CHAPTER 4

Systematic Sampling

4.1 INTRODUCTION

In systematic sampling, the first unit is selected by using a random number, and the remaining units are selected by a predetermined pattern. Systematic sampling is very convenient for selection of a sample and hence it is used extensively in most practical situations. In this chapter we will discuss various types of systematic sampling procedures and methods of estimation of the population mean and its variances for such procedures. The efficiency of systematic sampling will also be compared with simple random and stratified sampling.

4.2 LINEAR SYSTEMATIC SAMPLING

Suppose we are to select a sample of size n from a population U of size N by a linear systematic sampling procedure. We consider the following two situations.

4.2.1 Linear Systematic Sampling With $N/n = k$ an Integer

We first select one unit, r, at random from 1 to k, then every kth unit. The initial unit r selected is called "random start" and k is called "sampling interval." Thus with the selection of a random start r $(=1,\ldots, k)$, a systematic sample $s_r = \{r, r+k, r+2k,\ldots, r+(n-1)k\}$ of size n is selected. The probability of selection of each of the k possible systematic samples is equal to $1/k$. The systematic samples partition the population U into mutually exclusive and exhaustive samples, i.e., $\bigcup_{r=1}^{k} s_r = U$ and $s_i \cap s_j = \phi$ for $i \neq j = 1,\ldots, k$. Since a unit can occur in one and only one sample, the inclusion probability of ith unit is $\pi_i = 1/k$ for $\forall\, i \in U$ and the inclusion probability for the ith and j $(\neq i)$th unit is $\pi_{ij} = 1/k$ if ith and jth unit belong to the same sample, and zero if the units do not belong to the same systematic sample. Thus for $N = 12$, and $n = 4$, we get $k = 12/4 = 3$ and we have k $(=3)$ possible systematic samples given in Table 4.2.1.

Here inclusion probabilities $\pi_i = 1/3$ for $i = 1,\ldots, 12$. Since the units 1 and 4, 1 and 7, and 1 and 10 belong to the same sample s_1, we get $\pi_{1,4} = \pi_{1,7} = \pi_{1,10} = 1/3$. On the other hand, the units 1 and 5, 1 and 6, and 1 and 8 do not belong to the same sample, hence $\pi_{1,5} = \pi_{1,6} = \pi_{1,8} = 0$.

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00004-2

Table 4.2.1 LSS with *N/n* is integer

Random start (*r*)	**1**	**2**	**3**
Sample (s_r)	$s_1 = (1, 4, 7, 10)$	$s_2 = (2, 5, 8, 11)$	$s_3 = (3, 6, 9, 12)$
Probability	1/3	1/3	1/3

4.2.2 Linear Systematic Sampling With *N/n* = *k* Not an Integer

In case N/n is not an integer, k is chosen as an integer close to N/n. In this situation we cannot always realize the fixed sample size n. For example, if $N = nk + t$ with $0 < t < k$, a systematic sample corresponding to random start r less than or equal to t produces sample of size $(n + 1)$, whereas a random start r greater than t produces a sample of size n. Since the probability of selection of each sample is $1/k$, the expected sample size is $(n+1)\dfrac{t}{k} + n\dfrac{k-t}{k} = \dfrac{N}{k}$. If $N = (n-1)k + t$ with $0 < t < k$, we get t systematic samples each of size n and $(k - t)$ samples each of size $(n - 1)$. For more clarity, let us consider Table 4.2.2.

Since the selection of a systematic sample is very easy and ensures even distribution of units over the population, it is used in most practical situations. Suppose an auditor decides to check 20 vouchers out of 100 vouchers. Here $N = 100$, $n = 20$, and $k = 100/20 = 5$. So, the auditor should select first a voucher at random from 1 to 5 ($=k$), then every fifth voucher. Systematic sampling can be used for selection of every kth row of plants in an agricultural experiment, checking of driver's license for every kth car on the road, reexamining every kth answer scripts, and so on.

Table 4.2.2 LSS with *N/n* not an integer

N	***n***	***k***	**Possible systematic samples**
13	4	3	$s_1 = (1, 4, 7, 10, 13)$, $s_2 = (2, 5, 8, 11)$, $s_3 = (3, 6, 9, 12)$
14	3	5	$s_1 = (1, 6, 11)$, $s_2 = (2, 7, 12)$, $s_3 = (3, 8, 13)$, $s_4 = (4, 9, 14)$, $s_5 = (5, 10)$

4.2.3 Estimation of the Population Mean and Its Variance

It has been discussed in Chapter 2 that the systematic sampling design is a unicluster sampling design. For a unicluster sampling design the Horvitz–Thompson estimator $\widehat{Y}_{ht} = \sum_{i \in s} \dfrac{y_i}{\pi_i}$ is the only unbiased estimator of the population total Y in the class of linear unbiased estimators and hence it is trivially the best (MVUE) unbiased estimator in that class. Since for systematic sampling scheme $\pi_i = 1/k$, the Horvitz–Thompson estimator, $\widehat{Y}_{ht} = k \sum_{i \in s_r} y_i$ based on the systematic sample s_r, is unbiased for the population total. Hence we have the following theorem.

Theorem 4.2.1

For a systematic sampling, $\widehat{\overline{Y}}_{\text{LSS}} = \dfrac{\hat{Y}_{ht}}{N} = \dfrac{k}{N} \sum_{i\in s_r} y_i$ is an unbiased estimator for the population mean $\overline{Y}$.

Corollary 4.2.1

In particular, when $N/n = k$ is an integer, $\widehat{\overline{Y}}_{\text{LSS}}$ reduces to $\dfrac{\hat{Y}_{ht}}{N} = \dfrac{1}{n}\sum_{i\in s_r} y_i = \overline{y}(s_r) =$ sample mean, which is an unbiased estimator for the population mean.

Theorem 4.2.2

For $N/n = k$ an integer, the variance of $\overline{y}(s_r)$ is given by

$$V_{sy} = V[\overline{y}(s_r)] = \frac{1}{N}(\text{TSS} - \text{WSS}) = \frac{\sigma_y^2}{n}[1 + (n-1)\rho]$$

where $\text{TSS} = \sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2 =$ total sum of square,

$\text{WSS} = \sum_{r=1}^{k}\sum_{j\in s_r}\left\{y_j - \overline{y}(s_r)\right\}^2 =$ sum of square within the systematic samples, $\rho = \dfrac{\sum_{r=1}^{k}\sum_{i\neq}\sum_{j\in s_r}\left(y_i - \overline{Y}\right)\left(y_j - \overline{Y}\right)}{kn(n-1)\sigma_y^2} =$ intraclass correlation between the units of the same systematic sample, and $\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2$.

