Chapter 7: Cluster sampling design 2: Two-stage sampling

Jae-Kwang Kim

Iowa State University

Fall, 2014
Two-stage sampling

- **Setup:**
  1. Stage 1: Draw $A_I \subset U_I$ via $p_I(\cdot)$
  2. Stage 2: For every $i \in A_I$, draw $A_i \subset U_i$ via $p_i(\cdot \mid A_i)$

Sample of elements: $A = \bigcup_{i \in A_I} A_i$

- **Some simplifying assumptions**
  1. **Invariance** of the second-stage design $p_i(\cdot \mid A_I) = p_i(\cdot)$ for every $i \in U_I$ and for every $A_I$ such that $i \in A_I$
  2. **Independence** of the second-stage design

\[
P \left( \bigcup_{i \in A_I} A_i \mid A_I \right) = \prod_{i \in A_I} Pr (A_i \mid A_I)
\]
Remark: A non-invariant design is two-phase sampling design.

1. Phase 1: Select a sample and observe $x_i$
2. Phase 2: Based on the observed value of $x_i$, the second-phase sampling design is determined. The second-phase sample is selected by the second-phase sampling design.
Notation: Sample size
- $n_I$: Number of PSU's in the sample.
- $m_i$: Number of sampled elements in $A_i$.
- $\sum_{i \in A_I} m_i = |A|$: The number of sampled elements.

Notation: Inclusion probability
- Cluster inclusion probability: $\pi_{i}$ and $\pi_{ij}$ (same as in the single-stage cluster sampling)
- Conditional inclusion probability:

$$
\pi_{k|i} = Pr[k \in A_i \mid i \in A_I] \\
\pi_{kl|i} = Pr[k, l \in A_i \mid i \in A_I] \\
\Delta_{kl|i} = \pi_{kl|i} - \pi_{k|i}\pi_{l|i}.
$$

In general, $\pi_{k|i}$ is a random variable (in the sense that it is a function of $A_I$). Under invariance, it is fixed.
Element inclusion probability

- First order inclusion probability
  \[ \pi_{ik} = Pr \{ (ik) \in A \} = Pr (k \in A_i \mid i \in A_I) Pr (i \in A_I) = \pi_{k|i} \pi_{i}. \]

- Second order inclusion probability
  \[ \pi_{ik,jl} = \begin{cases} 
  \pi_{li} \pi_{k|i} & \text{if } i = j \text{ and } k = l \\
  \pi_{li} \pi_{kl|i} & \text{if } i = j \text{ and } k \neq l \\
  \pi_{lij} \pi_{k|i} \pi_{l|j} & \text{if } i \neq j 
\end{cases} \]
1 Introduction

2 Estimation

3 Examples
HT estimation

- HT estimation for $Y = \sum_{i \in U_i} Y_i = \sum_{i \in U_i} \sum_{k \in U_i} y_{ik}$:

  $$\hat{Y}_{HT} = \sum_{i \in A_i} \frac{\hat{Y}_i}{\pi_{li}} = \sum_{i \in A_i} \sum_{k \in A_i} \frac{y_{ik}}{\pi_{k|i} \pi_{li}}$$

- Properties of $\hat{Y}_{HT}$
  1. Unbiased
  2. Variance

  $$V\left(\hat{Y}_{HT}\right) = V_{PSU} + V_{SSU}$$

  where

  $$V_{PSU} = \sum_{i \in U_i} \sum_{j \in U_i} \Delta_{ij} \frac{Y_i}{\pi_{li}} \frac{Y_j}{\pi_{lj}}$$

  $$V_{SSU} = \sum_{i \in U_i} \frac{V_i}{\pi_{li}}, \quad V_i = V\left(\hat{Y}_i \mid A_i\right) = \sum_{k \in U_i} \sum_{l \in U_i} \Delta_{kl|i} \frac{y_{ik}}{\pi_{k|i}} \frac{y_{il}}{\pi_{l|i}}.$$
Remark

1. If $A_i = U_i$, then the design is a stratified sampling. Note that $\pi_{ii} = 1$, $\pi_{ij} = 1$, and $\Delta_{ij} = 0$ for all $i, j$. Thus, $V\left(\hat{Y}_{HT}\right) = \sum_{i \in U_i} V_i / 1$.

