VIVEKANANDA COLLEGE
THAKURPUKUR
KOLKATA-700063

NAAC ACCREDITED 'A' GRADE

Topic: Theory of Systematic Sampling

Course Title: Sample Survey

Paper: CC-8

Unit: 2

Semester: 4

Name of the Teacher: Riddhi Das Majumder

Name of the Department: Statistics

Systematic Sampling:

Description: Linear Systematic Sampling:

Suppose that the N units in the population are numbered 1 to N in some order. Suppose N = nk, where n is the sample size desired and k is an integer. A number is taken at random from the numbers 1 to k (using a table of random numbers). Suppose the random number is i. Then starting from the i^{th} unit in the popln, every k^{th} unit is selected till a sample of size n is obtained. Then the sample contains n units with serial numbers $i, i+k, i+2k, \ldots, i+(n-1)k$. Thus the sample consists of the first unit selected at random and every k^{th} unit thereafter. It is therefore called a systematic sample (with k as the sampling interval), and the procedure of selection is known as Systematic Sampling or linear systematic sampling.

For example, when N = 24, n = 6 and k = 4, the four possible linear systematic samples are:

Sample Number	Random start	Sampled units
1	1	1, 5, 9, 13, 17, 21
2	2	2, 6, 10, 14, 18, 22
3	3	3, 7, 11, 15, 19, 23
4	4	4, 8, 12, 16, 20, 24

The linear systematic sampling scheme described above can be regarded as dividing the popln of N units into k mutually exclusive and exhaustive groups (or clusters) $\{S_1, S_2, \ldots, S_k\}$ of n units each and choosing one of them at random.

Composition of the K-Systematic Samples:

	Sample number			
	1	2	... i ...	k
	y_1	y_2	y_i	y_k
	y_{k+1}	y_{k+2}	y_{k+i}	y_{2k}
	⋮	⋮	⋮	⋮
	y_{n-1k+1}	y_{n-1k+2}	y_{n-1k+i}	y_{nk}
Means	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_i$	$\bar{y}_k$

Thus the systematic sampling amounts to the selection of a single group or cluster with probability k^{-1}, from the k groups forming the entire popln.

Each unit U_i in the popln belongs to one and only one cluster. The probability of selecting a cluster is $1/k$, which is therefore

the probability with which any member of the cluster is selected in the sample i.e. $\pi_i = 1/k$. This shows that a systematic sample procedure is a probability sampling procedure. A linear systematic sample is a simple random sample of one cluster unit from a pop^n of k cluster units.

A linear systematic sampling is a mixed sampling, which is partly probabilistic and partly non-probabilistic. This is probabilistic since the 1st member of the sample is selected at random (with equal probabilities) from the 1st k units and non-probabilistic since the other members in the sample are fixed by the choice of the first member.

Unbiased Estimator for the Population Total and its Variance:

Theorem: An unbiased estimator for the population total Y under linear systematic sampling corresponding to the random start r is given by $\hat{Y}_{st} = \frac{N}{n}\sum_{j=1}^{n} Y_{r+(j-1)k}$ and its variance is $V(\hat{Y}_{st}) = \frac{1}{k}\sum_{r=1}^{k}(\hat{Y}_r - \bar{Y})^2$ where $\hat{Y}_r$ is the value of $\hat{Y}_{st}$ corresponding to the random start r.

Proof: Note that $\hat{Y}_{st}$ can take any one of k values $\hat{Y}_r$, r=1(1)k with prob. $\frac{1}{k}$. Therefore $E(\hat{Y}_{st}) = \sum_{r=1}^{k} \hat{Y}_r \cdot \frac{1}{k} = \frac{1}{k}\sum_{r=1}^{k}\frac{N}{n}\sum_{j=1}^{n} Y_{r+(j-1)k}$

$= \frac{N}{nk}\sum_{r=1}^{k}\sum_{j=1}^{n} Y_{r+(j-1)k} = \sum_{i=1}^{N} Y_i = Y$. Hence $\hat{Y}_{st}$ is unbiased for the population total Y.

