SAMPLING

Population:

In statistics, population does not refer only to persons but it is the aggregate or collection of a specified group of similar objects or individuals that have some common observable characteristic**.**

Sample:

A part of statistical population selected according to some rule or plans for drawing conclusions regarding the population is called sample. The number of individual in sample is called sample size.

Sampling units:

The elementary unit or group of elementary units in the population which is used as the basis of selection is called sampling units. (it changes according to the selection procedure)

Sampling frame: (is the complete list of sampling units)

A complete list of sampling units or other accepted materials which represents the population to be covered is called the sampling frame.

Types of survey:

1. **Complete Enumeration (Census survey):-**
   When each and every unit in the population is examined for the characteristics under study, we call it a complete enumeration or Census survey.
2. **Sample survey:-**
   When only a part called sample is selected from the population and is examined then we have a sample enumeration or sample survey.

Need of Sampling:

1. There is a reduction of cost either in terms of money or in terms of man-powers in a sample survey. In many cases, our resources may be limited or it may be necessary that the result of the survey should be available within a specified time. In such cases, it is imperative to adopt a sample survey rather than complete enumeration.

2. There is generally greater scope in a sample survey than in census. Some enquires may require highly trained person specialized equipment for collection of data. Thus in a sample survey may have greater coverage both in respect of the information collected and in respect of geographical boundaries taken into account.

3. A sample survey generally gives data of a better quality than a complete census because in a sample survey, it may be possible to employ better trained personal or better equipment than is possible or feasible in a complete enumeration.

Basic principle of survey:

1. Statistical regularity
2. Validity
3. Optimization

1. **Statistical regularity:-**

The principle of regularity stresses the desirability and importance of selecting a sample at-random. So that each and every unit in population has an equal chance of being selected in the sample.

2. **Validity:-**

By validity of sample design, we mean that the sample should be so selected that the results could be interpreted objectively in terms of probability. That means it should enable us to obtain valid test and estimates about the population parameters

3. **Optimization:-**

The principle of optimization ensures that a given level of efficiency will be reached with minimum cost or that the maximum possible efficiency will be obtained with a given level of cost.

Errors in survey:

1. **Sampling errors**

   An error which arises due to sampling is called sampling error. In other words, sampling errors arise from the fact that only a part of population which is used for estimation. The discrepancy observed between different estimates of the same population parameter or between value of parameter and its estimate derived from the sample are called sampling errors. Occurs only a sample survey.

2. **Non-Sampling errors**

   i. Non-Response error
   ii. Response error (or measurement error or observational error)
   iii. Tabulation error
   iv. Computational error

i. **Non-Response error**

The error which arises when we fail to get the information is called non-response error and phenomenon is called non-response. This error arises because of fact that we are not able to cover the whole sample.

ii. **Response error / measurement error /observational error**

The error that we bring in measuring the character is called measurement error.

iii. **Tabulation error**

The error which arises due to missing some numbers due to non-availability of data or recording some numbers wrongly making a table is called a tabulation error.

iv. **Computational error**

After the table is formed, we start our calculation .The errors committed in computation are known as computational error.

Different step in a survey:

1. **Defining the objective**

   The objective of the survey must be clearly defined. Along with objectives, the planner should take into account the available resources in terms of money and manpower, the time limit within which the survey result must be available and the accuracy designed in the set of estimates to be prepared.

2. **Defining the population**

   The population which the survey result would apply must be clearly and unambiguously defined. The geographical, demography and other boundaries of the population must be specified that no ambiguity arises regarding the coverage of the survey.

3. **Preparing a questionnaire**

   A list of questions which is known as questionnaire must be prepared. The questionnaire is revised and finalized in the light of the trail data. The questions should be brief, practical and as objectives possible and they must not leave much scope for guessing (gazing) on the part of the interviewer.

4. **Choice of sample unit**

   The sampling frame must be available for selecting the units in the sample from the population. If it is available, it must be scruitanised to see whether it is adequate, complete, accurate, up-to-date and not subject to any duplication .

5. **Drawing the sample**

   In order to select the sample from the population, the technique of random sampling should be applied.

6. **Deciding the method of collection of data**

   One must decide whether the interview method or the mail questionnaire method is to be adapted. Although the later method is less costly, there is large scope for non-response and it is only practicable among educated people who are interested in the population survey

