CHAPTER 6

Systematic Sampling

6.1 LINEAR SYSTEMATIC SAMPLING

In the preceding chapters, we have considered methods of sampling in which successive units are selected at random. In this chapter, an alternative sampling procedure is considered. The scheme, besides ensuring for each unit equal probability of inclusion in the sample, selects the whole sample with just one random number.

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 :

1	2	3	...	r	...	k
k+1	k+2	k+3	...	k+r	...	2k
2k+1	2k+2	2k+3	...	2k+r	...	3k
.	.	.	.	.		
.	.	.	.	.		
.	.	.	.	.		
(n-1)k+1	(n-1)k+2	(n-1)k+3	...	(n-1)k+r	...	nk

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.

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.

The systematic sampling has the nice feature of operational convenience. It lies in the fact that the selection of the first unit determines the whole sample. This operation is easier to understand, and can be speedily executed in relation to simple random sampling. The sampling procedure is particularly suited for situations, where the sample is to be selected by the field staff themselves. Because of simplicity in the execution of systematic sampling, it would be easy to train persons in using it. Thus, it would be desirable to employ this procedure, whenever the sampling work has to be carried out by a large number of persons stationed in different areas.

Systematic samples are well spread over the population, and there is no risk that any large contiguous part of the population will be left unrepresented. This scheme provides more efficient estimates of population mean/total in comparison to simple random sampling for populations with *linear trend*, where study variable values in the population tend to be linearly related with the serial numbers of units in the frame . It is also efficient for populations with autocorrelation.

Systematic sampling should, however, be used with considerable care in case of periodicity in the population where the values of the study variable for the units listed in the frame, tend to increase and then decrease and increase again in a cyclic manner with a definite period. For *periodic populations*, if the sampling interval is an odd multiple of half the period of the cycle, systematic sampling provides estimator of mean with considerably small variance. On the other hand, if the sampling interval is an integral multiple of the period of the cycle, systematic sampling is no better than usual simple random sampling. The periodic variation is likely to occur when the population consists of groups of equal or approximately equal number of units, and the units within each group are arranged according to some definite sequence. For instance, in a population census, the households are the sampling units, and the individuals in the family are arranged according to a set pattern, such as, head of the family first, then his wife and their children in order of their age. In such a case, systematic sampling with an interval equal to the group size, or its multiple, may lead to inefficient estimates since the units selected in a sample will tend to be more or less similar in respect of the characteristic under study. Other situations, where the populations may exhibit periodicity, could be the flow of road traffic past a point over 24 hours of the day, or the store sales over seven days of the week. When estimating an average over a time period, a systematic sample daily at 6 p.m. would obviously be injudicious. Instead, the investigator should see that the time and week days are equally represented for such estimation situations. The point to stress is that one should carefully examine the possibility of existing periodicity in the population. If it exists, it can rather be helpful in reducing the variation in sample estimates. A serious disadvantage of this scheme lies in its use with populations having unforeseen periodicity, which may substantially contribute to the bias in the estimate of mean/total. Stephan *et al.* (1940) and Lahiri (1954) have discussed, in detail, the pitfalls involved in using systematic sampling in case of populations having cyclic

variation.

Another serious disadvantage of the sampling scheme is that the variance of the estimator can not be estimated unbiasedly from a single sample.

This sampling procedure has been found to be very useful in forest surveys for estimating the volume of timber. Systematic samples of boats returning from sea, are used for estimating the total catch of fish. The sampling scheme has also been used in milk yield surveys for estimating the lactation yield.

We now consider an example to illustrate the use of linear systematic sampling for selection of samples.

