COMP 680
Statistics for Computing and Data Science

Week 9: Bayesian Inference

Su Chen, Assistant Teaching Professor,
Rice D2K Lab

Outline

Overview

- Bayesian v.s. Classical Statistics
  - Two major schools of statistics
  - Solve the same problem in two completely different ways
  - Somewhat philosophical

Overview

- Bayesian v.s. Classical Statistics
  - Two major schools of statistics
  - Solve the same problem in two completely different ways
  - Somewhat philosophical
- Classical Statistics
  - The unknown parameter is a fixed quantity
  - Rely on repeated experiments to make inference
  - Thus the name "Frequentist"

Overview

- Bayesian v.s. Classical Statistics
  - Two major schools of statistics
  - Solve the same problem in two completely different ways
  - Somewhat philosophical
- Classical Statistics
  - The unknown parameter is a fixed quantity
  - Rely on repeated experiments to make inference
  - Thus the name "Frequentist"
- Bayesian Statistics
  - The unknown parameter is a random variable follows some distribution
  - Has a prior belief about what that random distribution should be
  - Data comes in to update that prior belief to a posterior belief

Your Favorite Example

Consider the coin-flipping example:

- Goal: to estimate the probability that the coin flip is a head

Your Favorite Example

Consider the coin-flipping example:

- Goal: to estimate the probability that the coin flip is a head
- The Frequentist way:
  - repeated experiment $\rightarrow$ data $\rightarrow$ maximum likelihood estimator
  - sampling distribution quantify uncertainty

Your Favorite Example

Consider the coin-flipping example:

- Goal: to estimate the probability that the coin flip is a head
- The Frequentist way:
  - repeated experiment $\rightarrow$ data $\rightarrow$ maximum likelihood estimator
  - sampling distribution quantify uncertainty
- The Bayesian way:
  - prior distribution $\rightarrow$ data $\rightarrow$ posterior distribution
  - posterior distribution quantify uncertainty!

History

Thomas Bayes (1701 - 1761)

History

Thomas Bayes (1701 - 1761)

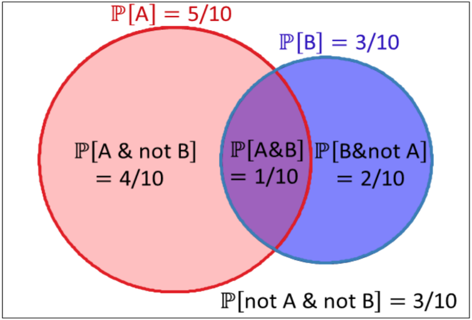
- The "Original" Bayes paper: An essay towards solving a problem in the doctrine of chances
  - use Binomial data comprising r successes out of n attempts
  - learn about the underlying chance θ of each attempt succeeding
  - use a probability distribution to represent uncertainty about θ

Outline

1. The Big Picture
2. The Bayesian "Recipe"
3. Conjugate Family
4. Posterior Inference
5. Code Demo

The Bayes Theorem

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

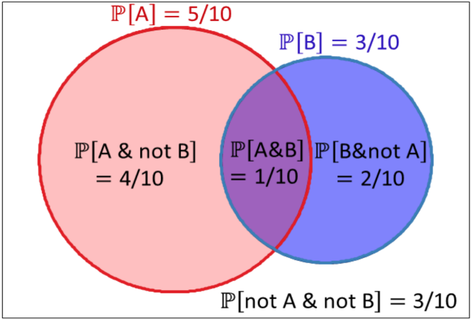


In this example;

- $\mathbb{P}[A|B] = \frac{1/10}{3/10} = 1/3$
- $\mathbb{P}[B|A] = \frac{1/10}{5/10} = 1/5$
- And $1/3 = 1/5 \times \frac{5/10}{3/10}$ ($\checkmark$)

The Bayes Theorem

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$



In this example;

- $\mathbb{P}[A|B] = \frac{1/10}{3/10} = 1/3$
- $\mathbb{P}[B|A] = \frac{1/10}{5/10} = 1/5$
- And $1/3 = 1/5 \times \frac{5/10}{3/10}$ (✓)

- the conditional probability of A given B is the conditional probability of B given A scaled by the relative probability of A compared to B.

