Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 10: Cluster Sample Design

UNIVERSITY OF MICHIGAN

Two stage sampling

- Most practical sample designs involve selecting a cluster of units and measure a subset of units within the selected cluster
- Two stage sample is very efficient and cost effective
- Sampling Indicators are correlated which then leads to statistics dependent on multiple units within the same cluster
- Why is this important for Bayesian Inference? How is this different from Stratified sampling?

Ex 4. Two-stage samples

- Sample design:
  - Stage 1: Sample c clusters from C clusters
  - Stage 2: Sample k_i units from the selected cluster $i=1,2,\ldots,c$

$$K_i = \text{Population size of cluster } i$$

$$N = \sum_{i=1}^{C} K_i$$

- Estimand of interest: Population mean Q
- Infer about excluded clusters and excluded units within the selected clusters

Models for two-stage samples

- Model for observables

$$Y_{ij} \sim N(\mu_i, \sigma^2); i = 1,...,C; j = 1,2,...,K_i$$

$$\mu_i \sim iid\ N(\theta, \tau^2)$$

$$Assume\ \sigma\ and\ \tau\ are\ known$$

- Prior distribution

$$\pi(\theta) \propto 1$$

- Joint model for observations within cluster as well as joint model for cluster means

Estimand of interest and inference strategy

- The population mean can be decomposed as

$$NQ = \sum_{i=1}^{c}[k_i \bar{y}_i + (K_i - k_i)\bar{Y}_{i,\text{exc}}] + \sum_{i=c+1}^{C} K_i \bar{Y}_i$$

- Posterior mean given Y_{inc}

$$E(NQ \mid Y_{\text{inc}}, \mu_i, i = 1, 2, ..., c; \theta) = \sum_{i=1}^{c}[k_i \bar{y}_i + (K_i - k_i)\mu_i] + \sum_{i=c+1}^{C} K_i \theta$$

$$E(NQ \mid Y_{\text{inc}}) = \sum_{i=1}^{c}[k_i \bar{y}_i + (K_i - k_i)E(\mu_i \mid Y_{\text{inc}})] + \sum_{i=c+1}^{C} K_i E(\theta \mid Y_{\text{inc}})$$

where $E(\mu_i \mid Y_{\text{inc}}) = \dfrac{\bar{y}_i \times (k_i / \sigma^2) + \hat{\theta} \times (1/\tau^2)}{k_i / \sigma^2 + 1/\tau^2}$

$$\hat{\theta} = E(\theta \mid Y_{\text{inc}}) = \frac{\sum_i \bar{y}_i / (\tau^2 + \sigma^2 / k_i)}{\sum_i 1/(\tau^2 + \sigma^2 / k_i)}$$

Posterior Variance

- Posterior variance can be easily computed

$$Var(NQ \mid Y_{\text{inc}}) = \sum_{i=1}^{c} (K_i - k_i)(\sigma^2 + (K_i - k_i)\tau^2) + \sum_{i=c+1}^{C} K_i(\sigma^2 + K_i\tau^2)$$

$$Var(\bar{Y}_{i,\text{exc}} \mid Y_{\text{inc}}) = E[Var(\bar{Y}_{i,\text{exc}} \mid Y_{\text{inc}}, \mu_i) \mid Y_{\text{inc}}] + Var[E(\bar{Y}_{i,\text{exc}} \mid Y_{\text{inc}}, \mu_i) \mid Y_{\text{inc}}]$$

$$= \frac{\sigma^2}{K_i - k_i} + \tau^2, i = 1, 2, \cdots, c$$

$$Var(\bar{Y}_i \mid Y_{\text{inc}}) = E[Var(\bar{Y}_i \mid Y_{\text{inc}}, \mu_i) \mid Y_{\text{inc}}] + Var[E(\bar{Y}_i \mid Y_{\text{inc}}, \mu_i) \mid Y_{\text{inc}}]$$

$$= \sigma^2 / K_i + \tau^2, i = c+1, c+2, \cdots, C$$

Inference with unknown σ and τ

- For unknown σ and τ
  - Option 1: Plug in maximum likelihood estimates. These can be obtained using PROC MIXED in SAS. PROC MIXED actually gives estimates of θ, σ, τ and $E(\mu_\iota | Y_{\text{inc}})$ (Empirical Bayes)
  - Option 2: Fully Bayes with additional prior

$$\pi(\theta, \sigma^2, \tau^2) \propto \sigma^{-2} \tau^{-2-v} \exp\left(-b / (2\tau^2)\right)$$

where b and v are small positive numbers

Extensions and Applications

- Relaxing equal variance assumption

$$Y_{il} \sim N(\mu_i, \sigma_i^2)$$

$$(\mu_i, \log \sigma_i) \sim \text{iid } BVN(\theta, \Omega)$$

- Incorporating covariates (generalization of ratio and regression estimates)

$$Y_{il} \sim N(x_{il}\beta_i, \sigma_i^2)$$

$$(\beta_i, \log \sigma_i) \sim \text{iid } MVN(\theta, \Sigma)$$

- Small Area estimation. An application of the hierarchical model. Here the quantity of interest is

$$E(\bar{Y}_i \mid Y_{\text{inc}}) = (k_i \bar{y}_i + (K_i - k_i) E(\bar{Y}_{i,\text{exc}} \mid Y_{\text{inc}})) / K_i$$

Extensions

- Relaxing normal assumptions

$$Y_{il} \mid \mu_i \sim \text{Glim}(\mu_i = h(x_{il}\beta_i), \sigma^2 v(\mu_i))$$

$$v: \text{ a known function}$$

$$\beta_i \sim iid\ \ MVN(\theta, \Omega)$$

- Incorporate design features such as stratification and weighting by modeling explicitly the sampling mechanism.

Non-parametric Bayes

- Working Model
  - Bayesian bootstrap (refer to Module 2) to generate nonsampled clusters
  - Use Weighted Polya-posterior to generate nonampled units within each cluster
  - Separate process for each stratum
  - Repeat to generate several pseudo-populations
- Use Target model to compute the population quantity of interest from each pseudo- population
- For details see Dong, Elliott and Raghunathan (2014) and Zhou, Elliott and Raghunathan (2016)

Summary

- Bayes inference for surveys must incorporate design features appropriately
- Stratification and clustering can be incorporated in Bayes inference through design variables
- Unlike design-based inference, Bayes inference is not asymptotic, and delivers good frequentist properties in small samples