CHAPTER 11

Multistage Sampling

11.1 INTRODUCTION

In chapter 10, we have considered sampling procedures in which all the elements of the selected clusters are enumerated. It was seen that though cluster sampling is generally economical, but it is usually less efficient than sampling of same number of ultimate units directly from the population. This is because the former strategy restricts the spread of the sample over the population. It can, therefore, be logically expected that, for a given number of units in the sample, greater precision can be attained if (1) the units are distributed over a larger number of clusters, and (2) instead of completely enumerating all the units in each selected cluster, only a sample of units is observed. This logic gives rise to the following definition :

Definition 11.1 The procedure of sampling, which consists in first selecting the clusters and then randomly choosing a specified number of units from each selected cluster, is known as *two-stage sampling.*

The clusters that form the units of sampling at the first stage are called the *first stage units*, or *primary stage units*, and the elements within the clusters which form the units of sampling at the second stage are called *subunits*, or *secondary units,* or *second stage units.* The procedure can be generalized to three or more stages, and is then termed *multistage sampling.* As an example of four-stage sampling, we consider surveys for estimating yield of a crop in a particular state. Here, the development blocks may be considered as primary stage units, villages within blocks the second stage units, fields within villages the third stage units, and the small plots within fields, which are harvested to record yield, as the fourth stage units. It may be mentioned that multistage sampling may be the only feasible procedure in a number of practical situations, where a satisfactory sampling frame of ultimate observational units is not readily available and the cost of obtaining such a frame is considerable.

Multistage sampling is being currently used in a number of surveys. Among the early workers, Mahalanobis (1940) used this sampling procedure in estimating area under jute. The use of this procedure in agriculture and population surveys respectively has been considered by Cochran (1939) and Hansen and Hurwitz (1943). Lahiri (1954) has discussed the use of this procedure in the Indian National Sample Surveys.

Keeping the scope of the book in view, we shall only consider two-stage sampling. For the extension of the estimation procedure to more than two-stage sampling, the reader may refer to Sukhatme *et al.* (1984), Murthy (1967), and Cochran (1977).

11.2 NOTATIONS

For the two-stage estimation procedure, we shall use the following notations :

N = number of primary stage units (psu's) in the population

n = number of psu's selected in the sample

M_i = number of second stage units (ssu's) in the i-th psu

m_i = number of ssu's selected from M_i ssu's

$M_o = \sum_{i=1}^{N} M_i$ = total number of ssu's in the population

$\overline{M}$ = M_o/N = average number of ssu's per psu

Y_{ij} = value of the estimation variable y for j-th ssu of the i-th psu, $j = 1, 2, ..., M_i$; $i = 1, 2, ..., N$

y_{ij} = value of the study variable for j-th selected ssu of the i-th selected psu, $j = 1, 2, ..., m_i$; $i = 1, 2, ..., n$

$Y_{i.} = \sum_{j=1}^{M_i} Y_{ij}$ = total of y-values for the i-th psu

$Y_{..} = \sum_{i=1}^{N} Y_{i.}$ = population total of y-values

$\overline{Y}_i = \frac{Y_{i.}}{M_i}$ = population mean for i-th psu

$\overline{Y}_N = \frac{1}{N} \sum_{i=1}^{N} \overline{Y}_i$ = population mean of psu means

$\overline{Y} = \frac{1}{M_o} \sum_{i=1}^{N} \overline{Y}_i M_i$ = population mean per ssu for the study variable

$y_{i.} = \sum_{j=1}^{m_i} y_{ij}$ = sample total for the i-th psu

$y_{..} = \sum_{i=1}^{n} y_{i.}$ = total of y-values for the whole sample

$\overline{y}_i = \frac{y_{i.}}{m_i}$ = sample mean for i-th psu

11.3 ESTIMATION OF MEAN/ TOTAL IN TWO-STAGE SAMPLING USING SRSWOR AT BOTH THE STAGES

We consider two-stage sampling where the first stage units are of unequal size, and the units are selected using equal probabilities WOR method at both the stages. To select the sample, the investigator must have a frame listing all the N psu's in the population. A WOR simple random sample of n psu's is drawn using procedures described in chapter 3. Then, the frames that list all the M_i second stage units in i-th selected psu (i =1, 2 ,..., n) are obtained. Finally, a WOR simple random sample of m_i units is drawn from the i-th selected psu, containing M_i second stage units, for i = 1, 2,.., n. Several estimators of mean and total are available for this sampling procedure. However, we shall consider only three of these estimators as they are used quite frequently.

11.3.1 *Estimator 1*

Let us define

$$\left.\begin{aligned} S^2_{1b} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i\,\overline{Y}_i}{\overline{M}} - \overline{Y}\right)^2 \\ &= \frac{1}{\overline{M}^2\,(N-1)}\left(\sum_{i=1}^{N} Y^2_{i.} - \frac{Y..^2}{N}\right) \end{aligned}\right] \tag{11.1}$$

$$\left.\begin{aligned} S^2_i &= \frac{1}{M_i-1}\sum_{j=1}^{M_i}(Y_{ij}-\overline{Y}_i)^2 \\ &= \frac{1}{M_i-1}\left(\sum_{j=1}^{M_i} Y^2_{ij} - M_i\,\overline{Y}^2_i\right) \end{aligned}\right] \tag{11.2}$$

$$\left.\begin{aligned} s^2_{1b} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i\overline{y}_i}{\overline{M}} - \overline{y}_{ml}\right)^2 \\ &= \frac{1}{\overline{M}^2\,(n-1)}\left[\sum_{i=1}^{n}(M_i\,\overline{y}_i)^2 - \frac{1}{n}\left(\sum_{i=1}^{n} M_i\,\overline{y}_i\right)^2\right] \end{aligned}\right] \tag{11.3}$$

$$\left.\begin{aligned} s^2_i &= \frac{1}{m_i-1}\sum_{j=1}^{m_i}(y_{ij}-\overline{y}_i)^2 \\ &= \frac{1}{m_i-1}\left(\sum_{j=1}^{m_i} y^2_{ij} - m_i\overline{y}^2_i\right) \end{aligned}\right] \tag{11.4}$$

where $\overline{y}_{ml}$ is given in (11.5). Then we have expressions (11.5) to (11.7).

Unbiased estimator of population mean $\overline{\mathbf{Y}}$:

$$\overline{y}_{ml} = \frac{N}{nM_o}\sum_{i=1}^{n} M_i\,\overline{y}_i \tag{11.5}$$

Variance of the estimator $\overline{\mathbf{y}}_{\mathbf{ml}}$:

$$V(\overline{y}_{ml}) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2_{1b} + \frac{1}{nN}\sum_{i=1}^{N}\frac{M^2_i}{\overline{M}^2}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S^2_i \tag{11.6}$$

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

$$v(\bar{y}_{ml}) = (\frac{1}{n} - \frac{1}{N}) s_{lb}^2 + \frac{1}{nN} \sum_{i=1}^{n} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 \qquad (11.7)$$

where S_{lb}^2, S_i^2, s_{lb}^2, and s_i^2 are defined in (11.1) through (11.4).

Example 11.1

The co-operative societies in an Indian state, provide loans to farmers in terms of cash and fertilizer within the sanctioned limit, which depends on the share of the individual in the co-operative society. The society declares an individual defaulter, if he/she does not repay the loan within the specified time limit. An investigator is interested in estimating the average amount of loan, per society, standing against the defaulters. The total number of co-operative societies in the state is 10126. However, the list of all the societies is not available at the state headquarter but the same is available at development block level. Therefore, it seems appropriate to use two-stage sampling for selecting a sample of societies. Keeping in view the budget and time constraints, it was decided to select 12 blocks from the total of 117 blocks and approximately 10 percent of the societies from each of the sample blocks. The information obtained from the selected societies is given in table 11.1

Table 11.1 Dues (in '000 rupees) standing against the defaulters

Block	M_i	m_i	Amount due from defaulters						Total
1	60	6	12.5	36.4	26.0	55.6	58.1	40.8	229.4
2	102	10	57.4	16.8	20.3	70.1	34.6	22.6	346.3
			44.9	28.4	17.5	33.7			
3	48	5	12.9	41.6	34.7	30.8	61.1		181.1
4	113	11	28.7	82.4	37.3	41.9	24.7	36.6	494.9
			39.3	49.6	26.0	76.8	51.6		
5	92	9	44.8	42.9	51.7	28.8	36.4	40.1	431.5
			61.6	47.8	77.4				
6	57	6	31.6	24.8	69.9	44.9	59.7	38.6	269.5
7	82	8	49.6	36.9	27.3	63.6	73.0	44.9	443.6
			87.1	61.2					
8	96	10	53.7	34.9	41.5	43.4	56.6	28.9	423.0
			23.4	32.8	60.2	47.6			
9	53	5	41.7	54.9	33.9	27.9	46.3		204.7
10	71	7	24.4	38.9	47.8	45.0	32.6	66.5	313.5
			58.3						
11	77	8	42.9	37.3	30.8	51.9	60.1	34.6	324.3
			28.4	38.3					
12	56	6	44.7	34.9	61.7	74.6	37.4	49.2	302.5

Estimate the average amount, per society, standing against defaulters, and also compute the confidence interval for it.