The Bayes Theorem

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \quad \Longrightarrow \quad \pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{p(X)}$$

The Bayes Theorem

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \quad \Longrightarrow \quad \pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{p(X)}$$

- Notation:
  - unknown parameter θ (can be multi-dimension)
  - data $X = (X_1, X_2, \cdots X_n)$ (usually assume i.i.d)

The Bayes Theorem

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \quad \Longrightarrow \quad \pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{p(X)}$$

- Notation:
  - unknown parameter θ (can be multi-dimension)
  - data $X = (X_1, X_2, \cdots X_n)$ (usually assume i.i.d)
- Likelihood: $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$

The Bayes Theorem

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \quad \Longrightarrow \quad \pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{p(X)}$$

- Notation:
  - unknown parameter θ (can be multi-dimension)
  - data $X = (X_1, X_2, \cdots X_n)$ (usually assume i.i.d)
- Likelihood: $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$
- Prior distribution: $\pi(\theta)$

The Bayes Theorem

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \implies \pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{p(X)}$$

- Notation:
  - unknown parameter θ (can be multi-dimension)
  - data $X = (X_1, X_2, \cdots X_n)$ (usually assume i.i.d)
- Likelihood: $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$
- Prior distribution: $\pi(\theta)$
- Posterior: $\pi(\theta|X)$

The Bayes Theorem

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \quad \Longrightarrow \quad \pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{p(X)}$$

- Notation:
  - unknown parameter θ (can be multi-dimension)
  - data $X = (X_1, X_2, \cdots X_n)$ (usually assume i.i.d)
- Likelihood: $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$
- Prior distribution: $\pi(\theta)$
- Posterior: $\pi(\theta|X)$
- Marginal Likelihood: (aka the "normalizing constant")
  - $p(X) = \int p(X|\theta)\pi(\theta)d\theta$
  - $p(X)$ is a constant regarding the posterior $\pi(\theta|X)$

Bayesian Inference

$$\pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{p(X)} \quad \Longrightarrow \quad \pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{\int p(X|\theta)\pi(\theta)d\theta}$$

How to update our belief about θ, as data is obtained?

Bayesian Inference

$$\pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{p(X)} \implies \pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{\int p(X|\theta)\pi(\theta)d\theta}$$

How to update our belief about θ, as data is obtained?

- Prior distribution: what you know about parameter θ, excluding the information in the data - denoted $\pi(\theta)$

Bayesian Inference

$$\pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{p(X)} \quad \Longrightarrow \quad \pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{\int p(X|\theta)\pi(\theta)d\theta}$$

How to update our belief about θ, as data is obtained?

- Prior distribution: what you know about parameter θ, excluding the information in the data - denoted $\pi(\theta)$
- Likelihood: based on modeling assumptions, how (relatively) likely the data X are if the true parameter is θ - denoted $p(X|\theta)$

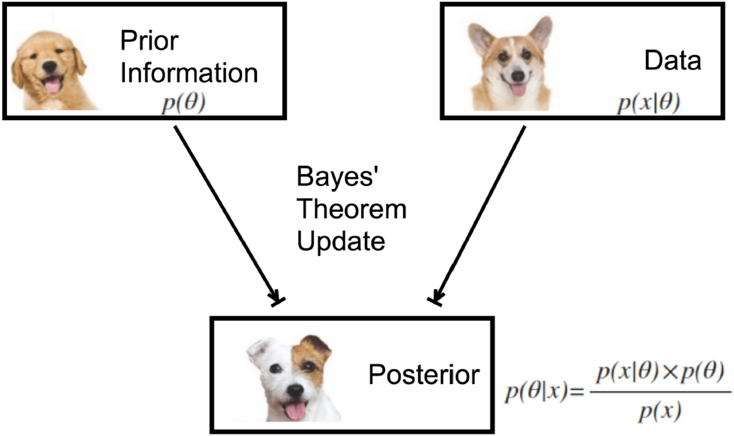
Bayesian Inference

$$\pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{p(X)} \quad \Longrightarrow \quad \pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{\int p(X|\theta)\pi(\theta)d\theta}$$

