Special Topic: Bayesian Finite Population Survey Sampling

Sudipto Banerjee

Division of Biostatistics
School of Public Health
University of Minnesota

September 15, 2010

Estimating the population mean

Let $Y_I$ be the collection of units selected in the sample; we often write $Y_I$ as $y = (y_1, \ldots, y_n)$. Let $Y_I'$ be the mean of $Y_I'$. Then $\bar{Y} = f \bar{y} + (1 - f) \bar{Y}_I'$; $f = \frac{n}{N}$ is the sampling fraction.

Simple random sampling

Simple Random Sampling With Replacement (SRSWR): We draw a unit from the population, put it back and draw again.

Simple Random Sampling Without Replacement (SRSWOR): We draw a unit from the population, keep it aside and draw again.

SRSWR leads to binomial distribution; chances of units being drawn remain the same between draws.

SRSWOR leads to hypergeometric distribution: chances of units being drawn change from draw to draw.

Key questions:
- What are the random variables?
- What are the parameters?

Key questions answered:
- What are the random variables? Answer: The $I_i$'s.
- What are the parameters? Answer: The $Y_i$'s.

For SRSWOR: $E[I_i] = P(I_i = 1) = \frac{\binom{N - 1}{n - 1}}{\binom{N}{n}} = \frac{n}{N}$ for each $i$.

Unbiasedness of the sample mean $\bar{y}$: Write $\bar{y} = \frac{1}{n} \sum_{i=1}^{N} I_i Y_i$ and see:

$$E[\bar{y}] = \frac{1}{n} \sum_{i=1}^{N} E[I_i] Y_i = \frac{1}{n} \sum_{i=1}^{N} \frac{n}{N} Y_i = \bar{Y}.$$

For SRSWR: $\bar{y} = \frac{1}{n} \sum_{i=1}^{N} W_i Y_i$, where $W_i$ is the number of times unit $i$ appears in the sample. Then, $W_i \sim Binom(n, \frac{1}{N})$, so $E[W_i] = \frac{n}{N}$, hence

$$E[\bar{y}] = \frac{1}{n} \sum_{i=1}^{N} E[W_i] Y_i = \frac{1}{n} \sum_{i=1}^{N} W_i Y_i = \bar{Y}.$$
Bayesian modelling

- We need a prior for each population unit: \(Y_i\)
- Suppose \(Y_i \sim N(\mu, \sigma^2)\)
- Let us first assume \(\sigma^2\) is known and \(P(\mu) \propto 1\).
- What is “seen”? \(D = (Y_1, 1)\)
- Note that \(P(D | \mu, \sigma^2) \propto P(Y_1 | \mu, \sigma^2) \propto P(\mu | D, \sigma^2)\); so likelihood is product of \(n\) independent \(N(\mu, \sigma^2)\).
- We need the posterior distribution
  \[
P(\mu | D, \sigma^2) = \frac{P(Y_1 | \mu, \sigma^2)P(\mu | D, \sigma^2)}{P(D)}
  \]
- Let us take a closer look at \(P(\bar{Y} | D, \sigma^2)\).
- Observe that \(\bar{Y} = f\bar{y} + (1 - f)\bar{Y}_I\).
- Given \(D, \bar{y}\) is fixed; so the posterior distribution of \(\bar{Y}_I\) will determine the posterior distribution of \(\bar{Y}\).
- So, what is the posterior distribution \(P(\bar{Y}_I | D, \sigma^2)\)?
  \[
P(\bar{Y}_I | D, \sigma^2) = \int P(\bar{Y}_I | \mu, \sigma^2)P(\mu | D, \sigma^2)d\mu,
  \]
- where we use \(P(\bar{Y}_I | \mu, \sigma^2) = P(\bar{Y}_I | \mu, \sigma^2)\). Note \(P(\bar{Y}_I | \mu, \sigma^2) = N\left(\mu, \frac{\sigma^2}{n^{1/2}}\right)\).
- Can we simplify this posterior without integrating?

Bayesian modelling

- Note that the marginal posterior distribution
  \[
P(\sigma^2 | D) = IG\left(\frac{n - 1}{2}, \frac{n - 1}{2}s^2\right),
  \]
  where \(s^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2\) is the sample variance.
- We can perform exact posterior sampling as follows. For \(l = 1, \ldots, L\) do the following:
  - Draw \(\sigma_{(l)}^2 \sim IG\left(\frac{n - 1}{2}, \frac{n - 1}{2}s^2\right)\)
  - Draw \(\mu(\cdot) \sim N\left(\bar{y}, \frac{s^2}{n} (1 - f)\right)\)
  - Draw \(Y_{i(l)}^\circ \sim N\left(\mu(\cdot), \sigma_{(l)}^2\right)\)
  - Set \(Y_{(l)}^\circ = f\bar{y} + (1 - f)Y_{i(l)}^\circ\).
- The resulting collection \(\{Y_{(l)}^\circ\}_{l=1}^{L}\) is a sample from the posterior distribution of the population mean.

