(*b*) ***How to improve Systematic Sampling in the Presence of Linear Trend ?***

Even though systematic sampling is not very efficient in the presence of linear trend, better results can be obtained by applying certain corrections, known as ***"Yates' End Corrections (1948)"*** to the systematic sample.

This consists in giving the weight $(1/n)$ to all the sample values except the first and the last values which receive the weights :

$$\frac{1}{n} + \frac{2i-k-1}{2(n-1)k} \text{ and } \frac{1}{n} - \frac{2i-k-1}{2(n-1)k} \text{ respectively.}$$

Using these weights, the weighted mean of the systematic sample becomes :

$$\bar{y}_{sys} = \left[\frac{1}{n} + \frac{2i-k-1}{2(n-1)k}\right] i + \frac{1}{n}\left[(i+k) + (i+2k) + \ldots + \{i + (n-2)k\}\right] + \left[\frac{1}{n} - \frac{2i-k-1}{2(n-1)k}\right] \times [i + (n-1)k]$$

$$= \frac{i}{n} + \left[\frac{2i-k-1}{2(n-1)k} - \frac{2i-k-1}{2(n-1)k}\right] i + \frac{1}{n}\left[(n-2)i + \frac{k(n-2)(n-1)}{2}\right] + \frac{1}{n}[i + (n-1)k] - \frac{(2i-k-1)}{2}$$

$$= \frac{i}{n}\left[1 + (n-2) + 1\right] + \frac{k}{2n}(n-1)(n-2) + \frac{(n-1)k}{n} - \frac{2i-k-1}{2}$$

$$= \frac{k(n-1)}{n}\left[\frac{n-2}{2} + 1\right] + \frac{k+1}{2} = \frac{k(n-1)}{2} + \frac{k+1}{2}$$

$$= \frac{nk+1}{2} = \frac{N+1}{2} = \bar{Y}_N = \bar{y}_{..}$$

Thus, after applying Yates' end correction, we get

$\bar{y}_{i\cdot} = \bar{y}_{..} \Rightarrow V(\bar{y}_{sys}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_{i\cdot} - \bar{y}_{..})^2 = 0$, thus making systematic sampling 100% efficient.

7·11·5. Merits and Demerits of Systematic Sampling.

Merits **1.** Systematic sampling is operationally more convenient than simple random sampling or stratified random sampling. Time and work involved is also relatively much less. Moreover, systematic sampling yields a sample which is evenly spread over the entire population. Some of the practical situations where systematic sampling has been found very usual are given below.

(*i*) The selection of every *k*th strip in forest survey for estimation of timber.

(*ii*) The selection of every *k*th village in rural surveys.

(*iii*) The selection of cornfields every *k*th mile/km apart for observation on incidence of borers.

(*iv*) The selection of every *k*th time interval for the estimation of the total catch of fish in fisheries [after the first unit is chosen at random].

Because of its operational convenience, the job of collecting the systematic sample can be entrusted to the field workers.

2. Systematic sampling may be more efficient than simple random sampling provided the frame (the list from which sample units are drawn) is arranged wholly at random. The most

Scanned with CamScanner

common approach to randomness is provided by alphabetical lists such as names in telephone directory, although even these may have certain non-random characteristics.

Demerits. Systematic sampling has a number of limitations and disadvantages as enumered below :

1. The main disadvantage of systematic sampling is that systematic samples are not in general random samples since the requirement in merit two is rarely fulfilled.

2. If N is not a multiple of n, then

(*i*) the actual sample size is different from that required, and

(*ii*) sample mean is not an unbiased estimate of the population mean.

However, these disadvantages can be overcome by adopting a technique known as circular systematic sampling (C.S.S.) . [See § 7·11·5]

3. In systematic sampling, $\text{Var}(\bar{y}_{sys}) = \frac{1}{k}\sum_{i=1}^{n}(\bar{y}_{i.} - \bar{y}_{..})^2$.

However, it is not possible to obtain an unbiased estimate of this variance, on the basis of a single sample because a systematic sample is regarded because a systematic sample is regarded as a sample of one unit (cluster). This obviously is a great drawback, since one important requirement for adopting any sampling method is that it should provide an estimate of the sampling error. However, it is possible to build up some biased but useful variance estimates on the basis of a systematic sample. [See § 7·11·6]

4. Systematic sampling may yield highly biased estimates if there are periodic features associated with the sampling interval, *i.e.*, if the frame (list) has a periodic feature and k is equal to or a multiple of the period. If the population exhibits a periodic trend (of a sine curve, say) the efficiency of the systematic sample depends on the value of k, the sampling interval. If k is equal to the period or an integral multiple of it, the systematic sampling becomes highly inefficient. In fact, in this case, systematic sampling is no better than selecting one unit at random.