How to update our belief about θ, as data is obtained?

- Prior distribution: what you know about parameter θ, excluding the information in the data - denoted $\pi(\theta)$
- Likelihood: based on modeling assumptions, how (relatively) likely the data X are if the true parameter is θ - denoted $p(X|\theta)$
- So how to get a posterior distribution: starting what we know about θ, combining the prior with the data, Bayes Theorem used for inference tells us to multiply and scale ... and that is it! (essentially!)

The Bayesian “Recipe”



Puppies borrowed by Kruschke J., Doing Bayesian Data Analysis, A tutorial with R, JAGS and STAN, Academic Press

The Bayesian "Recipe"

$$\pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{\int p(X|\theta)\pi(\theta)d\theta} \implies \pi(\theta|X) \propto p(X|\theta)\pi(\theta)$$

The Bayesian "Recipe"

$$\pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{\int p(X|\theta)\pi(\theta)d\theta} \quad \Longrightarrow \quad \pi(\theta|X) \propto p(X|\theta)\pi(\theta)$$

- Bayes Theorem provides the basis for Bayesian inference.

The Bayesian "Recipe"

$$\pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{\int p(X|\theta)\pi(\theta)d\theta} \implies \pi(\theta|X) \propto p(X|\theta)\pi(\theta)$$

- Bayes Theorem provides the basis for Bayesian inference.
- The "prior" distribution $\pi(\theta)$ is combined with "likelihood" $p(X|\theta)$ to provide a "posterior" distribution $\pi(\theta|X)$.

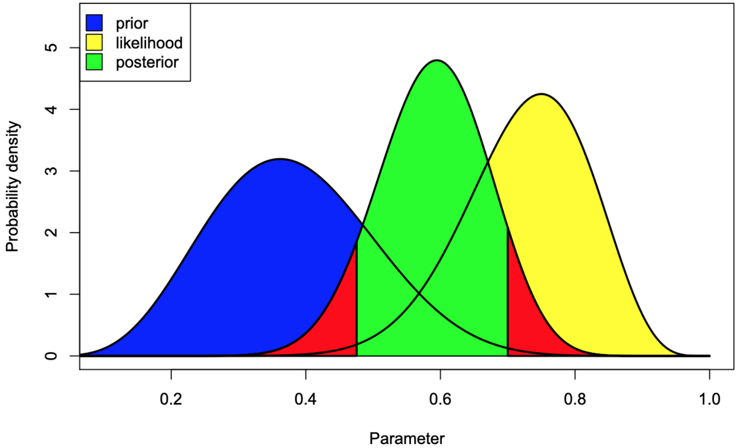
The Bayesian "Recipe"

$$\pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{\int p(X|\theta)\pi(\theta)d\theta} \quad \Longrightarrow \quad \pi(\theta|X) \propto p(X|\theta)\pi(\theta)$$

- Bayes Theorem provides the basis for Bayesian inference.
- The "prior" distribution $\pi(\theta)$ is combined with "likelihood" $p(X|\theta)$ to provide a "posterior" distribution $\pi(\theta|X)$.
- The likelihood is derived from an sampling model $p(X|\theta)$ but considered as function of θ for fixed X.

Illustration

$$\pi(\theta|X) \propto p(X|\theta)\pi(\theta) \quad \Longrightarrow \quad \text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$



The Challenge

This sounds too easy! What's the challenge?

