Introduction to Sampling Theory

Lecture 38
Systematic Sampling

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Systematic Sampling:

The systematic sampling technique is operationally more convenient than the simple random sampling. It also ensures at the same time that each unit has equal probability of inclusion in the sample.

In this method of sampling, the first unit is selected with the help of random numbers and the remaining units are selected automatically according to a predetermined pattern.

This method is known as systematic sampling.

Systematic Sampling:

Suppose the units in the population are numbered 1 to N in some order. Suppose further that N is expressible as a product of two integers n and k, so that $N = nk$.

To draw a sample of size n,

- **select a random number between 1 and k.**
- **Suppose it is i.**
- **Select the first unit whose serial number is i.**
- **Select every k^{th} unit after i^{th} unit.**
- **Sample will contain $i, i + k, , i + 2k, \ldots, , i + (n - 1)k$ serial number units.**

Systematic Sampling:

So first unit is selected at random and other units are selected systematically.

This systematic sample is called k^{th} systematic sample and k is termed as sampling interval.

This is also known as the linear systematic sampling.

Systematic Sampling:

The observations in the systematic sampling are arranged as in the following table:

Systematic Sample number		1	2	3	...	*i*	...	*k*
Sample composition	1	y_1	y_2	y_3	$\cdots$	y_i	$\cdots$	y_k
	2	y_{k+1}	y_{k+2}	y_{k+3}	$\cdots$	y_{k+i}	$\cdots$	y_{2k}
	. . .	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n	$y_{(n-1)k+1}$	$y_{(n-1)k+2}$	$y_{(n-1)k+3}$	$\cdots$	$y_{(n-1)k+i}$	$\cdots$	y_{nk}
Probability		$\frac{1}{k}$	$\frac{1}{k}$	$\frac{1}{k}$	$\cdots$	$\frac{1}{k}$	$\cdots$	$\frac{1}{k}$
Sample mean		$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\cdots$	$\bar{y}_i$	$\cdots$	$\bar{y}_k$

Systematic Sampling: Example

Let N = 50 and n = 5. So k = 10.

Suppose first selected number between 1 and 10 is 3.

Then systematic sample consists of units with following serial number 3, 13, 23, 33, 43.

Advantages of Systematic Sampling:

1. It is easier to draw a sample and often easier to execute it without mistakes. This is more advantageous when the drawing is done in fields and offices as there may be substantial saving in time.

2. The cost is low and the selection of units is simple. Much less training is needed for surveyors to collect units through systematic sampling.

Advantages of Systematic Sampling:

3. **The systematic sample is spread more evenly over the population.**

   So no large part will fail to be represented in the sample.

   The sample is evenly spread and cross section is better.

   Systematic sampling fails in case of too many blanks.

Relation to the Cluster Sampling:

The systematic sample can be viewed from the cluster sampling point of view.

With $N = nk$, there are k possible systematic samples.

The same population can be viewed as if divided into k large sampling units, each of which contains n of the original units.

Relation to the Cluster Sampling:

The operation of choosing a systematic sample is equivalent to choosing one of the large sampling unit at random which constitutes the whole sample.

A systematic sample is thus a simple random sample of one cluster unit from a population of *k* cluster units.

Estimation of Population Mean : When *N = nk*

Let y_{ij} : observation on the unit bearing the serial number

$i + (j - 1)k$ in the population, $i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, n.$

Suppose the drawn random number is $i \leq k$.

Sample consists of ith column (in earlier table).

Consider the sample mean

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

as an estimator of the population mean given by

$$\bar{Y} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{k}\sum_{i=1}^{k}\bar{y}_i.$$

Estimation of Population Mean: When *N = nk*

Probability of selecting *i^th^* column as systematic sample $= \frac{1}{k}$.

So

$$E(\bar{y}_{sy}) = \frac{1}{k}\sum_{i=1}^{k}\bar{y}_i = \bar{Y}.$$

Thus $\bar{y}_{sy}$ **is an unbiased estimator of** $\bar{Y}$.

Estimation of Population Mean: When *N = nk*

Further,

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

Consider

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

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

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2 + n\sum_{i=1}^{k}(\bar{y}_i - \bar{Y})^2$$

$$= k(n-1)S_{wsy}^2 + n\sum_{i=1}^{k}(\bar{y}_i - \bar{Y})^2$$

where $S_{wsy}^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$ **is the variation among the units that lie within the same systematic sample.**

Estimation of Population Mean: When *N = nk*

Thus

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

$$= \underbrace{\frac{N-1}{N} S^2}_{\text{Variation as a whole}} - \underbrace{\frac{(n-1)}{n} S^2_{wsy}}_{\text{Pooled within variation of the } k \text{ systematic sample}}$$

This expression indicates that when the within variation is large, then $Var(\bar{y}_i)$ becomes smaller.

