Cluster Sampling

In Random Sampling, the popln has been divided into a finite no. of distinct and identifiable units defined as sampling Units. The Smallest unit into which the popln can be eliminated is called an element of the popln. A group of such elements is known as Cluster. when the Sampling Units in the cluster the Procedure is called cluster sampling. If the entire area entaing popln under study is divided in to smaller segment and each element in the popln belongs to one and only one segment, the procedure is called area sampling.

Identification and location of element requires Considerable time area an element has been located, the time taken for surveying a few neighboring elements is small, Thus the main fun. in cluster Sampling is to specify the clusters or to divided the popln into oppropiriate clusters. The clusters are generaly made up of neighboring elements and the elements with in cluster 10 to have similar chorcterstic. as the sample rule, the no. of. elements in the cluster should be small, and no. of. cluster should be large.

After dividing the popln into specified clusters, the required no. of clusters and can be selected either by equal or unequal prob. of selection. all the elements in selected in clusters are enumerated.

For a given no. of sampling units, cluster sampling is more convenient and less cost the advantages of cluster sampling

Advantages:-

1) Collection of data for neighbouring elements is faster, cheaper, easier and operationly more convenient that observing unit spread over the region.

2) It is less costly than SRS due to the saving of time in journey, identification, etc.

3) when the sampling farme of elements may not be readily available.

disadvantages:-

1) Even if the Sample farme is available it could be expensive to conduct an enquary on a SRS of elements.

2) From the point of view of statistical efficiency, Cluster sampling is generaly less efficient then SRS due to the tendency of units in cluster. The efficience of cluster sampling is likely to decrease with increase in cluster size.

For a given sample size, the smaller sampling unit will bring more precise results than the large Sampling unit. The last efficiency may be balanced by the reduction in the cost, the efficiency per unit

Cost may be more in cluster sampling than in SRS

The selection of cluster can be random or by first selecting a unit call a key unit at random and then randomly taking the required no. of neighbouring clusters to estimate the mean production, the cluster of three villages formed by first selecting a key village in the random and then taking two more villages from a block of some specified area.

Notations:-

Suppose the popln consits of N clusters, each of M elements and that a sample of n clusters is drawn by the method of SRS.

N = no. of. clusters in the popln.

n = no. of. clusters in the sample.

M = no. of. elements in the cluster.

y_{ij} = the value of the characteristic under study for the j^{th} elements $(j=1,2...M)$ in the i^{th} cluster $(i=1,2...N)$

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

$\bar{y}_n = \sum_{i} \frac{\bar{y}_{i.}}{n}$ = the mean of cluster means in a sample of n cluster.

$\bar{X}_N = \sum_{i}^{N} \bar{Y}_{i.}/N$ = the mean of cluster means in the Popln.

$\bar{Y} = \sum_{i}^{N} \sum_{j}^{M} y_{ij}/NM$ = the mean per element in the popln.

$S_i^2 = \sum_{j}^{M} (y_{ij} - \bar{y}_{i.})^2/(M-1)$ = the mean square b/w

elements within the i^{th} cluster ($i = 1, 2, \ldots N$)

$S_w^2 = \sum_i^N S_i^2/N$ = the mean square within cluster, (w for within)

$S_b^2 = \sum_i^N (\bar{y}_{i.} - \bar{y}_N)^2/N-1$: the mean square blw clusters means in the popln (b for btw)

$S^2 = \sum_i^N \sum_j^M (y_{ij} - \bar{y})^2/(NM-1)$ = the mean square blw elements in the popln.

$$\rho = \frac{E(y_{ij} - \bar{y})(y_{ik} - \bar{y})}{E(y_{ij} - \bar{y})^2}$$

$$= \frac{\sum_i^N \sum_j^M \sum_{j \neq k} (y_{ij} - \bar{y})(y_{ik} - \bar{y})}{(M-1)(NM-1)S^2}$$

the intra cluster correlation co-efficient blw elements within clusters.

Equal Cluster Sampling :-

Estimator of mean and its variance :-

No new principles are involved in making estimator when a prob. sample of 'n' equal sized clusters has been taken and each cluster in enumerated completely since the clusters are of equal size, it is clear that $\bar{y}_N = \bar{y}$. For the sampling variance of the estimator $\bar{y}_n$ in the form developed by Hansen and Hurwitz (1942) we shall begin by Proving the following theorem.

