# Cluster sampling

Suppose a population of size $N$ is divided into $K = N/M$ groups of size $M$. We select a sample of size $km$ the following way:

- First we select $k$ groups out of $K$ groups by simple random sampling *with* replacement.
- We then select $m$ units in each group selected on the first step by simple random sample *with* replacement.
- The estimate of the population mean is the average $\bar{Y}$ of the sample.

Let $\mu_i$ be the population average in the $i$-the group for $i = 1, 2, \ldots, K$. Let

$$\sigma_u^2 = \frac{1}{K} \sum_{i=1}^{K} (\mu_i - \mu)^2 ,$$

where $\mu = \sum_{i=1}^{K} \mu_i / K$. Let

$$\sigma_w^2 = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{M} (y_{ij} - \mu_i)^2 ,$$

where $y_{ij}$ denotes the value of the variable for the $j$-the unit in the $i$-th group.

a. Let $k = 1$. Show that we can write the estimator as

$$\bar{Y} = \sum_{i=1}^{K} I_i Y_i ,$$

where

$$I_i = \begin{cases} 1 & \text{if the } i\text{-th group is selected.} \\ 0 & \text{otherwise} \end{cases}$$

and $\mathrm{var}(Y_i) = \sigma_i^2/m$. Argue that it is reasonable to assume that $Y_i$ and $I_i$ are all independent. Let $\sigma_i^2$ be the population variance for the $i$-th subgroup. Compute $\mathrm{var}(\bar{Y})$.

b. If we repeat the procedure we get independent estimators $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$, and estimate the population average by

$$\bar{Y} = \frac{1}{k} \sum_{i=1}^{k} \bar{Y}_k \,.$$

Show that

$$\mathrm{var}(\bar{Y}) = \frac{\sigma_u^2}{k} + \frac{\sigma_w^2}{km} \,.$$

Argue that this expression is the variance of the estimator described in the introduction.

c. The assumption that we sample with replacement is unrealistic. Let $k = 1$ and assume that the sample of size $m$ is selected by simple random sample *without* replacement. Argue that

$$\bar{Y} = \sum_{i=1}^{K} I_i Y_i \,,$$

where

$$I_i = \begin{cases} 1 & \text{if we select the } i\text{th subgroup.} \\ 0 & \text{otherwise} \end{cases}$$

Compute the variance of the estimator in this case.

d. Assume that the $k$ groups are selected by simple random sample *without* replacement. In this case the estimator is

$$\bar{Y} = \frac{1}{k} \sum_{i=1}^{K} I_i Y_i \,,$$

where

$$I_i = \begin{cases} 1 & \text{if we select the } i\text{th subgroup.} \\ 0 & \text{otherwise} \end{cases}$$

Argue that it is reasonable to assume that $I_1, \ldots, I_K$ and $Y_1, \ldots, Y_K$ are independent.

e. Explain why the sampling distribution in d. is approximately normal. Do a simulation and compare the standard error given by the formula with the standard error you get from simulations.