Chapter 7 Systematic Sampling

Definition: Given a list of \( N \) sampling units, a \( 1 \text{-in-} k \) systematic sample chooses at random one of the first \( k \) units, then takes every \( k \text{th} \) unit after that.

Example: Given \( N=100 \), suppose I want a systematic sample of size \( n=25 \). Do a \( 1 \text{-in-k} \) systematic sample.
1. Choose a random start between 1 and 4, say 3.
2. Choose every 4th unit thereafter.

Note: \( k \) should be approximately \( \frac{N}{n} \)

Examples of situations where systematic sampling is appropriate
1. Quality control at a production time
2. Exit polls
3. Customers at a checkout
4. Census does a systematic sample to send a more detailed questionnaire
5. Phone book, employee list

Exercise: What if \( N=100 \), \( n=15 \)? Conduct a \( 1 \text{-in-6} \) systematic sample? how?

Estimator of the population mean \( \mu \)

$$ \hat{\mu}_y = \bar{Y}_{sy} = \frac{\sum_{i=1}^{n} y_i}{n} $$

(7.1)

Estimated variance of \( Y_{sy} \)

$$ \hat{V}(Y_{sy}) = \frac{s^2}{n} \left( 1 - \frac{n}{N} \right) $$

(7.2)

Comment: Even though we use the same estimates for \( V(Y_{sy}) \) and \( V(Y_{SRS}) \), the true variances may not be equal.

Example: Consider the population of \( N = 469 \) private colleges and universities in the U.S. Suppose we want to estimate the average size of incoming class, by conducting a \( 1 \text{-in-10} \) systematic sample. In this situation, \( V(Y_{sy}) < V(Y_{SRS}) \) because systematic sampling prevents getting too many small schools or too many large schools.

\[
\begin{array}{c|c}
\text{small schools} & \text{large schools} \\
1 & 469 \\
2 & \\
3 & \\
\vdots & \\
469 & \\
\end{array}
\]
Example: Suppose we do a 1-in-12 systematic sampling of monthly sales, then $V(\bar{Y}_{sy}) > V(\bar{Y}_{sy})$ because some samples will contain all Decembers, others will contain no Decembers. Better if choose $k$ to avoid periodic length or fractions of periodic length.

Theory:

$$V(\bar{Y}_{sy}) = \frac{\sigma^2}{n} \left[ 1 + (n-1)\rho \right]$$

(7.4)

where $\rho$ is a measure of the correlation between pairs of elements in the systematic sample.

Terminology (p.222)

- A population is *random* if the elements are in random order.
- A population is *ordered* if the elements trend upward or downward.
- A population is *periodic* if the elements tend to cycle.

If I know that a population is ordered, and I do not want to use SRS estimates because it is does not take advantage of smaller variance, how do I get a better estimates of $\mu$ and margin of error?

### 7.6 Repeated Systematic Sampling

Consider a population of size $N = 960$. If we want a sample of size $n = 60$, then we can do a 1-in-16 systematic sample, because $\frac{960}{60} = 16$. E.g.

$$11th \rightarrow 27th \rightarrow 43rd \rightarrow ...$$

Or, we can do $n_s = 10$ repeated systematic samples of size $n' = 6$ each. Then each systematic sample is 1-in-160, because $\frac{960}{6} = 160$. Suppose that the $n_s = 10$ systematic samples yield estimates $\bar{y}_1, \bar{y}_2, ..., \bar{y}_{10}$. Note that the sample means $\bar{y}_1, \bar{y}_2, ..., \bar{y}_{10}$ are like 10 observations from a SRS. The estimator of the population mean would be

$$\hat{\mu} = \frac{\bar{y}_1 + \bar{y}_2 + \cdots + \bar{y}_{10}}{10}$$

The estimated variance of $\hat{\mu}$ is

$$\hat{V}(\hat{\mu}) = s_y^2 \left( 1 - \frac{60}{960} \right)$$

where $s_y^2 = \frac{(\bar{y}_1 - \hat{\mu})^2 + \cdots + (\bar{y}_{10} - \hat{\mu})^2}{10}.$

Estimator of the population mean $\mu$, using $n_s$ 1-in-$k'$ systematic samples

$$\hat{\mu} = \frac{\bar{y}_1 + \bar{y}_2 + \cdots + \bar{y}_{n_s}}{n_s}$$

(7.12)

Estimated variance of $\hat{\mu}$

$$\hat{V}(\hat{\mu}) = s_y^2 \left( 1 - \frac{n}{N} \right)$$

(7.13)

where $s_y^2 = \frac{\sum^{n_s}_1 (\bar{y}_i - \hat{\mu})^2}{n_s - 1}$.

