\textbf{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

\textbf{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

\textbf{Given}

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

then:

\textit{Bayesian inference is determined regardless of the design of the collection of the data.}

This is a misplaced appeal of the likelihood principle

\textbf{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 as 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

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

\[ y_i = \theta + \epsilon_i, \quad i = 1, \ldots, 100 \]

\[ \epsilon_i \sim N(0,1) \]

\[ p(\theta) = \text{non-informative} \]

- \( \theta \) is the true weight
- \( y_i \) 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.

\[ I_i \sim \text{Bernoulli}(0.9) \]

\[ p(\theta \mid y_{obs}, I) = p(\theta \mid y_{obs}) = N(\theta \mid \bar{y}_{obs}, 1/91) \]

2. With probability \( \pi \) the scale fails to report a value, and we observe 91 values

Let \( I_i \sim \text{Bernoulli}(\pi) \)

\[ p(\theta, \pi \mid y_{obs}, I) \propto p(\theta, \pi) p(\theta \mid \bar{y}_{obs}, 1/91) \times \text{Bin}(n \mid 100, \pi) \]

if \( \theta \) and \( \pi \) are independent, same solution as 1).

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

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

given \( n=91 \) and \( \bar{y}_{obs} \), this density can be calculated numerically over a range of \( \theta \), and then simulations of \( \theta \) can be drawn using the inverse-cdf method.
3. Censoring: all values above 200 kg are reported as "too heavy"

\[ I_i | y_i = 1 \text{ if } y \geq 200 \text{ and } 0 \text{ otherwise.} \]

\[
p(\theta | y_{\text{obs}}, I) \propto p(\theta) \int p(y_{\text{obs}}, y_{\text{mis}}, I | \theta) dy_{\text{mis}}
\]

\[
\propto N(\theta | \bar{y}_{\text{obs}}, 1/91)[\Phi(\theta - 200)]^9
\]

4. Censoring point unknown: all values above \( \phi \) kg are reported as "too heavy"

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

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

\[
p(\theta | y_{\text{obs}}, I) \propto p(\theta)N(\theta | \bar{y}_{\text{obs}}, 1/91) \sum_{N=0}^{\infty} p(N | \theta) \times \left( \frac{N}{91} \right)[\Phi(\theta - \phi)]^{N-91}
\]

\[
p(\theta, N) \propto \frac{1}{N}
\]

\[
p(\theta | y_{\text{obs}}, I) = N(\theta | \bar{y}_{\text{obs}}, 1/91)[1 - \Phi(\theta - \phi)]^{-91}
\]

---

- Do these data collection mechanisms influence the posterior distribution of \( \theta \)?
- 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
  \[
  I_{ij} = 1 \text{ if } y_{ij} \text{ is observed}
  \]
  \[
  I_{ij} = 0 \text{ if } y_{ij} \text{ is missing}
  \]
- \( \text{obs} = \{i, j : I_{ij} = 1\} \)
- \( \text{mis} = \{i, j : I_{ij} = 0\} \)
- \( y_{\text{obs}}, y_{\text{mis}} \) collection of elements of \( y \) that are observed and missing
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 \( I \)
- \( \theta \) parameter of interest
- \( \phi \) index of missingness of the model

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

Observed data likelihood

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

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

Censored data

Example 4: censoring point unknown

- \( y = (y_1, \ldots, y_{100}) \) original uncensored weighings
- \( y_{\text{obs}} = (y_{1,\text{obs}}, \ldots, y_{91,\text{obs}}) \) 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

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

\[
\prod_{i=1}^{100} \left\{
\begin{array}{ll}
1 & \text{if}(I_i = 1) \text{ and } y_i \leq \phi \text{ or } \\
0 & \text{if}(I_i = 0) \text{ and } y_i \geq \phi \\
\end{array}
\right.
\]

Joint posterior distribution of \( \theta \) and \( \phi \)

- 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

\[ p(\theta, \phi \mid x, y_{\text{obs}}, I) \propto p(\theta, \phi \mid x) \]
\[ \times \int p(y \mid x, \theta)p(I \mid x, y, \theta)dy_{\text{mis}} \]

- marginal posterior

\[ p(\theta \mid x, y_{\text{obs}}, I) \propto p(\theta \mid x) \]
\[ \times \int \int p(\theta, \phi \mid x)p(y \mid x, \theta) \]
\[ p(I \mid x, y, \theta)dy_{\text{mis}}d\phi \]

- Goal: draw posterior simulations of the joint vector of the unknowns \((y_{\text{mis}}, \theta, \phi)\)

- Joint likelihood of \( y_{\text{obs}} \) and \( I \)