Again $V(\hat{Y}_{st}) = E[\hat{Y}_{st} - E(\hat{Y}_{st})]^2 = E[\hat{Y}_{st} - \bar{Y}]^2$

$= \frac{1}{k}\sum_{r=1}^{k}[\hat{Y}_r - \bar{Y}]^2$.

An Alternative Expression for $V(\hat{Y}_{st})$:

Th: In linear systematic sampling interval of k, from a population of size N=nk, the variance of $\hat{Y}_{st}$ is given by

$V(\hat{Y}_{st}) = \frac{N(N-1)}{n} S^2\{1+(n-1)\rho\}$,

where $\rho = E[(Y_{ij}-\bar{Y})(Y_{ij'}-\bar{Y})] / E[Y_{ij}-\bar{Y}]^2$, is the intra-class correlation coefficient.

(12)

Ans: we have $V(\bar{Y}_{sy}) = \frac{1}{k}\sum_{r=1}^{k}(\bar{Y}_r - \bar{Y})^2 = \frac{1}{k}\sum_{r=1}^{k}\left\{\frac{1}{n}\sum_{j=1}^{n}(Y_{rj}-\bar{Y})^2\right\}$.

$= \frac{1}{k}\sum_{r}\sum_{j}(Y_{rj}-\bar{Y})^2 + (2/k)\sum_{r}\sum_{j}\sum_{k>j}(Y_{rj}-\bar{Y})(Y_{rk}-\bar{Y})$

By definition, $\rho = \frac{2}{kn(n-1)} \cdot \sum_{r}\sum_{j}\sum_{k>j}(Y_{rj}-\bar{Y})(Y_{rk}-\bar{Y}) \Big/ V(y)$

$\Rightarrow \sum_{r}\sum_{j}\sum_{k>j}(Y_{rj}-\bar{Y})(Y_{rk}-\bar{Y}) = \frac{kn(n-1)}{2} V(y).\rho$

$= \frac{1}{2}kn(n-1)\rho \cdot \frac{(N-1)}{N}S_y^2$

Hence, $V(\bar{Y}_{sy}) =$ [redacted]

$= k(N-1)S_y^2 + \frac{2}{k}\cdot\frac{kn(n-1)(N-1)\rho S_y^2}{2N}$

$= k(N-1)S_y^2 + k(n-1)(N-1)\rho S_y^2$. $[\because \frac{n}{N}=k]$

$= k(N-1)S_y^2\left\{1+(n-1)\rho\right\} = N(N-1)\bar{S}_y^2\cdot\left\{\frac{1+n-1\rho}{n}\right\}$

Corollary: $V(\bar{Y}_{sy})$ in systematic sampling be smaller than $V(\bar{Y})$ in SRSWOR if: $k(N-1)S_y^2 \cdot \frac{1+n-1\rho}{n} < N(N-n)\frac{S_y^2}{n}$.

if: $\rho < -\frac{1}{N-1}$.

Population with linear trend:

If the values $Y_1, Y_2, \ldots, Y_N$ of the units with labels $1, \ldots, N$ are related by $Y_i = \alpha + \beta i$, $i = 1(1)N$, i.e. the population constitutes solely a linear trend.

$\rho = \frac{1}{kn(n-1)}\cdot\sum_{r}\sum_{j}\sum_{k>j}(Y_{rj}-\bar{Y})(Y_{rk}-\bar{Y})$

Theorem: For populations possessing linear trend, $V(\bar{Y}_{sy}) < V(\bar{Y}_{SRS})$ where $\bar{Y}_{sy}$ and $\bar{Y}_{SRS}$ are the usual estimators under linear systematic sampling and simple random sampling, respectively.

Proof: $V(\bar{Y}_{SRS}) = N^2\cdot\frac{N-n}{nN}\cdot\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2$.

