Bayesian inference for sample surveys

Roderick Little

Module 1: Introduction

Learning Objectives

1. Understand basic features of alternative modes of inference for sample survey data.
2. Understand the mechanics of model-based and Bayesian inference for finite population quantitities under simple random sampling.
3. Understand the role of the sampling mechanism in sample surveys and how it is incorporated in model-based and Bayesian analysis.
4. More specifically, understand how survey design features, such as weighting, stratification, post-stratification and clustering, enter into a model-based or Bayesian analysis of sample survey data.
5. Be aware of Bayesian tools for computing posterior distributions of finite population quantities, and associated model checking and averaging.
6. The Bayesian perspective on survey nonresponse.

Acknowledgement and Disclaimer

- These slides are based in part on a short course on Bayesian methods in surveys presented by Dr. Trivellore Raghunathan and I at the 2010 Joint Statistical Meetings.
- While taking responsibility for errors, I'd like to acknowledge Dr. Raghunathan's major contributions to this material

Models for complex surveys

- Module 1: Introduction
- Module 2: Bayesian models for simple random samples
- Module 3: Bayesian models for complex sample designs
- Module 4: Bayesian computation and model assessment
- Module 5: Missing data

Module 1: Introduction

- Distinguishing features of survey sample inference
- Alternative modes of survey inference
  - Design-based, superpopulation models, Bayes
- Superpopulation modeling: basics of maximum likelihood estimation
- The Bayesian approach applied to simple random samples
  - Simple examples: binomial, normal, nonparametric, ratio/regression estimation

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 analytic model parameters:

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).

Approaches to Survey Inference

- Design-based (Randomization) inference
- Superpopulation Modeling
  - Specifies model conditional on fixed parameters
  - Frequentist inference based on repeated samples from superpopulation and finite population (hybrid approach)
- Bayesian modeling
  - Specifies full probability model (prior distributions on fixed parameters)
  - Bayesian inference based on posterior distribution of finite population quantities
  - argue that this is most satisfying approach

Design-Based Survey Inference

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

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

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

$Y = (Y_1, ..., Y_N)$ = population values,

recorded only for sample

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

Note: here I is random variable, (Y, Z) are fixed

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

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

$\hat{q} = \hat{q}(I, Y_{\text{inc}}, Z)$ = sample estimate of Q

$\hat{V}(I, Y_{\text{inc}}, Z)$ = sample estimate of V

$\left(\hat{q} - 1.96\sqrt{\hat{V}}, \hat{q} + 1.96\sqrt{\hat{V}}\right)$ = 95% confidence interval for Q

Random Sampling

- Random (probability) sampling characterized by:
  - Every possible sample has known chance of being selected
  - Every unit in the sample has a non-zero chance of being selected
  - In particular, for simple random sampling with replacement: "All possible samples of size *n* have same chance of being selected"

$Z = \{1, ..., N\}$ = set of units in the sample frame

$$\Pr(I \mid Z) = \begin{cases} 1/\binom{N}{n}, \sum_{i=1}^{N} I_i = n, \\ 0, \text{ otherwise} \end{cases}; \quad \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

$$E(I_i \mid Z) = \Pr(I_i = 1 \mid Z) = n / N$$

Example 1: Mean for Simple Random Sample

$$Q = \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \text{ population mean}$$

Random variable

$$\hat{q}(I) = \bar{y} = \sum_{i=1}^{N} I_i y_i / n, \text{ the sample mean}$$

Fixed quantity, not modeled

$$\text{Unbiased for } \bar{Y}: E_I\left(\sum_{i=1}^{N} I_i y_i / n\right) = \sum_{i=1}^{N} E_I(I_i) y_i / n = \sum_{i=1}^{N} (n/N) y_i / n = \bar{Y}$$

$$\text{Var}_I(\bar{y}) = V = (1 - n/N) S^2 / n, \; S^2 = \frac{1}{N-1}\sum_{i=1}^{N} (y_i - \bar{Y})^2$$

$(1 - n/N)$ = finite population correction

$$\hat{V} = (1 - n/N) s^2 / n, \; s^2 = \text{sample variance} = \frac{1}{n-1}\sum_{i=1}^{N} I_i (y_i - \bar{y})^2$$

$$95\% \text{ confidence interval for } \bar{Y} = \left(\bar{y} - 1.96\sqrt{\hat{V}}, \bar{y} + 1.96\sqrt{\hat{V}}\right)$$

Example 2: Horvitz-Thompson estimator

$$Q(Y) = T \equiv Y_1 + \ldots + Y_N$$

$$\pi_i = E(I_i \mid Y) = \text{inclusion probability} > 0$$

$$\hat{t}_{\mathrm{HT}} = \sum_{i=1}^{N} I_i Y_i / \pi_i, \; \mathrm{E}_I(\hat{t}_{\mathrm{HT}}) = \sum_{i=1}^{N} E(I_i) Y_i / \pi_i = \sum_{i=1}^{N} \pi_i Y_i / \pi_i = T$$

$$\hat{v}_{\mathrm{HT}} = \text{Variance estimate, depends on sample design}$$

$$\left(\hat{t}_{\mathrm{HT}} - 1.96\sqrt{\hat{v}_{\mathrm{HT}}}, \hat{t}_{\mathrm{HT}} + 1.96\sqrt{\hat{v}_{\mathrm{HT}}}\right) = 95\% \text{ CI for } T$$

- Pro: unbiased under minimal assumptions
- Cons:
  - variance estimator problematic for some designs (e.g. systematic sampling)
  - can have poor confidence coverage and inefficiency -- Basu “weighs in” with the following amusing example

Ex 2. Basu's inefficient elephants

$(y_1, ..., y_{50}) =$ weights of $N = 50$ elephants

Objective: $T = y_1 + y_2 + ... + y_{50}$. Only one elephant can be weighed!

- Circus trainer wants to choose "average" elephant (Sambo)
- Circus statistician requires "scientific" prob. sampling:

  Select Sambo with probability 99/100

  One of other elephants with probability 1/4900

  Sambo gets selected! Trainer: $\hat{t} = y_{(\text{Sambo})} \times 50$

  Statistician requires unbiased Horvitz-Thompson (1952) estimator: $\hat{T}_{HT} = \begin{cases} y_{(\text{Sambo})} / 0.99 \text{ (!!)}; \\ 4900\, y_{(i)}, \text{if Sambo not chosen (!!!)} \end{cases}$

HT estimator is unbiased on average but always crazy!

Circus statistician loses job and becomes an academic

Role of Models in Classical Approach

- Models are often used to motivate the choice of estimator. For example:
  - Regression model → regression estimator
  - Ratio model ⟶ ratio estimator
  - Generalized Regression estimation: model estimates adjusted to protect against misspecification, e.g. HT estimation applied to residuals from the regression estimator (Cassel, Sarndal and Wretman book).
- Estimates of standard error are then based on the randomization distribution
- This approach is design-based, model-assisted

Model-Based Approaches

- In our approach models are used as the basis for the entire inference: estimator, standard error, interval estimation
- This approach is more unified, but models need to be carefully tailored to features of the sample design such as stratification, clustering.
- One might call this model-based, design-assisted
- Two variants:
  - Superpopulation Modeling
  - Bayesian (full probability) modeling
- Common theme is “Infer” or “predict” about non-sampled portion of the population conditional on the sample and model

Superpopulation Modeling

- Model distribution *M:*

  $Y \sim f(Y \mid Z, \theta), Z$ = design variables, $\theta =$ fixed parameters

- Predict non-sampled values $\hat{Y}_{\text{exc}}$:

$\hat{y}_i = E(y_i \mid z_i, \theta = \hat{\theta}), \hat{\theta} =$ model estimate of θ

$$\hat{q} = Q(\tilde{Y}), \tilde{y}_i = \begin{cases} y_i, \text{ if unit sampled;} \\ \hat{y}_i, \text{ if unit not sampled} \end{cases}$$

$\hat{v} = \hat{mse}(\hat{q})$, over distribution of I and M

$(\hat{q} - 1.96\sqrt{\hat{v}}, \hat{q} + 1.96\sqrt{\hat{v}})$ = 95% CI for Q

I	Z	Y
1		Y_{inc}
1		
1		
0		$\hat{Y}_{\text{exc}}$
0		
0		
0		
0		

In the modeling approach, prediction of nonsampled values is central

In the design-based approach, weighting is central: “sample represents … units in the population”

Bayesian Modeling

- Bayesian model adds a prior distribution for the parameters:

$$(Y,\theta) \sim \pi(\theta \mid Z)\, f(Y \mid Z,\theta),\ \ \pi(\theta \mid Z) = \text{ prior distribution}$$

Inference about θ is based on posterior distribution from Bayes Theorem:

$$p(\theta \mid Z, Y_{\text{inc}}) \propto \pi(\theta \mid Z) L(\theta \mid Z, Y_{\text{inc}}), L = \text{likelihood}$$

Inference about finite population quantitity $Q(Y)$ based on

$$p(Q(Y) \mid Y_{\text{inc}}) = \text{ posterior predictive distribution}$$

of Q given sample values Y_{inc}

$$p(Q(Y) \mid Z, Y_{\text{inc}}) = \int p(Q(Y) \mid Z, Y_{\text{inc}}, \theta)\, p(\theta \mid Z, Y_{\text{inc}})\, d\theta$$

(Integrates out nuisance parameters θ)

I	Z	Y
1		Y_{inc}
1		
1		
0		$\hat{Y}_{\text{exc}}$
0		
0		
0		
0		

In the super-population modeling approach, parameters are considered fixed and estimated

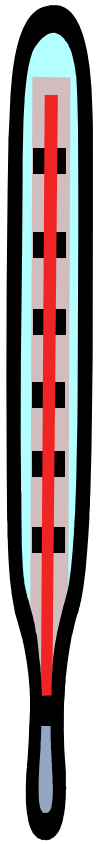
In the Bayesian approach, parameters are random and integrated out of posterior distribution – leads to better small-sample inference

Summary of design-based approach

- Avoids need for models for survey outcomes
- Robust approach for large probability samples
- Models needed for nonresponse, response errors, small areas
- Not well suited for small samples – inference basically assumes large samples, and models are needed for better precision in small samples
  - leading to “inferential schizophrenia”…

Inferential Schizophrenia

n-ometer



Design-based inference

Model-based inference

n_0 = "Point of inferential schizophrenia"

How do I choose n_0?
If $n_0 = 35$, should my entire statistical philosophy be different when n=34 and n=36?

Limitations of design-based approach

- Inference is based on probability sampling, but true probability samples are harder and harder to come by:
- Noncontact, nonresponse is increasing
- Face-to-face interviews increasingly expensive
- Can’t do “big data” (e.g. internet, administrative data) from the design-based perspective

Advantages of Bayesian approach

- Unified approach for large and small samples, nonresponse and response errors, data fusion, “big data”.
- Frequentist superpopulation modeling has the limitation that uncertainty in predicting parameters is not reflected in prediction inferences
- Bayes propagates uncertainty about parameters, yielding better frequentist properties in small samples
- Statistical modeling is the standard approach to statistics in substantive disciplines – having a design-based paradigm for surveys is divisive and confusing to modelers

Models bring survey inference closer to the statistical mainstream

Follow my design-based statistical standards

Why? I am an economist, I build models!

Challenges of the model-based perspective

- Explicit dependence on the choice of model, which has subjective elements (but assumptions are explicit)
- Bad models provide bad answers – justifiable concerns about the effect of model misspecification
  - In particular, models need to reflect features of the survey design, like clustering, stratification and weighting
- Models are needed for all survey variables – need to understand the data
- Potential for more complex computations

Overarching philosophy: calibrated Bayes

- Survey inference is not fundamentally different from other problems of statistical inference
  - But it has particular features that need attention
- Statistics is basically prediction: in survey setting, predicting survey variables for non-sampled units
- Inference should be model-based, Bayesian
- Seek models that are "frequency calibrated" (Box 1980, Rubin 1984, Little 2006):
  - Incorporate survey design features
  - Properties like design consistency are useful
  - "objective" priors generally appropriate
    - Little (2004, 2006, 2012); Little & Zhang (2007)

Calibrated Bayes

"The applied statistician should be Bayesian in principle and calibrated to the real world in practice – appropriate frequency calculations help to define such a tie."

"… frequency calculations are useful for making Bayesian statements scientific, … in the sense of capable of being shown wrong by empirical test; here the technique is the calibration of Bayesian probabilities to the frequencies of actual events."

Rubin (1984)

Bayesian inference for sample surveys

Roderick Little

Module 2: Bayesian models for simple random samples

Superpopulation Modeling: Estimating parameters

- Various principles: least squares, method of moments, maximum likelihood
- Sketch main ideas of maximum likelihood, an important approach that underlies statistical inferences for many common models:
  - Linear and nonlinear regression
  - Generalized linear models (logistic, Poission regression)
  - Repeated measures models (SAS PROC MIXED, NLMIXED
  - Survival analysis – proportional hazards models

Likelihood methods

- Statistical model + data $\Rightarrow$ Likelihood
- Two general approaches based on likelihood
  - maximum likelihood inference (for large samples)
  - Bayesian inference (better for small samples):

    log(likelihood) + log(prior) = log(posterior)
- Methods do not require rectangular data sets
  - can be applied to incomplete data

Definition of Likelihood

- Data Y
- Statistical model yields probability density $f(Y \mid \theta)$

  for Y with unknown parameters θ
- Likelihood function is then a function of θ

$$L(\theta \mid Y) = const \times f(Y \mid \theta)$$

- Loglikelihood is often easier to work with:

$$\ell(\theta \mid Y) = \log L(\theta \mid Y) = const + \log\{ f(Y \mid \theta)\}$$

Constants can depend on data but not on parameter θ

Example: Normal sample

- $Y = (y_1, \ldots, y_n)$ univariate iid normal sample

$$\theta = (\mu, \sigma^2)$$

$$f(Y \mid \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu\right)^2\right)$$

$$\ell(\mu, \sigma^2 \mid Y) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu\right)^2$$

Example: Multinomial sample

- $Y = (y_1, \ldots, y_n)$ univariate K-category multinomial sample

  n_j = number of y_i equal to j (j=1,…,K)

$$\theta = (\pi_1, \ldots, \pi_{K-1}); \;\; \pi_K = 1 - \pi_1 - \ldots - \pi_{K-1}$$

$$f(Y \mid \pi_1, \ldots, \pi_{K-1}) = \frac{n!}{n_1! \ldots n_K!} \left(\prod_{j=1}^{K-1} \pi_j^{n_j} \right) (1 - \pi_1 - \ldots - \pi_{K-1})^{n_K}$$

$$\ell(\pi_1, \ldots, \pi_{K-1} \mid Y) = \left(\sum_{j=1}^{K-1} n_j \log \pi_j \right) + n_K \log(1 - \pi_1 - \ldots - \pi_{K-1})$$

Maximum Likelihood Estimate

- The maximum likelihood (ML) estimate $\hat{\theta}$ of θ maximizes the likelihood, or equivalently the log-likelihood

$$L(\hat{\theta} \mid Y) \geq L(\theta \mid Y) \text{ for all } \theta$$

- The ML estimate is the

  "value of the parameter that makes the data most likely"
- The ML estimate is not necessarily unique, but is for many regular problems given enough data

Computing the ML estimate

- In regular problems, the ML estimate can be found by solving the likelihood equation

$$S(\theta \mid Y) = 0$$

where S is the score function, defined as the first derivative of the loglikelihood:

$$S(\theta \mid Y) \equiv \frac{\partial \log L(\theta \mid Y)}{\partial \theta}$$

For some models (e.g. multiple linear regression), likelihood equation has an explicit solution; for others (e.g. logistic regression) numerical optimization methods are needed

Normal Examples

- Univariate Normal sample $Y = (y_1, ..., y_n)$ $\theta = (\mu, \sigma^2)$

$$\hat{\mu} = \bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2$$

(Note the lack of a correction for degrees of freedom)

- Multivariate Normal sample

$$\hat{\mu} = \bar{y}, \ \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^T$$

- Normal Linear Regression (possibly weighted)

$$(y_i \mid x_{i1}, ..., x_{ip}) \sim N(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2 / u_i)$$

$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p) =$ weighted least squares estimates

$\hat{\sigma}^2 =$ (weighted residual sum of squares)/n

Multinomial Example

$$Y = (y_1, ..., y_n);\ y_i \sim \text{MNOM}(\pi_1, ..., \pi_K)$$

$$n_j = \text{ number of } y_i \text{ equal to } j\ (j = 1, ..., K)$$

Likelihood Equations:

$$\frac{\partial l}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_K}{1 - \pi_1 - ... - \pi_{K-1}} = 0,\ j = 1, ..., K-1$$

