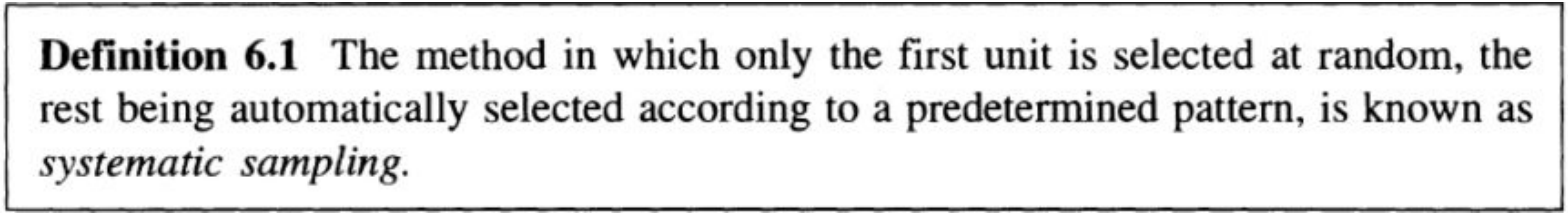
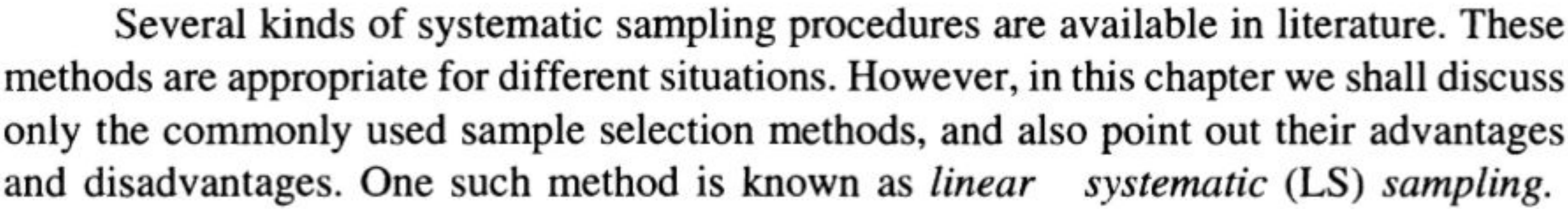
Systematic Sampling



Definition 6.1 The method in which only the first unit is selected at random, the rest being automatically selected according to a predetermined pattern, is known as *systematic sampling*.

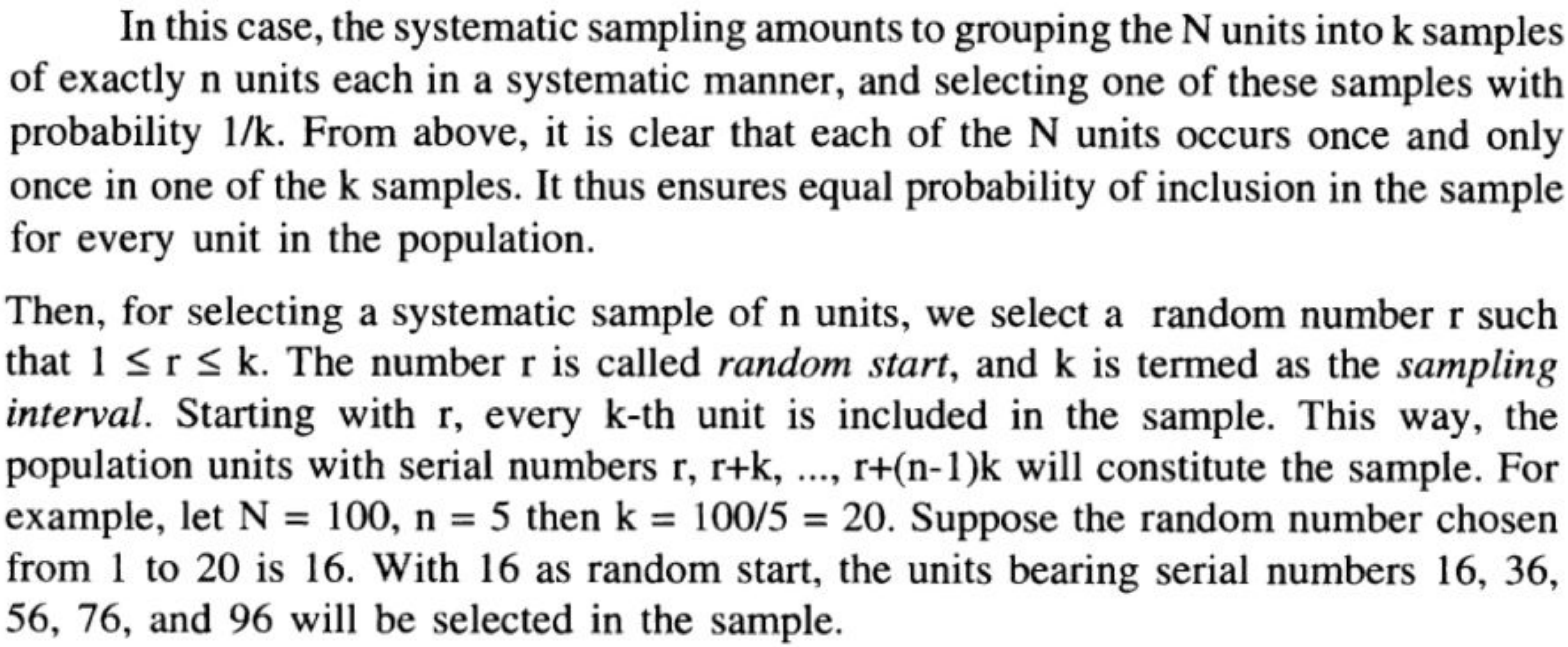


Several kinds of systematic sampling procedures are available in literature. These methods are appropriate for different situations. However, in this chapter we shall discuss only the commonly used sample selection methods, and also point out their advantages and disadvantages. One such method is known as *linear systematic* (LS) *sampling*.

Suppose we want to select a systematic sample of size n from a population consisting of N units. The method of LS sampling is employed when N is a multiple of n, that is, N=nk where k is an integer. For explaining the procedure, let us assume that the nk serial numbers of the population units in the frame are rearranged in k columns as follows :

$$\begin{array}{ccccccc}
1 & 2 & 3 & \cdots & r & \cdots & k \\
k+1 & k+2 & k+3 & \cdots & k+r & \cdots & 2k \\
2k+1 & 2k+2 & 2k+3 & \cdots & 2k+r & \cdots & 3k \\
\vdots & \vdots & \vdots & \vdots & \vdots & & \\
(n-1)k+1 & (n-1)k+2 & (n-1)k+3 & \cdots & (n-1)k+r & \cdots & nk
\end{array}$$

Linear systematic sampling (LS)



In this case, the systematic sampling amounts to grouping the N units into k samples of exactly n units each in a systematic manner, and selecting one of these samples with probability 1/k. From above, it is clear that each of the N units occurs once and only once in one of the k samples. It thus ensures equal probability of inclusion in the sample for every unit in the population.

Then, for selecting a systematic sample of n units, we select a random number r such that $1 \leq r \leq k$. The number r is called *random start*, and k is termed as the *sampling interval*. Starting with r, every k-th unit is included in the sample. This way, the population units with serial numbers r, r+k, ..., r+(n-1)k will constitute the sample. For example, let N = 100, n = 5 then k = 100/5 = 20. Suppose the random number chosen from 1 to 20 is 16. With 16 as random start, the units bearing serial numbers 16, 36, 56, 76, and 96 will be selected in the sample.

