UNIT - V

SYSTEMATIC SAMPLING

Systematic sampling is one of the commonly used Probability Sampling design. It is operationally more convenient than Simple random Sampling and at the same time ensures for each unit equal probability of exclusion in the sample. Systematic sampling technique consist of selecting the first unit at random and the remaining units are selected automatically in a pre-define manner.

Systematic Sampling is more suitable when a complete list of units in the population is available.

In forest Surveys, this technique is used to estimate the total volume of timber and in fisheries, it is used to estimate the total catch of

Definition :

Consider the population consisting of 'N' units Numbered 1 to N and arranged in Some order. To select a random sample of size n, first we determine the number k. where k is taken to be the integer nearest to N/n. The next step is to select a number 'i' at random from 1 to k and the unit corresponding to this number is selected. After this every k^{th} unit is selected till the desired sample is drawn. A sample Selected by this procedure is known as linear Systematic Sampling.

Ex : Suppose that we want to select a sample of size 5, from the population of size 40. Here

$$k = \frac{N}{n} = \frac{40}{5} = 8.$$

Now we select a random number from 1 to 8 say 4

The members in the sample are having the serial no/-

4, 12, 20, 28, 36.

Advantages :

(i) It is easier to draw a sample and often easier to execute without mistakes because only one random number is require to select the sample and it distributes the sample more evenly over the listed population

(ii) If the population doesnot possess periodicities, systematic sampling provides estimates as compared to any other sampling designs.

Disadvantages :

(i) If the population contains the periodic type of variation and the sampling ratio coincide with the period, then we get a biased sample and the inferences drawn from the sample leads to strong bias (error)

(ii) It has the drawback of estimating the sampling variance of the estimate with a single sample.

Circular Systematic Sampling :

If $N \neq nk$, and every k^{th} unit be included in a circular manner till the whale list is exhausted, it will be called circular systematic sampling.

For ex. with $N = 17$, $n = 3$, $k = \frac{17}{3} = 5.67 \simeq 6$

The possible systematic samples are as follows

1)	1	7	13
2)	2	8	14
3)	3	9	15
4)	4	10	16
5)	5	11	17
6)	6	12	

From the above table we observe the first 5 samples are of size 3 and the last one is of size 2.

To overcome this difficulty, Lahiri propose a method which provides both a constant sample size and an unbiased estimate of the population mean. This method is known as circular systematic sampling. In this method a units in the population are arranged in the form of a circle and then determine k the nearest integer to N/n.

Select a random number 1 to k and taking every k^{th} unit thereafter going around the circle till the desired 'n' units have been choosen.

Ex :

Suppose we want to take a sample of size 3 from the population of size 11.

Here $k = \frac{N}{n} = {}^{11}/_{3} = 3.67 \simeq 4$

If the random number selected is 6, then we include the units in the sample corresponding to the numbers 6, 10, 3

Advantages

i) It provides every unit has an equal probability of inclusion in the sample.

ii) It provides a constant sample size so that we can obtain an unbiased estimate of the population mean

Notations :

Let y_{ij} denote the observation on the j^{th} unit of the i^{th} sample $(i = 1, 2 \ldots k)$, $j = 1, 2 \ldots n$

Sample mean of i^{th} systematic sample

$$\bar{y}_{i.} = \frac{1}{n} \sum_{j=1}^{n} y_{ij}, \quad i = 1, 2 \ldots k$$

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

Sample Variance $s_{i.}^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2$

Population mean $\bar{Y}_{..} = \bar{Y} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}}{N}$

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

Population mean square $S^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{Y}_{..})^2$

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

S^2_{wsy} = Mean square among the units lies within the Same systematic Sample

$$= \frac{1}{k} \sum_{i=1}^{k} s_i^2$$

$$= \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2$$

ρ_w = correlation coefficient between the units that are in the same Systematic sample.

$$= \frac{\sum_{i=1}^{k} \sum_{j \neq j'=1}^{n} (y_{ij} - \bar{y}_{..})(y_{ij'} - \bar{y}_{..})}{(n-1)(N-1)S^2}$$

Theorem - 1. :

In systematic Sampling the sample mean $\bar{y}_{sys}$ is an unbiased estimate of the Population mean.

