

# Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 1: Course preliminaries,  
introduction to Bayesian inference



# Instructors

- Rod Little, Professor of Biostatistics and Research Professor, Michigan Program in Survey Methodology, Institute for Social Research
- T. E. Raghunathan, Professor of Biostatistics and Director, Survey Research Institute and Research Professor, Michigan Program in Survey Methodology, Institute for Social Research.

# Locations

- Class Meeting time: Mondays and Wednesdays 10:30am-12pm
- First meeting: Wednesday Jan 4, 2017
- Location: Room 368 (Basement), ISR. This course is offered jointly with the Joint Program in Survey Methodology (JPSM) through an interactive video transmission system.
- ISR 4036 10:30-12:00 on following dates:
- Jan 9<sup>th</sup>, Feb 13<sup>th</sup>, March 13<sup>th</sup>, April 10<sup>th</sup>

# Course Objectives

- Bayesian methods in statistics are increasingly popular, spurred by advances in computational power and tools.
- Bayesian inference provides solutions to problems that cannot be solved exactly by standard frequentist methods.
- Bayesian approach provides new analysis tools, and a deeper understanding of competing systems of statistical inference, including the frequentist approach.
- Describe the application of the Bayesian approach to survey sampling, where the focus of inference is on finite population quantities. This course will emphasize both theoretical and applied aspects of Bayesian inference, in general, and to sample surveys, in particular.

# Prerequisites

- This is an advanced course in statistics.
  - Students should be familiar with standard statistical methods based on likelihood and other statistical concepts in probabilities and distributions, sufficient statistics, point and interval estimation, testing of hypothesis and basic asymptotic theory. Students should also be well versed in calculus.
- The course will be a mix of theory and computations.
  - Computations will be performed using
  - WinBugs -- free software for Bayesian Analysis that can be downloaded from the website ([www.mrc-bsu.cam.ac.uk/bugs](http://www.mrc-bsu.cam.ac.uk/bugs)),
  - R (free software similar to S-plus) ([www.r-project.org](http://www.r-project.org))
  - IVEware, imputation and analysis software ([www.iveware.org](http://www.iveware.org)).
  - During the first week of classes, download these three software packages and install them on your computers.

# Grading

- Letter grades for the course will be based on
  - homework assignments, given periodically, a mix of theoretical and applied problems. Grade will be based on completing and handing in assignments (20%)
  - a midterm take- home examination (40%)
  - a final project. Either a theoretical project demonstrating an advantage (or disadvantage) of the Bayesian approach to a particular problem or an applied survey data analysis using the Bayesian approach. The final project will be graded based on a written report not exceeding 15 pages including tables, figures and references. (40%)

# Textbook

- Though there are no required textbooks, you might want to buy the third edition of the book “Bayesian Data Analysis” by Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin · CRC Press
- Course notes will be provided from time to time

# Course Topics

1. Modes of inference in surveys; Fundamentals of Bayesian Inference; Application to surveys using simple random sample (srs) design
2. Basic Monte-Carlo methods (Non-iterative)
3. Role of complex sample designs in Bayesian inference; unequal probabilities of selection and stratification
4. Auxiliary variables and nonparametric methods
5. Hierarchical models for cluster sample designs
6. Bayesian models for complex sample survey designs
7. Advanced computational methods; Marko Chain Monte-Carlo Methods
8. Bayesian inference with unit and item nonresponse

# Lecture Schedule (subject to change):

1. Jan 4	Introduction 1	
2. Jan 9	Introduction 2	ISR 4036
3. Jan 11	Maximum Likelihood for SRS	
Jan 16	MLK Day no class	
4. Jan 18	Bayes for Simple Random Samples	
5. Jan 23	Bayesian Computation 1: Principles	
6. Jan 25	Bayesian Computation 2: Software	
7. Jan 30	Complex Survey Designs	
8. Feb 1	Models for Complex Survey Designs	
9. Feb 6	Bayes for Stratified Samples 1	
10. Feb 8	Bayes for Stratified Samples 2	
11. Feb 13	Bayes for PPS Samples	ISR 4036
12. Feb 15	Role of Weights, Classical and Bayes	
13. Feb 20	Bayes for Clustered Data 1	
14. Feb 22	Bayes for Clustered Data 2	
Feb 27, Mar 1	Winter Break, no class	

# Lecture Schedule (continued)

15. Mar 6	Midterm review	
16. Mar 8	Midterm exam	
17. Mar 13	Multistage Sampling	ISR 4036
18. Mar 15	Missing Data 1	
19. Mar 20	Missing Data 2	
20. Mar 22	Missing Data 3	
21. Mar 27	Measurement error as missing data	
22. Mar 29	Applications 1	
23. Apr 3	Applications 2	
24. Apr 5	Applications 3	
25. Apr 10	Summary of Course material	ISR 4036
26. Apr 12	Project presentations 1	
27. Apr 17	Project presentations 2	
28. Apr 21	Extra Class for Project presentations 3 (if needed)	

# Bayesian inference for sample surveys

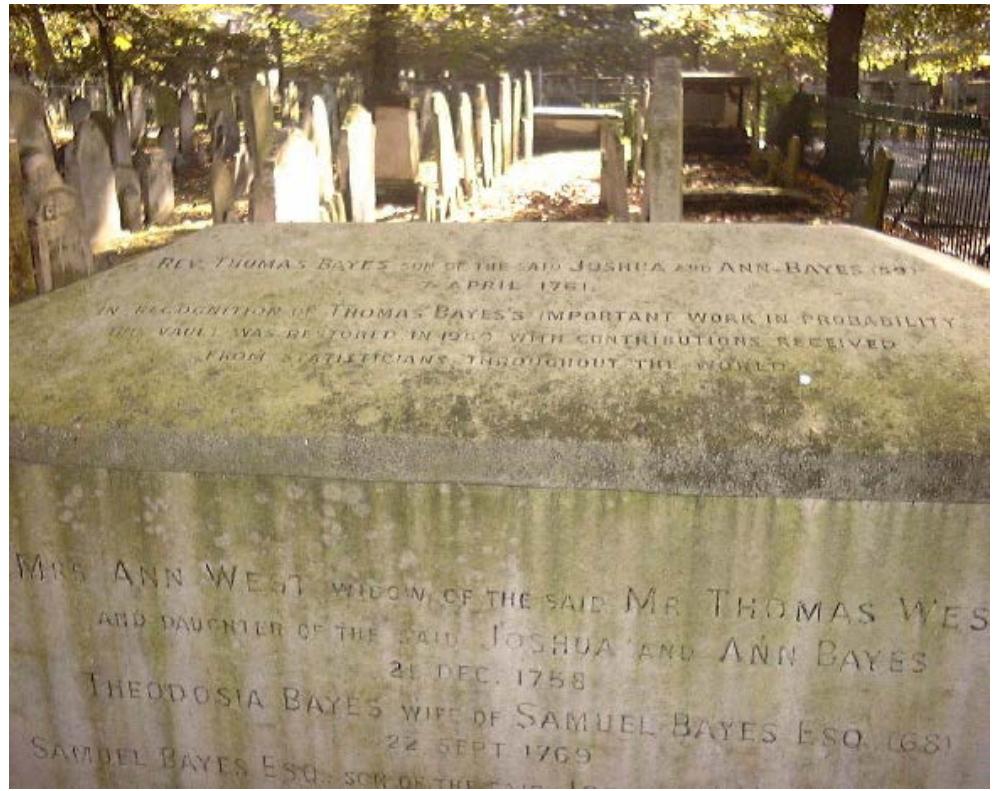
# Approaches to Statistical Inference

- Classical (Randomization, Frequentist) Inference
  - parameters are treated as *fixed*
  - Inferences -- *P-Values, confidence intervals* -- are based on distribution of statistics in repeated sampling
- Bayesian Inference
  - Key elements contained in a posthumous essay by Rev. Thomas Bayes (1702?-1761)
  - Gaining popularity, particularly for complex statistical modeling
  - parameters are treated as *random*, and assigned a prior distribution to represent prior knowledge
  - Bayesian inference based on *posterior distribution* of parameters given data, computed via *Bayes Rule*

# Rev Thomas Bayes (1702?-1761)



REV. T. BAYES



# Publications in his lifetime

1. Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures ([1731](#)) (Theological Work).
2. An Introduction to the Doctrine of Fluxions, and a Defense of the Mathematicians Against the Objections of the Author of the Analyst (published anonymously in [1736](#)) , Fellow of Royal Society based on this work in 1742 (speculative).

# Bayes' famous posthumous essay

“I now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and deserves to be preserved”

Richard Price, read to the Royal Society Dec 23, 1763

# Bayes' rule

- Bayesian statistics is founded on Bayes' rule, which is a simple consequence of basic rules of probability:

If  $A$  and  $B$  are two events, then the product rule is:

$$\Pr(A, B) = \Pr(A) \times \Pr(B | A)$$

Special case:  $A$  and  $B$  are independent if

$$\Pr(A, B) = \Pr(A) \times \Pr(B)$$

- This rule can be applied to make inferences about (a) hypotheses, (b) parameters, or (c) predictions of nonsampled values

# Bayes' rule for hypotheses

- Applying the product rule with  $A = \text{hypothesis } H, B = \text{data } D$ :

$$\Pr(H, D) = \Pr(D) \times \Pr(H | D) = \Pr(H) \times \Pr(D | H)$$

$$\Pr(H | D) = \Pr(H) \times \Pr(D | H) / \Pr(D)$$

Hence,

$$\frac{\Pr(H | D)}{\Pr(H' | D)} = \frac{\Pr(H)}{\Pr(H')} \times \frac{\Pr(D | H)}{\Pr(D | H')}$$

That is, posterior odds = prior odds  $\times$  Bayes factor

# $\Pr(D|H)$ or $\Pr(H|D)$ ?

- Examples of  $H$ :
  - The existence of god, or life in other universes
  - Does environmental factor X cause disease Y?
  - Will Donald Trump win the election?
  - Will sales for product X meet some target value?
- Examples of  $D$ :
  - Astronomical measurements
  - Opinion polls
  - Market research data
  - National probability samples

# A simple application of Bayes: Screening Tests

- A friend is diagnosed by a screening test ( $D = \text{result of test, + or -}$ ) to have an extremely rare form of cancer ( $H = \text{has cancer}$ ). Only one out of a million people in his age group have the cancer.
- Naturally he is very upset as the test is pretty accurate:  
Sensitivity:  $\Pr(+|\text{ has cancer})=0.99$ , implying  
 $\Pr(-|\text{ has cancer})=0.01$  (False negative)  
Specificity:  $\Pr(-|\text{no cancer})=0.999$ , implying  
 $\Pr(+|\text{ no cancer})=0.001$  (False positive)

# False Positive

- The probability that matters is the positive predictive value, which by Bayes Rule is

$$\begin{aligned}\Pr(\text{has cancer} | +) &= \frac{\Pr(+ | \text{has cancer})\Pr(\text{has cancer})}{\Pr(+)} \\ &= \frac{(0.99)(1/1000000)}{(0.99)(1/1000000) + (0.001)(999999/1000000)} \\ &= 0.001 \quad (!)\end{aligned}$$

# False Positive

Very likely, the friend does not have cancer.

# Screening Tests: output

Sensitivity:  $\Pr(+|\text{ has cancer})=0.99$ ,

Positive predictive value:  $\Pr(\text{has cancer}|+)=0.001$

Specificity:  $\Pr(-|\text{no cancer})=0.999$ ,

Negative predictive value:  $\Pr(\text{no cancer}|-)=0.999999$

Sensitivity and specificity are properties of the test

PPV and NPV are what we really care about – though they require prior information about probability of disease (which may be hard to pin down)

# History of Bayes

- Much maligned in the last century, Bayesian statistics has since experienced a dramatic revival
- Excellent and fun book: “The theory that would not die” by Sharon McGrayne

# Statisticians Impacting Science: Bayesians in red

## Most-cited mathematicians in science (**Science Watch 02**)

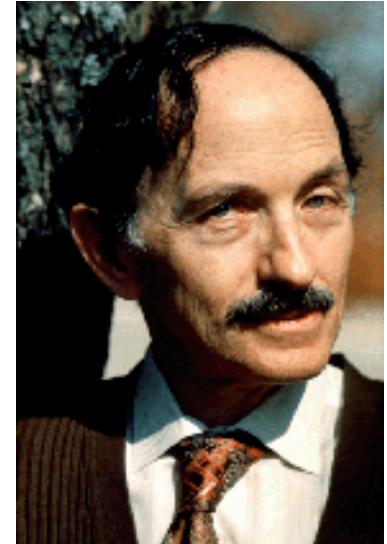
- 2 D. L. Donoho Stanford Stat;
- 3 A.F.M. Smith London Stat
- 4 E. A. Thompson Washington Biostat;
- 5 I.M.Johnstone Stanford Stat
- 6 J. Fan Hong Kong Stat;
- 7 D.B. Rubin Harvard Stat.
- 9 A. E. Raftery Washington Stat;
- 10 A.E. Gelfand U. Conn Stat.
- 11 S-W Guo Med. Coll. Wisc Biostat;
- 12 S.L. Zeger JHU Biostat.
- 13 P.J. Green Bristol Stat; 14 B.P. Carlin Minnesota Biostat
- 15 J. S. Marron UNC Stat; 16 D.G. Clayton Cambridge Biostat
- 16 G.O. Roberts Lancaster Stat; 20. X-L Meng Chicago Stat
- 21. M. P. Wand Harvard Biostat; 22.W.R. Gilks MRC Biostat
- 23 M. Chris Jones Open U Stat; 25.N. E. Breslow Washington Biostat

# Bayes Impacting Society

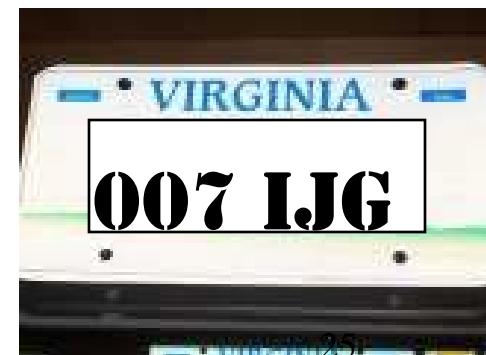


Alan Turing sculpture by Stephen Kettle, Bletchley Park. Photo by Jon Callas

Alan Turing and Jack Good's Bayesian statistical methods helped decode German naval ciphers, arguably reducing the length of World War II by two years or more, saving millions of lives.



I. Jack Good (IJG)



... and while on scientists who are subjects of current films...

“One can make the Anthropic Principle precise, by using **Bayes** statistics... One weights the a-priori probability (of a class of histories) ... with the probability that the class of histories contain intelligent life...  
Stephen Hawking



# Bayes for comparing hypotheses

$$\Pr(H_1 | D) = \text{Pr}(H_1) \times \Pr(D | H_1) / \Pr(D)$$

$$\Pr(H_2 | D) = \text{Pr}(H_2) \times \Pr(D | H_2) / \Pr(D)$$

Hence  $\frac{\Pr(H_1 | D)}{\Pr(H_2 | D)} = \frac{\text{Pr}(H_1)}{\text{Pr}(H_2)} \times \frac{\Pr(D | H_1)}{\Pr(D | H_2)}$

Posterior odds = prior odds x Bayes factor

“Prior odds are modified by the relative probability of observing the data under the two hypotheses”

Fun fact: according to Jack Good, Alan Turing coined the term “Bayes factor” (without the Bayes)

# Bayesian inference for population parameters

- Bayes rule also provides inferences for population quantities, like population means

$\theta$  = population parameters

$$p(\theta | data) = \frac{\pi(\theta) p(data | \theta)}{p(data)}$$

↑  
Posterior distribution of  $\theta$

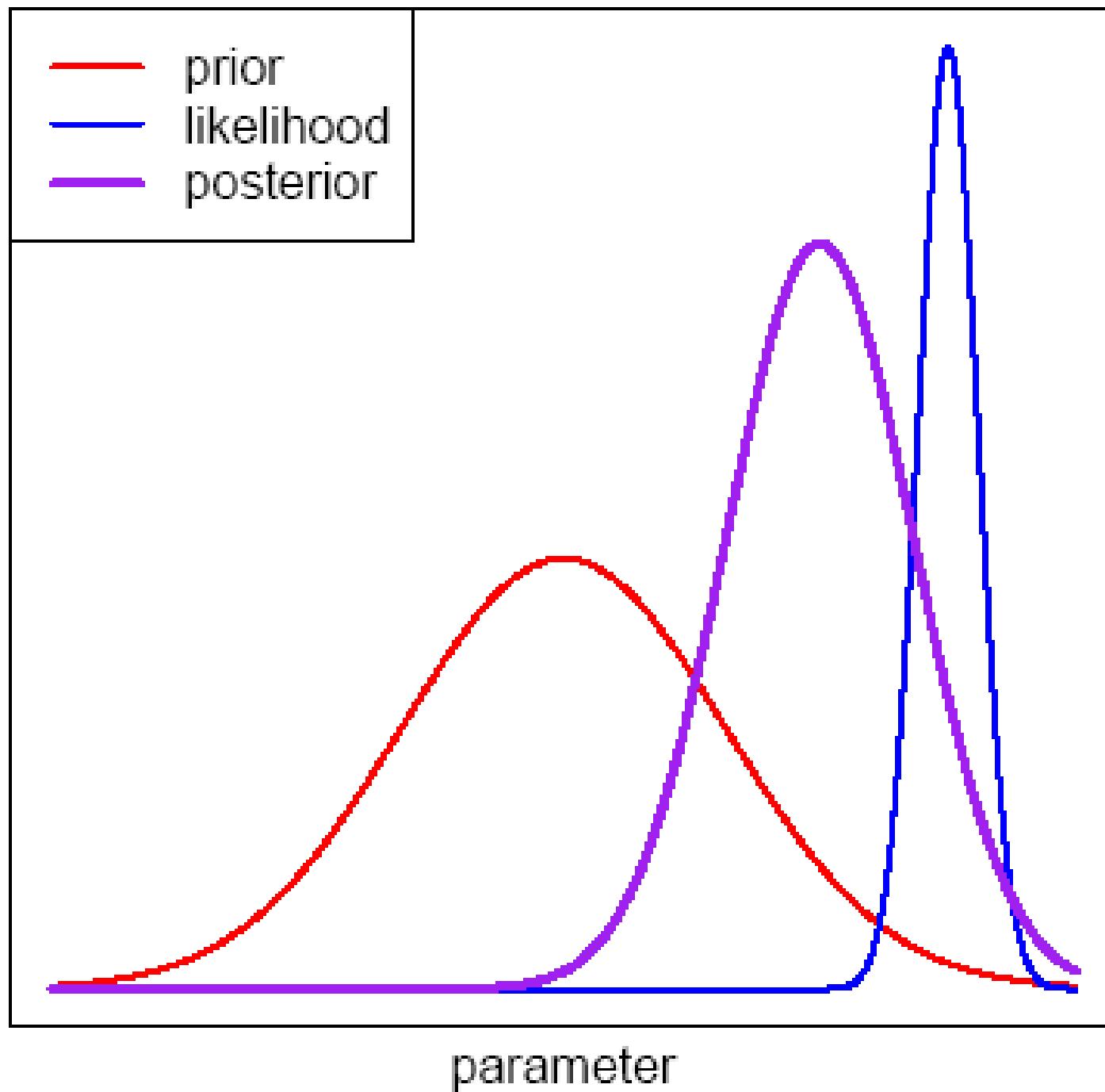
↑  
Prior distribution of  $\theta$

↓  
Likelihood  $L(\theta | data)$

Normalizing constant

The diagram illustrates the components of Bayes' theorem. At the top, the formula  $p(\theta | data) = \frac{\pi(\theta) p(data | \theta)}{p(data)}$  is shown. Below it, a horizontal line represents the posterior distribution  $p(\theta | data)$ . Four arrows point to different parts of the formula:

- An arrow from the left points to the term  $\pi(\theta)$ , labeled "Posterior distribution of  $\theta$ ".
- An arrow from the left points to the term  $p(data | \theta)$ , labeled "Prior distribution of  $\theta$ ".
- An arrow from the right points to the term  $p(data)$ , labeled "Likelihood  $L(\theta | data)$ ".
- An arrow from the right points to the denominator  $p(data)$ , labeled "Normalizing constant".



# Bayesian inference for predictions

- Bayes rule also provides inferences for predictions of future values (or finite population quantities, which are functions of sampled and nonsampled values)

$y^*$  = future value,  $\theta$  = population parameters

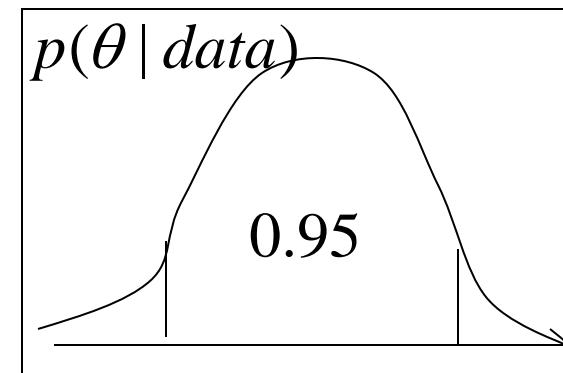
$$p(y^* | \text{data}) = \int p(y^* | \theta) p(\theta | \text{data}) d\theta$$

*Posterior predictive  
distribution*

- In a sense, all of statistics is about prediction, and quantifying the associated uncertainty
- This use of Bayes is a strong emphasis in this course

# Forms of Bayesian inference

- The posterior distribution summarizes information about an unknown quantity given the data
- Parameters can be *estimated* by the mean or median of the posterior distribution
- The spread of the posterior distribution indicates uncertainty, e.g. the posterior standard deviation
- A 95% posterior probability or credibility interval replaces the 95% confidence interval in frequentist statistics



# Prior distribution

- The novel feature of Bayesian statistics is the prior distribution, which formalizes prior knowledge about the parameters before the data are collected.
- Sometimes prior distributions are developed using information from previous studies (eg meta-analysis)
- Another approach is to use “noninformative” priors that correspond to limited prior information
  - There are standard “noninformative” or “reference” priors for common problems
- Subjective Bayesian approach “elicits” priors from experts

# Properties of Bayesian statistics

1. Bayes with noninformative reference priors and frequentist solutions agree on many standard problems.

Example: independent normal sample

$$y_1, \dots, y_n | \theta \sim_{iid} N(\theta, \sigma^2)$$

Frequentist inference: The standard 95% confidence interval for  $\theta$  is :

$$I = \bar{y} \pm t_{.975, n-1} s / \sqrt{n}, \text{ where}$$

$\bar{y}$  = sample mean,  $s$  = sample sd

We'll see that  $I$  is also the 95% posterior credibility interval for  $\theta$ , with a particular choice of reference prior distribution

# Agreement of Bayes and classical inferences

- More generally, in large samples, Bayesian inference yields similar answers to maximum likelihood (ML) inference, a very important method in classical inference
- The underlying principle behind many standard analysis methods in epidemiology is ML – linear and logistic regression, survival analysis, etc.
- I'll review the basic ideas of maximum likelihood and how they relate to Bayes inference later

# Properties of Bayes

2. Bayesian inference is more direct, less mysterious than frequentist inference

- confidence intervals have a tricky interpretation
- hypothesis testing is worse:

Frequentist: P-Value =  $\Pr(\text{data}|\mathcal{H})$

P-value is not the probability that  $\mathcal{H}$  is true!

Bayes:  $\Pr(\mathcal{H}|\text{data})$

# Properties of Bayes

3. There are many problems that have no exact frequentist answers in small samples. For example
  - (a) Comparing means of two independent samples with different means and variances. There is no exact frequentist solution, only approximate solutions (e.g. Welch approximation)
  - (b) Inference for nonlinear models (e.g. logistic regression, Cox model) assume large samples. Bayesian methods yield answers to such problems

# Properties of Bayes

4. Bayesian inference allows for prior information to be incorporated in the analysis in a simple and clear way, via informative prior distributions.
  - In large-sample survey applications, we often use weak, relatively noninformative priors – most of the information is contained in the likelihood – this is sometimes called “objective Bayes”
  - However, informative priors can play a useful role in situations where some parameters of a model are not identified, or weakly identified
  - One example of this is for nonresponse that is missing not at random (as discussed in the material on nonresponse later)

# Properties of Bayes

- 5. Bayesian inferences under a carefully-chosen model should have good frequentist properties.
- 6. Theory says you are best to act (e.g. bet) like a Bayesian.
- 7. Modern computation tools make Bayesian analysis much more feasible than in the past.

Markov Chain Monte Carlo (MCMC) methods for computing posterior distributions

# Frequentist and Bayes methods have pluses and minuses...

## Frequentist

No need for prior – limited specification

Good repeated sampling properties

Not prescriptive, can be ambiguous

Not enough exact answers

Fails the likelihood principle

## Bayes

Needs detailed specification of prior and likelihood

Bad model may have poor repeated sampling properties

Prescriptive, unambiguous

Plenty of answers

Satisfies the likelihood principle

# The compromise: calibrated Bayes

Activity	Bayes	Frequentist
Inference under assumed model	Strong	
Model formulation / assessment		Strong

Bayesian for inference

Frequentist for model assessment (enriched by Bayesian ideas)

Attempt to capitalize on strengths of both paradigms (Little, 2006 American Statistician)

# Bayes/frequentist compromises

“I believe that ... sampling theory is needed for exploration and ultimate *criticism* of the entertained model in the light of the current data, while Bayes' theory is needed for *estimation* of parameters conditional on adequacy of the model.”

George Box (1980)



# Calibrated Bayes

“... frequency calculations are useful for making Bayesian statements scientific, 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.”  
Don Rubin (1984 Annals of Statistics)



# Summary

- Bayes is flexible and principled
- Bayesian software is increasingly available
- Needs to specify prior distributions for parameters – approaches to be discussed
- For survey applications, emphasis is on predicting or “filling in” the nonsampled and nonresponding values – inference is based on the posterior predictive distribution of finite population quantities

# Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 2: Complex survey designs



# Bayesian inference for sample surveys

# Survey sampling

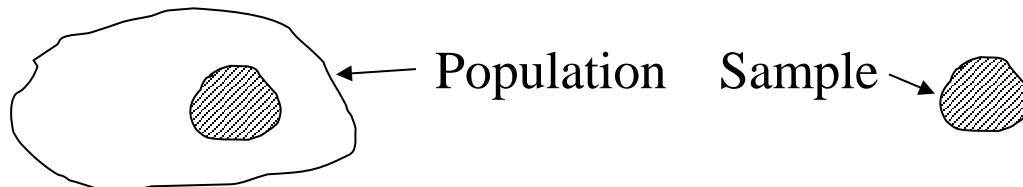
- So far we have discussed Bayes and frequentist inference for statistics in general
- We consider in this course the specific application of Bayes to survey sampling
- In this lecture we describe
  - Probability sample designs, in particular simple random sampling and more complex designs
  - Distinguishing features of survey sample inference
- In the next lecture we discuss in broad terms alternative modes of survey inference
  - Design-based, superpopulation models, Bayes

# Distinguishing Statistical Features of Survey Sampling

- Major interest in *descriptive* inference about *finite population quantities*, as opposed to parameters of models (though *analytical inference* for parameters can also be of interest).
- Probability sampling – method of sampling from the population that avoids selection biases.
- Prevailing orthodoxy is *design-based* (randomization) inference: survey outcomes are treated as fixed quantities, and statistical uncertainty derives from the probability distribution that determines sample selection

# Inference for a population based on a sample

- Parameters: population is thought of as drawn from an infinite “superpopulation”; Parameters are summary characteristics of this super-population, in superpopulation models (greek symbols)
- Population quantities: descriptive quantities of the population, such as means and totals (cap roman symbols)
- Statistical inference: the process of making inferences about model parameters and population quantities based on sample data.



	Parameter	Population	Sample
--	-----------	------------	--------

Mean	$\mu$	$\bar{X}$	$\bar{x}$
SD	$\sigma$	$S$	$s$

- Inference crucially requires that sample is randomly selected from population (or an assumption that it is)

# Properties of a good sampling scheme

- "representative" of the population (... whatever that means)
- demonstrably free of selection bias
- repeatable (at least in principle)
- efficient: low cost for given level of precision
- measurable precision: e.g., can quantify how close the sample estimate is to the population quantity it is estimating.
- Only probability (or random) sampling designs have these properties. Probability samples are characterized by the following two properties:
- every sample has a known (maybe zero) probability of selection
- every unit in the population has a (known) positive probability of selection.

# Simple Random Sampling

- The most familiar form of probability sampling
- Simple random sampling without replacement corresponds to selecting  $n$  balls out of a well-mixed urn containing  $N$  balls (like some lotteries). For this method:
  - All possible samples of size  $n$  have an equal probability of being selected.
  - All samples of size not equal to  $n$  have zero probability of selection
  - every unit has probability  $n/N$  of selection
- SRS with replacement – units are replaced after selection, can be selected more than once
  - Impractical, but simplifies design-based theory

# SRS example

- Example. Suppose the urn contains  $N = 5$  balls, labeled {A B C D E}; this is our population. We select a simple random sample of  $n = 2$  balls. There are 10 possible samples of size 2, namely:
  - AB, AC, AD, AE, BC, BD, BE, CD, CE, DE
  - Since all these samples have the same chance of being selected,
    - $\Pr(\text{any size 2 sample selected}) = 0.1$
    - $\Pr(\text{any other sample selected}) = 0$
    - $\Pr(\text{any particular ball is included}) = 0.4$

# Formalizing Sampling Distributions

Population units  $i = 1, \dots, N$

Sample indicator  $S_i = \begin{cases} 1, & \text{unit } i \text{ selected} \\ 0, & \text{unit } i \text{ not selected} \end{cases}$

Probability sampling puts known distribution on  $S = (S_1, \dots, S_N)$

This distribution can depend on design variables  $Z$

But not on survey outcomes  $Y$

Simple random sampling of size  $n$  without replacement:

$$\Pr(S = s | Y) = 1 / \binom{N}{n}, \quad \sum_{i=1}^N S_i = n; \quad \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

$$\Pr(S = s | Y) = 0, \quad \sum_{i=1}^N S_i \neq n$$

# Non-random sampling methods

- Some examples of sampling methods that do not yield random samples are:
  - Sample readily accessible individuals
  - Purposive or judgmental sampling
  - Self-selected samples - volunteers, phone polls
  - Quota sampling
- These methods are less scientific and less trustworthy than probability sampling, since they are subject to hidden biases.

