Statistical Methods for Analysis with Missing Data

Lecture 8: introduction to Bayesian inference

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY of WASHINGTON

Previous Lecture

EM algorithm for maximum likelihood estimation with missing data:

- Derivation and implementation of EM algorithm for categorical data
- For two categorical variables, EM under MAR is intuitive and simple

$$\pi_{kl}^{(t+1)} = \frac{1}{n}\left(n_{11kl} + \frac{\pi_{kl}^{(t)}}{\pi_{k+}^{(t)}} n_{10k+} + \frac{\pi_{kl}^{(t)}}{\pi_{+l}^{(t)}} n_{01+l} + \pi_{kl}^{(t)} n_{00++}\right),$$

where

$$n_{11kl} = \sum_i W_{ikl} I(r_i = 11), \quad n_{10k+} = \sum_i W_{ik+} I(r_i = 10),$$
$$n_{01+l} = \sum_i W_{i+l} I(r_i = 01), \quad n_{00++} = \sum_i I(r_i = 00)$$

- Bootstrap confidence intervals

Today's Lecture

- Introduction to Bayesian inference. Why?
  - Technique called *multiple imputation* is derived from a Bayesian point of view
  - Variation of multiple imputation called *multiple imputation by chained equations* imitates *Gibbs sampling* – a procedure commonly used for Bayesian inference
  - Bayesian inference has its own ways of handling missing data

Outline

Bayes' Theorem

Bayesian Inference

Bayes' Theorem

For events A and B:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- $P(A \mid B)$: conditional probability of event A given that B is true
- $P(B \mid A)$: conditional probability of event B given that A is true
- $P(A)$ and $P(B)$: unconditional/marginal probabilities of events A and B

Example: Using Bayes' Theorem

- A: person has a given condition
- B: person tests positive for condition using cheap test
- $P(B \mid A)$, $P(B \mid A^c)$: known from experimental testing
- $P(A)$: known from study based on gold standard
- $P(A \mid B) = P(B \mid A)P(A)/P(B)$: probability of having the condition given positive result in cheap test

Example: Using Bayes' Theorem

- A: person has a given condition
- B: person tests positive for condition using cheap test
- $P(B \mid A) = 0.99$: *sensitivity* of the test
- $P(B^c \mid A^c) = 0.99$: *specificity* of the test
- $P(A) = 0.01$: rare condition
- *Given that a generic person tests positive, what is the probability that they have the condition?*

$$\begin{aligned} P(A \mid B) &= \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)} \\ &= \frac{.99 \times .01}{.99 \times .01 + .01 \times .99} = 0.5 \end{aligned}$$

Bayes' Theorem

In the above example, we are implicitly working with two binary variables

- $Y = I(\text{having medical condition})$
- $X = I(\text{testing positive})$

with a joint density $p(X = x, Y = y)$

- Bayes' theorem simply relates the conditional probabilities $p(X = x \mid Y = y)$ and $p(Y = y \mid X = x)$

Bayes' Theorem

In general

- Random vectors X and Y
- Joint density $p(x, y)$
- Relationship between conditionals is given by Bayes' theorem

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

- This is a mathematical fact about conditional probabilities/densities
- Applying Bayes' theorem doesn't make you *Bayesian*!
- The *Bayesian* approach arises from applying this theorem to obtain statistical inferences on model parameters!

Bayes' Theorem

In general

- Random vectors X and Y
- Joint density $p(x, y)$
- Relationship between conditionals is given by Bayes' theorem

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

- This is a mathematical fact about conditional probabilities/densities
- Applying Bayes' theorem doesn't make you *Bayesian*!
- The *Bayesian* approach arises from applying this theorem to obtain statistical inferences on model parameters!

Bayes' Theorem

In general

- Random vectors X and Y
- Joint density $p(x, y)$
- Relationship between conditionals is given by Bayes' theorem

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

- This is a mathematical fact about conditional probabilities/densities
- Applying Bayes' theorem doesn't make you *Bayesian*!
- The *Bayesian* approach arises from applying this theorem to obtain statistical inferences on model parameters!

Outline

Bayes' Theorem

Bayesian Inference

Philosophical Motivation

- Bayesian paradigm: model parameters are treated as *random variables* to represent uncertainty about their true values
- Compare with the frequentist paradigm: model parameters are unknown *constants*
- Bayesian analysis requires:
  - Likelihood function, coming from a model for the distribution of the data
  - *Prior distribution* on model parameters, coming from previous knowledge about the phenomenon under study
- Prior distribution is *updated* using the likelihood via *Bayes' theorem*, resulting in a *posterior distribution*
- Bayesian inferences are drawn from posterior distribution (updated knowledge)

Philosophical Motivation

- Bayesian paradigm: model parameters are treated as *random variables* to represent uncertainty about their true values
- Compare with the frequentist paradigm: model parameters are unknown *constants*
- Bayesian analysis requires:
  - Likelihood function, coming from a model for the distribution of the data
  - *Prior distribution* on model parameters, coming from previous knowledge about the phenomenon under study
- Prior distribution is *updated* using the likelihood via *Bayes' theorem*, resulting in a *posterior distribution*
- Bayesian inferences are drawn from posterior distribution (updated knowledge)

Philosophical Motivation

- Bayesian paradigm: model parameters are treated as *random variables* to represent uncertainty about their true values
- Compare with the frequentist paradigm: model parameters are unknown *constants*
- Bayesian analysis requires:
  - Likelihood function, coming from a model for the distribution of the data
  - *Prior distribution* on model parameters, coming from previous knowledge about the phenomenon under study
- Prior distribution is *updated* using the likelihood via *Bayes' theorem*, resulting in a *posterior distribution*
- Bayesian inferences are drawn from posterior distribution (updated knowledge)

