CHAPTER 10

Cluster Sampling

10.1 INTRODUCTION

Let us consider a situation where a study is to be carried out regarding the indebtedness of farmers of a particular region. For this purpose, a sample of farmers is to be selected. In case, the list of all the farmers (frame) in the region is not available, a simple random sample or a systematic sample of farmers, can not be selected. Even if the frame of all farmers in the region was available, a simple random sample of farmers will result in the sampled units (farmers) being scattered all over the region. This will require a good amount of travel to reach all the selected farmers for collecting information about the indebtedness and will, therefore, involve a formidable amount of travel expenditure.

On the other hand, the list of all the villages of the region is usually available and the characteristics of farmers in one village do not differ appreciably from the farmers of the other village. Therefore, if a simple random sample of villages, which could be considered as groups or *clusters* of farmers, is selected and all the farmers in the selected villages are enumerated, then a considerably reduced number of villages will account for the given number of farmers to be selected in the sample. In the new set-up, the total travel expenditure will be reduced to a great extent as the investigator will be required to travel between the few sample villages only. Besides, contacting of various farmers in any selected village involves comparatively very small cost. Also, since the characteristics of farmers in one village of the region may not be much different from the characteristics of farmers of the other village, not much loss of total information in the sample is expected.

All these considerations point to the need of selecting groups (clusters) of units together, and examine all the units contained in the selected clusters. The importance of the idea is that once a cluster has been reached, the cost of surveying units within the cluster is negligible.

Definition 10.1 The *cluster sampling* consists of forming suitable clusters of contiguous population units, and surveying all the units in a sample of clusters selected according to an appropriate sampling scheme.

For a given total number of units in the sample, the cluster sampling is usually less efficient than sampling of individual units as the latter is likely to provide a better cross section of the population units than the former. This is because of the tendency of units in a cluster to be similar. Also, the efficiency of cluster sampling is likely to decrease with increase in cluster size. However, it is operationally convenient and economical than sampling of individual units. In many practical situations, the loss in efficiency from the view point of sampling variance is likely to be balanced by the reduction in cost. Hence, because of its operational convenience and possible reduction in cost, the survey tasks in many situations are facilitated by using nonoverlapping and collectively exhaustive clusters of units.

The clusters are usually formed by grouping neighboring units, or units which can be conveniently surveyed together. The construction of clusters, however, differs from the optimal construction of strata. Strata are to be as homogeneous as possible within themselves and differ as much as possible from one another with respect to the characteristic under study, and the units within a stratum need not be geographically contiguous. Clusters, on the other hand, should be as heterogeneous as possible within and as alike as possible between themselves. In such situations, cluster sampling is likely to be more efficient than the usual simple random sampling of same number of units from the population. Once appropriate clusters have been specified, a frame that lists all clusters in the population must be prepared. Various sampling procedures, viz., simple random sampling, varying probability sampling, stratified sampling, or systematic sampling, can be applied to cluster sampling by treating the clusters as sampling units. The expressions for the estimator, its variance, and estimator of variance can, therefore, be written in a straightforward manner. However, in this chapter, we shall only consider selection of clusters using simple random sampling and PPS with replacement sampling, and illustrate the steps involved in calculations of estimates of mean, total, and proportion.

10.2 NOTATIONS

In order to facilitate the understanding of the text, we first acquaint the reader with the notations to be used in the chapter. Let

N = number of clusters in the population

n = number of clusters in the sample

M_i = number of units in the i-th cluster of the population

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

$\overline{M} = M_o/N$ = average number of units per cluster in the population

Y_{ij} = value of the character under study for the j-th unit in the i-th cluster, $j = 1, 2, ..., M_i$; $i = 1, 2, ..., N$

$Y_{i.} = \sum_{j=1}^{M_i} Y_{ij}$ = i - th cluster total

$Y_{..} = \sum_{i=1}^{N} Y_{i.}$ = total of y - values for all the M_o units in the population

$$\overline{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij} = \text{per unit i - th cluster mean}$$

$$y_{i.} = \sum_{j=1}^{M_i} y_{ij} = \text{i - th sample cluster total}$$

$$\overline{y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} = \text{per unit i - th sample cluster mean}$$

$$\overline{y}_c = \frac{1}{n} \sum_{i=1}^{n} y_{i.} = \text{mean per cluster in the sample}$$

$$\overline{Y}_N = \frac{1}{N} \sum_{i=1}^{N} \overline{Y}_i = \text{mean of cluster means in the population}$$

$$\overline{Y} = \frac{1}{M_o} \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij} = \text{mean per unit of the population}$$

$$\overline{Y}_c = Y_{..}/N = \text{population mean per cluster}$$

10.3 ESTIMATION OF MEAN USING SIMPLE RANDOM SAMPLING

In practice, usually the clusters are of unequal sizes. For instance, households which are groups of persons, villages which are groups of households, fish catching centers which consist of groups of boats, could be considered as clusters for the purpose of sampling. In this chapter, we shall consider three estimators of population mean and total. The important results concerning these estimators are mentioned, assuming that a WOR simple random sample of n clusters has been drawn from N clusters, and all the units of the sample clusters are observed for the study variable. The results related to WR case can be obtained as particular cases from the results presented below for WOR sampling. We first consider the problem of estimating population mean per unit.

10.3.1 *Estimator 1*

Unbiased estimator of population mean when M_o is known :

$$\left.\begin{aligned}\overline{y}_{c1} &= \frac{N}{nM_o} \sum_{i=1}^{n} M_i \overline{y}_i \\ &= \frac{1}{\overline{M}n} \sum_{i=1}^{n} y_{i.}\end{aligned}\right] \qquad (10.1)$$

Variance of estimator $\overline{y}_{c1}$:

$$V(\overline{y}_{c1}) = \left(\frac{N-n}{Nn\overline{M}^2}\right) \frac{1}{N-1} \sum_{i=1}^{N} (Y_{i.} - \overline{Y}_c)^2 \qquad (10.2)$$

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

$$\left.\begin{aligned} v(\bar{y}_{cl}) &= \left(\frac{N-n}{NnM̄^2}\right)\frac{1}{n-1}\sum_{i=1}^{n}(y_{i.} - \bar{M}\bar{y}_{cl})^2 \\ &= \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{1}{n-1}\left[\sum_{i=1}^{n} y_{i.}^2 - n(\bar{M}\bar{y}_{cl})^2\right] \end{aligned}\right] \qquad (10.3)$$

If the clusters are selected using WR sampling, then fpc = (N-n)/(N-1) in relation (10.2) and the sampling fraction f = n/N in (10.3) are taken as 1 and 0 respectively to get the corresponding results for the with replacement case.

Example 10.1

The recommended dose of nitrogen for wheat crop is 120 kg per hectare. A survey project was undertaken by the Department of Agriculture with a view to estimate the amount of nitrogen actually applied by the farmers. For this purpose, 12 villages from a population of 170 villages of a development block were selected using equal probabilities WOR sampling, and the information regarding the nitrogen use was collected from all the farmers in the selected villages. The data collected are presented in table 10.1. The total number of farmers in these 170 villages is available from the *patwari's* record as 2890. Estimate the average amount of nitrogen used in practice by a farmer. Also, obtain standard error of the estimate, and place confidence limits on the population mean.

Table 10.1 Per hectare nitrogen (in kg) applied to wheat crop by farmers

Village	M_i	Nitrogen applied (in kg) by a farmer									$y_{i.}$
1	15	105	128	130	108	135	122	120	138	126	1843
		117	125	126	123	118	122				
2	18	135	128	105	130	120	125	114	128	121	2206
		109	128	122	129	112	133	117	119	131	
3	25	124	118	128	106	132	121	126	108	136	3085
		121	128	125	136	128	121	127	122	113	
		117	132	128	125	130	109	124			
4	21	108	116	111	129	119	137	129	121	118	2582
		126	131	128	134	125	112	121	116	114	
		129	127	131							
5	11	114	105	126	132	116	125	104	121	132	1292
		106	111								
6	13	128	116	132	136	121	122	129	123	127	1627
		118	134	126	115						
7	22	103	118	107	128	132	136	124	129	130	2686
		134	108	106	117	129	113	118	126	127	
		129	119	125	128						

Table 10.1 continued...