# Neyman's (1934) paper: compared Probability Sampling versus “Purposive Sampling”

- Definition of probability sampling:
  - every sample has a *known* probability of being selected
  - every individual in the population has a positive probability of being selected
- Initially, probability sampling was equated with its basic form, simple random sampling (SRS)
  - Every sample of size  $n$  has *equal* chance of being selected, hence an equal probability of selection method (*epsem*)
  - Samples of size other than  $n$  have no chance of being selected
  - With and without replacement

# “Purposive Sampling”

- “Non-probability sampling” – but hard to define a negative.
- Units are picked so that sample matches distribution of a characteristic known for the population.
- E.g. if we know distribution of age and gender in population, choose sample cases to match this distribution.
- A common form is *quota sampling*: interviewers are given a quota for each age group and gender and interview individuals until this quota is met

# The Controversy

- Let  $Z$  = characteristic known for all units in the population (age, gender, ...)
- Under simple random sampling, distribution of  $Z$  in the sample can deviate considerably from its (known) distribution in the population, purely by chance
- This “lack of representativeness” with respect to  $Z$  led some to prefer purposively picking the sample to match the population distribution of  $Z$

# Neyman's “Resolution”

- Neyman (1934) showed that we can get the best of both worlds by stratified sampling:
  - Create strata by the classifying population according to the known characteristics
  - Select a simple random sample of known size  $n_j$  from population of size  $N_j$  in stratum  $j$
- If  $f_j = n_j / N_j = \text{const.}$ , results in epsem sample, retains probabilistic selection, and sample matches distribution of strata in population
- Also one can vary  $f_j$  and weight sample cases by  $1/f_j$ : Neyman's optimal allocation

# More Complex Designs

- Neyman's paper helped to set the stage for extensions to cluster sampling, multistage sampling, greatly extending the practical feasibility and utility of probability sampling in practice
- E.g. simple random sampling of people in the US is not feasible – we do not have a complete list of everyone in the population from which to sample
- Work of Mahalanobis, Hansen, Cochran, Kish, ....

# Beyond simple random sampling

- Stratified Random Sampling
  - divide population into strata (e.g. based on race)
  - select units by srs within each stratum. Different sampling fractions are allowed within strata; for example, we may over-sample minorities
  - Generally, stratifying on a variable that is related to a survey outcome increases the precision for estimating distribution of that outcome
  - More strata the better, but a sample size of at least two in each stratum is needed to provide an estimate the sampling variance

# Systematic sampling from an ordered list

- For a continuous stratifying variable  $Z$ , order the population by values of  $Z$ .
- Choose a sampling interval  $I$  – the inverse of the sampling rate  $n/N$
- Choose a random start between 0 and  $I$ , say  $x$
- Sample units  $x, x+I, x+2I, \dots, x+(n-1)I$
- Creates  $n$  implicit strata of size  $I$ , sample one unit from each stratum
- Simple and convenient, but sampling variance requires modeling assumptions

# PPS sampling

- In certain applications, it is efficient to sample “large” units (firms, tax returns, transactions in an audit)...) with higher probability than “small” units – in particular when variability of outcome increases with size (as with variables like total sales, number of employees, ...)
- For a continuous stratifying size variable  $Z$ , this is conveniently achieved by probability proportional to size (pps) sampling
- Units in the population are first ordered, either randomly or by values of  $Z$ . Then:

# PPS sampling

- Associate unit  $i$  with interval  $(c_{i-1}, c_i]$ , where  $c_0 = 0$ ,  $c_i = z_1 + \dots + z_i$  are cumulated sizes up to  $i$ ,  $i = 1, \dots, n$ .
- Choose a sampling interval  $I = z_n/n$ .
- Choose a random start between 0 and  $I$ , say  $x$
- Units corresponding to the intervals that contain the values  $x, x+I, x+2I, \dots, x+(n-1)I$  are sampled
- Notes:
  - Units with size greater than  $I$  are selected with probability 1. They are pre-selected and removed from the list prior to sampling from the list
  - With units randomly ordered, creates a pps sample with no implicit stratification
  - With units sorted by size, creates a pps sample with implicit stratification on size, and  $n$  implicit strata of size 1. More efficient, but sampling variance requires models

# Cluster Sampling

- Group units into clusters (e.g. localities)
- Select a srs  $c$  of  $C$  clusters
- Sample all units within sampled clusters
- Useful for demographic surveys, since listing operations and interviews can focus on sampled clusters, saving on listing expense and travel time between households
- Less useful for telephone sampling, since there is no travel involved.

# Two-stage sampling

- Group units into clusters (e.g. localities)
- Select a sample of size  $c$  of  $C$  clusters
- Take a simple random sample of units within sampled clusters
- A common design is to sample clusters with probability proportional to estimated size, and units within clusters with probability inversely proportional to estimated size
  - Yields an epsem sample
  - If estimated size is the true size, this yields a constant number of units in each cluster, convenient for fieldwork

# Multistage sampling

- More than two stages are also possible
- E.g. sample households in two stages, and then take a subsample of individuals within households
- Or in a student sample, sample students within classes within schools
- This yields a more complex correlation structure
- The largest clusters in the hierarchy are called ultimate clusters, play an important role in design-based inference

# Multistage sampling with stratification

- Efficiency is increased by stratified sampling one or more stages of selection
- E.g. stratify clusters by cluster characteristics, and take random samples of clusters within strata
- A popular design is to sample two clusters per stratum, since it allows for design-based variances to be computed.

# Checking "Representativeness"

- One way of assessing representativeness is to compare distributions of known variables for the sample and the population
  - e.g. target population = U.S. Civilians
  - compare sample distribution of age, race, and sex with the population distribution from the nearest census.
  - should be done if possible, but often of limited value: really need to compare variables closely associated with the variables of interest

# Distinguishing Statistical Features of Survey Sampling

- A simple and brilliant idea: simple random sampling
- Study of *complex sample designs*: designs that go beyond simple random sampling, including features like stratification, weighting and clustering
  - Simple random sampling, though simple, is not optimal or even practical in many settings
- Many practical real-world sampling issues: sampling frames, making use of administrative information, alternative modes of survey administration

# Is Probability Sampling Optimal?

- Simple random sampling (or equal probability sampling in general) is an all-purpose strategy for selecting units to achieve representativeness “on average”
  - compare with randomized treatment allocation in clinical trials
- However, statisticians like optimal properties, and SRS is very suboptimal for some specific purposes...
- E.g. if distribution of  $X$  is known in population, and objective is slope of linear regression of  $Y$  on  $X$ , it’s obviously much more efficient to sample equally at the two extreme values of  $X$  – this minimizes the variance of the LS slope (Royall 1970)
- But this is not a probability sample– intermediate values of  $X$  have zero chance of selection!
- For linear regression through origin, optimal design is cut-off sampling, which is still applied in some business surveys

# Balanced Sampling

- BUT -- sampling the extremes of  $X$  does not allow checks of linearity, and lacks robustness.
- Royall and Herson (1974) argue that if linearity is a concern, choose fixed number of cases at intermediate values of  $X$ , rather leaving the sample to chance!
  - Their *balanced sampling* idea achieves robustness by matching moments of  $X$  in sample and population
- Even if sampling is random within categories of  $X$ , this is not probability sampling unless all values of  $X$  are included.

# 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:  
 $\theta$  = model parameter (meaningful only in context of the model)  
 $\tilde{\theta}(Y)$  = "estimate" of  $\theta$  from fitting model to whole population  $Y$   
(a finite population quantity, exists regardless of validity of model)  
A good estimate of  $\theta$  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.

- Design-based inference bases the inference directly on these repeated sampling properties
- Calibrated Bayes: model-based, but models should be chosen to have good frequentist properties
- This requires incorporating design features in the model (Little 2004, 2006).

# Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 3: Modes of survey inference



# 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

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

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

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

$Y = (Y_1, \dots, Y_N)$  = 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

$Q = Q(Y, Z)$  = 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  $Q$

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

# 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”

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

$$\Pr(I | Z) = \begin{cases} 1 / \binom{C_n^N}{n}, & \sum_{i=1}^N I_i = n, \\ & ; \quad C_n^N = \frac{N!}{n!(N-n)!}, n! = 1 \times 2 \dots \times n \\ 0, & \text{otherwise} \end{cases}$$

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

# Example 1: Mean for Simple Random Sample

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

Random variable

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

Fixed quantity, not modeled

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

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

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

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

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

# Example 2: Horvitz-Thompson estimator

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

$\pi_i = E(I_i | Y)$  = inclusion probability  $> 0$

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

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

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

- 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

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

Objective:  $T = y_1 + y_2 + \dots + y_{50}$ . 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

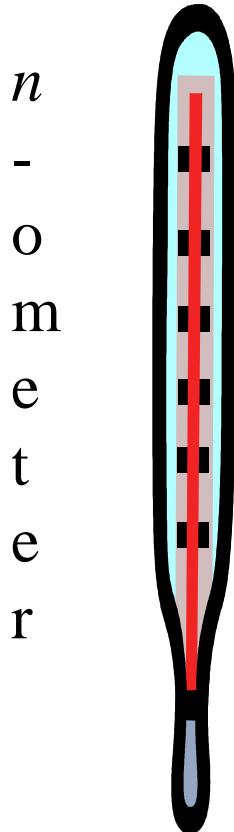
# Role of Models in Classical Approach

- 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

# 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



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

- Some raise theoretical objections to repeated-sampling inferences in general
  - Violates the likelihood principle (Birnbaum 1968)
  - Ambiguity about conditioning on ancillary statistics
- Inference 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

# Model-Based Approaches

- In the Bayesian 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  $M$ :

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

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

$$\hat{y}_i = E(y_i | z_i, \theta = \hat{\theta}), \hat{\theta} = \text{model estimate of } \theta \quad I \quad Z \quad Y$$

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

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

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

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

$I$	$Z$	$\hat{Y}_{\text{exc}}$

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:

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

Inference about  $\theta$  is based on posterior distribution from Bayes Theorem:

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

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

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

of  $Q$  given sample values  $Y_{\text{inc}}$

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

(Integrates out nuisance parameters  $\theta$ )

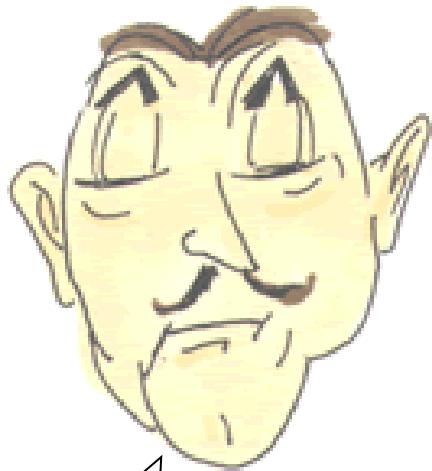
In the super-population modeling approach, parameters are considered fixed and estimated

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

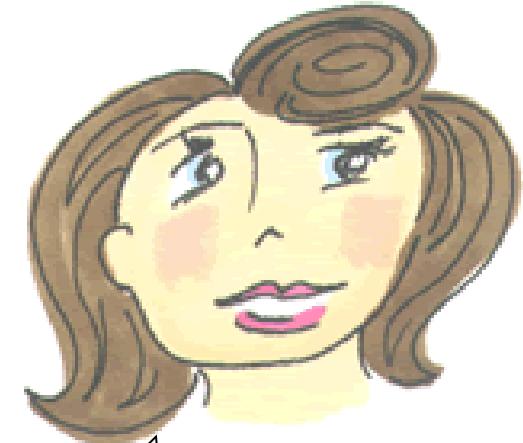
# Advantages of Bayesian approach

- 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. Simulation techniques greatly facilitate implementation

# 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

Trivellore Raghunathan

Module 4: Superpopulation models, and  
maximum likelihood



# 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

# Finite population inference

- Modeling takes a predictive perspective on statistical inference – predict the non-sampled values
  - ML models for the sampling/nonresponse weights lie outside this perspective
- Inference about parameters is intermediate step in predictive superpopulation model inference about finite population parameters

Predict non-sampled values  $\hat{y}_i = E(y_i | \hat{\theta}), \hat{\theta}$  ML estimate of  $\theta$

Estimate of total  $T = \sum_{i \in s}^n y_i + \sum_{i \notin s}^n \hat{y}_i$ , etc.

- Does not reflect uncertainty in ML estimate – Bayes incorporates this by intergrating over posterior distribution of parameters (as discussed later)

# Definition of Likelihood

- Data  $Y$
- Statistical model yields probability density  $f(Y | \theta)$  for  $Y$  with unknown parameters  $\theta$
- Likelihood function is then a function of  $\theta$

$$L(\theta | Y) = \text{const} \times f(Y | \theta)$$

- Loglikelihood is often easier to work with:

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

Constants can depend on data but not on parameter  $\theta$

# Example: Normal sample

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

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

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

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

# Example: Multinomial sample

- $Y = (y_1, \dots, y_n)$  univariate  $K$ -category multinomial sample  
 $n_j$  = number of  $y_i$  equal to  $j$  ( $j=1, \dots, K$ )

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

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

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

# Maximum Likelihood Estimate

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

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

- 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

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

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

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

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  $Y = (y_1, \dots, y_n)$   $\theta = (\mu, \sigma^2)$

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

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

- Multivariate Normal sample

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

- Normal Linear Regression (possibly weighted)

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

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

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

# Multinomial Example

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

$n_j$  = number of  $y_i$  equal to  $j$  ( $j = 1, \dots, K$ )

Likelihood Equations:

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

Hence ML estimate is

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

# Logistic regression

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

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

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

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

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

$\alpha$  are fixed effects;  $\beta$  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  $\theta$   
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,  $\theta$  as fixed; equation defines the sampling distribution of  $\hat{\theta}$
- Bayesian treats  $\theta$  as random,  $\hat{\theta}$  as fixed;  
equation defines posterior distribution of  $\theta$

# 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)

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

- Sandwich estimator (robust properties)

$$\hat{C}^* = I^{-1}(\hat{\theta}) \hat{K}(\hat{\theta}) I^{-1}(\hat{\theta}), \text{ where } \hat{K}(\hat{\theta}) = D_\ell(\hat{\theta}) D_\ell(\hat{\theta})^T$$

# Bootstrap variance estimate

- A bootstrap sample of a complete data set  $S$  with  $n$  observations is a sample of size  $n$  drawn with replacement from  $S$ 
  - Operationally, assign weight  $w_i$  to unit  $i$  equal to number of times it is included in the bootstrap sample

$$w_1, \dots, w_n \sim \text{MNOM}(n; \frac{1}{n}, \dots, \frac{1}{n})$$

# Bootstrap distribution

- Let  $\hat{\theta}^{(b)}$  be ML estimate from the  $b$ th bootstrap data set
- Inference can be based on the bootstrap distribution generated by values of  $\hat{\theta}^{(b)}$
- In particular the bootstrap estimate is

$$\hat{\theta}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^{(b)}$$

with variance

$$\hat{V}_{\text{boot}} = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^{(b)} - \hat{\theta}_{\text{boot}})^2$$

Asymptotic properties similar to sandwich estimator

# Interval estimation

- 95% (confidence, probability) interval for scalar  $\theta$  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:

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

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

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

# Significance Tests

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

$\theta = (\theta_{(1)}, \theta_{(2)})$ ;  $\theta_{(1)0}$  = null value of  $\theta_{(1)}$ ;  $\theta_2$  = 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:  $LR(\hat{\theta}, \tilde{\theta}) = 2 \left[ \ell(\hat{\theta} | Y) - \ell(\tilde{\theta} | Y) \right]$

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  $\theta_0$

$\chi_q^2$  = Chi-squared distribution with  $q$  degrees of freedom

# Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 5: Bayesian models for simple  
random samples



# Consulting Example

- In India (during the late 70's), 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

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

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

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

# General Setup

- Model for  $I = (I_1, I_2, \dots, I_N)$  : Sample Design
- Model for  $Y = (Y_1, Y_2, \dots, Y_N)$  : Prior
- Frame/Design Variables: Z
- Joint distribution:  $\Pr(Y, I | Z)$
- Observed Data:  $(Y_{inc}, I, Z)$
- Missing or Unobserved Data:  $Y_{exc}$
- Inference:  $\Pr(Y_{exc} | Y_{inc}, I, Z)$

$I$	$Z$	$Y$
1		$Y_{inc}$
1		
1		
0		
0		
0		$Y_{exc}$
0		
0		
0		

# Simple Random Sample

- Consider  $Z$  is not available
- $\Pr(Y, I) = \Pr(Y)\Pr(I)$
- Exchangeable Prior/Model for  $Y$ 
  - For any two permutations of the labels or index used in  $Y$   
 $(i_1, i_2, \dots, i_N)$  and  $(j_1, j_2, \dots, j_N)$   
 $\Pr(Y_{i_1}, Y_{i_2}, \dots, Y_{i_N}) = \Pr(Y_{j_1}, Y_{j_2}, \dots, Y_{j_N})$
- That is, the labels have no “information” relevant for the inference
- de Finetti (1937), Hewitt & Savage (1955) and Diaconis & Freedman (1980)
- Exchangeable distribution can be expressed as

$$\Pr(Y_1, Y_2, \dots, Y_N) = \int \prod_{i=1}^N \Pr(Y_i | \theta) \pi(\theta) d\theta$$

# Consulting example

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

$$Y_i | \theta \sim \text{iid Bernoulli}(\theta) \leftarrow \boxed{\text{Model for observable}}$$

$$\pi(\theta) = 1 \quad \theta \in (0, 1) \leftarrow \boxed{\text{Prior distribution}}$$

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

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

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

# Binomial Example

The posterior distribution is

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

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

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

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

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

# Infinite Population

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

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

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?

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

## Inverse CDF of Beta Distribution

Bayesian inference for surveys: simple random sampling

# Numerical Example

- $N=270$  households
- $n=20$  SRS sample
- $x=8$  out 20 had a device
- *Simulation*

- Write a R-code for  $\theta \sim Beta(9,13)$ ;  $X_{N-n} \sim Bin(250, \theta)$ ; and compute  $\bar{Y}_N = (x + X_{N-n}) / N$
- Generate Treat 250 households with missing values and use missing data package (for example IVEware or MICE in R or MI in Stata) which fits the model

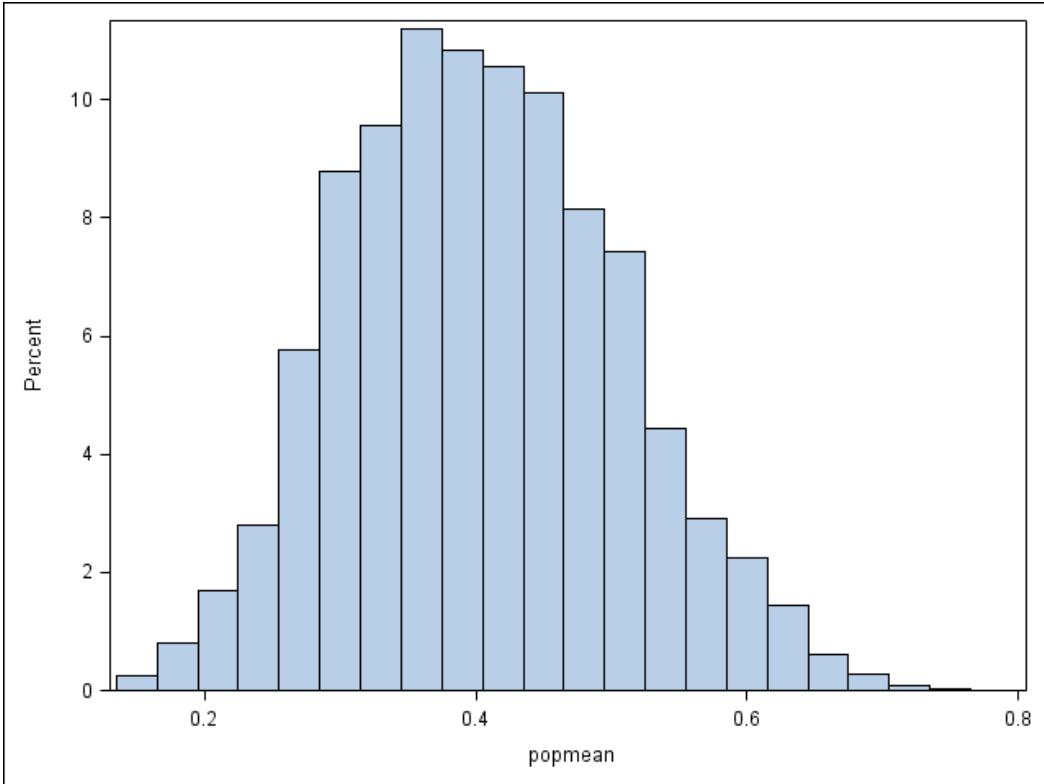
$$\Pr(Y = 1) = \exp(\beta) / (1 + \exp(\beta)), \pi(\beta) \propto 1$$

or

$$Y \sim Bern(\theta), \pi(\theta) \propto \theta^{-1}(1 - \theta)^{-1}$$

$$\begin{aligned}\Pr(\bar{Y}_N \geq 0.3 | x) &= \Pr(Y_{N-n} \geq 0.3 \times N - x | x) \\ &= \int_0^1 \Pr(Y_{N-n} \geq 0.3 \times N - x | \theta, x) \pi(\theta | x) d\theta\end{aligned}$$

# Histogram of the 2,500 Draws



Proportion of Draws exceeding  
0.3=84%  
Posterior mean: 0.4051  
Posterior standard deviation:  
0.1007  
Normal Approximation credible  
interval:  
(0.2077, 0.6025)

# 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  $1-\alpha$  in advance and determine the *highest posterior density* region  $C$  to include most likely values of  $Q$  totaling  $1-\alpha$  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  $1-\alpha$  as the posterior probability of this  $C$

# Interval Estimates

$C$  is such that

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

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

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

“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  $\theta$  used to determine the interval estimate of  $Q$

# Normal simple random sample

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

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

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

$$\begin{aligned} Q &= \bar{Y} = \frac{n\bar{y} + (N-n)\bar{Y}_{\text{exc}}}{N} \\ &= f \times \bar{y} + (1-f) \times \bar{Y}_{\text{exc}} \end{aligned}$$

Derive posterior distribution of  $Q$

# Normal simple random sample

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

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

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

$$\begin{aligned} Q &= \bar{Y} = \frac{n\bar{y} + (N-n)\bar{Y}_{\text{exc}}}{N} \\ &= f \times \bar{y} + (1-f) \times \bar{Y}_{\text{exc}} \end{aligned}$$

Derive posterior distribution of  $Q$

# Normal Example

Posterior distribution of  $(\mu, \sigma^2)$

$$\begin{aligned} p(\mu, \sigma^2 | Y_{\text{inc}}) &\propto (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i \in \text{inc}} \frac{(y_i - \mu)^2}{\sigma^2}\right) \\ &\propto (\sigma^2)^{-n/2-1} \exp\left(-\frac{1}{2} \left( \sum_{i \in \text{inc}} (y_i - \bar{y})^2 / \sigma^2 - n(\mu - \bar{y})^2 / \sigma^2 \right)\right) \end{aligned}$$

The above expressions imply that

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

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

# Posterior Distribution of $Q$

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

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

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

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

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

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

# 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  $(1-\alpha)100\%$  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

$$Population\ size = N_1$$

$$Sample\ size = n_1$$

$$Y_{1i} \sim \text{ind } N(\mu_1, \sigma_1^2)$$

$$\pi(\mu_1, \sigma_1^2) \propto \sigma_1^{-2}$$



$$Sample\ Statistics : (\bar{y}_1, s_1^2)$$

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

$$Population\ size = N_2$$

$$Sample\ size = n_2$$

$$Y_{2i} \sim \text{ind } N(\mu_2, \sigma_2^2)$$

$$\pi(\mu_2, \sigma_2^2) \propto \sigma_2^{-2}$$



$$Sample\ Statistics : (\bar{y}_2, s_2^2)$$

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

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

$$\text{Var}(Y_i | \theta) = \sigma^2 = \sum_k d_k^2 \theta_k - \mu^2$$

# 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  $\mu$  and  $\sigma^2$

# Nonparametric Inference (continued)

- Posterior distribution of  $\theta$  is Dirichlet:

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

- Posterior mean, variance and covariance of  $\theta$

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

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

# Inference for $Q$

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

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

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

Hence posterior mean and variance of  $Q$  are:

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

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

# Posterior Predictive Distribution

*Sample*:  $y_1, y_2, \dots, y_n$

*Non-sample*:  $y_{n+1}, y_{n+2}, \dots, y_N$

*Predictive distribution*:

$$\begin{aligned} \Pr(y_{n+1}, y_{n+2}, \dots, y_N \mid y_1, y_2, \dots, y_n) &= \Pr(y_{n+1} \mid y_1, y_2, \dots, y_n) \times \\ \Pr(y_{n+2} \mid y_{n+1}, y_1, y_2, \dots, y_n) \times \Pr(y_{n+3} \mid y_{n+2}, y_{n+1}, y_1, y_2, \dots, y_n) \times \\ \dots \times \Pr(y_N \mid y_{N-1}, \dots, y_{n+1}, y_{n+1}, y_1, y_2, \dots, y_n) \end{aligned}$$

The Polya Urn Model can be used to obtain draws from the posterior predictive distribution  
(Ghosh and Meeden (1997), Feller (1967))

# Simple random Sample with Auxiliary Variables

## Ratio and Regression Estimates

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

Objective: Infer about the population mean

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

Excluded  $Y$ 's are missing values

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

# Model Specification

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

$$i = 1, 2, \dots, N$$

$g$  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,

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

# Remarks (Continued)

- Bayesian Interpretation: Summary of the excluded portion of the population has approximate normal distribution conditional on the observed data. *That is  $Y_{\text{inc}}$  is fixed and  $Q$  is random.*
- Frequentist Interpretation: Under repeated sampling, the distribution of estimates of  $Q$ . *That is  $Q$  is fixed and  $Y_{\text{inc}}$  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

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

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

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

# 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 Surveys

Roderick Little and Trivellore Raghunathan  
Module 6: Computational Methods



- A Bayesian analysis uses the entire posterior distribution of the parameter of interest.
- 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 or intervals
- Conceptually, all these quantities can be expressed analytically in terms of integrals of functions of parameter with respect to its posterior distribution
- Computations
  - Numerical integration routines
  - Simulation techniques

# Numerical Integration

- Mean  $\int_a^b \theta \pi(\theta | Data) d\theta$
- Variance  $\int_a^b \theta^2 \pi(\theta | Data) d\theta - \left[ \int_a^b \theta \pi(\theta | Data) d\theta \right]^2$
- Probability
$$\Pr(\theta \leq c | Data) = \int_a^b I_{[\theta \leq c]} \pi(\theta | Data) d\theta$$
$$I_{[x \leq y]} = 1 \text{ if } x \leq y \text{ and } 0 \text{ otherwise}$$

- Gaussian quadrature approximates

$$\int_a^b w(x) f(x) dx = \sum_{i=1}^n w_i f(x_i);$$

$$\int_a^b f(x) dx = \int_a^b w(x) \times (f(x) / w(x)) dx$$

$x_i$  = Roots of the polynomials of order  $n$

$w_i$  = Weight function evaluated at the roots

$a = 0, b = \infty$ : Polynomial=Laguerre,  $w(x) = x^\alpha e^{-x}, \alpha > -1$

$a = -\infty, b = \infty$ : Polynomial=Hermite,  $w(x) = e^{-x^2}$

$a = -1, b = 1$ : Polynomial=Jacobi,

$w(x) = (1 - x)^\alpha + (1 + x)^\beta, \alpha, \beta > -1$

$a = -1, b = 1$ : Polynomial=Legendre,  $w(x) = 1$

- Abramovitz and Stegun give a table of values of the weight and abscissa.
- These can be computed in R. Download and install package “statmod” from the r-project web site.
- After installation, use the command
  - `library("statmod")`
  - `gauss.quad(n,polynomial=,a=,b=)`
  - See R manual for more help
- One can use simple SAS macro or even an Excel spread sheet to do these computations.