7. **Tabulation and analysis of data**

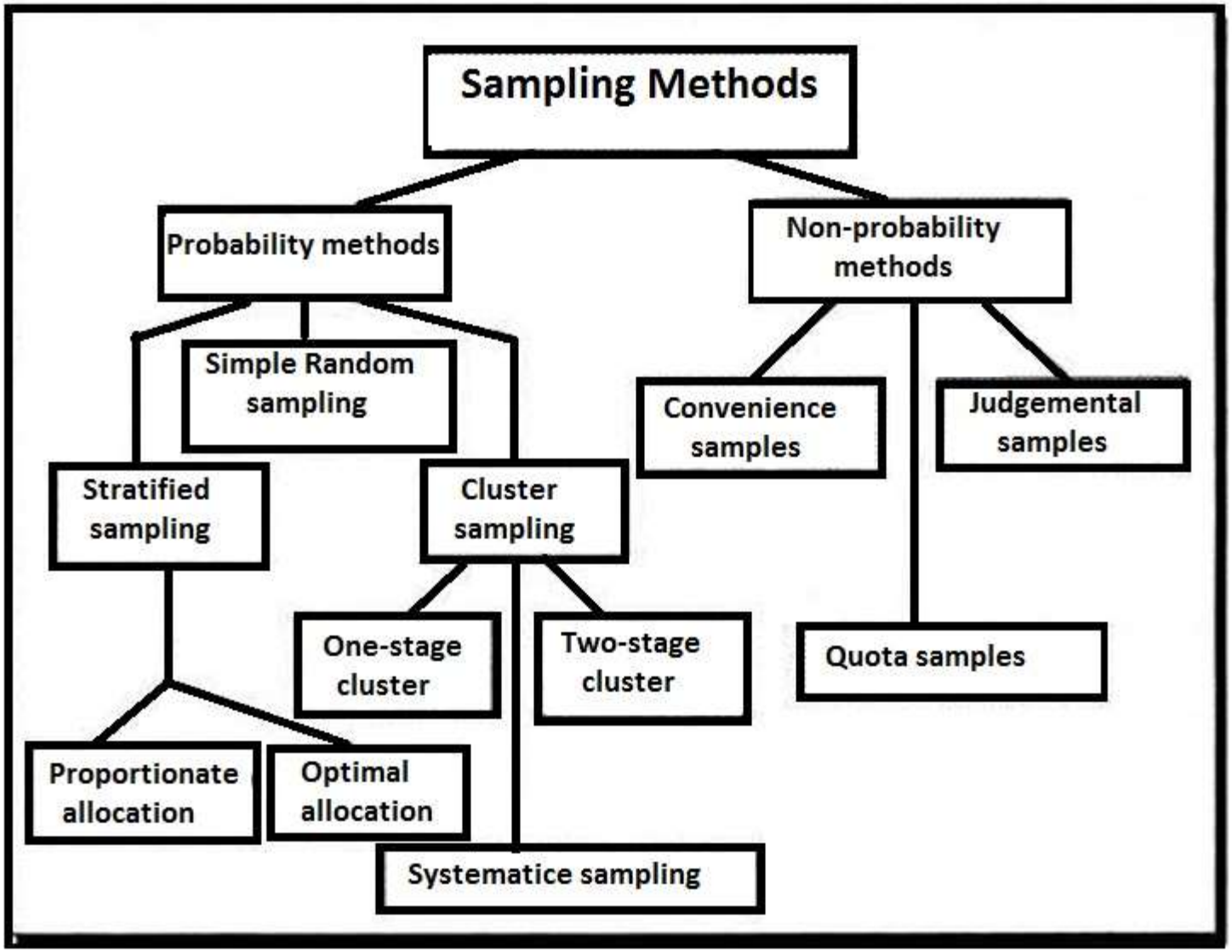
   The tabulation of the data can be done by hand or by machine. The primary tables may be further utilized for deriving necessary estimates for population characteristic or for testing of hypothesis.

National sample survey (NSS):

The NSS was initiated in 1950 to conduct the sampling inquiries with a view to preceding the government and other organization with socio-economic data which can be used for planning for national development and for research purposes. The measure portion of the field work in now conducted by NSSO (National sample survey organization) ,government of India . The technical work relating to the NSS, the processing and analysis of data and the preparation of the final reports were previously entrusted to the Indian Statistical Institute (ISI) but this too has now been taken over by NSSO.

Central Statistical organization(CSO):

Set up in 1951 by the government of India to co-ordinate statistical activities of the different ministries, state government and other statistical organization in the country. The CSO is responsible for deciding upon the coverage of the survey and methodology to be used.


Sampling Methods
Probability methods
Non-probability methods
Simple Random sampling
Stratified sampling
Cluster sampling
Convenience samples
Judgemental samples
One-stage cluster
Two-stage cluster
Quota samples
Proportionate allocation
Optimal allocation
Systematice sampling

Simple Random Sampling:

Two way are-

1- With replacement (SRSWR)

2- Without replacement (SRSWOR)

Random Sample:

The method of selecting a sample from population in which each and every unit of the population has equal probability of selection in the sample. In this method each of the possible sample has equal probability of selection and the sample selected by this method is called Random Sample.

(1) Simple Random Sampling With Replacement (SRSWR):

The unit drawn by the method of SRS is replaced in the population before the next draw and each of the unit is selected with same probability, the method is known as SRSWR scheme.

Let us consider a population consists of 'N' units and we select a sample of size 'n'

The probability of selecting a particular unit in **1st** draw $= \frac{1}{N}$

The probability of selecting a particular unit in **2nd** draw$= \frac{1}{N}$

. . . . .

. . . . .

. . . . .

The probability of selecting a particular unit in **rth** draw $= \frac{1}{N}$ (since $r \leq n$)

Properties:

1. The probability of selecting a particular unit at any draw is $\frac{1}{N}$.
2. The probability that a particular unit is included in the sample is $\frac{n}{N}$.
3. The total number of ordered sample is N^n.

