CHAPTER 3

Simple Random Sampling

3.1 WHAT IS SIMPLE RANDOM SAMPLING ?

In this book, we shall consider various sampling procedures (schemes) for selection of units in the sample. Since the objective of a survey is to make inferences about the population, a procedure that provides a precise estimator of the parameter of interest is desirable. Many sampling schemes have been developed to achieve this objective. To begin with, simple random sampling, the simplest and the most basic sample selection procedure, is discussed.

Definition 3.1 The sampling procedure is known as *simple random sampling* if every population unit has the same chance of being selected in the sample. The sample thus obtained is termed a *simple random sample.*

For selecting a simple random sample in practice, units from population are drawn one by one. If the unit selected at any particular draw is replaced back in the population before the next unit is drawn, the procedure is called *with replacement* (WR) *sampling*. A set of units selected at n such draws, constitutes a simple random with replacement sample of size n. In such a selection procedure, there is a possibility of one or more population units getting selected more than once. In case, this procedure is continued till n distinct units are selected, and all repetitions are ignored, it is called *simple random sampling* (SRS) *without replacement* (WOR). This method is equivalent to the procedure, where the selected units at each draw are not replaced back in the population before executing the next draw.

Another definition of simple random sampling, both with and without replacement, could be given on the basis of probabilities associated with all possible samples that can be selected from the population.

Definition 3.2 *Simple random sampling* is the method of selecting the units from the population where all possible samples are equally likely to get selected.

3.2 HOW TO DRAW A SIMPLE RANDOM SAMPLE ?

To draw a simple random sample from a population under study is not as trivial as it appears. If the investigator selects the sample by judgement, claiming the sample to be representative of population, it is subjected to investigator bias. Such a sampling leads

to estimators whose properties can not be evaluated. Therefore, one has to use a sampling mechanism which assigns to every population unit an equal probability of being selected in the sample. The most commonly used procedures for selecting a simple random sample are: (1) lottery method, and (2) through the use of random number tables.

3.2.1 *Lottery Method*

In this method, each unit of the population of N units is assigned a distinct identification mark (number) from 1 to N. This constitutes the population frame. Each of these numbers is then written on a different slip of paper. All the N slips of paper are identical in respect of size, color, shape, etc. Fold all these slips in an identical manner and put them in a container or drum, in which a thorough mixing of the slips is carried out before each blindfold draw. The paper slips are then drawn one by one. The units corresponding to the identification labels on the selected slips, are taken to be members of the sample. If the sampling is WR, each slip drawn is put back in the container after noting the identification label on that slip and refolding it. In WOR sampling, the paper slip once drawn is not replaced back in the container. Draws of paper slips, this way, are continued till a sample of required size is obtained.

One can also use a deck of cards, spherical balls or some other such items in place of slips of paper. This procedure of numbering units on slips and selecting slips after reshuffling, becomes tedious when the population size is large. To overcome this difficulty, tables of random numbers are used.

3.2.2 *Through the Use of Random Number Tables*

A *random number table* is an arrangement of ten digits from 0 to 9, occurring with equal frequencies (except for chance fluctuations) independently of each other and without any consistently recurring trends or patterns. Several standard tables of random numbers prepared by Tippett (1927), Fisher and Yates (1938), Kendall and Smith (1939), Rand Corporation (1955), and Rao *et al.* (1974) are available. However, in this book we shall make use of the random number tables due to Rao *et al.* (1974). Some of these random number tables are reproduced in appendix B to help in illustrating the use of random numbers for selecting a sample.

We discuss below three commonly used methods of *using random number tables* for selection of simple random samples.

Direct Approach. Again, the first step in the method is to assign serial numbers 1 to N to the N population units. If the population size N is made up of K digits, then consider K digit random numbers, either row wise or column wise, in the random number table. The sample of required size is then selected by drawing, one by one, random numbers from 1 to N, and including the units bearing these serial numbers in the sample.

This procedure may involve number of rejections of random numbers, since zero and all the numbers greater than N appearing in the table are not considered for selection. The use of random numbers has, therefore, to be modified. Two of the commonly used modified procedures are now discussed.

Remainder Approach. If N is a K digit number, determine the highest K digit multiple of N. Let it be N'. Then a random number r is selected, such that $1 \leq r \leq N'$. The unit bearing the serial number equal to the remainder (say) R, obtained on dividing r by N, is then considered as selected. If remainder is zero, the last unit is selected. As an illustration, let N=24. Here N is a two digit number. The highest two digit multiple of 24 is 96. Let the random number r, selected from 1 to 96, be 83. On dividing 83 by 24, we get remainder as 11. Therefore, the unit bearing serial number 11 is selected in the sample. The process is repeated till the sample of required size is selected. As before, the repeated selections of population units in the sample are permitted for with replacement sample, whereas they are rejected and only distinct units selected for a WOR sample.

Quotient Approach. As before, let N be a K digit number and N' be the highest K digit multiple of N, such that N'= Nm. Select a random number r from 0 to N'-1. Then the unit having serial number (Q+1) is included in the sample, where Q is the quotient when r is divided by m. For instance, if N=24 then N'=96, so that m=4. Let a random number (say) 49 be chosen from 0 to 95. Then Q=12. The unit bearing serial number (Q+1)=13 is then selected in the sample.

It may be noted here, that while using the random number tables, any starting point can be used, and one can move in any predetermined direction along the rows or columns. If more than one sample is to be selected in any problem, each should have its independent starting point.

Besides the above discussed methods, some more methods for sample selection are available in literature. However, being operationally inconvenient, they are usually not employed in practice.

Example 3.1

Appendix C gives data related to the number of tractors in 69 serially numbered villages of Doraha development block in Punjab (India). Select (1) WR and (2) WOR simple random sample of 10 villages using direct approach method.

Solution

Here village is the sampling unit. The villages in the population are already serially numbered which, otherwise, is the first step involved in the sample selection. Refer to appendix B, and use first column by dropping the last two digits of each four digit number. Then we see that the first random number thus formed is 34. Similarly, the subsequent random numbers are seen to be 61,58,...,35.

(1) By selecting the first 10 random numbers from 1 to 69, without discarding repetitions, we obtain the serial numbers of villages in the sample. These are given below along with their variable values (number of tractors).

Village :	34	61	58	62	47	34	11	43	5	35
Tractors :	14	8	15	39	9	14	11	19	12	18

One can see that 34th village has been selected twice in the with replacement simple random sample where repeated selection of units is permitted.

(2) In without replacement sample, any repetition (34th village in the present case) is omitted, and another random number is selected as its replacement. Next random number from 1 to 69 is 26, and it has not appeared earlier. Thus the WOR simple random sample of 10 villages from the population under study is with the following serial numbers :

Village	:	34	61	58	62	47	11	43	5	35	26
Tractors	:	14	8	15	39	9	11	19	12	18	22 ■

3.3 ESTIMATION OF POPULATION MEAN/TOTAL

Once the sample has been selected and the units in the sample observed for the study variable, the next step is to draw inferences about the population from the information contained in the sample. In sample surveys, usually we are interested in estimating certain specific population parameters. The parameters of common interest are mean, total, or the proportion. First we consider equal probability WR sampling.

3.3.1 *WR Simple Random Sampling*

Let y denote the study variable and $Y_1, Y_2,...,Y_N$ be the values of y for N units of the population. Further, let $y_1, y_2,...,y_n$ denote the values of y for n units selected in the sample. Then the sample mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

also defined in (2.5), is an unbiased estimator of population mean $\bar{Y}$ (given in 2.3), in case of both with and without replacement simple random sampling. In case of SRS with replacement, the *sampling variance* of mean is given by

$$V(\bar{y}) = \frac{\sigma^2}{n} \tag{3.1}$$

where the population variance σ^2 has already been defined in (2.4).

Variance of the sampling distribution of mean, given above in (3.1), depends on the population parameter σ^2. This value of σ^2 will not be known unless we know all $Y_1, Y_2,...,Y_N$. Since the values of y for all the population units are not known, the actual value of $V(\bar{y})$ can not be obtained. We have, therefore, to satisfy ourselves with only the estimated value of $V(\bar{y})$which we can get from the sample data. An unbiased estimator of $V(\bar{y})$ is then given by

$$v(\bar{y}) = \frac{s^2}{n} \tag{3.2}$$

with the sample mean square s^2 defined in (2.6).

We now summarize the above discussed formulas for the case of WR simple random sampling.