Proof

$$\begin{aligned} V_{sy} &= V[\overline{y}(s_r)] \\ &= E\left[\overline{y}(s_r) - \overline{Y}\right]^2 \\ &= \frac{1}{k}\sum_{r=1}^{k}\left(\overline{y}(s_r) - \overline{Y}\right)^2 \end{aligned}$$

Now

$$\begin{aligned} \text{TSS} &= \sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2 \\ &= \sum_{r=1}^{k}\sum_{j\in s_r}\left(y_j - \overline{Y}\right)^2 \\ &= \sum_{r=1}^{k}\sum_{j\in s_r}\left\{y_j - \overline{y}(s_r) + \overline{y}(s_r) - \overline{Y}\right\}^2 \\ &= \sum_{r=1}^{k}\sum_{j\in s_r}\left\{y_j - \overline{y}(s_r)\right\}^2 + n\sum_{r=1}^{k}\left\{\overline{y}(s_r) - \overline{Y}\right\}^2 \\ &= SSW + NV[\overline{y}(s_r)] \end{aligned}$$

Hence

$$V_{sy} = (\text{TSS} - \text{WSS})/N \tag{4.2.1}$$

Now

$$\begin{aligned}
V_{sy} &= \frac{1}{k}\sum_{r=1}^{k}\left[\overline{y}(s_r) - \overline{Y}\right]^2 \\
&= \frac{1}{k}\sum_{r=1}^{k}\left[\frac{1}{n}\sum_{i\in s_r}\left(y_i - \overline{Y}\right)\right]^2 \\
&= \frac{1}{kn^2}\left[\sum_{r=1}^{k}\sum_{i\in s_r}\left(y_i - \overline{Y}\right)^2 + \sum_{r=1}^{k}\sum_{i\neq}\sum_{j\in s_r}\left(y_i - \overline{Y}\right)\left(y_j - \overline{Y}\right)\right] \\
&= \frac{\sigma_y^2}{n}\left[1 + (n-1)\rho\right]
\end{aligned} \tag{4.2.2}$$

4.2.4 Nonexistence of Unbiased Variance Estimator

The variance of $\overline{y}(s_r)$ can be written as

$$\begin{aligned}
V_{sy} &= E\left[\overline{y}(s_r) - \overline{Y}\right]^2 \\
&= \sum_{r=1}^{k}\left\{\overline{y}(s_r)\right\}^2/k - \overline{Y}^2 \\
&= \sum_{r=1}^{k}\left(\sum_{i\in s_r} y_i\right)^2\Big/(kn^2) - \left\{\sum_{r=1}^{k}\left(\sum_{i\in s_r} y_i\right)\right\}^2\Big/N^2 \\
&= \frac{1}{N}\left(\frac{1}{n} - \frac{1}{N}\right)\sum_{r=1}^{k}\left\{\sum_{i\in s_r} y_i^2 + \sum_{i\neq}\sum_{j\in s_r} y_i y_j\right\} \\
&\quad - \frac{1}{N^2}\sum_{r\neq}^{k}\sum_{r'=1}^{k}\left(\sum_{j\in s_r} y_i\right)\left(\sum_{j\in s_{r'}} y_j\right)
\end{aligned}$$

Here we note that the quantity $\frac{1}{N}\left(\frac{1}{n} - \frac{1}{N}\right)\sum_{r=1}^{k}\left\{\sum_{i\in s_r} y_i^2 + \sum_{i\neq}\sum_{j\in s_r} y_i y_j\right\}$ can be estimated unbiasedly by $k\frac{1}{N}\left(\frac{1}{n} - \frac{1}{N}\right)\left\{\sum_{i\in s_r} y_i^2 + \sum_{i\neq}\sum_{j\in s_r} y_i\, y_j\right\}$ since the inclusion probability for the ith unit is $1/k$ for $i = 1,\ldots, N$ and the inclusion probability of the ith and jth $(i \neq j)$ unit belonging to the same systematic sample is also $1/k$. However, $\sum_{r\neq}^{k}\sum_{r'=1}^{k}\left(\sum_{i\in s_r}\sum_{j\in s_{r'}} y_i y_j\right)$ cannot be estimated unbiasedly because the inclusion probability for the pair of units i, j $(i \neq j)$ that do not belong to the same systematic sample is zero. Hence we cannot get an unbiased estimator of V_{sy} from a single systematic sample.

4.3 EFFICIENCY OF SYSTEMATIC SAMPLING

From Eqs. (4.2.1) and (4.2.2), we note that systematic sampling becomes efficient if WSS, the variation within the systematic sample, is large. In this case, the intraclass correlation ρ becomes negative. So, labeling the units plays an important role in ensuring the efficiency of the estimator. Thus systematic sampling will be efficient if one labels the units in such a way that units within the systematic samples become as heterogeneous as possible, allowing intraclass correlation to take a high negative value. This can be achieved if one labels the units so that the y-values increase with the label, i.e., $y_i \leq y_j$ for $i < j$. This type of labeling, however, is not possible as y-values are unknown before conducting a survey. However, if auxiliary information, x (such as area under cultivation), is known before the survey, and it is approximately proportional to y, one should label units in ascending values of x to employ a systematic sampling scheme effectively.

4.3.1 Comparison With Simple Random Sampling

In case k is an integer, the variances of the sample means based on the simple random sampling with (SRSWR) and without (SRSWOR) replacement methods of size n are given by $V_{wr} = \frac{1}{n}\sigma_y^2$ and $V_{wor} = \frac{(N - n)}{Nn}S_y^2 = \frac{(N - n)}{(N - 1)n}\sigma_y^2$, respectively. From Eq. (4.2.2), we note that the sample mean based on a systematic sampling procedure will be more, equal, or less efficient than the sample mean based on SSRWR, if the intraclass correlation ρ becomes negative, equal to $-1/(n - 1)$, or larger than $-1/(n - 1)$, respectively. Similarly, systematic sampling mean is more, equal, or less efficient than the sample mean based on SRSWOR if $\rho <, =$ or $> -1/(N - 1)$, respectively.

4.3.2 Comparison With Stratified Sampling

Let us write $y_{jr} = y_{(j-1)k+r}$; $j = 1,\ldots, n$; $r = 1,\ldots, k$ as the value of y for the jth unit of the rth systematic sample and arrange y_{jr}'s in an $n \times k$ table as follows:

	Systematic sample					
Stratum	**1**		***r***		***k***	**Stratum mean**
1	y_{11}	·	y_{1r}	·	y_{1k}	$\overline{y}_{1\cdot}$
	·	·	·	·	·	
j	y_{j1}	·	y_{jr}	·	y_{jk}	$\overline{y}_{j\cdot}$
	·	·	·	·	·	
n	y_{n1}	·	y_{nr}	·	y_{nk}	$\overline{y}_{n\cdot}$
Systematic sample mean	$\overline{y}_1$	·	$\overline{y}_r$	·	$\overline{y}_k$	

