Multistage Sampling (Chapter 13)

Multistage sampling refers to sampling plans where the sampling is carried out in stages using smaller and smaller sampling units at each stage. In a two-stage sampling design, a sample of primary units is selected and then a sample of secondary units is selected within each primary unit. This handout outlines the development of estimators under the general setting of two-stage sampling, considers the allocation question under the setting of equal sized primary and secondary units, and briefly examines three-stage sampling.

The simplest version of two-stage sampling is to use simple random sampling at each stage – an SRS of primary units, and an SRS of secondary units within each selected primary unit. The primary units do not need to be the same size and you do not need to select the same number of secondary units within each primary unit.

- Stratified random sampling and cluster sampling can be viewed as special cases of two-stage sampling. A stratified random sample is a census of the primary units (the strata) followed by an SRS of the secondary units within each primary unit. A cluster sample is an SRS of the primary units (the clusters) followed by a census of the secondary units within each selected primary unit.

- We can use any probability sampling plan at each stage of a multistage plan and the plans can be different at each stage. The formulas developed below are only for an SRS at each stage. It’s possible to derive formulas for other situations.

Example: In order to estimate the condition of highways under its jurisdiction and the cost of urgent repairs, the state Department of Transportation selected a number of “highway miles” in two stages. In the first stage, a number of highways were selected at random and without replacement from the list of all highways maintained by the Department. In the second stage, a number of one-mile segments were selected at random and without replacement from the total length of each selected highway; for example, if the length of highway 101 is 73 miles, it is seen as consisting of 73 one-mile segments (“highway miles”), from which a number are selected at random. Highway engineers then visit the selected segments, inspect the pavement condition, rate the condition of the segment, and estimate the cost of urgently needed repairs.

For the purpose of this problem, assume there are 352 highways in the state, with a total length of 28,950 miles. A simple random sample of five highways was selected without replacement. From each selected highway, approximately 10% of its one-mile segments were then selected. The inspection results were as follows:
<table>
<thead>
<tr>
<th>Highway Number</th>
<th>Length (miles)</th>
<th>Selected One-Mile Segments</th>
<th>Number Rated Excellent</th>
<th>Cost of Urgent Repairs (in $1,000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>155</td>
<td>85</td>
<td>10</td>
<td>2</td>
<td>90</td>
</tr>
<tr>
<td>489</td>
<td>120</td>
<td>15</td>
<td>1</td>
<td>110</td>
</tr>
<tr>
<td>283</td>
<td>47</td>
<td>5</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>698</td>
<td>98</td>
<td>10</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>311</td>
<td>34</td>
<td>5</td>
<td>1</td>
<td>30</td>
</tr>
</tbody>
</table>

For example, Highway 155 has a length of 85 miles. Ten of its 85 one-mile segments were selected and inspected. Two of these segments were rated Excellent. The total cost of urgent repairs on the 10 selected segments was $90,000.

(a) Estimate the proportion and number of state highway miles that are in Excellent condition.
(b) Estimate the average cost per highway mile and the total cost of urgently needed repairs.

- First, why would a two-stage sampling plan be adopted for this highway problem in the first place? Why not an SRS?

- Multistage samples are used primarily for cost or feasibility (practicality) reasons. For example, to select an SRS of households in the U.S. would be extremely difficult because no list of all households exists. However, we could proceed in stages: an SRS of counties in the U.S., an SRS of “blocks” within each county, and an SRS of households within each block. You would then only need to have a list of households within each block that was selected. Two-stage sampling also has the flexibility to sample more intensely in primary units which are larger or more variable. The disadvantage of two-stage sampling is that the variance of the resulting estimators are likely to be larger than for an SRS of the same total number of secondary units. This may well be more than offset by the cost efficiency of two-stage sampling.