Village	M_i	Nitrogen applied (in kg) by a farmer									$y_{i.}$
8	12	109	121	114	128	133	135	114	128	107	1471
		125	126	131							
9	10	119	128	117	131	105	128	136	113	127	1234
		130									
10	20	130	127	116	128	114	120	127	123	134	2449
		122	126	121	117	125	129	122	113	111	
		126	118								
11	10	126	117	124	121	131	133	126	120	128	1242
		116									
12	16	124	121	127	119	120	123	128	117	121	1935
		93	115	120	124	121	130	132			

Solution

Here we have N = 170, M_0 = 2890, and n = 12. It gives

$$\overline{M} = \frac{M_0}{N} = \frac{2890}{170} = 17$$

As the sample cluster totals $y_{i.}$ will be required for computing the estimates, these are worked out below and are presented in the last column of the table 10.1 above.

$$\begin{array}{lcl}
\text{Cluster 1} & : & y_{1.} = 105 + 128 + ... + 122 = 1843 \\
\text{Cluster 2} & : & y_{2.} = 135 + 128 + ... + 131 = 2206 \\
\cdot & & \cdot \\
\cdot & & \cdot \\
\cdot & & \cdot \\
\text{Cluster 12} & : & y_{12.} = 124 + 121 + ... + 132 = 1935
\end{array}$$

Estimate of the average amount of nitrogen used per hectare, by a farmer, follows from (10.1) as

$$\begin{aligned}
\overline{y}_{cl} &= \frac{1}{\overline{M}n} \sum_{i=1}^{n} y_{i.} \\
&= \frac{1}{(17)\,(12)} (1843 + 2206 + ... + 1935) \\
&= \frac{23652}{(17)\,(12)} \\
&= 115.941
\end{aligned}$$

We then work out the estimate of variance using (10.3). Thus,

$$v(\bar{y}_{c1}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{1}{n-1}\sum_{i=1}^{n}(y_{i.} - \bar{M}\,\bar{y}_{c1})^2$$

where

$$\bar{M}\,\bar{y}_{c1} = (17)\,(115.941) = 1970.997$$

Hence,

$$v(\bar{y}_{c1}) = \left(\frac{170-12}{(170)\,(12)\,(17)^2}\right)\frac{1}{11}\,[(1843-1970.997)^2 + (2206-1970.997)^2$$

$$+ \ldots + (1935-1970.997)^2]$$

$$= \left(\frac{170-12}{(170)\,(12)\,(17)^2}\right)\frac{1}{11}\,[(1843)^2 + (2206)^2 + \ldots + (1935)^2$$

$$-\ 12(1970.997)^2]$$

$$= \frac{(170-12)\,(4331060)}{(170)\,(12)\,(17)^2(11)}$$

$$= 105.519$$

Using above calculated estimate of variance, the standard error of mean will be

$$se(\bar{y}_{c1}) = \sqrt{105.519}$$

$$= 10.272$$

Following (2.8), the required confidence limits for population mean are obtained as

$$\bar{y}_{c1} \pm 2\sqrt{v(\bar{y}_{c1})}$$

$$= 115.941 \pm 20.544$$

$$= 95.397,\ 136.485$$

The above confidence limits reasonably ensure that per hectare average dose of nitrogen used by a farmer in the target population is likely to be within the range 95.397 to 136.485 kg. ■

10.3.2 *Estimator 2*

The estimator $\bar{y}_{c1}$ in (10.1) for population mean assumes the knowledge of M_o. When M_o is not known and the values of M_i are known only for the sample clusters, then $\bar{Y}$ can be estimated by using the estimator $\bar{y}_{c2}$.

Estimator of population mean which does not depend on M_o:

$$\bar{y}_{c2} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i \qquad (10.4)$$

Bias of the estimator $\bar{y}_{c2}$:

$$B(\bar{y}_{c2}) = - \frac{1}{\bar{\bar{M}}} \text{Cov}(\bar{y}_i, M_i) \qquad (10.5)$$

Variance of estimator $\bar{y}_{c2}$:

$$V(\bar{y}_{c2}) = \left(\frac{N-n}{Nn}\right) \frac{1}{N-1} \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y}_N)^2 \qquad (10.6)$$

Estimator of variance $V(\bar{y}_{c2})$:

$$\left.\begin{aligned} v(\bar{y}_{c2}) &= \left(\frac{N-n}{Nn}\right) \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{c2})^2 \\ &= \left(\frac{N-n}{Nn}\right) \frac{1}{n-1} \left(\sum_{i=1}^{n} \bar{y}_i^2 - n\bar{y}_{c2}^2\right) \end{aligned}\right] \qquad (10.7)$$

The bias for the estimator $\bar{y}_{c2}$, given by (10.5), is expected to be small when the cluster means $\bar{Y}_i$ and sizes M_i are not highly correlated. In such a case, it is advisable to use this estimator since the variance of the estimator $\bar{y}_{c2}$ is likely to be less than the variance $V(\bar{y}_{c1})$ given in (10.2). The bias of this estimator disappears if the cluster sizes M_1, M_2, ..., M_N are equal.

10.3.3 *Estimator 3*

An alternative estimator of population mean $\bar{Y}$ for the situation where M_i's are known only for sample clusters (irrespective of whether M_o is known or not), is the ratio type estimator. This estimator is also biased but the bias decreases with increase in n.

Estimator of population mean which does not depend on M_o:

$$\bar{y}_{c3} = \frac{\sum_{i=1}^{n} y_{i.}}{\sum_{i=1}^{n} M_i} \qquad (10.8)$$

Approximate bias of estimator $\bar{y}_{c3}$:

$$B(\bar{y}_{c3}) = \left(\frac{N-n}{Nn}\right) \frac{1}{\bar{M}^2} (\bar{Y}S_m^2 - S_{my}) \qquad (10.9)$$

where S_m^2 and S_{my} are defined in (7.2) with cluster size m replacing x.

Approximate variance of estimator $\bar{y}_{c3}$:

$$V(\bar{y}_{c3}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{1}{N-1}\sum_{i=1}^{N}(Y_{i.}-\bar{Y}M_i)^2 \qquad (10.10)$$

Estimator of variance V($\bar{y}_{c3}$):

$$v(\bar{y}_{c3}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{1}{n-1}\sum_{i=1}^{n}(y_{i.}-M_i\bar{y}_{c3})^2 \qquad (10.11)$$

$\bar{M}$ in (10.11) above can be replaced by $\hat{\bar{M}} = \frac{1}{n}\sum_{i=1}^{n} M_i$, if it is not already known.

The bias in the estimator $\bar{y}_{c3}$ becomes zero if the cluster sizes M_1, M_2, ..., M_N are equal. The estimator $\bar{y}_{c3}$ is also expected to be more efficient than the estimator $\bar{y}_{c1}$ when M_i and $y_{i.}$ are highly positively correlated.

Example 10.2

A state government wanted to estimate the extent of tax evasion, per passenger, by the private bus owners on a certain route. Being the busy route, it was decided to check the buses at random. The total number of buses that leave the terminal daily is 80. The buses were serially numbered depending on the time of their departure. Fifteen buses were then selected with SRS without replacement. The tickets with all the passengers of the selected buses were examined enroute, and the amount of tax evasion was recorded. The total of passenger tax evaded for each sampled bus was then computed, and is given in table 10.2 along with the total number of passengers in the bus. Estimate the average tax evaded per passenger by the private bus operators, and place a confidence interval on the population average.

Table 10.2 Tax evaded (in rupees) per sampled bus along with cluster mean

Bus	Passengers (M_i)	Tax evaded ($y_{i.}$)	Cluster (bus) mean ($\bar{y}_i$)
1	60	118.70	1.98
2	70	148.30	2.12
3	65	140.10	2.16
4	52	98.40	1.89
5	72	109.50	1.52
6	48	72.05	1.50
7	54	100.20	1.86
8	60	115.10	1.92
9	43	108.70	2.53

Table 10.2 continued ...

Bus	Passengers (M_i)	Tax evaded $(y_{i.})$	Cluster (bus) mean $(\bar{y}_i)$
10	69	135.45	1.96
11	58	117.30	2.02
12	74	150.70	2.04
13	55	126.40	2.30
14	69	95.30	1.38
15	66	111.65	1.69
Total	915	1747.85	28.87

Solution

In this problem, we have N = 80, n = 15, and M_0 is not known. Although the choice between estimators 2 and 3 depends on the value of the correlation coefficient as mentioned earlier, but for the sake of illustration we demonstrate the use of both the estimators.