The systematic sample mean corresponding to the random start r is given by $\overline{y}(s_r) = \overline{y}_r = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} y_{jr}$. Now let us assume that y_{jr} is a random sample of size 1 selected from the jth strata consisting of k units $y_{j1},\ldots, y_{jr},\ldots, y_{jk}$. Under this assumption, the variance of $\overline{y}_r$ is given by

$$\begin{aligned} V_{st} &= V\left(\frac{1}{n}\sum_{j=1}^{n} y_{jr}\right) \\ &= \frac{1}{n^2}\sum_{j=1}^{n} V(y_{jr}) \\ &= \frac{1}{n^2 k}\sum_{j=1}^{n}\sum_{r=1}^{k}\left(y_{jr} - \overline{y}_{j\cdot}\right)^2 \\ &= \frac{\sigma^2_{wst}}{n} \end{aligned} \tag{4.3.1}$$

where

$$\sigma^2_{wst} = \frac{\sum\limits_{r=1}^{k}\sum\limits_{j=1}^{n}\left(y_{jr} - \overline{y}_{j\cdot}\right)^2}{N(= nk)} \tag{4.3.2}$$

The variance of the systematic sample mean is

$$
\begin{aligned}
V_{sy} &= V[\bar{y}(s_r)] \\
&= \frac{1}{k}\sum_{r=1}^{k}\left(\bar{y}_r - \bar{y}\right)^2 \\
&= \frac{1}{k}\sum_{r=1}^{k}\left\{\frac{1}{n}\sum_{j=1}^{n}\left(y_{jr} - \bar{y}_{j\cdot}\right)\right\}^2 \\
&= \frac{1}{n^2k}\left[\sum_{r=1}^{k}\sum_{j=1}^{n}\left(y_{jr} - \bar{y}_{j\cdot}\right)^2 + \sum_{r=1}^{k}\left\{\sum_{j\neq}^{n}\sum_{j'=1}^{n}\left(y_{jr} - \bar{y}_{j\cdot}\right)\left(y_{j'r} - \bar{y}_{j'\cdot}\right)\right\}\right] \\
&= [1 + (n-1)\rho_{st}]\frac{\sigma^2_{wst}}{n}
\end{aligned}
\tag{4.3.3}
$$

where

$$
\rho_{st} = \frac{\sum_{r=1}^{k}\left\{\sum_{j\neq}^{n}\sum_{j\prime=1}^{n}\left(y_{jr} - \bar{y}_{\cdot j}\right)\left(y_{j'r} - \bar{y}_{\cdot j'}\right)\right\}}{kn(n-1)\ \ \sigma^2_{wst}}
$$

denotes the intraclass correlation coefficient between the units of the same systematic sample with deviation measured from stratum means. It should be noted that in ρ (defined in Theorem 4.2.2) the deviations are taken from the overall population mean.

Eqs. (4.3.2) and (4.3.3) indicate that systematic sampling is more efficient than stratified sampling if ρ_{st} is negative. The reverse is true when ρ_{st} is positive. Both are equally efficient if ρ_{st} is zero.

4.3.3 Random Arrangement of Units

If the labels of the units are attached at random, then systematic sampling becomes identical to SRSWOR sampling since, in this case, probability of selecting n distinct units is $1\Big/\binom{N}{n}$. Madaw and Madaw (1944) showed that if the labels of the units are attached at random, then the average variance of a systematic sample mean is exactly equal to the average variance of the sample mean based on SRSWOR sampling.

4.3.4 Population With Linear Trend

It is mentioned earlier that the systematic sampling procedure becomes efficient if y_i's increase or decrease with the label i ($\in U$). Here we will consider an artificial population with a linear trend viz. $y_i = \alpha + \beta i$ for $i = 1, \ldots, N$. In this situation

$$\begin{aligned}\overline{Y} &= \frac{1}{N}\sum_{i=1}^{N}(\alpha + \beta i)\\ &= \alpha + \frac{N+1}{2}\beta\\ \sigma_y^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2\\ &= \frac{\beta^2}{N}\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2\\ &= \beta^2\frac{(N^2-1)}{12}\end{aligned}$$

Let $\overline{y}(s_r)$ be the sample mean based on a linear systematic sample selected with a random start r and sampling interval $k = N/n$ is an integer. Then we have

$$\begin{aligned}\overline{y}(s_r) &= \frac{1}{n}\sum_{j=1}^{n} y_{r+(j-1)k}\\ &= \frac{1}{n}\sum_{j=1}^{n}[\alpha + \beta\,\{r + (j-1)k\}]\\ &= \alpha + \left\{r + k\frac{(n-1)}{2}\right\}\beta\end{aligned}$$

The variance of $\overline{y}(s_r)$ is

$$\begin{aligned}V_{sy} &= \frac{1}{k}\sum_{r=1}^{k}\left[\overline{y}(s_r) - \overline{Y}\right]^2\\ &= \frac{\beta^2}{k}\sum_{r=1}^{k}\left(r - \frac{k+1}{2}\right)^2\\ &= \beta^2\frac{(k^2-1)}{12}\end{aligned} \tag{4.3.4}$$

If we treat s_r as a stratified sample (see Chapter 7), the variance of $\overline{y}(s_r)$ becomes

$$\begin{aligned} V_{st} &= \frac{1}{n^2}\sum_{j=1}^{n}\left(1-\frac{1}{k}\right)\frac{1}{k-1}\sum_{r=1}^{k}\left(y_{jr}-\overline{y}_{j\cdot}\right)^2 \\ &= \frac{\beta^2}{n^2}n\left(\frac{1}{k}-1\right)\frac{1}{k-1}\,\frac{k(k^2-1)}{12} \qquad (4.3.5) \\ &= \beta^2\frac{k^2-1}{12n} \end{aligned}$$

The variances of the sample means based on SRSWOR and SRSWR sampling for this population are, respectively, given by

$$\begin{aligned} V_{swor} &= \frac{N-n}{n(N-1)}\sigma_y^2 = \beta^2\frac{(N-n)(N+1)}{12n} \qquad (4.3.6) \\ V_{swr} &= \frac{\sigma_y^2}{n} = \beta^2\frac{(N-1)(N+1)}{12n} \end{aligned}$$

From Eqs. (4.3.4)–(4.3.6), we note that the variance of stratified sampling is $1/n$ times of the variance of the systematic sampling scheme, and the systematic sampling scheme has a variance approximately $1/n$ times of that of SRSWOR sampling. Thus we have the following theorem.

Theorem 4.3.1

For a population with $y_i = \alpha + \beta i$ for $i = 1,\ldots, N$

(i) $V_{st} \le V_{sy} \le V_{swor} \le V_{swr}$

(ii) If the finite population correction term ($=n/N$) is ignored,

$$V_{st}: V_{sys}: V_{swor}: V_{swr} = \frac{1}{n^2}: \frac{1}{n}: 1: 1$$

4.3.4.1 End Corrections

From Theorem 4.3.1, we note that systematic sampling from a population with a linear trend yields a reduction in variance of a magnitude approximately $1/n$ over the SRSWOR sampling scheme. Yates (1948) showed for a population with linear trend, the variance of the systematic sample mean becomes exactly zero if we take a weighted systematic sample mean

$$\begin{aligned}\overline{y}_w(s_r) &= \sum_{j=1}^{n} w_j y_{r+(j-1)k}\\ &= \overline{Y} \qquad\qquad (4.3.7)\\ &= \alpha + \beta\frac{N+1}{2}\end{aligned}$$

with weights $w_1 = \frac{1}{n} + \frac{2r-k-1}{2(n-1)k}$, $w_n = \frac{1}{n} - \frac{2r-k-1}{2(n-1)k}$, and $w_j = \frac{1}{n}$ for $j = 2, \ldots, n-1$. The changes of weights for the first and last term are known as end corrections.