Hence ML estimate is

$$\hat{\pi}_j = n_j / n,\ j = 1, ..., K$$

Logistic regression

$$\Pr(y_i = 1 \mid x_{i1}, \ldots, x_{ip}) = \pi_i(\beta) = \frac{\exp\left(f_i(\beta)\right)}{1 + \exp\left(f_i(\beta)\right)}$$

$$f_i(\beta) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$\ell(\beta) = \sum_{i=1}^{n} \left(y_i \pi_i(\beta) + (1 - y_i)(1 - \pi_i(\beta)) \right)$$

ML estimation requires iterative methods like method of scoring

ML for mixed-effects models

$y_i = (y_{\text{obs},i}, y_{\text{mis},i})$: k-dimensional vector of repeated measures

$(y_i \mid X_i, \beta_i) \sim N_k(X_{1i}\alpha + X_{2i}\beta, \Sigma)$

α are fixed effects; β are random effects: $\beta_i \sim N_q(0, \Gamma)$

Missing Data Mechanism: missing at random

ML requires iterative algorithms

e.g. Harville (1977), Laird and Ware (1982), SAS Proc Mixed

- Very flexible mean and covariance structures
- Normality not a major assumption if *N* large, and recent programs allow for non-normal outcomes

Properties of ML estimates

- Under assumed model, ML estimate is:
  - Consistent (not necessarily unbiased)
  - Efficient for large samples
  - not necessarily the best for small samples
- ML estimate is transformation invariant
  - If $\hat{\theta}$ is the ML estimate of θ

    Then $\phi(\hat{\theta})$ is the ML estimate of $\phi(\theta)$

Large-sample ML Inference

- Basic large-sample approximation:

  for regular problems,

$$\theta - \hat{\theta} \sim N(0, C)$$

  where C is a covariance matrix estimated from the sample

  – Frequentist treats $\hat{\theta}$ as random, θ as fixed; equation defines the sampling distribution of $\hat{\theta}$

  – Bayesian treats θ as random, $\hat{\theta}$ as fixed;

    equation defines posterior distribution of θ

Forms of precision matrix

- The precision of the ML estimate is measured by C^{-1} Some forms for this are:
  - Observed information (recommended)

$$C^{-1} = I(\hat{\theta}|Y) = -\left.\frac{\partial^2 \log L(\theta|Y)}{\partial\theta\,\partial\theta}\right|_{\theta=\hat{\theta}}$$

  - Expected information (not as good, may be simpler)

$$C^{-1} = J(\hat{\theta}) = E\left[I(\hat{\theta}|Y,\theta)\right]_{\theta=\hat{\theta}}$$

  - Some other approximation to curvature of loglikelihood in the neighborhood of the ML estimate

Interval estimation

- 95% (confidence, probability) interval for scalar θ is:

  $\hat{\theta} \pm 1.96\ C^{1/2}$,where 1.96 is 97.5 pctile of normal distribution
- Example: univariate normal sample

$$I = J = \begin{bmatrix} n/\hat{\sigma}^2 & 0 \\ 0 & n/(2\hat{\sigma}^4) \end{bmatrix} \Rightarrow C = \begin{bmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 2\hat{\sigma}^4/n \end{bmatrix}$$

Hence some 95% intervals are:

$$\bar{y} \pm 1.96\, s/\sqrt{n} \quad \text{for } \mu$$

$$s^2 \pm 1.96\, s^2/\sqrt{n/2} \quad \text{for } \sigma^2$$

$$\ln(s) \pm 1.96\sqrt{2/n} \quad \text{for } \ln(\sigma)$$

Significance Tests

Tests based on likelihood ratio (LR) or Wald (W) statistics:

$\theta = (\theta_{(1)}, \theta_{(2)})$; $\theta_{(1)0}$ = null value of $\theta_{(1)}$; θ_2 = other parameters

$\hat{\theta}$ = unrestricted ML estimate

$\tilde{\theta} = (\theta_{(1)0}, \tilde{\theta}_{(2)})$; $\tilde{\theta}_{(2)}$ = ML estimate of $\theta_{(2)}$ given $\theta_{(1)} = \theta_{(1)0}$

LR statistic: $\mathrm{LR}(\hat{\theta}, \tilde{\theta}) = 2\left[\ell(\hat{\theta} \mid Y) - \ell(\tilde{\theta} \mid Y)\right]$

Wald statistic: $W(\hat{\theta}, \tilde{\theta}) = (\theta_{(1)0} - \hat{\theta}_{(1)})^T C_{(11)}^{-1} (\theta_{(1)0} - \hat{\theta}_{(1)})$

$C_{(11)}$ = covariance matrix of $(\theta_{(1)} - \hat{\theta}_{(1)})$

yield P-values $P = pr\left(\chi_q^2 > D(\hat{\theta}, \tilde{\theta})\right)$

D = LR or Wald statistic; q = dimension of θ_0

χ_q^2 = Chi-squared distribution with q degrees of freedom

Bayesian Modeling

- Bayesian model adds a prior distribution for the parameters:

$(Y,\theta) \sim \pi(\theta \mid Z) f(Y \mid Z,\theta)$, $\pi(\theta \mid Z) =$ prior distribution

Inference about θ is based on posterior distribution from Bayes Theorem:

$p(\theta \mid Z, Y_{\text{inc}}) \propto \pi(\theta \mid Z) L(\theta \mid Z, Y_{\text{inc}})$, L = likelihood

Inference about finite population quantitity $Q(Y)$ based on

$p(Q(Y) \mid Y_{\text{inc}}) =$ posterior predictive distribution

of Q given sample values Y_{inc}

$$p(Q(Y) \mid Z, Y_{\text{inc}}) = \int p(Q(Y) \mid Z, Y_{\text{inc}}, \theta)\, p(\theta \mid Z, Y_{\text{inc}})\, d\theta$$

(Integrates out nuisance parameters θ)

I	Z	Y
1		Y_{inc}
1		
1		
0		$\hat{Y}_{\text{exc}}$
0		
0		
0		
0		

In the super-population modeling approach, parameters are considered fixed and estimated

In the Bayesian approach, parameters are random and assigned prior distributions – leads to better small-sample inference

Bayes Inference

- Inferences about Q are based on its posterior predictive distribution:
  - "estimate" is posterior mean: $\hat{q} = E(Q|Y_{inc})$
  - "standard error" is posterior sd: $s = \sqrt{Var(Q \mid Y_{inc})}$
  - 95% posterior probability (or credibility) interval plays role of confidence interval (but with simpler interpretation)
  - In large samples, a 95% interval is $\hat{q} \pm 1.96s$
  - In small samples, can use highest posterior density (hpd) interval, or 2.5^{th} to 97.5^{th} percentiles of posterior distribution (often simulated using MCMC draws from the posterior distribution)

Inference about population quantities

- Inferences about Q are conveniently obtained by first conditioning on θ and then averaging over posterior of θ. In particular, the posterior mean is:

$$E(Q \mid Y_{\text{inc}}) = E\left[E(Q \mid Y_{\text{inc}}, \theta) \mid Y_{\text{inc}}\right]$$

and the posterior variance is:

$$Var(Q \mid Y_{\text{inc}}) = E\left[Var(Q \mid Y_{\text{inc}}, \theta) \mid Y_{\text{inc}}\right] + Var\left[E(Q \mid Y_{\text{inc}}, \theta) \mid Y_{\text{inc}}\right]$$

- Value of this technique will become clear in applications
- Finite population corrections are automatically obtained as differences in the posterior variances of Q and θ
- Inferences based on full posterior distribution useful in small samples (e.g. provides “t corrections”)

Example: linear regression

The normal linear regression model:

$$(y_i \mid x_{i1},...,x_{ip}) \sim N(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2)$$

with non-informative "Jeffreys'" prior:

$$p(\beta_0,...,\beta_p, \log\sigma^2) = const.$$

yields the posterior distribution of $(\beta_0,...,\beta_p)$ as multivariate *T* with mean given by the least squares estimates $(\hat{\beta}_0,...,\hat{\beta}_p)$ covariance matrix $(X^T X)^{-1} s^2$, where *X* is the design matrix, and degrees of freedom *n* - *p* - 1.

Resulting posterior probability intervals are equivalent to standard t confidence intervals.

Simulating Draws from Posterior Distribution

- With problems with high-dimensional θ , it is often easier to draw values from the posterior distribution, and base inferences on these draws
- For example, if

$$(\theta_1^{(d)} : d = 1, ..., D)$$

is a set of draws from the posterior distribution for a scalar parameter θ_1, then

$\bar{\theta}_1 = D^{-1} \sum_{d=1}^{D} \theta_1^{(d)}$ approximates posterior mean

$s_\theta^2 = (D-1)^{-1} \sum_{d=1}^{D} (\theta_1^{(d)} - \bar{\theta}_1)^2$ approximates posterior variance

$(\bar{\theta}_1 \pm 1.96 s_\theta)$ or 2.5th to 97.5th percentiles of draws approximates 95% posterior credibility interval for θ

Example: Posterior Draws for Normal Linear Regression

$(\hat{\beta}, s^2) =$ ls estimates of slopes and resid variance

$\sigma^{(d)2} = (n - p - 1)s^2 / \chi^2_{n-p-1}$

$\beta^{(d)} = \hat{\beta} + A^T z\sigma^{(d)}$

χ^2_{n-p-1} = chi-squared deviate with $n - p - 1$ df

$z = (z_1, ..., z_{p+1})^T$, $z_i \sim N(0,1)$

$A =$ upper triangular Cholesky factor of $(X^T X)^{-1}$:

$A^T A = (X^T X)^{-1}$

- Easily extends to weighted regression: see Example 6.19

Consulting Example

- In India, any person possessing a radio, transistor or television has to pay a license fee.
- In a densely populated area with mostly makeshift houses practically no one was paying these fees.
- It was determined that for enforcement to be fiscally meaningful, the proportion of households possessing one or more of these devices must exceed certain limit.

Consulting example (continued)

$N =$ Population Size

$$Y_i = \begin{cases} 1, & \text{if household } i \text{ has a device} \\ 0, & \text{otherwise} \end{cases}$$

$Q = \sum_{i=1}^{N} Y_i / N$ Proportion of households with a device

Question of Interest: $\Pr(Q \geq 0.3)$

- If the probability of Q exceeding 0.3 is very high then enforcement might be fiscally sensible
- Conduct a small scale survey to answer the question of interest
- Note that question only makes sense under Bayes paradigm

Consulting example

srs of size n, $Y_{\text{inc}} = \{Y_1, \ldots, Y_n\}$, $Y_{\text{exc}} = \{Y_{n+1}, \ldots, Y_N\}$

$Y_i \mid \theta \sim iid$ Bernoulli(θ)

$$x = \sum_{i=1}^{n} Y_i$$

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Model for observable

$$\pi(\theta) = 1 \quad \theta \in (0,1)$$

Prior distribution

$$Q = \sum_{i=1}^{N} Y_i / N = \left(x + \sum_{i=n+1}^{N} Y_i \right) / N$$

Estimand

Binomial Example

The posterior distribution is

$$p(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)d\theta} \propto f(x \mid \theta)\pi(\theta)$$

$$p(\theta \mid x) = \frac{\binom{n}{x}\theta^{x}(1-\theta)^{n-x} \times 1}{\int \binom{n}{x}\theta^{x}(1-\theta)^{n-x} d\theta}$$

$$\theta \mid x \sim Beta(x+1, n-x+1)$$

$$Q = (x + \sum_{i=n+1}^{N} Y_i) / N$$

$$\left(\sum_{i=n+1}^{N} Y_i \mid \theta, x \right) \sim \text{Bin}(N-n, \theta)$$

Infinite Population

$For\ N \to \infty,\ \bar{Y}_N \to \theta$

$\Pr(\bar{Y}_N \geq 0.3 \mid x) \approx \Pr(\theta \geq 0.3 \mid x)$

Compute using cumulative distribution function of a beta distribution which is a standard function in most software such as SAS, R

What is the maximum proportion of households in the population with devices that can be said with great certainty?

$$\Pr(\theta \leq ? \mid x) = 0.9$$

Inverse CDF of Beta Distribution

Point Estimates

- Point estimate is often used as a single summary "best" value for the unknown Q
- Some choices are the mean, mode or the median of the posterior distribution of Q
- For symmetrical distributions an intuitive choice is the center of symmetry
- For asymmetrical distributions the choice is not clear. It depends upon the "loss" function.

Interval Estimation

- Better summary is an interval estimate
- Fix the coverage rate 1-α in advance and determine the *highest posterior density* region C to include most likely values of Q totaling 1-α posterior probability
- Fix the value Q_o in advance, determine C by the collection of values of Q more likely than Q_o and calculate the coverage 1-α as the posterior probability of this C

Interval Estimates

C is such that

(1) $p(Q \mid Y_{\text{inc}}) > p(Q' \mid Y_{\text{inc}})$

$Q \in C, Q' \notin C$

(2) $\Pr(Q \in C \mid Y_{\text{inc}}) = 1 - \alpha$

"Most likely" is usually defined by highest posterior density

- Highest Posterior Density Region
- For symmetric unimodal posterior distributions, $(1-\alpha)$ HPD interval is $(q_{\alpha/2}, q_{1-\alpha/2})$ where $\Pr(Q \leq q_{\alpha/2}) = \alpha/2$
- In the Binomial example, the beta density of θ used to determine the interval estimate of Q

Normal simple random sample

$$Y_i \sim \text{iid } N(\mu, \sigma^2); i = 1, 2, ..., N$$

$$\pi(\mu, \sigma^2) \propto \sigma^{-2}$$

simple random sample results in $Y_{\text{inc}} = (y_1, ..., y_n)$

$$Q = \bar{Y} = \frac{n\bar{y} + (N - n)\bar{Y}_{\text{exc}}}{N}$$

$$= f \times \bar{y} + (1 - f) \times \bar{Y}_{\text{exc}}$$

Derive posterior distribution of Q

Normal Example

Posterior distribution of (μ,σ^2)

$$p(\mu,\sigma^2 \mid Y_{\text{inc}}) \propto (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i\in\text{inc}} \frac{(y_i-\mu)^2}{\sigma^2} \right)$$

$$\propto (\sigma^2)^{-n/2-1} \exp -\frac{1}{2}\left(\sum_{i\in\text{inc}} (y_i-\bar{y})^2/\sigma^2 - n(\mu-\bar{y})^2/\sigma^2 \right)$$

The above expressions imply that

$$(1)\ \sigma^2 \mid Y_{\text{inc}} \sim \sum_{i\in\text{inc}} (y_i-\bar{y})^2 / \chi^2_{n-1}$$

$$(2)\ \mu \mid Y_{\text{inc}}, \sigma^2 \sim N(\bar{y}, \sigma^2/n)$$

Posterior Distribution of Q

$$\bar{Y}_{\text{exc}} \mid \mu, \sigma^2 \sim N(\mu, \frac{\sigma^2}{N-n})$$

$$\bar{Y}_{\text{exc}} \mid \sigma^2, Y_{\text{inc}} \sim N\left(\bar{y}, \frac{\sigma^2}{N-n} + \frac{\sigma^2}{n} = \frac{\sigma^2}{(1-f)n}\right)$$

$$Q = f \times \bar{y} + (1-f) \times \bar{Y}_{\text{exc}}$$

$$Q \mid \sigma^2, Y_{\text{inc}} \sim N\left(\bar{y}, \frac{(1-f)\sigma^2}{n}\right)$$

$$\bar{Y}_{\text{exc}} \mid Y_{\text{inc}} \sim t_{n-1}\left(\bar{y}, \frac{s^2}{(1-f)n}\right)$$

$$Q \mid Y_{\text{inc}} \sim t_{n-1}\left(\bar{y}, \frac{(1-f)s^2}{n}\right)$$

HPD Interval for Q

Note the posterior t distribution of Q is symmetric and unimodal -- values in the center of the distribution are more likely than those in the tails.

Thus a (1-α)100% HPD interval is:

$$\bar{y} \pm t_{n-1,1-\alpha/2}\sqrt{\frac{(1-f)s^2}{n}}$$

Like frequentist confidence interval, but recovers the t correction

Some other Estimands

- Suppose *Q*=Median or some other percentile
- One is better off inferring about all non-sampled values
- As we will see later, simulating values of Y_{exc} add enormous flexibility for drawing inferences about any finite population quantity
- Modern Bayesian methods heavily rely on simulating values from the posterior distribution of the model parameters and predictive-posterior distribution of the nonsampled values
- Computationally, if the population size, *N*, is too large then choose any arbitrary value *K* large relative to *n*, the sample size
  - National sample of size 2000
  - US population size 306 million
  - For numerical approximation, we can choose *K=2000/f*, for some small *f*=0.01 or 0.001.