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

■

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.*

6.2 CIRCULAR SYSTEMATIC SAMPLING

This scheme can be used in both the cases, where N=nk or N≠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

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

and

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

$j = 0, 1, 2, ..., (n-1)$, 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 :

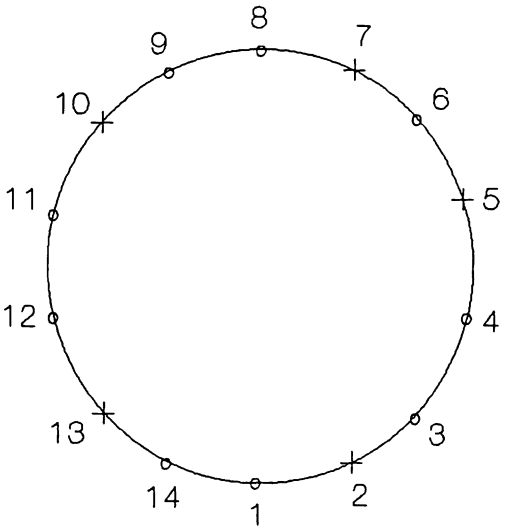

Fig. 6.1 Representation of CS sampling scheme

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
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

■

6.3 ESTIMATING MEAN/ TOTAL

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.

In this case, the expression for variance estimator is same as in (3.10). For the sake of completeness, it is also given in (6.5).

Estimator of population mean :

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

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

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

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

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

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

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

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

As usual, the estimate of population total Y will be $\hat{Y}_{sy} = N\,\bar{y}_{sy}$. The variance and its estimator will be given by $V(\hat{Y}_{sy}) = N^2V(\bar{y}_{sy})$ and $v(\hat{Y}_{sy}) = N^2v(\bar{y}_{sy})$.

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

$$v(\hat{Y}_{sy}) = N^2 v(\bar{y}_{sy}) = \frac{N(N-n)}{2n(n-1)}\sum_{i=1}^{n-1}(y_{i+1}-y_i)^2$$

$$= \frac{N(N-n)}{2n(n-1)}[(y_2-y_1)^2 + (y_3-y_2)^2 + ... + (y_{36}-y_{35})^2]$$

$$= \frac{3600(3600-36)}{2(36)(35)}[(1.29-1.72)^2+(1.08-1.29)^2+ ...+(1.60-1.80)^2]$$

$$= \frac{3600(3600-36)}{2(36)(35)}(10.8056)$$

$$= 55015.94$$

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

$$N\bar{y}_{sy} \pm 2N\sqrt{v(\bar{y}_{sy})}$$

$$= \hat{Y}_{sy} \pm 2\sqrt{v(\hat{Y}_{sy})}$$

$$= 6573 \pm 2\sqrt{55015.94}$$

$$= 6103.89, 7042.11$$

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 6.4 (for N≠nk)

On a particular day, 162 boats had gone to sea from the coast for fishing. It was desired to estimate the total catch of fish at the end of the day. As it was not possible to weigh the catch for all the 162 boats, it was decided to weigh fish for only 15 boats selected using circular systematic sampling. Discuss the selection procedure, and obtain the estimate of total catch of fish using data on the 15 sample boats given in table 6.4.

Table 6.4 Catch of fish (in quintals) for 15 selected boats

Serial No. of boat	Catch of fish	Serial No. of boat	Catch of fish	Serial No. of boat	Catch of fish
73	5.614	128	9.225	21	8.460
84	8.202	139	6.640	32	10.850
95	6.115	150	7.350	43	6.970
106	9.765	161	5.843	54	5.524
117	8.550	10	6.875	65	7.847

Solution

In this case, we have N=162 and n=15. Since N/n=162/15=10.8 is not a whole number, the value of sampling interval k is taken as 11, an integer nearest to 10.8, and circular systematic sampling is used for selection of boats. If the selected random number r, $1 \le r \le 162$, is 73, then the boats bearing serial numbers 73, 84,..., 65 will be included in the sample. The serial numbers of selected boats, along with the corresponding catch of fish, are presented in table 6.4. We now proceed to estimate the total catch of fish using (6.1). This estimate is

