AIMS

- Relevance of the study design in Bayesian analysis
- A collection of simple examples
- Complete-data likelihood and observed-data likelihood

  missing at random

  distinct parameters
- Ignorable designs with no covariates

  random sampling

  completely random experiments

Study Design in Bayesian Analysis

How should one account for the study-design of

sample-survey

an experiment

an observational study

in Bayesian analysis?

we must include the study-design as part of full probability modelling

Given

- A fixed model, including prior distribution for the underlying data
- fixed observed values of the data

then:

Bayesian inference is determined regardless of the design of the collection of the data.

This is a misplaced appeal of the likelihood principle

Key issues

- The pattern of what has been observed can be informative
- Sensitivity analysis to model specifications is part of the Bayesian analysis
- Thinking about design and the data one *could have observed* helps us structure inference about models and finite-population estimands such us the population mean in a simple survey or the average casual effect of an experimental treatment.

Data collection problem

- observed data
- complete data (potential data)
- missing data

Inference is conditional on observed data AND on the pattern of observed and missing observations

- unintentional missing data: they are due to unfortunate circumstances
- intentional missing data: data from units "appositely" not sampled in a survey and results of treatments "appositely" not applied in an experiment

Relevance of design of data collection

We illustrate a variety of possible missing-data mechanism by considering a series of variations on a simple example

$$\begin{aligned} y_i &= \theta + \epsilon_i,\ i = 1, \ldots, 100 \\ \epsilon_i &\sim N(0,1) \\ p(\theta) &= \text{non-informative} \end{aligned}$$

- θ is the true weigh
- y_i measurement with an electronic scale

1. *With probability .1 the scale fails to report a value, and we observe 91 values*

   Let $I_i = 1$ if y_i is observed and 0 otherwise, and let $\bar{y}_{obs}$ the mean of the observed measurements.

$$\begin{aligned} I_i &\sim \text{Bernoulli}(.9) \\ p(\theta \mid y_{obs}, I) &= p(\theta \mid y_{obs}) = N(\theta \mid \bar{y}_{obs}, 1/91) \end{aligned}$$

2. *With probability π the scale fails to report a value, and we observe 91 values*

$$\begin{aligned} I_i &\sim \text{Bernoulli}(\pi) \\ p(\theta, \pi \mid y_{obs}, I) &\propto p(\theta, \pi) N(\theta \mid \bar{y}_{obs}, 1/91) \\ &\quad \times \text{Bin}(n \mid 100, \pi) \end{aligned}$$

   if θ and π are independent, same solution as 1).

   if $\pi = \theta/(1+\theta)$, then

$$p(\theta, \pi \mid y_{obs}, I) \propto N(\theta \mid \bar{y}_{obs}, 1/91) \text{Bin}(n \mid 100, \tfrac{\theta}{1+\theta})$$

   given n=91 and $\bar{y}_{obs}$, this density can be calculated numerically over a range of θ, and then simulations of θ can be drawn using the inverse-cdF method.

3. *Censoring: all values above 200 kg are reported as "too heavy"*

$I_i \mid y_i = 1$ if $y \geq 200$ and 0 otherwise.

$$
\begin{aligned}
p(\theta \mid y_{obs}, I) &\propto p(\theta) \int p(y_{obs}, y_{mis}, I \mid \theta) dy_{mis} \\
&\propto N(\theta \mid \bar{y}_{obs}, 1/91)[\Phi(\theta - 200)]^9
\end{aligned}
$$

4. *Censoring point unknown: all values above ϕ kg are reported as "too heavy"*

$$p(\theta \mid y_{obs}, I) \propto p(\phi \mid \theta) N(\theta \mid \bar{y}_{obs}, 1/91)[\Phi(\theta - \phi)]^9$$

5. *Truncation: you have the 91 observed values, but not the number of times that the object was weighed. In addition you known that no values are reported by the scale over 200*

$$
\begin{aligned}
p(\theta \mid y_{obs}, I) &\propto p(\theta) N(\theta \mid \bar{y}_{obs}, 1/91) \sum_{N=91}^{\infty} p(N \mid \\
&\times \begin{pmatrix} N \\ 91 \end{pmatrix} [\Phi(\theta - \phi)]^{N-91} \\
p(\theta, N) &\propto \frac{1}{N} \\
p(\theta \mid y_{obs}, I) &= N(\theta \mid \bar{y}_{obs}, 1/91)[1 - \Phi(\theta - \phi)]^{-91}
\end{aligned}
$$

- Do these data collection mechanisms influence the posterior distribution of θ?
- We need to expand the sample space to include, in addition to the population data y, an indicator variable I for whether each element y is observed or not.

