Chapter 2: Horvitz-Thompson estimation

Jae-Kwang Kim

Iowa State University

Spring, 2013
1 Introduction

2 Basic setup

3 Simple random sampling
Sampling frame
- list frame
- area frame

Unit
- Sampling unit
- Reporting unit = Observational unit = Element

Two types of sampling
- Element sampling
- Cluster sampling

Parameter of interest: \( Y = \sum_{i=1}^{N} y_i \) (population total of \( y \))
1 Introduction

2 Basic setup

3 Simple random sampling
Definition

1. First order inclusion probability:
   \[ \pi_i = \Pr (i \in A) = \sum_{A; \ i \in A} P(A) \]

2. Second order inclusion probability:
   \[ \pi_{ij} = \Pr (i, j \in A) = \sum_{A; \ i, j \in A} P(A) \]

3. Probability sampling design: \( \pi_i > 0, \ \forall i \in U \)

4. Measurable sampling design: \( \pi_{ij} > 0 \ \forall i, j \in U \).
Remark

- $I_k$: indicator random variable with

\[ I_k = I_k(A) = \begin{cases} 
1 & \text{if } k \in A \\
0 & \text{if } k \not\in A 
\end{cases} \]

Note that

\[ E(I_k) = \pi_k \]
\[ E(I_k I_l) = \pi_{kl} \]
\[ V(I_k) = \pi_k (1 - \pi_k) \]
\[ C(I_k, I_l) = \pi_{kl} - \pi_k \pi_l \equiv \Delta_{kl} \]

- $n_A = \sum_{k=1}^{N} I_k(A)$: (realized) sample size. If $n_A$ does not depend on $A$, then it is fixed in the sense that $V(n_A) = 0$. 
Example (Bernoulli Sampling)

- Each unit is selected or not selected according to the outcome of a Bernoulli trial with inclusion probability $\pi_k = \pi$.
- Let $\epsilon_1, \epsilon_2, \cdots, \epsilon_N \sim iid \, Uniform(0, 1)$. If $\epsilon_k \leq \pi$ then accept unit $k$ in the sample. If $\epsilon_k > \pi$ then do not accept unit $k$ in the sample.
- Sample size $n_s$ : Binomial random variable.

$$Pr(n_A = x) = \binom{N}{x} \pi^x (1 - \pi)^{N-x} I_{\{0,1,2,\cdots,N\}}(x).$$

Thus,

$$E(n_A) = N\pi$$
$$V(n_A) = N\pi (1 - \pi)$$

- Sampling design

$$P(A) = \pi^{n_A} (1 - \pi)^{N-n_A}$$

where $n_A = |A|$.
- Inclusion probabilities
Lemma

(Properties of the inclusion probabilities)

1. \( \pi_{ii} = \pi_i \) and \( \pi_{ij} = \pi_{ji} \).

2. For a sampling design with (expected) sample size \( n \),
   \[
   \sum_{i=1}^{N} \pi_i = n.
   \]

3. For fixed sample size design \( (n_A = n) \),
   \[
   \sum_{i=1}^{N} \pi_{ij} = n\pi_j
   \]
   and, for \( \Delta_{ij} = \pi_{ij} - \pi_i\pi_j \),
   \[
   \sum_{i=1}^{N} \Delta_{ij} = 0.
   \]
Proof
Definition

Horvitz-Thompson estimator of \( Y = \sum_{i=1}^{N} y_i \):

\[
\hat{Y}_{HT} = \sum_{i \in A} \frac{y_i}{\pi_i}
\]

It is also called \( \pi \)-estimator.
Theorem
(Properties of HT estimator)

1. Unbiased

\[ E \left( \hat{Y}_{HT} \right) = Y \]

2. Variance

\[ \text{Var} \left( \hat{Y}_{HT} \right) = \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \frac{y_i \ y_j}{\pi_i \ \pi_j} \]

3. For fixed fixed sample size design \((V (n_s) = 0)\),

\[ \text{Var} \left( \hat{Y}_{HT} \right) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \]

This formula is called Sen-Yates-Grundy (SYG) variance formula.
Consider the following sampling design from a finite population $U = \{1, 2, 3\}$.

<table>
<thead>
<tr>
<th>Sample $(A)$</th>
<th>$Pr (A)$</th>
<th>HT estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1 = {1, 2}$</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>$A_2 = {1, 3}$</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>$A_3 = {2, 3}$</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>

1. Compute the first-order inclusion probability of each element in the population.
2. Find the HT estimator for each sample.
3. Check that the HT estimator is unbiased.
Basic setup

Variance estimation

- Unbiased variance estimator: want to find a statistic $\hat{V}$ such that $E(\hat{V}) = \text{Var}(\hat{Y}_{HT})$.

Idea:

If $Q = \sum_{i=1}^{N} \sum_{j=1}^{N} \Omega_{ij} y_i y_j$, then $\hat{Q} = \sum_{i \in A} \sum_{j \in A} \pi_{ij}^{-1} \Omega_{ij} y_i y_j$ is an unbiased estimator of $Q$.

