Stat 257: (Selected) Solutions to Assignment #5

(5.2) We begin with the observation that

$$\sum_{i=1}^{M}\sum_{j=1}^{L}(Y_{ij}-\overline{Y})^2 + 2\sum_{i=1}^{M}\sum_{j<k}(Y_{ij}-\overline{Y})(Y_{ik}-\overline{Y}) = \sum_{i=1}^{M}\sum_{j=1}^{L}\sum_{k=1}^{L}(Y_{ij}-\overline{Y})(Y_{ik}-\overline{Y}). \qquad (*)$$

The two expressions on the left side of $(*)$ are easy to deal with. From equation (5.3) on page 136 we have

$$\sum_{i=1}^{M}\sum_{j=1}^{L}(Y_{ij}-\overline{Y})^2 = (ML-1)S^2 \qquad (1)$$

and from the definition of ρ in equation (5.5) on page 137, we see that

$$2\sum_{i=1}^{M}\sum_{j<k}(Y_{ij}-\overline{Y})(Y_{ik}-\overline{Y}) = [(L-1)(ML-1)S^2]\rho. \qquad (2)$$

Well, it turns out that the right side of $(*)$ is also easy to deal with. Notice that the inner sum is over k and that $(Y_{ij}-\overline{Y})$ does NOT depend on k. Therefore,

$$\sum_{i=1}^{M}\sum_{j=1}^{L}\sum_{k=1}^{L}(Y_{ij}-\overline{Y})(Y_{ik}-\overline{Y}) = \sum_{i=1}^{M}\sum_{j=1}^{L}(Y_{ij}-\overline{Y})\sum_{k=1}^{L}(Y_{ik}-\overline{Y}).$$

But, notice that

$$\sum_{k=1}^{L}(Y_{ik}-\overline{Y}) = \left(\sum_{k=1}^{L}Y_{ik}\right) - L\overline{Y} = L\overline{Y}_i - L\overline{Y}$$

and similarly

$$\sum_{j=1}^{L}(Y_{ij}-\overline{Y}) = L\overline{Y}_i - L\overline{Y}$$

so that

$$\sum_{i=1}^{M}\sum_{j=1}^{L}\sum_{k=1}^{L}(Y_{ij}-\overline{Y})(Y_{ik}-\overline{Y}) = \sum_{i=1}^{M}\sum_{j=1}^{L}(Y_{ij}-\overline{Y})\sum_{k=1}^{L}(Y_{ik}-\overline{Y}) = L^2\sum_{i=1}^{M}(\overline{Y}_i-\overline{Y})^2.$$

We now use equation (5.3) on page 136 again to conclude

$$L^2\sum_{i=1}^{M}(\overline{Y}_i-\overline{Y})^2 = L\cdot L\sum_{i=1}^{M}(\overline{Y}_i-\overline{Y})^2 = L\cdot\left[(ML-1)S^2 - M(L-1)\overline{S}^2\right]. \qquad (3)$$

Finally, we can combine (1), (2), and (3) to give

$$(ML-1)S^2 + [(L-1)(ML-1)S^2]\rho = L\cdot\left[(ML-1)S^2 - M(L-1)\overline{S}^2\right]. \qquad (4)$$

Solving for ρ now gives

$$\rho = 1 - \left(\frac{ML}{ML-1}\right)\left(\frac{\overline{S}^2}{S^2}\right).$$

It now follows that

$$
\begin{aligned}
\mathrm{Var}(\overline{y}_{CL}) &= \frac{1}{m}(1-f)\sum_{i=1}^{M}\frac{(\overline{Y}_i-\overline{Y})^2}{M-1} \\
&= \frac{1}{m(M-1)}(1-f)\sum_{i=1}^{M}(\overline{Y}_i-\overline{Y})^2 \\
&= \frac{1}{m(M-1)}(1-f)\frac{L\left[(ML-1)S^2-M(L-1)\overline{S}^2\right]}{L^2} \quad \text{from (3)} \\
&= \frac{1}{m(M-1)}(1-f)\frac{(ML-1)S^2+[(L-1)(ML-1)S^2]\rho}{L^2} \quad \text{from (4)} \\
&= \frac{(1-f)}{m}\frac{ML-1}{L^2(M-1)}S^2[1+(L-1)\rho].
\end{aligned}
$$

We now find that

$$
\frac{\mathrm{Var}(\overline{y}_{CL})}{\mathrm{Var}(\overline{y})} = \frac{\frac{(1-f)}{m}\frac{ML-1}{L^2(M-1)}S^2[1+(L-1)\rho]}{\frac{(1-f)}{mL}S^2} = \frac{ML-1}{L(M-1)}[1+(L-1)\rho] \to 1+(L-1)\rho
$$

as $M \to \infty$ so that we may write

$$
\frac{\mathrm{Var}(\overline{y}_{CL})}{\mathrm{Var}(\overline{y})} \doteqdot 1+(L-1)\rho.
$$

