Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 7: Models for Complex Surveys

(Stratification)

University of Michigan

General Setup

$I: \text{ Inclusion indicator}$

$Y: \text{ Survey Variables}$

$Z: \text{ Design Variables}$

Prior

Sampling Mechanism

$$Model: \Pr(Y, I \mid Z) = \Pr(Y \mid Z)\Pr(I \mid Y, Z)$$

$$\boxed{\Pr(Y \mid Z) = \int \Pr(Y \mid Z, \theta)\pi(\theta \mid Z)d\theta}$$

Design variables include Weights, Clustering, Stratification.

There may be additional auxiliary variables not part of the design but predictive of Y and available for all subjects in the population

$$Observed\ Data: (Y_{inc}, I, Z)$$

Goal

Posterior (predictive) distribution:

$$\Pr(Y_{exc} \mid Y_{inc}, Z, I)$$

Two Stage construction:

$$\pi(\theta \mid Y_{inc}, Z, I)$$

$$\Pr(Y_{exc} \mid \theta, Z)$$

Particular Cases

- Scenario 1

$$\Pr(Y \mid Z) = \Pr(Y)$$

$$\Pr(I \mid Y, Z) = \Pr(I \mid Z)$$

- Scenario 2

$$\Pr(Y \mid Z)$$

$$\Pr(I \mid Y, Z) = \Pr(I \mid Z)$$

- Scenario 3

$$\Pr(Y \mid Z)$$

$$\Pr(I \mid Y, Z) = \Pr(I \mid Y_{inc}, Z)$$

Ignorable Sampling Mechanisms

- Scenario 4

$$\Pr(Y \mid Z)$$

$$\Pr(I \mid Y, Z)$$

Nonignorable Sampling Mechanism

Stratified Random Sample Design

- Population Setup

$$Z = \{1, 2, \ldots, H\}$$

$$Y_h = \{Y_{1h}, Y_{2h}, \ldots, Y_{N_h h}\}, h = 1, 2, \ldots, H$$

$$N = \sum_{h=1}^{H} N_h$$

- Model or Prior
  - Exchangeable within stratum (indexing within a stratum is arbitrary)

$$\prod_{h=1}^{H} \Pr(Y_{1h}, Y_{2h}, \ldots, Y_{N_h h}) = \prod_{h=1}^{H} \int \prod_{i=1}^{N_h} \Pr(Y_{ih} \mid \theta_h) \pi(\theta_h) d\theta_h$$

Examples

- Binary Outcome

$$
\begin{aligned}
&Y_{ih} \mid \theta_h \ : \ Bern(1, \theta_h) \\
&\theta_h \ : \ Beta(a_h, b_h) \\
&a_h, b_h : Known \\
&h = 1, 2, \mathrm{K} \ , H
\end{aligned}
$$

- Continuous (Normal) Outcome

$$
\begin{aligned}
&Y_{ih} \mid \mu_h, \sigma_h \ : \ N(\mu_h, \sigma_h^2) \\
&\pi(\mu_h, \sigma_h^2) \propto \\
&\left(\sigma_h^2\right)^{-d_h/2} \exp\left[-\frac{1}{2}\left(\frac{c_h}{\sigma_h^2} + \frac{b_h(\mu_h - a_h)^2}{\sigma_h^2}\right)\right] \\
&a_h, b_h, c_h, d_h : Known
\end{aligned}
$$

Numerical Example

- Binary Outcome (Yes/No)
- Number of Strata: 4
- Population sizes 44, 116,48 and 47
- Sample sizes: 9, 23, 10 and 9
- Number reporting Yes: 2, 8, 5 and 7
- Goal: Infer about the population total number of Yeses
- Approach: Fill-in 35, 93, 38 and 38 unobserved values in 4 strata

Example (Continued)

- Assume Jeffereys' prior: $a_h = b_h = 1/2$
- 4 posterior distributions

$$\theta_1 : Beta(1.5, 6.5) \qquad \theta_2 : Beta(7.5, 14.5)$$

$$Y_{exc,1} : Bin(35, \theta_1) \qquad Y_{exc,2} : Bin(93, \theta_2)$$

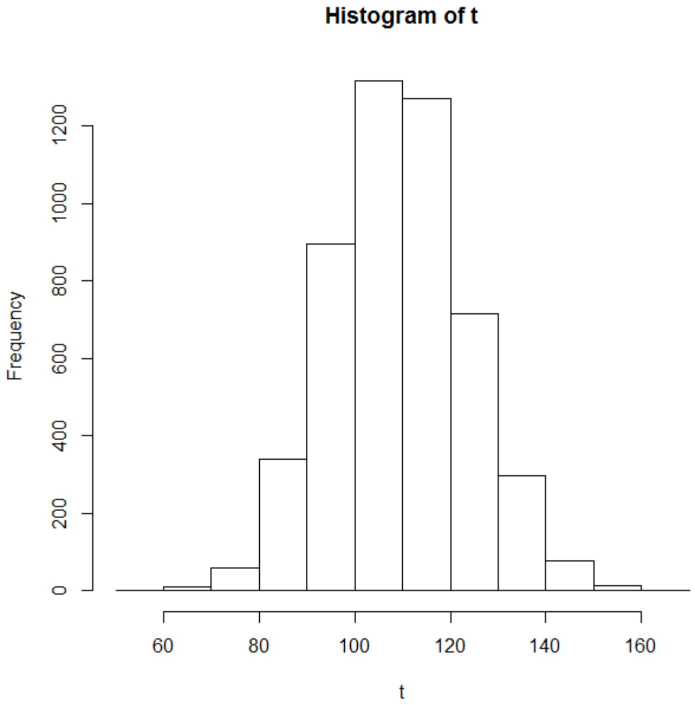
$$\theta_3 : Beta(4.5, 4.5) \qquad \theta_4 : Beta(6.5, 1.5)$$