Proof :

Consider, $E(\bar{y}_{i.}) = E(\bar{y}_{sys})$

$$= \sum_{i=1}^{k} \bar{y}_{i.} \text{ Prob}\{\bar{y}_{i.}\}$$

Since there are k systematic samples and they all are having equal chance for selection.

$$\therefore \text{Prob}\{\bar{y}_{i.}\} = \frac{1}{k} \quad \forall i$$

$$\therefore E(\bar{y}_{sys}) = \sum_{i=1}^{k} \bar{y}_{i.} \times \frac{1}{k}$$

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

$$= \bar{Y}$$

Theorem - 2 :

The variance of the mean of the systematic sample is given by

$$V(\bar{y}_{sys}) = \frac{N-1}{N} s^2 - \frac{k(n-1)}{N} s^2_{wsy}$$

Proof :

consider $s^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$

Add and substract $\bar{y}_{i.}$

$$(N-1) s^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} \left[(y_{ij} - \bar{y}_{i.})^2 + (\bar{y}_{i.} - \bar{y}_{..})^2 + 2 (y_{ij} - \bar{y}_{i.}) (\bar{y}_{i.} - \bar{y}_{..}) \right]$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 + n \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})^2 + 0 \rightarrow (1)$$

From the definition of s^2_{wsy} we can write

$$k(n-1) s^2_{wsy} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 \longrightarrow (2)$$

Also from the definition of $V(\bar{y}_{sys})$

$$k \cdot V(\bar{y}_{sys}) = \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})^2 \longrightarrow (3)$$

Using (2) & (3), eqn (1) can be written as

$$(N-1) s^2 = k(n-1) s^2_{wsy} + nk \, V(\bar{y}_{sys})$$

$$N \, V(\bar{y}_{sys}) = (N-1) s^2 - k(n-1) s^2_{wsy} \quad \because nk = N$$

$$V(\bar{y}_{sys}) = \frac{N-1}{N} s^2 - \frac{k(n-1)}{N} s^2_{wsy}$$

Comparison of systematic Sampling with SRSWOR

The mean of the systematic sample is more precise than SRSWOR estimate if and only if $S^2_{wsy} > S^2$

(or)

Systematic sampling is more precise than SRSWOR if variance within the systematic sample must be larger than the population variance as a whole.

Proof:

Consider

$$V(\bar{y}_{sys}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy} \longrightarrow (1)$$

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

From (1) & (2) we have

$$V(\bar{y}_{sys}) - V(\bar{y})_{SRSWOR} = \left[\frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy}\right] - \left[\frac{N-n}{N} \frac{S^2}{n}\right]$$

$$= \frac{S^2}{N}\left[(N-1) - \left(\frac{N-n}{n}\right)\right] - \frac{k(n-1)}{N} S^2_{wsy}$$

$$= \frac{S^2}{N}\left[\frac{n(N-1) - N + n}{n}\right] - \frac{k(n-1)}{N} S^2_{wsy}$$

$$= \frac{S^2}{N}\left[\frac{nN - \cancel{n} - N + \cancel{n}}{n}\right] - \frac{k(n-1)}{N} S^2_{wsy}$$

$$= \frac{S^2}{\cancel{N}}\left[\frac{\cancel{N}(n-1)}{n}\right] - \frac{k(n-1)}{N} S^2_{wsy}$$

$$= S^2\left(\frac{n-1}{n}\right) - \frac{\cancel{k}(n-1)}{n\cancel{k}} S^2_{wsy} = \frac{n-1}{n}\left[S^2 - S^2_{wsy}\right]$$

Systematic sampling is more precise than SRSWOR only when $V(\bar{y}_{sys}) < V(\bar{y})_{SRSWOR}$