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

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad (3.3.)$$

Sampling Variance of $\bar{y}$:

$$V(\bar{y}) = \frac{\sigma^2}{n} \qquad (3.4)$$

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

$$v(\bar{y}) = \frac{s^2}{n} \qquad (3.5)$$

where σ^2 and s^2 are defined in (2.4) and (2.6) respectively.

In order to clarify the concepts of expected value, variance, and estimator of variance, in relation to simple random sampling WR procedure, we consider an example.

Example 3.2

The height (in cm) of 6 students of M.Sc., majoring in statistics, from Punjab Agricultural University, Ludhiana was recorded during 1985. The data, so obtained, are given below :

Table 3.1 Heights of M.Sc. students

Student	Name	Height
1	Sarjinder Singh	168
2	Gurmeet Singh	175
3	Varinder Kumar	185
4	Sukhjinder Singh	173
5	Devinder Kumar	171
6	Gulshan Kumar	172

1. Calculate (a) population mean $\bar{Y}$, and (b) population variance σ^2.
2. Enumerate all possible SRS with replacement samples of size n=2. Obtain sampling distribution of mean, and hence show that

a. $E(\bar{y}) = \bar{Y}$

b. $V(\bar{y}) = \frac{\sigma^2}{n}$

c. $E(s^2) = \sigma^2$

d. $E[v(\bar{y})] = V(\bar{y})$

Solution

(1) Here N=6, and the study variable height is denoted by y. For making computations easily understandable, we present them in tabular form.

Table 3.2 Population data and other computations

Student	Y_i	Y_i^2
1	168	28224
2	175	30625
3	185	34225
4	173	29929
5	171	29241
6	172	29584
Total	1044	181828

It can be easily seen from the above table that

(a) Population mean

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

$$= \frac{1044}{6}$$

$$= 174$$

(b) Population variance

$$\sigma^2 = \frac{1}{N} \left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2 \right)$$

$$= \frac{1}{6} [181828 - 6 (174)^2]$$

$$= 28.667$$

(2) Number of all possible samples of size 2, in case of WR sampling, will be $N^n=6^2=36$. The students in various samples along with their corresponding height values, sample means, and sample mean squares are given in table 3.3.

Table 3.3 All possible WR samples and other related statistics

Sample	Sample units	Height of students	$\bar{y}$	s^2
1	(1, 1)	(168, 168)	168.0	0
2	(1, 2)	(168, 175)	171.5	24.5
3	(1, 3)	(168, 185)	176.5	144.5
4	(1, 4)	(168, 173)	170.5	12.5
5	(1, 5)	(168, 171)	169.5	4.5
6	(1, 6)	(168, 172)	170.0	8.0
7	(2, 1)	(175, 168)	171.5	24.5
8	(2, 2)	(175, 175)	175.0	0
9	(2, 3)	(175, 185)	180.0	50.0
10	(2, 4)	(175, 173)	174.0	2.0
11	(2, 5)	(175, 171)	173.0	8.0
12	(2, 6)	(175, 172)	173.5	4.5
13	(3, 1)	(185, 168)	176.5	144.5
14	(3, 2)	(185, 175)	180.0	50.0
15	(3, 3)	(185, 185)	185.0	0
16	(3, 4)	(185, 173)	179.0	72.0
17	(3, 5)	(185, 171)	178.0	98.0
18	(3, 6)	(185, 172)	178.5	84.5
19	(4, 1)	(173, 168)	170.5	12.5
20	(4, 2)	(173, 175)	174.0	2.0
21	(4, 3)	(173, 185)	179.0	72.0
22	(4, 4)	(173, 173)	173.0	0
23	(4, 5)	(173, 171)	172.0	2.0
24	(4, 6)	(173, 172)	172.5	.5
25	(5, 1)	(171, 168)	169.5	4.5
26	(5, 2)	(171, 175)	173.0	8.0
27	(5, 3)	(171, 185)	178.0	98.0
28	(5, 4)	(171, 173)	172.0	2.0
29	(5, 5)	(171, 171)	171.0	0
30	(5, 6)	(171, 172)	171.5	.5
31	(6, 1)	(172, 168)	170.0	8.0
32	(6, 2)	(172, 175)	173.5	4.5
33	(6, 3)	(172, 185)	178.5	84.5
34	(6, 4)	(172, 173)	172.5	.5
35	(6, 5)	(172, 171)	171.5	.5
36	(6, 6)	(172, 172)	172.0	0
Total			6264	1032

Column (4) in table 3.3, lists all possible values of sample mean $\bar{y}$. In case of SRS with replacement, each one of the 36 samples will have equal chance of getting selected and will, therefore, have a probability of $1/N^n=1/36$ associated with it. In case of other WR sampling procedures (other than simple random sampling), these probabilities may not be equal. More appropriately, these can be written in the form of a *sampling distribution* as given in table 3.4. The probability associated with any sample mean value is the number of samples that yield that particular sample mean value, times 1/36.

Table 3.4 Sampling distribution of mean

Serial No.	Sample mean ($\bar{y}$)	Frequency (f)	Probability (p)
1	168.0	1	1/36
2	169.5	2	2/36
3	170.0	2	2/36
4	170.5	2	2/36
5	171.0	1	1/36
6	171.5	4	4/36
7	172.0	3	3/36
8	172.5	2	2/36
9	173.0	3	3/36
10	173.5	2	2/36
11	174.0	2	2/36
12	175.0	1	1/36
13	176.5	2	2/36
14	178.0	2	2/36
15	178.5	2	2/36
16	179.0	2	2/36
17	180.0	2	2/36
18	185.0	1	1/36
Total		36	1

For verifying the other results in the statement, we use table 3.3.

(a) Average of all possible sample means, denoted by $E(\bar{y})$, is obtained from column (4) of table 3.3 as

$$E(\bar{y}) = \frac{1}{36}(168.0+171.5+...+172.0)$$

$$= \frac{6264}{36}$$

$$= 174$$

It is, therefore, seen that

$$E(\bar{y}) = \bar{Y}$$

Hence $\bar{y}$ is an unbiased estimator of $\bar{Y}$.

(b) By definition 2.7, the variance of all possible sample means in this case, is given by

$$V(\bar{y}) = \frac{1}{36} \sum_{i=1}^{36} \bar{y}_i^2 - [E(\bar{y})]^2$$

On using column (4) of table 3.3, one gets

$$V(\bar{y}) = \frac{1}{36} [(168.0)^2 + (171.5)^2 + ... + (172.0)^2] - (174)^2$$

$$= \frac{1090452}{36} - (174)^2$$

$$= 14.333$$

Now from part (b) of (1), we have

$$\frac{\sigma^2}{n} = \frac{28.667}{2} = 14.333$$

Hence the relation (b) of (2), that is, $V(\bar{y}) = \sigma^2/n$ stands verified.

The reader must note that the $E(\bar{y})$ and $V(\bar{y})$ can also be obtained from the sampling distribution of mean $\bar{y}$, given in table 3.4, by using the expressions

$$E(\bar{y}) = \Sigma\, \bar{y}_i p_i$$

and

$$V(\bar{y}) = \Sigma\, \bar{y}_i^2 p_i - (\Sigma\, \bar{y}_i p_i)^2$$

Here, p_i is the probability of the sample mean $\bar{y}$ taking value $\bar{y}_i$, and the summation is over all the values (18 in this case) taken by $\bar{y}$.
Thus,

$$p_i = \frac{f_i}{N^n}$$

f_i being the number of samples that yield $\bar{y}=\bar{y}_i$.

(c) From table 3.3, we find that the average of all possible sample mean squares is

$$= \frac{1}{36} (0 + 24.5 + ... + 0)$$

$$= \frac{1032}{36}$$

$$= 28.667$$

which equals population variance. It means that

$$E(s^2) = \sigma^2$$

This verifies relation (c) of (2).