Philosophical Motivation

- Bayesian paradigm: model parameters are treated as *random variables* to represent uncertainty about their true values
- Compare with the frequentist paradigm: model parameters are unknown *constants*
- Bayesian analysis requires:
  - Likelihood function, coming from a model for the distribution of the data
  - *Prior distribution* on model parameters, coming from previous knowledge about the phenomenon under study
- Prior distribution is *updated* using the likelihood via *Bayes' theorem*, resulting in a *posterior distribution*
- Bayesian inferences are drawn from posterior distribution (updated knowledge)

Philosophical Motivation

- Bayesian paradigm: model parameters are treated as *random variables* to represent uncertainty about their true values
- Compare with the frequentist paradigm: model parameters are unknown *constants*
- Bayesian analysis requires:
  - Likelihood function, coming from a model for the distribution of the data
  - *Prior distribution* on model parameters, coming from previous knowledge about the phenomenon under study
- Prior distribution is *updated* using the likelihood via *Bayes' theorem*, resulting in a *posterior distribution*
- Bayesian inferences are drawn from posterior distribution (updated knowledge)

Practical Motivation

Bayesian *machinery* might make this approach appealing

- Quantification of uncertainty in large discrete parameter spaces
  - Partitions in clustering problems
  - Graphs in graphical models
  - Binary vectors of variable inclusion in regression model selection
  - Bipartite matchings in record linkage
- Frequentist approach calls for constructing confidence sets, whereas Bayesian approach works with a posterior distribution from which we can sample to approximate summaries of interest
- Under some conditions, posteriors behave like sampling distribution of MLEs, but Bayesian machinery might be easier to implement than deriving estimate of asymptotic covariance matrix of MLE
- Implementing a Monte Carlo EM algorithm to obtain MLEs is similar to implementing Data Augmentation (coming soon), but the latter readily provide you with measures of uncertainty
- Convenient in hierarchical/multilevel models – priors just add another level to the hierarchy

Practical Motivation

Bayesian *machinery* might make this approach appealing

- Quantification of uncertainty in large discrete parameter spaces
  - Partitions in clustering problems
  - Graphs in graphical models
  - Binary vectors of variable inclusion in regression model selection
  - Bipartite matchings in record linkage
- Frequentist approach calls for constructing confidence sets, whereas Bayesian approach works with a posterior distribution from which we can sample to approximate summaries of interest
- Under some conditions, posteriors behave like sampling distribution of MLEs, but Bayesian machinery might be easier to implement than deriving estimate of asymptotic covariance matrix of MLE
- Implementing a Monte Carlo EM algorithm to obtain MLEs is similar to implementing Data Augmentation (coming soon), but the latter readily provide you with measures of uncertainty
- Convenient in hierarchical/multilevel models – priors just add another level to the hierarchy

Practical Motivation

Bayesian *machinery* might make this approach appealing

- Quantification of uncertainty in large discrete parameter spaces
  - Partitions in clustering problems
  - Graphs in graphical models
  - Binary vectors of variable inclusion in regression model selection
  - Bipartite matchings in record linkage
- Frequentist approach calls for constructing confidence sets, whereas Bayesian approach works with a posterior distribution from which we can sample to approximate summaries of interest
- Under some conditions, posteriors behave like sampling distribution of MLEs, but Bayesian machinery might be easier to implement than deriving estimate of asymptotic covariance matrix of MLE
- Implementing a Monte Carlo EM algorithm to obtain MLEs is similar to implementing Data Augmentation (coming soon), but the latter readily provide you with measures of uncertainty
- Convenient in hierarchical/multilevel models – priors just add another level to the hierarchy

Practical Motivation

Bayesian *machinery* might make this approach appealing

- Quantification of uncertainty in large discrete parameter spaces
  - Partitions in clustering problems
  - Graphs in graphical models
  - Binary vectors of variable inclusion in regression model selection
  - Bipartite matchings in record linkage
- Frequentist approach calls for constructing confidence sets, whereas Bayesian approach works with a posterior distribution from which we can sample to approximate summaries of interest
- Under some conditions, posteriors behave like sampling distribution of MLEs, but Bayesian machinery might be easier to implement than deriving estimate of asymptotic covariance matrix of MLE
- Implementing a Monte Carlo EM algorithm to obtain MLEs is similar to implementing Data Augmentation (coming soon), but the latter readily provide you with measures of uncertainty
- Convenient in hierarchical/multilevel models – priors just add another level to the hierarchy

Practical Motivation

Bayesian *machinery* might make this approach appealing

- Quantification of uncertainty in large discrete parameter spaces
  - Partitions in clustering problems
  - Graphs in graphical models
  - Binary vectors of variable inclusion in regression model selection
  - Bipartite matchings in record linkage
- Frequentist approach calls for constructing confidence sets, whereas Bayesian approach works with a posterior distribution from which we can sample to approximate summaries of interest
- Under some conditions, posteriors behave like sampling distribution of MLEs, but Bayesian machinery might be easier to implement than deriving estimate of asymptotic covariance matrix of MLE
- Implementing a Monte Carlo EM algorithm to obtain MLEs is similar to implementing Data Augmentation (coming soon), but the latter readily provide you with measures of uncertainty
- Convenient in hierarchical/multilevel models – priors just add another level to the hierarchy

Practical Motivation

Bayesian *machinery* might make this approach appealing

- Quantification of uncertainty in large discrete parameter spaces
  - Partitions in clustering problems
  - Graphs in graphical models
  - Binary vectors of variable inclusion in regression model selection
  - Bipartite matchings in record linkage
- Frequentist approach calls for constructing confidence sets, whereas Bayesian approach works with a posterior distribution from which we can sample to approximate summaries of interest
- Under some conditions, posteriors behave like sampling distribution of MLEs, but Bayesian machinery might be easier to implement than deriving estimate of asymptotic covariance matrix of MLE
- Implementing a Monte Carlo EM algorithm to obtain MLEs is similar to implementing Data Augmentation (coming soon), but the latter readily provide you with measures of uncertainty
- Convenient in hierarchical/multilevel models – priors just add another level to the hierarchy

