Survey samples from a finite population

Let $P = \{1, \ldots, N\}$ be a finite population of interest.
For each $i \in P$, let $y_i$ be a value of interest associated with unit $i$.

Let $s = \{i_1, \ldots, i_n\}$ be a subset of distinct elements of $P$, which is a sample selected with selection probability $p(s)$, where $p$ is known (sampling plan or sampling design).

The value $y_i$ is observed if and only if $i \in s$.

If $p(s)$ is constant, the sampling plan is called the simple random sampling without replacement.

Consider the estimation of $Y = \sum_{i=1}^{N} y_i$, the population total as the parameter of interest.

Issues to study

- How do we find an unbiased estimator of $Y$? Is $Y$ estimable?
- What is the variance of an unbiased estimator of $Y$?
- Is there a UMVUE under some conditions?
Theorem 3.15.

Define

$$\pi_i = \text{probability that } i \in s, \quad i = 1, \ldots, N.$$  

(i) (Horvitz-Thompson). If $\pi_i > 0$ for $i = 1, \ldots, N$ and $\pi_i$ is known when $i \in s$, then $\hat{Y}_{ht} = \sum_{i \in s} y_i / \pi_i$ is an unbiased estimator of the population total $Y$.

(ii) Define

$$\pi_{ij} = \text{probability that } i \in s \text{ and } j \in s, \quad i = 1, \ldots, N, j = 1, \ldots, N.$$  

Then

$$\text{Var}(\hat{Y}_{ht}) = \sum_{i=1}^{N} \frac{1 - \pi_i}{\pi_i} y_i^2 + 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j$$  

\hspace{1cm} (1)

$$= \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\pi_i \pi_j - \pi_{ij}) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2.$$  

\hspace{1cm} (2)

- Horvitz-Thompson’s idea: inverse probability weighting
- Extension: $P$ is a sample of size $N$ and $y_i$ is missing if $i \notin s$
Proof.

(i) Let \( a_i = 1 \) if \( i \in s \) and \( a_i = 0 \) if \( i \notin s \), \( i = 1, \ldots, N \).
Then \( E(a_i) = \pi_i \) and
\[
E(\hat{Y}_{ht}) = E\left( \sum_{i=1}^{N} \frac{a_i y_i}{\pi_i} \right) = \sum_{i=1}^{N} y_i = Y.
\]

(ii) Since \( a_i^2 = a_i \),
\[
\text{Var}(a_i) = E(a_i) - [E(a_i)]^2 = \pi_i (1 - \pi_i).
\]
\[
\text{Cov}(a_i, a_j) = E(a_i a_j) - E(a_i) E(a_j) = \pi_{ij} - \pi_i \pi_j, \quad i \neq j.
\]
Then
\[
\text{Var}(\hat{Y}_{ht}) = \text{Var}\left( \sum_{i=1}^{N} \frac{a_i y_i}{\pi_i} \right)
= \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2} \text{Var}(a_i) + 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{y_i y_j}{\pi_i \pi_j} \text{Cov}(a_i, a_j)
= \sum_{i=1}^{N} \frac{1 - \pi_i}{\pi_i} y_i^2 + 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j.
\]
Proof (continued)

Hence (1) follows.
To show (2), note that

\[
\sum_{i=1}^{N} \pi_i = n \quad \text{and} \quad \sum_{j=1, \ldots, N, j \neq i} \pi_{ij} = (n-1)\pi_i,
\]

which implies

\[
\sum_{j=1, \ldots, N, j \neq i} (\pi_{ij} - \pi_i \pi_j) = (n-1)\pi_i - \pi_i(n-\pi_i) = -\pi_i(1-\pi_i).
\]

Hence

\[
\sum_{i=1}^{N} \frac{1 - \pi_i}{\pi_i} y_i^2 = \sum_{i=1}^{N} \sum_{j=1, \ldots, N, j \neq i} (\pi_i \pi_j - \pi_{ij}) \frac{y_i^2}{\pi_i^2}
= \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\pi_i \pi_j - \pi_{ij}) \left( \frac{y_i^2}{\pi_i^2} + \frac{y_j^2}{\pi_j^2} \right)
\]

and, (2) follows from (1).
How do we get an unbiased estimator of $\text{Var}(\hat{Y}_{ht})$?

Using Horvitz-Thompson’s idea, the following estimators are unbiased:

$$
\begin{align*}
\nu_1 &= \sum_{i \in s} \frac{1 - \pi_i}{\pi_i^2} y_i^2 + 2 \sum_{i \in s} \sum_{j \in s, j > i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} y_i y_j \\
\nu_2 &= \sum_{i \in s} \sum_{j \in s, j > i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2.
\end{align*}
$$

**Simple random sampling**

For simple random sampling,

$$
\begin{align*}
\pi_i &= E(a_i) = P(a_i = 1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} \\
\pi_{ij} &= E(a_i a_j) = P(a_i = 1, a_j = 1) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)} \\
\hat{Y}_{ht} &= \frac{N}{n} \sum_{i \in s} y_i = \frac{N}{n} \sum_{i=1}^{N} a_i y_i = N(\text{the sample mean})
\end{align*}
$$
Simple random sampling

\[
\text{Var}(\hat{Y}_{ht}) = \sum_{i=1}^{N} \frac{1 - \frac{n}{N}}{\frac{n}{N}} y_i^2 + 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} y_i y_j
\]

\[
= \frac{N-n}{n} \sum_{i=1}^{N} y_i^2 - \frac{2(N-n)}{n(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} y_i y_j
\]

\[
= \frac{N-n}{n} \left[ \sum_{i=1}^{N} y_i^2 - \frac{1}{N-1} \sum_{i \neq j} y_i y_j \right]
\]

\[
= \frac{N-n}{n} \frac{N}{N-1} \sum_{i=1}^{N} \left( y_i - \frac{Y}{N} \right)^2
\]

\[
= \left( 1 - \frac{n}{N} \right) \frac{N^2}{n} \frac{1}{N-1} \sum_{i=1}^{N} \left( y_i - \frac{Y}{N} \right)^2.
\]

