Sample size selection for two-stage cluster sampling

As we noticed previously, if the number of clusters in the population, \( N \), is large, then \( \hat{V}(\hat{\mu}) \approx MSB/nm \), which estimates \( \frac{1}{n} [\sigma_b^2 + \sigma_w^2/m] \), where \( \sigma_b^2 \) is the variation among cluster means and \( \sigma_w^2 \) is the variation among elements within clusters. Also, MSW estimates \( \sigma_w^2 \), so to obtain an estimator of \( \sigma_b^2 \) we can solve these two equations to obtain \( \hat{\sigma}_b^2 = \frac{1}{m} [MSB - MSW] \). We can use these estimates to calculate a sample size, using variance estimates from a pilot study. We can either calculate a sample size for a fixed variance or a fixed cost. Let \( c_1 \) be the cost of sampling a cluster, and \( c_2 \) be the cost of sampling an element within a cluster. Then, we first calculate \( m \) via:

\[
m = \sqrt{\frac{\sigma^2_w c_1}{\sigma^2_b c_2}},
\]

and use either \( B = 2\sqrt{\frac{1}{n} [\sigma^2_b + \sigma^2_w/m]} \) (if using a fixed bound \( B \)), or \( c = nc_1 + nmc_2 \) (if using a fixed total cost \( c \)) to solve for the number of clusters \( n \).

Two-stage cluster sampling with Probabilities Proportional to Size

The idea here is to use PPS sampling to obtain a sample of clusters (with replacement, as before), then sample within those sampled clusters. From each sampled cluster we have a sample mean, \( \bar{y}_i \), and so an average of these cluster means is our estimator of \( \mu \):

\[
\hat{\mu}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i
\]

with estimated variance

\[
\hat{V}(\hat{\mu}_{pps}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\bar{y}_i - \hat{\mu}_{pps})^2.
\]

We can multiply by \( M \) to obtain an estimator of the total \( \tau \):
\[ \hat{\tau}_{pps} = \frac{M}{n} \sum_{i=1}^{n} \bar{y}_i \]

and its estimated variance:

\[ \hat{V}(\hat{\tau}_{pps}) = \frac{M^2}{n(n-1)} \sum_{i=1}^{n} (\bar{y}_i - \hat{\mu}_{pps})^2. \]