- Note that a two-stage sample can never be better than a cluster sample with the same number of primary units selected because a census within each primary unit is the best you can do.
Notation for Two-Stage Sampling:

\[ N = \text{the number of primary units in the population}, \]
\[ n = \text{the number of primary units in the sample}, \]
\[ M_i = \text{the number of secondary units in the } i^{th} \text{ primary unit}, \]
\[ m_i = \text{the size of the sample in the } i^{th} \text{ primary unit}, \]
\[ y_{ij} = \text{the response of the } j^{th} \text{ secondary unit within the } i^{th} \text{ primary unit}, \]
\[ y_i = \sum_{j=1}^{M_i} y_{ij} = \text{the total in the } i^{th} \text{ primary unit}, \]
\[ \mu_i = \frac{1}{M_i} y_i = \text{the mean response in the } i^{th} \text{ primary unit}, \]
\[ \bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} = \text{the sample mean response in the } i^{th} \text{ primary unit} \]

In two-stage problems, we are generally interested in:

\[ \tau = \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \text{total of the } y\text{-values for the population}, \]
\[ \mu = \frac{\tau}{M}, \text{ where } M = \sum_{i=1}^{N} M_i = \text{total } \# \text{ of secondary units in the population}. \]

- For stratified random sampling: \( n = N \) (census of primary units).
- For cluster sampling: \( m_i = M_i \) (census of secondary units).

Unbiased Estimation of the Population Total \( \tau \) and Mean \( \mu \).

- First, an unbiased estimator of the total in the \( i^{th} \) cluster (\( y_i \)) is \( \hat{y}_i = \)
- With these estimated cluster totals, estimators for the population total \( \tau \) and mean \( \mu \) are given by:

\[ \hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{y}_i, \quad \hat{\mu} = \frac{\hat{\tau}}{M} = \frac{N}{nM} \sum_{i=1}^{n} M_i \bar{y}_i. \]

R code to compute the estimated total for the highway example follows:

```r
N <- 352; n <- 5; M <- 28950
Mi <- c(85,120,47,98,34) # total no. segments on the highways sampled
mi <- c(10,15,5,10,5) # no. of segments sampled
yi <- c(2,1,0,0,1) # no. of excellent segments

# Unbiased estimation of total number of segments rated Excellent
# =============================================================
yhati <- (Mi/mi)*yi
```
> yhati # estimated no. excellent segments on each highway
[1] 17.0  8.0  0.0  0.0  6.8
>
> tauhat <- (N/n)*sum(yhati) # estimated total no. excellent
> tauhat
[1] 2238.72

So an unbiased estimate of the total number of highway segments rated as Excellent is \( \hat{\tau} = 2238.7 \). What is \( \hat{\mu} \)?

It can be shown that in two-stage sampling, the variance of the estimator of \( \mu \) is given as:

\[
\text{Var}(\hat{\mu}) = \frac{N^2}{M^2} \left( \frac{N-n}{N} \right) \frac{s_u^2}{n} + \frac{N}{nM^2} \sum_{i=1}^{N} M_i^2 \left( \frac{M_i - m_i}{M_i} \right) \frac{s_i^2}{m_i},
\]

where:
\[
\sigma_u^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu_1)^2,
\]
\[
\sigma_i^2 = \frac{1}{M_i-1} \sum_{j=1}^{M_i} (y_{ij} - \mu_i)^2.
\]

Also, \( \text{Var}(\hat{\tau}) = M^2 \text{Var}(\hat{\mu}) \).

- If \( n = N \), then the first term = 0, and the second term = \( \text{Var}(\hat{\mu}) \) for stratified random sampling.
- If \( m_i = M_i \), then the second term = 0, and the first term = \( \text{Var}(\hat{\mu}) \) for cluster sampling.
- The estimated variance of \( \hat{\mu} \) is given by:

\[
\hat{\text{Var}}(\hat{\mu}) = \frac{N^2}{M^2} \left( \frac{N-n}{N} \right) \frac{s_u^2}{n} + \frac{N}{nM^2} \sum_{i=1}^{N} M_i^2 \left( \frac{M_i - m_i}{M_i} \right) \frac{s_i^2}{m_i},
\]

where:
\[
s_u^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{y}_i - \hat{\mu}_1)^2, \quad s_i^2 = \frac{1}{m_i-1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2.
\]

There are two levels of approximation in \( s_u^2 \): we use \( n \) for \( N \) and \( \hat{y}_i \) for \( y_i \) (primary unit total).

- Similarly, \( \hat{\text{Var}}(\hat{\tau}) = M^2 \hat{\text{Var}}(\hat{\mu}). \) (since \( \hat{\tau} = M\hat{\mu} \))

- In the highway example, we are counting the number of 1-mile segments rated as Excellent; hence, we have binary data \( (y_{ij} = 0 \text{ or } 1) \). So the mean in this example is
the proportion of one-mile segments rated as Excellent. Here then, the within-primary unit sample variance is:

\[ s_i^2 = \frac{m_i}{m_i - 1} \hat{p}_i (1 - \hat{p}_i) \] (the binomial variance).