(d) Just now, we have numerically illustrated that

$$E(s^2) = \sigma^2$$

On dividing both sides by n, we get

$$E\left(\frac{s^2}{n}\right) = \frac{\sigma^2}{n}$$

That means, if we work out column s^2/n in table 3.3 and take average of values in that column over all possible samples, it will be same as σ^2/n. Thus $v(\bar{y})$ is seen to be an unbiased estimator of $V(\bar{y})$. These calculations are left as an exercise for the reader. ■

3.3.2 *WOR Simple Random Sampling*

In case of simple random sampling WOR, the sample mean $\bar{y}$ still remains unbiased estimator of the population mean $\bar{Y}$, and has the variance

$$\left.\begin{aligned} V(\bar{y}) &= \frac{N-n}{Nn} S^2 \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \end{aligned}\right] \qquad (3.6)$$

where the *population mean square* S^2 is defined as

$$\left.\begin{aligned} S^2 &= \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \\ &= \frac{1}{N-1} \left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right) \end{aligned}\right] \qquad (3.7)$$

It can be easily seen that the population mean square S^2 and the population variance σ^2 are related through the equation

$$(N-1)\, S^2 = N\sigma^2$$

The variance $V(\bar{y})$ in (3.6) can, therefore, also be written as

$$\left.\begin{aligned} V(\bar{y}) &= \left(\frac{N-n}{N-1}\right) \frac{\sigma^2}{n} \\ &= \left(1 - \frac{n-1}{N-1}\right) \frac{\sigma^2}{n} \end{aligned}\right] \qquad (3.8)$$

All the four forms of $V(\bar{y})$, two in (3.6) and two in (3.8) are equivalent, and one can use any one of them. We shall also not stick to any one particular form of variance $V(\bar{y})$ in the book, and may use any one of the four forms.

As discussed before the actual value of variance in (3.6) or (3.8) can not be found unless we know y values for all the population units. An unbiased estimator of this variance is, therefore, needed and is given by

$$v(\bar{y}) = \frac{N-n}{Nn} s^2$$

where s^2, to be obtained from sample information, is defined earlier in (2.6). Thus for WOR simple random sampling, we have :

Unbiased estimator of $\bar{Y}$ is same as in (3.3).

Variance of estimator $\bar{y}$:

$$V(\bar{y}) = \frac{N-n}{Nn} S^2 \qquad (3.9)$$

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

$$v(\bar{y}) = \frac{N-n}{Nn} s^2 \qquad (3.10)$$

where S^2 and s^2 are defined in (3.7) and (2.6) respectively.

It can be easily seen that the variance $V(\bar{y})$, in (3.8) or equivalently in (3.9), for WOR case reduces to the variance $V(\bar{y})$ in (3.4) for with replacement sampling when $N=\infty$. Thus, for an infinite population, both with and without replacement sampling procedures are equivalent. Because of this, we call the factor (N-n)/(N-1) as *finite population correction* (fpc).

Remark 3.1 $E(s^2)=\sigma^2$ for SRS with replacement, whereas in case of WOR equal probability sampling we have $E(s^2)=S^2$. For $N=\infty$, fpc takes value 1 whereas the sampling fraction reduces to zero.

Example 3.3

From the data given in example 3.2, enumerate all the SRS without replacement samples of size n=2, and write down sampling distribution of mean. Using this distribution, show that

a. $E(\bar{y}) = \bar{Y}$

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

c. $E(s^2) = S^2$

d. $E[v(\bar{y})] = V(\bar{y})$

Solution

Number of possible WOR samples of size n=2 will be $\binom{6}{2}$ =15. The heights of students included in various possible samples, along with sample means ($\bar{y}$) and sample mean squares (s^2), are given in table 3.5.

Table 3.5 All possible WOR samples and other related statistics

Serial No.	Sample units	Height of students	$\bar{y}$	s^2
1	(1 , 2)	(168 , 175)	171.5	24.5
2	(1 , 3)	(168 , 185)	176.5	144.5
3	(1 , 4)	(168 , 173)	170.5	12.5
4	(1 , 5)	(168 , 171)	169.5	4.5
5	(1 , 6)	(168 , 172)	170.0	8.0
6	(2 , 3)	(175 , 185)	180.0	50.0
7	(2 , 4)	(175 , 173)	174.0	2.0
8	(2 , 5)	(175 , 171)	173.0	8.0
9	(2 , 6)	(175, 172)	173.5	4.5
10	(3 , 4)	(185 , 173)	179.0	72.0
11	(3 , 5)	(185 , 171)	178.0	98.0
12	(3 , 6)	(185 , 172)	178.5	84.5
13	(4, 5)	(173 , 171)	172.0	2.0
14	(4 , 6)	(173 , 172)	172.5	.5
15	(5 , 6)	(171, 172)	171.5	.5
Total			2610	516

Column (4) in table 3.5 lists mean values for all possible samples. It can also be written in the form of a *sampling distribution* as shown in table 3.6. Probability (p) values, in this case also, are calculated in the same way as in table 3.4.

Table 3.6 Sampling distribution of mean

Serial No.	Sample mean ($\bar{y}$)	Frequency (f)	Probability (p)
1	169.5	1	1/15
2	170.0	1	1/15
3	170.5	1	1/15
4	171.5	2	2/15
5	172.0	1	1/15
6	172.5	1	1/15
7	173.0	1	1/15
8	173.5	1	1/15

Table 3.6 continued...

Serial No.	Sample mean ($\bar{y}$)	Frequency (f)	Probability (p)
9	174.0	1	1/15
10	176.5	1	1/15
11	178.0	1	1/15
12	178.5	1	1/15
13	179.0	1	1/15
14	180.0	1	1/15
Total		15	1

Using values computed in table 3.5, we proceed to verify the required results.

(a) The average of all possible 15 sample means is given as

$$E(\bar{y}) = \frac{1}{15}(171.5+176.5+...+171.5)$$

$$= \frac{2610}{15}$$

$$= 174$$

which is same as the population mean worked out in example 3.2. Hence $E(\bar{y}) = \bar{Y}$.

(b) By definition 2.7, the variance of $\bar{y}$ is given by

$$V(\bar{y}) = \frac{1}{15}\sum_{i=1}^{15} \bar{y}_i^2 - [E(\bar{y})]^2$$

$$= \frac{1}{15}[(171.5)^2 + (176.5)^2 + ... + (171.5)^2] - (174)^2$$

$$= \frac{454312}{15} - (174)^2$$

$$= 11.467$$

Also, the population mean square

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

$$= \frac{1}{5}[181828 - 6(174)^2]$$

$$= 34.400$$

Then we have

$$\frac{N-n}{Nn}S^2 = \left[\frac{6-2}{(6)(2)}\right](34.400)$$

$$= 11.467$$

This verifies the relation

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

(c) The average of all the 15 sample mean squares from table 3.5 is

$$E(s^2) = \frac{1}{15}(24.5+144.5+...+.5)$$

$$= \frac{516}{15}$$

$$= 34.400$$

which equals population mean square S^2 obtained above in (b). Thus, $E(s^2)=S^2$

(d) In (c) we have seen that

$$E(s^2) = S^2$$

On multiplying both sides by (N-n)/Nn, one gets

$$E[v(\bar{y})] = V(\bar{y})$$

which verifies the statement that $v(\bar{y})$ is an unbiased estimator of $V(\bar{y})$. Its numerical verification is left as an exercise for the reader. ■

Example 3.4

From the WOR sample of 10 villages selected in example 3.1, estimate the average number of tractors per village in the block along with its standard error. Also, set up confidence interval for the population mean.

Solution

For convenience, we display the data for 10 selected villages, and the other required computations in table 3.7.

Table 3.7 Sample data and other computations

Village	y	y^2
1	14	196
2	8	64
3	15	225
4	39	1521
5	9	81
6	11	121
7	19	361
8	12	144
9	18	324
10	22	484
Total	167	3521

Our estimate of $\overline{Y}$ average number of tractors per village in the block, is

$$\bar{y} = \frac{1}{10} \sum_{i=1}^{10} y_i$$

$$= \frac{167}{10}$$

$$= 16.7$$

$$\approx 17$$

To find the standard error, we first compute sample mean square as

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)$$

$$= \frac{1}{9} [3521 - 10 (16.7)^2]$$

$$= 81.344$$

The estimate of variance is then given by

$$v(\bar{y}) = \frac{N-n}{Nn} s^2$$

$$= \left[\frac{69-10}{(69)(10)}\right] (81.344)$$

$$= 6.956$$

Estimate of standard error of the sample mean $\bar{y}$ is, therefore,

$$se(\bar{y}) = \sqrt{v(\bar{y})}$$

$$= \sqrt{6.956}$$

$$= 2.637$$

The lower and upper limits of the confidence interval for $\overline{Y}$ are given by

$$\bar{y} \pm 2se(\bar{y})$$
$$= 16.7 \pm 2 (2.637)$$
$$= 11.426, 21.974$$
$$\approx 11, 22$$

Thus, the closed interval [11, 22] covers the average number of tractors per village in the block with probability approximately equal to .95. ■

The estimate of population total Y can be obtained by multiplying the estimate of mean by population size N. Its variance and estimator of variance are N^2 times the corresponding expression for mean $\bar{y}$.

