TWO-STAGE CLUSTER SAMPLING

How to draw a two-stage cluster sample

The first problem in selecting a two-stage cluster sample is the choice of appropriate clusters. Two conditions are desirable: (1) geographic proximity of the elements within a cluster and (2) cluster sizes that are convenient to administer.

The selection of appropriate clusters also depends on whether we want sample a few clusters and many elements from each or many clusters and a few elements from each. Ultimately the choice is based on costs. Large clusters tend possess heterogeneous elements and, hence a large sample is required from each in order to acquire accurate estimates of population parameters. In contrast, small clusters frequently contain relatively homogeneous elements, in which case accurate information on the characteristics of a cluster can be obtained by selecting a small sample from each cluster.

Consider the problem of sampling personal incomes in a large city. The city could be divided into large clusters, for example precincts, which contain a heterogeneous assortment of incomes. Thus a small number of precincts might yield a representative cross section of incomes within the city, but a fairly large sample of elements from each cluster would be required in order to accurately estimates its mean (due to the heterogeneity of incomes within the cluster). In contrast, the city could be divided into small, relatively homogeneous clusters, say city blocks. Then a small sample of people from each block would give adequate information on each cluster's mean, but it would require many blocks to obtain accurate information on the mean income for the entire city.

For another example, consider the university student opinion poll. If students within a university hold similar opinions on the question of interest but opinions differ widely from university to university, then the sample should contain a few representatives from many different universities. If the opinions vary greatly within each university, then the survey should include many representatives from each of a few universities.

To select the sample, we first obtain a frame listing all clusters in the population. We then draw a simple random sample of clusters, using the random sampling procedure. Third, we obtain frames that list all elements in each of the sampled clusters. Finally, we select a simple random sample of elements from each of these frames.

Unbiased estimation of a population mean and total

We are interested in estimating a population mean μ or a population total τ and placing a bound on the error of estimation. The following notation is used:

N = the number of clusters in the population
n = the number of clusters selected in a simple random sample
M_i = the number of elements in cluster i
m_i = the number of elements selected in a simple random sample from cluster i
$M = \sum_{i=1}^{N} M_i$ = the number of elements in the population

$\overline{M} = \frac{M}{N}$ = the average cluster size for the population

y_{ij} = the jth observation in the sample from the ith cluster

$\overline{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$ = the sample mean for the ith cluster

Unbiased estimator of the population mean μ:

$$\hat{m} = (\frac{N}{M})\frac{\sum_{i=1}^{n} M_i \overline{y}_i}{n} \quad (1)$$

Estimated variance of $\hat{m}$

$$\hat{V}(\hat{m}) = (\frac{N-n}{N})(\frac{1}{nN\overline{M}^2})s_b^2 + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2(\frac{M_i - m_i}{M_i})(\frac{s_i^2}{m_i}) \quad (2)$$

where

$$s_b^2 = \frac{\sum_{i=1}^{n}(M_i\overline{y}_i - \overline{M}\hat{m})^2}{n-1} \quad (3)$$

and

$$s_i^2 = \frac{\sum_{i=1}^{m_i}(y_{ij} - \overline{y}_i)^2}{m_i - 1} \qquad i = 1,...,n \quad (4)$$

Bound on the error of estimation:

$$2\sqrt{\hat{V}(\hat{m})} \quad (5)$$

The estimator, $\hat{m}$, shown in equation (1), depends on M, the number of elements in the population. A method of estimating μ when M is unknown is given in the next section.

Note that s_i^2 is the sample variance for the sample selected from cluster i.

Example 1

A garment manufacturer has 90 plants located throughout the United States and wants to estimate the average number of hours that the sewing machines were down for repairs in the past months. Because the plants are widely scattered, she decides to use cluster sampling, specifying each plant as a cluster of machines. Each plant contains many machines, and it would be time-consuming to check the repair record for each machine. Therefore, it seems appropriate to use two-stage sampling. Enough time and money are available to sample n = 10 plants and approximately 20% of the machines in each plant.

Table 1 *Downtime for sewing machines*

Plant	M_i	m_i	*Downtime (in hours)*	$\overline{y_i}$	s_i^2
1	50	10	5,7,9,0,11,2,8,4,3,5	5.40	11.38
2	65	13	4,3,7,2,11,0,1,9,4,3,2,1,5	4.00	10.67
3	45	9	5,6,4,11,12,0,1,8,4	5.67	16.75
4	48	10	6,4,0,1,0,9,8,4,6,10	4.80	13.29
5	52	10	11,4,3,1,0,2,8,6,5,3	4.30	11.12
6	58	12	12,11,3,4,2,0,0,1,4,3,2,4	3.83	14.88
7	42	8	3,7,6,7,8,4,3,2	5.00	5.14
8	66	13	3,6,4,3,2,2,8,4,0,4,5,6,3	3.85	4.31
9	40	8	6,4,7,3,9,1,4,5	4.88	6.13
10	56	11	6,7,5,10,11,2,1,4,0,5,4	5.00	11.8

Using the data in table 1, estimate the average downtime per machine and place a bound on the error of estimation. The manufacturer knows she has a combined total of 4500 machines in all plants.