Example:- Let $P=[1,2,3]$

$N=3, n=2$

So , Total number of ordered sample $\Rightarrow 3^2=9$

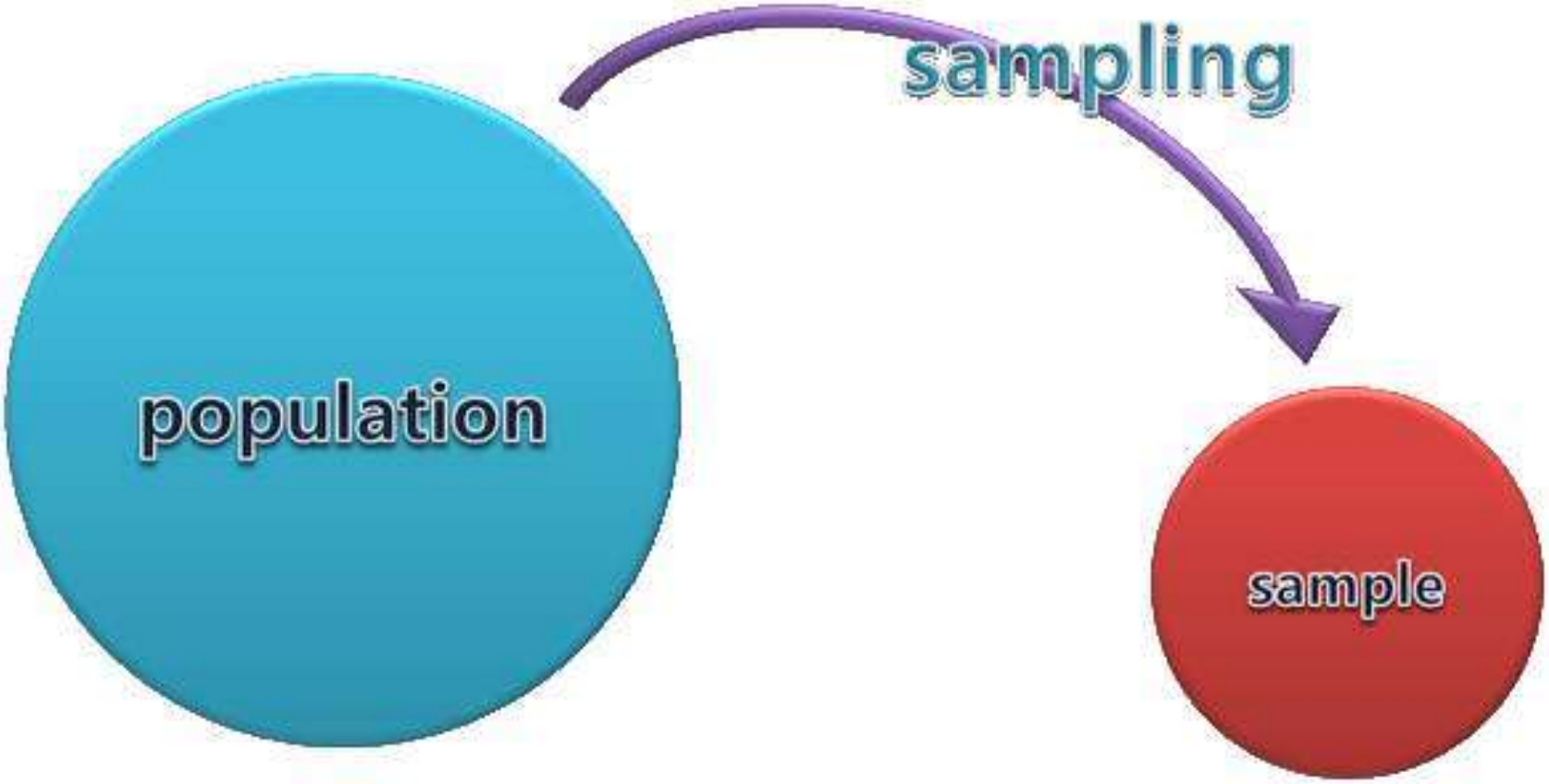
S = $\begin{Bmatrix}(1,1),(1,2),(1,3)\\(2,1),(2,2),(2,3)\\(3,1),(3,2),(3,3)\end{Bmatrix}$

note: (1,2) and (2,1) ,.. same sample but order is different

4. Total number of unordered sample is $\binom{N+n-1}{n}$.

   e.g. –for the given example number of unordered sample is 6.

SIMPLE RANDOM SAMPLING



(2) Simple Random Sampling Without Replacement (SRSWOR):

The unit draw by the method of SRS is not replaced in the population before the next draw, and each draw of the unit is selected with equal probability from the remaining units, the method is known as SRSWOR.

The probability of selecting a particular unit in **1st** draw $= \frac{1}{N}$

The probability of selecting a particular unit in **2nd** draw

=(The probability that unit is not selected in 1st draw)*(The probability that unit is selected in 2nd draw)

$$= \left(1 - \frac{1}{N}\right) * \left(\frac{1}{N-1}\right)$$

$$= \frac{1}{N}$$

The probability of selecting a particular unit in **3rd** draw $= \left(1-\frac{1}{N}\right)*\left(1-\frac{1}{N-1}\right)*\left(\frac{1}{N-2}\right)$

$= \frac{1}{N}$

. . . . .

. . . . .

. . . . .

The probability of selecting a particular unit in **rth** draw

$= \left(1-\frac{1}{N}\right)*\left(1-\frac{1}{N-1}\right)*\left(1-\frac{1}{N-2}\right).....\left(1-\frac{1}{N-r+2}\right)*\left(\frac{1}{N-r+1}\right)$

$= \frac{1}{N}$

Properties:

1. The probability of selecting a particular unit at any draw is $\frac{1}{N}$.
2. The probability that a particular unit is included in the sample is $\frac{n}{N}$.
3. The total number of ordered sample is ${}^{N}P_{n}$.
4. Total number of unordered sample is ${}^{N}C_{n}$

Example:-

$N = 3, n = 2, P = [1,2,3]$

${}^{N}P_{n} = {}^{3}P_{2} = 6$

Ordered:

$s = (1,2),\ (1,3), (2,1), (2,3), (3,1), (3,2)$

Unordered:

${}^{N}C_{n} = {}^{3}C_{2} = 3$

$s = (1,2),\ (1,3), (2,3)$

Let P ($u_1, u_2, u_3 \ldots\ldots u_N$) denotes a population of 'N' units, from which a sample of size 'n' is selected with the help of SRS scheme. Let Y be the character under study and y_i (i=1,2,…… N) be the observation on the i^{th} unit.

