Chapter 6: Cluster sampling design 1

Jae-Kwang Kim

Iowa State University

Fall, 2014
1 Introduction

2 Single-Stage Cluster Sampling: Equal size case

3 Single-stage cluster sampling: General case
Setup:
1. Frame list clusters (disjoint groups of population elements)
2. Select clusters by a probability sampling.

Why cluster sampling?
1. Frame inadequacy: No way to get a list of elements (very expensive) but relatively easy or cheap to list clusters
2. Convenience: By grouping elements into “close” subgroups, you can save time or money
• Single stage cluster sampling: Sample clusters and observe all elements in each selected cluster.

• Two-stage sampling:
  • [Stage 1] Population is divided into clusters, called Primary Sampling Units (PSU's). A probability sample of PSU's is drawn.
  • [Stage 2] Each selected PSU is divided into clusters or elements, called Secondary Sampling Units (SSU's). A probability sample of SSU's is drawn in each selected PSU.

  If SSU=cluster, it is called two-stage cluster sampling.
  If SSU=element, it is called two-stage element sampling

• Multi-stage sampling: PSU, SSU, ..., USU (Ultimate Sampling Unit)
  If USU=cluster, it is called multi-stage cluster sampling.
  If USU=element, it is called multi-stage element sampling.

  Probably, don’t know \( N \). So, estimation of \( \bar{y}_U \) is more difficult.

Notation (Population level)

- \( U_i = \{1, \cdots, N_i\} \): index set of clusters in the population
- \( U_i \): the set of elements in the \( i \)-th cluster of size \( M_i \), \( i = 1, 2, \cdots, N_i \)
- \( y_{ij} \): measurement of item \( y \) at the \( j \)-th element \( (j = 1, 2, \cdots, M_i) \) in cluster \( i \), \( i = 1, 2, \cdots, N_i \).
- Population total: \( Y = \sum_{i=1}^{N_i} \sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N_i} Y_i = \sum_{i=1}^{N_i} M_i \bar{Y}_i \) where \( Y_i = \sum_{j=1}^{M_i} y_{ij} = M_i \bar{Y}_i \)
- Population size: \( N = \sum_{i=1}^{N_i} M_i \)
Notation (Sample level)

- $A_I$: index set of clusters in the sample
- $n_I = |A_I|$: the number of sampled clusters
- $A = \bigcup_{i \in A_I} U_i$: index set of elements in the sample
- $n_A = |A| = \sum_{i \in A_I} M_i$: the number of sampled elements

Usually, $n_A$ is not fixed even if $n_I$ is fixed.
Introduction

Single-Stage Cluster Sampling: Equal size case

Single-stage cluster sampling: General case
Single-Stage Cluster Sampling: Equal size case

- $M_i$ are all equal (denoted by $M = M_i$).
- Sampling design: Simple random cluster sampling
  1. Simple random sampling of $n_I$ clusters from $N_I$ clusters.
  2. Observe all the elements in the selected clusters.
- Estimation of mean:

$$
\hat{Y}_U = \frac{\hat{Y}_{HT}}{N_1 M} = \frac{1}{n_I} \frac{1}{M} \sum_{i \in A_I} \sum_{j=1}^{M} y_{ij} = \frac{1}{n_I} \sum_{i \in A_I} \bar{Y}_i
$$

- Variance

$$
Var(\hat{Y}_U) = \frac{1}{n_I} \left(1 - \frac{n_I}{N_I}\right) \frac{1}{N_I - 1} \sum_{i=1}^{N_I} (\bar{Y}_i - \bar{Y}_U)^2
$$

where $\bar{Y}_i = \sum_{j=1}^{M} y_{ij}/M$ and $\bar{Y}_U = Y/(N_I M) = \sum_{i=1}^{N_I} \bar{Y}_i / N_I$. 

