Chapter 11

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.

Suppose the N 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 $i, i+k, 1+2k, ..., i+(n-1)k$ serial number units.

So first unit is selected at random and other units are selected systematically. This systematic sample is called $\boldsymbol{k^{th}}$ **systematic sample** and k is termed as **sampling interval.** This is also known as **linear 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	⋯	y_i	⋯	y_k
	2	y_{k+1}	y_{k+2}	y_{k+3}	⋯	y_{k+i}	⋯	y_{2k}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	n	$y_{(n-1)k+1}$	$y_{(n-1)k+2}$	$y_{(n-1)k+3}$	…	$y_{(n-1)k+i}$	…	y_{nk}
Probability		$\frac{1}{k}$	$\frac{1}{k}$	$\frac{1}{k}$	⋯	$\frac{1}{k}$	⋯	$\frac{1}{k}$
Sample mean		$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	⋯	$\bar{y}_i$	⋯	$\bar{y}_k$

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.

Systematic sampling in two dimensions:

Assume that the units in a population are arranged in the form of $m\ell$ rows and each row contains nk units. A sample of size mn is required. Then

- select a pair of random numbers (i, j) such that $i \le \ell$ and $j \le k$.
- Select the $(i, j)^{th}$ unit, i.e., j^{th} unit in i^{th} row as the first unit.
- Then the rows to be selected are

  $$i,\ i+\ell, i+2\ell, ..., i+(m-1)\ell$$

  and columns to be selected are

  $$j,\ j+k,\ j+2k, ...,\ j+(n-1)k.$$
- The points at which the m selected rows and n selected columns intersect determine the position of mn selected units in the sample.

Such a sample is called an **aligned sample.**

Alternative approach to select the sample is

- independently select n random integers $i_1, i_2, ..., i_n$ such that each of them is less than or equal to ℓ.
- Independently select m random integers $j_1, j_2, ..., j_m$ such that each of them is less than or equal to k.
- The units selected in the sample will have following coordinates: $(i_1 + r\ell,\ j_{r+1}), (i_2 + r\ell,\ j_{r+1} + k), (i_3 + r\ell,\ j_{r+1} + 2k), ..., (i_n + r\ell,\ j_{r+1} + (n-1)k)$.

Such a sample is called an **unaligned sample.**

Under certain conditions, an unaligned sample is often superior to an aligned sample as well as a stratified random sample.

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 .
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. 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,

$i = 1, 2, ..., k,\ j = 1, 2, ..., n.$

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

Sample consists of i^{th} column (in earlier table).

Consider the sample mean given by

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

as an estimator of the population mean given by

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

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

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 $\bar{Y}$.

Further,

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

Consider

$$\begin{aligned}(N-1)S^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{Y})^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n}\left[(y_{ij}-\bar{y}_i)+(\bar{y}_i-\bar{Y})\right]^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2 + n\sum_{i=1}^{k}(\bar{y}_i-\bar{Y})^2 \\ &= k(n-1)S_{wsy}^2 + n\sum_{i=1}^{k}(\bar{y}_i-\bar{Y})^2\end{aligned}$$

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 . Thus

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

↓	↓
Variation as a whole	Pooled within variation of the k systematic sample

with $N = nk$. 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 is well expected in systematic sample.

Alternative form of variance:

$$
\begin{aligned}
Var(\bar{y}_{sy}) &= \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i - \bar{Y})^2 \\
&= \frac{1}{k}\sum_{i=1}^{k}\left[\frac{1}{n}\sum_{j=1}^{n} y_{ij} - \bar{Y}\right]^2 \\
&= \frac{1}{kn^2}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{ij} - \bar{Y})\right] \\
&= \frac{1}{kn^2}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{ij} - \bar{Y})^2 + \sum_{j(\neq\ell)=1}^{n}\sum_{\ell=1}^{n}(y_{ij} - \bar{Y})(y_{i\ell} - \bar{Y})\right] \\
&= \frac{1}{kn^2}\left[(nk-1)S^2 + \sum_{i=1}^{k}\sum_{j(\neq\ell)=1}^{n}\sum_{\ell=1}^{n}(y_{ij} - \bar{Y})(y_{i\ell} - \bar{Y})\right].
\end{aligned}
$$