Example 6.1

An insurance company's claims, in dollars, for one day are 400, 600, 570, 960, 780, 800, 460, 650, 440, 530, 470, 810, 625, 510, and 700. List all possible systematic samples of size 3, that can be drawn from this set of claims using linear systematic sampling. Also, obtain corresponding sample means.

Solution

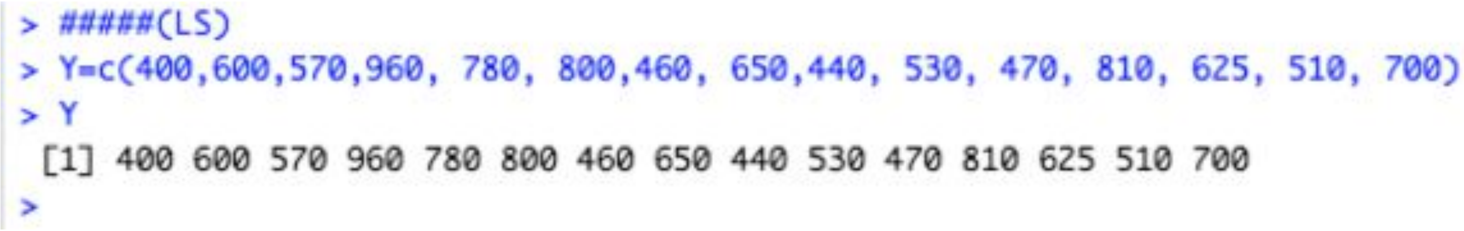
Here the population size N=15, and the size of the sample to be selected is n = 3. The sampling interval k will thus be 15/3 = 5. The random number r to be selected from 1 to k can, therefore, take any value in the closed interval [1, 5]. Each random start from 1 to 5 will yield corresponding systematic sample. In all, there will be k=5 possible samples. These are given below in table 6.1 along with their means.

Table 6.1 Possible systematic samples and their means

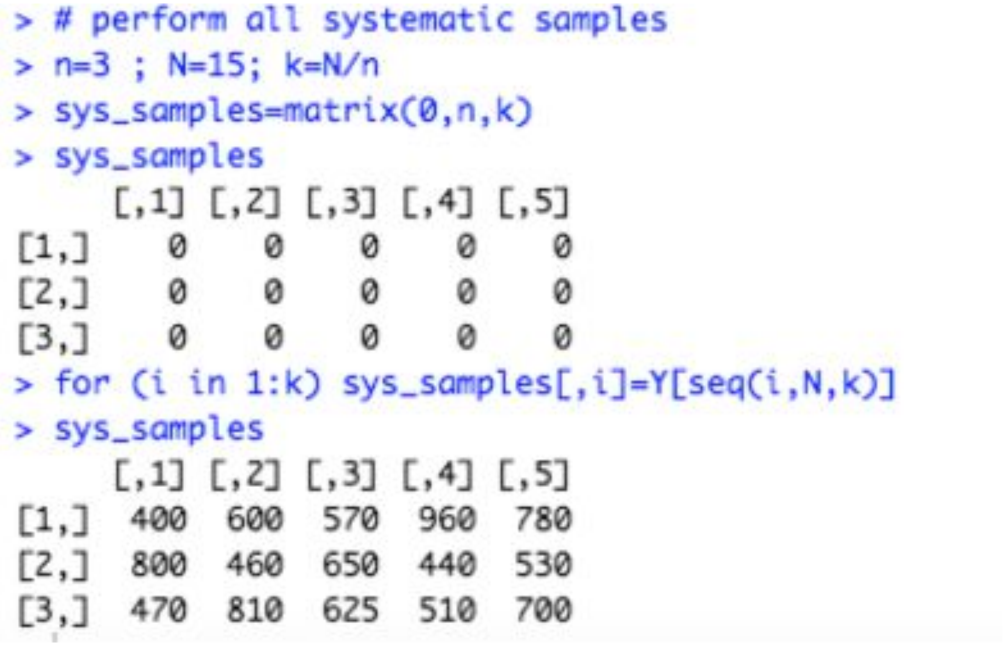
Random start (r)	Serial No. of sample units	y-values for sample units	Sample mean
1	(1, 6, 11)	400, 800, 470	556.67
2	(2, 7, 12)	600, 460, 810	623.33
3	(3, 8, 13)	570, 650, 625	615.00
4	(4, 9, 14)	960, 440, 510	636.67
5	(5, 10, 15)	780, 530, 700	670.00

■

R Codes

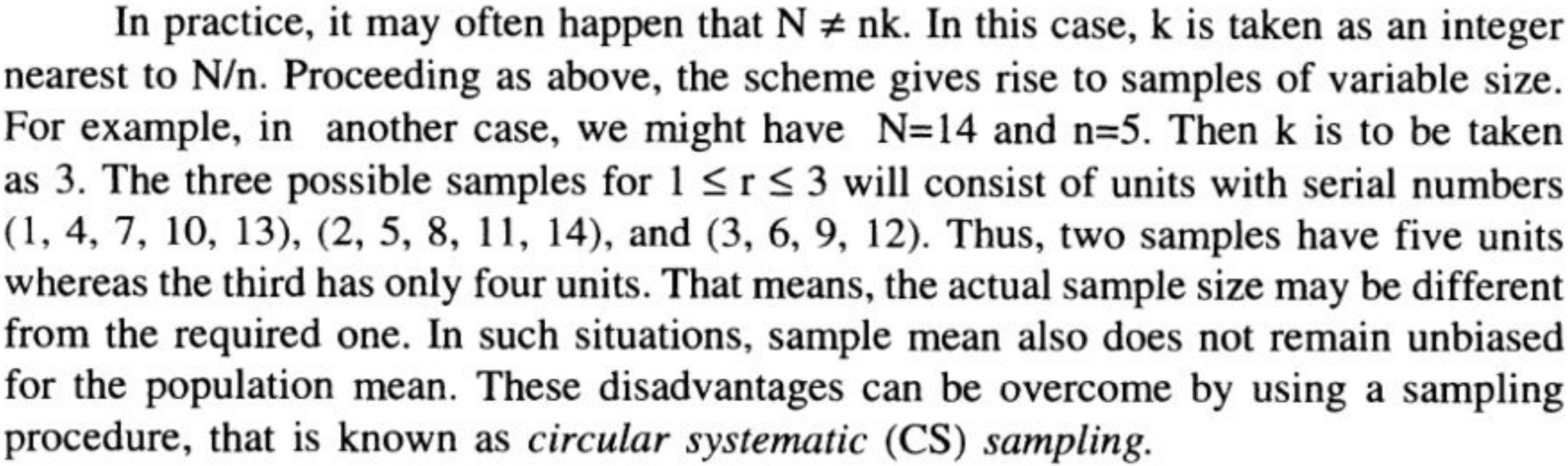


```
> #####(LS)
> Y=c(400,600,570,960, 780, 800,460, 650,440, 530, 470, 810, 625, 510, 700)
> Y
 [1] 400 600 570 960 780 800 460 650 440 530 470 810 625 510 700
>
```



