Two-stage cluster sampling; unbiased estimation of a population mean and total

Two-stage cluster sampling: In this chapter we examine two-stage cluster sampling, a very simple kind of multi-stage sampling. In real-life surveys, it is common to have several levels of sampling, including use of stratification and ratio estimators. Here we will consider two-stage cluster sampling where a simple random sample of clusters is taken, then a simple random sample of elements in the selected clusters is taken. One example would be to estimate the proportion of U.S. college students who like Korean food by taking a SRS of U.S. colleges, then taking a SRS of students in each selected college.

Notation: The extra stage of sampling here gives us notation that is slightly changed from that used for single-stage cluster sampling:

N = the number of clusters in the population
n = the number of clusters selected in a SRS
M_i = the number of elements in cluster i
m_i = the number of elements selected in a SRS from cluster i
$M = \sum_{i=1}^{N} M_i$ = the number of elements in the population
$\overline{M}/\overline{N}$ = 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 estimation of the population mean μ: In single-stage cluster sampling the unbiased estimator of τ was $\frac{N}{n} \sum_{i=1}^{n} y_i$, where the y_i term was the total of observations in the ith cluster. We are now sampling from each cluster, so we do not know these totals. However, we can estimate the cluster total by multiplying the cluster average ($\overline{y}_i$) by the number of elements in the cluster (M_i). We can divide by M to estimate μ. Then we have:

$$\frac{N}{Mn} \sum_{i=1}^{n} y_i \text{ is estimated by } \frac{N}{Mn} \sum_{i=1}^{n} M_i \overline{y}_i, \text{ so } \widehat{\mu} = \frac{N}{Mn} \sum_{i=1}^{n} M_i \overline{y}_i = \frac{1}{\overline{M}} \frac{\sum_{i=1}^{n} M_i \overline{y}_i}{n},$$

and the estimated variance is:

$$\widehat{V}(\widehat{\mu}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{\overline{M}^2}\right)\frac{s_b^2}{n} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\frac{s_i^2}{m_i},$$

where

$$s_b^2 = \frac{\sum_{i=1}^{n}(M_i\overline{y}_i - \overline{M}\widehat{\mu})^2}{n-1} \text{ and } s_i^2 = \frac{\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_i)^2}{m_i - 1}.$$

In the variance estimator, s_b^2 measures variation between clusters, and s_i^2 measures variation within cluster i. We can obtain an unbiased estimator of τ and a variance estimator by multiplying the above expressions by M:

$$\widehat{\tau} = \frac{N}{n}\sum_{i=1}^{n} M_i\overline{y}_i = N\,\frac{\sum_{i=1}^{n} M_i\overline{y}_i}{n},$$

with variance estimator:

$$\widehat{V}(\widehat{\tau}) = \left(\frac{N-n}{N}\right)N^2\frac{s_b^2}{n} + \frac{N}{n}\sum_{i=1}^{n} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\frac{s_i^2}{m_i},$$

where s_b^2 and s_i^2 are defined as above. See the text and SAS code on the web for example calculations.