# Example

$$\int_0^{\infty} x^2 e^{-x^2} dx = \frac{1}{2} \int_{-\infty}^{\infty} x^2 e^{-x^2} dx = \int_0^{\infty} x^2 e^{-x} e^{-x^2+x} dx$$

Substitute  $u = x^2$ ,

$$\frac{1}{2} \int_0^{\infty} u^{1/2} e^{-u} du$$

```
> y=gauss.quad(10,kind="laguerre",alpha=0.5)
> w=y$weight
> a=sum(w)/2
> a
> [1] 0.4431135
```

```
y=gauss.quad(10,kind="hermite")
> w=y$weight
> x=y$nodes
> a=sum(w*x*x)/2
> a
[1] 0.4431135
> sqrt(pi)/4
[1] 0.4431135
```

# Types of Simulation

- Direct simulation (Binomial and normal examples)
- Approximate direct simulation
  - Discrete approximation of the posterior density
  - Rejection sampling
  - Sampling Importance Resampling
- Iterative simulation techniques
  - Gibbs sampler
  - Metropolis Algorithm

# Simulation Techniques

- Numerical integration though can be extended to multidimensional integrals but can be quite time consuming.
- Error in approximation can be large
- Alternative is to draw samples from the posterior distribution and use the sample to characterize the features of the posterior distribution

$$X \sim F$$

Density :  $f(x)$

- Objective: Compute  $E(t(X))$

$x_1, x_2, \dots, x_K \sim$  draws from  $f(x)$

$$E(t(X)) \approx \bar{t} = \frac{1}{K} \sum_{i=1}^K t_i, t_i = t(x_i)$$

Monte-Carlo Error

(in the approximation)

$$e = \sqrt{\frac{1}{K(K-1)} \sum_{i=1}^K (t_i - \bar{t})^2}$$

$$\Pr(t(X) \geq t_o) \approx p_o = \frac{1}{K} \sum_{i=1}^K I_{[t_i \geq t_o]}$$

$I_A = 1$  if  $A$  is true  
 $= 0$  otherwise

$$MCSE = \sqrt{p_o(1 - p_o)/K}$$

Estimation of distribution function

Estimation of percentiles

$$\Pr(t(X) \leq ?) = p_o$$

Order statistics :  $t_{(1)} \leq t_{(2)} \leq \dots \leq t_{(K)}$

$$[Kp_o] \leq Kp_o \leq [Kp_o] + 1$$

$$? \approx t_{([Kp_o])}(Kp_o - [Kp_o]) + t_{([Kp_o] + 1)}([Kp_o] + 1 - Kp_o)$$

- Equal tail probability interval

$$\Pr(t(X) \leq ?_L) = \alpha / 2$$

$$\Pr(t(X) \leq ?_U) = 1 - \alpha / 2$$

- Highest posterior density interval (approximation)
  - Smooth density estimates and then compute highest HPD interval
  - Numerical approximation (assuming that  $K$  is large)

# Unimodal

- Order the values and construct intervals

$$t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(K)}$$

$$R_j : (t_{(j)}, t_{(\lceil j + (1-\alpha)K \rceil)})$$

- Each  $R_j$  is an posterior interval
- Choose the interval that is shortest
- Need a more general approach for multimodal situation. See R manual