Bayesian computation

- Note that the marginal posterior distribution
  \[
P(\mu, \sigma^2 | D) = \frac{1}{\text{det}A} \cdot \text{Beta}(\frac{n - 1}{2}, \frac{n - 1}{2}) \cdot \text{normal}(\mu | \bar{y}, \frac{s^2}{n}) \cdot \text{gamma}(\sigma^2 | \frac{n - 1}{2}, \frac{n - 1}{2}s^2),
  \]
  where \(A = \begin{bmatrix} 2 & 0 & \frac{n - 1}{2}s^2 \n & 2 & n - 1 \frac{n - 1}{2} s^2 \end{bmatrix}\).
- We can perform MCMC sampling as follows. For \(l = 1, \ldots, L\) do the following:
  - Draw \(\sigma_{(l)}^2 \sim IG\left(\frac{n - 1}{2}, \frac{n - 1}{2}s^2\right)\)
  - Draw \(\mu(\cdot) \sim N\left(\bar{y}, \frac{s^2}{n} (1 - f)\right)\)
  - Draw \(Y_{i(l)}^\circ \sim N\left(\mu(\cdot), \sigma_{(l)}^2\right)\)
  - Set \(Y_{(l)}^\circ = f\bar{y} + (1 - f)Y_{i(l)}^\circ\).
- The resulting collection \(\{Y_{(l)}^\circ\}_{l=1}^{L}\) is a sample from the posterior distribution of the population mean.

Homework

A simple random sample of 30 households was drawn from a township containing 576 households. The numbers of persons per household in the sample were as follows:

\[
\begin{align*}
5, 6, 4, 3, 3, 2, 3, 3, 4, 4, 3, 2, 7, 4, 3, 5, 4, 3, 3, 3, 2, 4, 3, 2, 4, 4
\end{align*}
\]

Use a non-informative Bayesian analysis to estimate the total number of people in the area and compute the posterior probability that the population total lies within 10% of the sample estimate.

From a list of 464 small 3-year colleges in the North-Eastern United States, a simple sample of 100 colleges was drawn. Data for the number of students (\(x\)) and the number of teachers (\(y\)) for these colleges were summarized as follows:

\[
\begin{align*}
\text{The total number of students in the sample was: } & 4,679 \\
\text{The total number of teachers in the sample was: } & 1,073 \\
\text{Also given are the sample sums of squares: } & \sum_{i=1}^{100} x_i^2 = 29,841, 219 \text{ and } \\
& \sum_{i=1}^{100} y_i^2 = 111,000.
\end{align*}
\]

Assuming a non-informative Bayesian setting and where the population of students and teachers are independent, find the posterior mean and 95% credible interval for the student-teacher ratio in the population (i.e., all the 464 colleges combined).
A general theory for Bayesian missing data analysis

Let $y = (y_1, \ldots, y_N)$ be the units in the population. We assume a likelihood for this population, say $p(y | \theta)$, where $\theta$ are the parameters in the model.

Let $I = (I_1, \ldots, I_N)$ be the set of “inclusion indicators”.

Let us denote by “obs” the index set $\{i : I_i = 1\}$ and by “mis” the index set $\{i : I_i = 0\}$.

So, $y_{obs}$ is the set of observed data points and $y_{mis}$ is for the unobserved data. We will often write $y = (y_{obs}, y_{mis})$.

We will break the joint probability model into two parts:

- The model for the “complete data” $p(y | \theta)$ – including observed and missing data
- The model for the inclusion vector $I$, say $p(I | y, \phi)$

We define the complete-likelihood as:

$$p(y, I | \theta, \phi) = p(y | \theta) p(I | y, \phi)$$

The joint posterior distribution is

$$p(\theta, \phi | y_{obs}, I) \propto p(\theta, \phi) p(y_{obs}, I | \theta, \phi)$$

These integrals are evaluated by drawing samples $(y_{mis}, \theta, \phi)$ from $p(y_{mis}, \theta, \phi | y_{obs}, I)$

Typically, we first compute samples $(\theta^{(l)}, \phi^{(l)})$ from $p(\theta, \phi | y_{obs}, I)$

Then we draw $y_{mis}^{(l)} \sim p(y_{mis} | y_{obs}, I, \theta^{(l)}, \phi^{(l)})$

Computations often simplify from assumptions:

- $p(I | y, \phi) = p(I | y_{obs}, \phi)$
- $p(I | y, \phi) = p(I | \phi) = p(I)$
- $p(\phi | \theta) = p(\phi)$
- $p(y_{mis} | \theta^{(l)}, y_{obs}) = p(y_{mis} | \theta^{(l)})$. 

The “complete-likelihood” is not the data likelihood as it involves the missing data component as well.

The actual information available is $(y_{obs}, I)$. So the data likelihood is:

$$p(y_{obs}, I | \theta, \phi) = \int p(y, I | \theta, \phi) dy_{mis}$$

The joint posterior distribution is

$$p(\theta, \phi | y_{obs}, I) \propto p(\theta, \phi) p(y_{obs}, I | \theta, \phi)$$

$$\propto p(\theta, \phi) \int p(y | \theta) p(I | y, \phi) dy_{mis}$$