Use of estimator 2. The estimate of the tax evaded per passenger is obtained by using (10.4) as

$$\bar{y}_{c2} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$$

$$= \frac{1}{15}(1.98 + 2.12 + \dots + 1.69)$$

$$= \frac{28.87}{15}$$

$$= 1.92$$

We then compute the estimate of variance using (10.7).

$$v(\bar{y}_{c2}) = \left(\frac{N-n}{Nn}\right)\frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{y}_{c2})^2$$

$$= \left(\frac{80-15}{(80)(15)}\right)\frac{1}{14}[(1.98-1.92)^2 + (2.12-1.92)^2 + \dots + (1.69-1.92)^2]$$

$$= \left(\frac{80-15}{(80)(15)}\right)\frac{1}{14}[(1.98)^2 + (2.12)^2 + \dots + (1.69)^2 - (15)(1.92)^2]$$

$$= \frac{(80-15)(1.5979)}{(80)(15)(14)}$$

$$= .006182$$

The confidence interval for the population average can be derived from

$$\begin{aligned}&\bar{y}_{c2} \pm 2\sqrt{v(\bar{y}_{c2})}\\ &= 1.92 \pm 2\sqrt{.006182}\\ &= 1.92 \pm .16\\ &= 1.76,\ 2.08\end{aligned}$$

The confidence limits computed above, indicate that the daily per passenger evasion of tax by the population of private bus owners is likely to fall in the closed interval [1.76, 2.08] rupees.

Use of estimator 3. The estimate of per passenger tax evaded by the private bus operators is given by

$$\bar{y}_{c3} = \frac{\sum\limits_{i=1}^{n} y_{i.}}{\sum\limits_{i=1}^{n} M_i}$$

Substituting the values from table 10.2, one gets

$$\bar{y}_{c3} = \frac{1747.85}{915} = 1.91$$

For computing the estimate of variance, we use expression (10.11). Thus,

$$v(\bar{y}_{c3}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{1}{n-1}\sum_{i=1}^{n}(y_{i.} - M_i\,\bar{y}_{c3})^2$$

Since $\bar{M}$ is unknown, we use its estimate $\hat{\bar{M}}$, where

$$\hat{\bar{M}} = \frac{1}{n}\sum_{i=1}^{n} M_i = \frac{915}{15} = 61$$

Then,

$$\begin{aligned}v(\bar{y}_{c3}) &= \left(\frac{80-15}{(80)\,(15)\,(61)^2}\right)\frac{1}{14}\,[\{118.70-(60)\,(1.91)\}^2\\ &\quad + \{148.30-(70)\,(1.91)\}^2 + \dots + \{111.65-(66)\,(1.91)\}^2]\\ &= \frac{(80-15)\,(4509.041)}{(80)\,(15)\,(61)^2\,(14)}\\ &= .004688\end{aligned}$$

We now compute confidence limits for daily per passenger tax evasion by all the private bus operators on the route under consideration. These we find as

$$
\begin{aligned}
& \bar{y}_{c3} \pm 2\sqrt{v(\bar{y}_{c3})} \\
&= 1.91 \pm 2\sqrt{.004688} \\
&= 1.91 \pm .14 \\
&= 1.77, 2.05
\end{aligned}
$$

These confidence limits are very close to those obtained earlier by using estimator 2. ■

10.4 ESTIMATION OF TOTAL USING SIMPLE RANDOM SAMPLING

An estimator of population total can be easily obtained by multiplying any one of the corresponding estimators of mean given in (10.1), (10.4), and (10.8) by M_o.

Estimators of population total Y:

$$\hat{Y}_{c1} = \frac{N}{n} \sum_{i=1}^{n} y_{i.} \tag{10.12}$$

$$\hat{Y}_{c2} = \frac{M_o}{n} \sum_{i=1}^{n} \bar{y}_i \tag{10.13}$$

$$\hat{Y}_{c3} = \frac{M_o \sum_{i=1}^{n} y_{i.}}{\sum_{i=1}^{n} M_i} \tag{10.14}$$

Expressions for variances and their estimators for the above estimators of population total, can be easily obtained by multiplying their counterparts for mean by M_o^2.

The estimator $\hat{Y}_{c1}$ in (10.12) can be used even when M_o is not known while the estimators $\hat{Y}_{c2}$ or $\hat{Y}_{c3}$ can be used only when M_o, or equivalently $\bar{M}$, is known. If the correlation between the cluster means and cluster size is low, the estimator $\hat{Y}_{c2}$ may be preferred. However, in case of large samples when correlation between the cluster total and cluster size is positive and high, use of estimator $\hat{Y}_{c3}$ is advised.

Example 10.3

Along the sea coast of an Indian state, there are 120 small villages. Some of the residents of these villages resort to fishing for their livelihood. The list of these villages is available. However, no information is available about the number of families (M_i) involved in the said profession in these villages. For estimating the total catch of fish by the villagers, 16 villages were selected using SRS without replacement. The information collected from all the families of sample villages about the catch of fish on a particular day, is given in table 10.3. Estimate total catch of fish on this day for the entire population of 120 villages. Also, build up confidence interval for the population total.

Table 10.3 Catch of fish (in quintals) per family for the selected villages

Village	M_i	Catch of fish by families											$y_{i.}$
1	11	6.8	5.0	2.4	7.5	3.2	4.4	4.0	3.0	6.8	5.6	9.2	57.9
2	7	4.3	5.9	6.0	1.7	2.3	7.3	2.6					30.1
3	8	2.0	4.1	6.6	7.0	8.2	9.9	1.4	1.2				40.4
4	6	5.3	6.8	2.9	3.3	4.8	1.8						24.9
5	4	1.8	10.5	1.0	2.7								16.0
6	10	8.4	1.5	6.7	8.8	5.6	2.8	1.2	4.5	6.8	8.1		54.4
7	7	5.5	2.8	4.6	1.0	10.9	6.7	3.8					35.3
8	8	5.4	6.4	3.3	8.1	7.4	6.5	1.9	2.0				41.0
9	7	6.0	7.1	2.1	1.8	6.7	8.3	2.9					34.9
10	6	2.8	3.4	6.9	1.1	6.6	4.8						25.6
11	3	5.6	6.8	1.6									14.0
12	5	9.9	6.4	1.2	2.0	3.6							23.1
13	9	8.5	2.0	1.8	4.9	6.7	5.2	7.7	5.9	1.6			44.3
14	8	1.2	4.5	6.8	1.0	2.3	3.7	6.9	8.1				34.5
15	7	3.8	6.4	7.6	2.5	6.1	3.5	1.9					31.8
16	8	2.6	10.0	6.3	9.8	1.0	1.8	7.8	2.7				42.0
Total													550.2

Solution

First, we work out sample cluster (which is village in this case) totals as

Cluster 1 : $y_{1.} = 6.8 + 5.0 + \ldots + 9.2 = 57.9$
Cluster 2 : $y_{2.} = 4.3 + 5.9 + \ldots + 2.6 = 30.1$

. .
. .
. .

Cluster 16 : $y_{16.} = 2.6 + 10.0 + \ldots + 2.7 = 42.0$

The cluster totals obtained this way, are presented in table 10.3.

Since M_0 is not known, the estimate of total catch of fish is obtained by using (10.12). Thus,

$$\hat{Y}_{cl} = \frac{N}{n}\sum_{i=1}^{n} y_{i.}$$

$$= \frac{(120)\,(550.2)}{16}$$

$$= 4126.50$$

The estimate of variance of $\hat{Y}_{cl}$ can be worked out from the estimate of variance of mean given in (10.3), after multiplying it by M_0^2. This means

$$v(\hat{Y}_{cl}) = \frac{N(N-n)}{n\,(n-1)}\sum_{i=1}^{n}(y_{i.} - \overline{M}\,\bar{y}_{cl})^2$$

Also,

$$\overline{M}\,\bar{y}_{cl} = \frac{\hat{Y}_{cl}}{N} = \frac{4126.50}{120} = 34.3875$$

Hence,

$$v(\hat{Y}_{cl}) = \frac{(120)\,(120-16)}{(16)\,(15)}\,[(57.9-34.3875)^2 + (30.1-34.3875)^2$$

$$+ \ldots + (42.0-34.3875)^2]$$

$$= \frac{(120)\,(120-16)}{(16)\,(15)}\,[(57.9)^2 + (30.1)^2 + \ldots + (42.0)^2 - 16(34.3875)^2]$$

$$= \frac{(120)\,(120-16)\,(2263.9974)}{(16)\,(15)}$$

$$= 117727.86$$

The required confidence interval for population total is given by

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

$$= 4126.50 \pm 2\sqrt{117727.86}$$

$$= 4126.50 \pm 686.23$$

$$= 3440.27,\ 4812.73$$

This means that the total catch of fish for this day, for all the 120 villages under study, is likely to take a value in the interval 3440.27 to 4812.73 quintals, with confidence coefficient as .95. ■

Example 10.4

A state is planning to set up a small spinning mill in a hilly area. Before finalizing the capacity, layout, etc., the administration thinks it appropriate to have information regarding the number of unemployed males/females that are educated up to at least 5th grade but have not attained the age of 45 years, in the villages falling within a radius of 15 km. The number of such villages is 60, and the total number of households, comprising these villages, is known to be 560. Twelve villages were selected using SRS without replacement method. The information collected from all the households in the sample villages, is given below in table 10.4.