Let $Y_i = \alpha + \beta i$, $i = 1(1)N$. Then $\bar{Y} = \alpha + \beta\left(\frac{N+1}{2}\right)$.

Now, $\sum_{i=1}^{N}(Y_i-\bar{Y})^2 = \sum_{i=1}^{N}\left\{\alpha+\beta i - \alpha - \beta\frac{N+1}{2}\right\}^2 = \beta^2\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^2$

$= \beta^2\left\{\sum_{i=1}^{N}i^2 - N\left(\frac{N+1}{2}\right)^2\right\} = \beta^2\frac{N(N^2-1)}{12}$

$$\text{Var}(\bar{y}_{lin}) = \frac{N^2(N-n)}{Nn} \cdot \frac{1}{N-1} \beta^2 \frac{N(N^2-1)}{12} = N^2\beta^2 \frac{(K-1)(nK+1)}{12}$$

using $N = nK$.

Again, $\sum_{r=1}^{K} [\bar{Y}_r - \bar{Y}]^2 = \sum_{r=1}^{K} N^2\beta^2 \left[r - \frac{K+1}{2}\right]^2 = N^2\beta^2 \cdot \frac{K(K^2-1)}{12}$

[Since $\bar{y}_r = \frac{N}{n} \sum_{j=1}^{n} Y_{r+j-1K} = \frac{N}{n} \sum_{j=1}^{n} \{\alpha + \beta(r + j - 1K)\}$

$= N\left(\alpha + \beta r + \beta K \frac{n-1}{2}\right)$ and $\bar{Y} = \sum_{i=1}^{N} Y_i = N\alpha + \beta \frac{N(N+1)}{2}$

Then $\bar{y}_r - \bar{Y} = \beta N \left[r + \frac{nK - K - nK - 1}{2}\right] = N\beta\left[r - \frac{K+1}{2}\right]$]

Therefore, $\text{Var}(\bar{y}_{sy}) = N^2\beta^2 \frac{(K^2-1)}{12}$.

Note that $\frac{\text{Var}(\bar{y}_{sy})}{\text{Var}(\bar{y}_{srs})} = \frac{K^2-1}{(K-1)(nK+1)} = \frac{K+1}{nK+1} < 1$, for all $n > 1$.

Hence, the linear systematic sampling is more precise than SRS in the presence of linear trend.

Estimation of Variance in Systematic Sampling:

Problems in estimation:

From the results of a simple random sample with $n > 1$, we can calculate an unbiased estimator of the variance of the sample mean, the estimator being unbiased whatever the form of the popln. Since a systematic sample can be regarded as a simple random sample with $n = 1$, hence the above useful property does not hold for the systematic sample. In other words, since the variance of $\bar{y}_{sys}$ in systematic sampling is $V(\bar{y}_r)$ where $\bar{y}_r$ is the value of $\bar{y}_{sys}$ corresponding to the random start r, an unbiased estimator of $V(\bar{y}_{sys})$ can not be obtained by making just one observation on $\bar{y}_r$, that is, by taking just one systematic sample.

[It has been pointed out earlier that $\pi_i = \frac{1}{K}$ and π_{ij}

$= \begin{cases} \frac{1}{K}, & \text{if } U_i \text{ and } U_j \text{ are in the same cluster or group.} \\ 0, & \text{if } U_i \text{ and } U_j \text{ are in the different clusters or groups; that is,} \end{cases}$

in linear systematic sampling the second order inclusion probabilities are not positive for all pairs of units in the population this makes unbiased estimation of the variance of the estimator.

A way out is to make use of the method interpenetrating Subsamples.

Interpenetrating Subsamples:

This technique, particularly useful for the study of correlated errors, was proposed by Mahalanobis (1946). To present it in the simplest terms, a random sample of n units is divided at random into k_1 subsamples, each subsample containing $m = \frac{n}{k_1}$ units. The field work and processing of the sample are planned so that there is no correlation between the errors of measurement of any two units in different subsamples. For instance, suppose that the correlation with which we have to deal arises solely from biases of the interviewers. If each of k_1 interviewers is assigned to a different subsample and if there is no correlation between errors of measurement for different interviewers, we have an example of the technique.