Practical Motivation

Bayesian *machinery* might make this approach appealing

- Quantification of uncertainty in large discrete parameter spaces
  - Partitions in clustering problems
  - Graphs in graphical models
  - Binary vectors of variable inclusion in regression model selection
  - Bipartite matchings in record linkage
- Frequentist approach calls for constructing confidence sets, whereas Bayesian approach works with a posterior distribution from which we can sample to approximate summaries of interest
- Under some conditions, posteriors behave like sampling distribution of MLEs, but Bayesian machinery might be easier to implement than deriving estimate of asymptotic covariance matrix of MLE
- Implementing a Monte Carlo EM algorithm to obtain MLEs is similar to implementing Data Augmentation (coming soon), but the latter readily provide you with measures of uncertainty
- Convenient in hierarchical/multilevel models – priors just add another level to the hierarchy

Practical Motivation

Bayesian *machinery* might make this approach appealing

- Quantification of uncertainty in large discrete parameter spaces
  - Partitions in clustering problems
  - Graphs in graphical models
  - Binary vectors of variable inclusion in regression model selection
  - Bipartite matchings in record linkage
- Frequentist approach calls for constructing confidence sets, whereas Bayesian approach works with a posterior distribution from which we can sample to approximate summaries of interest
- Under some conditions, posteriors behave like sampling distribution of MLEs, but Bayesian machinery might be easier to implement than deriving estimate of asymptotic covariance matrix of MLE
- Implementing a Monte Carlo EM algorithm to obtain MLEs is similar to implementing Data Augmentation (coming soon), but the latter readily provide you with measures of uncertainty
- Convenient in hierarchical/multilevel models – priors just add another level to the hierarchy

Practical Motivation

Bayesian *machinery* might make this approach appealing

- Quantification of uncertainty in large discrete parameter spaces
  - Partitions in clustering problems
  - Graphs in graphical models
  - Binary vectors of variable inclusion in regression model selection
  - Bipartite matchings in record linkage
- Frequentist approach calls for constructing confidence sets, whereas Bayesian approach works with a posterior distribution from which we can sample to approximate summaries of interest
- Under some conditions, posteriors behave like sampling distribution of MLEs, but Bayesian machinery might be easier to implement than deriving estimate of asymptotic covariance matrix of MLE
- Implementing a Monte Carlo EM algorithm to obtain MLEs is similar to implementing Data Augmentation (coming soon), but the latter readily provide you with measures of uncertainty
- Convenient in hierarchical/multilevel models – priors just add another level to the hierarchy

Practical Motivation

Bayesian *machinery* might make this approach appealing

- Quantification of uncertainty in large discrete parameter spaces
  - Partitions in clustering problems
  - Graphs in graphical models
  - Binary vectors of variable inclusion in regression model selection
  - Bipartite matchings in record linkage
- Frequentist approach calls for constructing confidence sets, whereas Bayesian approach works with a posterior distribution from which we can sample to approximate summaries of interest
- Under some conditions, posteriors behave like sampling distribution of MLEs, but Bayesian machinery might be easier to implement than deriving estimate of asymptotic covariance matrix of MLE
- Implementing a Monte Carlo EM algorithm to obtain MLEs is similar to implementing Data Augmentation (coming soon), but the latter readily provide you with measures of uncertainty
- Convenient in hierarchical/multilevel models – priors just add another level to the hierarchy

The Likelihood Function

Same as before:

- $Z = (Z_1, \ldots, Z_K)$: generic vector of study variables
- We work under a *parametric model* for the distribution of Z

$$\{p(z \mid \theta)\}_\theta, \quad \theta = (\theta_1, \theta_2, \ldots, \theta_d)$$

- Data from random i.i.d. vectors $\{Z_i\}_{i=1}^n \equiv \mathbf{Z}$
- Under our parametric model, the joint distribution of $\{Z_i\}_{i=1}^n$ has a density function

$$p(\mathbf{z} \mid \theta) = \prod_{i=1}^n p(z_i \mid \theta)$$

- This, seen as a function of θ, is the likelihood function

$$L(\theta \mid \mathbf{z}) = \prod_{i=1}^n p(z_i \mid \theta)$$

The Likelihood Function

Same as before:

- $Z = (Z_1, \dots, Z_K)$: generic vector of study variables

- We work under a *parametric model* for the distribution of Z

$$\{p(z \mid \theta)\}_\theta, \quad \theta = (\theta_1, \theta_2, \dots, \theta_d)$$

- Data from random i.i.d. vectors $\{Z_i\}_{i=1}^n \equiv \mathbf{Z}$

- Under our parametric model, the joint distribution of $\{Z_i\}_{i=1}^n$ has a density function

$$p(\mathbf{z} \mid \theta) = \prod_{i=1}^n p(z_i \mid \theta)$$

- This, seen as a function of θ, is the likelihood function

$$L(\theta \mid \mathbf{z}) = \prod_{i=1}^n p(z_i \mid \theta)$$

The Likelihood Function

Same as before:

- $Z = (Z_1, \dots, Z_K)$: generic vector of study variables
- We work under a *parametric model* for the distribution of Z

$$\{p(z \mid \theta)\}_\theta, \quad \theta = (\theta_1, \theta_2, \dots, \theta_d)$$

- Data from random i.i.d. vectors $\{Z_i\}_{i=1}^n \equiv \mathbf{Z}$
- Under our parametric model, the joint distribution of $\{Z_i\}_{i=1}^n$ has a density function

$$p(\mathbf{z} \mid \theta) = \prod_{i=1}^n p(z_i \mid \theta)$$

- This, seen as a function of θ, is the likelihood function

$$L(\theta \mid \mathbf{z}) = \prod_{i=1}^n p(z_i \mid \theta)$$

The Prior Distribution

- Prior to observing the realizations of $\mathbf{Z} = \{Z_i\}_{i=1}^{n}$, do we have any information on the parameters θ?