Solution

The statement of the problem shows that N = 117, n = 12, and M_0 = 10126. Hence,

$$\overline{M} = \frac{10126}{117} = 86.55$$

In order to illustrate the various steps involved in the calculations, we prepare table 11.2. The mean $\overline{y}_i$ and the sample mean square s_i^2 for the societies included in the sample, for each selected block, are calculated. These are given along with other computations in table 11.2.

Table 11.2 Calculations for various terms involved in the estimation of population mean

Block	M_i	m_i	$\overline{y}_i = \frac{y_{i.}}{m_i}$	s_i^2	$M_i^2 (\frac{1}{m_i} - \frac{1}{M_i}) s_i^2$	$M_i\overline{y}_i$
1	60	6	38.23	303.62	163954.80	2293.80
2	102	10	34.63	320.35	300616.44	3532.26
3	48	5	36.22	305.87	126263.13	1738.56
4	113	11	44.99	368.54	386162.91	5083.87
5	92	9	47.94	208.11	176569.77	4410.48
6	57	6	44.92	292.93	141924.58	2560.44
7	82	8	55.45	384.47	291620.49	4546.90
8	96	10	42.30	151.84	125359.10	4060.80
9	53	5	40.94	110.95	56451.36	2169.82
10	71	7	44.79	210.33	136534.21	3180.09
11	77	8	40.54	115.75	76872.47	3121.58
12	56	6	50.42	231.30	107940.00	2823.52
Total	907				2090269.26	39522.12

Since M_0 is available, we use (11.5) for obtaining unbiased estimator of population mean. Thus, the estimate of average dues, per society, standing against defaulters is given by

$$\overline{y}_{ml} = \frac{N}{nM_0} \sum_{i=1}^{n} M_i\overline{y}_i$$

$$= \frac{117}{(12)(10126)} [(60)(38.23) + (102)(34.63) + ... + (56)(50.42)]$$

$$= \frac{(117)(39522.12)}{(12)(10126)}$$

$$= 38.05$$

The estimator of variance $V(\bar{y}_{ml})$ is given by (11.7). Hence,

$$v(\bar{y}_{ml}) = (\frac{1}{n} - \frac{1}{N}) s_{1b}^2 + \frac{1}{nN} \sum_{i=1}^{n} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2$$

The expression for s_{1b}^2 is given in (11.3). We calculate it by using the last column of table 11.2 as

$$s_{1b}^2 = \frac{1}{(86.55)^2\,(12-1)}[(2293.80)^2 + (3532.26)^2 + ... + (2823.52)^2 - \frac{1}{12}(39522.12)^2]$$
$$= 147.46$$

On making substitutions from table 11.2, one gets

$$v(\bar{y}_{ml}) = (\frac{1}{12} - \frac{1}{117})\,(147.46) + \frac{2090269.26}{(12)\,(117)\,(86.55)^2}$$
$$= 11.03 + .20$$
$$= 11.23$$

The confidence limits for the population mean are given by

$$\bar{y}_{ml} \pm 2\sqrt{v(\bar{y}_{ml})}$$
$$= 38.05 \pm 2\sqrt{11.23}$$
$$= 38.05 \pm 6.70$$
$$= 31.35,\ 44.75$$

These limits yield the confidence interval as [31.35, 44.75]. It means that the investigator can reasonably believe that if whole of the population is enumerated, the average dues, per society, standing against the defaulters are likely to be in the range of 31.35 to 44.75 thousand rupees. ■

The estimator $\bar{y}_{ml}$ in (11.5), though unbiased, uses the knowledge of M_0. In practice, the situations may arise where the value of M_0 (or equivalently $\bar{M}$) is not available. For such cases, the estimators that do not depend on M_0 are needed. We shall now consider two such estimators of mean $\bar{Y}$.

11.3.2 *Estimator 2*

Again, let

$$\left.\begin{aligned} S_{2b}^2 &= \frac{1}{N-1} \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y}_N)^2 \\ &= \frac{1}{N-1} (\sum_{i=1}^{N} \bar{Y}_i^2 - N\bar{Y}_N^2) \end{aligned}\right] \qquad (11.8)$$

$$\left.\begin{aligned} s^2_{2b} &= \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{y}_{m2})^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\bar{y}_i^2 - n\bar{y}^2_{m2}\right) \end{aligned}\right] \tag{11.9}$$

where $\bar{y}_{m2}$ is obtained from (11.10). The expressions for bias, variance, and estimator of variance for this estimator are given in relations (11.11) through (11.13).

Estimator of population mean which does not depend on M_o :

$$\bar{y}_{m2} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i \tag{11.10}$$

Bias of the estimator $\bar{y}_{m2}$:

$$B(\bar{y}_{m2}) = -\frac{1}{N\bar{M}}\sum_{i=1}^{N}(M_i - \bar{M})\,\bar{Y}_i \tag{11.11}$$

Variance of the estimator $\bar{y}_{m2}$:

$$V(\bar{y}_{m2}) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2_{2b} + \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_i^2 \tag{11.12}$$

Estimator of variance $V(\bar{y}_{m2})$:

$$v(\bar{y}_{m2}) = \left(\frac{1}{n} - \frac{1}{N}\right)s^2_{2b} + \frac{1}{nN}\sum_{i=1}^{n}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_i^2 \tag{11.13}$$

The expressions for S^2_{2b}, S^2_i, s^2_{2b}, and s^2_i are given in (11.8), (11.2), (11.9), and (11.4) respectively.

It can be seen that the first term in (11.6) depends on the variation between psu totals. This component is larger than the corresponding component in (11.12), provided the correlation between psu size and the psu mean is positive, and the bias is small. The second term of (11.6) is also likely to be more than the second term of (11.12) as there is expected to be a positive correlation between M_i and S_i^2. Because of these considerations, the estimator $\bar{y}_{m2}$ should be preferred over the estimator $\bar{y}_{m1}$ unless the bias in $\bar{y}_{m2}$ is serious. We now consider another estimator of mean which does not depend on the knowledge of M_o. Thus, it can be used irrespective of the availability of information on M_o.

11.3.3 *Estimator 3*

We further define

$$\left.\begin{aligned} S^2_m &= \frac{1}{N-1}\sum_{i=1}^{N}(M_i - \bar{M})^2 \\ &= \frac{1}{N-1}\left(\sum_{i=1}^{N}M_i^2 - N\bar{M}^2\right) \end{aligned}\right] \tag{11.14}$$

$$\left.\begin{aligned} S_{my} &= \frac{1}{N-1} \sum_{i=1}^{N} (M_i - \overline{M})(M_i \overline{Y}_i - \overline{Y}\,\overline{M}) \\ &= \frac{1}{N-1} \left(\sum_{i=1}^{N} M_i Y_{i.} - N\overline{Y}\,\overline{M}^2\right) \end{aligned}\right] \quad (11.15)$$

$$S_{3b}^2 = \frac{1}{\overline{M}^2 (N-1)} \sum_{i=1}^{N} M_i^2 (\overline{Y}_i - \overline{Y})^2 \quad (11.16)$$

$$s_{3b}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \frac{M_i^2}{\overline{M}^2} (\overline{y}_i - \overline{y}_{m3})^2 \quad (11.17)$$

where $\overline{y}_{m3}$ has been defined in (11.18).