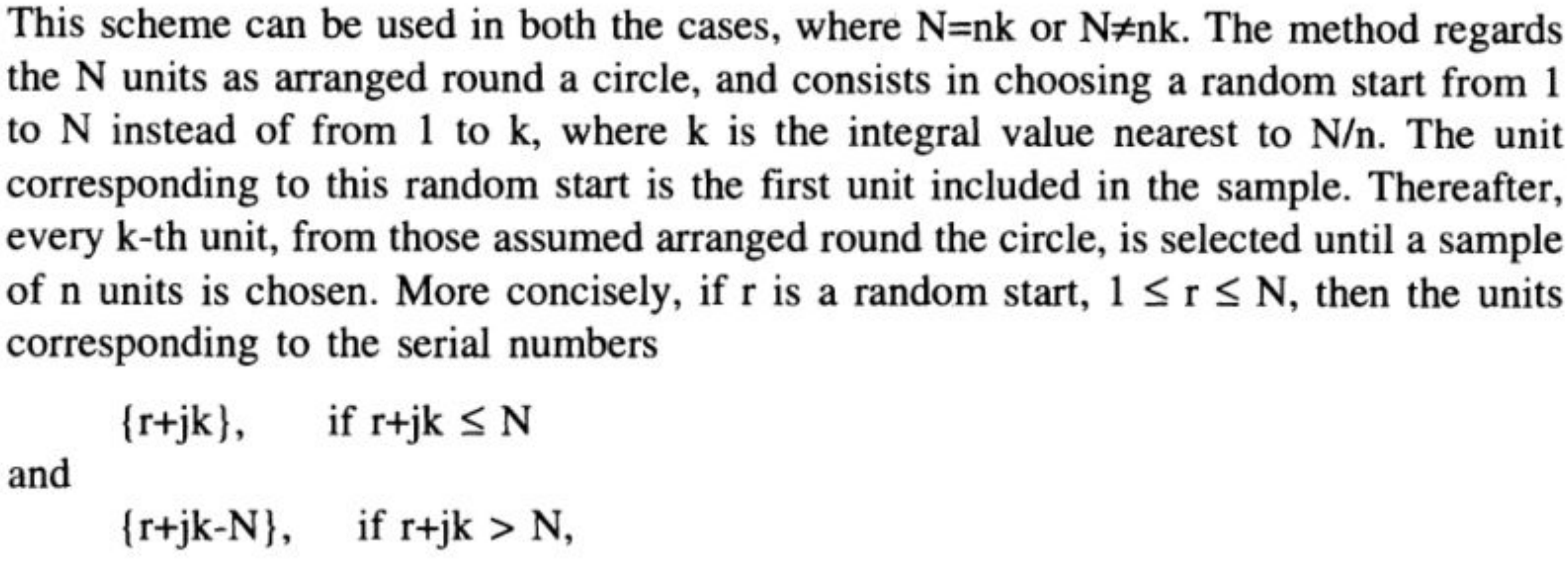
```
> # perform all systematic samples
> n=3 ; N=15; k=N/n
> sys_samples=matrix(0,n,k)
> sys_samples
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    0    0    0    0
[2,]    0    0    0    0    0
[3,]    0    0    0    0    0
> for (i in 1:k) sys_samples[,i]=Y[seq(i,N,k)]
> sys_samples
     [,1] [,2] [,3] [,4] [,5]
[1,]  400  600  570  960  780
[2,]  800  460  650  440  530
[3,]  470  810  625  510  700
```

Circular Systematic Sampling (CS)



In practice, it may often happen that $N \neq nk$. In this case, k is taken as an integer nearest to N/n. Proceeding as above, the scheme gives rise to samples of variable size. For example, in another case, we might have N=14 and n=5. Then k is to be taken as 3. The three possible samples for $1 \leq r \leq 3$ will consist of units with serial numbers (1, 4, 7, 10, 13), (2, 5, 8, 11, 14), and (3, 6, 9, 12). Thus, two samples have five units whereas the third has only four units. That means, the actual sample size may be different from the required one. In such situations, sample mean also does not remain unbiased for the population mean. These disadvantages can be overcome by using a sampling procedure, that is known as *circular systematic* (CS) *sampling*.

Circular Systematic Sampling (CS) cont.



This scheme can be used in both the cases, where $N=nk$ or $N \neq nk$. The method regards the N units as arranged round a circle, and consists in choosing a random start from 1 to N instead of from 1 to k, where k is the integral value nearest to N/n. The unit corresponding to this random start is the first unit included in the sample. Thereafter, every k-th unit, from those assumed arranged round the circle, is selected until a sample of n units is chosen. More concisely, if r is a random start, $1 \leq r \leq N$, then the units corresponding to the serial numbers

$$\{r+jk\}, \qquad \text{if } r+jk \leq N$$

and

$$\{r+jk-N\}, \qquad \text{if } r+jk > N,$$

$j = 0, 1, 2, \ldots, (n-1)$, will be selected in the sample. Theoretically, there is no problem in choosing any other smaller value of k, but it will only restrict the spread of the sample over a segment of the population. To illustrate, let N=14, n=5, and k be taken as 3. If random start r, $1 \leq r \leq 14$, is 7, then the units with serial numbers 7, 10, 13, 2, and 5 are included in the sample. Diagrammatically, this selection can be represented as below :

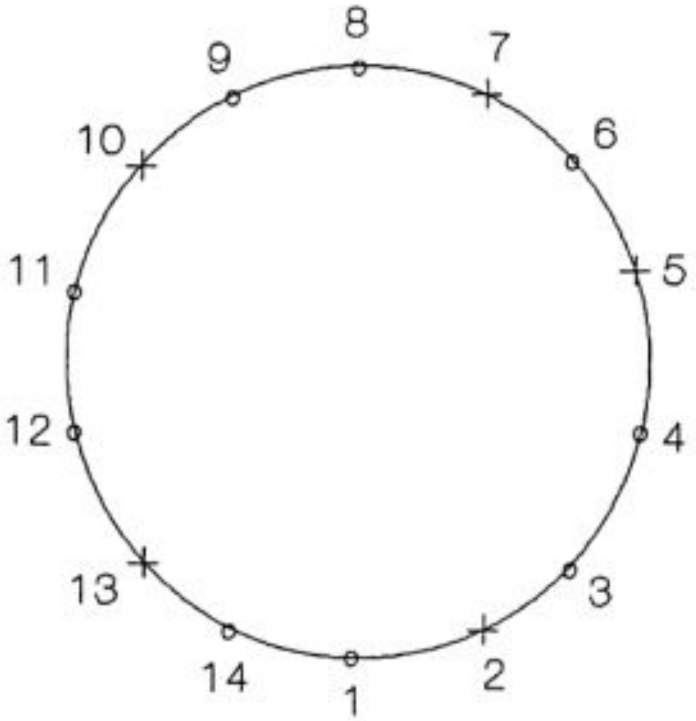
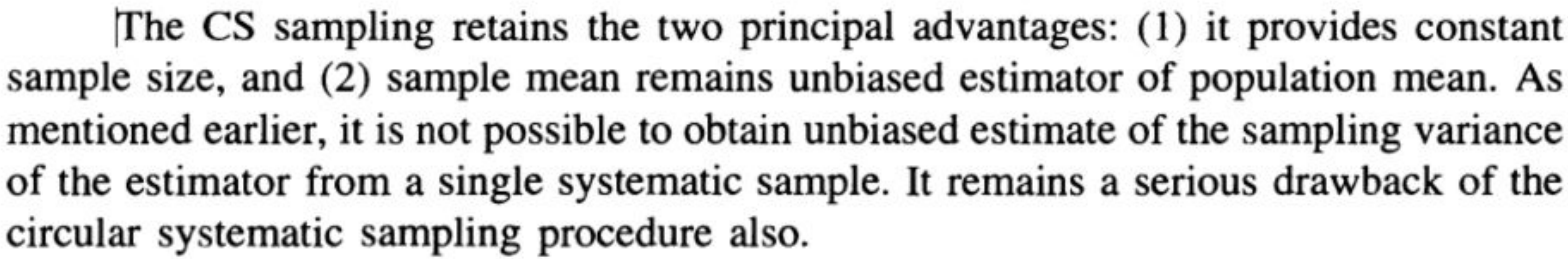


Fig. 6.1 Representation of CS sampling scheme

Circular Systematic Sampling (CS) advantages:



The CS sampling retains the two principal advantages: (1) it provides constant sample size, and (2) sample mean remains unbiased estimator of population mean. As mentioned earlier, it is not possible to obtain unbiased estimate of the sampling variance of the estimator from a single systematic sample. It remains a serious drawback of the circular systematic sampling procedure also.