Then we have

$$\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad \text{: Population mean}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \bar{Y}_N\right)^2 \qquad \text{: Population variance}$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}_N\right)^2 \qquad \text{: Population mean square}$$

$$\Rightarrow \quad \boxed{N\sigma^2 = (N-1)S^2}$$

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{: Sample mean}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}_n\right)^2 \qquad \text{: Sample mean square}$$

Theorem:

The sample mean is an unbiased estimator of population mean.

$$E\left(\bar{y}_n\right)=\bar{Y}_N$$

Population Total:-

$$T=\sum_{i=1}^{N} y_i = N\bar{Y}_N$$

Therefore, the estimator of population total can be obtained as

$$T^* = N\bar{y}_n$$

$$E\left(N\bar{y}_n\right)= NE\left(\bar{y}_n\right)= N\bar{Y}_N \text{ .}$$

Theorem:

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

$$=\frac{N-n}{Nn}S^2 \text{(SRSWOR)}$$

$$\Rightarrow \quad \boxed{V\left(\bar{y}_n\right)_{SRSWOR} \le V\left(\bar{y}_n\right)_{SRSWR}}$$

That is why SRSWOR is more preferable. If 'n' is increased ,Variance will be lesser.

Theorem (i)

The sample mean square is an unbiased estimator of population variance σ^2 in SRSWR .

i.e.-

$$E(s^2)=\sigma^2$$

(ii) The sample mean square is an unbiased estimator of population mean square in SRSWOR.

i.e.-

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

Theorem:

$$Cov(y_i, y_j)=0 \qquad \rightarrow SRSWR$$

$$Cov(y_i, y_j)=-\frac{\sigma^2}{N-1} \qquad \rightarrow SRSWOR$$

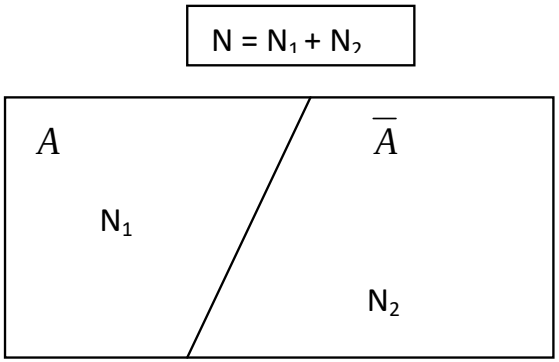
Theorem:

If X and Y be the two random variable then we have covariance between $\overline{y_n}$ and $\overline{x_n}$ is,

$$Cov(\overline{y_n}, \overline{x_n})=\frac{N-n}{Nn}S_{XY}$$

Where $$S_{XY}=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{Y_N})(x_i-\overline{X_N})$$

Estimation of population proportion in SRSWR:



$$p = \frac{N_1}{N} \qquad q = \frac{N_2}{N}, \qquad \hat{p} = \frac{n_1}{n}$$

$\hat{p} \Rightarrow$ Estimator of population proportion

Let us consider a population consists of 'N' units is divided in to mutually exclusive classes. Class 1 consisting 'N_1' units possessing the given attribute A and class 2 consisting 'N_2' units not possessing the attribute A.

A: Presence of A

$\overline{A}$: Absence of A

Let p and q be the population proportions of the units possessing the attribute A & not possessing the attribute A, respectively.

$$p = \frac{N_1}{N} \quad \text{and} \quad q = \frac{N_2}{N}$$

Now we select a sample of size 'n' and we observe that there are n_1 units possessing the attribute A & n_2 units not possessing the attribute A in the sample.

Let us consider the estimator of population proportion $\hat{p} = \frac{n_1}{n}$ (sample proportion)

Now $$E(\hat{p}) = E\left[\frac{n_1}{n}\right]$$

$$= \frac{1}{n} E[n_1]$$

Here $$E[n_1] = \sum n_1 p(n_1)$$

Where $p(n_1)$= probability that there are exactly n_1 units in the sample possessing the given attribute A.

$$p(n_1) = \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}}{\binom{N}{n}} = \frac{\binom{Np}{n_1}\binom{Nq}{n_2}}{\binom{N}{n}} \quad ; \text{n=0,1,2,3…….n}$$

$\therefore \quad E(n_1) = np$

So that $E(\hat{p}) = \frac{np}{n} = p$

Therefore we can say that the sample proportion $(\hat{p})$ is an unbiased estimator of population proportion ‘ p ’.