# Simulation for the Normal Example

- Revisit normal example

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

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

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

```
# Draws for the normal case
sampszie=20
k=5
ybar=10
ssquare=5
nsimul=1000
result=matrix(0,nsimul,3)
for (i in 1:nsimul){
  tmp=rnorm(sampszie-1)
```

```
chisq=sum(tmp*tmp)
sigmasq=(sampszie-1)*ssquare/chisq;
mu=ybar+sqrt(sigmasq/
sampszie)*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  $x_1, x_2, x_3 \dots, x_m$
- Objective is to infer about the largest of the  $m$  correlation coefficients

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

# Example: Model

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

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

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

- It is analytically difficult to derive the posterior distribution of  $\theta = \max_{1 \leq j \leq m} (\rho_{y, x_j})$
- Even more interesting is to find the posterior mean of

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

- Likelihood

$$\begin{aligned}
 & \prod_{i=1}^n |\Sigma|^{-1/2} \exp[-(U_i - \mu)^t \Sigma^{-1} (U_i - \mu)/2] \\
 & = |\Sigma|^{-n/2} \exp \left[ -\sum_i (U_i - \bar{U})^t \Sigma^{-1} (U_i - \bar{U})/2 \right] \times \\
 & \quad \exp \left[ -n(\mu - \bar{U})^t \Sigma^{-1} (\mu - \bar{U})/2 \right]
 \end{aligned}$$

- Posterior distribution

$$\begin{aligned}
 & \left[ |\Sigma^{-1}|^{(n-m-2)/2} \exp \left[ -Tr(S\Sigma^{-1})/2 \right] \right] \times \\
 & \left[ |\Sigma/n|^{-1/2} \exp \left[ -(\mu - \bar{U})^t (\Sigma/n)^{-1} (\mu - \bar{U})/2 \right] \right]
 \end{aligned}$$

# 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

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

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

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

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

# Example: Simulation

- It is easy to simulate from the posterior distribution of  $m$  and  $S$ .

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

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

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

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

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  $(\mu_*, \Sigma_*)$
- Repeat the above steps to simulate several draws from the posterior distribution.

# Approximate Direct Simulation

- Approximating the posterior distribution by a normal distribution by matching the posterior mean and variance.
  - Posterior mean and variance computed using numerical integration techniques
- An alternative is to use the mode and a measure of curvature at the mode
  - Mode and the curvature can be computed using many different methods
- Approximate the posterior distribution using a grid of values of the parameter and compute the posterior density at each grid and then draw values from the grid with probability proportional to the posterior density

# Normal Approximation

Posterior density :  $\pi(\theta | x)$

Easy to work with log-posterior density

$$l(\theta) = \log(\pi(\theta | x))$$

At the mode,  $f(\theta) = l'(\theta) = 0$

Curvature :  $f'(\theta) = l''(\theta)$

For logarithm of the normal density

Mode is the mean and

the curvature at the mode

is negative of the precision

(Precision:reciprocal of variance)

# Rejection Sampling

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

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

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

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

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

# Sampling Importance Resampling

- Target density from which to draw  $\pi(\theta | \text{data})$
- Candidate density from which it is easy to draw  $g(\theta)$ , such that  $g(\theta) > 0$  for all  $\theta$  with  $\pi(\theta | \text{data}) > 0$
- The importance ratio  $w(\theta) \propto \frac{\pi(\theta | \text{data})}{g(\theta)}$
- Sample M values of  $\theta$  from  $g$   $\theta_1^*, \theta_2^*, \dots, \theta_M^*$
- Compute the M importance ratios and resample with probability proportional to the importance ratios.  $w(\theta_i^*); i = 1, 2, \dots, M$

# Markov Chain Simulation

- In real problems it may be hard to apply direct or approximate direct simulation techniques.
- The Markov chain methods involve a random walk in the parameter space which converges to a stationary distribution that is the target posterior distribution.
  - Metropolis-Hastings algorithms
  - Gibbs sampling

# Gibbs sampling

- Gibbs sampling a particular case of Markov Chain Monte Carlo method suitable for multivariate problems

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

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

Gibbs sequence :

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

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

M

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

M

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

1. This is also a Markov Chain whose stationary Distribution is  $f(\underline{x})$
2. This is an easier Algorithm, if the conditional densities are easy to work with
3. If the conditionals are harder to sample from, then use MH or Rejection technique within the Gibbs sequence

# Metropolis-Hastings Approach

- A Markov Chain can be constructed whose stationary distribution is the desired posterior distribution
- Metropolis et al (1953) showed how and the procedure was later generalized by Hastings (1970). This is called Metropolis-Hastings algorithm.
- Algorithm:
  - Step 1 At iteration  $t$ , draw

$$y \sim p(y | x^{(t)})$$

$y$ : Candidate Point

$p$ : Candidate Density

- Step 2: Compute the ratio

$$w = \text{Min} \left\{ 1, \frac{f(y) / p(y | x^{(t)})}{f(x^{(t)}) / p(x^{(t)} | y)} \right\}$$

- Step 3: Generate a uniform random number,  $u$

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

$$X^{(t+1)} = X^{(t)} \text{ otherwise}$$

- This Markov Chain has stationary distribution  $f(x)$ .
- Any  $p(y|x)$  that has the same support as  $f(x)$  will work
- If  $p(y|x)=f(x)$  then we have independent samples
- Closer the proposal density  $p(y|x)$  to the actual density  $f(x)$ , faster will be the convergence.

# Remarks

- To reduce the impact of starting point usually a few draws are ignored (“Burn-in period”)
- Each successive draws are dependent. Sample every  $k^{th}$  observation to get approximate draws.
- Assessing whether or not the chain has converged is difficult.
- Several sequences starting at different points in the parameter space may be useful to assess convergence.

# Convergence

- Suppose that there are  $J$  parallel sequences each with  $n$  iterations. Let  $q$  be the scalar parameter of interest.
- Between variance

$$B = \frac{n}{J-1} \sum_{j=1}^J (\bar{\theta}_{+j} - \bar{\theta}_{++})^2$$

$$\bar{\theta}_{+j} = \sum_i \theta_{ij} / n; \quad \bar{\theta}_{++} = \sum_j \bar{\theta}_{+j} / J$$

- Within variance

$$W = \frac{\sum_i \sum_j (\theta_{ij} - \bar{\theta}_{+j})^2}{J(n-1)}$$

# Convergence

- 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.

# Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 7: Models for Complex Surveys  
(Stratification)



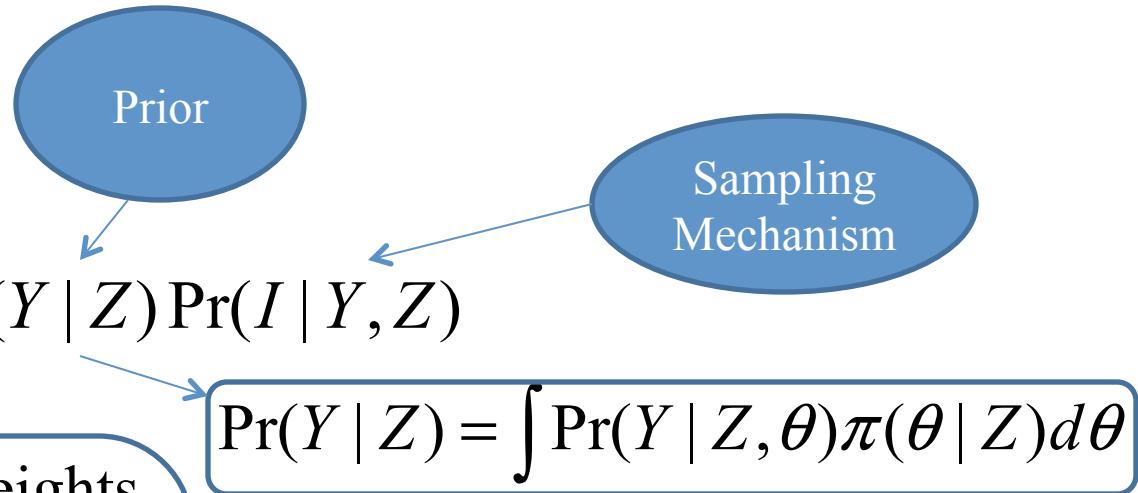
# General Setup

$I$  : Inclusion indicator

$Y$  : Survey Variables

$Z$  : Design Variables

$$\text{Model} : \Pr(Y, I | Z) = \Pr(Y | Z) \Pr(I | Y, Z)$$



Design variables include Weights,  
Clustering, Stratification.

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

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

# Goal

Posterior (predictive) distribution:

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

Two Stage construction:

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

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

# Particular Cases

- Scenario 1

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

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

- Scenario 2

$$\Pr(Y | Z)$$

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

- Scenario 3

$$\Pr(Y | Z)$$

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

- Scenario 4

$$\Pr(Y | Z)$$

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



Ignorable  
Sampling  
Mechanisms



Nonignorable  
Sampling  
Mechanism

# Stratified Random Sample Design

- Population Setup

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

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

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

- Model or Prior

- Exchangeable within stratum (indexing within a stratum is arbitrary)

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

# Examples

- Binary Outcome

$$Y_{ih} | \theta_h : \text{Bern}(1, \theta_h)$$

$$\theta_h : \text{Beta}(a_h, b_h)$$

$a_h, b_h : \text{Known}$

$h = 1, 2, K, H$

- Continuous (Normal) Outcome

$$Y_{ih} | \mu_h, \sigma_h : N(\mu_h, \sigma_h^2)$$

$$\pi(\mu_h, \sigma_h^2) \propto$$

$$(\sigma_h^2)^{-d_h/2} \exp\left[-\frac{1}{2}\left(\frac{c_h}{\sigma_h^2} + \frac{b_h(\mu_h - a_h)^2}{\sigma_h^2}\right)\right]$$

$a_h, b_h, c_h, d_h : \text{Known}$

# Numerical Example

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

# Example (Continued)

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

$$\theta_1 : \text{Beta}(1.5, 6.5) \quad \theta_2 : \text{Beta}(7.5, 14.5)$$

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

$$\theta_3 : \text{Beta}(4.5, 4.5) \quad \theta_4 : \text{Beta}(6.5, 1.5)$$

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

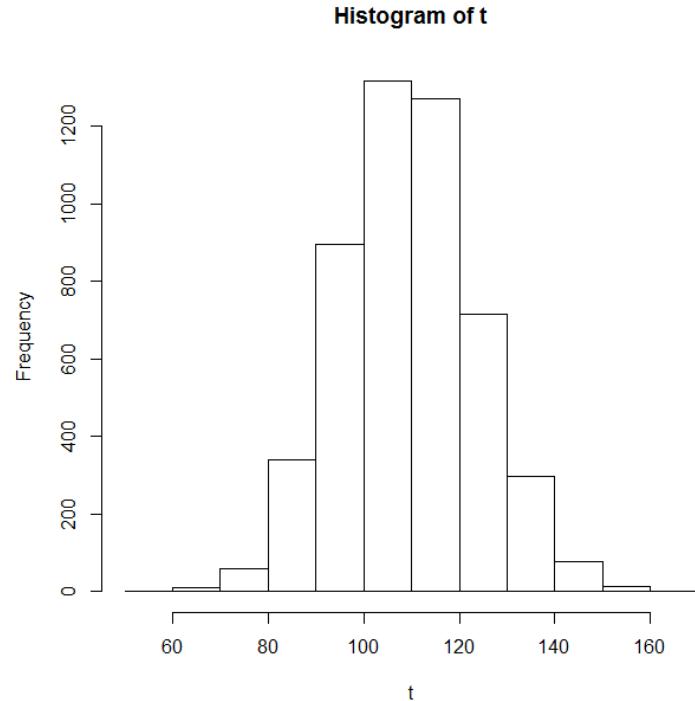
$$T = 2 + 8 + 5 + 7 + Y_{exc,1} + Y_{exc,2} + Y_{exc,3} + Y_{exc,4}$$

# R-Code and Results

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

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

(82,138)



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

(82,137)

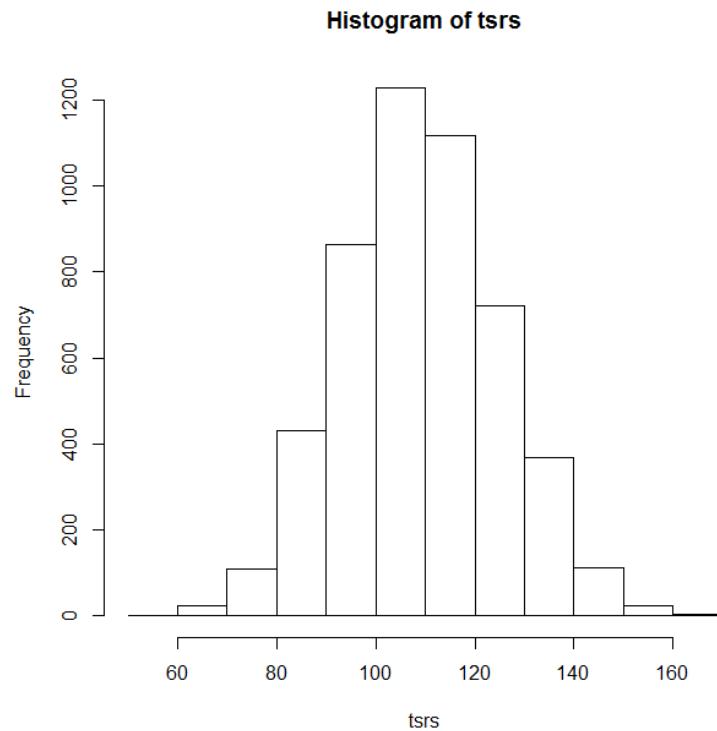
# Ignoring Stratification?

$\theta : Beta(21.5, 28.5)$

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

$t = 22 + Y_{exc}$

**HPD interval: (77,138)**



# Analysis Using a Missing Data Package

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

# Normal Continuous

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

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

$$W_h = N_h / N$$

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

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

$$f_h = n_h / N_h$$

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

# Generalization

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

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

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

- Exchangeability of strata indices implies

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

(Random effect models)

# Bayesian Infrence for Surveys

Module 7 (Continued)  
Models for Stratified Sample Design

# Model Refinements

- A popular design: 2 units sampled per stratum
- Too few to estimate the variance
- Option 1
  - Pooled variance

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

$$\pi(\mu_1, \dots, \mu_H, \sigma) \propto \sigma^{-2}$$

$$i = 1, 2, \dots, N_h; h = 1, 2, \dots, H$$

# Implementation

- Missing Data Approach
  - Create Two Variables Y, Z (Strata indicators)
  - Set all the unobserved values to missing
  - Multiply Impute the unobserved values using a regression model of Y on dummy variables based on Z
  - This is akin to fitting a one-way analysis of variance with Strata as Groups

- Option 2
  - Proper prior for variance parameters across strata

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

$$\pi(\mu_h, h=1, 2, \dots, H) \propto 1$$

$$\sigma_h^{-2} \sim iid Gamma(a, b)$$

$a, b : Known$

- Option 3
  - Random effects on both Mean and Variance

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

$$\mu_h | \sigma_h^2 \sim N(\mu, c_h \sigma_h^2)$$

$$\sigma_h^{-2} \sim iid Gamma(a, b)$$

$$c_h, a, b : Known$$

Options 2 and 3 require Gibbs sampling approach and can be implemented using Openbugs, Stan or Winbugs, Proc MCMC etc

# Systematic Sampling

- Population size  $N=nk$
- Sample size= $n$
- Elements sequenced into  $n$  groups each of size  $k$
- Choose a random number between 1 and  $k$  (say,  $L$ )
- Sample :  $L, L+k, L+2k, \dots, L+(n-1)k$

**Can be viewed as sampling 1 element from each of the  $n$  strata of size  $k$ . However, only one random start determines all the selection**

(Refinements for  $N$  that is not multiple of  $n$  are available)

# Model

- Treat as SRS
- Combine adjacent 2 groups to create  $n/2$  strata with 2 selections per stratum
- Assume some model for Y as a function index (ordered values)

(Variance estimation is also a problem in the design based inference)

# Bayesian inference for sample surveys

Roderick Little and Trivellore Raghunathan

Module 9: Bayesian models for  
stratified 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)

# General set-up: models that include data collection

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

$Z$  = fully-observed covariates, design variables

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

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

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

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

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

# Full model for $Y$ and $I$

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

Model for  
Population

Model for  
Inclusion

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

- Observed-data likelihood:

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

- Posterior distribution of parameters:

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

# Ignoring the data collection process

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

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

- The corresponding posteriors for  $\theta$  and  $Y_{\text{exc}}$  are:

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

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

Posterior predictive distribution of  $Y_{\text{exc}}$

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

# 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):

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

Bayesian Distinctness:

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

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

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

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

# Bayes inference for probability samples

- In probability sampling designs, selection does not depend on values of  $Y$  and the mechanism is known, that is:
$$p(I | Y, Z, \phi) = p(I | Z) \text{ for all } Y.$$
- 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 variables  $Y$  with the design variables  $Z$ .

# Stratified and PPS samples

- For **stratified samples**,  $Z$  consists of stratum indicators, so models for  $Y$  need to include stratum indicators as covariates
  - Same selection fraction across strata yields epsem design
  - Different selection fractions across strata yields unequal probability design – sampling weights are the inverse of selection fractions
- For **PPS sampling**,  $Z$  is the size variable, and models for  $Y$  need to include size as a covariate
  - Sampling weight is then proportional to *inverse* of  $Z$
- In either case, other auxiliary variables can be included, but correctly modeling the relationship between  $Y$  and  $Z$  is particularly important to avoid bad inferences because of model misspecification

# Design-based weighting

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

$$w_i = w_{is} \times w_{in} \times w_{ip}$$

$w_{is}$  = sampling weight

$w_{in}$  = nonresponse weight

$w_{ip}$  = post-stratification weight

- We'll compare Bayesian (predictive) inference with weighting

# Weighting and models

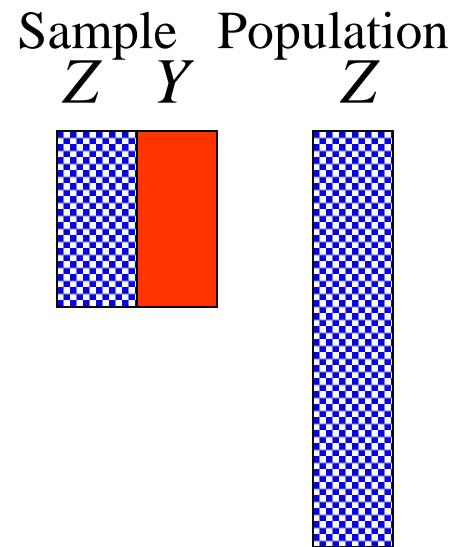
- The weights can't generally be ignored from a modeling perspective
  - Ignores different selection effects that bias estimates
- Weights are auxiliary covariates from a modeling perspective
- Design: weight the respondents
  - Simple: same weights for all  $Y$  variables, but:
  - Weighting adds noise for  $Y$ 's unrelated to weights
- Model: use weights as covariates to predict non-sampled and non-responding values
  - More flexible, but need a good model

# Ex 1: stratified random sampling

- Population is divided into  $J$  strata
- Simple random sample of  $n_j$  units selected from population of  $N_j$  units in stratum  $j$ .

- $Z$  is a variable indicating stratum:

$$z_i = j, \text{if unit } i \text{ is in stratum } j \quad (j = 1, \dots, J)$$



- In a regression model,  $Z$  is represented by a set of  $J - 1$  binary indicators for stratum, the stratum left out being the reference stratum (dummy variable regression)
- This design is ignorable *providing* model for survey variable  $Y$  conditions on the stratum indicators  $Z$ .

# Inference for a mean from a stratified sample

- A normal model that includes stratum effects is:

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

- For simplicity assume  $\sigma_j^2$  is known and the flat prior:

$$p(\theta_j \mid Z) \propto \text{const.}$$

- 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  $\log(\text{variances})$  yields useful t-like corrections for small samples (See module 7)

# 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

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

- 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  $\bar{Y}$ , and hence is design-consistent
- For the normal model that ignores stratum effects, the posterior mean  $\bar{y}$  converges to
$$\bar{Y}_\pi = \sum_{j=1}^J \pi_j N_j \bar{Y}_j / \sum_{j=1}^J \pi_j N_j$$
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

# Ex 2: PPS sampling

- In certain applications, it is efficient to sample “large” units (firms, tax returns, transactions in an audit)...) with higher probability than “small” units – in particular when variability of outcome increases with size (as with variables like total sales, number of employees, ...)
- For a continuous stratifying size variable  $Z$ , this is conveniently achieved by probability proportional to size (pps) sampling
- Units in the population are first ordered, either randomly or by values of  $Z$ . Then:

# PPS sampling

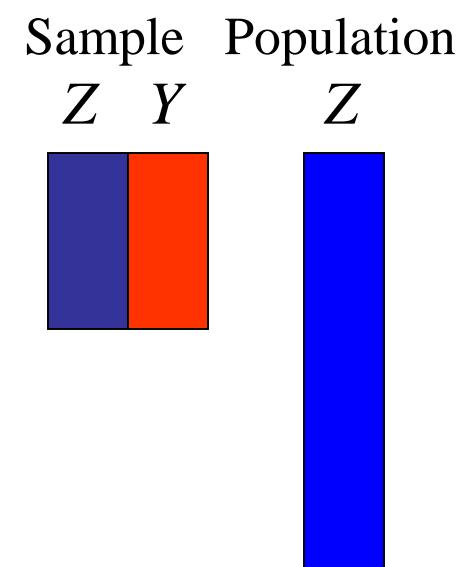
- Associate unit  $i$  with interval  $(c_{i-1}, c_i)$ , where  $c_0 = 0$ ,  $c_i = z_1 + \dots + z_i$  are cumulated sizes up to  $i$ ,  $i = 1, \dots, n$ .
- Choose a sampling interval  $I = z_n/n$ .
- Choose a random start between 0 and  $I$ , say  $x$
- Units corresponding to the intervals that contain the values  $x, x+I, x+2I, \dots, x+(n-1)I$  are sampled
- Notes:
  - Units with size greater than  $I$  are selected with probability 1. They are pre-selected and removed from the list prior to sampling from the list
  - With units randomly ordered, creates a pps sample with no implicit stratification
  - With units sorted by size, creates a pps sample with implicit stratification on size, and  $n$  implicit strata of size 1. More efficient, but sampling variance requires models

## Ex 2. PPS sampling, $Z = \text{size}$

Consider PPS sampling,  $Z = \text{measure of size}$

Standard design-based estimator is weighted  
Horvitz-Thompson estimate

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



Question: is there a model for  $Y$  for which the predictions yield the HT estimate of the mean?

An alternative to HT: Hajek:  $\bar{y}_{HK} = \frac{1}{\hat{N}} \sum_{i=1}^n (y_i / \pi_i), \hat{N} = \sum_{i=1}^n (1 / \pi_i)$

$$(\bar{y}_{HK} = \bar{y}_{HT} \text{ when } \hat{N} = N)$$

Question: when  $\bar{Y}_{HK} \neq \bar{Y}_{HT}$ , which is better? More on this later...

# Projection vs Prediction

Sample  $i = 1, \dots, n$ , non-sample  $i = n + 1, \dots, N$

$$\bar{Y} = \text{population mean} = \sum_{i=1}^n y_i / N$$

$\hat{y}_i$  = prediction of  $y_i$  from a model

$$\text{prediction estimator: } \bar{Y}_{\text{pred}} = \left( \sum_{i=1}^n y_i + \sum_{i=n+1}^N \hat{y}_i \right) / N$$

$$\text{projection estimator: } \bar{Y}_{\text{proj}} = \sum_{i=1}^N \hat{y}_i / N = \bar{Y}_{\text{pred}} + \frac{n}{N} \sum_{i=1}^n (y_i - \hat{y}_i) / n$$

Similar, particularly if  $n \ll N$

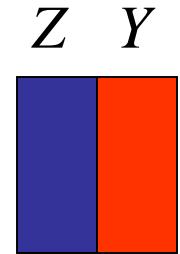
## Ex 2. PPS sampling, $Z = \text{size}$

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

$r_i = y_i / \pi_i \sim \text{Nor}(\beta, \sigma^2)$ , so

$\hat{\beta} = \bar{r} = \frac{1}{n} \sum_{i=1}^n (y_i / \pi_i)$ , yielding prediction  $\hat{y}_j = \hat{\beta}\pi_j$

Sample   Population



$$\bar{Y}_{\text{proj}} = \frac{1}{N} \sum_{j=1}^N \hat{y}_j = \frac{\hat{\beta}}{N} \sum_{j=1}^N \pi_j = \frac{1}{Nn} \sum_{i=1}^n (y_i / \pi_i) \sum_{j=1}^N \pi_j$$

$$= \frac{1}{N} \sum_{i=1}^n (y_i / \pi_i), \left( \text{since } \sum_{j=1}^N \pi_j = n \right) = \bar{y}_{\text{HT}}$$

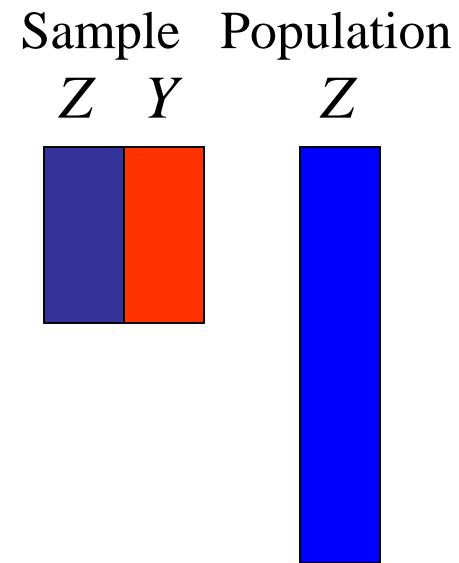
That is, the HT estimator is the projection estimator under the HT model.

## Ex 2. PPS sampling, $Z = \text{size}$

Implication:

When the relationship between  $Y$  and  $Z$  is well described by the HT model, the HT estimate performs well

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



# Ex. Basu's inefficient elephants

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

Objective:  $T = y_1 + y_2 + \dots + y_{50}$ . 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:

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

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  $y_i / \pi_i$  have the same distribution
- That is clearly a bad model given this design ...
- Which is why the estimator is silly

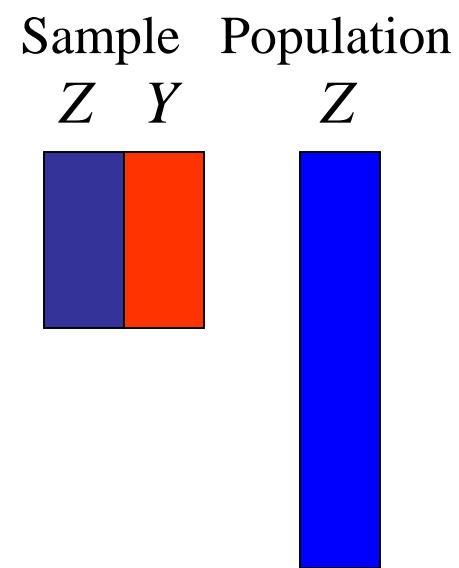
## Ex 2. PPS Sampling, $Z = \text{size}$

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

A modeling alternative to  $\bar{y}_{\text{HT}}$  is  
the prediction estimate

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

from a more flexible model relating  $Y$  to  $Z$



Zheng and Little (2004, 2005) fit a penalized spline model, and show superior performance to HT in simulations