Thus higher heterogeneity makes the estimator more efficient and higher heterogeneity in well expected is systematic sample.

Alternative Form of Variance in terms of Intraclass Correlation:

$$Var(\bar{y}_{sy}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i - \bar{Y})^2$$

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

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

$$= \frac{1}{kn^2}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{ij} - \bar{Y})^2 + \sum_{j(\neq\ell)=1}^{n}\sum_{\ell=1}^{n}(y_{ij} - \bar{Y})(y_{i\ell} - \bar{Y})\right]$$

$$= \frac{1}{kn^2}\left[(nk-1)S^2 + \sum_{i=1}^{k}\sum_{j(\neq\ell)=1}^{n}\sum_{\ell=1}^{n}(y_{ij} - \bar{Y})(y_{i\ell} - \bar{Y})\right].$$

Alternative Form of Variance in terms of Intraclass Correlation:

The intraclass correlation between the pairs of units that are in the same systematic sample is

$$\rho_w = \frac{E(y_{ij}-\bar{Y})(y_{i\ell}-\bar{Y})}{E(y_{ij}-\bar{Y})^2}; \qquad -\frac{1}{nk-1} \le \rho_w \le 1$$

$$= \frac{\frac{1}{nk(n-1)}\sum_{i=1}^{k}\sum_{j(\neq\ell)=1}^{n}\sum_{\ell=1}^{n}(y_{ij}-\bar{Y})(y_{i\ell}-\bar{Y})}{\left(\frac{nk-1}{nk}\right)S^2}.$$

So substituting

$$\sum_{i=1}^{k}\sum_{j(\neq\ell)=1}^{n}\sum_{\ell=1}^{n}(y_{ij}-\bar{Y})(y_{i\ell}-\bar{Y}) = (n-1)(nk-1)\rho_w S^2$$

in $Var(\bar{y}_{sy})$ **gives**

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

Comparison with SRSWOR:

For a SRSWOR sample of size *n*,

$$Var(\bar{y}_{SRS}) = \frac{N-n}{Nn} S^2$$

$$= \frac{nk-n}{Nn} S^2$$

$$= \frac{k-1}{N} S^2.$$

and

$$Var(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{n-1}{n} S_{wsy}^2.$$

Comparison with SRSWOR:

Since *N* = *nk*, so

$$Var(\bar{y}_{SRS}) - Var(\bar{y}_{sy}) = \left(\frac{k-1}{N} - \frac{N-1}{N}\right) S^2 + \frac{n-1}{n} S_{wsy}^2$$

$$= \frac{n-1}{n}(S_{wsy}^2 - S^2).$$

Thus $\bar{y}_{sy}$ **is**

- **more efficient than** $\bar{y}_{SRS}$ **when** $S_{wsy}^2 > S^2$.
- **less efficient than** $\bar{y}_{SRS}$ **when** $S_{wsy}^2 < S^2$.
- **equally efficient as** $\bar{y}_{SRS}$ **when** $S_{wsy}^2 = S^2$.

Comparison with SRSWOR:

Also, the relative efficiency of $\bar{y}_{sy}$ relative to $\bar{y}_{SRS}$ is

$$RE = \frac{Var(\bar{y}_{SRS})}{Var(\bar{y}_{sy})} = \frac{\frac{N-n}{Nn} S^2}{\frac{N-1}{Nn} S^2 \left[1+\rho_w(n-1)\right]}$$

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

$$= \frac{n(k-1)}{(nk-1)}\left[\frac{1}{1+\rho_w(n-1)}\right]; \qquad -\frac{1}{nk-1} \le \rho_w \le 1.$$

Thus $\bar{y}_{sy}$ is

- **more efficient than $\bar{y}_{SRS}$ when** $\rho_w < -\frac{1}{nk-1}$.
- **less efficient than $\bar{y}_{SRS}$ when** $\rho_w > -\frac{1}{nk-1}$.
- **equally efficient as $\bar{y}_{SRS}$ when** $\rho_w = -\frac{1}{nk-1}$.