$$V(\hat{p}) = V\left(\frac{n_1}{n}\right)$$

$$= \frac{1}{n^2} V(n_1)$$

$\because \quad V(n_1) = \frac{npq(N-n)}{N-1}$

$$V(\hat{p}) = \frac{pq(N-n)}{n(N-1)}$$

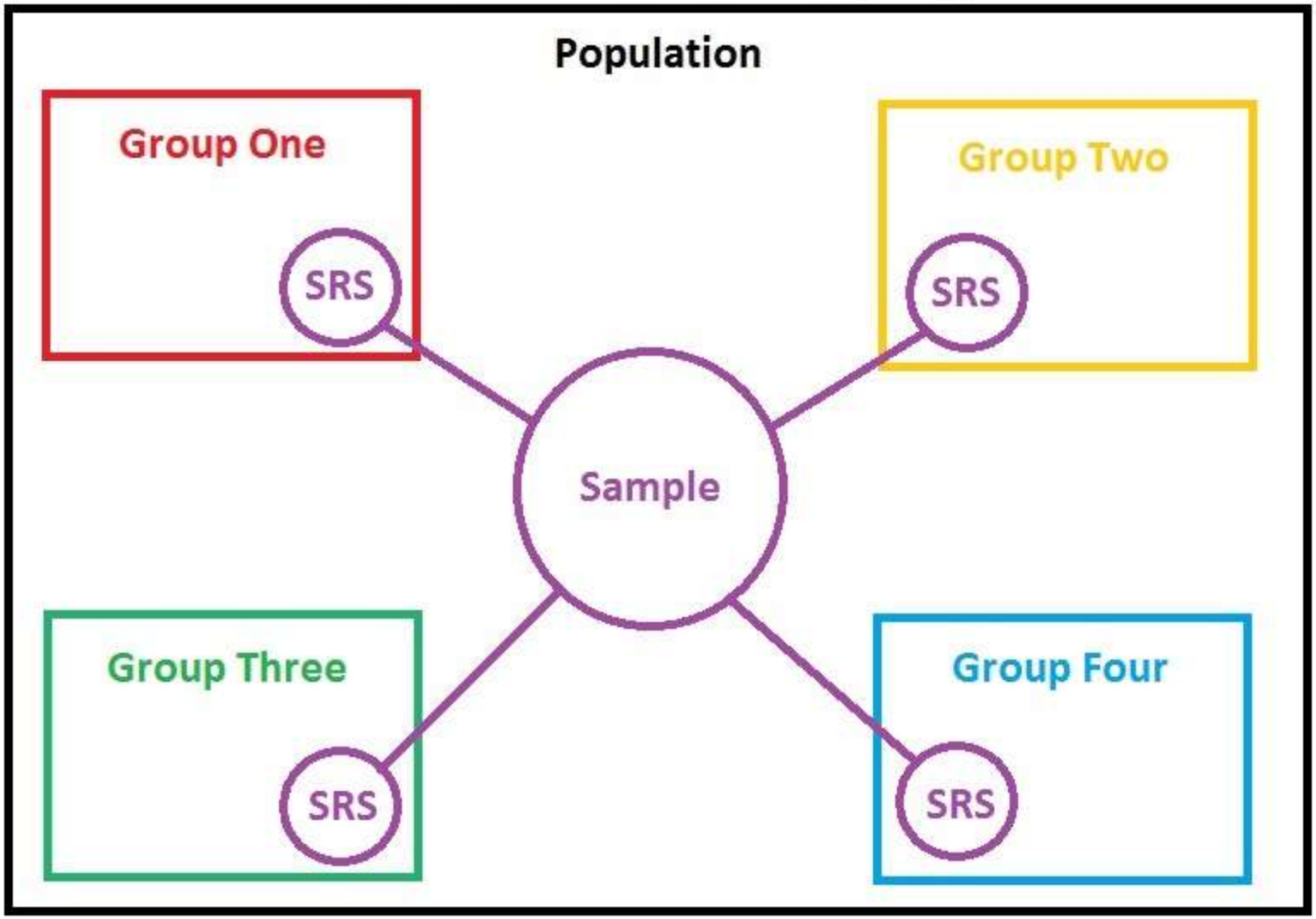
*Standard error is the standard deviation of a statistics.

$$S.E.(\hat{p}) = \sqrt{\frac{(N-n)pq}{(N-1)n}}$$

In SRSWR :

$$V(\hat{p}) = \frac{pq}{n}$$

Stratified Random Sampling



Stratified Random Sampling:

Let us consider a population P of N units is divided into K homogenous group (strata) such that i^{th} stratum consists of N_i units where i=1,2,3,……K and $\sum_{i=1}^{K} N_i = N$. If we select a sample of size n from the entire population in such a way that n_i units are selected randomly from the i^{th} stratum such that $\sum_{i=1}^{K} n_i = n$. Let y_{ij} (i=1,2,3,……k; j=1,2,3,…….. N_i) be the observation on j^{th} unit in the i^{th} stratum, than we have-

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

$$\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i} y_{ij} \quad \text{:Population mean}$$

$$= \sum_{i=1}^{K} p_i \bar{y}_{N_i} \quad \text{where } p_i = \frac{N_i}{N}$$

$$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(y_{ij} - \bar{y}_{N_i}\right)^2$$:Population mean square of ith stratum

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{K}\sum_{j=1}^{N_i}\left(y_{ij} - \bar{y}_{N_i}\right)^2$$:Population mean square

$$\bar{y}_{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$: sample mean of ith stratum

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{n_i}\right)^2$$: sample mean square of ith stratum

Let us define an estimator of population mean as,

$$\bar{y}_{st} = \sum_{i=1}^{K} p_i \bar{y}_{n_i}$$

Expectation of $\bar{y}_{st}$:

$$E\left(\bar{y}_{st}\right)=\bar{y}_N$$

$\bar{y}_{st}$ is an unbiased estimator of population mean $\bar{y}_N$.

variance of $\bar{y}_{st}$:

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

Allocation of n_i 's:

$$\text{If } n_i=\frac{n}{K} \qquad (2)$$

So from equation 1 we get,

$$V\left(\bar{y}_{st}\right)_{EA}=\frac{1}{nN^2}\sum_{i=1}^{K}\left(KN_i-n\right)N_i S_i^2 \qquad (3)$$

(i) Proportional allocation:

$$\text{If } \quad n_i \,\alpha\, \frac{n}{K}$$

Then

$$n_i=C_i N_i \qquad (4)$$

$$V\left(\bar{y}_{st}\right)_{PA}=\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{K} p_i S_i^2 \qquad (5)$$

(i) **Neyman allocation :**

$$n_i \ \alpha \ N_i$$

Or, $$n_i = C_2 N_i S_i \qquad (6)$$

$$V\left(\bar{y}_{st}\right)_{NA} = \frac{1}{n}\left(\sum_{i=1}^{K} \frac{p_i S_i}{n} - \frac{S_i}{N}\right) p_i S_i$$

(ii) **Optimum allocation:** let C_i be the cost per unit in ith stratum then the total cost of the survey is given by-

$$C = \sum_{i=1}^{K} n_i C_i \qquad (A)$$

There may be two cases:-

(a) **Total cost is fixed:** In this case, we fix the total cost and minimize the variance.