# Making the HT model more flexible

Mean of HT model:  $E(y_i | z_i) = \beta z_i$  (linear through origin)

A. Polynomial regression:  $E(y_i | z_i) + \beta_0 + \beta_1 z_i + \dots + \beta_k z_i^k$

B. Model mean of  $Y$  as a smooth flexible function of  $Z$

e.g. Penalized Spline (Ruppert and Carroll, 2000) with linear basis:

Set  $m$  knots  $\kappa_1, \dots, \kappa_m$  at known values of  $Z$

(e.g. equally-spaced percentiles of distribution of  $Z$ )

$$E(y_i | z_i) = \beta_0 + \beta_1 z_i + \sum_{j=1}^m \alpha_j (z_i - \kappa_j)_+,$$

$$u_+ = \begin{cases} u, & \text{if } u > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\alpha_j \stackrel{\text{iid}}{\sim} N(0, \tau^2), j = 1, \dots, m.$$

This is a linear mixed model:

$\beta_0, \beta_1$  are *fixed* effects

$\alpha_1, \dots, \alpha_m$  are *random* effects

# Making the HT model more flexible

Variance of HT model:  $\text{Var}(y_i | z_i) = \sigma^2 z_i^2$

Replace by  $\text{Var}(y_i | z_i) = \sigma^2 z_i^k$ ,

$k$  an unknown parameter to be estimated

# Fully Bayes model specification

Set  $m$  knots  $\kappa_1, \dots, \kappa_m$  at known values of  $Z$

$$(y_i | z_i, \{\alpha_j\}, \beta_0, \beta_1, \sigma^2, \tau^2, k)$$

$$\sim_{\text{iid}} N\left(\beta_0 + \beta_1 z_i + \sum_{j=1}^m \alpha_j (z_i - \kappa_j)_+, \sigma^2 z_i^k\right)$$

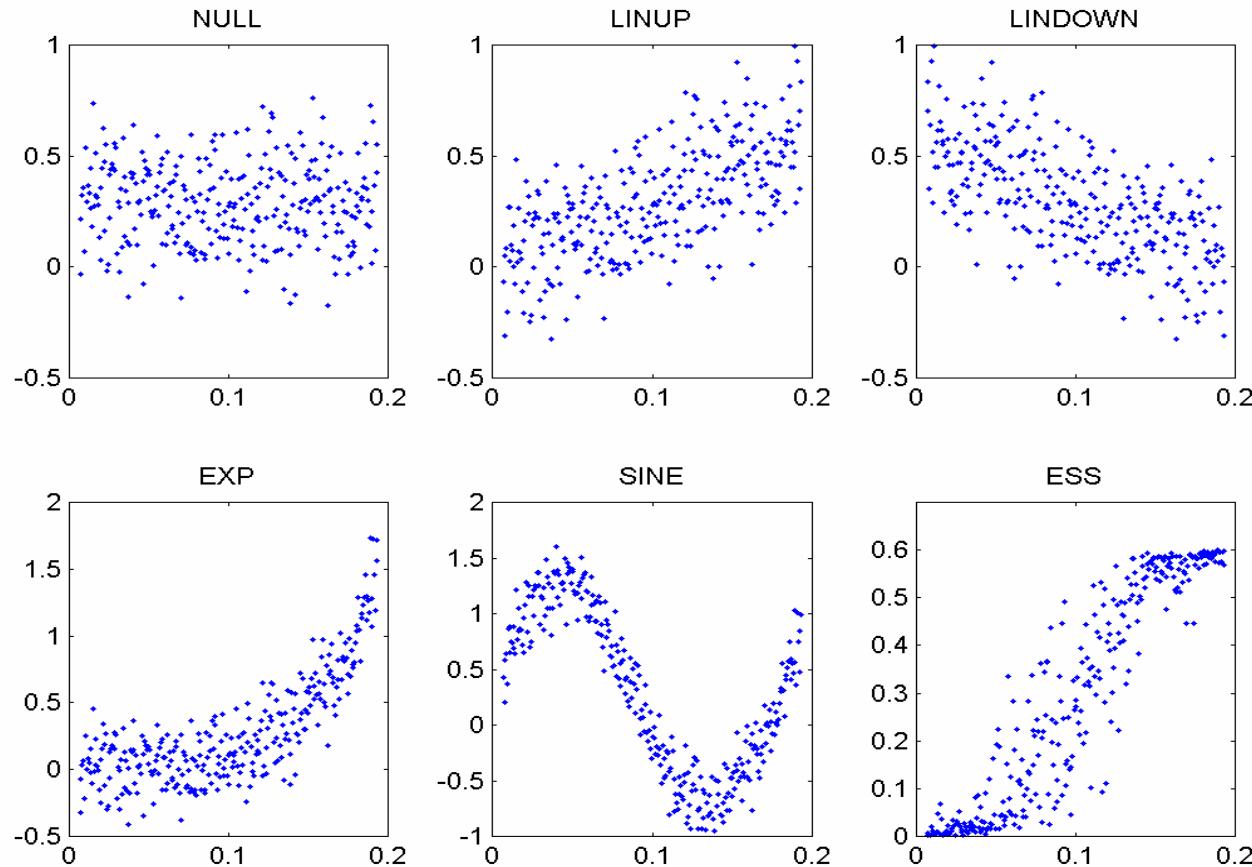
$$\alpha_j | \tau^2 \sim_{\text{iid}} N(0, \tau^2), j = 1, \dots, m.$$

Priors:

$$\pi(\beta_0, \beta_1, \sigma^2, \tau^2) = \text{const. } \sigma^{-2}$$

$$\pi(k | \beta_0, \beta_1, \sigma^2, \tau^2) = 1/4 \quad (-2 \leq k \leq 2)$$

# Simulation: PPS sampling in 6 populations



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

Population		model	wt	gr
NULL	Normal	<b>20</b>	33	21
	Lognormal	32	44	<b>31</b>
LINUP	Normal	<b>23</b>	24	25
	Lognormal	<b>25</b>	30	30
LINDOWN	Normal	30	66	<b>29</b>
	Lognormal	<b>24</b>	65	28
SINE	Normal	<b>35</b>	134	90
	Lognormal	<b>53</b>	130	84
EXP	Normal	<b>26</b>	32	57
	Lognormal	<b>40</b>	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	<b>88.0</b>	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 spline 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

# Hajek (ratio) estimator: A common alternative to HT

Horvitz-Thompson:  $\bar{y}_{HT} = \frac{1}{N} \sum_{i=1}^n (y_i / \pi_i)$

Hajek:  $\bar{y}_{HK} = \frac{1}{\hat{N}} \sum_{i=1}^n (y_i / \pi_i), \hat{N} = \sum_{i=1}^n (1 / \pi_i)$

( $\bar{y}_{HK} = \bar{y}_{HT}$  when  $\hat{N} = N$ )

Question: when  $\bar{y}_{HK} \neq \bar{y}_{HT}$ , which is better?

# A common alternative to HT

$\bar{y}_{HK}$  is projection estimator for Hajek model:

$$y_i \mid \pi_i \sim_{iid} N(\mu, \sigma^2 \pi_i)$$

$$(\bar{Y}_{\text{proj}} = \hat{\mu} = \sum_{i=1}^n (y_i / \pi_i) / \sum_{i=1}^n (1 / \pi_i) = \bar{y}_{HK})$$

So Hajek is better when this model is a better fit to the data

Note that the more general model

$$y_i \mid \pi_i \sim_{iid} N(\beta_0 + \beta_1 \pi_i, \sigma^2 \pi_i^k)$$

includes HT and HK model as special cases

Could fit this model and let the data decide ...

Zheng and Little spline model is even more flexible...

at the expense of more parameters to estimate

## Ex 3: Bayes for binary $Y$

- Inference for finite population proportion – slides from Qixuan Chen's thesis presentation (Chen, Elliott and Little, 2010)
- She also worked on estimating percentiles of a distribution (Chen, Elliott and Little, 2012)

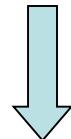
# Design-based vs. model-based (cont.)

## Design-based estimators

- design unbiased
- potentially very inefficient
- variance estimation is cumbersome, and CI may deviate from nominal level at small sample size

## Parametric model-based estimators

- subject to bias when the underlying model is misspecified
- efficient if model is correct
- variance estimation is more straightforward



Zheng and Little  
(2003, 2005)

## Robust Bayesian predictive estimators

- robust to model misspecification
- efficient
- variance or CI is estimated from posterior distribution, and the confidence coverage is close to the nominal level

# Probit p-spline regression model

- Probit truncated polynomial p-spline model (Ruppert, Wand, and Carroll 2003):

$$\Phi^{-1}\left(P(y_i = 1)\right) = \beta_0 + \sum_{k=1}^p \beta_k \pi_i^k + \sum_{l=1}^m b_l (\pi_i - k_l)_+^p$$
$$b_l \sim N(0, \tau^2), l = 1, \dots, m; i = 1, \dots, N$$

- the constants  $k_1 < \dots < k_m$  are  $m$  selected fixed knots.

# BPSP model for binary $Y$

- Gibbs sampling to obtain posterior distributions

- Model  $y$  via a normal latent variable

$$y_i^* \sim N\left(\left(X\beta + Zb\right)_i, 1\right), \quad y_i = I\left(y_i^* > 0\right)$$

- prior distributions  $\beta_i \sim N(0, 10^6)$ , or  $\{\beta_i \propto 1\}$

$$\tau^2 \sim IG(A, B), \text{ or } \{\tau \propto 1\}$$

- posterior distributions:

$$(\beta, b) | \tau^2, y^* \sim MVN_{m+p+1}\left(\left(C^T C + D / \tau^2\right)^{-1} C^T y^*, \left(C^T C + D / \tau^2\right)^{-1}\right)$$

$$\tau^2 | \beta, b \sim IG\left(A + m / 2, B + \|b\|^2 / 2\right), \quad C = [X, Z], \quad D \text{ a diagonal matrix}$$

with  $p+1$  values of  $10^{-6}$  followed by  $m$  1's on the diagonal

- can also be implemented using WinBUGS. (Crainiceanu, Ruppert, and Wand 2005)

# BPSP estimator (cont.)

- The posterior distribution of the population proportion can be simulated by generating a large number  $D$  of draws of the form

$$p^{(d)} = N^{-1} \left( \sum_{i \in s} y_i + \sum_{j \notin s} \hat{y}_j^{(d)} \right)$$

- Bayesian p-spline predictive (BPSP) estimator: average of these draws.
- The  $100(1 - \alpha)\%$  credible interval: split the tail area  $\alpha$  equally between the upper and lower endpoints.

# Other estimators

- The Hájek estimator (discussion of Basu (1971))

$$\hat{p}_{HK} = \left( \sum_{i \in s} y_i / \pi_i \right) / \left( \sum_{i \in s} 1 / \pi_i \right)$$

- The parametric model-based estimators

$$\hat{p}_M = N^{-1} \left( \sum_{i \in s} y_i + \sum_{j \notin s} \hat{y}_j \right)$$

$\hat{y}_j$  = prediction from linear logistic or probit model

- The generalized regression (GR) estimators (Lehtonen and Veijanen 1998)

$$\hat{p}_{GR} = N^{-1} \sum_{j=1}^N \hat{y}_j + \left( \sum_{i \in s} (y_i - \hat{y}_i) / \pi_i \right) / \left( \sum_{i \in s} 1 / \pi_i \right)$$

$\hat{y}_j$  = prediction from linear logistic or probit model

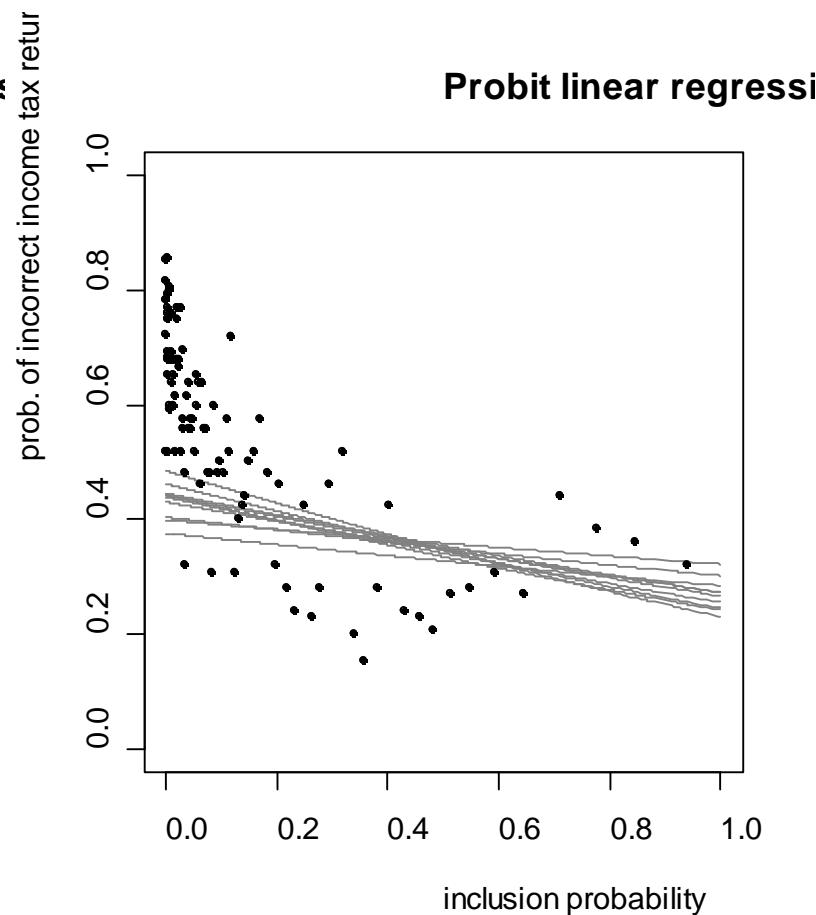
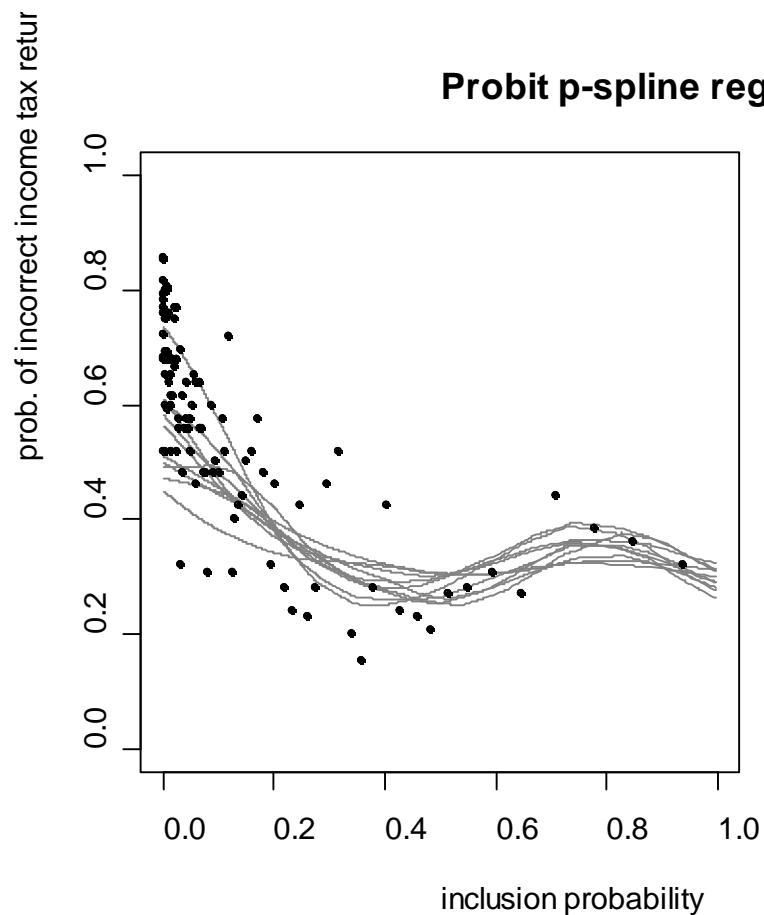
# Simulation studies

- Comparison study
  - **HK**, design-based Hájek estimator
  - **LR**, design-consistent predictive estimator with the ML predictions from the model  $\text{logit}(p_i) = \beta_0 + \beta_1 \pi_i^{-1}$  (Firth and Bennett 1998)
  - **PR**, predictive estimator with predictions from the Bayesian probit model  $\Phi^{-1}(p_i) = \beta_0 + \beta_1 \pi_i$
  - **PR\_GR**, the generalized regression (GR) estimator with the weighted ML predictions from the model  $\Phi^{-1}(p_i) = \beta_0 + \beta_1 \pi_i$
  - **BPS**, the BPS estimator ( $p = 1$  and 15 knots)
  - **BPS\_GR**, the GR estimator with the posterior means of  $\Pr(Y_i = 1 | \pi_i)$  from the BPS model as predictions

## Simulation study (2)

- Tax auditing data (Compumine 2007)
  - 3,119 income tax returns
  - $Y$ : whether the income tax return is incorrect ( $p=0.517$ )
  - $X$ : the amount of the realized profit
  - PPS sampling using  $X$  as the size variable
  - $n = 300$  or  $600$
  - 1,000 replicates of simulation

# Simulation study (2): Tax auditing data



# Simulation study (2): results

Table 3 Comparison of various estimators for empirical bias, root mean squared error, and average width and noncoverage rate of 95% CI, in the tax return example

Methods	bias*100		RMSE*100		average width*100		noncoverage*100	
	300	600	300	600	300	600	300	600
HK	-2.4	-1.8	12.4	10.2	36	29	14.1	10.2
LR	6.7	5.5	11.9	9.2	27	21	43.5	45.6
PR	-11.6	-10.1	12.4	10.6	18	14	69.8	83.4
PR_GR	-1.2	-0.3	11.5	8.8	33	26	16.1	11.4
B PSP	-6.8	-2.7	9.3	5.2	27	19	14.2	5.0
B PSP_GR	-0.7	0.2	12.0	10.1	34	26	15.9	12.8

\* The variance of GR estimator is estimated using linearization

★ BPSP estimator performs well; PR estimator is biased and has poor confidence coverage because of model misspecification

# Discussion

- The BPSP estimator yields smaller RMSE than the Hájek and GR estimators, despite slightly higher empirical bias.
- The BPSP estimator achieves robustness to model misspecification compared to parametric model-based estimators.
- The BPSP estimator has closer to nominal level confidence coverage and shorter average length of 95% CI than the Hájek and GR estimators.
  - especially when  $p$  is closer to zero or one and few data are selected into the sample in the tails.
  - This suggests the importance of the current research in estimating finite population prevalence of rare events.

# Ex 4. One stratifier $Z_1$ , one post-stratifier $Z_2$

## Design-based approaches

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

Notes: (1)  $Z_1$  proportions are not matched!

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

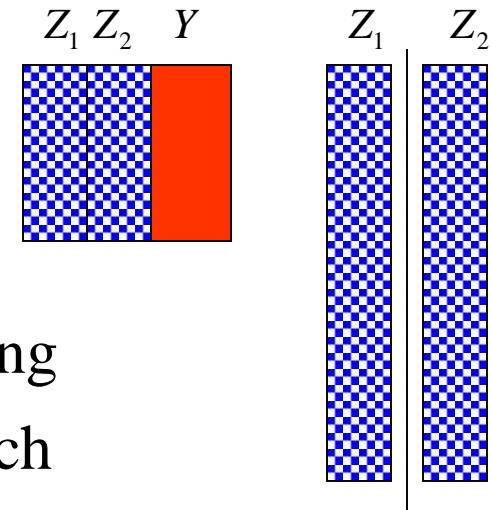
(B) Deville and Sarndal (1992) modifies sampling weights  $\{w_{is}\}$  to adjusted weights  $\{w_i\}$  that match poststratum margin, but are close to  $\{w_{is}\}$  with respect to a distance measure  $d(w_{is}, w_i)$ .

Questions:

What is the principle for choosing the distance measure?

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

Sample   Population



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

## Model-based approach

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

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

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

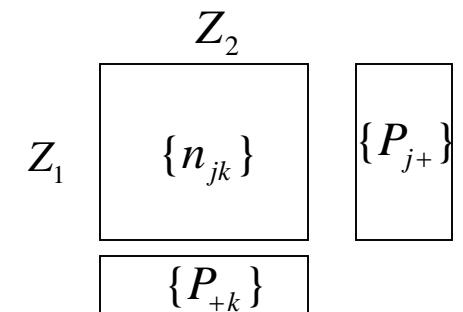
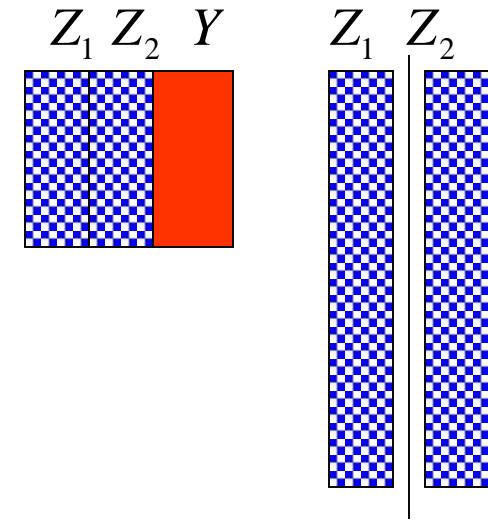
$n_{jk}$  = sample count,  $\bar{y}_{jk}$  = sample mean of  $Y$

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

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

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

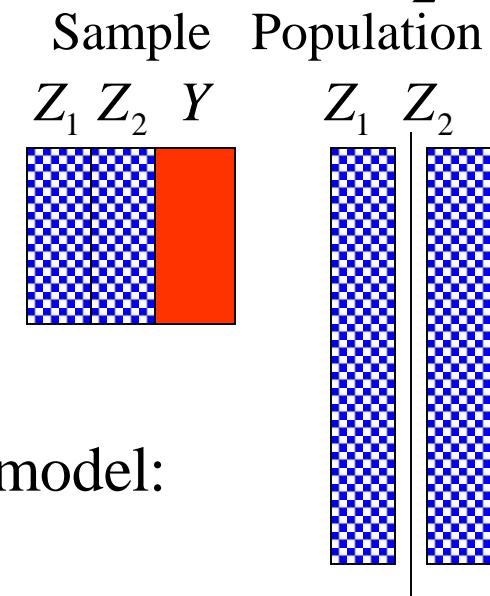
Sample   Population



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

## Model-based approach

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



What to do when  $n_{jk}$  is small?

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

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

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

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

otherwise shrinks towards additive model

Design: arbitrary collapsing, ad-hoc modification of weight

# Summary

- HT estimate is design-unbiased, but does not have good (design-based properties) when the “implied” underlying HT model is not a good fit to the data
- Bayes inference under a more flexible model relating  $Y$  to  $Z$  yields better design-based inferences
  - More efficient estimates
  - Better confidence coverage in moderate samples
- Unlike design-based inference, Bayes inference is not asymptotic, and can deliver good frequentist properties in small samples

# References

- Chen, Q., Elliott, M.R. & Little, R.J. (2010). Bayesian Penalized Spline Model-Based Estimation of the Finite Population Proportion for Probability-Proportional-to-Size Samples. *Survey Methodology*, 36, 23-34.
- Chen, Q., Elliott, M.R. & Little, R.J. (2012). Bayesian Inference for Finite Population Quantiles from Unequal Probability Samples. *Survey Methodology*, 38, 2, 203-214
- Ruppert, D. and Carroll, R.J. (2000). Spatially Adaptive Penalties for Spline Fitting. *Australia and New Zealand Journal of Statistics*, 42, 205–223.
- Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). Semiparametric Regression. Cambridge, UK: Cambridge University Press.
- Zheng, H. & Little, R.J. (2004). Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two-Stage Samples. *Survey Methodology*, 30, 2, 209-218.
- 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.

# Bayesian inference for sample surveys

Roderick Little and Trivellore Raghunathan

Module 9: Role of sampling weights in regression



# 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)

# Weighting in multiple regression

- Model-based: standard method of estimation is ordinary least squares (OLS)
  - OLS if the residual variance is constant
  - Or WLS, weighting by the inverse of the residual variance, if the residual variance is not constant:
$$y_i | x_i \sim N(\beta_0 + \beta^T x_i, \sigma^2 / u_i), w_i \propto u_i$$
- Design-based: WLS, weighting cases by inverse of probability of selection,  $w_i = 1 / \pi_i$
- These approaches to weighting are in general different, so which is right?
  - Much debated in the literature. See for example Brewer and Mellor (1973), DuMouchel and Duncan (1983)

# Key concepts

- Superpopulation parameters: parameters included in a superpopulation model (e.g. the regression coefficients in a multiple regression model)
- Finite population quantities: population quantities defined by fitting model to the whole population, by some specified method (e.g. ordinary least squares)
- Target model: a model that defines the target quantity (or quantities) of interest
- Working model: a model used for predicting non-sampled units in the population
  - Could be Bayesian

# Target model and working model 1

- Design: stratified sampling,  $Z = \text{strata}$

Target quantity:  $\bar{Y}$

Target model:  $y_i \sim_{\text{iid}} N(\mu, \sigma^2)$

( $\bar{Y}$  is estimate of  $\mu$  from fitting this model to population)

Working model:  $(y_i \mid z_i = j) \sim_{\text{iid}} N(\mu_j, \sigma^2)$

(Prediction model needs to condition on strata)

Resulting estimate of  $\bar{Y}$  weights cases by their sampling weights

# Target model and working model 2

- Design: stratified sampling,  $Z = \text{strata}$

Target quantity:  $\bar{Y}_j = \text{mean of } Y \text{ in stratum } j$

Target = working model:  $(y_i \mid z_i = j) \sim N(\mu_j, \sigma^2)$

Estimate of  $\bar{Y}_j = \bar{y}_j$ , sample mean in stratum  $j$

(Not weighted since weights are constant within strata)

# Weighting in Regression

- 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 regression coefficients

# Regression with sample weights

- Target model:

$$y_i | x_i \sim N(\beta_0 + \beta^T x_i, \sigma^2 / u_i), u_i \text{ known (constant for OLS)}$$

- Target parameter:  $\beta$
- 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

$$E(y_i | x_i, z_i) = \gamma_0 + \gamma_1 x_i + \gamma_2 z_i \text{ then}$$

$$E(y_i | x_i) = \gamma_0 + \gamma_1 x_i + \gamma_2 E(z_i | x_i, \psi), \text{ etc.}$$

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  $\beta$

# 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  $\beta$ . (Little 2004, Example 11)

# Regression with sample weights

(b)  $z_i$  not a part of  $x_i$

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.

Dumouchel and Duncan (1983): provides a test of equality of coefficients from weighted and unweighted least squares  
Uses weighted least squares as a specification check on the target model

# 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  $E(y_i | x_i, z_{i2}) = \gamma_0 + \gamma_1 x_i + \gamma_2 z_{i2}$  then

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

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  $\beta$

# 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.

# Summary

- Calibrated Bayes approach
  - Sampling weights as predictors
  - Flexible models for relationship between outcomes and sampling weights: eg penalized spline or propensity model
  - For regression, impute nonsampled cases using a target model that includes sampling weights (or stratum) as predictor; then fit target model to fitted population
- Next: cluster, multistage sampling

# Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 10: Cluster Sample Design



# Two stage sampling

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

# Ex 4. Two-stage samples

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

$K_i$  = 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

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

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

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

- Prior distribution

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

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

# Estimand of interest and inference strategy

- The population mean can be decomposed as

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

- Posterior mean given  $Y_{\text{inc}}$

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

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

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

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

# Posterior Variance

- Posterior variance can be easily computed

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

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

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

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

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

# Inference with unknown $\sigma$ and $\tau$

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

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

where  $b$  and  $\nu$  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

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

# 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.

# Non-parametric Bayes

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

# Summary

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

# Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 11: Hierarchical Models for Cluster  
Sample Designs



# Models for Cluster Sample Design

- Hierarchical models (two-stage)

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

$$j = 1, 2, \dots, K_i$$

$$i = 1, 2, \dots, C$$

$$prior : \pi(\mu, \sigma_\alpha, \sigma_\varepsilon)$$

$$Data : \{y_{ij}, j = 1, 2, \dots, k_i; i = 1, 2, \dots, c\}$$

*Draws :*

$$(1) \mu, \sigma_\alpha, \sigma_\varepsilon, \alpha_i, i = 1, 2, \dots, c$$

$$(2) y_{ij} \sim N(\mu + \alpha_i, \sigma_\varepsilon^2),$$

$$j = k_i + 1, k_i + 2, \dots, K_i$$

$$i = 1, 2, \dots, c$$

$$(3) \alpha_i \sim N(0, \sigma_\alpha^2), i = c + 1, \dots, C$$

$$Y_{ij} \sim N(\mu + \alpha_i, \sigma_\varepsilon^2), j = 1, 2, \dots, K_i$$

# Implementation

- Use any standard Bayesian software package (Winbugs, Openbugs, STAN, JAGS, PROC MCMC, PROC MIXED etc) to obtain the draws in Step (1)
- Step (2) involves drawing normal random variables using the parameters from Step 1
- Step (3) Involves drawing normal random variables using the parameters from Step 1
- Once the population is filled-in compute the finite population quantity of interest

# Incorporating other design features

- Include weights as covariates

$$Y_{ij} = \mu + \alpha_i + \beta f(w_i) + g(w_i)\varepsilon_{ij}$$

$f()$  and  $g()$  *known functions*

- Stratification
  - Analyze each stratum separately and fill-in the non-sampled values
  - If the number of clusters within a stratum is small then treat the stratum specific parameters as random effects

# Multistage designs

- Typically more than 2-stages are involved in selecting the elements from the population
- Example:
  - Goal: A national probability sample of adults with representation from every State
    - Draw a sample of Counties from every State
    - Draw a sample of census tracts within the sampled counties
    - Draw a sample of block groups within the sampled tracts
    - Draw a sample of blocks within the sampled block groups
    - Draw a sample of households within the sample blocks
    - Draw an adult from the sampled households
- Often, for confidentiality reasons, only the first stage (counties) may be released.

# Models for three stage design

- Nested Hierarchical models

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

$i$ : Stage 1

$j$ : Stage 2 nested within Stage 1

$k$ : Individuals

$k = 1, 2, \dots, N_{ij}$

$j = 1, 2, \dots, S_i$

$i = 1, 2, \dots, P$

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

$$\beta_{j(i)} \sim N(0, \sigma_\beta^2)$$

$$\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$$

$$prior: \pi(\mu, \sigma_\alpha, \sigma_\beta, \sigma_\varepsilon)$$

- (1) Draw parameters
- (2) Fill in non-sampled elements in the sampled and nonsampled first and second stage units

# Binary Outcomes

- Mixed Effects Logistic Model

$$Y_{ijk} \sim Ber(\theta_{ijk})$$

$$\theta_{ijk} = \Pr(Y_{ijk} = 1)$$

$$logit(\theta_{ijk}) = \mu + \alpha_i + \beta_{j(i)}$$

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

$$\beta_{j(i)} \sim N(0, \sigma_\beta^2)$$

$$prior: \pi(\mu, \sigma_\alpha, \sigma_\beta)$$

# Poisson Outcomes

- Count type variables

$$Y_{ijk} \sim Poisson(\theta_{ijk})$$

$$\log(\theta_{ijk}) = \mu + \alpha_i + \beta_{j(i)}$$

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

$$\beta_{j(i)} \sim N(0, \sigma_\beta^2)$$

$$prior: \pi(\mu, \sigma_\alpha, \sigma_\beta)$$

# Remarks

- All Bayesian analysis of survey data involves 3 step process:
  1. Generate draws of the parameters from their posterior distribution (this is a traditional Bayesian analysis step)
  2. Fill-in the non-sampled values conditional the draws of the parameters (just use the model, treating parameters as known)
  3. Compute the population quantity of interest
- Step 2 involves book keeping of whether the unit is sampled or not, and to use appropriate draws of the random effects

# Remarks

- Assumed that cluster sizes for the sampled and non-sampled clusters are known
- This may not be true. The following approximation may be used
- For sampled clusters define  $K_i = k_i / f$  where  $f$  is some small number , for example, 0.01 or 0.005
- For non-sampled clusters, bootstrap from the sampled cluster sizes  $K_1, K_2, \dots, K_c$

# Bayesian Inference for Sample Surveys

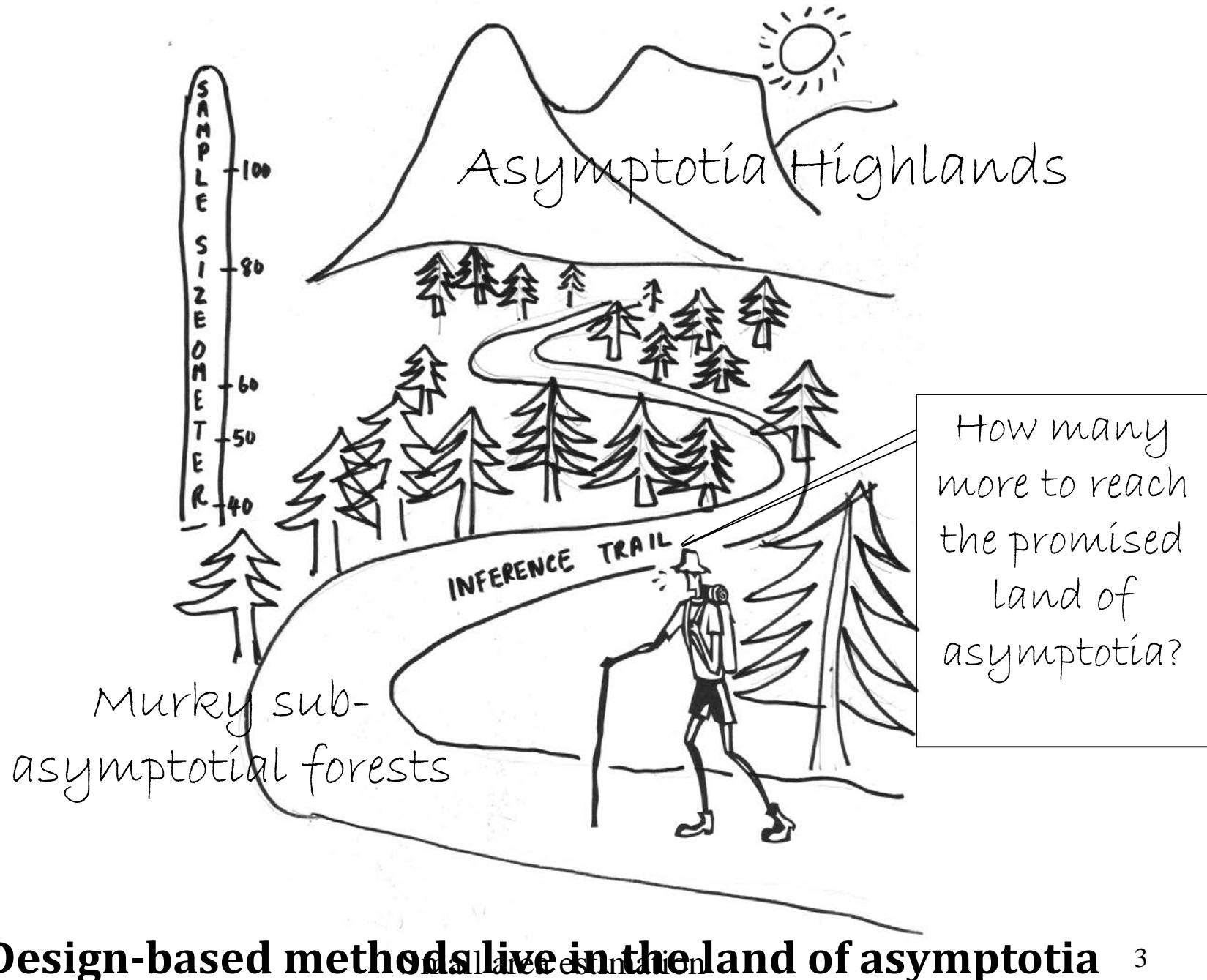
## Module 12: Small Area Estimation

Roderick Little and T. Rathunathan



# 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
- Theory is basically asymptotic -- limited tools for small samples, e.g. small area estimation



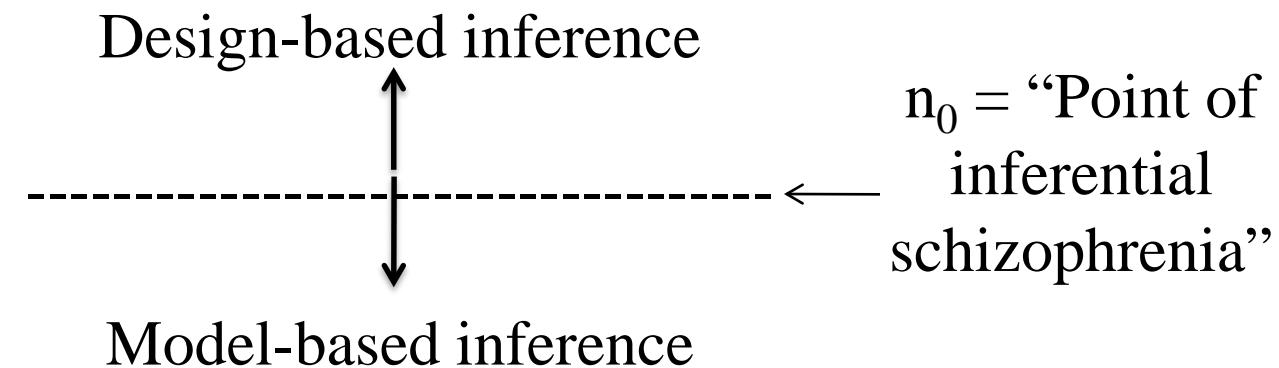
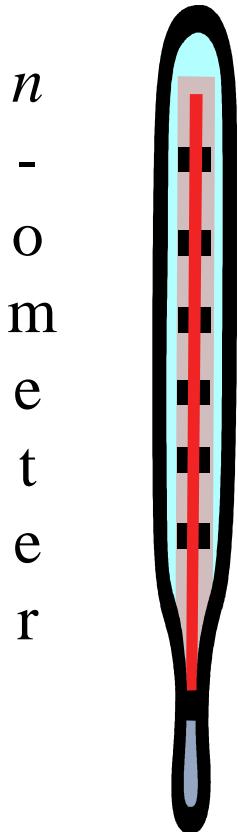