- Represent this prior information in terms of a distribution

$$p(\theta)$$

The Posterior Distribution

Now, "*simply*" use Bayes' theorem

$$p(\theta \mid \mathbf{z}) = \frac{L(\theta \mid \mathbf{z})p(\theta)}{p(\mathbf{z})}$$

$$= \frac{L(\theta \mid \mathbf{z})p(\theta)}{\int L(\theta \mid \mathbf{z})p(\theta)d\theta}$$

$$\propto L(\theta \mid \mathbf{z})p(\theta)$$

"*That's it!*"

The Posterior Distribution

For simple problems, we typically have two ways of computing the posterior $p(\theta \mid \mathbf{z})$

- Compute the integral $\int L(\theta \mid \mathbf{z})p(\theta)d\theta$, and then compute

$$p(\theta \mid \mathbf{z}) = \frac{L(\theta \mid \mathbf{z})p(\theta)}{\int L(\theta \mid \mathbf{z})p(\theta)d\theta}$$

- Stare at / manipulate the expression $L(\theta \mid \mathbf{z})p(\theta)$ seen as a function of θ alone
  - If $L(\theta \mid \mathbf{z})p(\theta) = a(\theta, \mathbf{z})b(\mathbf{z})$, then $p(\theta \mid \mathbf{z}) \propto a(\theta, \mathbf{z})$
  - If $a(\theta, \mathbf{z})$ looks like a known distribution except for a constant, then we have identified the posterior

The Posterior Distribution

For simple problems, we typically have two ways of computing the posterior $p(\theta \mid \mathbf{z})$

- Compute the integral $\int L(\theta \mid \mathbf{z})p(\theta)d\theta$, and then compute

$$p(\theta \mid \mathbf{z}) = \frac{L(\theta \mid \mathbf{z})p(\theta)}{\int L(\theta \mid \mathbf{z})p(\theta)d\theta}$$

- Stare at / manipulate the expression $L(\theta \mid \mathbf{z})p(\theta)$ seen as a function of θ alone
    - If $L(\theta \mid \mathbf{z})p(\theta) = a(\theta, \mathbf{z})b(\mathbf{z})$, then $p(\theta \mid \mathbf{z}) \propto a(\theta, \mathbf{z})$
    - If $a(\theta, \mathbf{z})$ looks like a known distribution except for a constant, then we have identified the posterior

Example: Binomial Data, Beta Prior

Let $Z \mid \theta \sim \text{Binom}(n, \theta)$, and $\theta \sim \text{Beta}(a, b)$

- $L(\theta \mid z) = \binom{n}{z}\theta^z(1-\theta)^{n-z}$, $\quad p(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$

- The proportionality constant is

$$\int L(\theta \mid z)p(\theta)d\theta = \frac{\binom{n}{z}}{B(a,b)}\int \theta^{z+a-1}(1-\theta)^{n-z+b-1}d\theta$$
$$= \binom{n}{z}\frac{B(z+a, n-z+b)}{B(a,b)}$$

- And the posterior is

$$p(\theta \mid z) = \frac{L(\theta \mid z)p(\theta)}{\int L(\theta \mid z)p(\theta)d\theta}$$
$$= \frac{1}{B(z+a, n-z+b)}\theta^{z+a-1}(1-\theta)^{n-z+b-1}$$

- Therefore, $\theta \mid z \sim \text{Beta}(z+a, n-z+b)$

Example: Binomial Data, Beta Prior

Let $Z \mid \theta \sim \text{Binom}(n, \theta)$, and $\theta \sim \text{Beta}(a, b)$

- $L(\theta \mid z) = \binom{n}{z}\theta^z(1-\theta)^{n-z}$, $p(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$

- The proportionality constant is

$$\int L(\theta \mid z)p(\theta)d\theta = \frac{\binom{n}{z}}{B(a,b)}\int \theta^{z+a-1}(1-\theta)^{n-z+b-1}d\theta$$
$$= \binom{n}{z}\frac{B(z+a, n-z+b)}{B(a,b)}$$

- And the posterior is

$$p(\theta \mid z) = \frac{L(\theta \mid z)p(\theta)}{\int L(\theta \mid z)p(\theta)d\theta}$$
$$= \frac{1}{B(z+a, n-z+b)}\theta^{z+a-1}(1-\theta)^{n-z+b-1}$$

- Therefore, $\theta \mid z \sim \text{Beta}(z+a, n-z+b)$

Example: Binomial Data, Beta Prior

Let $Z \mid \theta \sim \text{Binom}(n, \theta)$, and $\theta \sim \text{Beta}(a, b)$

- $L(\theta \mid z) = \binom{n}{z} \theta^z (1-\theta)^{n-z}$, $\quad p(\theta) = \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}$

- The proportionality constant is

$$\begin{aligned} \int L(\theta \mid z) p(\theta) d\theta &= \frac{\binom{n}{z}}{B(a,b)} \int \theta^{z+a-1}(1-\theta)^{n-z+b-1} d\theta \\ &= \binom{n}{z} \frac{B(z+a, n-z+b)}{B(a,b)} \end{aligned}$$

- And the posterior is

$$\begin{aligned} p(\theta \mid z) &= \frac{L(\theta \mid z) p(\theta)}{\int L(\theta \mid z) p(\theta) d\theta} \\ &= \frac{1}{B(z+a, n-z+b)} \theta^{z+a-1}(1-\theta)^{n-z+b-1} \end{aligned}$$

- Therefore, $\theta \mid z \sim \text{Beta}(z+a, n-z+b)$

Example: Binomial Data, Beta Prior

Let $Z \mid \theta \sim \text{Binom}(n, \theta)$, and $\theta \sim \text{Beta}(a, b)$

- $L(\theta \mid z) = \binom{n}{z}\theta^z(1-\theta)^{n-z}$, $\quad p(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$

- The proportionality constant is

$$\int L(\theta \mid z)p(\theta)d\theta = \frac{\binom{n}{z}}{B(a,b)}\int \theta^{z+a-1}(1-\theta)^{n-z+b-1}d\theta$$
$$= \binom{n}{z}\frac{B(z+a, n-z+b)}{B(a,b)}$$