4.3.4.2 Balanced Systematic Sampling

Balanced systematic sampling was introduced by Sethi (1965) and Murthy (1967). In this sampling scheme the units of a sample are selected from the hypothetical population with a linear trend in such a way that the sample mean becomes exactly equal to the population mean $\overline{Y}$.

In the balanced sampling scheme, we assume that the sample size n is even and k $(=N/n)$ is an integer. Here the population is divided into $n/2$ groups of $2k$ consecutive units each. Here we select a random number r from 1 to k. Corresponding to the random start r, the selected balanced systematic sample is

$$\{r + 2jk, 2(j+1)k - (r-1)\},\ j = 0, 1, \ldots, \frac{n}{2} - 1$$

The sample mean for the balanced systematic sample is

$$\overline{y}_s = \frac{1}{n}\sum_{j=0}^{n/2-1}\left\{y_{r+2jk} + y_{2(j+1)k-(r-1)}\right\} \qquad (4.3.8)$$

In presence of linear trend $y_i = \alpha + \beta i$, $\overline{y}_s$ becomes $\alpha + (N+1)\beta/2$, a constant. Hence in the presence of a linear trend, $\overline{y}_s$ is unbiased with a variance zero.

Singh et al. (1968) proposed an alternative method where pairs of units equidistant from the end points are selected in the sample. Thus when n is even, the random start r $(=1,\ldots, k)$ selects the sample

$$\{r + jk, (N - jk) - (r-1)\},\ j = 0, 1, \ldots, \frac{n}{2} - 1$$

The sample mean of Singh's balanced sample is

$$\begin{aligned}\overline{y}_s &= \frac{1}{n}\sum_{j=0}^{n/2-1}\left\{y_{r+jk} + \ y_{N-jk-(r-1)}\right\}\\ &= \frac{1}{n}\sum_{j=0}^{n/2-1}\{2\alpha + \beta(N+1)\}\\ &= \alpha + (N+1)\beta/2\end{aligned}$$

This method also eliminates the effect of linear trend because the sample mean is equal to the population mean $\alpha + (N+1)\beta/2$.

4.3.5 Population With Periodic Variation

Consider a hypothetical population that is strictly periodic with a period λ, i.e., $y_i = y_{i+\lambda}$. If we select a systematic sample with a sampling interval that is equal to the period $k = \lambda$ or its multiple, then all y_j values belonging to the same systematic samples will be equal and hence the variance of the systematic sample mean will be large. On the other hand, if we choose $k = \lambda/2$ or an odd multiple of it, then the systematic sample mean will be equal to the population mean $\overline{Y}$, and in this situation, the variance of the estimator is exactly zero. In practice, a population that is strictly periodic in nature may not be available, but approximately periodic populations are not uncommon. Examples are number of accidents, telephone calls during the 24 h in a day, and sales over 7 days in department stores. In this situation systematic sampling with an appropriate sampling interval produces an efficient estimator for the population mean.

4.3.6 Autocorrelated Population

In some population units, closer units are well related with respect to the study variable y. For example, the number of illegal immigrants in the villages decreases with the distance of the border from the neighboring country. In this situation, to accommodate relationship of the y-values, some statistical model (superpopulation model) is generally assumed. Details of the superpopulation model are given in Chapter 6. Cochran (1946) assumed the following autocorrelation model to study the performance of systematic sampling with other sampling designs:

$$E_m(y_i) = \mu, E_m(y_i - \mu)^2 = \sigma^2, E_m(y_i - \mu)(y_{i+u} - \mu) = \rho_u\sigma^2 \qquad (4.3.9)$$

where E_m denotes expectation over superpopulation model (4.3.9) and ρ_u, the autocorrelation of order (lag) u, satisfies $\rho_u \geq \rho_v \geq 0$ for $u < v$.

Let $E_m(V_{sy})$, $E_m(V_{st})$, and $E_m(V_{ran})$ denote average variance of the sample mean of the systematic sample, stratified sample, and simple random sample without replacement, respectively. Cochran (1946) derived the following theorems:

Theorem 4.3.3

Under superpopulation model (4.3.9) with $\rho_u \geq \rho_v \geq 0$, $E_m(V_{st}) \leq E_m(V_{ran})$

Proof

We can write $\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2$ as

$$\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2 = \frac{1}{N}\sum_{i<}^{N}\sum_{j=1}^{N}\left(y_i - y_j\right)^2 \tag{4.3.10}$$

Taking expectation both sides with respect to the model (4.3.9), we get

$$\begin{aligned}
E_m\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2 &= \frac{1}{N}\sum_{i<}^{N}\sum_{j=1}^{N}E_m\left(y_i - y_j\right)^2 \\
&= \frac{2\sigma^2}{N}\sum_{i<}^{N}\sum_{j=1}^{N}\left(1 - \rho_{j-i}\right) \\
&= \frac{2\sigma^2}{N}\left[\frac{N(N-1)}{2} - \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\rho_{j-i}\right] \\
&= \frac{2\sigma^2}{N}\left[\frac{N(N-1)}{2} - \sum_{i=1}^{N-1}\left(\rho_1 + \rho_2 + \cdots + \rho_{N-i}\right)\right] \\
&= \frac{2\sigma^2}{N}\left[\frac{N(N-1)}{2} - \sum_{j=1}^{N-1}(N-j)\rho_j\right] \\
&= \sigma^2(N-1)\left[1 - \frac{2}{N(N-1)}\sum_{j=1}^{N-1}(N-j)\rho_j\right]
\end{aligned} \tag{4.3.11}$$

Eq. (4.3.11) yields

$$E_m(V_{ran}) = \frac{N-n}{Nn}\frac{1}{N-1}E_m\sum_{i=1}^{N}\left(y_i-\overline{Y}\right)^2$$
$$= \sigma^2\frac{k-1}{nk}\left[1-\frac{2}{N(N-1)}\sum_{j=1}^{N-1}(N-j)\rho_j\right] \tag{4.3.12}$$
$$(\text{noting } N = nk)$$

Let us assume that the systematic sample of size n is a stratified sample where one unit is selected from each of the stratum. Then the variance of the stratified sample mean is obtained from Eq. (4.3.1) as

$$V_{st} = \frac{1}{n^2k}\sum_{j=1}^{n}\sum_{r=1}^{k}\left(y_{jr}-\overline{y}_{j\cdot}\right)^2 \tag{4.3.13}$$

Now using (4.3.11) we get

$$E_m\sum_{r=1}^{k}\left(y_{rj}-\overline{y}_{j\cdot}\right)^2 = \sigma^2(k-1)\left[1-\frac{2}{k(k-1)}\sum_{j=1}^{k-1}(k-j)\rho_j\right] \tag{4.3.14}$$

Now substituting (4.3.14) in (4.3.13), we have

$$E_m(V_{st}) = \sigma^2\frac{(k-1)}{nk}\left[1-\frac{2}{k(k-1)}\sum_{j=1}^{k-1}(k-j)\rho_j\right] \tag{4.3.15}$$

Eqs. (4.3.12) and (4.3.15) yield

$$E_m(V_{ran}) - E_m(V_{st}) = 2\sigma^2\frac{k-1}{nk}[L(N)-L(k)] \tag{4.3.16}$$

where

$$L(t) = 1-\frac{2}{t(t-1)}\sum_{j=1}^{t-1}(t-j)\rho_j$$
$$= 1-\frac{2}{t(t-1)}\sum_{j=1}^{t}(t-j)\rho_j \tag{4.3.17}$$

Now

$$L(t)-L(t+1) = -\frac{2}{t(t^2-1)}\sum_{j=1}^{t}a_j\rho_j \tag{4.3.18}$$

where $a_j = (t+1-2j)$.