Estimator of population mean which does not depend on M_o :

$$\overline{y}_{m3} = \frac{\sum_{i=1}^{n} M_i \overline{y}_i}{\sum_{i=1}^{n} M_i} \quad (11.18)$$

Approximate bias of the estimator $\overline{y}_{m3}$:

$$B(\overline{y}_{m3}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{\overline{M}^2} (\overline{Y} S_m^2 - S_{my}) \quad (11.19)$$

Approximate variance of the estimator $\overline{y}_{m3}$:

$$V(\overline{y}_{m3}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{3b}^2 + \frac{1}{\overline{M}^2 nN} \sum_{i=1}^{N} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 \quad (11.20)$$

Estimator of variance $V(\overline{y}_{m3})$:

$$v(\overline{y}_{m3}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{3b}^2 + \frac{1}{\overline{M}^2 nN} \sum_{i=1}^{n} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 \quad (11.21)$$

where S_m^2, S_{my}, S_{3b}^2, and s_{3b}^2 are defined in (11.14) to (11.17), whereas S_i^2 and s_i^2 are defined in (11.2) and (11.4) respectively.

The estimator $\overline{y}_{m3}$ is likely to be more efficient than $\overline{y}_{m1}$ and $\overline{y}_{m2}$, provided n is large and correlation coefficient between $M_i\overline{Y}_i$ and M_i is positive and greater than $CV(M_i)/2\ CV(M_i\overline{Y}_i)$.

Example 11.2

An orchard owner is to sell a truckload of oranges. The oranges are packed into 140 cartons. The number of oranges per carton may vary, and the total number of oranges in the truck is also not known exactly. Before striking the deal, the buyer thinks it wise to have an idea regarding the quantity of juice in the oranges. To do this, the buyer selects 10 cartons at random and then selects approximately 5 percent of the oranges from each selected carton. The information in respect of total number of oranges, and the juice obtained from the selected oranges, is given in table 11.3.

Table 11.3 Quantity of juice (in ml) for sample oranges

Carton	M_i	m_i	Juice (in ml)							Total
1	105	5	90	103	76	84	89			442
2	120	6	107	80	72	110	70	84		523
3	95	5	104	93	83	76	91			447
4	132	7	86	93	101	81	77	99	109	646
5	111	6	91	97	85	110	101	80		564
6	117	6	88	84	99	106	92	78		547
7	86	4	101	100	91	113				405
8	122	6	109	78	89	91	90	98		555
9	130	7	114	101	96	108	80	103	108	710
10	119	6	87	94	90	109	102	84		566

Estimate the average quantity of juice per orange, and also work out the probable range wherein the population mean would have fallen if all the oranges in the truck were observed.

Solution

In this problem, we are given that N = 140 and n = 10. Since total number of oranges M_0 (or equivalently $\bar{M}$) is not known, we can use either of the estimators given in (11.10) and (11.18). In this illustration, we proceed with estimator 2 given in (11.10). As in example 11.1, we prepare table 11.4 which provides details for the various steps involved in the solution.

Table 11.4 Calculations required for estimating population mean

Carton	M_i	m_i	$y_{i.}$	$\bar{y}_i = \frac{y_{i.}}{m_i}$	s_i^2	$(\frac{1}{m_i} - \frac{1}{M_i}) s_i^2$
1	105	5	442	88.40	97.30	18.53
2	120	6	523	87.17	300.17	47.53
3	95	5	447	89.40	112.30	21.28
4	132	7	646	92.29	133.57	18.07
5	111	6	564	94.00	120.00	18.92
6	117	6	547	91.17	103.37	16.34
7	86	4	405	101.25	81.58	19.45
8	122	6	555	92.50	106.70	16.91
9	130	7	710	101.43	122.62	16.57
10	119	6	566	94.33	90.67	14.35
Total	1137			931.94		207.95

From (11.10) and the table 11.4, we have the estimator of average quantity of juice per orange as

$$\bar{y}_{m2} = \frac{931.94}{10} = 93.19$$

Using (11.9), let us first compute s^2_{2b}. Thus we have

$$\begin{aligned} s^2_{2b} &= \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{m2})^2 \\ &= \frac{1}{10-1} [(88.40-93.19)^2 + (87.17-93.19)^2 + \dots + (94.33-93.19)^2] \\ &= \frac{1}{10-1} [(88.40)^2 + (87.17)^2 + \dots + (94.33)^2 - 10(93.19)^2] \\ &= 23.75 \end{aligned}$$

Making use of the calculated value of s^2_{2b} and column (7) of table 11.4, we work out the estimate of variance from (11.13). Therefore,

$$\begin{aligned} v(\bar{y}_{m2}) &= (\frac{1}{10} - \frac{1}{140}) (23.75) + \frac{207.95}{(10)(140)} \\ &= 2.354 \end{aligned}$$

As required in the statement of the problem, the lower and upper limits of the range, wherein the population mean is expected to fall, are obtained as

$$\begin{aligned} &\bar{y}_{m2} \pm 2\sqrt{v(\bar{y}_{m2})} \\ &= 93.19 \pm 2\sqrt{2.354} \\ &= 93.19 \pm 3.07 \\ &= 90.12,\ 96.26 \end{aligned}$$

It can, therefore, be said that the actual population average will be covered by the interval [90.12, 96.26] ml, with probability approximately .95. ■

Example 11.3

Using data of example 11.2, estimate the parameter in question through the alternative estimator $\bar{y}_{m3}$ defined in (11.18).

Solution

For working out the required estimate, some of the values computed in table 11.4 will be used. The estimator $\bar{y}_{m3}$ of the population mean, given in (11.18), is

$$\bar{y}_{m3} = \frac{\sum\limits_{i=1}^{n} M_i \bar{y}_i}{\sum\limits_{i=1}^{n} M_i}$$

On making use of columns (2) and (5) of table 11.4, we get

$$\bar{y}_{m3} = \frac{(105)(88.40) + (120)(87.17) + \ldots + (119)(94.33)}{(105+120 + \ldots + 119)}$$

$$= \frac{105922.24}{1137}$$

$$= 93.16$$

Thus, on using the estimator 3, the estimate of juice per orange is obtained as 93.16 ml.

Since $\bar{M}$ is not available, we shall use its sample estimate $\hat{\bar{M}}$ wherever necessary. For this we have

$$\hat{\bar{M}} = \frac{1}{n} \sum_{i=1}^{n} M_i = \frac{1137}{10} = 113.7$$

The estimator of variance from (11.21) is

$$v(\bar{y}_{m3}) = (\frac{1}{n} - \frac{1}{N}) s_{3b}^2 + \frac{1}{nN} \sum_{i=1}^{n} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2$$

For obtaining the value of $v(\bar{y}_{m3})$, we first compute the two terms involved in it. For working out the first term, we make use of columns (2) and (5) of table 11.4. After replacing $\bar{M}$ by $\hat{\bar{M}}$, the first term is

$$\frac{1}{\hat{\bar{M}}^2 (n-1)} \sum_{i=1}^{n} M_i^2 (\bar{y}_i - \bar{y}_{m3})^2 = \frac{1}{(113.7)^2 (10-1)} [(105)^2 (88.40-93.16)^2 + (120)^2$$

$$(87.17-93.16)^2 + \ldots + (119)^2 (94.33-93.16)^2]$$

$$= 22.655$$

The second term in $v(\bar{y}_{m3})$, after replacing $\bar{M}$ by $\hat{\bar{M}}$, is calculated by making use of column (7) of table 11.4. Therefore,

$$\sum_{i=1}^{n} \frac{M_i^2}{\hat{\bar{M}}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 = \frac{1}{(113.7)^2} [(105)^2 (18.53) + (120)^2 (47.53)$$

$$+ \ldots + (119)^2 (14.35)]$$

$$= 211.268$$

On substituting the values calculated above in the expression for $v(\bar{y}_{m3})$, one gets

$$\begin{aligned} v(\bar{y}_{m3}) &= (\frac{1}{10} - \frac{1}{140})(22.655) + \frac{211.268}{(10)(140)} \\ &= 2.104 + .151 \\ &= 2.255 \end{aligned}$$

The confidence limits for the average quantity of juice per orange, for whole of the target population, are given by

$$\begin{aligned} &\bar{y}_{m3} \pm 2\sqrt{v(\bar{y}_{m3})} \\ &= 93.16 \pm 2\sqrt{2.255} \\ &= 93.16 \pm 3.00 \\ &= 90.16,\ 96.16 \end{aligned}$$

One can, therefore, be reasonably sure that the average quantity of juice per orange for the whole lot of oranges in the truck is likely to be in the range of 90.16 to 96.16 ml. ■

Estimation of population total Y is a straightforward exercise from the estimators presented for mean. Estimators of total can be obtained by multiplying the estimators of mean given in (11.5), (11.10), and (11.18) by M_o.