- And the posterior is

$$p(\theta \mid z) = \frac{L(\theta \mid z)p(\theta)}{\int L(\theta \mid z)p(\theta)d\theta}$$
$$= \frac{1}{B(z+a, n-z+b)}\theta^{z+a-1}(1-\theta)^{n-z+b-1}$$

- Therefore, $\theta \mid z \sim \text{Beta}(z+a, n-z+b)$

Example: Binomial Data, Beta Prior

Let $Z \mid \theta \sim \text{Binom}(n, \theta)$ and $\theta \sim \text{Beta}(a, b)$

- $L(\theta \mid z) = \binom{n}{z}\theta^z(1-\theta)^{n-z}, \qquad p(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$

- We could also have noticed

$$\begin{aligned} p(\theta \mid z) &\propto L(\theta \mid z)p(\theta) \\ &\propto \theta^{z+a-1}(1-\theta)^{n-z+b-1} \end{aligned}$$

- This is the non-constant part (*the kernel*) of the density function of a beta random variable with parameters $z+a$ and $n-z+b$, therefore $\theta \mid z \sim \text{Beta}(z+a, n-z+b)$

Example: Binomial Data, Beta Prior

Let $Z \mid \theta \sim \text{Binom}(n, \theta)$ and $\theta \sim \text{Beta}(a, b)$

- $L(\theta \mid z) = \binom{n}{z}\theta^z(1-\theta)^{n-z}, \qquad p(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$

- We could also have noticed

$$\begin{aligned} p(\theta \mid z) &\propto L(\theta \mid z)p(\theta) \\ &\propto \theta^{z+a-1}(1-\theta)^{n-z+b-1} \end{aligned}$$

- This is the non-constant part (*the kernel*) of the density function of a beta random variable with parameters $z + a$ and $n - z + b$, therefore $\theta \mid z \sim \text{Beta}(z + a, n - z + b)$

Example: Binomial Data, Beta Prior

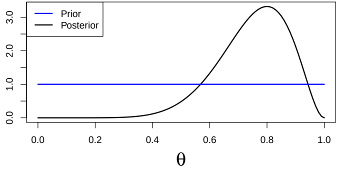
To illustrate the idea, say:

- Someone is flipping a coin $n = 10$ times in an independent and identical fashion
- Number of heads $Z \sim \text{Binomial}(n, \theta)$
- What is the value of θ?
- We use a $\text{Beta}(a, b)$ to express our *prior belief* on θ
- We observe $Z = 8$

Example: Binomial Data, Beta Prior

Possible scenarios of prior information on θ; posteriors with $Z = 8$:

- No idea of what θ could be: Beta(1,1)



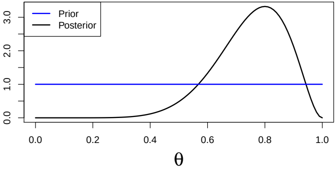
- The person flipping the coin looks like a trickster: Beta(9,1)

- Coin flipping usually has 50/50 chance of heads/tails: Beta(100,100)

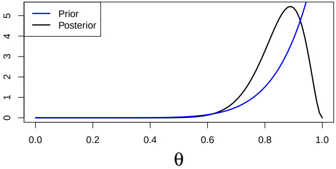
Example: Binomial Data, Beta Prior

Possible scenarios of prior information on θ; posteriors with $Z = 8$:

- No idea of what θ could be: Beta(1,1)



- The person flipping the coin looks like a trickster: Beta(9,1)

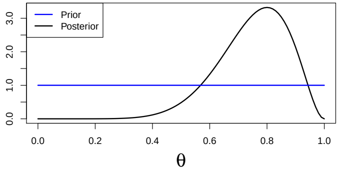


- Coin flipping usually has 50/50 chance of heads/tails: Beta(100,100)

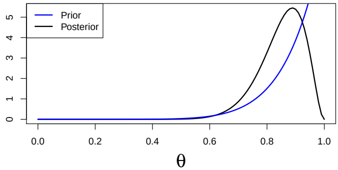
Example: Binomial Data, Beta Prior

Possible scenarios of prior information on θ; posteriors with $Z = 8$:

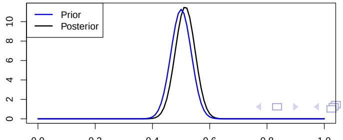
- No idea of what θ could be: Beta(1,1)



- The person flipping the coin looks like a trickster: Beta(9,1)



- Coin flipping usually has 50/50 chance of heads/tails: Beta(100,100)



Comments So Far

- Bayesian approach allows you to incorporate side information based on context
  - Do you know what "flipping a coin" means?
  - Is the person flipping the coin someone you trust?

- It seems expressing prior information works nicely with some parametric families

- Quantification of prior information can be tricky, especially for complicated models

Comments So Far

- Bayesian approach allows you to incorporate side information based on context
  - Do you know what "flipping a coin" means?
  - Is the person flipping the coin someone you trust?
- It seems expressing prior information works nicely with some parametric families
- Quantification of prior information can be tricky, especially for complicated models

Comments So Far

- Bayesian approach allows you to incorporate side information based on context
  - Do you know what "flipping a coin" means?
  - Is the person flipping the coin someone you trust?
- It seems expressing prior information works nicely with some parametric families
- Quantification of prior information can be tricky, especially for complicated models

Example: Multinomial Data, Dirichlet Prior

Continuing our example from the previous class:

- Let $Z_i = (Z_{i1}, Z_{i2})$, $\quad Z_{i1}, Z_{i2} \in \{1, 2\}$, $\quad Z_i$'s are i.i.d.,

$$p(Z_{i1} = k, Z_{i2} = l \mid \theta) = \pi_{kl}$$

- $\theta = (\ldots, \pi_{kl}, \ldots)$, $\qquad W_{ikl} = I(Z_{i1} = k, Z_{i2} = l)$