Comparison of Two Populations

- Population 1

$$\textit{Population size} = N_1$$
$$\textit{Sample size} = n_1$$
$$Y_{1i} \sim ind\ N(\mu_1, \sigma_1^2)$$
$$\pi(\mu_1, \sigma_1^2) \propto \sigma_1^{-2}$$

$$\textit{Sample Statistics}: (\bar{y}_1, s_1^2)$$
$$\textit{Posterior distributions}:$$
$$(n_1 - 1)s_1^2 / \sigma_1^2 \sim \chi^2_{n_1 - 1}$$
$$\mu_1 \sim N(\bar{y}_1, \sigma_1^2 / n_1)$$
$$Y_{1i} \sim N(\mu_1, \sigma_1^2), i \in \text{exc}$$

- Population 2

$$\textit{Population size} = N_2$$
$$\textit{Sample size} = n_2$$
$$Y_{2i} \sim ind\ N(\mu_2, \sigma_2^2)$$
$$\pi(\mu_2, \sigma_2^2) \propto \sigma_2^{-2}$$

$$\textit{Sample Statistics}: (\bar{y}_2, s_2^2)$$
$$\textit{Posterior distributions}:$$
$$(n_2 - 1)s_2^2 / \sigma_2^2 \sim \chi^2_{n_2 - 1}$$
$$\mu_2 \sim N(\bar{y}_2, \sigma_2^2 / n_2)$$
$$Y_{2i} \sim N(\mu_2, \sigma_2^2), i \in \text{exc}$$

Estimands

- Examples
  - $\bar{Y}_1 - \bar{Y}_2$ (Finite sample version of Behrens-Fisher Problem)
  - Difference $\Pr(Y_1 > c) - \Pr(Y_2 > c)$
  - Difference in the population medians
  - Ratio of the means or medians
  - Ratio of Variances
- It is possible to analytically compute the posterior distribution of some these quantities
- It is a whole lot easier to simulate values of non-sampled $Y_1^{'s}$ in Population 1 and $Y_2^{'s}$ in Population 2

Bayesian Nonparametric Inference

- Population: $Y_1, Y_2, Y_3, ..., Y_N$
- All possible distinct values: $d_1, d_2, \ldots, d_K$
- Model: $\Pr(Y_i = d_k) = \theta_k$
- Prior: $\pi(\theta_1, \theta_2, ..., \theta_k) \propto \prod_k \theta_k^{-1}$ if $\sum_k \theta_k = 1$
- Mean and Variance:

$$E(Y_i \mid \theta) = \mu = \sum_k d_k \theta_k$$

$$Var(Y_i \mid \theta) = \sigma^2 = \sum_k d_k^2 \theta_k - \mu^2$$

Bayesian Nonparametric Inference (continued)

- SRS of size n with n_k equal to number of d_k in the sample
- Objective is to draw inference about the population mean: $Q = f \times \bar{y} + (1 - f) \times \bar{Y}_{\text{exc}}$
- As before we need the posterior distribution of μ and σ^2

Nonparametric Inference (continued)

- Posterior distribution of θ is Dirichlet:

$$\pi(\theta \mid Y_{\text{inc}}) \propto \prod_k \theta_k^{n_k - 1} \text{ if } \sum_k \theta_k = 1 \text{ and } \sum_k n_k = n$$

- Posterior mean, variance and covariance of θ

$$E(\theta_k \mid Y_{\text{inc}}) = \frac{n_k}{n}, Var(\theta_k \mid Y_{\text{inc}}) = \frac{n_k(n - n_k)}{n^2(n+1)}$$

$$Cov(\theta_k, \theta_l \mid Y_{\text{inc}}) = -\frac{n_k n_l}{n^2(n+1)}$$

Inference for Q

$$E(\mu \mid Y_{\text{inc}}) = \sum_k d_k \frac{n_k}{n} = \bar{y}$$

$$Var(\mu \mid Y_{\text{inc}}) = \frac{s^2}{n}\frac{n-1}{n+1}; s^2 = \frac{1}{n-1}\sum_{i \in \text{inc}} (y_i - \bar{y})^2$$

$$E(\sigma^2 \mid Y_{\text{inc}}) = s^2 \frac{n-1}{n+1}$$

Hence posterior mean and variance of Q are:

$$E(Q \mid Y_{\text{inc}}) = f \times \bar{y} + (1-f)E(\mu \mid Y_{\text{inc}}) = \bar{y}$$

$$Var(Q \mid Y_{\text{inc}}) = (1-f)\frac{s^2}{n}\frac{n-1}{n+1}$$

Ratio and Regression Estimates

- Population: $(y_i, x_i;\ i=1,2,\ldots N)$
- Sample: $(y_i,\ i \in \text{inc},\ x_i,\ i=1,2,\ldots,N)$.

For now assume SRS

y_1	x_1
y_2	x_2
.	.
.	.
.	.
y_n	x_n
	x_{n+1}
	x_{n+2}
	.
	.
	.
	x_N

Objective: Infer about the population mean

$$Q = \sum_{i=1}^{N} y_i$$

Excluded Y's are missing values

Model Specification

$$(Y_i \mid x_i, \beta, \sigma^2) \sim \text{ind } N(\beta x_i, \sigma^2 x_i^{2g})$$

$$i = 1, 2, ..., N$$

g known

Prior distribution: $\pi(\beta, \sigma^2) \propto \sigma^{-2}$

g=1/2: Classical Ratio estimator. Posterior variance equals randomization variance for large samples

g=0: Regression through origin. The posterior variance is nearly the same as the randomization variance.

g=1: HT model. Posterior variance equals randomization variance for large samples.

Note that, no asymptotic arguments have been used in deriving Bayesian inferences. Makes small sample corrections and uses t-distributions.

Some Remarks

- For large samples, estimate and its variance under nonparametric model assumptions are very nearly the same as those under the normal model assumptions
- For large N, the population size, the finite population quantity is very nearly same as the model parameter ($Q \approx \mu$).
- For large samples,

$$\frac{Q - E(Q \mid Y_{\text{inc}})}{\sqrt{Var(Q \mid Y_{\text{inc}})}} \sim N(0,1)$$

Remarks (Continued)

- Bayesian Interpretation: Summary of the excluded portion of the population has approximate normal distribution conditional on the observed data. *That is* Y_{inc} *is fixed and Q is random.*
- Frequentist Interpretation: Under repeated sampling, the distribution of estimates of *Q. That is Q is fixed and* Y_{inc} *is random.*
- For large samples, the frequentist and Bayes will nearly give the same numerical answers but interpretations would differ.

Remarks

- In much practical analysis the prior information is diffuse, and the likelihood dominates the prior information.
- Jeffreys (1961) developed “noninformative priors” based on the notion of very little prior information relative to the information provided by the data.
- Jeffreys derived the noninformative prior requiring invariance under parameter transformation.
- In general,

$$\pi(\theta) \propto | J(\theta) |^{1/2}$$

where

$$J(\theta) = -E\left(\frac{\partial^2 \log f(y \mid \theta)}{\partial\theta\partial\theta^t} \right)$$

Examples of noninformative priors

Normal: $\pi(\mu, \sigma^2) \propto \sigma^{-2}$

Binomial: $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$

Poisson: $\pi(\lambda) \propto \lambda^{-1/2}$

Normal regression with slopes β: $\pi(\beta, \sigma^2) \propto \sigma^{-2}$

In simple cases these noninformative priors result in numerically same answers as standard frequentist procedures

Comments

- iid (normal) model for simple random sampling is based on exchangeability ideas of De Finetti
- other “off-the-shelf” models are more appropriate for other sample designs -- hence the design influences the choice of model, as we shall see
- Even in this simple normal problem, Bayes is useful:
  - t-inference is recovered for small samples by putting a prior on the unknown variance
- Bayes is even more attractive for more complex problems, as discussed later.

Summary

- Considered Bayesian predictive inference for population quantities
- Focused here on the population mean, but other posterior distribution of more complex finite population quantities Q can be derived
- Key is to compute the posterior distribution of Q conditional on the data and model
  - Summarize the posterior distribution using posterior mean, variance, HPD interval etc
- Modern Bayesian analysis uses simulation technique to study the posterior distribution
- Models need to incorporate complex design features like unequal selection, stratification and clustering

Bayesian inference for sample surveys

Roderick Little

Module 3: Bayesian models for complex sample designs

Modeling sample selection

- Role of sample design in model-based (Bayesian) inference
- Key to understanding the role is to include the sample selection process as part of the model
- Modeling the sample selection process
  - Simple and stratified random sampling
  - Cluster sampling, other mechanisms
  - See Chapter 7 of *Bayesian Data Analysis* (Gelman, Carlin, Stern and Rubin 1995)

Formal models that include data collection

$Y = (y_1, \ldots, y_N)$ = population data; y_i may be a vector

Z = fully-observed covariates, design variables

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

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

$$I_i = \begin{cases} 1, & y_i \text{ observed} \\ 0, & \text{otherwise} \end{cases}$$

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

Y_{inc} = included part of Y, Y_{exc} = excluded part of Y

- Notation implies *Stable Unit Treatment Value Assumption* (SUTVA): Values not affected by choice of inclusion vector I

Full model for Y and I

$$p(Y, I \mid Z, \theta, \phi) = p(Y \mid Z, \theta)\; p(I \mid Y, Z, \phi)$$

Model for Population

Model for Inclusion

- Observed data: (Y_{inc}, Z, I) (No missing values)
- Observed-data likelihood:

$$L(\theta, \phi \mid Y_{\text{inc}}, Z, I) \propto p(Y_{\text{inc}}, I \mid Z, \theta, \phi) = \int p(Y, I \mid Z, \theta, \phi) dY_{\text{exc}}$$

- Posterior distribution of parameters:

$$p(\theta, \phi \mid Y_{\text{inc}}, Z, I) \propto p(\theta, \phi \mid Z) L(\theta, \phi \mid Y_{\text{inc}}, Z, I)$$

Ignoring the data collection process

- The likelihood *ignoring the data-collection process* is based on the model for Y alone with likelihood:

$$L(\theta \mid Y_{\text{inc}}, Z) \propto p(Y_{\text{inc}} \mid Z, \theta) = \int p(Y \mid Z, \theta) dY_{\text{exc}}$$

- The corresponding posteriors for θ and Y_{exc} are:

$$p(\theta \mid Y_{\text{inc}}, Z) \propto p(\theta \mid Z) L(\theta \mid Y_{\text{inc}}, Z)$$

$$p(Y_{\text{exc}} \mid Y_{\text{inc}}, Z) \propto \int p(Y_{\text{exc}} \mid Y_{\text{inc}}, Z, \theta) p(\theta \mid Y_{\text{inc}}, Z) d\theta$$

Posterior predictive distribution of Y_{exc}

- When the full posterior reduces to this simpler posterior, the data collection mechanism is called *ignorable* for Bayesian inference about θ, Y_{exc}.

Conditions when data collection mechanism can be ignored

- Two general and simple sufficient conditions for ignoring the data-collection mechanism are:

  Selection at Random (SAR):

  $$p(I \mid Y, Z, \phi) = p(I \mid Y_{\text{inc}}, Z, \phi) \text{ for all } Y_{\text{exc}}.$$

  Bayesian Distinctness:

  $$p(\theta, \phi \mid Z) = p(\theta \mid Z)\, p(\phi \mid Z)$$

- It is easy to show that these conditions together imply that:

  $$p(\theta, Y_{\text{exc}} \mid Y_{\text{inc}}, Z) = p(\theta, Y_{\text{exc}} \mid Y_{\text{inc}}, Z, I)$$

  so the model for the data-collection mechanism does not affect inferences about the parameter θ or finite population quantities Q.

Ex: simple random sampling

- For *Simple Random Sampling*, the sampling distribution is:

$$p(I|Y,\phi) = \begin{cases} \binom{N}{n}^{-1}, & \text{if } \sum_{i=1}^{N} I_i = n; \\ 0, & \text{otherwise.} \end{cases}$$

- This is clearly ignorable, with Z null.
- This justifies ignoring the mechanism in Module 2

Bayes inference for probability samples

- In other probability sampling designs, selection does not depend on values of Y and the mechanism is known, that is:

$$p(I \mid Y, Z, \phi) = p(I \mid Z) \text{ for all } Y.$$

- This means that the data-collection mechanism is ignorable for Bayesian inference (with complete data)
- But the model needs to appropriately account for relationship of survey outcomes Y with the design variables Z.
- Consider how to do this for (a) unequal probability samples, and (b) clustered (multistage) samples

Models for unequal probability samples

- Appropriate analysis depends on how the variables leading to the design weights enter the model of substantive interest
  - (a) all are included
  - (b) some are included, others aren't
  - (c) none are included
- Consider these distinctions for (a) means and (b) regression coefficients

Design-based weighting

- A pure form of design-based estimation is to weight sampled units by inverse of inclusion probabilities π_i
  - Sampled unit *i* "represents" $w_i = 1/\pi_i$ units in the population
- More generally, a common approach is:

$$w_i = w_{is} \times w_{in}(w_{is}) \times w_{ip}(w_{is}, w_{in})$$

w_{is} = sampling weight

$w_{in}(w_{is})$ = nonresponse weight

$w_{ip}(w_{is}, w_{in})$ = post-stratification weight

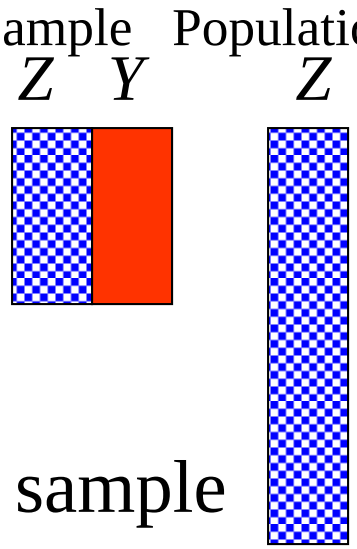
Weighting and models

- The weights can't generally be ignored from a modeling perspective
  - Ignores different selection effects that bias estimates
- Weights are auxiliary covariates from a modeling perspective
- Design: weight the respondents
  - One size fits all *Y* variables
- Model: use weights to help predict non-sampled and non-responding values
  - Weighting adds noise for *Y*'s unrelated to weights
- The model perspective is more flexible (but potentially more work)

Ex 1: stratified random sampling

- Population divided into J strata
- Z is set of stratum indicators:

$$z_i = \begin{cases} 1, & \text{if unit } i \text{ is in stratum } j; \\ 0, & \text{otherwise.} \end{cases}$$



- Stratified random sampling: simple random sample of n_j units selected from population of N_j units in stratum j.
- This design is ignorable *providing* model for outcomes conditions on the stratum variables Z.

Inference for a mean from a stratified sample

- Consider a model that includes stratum effects:

$$[y_i \mid z_i = j] \sim_{\text{ind}} N(\theta_j, \sigma_j^2)$$

- For simplicity assume σ_j^2 is known and the flat prior:

$$p(\theta_j \mid Z) \propto const.$$

- Standard Bayesian calculations lead to

$$[\bar{Y} \mid Y_{\text{inc}}, Z, \{\sigma_j^2\}] \sim N(\bar{y}_{\text{st}}, \sigma_{\text{st}}^2)$$

where:

$$\bar{y}_{\text{st}} = \sum_{j=1}^{J} P_j \bar{y}_j,\ P_j = N_j / N,\ \bar{y}_j = \text{ sample mean in stratum } j,$$

$$\sigma_{\text{st}}^2 = \sum_{j=1}^{J} P_j^2 (1 - f_j) \sigma_j^2 / n_j,\ f_j = n_j / N_j$$

Bayes for stratified normal model

- Bayes inference for this model is equivalent to standard classical inference for the population mean from a stratified random sample
- The posterior mean weights case by inverse of inclusion probability:

$$\bar{y}_{\text{st}} = N^{-1}\sum_{j=1}^{J} N_j \bar{y}_j = N^{-1}\sum_{j=1}^{J}\sum_{i:x_i=j} y_i / \pi_j,$$

where $\pi_j = n_j / N_j =$ selection probability in stratum j.

- With unknown variances, Bayes' for this model with flat prior on log(variances) yields useful t-like corrections for small samples

Suppose we ignore stratum effects?

- Suppose we assume instead that:

$$[y_i \mid z_i = j] \sim_{ind} N(\theta, \sigma^2),$$

  the previous model with no stratum effects.

- With a flat prior on the mean, the posterior mean of $\bar{Y}$ is then the unweighted mean

$$E(\bar{Y} \mid Y_{\text{inc}}, Z, \sigma^2) = \bar{y} \equiv \sum_{j=1}^{J} p_j \bar{y}_j,\ p_j = n_j / n$$

- This is potentially a very biased estimator if the selection rates $\pi_j = n_j / N_j$ vary across the strata
  - The problem is that results from this model are highly sensitive violations of the assumption of no stratum effects … and stratum effects are likely in most realistic settings.
  - Hence prudence dictates a model that allows for stratum effects, such as the model in the previous slide.