The expression $\sum_{j=1}^{t} a_j \rho_j$ can be written as

$$\sum_{j=1}^{t} a_j\rho_j = a_1(\rho_1 - \rho_2) + (a_1 + a_2)(\rho_2 - \rho_3) + \cdots + (a_1 + \cdots + a_i)$$
$$(\rho_i - \rho_{i+1}) + \cdots + (a_1 + \cdots + a_{t-1})(\rho_{t-1} - \rho_t) + (a_1 + \cdots + a_t)\rho_t$$

Now noting (i) $\rho_j - \rho_{j+1} \geq 0$, (ii) $a_1 + \ldots + a_i = i(t-i) \geq 0$ for $t \geq i$, and (iii) $a_1 + \ldots + a_t = 0$, we find

$$L(t) - L(t+1) \leq 0 \tag{4.3.19}$$

Finally the theorem follows from (4.3.16) and (4.3.19).

The expected variance of the systematic sample mean under the superpopulation model (4.3.9) was derived by Cochran (1946) as

$$E_m(V_{sy}) = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right)\left\{1 - \frac{2}{N(k-1)}\sum_{j=1}^{N-1}(N-j)\rho_j + \frac{2k}{n(k-1)}\sum_{j=1}^{n-1}(n-j)\rho_{kj}\right\} \tag{4.3.20}$$

Eqs. (4.3.15), (4.3.16), and (4.3.20) do not yield any general comparison of efficiency of the systematic sampling with random sampling and stratified sampling. However, Cochran (1946) reported for the model (4.3.9) that (i) systematic sample is more efficient than stratified sampling if $\rho_j = 0$ for $j > (k-1)$ while (ii) systematic sampling is less efficient than simple random sampling if $\rho_1 = \rho_2 = \ldots = \rho_k$ and $\rho_j = 0$ for $j > k$.

Cochran (1946) derived the following important results relating to the performances of systematic, stratified, and random sampling.

Theorem 4.3.4

Under superpopulation model (4.3.9) with (i) $\rho_i \geq \rho_{i+1} \geq 0$, for $i = 1, 2,\ldots,(N-1)$, and (ii) $\delta_i^2 = \rho_{i+1} - 2\rho_i + \rho_{i-1} \geq 0$, for $i = 2, 3,\ldots, N-2$,

$$E_m(V_{sy}) \leq E_m(V_{st}) \leq E_m(V_{ran})$$

for any sample size n.

Further, $E_m(V_{sy}) < E_m(V_{st})$ unless $\delta_i^2 = 0, i = 2, 3, \ldots, (N-2)$.

Further additional results are given by Quenouille (1949), Gautschi (1957), Jowett (1952), Hájek (1959), and Murthy and Rao (1988), among others.

4.4 LINEAR SYSTEMATIC SAMPLING USING FRACTIONAL INTERVAL

We have seen that linear systematic sampling cannot always produce a sample of the desired size n when N/n is not an integer. The above difficulty can be removed if we take $k = N/n$ without rounding it off. The method of selection is as follows. Firstly, select a random sample r, from a uniform distribution $(0, k)$. Then the random start r selects the ith unit in the sample if $i - 1 < r + jk \leq i$ for $j = 0, 1, \ldots, n - 1$. For instance, let $N = 13$, $n = 4$, and $k = 13/4 = 3.25$. Suppose $r = 2.15$ is a random sample selected from a uniform distribution (0, 3.25). Then the selected systematic sample includes the units 3, 6, 9, and 12. The first-order inclusion probability for each of the unit is equal to $1/k$ and hence an unbiased estimator of the population mean $\overline{Y}$ is

$$\widehat{\overline{Y}}_{ss} = \frac{k}{N} \sum_{i \in s} y_i = \overline{y}_s \tag{4.4.1}$$

4.5 CIRCULAR SYSTEMATIC SAMPLING

In a circular systematic sampling (CSS) scheme, initially a unit r (random start) is selected at random from all the N units of the population, then every kth unit is selected in a circular manner until the desired sample size n is obtained. Here k is taken as an integer nearest to N/n. Thus corresponding to a random start r, a circular systematic sample selects units

$$r + j\,k \quad \text{if} \quad r + j\;\;k \leq N$$

and

$$r + j\,k - N \quad \text{if} \quad r + j\;\;k > N$$
$$\text{for} \;\; j = 0, \ldots, n - 1.$$

The CSS method was introduced by Lahiri (1954), and it was extensively used by the Indian National Sample Survey Organization for large-scale sample surveys. Here we will consider the following cases.

4.5.1 Circular Systematic Sampling With $k = N/n$ as an Integer

In this situation we get N systematic samples, but all samples are not distinct. Each sample occurs exactly n times in different orders. To illustrate this phenomenon, let us consider the following example.

Example 4.5.1

Consider a population of size $N = 6$, the target sample size $n = 3$ and $k = 6/3 = 2$.

Random start	Systematic samples
1	1, 3, 5
2	2, 4, 6
3	3, 5, 1
4	4, 6, 2
5	5, 1, 3
6	6, 2, 4

Here we note that the random starts 1, 3, and 5 produce the same systematic sample but in different orders and they are (1, 3, 5), (3, 5, 1), and (5, 1, 3). Here also the sample mean $\overline{y}_s$ is an unbiased estimator of the population mean.

4.5.2 Circular Systematic Sampling With N/n is Not an Integer

In this case, if one chooses a value of k greater than N/n, then a circular systematic sample may contain a unit more than once, which is seen in the following example. In this case, the sample mean based on the distinct units and all the units including repetitions are unbiased for the population mean.

Example 4.5.2

Let $N = 30$ and the target sample size $n = 12$. Here $N/n = 2.5$, so both the integers 2 and 3 are nearest to 2.5. If we choose $k = 3$, then the random start 15 will select the following systematic sample:

$$s = (15, 18, 21, 24, 27, 30, 3, 6, 9, 12, 15, 18)$$

In the above systematic sample the units 15 and 18 occur twice.

To avoid the above difficulty, Sudakar (1978) pointed out that if k is taken as $\left[\frac{N}{n}\right]$, the largest integer less than or equal to N/n, then none of the units in the same systematic sample can occur more than once, i.e., all the units are distinct. In this case, the sample mean is an unbiased estimator of the population mean.