If k is an odd multiple of half the period, then systematic sampling becomes most effective because in that case, systematic sampling provides zero variance. Some of the situations exhibiting periodicity are :

(*i*) Sales of departmental stores over a week

(*ii*) Postal articles received in a post office over the week

(*iii*) Temperatures over 24-hour period

(*iv*) Number of vehicles passing over a bridge during a day.

7·11·6. Circular Systematic Sampling. If N is not a multiple of n, *i.e.*, $N \neq nk$, then the sampling interval k cannot be uniquely defined. In such a case, take k to be an integer nearest to (N/n). If we select the first unit, say, i, randomly between 1 and k, then the systematic sample is : $i, i+k, i+2k, \ldots, i+(n-1)k$; $1 \leq i \leq k$.

However, depending on the choice of the starting number i, sometimes we may have $i + (n-1)k > N$ and so, the sample size instead of n reduces to $(n-1)$.

Suppose we want a systematic sample of size 6 out of 22 units. We have $N = 22$ and $n = 6$ so that $(N/n) = 3{\cdot}67$ we take $k = 4$, the integer nearest to 3·67.

Scanned with CamScanner

Thus, the four systematic samples are :

Sample No.	Random Start	Sample Units	Sample Size
1	1	1, 5, 9, 13, 17, 21 ...	$n = 6$
2	2	2, 6, 10, 14, 18, 22 ...	$n = 6$
3	3	3, 7, 11, 15, 19, ...	$n = 5$
4	4	4, 8, 12, 16, 20, ...	$n = 5$

Thus, the sample size is not necessarily n (= 6) but in some cases it is $n - 1$ (= 5).

Hence, if $N \neq nk$, the sample size may differ from sample to sample, depending on the choice of the random start 'i'.

Moreover, in this case, the sample mean is not an unbiased estimate of the population mean.

The problem of variable sample size when $N \neq nk$ can be overcome by adopting a modified version of the above scheme introduced by **Prof. D.B. Lahiri** (1952) and known as *Circular Systematic Sampling* (CSS). This ensures a constant sample size.

The procedure consists in selecting the unit 'i' by random start from 1 to N, and thereafter select every kth unit in a circular way, k being an integer nearest to (N/n). The systematic sample is then specified by the units corresponding to the numbers :

$$\left.\begin{array}{ll} i + jk, & \text{if } i + jk \leq N \\ \text{and} \quad i + jk - N, & \text{if } i + jk > N \end{array}\right\}, \quad j = 0, 1, 2, \ldots, (n-1)$$

Using this technique in the above illustration, for the random starts $i = 3$ and $i = 4$, the corresponding systematic samples of size 6 are given below.

$i = 3$; Sample units are : 3, 7, 11, 15, 19, 1 (= 23 – 22)

$i = 4$; Sample units are : 4, 8, 12, 16, 20, 2 (= 24 – 22);

each with $n = 6$, the desired sample size.

Remarks **1.** If $N = 129$, $n = 20$, then $(N/n) = 6{\cdot}45$ and we take $k = 6$, the integer nearest to 6·45. If $N = 131$, $n = 20$, then $(N/n) = 6{\cdot}56$ and we take $k = 7$, the integer nearest to 6·56.

2. If $N \neq nk$, then an unbiased estimate of $\bar{Y}_N$ is provided by :

$$\hat{\bar{Y}}_N = \frac{k}{N} \sum_{j=1}^{n'} y_{ij},$$

where n' is the number of units that can be expected in the sample, and not by : $\frac{1}{n} \sum_{j=1}^{n} y_{ij}$, the sample mean.

7·11·7. Estimation of the Variance of Estimates. As already pointed out [*c.f.* Disadvantages of systematic sampling], it is not possible to obtain an unbiased estimate of $V(\bar{y}_{sys})$ on the basis of a single sample. There are no easy methods for estimating $V(\bar{y}_{sys})$. In the absence of direct method, the following methods are used for estimating it.