(ie) $V(\bar{y}_{sys}) - V(\bar{y})_{SRSWOR} < 0$

(ie) $\left(\frac{n-1}{n}\right)\left[S^2 - S^2_{wsy}\right] < 0$

$$\left[S^2 - S^2_{wsy}\right] < 0$$

$$S^2 < S^2_{wsy}$$

(or) $S^2_{wsy} > S^2$

Systematic sampling is more precise than simple random sampling only when the units within the systematic sample are hetrogeneous

Theorem :

In systematic sampling $V(\bar{y}_{sys})$ is given by

$$V(\bar{y}_{sys}) = \frac{N-1}{N}\frac{S^2}{n}\left[1 + (n-1)\rho_w\right]$$

where ρ_w is the intra-class correlation coefficient between the pair of units in the same systematic sample

Proof :

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

$$= \frac{1}{k}\sum_{i=1}^{k}\left[\frac{\sum_{j=1}^{n} y_{ij}}{n} - \frac{\sum_{j=1}^{n}\bar{y}_{..}}{n}\right]^2$$

By def of $\bar{y}_{i.}$ and $\bar{y}_{..} = \frac{n\bar{y}_{..}}{n}$

$$= \frac{1}{n^2 k}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})\right]^2$$

$$V(\bar{y}_{sys}) = \frac{1}{n^2 k} \sum_{i=1}^{k} \left[\sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 + \sum_{j \neq j'=1}^{n} (y_{ij} - \bar{y}_{..})(y_{ij'} - \bar{y}_{..}) \right.$$

$$\because (\Sigma a_i)^2 = \Sigma a_i^2 + \Sigma a_i a_j$$

$$n^2 k \, V(\bar{y}_{sys}) = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 + \sum_{i=1}^{k} \sum_{j \neq j'=1}^{n} (y_{ij} - \bar{y}_{..})(y_{ij'} - \bar{y}_{..}) \rightarrow (1)$$

From the definition of S^2, we have

$$\sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 = (N-1) S^2 \rightarrow (2)$$

From the definition of ρ_w, we have

$$\sum_{i=1}^{k} \sum_{j \neq j'=1}^{n} (y_{ij} - \bar{y}_{..})(y_{ij'} - \bar{y}_{..}) = (N-1)(n-1) S^2 \rho_w \rightarrow (3)$$

using (2) & (3), (1) can be written as

$$n^2 k \; V(\bar{y}_{sys}) = (N-1) S^2 + (N-1)(n-1) S^2 \rho_w$$

$$nN \; V(\bar{y}_{sys}) = (N-1) S^2 \left[1 + (n-1) \rho_w \right] \quad \because n^2 k = n$$

$$V(\bar{y}_{sys}) = \frac{N-1}{N} \cdot \frac{S^2}{n} \left[1 + (n-1) \rho_w \right]$$

Hence the proof

Note :

1. when $\rho_w = 0$, then $V(\bar{y}_{sys}) = \frac{N-1}{N} \frac{S^2}{n}$

   (ie) $V(\bar{y}_{sys}) = V(\bar{y})_{SRSWOR}$

2. If $\rho_w > 0$ (+ve correlation) then SRS is better than systematic sampling

   If $\rho_w < 0$ (-ve correlation) then systematic sampling is better than SRS

Efficiency of Systematic Sampling over SRSWOR :

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

$$V(\bar{y})_{sys} = \frac{N-1}{N} \frac{S^2}{n} \left[1 + (n-1) \rho_w \right]$$

Relative Efficiency is given by

$$E = \frac{V(\bar{y})_{SRSWOR}}{V(\bar{y})_{sys}}$$

$$E = \frac{\frac{N-n}{N} \cdot \frac{S^2}{n}}{\frac{N-1}{N} \cdot \frac{S^2}{n} \left[1 + (n-1) \rho_w \right]}$$