In Example 4.5.2, if we take $k = 2$, then the random start 15 would select the following systematic sample where no unit is repeated twice.

$$s = (15, 17, 19, 21, 23, 25, 27, 29, 1, 3, 5, 7)$$

The choice of the sampling interval k depends on the sample size n. Sometimes due to budgetary considerations and inadequate knowledge of sampling cost, the optimal sample size cannot be determined in advance. Hence it is of interest to know for a given population of size N, if there exists a sampling plan that can select a circular systematic sample of any size from 1 to N. Sudakar (1978) established a necessary and sufficient condition on N and k, which results in all n ($\leq N$) units of a sample being distinct. The condition is stated in the following theorem.

Theorem 4.5.1

Let N be the population size, k ($<N$) be an integer chosen as the sampling fraction, and the circular systematic sample selected with the random start r be denoted as

$$s(r, n) = (i_1, i_2, \ldots, i_n)$$

where

$$i_j = r + (j-1)k \ \ \mathrm{mod}(N) \ \ \text{where } j = 1, \ldots, n.$$

A necessary and sufficient condition for the units of the systematic samples s (r, n) to be distinct for all r, $n \leq N$ is that N and k are relatively coprimes.

Proof

Here both the necessary and sufficient conditions are proved by contradiction as follows:

(i) Sufficiency: Suppose N and k are relatively coprimes and there exist r and n for which two elements of s (r, n) are equal. Without loss of generality let us suppose that $i_1 = r$ and i_{j+1} are equal. Now $i_1 = r = i_{j+1} = (r + jk)$ mod (N) implies jk is a multiple of N; $j < n \leq N$. This contradicts that k and N are coprimes.

(ii) Necessary *j*th ($i \neq j$): Suppose for all r, $n \leq N$, all elements of s (r, n) are distinct and N and k are not coprimes. Let the greatest common factor of N and k be α so that $k = \beta\alpha$, $N = \gamma\alpha$, and β and γ are integers smaller than N. Then

$$\begin{aligned} i_{\gamma+1} &= (r + \gamma k)\text{mod}(N) \\ &= (r + \gamma\beta\alpha)\text{mod}(N) \\ &= (r + N\beta)\text{mod}(N) \\ &= r = i_1 \end{aligned}$$

This contradicts assumptions that all elements of s (r, n) are distinct.

Remark 4.5.1

The variance expressions for the estimator of the population mean $\overline{Y}$ based on fractional interval and CSS are not straightforward because all the possible samples may not be disjoint (see Särndal et al., 1992). The variance of $\widehat{\overline{Y}}_{ss}$ cannot be estimated from a single systematic sample as the second-order inclusion probabilities of some of the units are zero.

4.6 VARIANCE ESTIMATION

There are two types of variance estimators. One based on a single systematic sample of size n and the other based on several systematic samples (t) each of size m ($=n/t$). The variance estimators based on a single systematic sample produce biased estimators. Although the variance estimators based on several samples produce unbiased estimators of the variance, the convenience and efficiency of the systematic sampling are lost. The technique of variance estimation based on several samples was introduced by Mahalanobis (1946) and it is known as interpenetrating subsamples. Details of the variance estimations are given by Cochran (1946), Yates (1949), Koop (1971), Wolter (1984), Murthy and Rao (1988), and Bellhouse (1988), among others. Here we shall assume that the population size N is an integer multiple of the sample size n, i.e., $N = nk$ where k is the sampling interval.

4.6.1 Single Systematic Sample

Here the variance estimators are obtained under some assumptions on the population.

4.6.1.1 Random Arrangements of Units

Systematic sampling can be treated as SRSWOR sampling if the arrangement of the units in the population is random. So, treating the selected systematic sample s as an SRSWOR sample of size n, the variance of the systematic sample mean V_{sy} can be estimated as

$$\widehat{v}_1 = \frac{(1-f)}{n} s_y^2 \tag{4.6.1}$$

where $f = n/N$ and $s_y^2 = \sum_{i\in s} \{y_i - \overline{y}(s)\}^2/(n-1)$.

The estimator $\widehat{v}_1$ performs well if the arrangement of the units is random, i.e., if no trend or periodicity in the arrangements of the units persists. However, if the arrangement of the units is such that the systematic sampling is more efficient than SRSWOR sampling, then $\widehat{v}_1$ certainly overestimates the variance. On the other hand, $\widehat{v}_1$ underestimates the variance if in reality systematic sampling is less efficient than SRSWOR.

4.6.1.2 Stratified Sampling With One Unit Per Stratum

Systematic sampling stratifies the population into k strata each of size n. Let n be even and equal to $2q$. Now consider the units $(r,\ r+k)$ as a stratified sample of size 2 selected by the SRSWOR method from a stratum comprising the first $2k$ units viz. $1,\ldots, 2k$; $(r+2k,\ r+3k)$ as a stratified sample of size 2 from a stratum comprising the next $2k$ viz. $2k+1,\ldots, 4k$; in general, units $\{r+2(j-1)k,\ r+(2j-1)k\}$ as a stratified sample of size 2 from the jth stratum comprising units $2(j-1)k+1,\ldots, 2jk$; $j=1,\ldots, q$.

An approximate estimate of V_{sy} is obtained as

$$\widehat{v}_2 = \frac{(1-f)}{n^2}\sum_{j=1}^{n/2}\delta_j^2 \tag{4.6.2}$$

where $\delta_j = y_{r+(2j-1)k} - y_{r+2(j-1)k}$.

4.6.1.3 Presence of Linear Trend

Let us construct the following difference table from the selected systematic sample s_r:

s_r	Y	Δ	Δ^2
r	y_r		
$r+k$	y_{r+k}	$y_{r+k} - y_r = \Delta(1)$	
$r+2k$	y_{r+2k}	$y_{r+2k} - y_k = \Delta(2)$	$\Delta(2) - \Delta(1) = \Delta^2(1)$
.	.	.	
.	.	.	
$r+(n-1)k$	$y_{r+(n-1)k}$	$y_{r+(n-1)k} - y_{r+(n-2)k} = \Delta(n-1)$	$\Delta(n-1) - \Delta(n-2) = \Delta^2(n-2)$

The variance of the systematic sample mean V_{sy} is estimated by using the following formulae:

$$\widehat{v}_3 = \frac{(1-f)}{n}\frac{1}{2(n-1)}\sum_{j=1}^{n-1}\{\Delta(j)\}^2 \tag{4.6.3}$$

and

$$\widehat{v}_4 = \frac{(1-f)}{n}\frac{1}{6(n-2)}\sum_{j=1}^{n-2}\left\{\Delta^2(j)\right\}^2 \tag{4.6.4}$$

The estimators $\widehat{v}_3$ and $\widehat{v}_4$ perform well if the ordering of units produces a linear trend. Otherwise, they underestimate variance. Several variance estimators based on higher-order differences are also proposed, e.g., Yates (1949).