Method I. Regard the ith systematic sample selected as a simple random sample of size n units and then :

$$\hat{V}(\bar{y}_{sys}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_i^2, \text{ where } S_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot})^2 \qquad \ldots (7{\cdot}137)$$

Scanned with CamScanner

Method II. In this method, we take into account the successive differences of the sample values so that :

$$\hat{V}(\bar{y}_{sys}) \simeq \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2 \quad \ldots (7\cdot138)$$

Method III. This method consists in selecting the sample of required size n in the form of two or more, say, m systematic independent inter-penetrating sub-samples of the same size (n/m), each selected with independent random starts. If $\hat{\bar{y}}$, $i = 1, 2, \ldots, m$ are the estimates of the population mean $\bar{Y}_N$ based on these m sub-samples, then an unbiased estimator of the variance of the combined estimator

$$\hat{\bar{y}}_c = \frac{1}{m} \sum_{i=1}^{m} \hat{\bar{y}}_i ,$$

is given by : $\ldots (7\cdot139)$

$$\hat{V}(\hat{\bar{y}}_c) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\hat{\bar{y}}_i - \hat{\bar{y}}_c\right)^2 \quad \ldots (7\cdot140)$$

When $m = 2$, this expression reduces to $\left(\hat{\bar{y}}_1 - \hat{\bar{y}}_2\right)^2 / 4$.

7·12. CLUSTER SAMPLING

In this case the total population is divided, depending on problem under study, into some recognisable sub-divisions which are termed as *clusters* and a simple random sample of these clusters is drawn. We than observe, measure and interview each and every unit in the selected clusters.

For example, if we are interested in obtaining the income of opinion data in a city, the whole city may be divided into N different blocks or localities (which determine the clusters) and a simple random sample of n blocks is drawn. The individuals in the selected blocks determine the cluster sample.

Notations :

N = Total number of clusters ; n = Number of sampled clusters.

M_i = Number of sampling units in the *ith* cluster.

$M = \sum_{i=1}^{N} M_i$ is total number of units in the population.

Y_{ij} = jth observation in the ith cluster ($j = 1, 2, \ldots, M_i$; $i = 1, 2, \ldots, N$).

y_{ij} = jth observation in the ith sampled cluster ($j = 1, 2, \ldots, M$; $i = 1, 2, \ldots, n$).

In the typical situation, M_i as well as M are not known.

An unbiased estimate of the population total is given by :

$$\left(\hat{Y}\right)_c = \frac{N}{n} \sum_i \sum_j y_{ij} = \frac{N}{n} \sum_{i=1}^{n} T_i \quad \ldots (7\cdot141)$$

where T_i is the observed total for the ith sampled cluster and suffix 'c' stands for cluster sampling. The variance of the estimate is given by :

$$\text{Var}\left(\hat{Y}\right)_c = \frac{N(N-n)}{n(n-1)} \sum_{i=1}^{n} (T_i - T)^2 \quad \ldots (7\cdot142)$$

$$\ldots (7\cdot142a)$$

where $$\bar{T} = \frac{1}{n} \sum_{i=1}^{n} T_i$$

is the arithmetic mean of the sampled cluster totals.

Scanned with CamScanner

Remarks **1.** As pointed out earlier, $M = \sum_{i=1}^{N} M_i$, the total number of units in the population is generally unknown prior to sampling. However, on the basis of the sample, we may obtain its estimate by the formula :

$$\hat{M} = \frac{N}{n} \sum_{i=1}^{n} M_i \qquad \ldots (7\cdot142)$$

which is same as (7·141) with T_i replaced by M_i and consequently the variance of this estimate is given by (*c.f.* 7·142)

$$\text{Var}(\hat{M}) = \frac{N(N-n)}{n\,n-1} \sum_{i=1}^{n} (M_i - \bar{M})^2 \qquad \ldots (7\cdot142a)$$

2. It may be observed from expression in (7·142) that the variance of the estimated total $\hat{Y}_c$ depends upon the heterogeneity of the cluster totals T_i ($t = 1, 2, \ldots, n$). It is thus obvious that even if the units within a cluster are homogeneous Var $(\hat{Y})_C$ may be sufficiently large if the number of units varies considerably from cluster to cluster. Thus, in using cluster sampling, the following points should be borne in mind :

(*i*) Clusters should be as small as possible consistent with the cost and limitations of the survey, and

(*ii*) The number of sampling units in each cluster should be approximately same.

Thus, cluster sampling is not to be recommended if we are sampling areas in city where there are private residential houses, business and industrial complexes, apartment buildings, etc., with widely varying number of persons or households.

7·13. MULTISTAGE SAMPLING

Instead of enumerating all the sampling units in the selected clusters one can obtain better and more efficient estimators by resorting to subsampling within the clusters. This technique is called *two stage sampling*, clusters being termed as *primary units* and the units within the clusters as *secondary units*.

The above technique may be generalised to what is called *multistage sampling*. As the name suggests, multistage sampling refers to a sampling technique which is carried out in various stages. Here the population is regarded as made of a number of primary units each of which is further composed of a number of secondary stage units, and so on, till we ultimately reach the desired sampling unit in which we are interested. For example, if we are interested in obtaining a sample of, say, n households from a particular State the first stage units may be district, the second stage units may be villages in the districts and third stage units will be households in the villages. Each stage thus results in a reduction of the sample size.