- The likelihood of the study variables is

$$\begin{aligned} L(\theta \mid \mathbf{z}) &= \prod_i \left[\prod_{k,l} \pi_{kl}^{W_{ikl}} \right] \\ &= \prod_{k,l} \pi_{kl}^{n_{kl}} \end{aligned}$$

where

$$n_{kl} = \sum_i W_{ikl}, \quad k, l \in \{1, 2\}$$

Example: Multinomial Data, Dirichlet Prior

Continuing our example from the previous class:

- Let $Z_i = (Z_{i1}, Z_{i2})$, $\quad Z_{i1}, Z_{i2} \in \{1, 2\}$, $\quad Z_i$'s are i.i.d.,

$$p(Z_{i1} = k, Z_{i2} = l \mid \theta) = \pi_{kl}$$

- $\theta = (\dots, \pi_{kl}, \dots)$, $\qquad W_{ikl} = I(Z_{i1} = k, Z_{i2} = l)$

- The likelihood of the study variables is

$$\begin{aligned} L(\theta \mid \mathbf{z}) &= \prod_i \left[\prod_{k,l} \pi_{kl}^{W_{ikl}} \right] \\ &= \prod_{k,l} \pi_{kl}^{n_{kl}} \end{aligned}$$

where

$$n_{kl} = \sum_i W_{ikl}, \quad k, l \in \{1, 2\}$$

Example: Multinomial Data, Dirichlet Prior

- Inference on multinomial parameters is convenient using the *Dirichlet* prior

- $\theta = (\dots, \pi_{kl}, \dots) \sim \text{Dirichlet}(\alpha), \quad \alpha = (\dots, \alpha_{kl}, \dots),$

$$p(\theta) = \frac{\Gamma(\sum \alpha_{kl})}{\prod_{k,l} \Gamma(\alpha_{kl})} \prod_{k,l} \pi_{kl}^{\alpha_{kl}-1}$$

- The posterior is given by

$$\begin{aligned} p(\theta \mid \mathbf{z}) &\propto L(\theta \mid \mathbf{z}) p(\theta) \\ &\propto \prod_{k,l} \pi_{kl}^{n_{kl}+\alpha_{kl}-1} \end{aligned}$$

- Therefore, $\theta \mid \mathbf{z} \sim \text{Dirichlet}(\alpha'), \quad \alpha' = (\dots, \alpha_{kl} + n_{kl}, \dots)$

Example: Multinomial Data, Dirichlet Prior

- Inference on multinomial parameters is convenient using the *Dirichlet* prior

- $\theta = (\dots, \pi_{kl}, \dots) \sim \text{Dirichlet}(\alpha), \quad \alpha = (\dots, \alpha_{kl}, \dots),$

$$p(\theta) = \frac{\Gamma(\sum \alpha_{kl})}{\prod_{k,l} \Gamma(\alpha_{kl})} \prod_{k,l} \pi_{kl}^{\alpha_{kl}-1}$$

- The posterior is given by

$$\begin{aligned} p(\theta \mid \mathbf{z}) &\propto L(\theta \mid \mathbf{z}) p(\theta) \\ &\propto \prod_{k,l} \pi_{kl}^{n_{kl}+\alpha_{kl}-1} \end{aligned}$$

- Therefore, $\theta \mid \mathbf{z} \sim \text{Dirichlet}(\alpha'), \quad \alpha' = (\dots, \alpha_{kl} + n_{kl}, \dots)$

Comments on Priors

- There is no guarantee that the combination of arbitrary likelihoods and priors will lead to posteriors that are easy to work with
- *Conjugate priors* are distributions that lead to posteriors in the same family, as in the previous examples – typically easier to work with, but not always available
- Non-conjugate priors can be used, but we require more advanced techniques for handling them
- Lists of conjugate priors are available in multiple books and online resources

Comments on Priors

- There is no guarantee that the combination of arbitrary likelihoods and priors will lead to posteriors that are easy to work with
- *Conjugate priors* are distributions that lead to posteriors in the same family, as in the previous examples – typically easier to work with, but not always available
- Non-conjugate priors can be used, but we require more advanced techniques for handling them
- Lists of conjugate priors are available in multiple books and online resources

Comments on Priors

- There is no guarantee that the combination of arbitrary likelihoods and priors will lead to posteriors that are easy to work with
- *Conjugate priors* are distributions that lead to posteriors in the same family, as in the previous examples – typically easier to work with, but not always available
- Non-conjugate priors can be used, but we require more advanced techniques for handling them
- Lists of conjugate priors are available in multiple books and online resources

Comments on Priors

- There is no guarantee that the combination of arbitrary likelihoods and priors will lead to posteriors that are easy to work with
- *Conjugate priors* are distributions that lead to posteriors in the same family, as in the previous examples – typically easier to work with, but not always available
- Non-conjugate priors can be used, but we require more advanced techniques for handling them
- Lists of conjugate priors are available in multiple books and online resources

Example: Multivariate Normal

- Distribution of the data

$$\mathbf{Z} = \{Z_i\}_{i=1}^n \mid \mu, \Lambda \overset{i.i.d.}{\sim} \text{Normal}(\mu, \Lambda^{-1})$$

where $Z_i \in \mathbb{R}^K$, μ is the vector of means, Λ^{-1} is the covariance matrix, and Λ is the inverse covariance matrix (the *precision matrix*)

- Conjugate prior is constructed in two steps

$$\mu \mid \Lambda \sim \text{Normal}(\mu_0, (\kappa_0 \Lambda)^{-1})$$
$$\Lambda \sim \text{Wishart}(\upsilon_0, W_0)$$

Joint distribution of (μ, Λ) is called *Normal-Wishart*. The parameterization is such that $E(\Lambda) = \upsilon_0 W_0$.

