Cluster Sampling

A cluster sample is a probability sample in which each sampling unit is a collection of elements. Two common reasons for using cluster sampling are i) a frame of elements is either impossible or very costly, and ii) the cost of sampling increases with the distance between the elements. When using cluster sampling, the first decision is what to use as a cluster, several examples of these considerations are discussed in the text. Once the clusters are chosen, a frame of clusters is obtained and then a simple random sample of clusters is taken.

Notation for cluster sampling:

N = the number of clusters in the population,

n = the number of clusters sampled,

m_i = the number of elements in cluster i, $i = 1, 2, 3, ..., N$

$\overline{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ = the average cluster size for the sample,

$M = \sum_{i=1}^{N} m_i$ = the number of elements in the population,

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

y_i = the total of all observations in the ith cluster.

Estimation of a population mean μ :

Our estimator of the population mean is just the total of all elements in the sample divided by the number of elements in the sample:

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} \text{ with } \widehat{V}(\overline{y}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{\overline{M}^2}\right)\frac{s_r^2}{n}, \text{ where } s_r^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y}m_i)^2}{n-1}.$$

Note that the estimator $\overline{y}$ is a ratio estimator. The estimated variance above is biased, so it is advisable to have $n \geq 20$ unless the m_i are equal.

Example: the number of hours of television watched per day

Suppose we visit a small community of $N = 150$ households, and we randomly sample $n = 10$ households. For each sampled household we find out how many people live at the household, and how many hours of TV are watched per day by each of them:

Household	m_i	hours	y_i
1	2	4,5	9
2	4	2,4,2,3	11
3	2	1,2	3
4	3	3,4,4	11
5	5	4,4,5,2,2	17
6	1	2	2
7	3	4,3,3	10
8	2	4,3	7
9	1	6	6
10	4	3,3,5,6	17
Totals	27		93

Now we have $\overline{y} = 93/27 = 3.44$, $\overline{m} = 27/10 = 2.7$, and $\sum_{i=1}^{n}(y_i - \overline{y}m_i)^2 = 46.89$ so that $s_r^2 = 5.21$. Then we have:

$$\widehat{V}(\overline{y}) = \left(\frac{140}{150}\right)\left(\frac{1}{2.7^2}\right)\frac{5.21}{10} = .0667 \text{ so that } B = .52$$

We can also plot y_i against m_i to check the linearity of the data, and if the regression line appears to go through the origin.