Formal models for data collection

We divide the modeling tasks into two parts

- modeling the complete data y
- modeling the observation variable I which indexes which potential data are observed

Notation for observed and missing data

- $y = (y_1, \ldots, y_N)$ matrix of potential data
- $I = (I_1, \ldots, I_N)$ matrix of indicators

  $I_{ij} = 1$ if y_{ij} is observed

  $I_{ij} = 0$ if y_{ij} is missing
- obs=$\{i, j : I_{ij} = 1\}$
- mis=$\{i, j : I_{ij} = 0\}$
- y_{obs}, y_{mis} collection of elements of y that are observed and missing

Complete data likelihood and observed data likelihood

Complete-data likelihood

$p(y, I \mid \theta, \phi) = p(y \mid \theta) p(I \mid y, \phi)$

- $p(y \mid \theta)$ is the model for the underlying data
- $p(I \mid y, \phi)$ is the model for the inclusion vector I
- θ parameter of interest
- ϕ index of missingness of the model

but.. the actual information available is (y_{obs}, I)

Observed-data likelihood

$p(y_{obs}, I \mid \theta, \phi) = \int p(y, I \mid \theta, \phi) dy_{mis}$

if fully observed covariates x are available, all these expressions are conditional on x.

Joint posterior distribution of θ and ϕ

- Complete-data likelihood with covariates x

$$p(y, I \mid x, \theta, \phi) = p(y \mid x, \theta) p(I \mid x, y, \phi)$$

- joint posterior

$$\begin{aligned} p(\theta, \phi \mid x, y_{obs}, I) &\propto p(\theta, \phi \mid x) \\ &\quad \times \int p(y \mid x, \theta) p(I \mid x, y, \theta) dy_{mis} \end{aligned}$$

- marginal posterior

$$\begin{aligned} p(\theta \mid x, y_{obs}, I) &\propto p(\theta, \mid x) \\ &\quad \times \int \int p(\theta, \phi \mid x) p(y \mid x, \theta) \\ &\qquad p(I \mid x, y, \theta) dy_{mis} d\phi \end{aligned}$$

- Goal: draw posterior simulations of the joint vector of the unknowns (y_{mis}, θ, ϕ)

Censored data

Example 4: censoring point unknown

- $y = (y_1, \ldots, y_{100})$ original uncensored weighings
- $y_{obs} = (y_{1,obs}, \ldots, y_{91,obs})$ observed information
- $I = (I_1, \ldots, I_{100})$ inclusion vector composed of 91 ones and 9 zero
- complete data-likelihood $\prod_{i=1}^{100} N(y_i \mid \theta, 1)$
- likelihood of the inclusion vector

$$p(I \mid y, \phi) = \prod_{i=1}^{100} p(I_i \mid y_i, \theta)$$

$$\prod_{i=1} \begin{cases} 1 & \text{if}(I_i = 1) \text{ and } y_i \leq \phi \text{ or} \\ & \text{if}(I_i = 0) \text{ and } y_i \geq \phi \\ 0 & \text{otherwise} \end{cases}$$

- Joint likelihood of y_{obs} and I

$$\begin{aligned} p(y_{obs}, I \mid \theta, \phi) &= \int p(y, I \mid \theta, \phi) dy_{mis} \\ &= \left[\prod_{i=1}^{91} N(y_{obs,i} \mid \theta, 1)\right] \\ &\quad \times [\Phi(\theta - \phi)]^9 \end{aligned}$$

$$p(\theta, \phi \mid y_{obs}, I) \propto p(\theta, \phi) p(y_{obs}, I \mid \theta, \phi)$$

the unknown ϕ cannot be ignored in making inferences about θ

Finite population and superpopulation inference

1. Finite population quantities: summaries of the complete data
2. Superpopulation quantities: summaries of the underlying parameters

It is usually convenient to divide our analysis and computation in 2 steps:

1. superpopulation inference: $p(\theta, \phi \mid x, y_{obs}, I)$
2. finite population inference: $p(y_{mis} \mid x, y_{obs}, \theta, \phi)$