2. If $A_i = U_i$ for every $i \in A_i$, then the design is single-stage cluster sampling and $V\left(\hat{Y}_{HT}\right) = V_{PSU}$.
Variance estimation

\[ \hat{V} \left( \hat{Y}_{HT} \right) = \hat{V}_{PSU} + \hat{V}_{SSU} \]

\[ = \sum_{i \in A_I} \sum_{j \in A_I} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\hat{Y}_i}{\pi_I} \frac{\hat{Y}_j}{\pi_I} + \sum_{i \in A_I} \frac{\hat{V}_i}{\pi_{II}}, \]

where

\[ \hat{V}_{PSU} = \sum_{i \in A_I} \sum_{j \in A_I} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\hat{Y}_i}{\pi_I} \frac{\hat{Y}_j}{\pi_I} - \sum_{i \in A_I} \frac{1}{\pi_{II}} \left( \frac{1}{\pi_I} - 1 \right) \hat{V}_i \]

\[ \hat{V}_{SSU} = \sum_{i \in A_I} \frac{\hat{V}_i}{\pi_{II}^2} \]

and \( \hat{V}_i \) satisfies \( E \left( \hat{V}_i \mid A_I \right) = V \left( \hat{Y}_i \mid A_I \right) \).
Here, we used the fact

\[ E \left( \hat{Y}_i \hat{Y}_j \mid A_l \right) = \begin{cases} Y_i Y_j & \text{if } i \neq j \\ V_i + Y_i^2 & \text{if } i = j. \end{cases} \]

by the independence of the second-stage sampling across the clusters.

Often, \( \sum_{i \in A_l} \frac{\hat{V}_i}{\pi_i} \) is ignored. (if \( n_l/N_l \equiv 0 \)).
Justification

Let

\[ \hat{V}^* = \sum_{i \in A} \sum_{j \in A} \frac{\Delta_{lij}}{\pi_{lij}} \frac{\hat{Y}_i}{\pi_i} \frac{\hat{Y}_j}{\pi_j}. \]

Then, we have

\[
E\{\hat{V}^*\} = E_1 \left\{ \sum_{i \in A} \sum_{j \in A} \frac{\Delta_{lij}}{\pi_{lij}} \frac{1}{\pi_i \pi_j} E_2(\hat{Y}_i \hat{Y}_j) \right\} \\
= E \left( \sum_{i \in A} \sum_{j \in A} \frac{\Delta_{lij}}{\pi_{lij}} \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} \right) + E \left( \sum_{i \in A} \frac{\Delta_{lii}}{\pi_{lii}} \frac{V_i}{\pi_i^2} \right) \\
= \sum_{i \in U} \sum_{j \in U} \Delta_{lij} \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} + \sum_{i \in U} \frac{\pi_{li} - \pi_{li}^2}{\pi_i^2} V_i = V(\hat{Y}_{HT}) - \sum_{i \in U} V_i.
\]

Biased downward. The bias is of order \( O(N_I) \).
Example 1

- **Two-stage sampling design**
  1. Stage One: Simple random sampling of clusters of size $n_i$ from $N_i$ clusters.
  2. Stage Two: Simple random sampling of size $m_i$ from $M_i$ elements in the sampled cluster $i$

- **HT estimator of**

$$\hat{Y} = N^{-1} \sum_{i=1}^{N_i} \sum_{j=1}^{M_i} y_{ij}, \text{ where } N = \sum_{i=1}^{N_i} M_i \text{ is assumed to be known:}$$

$$\hat{Y}_{HT} = \frac{N_i}{n_iN} \sum_{i \in A_i} \hat{Y}_i = \frac{1}{n_i \bar{M}} \sum_{i \in A_i} \frac{M_i}{m_i} \sum_{j \in A_i} y_{ij}$$

where $\bar{M} = N_i^{-1} \sum_{i=1}^{N_i} M_i.$
Example 1 (Cont’d)