Kim (ISU) Ch. 6: Cluster sampling design 1 Fall, 2014 8 / 26
### ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>D.F.</th>
<th>Sum of Squares</th>
<th>Mean S.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between cluster</td>
<td>$N_i - 1$</td>
<td>SSB</td>
<td>$S^2_b$</td>
</tr>
<tr>
<td>Within cluster</td>
<td>$N_i (M - 1)$</td>
<td>SSW</td>
<td>$S^2_w$</td>
</tr>
<tr>
<td>Total</td>
<td>$N_i M - 1$</td>
<td>SST</td>
<td>$S^2$</td>
</tr>
</tbody>
</table>

where

\[
SSB = \sum_{i=1}^{N_i} M (\bar{Y}_i - \bar{Y}_U)^2
\]

\[
SSW = \sum_{i=1}^{N_i} \sum_{j=1}^{M} (y_{ij} - \bar{Y}_i)^2
\]

\[
SST = \sum_{i=1}^{N_i} \sum_{j=1}^{M} (y_{ij} - \bar{Y}_U)^2
\]

and $\text{MSS} = \text{SS}/(\text{d.f.})$. 

---

Kim (ISU) Ch. 6: Cluster sampling design 1 Fall, 2014 9 / 26
Note that

\[
S^2 = \frac{(N_I - 1) S_b^2 + N_I (M - 1) S_w^2}{N_I M - 1} \approx \frac{S_b^2 + (M - 1) S_w^2}{M}.
\]

The variance is

\[
\text{Var} \left( \hat{\bar{Y}}_U \right) = \frac{1}{n_I M} \left( 1 - \frac{n_I}{N_I} \right) S_b^2.
\]

Note that, under simple random sampling,

\[
\text{Var}_{SRS} \left( \hat{\bar{Y}}_U \right) = \frac{1}{n_I M} \left( 1 - \frac{n_I M}{N_I M} \right) S^2,
\]

as \( n = n_I M \) is the size of the sampled elements.
Single-Stage Cluster Sampling: Equal size case

Design effect

- Want to compare the current sampling design $p(\cdot)$ with SRS of equal sample size
- Kish (1965) introduced design effect

$$\text{deff} \left( p, \hat{Y}_{HT} \right) = \frac{V_p(\hat{Y}_{HT})}{V_{SRS}(\hat{Y}_{HT})}$$

- Usage 1: Compare designs
  - If $\text{deff} > 1$, then $p(\cdot)$ is less efficient than SRS.
  - If $\text{deff} < 1$, then $p(\cdot)$ is more efficient than SRS.
Design effect (Cont’d)

- **Usage 2**: Determine sample size:
  1. Have some desired variance \( V^* \)
  2. Under SRS, you can easily find required sample size \( n^* \)
  3. Choose \( n_p^* = \text{deff} \cdot n^* \).

- Then,

  \[ V_p \left( \hat{Y}_{HT} \mid n_p^* \right) = V^*. \]

- \( n^* \) is often called *effective sample size*. It is the sample size required for the given \( V^* \) if the sample design is SRS.
Intracluster correlation coefficient

- A measure of within cluster homogeneity
- Assume $M_i = M$
- Intracluster correlation coefficient

$$\rho = \frac{\text{Cov}[y_{ij}, y_{ik} | j \neq k]}{\sqrt{V(y_{ij})} \sqrt{V(y_{ik})}}$$

$$= \frac{\sum_{i=1}^{N_i} \sum_{j \neq k} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y}) / \sum_{i=1}^{N_i} M(M - 1)}{\sum_{i=1}^{N_i} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2 / \sum_{i=1}^{N_i} M}$$
Intracluster correlation coefficient (Cont’d)

- Properties
  1. \[ \rho = 1 - \frac{M}{M - 1} \frac{SSW}{SST} = 1 - \frac{S^2_w}{S^2} \]

  Since \( 0 \leq SSW \leq SST \),
  \[ -\frac{1}{M - 1} \leq \rho \leq 1. \]

  For \( SSW = 0 \), \( \rho = 1 \): perfect homogeneity in cluster.
  For \( SSW = SST \), \( \rho = -1/(M - 1) \): perfect heterogeneity in cluster.
  Each cluster is like the whole population in terms of variability.
Intracluster correlation coefficient (Cont’d)

- Variance under simple random cluster (SRC) sampling satisfies

\[ V_{SRC}(\hat{Y}) = V_{SRS}(\hat{Y})[1 + (M - 1)\rho] \]

Thus,

\[ \text{deff} = 1 + (M - 1)\rho. \]

- Effective sample size

\[ n^* = \frac{n_A}{1 + (M - 1)\rho}. \]