Table 10.4 Unemployed persons below 45 years and educated up to at least 5th grade

Village	M_i	Number of unemployed males/females														$y_{i.}$	$\bar{y}_i$
1	9	1	2	0	1	2	1	1	3	2						13	1.4444
2	6	3	4	0	1	1	2									11	1.8333
3	7	2	4	1	2	2	1	1								13	1.8571
4	14	1	0	0	2	4	0	2	3	1	2	0	3	5	2	25	1.7857
5	8	2	1	0	1	1	0	4	2							11	1.3750
6	9	2	4	3	1	0	2	3	2	0						17	1.8889
7	8	1	1	5	0	2	1	2	1							13	1.6250
8	10	3	1	0	3	0	4	1	0	0	4					16	1.6000
9	11	2	0	3	4	0	1	0	2	3	1	0				16	1.4545
10	11	0	2	1	5	3	2	0	2	3	1	1				20	1.8182
11	9	0	1	3	5	0	3	1	1	1						15	1.6667
12	12	4	1	2	1	2	0	2	4	3	2	1	1			23	1.9167
Total	114															193	20.2655

Estimate the total number of unemployed persons in question, and construct confidence interval for this population total.

Solution

In this problem, the total number of households in the population of 60 villages is known, that is, $M_0 = 560$. Also, $N = 60$ and $n = 12$. Since M_0 is known, we can use both the estimators defined in (10.13) and (10.14). Although the magnitudes of the correlation coefficients between the pairs of variables ($\bar{y}_i$, M_i) and ($y_{i.}$, M_i) are to decide as to which of these two estimators is to be preferred in practice, we shall consider both for the purpose of illustration.

Use of estimator 2. As in the previous examples, we first work out cluster totals and cluster means. Thus,

$$
\begin{array}{llll}
y_{1.} & = 1 + 2 + ... + 2 = 13, & \bar{y}_1 & = 13/9 = 1.4444 \\
y_{2.} & = 3 + 4 + ... + 2 = 11, & \bar{y}_2 & = 11/6 = 1.8333 \\
\cdot & & \cdot & \\
\cdot & & \cdot & \\
\cdot & & \cdot & \\
y_{12.} & = 4 + 1 + ... + 1 = 23, & \bar{y}_{12} & = 23/12 = 1.9167
\end{array}
$$

The cluster totals and means, so calculated, are presented in table 10.4.

The estimate of population total from (10.13) is given by

$$\hat{Y}_{c2} = \frac{M_o}{n} \sum_{i=1}^{n} \bar{y}_i$$

Using $\Sigma \bar{y}_i$, i = 1, 2, ..., n, computed in table (10.4), one gets the estimate of the total number of unemployed persons with required qualifications as

$$\hat{Y}_{c2} = \frac{560}{12} (20.2655) = 945.72 \approx 946$$

So far as the estimate of variance $V(\hat{Y}_{c2})$ is concerned, it can be computed from (10.7), after multiplying it by M_o^2. Thus we get

$$
\begin{aligned}
v(\hat{Y}_{c2}) &= \frac{M_o^2 (N-n)}{Nn(n-1)} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{c2})^2 \\
&= \frac{M_o^2 (N-n)}{Nn(n-1)} (\sum_{i=1}^{n} \bar{y}_i^2 - n\bar{y}_{c2}^2)
\end{aligned}
$$

where

$$\bar{y}_{c2} = \frac{\hat{Y}_{c2}}{M_o} = \frac{945.72}{560} = 1.6888$$

This yields

$$
\begin{aligned}
v(\hat{Y}_{c2}) &= \frac{(560)^2 (60-12)}{(60)(12)(11)} [(1.4444 - 1.6888)^2 + (1.8333-1.6888)^2 \\
&\quad + ... + (1.9167-1.6888)^2] \\
&= \frac{(560)^2 (60-12)}{(60)(12)(11)} [(1.4444)^2 + (1.8333)^2 + ... + (1.9167)^2 \\
&\quad -12(1.6888)^2] = 746.03
\end{aligned}
$$

Confidence interval for the total unemployed work force is then given by

$$\begin{aligned}
&\hat{Y}_{c2} \pm 2\sqrt{v(\hat{Y}_{c2})}\\
&= 945.72 \pm 2\sqrt{746.03}\\
&= 945.72 \pm 54.63\\
&= 891.09,\ 1000.35\\
&\approx 891,\ 1000
\end{aligned}$$

One can, therefore, say with probability approximately .95, that the total number of unemployed persons of the desired kind ranges from 891 to 1000.

Use of estimator 3. The estimate of the total number of unemployed persons with desired qualifications is now given by

$$\hat{Y}_{c3} = \frac{M_o \sum_{i=1}^{n} y_{i.}}{\sum_{i=1}^{n} M_i} = \frac{(560)\,(193)}{114} = 948.07 \approx 948$$

Estimate of variance $V(\hat{Y}_{c3})$ is worked out by multiplying (10.11) by M_o^2. For this, we calculate

$$\bar{y}_{c3} = \frac{\hat{Y}_{c3}}{M_o} = \frac{948.07}{560} = 1.6930$$

Then,

$$\begin{aligned}
v(\hat{Y}_{c3}) &= \frac{N(N-n)}{n\,(n-1)} \sum_{i=1}^{n} (y_{i.} - M_i\,\bar{y}_{c3})^2\\
&= \frac{(60)\,(60-12)}{(12)\,(11)} [\{13-9(1.6930)\}^2 + \{11-6(1.6930)\}^2\\
&\quad + \ldots + \{23-12(1.6930)\}^2]\\
&= \frac{(60)\,(48)\,(35.4954)}{(12)\,(11)}\\
&= 774.4451
\end{aligned}$$

As usual, the confidence interval for the total number of unemployed persons with specified qualifications is given by

$$\begin{aligned}
&\hat{Y}_{c3} \pm 2\sqrt{v(\hat{Y}_{c3})}\\
&= 948.07 \pm 2\sqrt{774.4451}\\
&= 948.07 \pm 55.66\\
&= 892.41,\ 1003.73\\
&\approx 892,\ 1004
\end{aligned}$$

Thus, the confidence intervals yielded by the two estimators are quite similar. ■

10.5. RELATIVE EFFICIENCY OF CLUSTER SAMPLING

In this section, we consider the relative efficiency aspect of cluster sampling. For this purpose, we assume the simplest situation where all the clusters in the population are of equal size, that is, $M_i = M$ for $i = 1, 2, ..., N$. In this case, the estimators of population mean given in (10.1) and (10.4) become identical, and hence, expressions for their variances and estimators of variances are also same. Therefore, whatever we conclude about estimator (10.1) will also hold for estimator in (10.4). In case of cluster sampling considered in this section, we select a sample of nM units in the form of n clusters each consisting of M units. Thus, if the same number of units were selected from the population of NM units by SRS without replacement, the simple mean estimator $\bar{y}$ and its variance will be given by

$$\bar{y} = \frac{1}{nM} \sum_{i=1}^{nM} y_i \tag{10.15}$$

$$V(\bar{y}) = \left(\frac{1}{nM} - \frac{1}{NM}\right) S^2$$

$$= \left(\frac{N-n}{NnM}\right) \frac{1}{NM-1} \left(\sum_{i=1}^{NM} Y_i^2 - NM\bar{Y}^2\right) \tag{10.16}$$

The relative efficiency of the estimator $\bar{y}_{cl}$ in (10.1), in relation to the simple mean estimator $\bar{y}$, will, therefore, be given by

$$RE = \frac{V(\bar{y})}{V(\bar{y}_{cl})} \tag{10.17}$$

where the variance $V(\bar{y}_{cl})$ is available in (10.2).

