Bayesian Inference for Sample Surveys

Trivellore Raghunathan (Raghu)
Director, Survey Research Center
Professor of Biostatistics
University of Michigan

Distinctive features of survey inference

1. Primary focus on descriptive finite population quantities, like overall or subgroup means or totals
   - Bayes – which naturally concerns predictive distributions -- is particularly suited to inference about such quantities, since they require predicting the values of variables for non-sampled items
   - This finite population perspective is useful even for analytical quantities:

θ = model parameter (meaningful only in context of the model)

$\tilde{\theta}(Y)$ = "estimate" of θ from fitting model to whole population Y

(a finite population quantity, exists regardless of validity of model)

A good estimate of θ should be a good estimate of $\tilde{\theta}$

(if not, then what's being estimated?)

Distinctive features of survey inference

2. Analysis needs to account for "complex" sampling design features such as stratification, differential probabilities of selection, multistage sampling.
   - Samplers reject theoretical arguments suggesting such design features can be ignored if the model is correctly specified.
   - Models are always misspecified, and model answers are suspect even when model misspecification is not easily detected by model checks (Kish & Frankel 1974, Holt, Smith & Winter 1980, Hansen, Madow & Tepping 1983, Pfeffermann & Holmes (1985).
   - Design features like clustering and stratification can and should be explicitly incorporated in the model to avoid sensitivity of inference to model misspecification.

Distinctive features of survey inference

3. A production environment that precludes detailed modeling.
   - Careful modeling is often perceived as "too much work" in a production environment (e.g. Efron 1986).
   - Some attention to model fit is needed to do any good statistics
   - “Off-the-shelf" Bayesian models can be developed that incorporate survey sample design features, and for a given problem the computation of the posterior distribution is prescriptive, via Bayes Theorem.
   - This aspect would be aided by a Bayesian software package focused on survey applications.

Distinctive features of survey inference

4. Antipathy towards methods/models that involve strong subjective elements or assumptions.
   - Government agencies need to be viewed as objective and shielded from policy biases.
   - Addressed by using models that make relatively weak assumptions, and noninformative priors that are dominated by the likelihood.
   - The latter yields Bayesian inferences that are often similar to superpopulation modeling, with the usual differences of interpretation of probability statements.
   - Bayes provides superior inference in small samples (e.g. small area estimation)

Distinctive features of survey inference

5. Concern about repeated sampling (frequentist) properties of the inference.
   - Calibrated Bayes: models should be chosen to have good frequentist properties
   - This requires incorporating design features in the model (Little 2004, 2006).

Survey Inference Setup

$Z = (Z_1, ..., Z_N) =$ design variables, known for population

$Y = (Y_1, ..., Y_N)$ = population values,
recorded only for sample

$Q = Q(Y, Z)$ = target finite population quantity

$I = (I_1, \ldots, I_N)$ = Sample Inclusion Indicators

$$I_i = \begin{cases} 1, & \text{unit included in sample} \\ 0, & \text{otherwise} \end{cases}$$

$Y_{\text{inc}} = Y_{\text{inc}}(I) =$ part of Y included in the survey

$Y = (Y_{\text{inc}}, Y_{\text{exc}})$

I	Z	Y
1		
1		Y_{inc}
1		
0		
0		
0		$[Y_{\text{exc}}]$
0		
0		

Models

- Joint distribution of *(Y, I)* conditional on *Z*
- Two approaches

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

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

- Typically

(Sampling mechanism does not depend on the survey outcomes) $\rightarrow$ $\Pr(I \mid Y, Z) = \Pr(I \mid Z)$

(Same substantive model applies to both sampled and nonsampled Subjects) $\rightarrow$ $\Pr(Y \mid I, Z) = \Pr(Y \mid Z)$

Model Specification

- Indices used to identify subjects in the population (conditional on Z) is assumed to be arbitrary
- Exchangeable joint distribution

$$\Pr(Y_1, Y_2, \cdots, Y_N \mid Z) \equiv \Pr(Y_{i_1}, Y_{i_2}, \cdots, Y_{i_N} \mid Z)$$

$$(i_1, i_2, \cdots, i_N) \text{ is a permutation of } (1, 2, \cdots, N)$$

- Exchangeable distribution are of the form

$$Y_i \mid Z, \theta \sim independent$$
$$\pi(\theta) \equiv prior$$

Examples

- Assume SRS and no Z, binary Y

$$Y_i \mid \theta \sim iid\ \ Bern(1,\theta), i = 1, 2, \cdots, N$$
$$\theta \sim Beta(a,b); a, b\ known$$