(5.3) Recall that the variance of the cluster sample total $\overline{y}_{c(b)}$ is given by

$$
\mathrm{Var}(\overline{y}_{c(b)}) = \frac{(M-m)M}{(M-1)mN^2}\sum_{i=1}^{M}(Y_{iT}-\overline{Y}_T)^2
$$

where

$$
Y_{iT} = N_i\overline{Y}_i \quad \text{and} \quad \overline{Y}_T = \frac{N}{M}\overline{Y}.
$$

From the data given, we find that $M = 12$, and

$$
N = \sum_{i=1}^{M} N_i = 81.
$$

If we want to find the standard error of the unbiased cluster sample estimator $\overline{y}_{c(b)}$ for a cluster sample of 4 branches, then we take $m = 4$. Next, we calculate the population mean

$$
\begin{aligned}
\overline{Y} &= \frac{1}{M}\sum_{i=1}^{M}\overline{Y}_i \\
&= \frac{1}{12}(24.32+27.06+27.60+28.01+27.56+29.07+32.03+28.41+28.91+25.55+28.58+27.27) \\
&= \frac{334.37}{12} \\
&\approx 27.86
\end{aligned}
$$

so that

$$\overline{Y}_T = \frac{N}{M}\overline{Y} \approx \frac{81}{12} \cdot 27.86 \approx 188.08.$$

Thus,

$$\sum_{i=1}^{M}(Y_{iT} - \overline{Y}_T)^2 \approx \sum_{i=1}^{M}(N_i Y_i - 188.08)^2 \approx 144\ 538$$

so that

$$\operatorname{Var}(\overline{y}_{c(b)}) = \frac{(M-m)M}{(M-1)mN^2}\sum_{i=1}^{M}(Y_{iT} - \overline{Y}_T)^2 \approx \frac{(12-4)\cdot 12}{(12-1)\cdot 4 \cdot 81^2} \cdot 144\ 538 \approx 48.07.$$

The standard error of $\overline{y}_{c(b)}$ is therefore

$$\operatorname{SE}(\overline{y}_{c(b)}) = \sqrt{\operatorname{Var}(\overline{y}_{c(b)})} \approx 6.93.$$

On the other hand, the variance of the simple random sampling estimator $\overline{y}$ is given by

$$\operatorname{Var}(\overline{y}) = \frac{(1-f)}{n}S^2$$

where the overall sample variance is given by

$$\begin{aligned} S^2 &= \frac{1}{N-1}\left(\sum_{i=1}^{M}(N_i - 1)S_i^2 + \sum_{i=1}^{M} N_i\left(\overline{Y}_i - \overline{Y}\right)^2\right) \\ &\approx \frac{1}{81-1}\left(\sum_{i=1}^{12}(N_i - 1)S_i^2 + \sum_{i=1}^{12} N_i\left(\overline{Y}_i - 27.86\right)^2\right) \\ &\approx \frac{1}{81-1}\left(416.04 + 240.72\right) \\ &\approx 8.21. \end{aligned}$$

Thus, we find for a sample of size $n = 27$ that

$$\operatorname{Var}(\overline{y}) = \frac{(1-f)}{n}S^2 \approx \frac{1 - \frac{27}{81}}{27} \cdot 8.21 \approx 0.202$$

which gives the standard error of $\overline{y}$ as

$$\operatorname{SE}(\overline{y}) = \sqrt{\operatorname{Var}(\overline{y})} \approx 0.449.$$

The relative efficiency is therefore

$$\operatorname{RE}(\overline{y}, \overline{y}_{c(b)}) = \frac{\operatorname{Var}(\overline{y})}{\operatorname{Var}(\overline{y}_{c(b)})} \approx \frac{0.202}{48.07} \approx 0.42\%.$$

This shows that the simple random sampling estimator $\overline{y}$ is vastly more efficient. In fact, the cluster sampling estimator is less than one-half-of-one percent as efficient as the simple random sampling estimator.

Since the cluster sizes are not equal (i.e., there is no number L such that $N_1 = N_2 = \cdots = N_{12} = L$), the estimator $\overline{y}_{c(a)}$ cannot be used in this situation. As noted in on page 141, the cluster estimator $\overline{y}_{c(c)}$ is always a biased estimator. Since we have complete knowledge of the population data (as given in the problem) and are able to use $\overline{y}_{c(b)}$, there is no reason to use $\overline{y}_{c(c)}$.