Example 6.2

Using data of example 6.1, list all possible samples of size 4 along with their means, using circular systematic sampling.

Solution

We have N=15, n=4 and, therefore, k=4. In CS sampling, the random number r is selected from 1 to N. It will result in 15 random starts. Since corresponding to each random start there is one systematic sample, in all, one will obtain 15 possible systematic samples. These are listed in table 6.2 along with their means.

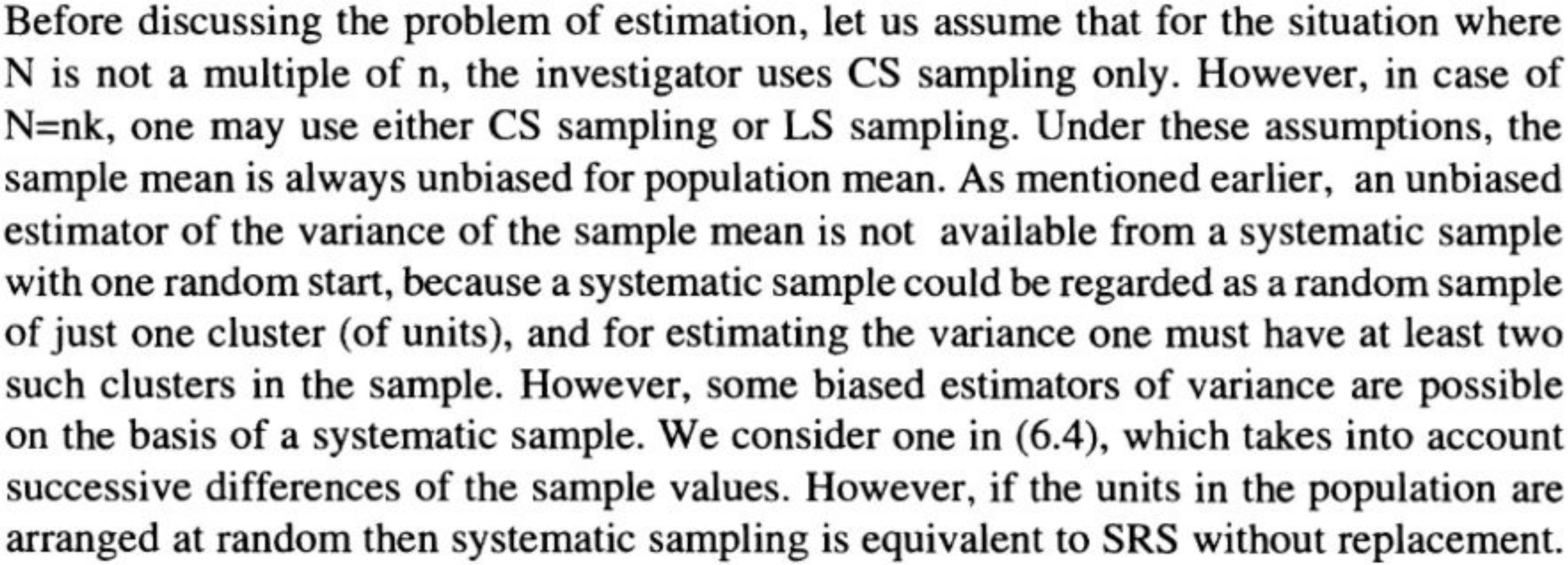
Table 6.2 Possible CS samples and their means

Random start (r)	Serial No. of sample units	y-values for sample units	Sample mean
1	(1, 5, 9, 13)	400, 780, 440, 625	561.25
2	(2, 6, 10, 14)	600, 800, 530, 510	610.00
3	(3, 7, 11, 15)	570, 460, 470, 700	550.00
4	(4, 8, 12, 1)	960, 650, 810, 400	705.00
5	(5, 9, 13, 2)	780, 440, 625, 600	611.25
6	(6, 10, 14, 3)	800, 530, 510, 570	602.50
7	(7, 11, 15, 4)	460, 470, 700, 960	647.50
8	(8, 12, 1, 5)	650, 810, 400, 780	660.00

Random start (r)	Serial No. of sample units	y-values for sample units	Sample mean
9	(9, 13, 2, 6)	440, 625, 600, 800	616.25
10	(10, 14, 3, 7)	530, 510, 570, 460	517.50
11	(11, 15, 4, 8)	470, 700, 960, 650	695.00
12	(12, 1, 5, 9)	810, 400, 780, 440	607.50
13	(13, 2, 6, 10)	625, 600, 800, 530	638.75
14	(14, 3, 7, 11)	510, 570, 460, 470	502.50
15	(15, 4, 8, 12)	700, 960, 650, 810	780.00