The relative efficiency, defined in (10.17), involves values of study variable for all population units. In practice, however, the investigator has only the sample observations on n clusters of M units each. Thus, he can only estimate the relative efficiency from the sample observations. For this, we shall need the estimates of two variances involved in (10.17). An unbiased estimator of $V(\bar{y})$, from a cluster sample, is given by

$$v_c(\bar{y}) = \frac{N-n}{(NM-1)\,n} \left[\frac{1}{nM} \sum_{i=1}^{n} \sum_{j=1}^{M} y_{ij}^2 + v(\bar{y}_{cl}) - \bar{y}_{cl}^2\right] \tag{10.18}$$

The estimate of relative efficiency in (10.17) will then be given as in (10.19).

The estimated RE of estimator $\bar{y}_{cl}$ with respect to the usual estimator $\bar{y}$, from a cluster sample :

$$RE = \frac{v_c(\bar{y})}{v(\bar{y}_{cl})} \tag{10.19}$$

where $v(\bar{y}_{cl})$ and $v_c(\bar{y})$ are given in (10.3) and (10.18) respectively.

Example 10.5

In a developing country, a certain company has 25 centers located at different places in a state. Each center has been provided with 4 telephones. A student attending a sample survey course was given an assignment to estimate the average number of calls per telephone made on a typical day for this company. The student did not have the telephone facility, and was also short of funds. Because of this, he selected 5 centers using SRS without replacement. The number of calls made on a typical working day from each telephone of the sample centers were recorded personally. The data so obtained are summarized in table 10.5.

Table 10.5 Number of calls made from selected centers

Center	M_i	Calls made				$y_{i.}$	$\bar{y}_i$
1	4	26	34	27	25	112	28
2	4	44	33	28	31	136	34
3	4	18	33	25	28	104	26
4	4	37	21	22	40	120	30
5	4	23	34	42	29	128	32

Estimate the average number of daily calls per telephone made from all the 25 centers, by using estimator (10.1). Also, estimate the relative efficiency of the estimator used with respect to the usual simple mean estimator, from the sample selected above.

Solution

Here N = 25, n = 5, and $M_i = M = 4$. The sample cluster means are given in the last column of table 10.5. The estimate of average number of daily calls is computed using estimator $\bar{y}_{cl}$ given in (10.1). Thus for $M_i = M$,

$$\bar{y}_{cl} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i \qquad \text{[same as } \bar{y}_{c2} \text{ in (10.4)]}$$

From the last column of table 10.5, we get

$$\bar{y}_{cl} = \frac{1}{5}(28+34+26+30+32)$$
$$= 30$$

For $M_i = M$, i = 1, 2, ..., N, the variance estimator $v(\bar{y}_{cl})$ in (10.3) becomes

$$v(\bar{y}_{cl}) = \frac{N-n}{Nn(n-1)} \left(\sum_{i=1}^{n} \bar{y}_i^2 - n\bar{y}_{cl}^2\right)$$

On making substitutions, one gets

$$v(\bar{y}_{cl}) = \frac{25-5}{(25)(5)(4)} [(28)^2 + (34)^2 + \ldots + (32)^2 - 5(30)^2]$$
$$= 1.6$$

Now from (10.18), the variance estimator of the simple mean estimator $\bar{y}$ in (10.15), from the selected cluster sample, will be

$$v_c(\bar{y}) = \frac{N-n}{(NM-1)n}\left[\frac{1}{nM}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij}^2 + v(\bar{y}_{cl}) - \bar{y}_{cl}^2\right]$$

We first compute the term involving sum of squares of all the individual observations. Thus,

$$\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij}^2 = (26)^2 + (34)^2 + \ldots + (29)^2$$

$$= 18962$$

Then,

$$v_c(\bar{y}) = \frac{25-5}{[(25)(4)-1](5)}\left[\frac{18962}{(5)(4)} + 1.6 - (30)^2\right]$$

$$= 2.0081$$

On using (10.19), the estimate of percent relative efficiency will be

$$RE = \frac{2.0081}{1.6}(100)$$

$$= 125.5 \blacksquare$$

We know that the mean square error/variance of any estimator is related to the number of units selected in the sample. In the following section, we, therefore, consider the problem of determining the required number of clusters to be included in the sample when one is to estimate the population mean or total with specified amount of tolerable error in the estimate.

10.6 DETERMINING THE SAMPLE SIZE FOR ESTIMATING MEAN/TOTAL

The total volume of information in cluster sampling is affected by two factors, namely, the number of clusters in the sample and the cluster size. Assuming that the cluster size has been fixed in advance, we consider the problem of determining the number of clusters required to be selected in the sample to obtain estimators with a given precision. Let a preliminary sample of n_1 clusters be selected initially. Based on the information obtained from these n_1 clusters, we compute for estimator 1

$$s_{cl}^2 = \frac{1}{(n_1-1)\overline{M}^2}\sum_{i=1}^{n_1}(y_{i.} - \overline{M}\,\bar{y}_{cl})^2$$

where

$$\overline{M}\,\bar{y}_{cl} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{i.}$$

Using s_{c1}^2 in place of

$$\frac{1}{(n-1)\,\overline{M}^2}\sum_{i=1}^{n}(y_{i.}-\overline{M}\,\overline{y}_{c1})^2$$

in (10.3), we solve the equation

$$2\sqrt{v(\overline{y}_{c1})} = B$$

or equivalently

$$2\sqrt{\frac{N-n}{Nn}\,s_{c1}^2} = B$$

for n. This gives us the required number of clusters to be selected in the overall sample. Here B is the half width of the confidence interval for population mean, and represents the error which the investigator is willing to tolerate in the estimate for population mean. Similarly, we can also obtain the formulas for the required sample size in case of other two estimators. All these formulas are listed below :

Sample size required to estimate mean/total with B as tolerable error :

$$n = \frac{Ns_{ci}^2}{ND+s_{ci}^2},\ i=1,2,3 \qquad (10.20)$$

where D and s_{ci}^2 for the three estimators are defined as follows:

Estimator 1 :

$$s_{c1}^2 = \frac{1}{\overline{M}^2(n_1-1)}\sum_{i=1}^{n_1}(y_{i.}\ -\ \overline{M}\,\overline{y}_{c1})^2 \qquad (10.21)$$

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

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

Estimator 2 :

$$s_{c2}^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(\overline{y}_i-\overline{y}_{c2})^2 \qquad (10.22)$$

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

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

Estimator 3 :

$$s_{c3}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_{i.} - M_i \, \bar{y}_{c3})^2 \qquad (10.23)$$

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

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

In all above cases, the estimators $\bar{y}_{ci}$, i = 1, 2, 3, are based on the preliminary sample, and $\bar{M}$, if unknown, is also to be estimated from the preliminary sample. If $n_1 \geq n$, then preliminary sample is sufficient, otherwise, $(n\text{-}n_1)$ additional clusters are to be selected to get the required overall sample.

Example 10.6

Suppose that the information of example 10.1 pertains to a preliminary sample of 12 villages. Using the data of this example, verify whether this sample size is sufficient to make the inference about the average amount of nitrogen used per hectare by farmers, with a permissible error of magnitude 12 kg ?