Theorem:

In simple random sampling, wor of n clusters each containing M elements from a population of N clusters, the sample mean $\bar{y}_n$ is an unbiased estimator of $\bar{Y}$ and its variance is given by

$$V(\bar{y}_n) = \frac{(1-f)}{n} S_b^2$$

$$\cong \frac{(1-f)}{n} S_M^2 [1+(M-1)\rho]$$

where ρ is the intra cluster correlation coefficien

Proof

W.k.T, $\bar{y}_n = \Sigma y_i / n$

$$E(\bar{y}_n) = \frac{1}{n} \Sigma E(y_i) \qquad \because E(\bar{y}_i) = \Sigma P_i y_i$$

$$E(\bar{y}_n) = \sum_{i=1}^{N} P_i y_i / N$$

$$= \sum_{i=1}^{N} \frac{y_i}{N}$$

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

Thus, $\bar{y}_n$ is an unbiased estimator of $\bar{Y}$.

Also we know that SRSWOR $V(\bar{y}) = \frac{1-f}{n} \cdot S^2$.

with this result we can write

$$V(\bar{y}_n) = \frac{(1-f)}{n} S_b^2$$

$$= \frac{(1-f)}{n} \sum_{i=1}^{N} \frac{(y_i - \bar{Y}_N)^2}{N-1} \quad \leftarrow (1)$$

After substituting the value of

$$\rho = \frac{\Sigma\Sigma\Sigma (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{(M-1)(NM-1) \cdot S^2}$$

in the above relation, we get

$$\sum_{i=1}^{N} (\bar{y}_i - \bar{Y}_N)^2 = \frac{1}{M^2} \sum_{i}^{N} \sum_{j}^{M} (\bar{y}_{ij} - \bar{y})^2 + \frac{1}{M^2} \sum_{i}^{N} \sum_{j}^{M} \sum_{j \neq k} (y_{ij} - \bar{y})(y_{ik} - \bar{y})$$

Sub in ①, we get

$$V(\bar{y}_n) = \left(\frac{1-f}{n}\right) \cdot \frac{NM-1}{M^2(N-1)} S^2 [1 + (M-1)\rho]$$

$$\simeq \left(\frac{1-f}{nM}\right) S^2 [1 + (M-1)\rho] \quad \text{for large } N$$

It is shown that the variance in cluster sampling depends on the no. of cluster in the ~~cluster~~ sample, the size of the cluster, the inter cluster correlation coefficient ρ and the variance S^2. If $M=1$ it gives the sampling variance of a simple random sample of nM elements taken individually.

Corollary 1:

In simple random sampling wor of n clusters each containing M elements from a population of N clusters. the population total is estimated unbiasedly by

$$\bar{Y}_c = NM \bar{y}_n$$

with its sampling variance

$$V(\hat{Y}_c) = (1-f) \frac{N^2}{n} M^2 S_b^2$$

$$\simeq MN^2 \frac{(1-f)}{n} S^2 [1 + (M-1)\rho]$$

Corollary 2:

Instead of sampling in clusters, if a simple random sample of nM elements be directly taken from the population of NM elements

the variance of the mean per element would be

$$V(\bar{y}_n) = \frac{(1-f)}{nM} S^2$$

Corollary 3:

In simple random sampling w.o.r. an estimator of variance of $\bar{y}_n$ is given by

$$v(\bar{y}_n) = \frac{(1-f)}{n} s_b^2$$

where, $s_b^2 = \sum_i \frac{(\bar{y}_i - \bar{y}_n)^2}{(n-1)}$

Corollary 4:

In simple random sampling w.o.r the cluster mean $\bar{y}_n$ is an unbiased estimator of $\bar{Y}$ and its variance is given by

$$V(\bar{y}_n) = \frac{S_b^2}{n}$$

Relative Efficiency of Cluster Sampling:

In sampling nM elements from the population by the simple random sampling; the variance of the sample mean $\bar{y}$ is given by

$$V(\bar{y}) = \frac{(1-f) S^2}{nM}$$

The relative efficiency of cluster sampling compared with simple random sample is

$$\text{Rel Efficiency} = \frac{V_{SR}(\bar{y})}{V_c(\bar{y}_n)} = \frac{S^2}{MS_b^2}$$

This shows that the efficiency of cluster sampling increases as the mean square between clusters decreases.