# The current “status quo” -- design-model compromise

- Design-based for large samples, descriptive statistics
  - But may be *model assisted*, e.g. regression calibration:

$$\hat{T}_{\text{REG}} = \sum_{i=1}^N \hat{y}_i + \sum_{i=1}^N I_i(y_i - \hat{y}_i) / \pi_i, \hat{y}_i = \text{model prediction}$$

- model estimates adjusted to protect against misspecification, (e.g. Särndal, Swensson and Wretman 1992).
  - Can incorporate auxiliary information, but does not borrow strength for small areas
- Model-based for small area estimation, nonresponse, time series,...
- Attempts to capitalize on best features of both paradigms... but ... at the expense of “inferential schizophrenia” (Little 2012)?

# Example: when is an area “small”?



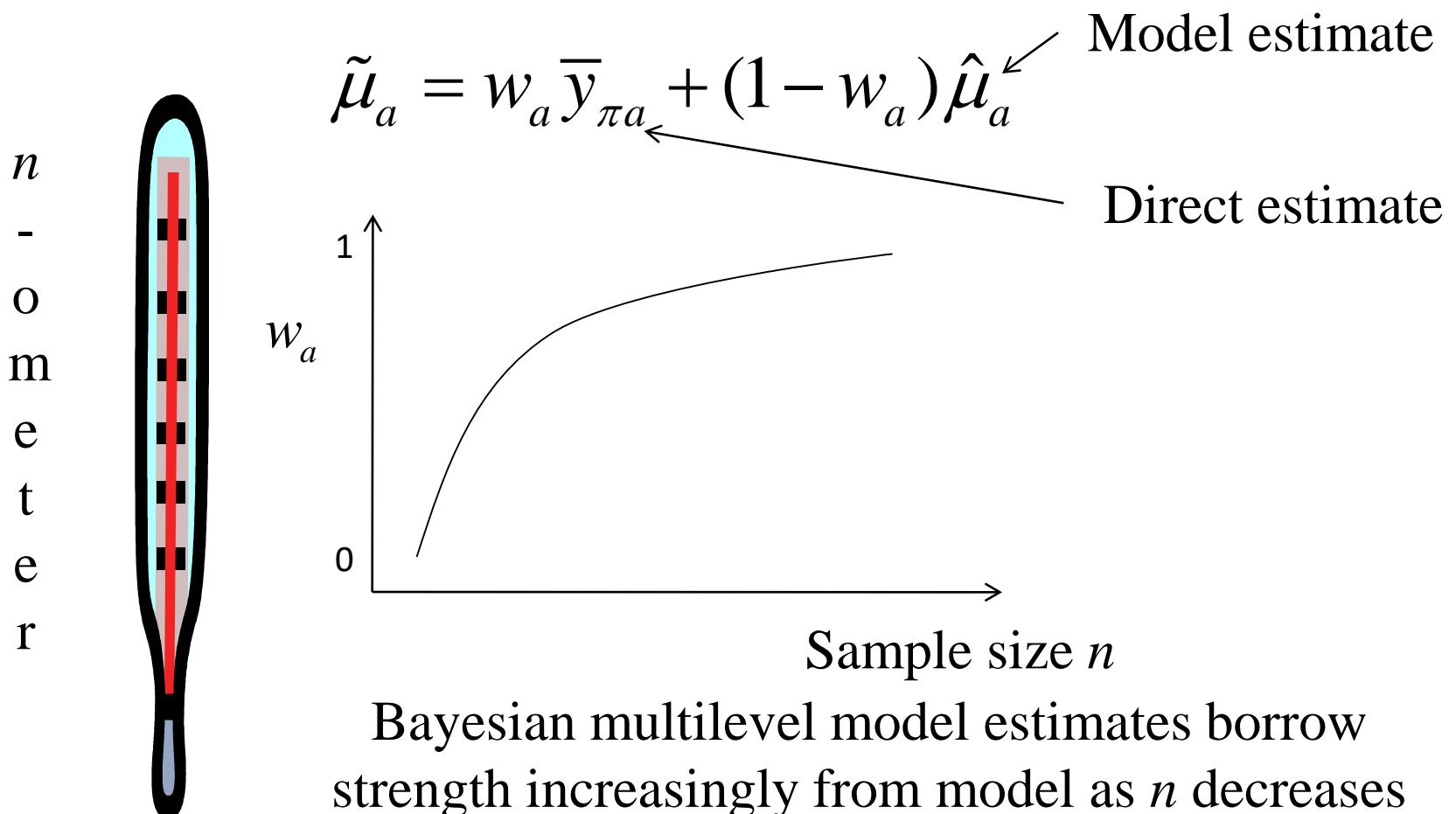
How do I choose  $n_0$ ?

If  $n_0 = 35$ , should my entire statistical philosophy and inference be different when  $n=34$  and  $n=36$ ?

$n=36$ , CI: [ ] (wider since based on direct estimate)

$n=34$ , CI: [ ] (narrower since based on model)

# Multilevel (hierarchical Bayes) models



# A hierarchical Bayes model for small areas

- Fixed-effects models have distinct parameters (means, variances) for small areas, e.g.

$$y_{ai} \mid \mu_a, \sigma_a^2 \sim N(\mu_a, \sigma_a^2), \text{ for unit } i \text{ in area } a$$

- Hierarchical Bayes models assign distributions to the parameters for each area, e.g.

$$(y_{ai} \mid \mu_a, \sigma_a^2, \beta, \tau^2) \sim N(\mu_a, \sigma_a^2);$$

$$\Rightarrow (\bar{y}_a \mid \mu_a, \sigma_a^2, \beta, \tau^2) \sim N(\mu_a, \sigma_a^2 / n_a)$$

$$(\mu_a \mid \mu_a, \sigma_a^2, \beta, \tau^2) \sim N(\beta z_a, \tau^2)$$

$z_a$  is a vector of covariates for area  $a$ , including constant term

# Hierarchical Bayes Models for small areas

$$(\bar{y}_a \mid \mu_a) \sim N(\mu_a, \sigma_a^2 / n_a)$$

$$(\mu_a) \sim N(\beta z_a, \tau^2)$$

$$p(\bar{y}_a, \mu_a) \propto \exp -\frac{1}{2} \left[ A(\bar{y}_a - \mu_a)^2 + B(\mu_a - \beta z_a)^2 \right]$$

$$\left( A = n_a / \sigma_a^2, B = 1 / \tau^2 \right)$$

$$p(\bar{y}_a, \mu_a) \propto \exp -\frac{1}{2}(A+B) \left[ (\mu_a - \frac{A\bar{y}_a + B\beta z_a}{A+B})^2 \right]$$

$$\text{Hence } E(\mu_a \mid \bar{y}_a, \beta) = \frac{A\bar{y}_a + B\beta z_a}{A+B} = w_a \bar{y}_a + (1-w_a) \beta z_a$$

$$\text{where weight on } \bar{y}_a \text{ is } w_a = A / (A+B) = \frac{n_a / \sigma_a^2}{n_a / \sigma_a^2 + 1 / \tau^2}$$

Proper prior ( $\tau^2 < \infty$ ) on  $\mu_a$  moves the direct area estimate  $\bar{y}_a$  towards the model prediction  $\beta z_a$ ;  $w_a$  increases with  $n_a$

Integrating over posterior of  $\beta$  replaces  $\beta$  by its posterior mean  
Small area estimation

# Hierarchical Bayes Models for small areas

Treatment of variances  $\sigma^2, \tau^2$ :

Empirical Bayes: replaces them by estimates  $\hat{\sigma}^2, \hat{\tau}^2$

(Maximum likelihood, or the simpler method of moments)

Bayes: assigns them dispersed prior distributions and  
integrates over their posterior distribution

(note that  $\tau^2$  cannot be assigned the Jeffreys' prior  $p(\tau^2) \propto 1/\tau^2$ )

Bayes approach is better, particularly if  $\hat{\tau}^2 = 0$

# A Basic Beta/Binomial small area model for binary outcomes

$n_a$  = count in area  $a$

$m_a$  = count with  $y = 1$  in area  $a$

$m_a \mid p_a \sim \text{Bin}(n_a; p_a)$

$p_a \sim \text{Beta}(\alpha, \beta)$  (conjugate prior distribution)

Prior mean of  $p_a$ :  $E(p_a) = \frac{\alpha}{\alpha + \beta}$

# A Basic Beta/Binomial small area model for binary outcomes

Posterior distribution of  $p_a$  is Beta:

$$(p_a | n_a, m_a) \sim \text{Beta}(\alpha + m_a, \beta + n_a - m_a)$$

Posterior mean of  $p_a$ :

$$E(p_a | n_a, m_a) = \frac{\alpha + m_a}{\alpha + \beta + n_a} = w_a \frac{m_a}{n_a} + (1 - w_a) \frac{\alpha}{\alpha + \beta}$$

where weight on sample proportion is  $w_a = \frac{n_a}{n_a + \alpha + \beta}$

"Prior adds  $\alpha$  successes and  $\beta$  failures to data"

# Applications

- U.S. Census Bureau SAIPE (Small Area Income and Poverty Estimates) Program
- Voting Rights Act special tabulation
- The American Community Survey (ACS) and the “standard error error”

# Example 1: SAIPE project

- Objective: Estimates of poverty for various age groups and median household income for all *states*, *counties*, and *school districts* in the U.S.
- Problem: Direct survey estimates (from Current Population Survey (CPS) or, later, American Community Survey (ACS) are too unreliable for many areas
  - CPS sample small for most states; no sample in  $\approx 2/3$  counties
  - ACS (single year) sample small for many counties and most school districts.
- Solution: Use small area model to integrate survey data with data from admin records (IRS, SNAP program) and previous census long form.

# Fay-Herriot (1979) Model

## (Hierarchical Bayesian Formulation)

$$y_i | \theta_i, v_i \sim N(\theta_i, v_i)$$

$$\theta_i | \beta, \sigma^2 \sim N(x_i' \beta, \sigma^2)$$

- $y_i$  = direct survey estimate of population quantity  $\theta_i$  for area  $i$
- $v_i$  = sampling variance of  $y_i$  (assumed known)
- $x_i$  = vector of regression variables for area  $i$
- $\beta$  = vector of regression parameters
- $\sigma^2$  = variance of small area random effects

# Example: State 5-17 poverty rate model

- Direct survey estimates  $y_i$  originally from CPS, but since 2005 from ACS
- Regression variables in  $x_i$  include a constant term and, for each state
  - Pseudo-poverty rate for children from tax return data
  - Tax “nonfiler rate”
  - SNAP (food stamp) participation rate
  - Previous census estimated state 5-17 poverty rate, or residuals from regressing previous census estimates on other elements of  $x_i$  for the census year.
  - Recent work explicitly Bayesian to propagate error in variance components

# Posterior Variances from State Model for 2004 CPS 5-17 Poverty Rates

Results for four states

State	$n_i$	$v_i$	$\text{Var}(Y_i \text{data})$	approx. wt. on $y_i$ in $E(Y_i \text{data})$
CA	5,834	1.1	0.8	.61
NC	1,274	4.6	2.0	.28
IN	904	8.1	2.0	.18
MS	755	12.0	3.9	.13



# Example 2: Voting Rights Tabulations

- Section 203 Language Provisions of the Voting Rights Act
- Determines counties and townships required to provide language assistance at the polls
- Determinations are based in part on the following “more than 5%” provision:
  - ... More than 5 percent of voting age citizens of political district are members of a single language minority and are LEP.

# Voting Rights Tabulations

- Previously used direct estimates from Long Form Decennial Census Data
- Use ACS 2005-2009 and 2010 Census data to produce a Federal Register Notice by mid-summer 2011
- Direct estimates for some districts are based on small ACS sample and hence have unacceptably high variance
- E.g. let  $P$  be proportion of voting age citizens in political district who are members of a single language minority and are LEP
- Suppose ACS was a simple random sample, a direct estimate of  $P$  is the sample proportion  $m/n$ 
  - District A with  $n=105, m=5, m/n < 0.05$
  - District B with  $n=105, m=6, m/n > 0.05$
  - Direct ACS estimation is more complex, but same idea applies

# Voting Rights Tabulations

- Alternative approach to the “more than 5%” provision:
- Build a district level regression model to predict  $P$  based on variables in the ACS
- Classify districts into classes with similar predicted  $P$  based on the model [predictive mean stratification]
- Within classes, apply a hierarchical random-effects model that pulls the direct ACS estimate of  $P$  towards the average  $P$  for districts in that class
- Compare HRE model estimate with 5% for this aspect of the determination
- Rationale: increased precision of HRE estimates in small samples increases the probability of getting the determination right, particularly in small districts

# Example 3: ACS tabulations

- American Fact Finder gives users access to a dazzling array of ACS tables, for small areas
- The move to make more data available is highly commendable, but the methodology remains design-based and assumes large samples.
- Need for methods appropriate for small samples...

B01001A. SEX BY AGE (WHITE ALONE) -  
Universe: **WHITE ALONE POPULATION**  
Data Set: 2005-2009 American  
Community Survey 5-Year Estimates  
Survey: American Community Survey

90% CI = Estimate  
+/- Margin of Error

Northfield township, Washtenaw County, Michigan		
	Estimate	Margin of Error
	8,062	+/-210
Male:	4,164	+/-239
Under 5 years	321	+/-108
5 to 9 years	239	+/-90
10 to 14 years	342	+/-171
15 to 17 years	151	+/-57
18 and 19 years	14	+/-21
20 to 24 years	332	+/-175
...	...	...

$90\% \text{ CI} = \text{Estimate}$   
 $+/- \text{ Margin of Error}$

Oops! →

Northfield township, Washtenaw County, Michigan		
	Estimate	Margin of Error
	8,062	+/-210
Male:	4,164	+/-239
Under 5 years	321	+/-108
5 to 9 years	239	+/-90
10 to 14 years	342	+/-171
15 to 17 years	151	+/-57
18 and 19 years	14	+/-21
20 to 24 years	332	+/-175
...	...	...

# Example: ACS tabulations

B01001B. SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE) - Universe: BLACK OR  
**AFRICAN AMERICAN ALONE POPULATION**

Data Set: 2005-2009 American Community Survey 5-Year Estimates  
Survey: American Community Survey

- 90% CI = Estimate  
+/- Margin of Error
- (1) Margin of Error gives a rough idea of uncertainty, but ...
  - (2) Intervals often contains negative values
  - (3) Truncated to be positive, CI still does not have stated coverage
  - (4) Simple fixes for SRS, less simple for complex designs

Northfield township, Washtenaw County, Michigan		
	Estimate	Margin of Error
Total:	43	+/-41
Male:	28	+/-32
Under 5 years	0	+/-109
5 to 9 years	0	+/-109
10 to 14 years	0	+/-109
15 to 17 years	0	+/-109
...	...	...
35 to 44 years	0	+/-109
45 to 54 years	28	+/-32
55 to 64 years	0	+/-109

Imagine a  
better table...

90% CI = ( LL, UL)  
LL, UL based on  
Bayesian model

...but more  
complex... and  
billions of cells!

Northfield township, Washtenaw County, Michigan			
	LL	Estimate	UL
Total:	0	?	?
Male:	0	?	?
Under 5 years	0	?	?
5 to 9 years	0	?	?
10 to 14 years	0	?	?
15 to 17 years	0	?	?
...	0	?	?
35 to 44 years	0	?	?
45 to 54 years	0	?	?
55 to 64 years	0	?	?

# American Community Survey

- US Census Bureau is making available thousands of ACS tables, with millions of cells
- A high fraction of these estimates are based on very little data, and hence are very noisy
  - Many people want information, not data, so ACS should produce information products, as well as data products
  - When noise swamps the signal, the information content is buried
  - Data products are highly constrained by confidentiality requirements, leading to incompleteness

# The Statistical Problem

- The ACS philosophy is essentially to produce “direct” (“design-based”) estimates, together with margins of error
- This works fine with large samples, but most of the ACS estimates are based on small samples
  - The estimates are often too noisy to be useful
  - The confidence intervals derived from the estimates and margins of error are known to be of poor quality, violating statistical standards
    - Intervals include proportions outside the range (0,1)
    - Intervals do not have nominal coverage

# The “standard error” error

- ACS reports estimates and margins of error that yield asymptotic 90% confidence intervals
- But in small samples, the implied confidence intervals do not have the stated coverage; so
- Calibrated Bayes: Seek to replaces estimates and margins of error by posterior means and 5% to 95% credibility intervals that have the approximately the nominal coverage
  - A non-Bayesian can interpret the posterior means as estimates, and the 90% credibility intervals as 90% confidence intervals.

# Pragmatic “pseudo-Bayes” approach

A fully Bayesian hierarchical model for proportions is feasible but beyond current Bureau of Census capabilities

Tom Louis suggested a simple “Bayes-like” approach, which “gets the Calibrated Bayes foot in the door” (my words)

# Pragmatic “pseudo-Bayes” approach

- A. Compute design-based estimate of proportion and standard error using existing design-based methods
- B. Pretend data are binomial with number of successes  $m_a^*$  and sample size  $n_a^*$  that lead to the estimates in A.
- C. Compute Beta posterior distribution with noninformative prior (e.g. uniform or Jeffreys)
- D. Compute 90% posterior credibility interval based on this Beta posterior (reflects asymmetry, always between 0 and 1)

Simple to implement and easily beats standard Wald-type confidence intervals in simulations (Franco, Little, Louis and Slud 2014)

# Approximate Beta posterior distribution for complex designs

SRS: Posterior distribution of  $p_a$  is Beta:

$$(p_a | n_a, m_a) \sim \text{Beta}(\alpha + m_a, \beta + n_a - m_a)$$

Complex Design: estimate  $\hat{p}_a$ , SE  $\hat{s}_a$  using a design-based approach

$$\hat{s}_a^2 = \frac{\hat{p}_a(1 - \hat{p}_a)}{n_a^*}, \quad n_a^* = \text{effective sample size}; \quad n_a / n_a^* = \text{design effect}$$

Hence define effective ss and count:  $n_a^* = \frac{\hat{p}_a(1 - \hat{p}_a)}{\hat{s}_a^2}$ ,  $m_a^* = n_a^* \hat{p}_a$ ,

Approximate posterior distribution as

$$(p_a | n_a^*, m_a^*) \sim \text{Beta}(\alpha + m_a^*, \beta + n_a^* - m_a^*)$$

# References

- Franco, C., Little, R., Louis, T. and Slud, E. (2014). Coverage Properties of Confidence Intervals for Proportions in Complex Sample Surveys . *ASA Proc Survey Research Methods Section*.
- Joyce, P.M., Malec, D., Little, R.J., Gilary, A., Navarro, A. and Asiala, M.E. (2014). Statistical Modeling Methodology for the Voting Rights Act Section 203 Language Assistance Determinations. *JASA*, 109, 36-47.
- Little, R.J. (2012). Calibrated Bayes: an alternative inferential paradigm for official statistics (with discussion and rejoinder). *JOS*, 28, 3, 309-372.
- Särndal, C.-E., Swensson, B. & Wretman, J.H. (1992), *Model Assisted Survey Sampling*, Springer Verlag: New York.

# Bayesian Inference for Surveys

Rod Little and Trivellore Raghunathan  
Approximations for Computations

# Two Stages

- Model for prediction
  - Joint distribution for the fine population values
    - Exchangeable on the index used to label units in the population
    - Introduce parameters, conditional on which , the units are independent
    - Prior on the parameters induces a joint distribution
- Analysis
  - Summary of the population values
  - Analytical models fitted to the entire population (linear regression). These also can be viewed as summary of the population values
- Model for prediction can lead to approximation of the summary of the population values

# Examples of Approximation

- Stratified population

$$N_h, h = 1, 2, \dots, H$$

$$Y_{ih}, i = 1, 2, \dots, N_h$$

$$\bar{Y} = \sum_h \sum_i Y_{ih} / N, N = \sum_h N_h$$

- Prediction or Population Model

$$\prod_h P(Y_{i1}, Y_{i2}, \dots, Y_{iN_h})$$

- Exchangeable within-stratum and independence across-stratum

$$\prod_h \left[ \int \left( \prod_{i=1}^{N_h} f(Y_{ih} | \theta_h) \right) \pi(\theta_h) d\theta_h \right]$$

- Estimand

$$Q(Y) = \Pr(Y \geq c)$$

# Analysis

- Conceptually, the straightforward approach is to generate several draws of nonsampled values, computed the proportion of the sampled and drawn values exceeding the constant  $c$ .
- Approximation
  - Prediction Model

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

$$\pi(\mu_h, \sigma_h) \propto \sigma_h^{-1}$$

- Model based approximation

$$\Pr(Y \geq c | \mu_h, \sigma_h, h=1, 2, \dots, H) \\ = \sum_h W_h \{1 - \Phi((c - \mu_h) / \sigma_h)\}$$

$$W_h = N_h / N$$

$$N_h \gg n_h$$

# Draws

- A standard Bayesian analysis from each stratum to obtain draws of
$$(\mu_h, \sigma_h) \mid y_{ih}, i = 1, 2, \dots, n_h$$
- Compute the approximation given on the previous slide to obtain approximate draws of
$$\Pr(Y \geq c)$$

- Though assumed normality, the approach works for any other parametric distribution with a known functional form
  - t-distribution
  - Chi-square or gamma
  - Beta
- Transformation to normality is another possible approach

$$\Pr(Y \geq c) \Rightarrow \Pr(g(Y) \geq g(c))$$

$$g(y) \sim Normal$$

# Imputation Based Approaches

- Suppose that any of the standard parametric forms don't fit the data well and none of the transformation works well
- Tukey introduced a general class of models that can accommodate a wide variety of distributions

$$Y = \mu + \sigma \times Z \times \frac{\exp(gZ) - 1}{gZ} \times \exp(hZ^2 / 2)$$

- Parameter

$\mu$  : *Location*

$\sigma$  : *Scale*

$g$  : *Skewness*

$h$  : *Kurtosis(tails)*

It is very difficult to write the density function of  $Y$  but it is easy to draw from.

Estimation of the parameters is easy using the percentiles

# Estimation

- Location parameter: Sample Median

$$p = \Pr(Y \leq Q_p)$$

$$A_p = \frac{Q_{1-p} - Q_{0.5}}{Q_{0.5} - Q_p} = \exp(-gZ_p)$$

$$B_p = \frac{g(Q_{1-p} - Q_{0.5})}{\exp(-gZ_p) - 1} = \sigma \exp(hZ_p^2 / 2)$$

# Estimation (contd.)

- Choose various values of  $p$  and obtain sample percentiles

*Regress  $\log(A_p)$  on  $-Z_p$  to estimate  $g$*

*Regress  $\log(B_p)$  on  $Z_p^2$  to estimate  $\sigma$  and  $h$*

# Imputation approach implemented in IVWare (Version 0.3 to be soon released)

- Draw an Approximate Bayesian bootstrap sample  
(Draw a bootstrap sample and again draw a bootstrap sample from this bootstrap sample)
- Estimate the percentiles and estimate the parameters
- Impute the unobserved values by drawing values of the standard normal random variables and using the estimated parameters from the previous step
- Stratum-specific imputations

# Implementation

- For each stratum create the missing values for the unobserved subjects
- Use the “gh” command in IVEware and “by Stratum” command
- Create multiple imputations
- Compute the population quantity of interest

# BBDESIGN

- IVWare (version 0.3) also implements a more general non-parametric approach for drawing the unobserved values in the population using Bootstrap and Polya urn model (that we discussed previously)
- This approach may be more preferable if good parametric models could not be found to fit the data well

# Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 14: missing data 1 -- overview



# 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

# Finite population inference

- Bayesian modeling takes a predictive perspective on statistical inference – predict the non-sampled values
- Inference about parameters is intermediate step in predictive superpopulation model inference about finite population parameters
- Bayesian approach extends naturally to handle missing data
  - Predict nonsampled and missing values
- Missingness not under control of sampler, so missing not at random mechanisms complicate inferences

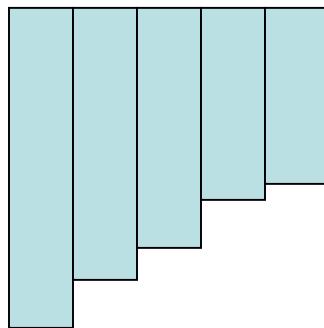
# Likelihood methods with missing data

- Likelihood methods do not require rectangular data, hence apply directly to missing-data problems
- Statistical model + incomplete data  $\square$  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

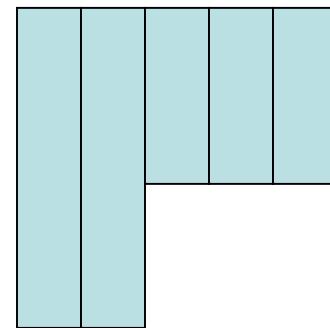
# Patterns of Missing Data 1

- special patterns

monotone

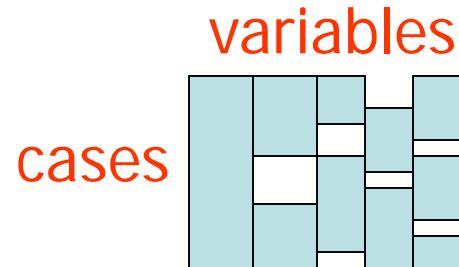


unit nonresponse



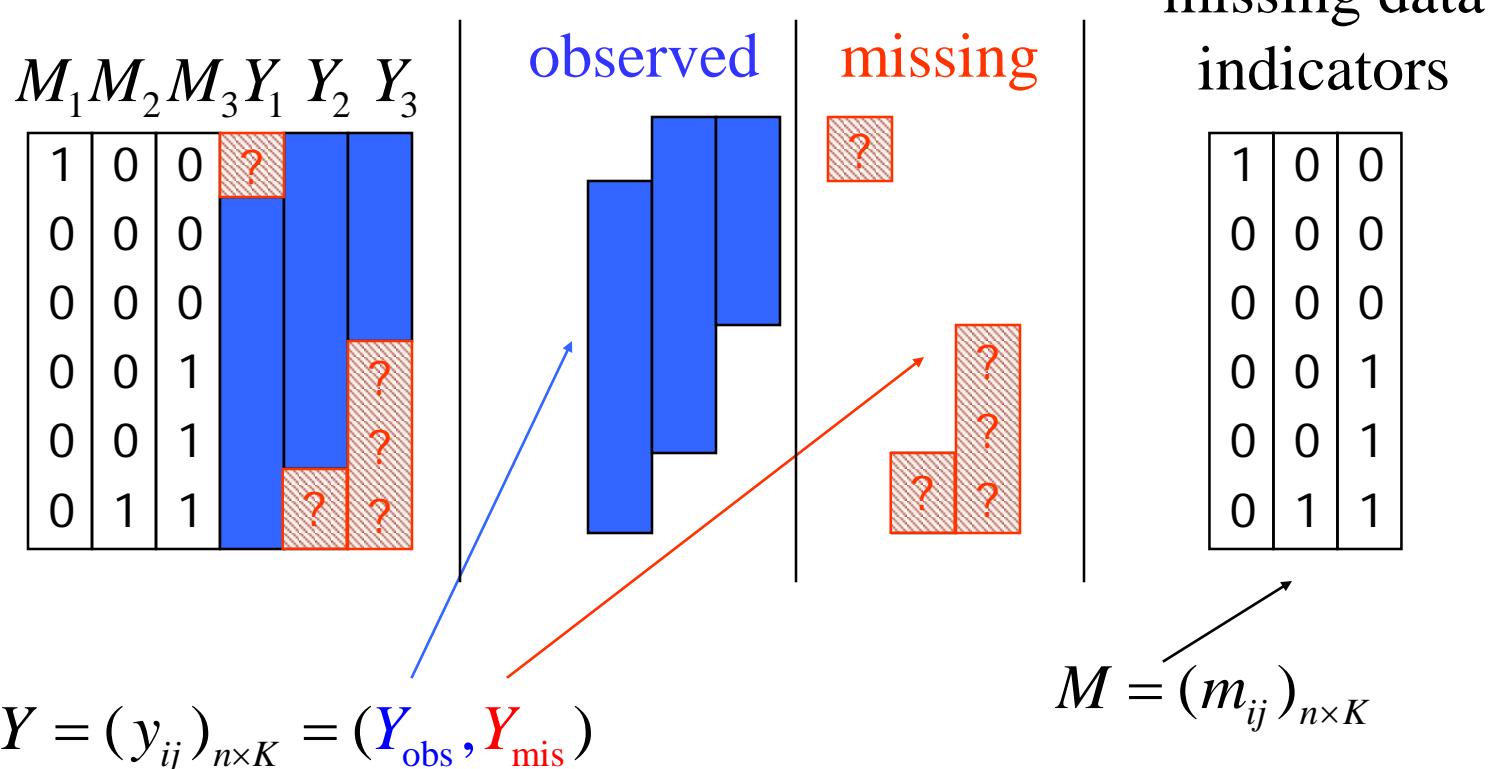
# Patterns of Missing Data 2

- Item nonresponse: general pattern



Bayes can handle general patterns with relative ease

# The Observed Data



# Likelihood methods with missing data

- Statistical model + incomplete data  $\square$  Likelihood
- Statistical models needed for:
  - data without missing values
  - missing-data mechanism
- Model for mechanism not needed if it is ignorable – missing at random (MAR) is the key condition
- With likelihood, proceed as before:
  - ML estimates, large sample standard errors
  - Bayes posterior distribution
  - Little and Rubin (2002, chapter 6)

# Bivariate Monotone Data: Model for $Y$ and $M$

$$f(Y, M | \theta, \psi) = f(Y | \theta) \times f(M | Y, \psi)$$

Complete-data model      model for mechanism

Example: bivariate normal monotone data

complete-data model:

$$(y_{i1}, y_{i2}) \sim_{iid} N_2(\mu, \Sigma)$$

model for mechanism:

$$(m_{i2} | y_{i1}, y_{i2}) \sim_{ind} Bern[\Phi(\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2})]$$

$M_1$	$M_2$	$Y_1$	$Y_2$
0	0		
0	0		
0	0		
0	1		?
0	1		?

$\Phi$  = Normal cumulative distribution function

# Two likelihoods

- *Full likelihood* - involves model for  $M$

$$f(Y_{\text{obs}}, M \mid \theta, \psi) = \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta) f(M \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) dY_{\text{mis}}$$
$$\Rightarrow L_{\text{full}}(\theta, \psi \mid Y_{\text{obs}}, M) = \text{const} \times f(Y_{\text{obs}}, M \mid \theta, \psi)$$

- Likelihood *ignoring the missing-data mechanism*  $M$ 
  - simpler since it does not involve model for  $M$

$$f(Y_{\text{obs}} \mid \theta) = \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta) dY_{\text{mis}}$$
$$\Rightarrow L_{\text{ign}}(\theta \mid Y_{\text{obs}}) = \text{const} \times f(Y_{\text{obs}} \mid \theta)$$

# Ignoring the missing-data mechanism

- Note that if:

$$L_{\text{full}}(\theta, \psi | Y_{\text{obs}}, M) = L(\psi | M, Y_{\text{obs}}) \times L_{\text{ign}}(\theta | Y_{\text{obs}})$$

where  $L(\psi | M, Y_{\text{obs}})$  does not depend on  $\theta$

then inference about  $\theta$  can be based on  $L_{\text{ign}}(\theta | Y_{\text{obs}})$

- The missing-data mechanism is then called *ignorable* for likelihood inference

# Ignoring the md mechanism continued

- Rubin (1976) showed that sufficient conditions for ignoring the missing-data mechanism are:

(A) Missing at Random (MAR):

$$f(M | Y_{\text{obs}}, \textcolor{red}{Y}_{\text{mis}}, \psi) = f(M | Y_{\text{obs}}, \psi) \text{ for all } \textcolor{red}{Y}_{\text{mis}}$$

(B) Distinctness:

$\theta$  and  $\psi$  have distinct parameter spaces

(Bayes: priors distributions are independent)

- If MAR holds but not distinctness, ML based on ignorable likelihood is valid but not fully efficient, so MAR is the key condition
- For frequentist inference, need MAR for all  $M, \textcolor{red}{Y}_{\text{mis}}$   
(Everywhere MAR, see Seaman et al. 2013)

# Two posterior distributions

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