Consider the mathematical model for errors of measurement:

Let y_{ijr} be the value obtained in the r^{th} repetition of the j^{th} member within the i^{th} subsample (interviewer). Then

$y_{ijr} = \mu_{ij'} + d_{ijr}$, where $\mu_{ij'}$ is the true mean of the unit and d_{ijr} is the response deviation on the unit or the fluctuating component of the measurement error.

[Since the i^{th} subsample is a random subsample, it is itself a simple random sample of size m. Hence, the variance of its mean is $V(\bar{y}_{i\cdot}) = \frac{1-f}{m} S_{\mu'}^2 + \frac{\sigma_d^2}{m}[1 + (m-1)\rho_w]$, assuming $\mathrm{cov}(\bar{d}_{i\cdot}, \bar{\mu}') = 0$, where ρ_w is the correlation coefficient between the d_{ijr} obtained by the same interviewer (subsample).

If $\frac{1-f}{m}$ is ignored, $V(\bar{y}_{i\cdot}) = \frac{1}{m}\left[\sigma_{\mu'}^2 + \sigma_d^2\{1 + (m-1)\rho_w\}\right]$

Since errors are independent in the different subsample

$$V(\bar{y}_{\cdot\cdot}) = \frac{1}{k_1} V(\bar{y}_{i\cdot}) = \frac{1}{n}\left[S_{\mu'}^2 + \sigma_d^2\{1 + (m-1)\rho_w\}\right]$$

From the sample results we can compute an ANOVA of the kn observations into components Between "Subsamples (interviewers)" with $(k_1 - 1)$ d.f.'s and "Within interviewers" with $k(m-1)$ d.f.'s.

ANOVA table [on a single unit basis]

Sources of Variation	d.f	m.s.	E(m.s)
Between interviewers (Subsamples)	K_1-1	$s_b^2 = \frac{m}{K_1-1}\sum_{i=1}^{K_1}(\bar{y}_{ir}-\bar{y}_r)^2$	$S\mu^2 + \sigma_1^2\{1+(m-1)\rho_w\}$
Within Subsamples (interviewers)	$K_1(m-1)$	$s_w^2 = \frac{1}{K_1(m-1)}\sum_{i=1}^{K_1}\sum_{j=1}^{m}(y_{ijr}-\bar{y}_{ir})^2$	$S\mu^2 + \sigma_1^2(1-\rho_w)$
Total	K_1m-1		

Table gives the important result: $\frac{s_b^2}{n}$ is an unbiased estimator of $\frac{1}{n}E(s_b^2) = V(\bar{y}_r)$, ignoring fpc. Thus interpenetrat Subsamples provide an estimator of $V(\bar{y}_r)$ that takes proper account of both the simple response variance and correlated component. Note that $\frac{s_b^2}{n} = \frac{m}{n(K_1-1)}\sum_{i=1}^{K_1}(\bar{y}_{ir}-\bar{y}_r)^2 = \frac{1}{K_1(K_1-1)}\sum_{i=1}^{K_1}(y_{ir}-\bar{y}_r)^2$

Here m subsamples are interpenetrating in the sense that each is a probability sample over the population.