Also $(N-1)\,M\,S_b^2 = (NM-1)\,S^2 - N(M-1)\,S_w^2$

Therefore, relative Efficiency will increases with increase in the mean square within clusters. Another way to express relative efficiency is to make use of the concept of intra cluster correlation coefficient.

For large N, the relative efficiency of cluster sampling interms of intracluster coefficient ρ is given by,

$$\text{Rel Efficiency } (E) = [1+(M-1)\rho]^{-1}$$

The Efficiency can easily calculated by estimating the value of ρ from the sample. An estimator of ρ can be written as

$$\hat{\rho} = \frac{(n-1)\,M\,s_b^2 - n s_w^2}{(n-1)\,M\,s_b - n(M-1)\,s_w^2}$$

where, $s_w^2 = \sum_i^n \sum_j^M \frac{(y_{ij} - \bar{y}_i)^2}{n(M-1)}$

Thus for large N, an estimate of the relative efficiency of the cluster sampling can be written as

$$\text{Est. Rel Efficiency } (e) = \frac{1}{M} + \frac{(M-1)\,S_w^2}{M^2 S_b^2}$$

Accordingly ρ can be estimated by

$$\hat{\rho} = \frac{(1-e)}{(M-1)e}$$

For a sample of n clusters, the results in the form of analysis of variance can be written in the form

Analysis of variance of sample of n clusters

Source of variation	Degrees of freedom	Mean Square
Between clusters	$n-1$	$MS_b^2 = \frac{\sum_{i}^{n} (\bar{y}_{i.} - \bar{y}_{n})^2}{(n-1)}$
within clusters	$n(M-1)$	$S_w^2 = \frac{\sum_{i}^{n} \sum_{j}^{M} (\bar{y}_{ij} - \bar{y}_{i.})^2}{n(M-1)}$
Total sample	$n(M-1)$	$S^2 = \frac{\sum_{i}^{n} \sum_{j}^{M} (\bar{y}_{ij} - \bar{y}_{n})^2}{(nM-1)}$ $= \frac{(n-1)MS_b + n(M-1)S_w^2}{(nM-1)}$

It should be noted clearly that in a random sample of n clusters, s_b^2 and s_w^2 will provide unbiased estimates of S_b^2 and S_w^2 respectively, while s^2 will not be an unbiased estimate of S^2. The reason is that a sample of nM elements is not taken randomly from the population of NM elements.

A unbiased estimate may be obtained easily by substituting the values in the relation, when

$$\hat{\sigma}^2 = \frac{(N-1)\,M s_b^2 + N(M-1)\,s_w^2}{(NM-1)}$$

Optimum cluster size

- For a given sample size the sampling variance increase with cluster size and decreases with increasing number of clusters.
- On otherhand, the cost decreade with the cluster size and increases with the number of clusters.
- It is necessary to determine a balancing Point by finding out the optimum cluster size and the number of cluster in the sample which can minimise the sampling variance for a given cost or alternatively, minimize the cost for a fixed variance.
- we shall discuss the problem of optimum size of cluster for which maximum precision is attained with a given cost.

The cost of a survey, apart from overhead cost, will be made up of two components.

i) Cost due to expenses in enumerating the elements in the sample and in travelling within cluster, which is proportional to the number of elements in the sample.

ii) Cost due to expenses on the travelling between clusters, it has been shown empirically that the expected value of minimum distance between n points located at random is proportional to $n^{1/2}$.

The Cost of the Survey can expressed as

$$C = c_1 nM + c_2 n^{1/2}$$

where,

c_1 = Cost of enumerating an element including the cost of travel between unit within the cluster.

c_2 = Cost per unit distance travelled between cluster.