Unbiased estimator of population total that does not depend on M_o :

$$\hat{Y}_{ml} = M_o\bar{y}_{ml} = \frac{N}{n}\sum_{i=1}^{n} M_i\,\bar{y}_i \tag{11.22}$$

Estimators of population total using M_o :

$$\hat{Y}_{m2} = M_o\bar{y}_{m2} = \frac{M_o}{n}\sum_{i=1}^{n}\bar{y}_i \tag{11.23}$$

$$\hat{Y}_{m3} = M_o\bar{y}_{m3} = \frac{M_o\sum_{i=1}^{n} M_i\,\bar{y}_i}{\sum_{i=1}^{n} M_i} \tag{11.24}$$

The expressions for bias in the estimators $\hat{Y}_{m2}$ and $\hat{Y}_{m3}$ can be obtained by multiplying the corresponding expressions for $\bar{y}_{m2}$ and $\bar{y}_{m3}$ respectively by M_o. Similarly, the expressions for variances and variance estimators can be arrived at by multiplying the corresponding expressions for mean by M_o^2.

Example 11.4

Assume that in example 11.1, the total number of societies in the state is not known. Using the data of that example, estimate the total amount standing against defaulters of co-operative societies in the state. Also, obtain the standard error of the estimate, and place confidence limits on the actual total amount.

Solution

Here M_o is not available. The estimator $\hat{Y}_{ml}$ given in (11.22) will, therefore, be used to estimate the population total. Thus

$$\hat{Y}_{ml} = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{y}_i$$

Making use of columns (2) and (4) of table 11.2, one gets

$$\hat{Y}_{ml} = \frac{117}{12} [(60)(38.23) + (102)(34.63) + \dots + (56)(50.42)]$$
$$= 385340.67$$

as an estimate of the total amount standing against all the defaulters in the state.

The estimator of the variance $V(\hat{Y}_{ml})$ can be written, from (11.7), as

$$v(\hat{Y}_{ml}) = M_o^2 v(\bar{y}_{ml}) = N^2\bar{M}^2 (\frac{1}{n} - \frac{1}{N}) s_{1b}^2 + \frac{N}{n} \sum_{i=1}^{n} M_i^2 (\frac{1}{m_i} - \frac{1}{M_i}) s_i^2 \quad (11.25)$$

Since M_o, or equivalently $\bar{M}$, is unknown, we instead use its sample estimate $\hat{\bar{M}}$, where

$$\hat{\bar{M}} = \frac{1}{n} \sum_{i=1}^{n} M_i$$
$$= \frac{1}{12} (60+102 + \dots + 56)$$
$$= \frac{907}{12}$$
$$= 75.58$$

For obtaining the value of s_{1b}^2, we shall now use $\hat{\bar{M}}$ in place of $\bar{M}$. Therefore, s_{1b}^2 is computed as

$$s_{1b}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{M_i\bar{y}_i}{\hat{\bar{M}}} - \bar{y}_{ml} \right)^2$$

where from example 11.1, $\bar{y}_{ml} = 38.05$. Thus on using the figures from table 11.2, we get

$$s_{1b}^2 = \frac{1}{12-1} \left[\left(\frac{(60)(38.23)}{75.58} - 38.05 \right)^2 + \left(\frac{(102)(34.63)}{75.58} - 38.05 \right)^2 + \dots + \left(\frac{(56)(50.42)}{75.58} - 38.05 \right)^2 \right]$$
$$= \frac{2493.55}{11}$$
$$= 226.69$$

Also, from table 11.2

$$\sum_{i=1}^{n} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 = 2090269.26$$

On using $\hat{M} = 75.58$ in place of $\bar{M}$ in (11.25), and making other substitutions, one obtains

$$\begin{aligned} v(\hat{Y}_{ml}) &= (117)^2 (75.58)^2 \left(\frac{1}{12} - \frac{1}{117}\right)(226.69) + \frac{(117)(2090269.26)}{12} \\ &= 1325684115 + 20380125 \\ &= 1346064240 \end{aligned}$$

The standard error of the estimate will, therefore, be

$$\begin{aligned} se(\hat{Y}_{ml}) &= \sqrt{1346064240} \\ &= 36688.75 \end{aligned}$$

The interval in which the total amount standing against defaulters all over the state would probably fall, is given by

$$\begin{aligned} &\hat{Y}_{ml} \pm 2\, se\,(\hat{Y}_{ml}) \\ &= 385340.67 \pm 73377.50 \\ &= 311963.17,\ 458718.17 \qquad (11.25) \end{aligned}$$

Thus the total amount due is likely to range from 311963.17 to 458718.17 thousand rupees. ■

11.4 ESTIMATION OF PROPORTION

As mentioned earlier, the estimators for population proportion can be obtained from those for population mean by allowing the study variable y_{ij} to take values 1, or 0, depending on whether the j-th unit in the i-th psu falls into the category of interest, or not. Let p_i be the proportion of sampled ssu's from i-th psu that fall into specified category. Again, we consider three estimators of population proportion corresponding to the estimators of mean in (11.5), (11.10), and (11.18).

11.4.1 *Estimator 1*

We find that for a variable y taking 0 and 1 values

$$\left.\begin{aligned} S_{1b}^2 &= \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{M_i P_i}{\bar{M}} - P\right)^2 \\ &= \frac{1}{\bar{M}^2(N-1)} \left[\sum_{i=1}^{N} (M_i P_i)^2 - \frac{1}{N}\left(\sum_{i=1}^{N} M_i P_i\right)^2\right] \end{aligned}\right] \qquad (11.26)$$

$$\left.\begin{aligned} s_{1b}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i p_i}{\bar{M}} - p_{ml}\right)^2 \\ &= \frac{1}{\bar{M}^2(n-1)}\left[\sum_{i=1}^{n}(M_i p_i)^2 - \frac{1}{n}\left(\sum_{i=1}^{n} M_i p_i\right)^2\right] \end{aligned}\right] \qquad (11.27)$$

where P_i and P are the proportions of units falling in the category under consideration in the i-th psu, and in the whole population respectively. Also, p_i and p are their sample analogs. The value of p_{ml} is computed from (11.28).

Unbiased estimator of population proportion when M_o is known :

$$p_{ml} = \frac{N}{nM_o}\sum_{i=1}^{n} M_i p_i \qquad (11.28)$$

Variance of the estimator p_{ml}:

$$V(p_{ml}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_{1b}^2 + \frac{1}{nN}\sum_{i=1}^{N}\frac{M_i^2}{\bar{M}^2}\left(\frac{M_i - m_i}{M_i - 1}\right)\frac{P_i Q_i}{m_i} \qquad (11.29)$$

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

$$v(p_{ml}) = \left(\frac{1}{n} - \frac{1}{N}\right)s_{1b}^2 + \frac{1}{nN\bar{M}^2}\sum_{i=1}^{n}\frac{M_i(M_i - m_i)\,p_i\,q_i}{m_i - 1} \qquad (11.30)$$

where $Q_i = 1-P_i$ and $q_i = 1-p_i$. Also, the terms S_{1b}^2 and s_{1b}^2 appearing in (11.29) and (11.30) are defined in (11.26) and (11.27) respectively.

Example 11.5

A state government has 4032 harvester combines. These were allocated 5 years ago to 96 centers in the state. These centers look after the operation of combines at their disposal. Exact number of combines at any point of time, with each center, is not known as the combines are transferred from one center to the other depending on the pressure of work. The state needs to estimate the proportion of the five year old combines that are operating at loss during the current paddy harvesting season, more than 80 percent of which is already over. For this purpose, a WOR random sample of 12 centers was drawn. For each center in the sample, about 20 percent of combines were selected and the number of combines operating at loss (COL) determined. The information thus collected is presented in table 11.5.