Here, $$C_0 = \sum_{i=1}^{K} n_i C_i \qquad (C_0 \text{ is fixed cost}) \qquad (i)$$

$$L = V\left(\bar{y}_{st}\right) + \mu\left(\sum_{i=1}^{K} n_i C_i - C_0\right) \qquad (ii)$$

μ is Lagrange multiplier

By differentiating the above equation w.r.t. n_i and equating the derivative to zero, we get

$$n_i = \left(\frac{p_i S_i}{\sqrt{\mu C_i}} \right) \qquad \text{(iii)}$$

Or,
$$n_i = \left(\frac{p_i S_i}{\sqrt{C_i}} \frac{C_0}{\sum_{i=1}^{K} p_i S_i \sqrt{C_i}} \right)$$

***particular case :**

(*) if $C_1 = C_2 = C_3 \ldots\ldots\ldots\ldots = C_K = C'$ (cost per unit is same in each stratum)

Then $C_0 = \sum_{i=1}^{K} n_i C_i = C_i' \sum_{i=1}^{K} n_i = C'n$

So , $V\left(\bar{y}_{st}\right)_{NA} = \sum_{i=1}^{K} \left(\frac{\sqrt{C_i} \sum_{i=1}^{K} p_i S_i \sqrt{C_i}}{C_0 p_i S_i} - \frac{1}{N_i} \right) p_i S_i^2$

(a) Variance is fixed:

In yhis case we fix the variance of the estimator and minimizing the total cost of the survey.

$$V_0 = V(\bar{y}_{st}) = \sum_{i=1}^{K}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)p_i^2 S_i^2$$

let us consider the lagrange function

$$L = C + \mu\left(V(\bar{y}_{st}) - V_0\right) \qquad \text{(i)}$$

μ is lagrange multiplier

By differentiating the above equation w.r.t. n_i and equating the derivative to zero, we get

$$n_i = \left(\frac{\sqrt{\mu}\, p_i S_i}{\sqrt{C_i}}\right) \qquad \text{(ii)}$$

Or,

$$C_{\min} = n\sum_{i=1}^{K}\left(\frac{p_i S_i \sqrt{C_i}}{\sum_{i=1}^{K}\frac{p_i S_i}{\sqrt{C_i}}}\right)$$

***comparison of Stratified random sampling with SRS:**

$$\frac{N-1}{N}S^2 = \sum_{i=1}^{K}\frac{(N_i-1)S_i^2}{N} + \sum_{i=1}^{K}\left(\bar{y}_{N_i} - \bar{y}_N\right)^2$$

If N and N_i is large:-

$$S^2 = \sum_{i=1}^{K} p_i S_i^2 + \sum_{i=1}^{K} p_i\left(\bar{y}_{N_i} - \bar{y}_N\right)^2$$

Now

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

By comparing we get,

$V\left(\bar{y}_n\right)_{SRSWOR} - V\left(\bar{y}_{st}\right)_{PA} \geq 0$; i.e., PA is better than SRSWOR.

This implies that the stratified random sampling under PA is more than SRS.

$$V\left(\bar{y}_{st}\right)_{PA} \geq V\left(\bar{y}_{st}\right)_{NA}$$

Stratified random sampling under neyman allocation is more précised than stratified random sampling under proportional allocation.i.e. the estimator under NA gives better estimate than SRS.

$$V\left(\bar{y}_{st}\right)_{NA} \leq V\left(\bar{y}_{st}\right)_{PA} \leq V\left(\bar{y}_n\right)_{SRSWOR} \leq V\left(\bar{y}_n\right)_{SRSWR}$$

SYSTEMATIC SAMPLING

Systematic sampling:

In this sampling scheme only the first unit is selected at-random and rest of the unit being automatically selected according to the pre determined pattern involving regular spacing of units.

Let us consider a population consists of N units. The unit is serially numbered from 1 to N and we select a sample of size n from a population, such that

$N = nK$

$k = \frac{N}{n}$; K is an integer called- sampling interval.

N is an integer multiple of sample size.

First of all we draw a number at random,say $i,(i \le K),i$ is called random start. Now, we select the unit corresponding to the number 'i' and every K^{th} unit is selected from the population in the sample. Thus the sample will consist of the unit $i, i+K, i+2K, ..., i+(n-1)K$.

Let Y be the characteristic under study and $y_{ij}(i=1,2,3.....,K;\ j=1,2,3,.....,n)$ be the observation on the j^{th} unit of the i^{th} sample.

$$\bar{y}_{i.}=\frac{1}{n}\sum_{j=1}^{n}y_{ij}$$: mean of the ith sample.

$$\bar{y}_{..}=\frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{n}y_{ij}$$: population mean

$$\bar{y}_{..}=\frac{1}{K}\sum_{i=1}^{K}\bar{y}_{i.}$$

There are K possible samples such as

random start	sample	prob.
1	$1, 1+K, 1+2K, 1+3K, \ldots\ldots\ldots, 1+jK, \ldots\ldots\ldots 1+(n-1)K$	$\frac{1}{K}$
2	$2, 2+K, 2+2K, 2+3K, \ldots\ldots\ldots, 2+jK, \ldots\ldots\ldots 2+(n-1)K$	$\frac{1}{K}$
.	. . . .	
.	. . . .	
.	. . . .	
i	$i, i+K, i+2K, i+3K, \ldots\ldots\ldots, i+jK, \ldots\ldots i+(n-1)K$	$\frac{1}{K}$
.	. . . .	
.	. . . .	
K	$K, 2K, 3K, 4K, \ldots\ldots\ldots\ldots\ldots\ldots, (i+j)K, \ldots\ldots\ldots\ldots, nK$	$\frac{1}{K}$

