Two stage cluster sampling

Some proofs

Assume the population consists of N clusters each of size M.

We select n clusters using srs and within the select clusters use srs to select independent samples each of size m.

Let $\bar{\bar{y}} = \frac{1}{n}\sum_{i \in smp} \bar{y}_i$. Then

$$E(\bar{\bar{y}}) = \bar{\bar{Y}} \quad \text{and} \quad V(\bar{\bar{y}}) = (1 - \frac{n}{N})\frac{\sigma_b^2}{n} + (1 - \frac{m}{M})\frac{\sigma_w^2}{mn}$$

Since an expectation can be written as the expectation of a conditional expectation we have

$$E(\bar{\bar{y}}) = E_1[E_2(\bar{\bar{y}})]$$

Here E_1 averages over all possible clusters that can appear in a first stage sample.

E_2 is a conditional expectation which averages over all possible units that can appear in a second stage of sampling given the clusters that appear in the first stage of the sampling.

Proof of first part

$$
\begin{aligned}
E(\bar{\bar{y}}) &= E_1[E_2(\bar{\bar{y}})] \\
&= E_1[E_2(\frac{1}{n}\sum_{i\in smp} \bar{y}_i)] \\
&= E_1[\frac{1}{n}\sum_{i\in smp} E_2(\bar{y}_i)] \\
&= E_1[\frac{1}{n}\sum_{i\in smp} \bar{Y}_i] \\
&= \bar{\bar{Y}}
\end{aligned}
$$

Proof of second part

$$V(\bar{\bar{y}}) = V_1(E_2(\bar{\bar{y}})) + E_1(V_2(\bar{\bar{y}}))$$

Consider the first term on the RHS.

$$V_1(E_2(\bar{\bar{y}})) = V_1(\frac{1}{n}\sum_{i\in smp}\bar{Y}_i) = (1-\frac{n}{N})\frac{\sigma_b^2}{n}$$

Now consider the second term on the RHS.

$$\begin{aligned} E_1(V_2(\bar{\bar{y}})) &= E_1\Big(\frac{1}{n^2}(1-\frac{m}{M})\sum_{i\in smp}\frac{\sigma_i^2}{m}\Big) \\ &= \frac{1}{nm}(1-\frac{m}{M})E_1\Big(\sum_{i\in smp}\frac{\sigma_i^2}{n}\Big) \\ &= \frac{1}{nm}(1-\frac{m}{M})\sum_{i=1}^{N}\frac{\sigma_i^2}{N} = (1-\frac{m}{M})\frac{\sigma_w^2}{nm} \end{aligned}$$

The unbiased estimator of $V(\bar{\bar{y}})$ is

$$\widehat{V(\bar{\bar{y}})} = (1 - \frac{n}{N})\frac{s_b^2}{n} + \frac{n}{N}(1 - \frac{m}{M})\frac{s_w^2}{mn}$$
$$= (1 - f_1)\frac{s_b^2}{n} + f_1(1 - f_2)\frac{s_w^2}{mn}$$

where $f_1 = n/N$ and $f_2 = m/M$.

The factor f_1 in the second term on the RHS is surprising.

It comes about because while s_w^2 is an unbiased estimator of σ_w^2 s_b^2 is a biased estimator of σ_b^2 and on average is an over estimate.

Showing the unbiasedness of s_w^2

$$
\begin{aligned}
E(s_w^2) &= E_1[E_2(s_w^2)] \\
&= E_1[E_2(\sum_{i\in smp}\sum_{j\in smp_i}(y_{ij}-\bar{y}_i)^2/(m-1)n)] \\
&= E_1[\frac{1}{n}\sum_{i\in smp} E_2(\sum_{j\in smp_i}(y_{ij}-\bar{y}_i)^2/(m-1))] \\
&= E_1[\frac{1}{n}\sum_{i\in smp}\sigma_i^2] \\
&= \sum_{i=1}^{N}\sigma_i^2/N \\
&= \sigma_w^2
\end{aligned}
$$

Showing that s_b^2 is biased

Recall $E_2(\bar{y}_i^2) = \bar{Y}_i^2 + (1 - f_2)\sigma_i^2/m$ and

$$\begin{aligned} E_2(\bar{\bar{y}}^2) &= [E_2(\bar{\bar{y}})]^2 + V_2(\bar{\bar{y}}) \\ &= [\sum_{i \in smp} \bar{Y}_i/n]^2 + \frac{1 - f_2}{n^2} \sum_{i \in smp} \sigma_i^2/m \end{aligned}$$

Let $\bar{\bar{Y}}_n = \sum_{i \in smp} \bar{Y}_i / n$. Then

$$(n-1)E_2(s_b^2) = E_2[\sum_{i \in smp} \bar{y_i}^2 - n\bar{\bar{y}}^2]$$

$$= \sum_{i \in smp} \bar{Y}_i^2 + \frac{1-f_2}{m} \sum_{i \in smp} \sigma_i^2$$

$$- n\bar{\bar{Y}}_n^2 - \frac{1-f_2}{nm} \sum_{i \in smp} \sigma_i^2$$

$$= \sum_{i \in smp} (\bar{Y}_i^2 - \bar{\bar{Y}}_n)^2 +$$

$$\frac{(n-1)(1-f_2)}{nm} \sum_{i \in smp} \sigma_i^2$$

Next we multiple both sides by $(1-f_1)/(n(n-1))$ to get

$$\frac{1-f_1}{n}E_2(s_b^2) = \frac{1-f_1}{n}\sum_{i\in smp}(\bar{Y}_i^2 - \bar{\bar{Y}}_n)^2/(n-1) + \frac{(1-f_1)(1-f_2)}{nm}\sum_{i\in smp}\sigma_i^2/n$$

Next we apply E_1 to both sides of the equation to get

$$\frac{1-f_1}{n}E(s_b^2) = \frac{1-f_1}{n}\sigma_b^2 + \frac{(1-f_1)(1-f_2)}{nm}\sigma_w^2$$

and we see that on the average s_b^2 will over estimate σ_b^2.

Since $(1-f_1)(1-f_2)+f_1(1-f_2) = 1-f_2$ the proof is complete.