So, to finish answering part (a) of the highway problem (where we already estimated the total number of segments in Excellent condition to be \( \hat{\tau} = 2238.7 \)), the estimated proportion of highway miles in Excellent condition, as well as standard errors for both this proportion and the total are given via the R code below:

```r
> su2 <- var(yhati)
> pi <- yi/mi # Proportion of segments rated excellent on each highway
> si2 <- (mi/(mi-1))*pi*(1-pi) # Estimated variance within each primary unit
> var1 <- (N*(N-n)*su2)/n # Term 1 of variance
> var2 <- (N/n)*sum((Mi*(Mi-mi)*si2)/mi) # Term 2 of variance
> var.tauhat <- var1 + var2
> SE.tauhat <- sqrt(var.tauhat) # SE of estimate of total
> phat <- tauhat/M # estimate of proportion Excellent
> SE.phat <- SE.tauhat/M # SE of estimate of proportion
> phat <- tauhat/M
```

Note that the confidence interval extends below 0. Since the estimated proportions within each highway are near 0, our sample sizes are too small to assume a normal sampling distribution for \( \hat{p} \). We might consider bootstrapping.
Ratio Estimation in Two-Stage Sampling: If the sizes of the primary units (highways) are linearly related (through the origin) with the values of the response (number rated as excellent, or cost of urgent repairs), a ratio estimator may provide a better estimator of the population total or mean.

- The estimators are given by:
  \[ \hat{\mu}_r = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}, \quad \hat{\tau}_r = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i} \]

- The variance of \( \hat{\mu}_r \) is given by:
  \[ \text{Var}(\hat{\mu}_r) = \frac{N^2}{M^2} \left( \frac{N-n}{N} \right) \frac{1}{n} \left( \frac{1}{N-1} \sum_{i=1}^{N} (y_i - M_i \mu)^2 \right) + \frac{N}{nM^2} \sum_{i=1}^{N} M_i^2 \left( \frac{M_i - m_i}{M_i} \right) \sigma_i^2 \]

  variability between clusters (var. in cluster (highway) totals) where the difference in highway lengths is accounted for

  Within-cluster variability (same as earlier)

- If \( M_i \) (cluster \( i \) size) is related to \( y_i \) (cluster \( i \) sample total) in the first term above (\( (y_i - M_i \mu)^2 \)), this term should be small. This is the situation where ratio estimation should be used. The approximate variance \( \text{Var}(\hat{\mu}_r) \) is given on page 147 of the text.

- The estimators and corresponding SE’s for ratio estimation are computed via R:

  > # Ratio Estimation of total segments and proportion rated Excellent
  > # ===================================
  > rhat <- sum(yhati)/sum(Mi)
  > rhat # estimate of the proportion
  > [1] 0.0828125
  > tauhat.r <- M*rhat
  > tauhat.r # estimate of the total
  > [1] 2397.422
  > sr2 <- (1/(n-1))*sum((yhati - Mi*rhat)^2)
  > sr2
  > [1] 49.96548
  > var.tauhat.r <- (N*(N-n)*sr2)/n + (N/n)*sum((Mi*(Mi-mi)*si2)/mi)
  > sqrt(var.tauhat.r) # SE of estimate of total
  > [1] 1111.438
  > sqrt(var.tauhat.r/M^2) # SE of estimate of proportion
  > [1] 0.03839164

  SE(\( \hat{\mu}_r \)) = .038, which is about the same as with the earlier unbiased estimator (.038), as there was no real relationship between the highway length and the number of Excellent segments on the highway.
• If appropriate, a ratio estimator will generally be better if there is a lot of variation in the highway lengths (same idea as before).

Part (b) of the Highway Example: Estimate the average cost per highway mile and the total cost of urgently needed repairs. Here, we have $N = 352$ total highways, $n = 5$ sampled highways, and $M = 28950$ total 1-mile segments, and:

- $M_i =$ the number of segments for highway $i$,
- $m_i =$ the number of sampled segments for highway $i$,
- $y_i =$ the cost of repairs for highway $i$  
  (This was the # of segments rated Excellent for highway $i$ earlier)
- $\bar{y}_i =$ the average cost per segment for highway $i$,
- $M_i\bar{y}_i =$ the estimated total cost for highway $i$.