For example, if \( \rho = 0.1 \) and \( M = 11 \), \( \text{deff} = 2 \). Thus, even if \( \rho \) is low, design effect can be large if \( M \) is large.
Estimation of intracluster correlation coefficient

Table: ANOVA table in the sample level (under SRC sampling)

<table>
<thead>
<tr>
<th>Source</th>
<th>D.F.</th>
<th>Sum of Squares</th>
<th>Mean SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>$n_I - 1$</td>
<td>Sample SSB (SSSB)</td>
<td>$s_b^2 = \frac{SSSB}{(n_I - 1)}$</td>
</tr>
<tr>
<td>Within</td>
<td>$n_I(M - 1)$</td>
<td>Sample SSW (SSSW)</td>
<td>$s_w^2 = \frac{SSSW}{{n_I(M - 1)}}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_IM - 1$</td>
<td>Sample SST (SSST)</td>
<td></td>
</tr>
</tbody>
</table>

Sum of Squares: Sample SST = Sample SSB + Sample SSW

Sample SSB $= \sum_{i \in A_I} M \left \{ \bar{Y}_i - \left( n_I^{-1} \sum_{k \in A_I} \bar{Y}_k \right) \right \}^2$

Sample SSW $= \sum_{i \in A_I} \sum_{j=1}^M (y_{ij} - \bar{Y}_i)^2$
Results

\[ E(s_b^2) = \frac{M}{N_i - 1} \sum_{i=1}^{N_i} (\bar{Y}_i - \bar{Y})^2 = S_b^2 \]

\[ E(s_w^2) = \frac{1}{N_i(M - 1)} \sum_{i=1}^{N_i} \sum_{j=1}^{M} (y_{ij} - \bar{Y}_i)^2 = S_w^2 \]

Estimated intracluster correlation

\[ \hat{\rho} = 1 - \frac{s_w^2}{\hat{\sigma}_y^2} \]

where

\[ \hat{\sigma}_y^2 = \frac{(N_i - 1)s_b^2 + N_i(M - 1)s_w^2}{N_iM} \]

\[ \approx \frac{1}{M} s_b^2 + \left(1 - \frac{1}{M}\right) s_w^2 \]
Introduction

Single-Stage Cluster Sampling: Equal size case

Single-stage cluster sampling: General case
Single-stage cluster sampling: General case

- **Meaning**
  1. Draw a probability sample $A_I$ from $U_I$ via $p_I(\cdot)$.
  2. Observe every elements in each selected clusters.

- **Cluster inclusion probability**
  
  \[
  \pi_{ii} = \Pr(i \in A_I) = \sum_{A_I; i \in A_I} p_I(A_I)
  \]
  \[
  \pi_{ij} = \Pr(i, j \in A_I) = \sum_{A_I; i, j \in A_I} p_I(A_I)
  \]

- **Element inclusion probability**
  
  \[
  \pi_{ik} = \Pr\{(ik) \in A\} = \Pr(i \in A_I) = \pi_{ii} \text{ where } k \in U_i
  \]
  \[
  \pi_{ik,jl} = \Pr\{(ik) \in A \& (jl) \in A\} = \begin{cases} 
  \Pr(i \in A_I) = \pi_{ii} & \text{if } i = j \\
  \Pr(i, j \in A_I) = \pi_{ij} & \text{if } i \neq j
  \end{cases}
  \]
Single-stage cluster sampling

- **HT estimation**
  1. Point estimator:

  \[
  \hat{Y}_{HT} = \sum_{i \in A_i} \frac{Y_i}{\pi_i}
  \]

  2. Variance

  \[
  \text{Var} \left( \hat{Y}_{HT} \right) = \sum_{i \in U_i} \sum_{j \in U_i} \Delta_{ij} \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j}
  \]

  where \( \Delta_{ij} = \pi_{ij} - \pi_i \pi_j \).