Table 11.5 Data regarding combines, and certain other computations

Center	M_i	m_i	COL	p_i	M_ip_i	$\frac{M_i(M_i - m_i)\, p_i\, q_i}{m_i - 1}$
1	50	10	2	.2000	10.00	35.56
2	35	7	1	.1429	5.00	20.00
3	42	8	3	.3750	15.75	47.81
4	46	9	3	.3333	15.33	47.28
5	37	7	2	.2857	10.57	37.75
6	55	11	4	.3636	20.00	56.00
7	40	8	1	.1250	5.00	20.00
8	35	7	2	.2857	10.00	33.33
9	47	9	2	.2222	10.44	38.58
10	31	6	0	0	0	0
11	44	9	4	.4444	19.55	47.53
12	48	10	3	.3000	14.40	42.56
Total	510				136.04	426.40

Estimate the proportion of combines operating at loss, and build up the confidence interval for this proportion in the state.

Solution

The statement of the example provides that N=96, $M_o = 4032$, and n = 12. Since M_o is known, we use estimator p_{ml} defined in (11.28). This is

$$p_{ml} = \frac{N}{nM_o} \sum_{i=1}^{n} M_i\, p_i$$

Making use of total of column (6) of table 11.5, one gets the estimate of proportion of the combines operating at loss, as

$$p_{ml} = \frac{(96)\,(136.04)}{(12)\,(4032)} = .2699$$

The estimate of variance $V(p_{ml})$ is obtained by using (11.30). The expression for the variance estimator is

$$v(p_{ml}) = (\frac{1}{n} - \frac{1}{N})\, s_{lb}^2 + \frac{1}{nN\overline{M}^2} \sum_{i=1}^{n} \frac{M_i(M_i - m_i)\, p_i\, q_i}{m_i - 1}$$

We then first compute

$$s_{lb}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{M_i p_i}{\overline{M}} - p_{ml} \right)^2$$

where

$$\overline{M} = \frac{4032}{96} = 42$$

Using figures from column (6) of table 11.5, we get

$$\begin{aligned} s_{1b}^2 &= \frac{1}{12-1}\left[\left(\frac{10.00}{42} - .2699\right)^2 + \left(\frac{5.00}{42} - .2699\right)^2 + \ldots + \left(\frac{14.40}{42} - .2699\right)^2\right] \\ &= \frac{.2274}{11} \\ &= .02067 \end{aligned}$$

The value of the summation term in the second component of $v(p_{m1})$, is the total of column (7) of table 11.5. Thus we get

$$\begin{aligned} v(p_{m1}) &= (\frac{1}{12} - \frac{1}{96})(.02067) + \frac{426.40}{(12)(96)(42)^2} \\ &= .001717 \end{aligned}$$

The required confidence interval for the population proportion is obtained from the limits

$$\begin{aligned} & p_{m1} \pm 2\sqrt{v(p_{m1})} \\ &= .2699 \pm 2\sqrt{.001717} \\ &= .2699 \pm .0829 \\ &= .1870, .3528 \end{aligned}$$

The above confidence limits indicate with approximate probability as .95, that 18.70 to 35.28 percent combines in the population of 4032 combines are incurring loss to the state. ■

For the situations where M_0 is unknown, we present estimators analogous to $\overline{y}_{m2}$ and $\overline{y}_{m3}$ given in (11.10) and (11.18) respectively.

11.4.2 *Estimator 2*

For a variable taking 0 and 1 values

$$\left.\begin{aligned} S_{2b}^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(P_i - \frac{1}{N}\sum_{i=1}^{N} P_i\right)^2 \\ &= \frac{1}{N-1}\left[\sum_{i=1}^{N} P_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} P_i\right)^2\right] \end{aligned}\right] \quad (11.31)$$

$$\left.\begin{aligned} s_{2b}^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (p_i - p_{m2})^2 \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{n} p_i^2 - np_{m2}^2\right) \end{aligned}\right] \qquad (11.32)$$

where p_{m2} is defined in (11.33) below. Then, we have the expressions (11.33) through (11.36).

Estimator of population proportion which does not depend on M_o :

$$p_{m2} = \frac{1}{n} \sum_{i=1}^{n} p_i \qquad (11.33)$$

Bias of the estimator p_{m2} :

$$B(p_{m2}) = -\frac{1}{N\overline{M}} \sum_{i=1}^{N} (M_i - \overline{M})\, P_i \qquad (11.34)$$

Variance of the estimator p_{m2} :

$$V(p_{m2}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{2b}^2 + \frac{1}{nN} \sum_{i=1}^{N} \left(\frac{M_i - m_i}{M_i - 1}\right) \frac{P_i Q_i}{m_i} \qquad (11.35)$$

Estimator of variance $V(p_{m2})$:

$$v(p_{m2}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{2b}^2 + \frac{1}{nN} \sum_{i=1}^{n} \left(\frac{M_i - m_i}{M_i}\right) \frac{p_i q_i}{m_i - 1} \qquad (11.36)$$

The terms S_{2b}^2 and s_{2b}^2 involved in (11.35) and (11.36) above are defined in (11.31) and (11.32) respectively.

11.4.3 *Estimator 3*

For the kind of variable under consideration,

$$\left.\begin{aligned} S_{my} &= \frac{1}{N-1} \sum_{i=1}^{N} (M_i - \overline{M})(M_i P_i - \overline{M}P) \\ &= \frac{1}{N-1} \left(\sum_{i=1}^{N} M_i^2 P_i - N\overline{M}^2 P\right) \end{aligned}\right] \qquad (11.37)$$

$$S_{3b}^2 = \frac{1}{\overline{M}^2 (N-1)} \sum_{i=1}^{N} M_i^2 (P_i - P)^2 \qquad (11.38)$$

$$s_{3b}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \frac{M_i^2}{\overline{M}^2} (p_i - p_{m3})^2 \qquad (11.39)$$

with p_{m3} defined in (11.40). We thus have :

Estimator of population proportion which does not depend on M_o :

$$p_{m3} = \frac{\sum_{i=1}^{n} M_i p_i}{\sum_{i=1}^{n} M_i} \tag{11.40}$$

Approximate bias of the estimator p_{m3} :

$$B(p_{m3}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{\bar{M}^2} (PS_m^2 - S_{my}) \tag{11.41}$$

Approximate variance of the estimator p_{m3} :

$$V(p_{m3}) = (\frac{1}{n} - \frac{1}{N}) S_{3b}^2 + \frac{1}{nN\bar{M}^2} \sum_{i=1}^{N} \left(\frac{M_i^2 (M_i - m_i)}{M_i - 1}\right) \frac{P_i Q_i}{m_i} \tag{11.42}$$

Estimator of variance $V(p_{m3})$:

$$v(p_{m3}) = (\frac{1}{n} - \frac{1}{N}) s_{3b}^2 + \frac{1}{nN\bar{M}^2} \sum_{i=1}^{n} \left(\frac{M_i^2 (M_i - m_i)}{M_i}\right) \frac{p_i q_i}{m_i - 1} \tag{11.43}$$

The terms S_m^2, S_{my}, S_{3b}^2, and s_{3b}^2 have already been defined in (11.14), (11.37), (11.38), and (11.39) respectively. When $\bar{M}$ is not known, it has to be replaced by $\hat{\bar{M}}$ in (11.43).

Example 11.6

Do only stray dog bites cause the rabies ? The answer to this question is required to frame a policy for giving dogs antirabies shots. A survey was undertaken for this purpose in an Indian state. The state has 70 hospitals where dog bite cases are treated. The addresses of persons, treated for dog bite, were available in these hospitals. However, the category of the dogs - stray or pet- who had bitten the patient, was not mentioned in the records. Keeping the resource constraints in view, a WOR simple random sample of 11 hospitals was drawn. About 5 percent of the treated persons were sampled using SRS without replacement. The data so collected, are presented in table 11.6. The observation is recorded as 1 if the rabies was caused by pet dog bite, 0 otherwise.