Solution

We have N = 170, M_o = 2890, and the preliminary sample size n_1=12. Further, from (10.21)

$$s_{c1}^2 = \frac{1}{\bar{M}^2 (n_1 - 1)} \sum_{i=1}^{n_1} (y_{i.} - \bar{M} \, \bar{y}_{c1})^2$$

Using information from intermediate computations for the estimate of variance in example 10.1, one gets

$$s_{c1}^2 = \frac{4331060}{(17)^2 \, (11)}$$

$$= 1362.397$$

Also,

$$D = \frac{B^2}{4} = \frac{(12)^2}{4} = 36$$

Then the sample size required to estimate the average amount of nitrogen used per hectare with a bound of 12 kg on the error of estimation, would be obtained by using (10.20). Thus,

$$\begin{aligned} n &= \frac{N\, s_{c1}^2}{ND + s_{c1}^2} \\ &= \frac{(170)\,(1362.397)}{(170)\,(36) + 1362.397} \\ &= 30.95 \\ &\approx 31 \end{aligned}$$

This shows that to draw the inference about the average dose of nitrogen actually used in practice by the farmers, with a tolerable error of 12 kg, the preliminary sample size of 12 villages is not sufficient. The investigator will need to select 31-12=19 more villages to achieve the required degree of precision. ■

10.7 ESTIMATION OF PROPORTION

Sometimes the investigator is interested in estimating proportion of units in a population, belonging to a specified category (say) A. For instance, he/she may wish to estimate proportion of heads of religious places who are graduates, or the proportion of persons who drink. On defining Y_{ij} as 1 if the j-th unit of i-th cluster belongs to the specified class, and 0 otherwise, $Y_{i.}$ yields the total number of units in the i-th cluster that belong to category A. Let a_i denote the number of such units in cluster i (we shall denote the corresponding random variable by a), then the population mean per unit reduces to the population proportion P, since

$$\begin{aligned} \overline{Y} &= \frac{1}{M_o} \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij} \\ &= \frac{M_A}{M_o} \\ &= P \end{aligned}$$

where $M_A = \Sigma a_i$, i = 1, 2, ..., N, denotes the total number of units in the population that belong to category A.

The estimators of proportion of units belonging to category A corresponding to the estimators $\overline{y}_{ci}$, i = 1, 2, 3, expressions of their variances, and estimators of variances can be obtained in a straightforward manner by replacing $y_{i.}$ by a_i and $\overline{Y}$ by P in (10.1) to (10.11).

10.7.1 *Estimator 1*

The reader should note that this estimator of population proportion can only be used when M_o, the total number of units in the population, is known.

Unbiased estimator of population proportion when M_o is known :

$$p_{c1} = \frac{1}{n\bar{\bar{M}}} \sum_{i=1}^{n} a_i \tag{10.24}$$

Variance of estimator p_{c1}:

$$V(p_{c1}) = \frac{N-n}{Nn(N-1)} \sum_{i=1}^{N} \left(\frac{a_i}{\bar{\bar{M}}} - P\right)^2 \tag{10.25}$$

Estimator of variance $V(p_{c1})$:

$$v(p_{c1}) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} \left(\frac{a_i}{\bar{\bar{M}}} - p_{c1}\right)^2 \tag{10.26}$$

10.7.2 *Estimator 2*

This estimator can be used in both the cases where M_o is known, or when it is not known.

Estimator of population proportion which does not depend on M_o :

$$p_{c2} = \frac{1}{n} \sum_{i=1}^{n} \frac{a_i}{M_i} \tag{10.27}$$

Bias of the estimator p_{c2} :

$$B(p_{c2}) = -\frac{1}{\bar{\bar{M}}} \operatorname{Cov}\left(\frac{a_i}{M_i}, M_i\right) \tag{10.28}$$

Variance of estimator p_{c2} :

$$V(p_{c2}) = \frac{N-n}{Nn(N-1)} \sum_{i=1}^{N} \left(\frac{a_i}{M_i} - \bar{P}\right)^2 \tag{10.29}$$

where $\bar{P}$ is the average of cluster proportions $P_i = \frac{a_i}{M_i}$, $i = 1, 2, ..., N$.

Estimator of variance $V(p_{c2})$:

$$\left.\begin{aligned} v(p_{c2}) &= \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} \left(\frac{a_i}{M_i} - p_{c2}\right)^2 \\ &= \frac{N-n}{Nn(n-1)} \left[\sum_{i=1}^{n} \left(\frac{a_i}{M_i}\right)^2 - np_{c2}^2\right] \end{aligned}\right] \tag{10.30}$$

10.7.3 *Estimator 3*

This estimator can also be used in both the cases where M_0 is known, or when it is not known. Let S^2_m and S_{am} be defined as in (7.2) with variables the cluster size m, and a, the number of units in a cluster belonging to category A, replacing x and y respectively. Then we have the following :

Estimator of population proportion that does not depend on M_0:

$$p_{c3} = \frac{\sum\limits_{i=1}^{n} a_i}{\sum\limits_{i=1}^{n} M_i} \qquad (10.31)$$

Approximate bias of estimator p_{c3} :

$$B(p_{c3}) = \frac{N-n}{Nn\overline{M}^2}(PS^2_m - S_{am}) \qquad (10.32)$$

Approximate variance of estimator p_{c3} :

$$V(p_{c3}) = \frac{N-n}{Nn\overline{M}^2\,(N-1)} \sum_{i=1}^{N} (a_i - PM_i)^2 \qquad (10.33)$$

Estimator of variance $V(p_{c3})$:

$$v(p_{c3}) = \frac{N-n}{Nn\overline{M}^2(n-1)} \sum_{i=1}^{n} (a_i - p_{c3}M_i)^2 \qquad (10.34)$$

$\overline{M}$, if unknown, is to be replaced by $\hat{\overline{M}} = \frac{1}{n}\sum\limits_{i=1}^{n} M_i$ in (10.34).

Example 10.7

An earlier survey conducted in a rural area, comprising a development block, showed that the proportion of infants vaccinated against polio was only .05. A vigorous campaign was then launched to make the people of this area aware of the need for vaccination against this disease. The Department of Health wanted to have an idea about the extent of impact the campaign had made. This information might be helpful in framing policies for future campaigns. To accomplish the task, 20 villages out of a population of 238 villages were selected using SRS without replacement. Children in a village, who should have been vaccinated but were not vaccinated before the campaign commenced, formed the units in the cluster. All such children in the 238 villages of the block were the target population. The data on the number of children vaccinated after the launch of campaign, are presented in table 10.6 along with other intermediate computations.

Table 10.6 Data regarding children vaccinated after the launching of campaign

Village	Target children (M_i)	Vaccinated children (a_i)	a_i/M_i
1	860	63	.07326
2	935	122	.13048
3	400	105	.26250
4	825	221	.26788
5	642	151	.23520
6	406	90	.22167
7	809	150	.18541
8	679	160	.23564
9	618	103	.16667
10	331	130	.39275
11	410	103	.25122
12	1060	198	.18679
13	576	76	.13194
14	318	113	.35535
15	845	117	.13846
16	921	212	.23018
17	308	44	.14286
18	218	82	.37615
19	880	171	.19432
20	770	130	.16883
Total	12811		4.34756

Using estimator 2, examine whether the campaign for vaccination against polio has been effective ? Also, build up confidence interval for population proportion.

Solution

The statement of the example provides that N = 238 and n = 20. Also, the calculated values for a_i/M_i are given in the last column of table 10.6. Using estimator p_{c2} in (10.27) for proportion P, we get from table 10.6,

$$p_{c2} = \frac{1}{n} \sum_{i=1}^{n} \frac{a_i}{M_i} = \frac{4.34756}{20} = .21738$$

as an estimate of the proportion of children vaccinated against polio in the target population.

Estimate of variance is calculated from (10.30). This is

$$v(p_{c2}) = \frac{N-n}{Nn\,(n-1)} \sum_{i=1}^{n} \left(\frac{a_i}{M_i} - p_{c2}\right)^2$$

$$= \frac{238-20}{(238)\,(20)\,(19)} [(.07326-.21738)^2 + (.13048-.21738)^2$$

$$+ \ldots + (.16883-.21738)^2]$$

$$= \frac{238-20}{(238)\,(20)\,(19)}[(.07326)^2 + (.13048)^2 + \ldots + (.16883)^2 - 20(.21738)^2]$$

$$= \frac{238-20}{(238)\,(20)\,(19)} (.13637)$$

$$= .000329$$

Then we compute the required confidence interval for population proportion P from

$$p_{c2} \pm 2\sqrt{v(p_{c2})}$$

$$= .21738 \pm 2\sqrt{.000329}$$

$$= .21738 \pm .03628$$

$$= .18110,\ .25366$$

The confidence limits above indicate that the proportion of children vaccinated in the target population of 238 villages, after the campaign was launched, is most likely to fall in the closed interval [.18110, .25366]. ■

10.8 SAMPLE SIZE REQUIRED FOR ESTIMATION OF PROPORTION

Analogous to sample size determination in case of continuous data, a preliminary sample of size n_1 is drawn. Proceeding in the same way as in section 10.6, one gets the required size for a sample of clusters to estimate the population proportion with a specified error that can be tolerated.