- Variance

\[
V\{\hat{Y}_{HT}\} = V\left\{ \frac{1}{n_l \bar{M}} \sum_{i \in A_l} Y_i \right\} \\
+ E \left\{ \frac{1}{n_l^2 \bar{M}^2} \sum_{i \in A_l} \frac{M_i^2}{m_i} \left(1 - \frac{m_i}{M_i}\right) \frac{1}{M_i - 1} \sum_{k \in U_i} (y_{ik} - \bar{Y}_i)^2 \right\} \\
= \frac{1}{n_l} \left(1 - \frac{n_l}{N_l}\right) S_{q1}^2 + \frac{1}{n_l N_l \bar{M}^2} \sum_{i = 1}^{N_l} \frac{M_i^2}{m_i} \left(1 - \frac{m_i}{M_i}\right) S_{2i}^2
\]

where \(S_{q1}^2 = (N_l - 1)^{-1} \sum_{i = 1}^{N_l} (q_i - \bar{q}_1)^2\) with \(q_i = Y_i / \bar{M}\), \(\bar{q}_1 = N_l^{-1} \sum_{i = 1}^{N_l} q_i\), and \(S_{2i}^2 = (M_i - 1)^{-1} \sum_{k \in U_i} (y_{ik} - \bar{Y}_i)^2\).
If the sampling rate for the second stage sampling is constant such that $m_i/M_i = f_2$, then we can write

$$V\{\hat{Y}_{HT}\} = \frac{1}{n_l}(1 - f_1)S^2_{q_1} + \frac{1}{n_l\bar{m}}(1 - f_2)\frac{1}{N}\sum_{i=1}^{N_l} M_iS^2_{2i}$$

$$= \frac{1}{n_l}(1 - f_1)B^2 + \frac{1}{n_l\bar{m}}(1 - f_2)W^2$$

where $f_1 = n_l/N_l$ and $\bar{m} = N_l^{-1}\sum_{i=1}^{N_l} m_i$. 
Minimizing

$$V(\hat{Y}_{HT}) = \frac{1}{n_I} B^2 + \frac{1}{n_I \bar{m}} W^2$$

subject to

$$C = c_0 + c_1 n_I + c_2 n_I \bar{m}$$

lead to

$$\bar{m}_{opt} = \sqrt{\frac{c_1}{c_2} \times \frac{W^2}{B^2}}$$
Example 1 (Cont’d)

If $M_i$ are equal ($M_i = M$), then the following results are satisfied

- We have $B^2 = S_b^2 / M$ and $W^2 = SSW \cdot M / (M - 1) \cdot 1 / (N_i M) = S_w^2$. Thus, for sufficiently large $M$, we have $S^2 = B^2 + W^2$.

- The homogeneity measure defined by

$$\delta = \frac{B^2}{B^2 + W^2}$$

is equal to the intracluster correlation coefficient $\rho$.

- Ignoring $f_1$ term,

$$V\{\hat{Y}_{HT}\} = \frac{1}{n_i \bar{m}} \{1 + (\bar{m} - 1)\delta\} S_y^2$$

$$= V_{SRS}(\hat{Y}_{HT}) \{1 + (\bar{m} - 1)\delta\}$$
Example 1 (Cont’d)

- Even when $M_i$ are unequal, we can still express

\[
V\{\hat{Y}_{HT}\} = \frac{1}{n_i \bar{m}} \{1 + (\bar{m} - 1)\delta\} kS^2
\]

\[
= V_{SRS}(\hat{Y}_{HT}) k \{1 + (\bar{m} - 1)\delta\}
\]

where $k = (B^2 + W^2)/S^2$. Thus,

\[
deff = k \{1 + (\bar{m} - 1)\delta\}.
\]

- If $M_i = M$, then $k = 1$. Otherwise, it is greater than one.
1 Introduction

2 Estimation

3 Examples
Example 2: Special case of Example 1

- **Sampling design**
  1. Stage 1: Select SRS of clusters of size $n_I$ from a population of $N_I$ clusters.
  2. Stage 2: Select SRS of elements of size $m$ from a cluster of $M_i = M$ elements at each cluster.