Table 11.6 Sample observations regarding type of dog causing rabies

Hospital	M_i	m_i	Observations	Total	p_i
1	140	7	0 1 1 0 0 0 1	3	.4286
2	203	10	1 0 0 0 0 0 1 1 0 0	3	.3000
3	91	5	0 1 0 0 1	2	.4000
4	176	9	1 0 0 1 0 1 0 0 0	3	.3333
5	121	6	0 1 0 0 0 0	1	.1667
6	263	13	1 0 1 1 0 0 0 1 0 0 0 1 0	5	.3846
7	118	6	0 1 0 0 0 1	2	.3333
8	144	7	1 0 1 1 1 0 0	4	.5714
9	236	12	0 0 0 1 0 1 0 0 0 0 0 0	2	.1667
10	184	9	1 0 0 1 0 0 1 0 1	4	.4444
11	137	7	0 0 1 0 0 0 0	1	.1429

Estimate the proportion of rabies caused by pet dogs. Also, place confidence limits on it.

Solution

We have in this case, N = 70 and n = 11. As in the foregoing examples, we prepare table 11.7.

Table 11.7 Certain intermediate calculations

Hospital	M_i	m_i	p_i	$\left(\frac{M_i - m_i}{M_i}\right)\frac{p_i q_i}{m_i - 1}$	M_i^2 Col (5)	$M_i^2(p_i\text{-}p_{m3})^2$
1	140	7	.4286	.0388	760.48	187.85
2	203	10	.3000	.0222	914.84	38.84
3	91	5	.4000	.0567	469.53	39.77
4	176	9	.3333	.0264	817.77	.21
5	121	6	.1667	.0264	386.52	393.78
6	263	13	.3846	.0187	1293.46	200.95
7	118	6	.3333	.0422	587.59	.09
8	144	7	.5714	.0388	804.56	1201.37
9	236	12	.1667	.0120	668.35	1498.00
10	184	9	.4444	.0294	995.37	437.68
11	137	7	.1429	.0194	364.12	661.96
Total	1813		3.6719	.3310	8062.59	4660.50

Since M_0 is not known, the estimate of proportion of rabies caused by pet dogs can, therefore, be obtained by using either of the two estimators in (11.33) and (11.40). In this example, we illustrate the use of estimator p_{m2} defined in (11.33). Thus,

$$p_{m2} = \frac{1}{n}\sum_{i=1}^{n} p_i$$

$$= \frac{3.6719}{11} \qquad \text{(from column (4) of table 11.7)}$$

$$= .3338$$

The estimate of variance of the above estimator is provided by (11.36) as

$$v(p_{m2}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{2b}^2 + \frac{1}{nN}\sum_{i=1}^{n}\left(\frac{M_i - m_i}{M_i}\right)\frac{p_i q_i}{m_i - 1}$$

where

$$s_{2b}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(p_i - p_{m2})^2$$

On using column (4) of table 11.7 and the value of p_{m2} calculated above, one obtains

$$s_{2b}^2 = \frac{1}{11-1}[(.4286-.3338)^2 + (.3000-.3338)^2 + ... + (.1429-.3338)^2]$$

$$= \frac{1}{10}[(.4286)^2 + (.3000)^2 + ... + (.1429)^2 - 11\,(.3338)^2]$$

$$= .0178$$

The value of second term in $v(p_{m2})$ is calculated by using column (5) of table 11.7. Thus we get

$$v(p_{m2}) = (\frac{1}{11} - \frac{1}{70})(.0178) + \frac{.3310}{(11)(70)}$$

$$= .001794$$

The confidence interval is given by the limits

$$p_{m2} \pm 2\sqrt{v(p_{m2})}$$

$$= .3338 \pm 2\sqrt{.001794}$$

$$= .3338 \pm .0847$$

$$= .2491,\ .4185$$

Thus had all the patients of rabies treated in 70 hospitals been contacted, the proportion of rabies caused by pet dogs would most probably have taken a value in the range of 24.91% to 41.85%. ■

Example 11.7

Using data of example 11.6, estimate the required proportion by using estimator p_{m3} given in (11.40). Also, work out the confidence interval for the parameter being estimated.

Solution

Again, we are given N = 70 and n = 11. The proportion of rabies caused by pet dogs is now estimated using estimator p_{m3}, where

$$p_{m3} = \frac{\sum_{i=1}^{n} M_i p_i}{\sum_{i=1}^{n} M_i}$$

Using columns (2) and (4) of table 11.7, the estimator p_{m3} is evaluated as

$$p_{m3} = \frac{1}{1813}[(140)(.4286) + (203)(.3000) + ... + (137)(.1429)]$$

$$= \frac{599.5844}{1813}$$

$$= .3307$$

The estimator of variance is provided by (11.43), where

$$v(p_{m3}) = (\frac{1}{n} - \frac{1}{N}) s_{3b}^2 + \frac{1}{nN\overline{M}^2} \sum_{i=1}^{n} \left(\frac{M_i^2(M_i - m_i)}{M_i} \right) \frac{p_i q_i}{m_i - 1}$$

As $\overline{M}$ is not available in this case, its sample estimate

$$\hat{\overline{M}} = \frac{1}{n} \sum_{i=1}^{n} M_i = \frac{1813}{11} = 164.82$$

will be used in its place. Thus, s_{3b}^2 can be put as

$$s_{3b}^2 = \frac{1}{\hat{\overline{M}}^2 (n-1)} \sum_{i=1}^{n} M_i^2 (p_i - p_{m3})^2$$

The use of column (7) of table 11.7 yields

$$s_{3b}^2 = \frac{4660.50}{(164.82)^2(10)} = .01716$$

The second component of $v(p_{m3})$ is evaluated by using total of column (6) of table 11.7. We thus obtain

$$v(p_{m3}) = (\frac{1}{11} - \frac{1}{70}) (.01716) + \frac{8062.59}{(11)(70)(164.82)^2}$$
$$= .0017$$

The required confidence limits, within which the population proportion of rabies caused by pet dogs is likely to fall, are arrived at by using

$$p_{m3} \pm 2 \sqrt{v(p_{m3})}$$
$$= .3307 \pm 2 \sqrt{.0017}$$
$$= .3307 \pm .0825$$
$$= .2482, .4132 \blacksquare$$

11.5 ESTIMATION OF MEAN / TOTAL USING PPSWR AND SRSWOR

It is possible to increase the efficiency of multistage sampling by making use of the auxiliary information that may be available for first and subsequent stage units. For instance, in a socio-economic survey where the ultimate sampling unit is the household, villages may be treated as psu's. These could be selected with probability proportional to size and with replacement, the size being the number of households, or population for the village. If the number of second stage units in the psu's differ considerably, it may be useful to select psu's with probability proportional to M_i values where M_i, as before, is the number of ssu's in the i-th psu, i = 1, 2,...,N. One may also follow more efficient strategy for selecting ssu's by utilizing any other auxiliary information available for them in the selected psu's.

For discussion in this section, we assume that a sample of n psu's is selected WR using unequal probabilities $\{P_i\}$. From the M_i ssu's of the i-th selected psu, a sample of m_i ssu's is drawn with WOR simple random sampling. This sampling strategy gives following formulas.

Unbiased estimator of population total Y :

$$\hat{Y}_{mp} = \frac{1}{n}\sum_{i=1}^{n} M_i \frac{\bar{y}_i}{p_i} \tag{11.44}$$

Variance of the estimator $\hat{Y}_{mp}$:

$$V(\hat{Y}_{mp}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_{i.}}{P_i} - Y\right)^2 P_i + \frac{1}{n}\sum_{i=1}^{N}\frac{M_i^2}{P_i}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_i^2 \tag{11.45}$$

where S_i^2 has been defined in (11.2).

Estimator of variance $V(\hat{Y}_{mp})$:

$$\left.\begin{aligned} v(\hat{Y}_{mp}) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{M_i\bar{y}_i}{p_i} - \hat{Y}_{mp}\right)^2 \\ &= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{M_i\bar{y}_i}{p_i}\right)^2 - n\hat{Y}_{mp}^2\right] \end{aligned}\right] \tag{11.46}$$

Corresponding estimator of population mean will be obtained by dividing $\hat{Y}_{mp}$ by M_o.

