \sum P_i$$

$$= \sum P_i S_i^2 - 2(\sum P_i S_i)^2 + (\sum P_i S_i)^2 \quad \because \sum P_i = 1 \text{ & } \bar{S} = \sum P_i S_i$$

$$= \sum P_i S_i^2 - (\sum P_i S_i)^2$$

Scanned with CamScanner

Source :

1. S.C. Gupta and V.K. Kapoor: Fundamental of Applied Statistics –Sultan Chand & Sons, Fourth Edition, 2015.