Multistage sampling consists of sampling first stage units by some suitable method of sampling. From among the selected first stage units, a sub-sample of secondary stage units is drawn by some suitable method of sampling which may be same as or different from the method used in selecting first stage units. Further stages may be added to arrive at a sample of the desired sampling units.

Merits and Limitations. Multistage sampling is more flexible as compared to other methods of sampling. It is simple to carry out and results in administrative convenience by permitting the field work to be concentrated and yet covering large area.

Scanned with CamScanner

Unit IV : Cluster Sampling

Introduction

In random sampling, it is presumed that the population has been divided into a finite number of distinct and identifiable units defined as sampling units. The smallest unit into which the population can be divided is called an element of the population. A group of such elements is known as a cluster. When the sampling unit is a cluster, the procedure is called cluster sampling.

Generally, identification and location of an element requires considerable time. However, once an element has been located, the time taken for surveying a few neighbouring elements is small. Thus the main function in cluster sampling is to specify clusters (or) to divide the population into appropriate clusters.

clusters are generally made up of neighbouring elements and, therefore, the elements within a cluster tend to have similar characteristics. As a simple rule, the number of elements in a cluster should be small and the number of clusters should be large. After dividing the population into specified clusters, the required number of clusters can be selected either by equal or unequal probabilities of selection. All the elements in selected clusters are enumerated.

Scanned with CamScanner

Equal Cluster Sampling

Notations : (For equal clusters)

Suppose the population consist of 'N' clusters, each of M elements, and that a sample of 'n' clusters is drawn by the method of simple random sampling.

N = no/. of clusters in the population

n = " " " sample

M = no/. of elements in the cluster

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

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

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

$\bar{y}_n = \frac{\sum_{i=1}^{n} \bar{y}_{i.}}{n}$ = the mean of cluster means in a sample of 'n' clusters

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

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

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

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

$S_w^2 = \frac{\sum_{i=1}^{N} S_i^2}{N}$ = the mean square within clusters (w for within)

$S_b^2 = \frac{\sum_{i=1}^{N} (\bar{y}_{i.} - \bar{Y}_N)^2}{N-1}$

= the mean square between clusters mean in the population (b for between)

Scanned with CamScanner

$$S^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{Y})^2}{NM-1}$$ = the mean square between elements in the population

$$\rho = \frac{\sum_{i=1}^{N}\sum_{j\neq j'=1}^{M}(y_{ij}-\bar{Y})(y_{ij'}-\bar{Y})}{(M-1)(NM-1)S^2}$$

= the intracluster Correlation Co-efft between elements within clusters.

Theorem: 1

In cluster sampling, P.T. the sample mean is an unbiased estimate of the population mean.

Proof.

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

The mean of cluster means in a sample of 'n' clusters is given by

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

$$\therefore E(\bar{y}_n) = E\left[\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}E(\bar{y}_{i.}) \quad \text{——— (1)}$$

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

Every cluster will gives the following frequency distn. with probability $1/N$.

ie.,

Cluster mean	$\bar{y}_1$.	$\bar{y}_2$.	...	$\bar{y}_N$.
Probability	$1/N$	$1/N$	...	$1/N$

Scanned with CamScanner

$$\therefore E(\bar{y}_{i.}) = \sum_{i=1}^{N} \bar{y}_{i.} P(\bar{y}_{i.}) \quad \text{(illegible)}$$

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

$$= \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{M} \sum_{j=1}^{M} y_{ij} \right]$$

$$= \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} = \bar{Y} \quad \text{——} (2)$$

Put (2) in (1), we get

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

$\therefore E(\bar{y}_n) = \bar{Y}$ and hence the proof.

Theorem: 2

In equal cluster sampling, show that

$$V(\bar{y}_n) = \frac{1-f}{n} S_b^2 = \frac{N-n}{Nn} S_b^2 \simeq \frac{1-f}{nM} S^2 [1 + (M-1)\rho]$$

Proof:

Suppose that a population consists of 'N' clusters and from that a sample of 'n' clusters is drawn by the method of SRS.