Sample size required to estimate proportion P with B as tolerable error :

$$n = \frac{Ns_{ci}^2}{ND + s_{ci}^2} \qquad (10.35)$$

where D and s_{ci}^2 for the three estimators are defined below :

Estimator 1 :

$$s_{c1}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(\frac{a_i}{\overline{M}} - p_{c1}\right)^2 \qquad (10.36)$$

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

Estimator 2 :

$$s_{c2}^2 = \frac{1}{n_l - 1} \sum_{i=1}^{n_l} \left(\frac{a_i}{M_i} - p_{c2} \right)^2 \qquad (10.37)$$

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

Estimator 3 :

$$s_{c3}^2 = \frac{1}{n_l - 1} \sum_{i=1}^{n_l} (a_i - M_i \, p_{c3})^2 \qquad (10.38)$$

$$D = \frac{B^2 \overline{M}^2}{4}$$

As before, if $\overline{M}$ is unknown, it is to be estimated from the preliminary sample. Also, if $n_l \geq n$, no additional cluster need to be selected, otherwise, $(n-n_l)$ more clusters will be selected to get the required overall sample.

Example 10.8

The data of example 10.7 had been collected from a sample of 20 villages. Assuming this sample as a preliminary sample, determine the sample size required to estimate the proportion of children vaccinated with a bound of magnitude .03 on the error of estimation.

Solution

In this problem, we have N=238 and the preliminary sample size n_l = 20. Then on using (10.37) and the other intermediate computations for estimate of variance from example 10.7, we work out

$$\begin{aligned} s_{c2}^2 &= \frac{1}{n_l - 1} \sum_{i=1}^{n_l} \left(\frac{a_i}{M_i} - p_{c2} \right)^2 \\ &= \frac{.13637}{19} \\ &= .007177 \end{aligned}$$

Further,

$$D = \frac{B^2}{4} = \frac{(.03)^2}{4} = .000225$$

Then the sample size needed to estimate the population proportion under study with a permissible error of .03, can be obtained by using (10.35) as

$$\begin{aligned} n &= \frac{N s_{c2}^2}{ND + s_{c2}^2} \\ &= \frac{238\,(.007177)}{(238)\,(.000225) + .007177} \end{aligned}$$

$$= 28.13$$
$$\approx 28$$

The optimum sample size, obtained above, means that if the investigator wishes to estimate the population proportion with a tolerable error of magnitude .03, he/she will have to select 28-20 = 8 more villages. ■

10.9 SELECTION OF CLUSTERS WITH UNEQUAL PROBABILITIES

In many practical situations, clusters appreciably differ in their size and the cluster total for the study variable is likely to be positively correlated with the number of units in the cluster. In such cases, it may be advantageous to select the clusters with probability proportional to the number of units in the cluster, instead of with equal probability. If the i-th cluster contains M_i units, the probability of selection for this cluster will be $P_i = M_i/M_0$, i = 1, 2, ..., N. Let a sample of n clusters be selected using probability proportional to size and WR method, the size being the number of units in the cluster. The estimator

$$\bar{y}_{c2} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$$

of mean $\bar{Y}$, given in (10.4), becomes unbiased in this case.

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

$$\bar{y}_{c2} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i \qquad (10.39)$$

Variance of estimator $\bar{y}_{c2}$:

$$V_p(\bar{y}_{c2}) = \frac{1}{nM_0}\sum_{i=1}^{N} M_i(\bar{Y}_i - \bar{Y})^2 \qquad (10.40)$$

Estimator of variance $V_p(\bar{y}_{c2})$:

$$v_p(\bar{y}_{c2}) = \frac{1}{n(n-1)}\sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{c2})^2 \qquad (10.41)$$

In certain other situations, the investigator may find some other more appropriate measures of cluster size. The selection probabilities for the clusters could then be taken proportional to this measure of size. In relations (10.42) to (10.44), we give an unbiased estimator of population mean $\bar{Y}$, and other related results, for any arbitrary set of selection probabilities $\{P_i\}$.

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

$$\bar{y}_{c4} = \frac{1}{nM_o} \sum_{i=1}^{n} \frac{y_{i.}}{p_i} \qquad (10.42)$$

Variance of estimator $\bar{y}_{c4}$:

$$V(\bar{y}_{c4}) = \frac{1}{M_o^2 n} \sum_{i=1}^{N} \left(\frac{Y_{i.}}{P_i} - Y_{..} \right)^2 P_i \qquad (10.43)$$

Estimator of variance $V(\bar{y}_{c4})$:

$$v(\bar{y}_{c4}) = \frac{1}{M_o^2 \, n(n-1)} \left(\sum_{i=1}^{n} \frac{y_{i.}^2}{p_i^2} - nM_o^2 \, \bar{y}_{c4}^2 \right) \qquad (10.44)$$

Estimators of population total Y.. are obtained by multiplying the estimators of population mean $\bar{Y}$ by M_o. The expressions for variance and estimators of variance for these estimators are obtained, as before, by multiplying corresponding expressions for $\bar{Y}$ by M_o^2.

Example 10.9

An agricultural university consists of N = 56 departments including 22 research stations. There are M_o= 570 fast moving vehicles (cars, jeeps, etc.) in the university. The number of vehicles in a department (M_i) vary with the strength of faculty and the nature of research work being carried out. The objective of the survey is to estimate the number of unsafe mounted tires for all the vehicles. For this purpose, a WR sample of n = 14 departments /research stations (D/RS) was selected with probability proportional to the number of vehicles in the department. A team of experts examined all the vehicles in the selected departments /research stations. The information thus collected, is given in the table below :

Table 10.7 Number of vehicles (M_i) and unsafe mounted tires ($y_{i.}$) for the selected D/RS

D/RS	M_i	$y_{i.}$	D/RS	M_i	$y_{i.}$
1	9	4	8	14	9
2	5	3	9	19	12
3	11	8	10	10	6
4	17	9	11	8	3
5	10	6	12	6	2
6	6	0	13	9	3
7	12	3	14	19	9

Estimate the total number of unsafe mounted tires being used in all the 570 vehicles, and place confidence limits on this total.

Solution

We have N = 56, M_o = 570, and n = 14. Through (10.39), the estimator of total is written as

$$\hat{Y}_{c2} = M_o \bar{y}_{c2} = \frac{M_o}{n} \sum_{i=1}^{n} \frac{y_{i.}}{M_i}$$

On making substitutions, one gets the estimate of total number of unsafe tires as

$$\hat{Y}_{c2} = \frac{570}{14}\left(\frac{4}{9} + \frac{3}{5} + \dots + \frac{9}{19}\right)$$

$$= 266.31$$

$$\approx 266$$

The estimator of variance $V(\hat{Y}_{c2})$, from (10.41), will be

$$v_p(\hat{Y}_{c2}) = M_o^2 \, v_p(\bar{y}_{c2})$$

$$= \frac{M_o^2}{n(n-1)} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{c2})^2$$

$$= \frac{M_o^2}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_{i.}}{M_i} - \frac{\hat{Y}_{c2}}{M_o}\right)^2$$

$$= \frac{(570)^2}{14(14-1)}\left[\left(\frac{4}{9} - .467\right)^2 + \left(\frac{3}{5} - .467\right)^2 + \dots + \left(\frac{9}{19} - .467\right)^2\right]$$

$$= \frac{(570)^2\,(.492645)}{14(14-1)}$$

$$= 879.453$$

Following (2.8), we now calculate the confidence limits for the total number of unsafe tires being used in all the 570 vehicles. These limits are

$$\hat{Y}_{c2} \pm 2\sqrt{v_p(\hat{Y}_{c2})}$$

$$= 266.31 \pm 2\sqrt{879.453}$$

$$= 266.31 \pm 59.31$$

$$= 207.00,\ 325.62$$

$$\approx 207,\ 326$$

It can, therefore, be concluded that the total number of unsafe tires mounted on all the 570 vehicles will, most probably, range from 207 to 326. ■

10.10 SOME FURTHER REMARKS

10.1 For a given total number of units in a cluster sample, the sampling variance increases with the increase in cluster size (which consequently results in the decrease of the number of clusters in the sample). On the other hand, the survey cost decreases with the increase in cluster size. In surveys it is, therefore, imperative to strike a balance between the two opposing points. This problem has been considered by various workers. The details are available in Murthy (1967) and Sukhatme *et al.* (1984).

10.2 The cluster sampling, though cheaper, is generally less efficient as compared to the usual simple random sampling of population units. However, if auxiliary information in the form of values of the study variable from the recent past is available, the efficiency of cluster sampling could possibly be improved. Zarkovich and Krane (1965) have considered this aspect of cluster sampling.