Estimator of population mean $\bar{Y}$:

$$\bar{y}_{mp} = \hat{Y}_{mp}/M_o = \frac{1}{nM_o}\sum_{i=1}^{n}\frac{M_i\bar{y}_i}{p_i} \tag{11.47}$$

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

$$V(\bar{y}_{mp}) = \frac{V(\hat{Y}_{mp})}{M_o^2} \tag{11.48}$$

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

$$v(\bar{y}_{mp}) = \frac{v(\hat{Y}_{mp})}{M_o^2} \tag{11.49}$$

where $V(\hat{Y}_{mp})$ and $v(\hat{Y}_{mp})$ are given in (11.45) and (11.46) respectively.

Example 11.8

In order to estimate the total production of wheat in a certain district, 16 villages out of a total of 410 villages were selected using PPS with replacement sampling, the size measure being the net cropped area in hectares. The total net cropped area for the district was 140576 hectares. About 3 percent of farmers were selected from the sampled villages. The total produce of wheat (in quintals) for each of the sampled farmer, as reported by him, was recorded. These data are presented in table 11.8.

Table 11.8 Net cropped area (x_i) for the village and production of wheat (y_{ij}) for the selected farmers

Village	x_i	$p_i = \frac{x_i}{X}$	M_i	m_i	y_{ij}				Total ($y_{i.}$)	$\bar{y}_i = \frac{y_{i.}}{m_i}$
1	420	.00299	96	3	138	166	190		494	164.67
2	613	.00436	112	3	142	185	215		542	180.67
3	178	.00127	40	1	110				110	110.00
4	199	.00142	46	1	133				133	133.00
5	345	.00245	86	3	160	164	210		534	178.00
6	467	.00332	122	4	100	162	85	124	471	117.75
7	123	.00087	42	1	107				107	107.00
8	328	.00233	92	3	140	163	116		419	139.67
9	150	.00107	61	2	105	98			203	101.50
10	764	.00543	190	6	200	140	173	160	902	150.33
					101	128				
11	269	.00191	76	2	120	135			255	127.50
12	483	.00344	138	4	149	113	161	131	554	138.50
13	389	.00277	98	3	110	124	90		324	108.00
14	212	.00151	66	2	190	105			295	147.50
15	160	.00114	34	1	166				166	166.00
16	532	.00378	150	5	136	170	100	156	702	140.40
					140					

Estimate the total production of wheat in the district, and determine confidence limits for it.

Solution

We are given that X = 140576, N = 410, and n = 16. In table 11.8, the observations in column (7) are the total of observations in column (6) for a given village. The estimate of total production of wheat is given by (11.44) as

$$\hat{Y}_{mp} = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{y}_i}{p_i}$$

$$= \frac{1}{16}\left[\frac{(96)\,(164.67)}{.00299} + \frac{(112)\,(180.67)}{.00436} + \ldots + \frac{(150)\,(140.40)}{.00378}\right]$$

$$= \frac{81422932}{16}$$

$$= 5088933.2$$

We now obtain estimate of variance $V(\hat{Y}_{mp})$. For this we use (11.46). Therefore,

$$v(\hat{Y}_{mp}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{M_i\bar{y}_i}{p_i} - \hat{Y}_{mp}\right)^2$$

$$= \frac{1}{(16)\,(15)}\left[\left\{\frac{(96)\,(164.67)}{.00299} - 5088933.2\right\}^2 + \left\{\frac{(112)\,(180.67)}{.00436} - 5088933.2\right\}^2 + \ldots + \left\{\frac{(150)\,(140.40)}{.00378} - 5088933.2\right\}^2\right]$$

$$= \frac{1.00376\times 10^{13}}{(16)\,(15)}$$

$$= 4.18236\times 10^{10}$$

The confidence limits, within which the total production of wheat in the district is likely to fall, are

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

$$= 5088933.2 \pm 2\sqrt{4.18236\times 10^{10}}$$

$$= 5088933.2 \pm 409016.4$$

$$= 4679916.8,\ 5497949.6$$

Thus the desired confidence interval is [4679916.8, 5497949.6] quintals. ■

The reader will notice that in the above example only one second stage unit (farmer) was selected from some of the selected first stage units (villages). It is possible to estimate the variance of the estimator of population total, or mean, in such a case where the sample of psu's is selected with replacement. In case the psu's are selected without replacement, one needs a sample of at least two second stage units from each selected psu to get an estimator of variance.

11.6 SOME FURTHER REMARKS

11.1 The results of this chapter can be easily extended to more than two stages of sampling, or stratified sampling. For these, and various other variants, reader may refer to Sukhatme *et al.* (1984) for details.

11.2 In case the primary stage units are selected with PPS with replacement and if $\hat{y}_{i.}$ denotes an unbiased estimator of the i-th sample psu total $y_{i.}$, i = 1, 2, ..., n, then an unbiased estimator of population total is given by

$$\hat{Y}_{mp} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{y}_{i.}}{p_i} \tag{11.50}$$

Also, an unbiased estimator of variance $V(\hat{Y}_{mp})$ is given by

$$v(\hat{Y}_{mp}) = \frac{1}{n(n-1)} \left[\sum_{i=1}^{n} \left(\frac{\hat{y}_{i.}}{p_i} \right)^2 - n\hat{Y}^2_{mp} \right] \tag{11.51}$$

The results in (11.50) and (11.51) hold for any number of stages and all kinds of estimators used at different stages, in building the estimator $\hat{y}_{i.}$.

LET US DO

11.1 Define multistage sampling, and identify the situations where it is to be preferred over usual SRS for selecting a sample of ultimate units.

11.2 Suppose the objective of a survey is to estimate the total wheat production in a state. The state has a number of districts and each district has several development blocks. Many villages constitute a block while there are several farmers in each village. The frame of villages is also available at the state level. Suggest what should be the psu's and ssu's, if one is to go for a two-stage sampling design ?

11.3 Give expression for the unbiased estimator of mean for a two-stage sampling design. What will be its variance, and how will you estimate it ? Choose your own sampling schemes at both the stages of sampling.

11.4 A district is running 110 nursery schools (*anganwadi*) in the rural area. The total number of children in these schools, is known to be 8040. The Department of Foods and Nutrition of a university has undertaken a project to determine the quality of food intake by the children in these schools. For this purpose, a sample of 10 schools was selected using WOR equal probability sampling. From each selected school, about 5 percent of children were selected using same procedure. Elaborate diet records were kept for each selected child, and the average daily calory intake was then determined. The information thus collected, is presented in the following table.

School	M_i	m_i	Average calory intake for sample children				
1	100	5	650	680	712	766	770
2	93	5	745	690	703	740	671
3	40	2	736	692			
4	80	4	703	777	687	714	

Table continued ...

School	M_i	m_i	Average calory intake for sample children				
5	42	2	699	724			
6	62	3	760	704	680		
7	96	5	715	660	690	670	650
8	46	2	668	701			
9	99	5	704	712	692	697	687
10	104	5	714	716	704	725	730

Estimate unbiasedly the average daily calory intake for the children of all the 110 schools, and also obtain the confidence interval for it.

11.5 In case the total number of second stage units in the population is not known, discuss how will you estimate the population mean ? Also, give the expressions for the variance and estimator of variance for the estimator(s) you propose to use.

11.6 In addition to the crop cutting survey, the Department of Agriculture had decided to estimate per hectare wheat yield by taking cultivator as the ultimate sampling unit. The list of all the 300 villages comprising the district was available. However, information regarding the number of cultivators in each village was not available. Two-stage sampling was, therefore, thought to be an appropriate design. A WOR simple random sample of 10 villages was drawn. About 10 percent of the cultivators in each sample village were selected using same sampling scheme. The per hectare wheat yield (in quintals) obtained by the selected cultivators was asked for by the investigator. This information, along with the total number of cultivators (M_i) and the number of cultivators sampled (m_i), is presented in the following table :

Village	M_i	m_i	Per hectare wheat yield				
1	40	4	29.70	32.80	28.80	30.65	
2	28	3	33.18	32.60	33.41		
3	35	4	27.76	25.30	31.40	30.60	
4	24	2	34.17	31.69			
5	17	2	27.28	32.40			
6	30	3	35.20	33.40	29.50		
7	25	3	28.24	27.45	34.90		
8	21	2	30.45	29.12			
9	38	4	28.23	31.56	33.45	29.70	
10	50	5	33.60	28.70	32.07	32.65	28.40

Estimate average per hectare yield of wheat in the district using estimator $\bar{y}_{m2}$ in (11.10). Also, build up the confidence interval for this average.