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

$$= \frac{162}{15}(5.614 + 8.202 + \ldots + 7.847)$$

$$= \frac{(162)(113.83)}{15}$$

$$= 1229.364$$

The estimate of variance $V(\hat{Y}_{sy})$ is then computed by using the expression (6.4). Thus,

$$v(\hat{Y}_{sy}) = N^2 v(\bar{y}_{sy}) = \frac{N(N-n)}{2n(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2$$

$$= \frac{162(162-15)}{2(15)(14)} [(8.202 - 5.614)^2 + (6.115 - 8.202)^2 + ... + (7.847 - 5.524)^2]$$

$$= \frac{(162)(162-15)(67.596)}{2(15)(14)}$$

$$= 3832.693$$

The confidence interval, for the total catch of fish for 162 boats, can then be calculated from

$$\hat{Y}_{sy} \pm 2\sqrt{v(\hat{Y}_{sy})}$$

$$= 1229.364 \pm 2\sqrt{3832.693}$$

$$= 1105.547, 1353.181$$

Thus, the estimate of total catch of fish obtained from a single sample is 1229.364 quintals. The confidence limits, obtained above, indicate that the total catch from all the 162 boats is likely to fall in the interval [1105.547, 1353.181] quintals. ■

The variance estimator given in (6.4) is biased and, therefore, should be used with care as inferences based on these estimates may sometimes be misleading in practice. Various approaches have been suggested to obtain unbiased variance estimators. Using a mixture of systematic and simple random sampling, Zinger (1963, 1964) suggested an unbiased estimator $\bar{y}_z$ of population mean but his proposed unbiased estimator of variance $V(\bar{y}_z)$ could not be proved nonnegative. Subsequently, Rana and Singh (1989) proposed another unbiased estimator of population mean, and also gave an unbiased estimator of variance of this estimator which was proved to be nonnegative. Discussion of these methods is, however, beyond the scope of this book. An alternative approach which provides unbiased estimator of variance, is through the use of interpenetrating subsampling. This approach we discuss in the following section.

6.4 ESTIMATING MEAN/TOTAL THROUGH INTERPENETRATING SUBSAMPLES

A method of estimating the variance $V(\bar{y}_{sy})$ unbiasedly, consists in selecting the sample of required size n in the form of two or more (say m) systematic subsamples of same size with independent random starts. Let us assume that these m *interpenetrating subsamples,* each of size n/m, are to be selected from the population of N units. Also, let N/n = k. Then, for selecting the required m samples using linear systematic sampling, we select m random starts, either using simple random sampling WR, or using simple random sampling WOR, from 1 to mk (assumed to be integer). The subsample corresponding to a particular random start will thus include population units with serial numbers at intervals of mk. In case mk is not an integer, we can use circular systematic

sampling for selecting the subsamples. Let $\bar{y}_1, \bar{y}_2, ..., \bar{y}_m$ be the estimators of the population mean based on m such subsamples, each of size n/m.

When m random starts are selected through simple random sampling WR, we have the following results :

Unbiased estimator of population mean $\bar{Y}$:

$$\bar{y}_{sy} = \frac{1}{m} \sum_{i=1}^{m} \bar{y}_i \tag{6.6}$$

Unbiased estimator of variance $V(\bar{y}_{sy})$:

$$v(\bar{y}_{sy}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} (\bar{y}_i - \bar{y}_{sy})^2 \tag{6.7}$$

If the sampling interval is small, selection of m random starts with replacement may lead to repetition of samples. In such cases, it is, therefore, desirable to select m interpenetrating systematic subsamples of n/m units each, with random starts selected from 1 to mk using without replacement sampling. This results in the selection of m of the mk possible samples with SRS without replacement. The formulas corresponding to this procedure are listed below :

Unbiased estimator of population mean $\bar{Y}$:

$\bar{y}_{sy}$ is same as in (6.6).

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

$$V(\bar{y}_{sy}) = \frac{k-1}{km(km-1)} \sum_{i=1}^{mk} (\bar{y}_i - \bar{Y})^2 \tag{6.8}$$

where $\bar{y}_i$ is the i-th subsample mean, i=1,2,...,mk.

Unbiased variance estimator :

$$v(\bar{y}_{sy}) = \frac{k-1}{km(m-1)} \sum_{i=1}^{m} (\bar{y}_i - \bar{y}_{sy})^2 \tag{6.9}$$

where $\bar{y}_1, \bar{y}_2, ..., \bar{y}_m$ are estimators based on m systematic samples with random starts selected with SRS without replacement.