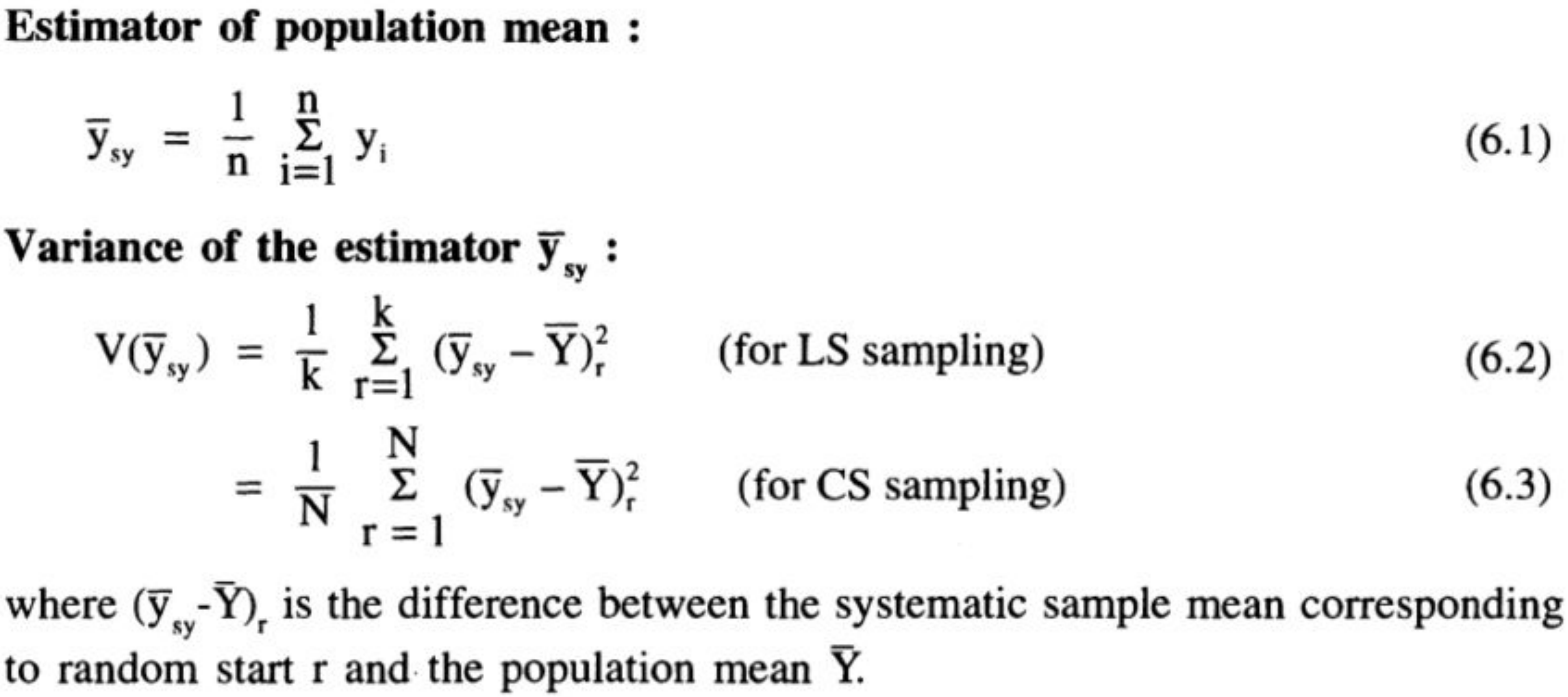
■

Estimation Issues:



Before discussing the problem of estimation, let us assume that for the situation where N is not a multiple of n, the investigator uses CS sampling only. However, in case of N=nk, one may use either CS sampling or LS sampling. Under these assumptions, the sample mean is always unbiased for population mean. As mentioned earlier, an unbiased estimator of the variance of the sample mean is not available from a systematic sample with one random start, because a systematic sample could be regarded as a random sample of just one cluster (of units), and for estimating the variance one must have at least two such clusters in the sample. However, some biased estimators of variance are possible on the basis of a systematic sample. We consider one in (6.4), which takes into account successive differences of the sample values. However, if the units in the population are arranged at random then systematic sampling is equivalent to SRS without replacement.

Estimating the mean:



Estimator of population mean :

$$\bar{y}_{sy} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{6.1}$$

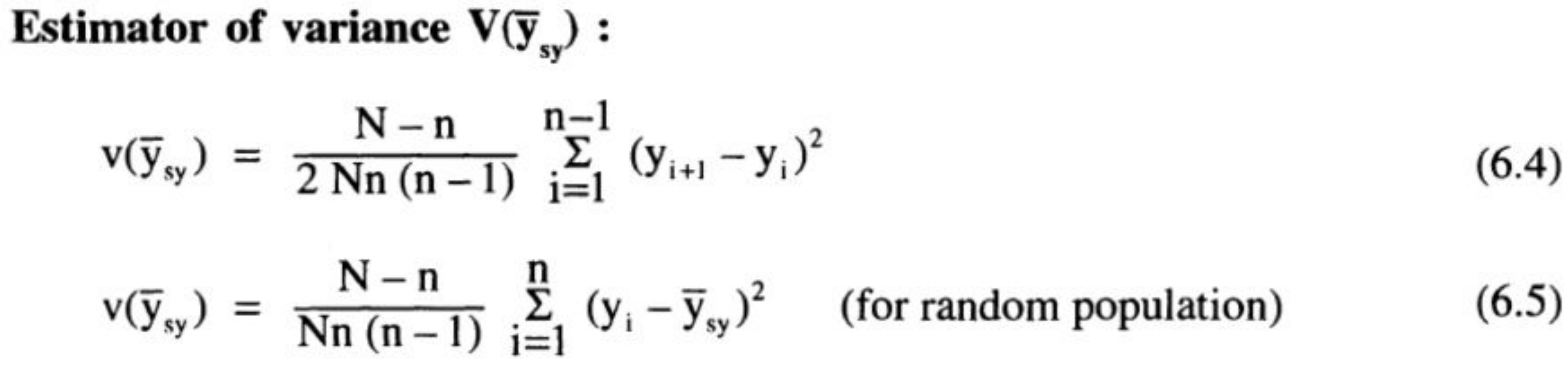
Variance of the estimator $\bar{y}_{sy}$:

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum_{r=1}^{k} (\bar{y}_{sy} - \bar{Y})_r^2 \qquad \text{(for LS sampling)} \tag{6.2}$$

$$= \frac{1}{N} \sum_{r=1}^{N} (\bar{y}_{sy} - \bar{Y})_r^2 \qquad \text{(for CS sampling)} \tag{6.3}$$

where $(\bar{y}_{sy}-\bar{Y})_r$ is the difference between the systematic sample mean corresponding to random start r and the population mean $\bar{Y}$.

Estimating the mean (Cont.)



Estimator of variance $V(\overline{y}_{sy})$:

$$v(\overline{y}_{sy}) = \frac{N-n}{2\,Nn\,(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2 \qquad (6.4)$$

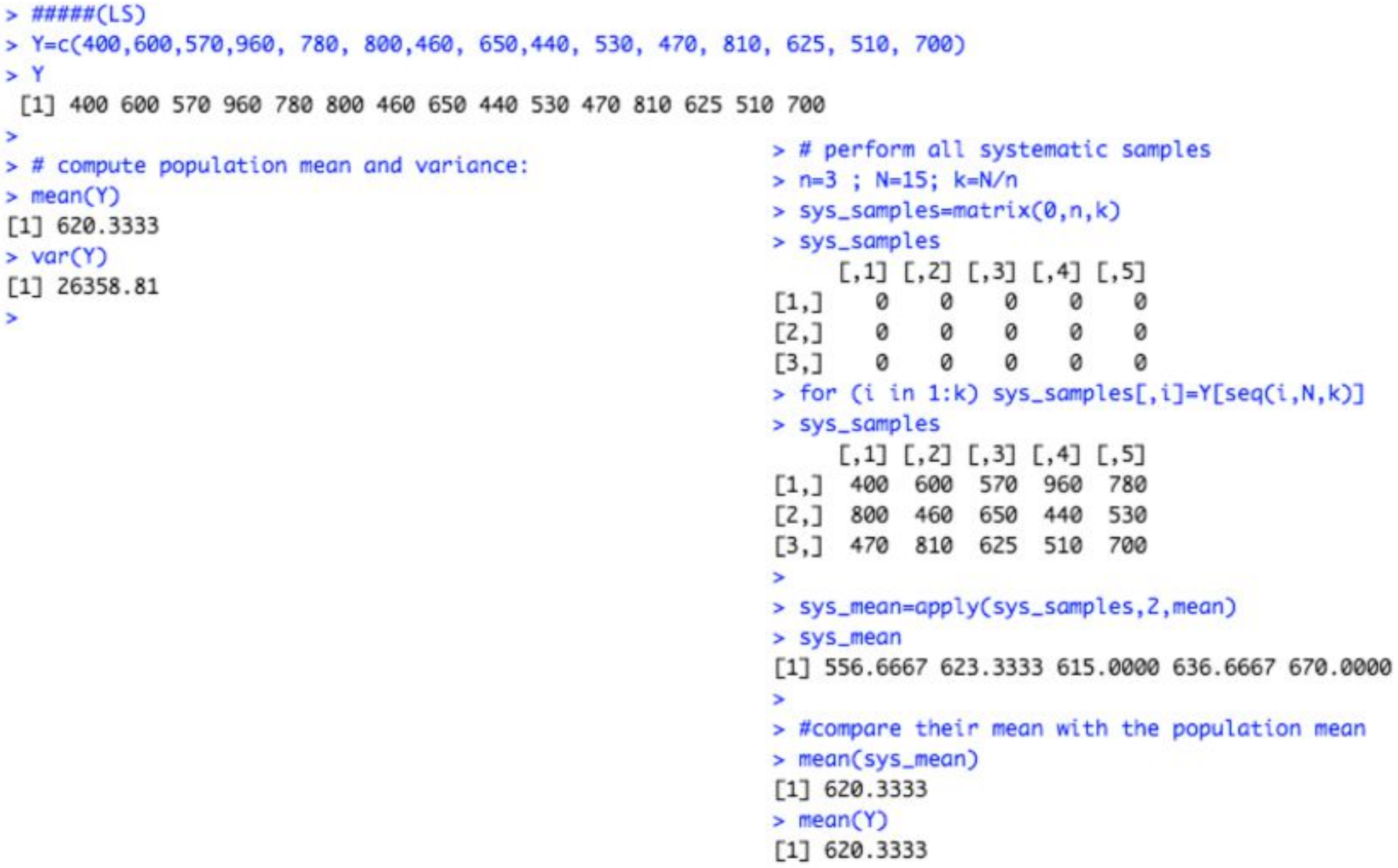
$$v(\overline{y}_{sy}) = \frac{N-n}{Nn\,(n-1)} \sum_{i=1}^{n} (y_i - \overline{y}_{sy})^2 \quad \text{(for random population)} \qquad (6.5)$$