Variance of estimator $\bar{y}_n$ is based on a sample of n clusters of size m each,

$$V(\bar{y}_n) = \frac{(1-f) S_b^2}{n}$$

S_b^2 can be obtained if we know

i) variance 'S^2' between all elements in the population

ii) The variance 'S_w^2' within cluster.

Hence, an approach has always been made to predict S_w^2 as it affected by the cluster size while S^2 remain unchanged by it.

On the basis of several agriculture survey, they observed that S_w^2 appears to bear a relation with m which can be written by the empirical relation

$$S_w^2 = aM^b \quad (b > 0)$$

where, a and b are positive constants are find by survey from ANOVA

$$S_b^2 = \frac{(MN-1)S^2 - N(M-1)S_w^2}{M(N-1)}$$

Substituting value of S_w^2

$$S_b^2 = S^2 - (M-1) a M^{b-1}$$

$$V(\bar{y}_n) = \frac{1}{n}\left[S^2 - (M-1)\,a M^{b-1}\right]$$

To calculate the values of n and M by minimising V for given C. To minimise $\phi = C + \lambda V$

Differentiate ϕ w.r.t n and M and equate to zero

$$n = \left[\frac{-C_2 + (C_2^2 + 4C_1 CM)^{1/2}}{2C_1 M}\right]^2$$

'M' can be obtained from the equation.

$$\frac{aM^{b-1}[bM - (b-1)]}{S^2 - (b-1)\,aM^{b-1}} = 1 - \left(1 + \frac{4C_1 CM}{C_2^2}\right)^{-1/2}$$

By iterative method. On substituting the value of M we can obtain the optimum value of 'n'.

M will change according to change in C_1, C_2 and C such that $C_1 CM / C_2^2$ is nearly constant.

Conclusions:

The optimum size of the unit will be smaller when

i) The cost of enumeration of an element increase.
ii) The cost of travel between units decrease;
iii) The cost of the survey is sufficiently large

Cluster Sampling For Proportions:-

Suppose it is required to estimate the proportion of elements belonging to a specified Class when the population consists of N Clusters, each of size M, and a random Sample, wor, of n clusters is selected. Suppose the M elements in any cluster can be classified into two classes. to the Class, and 0 otherwise. It can be be easily seen that

$y_{ij} = 1$, otherwise 0

$p_i = a_i / M$

a_i being the number of elements in the i^{th} Cluster belonging the Specified Class.

An unbiased estimator of the population proportion

$$P = \sum_{i}^{N} a_i / NM \qquad \therefore p_i = a_i / M$$

$$= \sum_{i}^{N} p_i / N$$

$$\hat{P}_c = \sum_{i}^{n} \frac{p_i}{n} = \bar{p}$$

where p_i is the proportion of elements belonging to the Specified Class in the i^{th} cluster of the Sample

The Sampling variance of $\hat{P}_c$ is give by

$$V(\hat{P}_c) = \frac{(1-f)}{n} S_b^2$$

Where S_b^2 is the variance between Cluster proportions and is given by

$$M S_b^2 = \sum_{i}^{N} \frac{(p_i - P)^2}{N} = PQ - \sum_{i}^{N} \frac{P_i Q_i}{N}$$

For large N, we have $S^2 \simeq S_b^2 + S_w^2 = PQ$

and the within-variance S_w^2 is given by $\sum_{i}^{N} P_i Q_i / N$. Hence, the intracluster correlation coefficient ρ can be written as

$$\rho = 1 - \frac{\sum_{i}^{N} MP_i Q_i}{(M-1)PQ}$$

∴ The sampling variance, in terms of the intracluster correlation coefficient, can be expressed as

$$V(\hat{P}_c) = \frac{(1-f)NPQ}{(N-1)nM}[1+(M-1)\rho] \longrightarrow (*)$$

An estimator of the total number of units belonging to the specified category can be obtained by multiplying $\hat{P}_c$ by NM and the expression for its sampling variance is N^2M^2 times that given by (*).