Estimation of Population mean:

Estimation of population mean $\bar{y}_{..}$ is given as

$$\bar{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

$$E\left(\bar{y}_{i.}\right) = \bar{y}_{..}$$

$\bar{y}_{i.}$ is an unbiased estimator of population ,mean $\bar{y}_{..}$.

$$\bar{y}_{i.} = \bar{y}_{sys}$$

$$V\left(\bar{y}_{i.}\right) = \frac{1}{K}\sum_{i=1}^{K}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$$

Or,

$$V\left(\bar{y}_{sys}\right) = \frac{1}{K}\sum_{i=1}^{K}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$$

Theorem:

$$V\left(\bar{y}_{sys}\right)=\frac{N-1}{N}\sigma^2-\frac{K(n-1)}{N}S_{wsy}{}^2$$

$$where, S_{wsy}{}^2=\frac{1}{K(n-1)}\sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{i.}\right)^2$$

$S_{wsy}{}^2$ is the mean square among the unit which lie within the same systematic sample.

Theorem:

$$V\left(\bar{y}_{sys}\right)=\frac{nK-1}{nK}\frac{S^2}{n}\left(1+(n-1)\rho\right)$$

$$where, \rho=\frac{E\left(y_{ij}-\bar{y}_{..}\right)\left(y_{ij'}-\bar{y}_{..}\right)}{E\left(y_{ij}-\bar{y}_{..}\right)^2}$$

ρ is the intraclass correlation coefficient between the unit of the same systematic sample.

Comparison of systematic sampling with SRSWOR:

we have ,

$$V\left(\overline{y_n}\right)_{SRSWOR} = \frac{N-n}{Nn} S^2$$

$$= \frac{nK-n}{n^2K} S^2$$

and,

$$V\left(\overline{y}_{sys}\right) = \frac{nK-1}{nK} \frac{S^2}{n} \left(1+(n-1)\rho\right)$$

The relative efficiency of the estimator of population mean in systematic sampling W.R.T. SRSWOR:-

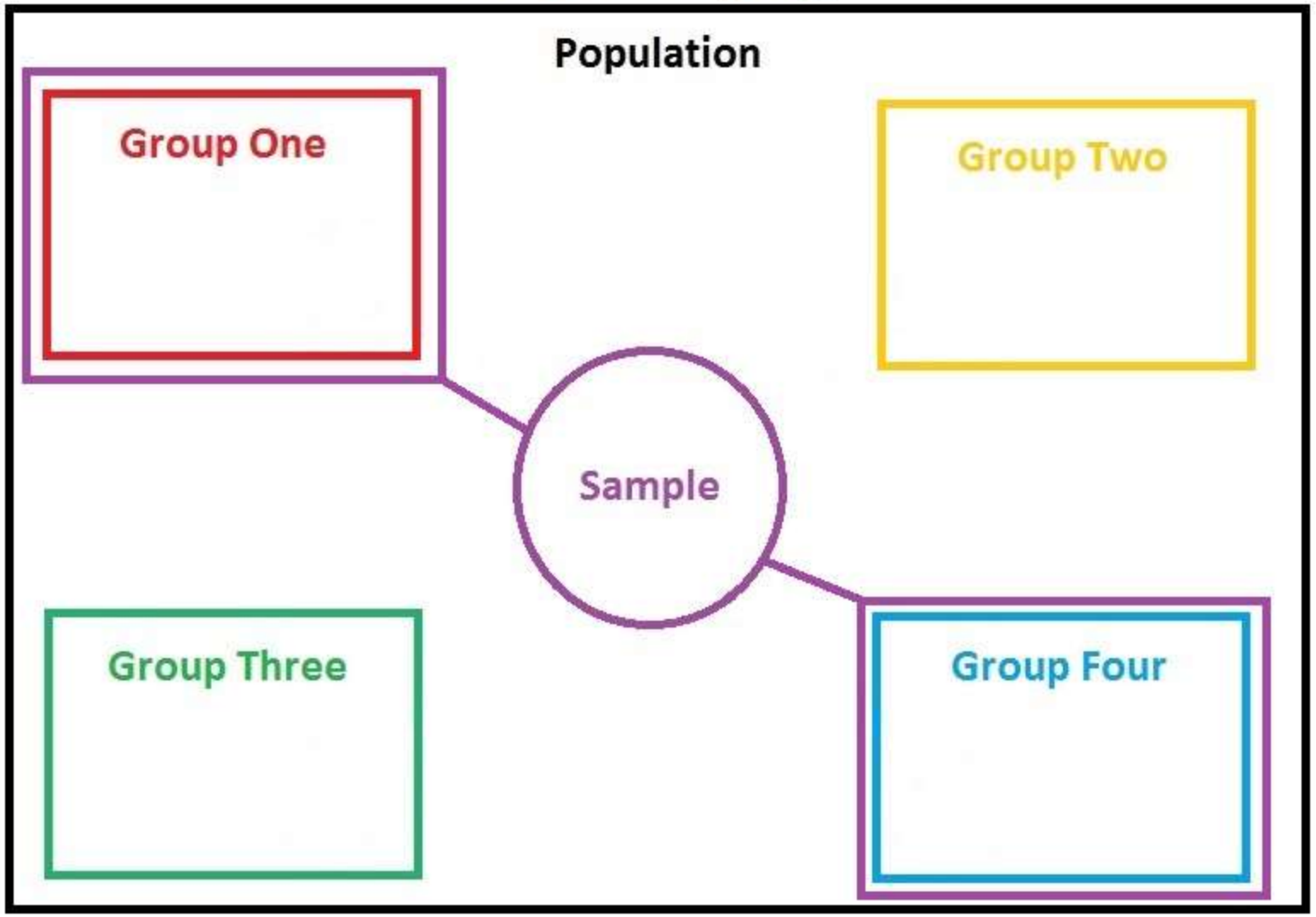
$$E = \frac{V\left(\overline{y_n}\right)_{SRSWOR}}{V\left(\overline{y_{sys}}\right)} = \frac{\frac{nk-n}{n^2K} S^2}{\frac{nK-1}{nK} \frac{S^2}{n} \left(1+(n-1)\rho\right)} = \frac{nk-n}{(nK-1)\left(1+(n-1)\rho\right)}$$