Example 7.6 (p.234)
Chapter 8: Cluster sampling

–each sampling unit is a cluster or collection of elements.

Example:
Classes - a cluster of students
Blocks - cluster of households or residents
Pages - cluster of words

Reasons for cluster sampling
1. Sometimes, a frame of clusters is available rather than a frame of elements.
2. May be cheaper and faster (e.g. hand out questionnaires in a class)

Example:
The Current Population Survey (CPS) divides the US into ultimate sampling units (USU) which are a cluster of households

Notation (p.255)
N=number of clusters in the population
n=number of clusters in sample
m_i=number of elements in cluster i, i = 1, 2, ..., N.
\( \overline{m} = \frac{1}{n} \sum_{i=1}^{n} m_i \) = average cluster size for sample
M= \( \sum_{i=1}^{N} m_i \) = number of elements in population
\( \overline{M} = M/N = \frac{1}{N} \sum_{i=1}^{N} m_i \) = average cluster size for population
y_i = total measurement for all observations in the ith cluster

Example 8.1:
N = 415 city blocks
n = 25 blocks sampled
y_i = total income for block i
m_i = total number of residents in block i

Estimate per capita income or
\( \mu = \frac{\text{(total income for sample)}}{\text{(total number of residents in sample)}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} \)

Note that this is a ratio estimator with \( x_i = m_i \), i.e. the cluster size is the subsidiary variable.

Cluster estimator of the population mean \( \mu \)
\( \bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} \) (8.1)

Estimated variance of \( \bar{y} \)
\( \hat{V}(\bar{y}) = \frac{1}{(\overline{M})^2} \frac{S_r^2}{n} (1 - \frac{n}{N}) \) (8.2)
where \( S_r^2 = \sum_{i=1}^{n} (y_i - \overline{y} m_i)^2 \). If M is unknown, it may be estimated by \( \overline{m} \).

Estimator of the population total \( \tau \) (e.g. total income)
• One Way:

\[ \hat{\tau} = M\overline{y} = (\text{number of residents in population})(\text{average income per resident}) \]

Then \( \hat{V}(M\overline{y}) = M^2\hat{V}(\overline{y}) \).

• Other way: (M unknown)

\[ \hat{\tau} = N\overline{y}_t = N\frac{\sum_{i=1}^n y_i}{n} = (\text{number of clusters in population})(\text{average income per cluster}) \]

Then \( \hat{V}(N\overline{y}_t) = N^2 \frac{S^2_t}{n} \left(1 - \frac{n}{N}\right) \) where \( S^2_t = \frac{\sum(y_i - \overline{y}_t)^2}{n-1} \).

Comments:

1. Cluster sampling works best when the variability \( S^2_r \) in residuals is small. i.e.

\[ y_i \approx \hat{\mu}m_i \]

2. When choosing clusters, we want clusters to be heterogeneous (different) versus choosing strata which we want to be homogeneous (the same).

Example: Major is a good stratification variable, but a bad cluster.

3. Ideally, each cluster is representative of the population.

8.8 Cluster sampling combined with stratification

Example 8.1

<table>
<thead>
<tr>
<th>Stratum 1</th>
<th>Stratum 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_1 = 415 ) (p.256)</td>
<td>( N_2 = 168 ) (p.272)</td>
</tr>
<tr>
<td>( n_1 = 25 )</td>
<td>( n_1 = 10 )</td>
</tr>
<tr>
<td>Cluster</td>
<td>( m_i )</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
</tr>
<tr>
<td>( \overline{m}_1 = 6.04 )</td>
<td>( \overline{y}_1 = 53,160 )</td>
</tr>
</tbody>
</table>

Q: What is the estimated per-capita income?