4.6.1.4 Presence of Autocorrelation Between Successive Units

Cochran (1946) provided an alternative variance estimator involving $\widehat{\rho}_k$ as an estimate of ρ_k, intraclass correlation of order k $(=N/n)$. Cochran's variance estimator is given by

$$\widehat{v}_5 = \begin{cases} \dfrac{(1-f)s_y^2}{n}\left[1+2\left\{\dfrac{1}{\text{In}(\widehat{\rho}_k)}+\dfrac{1}{(\widehat{\rho}_k)^{-1}-1}\right\}\right] & \text{if } \widehat{\rho}_k > 0 \\ \qquad\qquad \dfrac{(1-f)s_y^2}{n} \quad \text{if } \widehat{\rho}_k \le 0 \end{cases} \tag{4.6.5}$$

where $\widehat{\rho}_k = \left[\dfrac{1}{(n-1)}\displaystyle\sum_{j=0}^{n-2}\left\{y_{r+(j+1)k}-\overline{y}(s)\right\}\left\{y_{r+jk}-\overline{y}(s)\right\}\right]\Big/ s_y^2.$

The proposed variance estimators mentioned earlier are based on underlying certain superpopulation model. The variance estimators may not perform well if the underlying model fails. Wolter (1984) studied in detail the properties of bias, mean-square error, and confidence interval estimation theoretically and empirically.

4.6.1.5 Splitting of a Systematic Sample

Koop (1971) divided the systematic sample into two samples each of size $n/2$ (assuming integer) and proposed the following variance estimator

$$\widehat{v}_6 = \frac{1}{4}\left(\overline{y}_A - \overline{y}_B\right)^2 \tag{4.6.6}$$

where $\overline{y}_A$ and $\overline{y}_B$ are the sample means based on odd and even labels of the systematic sample *s*. Koop (1971) derived the expression of bias relative to its variance in terms of the intraclass correlation coefficient.

Now noting $\overline{y}_s = \frac{1}{2}\left(\overline{y}_A + \overline{y}_B\right)$, we find

$$\begin{aligned} V\left(\overline{y}_s\right) &= \frac{1}{2}(1+\rho_0)\sigma_0^2 \\ &= \frac{(1+\rho_0)}{(1-\rho_0)}E(\widehat{v}_6) \end{aligned}$$

where $\sigma_0^2 = V\left(\overline{y}_A\right) = V\left(\overline{y}_B\right)$ and ρ_0 is the correlation coefficient between the subsample means. In case correlation coefficient ρ_0 is known from the past survey, an almost unbiased estimator of $V\left(\overline{y}_s\right)$ is given by

$$\begin{aligned} \widehat{v}_7 &= \widehat{V}[\overline{y}(s)] \\ &= \frac{1+\rho_0}{1-\rho_0}\widehat{v}_6 \end{aligned} \tag{4.6.7}$$

4.6.2 Several Systematic Samples

Here instead of taking a single systematic sample of size *n*, several (*l*) independent systematic samples are selected. The sample sizes for the systematic samples need not be the same. Let t_i be an unbiased estimator of $\overline{Y}$ based on the *i*th (=1,..., *l*) systematic sample of size $m_i\left(\sum_{i=1}^{l} m_i = n\right)$. Then we have the following theorem.

Theorem 4.6.1

(i) $\overline{t} = \frac{1}{l}\sum_{i=1}^{l} t_i$ is an unbiased estimator of $\overline{Y}$

(ii) An unbiased estimator of the variance of $\overline{t}$ is $\widehat{v}_8 = \frac{1}{l(l-1)}\sum_{i=1}^{l}\left(t_i - \overline{t}\right)^2$

Proof

(i) $E(\overline{t}) = \frac{1}{l}\sum_{i=1}^{l} E(t_i) = \overline{Y}$

(ii) Variance of $\overline{t} = V(\overline{t}) = \frac{1}{l^2}\sum_{i=1}^{l} V(t_i)$ and

$$\begin{aligned} E(\widehat{v}_8) &= \frac{1}{l(l-1)}E\left(\sum_{i=1}^{l} t_i^2 - l\,\overline{t}^2\right) \\ &= \frac{1}{l(l-1)}\left[\sum_{i=1}^{l}\left\{V(t_i) + \overline{Y}^2\right\} - l\left\{V(\overline{t}) + \overline{Y}^2\right\}\right] \\ &= V(\overline{t}) \end{aligned}$$

For independent interpenetrating sampling, a unit may be repeated more than once. To avoid repletion of units, we first enumerate all possible $N/m = t$ (assuming integer) systematic samples of size m corresponding to t random starts. From the t random starts, a sample of l random starts is selected by SRSWOR and l systematic samples each of size m are selected using the chosen l random starts.

Let $\bar{y}(s_j)$ be the sample mean for the jth systematic sample and

$$\bar{y}(s) = \sum_{j=1}^{l} \bar{y}(s_j)/l \tag{4.6.8}$$

be the pooled sample mean.

The combined mean $\bar{y}(s)$ is unbiased for $\overline{Y}$ and an unbiased estimator of $V\{\bar{y}(s)\}$ is given by

$$\widehat{v}_9 = \frac{(1-g)}{l(l-1)} \sum_{j=1}^{l} \left\{\bar{y}(s_j) - \bar{y}(s)\right\}^2 \tag{4.6.9}$$

where $g = l/t$.

As an illustration, we give the following example.

Example 4.6.1

The following list gives claims against 30 vouchers.

Voucher number	Amount in $	Voucher number	Amount in $	Voucher number	Amount in $
1	128	11	453	21	303
2	499	12	395	22	112
3	196	13	334	23	256
4	142	14	400	24	256
5	158	15	342	25	229
6	126	16	317	26	225
7	101	17	316	27	225
8	282	18	210	28	488
9	368	19	335	29	117
10	473	20	286	30	329

(i) Select a sample of size six vouchers by systematic sampling procedures. Estimate the mean amount of the vouchers and estimate its variance by using the different procedures described in Section 4.6.1.

(ii) Select three independent samples of sizes 4, 6, and 7 by using the CSS procedure and obtain an estimate of average claim per voucher from each of the samples separately. Obtain an estimator for the mean amount of a voucher based on the three samples combined and find an unbiased estimator of the standard error of the proposed combined estimator.