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

where S_b^2 = the mean square between clusters mean in the population

$$= \frac{1}{N-1} \sum_{i=1}^{N} (\bar{y}_{i.} - \bar{Y})^2 \quad \text{——} (1)$$

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

$$= \sum_{i=1}^{N} \left[\frac{1}{M} \sum_{j=1}^{M} (y_{ij} - \bar{Y}) \right]^2$$

Scanned with CamScanner

$$= \sum_{i=1}^{N} \left[\frac{1}{M} \left(\sum_{j=1}^{M} y_{ij} - M\bar{Y} \right) \right]^2$$

$$= \sum_{i=1}^{N} \left[\frac{1}{M} \left(\sum_{j=1}^{M} y_{ij} - \sum_{j=1}^{M} \bar{Y} \right) \right]^2$$

$$= \sum_{i=1}^{N} \left[\frac{1}{M} \sum_{j=1}^{M} (y_{ij} - \bar{Y}) \right]^2$$

$$= \frac{1}{M^2} \sum_{i=1}^{N} \left[\sum_{j=1}^{M} (y_{ij} - \bar{Y})^2 + \sum_{j \neq j'=1}^{M} (y_{ij} - \bar{Y})(y_{ij'} - \bar{Y}) \right]$$

$$= \frac{1}{M^2} \left[\sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2 + \sum_{i=1}^{N} \sum_{j \neq j'=1}^{M} (y_{ij} - \bar{Y})(y_{ij'} - \bar{Y}) \right]$$

$$= \frac{1}{M^2} \left[(NM-1)S^2 + (M-1)(NM-1)S^2\rho \right]$$

since $S^2 = \dfrac{\sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2}{NM-1}$

and $\rho = \dfrac{\sum_{i=1}^{N} \sum_{j \neq j'=1}^{M} (y_{ij} - \bar{Y})(y_{ij'} - \bar{Y})}{(M-1)(NM-1)S^2}$

$$\therefore \sum_{i=1}^{N} (\bar{y}_{i\cdot} - \bar{Y})^2 = \frac{(NM-1)S^2}{M^2} \left[1 + (M-1)\rho \right] \quad \text{---------- (2)}$$

Put (2) in (1), we get

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

When N is very large,

Taking $N-1 \simeq N$ and $NM-1 \simeq NM$, we get

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

$$= \frac{S^2}{M} \left[1 + (M-1)\rho \right]$$

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

$$= \frac{1-f}{nM} S^2 \left[1 + (M-1)\rho \right]$$

Scanned with CamScanner

Note :

It has been shown that the variance in cluster sampling depends on the no. of clusters in the sample (n), the size of the cluster (m), the intracluster correlation co-efft (ρ) and the variance S^2.

If $m = 1$, it gives the sampling variance of a SRS of nm elements taken individually. In this situation

Cluster sampling = SRS

If $m > 1$ and ρ is positive, cluster sampling will give a higher variance than the mean per element. If ρ is -ve, then cluster sampling can be used.

Relative efficiency of cluster Sampling

In sampling nm elements from the popln. by SRS, the variance of the sample mean $\bar{y}$ is given by

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

In cluster sampling, $V(\bar{y})_c = \frac{1-f}{n} S_b^2$

$$\therefore \text{Relative Efficiency} = \frac{V(\bar{y})_R}{V(\bar{y})_c} = \frac{\frac{(1-f)\,S^2}{nm}}{\frac{1-f}{n} S_b^2} = \frac{S^2}{m S_b^2} \quad \text{---(1)}$$

This shows that the efficiency of cluster sampling increases as the mean square between cluster (S_b^2) decreases.

$$S^2 = \frac{(N-1)S_b^2 + N(m-1)S_w^2}{Nm-1}$$

Also $(N-1)mS_b^2 = (Nm-1)S^2 - N(m-1)S_w^2$ ---(2)

Scanned with CamScanner

This result suggest that the clusters should be so formed that variation within clusters is maximum while variation between clusters is minimum.

Relative efficiency using intra cluster Co-efficient ρ

For large N, the relative efficiency of cluster sampling in terms of intra-cluster Co-efft 'ρ' is given by,

$$\left.\begin{matrix}\text{Relative} \\ \text{Efficiency}\end{matrix}\right\} = \frac{V(\bar{y})_R}{V(\bar{y})_c} = \frac{\frac{1-f}{nm} S^2}{\frac{1-f}{nm} S^2 \left[1+(m-1)\rho\right]}$$

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

i) In Case of Complete homogeneity of clusters, $S_w^2 = 0$ and so $\rho = 1$ and $E = \frac{1}{m}$. ie. cluster sampling is not efficient

ii) In case of Complete heterogenity, $S_w^2 = S^2$, so $S_b^2 = 0$ and $\rho = \frac{-1}{m-1}$, ie., cluster sampling is very efficient.
$\because S_b^2 = \frac{S^2}{m}\left[1+(m-1)\rho\right]$

Hence it should be noted that 'ρ' lies in the range $\frac{-1}{m-1}$ to 1. ie., $-\frac{1}{m-1} \le \rho \le +1$. It also shows that cluster sampling will be more efficient if ρ is negative. In practice, ρ is usually +ve as neighbouring elements are grouped to form clusters. Generally, ρ decreases with increase in m. The efficiency of cluster sampling increases as the factor $(m-1)\rho$ increases with cluster size.