Again, we can use either the unbiased estimators of the mean and total or the ratio estimators. The ratio estimators would be expected to be better if there’s a linear relationship (through the origin) between the length of a highway and the estimated total repair costs, or, equivalently, if there’s little variation between the average repair costs per mile on the different highways.

There’s an important piece of information missing in the data table on p. 91 which we need to calculate the SE’s of our estimates for the cost data. Can you see what it is?

In the R analysis below, some values are assumed for the missing information, but we’ll also see that what values we assume makes little difference in the SE’s.

```r
> N <- 352; n <- 5; M <- 28950
> Mi <- c(85,120,47,98,34) # total no. segments on the highways sampled
> mi <- c(10,15,5,10,5) # no. of segments sampled
> yi <- c(90,110,60,100,30) # total cost on sampled segments
> si <- c(3.1,3.5,4.8,2.9,2.5)
> si2 <- si^2
>
> # Unbiased estimation of total cost of repairs and mean cost per segment
> # ==============================================================
> yhati <- (Mi/mi)*yi
> yhati # estimated total cost on each highway
[1] 765 880 564 980 204
>
> tauhat <- (N/n)*sum(yhati) # estimated total cost
> tauhat
[1] 238867.2
```
su2 <- var(yhati)
su2
[1] 94311.8
var1 <- (N*(N-n)*su2)/n # Term 1 of variance of tauhat
var2 <- (N/n)*sum((Mi*(Mi-mi)*si2)/mi) # Term 2 of variance of tauhat
c(var1,var2)
[1] 2303924100 2393449
var.tauhat <- var1 + var2
SE.tauhat <- sqrt(var.tauhat) # SE of estimate of total
SE.tauhat
[1] 48024.14
c(tauhat-qt(.975,n-1)*SE.tauhat,tauhat+qt(.975,n-1)*SE.tauhat) # 95% CI
[1] 105530.8 372203.6
muhat <- tauhat/M # estimate of mean cost per segment
muhat
[1] 8.251026
SE.muhat <- SE.tauhat/M # SE of estimate of proportion
SE.muhat
[1] 1.658865
c(muhat-qt(.975,n-1)*SE.muhat,muhat+qt(.975,n-1)*SE.muhat) # 95% CI
[1] 3.645279 12.856773
# Ratio Estimation of total cost and mean cost per segment
# ===================================
rhat <- sum(yhati)/sum(Mi)
tauhat.r <- M*rhat
tauhat.r # estimate of the total cost
[1] 255800.4
sr2 <- (1/(n-1))*sum((yhati - Mi*rhat)^2)
sr2
[1] 19283.25
var.tauhat.r <- (N*(N-n)*sr2)/n + (N/n)*sum((Mi*(Mi-mi)*si2)/mi)
sqrt(var.tauhat.r) # SE of estimate of total cost
[1] 21759.14
rhat# estimated mean cost per segment
[1] 8.835938
sqrt(var.tauhat.r/M^2)# SE of estimated mean cost per segment
[1] 0.751611
• We were not given the standard deviations of the costs for each sampled highway and we were not given the individual data values (the cost for each of the sampled sections on a highway) from which to compute them. However, we can also see that the within highway variability contributed little to the estimated variance of our estimators. If we had assumed a standard deviation of $10,000 on each highway (very high, considering that the average costs ranged from $6,000 to $12,000), the SE of the unbiased estimate of the total cost would have increased from $48,024 to only $48,214 (and would have decreased to $47,999 if all the standard deviations were 0). This points out something important in two-stage designs: it is generally the variability between the primary units and the sample size of the primary units that determines the accuracy of the estimators.

• Note that for the cost data, the ratio estimator decreased the SE’s of the estimates of the total and mean by over half. This was because there was a relationship between the lengths of the highways and the estimated total cost of repairs.

Comparison of Var(\(\hat{\tau}\)) and \(\hat{\text{Var}}(\hat{\tau})\): Recall:

\[
\text{Var}(\hat{\tau}) = N(N-n)\frac{\sigma_u^2}{n} + \frac{N}{n} \sum_{i=1}^{N} M_i(M_i - m_i) \frac{\sigma_i^2}{m_i},
\]

\[
\hat{\text{Var}}(\hat{\tau}) = N(N-n)\frac{s_u^2}{n} + \frac{N}{n} \sum_{i=1}^{N} M_i(M_i - m_i) \frac{s_i^2}{m_i}.
\]

• It was mentioned earlier that \(\hat{\text{Var}}(\hat{\tau})\) is an unbiased estimator of \(\text{Var}(\hat{\tau})\).