Solution

The best estimate of m *is* $\hat{m}$*, shown in equation (1), which yields*

$$\hat{m} = \frac{N}{Mn}\sum_{i=1}^{n} M_i\overline{y}_i$$

$$= \frac{90}{(4500)(10)}[(50)(5.40) + (65)(4.00) + \ldots + (56)(5.00)]$$

$$= \frac{90}{(4500)(10)}(2400.59) = 4.80$$

In order to estimate the variance of $\hat{m}$*, we must calculate*

$$s_b^2 = \frac{1}{\mathrm{n} - 1}\sum_{i=1}^{n}(M_i\overline{y}_i - \overline{M}\hat{m})^2$$

$$= \frac{1}{\mathrm{n} - 1}[\sum_{i=1}^{n}(M_i\overline{y}_i)^2 - 2\overline{M}\hat{m}\sum_{i=1}^{n} M_i\overline{y}_i + n(\overline{M}\hat{m})^2]$$

$$= \frac{1}{9}[583{,}198.6721 - 2(50)(4.80)(2400.59) + 10(240)^2]$$

$$= 768.38$$

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

$$= (50)^2(\frac{50-10}{50})(\frac{11.38}{10}) + \ldots + (56)^2(\frac{56-11}{56})(\frac{11.80}{11})$$

$$= 21{,}990.96$$

Then from equation (2),

$$\hat{V}(\hat{m}) = (\frac{N-n}{N})(\frac{1}{n\overline{M}^2})s_b^2 + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2(\frac{M_i-m_i}{M_i})(\frac{s_i^2}{m_i})$$

$$= (\frac{90-10}{90})[\frac{1}{(10)(50)^2}](768.38) + \frac{1}{(10)(90)(50)^2}(21{,}990.96)$$

$$= 0.037094$$

The estimate of m *with a bound on the error of estimation is given by*

$$\hat{m} \pm 2\sqrt{\hat{V}(\hat{m})}, \quad \text{or} \quad 4.80 \pm 2\sqrt{0.037094}, \quad \text{or} \quad 4.80 \pm 0.38$$

Thus the average downtime is estimated to be 4.80 hours. The error of estimation should be less than 0.38 hours with a probability of approximately 95%.

An unbiased estimator of a population total can be found by taking an unbiased estimator of the population mean and multiplying by the number of elements in the population in a manner similar to that used in simple random sampling. Thus $M\hat{m}$ is an unbiased estimator of τ for two-stage cluster sampling.

Estimation of the population total t:

$$\hat{t} = M\hat{m} = N\frac{\sum_{i=1}^{n} M_i\bar{y}_i}{n} \qquad (6)$$

Estimated variance of $\hat{t}$:

$$\hat{V}(\hat{t}) = M^2\hat{V}(\hat{m})$$

$$= (\frac{N-n}{N})(\frac{N^2}{n})s_b^2 + \frac{N}{n}\sum_{i=1}^{n} M_i^2(\frac{M_i-m_i}{M_i})(\frac{s_i^2}{m_i}) \qquad (7)$$

where s_b^2 is given by equation (3) and s_i^2 is given by equation (4).

Bound on the error of estimation:

$$2\sqrt{\hat{V}(\hat{t})} = 2\sqrt{M^2\hat{V}(\hat{m})} \qquad (8)$$

Note that we do not need to know M in order to calculate $\hat{t}$ or the estimated variance of $\hat{t}$, since the M's cancel out in the formula for $\hat{t}$ and s_b^2 [see equations (6) and (7)].

Example 2
Estimate the total amount of downtime during the past month for all machines owned by the manufacturer in example 1. Place a bound on the error of estimation.

Solution
The best estimate of t *is*

$$\hat{t} = M\hat{m} = \frac{N}{n}\sum_{i=1}^{n} M_i \overline{y_i} = \frac{90}{10}(2400.59) = 21{,}605.31$$

The estimated variance of $\hat{t}$ *is found by using the value of* $\hat{V}(\hat{m})$ *calculated in example 1 and substituting as follows:*

$$\hat{V}(\hat{t}) = M^2\hat{V}(\hat{m}) = (4{,}500)^2(.037094)$$

The estimate of t *with a bound on the error of estimation is*

$$\hat{t} \pm 2\sqrt{\hat{V}(\hat{t})}, \qquad \text{or} \qquad 21{,}605.31 \pm 2\sqrt{(4{,}500)^2(.037094)}$$

$$21{,}605.31 \pm 1{,}733.4$$

Thus is estimate of total downtime is 21,605.31 hours. We are fairly confident that the error of estimation is less than 1,733.4 hours.