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

$$
\begin{aligned}
\rho_w &= \frac{E(y_{ij} - \bar{Y})(y_{i\ell} - \bar{Y})}{E(y_{ij} - \bar{Y})^2}; \qquad -\frac{1}{nk-1} \le \rho \le 1 \\
&= \frac{\dfrac{1}{nk(n-1)}\displaystyle\sum_{i=1}^{k}\sum_{j(\neq\ell)=1}^{n}\sum_{\ell=1}^{n}(y_{ij} - \bar{Y})(y_{i\ell} - \bar{Y})}{\left(\dfrac{nk-1}{nk}\right)S^2}.
\end{aligned}
$$

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 $Var(\bar{y}_i)$ gives

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

Comparison with SRSWOR:

For a SRSWOR sample of size n,

$$
\begin{aligned}
Var(\bar{y}_{SRS}) &= \frac{N-n}{Nn}S^2 \\
&= \frac{nk-n}{Nn}S^2 \\
&= \frac{k-1}{N}S^2.
\end{aligned}
$$

Since

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

$$N = nk$$

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

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

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

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

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

$$RE = \frac{Var(\bar{y}_{SRS})}{Var(\bar{y}_{sy})}$$

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

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

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

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 $\rho_w = -\frac{1}{nk-1}$.

Comparison with stratified sampling:

The systematic sample can also be viewed as if arising as a stratified sample. If population of $N = nk$ units

is divided into n strata and suppose one unit is randomly drawn from each of the strata. Then we get a stratified sample of size n. In doing so, just consider each row of the following arrangement as a stratum.

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

Recall that in case of stratified sampling with k strata, the stratum mean

$$\bar{y}_{st} = \frac{1}{N}\sum_{j=1}^{k} N_j \bar{y}_j$$

is an unbiased estimator of population mean.

Considering the set up of stratified sample in the set up of systematic sample, we have

- Number of strata = n
- Size of strata = k (row size)
- Sample size to be drawn from each stratum = 1

and $\bar{y}_{st}$ becomes

$$\bar{y}_{st} = \frac{1}{nk}\sum_{j=1}^{n} k\bar{y}_j$$
$$= \frac{1}{n}\sum_{j=1}^{n} \bar{y}_j$$

$$
\begin{aligned}
Var(\bar{y}_{st}) &= \frac{1}{n^2}\sum_{j=1}^{n} Var(\bar{y}_j) \\
&= \frac{1}{n^2}\sum_{j=1}^{n} \frac{k-1}{k.1} S_j^2 \quad \left(\text{using } Var(\bar{y}_{SRS}) = \frac{N-n}{Nn} S^2 \right) \\
&= \frac{k-1}{kn^2}\sum_{j=1}^{n} S_j^2 \\
&= \frac{k-1}{nk} S_{wst}^2 \\
&= \frac{N-n}{Nn} S_{wst}^2
\end{aligned}
$$

where

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

is the mean sum of squares of units in the j^{th} stratum.

$$
\begin{aligned}
S_{wst}^2 &= \frac{1}{n}\sum_{j=1}^{n} S_j^2 \\
&= \frac{1}{n(k-1)}\sum_{i=1}^{k}\sum_{j=1}^{n} (y_{ij} - \bar{y}_j)^2
\end{aligned}
$$

is the mean sum of squares within strata (or rows).