The Challenge

This sounds too easy! What's the challenge?

- How to choose the prior for the unknown parameter θ?

The Challenge

This sounds too easy! What's the challenge?

- How to choose the prior for the unknown parameter θ?
- How to calculate the posterior, in particular, the normalizing constant $p(X) = \int p(X|\theta)d\theta$ where you have to do an integral (possibly high dimension integral)!

Where do priors come from?

"There's nothing wrong, dirty, unnatural or even unusual about making assumptions - carefully. Scientists and statisticians all make assumptions... even if they don't like to talk about them."

Where do priors come from?

"There's nothing wrong, dirty, unnatural or even unusual about making assumptions - carefully. Scientists and statisticians all make assumptions... even if they don't like to talk about them."

- Priors come from all data external to the current study i.e. everything else.

Where do priors come from?

"There's nothing wrong, dirty, unnatural or even unusual about making assumptions - carefully. Scientists and statisticians all make assumptions... even if they don't like to talk about them."

- Priors come from all data external to the current study i.e. everything else.
- "Boil down" to what subject-matter experts know/think is known as eliciting a prior.

Where do priors come from?

"There's nothing wrong, dirty, unnatural or even unusual about making assumptions - carefully. Scientists and statisticians all make assumptions... even if they don't like to talk about them."

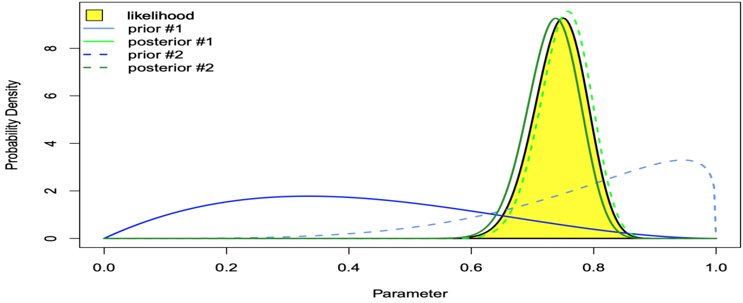
- Priors come from all data external to the current study i.e. everything else.
- "Boil down" to what subject-matter experts know/think is known as eliciting a prior.
- It is not easy!

When don't priors matter?

When the data provide a lot more information than the prior, this happens;

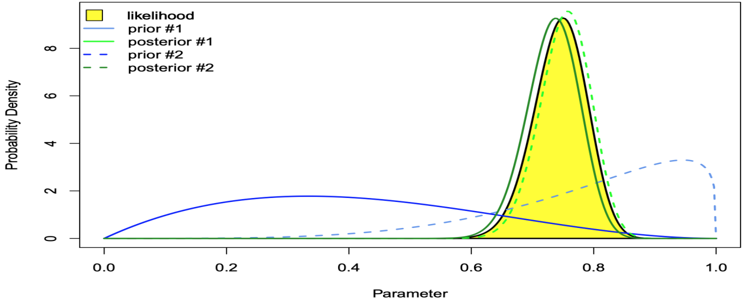
When don't priors matter?

When the data provide a lot more information than the prior, this happens;



When don't priors matter?

When the data provide a lot more information than the prior, this happens;



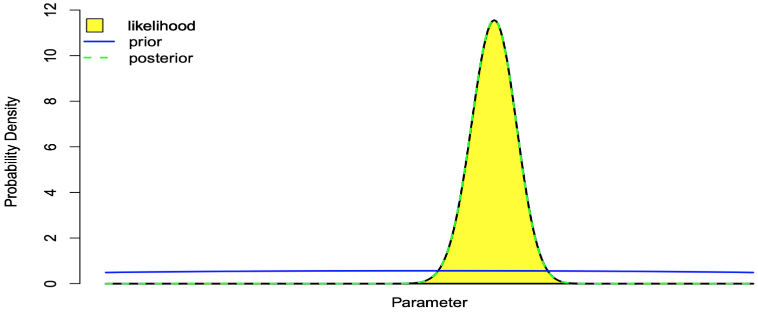
These two priors (and many more) are dominated by the likelihood, and they give very similar posteriors - i.e. everyone agrees. (Phew!)