Prior dn    Complete-data model    model for mechanism

- Full posterior distribution - involves model for  $M$

$$p_{\text{full}}(\theta, \psi | Y_{\text{obs}}, M) \propto \pi(\theta, \psi) \times f(Y_{\text{obs}}, M | \theta, \psi)$$
$$f(Y_{\text{obs}}, M | \theta, \psi) = \int f(Y_{\text{obs}}, Y_{\text{mis}} | \theta) f(M | Y_{\text{obs}}, Y_{\text{mis}}, \psi) dY_{\text{mis}}$$

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

$$p_{\text{ign}}(\theta | Y_{\text{obs}}) \propto \pi(\theta) \times f(Y_{\text{obs}} | \theta)$$
$$f(Y_{\text{obs}} | \theta) = \int f(Y_{\text{obs}}, Y_{\text{mis}} | \theta) dY_{\text{mis}}$$

# Ignoring the md mechanism continued

- Sufficient conditions for ignoring the missing-data mechanism and basing inference on  $p_{\text{ign}}(\theta | Y_{\text{obs}})$  are:
- MAR:  $f(M | Y_{\text{obs}}, \textcolor{red}{Y}_{\text{mis}}, \psi) = f(M | Y_{\text{obs}}, \psi)$  for all  $\textcolor{red}{Y}_{\text{mis}}$
- 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
- Missing Not at Random (MNAR) – missingness depends on missing data – harder problem

# Remarks

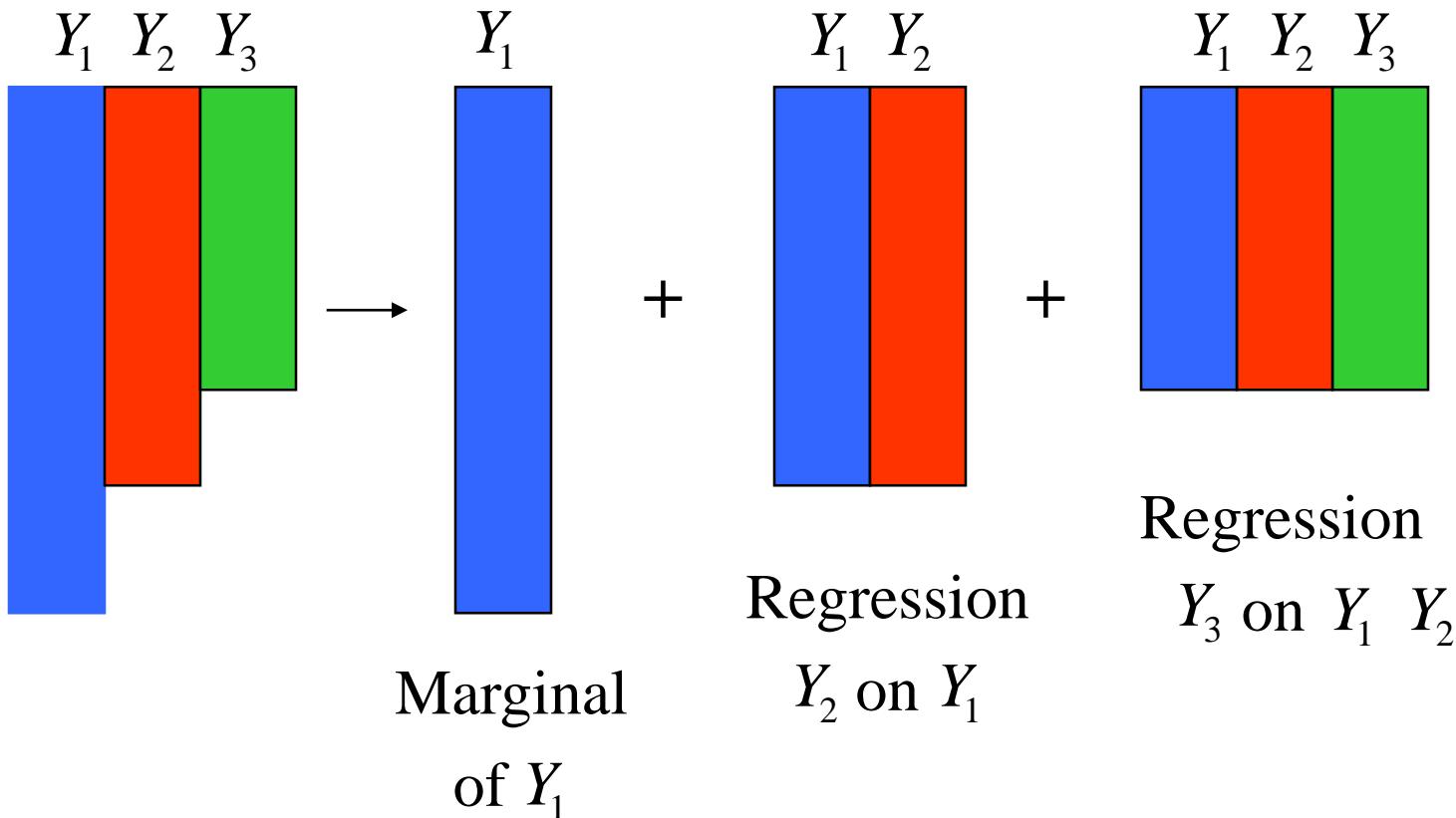
- When data are assumed to be MNAR. One has to specify the relationship between  $M$ , the response indicator, and the unobserved portion of substantive data,  $y_{\text{mis}}$ 
  - This assumption cannot be verified from the observed data in hand
  - Some external information is needed to specify this 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  $y_{\text{mis}}$
- 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)

# Computational tools

- Tools for monotone patterns
  - Maximum likelihood based on factored likelihood
  - Draws from Bayesian posterior distribution based on factored posterior distributions
- Joint distribution is factored into sequence of conditional distributions
- ML/Bayes is then a set of complete data problems (at least under MAR)

# Monotone Data, Three blocks

Regress current on more observed variables using available cases; e.g. 3 variables:



# ML: computational tools for general patterns

- ML usually requires iterative algorithms – general optimization methods like Newton Raphson and scoring, the EM algorithm and extensions (ECM, ECME, PXEM, etc.), or combinations
- Software – mixed model software like PROC MIXED, NLMIXED can handle missing values in outcomes, under MAR assumption
- This does not handle missing data in predictors
- MI has some advantages over this approach, as discussed later

# Bayes: computational tools for general patterns

- Iterative algorithms are usually needed
- Bayes based on Gibbs' sampler (which also provides multiple imputations of missing values)
- Gibbs' is essentially a stochastic version of ECM algorithm, yielding draws from posterior distribution of the parameters
- These draws from an intermediate step for creating multiple imputations from the posterior predictive distribution of the missing values
- Chained equation MI: logic of the Gibbs' sampler, with flexible modeling of sequence of conditional distributions. Trades rigor for practical flexibility

# Conclusion

- Bayesian prediction extends naturally to handle unit and item nonresponse
- MAR: missingness depends only on observed variables – key assumption for many methods

# Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

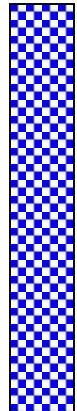
Module 15: missing data 2 -- unit  
nonresponse



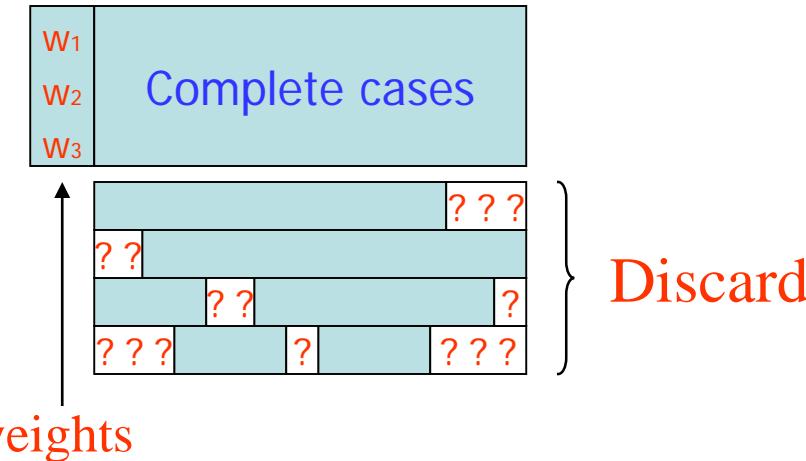
# Unit nonresponse

- Predict nonrespondents by regression on observed survey variables  $X$
- For bias reduction, predictors  $X$  should be related to  $M$  and outcome  $Y$
- In particular, consider categorical predictor  $X$  and flat priors, for inference about mean:
  - Bayes corresponds to weighting by inverse of estimated response rate in each category

Sample			Pop
$X$	$Y$	$M$	$X$
■■■■■	■■■■■	■■■■■	0
■■■■■			1



# Design-based approach: weighting



- One way to reduce the potential bias of CC analysis is to **weight** respondents differentially e.g. a CC mean becomes a weighted mean
- Common for unit nonresponse in surveys
- “Quasi-randomization inference” : extends ideas of randomization inference in surveys  
unit nonresponse

# Unit nonresponse

- Suppose we have:
- Design variables  $Z$  observed for whole population
- Fully observed survey variables  $X$ , measured for respondents and nonrespondents
- Survey variables  $Y$  measured only for respondents (see diagram)
- Goal of weighting is to use information in  $X$  and  $Z$  to weight the respondents, improve estimates

Sample				Pop
$Z$	$X$	$Y$	$M$	$Z$
1	1	0	1	1
1	1	1	0	1



# Sampling weights

- In a probability survey, each sampled unit  $i$  “represents”  $w_i$  units of the population, where

$$w_i = \frac{1}{\Pr(\text{unit } i \text{ sampled})}$$

$w_i$  is determined by the sample design and hence *known*

- Extend this idea to unit nonresponse ...

# Unit Nonresponse Weights

- If probability of response was known, could obtain weight for units that are sampled and respond:

$$\begin{aligned} w_i &= \frac{1}{\Pr(\text{unit } i \text{ is sampled and responds})} \\ &= \frac{1}{\Pr(i \text{ sampled})} \times \frac{1}{\Pr(i \text{ responds|sampled})} \\ &= (\text{sampling weight}) \times (\text{response weight}) \end{aligned}$$

Since prob of response is not known, we need to estimate it.

# Adjustment Cell method

- Group respondents and nonrespondents into adjustment cells with similar values on variables recorded for both:
    - e.g. white females aged 25-35 living in SW

100 in sample < 80 respondents  
20 nonrespondents

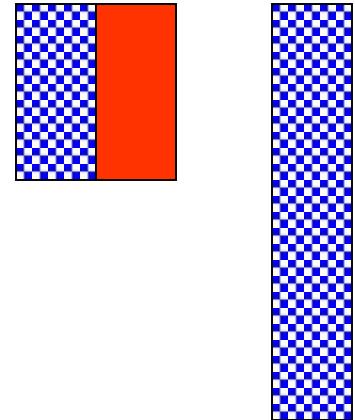
$\text{pr}(\text{response in cell}) = 0.8$

response weight = 1.25

# Ex 1: nonresponse with categorical Z

- Population divided into  $J$  adjustment cells  

Respondents	$X$	Sample
$X$	$Y$	$X$
- $X$  is set of indicators, known for sample:  
$$x_i = \begin{cases} 1, & \text{if unit } i \text{ is in cell } j; \\ 0, & \text{otherwise.} \end{cases}$$
- Simple random sampling  
of  $n_j$  units selected from cell  $j$ ,  $r_j$  respond.
- Assume MAR:  $M$  and  $Y$  are independent given  $X$ .



# Bayes inference for adjustment cell model

- Model for  $Y$  given  $X$ :

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

- For simplicity assume  $\sigma_j^2$  is known and the flat prior:

$$p(\theta_j \mid X) \propto \text{const.}$$

- Standard Bayesian calculations lead to

$$[\bar{Y} \mid Y_{\text{inc}}, X, \{\sigma_j^2\}] \sim N(\bar{y}_w, \sigma_w^2)$$

where:

$$\bar{y}_w = \sum_{j=1}^J p_j \bar{y}_j, \quad p_j = n_j / n, \quad \bar{y}_j = \text{respondent mean in cell } j$$

Equivalent to weighting  $n_j/r_j$ , inverse of response rate in cell  $j$

# Bayes inference for adjustment cell model

- If  $X$  is known for population (not just sample) posterior mean weights respondents by  $N_j/r_j$  (post-stratification)
- Unlike stratified mean, the counts  $r_j$  are not under control of sampler, and weights can be very large if  $r_j$  is small
- Hence may want to put a proper prior on adjustment cell means, to pull in extreme weights
- Posterior variance accounts for uncertainty in cell proportions

# Impact of weighting for nonresponse

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

	Low	High
Low	---	var $\downarrow$
High	var $\uparrow$	var $\downarrow$ bias $\downarrow$

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

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

# Weight the response rates?

- Nonresponse weights are often computed using units weighted by their sampling weights  $\{w_{1i}\}$

$$w_{2j}^{-1} = \left( \sum_{r_i=1, x_i=j} w_{1i} \right) / \left( \sum_{r_i=1, x_i=j} w_{1i} + \sum_{r_i=0, x_i=j} w_{1i} \right)$$

- Gives unbiased estimate of response rate in each adjustment cell defined by  $X$
- Not correct from a prediction perspective

# Arguments against weighting response weights

- Unnecessary if cells are created properly!
  - Adjustment cells should be homogeneous with respect to response propensity
  - In that case, weighting the nonresponse rates is unnecessary and adds variance to estimates
- Doesn't work if cells are created improperly!
  - If adjustment cells are not homogeneous with respect to response propensity, then weighting RR's does not yield unbiased estimates survey estimands.
- The right approach is to create adjustment cells based on classification of the observed variables and the survey design variables.
  - Then weighting is unnecessary

# Simulation Study

- Simulations in Little & Vartivarian (2003 Statistics in Medicine) consider the variance and bias of estimators of weighted and unweighted rates and alternative estimators, under a variety of population structures and nonresponse mechanisms.
- Binary outcome  $Y$ , stratum  $Z$ , adjustment cell  $X$  to avoid distributional assumptions such as normality.
- 25 populations to cover the factor space

# Simulation Results

- ML for the model used to generate the data is always best or close to best.
- Unweighted-RR( $x, z$ ) is best overall: form cells based on  $X$  and  $Z$
- Additive model theoretically biased when the data-generating model includes  $XZ$  interaction, but in these simulations the bias is modest.
- Unweighted-RR ( $x$ ) is biased when both  $Y$  and  $R$  depend on  $Z$ .
- Weighted RR ( $x$ ) does not generally correct the bias in these situations: similar to Unweighted-RR( $x$ ) overall

# Remarks

- Don't weight response rates! Rather
  - Condition on design variables when creating adjustment cells
- Too many strata? Response propensity stratification
- To improve efficiency: weight shrinkage by multilevel modeling

# Response propensity stratification

- $X$  = covariates observed for respondents and nonrespondents,  $Y$  missing
- $M$  = missing-data indicator
  - nonrespondent = 1, respondent = 0
- (A) Regress  $M$  on  $X$  (probit or logistic), using respondent and nonrespondent data  $\hat{p}(M = 0 | X)$  = propensity score
- (B1) Weight respondents by inverse of propensity score from (A),  $1/\hat{p}(M = 0 | X)$ , or:
- (B2) form adjustments cells by categorizing  $1/\hat{p}(M = 0 | X)$
- Note that this method is only effective if propensity is also related to the outcome

$X_1$	$X_2$	$\dots$	$X_p$	$Y$	$M$
Complete cases					0 0 0
				?	1
				?	1
				?	1

# Construction of Weights

- One should think about missing data at the design stage and collect information predictive of participation and key variables of interest
- Common sources for data
  - Administrative data from which the sample has been drawn
    - Medicare enrollment files
    - Driver license file
    - A large study may be used to sample subjects for a supplement or a smaller study.; Two-stage, nested case-control study etc.

- Data collected during the study conduct phase (“Paradata”)
  - Statements made by sampled subjects to the interviewer
  - Interviewer observations
- Data at Neighborhood level
  - Block or Block-group level characteristics
- Example
  - Assets and Health Dynamics (AHEAD) Study
  - Primary outcome variable: 5 point Self-reported health status
  - Interviewer noted the following four variables
    - Time delay statements
    - Negative statements about participation in study
    - Positive statements
    - Statements about being Old to participate etc

- Subject level variable
  - Age and Sex
- Neighborhood level variables
  - Large urban area (yes/no)
  - Barriers to contact (yes/no)
  - Block: Persons per sq-mile
  - Block : % persons 70 or older
  - Block: % Minority populations
  - Block: % Multi-unit structures (10+)
  - Block: % Occupied Housing Units
  - Block: % Single person Hus
  - Block: % vacant Hus
  - Block: Persons per occupied Hu
- Sample size: n=10,173, respondents=8,212 (80.7%)

# Comparison of “Paradata” between respondents and nonrespondents

Variable	Respondent	Nonrespondent	Effect Size
Age	77.41 (6.81)	77.83 (6.12)	0.065
Sex (% Female)	63.3%	63.6%	0.006
Mention age or illness	10.2%	16.1%	0.175
Negative Statement	9.5%	37.7%	0.703
Positive Statement	11.6%	2%	0.388
Time delay statement	8.1%	13.8%	0.183

*Effect size*

$$D = \frac{|\bar{x}_R - \bar{x}_{NR}|}{\sqrt{(s_R^2 + s_{NR}^2)/2}}$$

For proportions,  $s^2 = p(1-p)$

unit nonresponse

*Small* :  $D \leq 0.25$

*Medium* :  $0.25 < D \leq 0.5$

*Large* :  $0.5 < D \leq 0.75$

*Very Large* :  $D > 0.75$

# Comparison of Neighborhood level data between respondents and nonrespondents

Variable	Respondents	Nonrespondents	Effect size
% from Large urban areas	15.4%	22.7%	0.19
% with Barriers to contact	12.7%	20.0%	0.20
Block: Persons per sq-mile	6162.20 (14318.64)	7750.92 (15086.18)	0.11
Block : % persons 70 or older	10.98 (8.18)	11.16 (7.96)	0.02
Block: % Minority populations	22.76 (30.16)	24.60 (31.94)	0.06
Block: % Multi-unit structures (10+)	10.96 (20.24)	12.65 (21.74)	0.08
Block: % Occupied Housing Units	33.50 (23.76)	35.43 (24.68)	0.08
Block: % Single person HUs	24.48 (11.54)	25.17 (11.75)	0.06
Block: % vacant HUs	9.05 (9.18)	8.82 (8.90)	0.03
Block: Persons per occupied Hu	2.64 (0.47)	2.62 (0.48)	0.03

unit nonresponse

# Building a Response Propensity Model

- Goal is to obtain a well fitting (logistic or probit) model that balances the covariates between respondents and nonrespondents

$$e(x) = \text{Propensity score}$$

$$X \perp R \mid e(x)$$

- One useful measure of goodness of fit: Hosmer-Lemeshow test
  - Create deciles based on the estimated propensity score , compare the observed and expected frequencies /counts in these 10 classes
  - Chisquare statistics with 9 degrees of freedom as a measure of lack of it

- Checking the balancing property
  - Once satisfied with the logistic or probit model create classes (or strata or groups) based on quartiles or quintiles or deciles (depending upon your sample size) of the estimated propensity score.
  - In each class compute the effect size as we did for the overall sample.
  - Good balancing is indicated by the effect size being very small in each class
- Developing the propensity score model and checking its balancing property is an iterative process

- One can start with the main effect model and check the Hosmer-Lemeshow chisquare statistic
- Add two factor interactions to reduce the value of chisquare statistic
- Use stepwise or other selection procedure to reduce the model complexity
- Add higher order interactions if necessary
- Transform the covariates or standardized the continuous covariates (i.e. subtract the overall mean and divide by the overall standard deviation)

# AHEAD example

- First model with 16 main effects results in Hosmer-Lemeshow chisquare statistic of 28.68 with the p-value of 0.0004
- Next tried the model with all two factor interactions which reduced the chisquare statistic as 17.33 with the p-value of 0.03.
- As is typical, the full interaction model may be overmatching and result in poor fit.
- Next tried stepwise selection with different entry and exit probabilities. Finally with 0.5 as the probability for both entry and exit, obtained the model with Chisquare statistic of 4.75 with the p-value 0.784

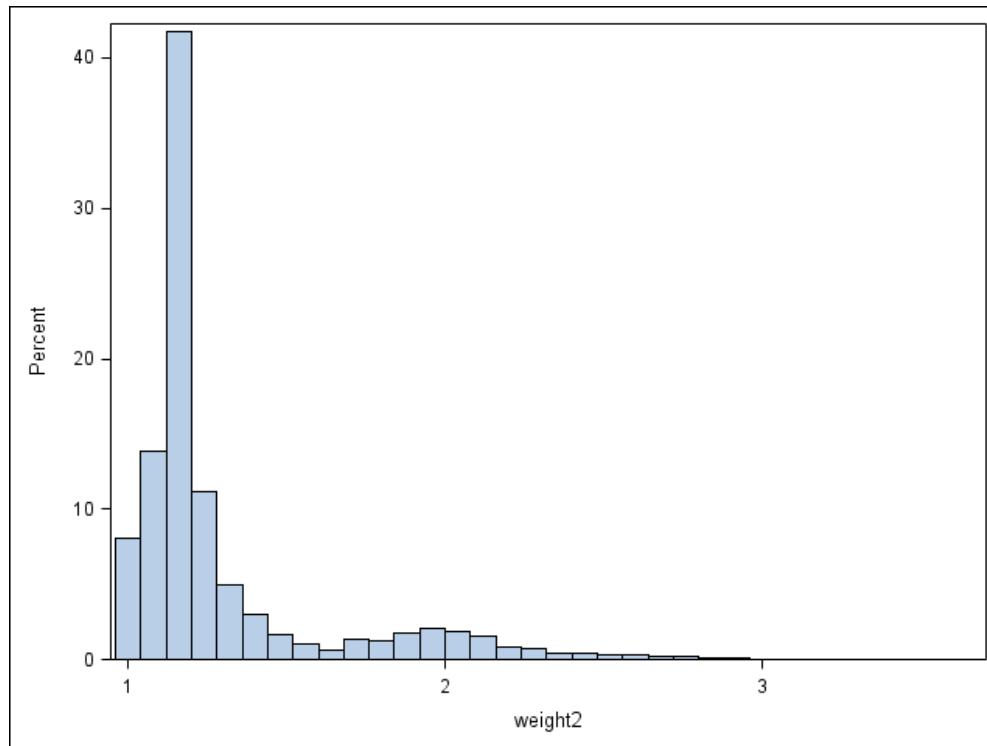
# Checking the Balance

- Created 4 groups (strata or classes) based on the quartiles of the estimated propensity squares.
- Computed the effect sizes for the 16 variables in each class. All were smaller than 0.05 indicating good balance.
- Propensity score stratification weights

Group	Respondents	Nonrespondents	Weight
1	1505	1037	1.6890
2	2100	445	1.2119
3	2237	306	1.1368
4	2370	173	1.0730

unit nonresponse

- Inverse probability weighting :  
weight=1/estimated propensity score
- Histogram of the nonresponse adjustment weight



# Weighted analysis

- Survey analysis software need to be used to correctly estimate the standard error. SAS procedures surveymeans, surveyfreq, surveyreg and surveylogistic are a few examples in SAS.
- In STATA, use svy commands
- In R, Thomas Lumley has developed packages and can be downloaded from the r library.

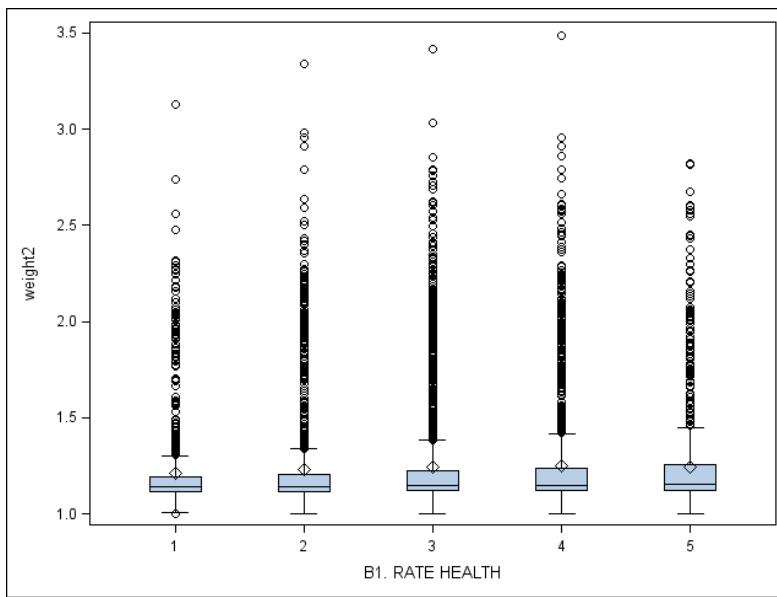
# Weighted and Unweighted frequencies

Self-rated health	Unweighted	Propensity score Class weighted	inverse probability weighted
1	10.75	10.51 (0.34)	10.52 (0.34)
2	22.80	22.53 (0.47)	22.61 (0.47)
3	30.35	30.41 (0.52)	30.50 (0.52)
4	23.08	23.29 (0.48)	23.31 (0.48)
5	13.03	13.26 (0.38)	13.06 (0.38)

unit nonresponse

# Weighted and Unweighted analysis

- The effect of weighting will depend upon the correlation between weight and the survey variable of interest



- Weighting is not that effective

# Inference from Weighted Data

- Role of weights in analytical inference (regression, factor analysis, ...) is controversial
- Can use packages for computing standard errors for complex sample designs -- but often these do not take into account sampling uncertainty in weights
- Bootstrap/Jackknife of weighting procedure propagates uncertainty in weights – but weights need to be recalculated on each BS/JK sample

# Penalized Spline of Propensity Prediction (PSPP)

- PSPP (Little & An 2004, Zhang & Little 2009, 2011).
- Regression imputation that is
  - Non-parametric (spline) on the propensity to respond
  - Parametric on other covariates
- Exploits the key property of the propensity score that conditional on the propensity score and assuming missing at random, missingness of  $Y$  does not depend on other covariates
- This property leads to a form of double robustness.
- Similar to previous PSPP for continuous stratifier, with propensity for selection replaced by estimated propensity to respond

# PSPP method

Estimate:  $Y^* = \text{logit} (\Pr(M=0/X_1, \dots, X_p))$

Impute using the regression model:

$$(Y | Y^*, X_1, \dots, X_p; \beta) \sim$$

$$N(s(Y^*) + g(Y^*, X_2, \dots, X_p; \beta), \sigma^2)$$

- Nonparametric part
- Need to be correctly specified
- We choose penalized spline

- Parametric part
- Misspecification does not lead to bias
- Increases precision
- $X_1$  excluded to prevent multicollinearity

# Alternative method: Predictive mean stratification

- $X$  = covariates observed for respondents and nonrespondents
- $Y$  = outcome with missing data
- (A) Regress  $Y$  on  $X$  (linear, other as appropriate) using respondent data only
- (B) Form adjustments cells with similar values of predictions from the regression  $\hat{Y}(X)$

$X_1$	$X_2$	$\dots$	$X_p$	$Y$	$M$
Complete cases					0 0 0
				?	1
				?	1
				?	1

- This method has potential to reduce both bias and variance
- Note that the adjustment cells depend on outcome, so this method yields different cells for each outcome, hence is more complex with many outcomes with missing values

# Conclusion

- Unit nonresponse: standard design-based approach is to weight by inverse of estimated response propensity
- Bayes for adjustment cells yields a posterior mean that is equivalent to weighting
- More generally, Bayes can use estimated propensity to respond as a predictor, as in PSPP

# Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 16: missing data 3 – item  
nonresponse



# 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
- Two variants of Bayes for item nonresponse:
  - Compute posterior predictive distribution of population quantities, given the observed data
  - Multiple imputation of draws from predictive distribution of missing values
- 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:

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

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

As  $t$  tends to infinity, this sequence converges to a draw from the joint posterior distribution of  $(Y_{\text{mis}}, \theta)$ , 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						
		?				$\hat{\mu}_1 = \text{mean based on all cases}$
		?				$\hat{\beta}_{51.1234} = ?$
	?		?			Impute to recover information in incomplete cases

# Single Imputation

- Impute missing values with predictions

Estimate ( $se^2$ )

Dataset ( $l$ )	$\mu_1$	$\beta_{51.1234}$
-----------------	---------	-------------------

$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$
-------	-------	-------	-------	-------

1	12.6 (3.6 $^2$ )	4.32 (1.95 $^2$ )
---	------------------	-------------------

$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$
24	1			
		2.1		
		4.5		

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

					Estimate ( $se^2$ )		
				Dataset ( $l$ )	$\mu_1$	$\beta_{51.1234}$	
$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$	1	12.6 ( $3.6^2$ )	4.32 ( $1.95^2$ )
				2	12.6 ( $3.6^2$ )	4.15 ( $2.64^2$ )	
					2.7		
					5.1		
					31	1	

# Third imputed dataset

					Dataset ( $l$ )	Estimate ( $se^2$ )	
$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$		$\mu_1$	$\beta_{51.1234}$
					1	12.6 (3.6 <sup>2</sup> )	4.32 (1.95 <sup>2</sup> )
					2	12.6 (3.6 <sup>2</sup> )	4.15 (2.64 <sup>2</sup> )
					3	12.6 (3.6 <sup>2</sup> )	4.86 (2.09 <sup>2</sup> )
1.9							
5.8							
32		2					



# Fourth imputed dataset

					Estimate ( $se^2$ )		
					Dataset ( $l$ )		
					$\mu_1$	$\beta_{51.1234}$	
$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$	1	12.6 (3.6 <sup>2</sup> )	4.32 (1.95 <sup>2</sup> )
					2	12.6 (3.6 <sup>2</sup> )	4.15 (2.64 <sup>2</sup> )
					3	12.6 (3.6 <sup>2</sup> )	4.86 (2.09 <sup>2</sup> )
					4	12.6 (3.6 <sup>2</sup> )	3.98 (2.14 <sup>2</sup> )
2.5							
3.9							
18			1				

# Fifth imputed dataset

					Estimate ( $se^2$ )		
				Dataset ( $l$ )	$\mu_1$	$\beta_{51.1234}$	
$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$	1	12.6 (3.6 <sup>2</sup> )	4.32 (1.95 <sup>2</sup> )
					2	12.6 (3.6 <sup>2</sup> )	4.15 (2.64 <sup>2</sup> )
		2.3			3	12.6 (3.6 <sup>2</sup> )	4.86 (2.09 <sup>2</sup> )
			4.2		4	12.6 (3.6 <sup>2</sup> )	3.98 (2.14 <sup>2</sup> )
				25	5	12.6 (3.6 <sup>2</sup> )	4.50 (2.47 <sup>2</sup> )
					Mean	12.6 (3.6 <sup>2</sup> )	4.36 (2.27 <sup>2</sup> )
					Var	0	0.339

# MI combining rules

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

$$E(\theta | \mathbf{Y}_{\text{obs}}) \approx \frac{1}{D} \sum_{d=1}^D E(\boldsymbol{\theta} | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(d)}) = \frac{1}{D} \sum_{d=1}^D \hat{\theta}_d$$

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

$$\text{Var}(\theta | \mathbf{Y}_{\text{obs}}) \approx \frac{1}{D} \sum_{d=1}^D W_d + (1 + 1/D) \frac{1}{D-1} \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta}_D)^2$$