Unbiased estimator of population total Y :

$$\hat{Y} = N\bar{y} \tag{3.11}$$

Variance of estimator $\hat{Y}$:

$$V(\hat{Y}) = N^2V(\bar{y}) \tag{3.12}$$

Unbiased estimator of variance $V(\hat{Y})$

$$v(\hat{Y}) = N^2v(\bar{y}) \tag{3.13}$$

where $V(\bar{y})$ and $v(\bar{y})$ will be used respectively from (3.4) and (3.5) in case of WR sampling, and from (3.9) and (3.10) if the sampling is WOR.

Confidence interval for total Y can also be obtained accordingly following (2.8).

Example 3.5

From the data of WOR sample comprising of 10 villages in example 3.1, estimate the total number of tractors in the development block of 69 villages. Also, set up the confidence interval for it.

Solution

We have N=69 and n=10. For the sake of convenience, the data has been presented in table 3.7 along with some other computations. From (3.11), the estimate of total number of tractors in the block is

$$\begin{aligned}\hat{Y} &= N\bar{y} \\ &= (69)\,(16.7) \qquad \text{(from example 3.4)} \\ &= 1152.3 \\ &\approx 1152\end{aligned}$$

The estimate of variance of $\hat{Y}$ is provided by (3.13). Thus,

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

Substituting the value of $v(\bar{y})$ from example 3.4, one gets

$$\begin{aligned}v(\hat{Y}) &= (69)^2\,(6.956) \\ &= 33117.5\end{aligned}$$

We now work out confidence interval for the population total. Following (2.8), the required confidence limits will be

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

$$= 1152.3 \pm 2\sqrt{33117.5}$$

$$= 788.3, 1516.3$$

$$\approx 788, 1516$$

From the limits of the confidence interval obtained above, the investigator is quite confident that the total number of tractors in the block are likely to be in the range of 788 to 1516. ■

The concept of *confidence interval* used in examples 3.4 and 3.5 is further elaborated in the next example.

Example 3.6
Refer to data in appendix C considered for example 3.1. Examine the behavior of approximately 95% level confidence intervals for total number of tractors by selecting 50 WOR simple random samples each of size n=20. Total number of tractors in the block are 1465.

Solution
Fifty different WOR simple random samples of size 20 drawn from the population of 69 villages (appendix C) are presented in appendix D. The estimate $\hat{Y}$ of total and estimated mean square s^2, along with *lower* and *upper confidence limits* (LCL and UCL) computed from these 50 samples using the relation $\hat{Y} \pm 2\sqrt{v(\hat{Y})}$, are exhibited in table 3.8.

Table 3.8 $\hat{Y}$, s^2, LCL, and UCL for n=20 and N=69

Sample	$\hat{Y}$	s^2	LCL	UCL	CI covers Y?
1	1559.4	306.04	1104.5	2014.3	Yes
2	1666.4	381.71	1158.3	2174.4	Yes
3	1607.7	566.64	988.7	2226.7	Yes
4	1373.1	359.67	879.9	1866.3	Yes
5	1197.2	69.29	980.7	1413.6	No
6	1390.4	199.08	1023.4	1757.3	Yes
7	1321.4	129.19	1025.8	1616.9	Yes
8	1183.4	314.34	722.3	1644.4	Yes
9	1566.3	346.33	1082.4	2050.2	Yes
10	1480.1	258.58	1061.9	1898.2	Yes
11	1711.2	565.54	1092.8	2329.6	Yes
12	1666.4	380.77	1158.9	2173.8	Yes
13	1649.1	459.15	1091.9	2206.3	Yes
14	1300.7	187.71	944.4	1656.9	Yes

Table 3.8 continued...

Sample	$\hat{Y}$	s^2	LCL	UCL	CI covers Y?
15	1956.2	536.03	1354.1	2558.2	Yes
16	1680.2	499.92	1098.7	2261.6	Yes
17	1859.6	397.10	1341.4	2377.7	Yes
18	1452.5	329.21	980.6	1924.3	Yes
19	1235.1	307.57	779.1	1691.1	Yes
20	1400.7	328.01	929.7	1871.7	Yes
21	1438.7	248.34	1028.9	1848.4	Yes
22	1276.5	335.21	800.4	1752.6	Yes
23	1518.0	347.16	1033.5	2002.5	Yes
24	1549.1	366.89	1051.0	2047.1	Yes
25	1835.4	362.36	1340.4	2330.4	Yes
26	1280.0	171.63	939.3	1620.6	Yes
27	1814.7	489.48	1239.4	2390.0	Yes
28	1307.6	211.00	929.8	1685.3	Yes
29	1952.7	582.85	1324.9	2580.5	Yes
30	1745.7	406.54	1221.4	2270.0	Yes
31	1335.2	151.71	1014.9	1655.4	Yes
32	1483.5	271.74	1054.8	1912.2	Yes
33	1752.6	571.62	1130.9	2374.3	Yes
34	1518.0	549.89	908.2	2127.8	Yes
35	1500.8	307.57	1044.7	1956.8	Yes
36	1445.6	305.94	990.7	1900.4	Yes
37	1569.8	275.04	1138.5	2001.0	Yes
38	1600.8	346.17	1117.0	2084.6	Yes
39	1169.6	304.89	715.5	1623.6	Yes
40	1649.1	462.62	1089.8	2208.4	Yes
41	1566.3	275.69	1134.5	1998.1	Yes
42	1583.6	234.79	1185.1	1982.0	Yes
43	1656.0	410.63	1129.1	2182.9	Yes
44	1369.7	144.87	1056.7	1682.6	Yes
45	1414.5	368.37	915.4	1913.6	Yes
46	1193.7	132.01	894.9	1492.5	Yes
47	1386.9	76.83	1159.0	1614.8	Yes
48	897.0	39.26	734.1	1060.0	No
49	1411.1	387.31	899.3	1922.8	Yes
50	1145.4	165.94	810.4	1480.4	Yes

This example shows that in about 4% cases the population total Y=1465 falls outside the confidence interval, otherwise, in 96% cases it is covered by the said interval. If many more samples are examined it will tend to reach 95% cases covering the population total. ■

3.4 ESTIMATION OF MEAN/TOTAL USING DISTINCT UNITS

In case of SRS with replacement, the units which get repeated while selecting the sample do not provide any additional information. Therefore, the information obtained from distinct units is sufficient to estimate population mean/total. Let $y_1, y_2, \ldots, y_d$ be the values of the study variable y for the d distinct units in a WR simple random sample of n units. Then, an unbiased estimator of population mean, based on distinct units, is due to Des Raj and Khamis (1958). This estimator, along with its variance and estimator of variance, is given in (3.14), (3.15), and (3.16).

Estimator of population mean $\bar{Y}$ based on distinct units :

$$\bar{y}_d = \frac{1}{d} \sum_{i=1}^{d} y_i \tag{3.14}$$

Variance of estimator $\bar{y}_d$:

$$V(\bar{y}_d) = \left[E\left(\frac{1}{d}\right) - \frac{1}{N} \right] S^2 \tag{3.15}$$

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

$$v(\bar{y}_d) = \left[\frac{1}{d} - \frac{1}{N} \right] s_d^2 \tag{3.16}$$

where for $d \geq 2$, $s_d^2 = \frac{1}{d-1} \sum_{i=1}^{d} (y_i - \bar{y}_d)^2$ and S^2 is same as in (3.7).

The estimator $\bar{y}_d$ is always more efficient than the usual WR estimator $\bar{y}$ given in (3.3).

The population total, in this case also, will be estimated by $\hat{Y}_d = N\bar{y}_d$. The expressions for variance $V(\hat{Y}_d)$ and its estimator are, as usual, given by $N^2V(\bar{y}_d)$ and $N^2v(\bar{y}_d)$. For detail, the reader may refer to Murthy (1967).

Example 3.7

Refer to data in part (1) of example 3.1 , where 10 units have been selected in the sample using WR simple random sampling. From observations related to distinct units in this sample, estimate the population total, and also obtain confidence limits for it.