Example 6.5

A dairy research institute is interested in estimating the total milk yield of a buffalo in connection with a breeding program. The milk yield of first five days was not recorded, being the colostrum period. The total lactation period was taken as 300 days. It was decided to select 3 systematic subsamples each of size 10 days, so as to arrive at a total

sample size of 30 days. The method of selecting the subsamples is discussed in the solution. Milk yield recorded for the selected days is given in table 6.5.

Table 6.5 Milk yield (in liters) for 30 selected days

Subsample I		Subsample II		Subsample III	
Selected days	Milk yield	Selected days	Milk yield	Selected days	Milk yield
1	8.10	12	9.30	26	11.15
31	12.00	42	13.50	56	14.70
61	15.20	72	14.40	86	14.60
91	14.00	102	14.35	116	12.80
121	11.25	132	9.80	146	10.65
151	10.10	162	10.00	176	10.60
181	9.80	192	8.60	206	8.30
211	8.75	222	8.40	236	7.50
241	7.25	252	6.10	266	4.30
271	4.10	282	3.10	296	2.20
Mean	10.055		9.755		9.680

Solution

In this problem, we have N=300, n=30, and m=3. This gives k=300/30=10. In order to avoid the possibility of repetition of subsamples, we select WOR three random starts from 1 to mk=30. Suppose we get the random starts as 1, 12, and 26. Corresponding to these three random starts, the selected days and corresponding milk yields are recorded in table 6.5. Subsample means are also given at the end of the table.

Estimate of total milk yield (in liters) can be obtained by using (6.6) as

$$\hat{Y}_{sy} = N\bar{y}_{sy} = \frac{N}{m}\sum_{i=1}^{m}\bar{y}_i$$

$$= \frac{300}{3}(10.055 + 9.755 + 9.680)$$

$$= 300\,(9.830)$$

$$= 2949$$

In this case, the estimate of variance is given by (6.9). Therefore,

$$v(\hat{Y}_{sy}) = N^2 v(\bar{y}_{sy})$$

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

$$= \frac{(300)^2(10-1)}{10(3)(3-1)}[(10.055-9.830)^2 + (9.755-9.830)^2 + (9.680-9.830)^2]$$

$$= 1063.125$$

Confidence limits for total milk yield are derived from

$$\hat{Y}_{sy} \pm 2\sqrt{v(\hat{Y}_{sy})}$$

$$= 2949 \pm 2\sqrt{1063.125}$$

$$= 2883.789,\ 3014.211$$

The confidence limits computed above indicate that the total milk yield, if all the 300 observations were recorded, would most probably fall in the closed interval [2883.789, 3014.211] liters. ■

6.5 SAMPLE SIZE DETERMINATION FOR ESTIMATING MEAN/TOTAL

Assuming the population in random order, let n_1 be the number of units selected, in preliminary sample, from the population of N units. Then from the sampled n_1 units, we compute

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_i - \bar{y}_{sy1})^2 \qquad (6.10)$$

where

$$\bar{y}_{sy1} = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i$$

It may be noted that a systematic sample from a population in random order is equivalent to a simple random sample drawn without replacement. Therefore, the formulas for finding required sample size, obtained in section 3.5, will be applicable in this case also. These formulas are reproduced below for readers' convenience.

Sample size for estimating mean/total with a permissible error B :

$$n = \frac{Ns_1^2}{ND + s_1^2} \qquad (6.11)$$

where

$$D = \frac{B^2}{4} \quad \text{(when estimating mean)}$$

$$D = \frac{B^2}{4N^2} \quad \text{(when estimating total)}$$

with s_1^2 defined in (6.10). If $n_1 \geq n$, the sample size n_1 is sufficient, otherwise, $(n-n_1)$ additional units need to be selected.