where  $W_d = \text{Var}(\theta | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(d)})$  is posterior variance from  $d$ th dataset

# MI Inferences (M=5)

	$\bar{\theta}$	$\bar{W}$	$B$	$\sqrt{V} = \sqrt{\bar{W} + (1 + \gamma_D)B}$	$R = \frac{(1+1/D)B}{V}$
$\mu_1$	12.6	$3.6^2$	0	3.6	0
$\beta_{51.1234}$	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:

$$\mathbf{X} \sim \mathbf{N}(0,1)$$

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

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

- 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:

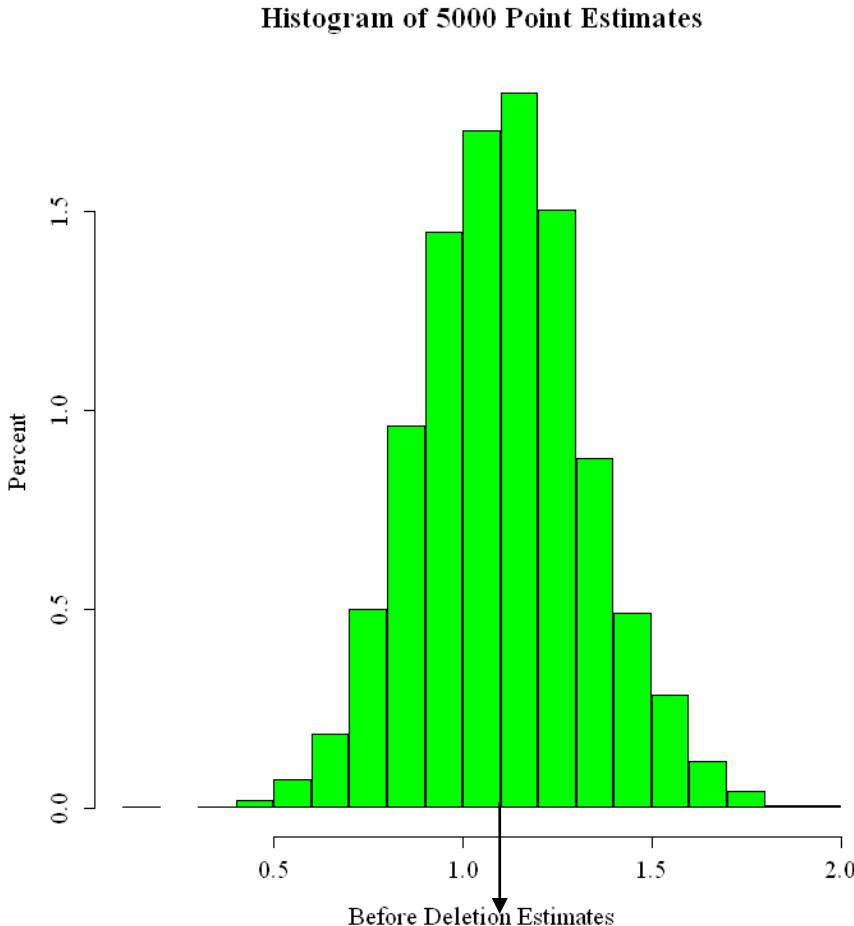
$$D=0, E=0: \quad p_{00}=0.19$$

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

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

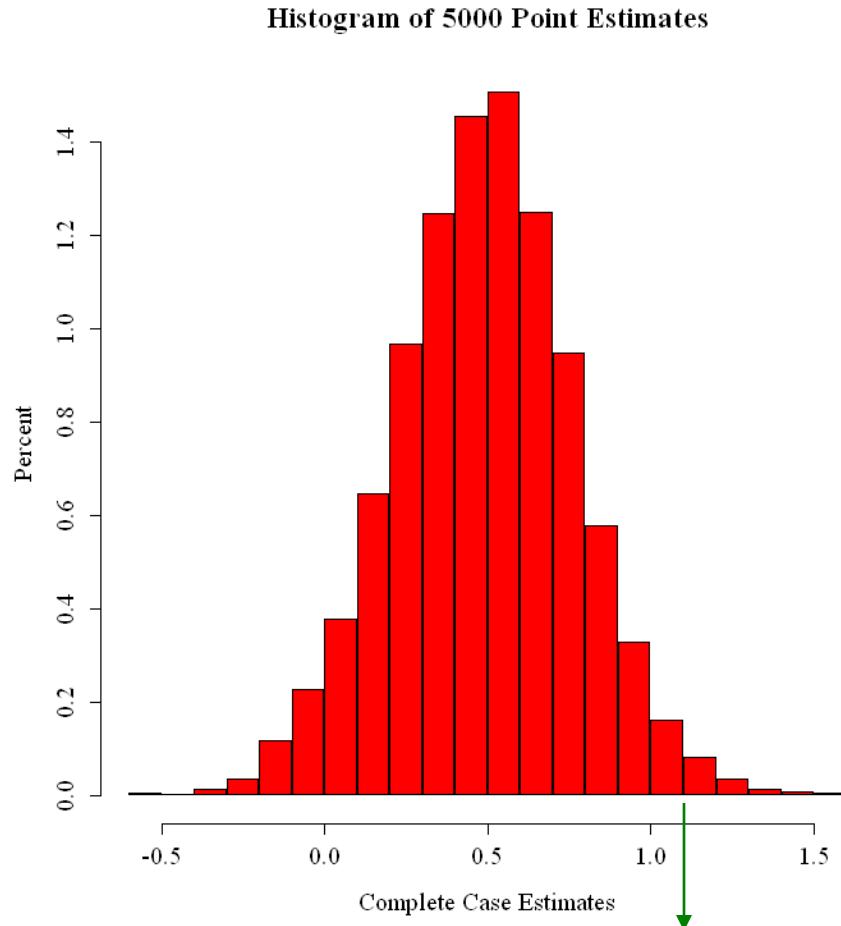
$$D=1, E=1: \quad p_{11}=0.055$$

# Before Deletion Estimates



- Histogram of 5000 estimates before deleting values of X
- logistic model  
$$\text{logit } \Pr(D=1|E, X) = \beta_0 + \beta_1 E + \beta_2 X$$

# 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  $X$ :

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

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

# Predictive Distributions

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

$WSS$  = residual sum of squares,

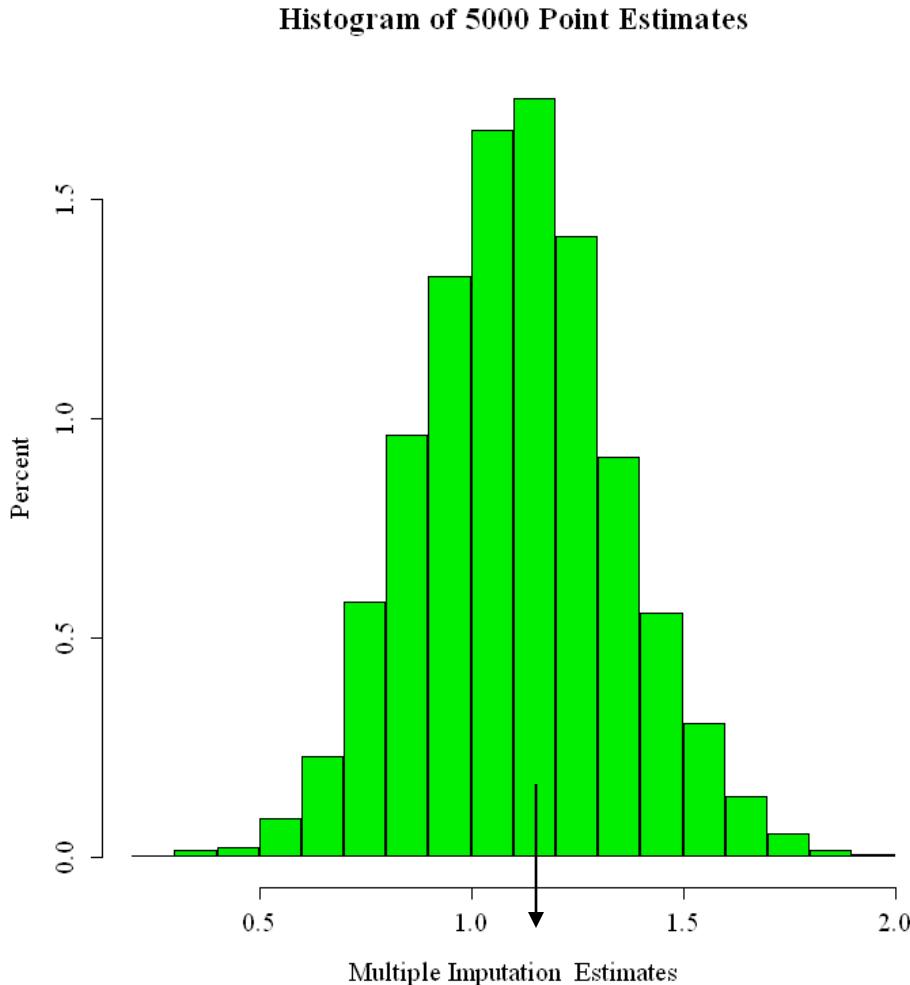
$r$  = number of complete cases

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

$\bar{x}_{ed}, r_{ed}$  = 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	90.28	0.0566
Single Imputation		
Multiple Imputation	94.56	0.0547
<i>Before Deletion</i>	<b>94.68</b>	<b>0.0494</b>

# Bayesian Theory of MI (Rubin, 1987)

For simplicity assume MAR -- MNAR also allowed

Model:  $f(Y | \theta) \Rightarrow$  Likelihood  $L(\theta | Y) \propto f(Y | \theta)$

Prior distribution:  $\pi(\theta)$ ; md mechanism: MAR

$Y = (Y_{\text{obs}}, Y_{\text{mis}})$ ,  $Y_{\text{obs}}$  = observed data,  $Y_{\text{mis}}$  = missing data

Complete-data posterior distribution,

if there were no missing values:

$$p(\theta | Y_{\text{obs}}, Y_{\text{mis}}) \propto \pi(\theta) L(\theta | Y_{\text{obs}}, Y_{\text{mis}})$$

Posterior distribution given observed data:

$$p(\theta | Y_{\text{obs}}) \propto \pi(\theta) L(\theta | Y_{\text{obs}})$$

Theory relates these two distributions ...

# Relating the posteriors

- The posterior is related to the complete-data posterior by:

$$\begin{aligned} p(\theta | Y_{\text{obs}}) &= \int p(\theta | Y_{\text{obs}}, \mathbf{Y}_{\text{mis}}) p(\mathbf{Y}_{\text{mis}} | Y_{\text{obs}}) d\mathbf{Y}_{\text{mis}} \\ &\approx \frac{1}{D} \sum_{d=1}^D p(\theta | Y_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(d)}), \text{ where } \mathbf{Y}_{\text{mis}}^{(d)} \sim p(\mathbf{Y}_{\text{mis}} | Y_{\text{obs}}) \end{aligned}$$

$\mathbf{Y}_{\text{mis}}^{(d)}$  is a draw from the predictive distribution of the missing values

The accuracy of the approximation increases with  $D$  and the fraction of observed data

# MI approximation to posterior mean

- Similar approximations yield MI combining rules:

$$E(\theta | Y_{\text{obs}}) = \int E(\theta | Y_{\text{obs}}, \mathbf{Y}_{\text{mis}}) p(\mathbf{Y}_{\text{mis}} | Y_{\text{obs}}) d\mathbf{Y}_{\text{mis}}$$

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

where  $\hat{\theta}_d$  = is posterior mean from  $d$ th imputed dataset

# MI approximation to posterior variance

$$\text{Var}(\theta | Y_{\text{obs}}) = E(\theta^2 | Y_{\text{obs}}) - (E(\theta | Y_{\text{obs}}))^2$$

Apply above approx to  $E(\theta | Y_{\text{obs}})$  and  $E(\theta^2 | Y_{\text{obs}})$

Algebra then yields:

$$\text{Var}(\theta | Y_{\text{obs}}) \approx \bar{V} + B$$

$$\bar{V} = \frac{1}{D} \sum_{d=1}^D V_d = \text{within-imputation variance},$$

$V_d = \text{Var}(\theta | Y_{\text{obs}}, Y_{\text{mis}}^{(d)})$  is posterior variance from  $d$ th dataset

$$B = \frac{1}{D-1} \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta}_D)^2 = \text{between-imputation variance}$$

# Refinements for small $D$

(A):  $Var(\theta | Y_{\text{obs}}) \approx \bar{V} + (1 + 1/D) B$

(B) Replace normal reference distribution by t distribution with  $\nu$  df

$$\nu = (D - 1) \left( 1 + \frac{D}{D + 1} \frac{\bar{V}}{B} \right)^2$$

(C) For normal sample with variance based on  $\nu_{\text{com}}$  df, replace  $\nu$  by

$$\nu^* = \left( \nu^{-1} + \hat{\nu}_{\text{obs}}^{-1} \right)^{-1}, \hat{\nu}_{\text{obs}} = (1 - \hat{\gamma}_D) \left( \frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3} \right) \nu_{\text{com}}$$

$$\hat{\gamma}_D = \frac{(1 + D^{-1}) B}{\bar{V} + (1 + D^{-1}) B} = \text{estimated fraction of missing information}$$

# Why MI for surveys?

- Software is widely available (IVEware, MICE, etc.)
- MI based on Bayes for a joint model for the data has optimal asymptotic properties under that model.
- Propagates imputation uncertainty in a way that is practical for public use files
- Flexible, using models that fully condition on observed data – makes MAR assumption “as weak as possible”
- Applies to general patterns – weighting methods do not generalize in a compelling way beyond monotone patterns

# Why MI for surveys?

- Allows inclusion of auxiliary variables in the imputation model that are not in the final analysis
- “Design-based” methods can be applied to multiply-imputed data, with MI combining rules: model assumptions only used to create the imputations (where assumptions are inevitable).

# Arguments against MI for surveys

- It's model-based, and I don't want to make assumptions – but there is no assumption-free imputation method!
- Lack of congeniality between imputer model and analyst model
  - advice is to be inclusive of potential predictors, leading to at worst conservative inferences – parametric models allow main effects to be prioritized over high order interactions
  - Congeniality problem also applies to other methods that falsely claim to be assumption free
  - Perfection is the enemy of the good – in simulation studies, MI tends to work well, because it is propagating imputation uncertainty

# Arguments against MI for surveys

- Misspecified parametric models can lead to problems with the imputes – for example, imputing log-transformed data and then exponentiating can lead to wild imputations
- So, important to plot the imputations to check that they are plausible
- With large samples, chained equations with predictive mean matching hot deck has some attractions, since only actual values are imputed
- But hot deck methods are less effective in small samples where good matches are lacking (Andridge & Little, 2010)

# Missing Not at Random Models

- Difficult problem, since information to fit non-MAR is limited and highly dependent on assumptions
- Sensitivity analysis is preferred approach – this form of analysis is not appealing to consumers of statistics, who want clear answers
- Selection vs Pattern-Mixture models
  - Prefer pattern-mixture factorization since it is simpler to explain and implement
  - Offsets, Proxy Pattern-mixture analysis
- Missing covariates in regression
  - Subsample Ignorable Likelihood

# A simple pattern-mixture model

Giusti & Little (2011) extends this idea to a PM model for income nonresponse in a rotating panel survey:

- \* Two mechanisms (rotation MCAR, income nonresponse NMAR)
  - \* Offset includes as a factor the residual sd, so smaller when good predictors are available
  - \* Complex problem, but PM model is easy to interpret and fit
- Readily implemented extension of chained equation MI to MNAR models

# An Alternative: Proxy Pattern-Mixture Analysis

$$[y_i | x_i, r_{2i} = k] \sim G(\beta^{(k)} x_i, \tau^{2(k)})$$

$$\Pr(r_i = 1 | x_i, y_i) = g(y_i^*(\lambda)), \quad y_i^*(\lambda) = \hat{y}(x_i) + \lambda y_i$$

$\hat{y}(x_i)$  = best predictor of  $y_i$

MAR:  $\lambda = 0$ , MNAR:  $\lambda \neq 0$

(Andridge and Little 2011)

(\*) implies that  $[y_i \text{ indep } r_i | y_i^*(\lambda)]$ , which identifies the model

Interesting feature:  $g()$  is arbitrary, unspecified

NMAR model that avoids specifying missing data mechanism

PPMA: Sensitivity analysis for different choices of  $\lambda$

If  $x_i$  is a noisy measure of  $y_i$ , it may be plausible to assume  $\lambda = \infty$   
(West and Little, 2013)

# 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

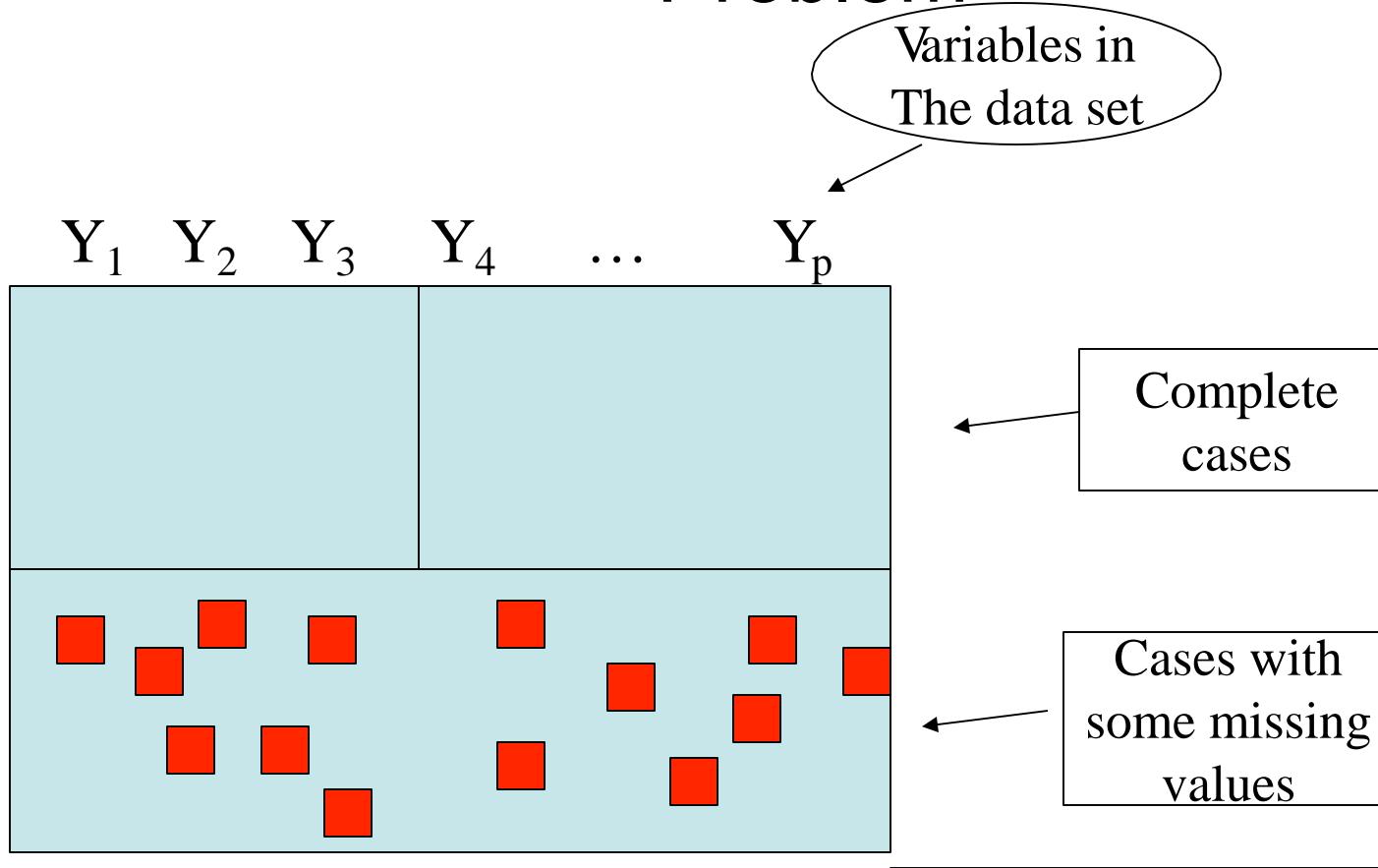
# Bayesian Inference for Surveys

Roderick Little and Trivellore Raghunathan

Multiple Imputation using Sequential  
Regression/Chained Equations



# Problem



$D_{obs}$  = Observed data: 

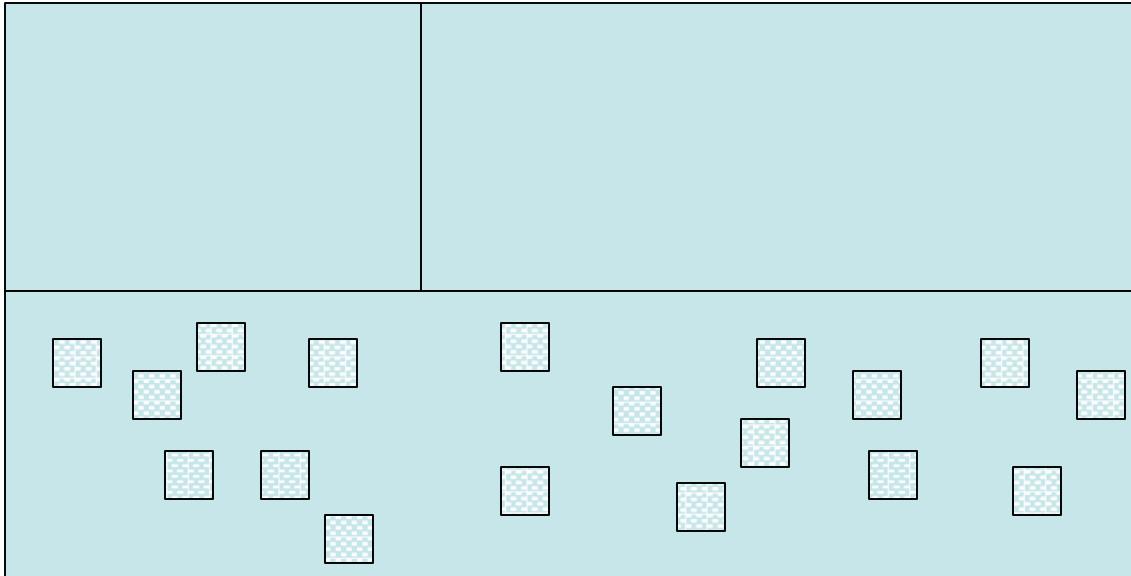
$D_{miss}$  = Missing data: 

$Y$ : Discrete, continuous or semi-continuous as well as multivariate

# Setting

- Multiple users analyzing different subsets of variables
- Multiple analytical techniques
- Different skill levels dealing with incomplete data
- Analysis to be performed with complete data is known
- Software to perform complete data analysis is available
- Assume missing at random.
  - That is conditional on the observed characteristics the residual differences between those with missing and those with no missing values are random

# Imputation



"Ideal" imputations :

Draws from  $\Pr(D_{miss} | D_{obs})$

Important issues:

Imputations are not real values

Uncertainties associated with imputes

# Practical Issues

- Hot deck imputation is limited
  - Variables have to be completely observed
  - Continuous variables have to be categorized
- Explicit Model is difficult
  - Large number of variables of different types
  - Restrictions
    - Question is valid only for certain subjects
    - Skip pattern
  - Bounds
    - Variables are bounded. *Example: Years smoked cannot exceed Age for current smokers and (Age-Years since Quit smoking ) for former smokers. It can become more complex, if a question about teen age smoking was asked and age when started smoking was also asked*
    - Bracketed responses

# Sequential Regression/Chained Equation/Flexible Conditional Specification Approach

Variables With Missing Values:

$$Y_1, Y_2, \dots, Y_p$$

Variables With No Missing Values:  $U$

Iteration 1:

$$Y_1 | U$$

$$Y_2 | Y_1^{(1)}, U$$

$\vdots$

$$Y_j | U, Y_1^{(1)}, \dots, Y_{j-1}^{(1)}$$

$\vdots$

$$Y_p | U, Y_1^{(1)}, Y_2^{(1)}, \dots, Y_{p-1}^{(1)}$$

Iteration  $t=2,3,\dots$ :

$$Y_1 | U, Y_2^{(t-1)}, \dots, Y_p^{(t-1)}$$

$$Y_2 | U, Y_1^{(t)}, Y_3^{(t-1)}, \dots, Y_p^{(t-1)}$$

$\vdots$

$$Y_j | U, Y_1^{(t)}, \dots, Y_{j-1}^{(t)}, Y_{j+1}^{(t-1)}, \dots, Y_p^{(t-1)}$$

$\vdots$

$$Y_p | U, Y_1^{(t)}, \dots, Y_{p-1}^{(t)}$$

Sequential Regression/Chained Equation

Each step involves draws from the predictive distribution

- Ability to specify individual regression model
- Types of variables
  - Continuous (Normal)
  - Categorical (Logistic or generalized logistic)
  - Count (Poisson)
  - Mixed or semi-continuous (Logistic/Normal)
  - Ordinal (ordered probit)
- Parametric or semi-parametric regression models
- Restrictions
  - Regression model is fitted only to the relevant subset
- Bounds
  - Draws from a truncated distribution from the corresponding regression model
- Models each conditional distribution. There is no guarantee that a joint distribution exists with these conditional distributions
- How many iterations?
  - Empirical studies show that nothing much changes after 5 or 6 iterations

# Software

- Sequential regression imputations
  - R and Stata (MICE, ICE, MI)
  - Standalone (SRCWARE)
  - SAS (IveWare), PROC MI
- MI-Analysis
  - PROC MIANALYZE
  - IveWare (can handle complex sample survey)
  - SRCWARE
  - MICOMBINE/MITOOLS (STATA)
  - SUDAAN

# Software for Multiple Imputation Analysis

## For Creating Imputations

- SAS
  - PROC MI
  - IVEware
- Standalone
  - SRCware
- STATA
  - MI IMPUTE
  - IVEware
- R
  - MICE
  - IVEware
- SOLAS
- SPSS (Version 22)
  - IVEware

## For Analysis of multiply imputed data

- SAS
  - PROC MIANALYZE
  - IVEware
- Standalone
  - SRCware
- STATA
  - MI ESTIMATE
- SUDAAN
- R
- SPSS (Version 22)

# IveWare

- SAS, R, Stata, SPSS interface
  - A collection of C and Fortran routines
  - Handles linear (Continuous), logistic (Binary), multinomial logistic (categorical), Poisson (Count) and two-stage linear/logistic (Mixed or semi-continuous)
  - Predictive mean matching using Approximate Bayesian Bootstrap and Tukey's gh distribution
  - Stepwise selection possible at each step to save computation time (use with caution and only if it is absolutely necessary )
  - Add interaction terms
  - Specify bounds
  - Specify logical restrictions and skip patterns

- Uses Normal approximation for the posterior distribution of the parameters
- Sampling Importance Resampling to handle non-normal posterior
- Non-informative prior
- Single chain or multiple chain (starting with different seeds)
- Iterations and Multiples control the length of the chain
- Built-in diagnostics to assess imputed values
- Complex Survey Data Analysis
- Download: [www.isr.umich.edu/src/smp/ive/dev](http://www.isr.umich.edu/src/smp/ive/dev)

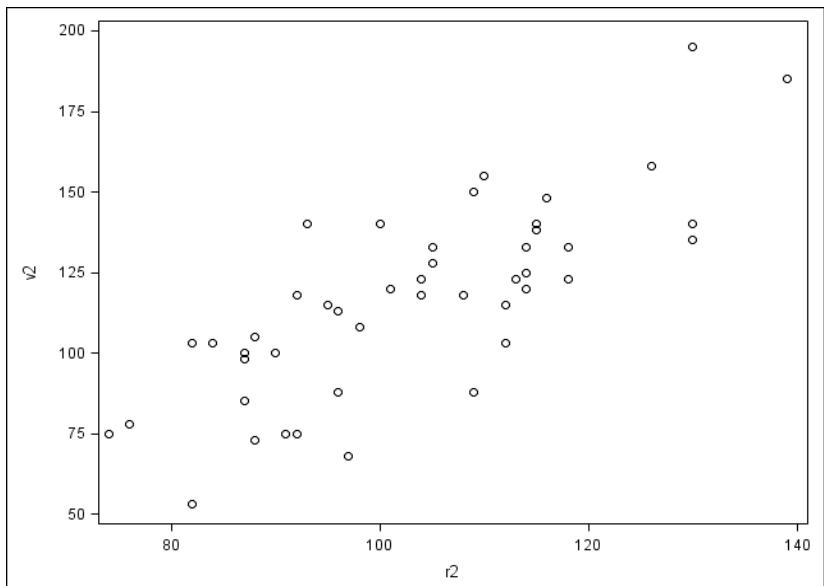
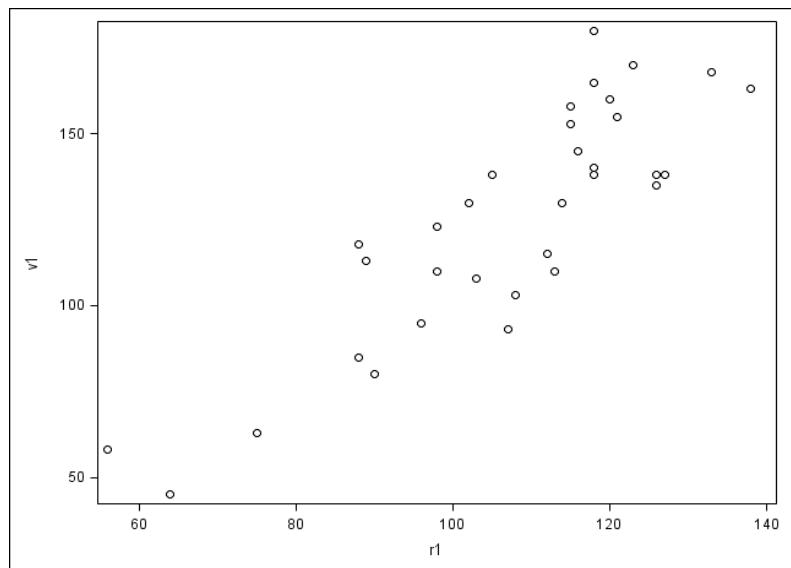
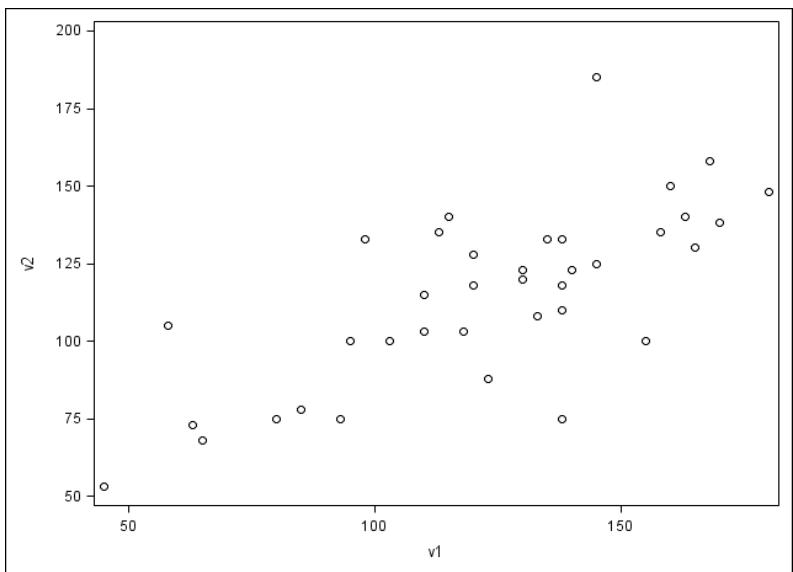
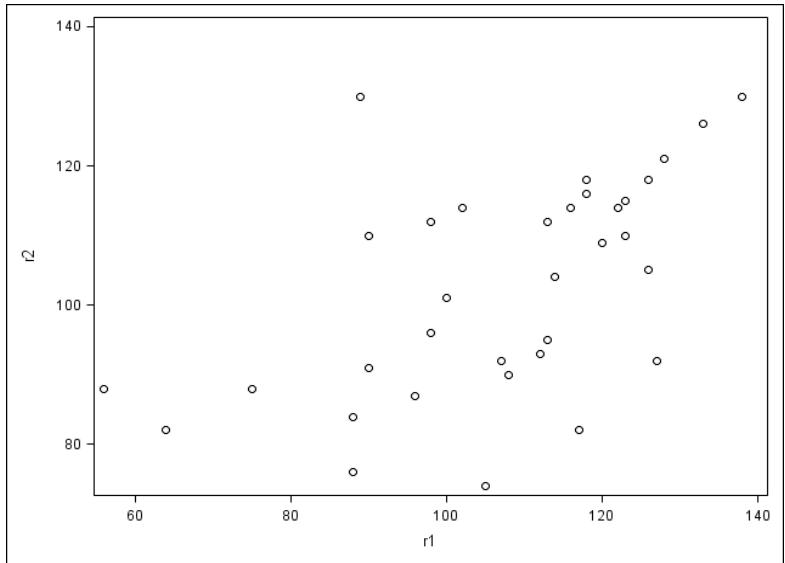
- Issues
  - Convergence
  - Several completed data statistics seem to converge to the same value regardless of seeds
  - Zhu and Raghunathan (JASA 2015) establish conditions for convergence
  - Good fitting models are needed to get results with desirable repeated sampling properties

# St. Louis Risk Study

(Little and Rubin, 2002)

- A study was conducted to evaluate the effects parental psychological disorders on various aspects of the development of the children. Data from 69 families with two children were collected. Families were classified into risk group of the parent (G) with
  - G=1 normal or control group
  - G=2 Moderate risk group with one parent having some psychiatric illness
  - G=3 High risk group with one or more parent having schizophrenia or affective mental dis order

- Variables measured on Child 1
  - D1= Number of symptoms (1=Low, 2=High)
  - V1= Standardized verbal comprehension score
  - R1=Standardized Reading score
- Variables Measured on Child 2
  - D2, V2, R2
- G is always observed and other variables are missing with variety of different combinations



# St Louis Risk Study (Contd.)

- Sequential regression approach used to impute the missing values R1, R2, V1, V2 using normal linear regression model and D1, D2 using the logistic regression model
- Analysis
  - Regress R on G and D , treating the Family ID as “cluster” or “Repeated” factor
  - Regress V on G and D, treating the Family ID as “cluster” or “Repeated” factor
  - Regress D on G, treating Family ID as “cluster” or “Repeated” factor

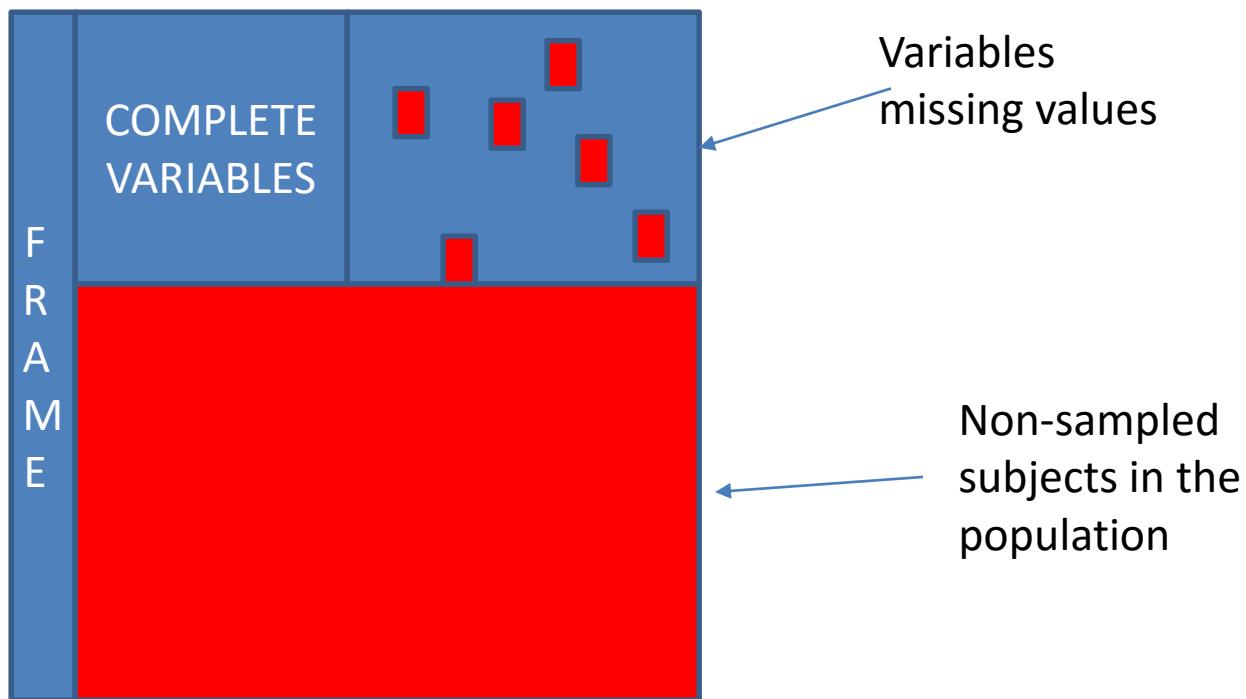
# Results

## Multiple Imputation Analysis

Parameter	Reading	Verbal	Symptoms
Intercept	114.23 (5.49)	152.67 (15.50)	-0.32 (0.41)
Group 2 vs 1	-9.85 (3.93)	-25.14 (13.56)	1.05 (0.74)
Group 3 vs 1	-9.69 (5.09)	-19.41 (11.03)	0.47 (0.51)
Symptoms	-1.02 (3.35)	-10.86 (10.97)	

# Prediction of the population

- Schematic Display



# Options

- Option 1: Impute the missing values in the sample and then predict the non-sampled portion of the population
- Option 2: Simultaneously impute all the missing values including the non-sampled portion of the population
- Model

$$\begin{aligned}\Pr(Y, I, M | Z) &= \Pr(Y | Z) \Pr(I | Y, Z) \Pr(M | Y, I, Z) \\ &= \Pr(Y | Z) \Pr(I | Z) \Pr(M | Y_{obs}, I, Z) \\ Y &= \{Y_{obs}, Y_{mis}, Y_{exc}\}\end{aligned}$$

- Option 1

$$\Pr(Y_{exc} | Y_{obs}, Y_{mis}) \Pr(Y_{mis} | Y_{obs})$$

- Not all Y's to be predicted for the population
- Simpler

- Option 2

$$\Pr(Y_{exc}, Y_{mis} | Y_{obs})$$

- All Y's in the population are to be predicted
- May be useful as a public-use file

# MI Applications

- Survey of Consumer Finances, 1992
  - 5 multiply imputed data sets
- National Health and Nutritional Examination Survey
  - 5 multiply imputed data sets for a selected set of variables in NHANES-III. Uses general location model.
- National Health Interview Survey 1997-Present
  - Multiple imputation of missing family income and personal earnings.
- Numerous applications in a variety of fields. Becoming a very common approach.

# Conclusion

- Sequential Regression/Chained Equation is a flexible approach for handling missing data with varying type of variables and complex structure
- Standard regression diagnostics can be used to fine tune the model to fit the observed data well
- Models can be parametric, semi-parametric or non-parametric
- Many software available to implement the method
- It is easy to program using a macro environment

# Bayesian Inference for Surveys

Roderick Little and Trivellore  
Raghunathan

Model Checking, Comparison and  
Averaging



# Model Checking

- Model checking is an important step in a Bayesian analysis
- Two approaches
  - Generate new data from the marginal distribution of observables under the model and compare with the observed data. Prior-predictive check
  - Generate new data from the posterior predictive distribution and compare with the observed data. Posterior-predictive check

$y$  = Observed data

$\theta$  = Parameters in the model

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

$\pi(\theta)$ =Prior density

# Prior Predictive Check

- Prior predictive distribution (marginal distribution of observables)

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

- Generate new data from the prior-predictive distribution and compare with the observed data  $y_{\text{inc}}$

$\theta^*$  = draw from  $\pi(\theta)$

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

Not possible to implement  
if the prior distribution is  
improper

# Posterior Predictive Check

- Posterior predictive distribution

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

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

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

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

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

# Modification of Posterior Predictive check

- Use a subsample to construct posterior distribution

$$y_{inc} = (y_{inc}^{(1)}, y_{inc}^{(2)})$$

$$\theta \sim \pi(\theta | y_{inc}^{(1)})$$

$$y_{new}^{(2)} \sim f(y | \theta)$$

*Compare :*  $y_{new}^{(2)}, y_{inc}^{(2)}$

This is useful when a survey is conducted by drawing replicates

Using half the replicates for model fitting and the other half for model checking

# Example 1: Prior and Posterior Predictive Distributions

Objective: To check the independence in a sequence of Bernoulli trials

Model :

$$y_i \sim \text{iid} \text{ Bin}(1, \theta)$$

$$\theta \sim \text{Unif}(0,1)$$

Observed sequence:

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

$$T(y_{\text{inc}}) = 3$$

Prior Predictive

$$\theta^{\text{rep}} \sim \text{Unif}(0,1)$$

$$y_i^{\text{rep}} \sim \text{Bin}(1, \theta^{\text{rep}})$$

$$i = 1, 2, \dots, 20$$

$$T(y^{\text{rep}})$$

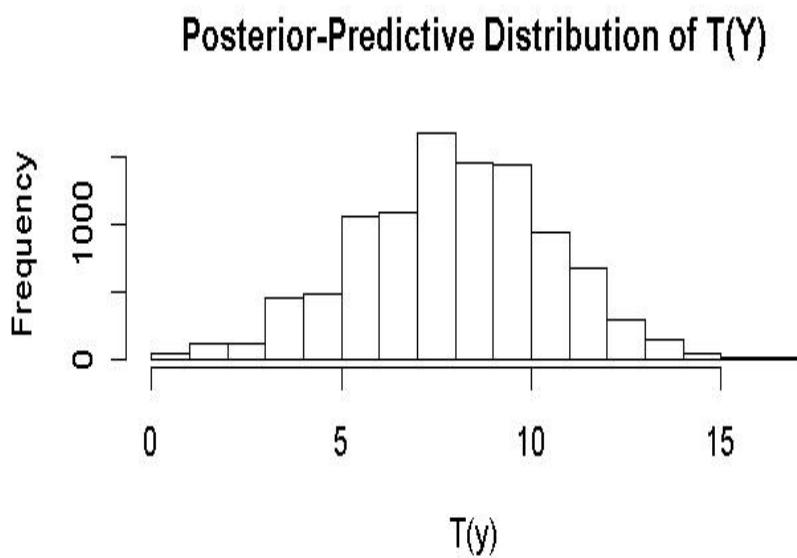
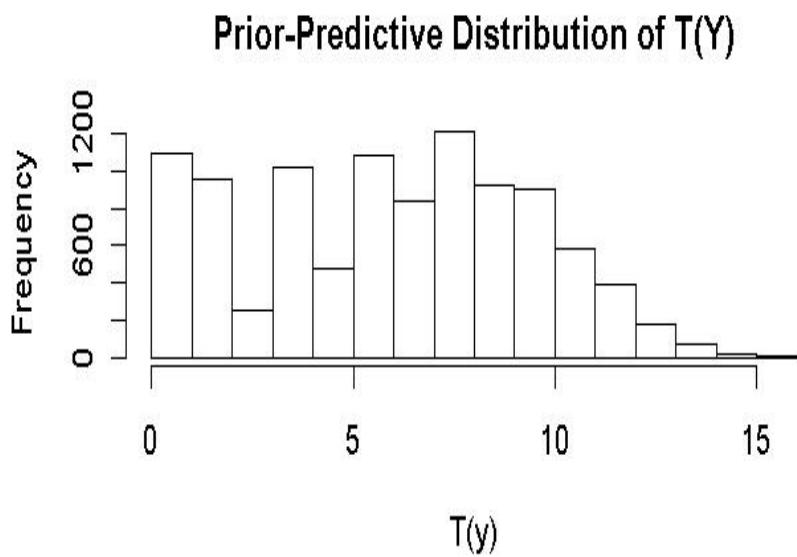
Posterior predictive

$$\theta^{\text{rep}} \sim \text{Beta}(8, 14)$$

$$y_j^{\text{rep}} \sim \text{Bin}(1, \theta^{\text{rep}})$$

$$j = 1, 2, \dots, 20$$

$$T(y^{\text{rep}})$$



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
Checking 3	0	17
		7

# Comparing new and observed data

- Develop “discrepancy measure”  $T(y)$  or  $T(y, q)$ .  
Note that the discrepancy measure can depend upon the parameter  $q$ .
- Compare  $T(y_{\text{inc}})$  with  $T(y_{\text{new}})$  or  
 $T(y_{\text{inc}}, \theta^*)$  with  $T(y_{\text{new}}, \theta^*)$
- Discrepancy measures depends upon the problem
- General goodness of fit measure:

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

# Model Checking: Quality Control Example

- In the quality control example with beta-binomial model, generating replicates is fairly straightforward given the draws of  $(a, b, q)$  (independent binomial samples)
- The discrepancy measure used:

$$T = (\text{Min}(y_i - k_i \theta_i), \text{Max}(y_i - k_i \theta_i))$$

- Computed fraction of times, out of 2500 replicates, the observed minimum was less than the replicate minimum and/or the replicate maximum was greater than the observed maximum (Bayesian p-value)
- 2247 replicates did not capture the observed spread (p-value=0.1012)

# 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

- Model checking requires thought about measuring deviations between observed data and what one would expect under the model.
- Many such choices have to be investigated to ensure that the model is an adequate abstraction for glean useful information about the population based on the observed values and express the associated uncertainty.
- An alternative mechanism: To compute the posterior probability for the model being “correct”.
  - Bayes factor

# Model Comparisons 2

- $K$  models under consideration

$$M_j, j = 1, 2, \dots, K$$

- Prior probabilities for model  $j$  being correct

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

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

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

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

# Model Comparisons 3

Posterior odds = Prior odds x Bayes Factor:

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

Bayes factor

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

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

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

ML estimate under model  $M$

$k$  = number of parameters

# Model Comparisons 4

- Bayes Information Criterion

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

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

- 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

$$\begin{aligned}\pi(\theta | y) &= \int \pi(\theta, \phi | y) d\phi \\ &\propto \int f(y | \theta, \phi) \pi(\theta | \phi) p(\phi) d\phi\end{aligned}$$

# Example

- Original model

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

- Expanded model

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

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

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

- 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