Solution

One can see that in the WR sample of 10 villages drawn in part (1) of example 3.1, the village bearing serial number 34 has been selected twice. On discarding the repetition, the data for 9 distinct villages in the sample is displayed below :

Village	:	34	61	58	62	47	11	43	5	35
Tractors	:	14	8	15	39	9	11	19	12	18

Now we have d = 9 and N = 69. From (3.14), the estimate of the average number of tractors per village in the block is seen to be

$$\bar{y}_d = \frac{1}{d} \sum_{i=1}^{d} y_i$$

$$= \frac{1}{9} (14+8+...+18)$$

$$= 16.11$$

Using the value of $\bar{y}_d$ obtained above, one gets the estimate of the total number of tractors in the block as

$$\hat{Y}_d = N\bar{y}_d$$

$$= 69\ (16.11)$$

$$= 1111.6$$

$$\approx 1112$$

In order to obtain estimated variance of $\hat{Y}_d$, we first compute

$$s_d^2 = \frac{1}{d-1} \sum_{i=1}^{d} (y_i - \bar{y}_d)^2$$

$$= \frac{1}{d-1} \left(\sum_{i=1}^{d} y_i^2 - d\bar{y}_d^2\right)$$

$$= \frac{1}{9-1} [(14)^2 + (8)^2 + ... + (18)^2 - 9(16.11)^2]$$

$$= \frac{1}{8} [3037 - 9(16.11)^2]$$

$$= 87.65$$

From (3.16), the estimated variance of $\hat{Y}_d$ will be

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

$$= N^2 \left(\frac{1}{d} - \frac{1}{N}\right) s_d^2$$

Making substitutions for different terms, yields

$$v(\hat{Y}_d) = (69)^2 \left(\frac{1}{9} - \frac{1}{69}\right) (87.65)$$

$$= 40319$$

Following (2.8), the confidence limits are obtained as

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

$$= 1111.6 \pm 2\sqrt{40319}$$

$$= 710.0, 1513.2$$

$$\approx 710, 1513$$

The above values indicate that the total number of tractors in 69 villages of Doraha block would probably fall in the closed interval [710, 1513] with probability approximately equal to .95. ■

3.5 DETERMINING SAMPLE SIZE FOR ESTIMATING POPULATION MEAN/TOTAL

So far, in this chapter, we have discussed methods of selecting simple random samples and the point and interval estimation of population mean (or total). The next important topic that merits consideration is the determination of number of units to be included in the sample. If the sample is too large then time, effort, money, and talent are wasted. Conversely, if the number of units included in the sample is too small, we have collected inadequate information which diminishes the utility of the results. This problem can, however, be solved by using the framework of sampling theory.

Though, the required sample size can be determined by using prior information on variance or coefficient of variation of the population (Cochran, 1977; Sukhatme *et al.*, 1984), but usually, it is difficult to obtain reliable information on these parameters. We, therefore, discuss below a two step approach which does not require prior information on the value of any population parameter. Here, a small preliminary sample is used to estimate the population parameter values, which in turn are used to determine final sample size. The preliminary sample is then augmented by drawing additional units from the population, so that, the size of the augmented sample is same as the required final sample size.

Let n_1 be the size of preliminary sample selected using SRS without replacement and

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2$$

where

$$\bar{y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i$$

is the mean of the preliminary sample. Using s_1^2 in place of s^2 in (3.10), and then equating half width of the confidence interval in (2.8) to the permissible error B, one gets the required sample size as

$$n = \frac{Ns_1^2}{ND + s_1^2}$$

where

$$D = \frac{B^2}{4} \tag{3.17}$$

The above equation in case of SRS with replacement reduces to

$$n = \frac{s_1^2}{D}$$

The rule for selecting sample size can then be stated as follows :

Sample size required for estimating population mean with permissible error B :

$$n = \frac{Ns_1^2}{ND + s_1^2} \quad \text{(for SRS without replacement)} \tag{3.18}$$

$$n = \frac{s_1^2}{D} \quad \text{(for SRS with replacement)} \tag{3.19}$$

where D is defined in (3.17). If $n_1 \geq n$, then n_1 is the sufficient sample size and no additional units need to be selected, otherwise, $(n-n_1)$ additional units are to be selected to get the required sample.

Note that, sometimes it is more convenient to express the permissible margin of error as a fraction of true value. In this case $B = \in \overline{Y}$, and in the formulas (3.18) and (3.19), B will have to be replaced by its estimated value $\in \overline{y}_1$.

Example 3.8

An investigator is interested in estimating average number of tractors per village in Doraha development block of Punjab state. The block consists of 69 villages. Using WOR sample of 10 villages selected in example 3.1, as preliminary sample, determine the sample size needed to estimate the said population mean with a margin of error not exceeding 3.

Solution

Here N=69, n_1=10, and B=3. From example 3.4, we have $s_1^2 = 81.344$. Thus, on using (3.17), it is observed that

$$D = \frac{B^2}{4} = 2.25$$

Then from (3.18),

$$n = \frac{Ns_1^2}{ND + s_1^2}$$

$$= \frac{69\,(81.344)}{69\,(2.25) + 81.344}$$

$$= 23.72$$

$$\approx 24$$

It means that the size of preliminary sample is not sufficient for estimating the population mean with desired precision. Therefore, $n-n_1=14$ more villages are required to be selected. ■

In a like manner, we can determine the sample size needed to estimate a population total with desired precision. On equating half width of confidence interval for total, to permissible error and then solving the equation for n, we get the rule for determining sample size analogous to one obtained in (3.18) and (3.19). However, the value of D is different. For reader's convenience, these results are reproduced in (3.20) and (3.21).

Sample size required to estimate population total with permissible error B :

$$n = \frac{Ns_1^2}{ND + s_1^2} \qquad \text{(for SRS without replacement)} \qquad (3.20)$$

$$n = \frac{s_1^2}{D} \qquad \text{(for SRS with replacement)} \qquad (3.21)$$

where $D=B^2/4N^2$. If $n_1 \geq n$, then sample size n_1 is sufficient, otherwise, select $(n-n_1)$ more units to augment the preliminary sample.

Example 3.9

The owner of a poultry farm is interested in estimating the total weight gain, in a period of one month, for N=1500 chicks kept on a new feed. For this purpose, a simple random WOR sample of $n_1=25$ chicks is observed for weight gain. The sample data yielded $s_1^2=45$ gm^2. Determine the sample size required to estimate total weight gain with two kg (=2000 gm) as margin of error.

Solution

In this case, we have N=1500, $n_1=25$, B=2000, $s_1^2=45$, and

$$D = \frac{B^2}{4N^2} = \frac{(2000)^2}{4(1500)^2} = .4444$$

From (3.20), we find that

$$n = \frac{1500\,(45)}{1500\,(.4444) + 45}$$
$$= 94.86$$
$$\approx 95$$

Thus, $n-n_1=95-25=70$ additional chicks are to be selected to get a total sample of required size. ■

3.6 ESTIMATION OF POPULATION PROPORTION

The investigator conducting a sample survey is, sometimes, interested in estimating the *proportion* of the population that possesses a specified attribute. For example, the government might be interested in estimating the proportion of factories using environmental pollution control measures. The interest of the state could also be in estimating the proportion of population in favor of a particular bill, which is to be introduced in parliament for making it a law. A political party might be interested in estimating the proportion of voters likely to vote for it, in the coming elections.

All these examples exhibit a characteristic of dichotomized or binomial population, where an observation either belongs, or does not belong, to the category of interest. The value 1 or 0 is assigned to each unit according as the unit belongs or does not belong to the desired category. The population total for such a variable becomes equal to the number of units in the population possessing the specified attribute. Also, the mean reduces to the proportion P of such units in the population. Similarly, all the other results presented earlier for continuous data, yield corresponding results for this particular case.

Let N_1 units out of N possess the attribute of interest. Then population proportion $P=N_1/N$ and $Q=1-P$. If n_1 units out of n sample units possess this attribute, sample proportion is given by $p=n_1/n$ and $q=1-p$. It can be easily seen that for a random variable y taking values 1 and 0, the sample mean $\bar{y}$ reduces to sample proportion p. Hence, $E(p)=P$ is true for both with and WOR simple random sampling. The important results for the case of estimation of proportion are listed in (3.22) through (3.26).