Scanned with CamScanner

Advantages

For a given number of sampling units, cluster sampling is more convinient and less costly. The advantages of cluster sampling are that

i) Collection of data for neighbouring elements is easier, cheaper, faster and operationally more convinient than observing units spread over the region

ii) it is less costly than SRS due to the saving of time in journeys, identification, Contacts, etc.

iii) When the sampling frame of elements may not be readily available.

Disadvantages

From the point of view of statistical efficiency, cluster sampling is generally less efficient than SRS due to the usual tendency of units in a cluster to be similar.

Scanned with CamScanner

Two-stage Sampling

Introduction

In cluster sampling, clusters were considered as sampling units and all the elements in the selected clusters were enumerated completely. It has been stated that, cluster sampling is economical under certain circumstances but the method restricts the spread of the sample over the population which results generally in increasing the variance of the estimator. It is, therefore, logical to expect that the efficiency of the estimator will be increased by distributing a sample of units in each selected cluster instead of completely enumerating all the elements in the sample of clusters. This type of sampling which consists in first selecting the clusters and then selecting a specified no. of elements from each selected cluster is known as sub-sampling or two-stage sampling.

In two-stage sampling, clusters which form the units of sampling at the first stage are called the first stage units (fsu) or primary sampling units (psu) and the elements within clusters are called second-stage units (ssu). This procedure can be generalized to three or more stages and is termed multi-stage sampling. For example, in crop surveys for estimating yield of a crop in a district, a block may be considered as a primary sampling unit, the villages the second stage units, the crop fields the third stage units, and a plot of fixed size the ultimate unit of sampling.

Scanned with CamScanner

Advantages

The multi-stage sampling procedure may be taken to be a better combination of random sampling and cluster sampling procedures. The following are the advantages of multi-stage sampling procedure

1. Multi-stage sampling has been found to be very useful in practice and this procedure is being commonly used in large-scale surveys
2. It is less efficient than single stage random sampling and more efficient than cluster sampling from the sampling variability point of view.
3. From the cost and operational point of view, it is more efficient than single stage random sampling and less efficient than cluster sampling
4. Since the first stage units (fsu) are selected in the first stage itself, it is very easy to select the second stage units (ssu) from the first stage units.
5. This design is more flexible as it permits the use of different selection procedures in different stages.
6. Multi-stage sampling may be the only choice in a number of practical situations where a satisfactory sampling frame of ultimate stage units is not readily available and the cost of obtaining such a frame is large.

Scanned with CamScanner

Two-Stage Sampling with Equal First-Stage Units

Notations

Let us assume that the population consists of NM elements grouped into 'N' fsu's of M ssu's each.

Let 'n' be the no. of fsu's in the sample and 'm' be the no. of ssu's to be selected from each sampled first stage unit. Also we assume that the units at each stage are selected with equal probability.

y_{ij} = the value obtained for the j^{th} ssu in the i^{th} fsu

$\bar{Y}_{i.} = \frac{\sum_{j=1}^{M} y_{ij}}{M}$ = mean per element in the ~~population~~ i^{th} fsu

$\bar{Y} = \frac{\sum_{i}^{N} \bar{Y}_{i.}}{N}$ = mean per element in the population

$S_b^2 = \frac{\sum_{i}^{N} (\bar{Y}_{i.} - \bar{Y})^2}{N-1}$ = true variance between first stage unit means

$S_w^2 = \frac{\sum_{i}^{N} \sum_{j}^{M} (y_{ij} - \bar{Y}_{i.})^2}{N(M-1)}$ = true variance within first stage units

$\bar{y}_{i.} = \frac{\sum_{j}^{m} y_{ij}}{m}$ = sample mean per ssu in the i^{th} fsu

$\bar{y} = \frac{\sum_{i}^{n} \bar{y}_{i.}}{n}$ = overall sample means per element

Theorem used

1. $E(t) = E_1 E_2(t)$ ——— (1)
2. $V(t) = V_1 E_2(t) + E_1 V_2(t)$ ——— (2)

where E_1 and V_1 are expectation and variance over the first stage and E_2 and V_2 are the conditional expectation and variance over the second stage for a given sample of fsu's.