Relative efficiency:

- $RE = \frac{V(y)}{V(y_{sys})} * 100$
- where the variances $V(y)$ is simple random sampling variance and $V(y_{sys}\ )$ is the variance of systematic sampling.
- If $RE < 100$, then simple random sampling is more efficient
- If $RE > 100$, then systematic sampling is more efficient

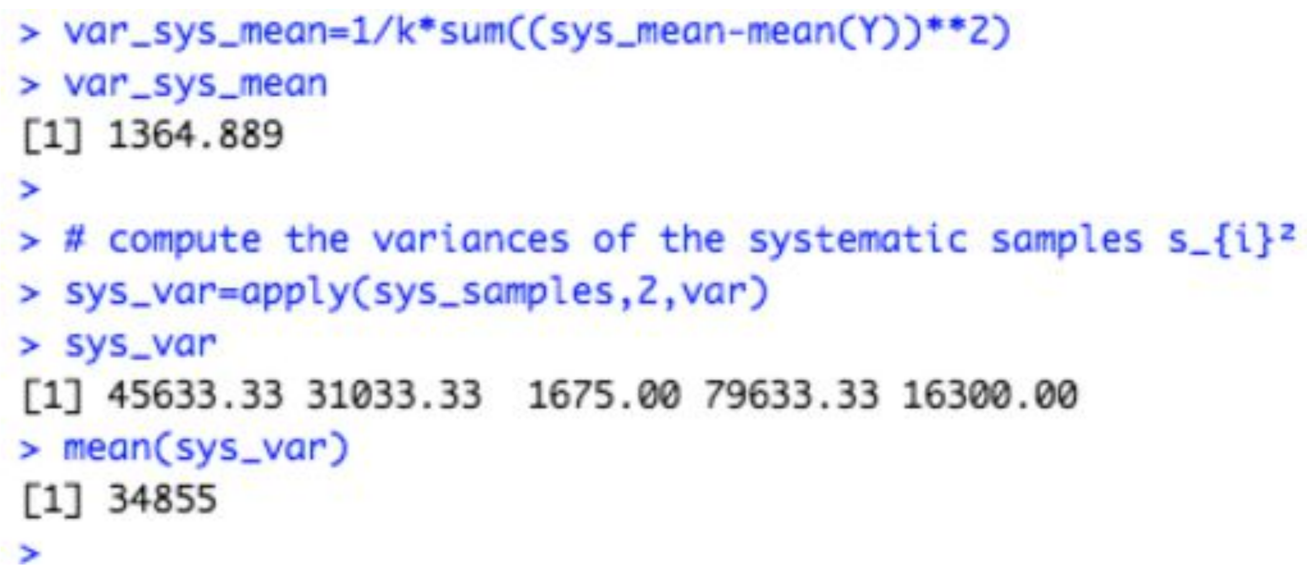
Example

- In 6.1, compute the population mean
- Perform all 1-in-5 systematic samples (LS).
- Compute their means.
- Compute their variances.
- Compute all systematic sample mean.
- Verify that systemic mean is unbiased estimator of the population mean.
- Compute the variance of the systematic sample mean $Var(\bar{y}_{sys})$
- Compute the variance of the SRS mean $Var(\bar{y})$ with $n = 20$.
- Find the relative efficiency of the variance of the simple random mean, $Var(\bar{y})$, and the variance of the systematic mean, $Var(\bar{y}_{sys})$.

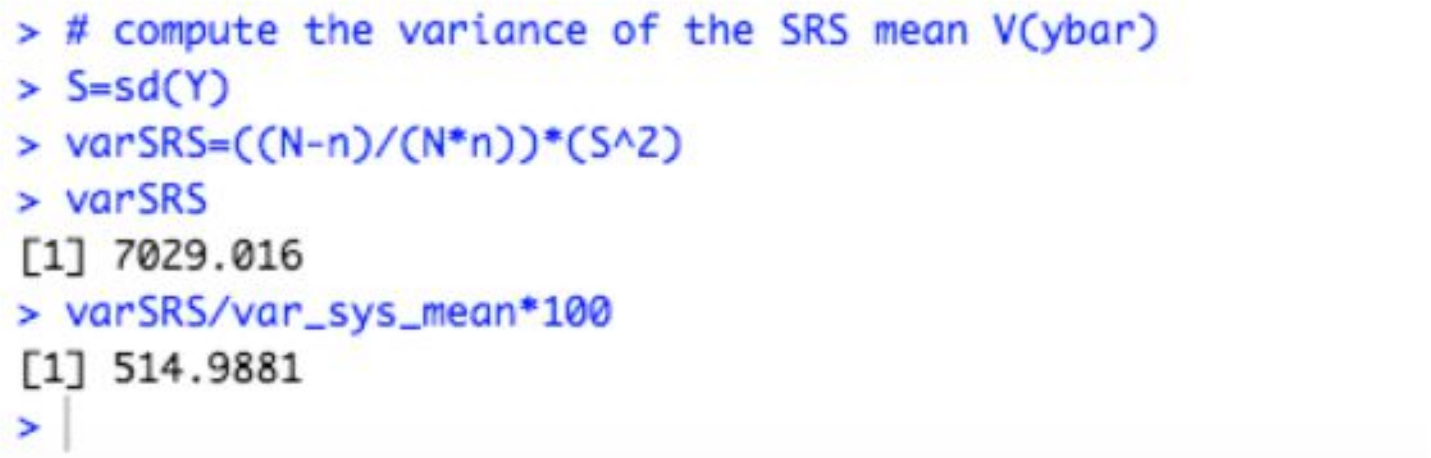


```
> #####(LS)
> Y=c(400,600,570,960, 780, 800,460, 650,440, 530, 470, 810, 625, 510, 700)
> Y
 [1] 400 600 570 960 780 800 460 650 440 530 470 810 625 510 700
>
> # compute population mean and variance:
> mean(Y)
[1] 620.3333
> var(Y)
[1] 26358.81
>
```

```
> # perform all systematic samples
> n=3 ; N=15; k=N/n
> sys_samples=matrix(0,n,k)
> sys_samples
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    0    0    0    0
[2,]    0    0    0    0    0
[3,]    0    0    0    0    0
> for (i in 1:k) sys_samples[,i]=Y[seq(i,N,k)]
> sys_samples
     [,1] [,2] [,3] [,4] [,5]
[1,]  400  600  570  960  780
[2,]  800  460  650  440  530
[3,]  470  810  625  510  700
>
> sys_mean=apply(sys_samples,2,mean)
> sys_mean
[1] 556.6667 623.3333 615.0000 636.6667 670.0000
>
> #compare their mean with the population mean
> mean(sys_mean)
[1] 620.3333
> mean(Y)
[1] 620.3333
```



```
> var_sys_mean=1/k*sum((sys_mean-mean(Y))**2)
> var_sys_mean
[1] 1364.889
>
> # compute the variances of the systematic samples s_{i}²
> sys_var=apply(sys_samples,2,var)
> sys_var
[1] 45633.33 31033.33  1675.00 79633.33 16300.00
> mean(sys_var)
[1] 34855
>
```



```
> # compute the variance of the SRS mean V(ybar)
> S=sd(Y)
> varSRS=((N-n)/(N*n))*(S^2)
> varSRS
[1] 7029.016
> varSRS/var_sys_mean*100
[1] 514.9881
>
```