Unbiased estimator of population proportion P for WR case :

$$p = \frac{n_1}{n} \tag{3.22}$$

Variance of estimator p :

$$V(p) = \frac{PQ}{n} \tag{3.23}$$

Unbiased estimator of variance V(p) :

$$v(p) = \frac{pq}{n-1} \tag{3.24}$$

where $q = 1 - p$, $P = N_1/N$, and $Q = 1 - P$.

Unbiased estimator of P for WOR case is same as in (3.22).

Variance of estimator p :

$$V(p) = \frac{N-n}{N-1}\left(\frac{PQ}{n}\right) \tag{3.25}$$

Unbiased estimator of variance V(p) :

$$v(p) = \left(1-\frac{n}{N}\right)\left(\frac{pq}{n-1}\right) \tag{3.26}$$

For large samples, the sample proportion p can be considered to be approximately normally distributed with mean P and variance V(p). Since the variance V(p) is usually not available in practice, the estimate of variance would be used for setting the confidence interval for P.

Example 3.10

Punjab Agricultural University, Ludhiana, is interested in estimating the proportion P of teachers who consider semester system to be more suitable as compared to the trimester system of education. A with replacement simple random sample of n=120 teachers is taken from a total of N=1200 teachers. The response is denoted by 0 if the teacher does not think the semester system suitable, and 1 if he/she does. From the sample observations given below, estimate the proportion P along with the standard error of your estimate. Also, work out the confidence interval for P.

Teacher :	1	2	3	4	5	6	. . .	119	120	Total
Response :	1	0	1	1	0	1	. . .	0	1	72

Solution

Estimate of P is given by (3.22). Thus,

$$p = \frac{72}{120} = .6$$

Estimate of the standard error of p will then be obtained as

$$\begin{aligned} se(p) &= \sqrt{v(p)} \\ &= \sqrt{\frac{pq}{n-1}} \\ &= \sqrt{\frac{(.6)\,(.4)}{119}} \\ &= .04491 \end{aligned}$$

The confidence limits for P would be obtained following (2.8). Thus,

$$p \pm 2\sqrt{v(p)}$$
$$= .6 \pm 2(.04491)$$
$$= .5102, .6898$$

The proportion of teachers in the university favoring semester system is, therefore, likely to be in the closed interval [.5102, .6898]. ■

In certain situations, the objective could be to estimate the number N_1 of units possessing the attribute of interest. The unbiased estimator $\hat{N}_1$ of N_1 would be N times the sample proportion p defined in (3.22). The variance and its estimator respectively for $\hat{N}_1$, would be N^2 times the variance and the estimator of variance for p.

3.7 SAMPLE SIZE FOR ESTIMATION OF PROPORTION

As in case of quantitative data, let n_1 be the number of units selected in the preliminary sample, and p_1 denote the proportion of units in this sample possessing the attribute under consideration. Also, let $q_1=1-p_1$. Using p_1 and q_1 in place of p and q in (3.26), and then equating half width of confidence interval to the permissible error B, one gets for WOR simple random sampling

$$n = \frac{n_0}{1+(n_0-1)/N}$$

where

$$n_0 = \frac{4p_1q_1}{B^2} + 1 \qquad (3.27)$$

For SRS with replacement case, the above expression for n becomes

$$n = n_0$$

Thus, the formulas for determining the required sample size can be stated as in (3.28) and (3.29).

Sample size required for estimating P with tolerable error B :

In addition to n_1 units included in the preliminary sample, select $n-n_1$ more units, where

$$n = \frac{n_0}{1+(n_0-1)/N} \quad \text{(for SRS without replacement)} \qquad (3.28)$$

$$n = n_0 \quad \text{(for SRS with replacement)} \qquad (3.29)$$

with n_0 defined in (3.27). If $n_1 \geq n$, then n_1 is the required sample size.

Example 3.11
In example 3.10, while estimating P, the investigator feels that the tolerable error could be taken as .08. Do you think the sample size 120 is sufficient ? If not, how many more units should be included in the sample ?

Solution
From example 3.10, we observe that $n_1=120$, $p_1=.6$, and $q_1=.4$. Now from (3.27) and (3.29),

$$\begin{aligned} n &= n_0 \\ &= \frac{4p_1q_1}{B^2} + 1 \\ &= \frac{4(.6)(.4)}{(.08)^2} + 1 \\ &= 151 \end{aligned}$$

The already selected sample of size 120 is thus not sufficient for achieving the given precision. Therefore, $n-n_1=31$ more teachers need to be selected. ■

3.8 ESTIMATION OF PROPORTION USING INVERSE SAMPLING

In practice, a situation may arise, where attribute under consideration prevails with rare frequency. In such cases, the proportion P to be estimated is very small, and estimation procedure described in section 3.6 may not be satisfactory. Even a large sample may not be enough to estimate P with a reasonable degree of precision . The appropriate sample selection procedure for such type of attributes, is known as inverse sampling. The procedure is due to Haldane (1946).

Definition 3.3 The procedure where sampling is continued until a predetermined number of units possessing the attribute are included in the sample, is known as *inverse sampling.*

Let n be the number of units required to be selected to obtain a predetermined number m of units possessing the rare attribute. Though a biased maximum likelihood estimator is also available, we consider an unbiased estimator of P. The variance expressions, for this unbiased estimator, given by Haldane (1946) and Best (1974) are complicated. However, estimator of variance due to Finney (1949) takes a simple form. If the selection of the units is with SRS without replacement, then the unbiased estimator of P in (3.30) follows *negative hypergeometric* distribution. In case of simple random WR selection of units, it is distributed as *negative binomial* (also known as *inverse binomial*).

Unbiased estimator of population proportion P :

$$p = \frac{m-1}{n-1} \tag{3.30}$$

Estimator of variance V(p) :

$$v(p) = \frac{p(1-p)}{n-2}\left(1-\frac{n-1}{N}\right) \quad \text{(when sampling is WOR)} \tag{3.31}$$

$$v(p) = \frac{p(1-p)}{n-2} \quad \text{(when sampling is WR)} \tag{3.32}$$

Example 3.12

A survey conducted by a student of a medical college in Ludhiana town showed that a proportion .008 of adults over 18 years of age, living in a posh colony, are suffering from tuberculosis. Another student of the same college was subsequently given an assignment to examine whether the incidence of tuberculosis infection in the adults of the same age group, living in a slum area, is on the higher side of .008 ? For conducting this survey, voters' lists were used as frame, and voters as the sampling units. It was decided in advance to continue WR simple random sampling of individuals till 10 cases of tuberculosis infection were detected. To arrive at this predetermined number of 10, the investigator had to select 380 adults from the slum area. Besides estimating the proportion in question, work out the confidence limits within which this parameter is expected to lie.

Solution

Here m=10 and n=380. Estimate of proportion of adults suffering from tuberculosis in slum area is, therefore, given by (3.30). Thus,

$$p = \frac{m-1}{n-1}$$

$$= \frac{10-1}{380-1}$$

$$= .02375$$

Proportion of tuberculosis infection among adults of the slum area is approximately three times the proportion of infected adults in the posh colony. We now compute the estimate of variance and the confidence interval using the relations (3.32) and (2.8) respectively. Let us first work out the estimate of variance. From (3.32),

$$v(p) = \frac{p(1-p)}{n-2}$$

$$= \frac{.02375\,(1-.02375)}{378}$$

$$= .000061$$

Confidence limits for the proportion of tuberculosis infected adult population in slum area is obtained as

$$
\begin{aligned}
& p \pm 2\sqrt{v(p)} \\
&= .02375 \pm 2\sqrt{.000061} \\
&= .00809, .03941 \blacksquare
\end{aligned}
$$

3.9 ESTIMATION OVER SUBPOPULATIONS

It may often be impossible to obtain a frame that lists only those units in the population which are of interest. For instance, the investigator is interested in sampling households, where both husband and wife work, or he/she wants to sample households having adults over 50 years of age. However, the best frame available in both the cases is the list of all households in the target area. In this case, before any sample unit is observed, the investigator has no way of knowing whether any particular selected unit is a member of the *subpopulation* under consideration, or not. The procedure for estimating mean or total, therefore, needs to be modified. This modification consists in taking the values of the study variable, for units not belonging to the class of interest, as zero. This indirectly amounts to using only those sample units that belong to the subpopulation of interest.