• One-way: Separate Estimator

\[ \hat{\mu}_s = W_1\hat{\mu}_1 + W_2\hat{\mu}_2 = \left( \frac{415}{415 + 168} \right) \frac{53,160}{6.04} + \left( \frac{168}{415 + 168} \right) \frac{54,700}{4.90} = 9482 \]

\[ \hat{V}(\hat{\mu}) = W_1^2\hat{V}(\hat{\mu}_1) + W_2^2\hat{V}(\hat{\mu}_2) \]

\...
• **Other way:** Combined Estimator

\[ \hat{\mu}_c = \frac{\text{total income over all strata}}{\text{total number of residents over all strata}} = \frac{\hat{\tau}_1 + \hat{\tau}_2}{M_1 + M_2} \]

where \( M = M_1 + M_2 \) may be estimated by \( \frac{N_1 m_1 + N_2 m_2}{N_1 m_1 + N_2 m_2} \).

\[ \Rightarrow \hat{\mu}_c = \frac{N_1 \bar{y}_1 + N_2 \bar{y}_2}{N_1 \bar{m}_1 + N_2 \bar{m}_2} \quad (p.272) \]

Using formula from p.188 (Ratio estimation)

\[ \hat{V}(\hat{\mu}_c) = \frac{1}{M^2} \left\{ N_1^2 \frac{S^2_{c_1}}{n_1} \left( 1 - \frac{n_1}{N_1} \right) + N_2^2 \frac{S^2_{c_2}}{n_2} \left( 1 - \frac{n_2}{N_2} \right) \right\} \]

where \( S^2_{ci} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_c m_i)^2}{n_i - 1} \). Following calculations on p.273,

\[ 2\sqrt{\hat{V}(\hat{\mu}_c)} = \$1285 \text{ per resident.} \]

### 8.9 Cluster sampling with probabilities proportional to size

Before: choose an SRS of \( n \) clusters from \( N \) (i.e. w/equal probability)

Now: choose larger clusters with higher probability

First, some theory (Horvitz-Thompson)

**Example 1:** Draw \( n=2 \) without replacement from

\[
\begin{array}{c|cccc}
\text{Outcome} & 1 & 2 & 3 & 4 & 5 \\
\hline
\text{\( \bar{y} \)} & 1.5 & 2.0 & 3.0 & 3.5 & 4.0 \\
\text{\( p(\bar{y}) \)} & 1/10 & 1/10 & 2/10 & 2/10 & 1/10 \\
\text{element} & 1 & 2 & 3 & 4 & 5 \\
\text{\( R_i = \text{"inclusion prob. of element"} \)} & \frac{4}{10} & \frac{4}{10} & \frac{4}{10} & \frac{2}{10} & \frac{1}{10} \\
\end{array}
\]

\( \mu = 3.0 \)

\[ y_{1.5} = 2.0 \text{, } y_{2.0} = 2.5 \text{, } y_{3.0} = 3.0 \text{, } y_{3.5} = 3.5 \text{, } y_{4.0} = 4.0 \text{, } y_{4.5} = 4.5 \]

Note that \( \mu_\bar{y} = 3.0 \) so \( \bar{y} \) is unbiased.
Example 2: Suppose we create two strata and draw one element from each:

<table>
<thead>
<tr>
<th>Stratum 1</th>
<th>Stratum 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3</td>
<td>4 5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome</th>
<th>( \bar{\eta} )</th>
<th>( R_i )</th>
<th>( p(\bar{\eta}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,4)</td>
<td>2.5</td>
<td>1</td>
<td>2/6</td>
</tr>
<tr>
<td>(1,5)</td>
<td>3.0</td>
<td>2</td>
<td>2/6</td>
</tr>
<tr>
<td>(2,4)</td>
<td>3.0</td>
<td>3</td>
<td>2/6</td>
</tr>
<tr>
<td>(2,5)</td>
<td>3.5</td>
<td>4</td>
<td>3/6</td>
</tr>
<tr>
<td>(3,4)</td>
<td>3.5</td>
<td>5</td>
<td>3/6</td>
</tr>
<tr>
<td>(3,5)</td>
<td>4.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \mu_{\bar{\eta}} = 3.25 \) (biased!)