Example 6.3 (for N=nk)

About 70 years back, *Dalbergia sissoo* trees were planted in a single row on both sides of a road. The total number of trees are 3600. The Department of Public Works of a state is interested in estimating the total timber volume. A 1-in-100 systematic sample is selected. The data on estimated timber volume for the sampled trees (procedure of selection is given in solution) are presented in table 6.3. Estimate the total timber volume, and also construct the confidence interval for it.

Table 6.3 Timber volume (in cubic meters) for 36 selected trees

Serial No. of tree	Timber volume	Serial No. of tree	Timber volume	Serial No. of tree	Timber volume
28	1.72	1228	2.17	2428	1.89
128	1.29	1328	1.63	2528	1.63
228	1.08	1428	1.91	2628	2.23
328	2.29	1528	1.66	2728	2.40
428	2.01	1628	1.56	2828	2.51
528	1.77	1728	2.26	2928	2.57

Table 6.3 continued ...

Serial No. of tree	Timber volume	Serial No. of tree	Timber volume	Serial No. of tree	Timber volume
628	1.63	1828	2.49	3028	1.26
728	1.20	1928	2.26	3128	1.46
828	2.03	2028	2.31	3228	1.00
928	1.17	2128	1.60	3328	1.94
1028	2.47	2228	1.64	3428	1.80
1128	1.86	2328	1.43	3528	1.60

Solution

In this case, population size N=3600 and sampling interval k=100. We use linear systematic sampling for the selection of trees. Let the random number r selected from 1 to k(=100) be 28. The trees bearing serial numbers 28, 128, 228, 328, ..., 3528 will, therefore, be selected in the sample. The timber volume observed for the sample trees is given in table 6.3.

Estimate of total timber volume from (6.1) is

$$\hat{Y}_{sy} = N\,\bar{y}_{sy} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

$$= \frac{3600}{36}(1.72 + 1.29 + ... + 1.60)$$

$$= \frac{3600}{36}(65.73)$$

$$= 6573$$

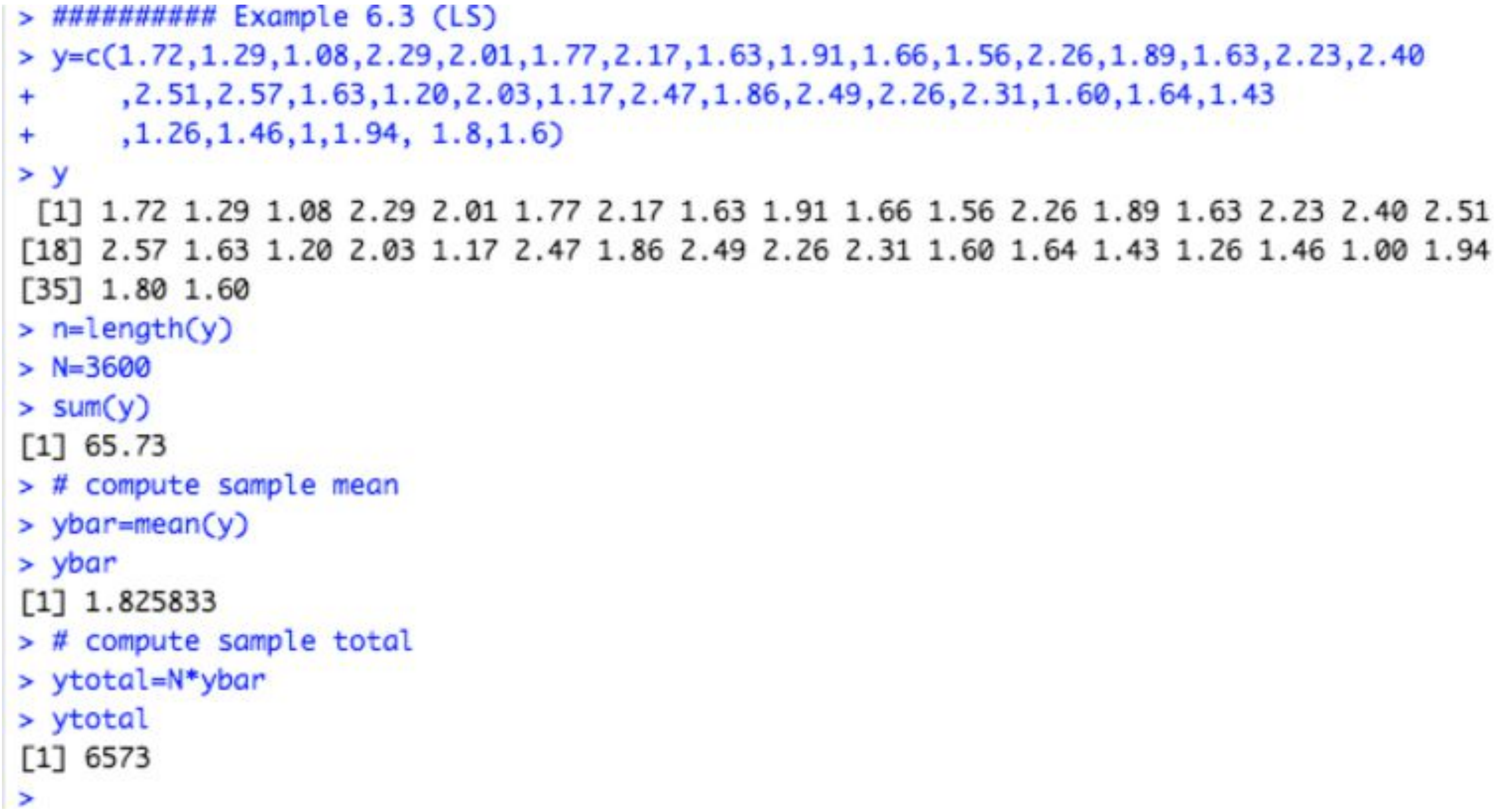
We now work out the estimate of variance $V(\hat{Y}_{sy})$ from (6.4) as

$$
\begin{aligned}
v(\hat{Y}_{sy}) &= N^2 v(\overline{y}_{sy}) = \frac{N(N-n)}{2n(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2 \\
&= \frac{N\,(N-n)}{2n(n-1)} [(y_2 - y_1)^2 + (y_3 - y_2)^2 + \ldots + (y_{36} - y_{35})^2] \\
&= \frac{3600(3600-36)}{2(36)\,(35)} [(1.29 - 1.72)^2 + (1.08 - 1.29)^2 + \ \ldots + (1.60 - 1.80)^2] \\
&= \frac{3600(3600-36)}{2\,(36)\,(35)} (10.8056) \\
&= 55015.94
\end{aligned}
$$