If a simple random sample of nM elements could be taken, the variance of the sample proportion $\hat{P}$ would be given by

$$V(\hat{P}) = (1-f)\frac{NPQ}{nM(N-1)} \longrightarrow (**)$$

The efficiency of cluster sampling as compared to simple random sampling, usually can be obtained as

$$RE = \frac{(N-1)}{(NM-1)} \frac{NPQ}{NPQ - \sum_{i}^{N} P_i Q_i}$$

An estimator of $V(\hat{P}_c)$ is given by

$$v(\hat{P}_c) = (1-f)\frac{s_b^2}{n}$$

$$= (1-f)\frac{\sum_i (P_i - \bar{P})^2}{n(n-1)} \rightarrow (***)$$

An estimator for the Sampling Variance of the total number of units belonging to a Specified Class can be obtained by Multiplying $N^2 M^2$ to the Value in relation (***)

Unequal cluster Sampling: Estimators of Mean & their Variance:

Consider the case when the size of all the cluster is the same. But in many practical situation cluster, sizes vary. Now we shall discuss the case of ~~unclustered~~ unequal cluster.

Suppose N cluster. let i^{th} cluster of M_i element $(i = 1, 2 \ldots N)$ and $\sum_{i}^{N} M_i = M_0$. The population mean is defined by

$$\bar{Y} = \frac{\sum_{i}^{N} \sum_{j}^{M_i} y_{ij}}{\sum_{i}^{N} M_i} = \frac{\sum_{i}^{N} M_i \bar{y}_i}{\sum_{i}^{N} M_i} = \frac{\sum_{i}^{N} M_i \bar{y}_i}{M_0}.$$

$\bar{y}_i$ = mean per element of the i^{th} cluster.

The Pooled mean of the cluster means in the popn.

$$\bar{Y}_N = \sum_{i}^{N} \bar{y}_i / N.$$

Let a random sample, wor, of n cluster be drawn and all elements of the cluster surveyed.

i) $\bar{y}_n = \sum_{i}^{n} \bar{y}_i / n$

ii) $\bar{y}'_n = \sum_{i}^{n} \frac{M_i \bar{y}_i}{\sum_{i}^{n} M_i}$

iii) $\bar{y}^{*}_n = \frac{N}{M_0} \frac{\sum_{i}^{n} M_i \bar{y}_i}{n} = \frac{N}{nM_0} \sum_{i}^{n} M_i \bar{y}_i$

$$\left[\because \bar{M} = \sum_{i}^{N} M_i / N = 1/n \sum_{i}^{n} \frac{M_i \bar{y}_i}{\bar{M}} \right.$$

$$\bar{M} = \frac{M_0}{N}$$

$$\left. \frac{1}{\bar{M}} = \frac{N}{M_0} \right]$$

Theorem:

S.T the simple arithmetic mean given by relation (i) is not an unbiased estimator. The bias and sampling variance of the estimator are given by.

$$B(\bar{y}_n) = -\frac{Cov(\bar{y}_i, M_i)}{\bar{M}} \quad \text{and}$$

$$V(\bar{y}_n) = \frac{1-f}{n} S^2_b$$

Where $S^2_b = \sum_i^N \frac{(\bar{y}_i - \bar{Y}_N)^2}{N-1}$

Pf

To Prove $\bar{y}_n$ is not unbiased

$$E(\bar{y}_n) = E\left[\frac{1}{n} \sum_i^n \bar{y}_i\right]$$

$$= \frac{1}{n} E\left[\sum_i^n \bar{y}_i\right]$$

$$= \frac{1}{n} \sum_{i=1}^n E(\bar{y}_i)$$

$$= \frac{1}{n} \sum_{i=1}^n \frac{1}{N} \sum_{i=1}^N \bar{y}_i$$

$$= \bar{Y}_N \neq \bar{Y}$$

Thus, $\bar{y}_n$ is a biased estimator of the Population mean $\bar{Y}$.