If

$$\begin{cases} E>1 \Rightarrow \rho > \dfrac{1}{nK-1} \\ or, \rho < \dfrac{-1}{nK-1} \end{cases}$$

Thus, the systematic sampling is more efficient than SRSWOR if $\rho < \frac{-1}{nK-1}$ and systematic sampling is less efficient than SRSWOR if $\rho > \frac{-1}{nK-1}$.

Cluster Sampling



Cluster Sampling:

Suppose that a population consists of N clusters and each cluster have M elementary units.The procedure in which a sample of n clusters is selected from the population by the method of SRS is knows as Cluster sampling (equal size clusters).

Let Y be the characteristic under study and $y_{ij}(i=1,2,3.....,N;\ j=1,2,3....,M)$ be the j^{th} observation in the i^{th}, Now we have.

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

$\bar{\bar{Y}}_{N.}=\frac{1}{N}\sum_{i=1}^{N}\bar{y}_{i.}$: mean of the cluster means in the population.

$\bar{y}_{..}=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}$: mean per element in the population.

$\bar{y}_{..}=\bar{\bar{y}}_{N.}$: (only in case of equal size cluster).

$S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(y_{ij} - \bar{y}_{i.}\right)^2$: Mean square between the elements in the i^{th} cluster.

$S_w^2 = \frac{1}{N}\sum_{i=1}^{N} S_i^2$: Mean square within cluster.

$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{y}_{i.} - \bar{\bar{y}}_{N.}\right)^2$: Mean square between cluster means

$\bar{\bar{y}}_{n.} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}$: Mean of cluster mean in the sample.

Estimation of population mean:

Let us define the estimator of population mean $\bar{\bar{y}}_{..}$ as-

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

$$E\left(\bar{\bar{y}}_{n.}\right) = \bar{\bar{Y}}_{N.}$$

$\bar{\bar{y}}_{n.}$ is an unbiased estimator of population mean $\bar{\bar{y}}_{..}$.

Variance of $\bar{\bar{y}}_{..}$:-

$$V\left(\bar{\bar{y}}_{n.}\right) = \frac{N-n}{Nn} S_b^2$$

$$where, S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{y}_{i.} - \bar{\bar{y}}_{N.}\right)^2$$

Comparison of cluster sampling w.r.t. SRSWOR:-

we have ,

$$V\left(\overline{y_{nM}}\right)_{SRSWOR} = \frac{NM - nM}{NMnM} S^2$$

$$= \frac{N-n}{NnM} S^2 \text{, and } V\left(\bar{\bar{y}}_{n.}\right) = \frac{N-n}{Nn} S_b^2$$

The relative efficiency of cluster sampling w.r.t. SRSWOR is given by-

$$E = \frac{V\left(\overline{y_{nM}}\right)_{SRSWOR}}{V\left(\bar{\bar{y}}_{n.}\right)} = \frac{1}{M} \frac{S^2}{S_b^2}$$

From the above, we conclude that the relative efficiency of cluster sampling increases if mean square of cluster increase.

Now , we have

$$S^2 = \frac{1}{NM-1} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(y_{ij} - \bar{\bar{y}}_{..}\right)^2$$; by expanding this equation we can get,

$$MS_b^2 = \frac{1}{N-1}\left[(NM-1)S^2 - N(M-1)\bar{S}_w^2\right]$$

$$E = \frac{1}{M}\frac{S^2}{S_b^2}$$

$$\therefore E = \frac{(N-1)S^2}{\left[(NM-1)S^2 - N(M-1)\bar{S}_w^2\right]S_b^2}$$

From the above, we have seen that the relative efficiency of cluster sampling will increase with increase of mean square within cluster.

Efficiency of cluster sampling in terms of intra-class correlation:

$$\rho = \frac{\left(\frac{N-1}{N}\right)S_b^2 - \frac{\bar{S}_w^2}{M}}{\frac{(NM-1)}{Nm}S^2}$$

$$\Rightarrow S_b^2 = \frac{(NM-1)[1+(m-1)\rho]S^2}{M^2(N-1)}$$

$$\because V\left(\bar{\bar{y}}_{n.}\right)=\frac{N-n}{Nn}S_b^2$$

$$\therefore V\left(\bar{\bar{y}}_{n.}\right)=\left[1+(M-1)\rho\right]\frac{S^2}{nM}$$;if N is sufficient large

Relative efficiency:-

Relative efficiency of cluster sampling w.r.t. SRSWOR in terms of intra-class correlation-

$$E=\frac{V\left(\bar{y}_{nM}\right)_{SRSWOR}}{V\left(\bar{\bar{y}}_{n.}\right)}=\frac{S^2}{MS_b^2}$$

$$E=\frac{1}{1+(m-1)\rho}$$

Thank you