Posterior simulations of y_{mis} and y_{obs}

- Posterior simulations of y_{mis} from its posterior distribution are called "multiple imputations"

$$\begin{aligned} \theta^l, \phi^l &\sim p(\theta, \phi \mid x, y_{obs}, I) \\ y_{mis} &\sim p(y_{mis} \mid x, y_{obs}, I, \theta^l, \phi^l) \end{aligned}$$

Posterior predictive distributions

- predicting future complete data $\tilde{y}$ — it depends only on the available data distribution $p(y \mid x, \theta)$ and the posterior distribution of θ
- predicting future observed data $\tilde{y}_{obs}$ — it depends also on the data collection mechanism $p(I \mid x, y, \phi)$

Ignorability

Study design or the missing data mechanism is Ignorable if $p(\theta \mid x, y_{obs}, I) = p(\theta \mid x, y_{obs})$.

In this case the posterior inferences on θ and the predictive distribution of y_{mis} are determinated by the specification of the model, $p(\theta \mid x)p(y \mid \theta, x)$ and by the observed values y_{obs}:

$$\begin{aligned} p(\theta \mid x, y_{obs}) &= p(\theta \mid x)p(y_{obs} \mid x, \theta) \\ &\propto p(\theta \mid x) \int p(y \mid x, \theta) dy_{mis} \end{aligned}$$

The censored data example with unknown censoring point is not an Ignorable design.... show that

Missing at random and distinct parameters

Two conditions are sufficient to ensure Ignorability of the missing data mechanism for the Bayesian analysis

Missing at random: $p(I \mid x, y, \phi) = p(I \mid x, y_{obs}, \phi)$.

Distinct parameters: $$\begin{aligned} p(\phi, \theta) &= p(\phi)p(\theta) \\ p(\phi \mid x, \theta) &= p(\phi \mid x) \end{aligned}$$

Missing at random $+$

Distinct parameters $=$

Ignorability

Ignorability and Bayesian inference: summary

> For all Ignorable designs, with fixed data and fixed models for the data, the data collection process does not influence Bayesian inference

- The data collection does not influence Bayesian inference only for all the Ignorable designs
- Even with fixed likelihood function $p(y \mid \theta)$ and with fixed data y, the posterior distribution does vary with different "non-Ignorable" mechanism. *example 7.2*
- known mechanisms: data collection processes that follow a known parametric family

Ignorable and known designs

- $p(I \mid x, y, \phi) = p(I \mid x, y_{obs})$
- there is not an unknown parameter ϕ because there is a single acceptable specification for $p(I \mid x, y)$
- we need only $p(y \mid \theta)$ and $p(\theta)$ for inference

Ignorable and known design: random sampling

- y_i weekly amount spent on food by the i-th person, $i = 1, \dots, N$
- Aim: to estimate the average weekly spending on food in the population: $\bar{y} = 1/N \sum_{i=1}^{N} y_i$
- we estimate $\bar{y}$ by a sample of $n < N$ persons
- $I = (I_1, \dots, I_N)$ vector of indicators for whether or not person i is included in the sample
- Formally, simple random sampling is defined by

$$p(I \mid y, \phi) = p(I) = \begin{cases} \binom{N}{n}^{-1} & \text{if} \sum I_i = n \\ 0 & \text{otherwise} \end{cases}$$

This method is Ignorable and known

Bayesian Inference for superpopulation and population estimands

- posterior inference is based on

$$p(\theta \mid y_{obs}, I) \propto p(\theta) p(y_{obs} \mid \theta)$$

- the estimands $\bar{y}$ can be expressed as

$$\bar{y} = \frac{n}{N} \bar{y}_{obs} + \frac{N-n}{N} \bar{y}_{mis} \quad (\star)$$

Simulations from $p(\bar{y} \mid \bar{y}_{obs}, \theta)$

1. draw $\theta^l \sim p(\theta \mid y)$
2. draw $y_{mis}^l \sim p(y_{mis} \mid \theta^l, y_{obs}) = \prod_{i: I_i = 0} p(y_i \mid \theta^l)$. Then $\bar{y}_{mis}^l \sim p(y_{mis} \mid y_{obs})$
3. because $\bar{y}_{obs}$ is known we can easy compute draws $\bar{y}^l$ from $p(\bar{y} \mid y_{obs})$ by using $(\star)$