Note that \( \frac{n}{N} \) is called the finite sample fraction and \( 1 - \frac{n}{N} \) is called the finite sample correction.
UMVUE under simple random sampling

We now show that $\hat{Y}_{ht}$ is in fact the UMVUE of $Y$ under simple random sampling.

Let $X = (X_i, i \in s)$ be the vector such that

$$P(X_1 = y_{i_1}, ..., X_n = y_{i_n}) = \frac{p(s)}{n!}$$

Let $\mathcal{Y}$ be the range of $y_i$, $\theta = (y_1, ..., y_N)$ and $\Theta = \prod_{i=1}^{N} \mathcal{Y}$.

Under simple random sampling, the population under consideration is a parametric family indexed by $\theta \in \Theta$.

Theorem 3.13 (Watson-Royall theorem)

(i) If $p(s) > 0$ for all $s$, then the vector of order statistics $X_{(1)} \leq \cdots \leq X_{(n)}$ is complete for $\theta \in \Theta$.

(ii) Under simple random sampling, the vector of order statistics is sufficient for $\theta \in \Theta$.

(iii) Under simple random sampling, for any estimable function of $\theta$, its unique UMVUE is the unbiased estimator $g(X_1, ..., X_n)$, where $g$ is symmetric in its $n$ arguments.
Proof.

(i) Let $h(X)$ be a function of the order statistics. Then $h$ is symmetric in its $n$ arguments. We need to show that if

$$E[h(X)] = \sum_{s=\{i_1,...,i_n\} \subset \{1,...,N\}} p(s) h(y_{i_1},...,y_{i_n}) / n! = 0 \quad (3)$$

for all $\theta \in \Theta$, then $h(y_{i_1},...,y_{i_n}) = 0$ for all $y_{i_1},...,y_{i_n}$.

First, suppose that all $N$ elements of $\theta$ are equal to $a \in \mathcal{Y}$. Then (3) implies $h(a,...,a) = 0$.

Next, suppose that $N - 1$ elements in $\theta$ are equal to $a$ and one is $b > a$. Then (3) reduces to

$$q_1 h(a,...,a) + q_2 h(a,...,a,b),$$

where $q_1$ and $q_2$ are some known numbers in $(0,1)$. Since $h(a,...,a) = 0$ and $q_2 \neq 0$, $h(a,...,a,b) = 0$.

Using the same argument, we can show that $h(a,...,a,b,...,b) = 0$ for any $k$ $a$’s and $n-k$ $b$’s.
Proof (continued)

Suppose next that elements of \( \theta \) are equal to \( a, b, \) or \( c \), \( a < b < c \).
Then we can show that \( h(a, ..., a, b, ..., b, c, ..., c) = 0 \) for any \( k \) \( a \)'s, \( l \) \( b \)'s, and \( n − k − l \) \( c \)'s.
Continuing inductively, we see that \( h(y_1, ..., y_n) = 0 \) for all possible \( y_1, ..., y_n \).
This completes the proof of (i).

(ii) The result follows from the factorization theorem (Theorem 2.2), the fact that \( p(s) \) is constant under simple random sampling, and

\[
P(X_1 = y_{i_1}, ..., X_n = y_{i_n}) = P(X_{(1)} = y_{(i_1)}, ..., X_{(n)} = y_{(i_n)})/n!,
\]

where \( y_{(i_1)} \leq ... \leq y_{(i_n)} \) are the ordered values of \( y_{i_1}, ..., y_{i_n} \).

(iii) The result follows directly from (i) and (ii).
Remark

It is interesting to note the following two issues.

- Although we have a parametric problem under simple random sampling, the sufficient and complete statistic is the same as that in a nonparametric problem (Example 2.17).
- For the completeness of the order statistics, we do not need the assumption of simple random sampling.

Example 3.19.

Under simple random sampling, \( \hat{Y}_{ht} = N \bar{X} \) is unbiased for \( Y \). Since \( \hat{Y}_{ht} \) is symmetric in its arguments, it is the UMVUE of \( Y \). We now derive the UMVUE for \( \text{Var}(\hat{Y}_{ht}) \).

From the previous discussion,

\[
\text{Var}(\hat{Y}_{ht}) = \frac{N^2}{n} \left( 1 - \frac{n}{N} \right) \sigma^2
\]

where
Example 3.19

\[
\sigma^2 = \frac{1}{N - 1} \sum_{i=1}^{N} \left( y_i - \frac{Y}{N} \right)^2 .
\]

It can be shown (exercise) that \( E(S^2) = \sigma^2 \), where \( S^2 \) is the usual sample variance

\[
S^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n - 1} \sum_{i \in s} \left( y_i - \frac{\hat{Y}_{ht}}{N} \right)^2 .
\]

Since \( S^2 \) is symmetric in its arguments,

\[
\frac{N^2}{n} \left( 1 - \frac{n}{N} \right) S^2
\]

is the UMVUE of \( \text{Var}(\hat{Y}_{ht}) \).