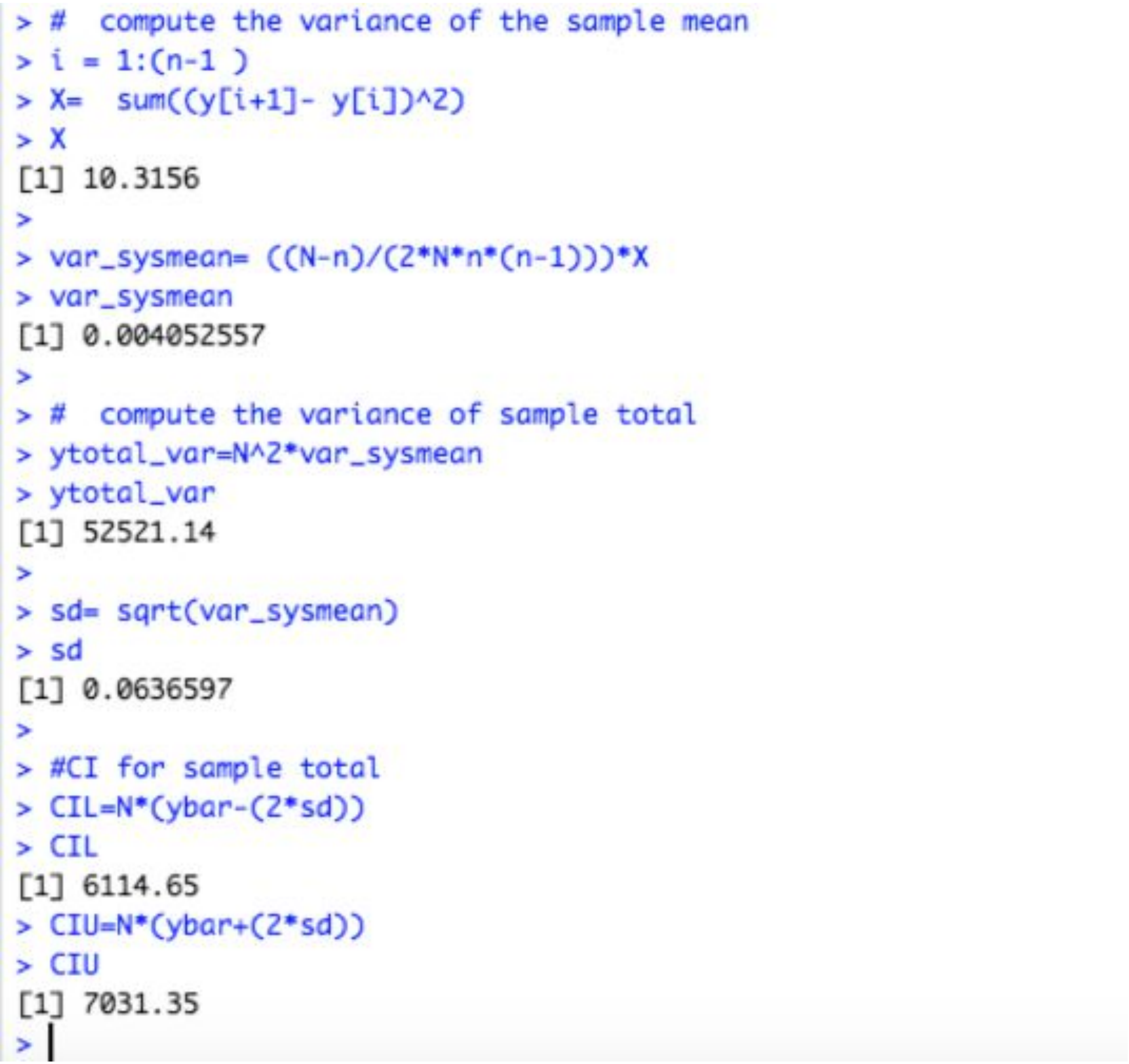
Using the estimate for total timber volume and the estimate of its variance, we now calculate the confidence interval for population total from (2.8). It is given by

$$\begin{aligned} & N\overline{y}_{sy} \pm 2N\sqrt{v(\overline{y}_{sy})} \\ &= \hat{Y}_{sy} \pm 2\sqrt{v(\hat{Y}_{sy})} \\ &= 6573 \pm 2\sqrt{55015.94} \\ &= 6103.89,\ 7042.11 \end{aligned}$$

To summarize, the estimate of total timber volume obtained from the selected sample is 6573 cubic meters. It can be said with probability approximately equal to .95, that the actual total timber volume that can be had from all the 3600 trees, would be in the range of 6103.89 to 7042.11 cubic meters. ■



```
> ########## Example 6.3 (LS)
> y=c(1.72,1.29,1.08,2.29,2.01,1.77,2.17,1.63,1.91,1.66,1.56,2.26,1.89,1.63,2.23,2.40
+     ,2.51,2.57,1.63,1.20,2.03,1.17,2.47,1.86,2.49,2.26,2.31,1.60,1.64,1.43
+     ,1.26,1.46,1,1.94, 1.8,1.6)
> y
 [1] 1.72 1.29 1.08 2.29 2.01 1.77 2.17 1.63 1.91 1.66 1.56 2.26 1.89 1.63 2.23 2.40 2.51
[18] 2.57 1.63 1.20 2.03 1.17 2.47 1.86 2.49 2.26 2.31 1.60 1.64 1.43 1.26 1.46 1.00 1.94
[35] 1.80 1.60
> n=length(y)
> N=3600
> sum(y)
[1] 65.73
> # compute sample mean
> ybar=mean(y)
> ybar
[1] 1.825833
> # compute sample total
> ytotal=N*ybar
> ytotal
[1] 6573
> 
```



```
> #  compute the variance of the sample mean
> i = 1:(n-1 )
> X=  sum((y[i+1]- y[i])^2)
> X
[1] 10.3156
>
> var_sysmean= ((N-n)/(2*N*n*(n-1)))*X
> var_sysmean
[1] 0.004052557
>
> #  compute the variance of sample total
> ytotal_var=N^2*var_sysmean
> ytotal_var
[1] 52521.14
>
> sd= sqrt(var_sysmean)
> sd
[1] 0.0636597
>
> #CI for sample total
> CIL=N*(ybar-(2*sd))
> CIL
[1] 6114.65
> CIU=N*(ybar+(2*sd))
> CIU
[1] 7031.35
> |
```

HW

- Page 161 exercises 6.4, 6.5, and 6.6
- Assume the data that we have from 100 observations as follows and consider the simple random sampling without replacement (SRS) with $n = 20$.

RNGkind(sample.kind = "Rejection")

$$\boldsymbol{set.seed}(\mathbf{111})$$

$$\boldsymbol{Y} = \boldsymbol{sample}(\mathbf{1}{:}\,\mathbf{40}, \mathbf{100}, \boldsymbol{replace} = \boldsymbol{TRUE})$$

- Compute the population mean,
- Perform all 1-in-5 systematic samples (LS).
- Compute their means.
- Compute their variances.
- Compute all systematic sample mean.
- Verify that systemic mean is unbiased estimator of the population mean.
- Compute the variance of the systematic sample mean $Var(\bar{y}_{sys})$
- Compute the variance of the SRS mean $Var(\bar{y})$ with $n = 20$.
- Find the relative efficiency of the variance of the simple random mean, $Var(\bar{y})$, and the variance of the systematic mean, $Var\left(\bar{y}_{sys}\right)$.