• Although \(s_i^2\) is an unbiased estimator of \(\sigma_i^2\), the second term in the expression for \(\hat{\text{Var}}(\hat{\tau})\) above, \(\frac{N}{n} \sum_{i=1}^{N} M_i(M_i - m_i) \frac{s_i^2}{m_i}\) is not an unbiased estimator of the second term in \(\text{Var}(\hat{\tau})\), but is an unbiased estimator of \(\sum_{i=1}^{N} M_i(M_i - m_i) \frac{\sigma_i^2}{m_i}\) without the \(N/n\) constant.

• So the 2nd piece in \(\hat{\text{Var}}(\hat{\tau})\) underestimates the corresponding 2nd piece in \(\text{Var}(\hat{\tau})\). And the 1st piece in \(\hat{\text{Var}}(\hat{\tau})\) overestimates the corresponding 1st piece in \(\text{Var}(\hat{\tau})\). Why?

• It is easy to show with an example that \(s_u^2 \geq \sigma_u^2\). How?

So \(s_u^2\) overestimates \(\sigma_u^2\) because it includes both the variability between primary units and the variability within primary units.
Allocation in Two-Stage Sampling: A practical question in developing a two-stage sampling plan is how to allocate resources to the sampling of primary units versus secondary units. Here, we consider the special case of:

1. Equal-sized primary units: \( M_1 = \ldots = M_N = \overline{M} \).
2. Equal-sized samples within primary units: \( m_1 = \ldots m_n = m \).

The total sample size then is \( mn \), and \( MN = M \) where:

- \( M \) = the number of secondary units per primary unit,
- \( N \) = the total number of primary units,
- \( M \) = the total number of secondary units.

- Under the allocation assumptions of equal-sized primary units and equal-sized samples within primary units, the unbiased and ratio estimators of the total are identical, where:

  \[
  \hat{\tau}_r = \frac{\overline{M}}{M} \bar{y} = \frac{M}{n} \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i} = \frac{N}{n} \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i} = \frac{N}{n} \frac{\sum_{i=1}^{n} \bar{y}_i}{\sum_{i=1}^{n} M_i} = \frac{N}{n} \frac{\sum_{i=1}^{n} y_i}{M} = \frac{\overline{M}}{M} \bar{y} = \hat{\tau}.
  \]

  So, we take the average of all responses and multiply it by the number of secondary units.

- Also: \( \hat{\mu} = \frac{\overline{\tau}}{M} = \overline{y} \).

Working with the variance of \( \hat{\mu} \):

\[
\text{Var}(\hat{\mu}) = \frac{\text{Var}(\hat{\tau})}{M^2} = \frac{N(N-n)}{M^2} \frac{\sigma^2_u}{n} + \frac{N}{M^2 n} \sum_{i=1}^{N} (M_i \mu_i - \overline{M} \mu)^2
\]

[Note: \( \sigma^2_u = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu_1)^2 = \frac{1}{N-1} \sum_{i=1}^{N} (M_i \mu_i - \overline{M} \mu)^2 \)]

[where: \( \sigma^2_b = \text{variability between primary units} \)]

\[
= \frac{N(N-n)}{M^2 N^2} \frac{\sigma^2_u}{n} + \frac{N \overline{M} (M - m)}{M^2 N^2 mn} \sum_{i=1}^{N} \sigma^2_i
\]

\[
= \left( \frac{N}{N-n} \right) \frac{\sigma^2_u}{n} + \left( \frac{M - m}{M} \right) \frac{1}{nm} \left( \frac{1}{N} \sum_{i=1}^{N} \sigma^2_i \right)
\]

\[
= \left( 1 - f_1 \right) \frac{\sigma^2_u}{n} + \left( 1 - f_2 \right) \frac{\sigma^2_w}{mn},
\]

where: \( \sigma^2_w \) = the average within primary unit variability, and \( f_1 \) and \( f_2 \) are the sampling fractions at the first and second stages, respectively.

Note: If we increase \( m \) (the # of secondary units), we can drive the 2nd term in the variance above to zero, but this will have no effect on the 1st term.
Goal: We want to find those values of \( n \) and \( m \) which minimize the \( \text{Var}(\hat{\mu}) \). Suppose we fix \( nm = c \); this assumes the cost is the same for all possible choices of \( n \) and \( m \).