$$= \frac{N-n}{(N-1)\left[1 + (n-1) \rho_w \right]}$$

Obviously this depends on the value of ρ_w

$$E > 1 \Rightarrow \frac{N-n}{N-1 \left[1 + (n-1) \rho_w \right]} > 1$$

$$N-n > (N-1) \left[1 + (n-1) \rho_w \right]$$

$$\frac{N-n}{N-1} > 1 + (n-1) \rho_w$$

$$\left(\frac{N-n}{N-1} \right) - 1 > (n-1) \rho_w$$

$$\frac{N - n - N + 1}{N-1} > (n-1) \rho_w$$

$$- \frac{(n-1)}{(N-1)} > (n-1) \rho_w$$

$$\rho_w < -\frac{1}{N-1}$$

Thus systematic sampling would be more efficient as compared with SRSWOR if

$$\rho_w < -\frac{1}{N-1}$$

On the other hand, SRSWOR would be superior to systematic sampling if

$$\rho_w > -\frac{1}{N-1}$$

Systematic Sampling Viewed as Stratified Random Sampling :

Let us suppose that the population of size $N = nk$ units divided into n strata each consisting of k units. The first stratum consisting of first k units namely $1, 2 \ldots k$ and the second stratum consisting of next units namely $k+1, k+2, \ldots 2k$ and so on.

The systematic sampling can be treated as the stratified random sampling by selecting one unit from each stratum. The sample composition of nk units represented in the following table

Sys. sample No	Stratum Number 1	2	...	j	...	n	Sys. sample mean
1	y_{11}	y_{12}	..	y_{1j}	.	y_{1n}	$\bar{y}_{1.}$
2	y_{21}	y_{22}	...	y_{2j}	..	y_{2n}	$\bar{y}_{2.}$
$\vdots$							
i	y_{i1}	y_{i2}	...	y_{ij}	..	y_{in}	$\bar{y}_{i.}$
$\vdots$							
k	y_{k1}	y_{k2}		y_{kj}	..	y_{kn}	$\bar{y}_{k.}$

Stratum Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	... $\bar{y}_{.j}$	...	$\bar{y}_{.n}$	$\bar{y}_{..} = \bar{Y}$

Mean of j^{th} stratum $\bar{y}_{.j} = \frac{\sum_{i=1}^{k} y_{ij}}{k}$

Stratum Mean square $S_j^2 = \frac{\sum_{i=1}^{k} (y_{ij} - \bar{y}_{.j})^2}{k-1}$

S^2_{wst} = Combined Mean square between the units within the stratum

$$= \frac{1}{n} \sum_{j=1}^{n} S_j^2 = \frac{1}{n(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{.j})^2$$

ρ_{wst} = correlation between deviation from stratum means of pair of observation that are in the same systematic sampling.

$$= \frac{\sum_{i=1}^{k} \sum_{j \neq j'=1}^{n} (y_{ij} - \bar{y}_{.j})(y_{ij'} - \bar{y}_{.j'})}{n(n-1)(k-1) S^2_{wst}}$$

Theorem :

In systematic sampling.

$$V(\bar{y}_{sys}) = \frac{k-1}{nk} S^2_{wst} \left[1 + (n-1) \rho_{wst} \right]$$

Proof :

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

$$= \frac{1}{k} \sum_{i=1}^{k} \left[\frac{\sum_{j=1}^{n} y_{ij}}{n} - \frac{1}{n} \sum_{j=1}^{n} \bar{y}_{.j} \right]^2$$

$$= \frac{1}{n^2 k} \sum_{i=1}^{k} \left[\sum_{i=1}^{n} (y_{ij} - \bar{y}_{.j}) \right]^2 \qquad \because \bar{y}_{..} = \frac{1}{n} \sum_{j=1}^{n} \bar{y}_{.j}$$

$$n^2 k\, V(\bar{y}_{sys}) = \sum_{i=1}^{k} \left[\sum_{j=1}^{n} (y_{ij} - \bar{y}_{.j})^2 + \sum_{j \neq j'=1}^{n} (y_{ij} - \bar{y}_{.j})(y_{ij'} - \bar{y}_{.j'}) \right]$$

$$n^2 k\, V(\bar{y}_{sys}) = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{.j})^2 + \sum_{i=1}^{k} \sum_{j \neq j'=1}^{n} (y_{ij} - \bar{y}_{.j})(\bar{y}_{ij'} - \bar{y}_{.j'}) \longrightarrow (1)$$