- Z: H Strata, SRS within stratum, Continuous Y

$$Y_{ih} \mid Z = h \sim iid\ \ N(\mu_h, \sigma_h^2)$$
$$\pi(\mu_h, \log\sigma_h) \sim BVN$$

- Cluster sampling, Count Y

$$Y_{ic} \sim iid\ \ Poisson(\lambda_c), i = 1, 2, \cdots, N_c$$
$$\log\lambda_c \sim iid\ \ N(\mu, \sigma^2), c = 1, 2, \cdots, C$$
$$\pi(\mu, \log\sigma) \sim BVN$$

Inference

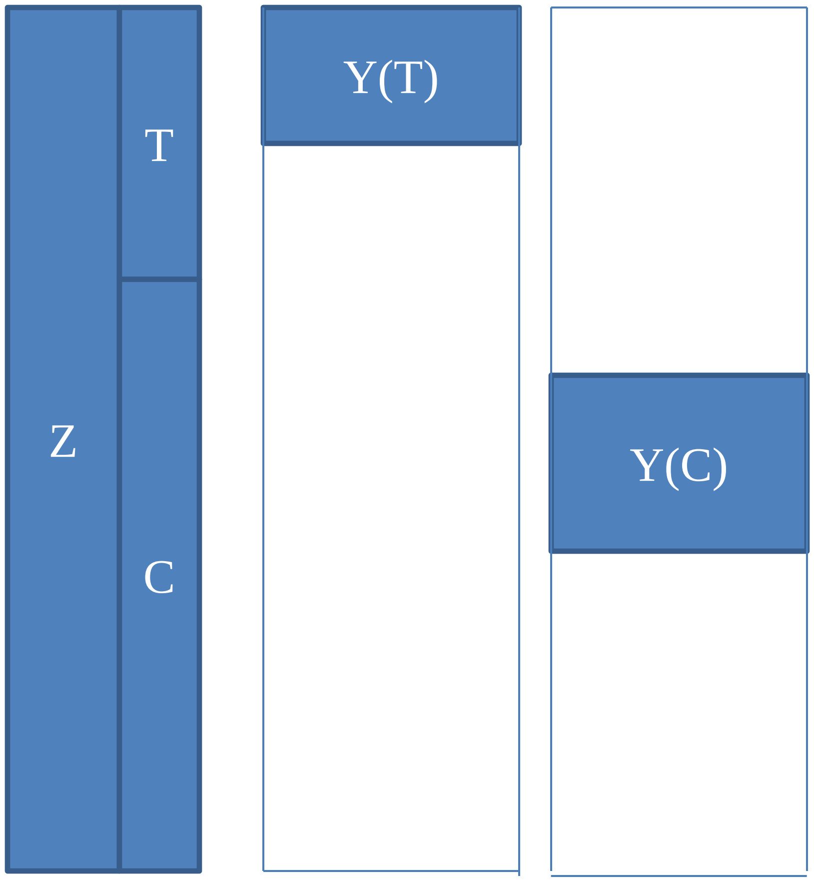
- Observed data: $\{Y_{inc}, Z, I\}$
- Unobserved or missing data: Y_{exc}
- Model: $\Pr(Y \mid Z)$
- Inference: $\Pr(Y_{exc} \mid Z, I, Y_{inc})$
- Goal: Simulate copies of Y_{exc} by drawing from the above predictive distribution and compute the estimand of interest *Q(Y,Z)*
- Multiple Imputation of Missing Values or create synthetic populations