The variance of systematic sample mean is

$$
\begin{aligned}
Var(\bar{y}_{sy}) &= \frac{1}{k}\sum_{i=1}^{k} (\bar{y}_i - \bar{Y})^2 \\
&= \frac{1}{k}\sum_{i=1}^{k} \left[\frac{1}{n}\sum_{j=1}^{n} y_{ij} - \frac{1}{n}\sum_{j=1}^{n} \bar{y}_j \right]^2 \\
&= \frac{1}{n^2 k}\sum_{i=1}^{k} \left[\sum_{j=1}^{n} (y_{ij} - \bar{y}_j) \right]^2 \\
&= \frac{1}{n^2 k} \left[\sum_{i=1}^{k}\sum_{j=1}^{n} (y_{ij} - \bar{y}_j)^2 + \sum_{i=1}^{k}\sum_{j\neq \ell=1}^{n} (y_{ij} - \bar{y}_j)(y_{i\ell} - \bar{y}_\ell) \right].
\end{aligned}
$$

Now we simplify and express this expression in terms of intraclass correlation coefficient. The intraclass correlation coefficient between the pairs of deviations of units which lie along the same row measured from their stratum means is defined as

$$\rho_{wst} = \frac{\frac{1}{nk(n-1)}\sum_{i=1}^{k}\sum_{j\neq\ell=1}^{n}(y_{ij}-\bar{y}_j)(y_{i\ell}-\bar{y}_\ell)}{\frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_j)^2}$$

$$= \frac{\sum_{i=1}^{k}\sum_{j\neq\ell=1}^{n}(y_{ij}-\bar{y}_j)(y_{i\ell}-\bar{y}_\ell)}{(N-n)(n-1)S_{wst}^2}$$

So

$$Var(\bar{y}_{sy}) = \frac{1}{n^2k}\left[(N-n)S_{wst}^2 + (N-n)(n-1)\rho_{wst}S_{wst}^2\right]$$

$$= \frac{N-n}{Nn}S_{wst}^2\left[1+(n-1)\rho_{wst}\right]. \text{ (using } N=nk)$$

Thus

$$Var(\bar{y}_{st}) - Var(\bar{y}_{sy}) = -\frac{N-n}{Nn}(n-1)\rho_{wst}S_{wst}^2$$

and the relative efficiency of systematic sampling relative to equivalent stratified sampling is given by

$$RE = \frac{1}{1+(n-1)\rho_{wst}}.$$

So the systematic sampling is

- more efficient than the corresponding equivalent stratified sample when $\rho_{wst} > 0$.
- less efficient than the corresponding equivalent stratified sample when $\rho_{wst} < 0$
- equally efficient than the corresponding equivalent stratified sample when $\rho_{wst} = 0.$

Comparison of systematic sampling, stratified sampling and SRS with population with linear trend:

We assume that the values of units in the population increase according to linear trend.

So the values of successive units in the population increase in accordance with a linear model so that

$$y_i = a + bi,\ i = 1, 2, ..., N.$$

Now we determine the variances of $\bar{y}_{SRS}, \bar{y}_{sy}$ and $\bar{y}_{st}$ under this linear trend.

Under SRSWOR

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

Here $N = nk$

$$\bar{Y} = a + b\frac{1}{N}\sum_{i=1}^{N} i$$

$$= a + b\frac{1}{N}\frac{N(N+1)}{2}$$

$$= a + b\frac{N+1}{2}$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left[a + bi - a - b\frac{N+1}{2}\right]^2$$

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

$$= \frac{b^2}{N-1}\left[\sum_{i=1}^{N} i^2 - N\left(\frac{N+1}{2}\right)^2\right]$$

$$= \frac{b^2}{N-1}\left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4}\right]$$

$$= b^2\frac{N(N+1)}{12}$$

$$Var(\bar{y}_{SRS}) = \frac{nk-n}{nk.n}\, b^2\frac{nk(nk+1)}{12}$$

$$= \frac{b^2}{12}(k-1)(nk+1).$$

Under systematic sampling

Earlier y_{ij} denoted the value of study variable with the j^{th} unit in the i^{th} systematic sample. Now y_{ij} represents the value of $\left[i+(j-1)k\right]^{th}$ unit of the population, so