As stated above, the assumption of population being in random order amounts to reducing systematic sampling to simple random sampling. Because of this, the expression in (6.11) above is same as the expressions in (3.18) and (3.20). In absence of this assumption, (6.11) could give an extra large sample for populations with linear trend and

too small a sample for periodic populations. Also, when the preliminary systematic sample of n_1 units is augmented by another systematic sample of $(n-n_1)$ units, the composite sample does not strictly remain a systematic sample.

Example 6.6

Assuming the population in example 6.3 to be in random order, and treating the sample of 36 trees selected there as the preliminary sample, determine the sample size required to estimate total timber volume with a tolerable error of 400 cubic meters.

Solution

Here, we are given B=400 cubic meters. From example 6.3, we have N=3600, n_1=36, and

$$\bar{y}_{sy1} = \frac{1}{36}(1.72+1.29+\ldots+1.60)$$
$$= 1.826$$

so that,

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_i-\bar{y}_{sy1})^2$$
$$= \frac{1}{n_1-1}\left(\sum_{i=1}^{n_1} y_i^2 - n_1\bar{y}_{sy1}^2\right)$$
$$= \frac{1}{36-1}[(1.72)^2+(1.29)^2+\ldots+(1.60)^2-36(1.826)^2]$$
$$= .1919$$

Also,

$$ND = \frac{B^2}{4N}$$
$$= \frac{(400)^2}{4(3600)} = 11.1111$$

Then, from (6.11), we can determine the required sample size as

$$n = \frac{Ns_1^2}{ND+s_1^2}$$
$$= \frac{(3600)(.1919)}{11.1111+(.1919)}$$
$$= 61.12$$
$$\approx 61$$

Since the sample size required to estimate the total timber volume with a permissible error of 400 cubic meters is 61, the investigator will, therefore, need to select 61-36=25 more trees to get the estimate with specified magnitude of tolerable error. ■

6.6 ESTIMATION OF PROPORTION

Sometimes, the investigator is interested in estimating the proportion P of population possessing a particular attribute, from a systematic sample. For instance, he/she may wish to estimate the proportion of voters who are satisfied with the functioning of Municipal Committee, an elected body of a particular town. In such a situation, it would be convenient to select a 1-in-k systematic sample from the list of registered voters in place of the usual simple random sample. All relevant formulas corresponding to the estimator of P can be obtained from the formulas for $\bar{y}_{sy}$, by taking $y_i=0$ if the i-th sample unit does not possess the specified attribute, and $y_i=1$ if it does. The estimator p_{sy} would thus be the average of the 0 and 1 values assigned to the units in the sample. Assuming that n_1 units in the systematic sample of n units possess the attribute under study, expressions for the estimator p_{sy} of the population proportion, variance $V(p_{sy})$, and the estimator of this variance can be obtained from (6.1) to (6.5). These are given as follows :

Estimator of proportion P :

$$p_{sy} = \frac{n_1}{n} \tag{6.12}$$

Variance of estimator p_{sy} :

$$V(p_{sy}) = \frac{1}{k} \sum_{r=1}^{k} (p_{sy} - P)_r^2 \qquad \text{(for LS sampling)} \tag{6.13}$$

$$= \frac{1}{N} \sum_{r=1}^{N} (p_{sy} - P)_r^2 \qquad \text{(for CS sampling)} \tag{6.14}$$

Estimator of variance $V(p_{sy})$:

$$v(p_{sy}) = \frac{(N-n)\,R}{2Nn\,(n-1)} \tag{6.15}$$

where R is the total number of times that 0 follows 1 or 1 follows 0 in the ordered sequence of observations for the n sample units.

Estimator of $V(p_{sy})$ for population in random order :

$$v(p_{sy}) = \frac{(N-n)}{N} \left[\frac{p_{sy}(1-p_{sy})}{(n-1)} \right] \tag{6.16}$$

In case one wishes to estimate the total number N_1 of units in the population that possess the desired attribute, the estimator $\hat{N}_1$ is obtained by multiplying p_{sy} by N. Also, the variance $V(\hat{N}_1)$ and its estimator $v(\hat{N}_1)$ are N^2 times the corresponding expressions for p_{sy}.

We now take up an example to illustrate the various steps involved in estimating proportion and the number of units in the population possessing the attribute under study.