11.7 The excessive rains have caused damage to cotton crop in a certain area consisting of 132 villages. Before deciding on the extent of relief to be given to cultivators, the administration has decided to estimate the average per hectare loss for the area. For this purpose, a sample of 10 villages was selected using without replacement SRS, and about 5 percent of cultivators from the sample villages were selected using the same sampling scheme. Per hectare loss (in '00 rupees) incurred by each selected cultivator was assessed by visiting his fields. The results are as follows :

Village	M_i	m_i	Loss per hectare				
1	43	2	8	12			
2	76	4	16	7	11	13	
3	46	2	14	9			
4	67	3	17	10	12		
5	97	5	14	10	15	17	11
6	75	4	16	7	9	14	
7	83	4	9	13	11	13	
8	54	3	15	15	13		
9	69	3	12	9	11		
10	58	3	14	15	14		

Work out the estimated average per hectare loss, and also construct the confidence interval for it.

11.8 A project has been undertaken to study the feeding and rearing practices of sheep in Rajasthan state of India. As a first step, a survey was conducted to estimate the total sheep population in the state. For this purpose, a WOR simple random sample of 12 development blocks, from a total of 200 blocks, was selected. About 3 percent of the villages in the sample blocks were chosen using same sampling scheme. The following table presents the number of sheep in the selected villages along with the total number of villages (M_i) and the number of selected villages (m_i), in the sample blocks.

Block	M_i	m_i	Number of sheep				
1	160	5	141	376	267	201	55
2	130	4	60	228	270	80	
3	128	4	177	95	265	301	
4	114	3	225	107	119		
5	70	2	80	101			
6	105	3	201	69	134		
7	89	3	135	240	405		
8	71	2	309	111			
9	97	3	250	280	170		
10	106	3	96	269	118		
11	128	4	314	246	107	80	
12	69	2	307	286			

Using estimator $\hat{Y}_{m2}$ in (11.23), estimate the total number of sheep in the state. Also work out the standard error of the estimate, and construct the confidence interval for this total. The total number of villages in the state is 24000.

11.9 From the sample data given in exercise 11.8, estimate total number of sheep in the state unbiasedly. Also, work out the standard error of your estimate.

11.10 A soap factory is planning to use *neem (Azadirachta indica)* oil in manufacturing a new brand of toilet soap. For extracting oil, the *neem* seeds are to be collected at 100 centers established for the purpose. It is felt that the persons from villages, falling within a radius of 20 km of a *neem* seed collection center, will find it remunerative to sell *neem* seeds to the collection centers. For getting an idea about the *neem* seed supply position, the factory wants to estimate the total number of *neem* trees in the procurement area for all the 100 collection centers. For this purpose, an SRS without replacement sample of 10 centers was selected. Using same sampling procedure, about 10 percent of villages falling within the 20 km radius of the sample centers were selected as second stage units. Number of *neem* trees in the selected villages are given below along with the values of M_i and m_i.

Center	M_i	m_i	No. of *neem* trees					
1	30	3	40	61	76			
2	44	4	88	25	49	33		
3	56	6	75	32	56	44	37	55
4	36	4	42	25	34	39		
5	27	3	33	47	29			
6	51	5	55	43	61	40	36	
7	42	4	27	42	36	58		
8	30	3	60	46	39			
9	28	3	51	37	42			
10	33	3	64	34	56			

Estimate the total number of *neem* trees falling in the procurement area for all the 100 centers by using an appropriate estimator. Also, place confidence limits on this total.

11.11 Give an unbiased estimator for population proportion for a two-stage sampling design. Also, write expressions for its variance and estimator of variance.

11.12 The objective of a study was to estimate the proportion of liquor shops in a state selling spurious liquor. The state consists of 117 development blocks and has 8068 liquor shops. A WOR simple random sample of 10 blocks was selected for this study. About 10 percent of liquor shops were selected from the sampled development blocks. The liquor in the selected shops was examined. The collected information, along with the total number of liquor shops in the sample blocks (M_i) and the number of shops selected (m_i), is given as follows.

Block	:	1	2	3	4	5	6	7	8	9	10
M_i	:	104	96	60	102	101	91	87	90	76	89
m_i	:	10	10	6	10	10	9	9	9	8	9
Shops selling spurious liquor	:	4	5	4	6	7	5	4	3	5	6

Estimate unbiasedly the proportion of shops selling spurious liquor in the state. Also, work out the standard error of your estimate, and place confidence limits on the value of the parameter under study.

11.13 There are 76 senior citizen homes (SCH's) in a state which care for the old persons who are unable to stay at their native homes due to some reasons. It is felt that once a person is admitted to an SCH and stays there for sometime, he/she does not like to return to his/her native home even if the circumstances that had compelled him/her to live in the SCH have changed and are favorable for his/her return. For verifying this belief, a WOR simple random sample of 9 SCH's was drawn. About 10 percent of the persons living in the sample homes were interviewed. The information, so collected, is given in the following table, where 1 indicates that the person is not willing to return to his/her native home, and 0 otherwise. The total number of old persons staying in a sample SCH (M_i), and the number of persons selected for interview (m_i), are also given in the table.

SCH	M_i	m_i	Scores										
1	110	11	1	1	0	1	1	1	0	0	1	1	1
2	36	4	1	1	1	0							
3	70	7	0	0	1	1	1	1	1				
4	82	8	1	1	1	0	1	1	0	1			
5	66	7	1	1	0	0	0	1	1				
6	97	10	1	1	1	0	1	1	1	1	1	0	
7	85	9	0	1	1	1	1	1	0	1	1		
8	67	7	1	1	1	1	1	0	1				
9	56	6	1	0	0	0	1	1					

Using estimator p_{m2} in (11.33), estimate the proportion of persons not willing to return to their native homes. Also, work out the standard error of the estimate, and place confidence limits on the population value.

11.14 A district transport office is concerned about the proportion of tractors plying without valid documents in the district. The total number of tractors operating in the district is not exactly known. This is because some of the tractors now plying in the district might have been registered in some other districts/states, whereas some other tractors registered with the office might have been sold outside the district. It was, therefore, thought appropriate to use two-stage sampling design to estimate this proportion. For this purpose, 16 villages out of the total of 400 villages in the district, were selected using SRS without replacement. About 20 percent of tractors in the sample villages were selected using same sampling scheme. Documents of the selected tractors were then examined. The information thus collected is presented below, along with the total number of tractors in sample villages (M_i) and number of tractors examined (m_i) for validity of documents. The abbreviation TWD in the following table stands for tractors without valid documents.

Village	M_i	m_i	TWD	Village	M_i	m_i	TWD
1	40	8	6	9	36	7	5
2	13	3	3	10	27	5	3
3	36	7	5	11	15	3	2
4	48	10	7	12	62	12	9
5	32	6	4	13	33	7	5
6	42	8	7	14	44	9	6
7	60	12	8	15	38	8	6
8	28	6	4	16	55	11	7

Using estimator p_{m3} in (11.40), estimate the proportion of tractors in the district plying without valid documents. Also, place confidence limits on this proportion.

11.15 A with replacement PPS sample of 10 wards out of a total of 63 wards comprising a town was selected, the size measure being the number of households in a ward available from the population census records. From the sample wards, about 5 percent of households were selected through WOR simple random sampling. The per month expenses (in rupees) on vegetables were recorded for the selected households. This information is given in the table below along with the values of M_i and m_i which carry their usual meaning. The total number of households in all the 63 wards is 12800.

Ward	M_i	m_i	Total monthly expenses for m_i households
1	160	8	1600
2	240	12	2000
3	218	11	2350
4	148	7	1760
5	276	14	3500
6	238	12	3320
7	196	10	3035
8	256	13	2600
9	217	11	3730
10	177	9	3100

What is the estimated per month expenditure on vegetables for a household ? Work out the standard error of your estimate, and place confidence limits on the population value.