The bias of the Estimator is given by

$$B(\bar{y}_n) = E(\bar{y}_n) - \bar{Y} = \frac{\sum_i^N \bar{y}_i}{N} - \frac{\sum_i^N M_i \bar{y}_i}{N\bar{M}}$$

$$= \frac{\sum_{i}^{N} (\bar{y}_i)(\bar{M}) - \sum_{i}^{N} M_i \bar{y}_i}{N\bar{M}}$$

$$= \sum_{i}^{N} (\bar{M} - M_i)\, \bar{y}_i \,/\, N\bar{M}$$

$$= \frac{-\mathrm{cov}(y_i, M_i)}{\bar{M}}$$

For a Popn in which M_i's do not Appreciably vary from one cluster to Another. The bias may be materially Significant. If M_i and y_i are uncorrelated the bias is zero and $\bar{y}_n$ is an unbiased estimate in the case

This shows bias is Expected to be Small when M_i and $\bar{y}_i$ are not highly correlated In such a case. It is advisable to use this estimater. Its Sampling variance is given by

$$V(\bar{y}_n) = E[\bar{y}_n - E(\bar{y}_n)]^2 = E(\bar{y}_n - \bar{\bar{y}}_n)^2$$

$$= \frac{(1-f)}{n} \frac{\sum_{i}^{N} (\bar{y}_i - \bar{\bar{y}}_n)^2}{(N-1)} = \frac{(1-f)}{n} s^2_b$$

Corrollary An unbiased estimator of $V(\bar{y}_n)$ is given by

$$v(\bar{y}_n) = \frac{(1-f)}{n} s^2_b$$

Where $s^2_b = \sum_{i}^{n} \frac{(\bar{y}_i - \bar{\bar{y}}_n)^2}{(n-1)}$

Theorem

Show that the estimator of the mean given by relation (iii) is unbiased and its variance is given by.

$$V(\bar{y}^*_n) = \frac{(1-f)}{n} S_b^{*2}$$

where $$S_b^{*2} = \frac{1}{(N-1)} \sum_i^N \left(\frac{M_i}{\bar{M}} \bar{y}_i - \bar{Y} \right)^2$$

Pf

Let us consider the estimator $\bar{Y}^*_n = \frac{\sum_i^n M_i \bar{y}_i}{n\bar{M}}$

$$E(\bar{y}^*_n) = E\left(\sum_i^n M_i \bar{y}_i / n\bar{M} \right) = \frac{\sum_i^n E(M_i \bar{y}_i)}{n\bar{M}} = \bar{Y}$$

Hence $\bar{y}^*_n$ is an unbiased estimator. The sampling variance of the estimator is given by.

$$V(\bar{y}^*_n) = \frac{(1-f)}{n} \sum_i^n \left(\frac{M_i}{\bar{M}} \bar{y}_i - \bar{Y} \right)^2$$

$$V \qquad (N-1)$$

$$= \frac{(1-f)}{n} S_b^{*2}$$

If may be noticed that the estimator $\bar{y}^*_n$ will often be less precise this occurs because the variance depends upon the variation of the product $M_i \bar{y}_i$ and is likely to be larger than $\bar{y}_n$ unless $\bar{y}_i$ and M_i vary in such a way that

their product is almost constant

Corollary An unbiased estimator of $V(\bar{y}^*_n)$ is given by.

$$v(\bar{y}_n^*) = \frac{(1-f)}{n} s_L^{*2}$$

where

$$s_L^{*2} = \frac{\sum_i^n \left(\frac{M_i \bar{y}_i}{\bar{M}} - \bar{y}^*_n \right)^2}{(n-1)}$$

Relative efficiency of unequal cluster sampling

In a number of situations it is easier to take some naturally formed groups of elements usually in such cases cluster size would be unequal for example villages which are groups of households or households which are groups of persons are usually taken as clusters for the purpose of sampling on account of operational convenience thus unequal cluster sampling is the most practial situation and its relative efficiency with respect to simple random sampling should be worked out.

In unequal cluster sampling the total number of elements $\sum_1^n M_i$ in the sample is a ~~value~~ random variable with expected value nM.

If an equivalent simple random sample of size $n\bar{M}$ had been selected directly from the population of $N\bar{M}$ elements the variance of the mean per element would be given by

$$V_{SR}(\bar{y}) = \frac{(N\bar{M} - n\bar{M})}{N\bar{M}\, n\bar{M}} S^2 = \frac{(1-f)}{n\bar{M}} S^2$$

Comparing this with the value given by,

$$R.E = \frac{S^2}{\bar{M} S_b^2}$$

Hence it is observed that the efficiency increases as the variation ween clusters decreases in general cluster sampling will be efficient only when the variation between clusters is as small as possible.