$$y_{ij} = a + b\left[i+(j-1)k\right],\ i = 1,2,...,k;\ j = 1,2,...,n.$$

$$\overline{y}_{sy} = \frac{1}{k}\sum_{i=1}^{k}\overline{y}_i$$

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

$$\overline{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

$$= \frac{1}{n}\sum_{j=1}^{n}\left[a + b\left\{i+(j-1)k\right\}\right]$$

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

$$\sum_{i=1}^{k}(\overline{y}_i - \overline{Y})^2 = \sum_{i=1}^{k}\left[a + b\left(i + \frac{n-1}{2}k\right) - a - b\frac{nk+1}{2}\right]^2$$

$$= b^2\sum_{i=1}^{k}\left(i - \frac{k+1}{2}\right)^2$$

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

$$= b^2\left[\frac{k(k+1)(2k+1)}{6} + \left(\frac{k+1}{2}\right)^2 - (k+1)\frac{k(k+1)}{2}\right]$$

$$= \frac{b^2}{12}k(k^2-1)$$

$$Var(\overline{y}_{sy}) = \frac{1}{k}\frac{b^2}{12}k(k^2-1)$$

$$= \frac{b^2}{12}(k^2-1).$$

Under stratified sampling

$$y_{ij} = a + b\left[i + (j-1)k\right],\ i = 1,2,...,k,\ \ j = 1,2,...,n$$

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i$$

$$Var(\bar{y}_{st}) = \frac{N-n}{Nn} S_{wst}^2 = \frac{k-1}{nk} S_{wst}^2$$

where $S_{wst}^2 = \frac{1}{n}\sum_{j=1}^{n} S_j^2$

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

$$= \frac{1}{n(k-1)} \sum_{i=1}^{k}\sum_{j=1}^{n} \left[a + b\{i + (j-1)k\} - a - b\left\{\frac{k+1}{2} + (j-1)k\right\}\right]^2$$

$$= \frac{b^2}{n(k-1)} \sum_{i=1}^{k}\sum_{j=1}^{n} \left(i - \frac{k+1}{2}\right)^2$$

$$= \frac{b^2}{n(k-1)} \frac{nk(k^2-1)}{12}$$

$$= b^2 \frac{k(k+1)}{12}$$

$$Var(\bar{y}_{st}) = \frac{k-1}{nk} b^2 \frac{k(k+1)}{12}$$

$$= \frac{b^2}{12}\left(\frac{k^2-1}{n}\right)$$

If k is large, so that $\frac{1}{k}$ is negligible, then comparing $Var(\bar{y}_{st}), Var(\bar{y}_{sy})$ and $V(\bar{y}_{SRS})$,

$$Var(\bar{y}_{st}) \ : \ Var(\bar{y}_{sy}) \ : \ Var(\bar{y}_{SRS})$$

or $\quad \frac{k^2-1}{n} \ : \ k^2 - 1 \ : \ (k-1)(1+nk)$

or $\quad \frac{k+1}{n} \ : \ k+1 \ : \ nk+1$

or $\quad \frac{k+1}{n(k+1)} \ : \ \frac{k+1}{k+1} \ : \ \frac{nk+1}{k+1}$

$\simeq \quad \frac{1}{n} \qquad 1 \ : \ n$

Thus

$$Var(\bar{y}_{st}) \ : \ Var(\bar{y}_{sy}) \ : \ Var(\bar{y}_{SRS}) \ :: \ \frac{1}{n} \ : \ 1 \ : \ n$$

So stratified sampling is best for linearly trended population. Next best is systematic sampling.

Estimation of variance:

As such there is only one cluster, so variance in principle, cannot be estimated.

Some approximations have been suggested.