HT variance estimator:

$$\hat{V} = \sum_{i \in A} \sum_{j \in A} \left( \pi_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} - \pi_{ij} \right)$$

SYG variance estimator (for fixed-size design):

$$\hat{V}_{SYG} = -\frac{1}{2} \sum_{i \in A} \sum_{j \in A} \left( \pi_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} - \pi_{ij} \right)^2$$

Under what condition does an unbiased variance estimator of HT estimator exist?
Consider the following sampling design from a finite population $U = \{1, 2, 3\}$ with $y_1 = 16$, $y_2 = 21$, $y_3 = 18$.

<table>
<thead>
<tr>
<th>Sample (A)</th>
<th>$Pr (A)$</th>
<th>HT estimate</th>
<th>HT var. est.</th>
<th>SYG var. est.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1 = {1, 2}$</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A_2 = {1, 3}$</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A_3 = {2, 3}$</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Check the unbiasedness of the variance estimates.
Remark

1. For unbiased estimation, HT estimator can be used if the sampling design is a probability sampling design.
2. When the HT estimator is efficient (has smaller variance)?
   1. Note
      \[ \text{Var} \left( \hat{Y}_{HT} \right) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \]
      When the variance is small? (That is, how to choose \( \pi_k \) to make the variance small?)
   2. We don’t know \( y_k \) in advance.
   3. Hence, if \( x_k \) is correlated with \( y_k \), then we can choose \( \pi_k \propto x_k \).
   4. On the other hand, HT estimator works poorly if \( \pi_k \) is not correlated with \( y_k \). (Basu’s elephant example)
Basu’s elephant example

- Circus with $N=50$ elephants. Want to estimate the total weights of the elephants using a sample of size $n = 1$
- About three years ago, every elephant is weighted and “Sambo” was in the middle in terms of the weight. (and “Jumbo” was the largest one.)
- Circus owner’s idea: measure Sambo’s weight and multiply it by 50.
- Statistician: No! It’s not a probability sampling.
- Circus owner: Well, what is your sampling scheme?
- Statistician: Let’s select Sambo with high probability. Say, select Sambo with probability $99/100$, and select the other 49 elephants with probability $(1/49)(1/100)$. 
Basic setup

Basu’s elephant example (Cont’d)

- Circus owner: OK. Let’s select one with this scheme. (Sambo is selected.) OK. Let’s multiply 50 to sambo’s weight.
- Statistician: No! You should multiply the inverse of the first order inclusion probability. So, you should multiply by 100/99, not by 50.
- Circus owner: ????? What if Jumbo was selected? What number should we multiply?
- Statistician: Well, it is 4,900.
- Circus owner: What??? You are fired!
Remark

- HT estimator is not location invariant.

Definition

Write $\hat{\theta} = \hat{\theta}(A) = \hat{\theta}(y_i; i \in A)$. $\hat{\theta}$ is location invariant if

$$\hat{\theta}(a + y_i; i \in A) = a + \hat{\theta}(y_i; i \in A)$$

for all $a$.

Thus, changing from “Celsius” to “Fahrenheit”: the estimates is not the same.
1 Introduction

2 Basic setup

3 Simple random sampling
Simple Random Sampling

**Motivation:** Choose \( n \) units from \( N \) units without replacement.

1. Each subsect of \( n \) distinct units is equally likely to be selected.
2. There are \( \binom{N}{n} \) samples of size \( n \) from \( N \).
3. Give equal probability of selection to each subset with \( n \) units.

**Definition**

Sampling design for SRS:

\[
P(A) = \begin{cases} 
1 / \binom{N}{n} & \text{if } |A| = n \\
0 & \text{otherwise.}
\end{cases}
\]
Lemma

Under SRS, the inclusion probabilities are

\[ \pi_i = \frac{n}{N} \]
\[ \pi_{ij} = \frac{n(n-1)}{N(N-1)} \quad \text{for } i \neq j. \]
**Theorem**

*Under SRS design, the HT estimator*

\[
\hat{Y}_{HT} = \frac{N}{n} \sum_{i \in A} y_i = N \bar{y}
\]

*is unbiased for Y and has variance of the form*

\[
V\left(\hat{Y}_{HT}\right) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S^2
\]

*where*

\[
S^2 = \frac{1}{2} \frac{1}{N} \frac{1}{N-1} \sum_{i=1}^{N} \sum_{j=1}^{N} (y_i - y_j)^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2.
\]

*Also, the SYG variance estimator is*

\[
\hat{V}\left(\hat{Y}_{HT}\right) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) s^2
\]
Proof
Remark (under SRS)

- $1 - n/N$ is often called the finite population correction (FPC) term. The FPC term can be ignored ($\text{FPC} \approx 1$) if the sampling rate $n/N$ is small ($\leq 0.05$) or for conservative inference.

- For $n = 1$, the variance of the sample mean is

$$\frac{1}{n} \left(1 - \frac{n}{N}\right) S^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{Y})^2 \equiv \sigma_Y^2$$

- Central limit theorem: under some conditions,

$$\sqrt{n} \left(\hat{Y}_{HT} - Y\right) = \frac{\bar{y} - \bar{Y}}{\sqrt{\frac{1}{n} \left(1 - \frac{n}{N}\right) S^2}} \rightarrow N(0, 1).$$
Remark (under SRS)

Sample size determination

1. Choose the target variance $V^*$ of $V(\bar{y})$.
2. Choose $n$ the smallest integer satisfying

$$\frac{1}{n} \left(1 - \frac{n}{N}\right) S^2 \leq V^*.$$