Example: Multivariate Normal

- Distribution of the data

$$\mathbf{Z} = \{Z_i\}_{i=1}^{n} \mid \mu, \Lambda \overset{i.i.d.}{\sim} \text{Normal}(\mu, \Lambda^{-1})$$

  where $Z_i \in \mathbb{R}^K$, μ is the vector of means, Λ^{-1} is the covariance matrix, and Λ is the inverse covariance matrix (the *precision matrix*)

- Conjugate prior is constructed in two steps

$$\mu \mid \Lambda \sim \text{Normal}(\mu_0, (\kappa_0 \Lambda)^{-1})$$
$$\Lambda \sim \text{Wishart}(\upsilon_0, W_0)$$

  Joint distribution of (μ, Λ) is called *Normal-Wishart*. The parameterization is such that $E(\Lambda) = \upsilon_0 W_0$.

Example: Multivariate Normal

Posterior is also Normal-Wishart

$$\mu \mid \Lambda, \mathbf{z} \sim \text{Normal}(\mu', (\kappa'\Lambda)^{-1})$$
$$\Lambda \mid \mathbf{z} \sim \text{Wishart}(\upsilon', W')$$

where

$$\begin{aligned}
\mu' &= (\kappa_0\mu_0 + n\bar{z})/\kappa' \\
\kappa' &= \kappa_0 + n \\
\upsilon' &= \upsilon_0 + n \\
W' &= \{W_0^{-1} + n[\hat{\Sigma} + \frac{\kappa_0}{\kappa'}(\bar{z} - \mu_0)(\bar{z} - \mu_0)^T]\}^{-1} \\
\bar{z} &= \sum_{i=1}^{n} z_i/n \\
\hat{\Sigma} &= \sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^T/n
\end{aligned}$$

Bayesian Point Estimation

- $\mathcal{L}(\theta, \theta')$: loss of estimating a parameter to be θ' when the true value is θ
- Bayes estimator minimizes the expected posterior loss

$$\hat{\theta}_{\text{Bayes}} = \arg\min_{\theta'} \int \mathcal{L}(\theta, \theta') p(\theta \mid \mathbf{z}) d\theta$$

- For univariate θ it is common to choose
  - $\mathcal{L}(\theta, \theta') = (\theta - \theta')^2 \implies \hat{\theta}_{\text{Bayes}}$ is posterior mean
  - $\mathcal{L}(\theta, \theta') = |\theta - \theta'| \implies \hat{\theta}_{\text{Bayes}}$ is posterior median
  - $\mathcal{L}(\theta, \theta') = I(\theta \neq \theta') \implies \hat{\theta}_{\text{Bayes}}$ is posterior mode

Bayesian Point Estimation

- $\mathcal{L}(\theta, \theta')$: loss of estimating a parameter to be θ' when the true value is θ
- Bayes estimator minimizes the expected posterior loss

$$\hat{\theta}_{\text{Bayes}} = \arg\min_{\theta'} \int \mathcal{L}(\theta, \theta') p(\theta \mid \mathbf{z}) d\theta$$

- For univariate θ it is common to choose
  - $\mathcal{L}(\theta, \theta') = (\theta - \theta')^2 \implies \hat{\theta}_{\text{Bayes}}$ is posterior mean
  - $\mathcal{L}(\theta, \theta') = |\theta - \theta'| \implies \hat{\theta}_{\text{Bayes}}$ is posterior median
  - $\mathcal{L}(\theta, \theta') = I(\theta \neq \theta') \implies \hat{\theta}_{\text{Bayes}}$ is posterior mode

Bayesian Point Estimation

- $\mathcal{L}(\theta, \theta')$: loss of estimating a parameter to be θ' when the true value is θ
- Bayes estimator minimizes the expected posterior loss

$$\hat{\theta}_{\text{Bayes}} = \arg\min_{\theta'} \int \mathcal{L}(\theta, \theta') p(\theta \mid \mathbf{z}) d\theta$$

- For univariate θ it is common to choose
  - $\mathcal{L}(\theta, \theta') = (\theta - \theta')^2 \implies \hat{\theta}_{\text{Bayes}}$ is posterior mean
  - $\mathcal{L}(\theta, \theta') = |\theta - \theta'| \implies \hat{\theta}_{\text{Bayes}}$ is posterior median
  - $\mathcal{L}(\theta, \theta') = I(\theta \neq \theta') \implies \hat{\theta}_{\text{Bayes}}$ is posterior mode

Bayesian Credible Sets/Intervals

- C is a $(1-\alpha)100\%$ *credible set* if

$$\int_C p(\theta \mid \mathbf{z}) d\theta \geq 1-\alpha$$

- For univariate θ, define the $(1-\alpha)100\%$ *credible interval* C as the interval within the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior $p(\theta \mid \mathbf{z})$

Bayesian Credible Sets/Intervals

- C is a $(1-\alpha)100\%$ *credible set* if

$$\int_C p(\theta \mid \mathbf{z})d\theta \geq 1-\alpha$$

- For univariate θ, define the $(1-\alpha)100\%$ *credible interval* C as the interval within the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior $p(\theta \mid \mathbf{z})$

Asymptotic Behavior of Posteriors: Bernstein - von Mises

Bernstein - von Mises theorem[1]

- Under some conditions, the posterior distribution asymptotically behaves like the sampling distribution of the MLE
- Heuristically, we say

$$p(\theta \mid \mathbf{z}) \approx \mathcal{N}(\hat{\theta}_{\mathsf{MLE}}, \mathcal{I}(\hat{\theta}_{\mathsf{MLE}})^{-1}/n)$$

- Therefore, for well-behaved models and with a good amount of data, Bayesian and frequentist inferences will be similar

[1] For details, see lecture notes of Richard Nickl: `http://www.statslab.cam.ac.uk/~nickl/Site/__files/stat2013.pdf`

Asymptotic Behavior of Posteriors: Bernstein - von Mises

Bernstein - von Mises theorem[1]

- Under some conditions, the posterior distribution asymptotically behaves like the sampling distribution of the MLE
- Heuristically, we say

$$p(\theta \mid \mathbf{z}) \approx \mathcal{N}(\hat{\theta}_{\mathrm{MLE}}, \mathcal{I}(\hat{\theta}_{\mathrm{MLE}})^{-1}/n)$$