1. Treat systematic sample as if it were a random sample. In this case, an estimate of variance is

$$\widehat{Var}(\bar{y}_{sy}) = \left(\frac{1}{n} - \frac{1}{nk}\right) s_{wc}^2$$

where $s_{wc}^2 = \frac{1}{n-1}\sum_{j=0}^{n-1}(y_{i+jk} - \bar{y}_i)^2.$

This estimator under-estimates the true variance.

2. Use of successive differences of the values gives the estimate of variance as

$$\widehat{Var}(\bar{y}_{sy}) = \left(\frac{1}{n} - \frac{1}{nk}\right)\frac{1}{2(n-1)}\sum_{j=0}^{n-1}\left(y_{i+jk} - y_{i+(j+1)k}\right)^2.$$

This estimator is a biased estimator of true variance.

3. Use the balanced difference of $y_1, y_2, ..., y_n$ to get the estimate of variance as

$$\widehat{Var}(\bar{y}_{sy}) = \left(\frac{1}{n} - \frac{1}{nk}\right)\frac{1}{5(n-2)}\sum_{i}^{n-2}\left[\frac{y_i}{2} - y_{i+1} + \frac{y_{i+2}}{2}\right]^2$$

or

$$\widehat{Var}(\bar{y}_{sy}) = \left(\frac{1}{n} - \frac{1}{nk}\right)\frac{1}{15(n-4)}\sum_{i}^{n-4}\left[\frac{y_i}{2} - y_{i+1} + y_{i+2} - y_{i+3} + \frac{y_{i+4}}{2}\right]^2.$$

4. The interpenetrating subsamples can be utilized by dividing the sample into C groups each of size $\frac{n}{c}$. Then the group means are $\bar{y}_1, \bar{y}_2, ..., \bar{y}_c$. Now find

$$\bar{y} = \frac{1}{c}\sum_{t=1}^{c}\bar{y}_t$$

$$\widehat{Var}(\bar{y}_{sy}) = \frac{1}{c(c-1)}\sum_{t=1}^{c}(y_t - \bar{y})^2.$$

Systematic sampling when $N \neq nk$.

When N is not expressible as nk then suppose N can be expressed as

$$N = nk + p;\ p < k.$$

Then consider the following sample mean as an estimator of population mean

$$\bar{y}_{sy} = \bar{y}_i = \begin{cases} \dfrac{1}{n+1}\sum_{j=1}^{n+1} y_{ij} & \text{if } i \le p \\ \dfrac{1}{n}\sum_{j=1}^{n} y_{ij} & \text{if } i > p. \end{cases}$$

In this case

$$E(\bar{y}_i) = \frac{1}{k}\left[\sum_{i=1}^{p}\left(\frac{1}{n+1}\sum_{j=1}^{n+1} y_{ij}\right) + \sum_{i=p+1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n} y_{ij}\right)\right]$$
$$\neq \bar{Y}.$$

So $\bar{y}_{sy}$ is a biased estimator of $\bar{Y}$.

An unbiased estimator of $\bar{Y}$ is

$$\bar{y}^*_{sy} = \frac{k}{N}\sum_j y_{ij}$$
$$= \frac{k}{N} C_i$$

where $C_i = n\bar{y}_i$ is the total of values of the i^{th} column.

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

$$Var(\bar{y}^*_{sy}) = \frac{k^2}{N^2}\left(\frac{k-1}{k}\right) S^{*2}_c$$

where $S^{*2}_c = \dfrac{1}{k-1}\sum_{i=1}^{k}\left(n\bar{y}_i - \dfrac{N\bar{Y}}{k}\right)^2$.

Now we consider another procedure which is opted when $N \neq nk$.

[Reference: Theory of Sample Surveys, A.K. Gupta, D.G. Kabe, 2011, World Scientific Publishing Co.]

When population size N is not expressible as the product of n and k, then let

$$N = nq + r.$$

Then take the sampling interval as

$$k = \begin{cases} q & \text{if } r \le \frac{n}{2} \\ q+1 & \text{if } r > \frac{n}{2} \end{cases}.$$

Let $\left[\frac{M}{g}\right]$ denotes the largest integer contained in $\frac{M}{g}$.