Example 6.7

On public complaint that some gas cylinders supplied for domestic use were underweight, an inquiry committee was set up. The committee decided to examine 1-in-50 cylinders from the 8000 cylinders stored in a warehouse, arranged in rows by the gas company. The committee found 18 cylinders to be underweight from the 160 sampled cylinders. Estimate the total number N_1, and also the proportion, of underweight cylinders in the warehouse. Also, build up confidence interval for these parameters.

Solution

We have N= 8000, n=160, and n_1=18. Then the estimate of proportion of underweight cylinders in the warehouse is

$$p_{sy} = \frac{n_1}{n}$$
$$= \frac{18}{160}$$
$$= .1125$$

Estimated total number of underweight cylinders would then be

$$\hat{N}_1 = Np_{sy}$$
$$= (8000)(.1125)$$
$$= 900$$

Assuming that the population units (gas cylinders) under study were placed in random order before drawing the sample, the estimate of variance $V(\hat{N}_1)$ is computed using (6.16) as

$$v(\hat{N}_1) = N^2 v(p_{sy})$$
$$= \frac{N(N-n)\, p_{sy}(1-p_{sy})}{(n-1)}$$
$$= \frac{8000(8000-160)(.1125)(1-.1125)}{159}$$
$$= 39384.90$$

Also, one can work out the confidence interval for N_1 from

$$\hat{N}_1 \pm 2\sqrt{v(\hat{N}_1)}$$
$$= 900 \pm 2\sqrt{39384.90}$$
$$= 503.09,\ 1296.91$$
$$\approx 503,\ 1297$$

Thus, the inquiry committee estimated 900 underweight cylinders from this particular sample. The committee also feels that the total number of underweight cylinders in the warehouse is likely to be between 503 and 1297.

From the above calculations, we find that $p_{sy} = .1125$ is the estimate of the proportion of underweight cylinders in the warehouse. Also,

$$v(p_{sy}) = \frac{v(\hat{N}_1)}{N^2}$$

$$= \frac{39384.90}{(8000)^2}$$

$$= .0006154$$

The required confidence interval for the proportion of underweight cylinders is given by

$$p_{sy} \pm 2\sqrt{v(p_{sy})}$$

$$= .1125 \pm 2\sqrt{.0006154}$$

$$= .0629, .1621$$

Thus, the proportion of underweight cylinders in the warehouse is most probably in the range .0629 to .1621. ■

It can be noted here that the lower and upper limits for the confidence interval of P can alternatively be obtained by dividing corresponding limits for the confidence interval for N_1 by N.

6.7 SOME FURTHER REMARKS

6.1 In case of $N \neq nk$, the sample mean $\bar{y}_{sy}$ based on a LS sample does not remain unbiased for the population mean $\bar{Y}$. This problem can also be resolved if the procedure in selecting the random start is slightly modified. The modification consists of selecting random start r, $1 \leq r \leq k$, with probability $P(r)=n_r/N$, where n_r is the number of units to be selected in the systematic sample corresponding to the random start r. For example, if we have N=14, n=5, so that k = 3, then random starts r = 1, 2 will yield 5 units in each sample, while r = 3 will result in the selection of 4 units. Thus, if we associate with r = 1 and r = 2 a probability of 5/14 while r = 3 is selected with probability 4/14, the sample mean $\bar{y}_{sy}$ will become unbiased for the population mean.

6.2 In case of populations with *linear trend*, the relative efficiency of systematic sample mean is very high in relation to the simple random sample mean for estimating the mean of such populations. Consider a hypothetical population in which the model $Y_i=a+b.i$, where a and b are constants and i is the serial number of the unit in the frame with study variable value Y_i, exactly holds. For such a population, the relative efficiency is approximately equal to the sample size.