Design consistency

- Loosely speaking, an estimator is *design-consistent* if (irrespective of the truth of the model) it converges to the true population quantity as the sample size increases, holding design features constant.
- For stratified sampling, the posterior mean $\bar{y}_{st}$ based on the stratified normal model converges to $\bar{\bar{Y}}$, and hence is design-consistent
- For the normal model that ignores stratum effects, the posterior mean $\bar{y}$ converges to

$$\bar{\bar{Y}}_{\pi} = \sum_{j=1}^{J} \pi_j N_j \bar{\bar{Y}}_j \Big/ \sum_{j=1}^{J} \pi_j N_j$$

  and hence is not design consistent unless $\pi_j = const.$
- We generally advocate Bayesian models that yield design-consistent estimates, to limit effects of model misspecification

Target and working models

- I think it's helpful to distinguish between
- *Target model:* the model that determines the target parameter/quantity of interest
- *Working model*: the model used to model the data (i.e. to predict the non-sampled values in the population)
- In our simple setting, target model does not condition on *Z:* $[y_i \mid z_i = j] \sim_{ind} N(\theta, \sigma^2)$
  - Target quantity, the overall population mean, results from fitting this model to whole population
- Working model needs to condition on *Z*

$$[y_i \mid z_i = j] \sim_{\text{ind}} N(\theta_j, \sigma_j^2)$$

Weighting in regression

In multiple linear regression, standard method of estimation is ordinary least squares (OLS)

- Model-based: If residual variance is not constant, weight by inverse of residual variance

  $\mathrm{Var}(y_i) = \sigma^2 / u_i \Rightarrow$ weighted LS with weight $\propto u_i$

- Design-based: OLS wrong, weight by inverse of probability of selection, $w_i = 1/\pi_i$
- Which is right? Need to consider variables leading to the sampling weight, and how they enter the regression model

Regression with sample weights

- Target model:

  $y_i \mid x_i \sim N(\beta_0 + \beta^T x_i, \sigma^2 / u_i)$, u_i known (constant for OLS)
- Target parameter: β
- Corresponding finite population parameter: B = result of fitting model to the entire population

  $z_i =$ design variables leading to sampling weights

  (stratum, size in pps sample)
- Consider three cases:
- (a) z_i included as part of x_i

  (b) z_i not a part of x_i

  (c) $z_i = (z_{i1}, z_{i2})$, z_{i1} a part of x_i, z_{i2} not a part of x_i

Regression with sample weights

(a) z_i included as part of x_i

If working model is correctly specified, then regression with weight u_i is correct – no need to include the sample weight

Design-weighted regression with weight $u_i w_i$ yields a design-consistent estimate of the target population quantity *B*. If this differs markedly from model estimate with weight u_i, this suggests model is misspecified, and assumptions need checking.

Regression with sample weights

(b) z_i not a part of x_i

Working model with weight u_i is subject to a known selection bias arising from the stratified design – only valid if this selection does not affect the target parameter estimate

Principled modeling approach is to regress y_i on x_i and z_i and then average over the distribution of z_i given x_i; e.g. if $E(y_i \mid x_i, z_i) = \gamma_0 + \gamma_1 x_i + \gamma_2 z_i$ then $E(y_i \mid x_i) = \gamma_0 + \gamma_1 x_i + \gamma_2 E(z_i \mid x_i, \psi)$, etc.

Bayes simulation: impute draws of the non-sampled values of Y based on regression of Y on X, Z, and then fit regression of Y on X to imputed population. Repeat to simulate posterior distribution of β

Regression with sample weights

(b) z_i not a part of x_i

Pragmatic approach: design-based regression of y_i on x_i with weights $w_i u_i$

Model-based justification: assume a working model with a different regression model for y_i on x_i within each stratum defined by Z. Regression of y_i on x_i with weight $w_i u_i$ then approximates the posterior mean of β. (Little 2004, Example 11)

Pragmatic approach B: compare regression of y_i on x_i with weights $w_i u_i$ with regression of y_i on x_i with weights u_i. If coefficients of interest are close, effects of selection may be ignored, leading to model-based solution.

Regression with sample weights

(c) $z_i = (z_{i1}, z_{i2})$, z_{i1} a part of x_i, z_{i2} not a part of x_i

Principled modeling approach is to regress y_i on x_i and z_{i2} and then average over the distribution of z_{i2} given x_i; e.g. if $E(y_i \mid x_i, z_{i2}) = \gamma_0 + \gamma_1 x_i + \gamma_2 z_{i2}$ then

$E(y_i \mid x_i) = \gamma_0 + \gamma_1 x_i + \gamma_2 E(z_{i2} \mid x_i, \psi)$, etc.

Bayes simulation: impute draws of the non-sampled values of Y based on regression of Y on X, Z_2, and then fit regression of Y on X to imputed population. Repeat to simulate posterior distribution of β

Regression with sample weights

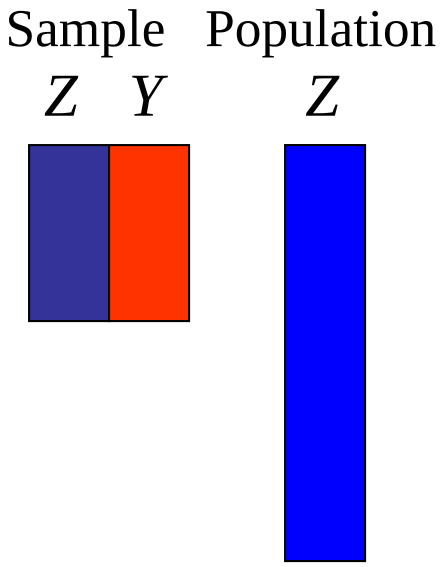
(c) $z_i = (z_{i1}, z_{i2})$, z_{i1} a part of x_i, z_{i2} not a part of x_i

Pragmatic approach: design-based regression of y_i on x_i with weights $w_{i2}u_i$, where w_{i2} is component of sampling weight attributable to z_{i2} (given z_{i1}).
(Weighting on w_iu_i is ok but inefficient)

Pragmatic approach B: compare regression of y_i on x_i with weights w_iu_i with regression of y_i on x_i with weights u_i. If coefficients of interest are close, effects of selection may be ignored, leading to model-based solution.

Ex 4. One continuous (post)stratifier Z

Consider PPS sampling, Z = measure of size



Standard design-based estimator is weighted Horvitz-Thompson estimate

$$\bar{y}_{\text{HT}} = \frac{1}{N}\left(\sum_{i=1}^{n} y_i / \pi_i\right); \pi_i = \text{ selection prob (HT)}$$

$\bar{y}_{\text{HT}} \approx$ model-based prediction estimate for

$y_i \sim \text{Nor}(\beta\pi_i, \sigma^2\pi_i^2)$ ("HT model")

When the relationship between Y and Z deviates a lot from the HT model, HT estimate is inefficient and CI's can have poor coverage

Ex. Basu's inefficient elephants

$(y_1, ..., y_{50}) =$ weights of $N = 50$ elephants

Objective: $T = y_1 + y_2 + ... + y_{50}$. Only one elephant can be weighed!

- Circus trainer wants to choose "average" elephant (Sambo)
- Circus statistician requires "scientific" prob. sampling:

  Select Sambo with probability 99/100

  One of other elephants with probability 1/4900

  Sambo gets selected! Trainer: $\hat{t} = y_{(\text{Sambo})} \times 50$

  Statistician requires unbiased Horvitz-Thompson (1952) estimator:

$$\hat{T}_{HT} = \begin{cases} y_{(\text{Sambo})} / 0.99 \text{ (!!)}; \\ 4900 y_{(i)}, \text{if Sambo not chosen (!!!)} \end{cases}$$

  HT estimator is unbiased on average but always crazy!

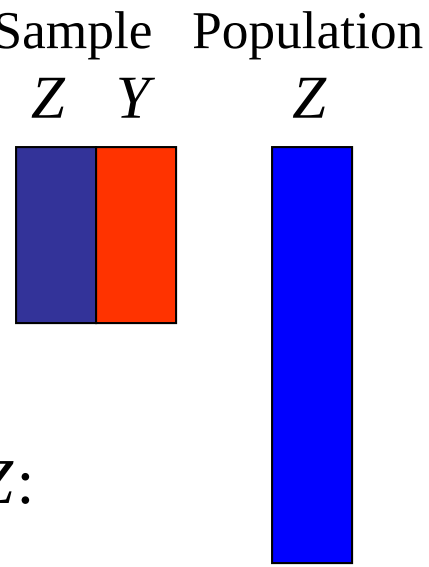
  HT model is clearly hopeless here …

What went wrong?

- HT estimator optimal under an implicit HT model that y_i / π_i have the same distribution
- That is clearly a silly model given this design …
- Which is why the estimator is silly

Ex 4. One continuous (post)stratifier Z

$$\bar{y}_{\text{wt}} = \frac{1}{N}\left(\sum_{i=1}^{n} y_i / \pi_i\right); \pi_i = \text{ selection prob (HT)}$$



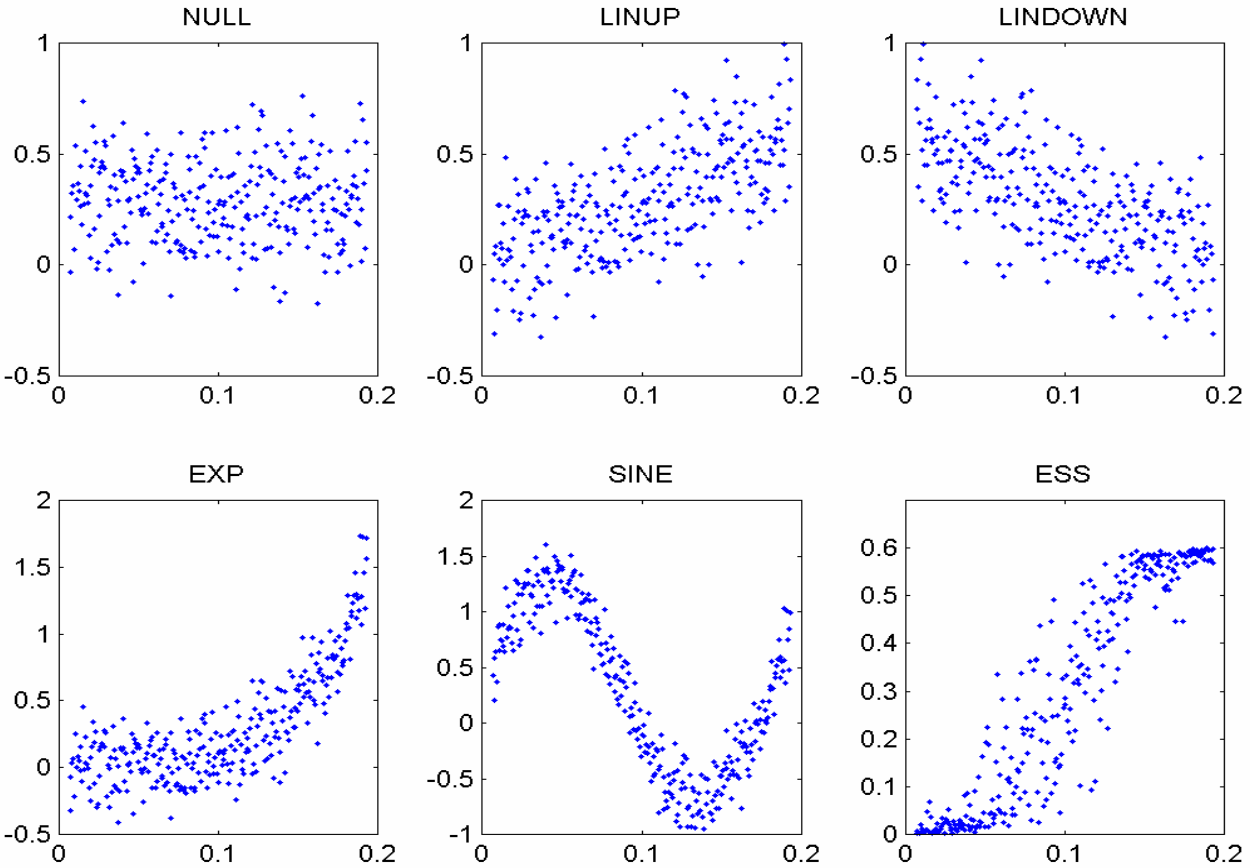
A modeling alternative to the HT estimator is create predictions from a more robust model relating Y to Z:

$$\bar{y}_{\text{mod}} = \frac{1}{N}\left(\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{y}_i\right), \ \hat{y}_i \text{ predictions from:}$$

$y_i \sim \text{Nor}(S(\pi_i), \sigma^2 \pi_i^k)$; $S(\pi_i)$ = penalized spline of Y on Z

(Zheng and Little 2003, 2005)

Simulation: PPS sampling in 6 populations



Estimated RMSE of four estimators for N=1000, n=100

Population		model	wt	gr
NULL	Normal	**20**	33	21
	Lognormal	32	44	**31**
LINUP	Normal	**23**	24	25
	Lognormal	**25**	30	30
LINDOWN	Normal	30	66	**29**
	Lognormal	**24**	65	28
SINE	Normal	**35**	134	90
	Lognormal	**53**	130	84
EXP	Normal	**26**	32	57
	Lognormal	**40**	41	58

95% CI coverages: HT

Population	V1	V3	V4	V5
NULL	90.2	91.4	90.0	90.4
LINUP	94.0	95.0	95.0	95.0
LINDOWN	89.0	89.8	90.0	90.6
SINE	93.2	93.4	93.0	93.0
EXP	93.6	94.6	95.0	95.0
ESS	95.0	95.6	95.4	95.2

V1 Yates-Grundy, Hartley-Rao for joint inclusion probs.
V3 Treating sample as if it were drawn with replacement
V4 Pairing consecutive strata
V5 Estimation using consecutive differences

95% CI coverages: B-spline

Population	V1	V2	V3
NULL	95.4	95.8	95.8
LINUP	94.8	97.0	94.6
LINDOWN	94.2	94.2	94.6
SINE	**88.0**	92.6	97.4
EXP	94.4	95.2	95.6
ESS	97.4	95.4	95.8

Fixed with more knots

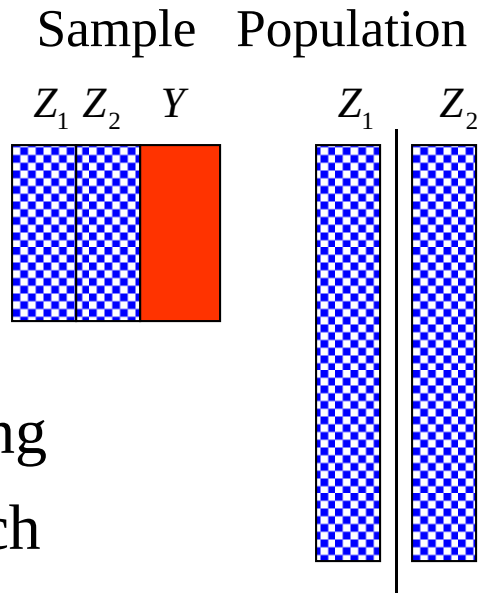
V1 Model-based (information matrix)
V2 Jackknife
V3 BRR

Why does model do better?

- Assumes smooth relationship – HT weights can "bounce around"
- Predictions use sizes of the non-sampled cases
  - HT estimator does not use these
  - Often not provided to users (although they could be)
- Little & Zheng (2007) also show gains for model when sizes of non-sampled units are not known
  - Predicted using a Bayesian Bootstrap (BB) model
  - BB is a form of stochastic weighting

Ex 3. One stratifier Z_1, one post-stratifier Z_2

<u>Design-based approaches</u>



(A) Standard weighting is $w_i = w_{is} \times w_{ip}(w_{is})$

Notes: (1) Z_1 proportions are not matched!

(2) why not $w_i^* = w_{ip} \times w_{is}(w_{ip})$?

(B) Deville and Sarndal (1992) modifies sampling weights $\{w_{is}\}$ to adjusted weights $\{w_i\}$ that match poststratum margin, but are close to $\{w_{is}\}$ with respect to a distance measure $d(w_{is}, w_i)$.

Questions:

What is the principle for choosing the distance measure?

Should the $\{w_i\}$ necessarily be close to $\{w_{is}\}$?