Large sample equivalence of superpopulation and population inferences

- $N-n$ large

$$p(\bar{y}_{mis} \mid \theta) \simeq N\left(\bar{y}_{mis} \mid \mu, \frac{\sigma^2}{N-n}\right)$$

$\mu = E(y_i \mid \theta),\ \sigma^2 = Var(y_i \mid \theta)$

- n large

$$\begin{aligned} p(\bar{y}_{mis} \mid y_{obs}) &\simeq N(\bar{y}_{mis} \mid \bar{y}_{obs}, \tfrac{N}{n(N-n)} s^2_{obs}) \\ p(\bar{y} \mid y_{obs}) &\simeq N(\bar{y} \mid \bar{y}_{obs}, \tfrac{N-n}{nN} s^2_{obs}) \end{aligned}$$

- Normal model + non informative prior

$$p(\bar{y} \mid y_{obs}) = t_{n-1}(\bar{y} \mid \bar{y}_{obs}, \frac{N-n}{nN} s^2_{obs})$$

Completely randomized experiments

- $n/2$ units receive treatement A or B
- outcomes $(y_i^A, y_i^B),]; i = 1, \ldots, n$
- casual effect of A versus B: $y_i^A - y_i^B$
- average casual effect: $E(y_i^A \mid \theta) - E(y_i^B \mid \theta)$
- finitre population effect: $\bar{y}^A - \bar{y}^B$
- data collection indicator $I_i = (I_i^A, I_i^B),\ i = 1, \ldots, n$
  where $I_i = \{(1,0), (0,1), (0,0)\}$
- this treatment assignment is known and Ignorable

$$p(I \mid y, \phi) = \begin{cases} \binom{n}{n/2}^{-1} & \text{if } \sum I_i^A = \sum I_i^B = n/2 \\ 0 & \text{otherwise} \end{cases}$$

Bayesian Inference for superpopulation and population estimands

- given that the treatment assignment is Ignorable

$$\begin{aligned} p(\theta \mid y_{obs}, I) &\propto p(\theta) p(y_{obs} \mid \theta) \\ p(y_{obs} \mid \theta) &= \textstyle\prod_{i:I_i=(1,0)} p(y_i^A \mid \theta) \\ &\times \textstyle\prod_{i:I_i=(0,1)} p(y_i^B \mid \theta) \end{aligned}$$

- simulations from $p(\bar{y}^A_{mis}, \bar{y}^B_{mis} \mid \bar{y}_{obs}, \theta)$

draw $\theta^l \sim p(\theta \mid y)$

draw $y^l_{mis} \sim p(y_{mis} \mid \theta, y_{obs})$

where

$$\begin{aligned} p(y_{mis} \mid \theta, y_{obs}) &\propto \textstyle\prod_{i:I_i=(1,0)} p(y_i^A \mid \theta^l, y_i^B) \\ &\times \textstyle\prod_{i:I_i=(0,1)} p(y_i^B \mid \theta^l, y_i^A) \end{aligned}$$

- need a model for $(y_i^A, y_i^B \mid \theta)$ ..*ex 7.9*

Large sample correspondence

- $\bar{y}^A - \bar{y}^B$ are sensitive to $p(y_i^A, y_i^B \mid \theta)$ and
  to $corr(y_i^A, y_i^B \mid \theta)$ for which no data are available
- if the n units are sampled from a large population of N units then the sensitivity vanishes, in fact
  for n and N/n large

$$p(\bar{y}^A - \bar{y}^B \mid y_{obs}) \simeq N(\bar{y}^A_{obs} - \bar{y}^B_{obs}, \tfrac{2}{n}(s^{2^A}_{obs} + s^{2^B}_{obs})$$

Home work

Suppose that someone else weigh an object 100 times on an electronic scale with known $N(\theta, 3)$ error distribution. This person provides to you also the 91 observed values, but not the number of times the object was weighed.

Also suppose that we known that no values are reported by the scale over ϕ, where the truncation point ϕ is unknown.

Perform a Bayesian analysis assuming $p(\theta, \phi) = N(\theta, 2)$ and $p(\theta, \phi)$ non informative.

Table 1: *Use of observed and missing data notation for various data structures*

Example	Observed data	Complete data
Sampling	Values from the n units in the sample	Values from all N units in the population
Experiment	Outcomes under the observed treatment for each unit treated	Outcomes under all treatment for all units
Rounded data	Rounded observations	Precise values of all observations
Unintentional missing data	observed data values	Complete data, both observed and missing