- **HT estimator of**
  
  $\bar{Y} = \frac{\sum_{i=1}^{N_I} \sum_{j=1}^{M} y_{ij}}{(N_I M)}$

  $\hat{Y} = \frac{1}{n_I m} \sum_{i \in A_I} \sum_{j \in A_i} y_{ij}$
Example 2 (Cont’d)

- **Variance**

\[
V(\hat{Y}) = \left(1 - \frac{n_l}{N_l}\right) \frac{S_b^2}{n_lM} + \left(1 - \frac{m}{M}\right) \frac{S_w^2}{n_l m}
\]

Or,

\[
V(\hat{Y}) = \left(1 - \frac{n_l}{N_l}\right) \frac{S_1^2}{n_l} + \left(1 - \frac{m}{M}\right) \frac{S_2^2}{n_l m}
\]

where

\[
S_1^2 = \frac{1}{N_l - 1} \sum_{i=1}^{N_l} (\bar{Y}_i - \bar{Y})^2 = \frac{S_b^2}{M}
\]

and

\[
S_2^2 = S_w^2 = \frac{1}{N_l (M - 1)} \sum_{i=1}^{N_l} \sum_{j=1}^{M} (y_{ij} - \bar{Y}_i)^2.
\]
Example 2 (Cont’d)

- Variance components:
  \[ S^2_b = S^2 \{1 + (M - 1) \rho\} \]
  \[ S^2_w = S^2 (1 - \rho) \]

- Ignoring \( f_1 = n_I / N_I \),
  \[ V(\hat{Y}) = \frac{S^2}{n_I m} \{1 + (m - 1) \rho\} \]

Thus, Design effect \( = 1 + (m - 1) \rho \).
Example 2 (Cont’d)

- Variance estimation

\[ \hat{V}(\hat{Y}) = \left(1 - \frac{n_I}{N_I}\right) \frac{s_1^2}{n_I} + \frac{n_I}{N_I} \left(1 - \frac{m}{M}\right) \frac{s_2^2}{n_I m} \]

where

\[ s_1^2 = (n_I - 1)^{-1} \sum_{i \in A_I} \left(\bar{y}_i - \hat{Y}\right)^2 \]

\[ s_2^2 = n_I^{-1} (m - 1)^{-1} \sum_{i \in A_I} \sum_{j \in A_i} (y_{ij} - \bar{y}_i)^2 \]

and \( \bar{y}_i = \sum_{j \in A_i} y_{ij} / m. \)

- If \( n_I / N_I \neq 0, \) then \( \hat{V}(\hat{Y}) = s_1^2 / n_I. \)
Example 3: Two-stage PPS sampling

- Sampling design
  1. Stage One: PPS sampling of $n_I$ clusters with MOS = $M_i$
  2. Stage Two: SRS sampling of $m$ elements in each selected clusters.

- Estimation of mean

$$\hat{Y}_{PPS} = \frac{1}{n_I} \sum_{k=1}^{n_I} z_k$$

where $z_k = \frac{\hat{t}_i}{M_i}$ if cluster $i$ is selected in the $k$-th PPS sampling and

$$\hat{t}_i = \frac{M_i}{m} \sum_{j \in A_i} y_{ij}.$$
Example 3 (Cont’d)

- We can express
  \[
  \hat{Y}_{PPS} = \frac{1}{n_I m} \sum_{i \in A_I} \sum_{j \in A_i} y_{ij}.
  \]

  Self-weighting design: equal weights

- Variance estimation
  \[
  \hat{V} \left( \hat{Y}_{PPS} \right) = \frac{1}{n_I} s_z^2
  \]
  where
  \[
  s_z^2 = \frac{1}{n_I - 1} \sum_{k=1}^{n_I} (z_k - \bar{z}_n)^2
  \]
  and \( z_k = \hat{t}_i / M_i = \hat{Y}_i \) if cluster \( i \) is selected in the \( k \)-th PPS sampling.

- Very popular in practice.
Example 3 (Cont’d)

- Very popular in practice for the following reasons:
  1. $p_i \propto M_i$: efficient
  2. Point estimation is easy (self-weighting)
  3. Interviewer workloads are equal
  5. Works for multi-stage sampling design.