In linear systematic sampling, $V(\bar{y}_{Lss}) = V(\hat{Y}_r) = V(N\bar{y}_r)$
$= N^2V(\bar{y}_r)$. Hence, by interpenetrating subsamples, an estimator of $V(\bar{y}_{Lss})$ is $v(\bar{y}_{Lss}) = N^2.\frac{s_b^2}{n} = N.K.s_b^2$,

$$= N^2.\frac{1}{K_1(K_1-1)}\sum_{i=1}^{K_1}(\bar{y}_{ir}-\bar{y}_r)^2.$$

— * — * —

Comparison of Linear Systematic with Stratified random Sampling:

Linear Systematic Sampling stratifies the population into n strata, which consist of the first K units, the second K units, and so on. We might therefore expect the linear systematic sample to be about as precise as the corresponding stratified random sample with one unit per stratum. The differen is that with the systematic sample the units occur at the same relative position in the stratum, whereas with the stratified random sample the position in the stratum is determined separately by randomisation within each stratum. The systematic sample is spread more evenly over the population and this fact has sometimes made systematic sampling considerably more precise than stratified random sampling.

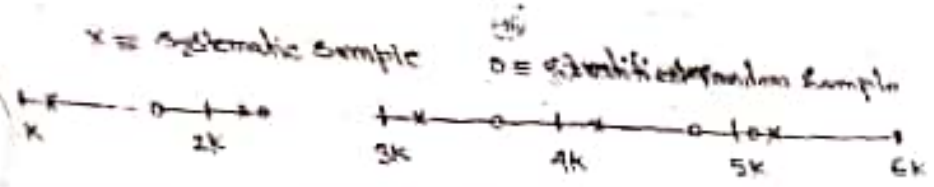

Fig: Systematic and Stratified random Sampling.

The performance of linear systematic sampling in relation to that of stratified on SRS is greatly dependent on the properties of the population. For some populations and some values of n, $V(\bar{y}_{sy}) = \frac{S^2}{n}\frac{(N-1)}{N}\{1+(n-1)\rho_w\}$ may even increase when a large sample is taken — even a small positive correlation may have a large effect because of the multiplier $(n-1)$.

Thus it is difficult to give advice about the situations in which systematic sampling is to be recommended — a knowledge of the structure of the population is necessary for its most effective use.

Circular Systematic Sampling:

Since N is not ingeneral an integral multiple of k, different systematic samples from the same finite population may vary by one unit in size. Thus, with N = 23, K = 5, the members of the units in the five systematic samples are shown in the table:

Systematic Sample number				
I	II	III	IV	V
1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	—	—

The first 3 samples have n=5 ~~and the last two have~~ n = 4. This fact introduces a disturbance into the theory of systematic sampling. (The disturbance is probably negligible if n > 50 and will be ignored, for simplicity.). In some cases, some units of the population will never appear in the sample. Thereby the estimation of the popln total (mean) becomes impossible. For example, when N = 30, n = 7, K = 4, the units labels 29 and 30 will never appear as sampled units.

These problems can be overcome by [illegible] a method, known as circular systematic sampling (C.S.S.) [illegible]

This method consists in choosing a random start from 1 to N and selecting the unit corresponding to the random start and thereafter every k^{th} unit in cyclical manner till a sample of size n units is obtained, k being the integer nearest to $\left(\frac{N}{n}\right)$. That is, if r is the number selected at random from 1 to N, the sample consists of the units corresponding to the numbers $r+jK$ if $r+jK \leq N$ and $(r+jK-N)$, if $r+jk > N$, for $j=0,1,\ldots,n-1$. It is to be noted that, if the sampling interval is taken as the integer closest to $\frac{N}{n}$, it is not always possible to get a sample of the given size as shown in the following example. Let $N=15$, $n=6$ and $k=3$. The sample with the random start 3 has only five distinct units namely 3, 6, 9, 12, 15.

Under circular systematic sampling, the conventional expansion estimator $\hat{Y}_{css} = \frac{N}{n}\sum_{j=1}^{n} y_j^i$, where y_j^i being the value of the j^{th} unit in CSS corresponding to the random start i, is unbiased for population total $Y = \sum_{i=1}^{N} Y_i$ (whenever N and K are relatively coprime) and its variance is given by

$$V(\hat{Y}_{css}) = \frac{1}{N}\sum_{i=1}^{N}[\hat{Y}_{ci} - Y]^2.$$