- **Variance estimation**

  \[
  \hat{V} \left( \hat{Y}_{HT} \right) = \sum_{i \in A_i} \sum_{j \in A_i} \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} \frac{\Delta_{ij}}{\pi_{ij}}
  \]

  provided \( \pi_{ij} > 0 \).
Remark

For fixed size design \((n_{A_i} = n_I)\),

\[
\text{Var}\left( \hat{Y}_{HT} \right) = -\frac{1}{2} \sum_{i \in U_I} \sum_{j \in U_I} \Delta_{ij} \left( \frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2
\]

1. If \(\pi_i \propto Y_i\), then \(\hat{Y}_{HT} = Y\)
2. If \(\pi_i \propto M_i\) and if \(\bar{Y}_i\) is constant, then \(\hat{Y}_{HT} = Y\).
3. Equal probability sampling design is generally inefficient.
## Optimal design

- Assume the following model

\[ y_{ij} = \mu + a_i + e_{ij} \]

where \( \mu \) is an unknown parameter, \( a_i \sim (0, \sigma_a^2) \), and \( e_{ij} \sim (0, \sigma_e^2) \).

- Total variance of \( \hat{Y}_{HT} \):

\[
V(\hat{Y}_{HT}) = V\left(\sum_{i \in A_I} \pi_{li}^{-1} M_i \mu\right) + E\left\{\sum_{i \in A_I} \pi_{li}^{-2} \gamma_i\right\}
\]

\[
= V\left(\sum_{i \in A_I} \pi_{li}^{-1} M_i \mu\right) + E\left\{\sum_{i \in U_I} \pi_{li}^{-1} \gamma_i\right\}
\]

where \( \gamma_i = V(Y_i) = M_i^2 \sigma_a^2 + M_i \sigma_e^2 \).
Optimal design (Cont’d)

- The second term of the total variance is minimized when

\[ \pi_{li} \propto \gamma_i^{1/2} = M_i \sigma_a \left( 1 + \frac{\sigma_e^2}{\sigma_a^2} \cdot \frac{1}{M_i} \right)^{1/2}. \]

- If \( \sigma_e^2 / \sigma_a^2 \) is small, then \( \pi_{li} \propto M_i \) lead to an optimal sampling design.
- If \( \sigma_e^2 / \sigma_a^2 \) is large, then \( \pi_{li} \propto M_i^{1/2} \) can be a better design.
Mean estimation

- Population size $N$ is generally unknown.
- If the parameter of interest is population mean, we may have two different concepts:

$$
\mu_1 = \frac{1}{N_I} \sum_{i=1}^{N_I} \frac{Y_i}{M_i} = \frac{1}{N_I} \sum_{i=1}^{N_I} \bar{Y}_i
$$

$$
\mu_2 = \frac{1}{N} \sum_{i=1}^{N_I} Y_i = \frac{\sum_{i=1}^{N_I} \sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{N_I} M_i}
$$

- If $M_i = M$, then $\mu_1 = \mu_2$. Otherwise, the two parameters have different meanings.
Mean estimation

- Suppose that we are interested in estimating $\bar{Y}_U = \mu_2$.
- From each sampled cluster $i$, we observe $(M_i, \bar{Y}_i)$.
- Ratio estimation

$$\hat{Y}_R = \frac{\sum_{i \in A_i} Y_i / \pi_i}{\sum_{i \in A_i} M_i / \pi_i} \equiv \frac{\hat{Y}_{HT}}{\hat{N}_{HT}}$$

- Variance

$$V \left( \hat{Y}_R \right) = \frac{1}{N^2} \left\{ V(\hat{Y}_{HT}) - 2\mu \text{Cov}(\hat{Y}_{HT}, \hat{N}_{HT}) + \mu^2 V(\hat{N}_{HT}) \right\}$$
Mean estimation

- Under SRC sampling, for example, the variance is

\[
V \left( \hat{Y}_R \right) = \frac{N_I^2}{n_I N^2} \left( 1 - \frac{n_I}{N_I} \right) \frac{1}{N_I - 1} \sum_{i=1}^{N_I} (Y_i - M_i \bar{Y}_U)^2
\]

\[
= \frac{1}{n_I \bar{M}^2} \left( 1 - \frac{n_I}{N_I} \right) \frac{1}{N_I - 1} \sum_{i=1}^{N_I} M_i^2 (\bar{Y}_i - \bar{Y}_U)^2
\]

with \( \bar{M} = N_I^{-1} \sum_{i=1}^{N_I} M_i = N/N_I \).