If $k = q^*(= q \text{ or } q+1)$, then the

$$\text{number of units expected in sample} = \begin{cases} \left[\frac{N}{q^*}\right] & \text{with probability } \left[\frac{N}{q^*}\right] + 1 - \left(\frac{N}{q^*}\right) \\ \left[\frac{N}{q^*}\right] + 1 & \text{with probability } \left(\frac{N}{q^*}\right) - \left[\frac{N}{q^*}\right]. \end{cases}$$

If $q = q^*$, then we get

$$n^* = \begin{cases} n + \left[\frac{r}{q}\right] & \text{with probability } \left(\frac{r}{q}\right) + 1 - \left[\frac{r}{q}\right] \\ n + \left[\frac{r}{q}\right] + 1 & \text{with probability } \left(\frac{r}{q}\right) - \left[\frac{r}{q}\right] \end{cases}.$$

Similarly, if $q^* = q+1,$ then

$$n^* = \begin{cases} n - \left(\frac{n-r}{q+1}\right) & \text{with probability } \left[\frac{(n-r)}{(q+1)}\right] + 1 - \left(\frac{n-r}{q+1}\right) \\ n + \left[\left(\frac{n-r}{q+1}\right) + 1\right] & \text{with probability } \left(\frac{n-r}{q+1}\right) - \left[\frac{(n-r)}{(q+1)}\right]. \end{cases}$$

Example: Let $N = 17$ and $n = 5.$ Then $q = 3$ and $r = 2$. Since $r < \frac{n}{2}$, $k = q = 3.$

Then sample sizes would be

$$n^* = \begin{cases} n + \left[\frac{r}{q}\right] = 5 & \text{with probability } \left[\frac{r}{q}\right] + 1 - \left(\frac{r}{q}\right) = \frac{1}{3} \\ n + \left[\frac{r}{q}\right] + 1 = 6 & \text{with probability } \left(\frac{r}{q}\right) - \left[\frac{r}{q}\right] = \frac{2}{3}. \end{cases}$$

This can be verified from the following example:

Systematic sample number	Systematic sample	Probability
1	$Y_1, Y_4, Y_7, Y_{10}, Y_{13}, Y_{16}$	1/3
2	$Y_4, Y_5, Y_8, Y_{11}, Y_{14}, Y_{17}$	1/3
3	$Y_3, Y_6, Y_9, Y_{12}, Y_{15}$	1/3

We now prove the following theorem which shows how to obtain an unbiased estimator of the population mean when $N \neq nk$.

Theorem: In systematic sampling with sampling interval k from a population with size $N \neq nk$, an unbiased estimator of the population mean $\bar{Y}$ is given by

$$\hat{\bar{Y}} = \frac{k}{N}\left(\sum^{n'} y\right)_i$$

where i stands for the i^{th} systematic sample, $i = 1, 2, ..., k$ and n' denotes the size of i^{th} systematic sample.

Proof. Each systematic sample has probability $\frac{1}{k}$. Hence

$$E(\hat{\bar{Y}}) = \sum_{i=1}^{k} \frac{1}{k} . \frac{k}{N}\left(\sum^{n'} y\right)_i$$
$$= \frac{1}{N}\sum_{i=1}^{k}\left(\sum^{n'} y\right)_i .$$

Now, each unit occurs in only one of the k possible systematic samples. Hence

$$\sum_{i=1}^{k}\left(\sum^{n'} y\right)_i = \sum_{i=1}^{N} Y_i ,$$

which on substitution in $E(\hat{\bar{Y}})$ proves the theorem.

When $N \neq nk$, the systematic samples are not of the same size and the sample mean is not an unbiased estimator of the population mean. To overcome these disadvantages of systematic sampling when $N \neq nk$, circular systematic sampling is proposed. Circular systematic sampling consists of selecting a random number from 1 to N and then selecting the unit corresponding to this random number. Thereafter every k^{th} unit in a cyclical manner is selected till a sample of n units is obtained, k being the nearest integer to $\frac{N}{n}$.