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

\[
= \left[ \prod_{i=1}^{91} N(y_{\text{obs},i} \mid \theta, 1) \right]
\]

\[
\times [\Phi(\theta - \phi)]^{91}
\]

\[ p(\theta, \phi \mid y_{\text{obs}}, I) \propto p(\theta, \phi)p(y_{\text{obs}}, I \mid \theta, \phi) \]

the unknown \( \phi \) cannot be ignored in making inferences about \( \theta \)
Posterior simulations of $y_{mis}$ and $y_{obs}$

- Posterior simulations of $y_{mis}$ from its posterior distribution are called "multiple imputations"
  
  $$\theta^l, \phi^l \sim p(\theta, \phi \mid x, y_{obs}, I)$$
  
  $$y_{mis} \sim p(y_{mis} \mid x, y_{obs}, I, \theta^l, \phi^l)$$

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 $\theta$
- 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 $\theta$ and the predictive distribution of $y_{mis}$ are determined by the specification of the model, $p(\theta \mid x)p(y \mid \theta, x)$ and by the observed values $y_{obs}$:

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

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:

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

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

Missing at random +

Distinct parameters = ignorability
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 design: random sampling

- \( y_i \) weekly amount spent on food by the \( i \)-th person, \( i = 1, \ldots, N \)
- Aim: to estimate the average weekly spending on food in the population: \( \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \)
- We estimate \( \bar{y} \) by a sample of \( n < N \) persons
- \( I = (I_1, \ldots, I_N) \) 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

\[
\bar{y} = \frac{n}{N} \bar{y}_{obs} + \frac{N-n}{N} \bar{y}_{mis} \quad (*)
\]

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

1. Draw \( \theta^l \sim p(\theta \mid y) \)
2. Draw \( y_{mis}^l \sim p(y_{mis} \mid \theta^l, y_{obs}) = \Pi_{i; I_i=0} p(y_i \mid \theta^l) \).
   Then \( \bar{y}_{mis}^l \sim p(y_{mis} \mid y_{obs}) \)
3. Because \( \bar{y}_{obs} \) is known, we can easily compute draws \( \bar{y}^l \) from \( p(\bar{y} \mid y_{obs}) \) by using \( (*) \).
Large sample equivalence of superpopulation and population inferences

- $N - n$ large

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

\[
\mu = E(y_i \mid \theta), \quad \sigma^2 = Var(y_i \mid \theta)
\]

- $n$ large

\[
p(\bar{y}_{mis} \mid y_{obs}) \sim N(\bar{y}_{mis} \mid \bar{y}_{obs}, \frac{N}{n(N - n)} s_{obs}^2)
\]

\[
p(\bar{y} \mid y_{obs}) \sim N(\bar{y} \mid \bar{y}_{obs}, \frac{N - n}{nN} s_{obs}^2)
\]

- Normal model + non informative prior

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

Bayesian Inference for superpopulation and population estimands

- given that the treatment assignment is ignorable

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

\[
p(y_{obs} \mid \theta) = \prod_{i \in I_{i=0}} p(y_i^A \mid \theta)
\]

\[
\quad \times \prod_{i \in I_{i=1}} p(y_i^B \mid \theta)
\]

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

\[
draw \theta^l \sim p(\theta \mid y)
\]

\[
draw y_{mis}^l \sim p(y_{mis} \mid \theta, y_{obs})
\]

where

\[
p(y_{mis} \mid \theta, y_{obs}) \propto \prod_{i \in I_{i=0}} p(y_i^A \mid \theta^l, y_i^B)
\]

\[
\quad \times \prod_{i \in I_{i=1}} p(y_i^B \mid \theta^l, y_i^A)
\]

- 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

\[
p(\bar{y}^A - \bar{y}^B \mid y_{obs}) \sim N(\bar{y}_{obs}^A - \bar{y}_{obs}^B, 2s_{obs}^2 + s_{obs}^2)
\]
Suppose that someone else weighs 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 $\phi$, where the truncation point $\phi$ is unknown.

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

<table>
<thead>
<tr>
<th>Example</th>
<th>Observed data</th>
<th>Complete data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling</td>
<td>Values from the $n$ units in the sample</td>
<td>Values from all $N$ units in the population</td>
</tr>
<tr>
<td>Experiment</td>
<td>Outcomes under the observed treatment for each unit treated</td>
<td>Outcomes under all treatment for all units</td>
</tr>
<tr>
<td>Rounded data</td>
<td>Rounded observations</td>
<td>Precise values of all observations</td>
</tr>
<tr>
<td>Unintentional missing data</td>
<td>observed data values</td>
<td>Complete data, both observed and missing</td>
</tr>
</tbody>
</table>