First, note that:
\[
\text{Var}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{\sigma_b^2}{n} + \left(1 - \frac{m}{M}\right) \frac{\sigma_w^2}{nm} = \frac{\sigma_b^2}{n} - \frac{\sigma_b^2}{N} + \frac{\sigma_w^2}{nm} - \frac{\sigma_w^2}{nM},
\]
where the middle two terms are fixed for any choice of \( n, m \) (\( nm = c \)).

So, we want to minimize \( \frac{1}{n} \left( \sigma_b^2 - \frac{\sigma_w^2}{M} \right) \) with respect to \( n \).

- If \( \sigma_b^2 > \frac{\sigma_w^2}{M} \), then we should:

- So if the primary unit size, \( M \), is large (it usually is), then \( \sigma_b^2 \) will be larger.

- If \( \sigma_b^2 < \frac{\sigma_w^2}{M} \), then we should:

- All of this ignores any differences in cost.

Cost Considerations in Allocation: Suppose we let \( C = c_0 + c_1n + c_2nm \) where:

\[
\begin{align*}
    c_0 &= \text{the startup cost}, \\
    c_1 &= \text{the cost of selecting a primary unit (travel, time, etc.)}, \\
    c_2 &= \text{the cost of sampling a secondary unit once we’ve selected a primary unit}.
\end{align*}
\]

Suppose we fix the total cost \( C \). Then, the optimal allocation for \( m \) (that which minimizes \( \text{Var}(\hat{\mu}) \) for fixed \( C \)) is:

\[
m_{\text{opt}} = \sqrt{\frac{c_1\sigma_w^2}{c_2\left(\sigma_b^2 - \frac{\sigma_w^2}{M}\right)}}.
\]

- Note that this optimal choice for \( m \) does not depend in any way on the total cost \( C \).

- If \( c_1 \) increases relative to \( c_2 \), it makes sense that \( m_{\text{opt}} \) will increase, because the cost of sampling primary units increases.

- Often, if \( M \) is large, then \( \frac{\sigma_w^2}{M} \approx 0 \), and \( m_{\text{opt}} \approx \sqrt{\frac{c_1\sigma_w^2}{c_2\sigma_b^2}} \). In this case, we need only know the relative costs and relative variabilities.

Back to the Highway Example: Suppose it takes 1/2 hour to actually sample a 1-mile segment (\( c_2 \)). It might be much more costly to select a primary unit, and suppose we guess: \( \frac{c_1}{c_2} \approx 25 \).
Suppose we have preliminary data (say on 4 highways with 5 segments each) and we conduct an analysis of variance (ANOVA) to estimate the two variance components $\sigma_b^2$ and $\sigma_w^2$, given below:

$$
\sigma_b^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (\mu_i - \mu)^2 \\
\sigma_w^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2, \text{ where: } \sigma_i^2 = \frac{1}{M - 1} \sum_{j=1}^{M} (y_{ij} - \mu_i)^2.
$$

Recall that $s_b^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2$ overestimates $\sigma_b^2$.

And $s_w^2 = \frac{1}{n} \sum_{i=1}^{n} s_i^2$ is an unbiased estimate of $\sigma_w^2$ (since $s_i^2$ is unbiased for $\sigma_i^2$).

Conducting an ANOVA on the $y_{ij}$'s with the primary units (highways) as factors yields the partitioning:

$$
\sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \bar{y})^2 = m \sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2.
$$

This gives the following ANOVA table:

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>Degrees of Freedom</th>
<th>Sums of Squares</th>
<th>Mean Squares</th>
<th>E(MS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$
E(MSW) = E \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m - 1} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2 \right] = E \left[ \frac{1}{n} \sum_{i=1}^{n} s_i^2 \right] = \sigma_w^2.
$$

We want to use the expected mean squares above to estimate $\sigma_b^2$. How?