What is a flat prior?

A related idea; use very flat priors to represent "ignorance";

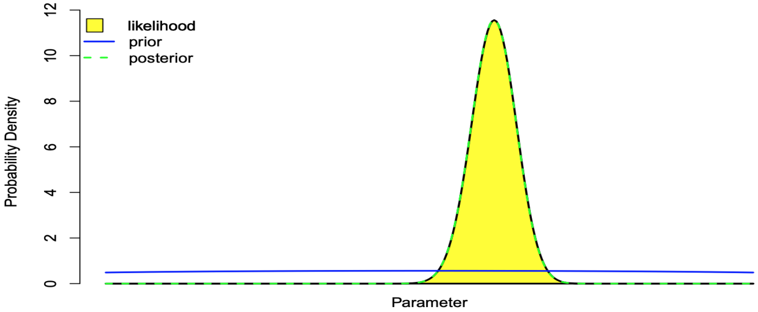
What is a flat prior?

A related idea; use very flat priors to represent "ignorance";



What is a flat prior?

A related idea; use very flat priors to represent "ignorance";



"Objective Bayes": use flat (non-informative) priors.

Outline

Conjugate Priors

A class of prior distributions for θ is called conjugate for a particular sampling model (likelihood) $p(X|\theta)$, if the posterior distribution $\pi(\theta|X)$ is in the same distribution family as the prior $\pi(\theta)$.

Conjugate Priors

A class of prior distributions for θ is called conjugate for a particular sampling model (likelihood) $p(X|\theta)$, if the posterior distribution $\pi(\theta|X)$ is in the same distribution family as the prior $\pi(\theta)$.

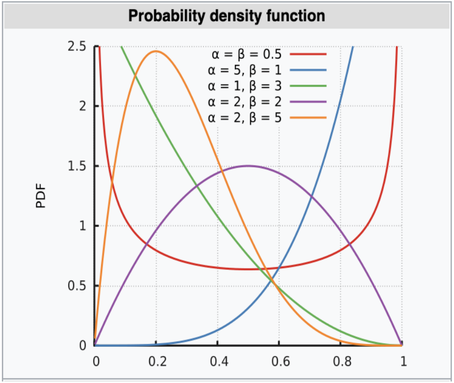
- This simplifies the posterior calculation because we will be able to "recognize" the posterior distribution without actually calculating the normalizing constant.

Conjugate Priors

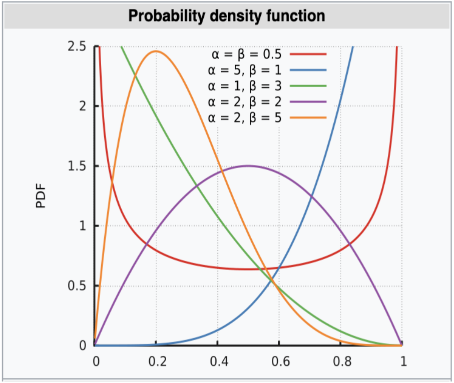
A class of prior distributions for θ is called conjugate for a particular sampling model (likelihood) $p(X|\theta)$, if the posterior distribution $\pi(\theta|X)$ is in the same distribution family as the prior $\pi(\theta)$.

- This simplifies the posterior calculation because we will be able to "recognize" the posterior distribution without actually calculating the normalizing constant.
- We almost always choose a conjugate prior if there is one!

