Cluster Sampling

Professor Ron Fricker
Naval Postgraduate School
Monterey, California

Reading Assignment: Scheaffer, Mendenhall, Ott, & Gerow
Chapter 8
Goals for this Lecture

• Introduction to cluster sampling
  – Single-stage and multi-stage
  – Advantages and disadvantages

• Estimators
  – Means and totals

• Design effect
  – Using the design effect to determine effective sample size

• Sampling with probability proportional to size
What is Cluster Sampling?

- **Cluster sampling**: a probability sample in which each sampling unit is a collection, or cluster, of elements
  - Elements for survey occur in groups (clusters)
  - So, sampling unit is the cluster, not the element
  - Aka single stage cluster sampling
  - When sampling clusters by region, called area sampling

- There are more complicated types of cluster sampling such as **two-stage cluster sampling**
  - First select primary sampling units (PSUs) by probability sampling
  - Then within each selected PSU, sample secondary sampling units (SSUs) via probability sampling
  - Survey all units in each selected SSU
List of all 3,143 counties in U.S. (PSUs)

Draw random sample of counties

For each PSU, list all Census blocks (SSUs)

Draw random sample of blocks from each county

Draw random sample of households from each block

For each SSU, enumerate all the households

Final sample (of households)
Advantages and Disadvantages

• Advantages:
  – For some populations, cannot construct explicit sampling frame
    • But can construct a frame by cluster (area, organizational, etc.) then sample within
  – For some efforts, too expensive to conduct a SRS
    • E.g., drawing a SRS from the US population for an in-person interview

• Disadvantage: Cluster sampling generally provides less precision than SRS or stratified sampling
When To Use Cluster Sampling

• Use cluster sampling only when economically justified
  – I.e., when cost savings overcome (or require) loss in precision

• Most likely to occur when
  – Constructing a complete list-based sampling frame is difficult, expensive, or impossible
  – The population is located in natural clusters (schools, city blocks, etc.)
Example: Small Village

Frame Population

Samples of Two Distinct Blocks (Proportion "⊗" Households in Realization)

<table>
<thead>
<tr>
<th>Block Numbers</th>
<th>Proportion &quot;⊗&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2</td>
<td>19/20 = .95 *</td>
</tr>
<tr>
<td>2,3</td>
<td>17/20 = .85 *</td>
</tr>
<tr>
<td>1,3</td>
<td>16/20 = .80 *</td>
</tr>
<tr>
<td>2,4</td>
<td>13/20 = .65</td>
</tr>
<tr>
<td>1,4</td>
<td>12/20 = .60</td>
</tr>
<tr>
<td>2,6</td>
<td>11/20 = .55</td>
</tr>
<tr>
<td>1,6</td>
<td>10/20 = .50</td>
</tr>
<tr>
<td>2,5</td>
<td>10/20 = .50</td>
</tr>
<tr>
<td>3,4</td>
<td>10/20 = .50</td>
</tr>
<tr>
<td>1,5</td>
<td>9/20 = .45</td>
</tr>
<tr>
<td>3,6</td>
<td>8/20 = .40</td>
</tr>
<tr>
<td>3,5</td>
<td>7/20 = .35</td>
</tr>
<tr>
<td>4,6</td>
<td>4/20 = .20 *</td>
</tr>
<tr>
<td>4,5</td>
<td>3/20 = .15 *</td>
</tr>
<tr>
<td>5,6</td>
<td>1/20 = .05 *</td>
</tr>
</tbody>
</table>

Figure 4.4 A bird's eye view of a population of 30 "⊕" and 30 "⊗" households clustered into six city blocks, from which two blocks are selected.