From the definition of S^2_{wst}, we have

$$n(k-1) S^2_{wst} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{.j})^2 \longrightarrow (2)$$

From the definition of ρ_{wst}, we have

$$n(n-1)(k-1) S^2_{wst}\, \rho_{wst} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{.j})(y_{ij'} - \bar{y}_{.j'}) \longrightarrow (3)$$

using (2) and (3), (1) can be written as

$$n^2 k\, V(\bar{y}_{sys}) = n(k-1) S^2_{wst} + n(n-1)(k-1) S^2_{wst}\, \rho_{wst}$$

$$n^2 k\, V(\bar{y}_{sys}) = n(k-1) S^2_{wst} \left[1 + (n-1) \rho_{wst} \right]$$

$$V(\bar{y}_{sys}) = \frac{k-1}{nk} S^2_{wst} \left[1 + (n-1) \rho_{wst} \right]$$

Comparison of systematic Sampling with stratified random sampling :

By dividing the population of $N = nk$ units into 'n' strata each consisting of k units, we have

$$L = n$$

$$N_h = k$$

$$n_h = 1$$

$$f_h = \frac{n_h}{N_h} = \frac{1}{k}$$

$$W_h = \frac{N_h}{N} = \frac{k}{nk} = \frac{1}{n}$$

$$S_h^2 = S_j^2$$

we know that

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} (1-f_h) \frac{W_h^2 S_h^2}{n_h}$$

$$= \sum_{j=1}^{n} (1-\frac{1}{k}) \frac{(1/n)^2 S_j^2}{1}$$

$$V(\bar{y}_{st}) = \frac{k-1}{n^2 k} \sum_{j=1}^{n} S_j^2$$

NATIONAL SAMPLE SURVEY ORGANISATION (NSSO) :

The Government needs statistical information for planning and development of the country. The data relating to the subject under study can be collected through census or sample survey. The NSS is a central organisation which collects the data, compiles, analyse and publish the reports.

Prof. P.C. Mahalonabis prepared a scheme for national sample survey and it was approved by the Ministry of Finance in 1950.

At the initial stage both the directorates of NSS and ISI join together in the field work design, data processing and report writing. In 1971, NSS organisation was setup to bring all the works of NSS under the unitied organisation. The work of collecting statistical data started on 1st October, 1950 and completed on March 1951. Since then the NSSO functioning on continual basis.

The functioning of NSSO are governed by a council which consists of a non-official chairman, four non-official and five official economists, four director of different divisions. NSSO consists of four divisions. They are

(i) Survey, Design and Research

(ii) Field operation

(iii) Data Processing and

(iv) Economic Analysis.

Functions of NSSO:

The three main functions of NSSO are as follows

(i) Collection of data on socio-economic conditions, production on small-scale house-hold enterprise, consumption etc on a continual basis in a comprehensive manner for the whole country.

(ii) Collection of data relating to the industrial sector of the country.

(iii) Supervision of the surveys conducted by States in agricultural sector through their own agencies and also giving guidance to them.

Methods of collecting Information:

The surveys are conducted in the form of rounds, each round covering some topics of interest. The rounds of NSS have been varying duration from 3-8 months. During the rounds the importance given to collection of statistical information relating to national income, small scale house hold enterprise, consumer expenditure, employment and un-employment distribution of land holdings population etc.

Method of selecting sampling unit

The commonly used procedure for selecting the sampling units in stratification. The detailed procedure for selecting the sampling units is described below.

(i) Stratification

All the states were divided into 160 strata on the basis of geographical position. Then four sub-groups were formed as follows.

a) Sub-group with a population of 1-499

b) " 500-999

c) " 1000-1999

d) " 2000 & above

(ii) Selection of villages

Within each sub-group villages were selected on the basis of the information available. In this selection, the following method was adopted.

(i) Equal probability of selection of each village

(ii) Probability proportional to village area

(iii) Probability proportional to village population

Selection of House-holds:

The step involved in the selection of house-holds from the village were as follows

(i) A complete list of all house-holds was made in the selected village

(ii) A random sample of 80 house-holds were selected and they were classified into agricultural and non-agricultural.