$$Y_{exc,3} : Bin(38, \theta_3) \qquad Y_{exc,4} : Bin(38, \theta_4)$$

$$T = 2 + 8 + 5 + 7 + Y_{exc,1} + Y_{exc,2} + Y_{exc,3} + Y_{exc,4}$$

R-Code and Results

```
theta1=rbeta(5000,1.5,6.5)
yexc1=rbinom(5000,35,theta1)
theta2=rbeta(5000,7.5,14.5)
yexc2=rbinom(5000,93,theta2)
theta3=rbeta(5000,4.5,4.5)
yexc3=rbinom(5000,38,theta3)
theta4=rbeta(5000,6.5,1.5)
yexc4=rbinom(5000,38,theta4)
t= 22+yexc1+yexc2+yexc3+yexc4
```



Mean=109.9336, SD=14.2269
95% Equal tail credible interval

(82,138)

```
library(HDInterval)
hdi(t)
```

(82,137)

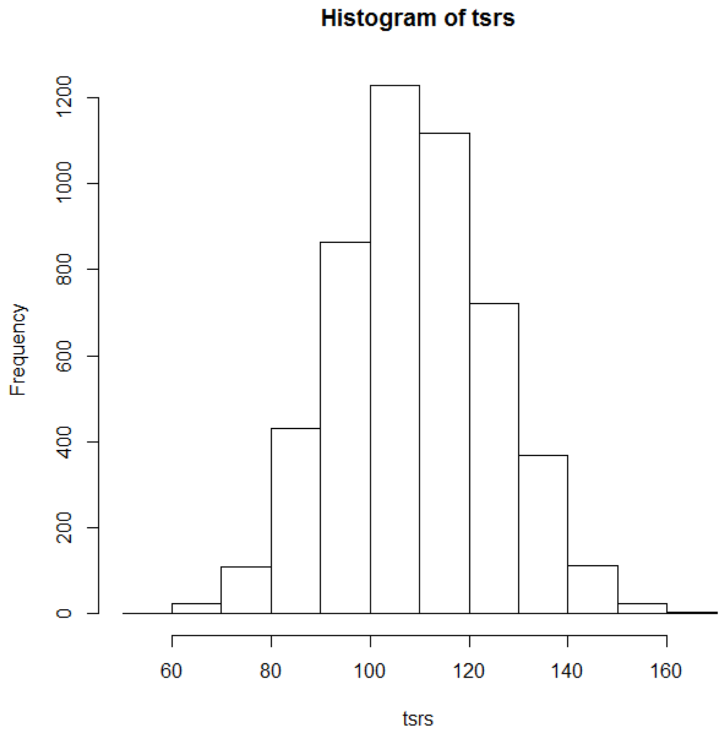
Ignoring Stratification?

$$\theta : \; Beta(21.5, 28.5)$$

$$Y_{exc} : \; Bin(204, \theta)$$

$$t = 22 + Y_{exc}$$

HPD interval: (77,138)



Analysis Using a Missing Data Package

- Z: Variable with 4 categories: 1,2,3 and 4
- Y:
  - For Z=1: 2 1's, 7 0's, 35 missing
  - For Z=2: 8 1's, 15 0's, 93 missing
  - For Z=3: 5 1's, 5 0's, 38 missing
  - For Z=4: 7 1's, 2 0's, 38 missing
- Logistic regression model: Y on Z (3 dummy variables) to multiply impute the missing values
- Compute the total from each completed data set

Normal Continuous

$$Y_{ih} \mid \mu_h, \sigma_h \sim iid \;\; N(\mu_h, \sigma_h^2)$$

$$\pi(\mu_h, \sigma_h^2) \propto \sigma_h^{-2}$$

$$W_h = N_h / N$$

$$Q = \sum_{h=1}^{H} W_h \overline{Y}_h$$

$$\overline{Y}_h \mid Y_{inc,h} \sim t_{n_h-1}(\overline{y}_h, (1 - f_h) s_h^2 / n_h)$$

$$f_h = n_h / N_h$$

$$Q \mid Y_{inc} \sim \sum_{h=1}^{H} W_h t_{n_h-1}$$

Generalization

- So far Exchangeability within Strata (in the models for outcomes) and Independence across strata (no connections across parameters)
- Connection across strata parameters

$$Y_{ih} \mid \theta_h \sim iid \ \Pr(Y_{ih} \mid \theta_h)$$

$$\pi(\theta_1, \theta_2, \cdots, \theta_H)$$

- Exchangeability of strata indices implies

$$\pi(\theta_1, \theta_2, \cdots, \theta_H) = \int \left(\prod_{h=1}^{H} \pi(\theta_h \mid \lambda) \right) \pi(\lambda) d\lambda$$

(Random effect models)