Solution:

(i) Here $N = 30$ and $n = 6$, so $k = 30/6 = 5$. Let the random start 3 (=r) be selected from 1 to 5 (=k). The random start select sample $s = (3, 8, 13, 18, 23, 28)$. So, the estimated mean amount of voucher is

$$\begin{aligned}\overline{y}(s) &= \frac{y_3 + y_8 + y_{13} + y_{18} + y_{23} + y_{28}}{6} \\ &= \frac{196 + 282 + 334 + 210 + 256 + 488}{6} \\ &= \$294.33\end{aligned}$$

The estimated variances based on the Eq. (4.6.1) to (4.6.6) are computed as follows:

$$\begin{aligned}\widehat{v}_1 &= \frac{(1-f)}{n}s_y^2 \\ &= \frac{(1-6/30)}{6} \times 11496.67 \\ &= 1532.889\end{aligned}$$

$$\begin{aligned}\widehat{v}_2 &= \frac{(1-f)}{n^2}\sum_{j=1}^{n/2}\delta_j^2 \\ &= \frac{(1-6/30)}{6^2} \times 76596 \\ &= 1702.133\end{aligned}$$

$$\begin{aligned}\widehat{v}_3 &= \frac{(1-f)}{n}\frac{1}{2(n-1)}\sum_{j=1}^{n-1}\{\Delta(j)\}^2 \\ &= \frac{(1-6/30)}{6}\frac{1}{2(6-1)} \times 81416 \\ &= 1085.54\end{aligned}$$

$$\begin{aligned}\widehat{v}_4 &= \frac{(1-f)}{n}\frac{1}{6(n-2)}\sum_{j=1}^{n-2}\{\Delta^2(j)\}^2 \\ &= \frac{(1-6/30)}{6}\frac{1}{6 \times (6-2)} \times 95628 \\ &= 192.2\end{aligned}$$

Here correlation coefficient $\widehat{\rho}_5$ between the pairs of observation (282, 196), (334, 282), (210, 334), (256, 210), and (448, 256) is -0.125 (negative). Hence the estimate variance

$$\begin{aligned}\widehat{v}_5 &= \frac{(1-f)}{n}s_y^2\\ &= 1532.889\end{aligned}$$

Sample mean based on odd and even labels are $\overline{y}_A = 262$ and $\overline{y}_B = 326.66$, respectively. Hence

$$\begin{aligned}\widehat{v}_6 &= \frac{1}{4}\left(\overline{y}_A - \overline{y}_B\right)^2\\ &= 1045.44\end{aligned}$$

(ii) For sample size $n = 4$, $N/n = 7.5$, is not an integer. So, we take $k = (30/4) = 7$ and select one unit at random from 1 to 30. Let the selected unit (random start) be 5. Then the selected sample is $s_1 = \{5, 12, 19, 26\}$. The estimator for the mean $\overline{Y}$ is $t_1 = (158 + 395 + 335 + 225)/4 = \$278.25$. For sample size $n = 6$, $k = N/n = 30/6 = 5$ an integer. Here also we select a random start from 1 to 30 and let it be 10. The selected CSS with the random start 10 is $s_2 = \{10, 15, 20, 25, 30, 5\}$; the estimator for $\overline{Y}$ is $t_2 = (473 + 342 + 286 + 229 + 329 + 158)/6 = \$302.83$. Similarly for the sample size $n = 7$, we choose $k = (30/7) = 4$. Let the selected random start from 1 to 30 be 25, then the selected CSS is $s_3 = \{25, 29, 3, 7, 11, 15, 19\}$ and estimator for $\overline{Y}$ is $t_3 = (229 + 117 + 196 + 101 + 453 + 342 + 335)/7 = \$253.29$. The combined estimator of $\overline{Y}$ is $\overline{t} = (t_1 + t_2 + t_3)/3 = \$278.12$ and an estimate of the standard error of t is

$$\sqrt{\widehat{V}(\overline{t})} = \sqrt{\frac{1}{3 \times 2}\sum_{i=1}^{3}(t_i - \overline{t})^2 = \$14.30}$$

4.7 TWO-DIMENSIONAL SYSTEMATIC SAMPLING

Linear systematic sampling can be generalized to two-dimensional systematic sampling. Consider a population of size $N = ab$ arranged in $a = ml$ rows and $b = tk$ columns. Suppose we want to take a systematic sample of size $n = mt$ units from N units. The procedure is as follows: Instead of selecting single random number as used in a one-dimensional systematic sample, we select a pair of random numbers (r, r'). The random number r is selected from 1 to l, and r' is selected from 1 to k. (r, r') may be regarded as a two-dimensional random start. The first coordinate (r) represents the position of the row and the second (r') the column of the initially chosen unit. Then the two-dimensional systematic sample consists of units with coordinates (x, y), where $x = r + (i - 1)l$ and $y = r' + (j - 1)k$; $i = 1,\ldots, m$; $j = 1,\ldots, t$. Methods of estimation of population mean and its variance can be obtained from the results derived for one-dimensional systematic sampling. Interested readers are referred to Quenouille (1949), Das (1951), Koop (1976), and Bellhouse (1977, 1981).

Example 4.7.1

Consider 180 units arranged into $a = 12$ rows and $b = 15$ columns. We want to select a sample of 20 units. Here we may take $m = 4$, $l = 3$ and $t = 5$, $k = 3$. Suppose a random number $r = 2$ is selected from 1 to l (=3) and another number $r' = 3$ is selected from 1 to t (=5) so that the selected random start is (2, 3). Then our selected two-dimensional systematic sample is as follows:

$$\begin{aligned} s = {} & (2,3), (2,6), (2,9), (2,12), (2,15) \\ & (5,3), (5,6), (5,9), (5,12), (5,15) \\ & (8,3), (8,6), (8,9), (8,12), (8,15) \\ & (11,3), (11,6), (11,9), (11,12), (11,15) \end{aligned}$$

	Columns														
Rows	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1															
2			√			√			√			√			√
3															
4															
5			√			√			√			√			√
6															
7															
8			√			√			√			√			√
9															
10															
11			√			√			√			√			√
12															

4.8 EXERCISES

4.8.1 Define the systematic sampling procedure and explain how it differs from random sampling. State some of the merits and demerits of systematic sampling with suitable examples.

4.8.2 Explain the LSS. Suggest an unbiased estimator for the population mean. Obtain the expression of its variance. Compute the first- and second-order inclusion probabilities when N/n is an integer. Explain why it is not possible to get an unbiased estimator of the variance based on a single systematic sample.

4.8.3 Compute the inclusion probabilities of the first two orders for an LSS when N/n is an integer. Explain the difficulty you may face for using LSS when N/n is not an integer.

4.8.4 Describe the CSS and explain how you will choose the sampling interval k when N/n is not an integer. Show that the first-order inclusion probabilities of LSS and CSS are the same when N/n is an integer.

4.8.5 The following data relate to the heights of 30 plants for an agricultural experiment.

Serial number of plant	Height in cm	Serial number of plant	Height in cm	Serial number of plant	Height in cm
1	28	11	53	21	33
2	99	12	95	22	12
3	96	13	34	23	56
4	42	14	40	24	56
5	58	15	42	25	29
6	26	16	17	26	25
7	11	17	16	27	25
8	82	18	10	28	88
9	68	19	35	29	17
10	73	20	86	30	29

(a) Write all possible LSSs of size 5 and compute the variance of the estimator of the population mean. Compare the performance of systematic sampling and SRSWOR sampling based on the same sample size 5.

(b) Select an LSS of size 8 by using a fractional interval and find an unbiased estimate of the population mean height of the 30 plants.

(c) Select a circular systematic sample of size 10 and estimate average height of the plants.

4.8.6 Consider the data given in Example 4.6.1. (a) Select three independent systematic samples of sizes 6, 7, and 8 using LSS using fractional intervals. Using the selected samples find an unbiased estimate of the population mean and its standard errors. (b) Select two independent samples of sizes 6 and 8 by CSS and compute a 90% confidence interval of the population mean height.