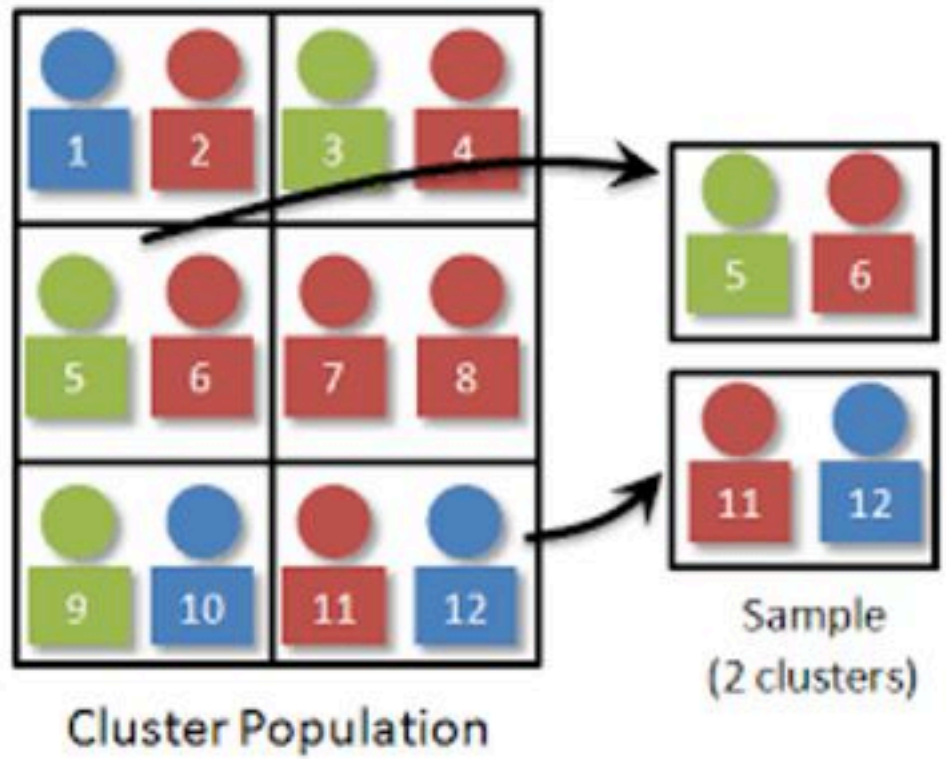
Definition 10.1 The *cluster sampling* consists of forming suitable clusters of contiguous population units, and surveying all the units in a sample of clusters selected according to an appropriate sampling scheme.

Cluster Random Sampling



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

$\bar{M} = M_o/N$ = 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

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

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

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

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

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

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

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

Estimator of population mean which does not depend on M_o:

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

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

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

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

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

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

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

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_o 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

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

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

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

$$= 1.92$$

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

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

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

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

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

$$= .006182$$

The confidence interval for the population average can be derived from

$$\bar{y}_{c2} \pm 2\sqrt{v(\bar{y}_{c2})}$$

$$= 1.92 \pm 2\sqrt{.006182}$$

$$= 1.92 \pm .16$$

$$= 1.76, 2.08$$

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.

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.

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

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.

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_o 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_o^2. This means

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

Also,

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

Hence,

$$v(\hat{Y}_{c1}) = \frac{(120)(120-16)}{(16)(15)}\left[(57.9-34.3875)^2 + (30.1-34.3875)^2\right.$$

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

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

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

$$= 117727.86$$

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. ■

HW

- Page 278, do questions 10.4, 10.6.
- Do examples 10.2 and 10.3 by R.