- Therefore, for well-behaved models and with a good amount of data, Bayesian and frequentist inferences will be similar

[1]For details, see lecture notes of Richard Nickl:
http://www.statslab.cam.ac.uk/~nickl/Site/__files/stat2013.pdf

Asymptotic Behavior of Posteriors: Bernstein - von Mises

Bernstein - von Mises theorem[1]

- Under some conditions, the posterior distribution asymptotically behaves like the sampling distribution of the MLE

- Heuristically, we say

$$p(\theta \mid \mathbf{z}) \approx \mathcal{N}(\hat{\theta}_{\text{MLE}}, \mathcal{I}(\hat{\theta}_{\text{MLE}})^{-1}/n)$$

- Therefore, for well-behaved models and with a good amount of data, Bayesian and frequentist inferences will be similar

[1] For details, see lecture notes of Richard Nickl: `http://www.statslab.cam.ac.uk/~nickl/Site/__files/stat2013.pdf`

Missing Data and Bayes

- With missing data, things get complicated

$$L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) = \prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(r_i \mid z_i, \psi) p(z_i \mid \theta) \; dz_{i(\bar{r}_i)}$$

$$\overset{\text{MAR}}{=} \underbrace{\left[\prod_{i=1}^{n} p(r_i \mid z_{i(r_i)}, \psi)\right]}_{\text{Can be ignored}} \underbrace{\left[\prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(z_i \mid \theta) \; dz_{i(\bar{r}_i)}\right]}_{\text{Likelihood for } \theta \text{ under MAR}}$$

- Under a Bayesian approach, we need to obtain

$$p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) \propto L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi)$$

- Under *ignorability* (MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori), we need

$$p(\theta \mid \mathbf{z}_{(\mathbf{r})}) \propto L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta)$$

- Integrals in $L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})})$ complicate things – what to do?

Missing Data and Bayes

- With missing data, things get complicated

$$L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) = \prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(r_i \mid z_i, \psi) p(z_i \mid \theta) \; dz_{i(\bar{r}_i)}$$

$$\overset{\text{MAR}}{=} \underbrace{\left[\prod_{i=1}^{n} p(r_i \mid z_{i(r_i)}, \psi) \right]}_{\text{Can be ignored}} \underbrace{\left[\prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(z_i \mid \theta) \; dz_{i(\bar{r}_i)} \right]}_{\text{Likelihood for } \theta \text{ under MAR}}$$

- Under a Bayesian approach, we need to obtain

$$p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) \propto L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi)$$

- Under *ignorability* (MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori), we need

$$p(\theta \mid \mathbf{z}_{(\mathbf{r})}) \propto L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta)$$

- Integrals in $L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})})$ complicate things – what to do?

Missing Data and Bayes

- With missing data, things get complicated

$$L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) = \prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(r_i \mid z_i, \psi) p(z_i \mid \theta) \; dz_{i(\bar{r}_i)}$$

$$\stackrel{\text{MAR}}{=} \underbrace{\left[\prod_{i=1}^{n} p(r_i \mid z_{i(r_i)}, \psi)\right]}_{\text{Can be ignored}} \underbrace{\left[\prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(z_i \mid \theta) \; dz_{i(\bar{r}_i)}\right]}_{\text{Likelihood for } \theta \text{ under MAR}}$$

- Under a Bayesian approach, we need to obtain

$$p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) \propto L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi)$$

- Under *ignorability* (MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori), we need

$$p(\theta \mid \mathbf{z}_{(\mathbf{r})}) \propto L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta)$$

- Integrals in $L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})})$ complicate things – what to do?

Missing Data and Bayes

- With missing data, things get complicated

$$L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) = \prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(r_i \mid z_i, \psi) p(z_i \mid \theta) \; dz_{i(\bar{r}_i)}$$

$$\overset{\text{MAR}}{=} \underbrace{\left[\prod_{i=1}^{n} p(r_i \mid z_{i(r_i)}, \psi)\right]}_{\text{Can be ignored}} \underbrace{\left[\prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(z_i \mid \theta) \; dz_{i(\bar{r}_i)}\right]}_{\text{Likelihood for } \theta \text{ under MAR}}$$

- Under a Bayesian approach, we need to obtain

$$p(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) \propto L_{obs}(\theta, \psi \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{r}) p(\theta, \psi)$$

- Under *ignorability* (MAR + separability + $\theta \perp\!\!\!\perp \psi$ a priori), we need

$$p(\theta \mid \mathbf{z}_{(\mathbf{r})}) \propto L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})}) p(\theta)$$

- Integrals in $L_{obs}(\theta \mid \mathbf{z}_{(\mathbf{r})})$ complicate things – what to do?

Summary

Main take-aways from today's lecture:

- Using Bayes' theorem doesn't make you a Bayesian
- Bayesian inference offers alternative framework for deriving inferences from data
  - Philosophical motivation: inclusion of prior belief or knowledge, uncertainty quantification in terms of distributions for parameters
  - Practical motivation: convenient in some problems, might lead to good frequentist performance
  - Complex problems are computationally challenging – posterior needs to be approximated (e.g., Markov chain Monte Carlo)

Next lecture:

- Gibbs sampling
- Data augmentation
- Introduction to multiple imputation

Summary

Main take-aways from today's lecture:

- Using Bayes' theorem doesn't make you a Bayesian
- Bayesian inference offers alternative framework for deriving inferences from data
    - Philosophical motivation: inclusion of prior belief or knowledge, uncertainty quantification in terms of distributions for parameters
    - Practical motivation: convenient in some problems, might lead to good frequentist performance
    - Complex problems are computationally challenging – posterior needs to be approximated (e.g., Markov chain Monte Carlo)

Next lecture:

- Gibbs sampling
- Data augmentation
- Introduction to multiple imputation