6.3 For populations with linear trend, Yates (1948) has made a suggestion known as *Yates end correction*, which helps in greatly reducing the error in the estimator $\bar{y}_{sy}$. We find that in estimator $\bar{y}_{sy}$ in (6.1), all the observations $(y_1, y_2,...,y_n)$ received a weight equal to 1/n. Yates has proposed certain corrections to the weights

associated with the end units (first and last) in the sample. In place of 1/n, the first unit receives a weight $\left(\frac{1}{n}+x\right)$ whereas the weight associated with the last unit equals $\left(\frac{1}{n}-x\right)$, where $x=\frac{2r-k-1}{2(n-1)k}$. Thus, the new estimator of population mean becomes

$$\bar{y}'_{sy} = \bar{y}_{sy} + \frac{2r-k-1}{2(n-1)k}(y_1 - y_n)$$

6.4 For the population with linear trend and N=nk, a sampling procedure known as *balanced systematic sampling*, is also helpful in reducing the error of the estimator $\bar{y}_{sy}$. The procedure assumes that the serial numbers of the units in the population are divided into n/2 groups of 2k (sampling interval) units each. A pair of units, equidistant from the end units of that group, is then systematically selected from each group. For example, if r is the random start selected from 1 to k, the units with serial numbers r and 2k-r+1 will be selected from the first group of 2k units. Then, the two units selected from the second group of 2k units, will be the units with serial numbers 2k+r and 4k-r+1, and so on. Thus, the balanced systematic sample of n (even) units with random start r will consist of units with serial numbers [r+2jk, 2(j+1)k-r+1], j=0,1,2,...,(n/2 - 1).

LET US DO

6.1 What is systematic sampling? Discuss its merits and demerits.

6.2 Differentiate between systematic sampling and simple random sampling. Do you think that in presence of periodicity in the population, systematic sampling can be used more efficiently ? If so, how could it be done ?

6.3 The circular systematic sampling is usually preferred over linear systematic sampling. Discuss, why is it so ?

6.4 The number of colleges in 12 districts of a state are 8, 10, 6, 7, 7, 9, 11, 5, 6, 8, 9, and 11. List all possible samples of size 3 that can be selected from this population of 12 units using LS and CS sampling. Also, determine the average of corresponding sample means in both the cases. Are the two averages equal to the population mean? If yes, what does it indicate about the bias in the two estimators ?

6.5 Many trees along a canal have been uprooted by a storm. This damage persists along a 35 km stretch. The Department of Irrigation is interested in estimating total number of these damaged trees. Each one kilometer segment along the canal has been divided into 5 equal parts by stone markers. Thus, the entire 35 km long stretch is divided into 175 equal segments. Twenty five of these segments are selected using LS sampling with a sampling interval of 7 segments. The information regarding number of uprooted trees (y) obtained from this 1-in-7 systematic sample is given in the following table :

Selected segment	y	Selected segment	y	Selected segment	y
6	4	62	3	118	23
13	17	69	8	125	12
20	11	76	5	132	8
27	6	83	13	139	17
34	8	90	9	146	6
41	16	97	16	153	5
48	21	104	17	160	8
55	13	111	9	167	10
				174	15

Estimate the total number of uprooted trees, and also determine the confidence interval for it.

6.6 It is desired to estimate the average per day rent for single occupancy rooms in well known hotels of a state. In all, there are 192 such hotels in the state and these are listed in a book entitled "A Guide to Visitors". The investigator selected a 1-in-8 sample of hotels and rang up the managers of sampled hotels. The information on rent (in rupees) so obtained is given below :

Hotel	Rent	Hotel	Rent	Hotel	Rent
1	100	9	90	17	125
2	120	10	110	18	85
3	125	11	125	19	90
4	115	12	80	20	105
5	110	13	70	21	130
6	80	14	125	22	95
7	130	15	130	23	135
8	120	16	105	24	140

Estimate the average per day rent along with the confidence limits for it.

6.7 The editor of a local daily newspaper is interested in estimating average number of misprints in the daily over a year of 365 days. All the editions of the daily corresponding to 365 days were labeled from 1 to 365. A systematic sample of 28 editions of the paper was selected using CS sampling, taking $k=365/28 \approx 13$. The selected editions were then carefully examined and misprints counted. These are given in the following table along with the serial numbers of the selected editions of the paper.