Ex 3. One stratifier Z_1, one post-stratifier Z_2

Model-based approach

Saturated model: $\{n_{jk}\} \sim \text{MNOM}(n, \pi_{jk})$;

$$y_{jki} \sim \text{Nor}(\mu_{jk}, \sigma^2_{jk})$$

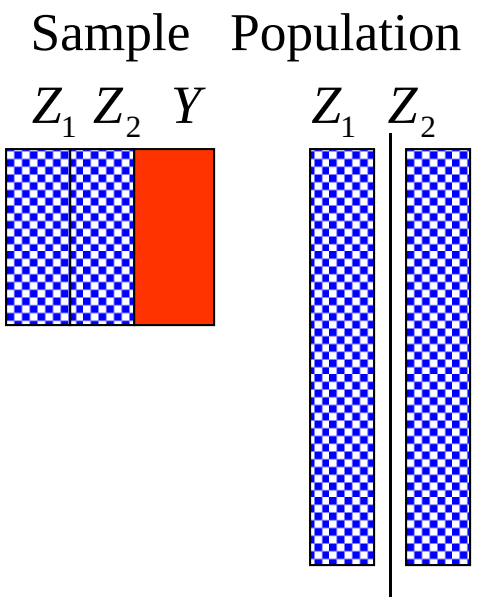
$$\bar{y}_{\text{mod}} = \sum_{j=1}^{J}\sum_{k=1}^{K} \hat{P}_{jk}\bar{y}_{jk} = \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk} n_{jk} \bar{y}_{jk} \Big/ \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk} n_{jk}$$

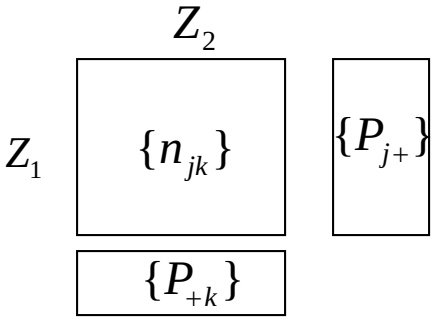
n_{jk} = sample count, $\bar{y}_{jk}$ = sample mean of Y

$\hat{P}_{jk}$ = proportion from raking (IPF) of $\{n_{jk}\}$

to known margins $\{P_{j+}\}, \{P_{+k}\}$

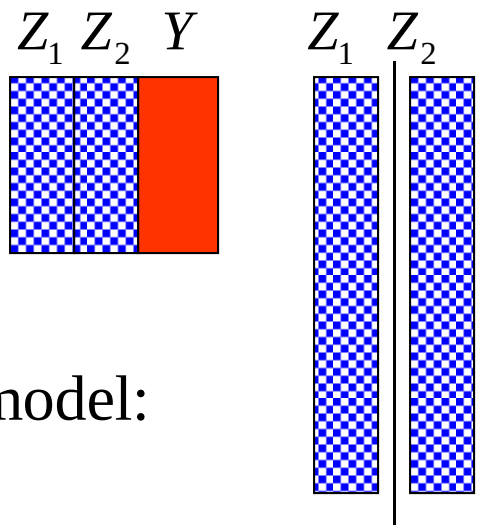
$w_{jk} = n\hat{P}_{jk} / n_{jk}$ = model weight





Ex 3. One stratifier Z_1, one post-stratifier Z_2

Sample Population



Model-based approach

$$\bar{y}_{\text{st}} = \sum_{j=1}^{J}\sum_{k=1}^{K} \hat{P}_{jk}\bar{y}_{jk} = \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk}n_{jk}\bar{y}_{jk} \Big/ \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk}n_{jk}$$

What to do when n_{jk} is small?

Model: replace $\bar{y}_{jk}$ by prediction from modified model:

e.g. $y_{jki} \sim \text{Nor}(\mu + \alpha_j + \beta_k + \gamma_{jk}, \sigma^2_{jk})$,

$$\sum_{j=1}^{J}\alpha_j = \sum_{k=1}^{K}\beta_k = 0,\ \gamma_{jk} \sim \text{Nor}(0, \tau^2)$$ (Gelman 2007)

Setting $\tau^2 = 0$ yields additive model,

otherwise shrinks towards additive model

Design: arbitrary collapsing, ad-hoc modification of weight

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
- But outcome on subjects within a cluster may be correlated (typically, positively).
- Models can easily incorporate the correlation among observations

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,…,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 σ *and* τ *are known*

- Prior distribution

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

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 $E(\mu_i | Y_{\text{inc}})$ (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: a known function

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

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

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

Bayesian Inference for Sample Surveys

Roderick Little

Module 4: Bayesian computation and model assessment

Bayesian computation

- As previously discussed, summaries of the posterior distribution are used for statistical inferences
  - Means, Median, Modes or measures of central tendency
  - Standard deviation, mean absolute deviation or measures of spread
  - Percentiles for intervals
- Conceptually, all these quantities can be expressed analytically in terms of multidimensional integrals
- Extensive work on methods for computing these integrals has made Bayesian methods practically feasible. I'll review some important computational approaches
  - Numerical integration routines
  - Simulation techniques

Simple Numerical Integration Methods

- Finite integrals can be approximated as sums
  - Trapezoidal rule, Simpson's rule, Newton-Cotes method
- Gaussian quadrature – applies to infinite integrals
- R package statmod
- Abramowitz, M and Stegun, C. A. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, New York: Dover

Simulation Techniques

- Numerical integration can be extended to multidimensional integrals but becomes impractical in high dimensions, and error in approximations can be large
- An attractive alternative is to *draw* samples from the posterior distribution and use the sample to characterize the features of the posterior distribution – many clever ways to do this
- Magically, this avoids multidimensional integration -- which is the key for models with many parameters
- These methods can be computer intensive – but that is no longer a problem these days, given fast computers!

Drawing model parameters

To simulate posterior distribution of $\phi(\theta)$, $\theta = (\theta_1, ..., \theta_K)$:

For $d = 1, ...D$ (D large):

Draw $\theta^{(d)}$ from posterior distribution

Compute $\phi^{(d)} = \phi(\theta^{(d)})$

Then approximate:

Posterior mean by $\bar{\phi} = \sum_{d=1}^{D} \phi^{(d)} / D$

Posterior sd by $s_\phi = \sum_{d=1}^{D} \left(\phi^{(d)} - \bar{\phi} \right)^2 / (D-1)$

95% credibility interval by $(\bar{\phi} \pm 1.96 s_\phi)$

Or 2.5 to 97.5 percentiles of sample distribution $\{\phi^{(d)} : d = 1, .., D\}$

Drawing population quantities

To simulate posterior distribution of $Q(Y)$, Y = population data:

For $d = 1,...D$ (D large):

Draw $\theta^{(d)}$ from posterior distribution

Draw $Y^{(d)}$ from $p(Y|$ data, $\theta^{(d)})$ (Fills in the population data)

Compute draw of Q, $Q^{(d)} = Q(Y^{(d)})$

Then approximate:

Posterior mean by $\bar{Q} = \sum_{d=1}^{D} Q^{(d)} / D$

Posterior sd by $s_Q = \sum_{d=1}^{D} \left(Q^{(d)} - \bar{Q}\right)^2 / (D-1)$

95% credibility interval by $(\bar{Q} \pm 1.96 s_Q)$

Or 2.5 to 97.5 percentiles of sample distribution $\{Q^{(d)} : d = 1,..,D\}$

Simulation methods

- Direct simulation
- Approximate direct simulation
  - Discrete approximation of the posterior density
  - Rejection sampling
  - Sampling Importance Resampling
- Iterative simulation techniques
  - Metropolis Algorithm
  - Gibbs sampler

Simulation for the Normal Example

- Revisit normal example

$$\sigma^2 \mid y_{\text{inc}} \sim (n-1)s^2 / \chi^2_{n-1}$$

$$\mu \mid \sigma^2, y_{\text{inc}} \sim N(\bar{y}, \sigma^2 / n)$$

$$\bar{Y}_k \mid \mu, \sigma^2, y_{\text{inc}} \sim N(\mu, \sigma^2 / k)$$

```
# Draws for the normal case
sampsize=20
k=5
ybar=10
ssquare=5
nsimul=1000
result=matrix(0,nsimul,3)
for (i in 1:nsimul){
tmp=rnorm(sampsize-1)
chisq=sum(tmp*tmp)
sigmasq=(sampsize-1)*ssquare/chisq;
mu=ybar+sqrt(sigmasq/sampsize)*rnorm
(1)
ybark=mu+sqrt(sigmasq/k)*rnorm(1)
result[i,1]=sigmasq
result[i,2]=mu
result[i,3]=ybark}
```

Multivariate Example

- In an investigation several versions of a question asking about an outcome Y were to be investigated. The true values of Y were known for a sample of subjects.
- The m versions of the questions were administered to the same sample resulting in measurements $x_1, x_2, x_3,\ldots,x_{\mathrm{m}}$
- Objective is to infer about the largest of the m correlation coefficients

$$\rho_{y,x_j}; j = 1,2,...,m$$

Example: Model

- Suppose that these measures are continuous and a multivariate normal model is posited:

$$U = (Y, X_1, X_2, \ldots, X_m) \sim MVN_{m+1}(\mu, \Sigma)$$

$$\pi(\mu, \Sigma) \propto |\Sigma^{-1}|^{-(m+1)/2}$$

- It is analytically difficult to derive the posterior distribution of $\theta = \underset{1 \le j \le 15}{\text{Max}}(\rho_{y,x_j})$
- Even more interesting is to find the posterior mean of

$$\lambda_j = \Pr(\rho_{y,x_j} \ge \rho_{y,x_i} \forall i \ne j)$$

- Likelihood

$$\prod_{i=1}^{n} |\Sigma|^{-1/2} \exp[-(U_i - \mu)^t \Sigma^{-1} (U_i - \mu)/2]$$

$$= |\Sigma|^{-n/2} \exp\left[-\sum_i (U_i - \bar{U})^t \Sigma^{-1} (U_i - \bar{U})/2\right] \times$$

$$\exp\left[-n(\mu - \bar{U})^t \Sigma^{-1} (\mu - \bar{U})/2\right]$$

- Posterior distribution

$$\left[|\Sigma^{-1}|^{(n-m-2)/2} \exp\left[-Tr(S\Sigma^{-1})/2\right]\right] \times$$

$$\left[|\Sigma/n|^{-1/2} \exp\left[-(\mu - \bar{U})^t (\Sigma/n)^{-1} (\mu - \bar{U})/2\right]\right]$$

Wishart and Inverse-Wishart Distributions

$Z =$ Positive definite symmetric random matrix of dimension p with $p(p+1)/2$ distinct random variables.

Z has a Wishart distribution if

$$pdf(Z) = C\,|B|^{-\nu/2}|Z|^{(\nu-p-1)/2}\exp[-Tr(B^{-1}Z)/2]$$

$$C^{-1} = 2^{\nu p/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma((\nu+1-i)/2)$$

$Z \sim \text{Wishart}(B,\nu)$

$U \sim \text{Inv-Wishart}(B,\nu)$ if $U^{-1} \sim \text{Wishart}(B,\nu)$

Example:Simulation

- It is easy to simulate from the posterior distribution of μ and Σ.

$$\Sigma^{-1} \mid \text{Data} \sim \text{Wishart}(S^{-1}, n-1)$$

$$S = \sum_{i=1}^{n} (u_i - \bar{u})(u_i - \bar{u})^t$$

$$\bar{u} = \sum_i u_i / n$$

Generate $z_j \sim N(0, S^{-1})$; $j = 1, 2, ..., n-1$

Define $\Sigma_*^{-1} = \sum_j z_j z_j^t$

Generate $\mu_* \sim N(\bar{u}, \Sigma_* / n)$

- Also $\mu \mid \text{Data}, \Sigma \sim N(\bar{u}, \Sigma / n)$
- Compute the desired function of (μ_*, Σ_*)
- Repeat the above steps to simulate several draws from the posterior distribution.

Rejection Sampling

- Actual Density from which to draw from
- Candidate density from which it is easy to draw
- The importance ratio is bounded
- Sample θ from g, accept θ with probability p otherwise redraw from g

$$\pi(\theta \mid \text{data})$$

$$g(\theta), \text{ with } g(\theta) > 0 \text{ for all } \theta$$

$$\text{with } \pi(\theta \mid \text{data}) > 0$$

$$\frac{\pi(\theta \mid \text{data})}{g(\theta)} \leq M$$

$$p = \frac{\pi(\theta \mid \text{data})}{M \times g(\theta)}$$

Sampling Importance Resampling

- Target density from which to draw

$$\pi(\theta \mid \text{data})$$

- Candidate density from which it is easy to draw

$$g(\theta), \text{ such that } g(\theta) > 0$$

$$\text{for all } \theta \text{ with } \pi(\theta \mid \text{data}) > 0$$

- The importance ratio

$$w(\theta) \propto \frac{\pi(\theta \mid \text{data})}{g(\theta)}$$

- Sample M values of θ from g

$$\theta_1^*, \theta_2^*, \ldots, \theta_M^*$$

- Compute the M importance ratios and resample with probability proportional to the importance ratios.

$$w(\theta_i^*); i = 1,2,\ldots, M$$

Markov Chain Monte Carlo

- In real problems it may be hard to apply direct or approximate direct simulation techniques.
- The Markov chain Monte Carlo (MCMC) methods involve a random walk in the parameter space which converges to a stationary distribution that is the target posterior distribution. Two popular methods are
  - Gibbs sampling (BUGS software)
  - Metropolis-Hastings algorithms
- Once a MCMC chain has converged to the target distribution, use subsequent draws to approximate the posterior distribution (this are dependent, but that is generally not a problem)

Gibbs sampling

- Gibbs sampling a particular MCMC method for sampling from multivariate problems

$$\underline{x} = (x_1, x_2, ..., x_p) \sim f(\underline{x})$$

$$f(x_i \mid x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_p)$$

Gibbs sequence:

$$x_1^{(t+1)} \sim f(x_1 \mid x_2^{(t)}, x_3^{(t)}, ..., x_p^{(t)})$$

$$x_2^{(t+1)} \sim f(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, ..., x_p^{(t)})$$

$$\vdots$$

$$x_i^{(t+1)} \sim f(x_i \mid x_1^{(t+1)}, ..., x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, ..., x_p^{(t)})$$

$$\vdots$$

$$x_p^{(t+1)} \sim f(x_p \mid x_1^{(t+1)}, ..., x_{p-1}^{(t+1)})$$

1. This is a Markov Chain whose stationary Distribution is $f(\underline{x})$
2. Useful if the conditional densities are easy to work with
3. If not, then use Metropolis-Hastings or rejection sampling within the Gibbs sequence

Metropolis-Hastings algorithm

To draw from a density $f(x)$

Choose a candidate distribution $p(y|x)$ that has the same support as $f(x)$ (preferably longer tails) and is easy to draw from

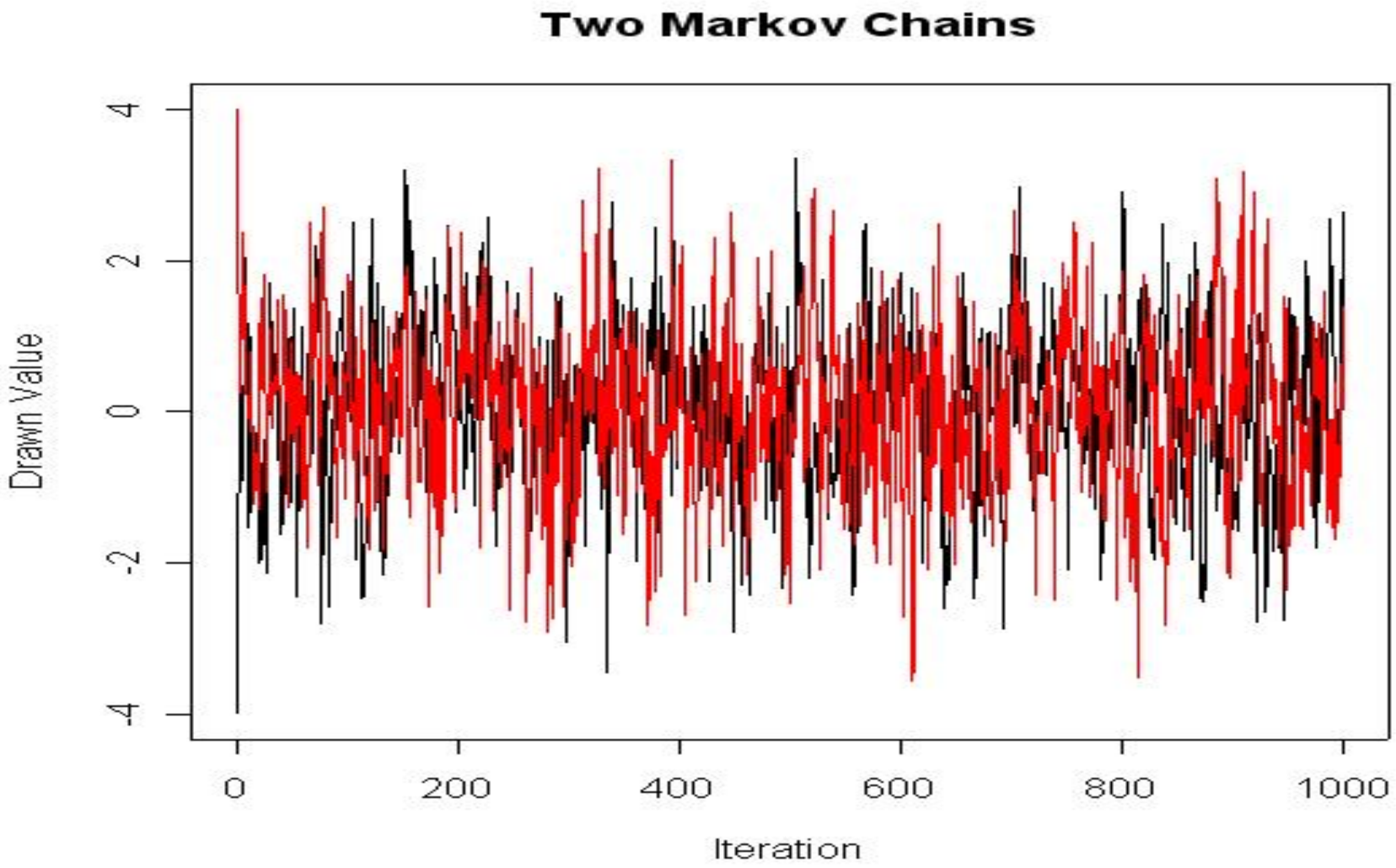
- Step 1 At iteration t, draw $y^{(d)} \sim p(y \mid x^{(t)})$
- Step 2: Compute the ratio $w = Min\left\{1, \frac{f(y) / p(y \mid x^{(t)})}{f(x^{(t)}) / p(x^{(t)} \mid y)}\right\}$
- Step 3: Generate a uniform random number, u and set

  $x^{(t+1)} = y$ if $u \leq w$

  $x^{(t+1)} = x^{(t)}$ otherwise
- This Markov Chain has stationary distribution $f(x)$.