Q: How do I get an unbiased estimate?
A: Horvitz-Thompson (1952) says that an unbiased estimator for population total \( \tau \) is

\[ \hat{\tau} = \sum_{i=1}^{n} \frac{y_i}{p_i} \]

The simplest case is SRS where \( p_i = \frac{n}{N} \). Then the HT estimator is \( \hat{\tau} = \sum_{i=1}^{n} \frac{y_i}{p_i} = N \bar{\eta} \). An estimator for population mean \( \mu \) is

\[ \hat{\mu}_{HT} = \frac{1}{N} \hat{\tau} = \frac{1}{N} \sum_{i=1}^{n} \frac{y_i}{p_i} \]

Going back to our example,

<table>
<thead>
<tr>
<th>Outcome</th>
<th>( \hat{\tau}_{HT} )</th>
<th>( \hat{\mu}_{HT} = \hat{\tau}/N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,4)</td>
<td>2.5</td>
<td>11/5=2.2</td>
</tr>
<tr>
<td>(1,5)</td>
<td>3.0</td>
<td>2.6</td>
</tr>
<tr>
<td>(2,4)</td>
<td>3.0</td>
<td>2.8</td>
</tr>
<tr>
<td>(2,5)</td>
<td>3.5</td>
<td>3.2</td>
</tr>
<tr>
<td>(3,4)</td>
<td>3.5</td>
<td>3.4</td>
</tr>
<tr>
<td>(3,5)</td>
<td>4.0</td>
<td>3.8</td>
</tr>
</tbody>
</table>

\( \hat{\mu}_{HT} \)

\[ \mu = 3.0 \] (unbiased!)

Cluster sampling with PPS:

PPS implies that we select cluster \( i \) with probability \( m_i/m_1+...+m_N = \frac{m_i}{M} \) at each draw. Since we draw \( n \) times, cluster \( i \) will be included in the sample with probability \( p_i = n \frac{m_i}{M} \). Under PPS, the HT estimate of \( \tau \) will be

\[ \hat{\tau}_{pps} = \sum_{i=1}^{n} \frac{Y_i}{p_i} = \sum_{i=1}^{n} \frac{Y_i}{\frac{n}{M} \bar{\eta}} = \frac{M}{n} \sum_{i=1}^{n} \frac{Y_i}{m_i} = \frac{M}{n} \sum_{i=1}^{n} \frac{y_i}{m_i} \]

where \( \bar{\eta}_i = \frac{Y_i}{m_i} \) is the average of observations in the \( i \)th cluster. Returning to our example, \( \hat{\tau}_{pps} \) may be interpreted as follows:

\[ \bar{\eta}_i = \frac{Y_i}{m_i} = \frac{\text{total income in } i \text{th block}}{\text{number of residents in } i \text{th block}} = \text{average income per resident} \]
\[ \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{m_i} = \text{average income per resident over sample} \]

\[ \frac{M}{n} \sum_{i=1}^{n} \frac{Y_i}{m_i} = \text{estimated total income of population} \]

The estimator of population mean is then \( \hat{\mu} = \frac{\hat{\tau}}{M} \).

**Estimator of the population mean \( \mu \)**

\[ \hat{\mu}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i = \bar{\bar{y}} \quad (8.19) \]

where \( \bar{y}_i = \frac{Y_i}{m_i} \) is the mean for the \( i \)th cluster.

**Estimated variance of \( \hat{\mu}_{pps} \)**

\[ \hat{V}(\hat{\mu}_{pps}) = \frac{1}{n} S_{\bar{y}}^2 = \frac{1}{n(n-1)} \sum (\bar{y}_i - \bar{\bar{y}})^2 \quad (8.20) \]

Q: How to draw PPS? **Example p.276 Table 8.3**

\[ \begin{align*}
\bar{y}_1 &= \frac{4320}{2100} = 2.06 \\
\bar{y}_2 &= \frac{4160}{1910} = 2.18 \\
\bar{y}_3 &= \frac{5790}{3290} = 1.81
\end{align*} \]

ave = 2.0167

SD = 0.1888

\[ \hat{\mu}_{pps} = 2.0167 \]

\[ \hat{V}(\hat{\mu}_{pps}) = \frac{s^2}{n} = \frac{(0.1888)^2}{3} = .0119 \]

MOE = 2\sqrt{.0119} = .2180