Edition No.	Number of misprints	Edition No.	Number of misprints	Edition No.	Number of misprints
73	3	190	9	307	11
86	11	203	4	320	9
99	4	216	8	333	6
112	2	229	6	346	10
125	8	242	5	359	8
138	0	255	7	7	1
151	6	268	16	20	5
164	13	281	10	33	0
177	5	294	8	46	3
				59	7

Estimate average number of misprints in a daily edition, and place confidence limits on it.

6.8 There are 280 wells in a region for recording water table depth. Owing to heavy rainfall in the area under study, the water table has gone up considerably and is likely to cause waterlogging in the area. For estimating the average water table depth, the irrigation department selected a 1-in-10 sample of wells, and made the following observations on the water table depth (y) in meters.

Well	y	Well	y	Well	y	Well	y
1	2.60	8	2.40	15	2.60	22	2.10
2	2.80	9	2.60	16	1.40	23	2.00
3	2.85	10	2.35	17	1.55	24	1.80
4	3.75	11	1.85	18	1.80	25	2.35
5	2.85	12	3.30	19	2.50	26	1.60
6	2.45	13	2.50	20	2.30	27	1.70
7	1.90	14	2.10	21	1.70	28	2.10

Estimate mean water table depth in the region, and place confidence limits on it.

6.9 A reputed public school has 800 children. The school management has changed the school uniform twice in a year. A sociologist wishes to gauge parents' reaction to this decision of the management. The sociologist has 5 skilled investigators. He selected 5 subsamples of 8 students each, so that, the overall sample was of size 40. Each investigator interviewed the parents of the students falling in the subsample assigned to him, and gave scores from 1 to 10 depending on the severity of respondent's reaction to the management's decision. The scores thus recorded are presented in the following table :

Subsample	Scores							
1	3	5	1	6	8	4	9	10
2	1	7	5	3	6	4	7	8
3	6	2	8	3	6	5	2	4
4	9	4	6	5	4	3	1	2
5	5	3	9	6	1	4	8	1

Assuming that the random starts for the subsamples were selected with replacement, estimate the average score for the parents of all the 800 students, and also build up the confidence interval for the true average score.

6.10 A graduate student in statistics was given an assignment to estimate average height of all the 900 graduate students at a certain university. The student selected an overall sample of 60 graduate students in the form of six subsamples, each consisting of 10 students, using systematic sampling. The average height (in cm) of students in each subsample was computed, and is given below :

$\bar{y}_1 = 160.8$ $\bar{y}_2 = 168.6$ $\bar{y}_3 = 169.4$
$\bar{y}_4 = 164.5$ $\bar{y}_5 = 163.4$ $\bar{y}_6 = 166.8$

Estimate the average height of all the 900 students, and construct confidence interval for it. Assume that the random starts for the subsamples were selected through SRS without replacement.

6.11 Assume that the sample of 24 hotels selected in exercise 6.6 is a preliminary sample. Examine whether this sample is sufficient to estimate the average per day rent with a permissible error of Rs 5 ? If not, how many additional units need to be selected.

6.12 Some of the school buildings in a district collapsed during last few years, and caused damage to life and property. The district administration decided to have a quick estimate of the proportion of unsafe school buildings in the district. For this purpose, a systematic sample of 84 buildings, out of a total of 1260 school buildings, was selected . The selected school buildings were examined by experts. The number of unsafe buildings was found to be 16. Estimate the proportion of unsafe buildings in the district, and work out the confidence interval for it.

6.13 District traffic police is concerned about the vehicle owners not carrying necessary documents with them. To estimate the seriousness of the problem, a check point was set up on the Grand Trunk Road. Due to heavy rush of traffic, every 20th vehicle was stopped and its papers examined. In all, 114 vehicles were checked, of which 19 were not having necessary documents. Estimate the proportion of vehicle owners who do not carry the required documents. Also, construct confidence interval for this proportion.

6.14 Consider a population of units exhibiting a linear trend. Would you prefer using systematic sampling, or the usual SRS, for estimating mean/total ? Give reasons in support of your decision.