Remarks

- To reduce the impact of starting point the initial few draws are ignored ("Burn-in period")
- Successive draws are dependent. Sample every k^{th} observation to get independent draws, but this does not appear to be necessary.
- Assessing whether or not the chain has converged is tricky – one approach is to run a set (say 10) chains with different starting values and asess when they are completely mixed



Two sequences are essentially indistinguishable after 5 or 6 iterations. This doesn't happen to all Markov Chains. The stationary distribution is unique only if the Markov Chain is irreducible and aperiodic.

Gelman-Rubin statistic for assessing convergence

- Suppose that there are J parallel sequences each with n iterations. Let θ be the scalar parameter of interest.
- Between variance $$B = n\sum_{j=1}^{J}(\bar{\theta}_{+j} - \bar{\theta}_{++})^2 / (J-1)$$
$$\bar{\theta}_{+j} = \sum_i \theta_{ij} / n; \qquad \bar{\theta}_{++} = \sum_j \bar{\theta}_{+j} / J$$
- Within variance $$W = \sum_i \sum_j (\theta_{ij} - \bar{\theta}_{+j})^2 / \left(J(n-1)\right)$$

Gelman-Rubin statistic

- At convergence the statistic

$$R = \sqrt{\frac{n-1}{n} + \frac{B}{nW}}$$

should be approximately equal 1. Note that "Between" variance cannot be computed without multiple sequences.

Model Checking

- Model checking is an important step in a Bayesian analysis
- Standard approaches to checking models, such as residual plots for regression, can be applied
- Posterior predictive checks generate new data from the posterior predictive distribution and compare generated values of statistics of interest with values computed on the observed data.

$y =$ Observed data

$\theta =$ Parameters in the model

$f(y|\theta)$: Conditional distribution of observables given the parameter θ

$\pi(\theta)$=Prior density

Posterior Predictive Check

- Posterior predictive distribution

$$f(y_{\text{new}} \mid y_{\text{inc}}) = \int f(y_{\text{new}} \mid \theta, y_{\text{inc}})\pi(\theta \mid y_{\text{inc}})d\theta$$

$$= \int f(y_{\text{new}} \mid \theta)\pi(\theta \mid y_{\text{inc}})d\theta$$

- Generate new data from the posterior-predictive distribution and compare with the observed data y_{inc}

$$\theta^* = \text{draw from } \pi(\theta \mid y_{\text{inc}})$$

$$y_{\text{new}} = \text{ draw from } f(y \mid \theta^*)$$

If the prior distribution is diffuse generate new data from "likelihood" and compare with the observed data

Comparing new and observed data

- Develop "discrepancy measure" $T(y)$ or $T(y,\theta)$. Note that the discrepancy measure can depend upon the parameter θ.
- Compare $T(y_{\text{inc}})$ with $T(y_{\text{new}})$ or

$$T(y_{\text{inc}},\theta^*) \text{ with } T(y_{\text{new}},\theta^*)$$

- Discrepancy measures depends upon the problem
- General goodness of fit measure:

$$T(y,\theta) = \sum_{i=1}^{n} \frac{(y_i - E(y_i \mid \theta))^2}{\text{var}(y_i \mid \theta)}$$

Example: checking the independence in a sequence of Bernoulli trials

Model : $y_i \sim$ iid $\text{Bin}(1,\theta)$, $\theta \sim \text{Unif}(0,1)$

Observed sequence:

$y_{\text{inc}} = \{1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0\}$

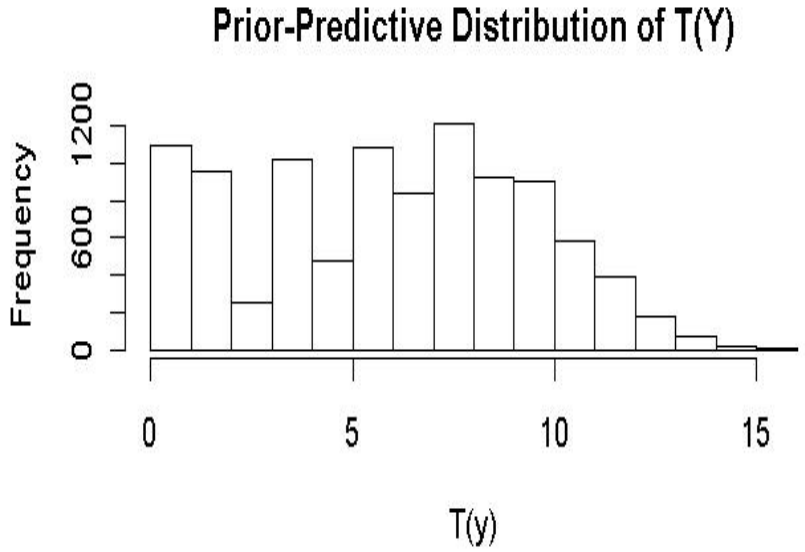
T = number of runs of same value, $T(y_{\text{inc}}) = 3$

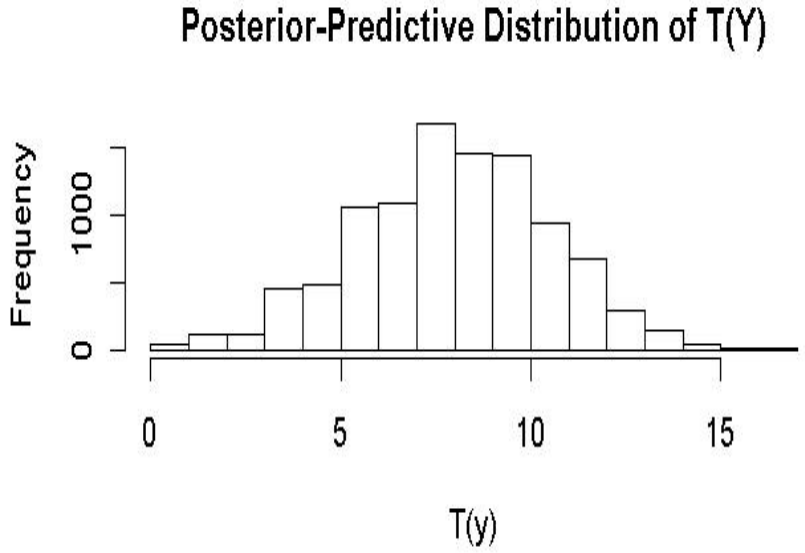
To simulate the posterior predictive distribution:

For $d = 1,...,D$

Draw $\theta^{\text{rep},d} \sim \text{Beta}(8,14)$, $y_j^{\text{rep},d} \sim \text{Bin}(1,\theta^{\text{rep}})$, $j = 1,2,...,20$

Draw $T^{(d)} = T(y^{\text{rep},d})$





Posterior Predictive frequency	Prior Predictive frequency	T(y)
31	994	0
7	99	1
120	955	2
111	253	3
459	1022	4
477	472	5
1064	1090	6
1084	838	7
1671	1210	8
1457	920	9
1435	900	10
934	579	11
669	394	12
296	173	13
137	74	14
33	21	15
12	6	16
3	0	17

Remarks

- Model checking is an important step. The posterior predictive check may give indication about lack of fit
- Prior knowledge is also very important in making judicious choice of the models.
- It may be difficult to pin down one model that may be satisfactory in every aspect
- It is better to consider a continuum of models and perform sensitivity analysis by inspecting inferences under these models
- Generally need some prior information to handle extrapolation outside the range of observed data

Model Comparisons 1

- K models under consideration

$$M_j, j = 1, 2, ..., K$$

- Prior probabilities for model j being correct

$$\{p_j, j = 1, 2, ..., K\}; \sum_{j=1}^{K} p_j = 1$$

- Model $M_j : [f_j(y \mid \theta_j), \pi_j(\theta_j)]$
- Posterior probability for model j being correct

$$\Pr(M_j \mid y) = \frac{p_j L_j(M_j \mid y)}{\sum_{j=1}^{K} p_j L_j(M_j \mid y)},$$

$$L(M_j \mid y) = \int f_j(y \mid \theta_j) \pi_j(\theta_j) d\theta_j$$

Model Comparisons 2

Posterior odds = Prior odds x Bayes Factor:

$$\frac{\Pr(M_j \mid y)}{\Pr(M_l \mid y)} = \frac{p_j}{p_l} \times \frac{L(M_j \mid y)}{L(M_l \mid y)}$$

Bayes factor

$$\log(\text{Bayes Factor}) \approx$$

$$\log(L(\hat{\theta}_j \mid y, M_j)) - \log(L(\hat{\theta}_l \mid y, M_l)) + (k_j - k_l)\log(n)/2$$

$L(\hat{\theta} \mid y, M) =$ likelihood evaluated at

ML estimate under model M

k = number of parameters

Model Comparisons 3

- Bayes Information Criterion

$$\mathrm{BIC}(M) = \log(L(\hat{\theta} \mid y, M)) - k \log(n)/2$$

$$\text{Bayes factor}(M_j \ vs \ M_l) \approx \mathrm{BIC}(M_j) - \mathrm{BIC}(M_l)$$

- The approximation has been derived for nested models
- The Bayes factor, however, can be applied for non-nested models
- The marginal distribution has to be proper. This is not guaranteed when a non-informative prior is used for the parameter

Model Averaging

- Original model: $\{f(y|\theta), \pi(\theta)\}$
- Consider an expanded class of model

$$\{f(y|\theta,\phi), \pi(\theta|\phi), p(\phi)\}$$

where the original model is a member of this expanded class

- Inference from

$$\pi(\theta|y) = \int \pi(\theta,\phi|y)d\phi$$

$$\propto \int f(y|\theta,\phi)\pi(\theta|\phi)\,p(\phi)d\phi$$

Example

- Original model

$$y \sim N(\mu, \sigma^2), \pi(\mu, \sigma^2) \propto \sigma^{-2}$$

- Expanded model

$$y \sim t_\nu(\mu, \sigma^2), \pi(\mu, \sigma^2, \nu) \propto \sigma^{-2}\nu^{-2}$$

$$t_\infty(\mu, \sigma^2) \equiv N(\mu, \sigma^2)$$

$$t_1(\mu, \sigma^2) \equiv \text{Cauchy}$$

- Prior distribution gives more weight towards normal range

Example

- Original model

$$y \sim N(\mu, \sigma^2),\ \pi(\mu, \sigma^2) \propto \sigma^{-2}$$

- Expanded model

$$\alpha N(\mu, \sigma^2) + (1-\alpha) g(\mu + \delta, \rho^2 \sigma^2)$$

where g is a member of some location-scale family of distribution

Bayesian inference for sample surveys

Roderick Little

Module 5: Missing Data

Missing data methods -- history

1. Before the EM algorithm (pre-1970's)
   - Ad-hoc adjustments (simple imputation)
   - ML for simple problems (Anderson 1957)
   - ML for complex problems too hard
2. ML era (1970's – mid 1980's)
   - Rubin formulates model for missing data mechanism, defines MAR (1976)
   - EM and extensions facilitate ML for complex problems
   - ML for more flexible models – beyond multivariate normal (see e.g. Little and Rubin 1987)

Missing data methods -- history

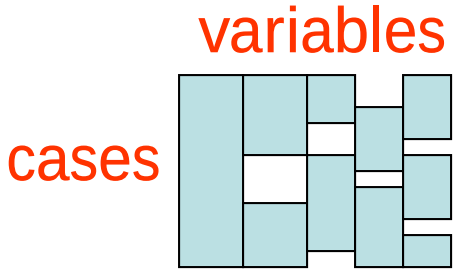
3. Bayes and Multiple Imputation (mid 1980's – present)
   - Tanner and Wong describes data augmentation for the multivariate normal problem (1984)
   - Rubin proposes MI, justified via Bayes (1977, 1987)
   - MCMC facilitates Bayes as an alternative to ML, with better small sample properties (see e.g. Little and Rubin 2002)
4. Robustness concerns (1990's – present)
   - Robins et al propose doubly robust methods for missing data
   - Robust Bayesian models, more attention to model checks

Likelihood methods with missing data

- Likelihood methods do not require rectangular data, hence apply directly to missing-data problems
- Statistical model + incomplete data $\Rightarrow$ Likelihood
- Approaches based on the likelihood:
  - ML estimates, large sample standard errors
  - Bayes: add priors, compute posterior distribution
  - Multiple imputation: multiple draws of missing values, apply MI combining rules
  - Flexible and general
  - Methods reflect added uncertainty from missing data

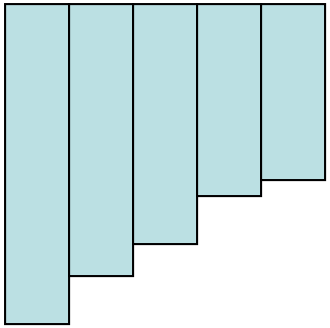
Patterns of Missing Data

- Item nonresponse: general pattern

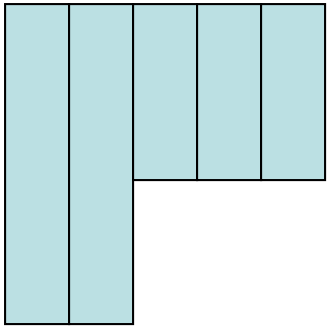


- special patterns

monotone

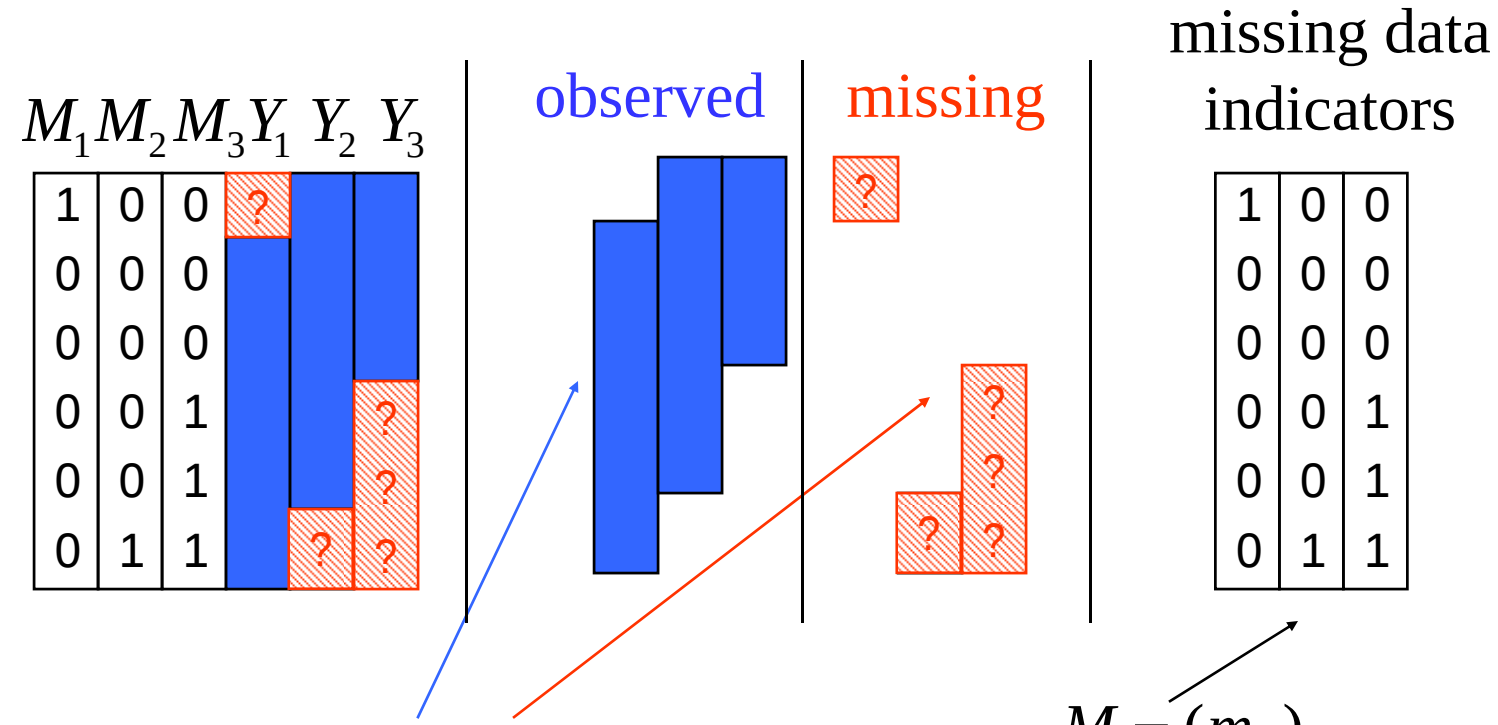


unit nonresponse



Bayes can handle general patterns with relative ease

The Observed Data



$$Y = (y_{ij})_{n\times K} = (Y_{\text{obs}}, Y_{\text{mis}})$$

$$M = (m_{ij})_{n\times K}$$

$$m_{ij} = \begin{cases} 0, y_{ij} \text{ observed} \\ 1, y_{ij} \text{ missing} \end{cases}$$