Scanned with CamScanner

Theorem 1

If the 'n' fsu's and the 'm' ssu's from each chosen fsu are selected by SRSWOR, $\bar{y}$ is an unbiased estimator of $\bar{Y}$.

Proof:

Applying relation $E(t) = E_1 E_2 (t)$ for getting expectation, we have

$$E(\bar{y}) = E_1 E_2 (\bar{y}/i)$$
$$= E_1 (\bar{Y}_{i.})$$
$$= \bar{Y}$$

Theorem: 2

S.T $V(\bar{y}) = \frac{N-n}{Nn} S_b^2 + \frac{M-m}{Mm} \frac{S_w^2}{n}$

Proof:

To obtain the variance of the estimator, by relation $V(t) = V_1 E_2 (t) + E_1 V_2 (t)$

where E_1 and V_1 are expectation and variance over the first stage and E_2 and V_2 are the conditional expectation and variance over the second stage for a given sample of fsu's., we have

$$V(\bar{y}) = V_1 \left[E_2 (\bar{y}_n / i) \right] + E_1 \left[V_2 (\bar{y}_n / i) \right]$$
$$= V_1 (\bar{Y}_{i.}) + E_1 \left[V_2 \left\{ \frac{1}{n} \sum_{i=1}^{n} \bar{y}_{i.}/i \right\} \right]$$
$$= \frac{N-n}{Nn} S_b^2 + E_1 \frac{1}{n^2} \left[V_2 \left\{ \sum_{i=1}^{n} \bar{y}_{i.}/i \right\} \right]$$
$$= \frac{N-n}{Nn} S_b^2 + \frac{1}{n^2} E_1 \sum_{i=1}^{n} \left(\frac{M-m}{Mm} S_i^2 \right)$$

Scanned with CamScanner

$$V(\bar{y}) = \frac{N-n}{Nn} S_b^2 + \frac{1}{n^2} \frac{m-m}{Mm} \sum_{i=1}^{n} E_1(S_i^2)$$

$$= \frac{N-n}{Nn} S_b^2 + \frac{1}{n^2} \times \frac{m-m}{Mm} \times n S_w^2$$

$$= \frac{N-n}{Nn} \frac{S_b^2}{n} + \frac{m-m}{Mm} \frac{S_w^2}{n} \quad \text{where } S_w^2 = \frac{1}{N} \Sigma S_i^2$$

Theory Questions

1. Describe cluster sampling and in what situations the cluster sampling be preferred? (Nov'07)
2. What are the merits of cluster sampling? (Nov'08)
3. Distinguish cluster and two stage sampling. State the main advantages of two stage sampling? (Nov'07)
4. Compare two-stage with one-stage sampling and give your comments (Apr'09)

↳ Continue...

If $f_1 = \frac{n}{N}$ and $f_2 = \frac{m}{m}$ are the sampling fractions in the first and second stages, the result can be written as

$$V(\bar{y}) = \frac{(1-f_1)}{n} S_b^2 + \frac{(1-f_2)}{nm} S_w^2$$

Scanned with CamScanner

UNEQUAL CLUSTER SAMPLING

We have considered the case when the size of all the clusters is the same. But in many practical situations, cluster sizes vary. Now we shall discuss the case of unequal clusters.

Suppose there are N clusters. Let the i^{th} cluster consist of M_i elements; $(i = 1, 2, \ldots, N)$ and

$$\sum_{i=1}^{N} M_i = M_0.$$

The population mean per element $\bar{Y}$ is defined by

$$\bar{Y} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N} M_i \bar{y}_{i.}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N} M_i \bar{y}_{i.}}{M_0},$$

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

where $\bar{y}_{i.}$ is the mean per element of the i^{th} cluster.

We may also define the pooled mean of the cluster means as

$$\bar{Y}_N = \sum_{i=1}^{N} \frac{\bar{y}_{i.}}{N}$$

Let a random sample, wor, of 'n' clusters be drawn and all elements of the clusters surveyed. Three estimators of $\bar{Y}$ may be considered.

$$\bar{y}_n = \sum_{i=1}^{n} \frac{\bar{y}_{i.}}{n} \quad \text{(A.M)} \qquad \text{— (1)}$$

$$\bar{y}'_n = \frac{\sum_{i=1}^{n} M_i \bar{y}_{i.}}{\sum_{i=1}^{n} M_i} \quad \text{(Weighted Mean)} \qquad \text{— (2)}$$

$$\text{and } \bar{y}^*_n = \frac{N}{nM_0} \sum_{i=1}^{n} M_i \bar{y}_{i.} = \frac{\sum_{i=1}^{n} M_i \bar{y}_{i.}}{n\bar{M}} \qquad \text{— (3)}$$

where $\bar{M} = \frac{\sum_{i=1}^{N} M_i}{N} = \frac{M_0}{N}$.