R Code to Estimate $\sigma_b^2$ and $\sigma_w^2$ via ANOVA: Reconsider the highway repair example with hypothetical data on repair costs.
> hwy <- rep(c("A","B","C","D"),rep(5,4))
> hwy
[1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "C" "C" "C" "C" "C" "C"
[16] "D" "D" "D" "D" "D"
> repcost <- c(3,6,7,9,4,12,8,14,9,10,6,8,10,7,10,5,4,8,6,6)
> repcost
[1] 3 6 7 9 4 12 8 14 9 10 6 8 10 7 10 5 4 8 6 6
> d <- data.frame(hwy,repcost)
> a <- aov(repcost~hwy,data=d) # Conducts an ANOVA of repair costs
> summary(a) # on the highway ID

Df  Sum Sq Mean Sq F value   Pr(>F)
  hwy  3  79.20 26.400  6.2485 0.005184 **
Residuals 16  67.60  4.225
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

> sigb2 <- (1/5)*(26.4 - 4.225)
> sigb2 # Estimate of sigma-b-sq
[1] 4.435
> 4.225/90 # Estimate of sigma-w-sq / Mbar
[1] 0.04694444

• Since \( \sigma^2_w / M = 0.0469 \) is small relative to \( \sigma^2_b \), and hence effectively negligible, then the approximate optimal allocation for \( m \) is given by:

\[
m_{\text{opt}} = \sqrt{\frac{c_1 \sigma^2_w}{c_2 (\sigma^2_b - \sigma^2_w)} / M} \approx \sqrt{\frac{c_1 \sigma^2_w}{c_2 \sigma^2_b}} = \sqrt{\frac{c_1 (4.225)}{c_2 (4.435)}}.
\]

• We had guessed that \( \frac{c_1}{c_2} = 25 \). Then \( m_{\text{opt}} = 4.8 \), so we might use 5 one-mile segments per highway.

• Had we guessed that \( \frac{c_1}{c_2} = 10 \), then \( m_{\text{opt}} = 3.09 \) and we would have used 3 one-mile segments per highway.

• The value of \( n \) is now determined by the overall budget (or cost). Recall that the total cost was given by: \( C = c_0 + c_1 n + c_2 nm \)

\[
\Rightarrow \quad C = c_0 + c_1 n + 5c_2 n \quad \Rightarrow \quad n = \frac{C - c_0}{c_1 + 5c_2}.
\]
PPS Sampling in Two-Stage Problems: As with cluster sampling, Hansen-Hurwitz estimation can be employed in two-stage sampling if selection of the primary units is made proportional to size with replacement. For details of the forms of the resulting unbiased estimators, see pages 148-149 in Chapter 13.

Horvitz-Thompson Estimator in Two-Stage Sampling: Recall that the Horvitz-Thompson estimator can be applied in virtually any sampling problem. Recall also that this estimator depends on the inclusion probabilities for the units in the population. How do we find these \( \pi_{ij} \), the inclusion probabilities?

The inclusion probability of the \( j^{th} \) secondary unit within the \( i^{th} \) primary unit is:

\[
\pi_{ij} = \left( \frac{n}{N} \right) \left( \frac{m_i}{M_i} \right)
\]

With these inclusion probabilities, the Horvitz-Thompson estimator of the population total is:

\[
\hat{\tau} = \frac{n}{N} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{y_{ij}}{\pi_{ij}} = \frac{N}{n} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij} = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{y}_i,
\]

which is what we found earlier as the SRS estimator.

Three-Stage Sampling: Suppose now that we sample:

- \( n \) out of \( N \) primary units
- \( m \) out of \( M \) secondary units
- \( t \) out of \( T \) tertiary units

(3 Stages)

For this three-stage sampling plan, the variance of the estimated mean is:

\[
\text{Var}(\hat{\mu}) = (1 - f_1) \frac{\sigma_1^2}{n} + (1 - f_2) \frac{\sigma_2^2}{mn} + (1 - f_3) \frac{\sigma_3^2}{nmt}
\]

\[
( f_1 = \frac{n}{N}, f_2 = \frac{m}{M}, f_3 = \frac{t}{T} )
\]

\[
\Rightarrow \text{Var}(\hat{\mu}) = (1 - f_1) \frac{s_1^2}{n} + f_1 (1 - f_2) \frac{s_2^2}{mn} + f_1 f_2 (1 - f_3) \frac{s_3^2}{nmt}.
\]

- Again, \( s_1^2 \) overestimates \( \sigma_1^2 \), \( s_2^2 \) overestimates \( \sigma_2^2 \), and \( s_3^2 \) underestimates \( \sigma_3^2 \), but \( \text{Var}(\hat{\mu}) \) is unbiased for \( \text{Var}(\hat{\mu}) \).