10.3 The discussion in this chapter so far has been limited to nonoverlapping clusters where every population unit belongs to one and only one cluster. However, there could be situations where certain population units may fall in more than one cluster. Such clusters are known as *overlapping clusters.* The problem of estimating population mean in such cases has been considered by Tracy and Osahan (1994).

LET US DO

10.1 Define cluster sampling. In which situation is it expected to work better in relation to the usual simple random sampling ?

10.2 In what way the cluster sampling is different from the stratified sampling ? Discuss.

10.3 Give expressions for the unbiased estimator of mean and the estimator of its variance when total number of units in the population is known. From these expressions, how will you write expressions for unbiased estimator of total and the estimator of its variance ?

10.4 '*Ramayana*', the famous religious TV serial of India, was telecast for about 1 year and 9 months. Later, a private company released videocassette of this serial. The company has 30 centers throughout India, and each center has its dealers. In all, there are 250 dealers. After 6 months, an investigator wanted to estimate the total number of times a videocassette of the serial was hired after its introduction in the market. For this purpose, six centers were selected using WOR simple random sampling, and all the dealers in the sample centers were enumerated for the number of times videocassette tapes of the serial were hired from the dealers. The information, so obtained, is given in the following table.

Center	Dealers	Number of times videocassette was hired								
1	7	10	19	7	4	16	21	13		
2	3	8	13	14						
3	6	6	9	22	5	17	8			
4	4	4	2	17	13					
5	9	12	9	21	24	18	7	12	19	10
6	4	6	11	12	7					

Estimate the average number of times a videocassette of the serial was hired from a dealer, and place confidence limits on its population value.

10.5 What are the two estimators of mean in cluster sampling that can be used in situations when the total number of population units is not known ? Also, give expressions for variances of these estimators.

10.6 A sociologist is interested in determining the average degree of mental alertness in persons of age over 80 years in a development block consisting of 70 villages. Since the number of such individuals was not known for each village, a WOR simple random sample of 9 villages was selected. All the persons, with age over 80 years in each selected village, were interviewed and then alertness was ranked from 0 to 10. The score of 10 was given to the persons with perfect mental alertness, whereas 0 score was meant for persons who were not in control of their mental faculties. The number of persons with age over 80 years (M_i) and the scores for their mental alertness, are given below for the sample villages.

Village	M_i	Scores										
1	7	4	6	1	3	7	9	2				
2	11	3	5	2	9	6	8	4	10	5	3	6
3	4	5	7	9	4							
4	5	6	9	8	2	5						
5	8	1	5	9	8	4	3	8	10			
6	6	6	8	4	2	5	7					
7	4	5	7	9	1							
8	7	2	4	9	7	10	1	10				
9	9	3	6	2	4	9	10	1	9	5		

Estimate the average mental alertness for the population of persons aged above 80 years in the development block, and place confidence limits on this average.

10.7 The Department of Education of a state has been providing fixed medical allowance at the rate of rupees 60 per head, for a quarter, to its teachers and their dependents for the last five years. With a view to examine the rationality of this policy today, when the price index has gone up about 1.5 times during the preceding five years, a simple random WOR sample of 10 schools was drawn from a total of 104 schools in a development block by the investigator. Since some of the teachers might be on long leave, the total number of teachers available could not be known in advance. All the teachers (M_i), except those on long leave, in the sample schools were interviewed. They were requested to give per head medical expenses (in rupees), for themselves and their dependents, during the past 3 months. The results are as follows :

School	M_i	Per head medical expenses for 3 months						
1	4	50	100	120	110			
2	4	90	70	40	140			
3	5	85	33	122	60	105		
4	6	55	80	130	70	240	80	
5	4	130	70	40	120			
6	3	85	65	45				
7	4	30	75	65	115			
8	6	150	105	0	25	185	100	
9	5	100	60	130	40	125		
10	7	50	45	110	120	60	140	95

Estimate the average per head money spent as medical expenses during the past 3 months, using the estimators given in (10.4) and (10.8). Also, build up the confidence interval for the population average in each case.

10.8 From the sample observations in exercise 10.4, estimate the total number of times the videocassettes of the serial '*Ramayana*' were hired from all the centers by using the estimator given in (10.14). Also, place confidence limits on this population total.

10.9 There are 300 ponds in a development block consisting of 120 villages. The block administration is planning to use these ponds for fish farming. It is felt that the actual area of ponds is less than that appearing in revenue records. It is perhaps due to the encroachment of the pond area by the residents. A survey was, therefore, undertaken to estimate the total pond area, available at present, in the block. A WOR simple random sample of 10 villages was drawn. Area of each pond in the sample villages was accurately measured. The number of ponds (M_i) in each sample village and the area of ponds are as follows :

Village	M_i	Area (in hectares)			
1	3	2.56	.43	1.62	
2	2	1.31	2.94		
3	4	1.05	2.66	.87	1.02
4	2	2.31	1.75		
5	4	.22	.85	1.93	.74
6	2	1.36	.99		
7	2	.34	1.41		
8	3	2.07	1.61	.73	
9	4	.73	1.82	1.16	.58
10	4	.42	.81	1.3	.75

Using the estimator in (10.14), determine the total pond area in the development block, and also place confidence limits on its population value.

10.10 How will you determine the optimal number of clusters to be included in the sample, in case the estimator given in (10.14) is to be used, for estimating population total with a margin of error of magnitude B ?

10.11 Assume that the sample of 15 buses selected in example 10.2 is a preliminary sample. Using the information obtained from that sample, determine how many more buses are to be included in the sample if the variance of mean is fixed at .007?

10.12 Which of the three estimators given in (10.24), (10.27), and (10.31) will you prefer for estimation of population proportion in case the total number of units in the population is known ? State your reasons.

10.13 A survey project was undertaken to estimate the proportion of railway trains reaching late at their terminal station, in a certain zone. For this purpose, 14 stations from a total of 105 terminal stations were selected using WOR equal probability sampling. The total number of trains terminating at all these 105 stations were counted from railway time table and was found to be 773. All the trains terminating at the selected terminal stations were examined for 24 hours, for late arrivals. A train was considered to be late if it moved into the station 10 or more minutes behind schedule. The information regarding the total number of trains terminating (M_i) and the number of trains arriving late (a_i) at the sample stations is given in the table below:

Station	M_i	a_i	Station	M_i	a_i
1	34	9	8	6	2
2	4	1	9	9	2
3	10	2	10	15	3
4	16	4	11	24	5
5	8	1	12	11	2
6	28	6	13	5	0
7	15	3	14	13	3

Estimate unbiasedly, the proportion/ total number of trains arriving late at all the 105 terminal stations. Also, work out the standard error of your estimate, and place confidence limits on the population parameter being estimated.

10.14 On a canal distributary, there are 56 outlets in the last 10 km length from where the water is supplied to farmers' fields for irrigation. The exact number of farmers using water from these 56 outlets, was not known because of sale and purchase of agricultural land over time. The objective of survey was to assess whether the farmers at the end of distributary were satisfied with the release of water during paddy season, or some more water was required to be released. A WOR simple random sample of 9 outlets was selected, and the views of all the users of the sample outlets were elicited. The response 1 indicates that the farmer was satisfied with the present supply of water, whereas 0 response indicates that additional water needs to be released. The data collected are given in the table below :

Outlets	Users	Response of users
1	4	1 0 0 1
2	7	0 1 1 1 0 1 0
3	8	1 0 0 1 1 1 1 0
4	9	1 0 1 0 1 1 1 1 0
5	6	0 1 0 1 0 0
6	4	1 1 1 0
7	7	1 0 0 1 1 1 1
8	5	0 1 1 1 1
9	9	1 0 0 1 0 1 1 1 1

Estimate proportion of farmers who are satisfied with the present supply of water, and also construct the confidence interval for it.

10.15 Taking the sample of 14 terminal stations drawn in exercise 10.13 as a preliminary sample, work out the required number of stations to be included in the sample, so that, the proportion of late arriving trains can be estimated with a margin of error .06.

10.16 Give expressions for the estimator of population mean and the estimator of its variance if the clusters are selected using varying probability WR sampling.

10.17 Suppose that the sample of 12 villages in example 10.1 was selected using PPS with replacement, the size measure being the number of farmers in the village. Using the sample data in table 10.1, estimate the parameter in question.