In other words, if i is a number selected at random from 1 to N, then the circular systematic sample consists of units with serial numbers

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

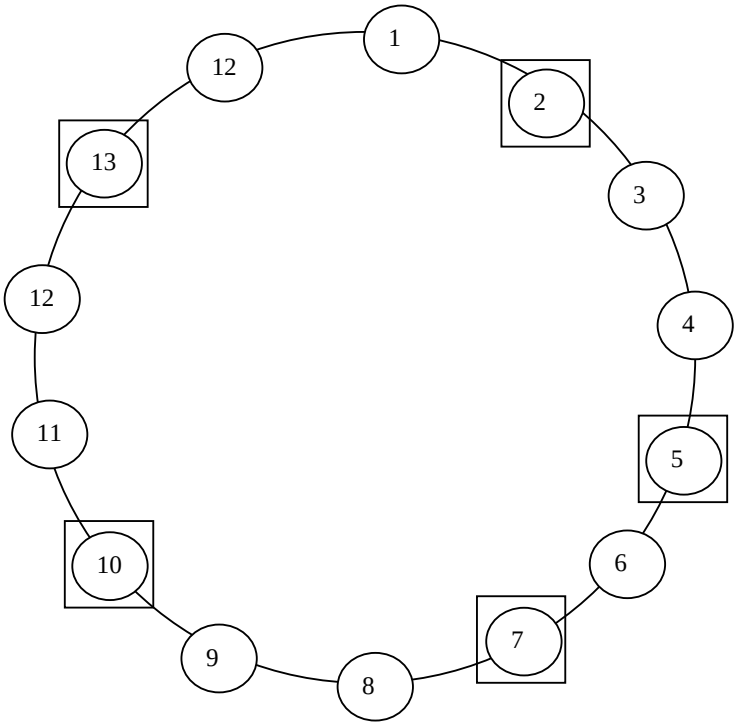
This sampling scheme ensures equal probability of inclusion in the sample for every unit.

Example:

Let $N=14$ and $n=5$. Then, $k=$ nearest integer to $\frac{14}{5}=3$. Let the first number selected at random from 1 to 14 be 7. Then, the circular systematic sample consists of units with serial numbers

7,10,13, 16-14=2, 19-14=5.

This procedure is illustrated diagrammatically in following figure.



Theorem: In circular systematic sampling, the sample mean is an unbiased estimator of the population mean.

Proof: If i is the number selected at random, then the circular systematic sample mean is

$$\bar{y} = \frac{1}{n}\left(\sum^{n} y\right)_i ,$$

where $\left(\sum^{n} y\right)_i$ denotes the total of y values in the i^{th} circular systematic sample, $i = 1, 2, ..., N$. We note here that in circular systematic sampling, there are N circular systematic samples, each having probability $\frac{1}{N}$ of its selection. Hence,

$$E(\bar{y}) = \sum_{i=1}^{N} \frac{1}{n}\left(\sum^{n} y\right)_i \times \frac{1}{N} = \frac{1}{Nn}\sum_{i=1}^{N}\left(\sum^{n} y\right)_i$$

Clearly, each unit of the population occurs in n of the N possible circular systematic sample means. Hence,

$$\sum_{i=1}^{N}\left(\sum^{n} y\right)_i = n\sum_{i=1}^{N} Y_i ,$$

which on substitution in $E(\bar{y})$ proves the theorem.

What to do when $N \neq nk$

One of the following possible procedures may be adopted when $N \neq nk$.

(i) Drop one unit at random if sample has $(n+1)$ units.

(ii) Eliminate some units so that $N = nk$.

(iii) Adopt circular systematic sampling scheme.

(iv) Round off the fractional interval k.