Effect of Cluster Sampling

• First, note that the true proportion is 0.5
  – Further, sampling 2 blocks out of 6 produces an unbiased estimate
• Now note that 6 of the 15 possible samples had proportions > 0.8 or < 0.2
  – That is, 40 percent of the samples are “far off”
• With a SRS*, the probability of getting 16 or more of the same type of household is about 0.3 percent…

* Assuming an infinitely sized population (or sampling with replacement) to make the calculations easy.
Cluster Sampling Notation

• Lots of notation for cluster sampling:
  • \( M \) is the number of clusters in the population
  • \( m \) is the number of clusters selected via SRS
  • \( n_i \) is the number of elements in cluster \( i \)
  • \( N = \sum_{i=1}^{M} n_i \), the total number of population elements
  • \( \bar{N} = N / M \), the average cluster size in the population
  • \( \bar{n} = \frac{1}{m} \sum_{i=1}^{m} n_i \), the average cluster size in the sample
  • \( y_{ij} \) is the \( j^{th} \) observation in the \( i^{th} \) cluster
  • \( t_i = \sum_{j=1}^{n_i} y_{ij} \), the total of the \( i^{th} \) sampled cluster

2/1/13

* I’m deviating from the notation in the text here. I will carry this change through in all of the formulas to follow.
Mean Estimation Summary

- Estimator for the mean: \( \bar{y}_{cl} = \frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} t_i = \frac{1}{m\bar{n}} \sum_{i=1}^{m} t_i \)

- Variance of \( \bar{y}_{cl} \): \( \text{Var}(\bar{y}_{cl}) = \left(1 - \frac{m}{M}\right) \frac{s_r^2}{m\bar{N}^2} \)

where \( s_r^2 = \frac{1}{m-1} \sum_{i=1}^{m} (t_i - \bar{y}_{cl} n_i)^2 \)

- Key idea: Only the clusters are random

If \( \bar{N} \) not known, estimate with \( \bar{n} \)
Total Estimation Summary

- Estimator for the total: \( \hat{t} = N \bar{y}_{cl} = \frac{N}{m} \sum_{i=1}^{m} t_i = \frac{N}{mn} \sum_{i=1}^{m} t_i \)

- Variance of \( \hat{t} \):

\[
\hat{\text{Var}}(\hat{t}) = N^2 \hat{\text{Var}}(\bar{y}_{cl})
\]

\[
= N^2 \left(1 - \frac{m}{M}\right) \frac{s_r^2}{mN^2}
\]

\[
= M^2 \left(1 - \frac{m}{M}\right) \frac{s_r^2}{m}
\]

where \( s_r^2 = \frac{1}{m-1} \sum_{i=1}^{m} (t_i - \bar{y}_{cl}n_i)^2 \)
What if \( N \) is Not Known?

- Alternate estimator for the total:
  \[
  \hat{t} = M \times \text{average cluster total} = M \times \frac{1}{m} \sum_{i=1}^{m} t_i \equiv M \times \bar{y}_t
  \]

- Where now, variance of \( \hat{t} \) is
  \[
  \widehat{\text{Var}}(\hat{t}) = M^2 \left( 1 - \frac{m}{M} \right) \frac{s_t^2}{m}
  \]
  with
  \[
  s_t^2 = \frac{1}{m-1} \sum_{i=1}^{m} (t_i - \bar{y}_t)^2
  \]

- Again, note that the calculations are based on the idea that only the clusters are random
Example: NAEP

• Assume:
  – 40,000 4th grade classrooms in US
  – \( n_i = 25 \) students per classroom

• Sampling procedure:
  – Select \( m \) classrooms
  – Visit each classroom and collect data on all students
    • If \( m = 8 \), will have data on 200 students

• Note the differences from SRS
  – All groups of 200 students cannot be sampled
  – Students in each classroom more likely to be alike
Mean & Variance Computations

• Calculate the mean test score as

\[ \bar{y}_{cl} = \frac{1}{mn} \sum_{i=1}^{m} t_i = \frac{1}{8 \times 25} \sum_{i=1}^{8} \sum_{j=1}^{25} y_{ij} \]

• The variance is

\[ \text{Var}(\bar{y}_{cl}) = \left(1 - \frac{8}{40,000}\right) \frac{s_r^2}{8 \times 25^2} \approx \frac{s_r^2}{8 \times 25^2} \]

where \( s_r^2 = \left(\frac{1}{7}\right) \sum_{i=1}^{8} \left(t_i - 25 \bar{y}_{cl}\right)^2 \)
The Design Effect

• Consider 8 classrooms with mean scores of 370, 370, 375, 375, 380, 380, 390, and 390
  – So, \( s_r^2 = 39,062.5 \)

• Then \( \bar{y}_{cl} = 378.75 \) and \( \text{Var}(\bar{y}_{cl}) = 7.81 \)