Scanned with CamScanner

Theorem:

In unequal cluster sampling, P.T. simple arithmetic mean $\hat{\bar{y}}_n$ is not an unbiased estimator of the population mean.

Proof:

In unequal cluster sampling, the simple arithmetic mean is given by

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

Taking expectation on both sides, we get

$$E(\bar{y}_n) = E\left[\sum_{i=1}^{n} \frac{\bar{y}_{i.}}{n}\right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} E(\bar{y}_{i.}) \quad \text{---(1)}$$

If $\bar{y}_{i.}$ be the mean per element of the i^{th} cluster, every cluster will gives the following frequency distribution with probability $1/N$.

Cluster mean	$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\bar{y}_{3.}$	...	$\bar{y}_{N.}$
Probability	$1/N$	$1/N$	$1/N$	...	$1/N$

$$\therefore E(\bar{y}_{i.}) = \sum_{i=1}^{N} \bar{y}_{i.} P(\bar{y}_{i.})$$

$$= \frac{1}{N} \sum_{i=1}^{N} \bar{y}_{i.} = \frac{1}{N} N \bar{Y}_N \quad \therefore \bar{Y}_N = \frac{\sum_{i=1}^{N} \bar{y}_{i.}}{N} \quad \text{---(2)}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{j=1}^{m_i} \frac{y_{ij}}{m_i}\right]$$

Put (2) in (1), we get

$$\therefore E(\bar{y}_n) = \frac{1}{n} \sum_{i=1}^{n} \bar{Y}_N = \bar{Y}_N \neq \bar{Y}$$

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

Scanned with CamScanner

Theorem: 2

In Unequal cluster sampling, Show that $V(\bar{y}_n) = \frac{1-f}{n} S_b^2$ where $S_b^2 = \frac{\sum_{i=1}^{N} (\bar{y}_{i.} - \bar{Y}_N)^2}{(N-1)}$

Proof:

Suppose that a population Consists of 'N' clusters and i^{th} cluster consist of M_i elements; $(i = 1, 2, \ldots, N)$ and from that a sample of 'n' clusters is drawn by the method of SRS.

$$\therefore V(\bar{y}_n) = \frac{N-n}{Nn} S_b^2 \quad \text{where } S_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{y}_{i.} - \bar{Y}_N)^2$$

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

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

Relative Efficiency of Unequal cluster Sampling

In a no. 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 convinience. Thus

Scanned with CamScanner

Unequal cluster sampling is the most practical situation and its relative efficiency w.r.to SRS should be worked out.

In unequal cluster sampling, the total no. of elements $\sum_{i=1}^{n} m_i$ in the sample is a r.v. with expected value $n\bar{m}$. If an equivalent SRS 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

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

$$= \frac{N\bar{m} - n\bar{m}}{N\bar{m}}\frac{S^2}{n\bar{m}}$$

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

$$= (1-f)\frac{S^2}{n\bar{m}} \quad \text{where } f = \frac{n}{N}$$

and we know that, in unequal cluster sampling

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

$$\therefore \text{Relative efficiency} = \frac{V(\bar{y}_n)_R}{V(\bar{y}_n)}$$

$$= \frac{(1-f)\frac{S^2}{n\bar{m}}}{(1-f)\frac{S_b^2}{n}}$$

$$= \frac{S^2}{\bar{m}S_b^2}$$

Scanned with CamScanner

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

Varying Probability cluster Sampling

In many practical situations, cluster size is positively correlated with the variable under study. In these cases it is advisable to select the clusters with probability proportional to the no. of elements in the cluster. Hansen and Hurwitz (1943) gave a technique to select the clusters with probability proportional to their sizes M_i. In some cases the sizes M_i are known only approximately.

Let p_i $(0 < p_i < 1)$ be the probability of selecting the i^{th} cluster of size M_i $(i = 1, 2, \dots N)$ at each draw, such that $\sum_{i=1}^{N} p_i = 1$.

Suppose that $z_{ij} = \frac{M_i y_{ij}}{M_0 p_i}$ for $j = 1, 2, \dots, M_i$; $i = 1, 2, \dots, N$

Further, suppose that 'n' clusters are selected by pps with replacement, so that

$$\bar{z}_{i.} = \frac{M_i \bar{y}_i}{M_0 p_i} \quad \text{for} \quad i = 1, 2, \dots, N.$$

Scanned with CamScanner

Source :

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