Beta Distribution



Beta Distribution



- A continuous random variable defined on $[0, 1]$ with PDF:

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

  - $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
  - $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$
  - $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Binomial-Beta Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \textbf{Bernoulli}(X|\theta)$

Binomial-Beta Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \textbf{Bernoulli}(X|\theta)$
  - $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{(1-X_i)}$
  - Remember $\hat{\theta}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$

Binomial-Beta Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \textbf{Bernoulli}(X|\theta)$
  - $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{(1-X_i)}$
  - Remember $\hat{\theta}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$
- Prior distribution: $\pi(\theta) = \textbf{Beta}(\theta|\alpha_0, \beta_0)$

Binomial-Beta Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \textbf{Bernoulli}(X|\theta)$
  - $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{(1-X_i)}$
  - Remember $\hat{\theta}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$
- Prior distribution: $\pi(\theta) = \textbf{Beta}(\theta|\alpha_0, \beta_0)$
- Posterior: $\pi(\theta|X) = \textbf{Beta}(\theta|\alpha_n, \beta_n)$
- Posterior is in the same family of distribution as prior, with information updated by data:

Binomial-Beta Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \textbf{Bernoulli}(X|\theta)$
  - $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{(1-X_i)}$
  - Remember $\hat{\theta}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$
- Prior distribution: $\pi(\theta) = \textbf{Beta}(\theta|\alpha_0, \beta_0)$
- Posterior: $\pi(\theta|X) = \textbf{Beta}(\theta|\alpha_n, \beta_n)$
- Posterior is in the same family of distribution as prior, with information updated by data:
  - $\alpha_n = \alpha_0 + n\bar{X}$
  - $\beta_n = \beta_0 + n - n\bar{X}$

Binomial-Beta Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \textbf{Bernoulli}(X|\theta)$
  - $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{(1-X_i)}$
  - Remember $\hat{\theta}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$
- Prior distribution: $\pi(\theta) = \textbf{Beta}(\theta|\alpha_0, \beta_0)$
- Posterior: $\pi(\theta|X) = \textbf{Beta}(\theta|\alpha_n, \beta_n)$
- Posterior is in the same family of distribution as prior, with information updated by data:
  - $\alpha_n = \alpha_0 + n\bar{X}$
  - $\beta_n = \beta_0 + n - n\bar{X}$
  - Posterior mean is a weighted average of prior mean and sample mean
  - $\frac{\alpha_n}{\alpha_n+\beta_n} = \frac{\alpha_0+n\bar{X}}{\alpha_0+\beta_0+n} = \frac{\alpha_0+\beta_0}{\alpha_0+\beta_0+n}\frac{\alpha_0}{\alpha_0+\beta_0} + \frac{n}{\alpha_0+\beta_0+n}\bar{X} = \omega_n\frac{\alpha_0}{\alpha_0+\beta_0} + (1-\omega_n)\bar{X}$

Gamma Distribution

- Exponential(β): a continuous, non-negative random variable
- PDF:

$$f_X(x) = \beta \exp(-\beta x)$$

  - $\mathbb{E}[X] = \frac{1}{\beta}$
  - $\mathrm{Var}(X) = \frac{1}{\beta^2}$

Gamma Distribution

- Exponential(β): a continuous, non-negative random variable
- PDF:

$$f_X(x) = \beta \exp(-\beta x)$$

  - $\mathbb{E}[X] = \frac{1}{\beta}$
  - $\text{Var}(X) = \frac{1}{\beta^2}$

- Gamma(α, β): sum of α i.i.d. Exponential(β)
- PDF:

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$$

  - $\mathbb{E}[X] = \frac{\alpha}{\beta}$
  - $\text{Var}(X) = \frac{\alpha}{\beta^2}$

Poisson-Gamma Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \textbf{Poisson}(X|\lambda)$

Poisson-Gamma Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \textbf{Poisson}(X|\lambda)$
  - $p(X|\lambda) = \prod_{i=1}^{n} p(X_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$
  - Remember $\hat{\lambda}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i / n$