• Suppose a SRS with \( n = 200 \) gives \( s^2 = 500 \) so that \( \text{Var}(\bar{y}_{SRS}) = 2.50 \)

• Design effect (\( d^2 \)) is the ratio of the variances:

\[
\hat{d}^2 = \left( \frac{\text{Var}(\bar{y}_{cl})}{\text{Var}(\bar{y}_{SRS})} \right) = \frac{7.81}{2.50} = 3.13
\]
Interpretation

- In example, the design effect says that clustering tripled the sampling variance.
- Means an increase in the standard error (and hence the confidence limits) of 77%.
  - Because $\sqrt{3.13} = 1.77$
- Says in this case we need almost twice the sample size as a SRS sample to get the same precision.
Effective Sample Size

- I like to think about design effects in terms of **effective sample size**
  - What size SRS would give the same precision as the clustered sample?

- In previous example, we had \( \sum n_i = 200 \) with \( d^2 = 3.13 \)
  - The effective sample size is \( n_{eff} = \frac{200}{3.13} = 64 \)
  - So we could have done a SRS of a sample of 64 and achieved the same precision
  - Would have meant going to (up to) \( \frac{64}{8} = 8 \) times as many sites – perhaps unaffordable
Sampling within Selected Clusters

• To reduce design effect, could sample within clusters

• In previous example, perhaps sample 10 kids per class over 20 classes
  – After some calculations that we’ll exclude here, we get $n_{eff} = 200 / 1.8 = 111$

• Sample mean estimate will be more precise
  – Comes at a cost of going to 12 more classrooms
  – Usually the between-cluster sampling costs are much more than the within-cluster sampling costs
Typically, we want to determine a sample size in terms of the number of clusters \( m \) to achieve a particular margin of error \( B \).

So, solving the following for \( m \)

\[
2 \sqrt{\left(1 - \frac{m}{M}\right) \frac{\sigma_r^2}{mN^2}} = B
\]

gives

\[
m = \frac{M \sigma_r^2}{MB^2 \bar{N}^2 / 4 + \sigma_r^2}
\]

Remember to inflate to allow for nonresponse (if relevant).
(Approximate) Sample Size Calculations for Estimating Totals

• Now solve the following for \( m \) (the number of clusters) using

\[
2M \sqrt{\left(1 - \frac{m}{M}\right) \frac{\sigma_r^2}{m}} = B
\]

which gives

\[
m = \frac{M \sigma_t^2}{B^2 / 4M + \sigma_t^2}
\]

• Again, remember to inflate to account for the nonresponse rate (if relevant)
Sampling with Probability Proportional to Size (pps)

• Sometimes it makes sense to sample clusters in proportion to their size
  – Puts more emphasis on the larger clusters where most of the observations are
    • E.g., sampling clusters equally when there are a lot of small clusters and only a few very large ones could be very inefficient

• For cluster sampling, sampling with probability proportional to size (pps) means

\[
\Pr(\text{select cluster } i) = \frac{n_i}{N}
\]
Estimation Summary With PPS

- Plugging $\pi_i = n_i/N$ into the H-T estimator gives:
  - **Mean**
    - Estimator for the mean: $\hat{\mu}_{pps} = \frac{1}{m} \sum_{i=1}^{m} \bar{y}_i$
    - Variance of $\hat{\mu}_{pps}$:
      $$\text{Var}(\hat{\mu}_{pps}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left( \bar{y}_i - \hat{\mu}_{pps} \right)^2$$
  - **Total**
    - Estimator for the total: $\hat{\tau}_{pps} = \frac{N}{m} \sum_{i=1}^{m} \bar{y}_i$
    - Variance of $\hat{\tau}_{pps}$:
      $$\text{Var}(\hat{\tau}_{pps}) = \frac{N^2}{m(m-1)} \sum_{i=1}^{m} \left( \bar{y}_i - \hat{\mu}_{pps} \right)^2$$
What We Have Covered

• Introduced cluster sampling
  – Single-stage and multi-stage
  – Advantages and disadvantages

• Developed estimators
  – Means and totals

• Discussed the design effect
  – Using the design effect to determine effective sample size

• Sampling with probability proportional to size