Two posterior distributions

$$p_{\text{complete}}(\theta,\psi \mid Y,M) = \pi(\theta,\psi) \times f(Y \mid \theta) \times f(M \mid Y,\psi)$$

Prior dn Complete-data model model for mechanism

- *Full* posterior distribution - involves model for *M*

$$p_{\text{full}}(\theta,\psi \mid Y_{\text{obs}},M) \propto \pi(\theta,\psi) \times f(Y_{\text{obs}},M \mid \theta,\psi)$$

$$f(Y_{obs},M \mid \theta,\psi) = \int f(Y_{\text{obs}},Y_{\text{mis}} \mid \theta) f(M \mid Y_{\text{obs}},Y_{\text{mis}},\psi) dY_{\text{mis}}$$

- Posterior dn *ignoring the missing-data mechanism M* (simpler since it does not involve model for *M*)

$$p_{\text{ign}}(\theta \mid Y_{\text{obs}}) \propto \pi(\theta) \times f(Y_{\text{obs}} \mid \theta)$$

$$f(Y_{\text{obs}} \mid \theta) = \int f(Y_{\text{obs}},Y_{\text{mis}} \mid \theta) dY_{\text{mis}}$$

Ignoring the md mechanism continued

- Sufficient conditions for ignoring the missing-data mechanism and basing inference on $p_{\text{ign}}(\theta \mid Y_{\text{obs}})$ are:
- MAR: $f(M \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) = f(M \mid Y_{\text{obs}}, \psi)$ for all Y_{mis}
- Independent priors for parameters of dns of Y and M

$$\pi(\theta, \psi) = \pi_1(\theta) \times \pi_2(\psi)$$

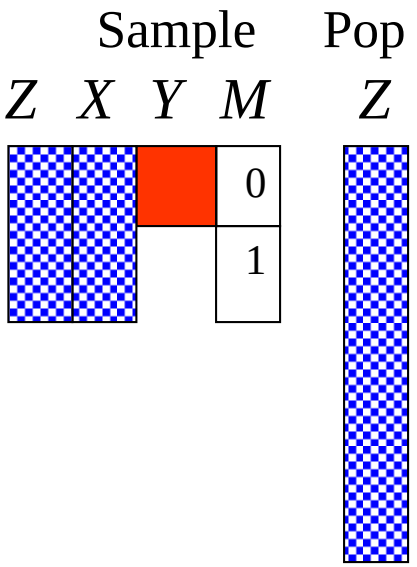
- MAR is the key condition in practice
- Main challenges are choice of model, computation

Remarks

- When data are assumed to be MNAR. One has to specify the relationship between M, the response indicator, and the unobserved portion of substantive data, y_{mis}
  - This assumption cannot be verified form the observed data in hand
  - Some external information is needed to specify this relationship.
  - The MNAR is more or less a subjective opinion and as such inferences are highly sensitive to assumption made about the relationship between M and y_{mis}
- MAR conditional on rich set of covariates may be the most reasonable approach rather than making some empirically unverifiable assumption
- Since missing data may be a problem, attempts should be made to collect a rich set of correlates of missing outcomes
- Use designs to reduce nonresponse bias (multiple matrix sampling)

Unit nonresponse

- Predict nonrespondents by regression on design variables Z and any observed survey variables X
- For bias reduction, predictors should be related to M and outcome Y
- In choosing from a set of predictors, good prediction of Y is more important than good prediction of M
- In particular, consider categorical predictors and flat priors
  - Bayes corresponds to weighting by inverse of estimated response rate in each category



Impact of weighting for nonresponse

$\text{corr}^2(X, Y)$

$\text{corr}^2(X, M)$	Low	High
Low	---	var $\Downarrow$
High	var $\uparrow$	var $\Downarrow$ bias $\Downarrow$

Too often adjustments do this?

- Standard "rule of thumb" $\text{Var}(\bar{y}_w) = \text{Var}(\bar{y}_u)(1 + \text{cv}(w))$ fails to reflect that nonresponse weighting can reduce variance
- Little & Vartivarian (2005) propose refinements

Item nonresponse

- Item nonresponse generally has complex "swiss-cheese" pattern
- Weighting methods are possible when the data have a monotone pattern, but are very difficult to develop for a general pattern
- Model-based multiple imputation methods are available for this situation
- By conditioning fully on all observed data, these methods weaken MAR assumption

Bayesian MCMC Computations

A convenient algorithmic approach for complex problems is to iterate between draws of the missing values and draws of the parameters:

$$(Y_{\text{mis}}^{(d,t+1)} \mid Y_{\text{obs}}, \theta^{(dt)}) \sim p(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta^{(dt)})$$

$$(\theta^{(d,t+1)} \mid Y_{\text{mis}}^{(d,t+1)}) \sim p(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}^{(d,t+1)})$$

As t tends to infinity, this sequence converges to a draw from the joint posterior distribution of (Y_{mis}, θ), as required.

- One of the first applications of the Gibbs' sampler (Tanner and Wong 1984)

  Unlike the related EM algorithm, yields full posterior distribution, not just an ML estimate.

- Draws $Y_{\text{mis}}^{(d,t)}$ of missing data can be used to create multiply-imputed data sets

Multiple imputation

- Imputes *draws*, not means, from the predictive distribution of the missing values
- Creates $D > 1$ filled-in data sets with different values imputed
- Bayesian MI combining rules yield valid inferences under well-specified models – propagate imputation uncertainty, and averaging of estimates over MI data sets avoids the efficiency loss from imputing draws
- MI can also be used for non-MAR models, particularly for *sensitivity analyses*

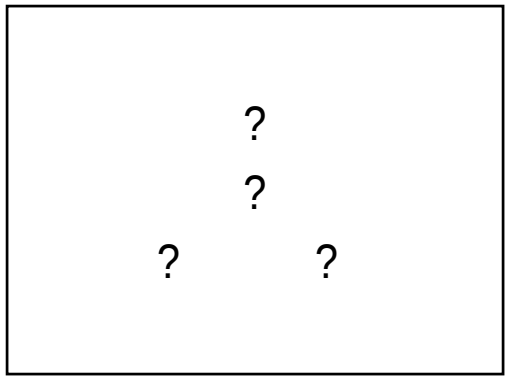
Idea of Multiple Imputation

- Data matrix with missing values

Variables

Y_1 Y_2 Y_3 Y_4 Y_5

Cases



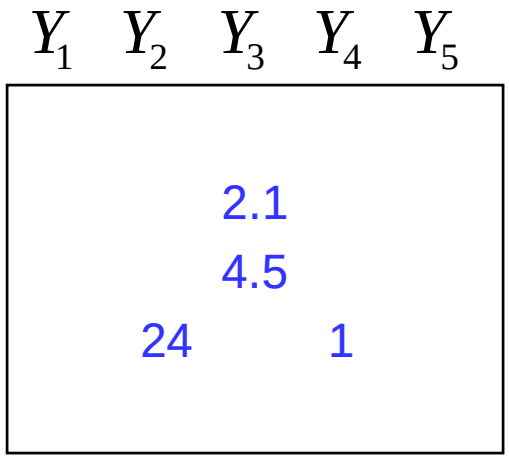
$\hat{\mu}_1 =$ mean based on all cases

$\hat{\beta}_{51 \cdot 1234} = ?$

Impute to recover information in incomplete cases

Single Imputation

- Impute missing values with predictions



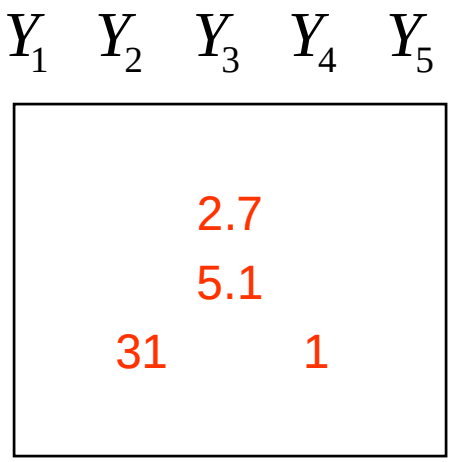
Dataset (l)	Estimate (se^2) μ_1	$\beta_{51\cdot1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)

Imputing best estimates biases slope - need to impute draws

SE of slope is too low – imputation error is not accounted for

MI: repeat with other draws

Second imputed dataset



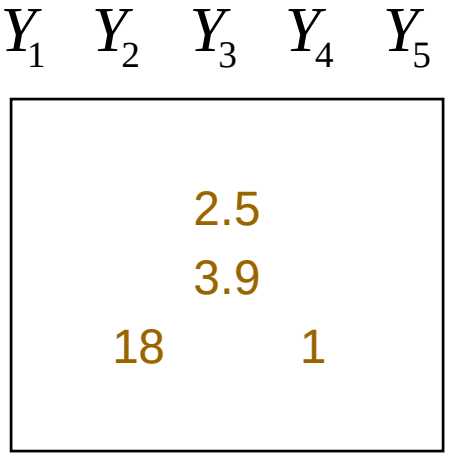
Dataset (l)	Estimate (se^2) μ_1	$\beta_{51\cdot1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)
2	12.6 (3.6^2)	4.15 (2.64^2)

Third imputed dataset

Y_1	Y_2	Y_3	Y_4	Y_5
		1.9		
		5.8		
32			2	

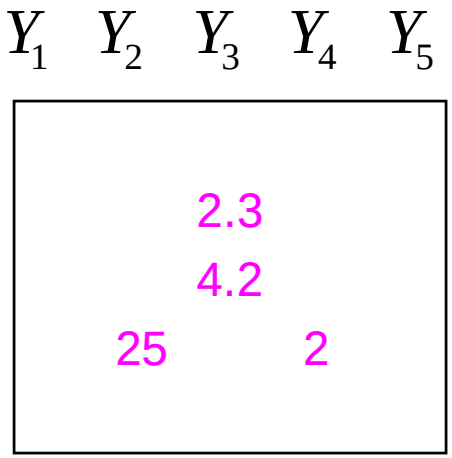
Dataset (l)	Estimate (se^2) μ_1	$\beta_{51\cdot1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)
2	12.6 (3.6^2)	4.15 (2.64^2)
3	12.6 (3.6^2)	4.86 (2.09^2)

Fourth imputed dataset



Dataset (l)	Estimate (se^2) μ_1	$\beta_{51\cdot1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)
2	12.6 (3.6^2)	4.15 (2.64^2)
3	12.6 (3.6^2)	4.86 (2.09^2)
4	12.6 (3.6^2)	3.98 (2.14^2)

Fifth imputed dataset



	Estimate (se^2)	
Dataset (l)	μ_1	$\beta_{51\cdot1234}$
1	12.6 (3.6^2)	4.32 (1.95^2)
2	12.6 (3.6^2)	4.15 (2.64^2)
3	12.6 (3.6^2)	4.86 (2.09^2)
4	12.6 (3.6^2)	3.98 (2.14^2)
5	12.6 (3.6^2)	4.50 (2.47^2)
Mean	12.6 (3.6^2)	4.36 (2.27^2)
Var	0	0.339

MI combining rules

Simulation approximations of posterior mean, variance yield the ML combining rules:

$$E(\theta \mid Y_{\text{obs}}) \approx \frac{1}{D}\sum_{d=1}^{D} E(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}^{(d)}) = \frac{1}{D}\sum_{d=1}^{D} \hat{\theta}_d$$

where $\hat{\theta}_d$ = is posterior mean from dth dataset

$$Var(\theta \mid Y_{\text{obs}}) \approx \frac{1}{D}\sum_{d=1}^{D} W_d + (1+1/D)\frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\theta}_d - \bar{\theta}_D\right)^2$$

where $W_d = \text{Var}(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}^{(d)})$ is posterior variance from dth dataset

MI Inferences (M=5)

	$\bar{\theta}$	$\bar{W}$	B	$\sqrt{V} = \sqrt{\bar{W} + (1 + 1/D)B}$	$R = \frac{(1+1/D)B}{V}$
μ_1	12.6	3.6^2	0	3.6	0
$\beta_{51 \cdot 1234}$	4.36	2.27^2	0.339	2.36	0.073

$\bar{\theta}$ = MI estimate

$\sqrt{V}$ = MI standard error

R = estimated fraction of missing information

Advantages of MI

- Imputation model can differ from analysis model
  - By including variables not included in final analysis
  - Promotes consistency of treatment of missing data across multiple analyses
  - MI combining rules can also be applied when the complete-data inference is not Bayesian (e.g. design-based survey inference).
  - Assumptions in imputation model are then confined to the imputations – with little missing data, simple methods suffice
- Public use data set users can be provided MI's, spared task of building imputation model
  - MI analysis of imputed data is easy, using complete-data methods (SAS PROC MIANALYZE)

MI for parametric models

- Principled, MCMC methods for creating draws have predictable properties
- Parametric assumptions can be improved by usual data-analytic strategies, e.g. transformations
- Analysis of MI data sets can be based on less parametric methods if desired
- For monotone pattern, flexibility is achieved by *factoring* the joint distribution
- However, for general patterns, the requirement for a coherent joint distribution limits flexibility
  - E.g. multivariate normality assumes regressions are linear and additive

Sequential regression MI (SRMI)

- Sequential regression MI (IVEware, MICE) regresses each variable with missing values in succession on all the other variables, with missing values of regressors filled in from earlier steps
- Iterates until imputations appear “stable”
- For parametric model, sequential imputation is essentially a form of Gibbs’ sampler
- Flexibility allowed in regressions – e.g. logit links for binary variables, nonlinear terms
- Conditionals may be incoherent – do not correspond to well-specified joint d/n – but gain in flexibility outweighs this theoretical drawback

Example. Logistic regression simulation study

- True model:

$$X \sim N(0,1)$$

$$\text{Logit}[Pr(E=1|X)]=0.5+X$$

$$\text{logit}[Pr(D=1|E,X)]=0.25+0.5X+1.1E$$

- Sample size: 500
- Number of Replicates: 5000
- Before Deletion Data Sets

Missing-Data Mechanism

- D and E : completely observed
- X : sometimes missing
- Missing Data Probabilities:

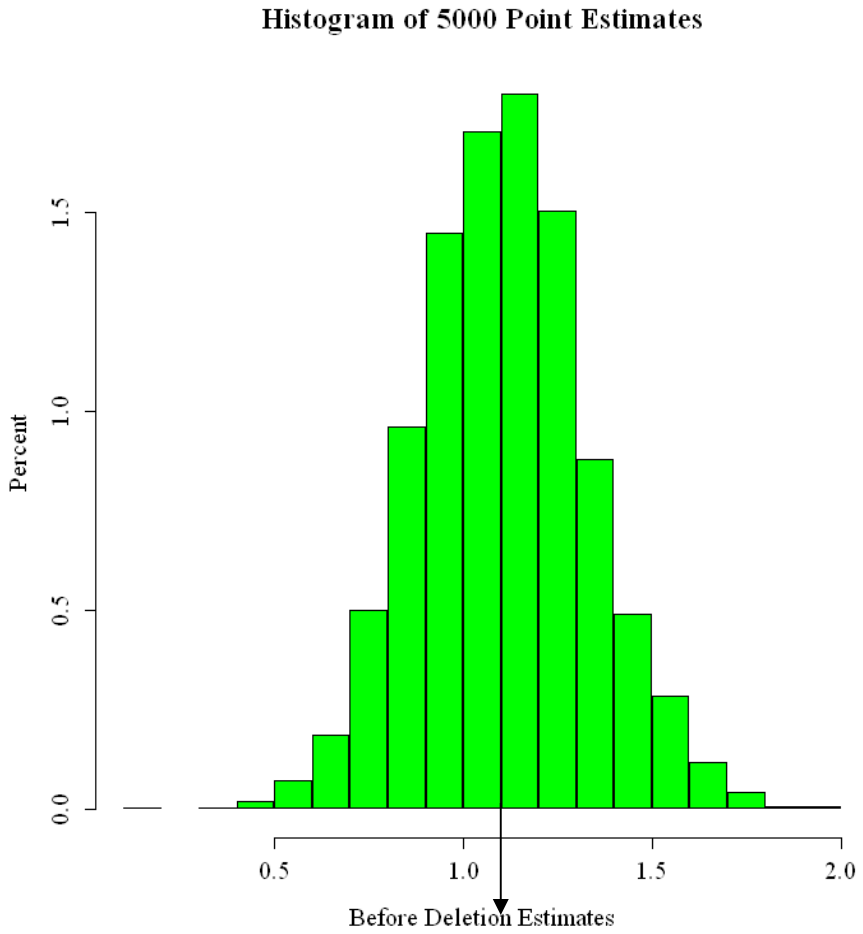
  D=0,E=0: $p_{00}=0.19$

  D=0,E=1: $p_{01}=0.09$

  D=1,E=0: $p_{10}=0.015$

  D=1,E=1: $p_{11}=0.055$

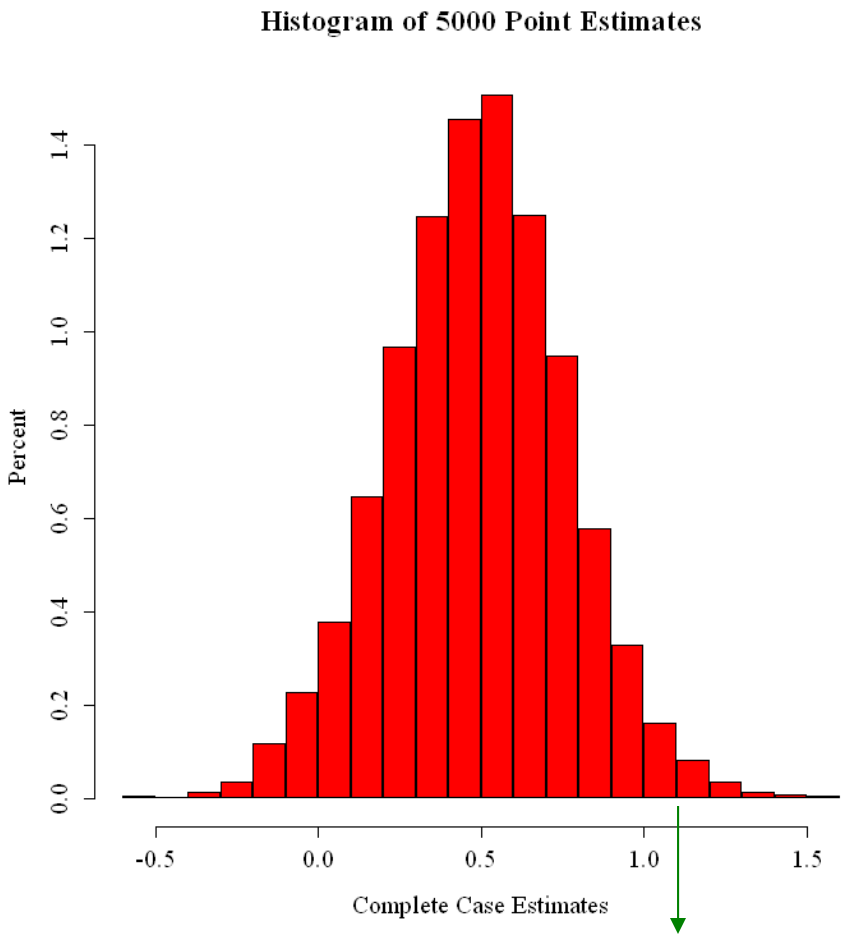
Before Deletion Estimates



- Histogram of 5000 estimates before deleting values of X
- logistic model

$$logit\ Pr(D=1|E,X) = \beta_0 + \beta_1 E + \beta_2 X$$

Complete-Case Estimates



Histogram of complete- case analysis estimates

Delete subjects with missing X values

True value = 1.1, serious negative bias

MI for logistic regression example

- The model for the data implies that for missing values of X:

$$(X_i \mid D_i = d, E_i = e, \mu_{ed}, \sigma^2) \sim N(\mu_{ed}, \sigma^2)$$

- Improper MI: substitute estimates of $\{\mu_{ed}\}, \sigma^2$
- Proper MI: Imputations are draws from the posterior predictive distribution
- Draw σ^2, then μ_{ed} and then missing X_i

Predictive Distributions

$\sigma^{2(\ell)} \sim WSS / \chi^2_{r-4},$

WSS = residual sum of squares,

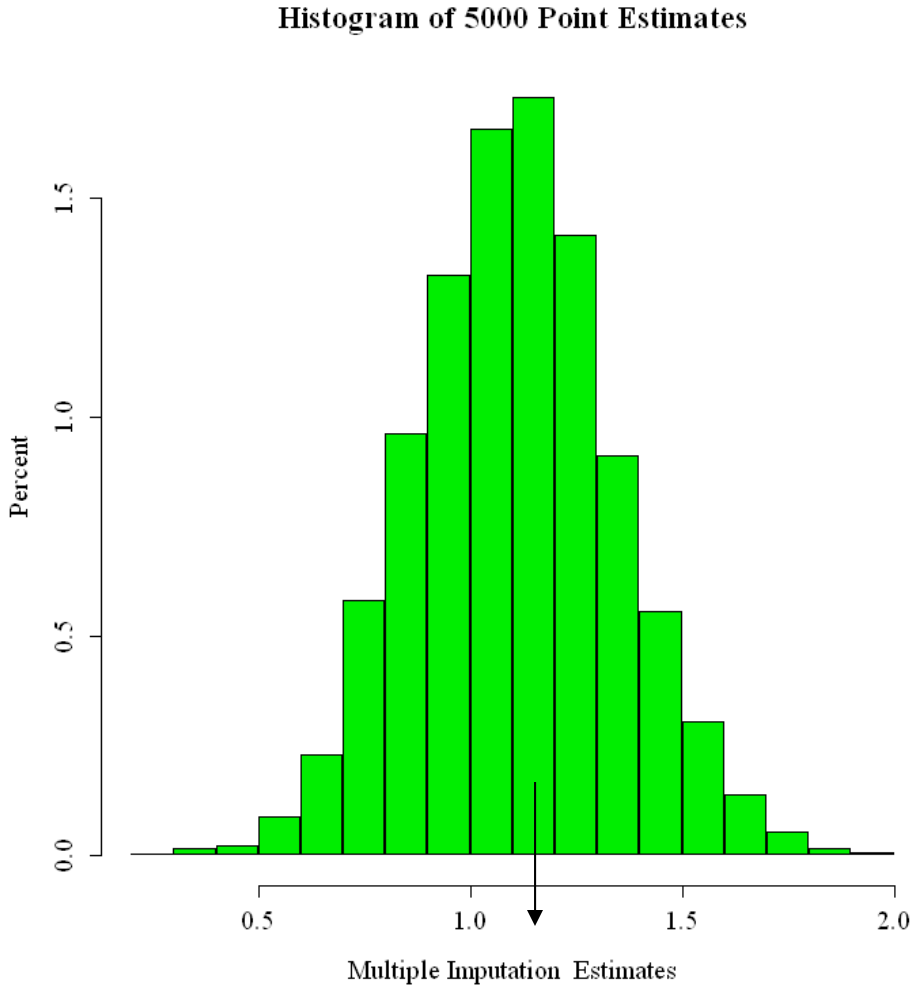
$r =$ number of complete cases

$\mu_{ed}^{(\ell)} \sim N(\bar{x}_{ed}, \sigma^2 / r_{ed})$

$\bar{x}_{ed}, r_{ed} =$ mean, complete cases in cell (e,d)

$X_{edi}^{(\ell)} \sim N(\mu_{ed}^{(\ell)}, \sigma^{2(\ell)})$

Histogram of Multiple Imputation Estimates



- 5 Imputations per missing value
- 5 completed Datasets
- Analyze each separately
- Combine using the formulae given earlier

Coverage and MSE of Various Methods

METHOD	COVERAGE (95% Nominal)	MSE
Complete-case	37.86	0.4456
Hot-Deck Single Imputation	90.28	0.0566
Multiple Imputation	94.56	0.0547
Before Deletion	***94.68***	***0.0494***

Summary

- Bayesian approach to missing data meshes seamlessly with Bayesian approach to survey inference
  - Predict missing values as well as non-sampled values
- MAR key condition: MAR methods much easier if they can be justified
- Multiple imputation provides flexibility, allows design-based complete-data methods to be applied

Books on Bayesian Methods

Box, G.E. P; Tiao, G. C.; Bayesian Inference in Statistical Analysis. Addison-Wesley, Massachusetts, 1973.

Gelman, A., Carlin, J. B, Stern, H. S., Dunson, D.B., Vehtari, A. and Rubin, D. B. (2013). Bayesian Data Analysis. 3rd edition. Chapman & Hall/CRC, London, 2013.

Rubin, D. B.; Multiple Imputation for Nonresponse in Surveys. Wiley, New York, 1987. (Chapter 2)

A partial list of articles on Bayesian inference in Surveys

Andridge, R.H. & Little, R.J. (2011). Proxy Pattern-Mixture Analysis for Survey Nonresponse. *Journal of Official Statistics*, 27, 2, 153-180.

Bolfarine, H. & Sandoval, M. C. (1993). Prediction of the finite population distribution function under Gaussian superpopulation models. *Australian Journal of Statistics*. 35 (1993), no. 2, 195--204.

Bolfarine, H. & Zacks, S. (1992). *Prediction theory for finite populations*. Springer-Verlag, New York, 1992. xii+207 pp.

Cocchi, D. & Mouchart, M. (1990). Linear Bayes estimation in finite populations with a categorical auxiliary variable. *Statistics, a Journal of Theoretical and Applied Statistics* 21 (1990), no. 3, 437--454.

Datta, G. S. & Ghosh, M. (1992). The Horvitz-Thompson estimate and Basu's circus revisited. *Bayesian analysis in statistics and econometrics* (Bangalore, 1988), 225--228, Springer, New York, 1992.

Datta, G. S. & Ghosh, M. (1993). Bayesian estimation of finite population variances with auxiliary information. *Sankhya Ser. B* 55 (1993), no. 2, 156--170.

Elliott, M. R. & Little, R.J.A. (2000). Model-Based Alternatives to Trimming Survey Weights. *Journal of Official Statistics*, 16, No. 3, 191-209.

Ericson, W. A (1988). Bayesian inference in finite populations. *Handbook of Statistics*, 6, 213--246, North-Holland, Amsterdam, 1988.

Fienberg, S.E. (2011). Bayesian Models and Methods in Public Policy and Government Settings. *Statistical Science*, 26, 2, 212–226.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22, 151-174 (with discussion).

Ghosh, M. & Meeden (1997). *Bayesian methods for finite population sampling*. Chapman & Hall, London, 1997. x+289 pp.

Godambe, V. P. & Thompson, M. E. (1971). Bayes, fiducial and frequency aspects of statistical inference in regression analysis in survey-sampling. With discussion by G. A. Barnard, M. Stone, A. W. F. Edwards, D. J. Bartholomew, G. N. Wilkinson, D. A. Sprott, E. F.

Harding and V. M. Joshi. *Journal of the Royal Statistical Society, Ser. B* 33 (1971), 361--390.

Koop, J. C. (1979). On statistical inference in sample surveys and the underlying role of randomization. *Annals of Institute of Statistical Mathematics* 31, 2, 253--269.

Krishnaiah, P. R. & Rao, C. R. (1988). Handbook of Statistics, 6. Edited by P. R. Krishnaiah and C. R. Rao. North-Holland Publishing Co., Amsterdam-New York, xvi+594 pp.

Little, R.J.A. (2003). The Bayesian Approach to Sample Survey Inference. In *Analysis of Survey Data*, R.L. Chambers & C.J. Skinner, eds., pp. 49-57. Wiley: New York.

Little, R.J.A. (2003). Bayesian Methods for Unit and Item Nonresponse. In *Analysis of Survey Data*, R.L. Chambers & C.J. Skinner, eds., pp. 289-306. Wiley: New York.

Little, R.J.A. (2004). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association,* 99, 546-556.

Little, R.J.A. (2006). Calibrated Bayes: A Bayes/Frequentist Roadmap. *The American Statistician*, 60, 3, 213-223.

Little, R.J. (2012). Calibrated Bayes: an Alternative Inferential Paradigm for Official Statistics (with discussion and rejoinder), *Journal of Official Statistics*, 28, 3, 309-372.

Little, R.J.A. & Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, 31, 161-168.

Lo, A. Y. (1986). Bayesian statistical inference for sampling a finite population. *Annals of Statistics* 14, 3, 1226--1233.

Lo, A. Y. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics* 16 , 4, 1684--1695.

Malec, D. & Sedransk, J. (1985). Bayesian inference for finite population parameters in multistage cluster sampling. *Journal of American Statistical Association* 80 (1985), 392, 897--902.

Meeden, G. & Vardeman, S. (1991). A noninformative Bayesian approach to interval estimation in finite population sampling. *Journal of American Statistical Association* 86, 972--980.

Nandram, B. (1994). Bayesian predictive inference for multivariate sample surveys. *Journal of Official Statistics* 10 1994 167-179.

Nandram, B.& Sedransk, J. (1993). Bayesian predictive inference for a finite population proportion: two-stage cluster sampling. *Journal of the Royal Statistical Society, Ser. B* 55 (1993), no. 2, 399--408.

Pfeffermann, D. &Nathan, G. (1977). Regression analysis of data from complex samples. In Bayes, fiducial and frequency aspects of statistical inference in regression analysis in survey-sampling... *Bulletin of Institute of International Statistics* 47, no. 3, 21--42, 58--64.

Raghunathan, T.E. & Grizzle, J.E. (1995). A split questionnaire survey design, *Journal of the American Statistical Association,* 90:55–63.

Raghunathan, T.E. & Rubin, D.B. (1997). Roles for Bayesian techniques in survey sampling, *Proceedings of Statistical Society of Canada*, 51–55.

Raghunathan, T.E. (2000). Bayesian analysis of quality level using simulation methods, *Journal of Quality Technology*, 32:172–82.

Raghunathan, T.E, Lepkowski, J.M., VanHoewyk & J,.Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models, *Survey Methodology*, 27:85–95.

Raghunathan, T.E., Xie, D., Schenker, N, Parsons, V., Davis, W., Rancourt, E., Dodd, K. (2007). Combining Information from Multiple Surveys for Small Area Estimation: A Bayesian Approach. *Journal of American Statistical Association,* 102, *474-486*.

Rodrigues, J., Bolfarine, H. & Rogatko, A. (1985). A general theory of prediction in finite populations. *International Statistical Review* 53, 3, 239--254.

Royall, R.M. & Pfeffermann, D. (1982). Balanced samples and robust Bayesian inference in finite population sampling. *Biometrika* 69, no. 2, 401--409.

Sedransk, J. (2008). Assessing the Value of Bayesian Methods for Inference about Finite Population Quantities. Journal of Official Statistics, 24**,** 495–506.

Smith, T. M. F. (1987).To weight or not to weight, that is the question. *Bayesian statistics, 3* (Valencia, 1987), 437--451, Oxford Univ. Press, New York, 1988.

Smouse, E. P. (1982). Bayesian estimation of a finite population total using auxiliary information in the presence of nonresponse. *Journal of American Statistical Association* 77 (1982), no. 377, 97--102.

Zanganeh, S.Z. & Little, R.J. (2015). Bayesian inference for the finite population total in heteroscedastic probability proportional to size samples. *Journal of Survey Statistics and Methodology*, 3, 162-192.

Zheng, H. & Little, R.J. (2005). Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model. *Journal of Official Statistics*, 21, 1-20.