For further discussion, we shall use the following notations :

- N = the total number of units in the population
- N_1 = the number of units in the subpopulation of interest
- n = the number of units in the WOR simple random sample drawn from the population of size N
- n_1 = the number of units in the sample of size n that belong to the subpopulation under consideration
- Y_{si} = the value of study variable y for the i-th unit of the subpopulation
- y_{si} = the value of y for the i-th sample unit from the subpopulation

The mean, total, and mean square error for the target subpopulation are then given by

$$
\left.
\begin{aligned}
\bar{Y}_s &= \frac{1}{N_1} \sum_{i=1}^{N_1} Y_{si} \\
Y_s &= \sum_{i=1}^{N_1} Y_{si} \\
S_s^2 &= \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} (Y_{si} - \bar{Y}_s)^2
\end{aligned}
\right] \qquad (3.33)
$$

The unbiased estimator of subpopulation mean, and other related results are given in the following box :

Unbiased estimator of the subpopulation mean $\bar{Y}_s$ when N_1 is known :

$$\bar{y}_s = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{si} \tag{3.34}$$

Variance of estimator $\bar{y}_s$:

$$V(\bar{y}_s) = \left[E\left(\frac{1}{n_1}\right) - \frac{1}{N_1}\right] S_s^2 \tag{3.35}$$

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

$$v(\bar{y}_s) = \left(\frac{1}{n_1} - \frac{1}{N_1}\right)\left(\frac{1}{n_1 - 1}\right) \left(\sum_{i=1}^{n_1} y_{si}^2 - n_1 \bar{y}_s^2\right) \tag{3.36}$$

In case N_1 is not known, it may be substituted by $n_1 N/n$.

The estimator $\hat{Y}_s$ for the total of the subpopulation is obtained by multiplying the estimator $\bar{y}_s$ by N_1, and the expressions for variance $V(\hat{Y}_s)$ and estimator of variance $v(\hat{Y}_s)$ are arrived at by respectively multiplying $V(\bar{y}_s)$ and $v(\bar{y}_s)$ by N_1^2.

Example 3.13

The family planning wing of the health department of a certain state wishes to conduct a survey at a university campus for estimating the average time gap between the births of children in families having two children. The frame available, of course, lists all the 800 families of the campus. As the prior identification of the families in the population having just two children was difficult, the investigator selected a WOR random sample of 80 families. In the sample families, 32 families were found having two children. These 32 families were interviewed, and the information collected is shown in table 3.9.

Table 3.9 Time gap (in months) between the births of two children

Family	Gap	Family	Gap	Family	Gap	Family	Gap
1	24	9	64	17	57	25	42
2	30	10	32	18	65	26	16
3	50	11	58	19	26	27	37
4	41	12	48	20	35	28	61
5	27	13	51	21	31	29	34
6	47	14	22	22	17	30	29
7	47	15	69	23	28	31	19
8	39	16	54	24	55	32	57

Estimate the average gap between the births of two children, and obtain confidence limits for it.

Solution

We have $n_1=32$, $n=80$, and $N=800$. Let us first work out the estimate of mean given by (3.34). We have

$$\begin{aligned}\bar{y}_s &= \frac{1}{n_1}\sum_{i=1}^{n_1} y_{si} \\ &= \frac{1}{32}(24+30+\ldots+57) \\ &= 41\end{aligned}$$

months as the estimate of average time gap between the births of children. The estimate of variance is provided by (3.36). The expression for it is

$$v(\bar{y}_s) = \left(\frac{1}{n_1} - \frac{1}{N_1}\right)\left(\frac{1}{n_1-1}\right)\left(\sum_{i=1}^{n_1} y_{si}^2 - n_1\bar{y}_s^2\right)$$

Since N_1 is unknown, it would be replaced by

$$\begin{aligned}\hat{N}_1 &= \frac{n_1 N}{n} \\ &= \frac{(32)(800)}{80} \\ &= 320\end{aligned}$$

On substituting $\hat{N}_1$ for N_1, one gets the estimate of variance as

$$\begin{aligned}v(\bar{y}_s) &= \left(\frac{1}{32}-\frac{1}{320}\right)\left(\frac{1}{32-1}\right)[24^2 + 30^2+\ldots+57^2-32(41)^2] \\ &= \frac{(320-32)(233.35)}{(320)(32)} \\ &= 6.56\end{aligned}$$

The required confidence limits are obtained, following (2.8), by using

$$\begin{aligned}&\bar{y}_s \pm 2\sqrt{v(\bar{y}_s)} \\ &= 41 \pm 2\sqrt{6.56} \\ &= 35.88,\ 46.12\end{aligned}$$

The confidence limits obtained above indicate with reasonable confidence that the average gap between the births of two children in the population of families having two children, is likely to be in the range of 35.88 to 46.12 months. ■

The objective in certain situations could be to estimate the proportion of units in a subpopulation possessing a specific attribute. For instance, one may wish to estimate the proportion of men over 70 years of age who are still actively contributing towards family income by way of doing some kind of work (business, farming, job, etc.). However, the frame of such individuals (above 70 years and alive) is not readily available. Instead, a voters' list prepared five years back is available. In this case, the frame

consisting of men expected to cross their 70th year could be prepared from the available voters' list. Since the voters' list was prepared five years back, it could be possible that some of the individuals included in the frame are no more. For a situation like this, the required estimator for the subpopulation proportion and other related expressions can be easily obtained from (3.34), (3.35), and (3.36) by assigning value 1 to the working men who have attained the age of 70, and 0 to the others.

Let us define

$$P_s = \frac{N_1'}{N_1} \text{ and } p_s = \frac{n_1'}{n_1}$$

where n_1' and N_1' are the number of units possessing the attribute of interest out of n_1 and N_1 units respectively.

Unbiased estimator of proportion in the subpopulation when N_1 is known :

$$p_s = \frac{n_1'}{n_1} \tag{3.37}$$

Variance of estimator p_s :

$$V(p_s) = \left[N_1 E\left(\frac{1}{n_1}\right) - 1\right] \frac{P_s(1-P_s)}{N_1 - 1} \tag{3.38}$$

Estimator of variance $V(p_s)$:

$$v(p_s) = \left(1 - \frac{n_1}{N_1}\right) \frac{p_s(1-p_s)}{n_1 - 1} \tag{3.39}$$

Substitute N_1 as $n_1 N/n$ in the above expression, when it is not known.

Example 3.14

Let us consider the example used for discussion above and assume that a sociologist is interested in estimating the proportion of men over 70 years, who are still contributing towards family income by way of doing some work. As mentioned before, the frame is prepared from a five years old voters' list which consists of 1500 men expected to cross their 70th year. Obviously, the frame also includes the names of those voters who expired before, or after reaching the age of 70 years, during the preceding five years period. A sample of n=120 persons was drawn from this frame following WOR simple random sampling. Out of these, 14 individuals were found to have died. On interviewing the remaining 106 persons, it was observed that 21 persons were still actively engaged in earning by doing some kind of work. Estimate the proportion in question, and also obtain the confidence interval for this parameter.

Solution

We are given that N=1500, n=120, n_1=120-14=106, and n_1' = 21. The required estimate of proportion is given by (3.37). Thus,

$$p_s = \frac{n_1'}{n_1} = \frac{21}{106} = .1981$$

We then work out the variance estimator. From (3.39), we have this as

$$v(p_s) = \left(1 - \frac{n_1}{N_1}\right) \frac{p_s (1 - p_s)}{n_1 - 1}$$

Since N_1 is not known, it is estimated by

$$\hat{N}_1 = \frac{n_1 N}{n}$$
$$= \frac{106\,(1500)}{120}$$
$$= 1325$$

On substituting the values of n_1, $\hat{N}_1$ and p_s, in the expression for $v(p_s)$ given above, one gets

$$v(p_s) = \left(1 - \frac{106}{1325}\right)\left[\frac{(.1981)\,(1 - .1981)}{106 - 1}\right]$$
$$= .0014$$

The confidence limits can be worked out following (2.8). These will, therefore, be

$$p_s \pm 2\sqrt{v(p_s)}$$
$$= .1981 \pm 2\sqrt{.0014}$$
$$= .1233,\ .2729$$

The investigator can, therefore, reasonably believe that the proportion of men over 70 years of age, who are actively engaged in supplementing their family income, is likely to be in the closed interval [.1233, .2729]. ■

3.10 SOME FURTHER REMARKS

3.1 A population contains N units, and the value of the study variable y is known for m units of this population. Let these be denoted as $y_1, y_2, \ldots, y_m$. A without replacement simple random sample of n units is selected from the remaining (N-m) units of the population. If $\bar{y}_n$ is the simple mean for the n units selected from the (N-m) units, the estimator

$$\hat{Y}_1 = \sum_{i=1}^{m} y_i + (N - m)\,\bar{y}_n$$

is unbiased for population total Y. Also, it has smaller variance than the estimator $\hat{Y} = N\,\bar{y}$ based on a WOR simple random sample of size n taken from the entire

population. Thus, the advance knowledge of y values for some of the population units can be used profitably.