Example

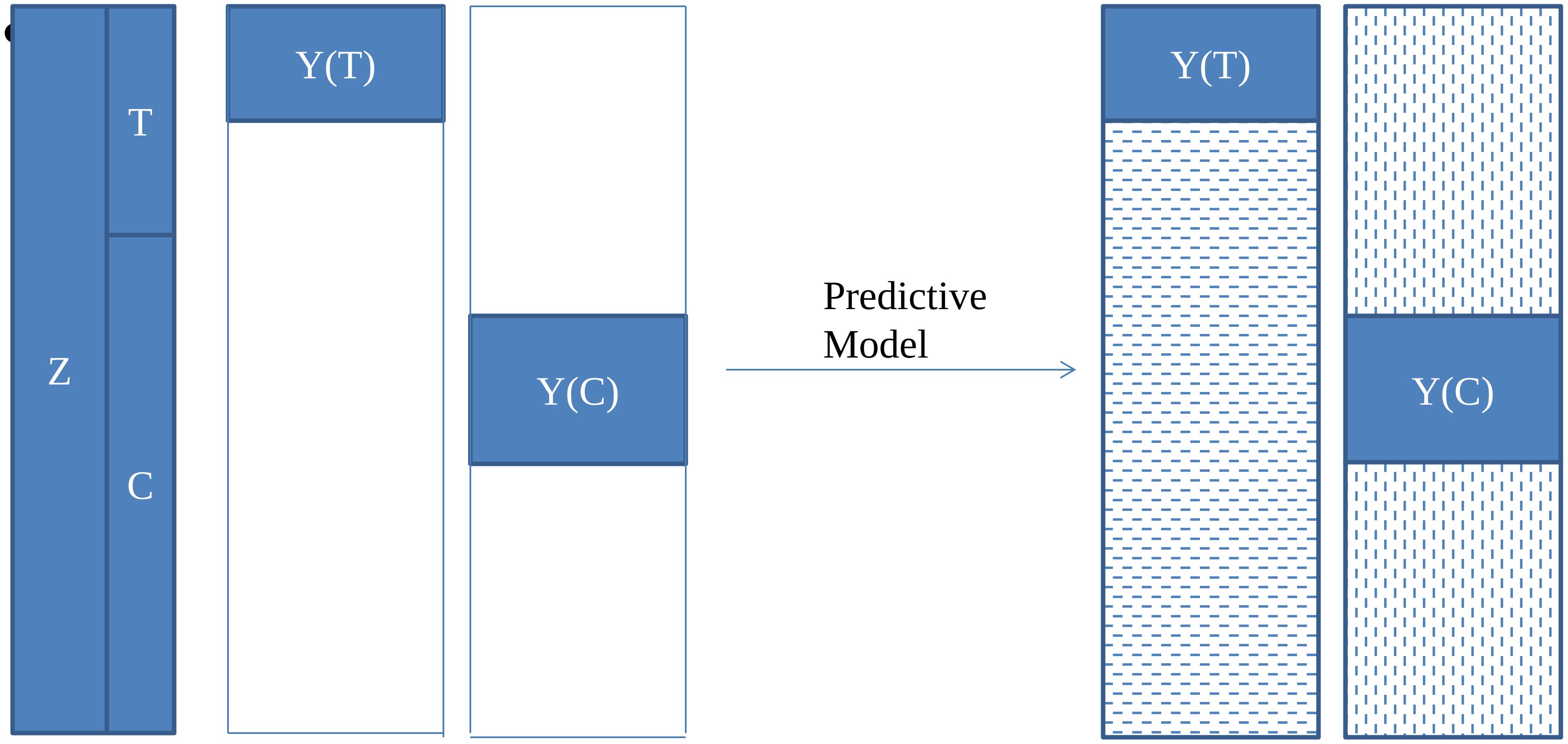
- Housing and Children Study to evaluate the effect of providing housing voucher on child development
- Population: All applicants for voucher
- Treatment: Random Selection
- Control: Rest of the population
- Survey: Samples of Treatment and Control subjects
- Two waves, Dried Blood spots, Child development measures, adult primary care giver

Data Setup

- Z: Data from sampling frame (from voucher application)
- T for Treatment and C for Control
- Y(T): Measures for Treatment subjects
- Y(C): Measures for Control Subjects



Fill-in Synthetic potential populations



Inference

- Create several potential synthetic populations under treatment and control conditions
- Compute summary measures (such as mean, median etc.)
- Compare the distribution of summary measures under treatment and control conditions
  - Numerical summaries
  - Graphical summaries like histogram or kernel densities
- Analyze the two sets of populations to discern treatment effects, heterogeneity of treatment effects etc.

Summary

- Bayes inference for surveys must incorporate design features such as stratification, weighting and clustering appropriately
- Bayes inference is not asymptotic, and delivers good frequentist properties in small samples
- Software like BUGS (PROC MCMC in SAS) can be used to implement fully model based framework
- Recasting the Bayesian inference problem as missing data problem allows the use of multiple imputation software
- Nonparametric Bayes allows incorporation of complex design features without making strong model assumptions
- Pseudo or synthetic population framework makes the inference problem easy (just compute any estimand of interest)
- Give it a try!! (you will love it ☺)