3.2 In practice, a situation may arise where the estimates obtained from different samples have to be combined to get a pooled estimate of population mean. Let $\bar{y}_1, \bar{y}_2,...,\bar{y}_k$ be the respective means obtained from k independent simple random samples consisting of $n_1, n_2,..., n_k$ units. The minimum variance pooled estimator, based on all the k samples, would be

$$\bar{y}_p = \sum_{i=1}^{k} w_i \bar{y}_i$$

where, for i=1, 2,..., k,

$$w_i = \begin{cases} n_i / \left(\sum_{i=1}^{k} n_i\right) \\ \text{or} \\ \left(\frac{n_i}{N-n_i}\right) / \sum_{i=1}^{k} \left(\frac{n_i}{N-n_i}\right) \end{cases}$$

depending on whether the k samples are drawn using SRS with or without replacement. The variance and estimator of variance for $\bar{y}_p$ are given by

$$V(\bar{y}_p) = \sum_{i=1}^{k} w_i^2 V(\bar{y}_i)$$

$$v(\bar{y}_p) = \sum_{i=1}^{k} w_i^2 v(\bar{y}_i)$$

where the expressions for $V(\bar{y}_i)$ and $v(\bar{y}_i)$ depend on the sampling procedure used for the selection of k samples. The estimator for population total, or proportion, can be obtained in the usual manner from the above estimator of population mean.

LET US DO

3.1 What do you understand by equal probability sampling ? What are the different methods commonly used for selecting a simple random sample ? Keeping in view the population size, which of the methods would you prefer, and why ?

3.2 Is the WOR simple random sample mean unbiased for population mean ? Write the expressions for variance of the sample mean and the unbiased estimator of this variance.

3.3 Using random number tables, select a WR sample of 15 villages from the population of 69 villages in appendix C. From this sample, estimate the total number of tractors in Doraha development block.

3.4 A population consists of 5 M.Sc. students having teaching load of 10, 14, 13,

12, and 15 credit hours in a semester. List all possible WR simple random samples of size 3 that can be drawn from this population. Also, show numerically that

a. $E(\bar{y}) = \bar{Y}$

b. $V(\bar{y}) = \frac{\sigma^2}{n}$

c. $E(s^2) = \sigma^2$

3.5 From the population given in exercise 3.4, list all possible WOR samples of size 3, and show numerically that

a. $E(\bar{y}) = \bar{Y}$

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

c. $E(s^2) = S^2$

3.6 Refer to data in appendix C, considered for example 3.1. Examine the behavior of the approximately 95% level confidence intervals for total number of tractors by selecting 50 WR simple random samples each of size 25.

3.7 A sociologist wishes to estimate the average age at the time of death for women (age 18 years or more) in a city. The frame used was the list of death records for the year 1992, available in the office of Registrar of Births and Deaths. There were 680 deaths of women during 1992. From these, a WR random sample of 32 deceased women was drawn. The below given information regarding the age of sample women was obtained by contacting their kith and kin. Serial number in the table is the serial number of unit in the population.

Serial No.	Age	Serial No.	Age	Serial No.	Age	Serial No.	Age
102	49	52	66	171	63	259	70
52	66	110	47	54	69	326	44
36	56	211	59	619	46	627	54
447	47	512	60	380	33	280	32
351	71	43	51	210	57	89	50
161	58	14	67	7	67	130	68
7	67	215	61	123	28	431	74
85	74	16	72	54	69	320	60

Estimate the average age at the time of death for women in 1992, and place confidence limits on it.

3.8 During the last decade, some farmers have started raising sunflower crop. Being a new crop, it is grown only by a few farmers in each village. The area under this crop, being quite small, is not entered in the revenue records. The Department of Agriculture is interested in estimating total area under this crop in a district having

900 villages. Since it is difficult to collect information for each village, a WOR simple random sample of 32 villages was drawn. The information collected on area (in hectares) under sunflower cultivation for the sample villages is presented in the table below :

Village	Area	Village	Area	Village	Area	Village	Area
1	2.0	9	0	17	2.0	25	1.0
2	1.5	10	0	18	2.5	26	0
3	1.7	11	1.2	19	1.5	27	1.8
4	2.5	12	1.8	20	1.5	28	0
5	3.5	13	1.0	21	2.0	29	2.4
6	0	14	2.6	22	2.8	30	2.1
7	1.0	15	1.5	23	4.0	31	1.3
8	1.3	16	3.1	24	2.5	32	1.5

Estimate total area under sunflower crop in the district, and place confidence limits on it.

3.9 Refer to data of exercise 3.7, where the sample units have been selected using SRS with replacement method. Making use of observations on distinct units only, estimate the parameter of exercise 3.7, and also obtain confidence interval for it.

3.10 Assume that the sample of 32 women drawn from the population of 680 deceased women in exercise 3.7 is a preliminary sample. Examine, whether this sample size is sufficient to estimate the average age with a margin of error of 5 years? If not, how many more deceased women need to be selected in the sample ?

3.11 A car dealer is feeling concerned over the complaints received in the office of the manufacturer regarding the free service provided by him to the newly purchased cars. To assess the seriousness of the problem, the dealer decided to draw a WR random sample of 70 buyers out of the total of 1400 individuals who had purchased cars through him during the last one year. Twenty one buyers included in the sample graded service provided by him as unsatisfactory. Estimate the percentage of buyers feeling unsatisfied with the service provided, and construct a suitable level confidence interval for it.

3.12 An investigator wishes to estimate the proportion of students in a university whose fathers are graduates. To arrive at the estimate, a WOR simple random sample of 67 students was drawn from a total of 1400 students. On contacting the sampled students, it was found that the fathers of 46 students had not graduated. Estimate the proportion of students whose fathers were at least graduates. Also, set the confidence interval for population proportion.

3.13 Assume the WOR sample of 67 students in exercise 3.12 as the preliminary sample. If the permissible error could be taken as .1, determine how many additional students will have to be selected to estimate the proportion in question with specified precision ?

3.14 The Mayor of a municipal corporation noticed an error in the calculations of general provident fund (GPF) account of an employee. Fearing that such errors might have also crept in the calculations of some other GPF accounts, he directed the audit unit of the corporation to estimate the proportion of such accounts. Expecting that the percentage of such accounts could be quite small, the investigator followed inverse sampling approach. He decided to go on sampling the accounts, using WOR method, from the list of all GPF accounts till 5 wrongly calculated accounts were detected. To arrive at this predetermined number of 5, he had to select 60 accounts out of a total of 1200 GPF accounts. Estimate the proportion of wrongly calculated accounts, and also find the confidence interval for it.

3.15 Earlier, an investigator had estimated the average annual expenses on uniform for school going children, studying in Punjabi (mother tongue) medium government primary schools, in a certain locality. This annual estimated expenses figure was found to be Rs 410. The investigator now wishes to estimate such expenses for children of the same locality studying in English medium public schools. The frame available lists all the families in the locality. The investigator selected 81 families from the population of 700 families. Twenty seven of the sample families were sending their children to English medium public schools. The money spent annually (in rupees) on school uniforms of these children is given below :

Family	Expenses	Family	Expenses	Family	Expenses
1	800	10	1120	19	700
2	1200	11	860	20	960
3	950	12	900	21	850
4	760	13	650	22	600
5	1100	14	750	23	750
6	1050	15	1160	24	800
7	950	16	1000	25	930
8	800	17	900	26	1020
9	1300	18	650	27	730

Estimate the average annual expenses on public school uniforms, and also work out the confidence interval for this average.

3.16 The objective of the survey to be undertaken by the Animal Science department, is to estimate the proportion of families in a town who are selling milk. The frame of families rearing milch cattle is, however, not available. Instead, a list of all the 3350 families in the town is available. To arrive at the estimate, a WOR simple random sample of 360 families was drawn. When contacted, 114 families were found rearing milch cattle. Out of these, 65 families reported that they were selling milk, whereas the others consumed all the milk produced in the family. Estimate the proportion of families selling milk, and place confidence limits on it.