Poisson-Gamma Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim}$ **Poisson**$(X|\lambda)$
  - $p(X|\lambda) = \prod_{i=1}^{n} p(X_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$
  - Remember $\hat{\lambda}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i / n$
- Prior distribution: $\pi(\lambda) =$ **Gamma**$(\lambda|\alpha_0, \beta_0)$

Poisson-Gamma Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \textbf{Poisson}(X|\lambda)$
  - $p(X|\lambda) = \prod_{i=1}^{n} p(X_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$
  - Remember $\hat{\lambda}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$
- Prior distribution: $\pi(\lambda) = \textbf{Gamma}(\lambda|\alpha_0, \beta_0)$
- Posterior: $\pi(\lambda|X) = \textbf{Gamma}(\lambda|\alpha_n, \beta_n)$
- Posterior is in the same family of distribution as prior, with information updated by data:

Poisson-Gamma Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \textbf{Poisson}(X|\lambda)$
  - $p(X|\lambda) = \prod_{i=1}^{n} p(X_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$
  - Remember $\hat{\lambda}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$
- Prior distribution: $\pi(\lambda) = \textbf{Gamma}(\lambda|\alpha_0, \beta_0)$
- Posterior: $\pi(\lambda|X) = \textbf{Gamma}(\lambda|\alpha_n, \beta_n)$
- Posterior is in the same family of distribution as prior, with information updated by data:
  - $\alpha_n = \alpha_0 + n\bar{X}$
  - $\beta_n = \beta_0 + n$

Poisson-Gamma Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim}$ **Poisson**$(X|\lambda)$
  - $p(X|\lambda) = \prod_{i=1}^{n} p(X_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$
  - Remember $\hat{\lambda}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$
- Prior distribution: $\pi(\lambda) =$ **Gamma**$(\lambda|\alpha_0, \beta_0)$
- Posterior: $\pi(\lambda|X) =$ **Gamma**$(\lambda|\alpha_n, \beta_n)$
- Posterior is in the same family of distribution as prior, with information updated by data:
  - $\alpha_n = \alpha_0 + n\bar{X}$
  - $\beta_n = \beta_0 + n$
  - Posterior mean is a weighted average of prior mean and sample mean
  - $\frac{\alpha_n}{\beta_n} = \frac{\alpha_0 + n\bar{X}}{\beta_0 + n} = \frac{\beta_0}{\beta_0 + n}\frac{\alpha_0}{\beta_0} + \frac{n}{\beta_0 + n}\bar{X} = \omega_n \frac{\alpha_0}{\beta_0} + (1 - \omega_n)\bar{X}$

Normal-Normal Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \mathbf{N}(X|\theta, \sigma^2)$

Normal-Normal Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \mathbf{N}(X|\theta, \sigma^2)$
  - $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta) = \prod_{i=1}^{n} \mathbf{N}(X_i|\theta, \sigma^2)$
  - Assume θ is the parameter of interest, and σ^2 is known
  - Remember $\hat{\theta}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$

Normal-Normal Model

- Likelihood: $X_1, X_2, \cdots X_n \stackrel{iid}{\sim} \mathbf{N}(X|\theta, \sigma^2)$
  - $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta) = \prod_{i=1}^{n} \mathbf{N}(X_i|\theta, \sigma^2)$
  - Assume θ is the parameter of interest, and σ^2 is known
  - Remember $\hat{\theta}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$
- Prior distribution: $\pi(\theta) = \mathbf{N}(\theta|\mu_0, \tau_0^2)$

Normal-Normal Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \mathbf{N}(X|\theta, \sigma^2)$
  - $p(X|\theta) = \prod_{i=1}^n p(X_i|\theta) = \prod_{i=1}^n \mathbf{N}(X_i|\theta, \sigma^2)$
  - Assume θ is the parameter of interest, and σ^2 is known
  - Remember $\hat{\theta}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^n X_i/n$
- Prior distribution: $\pi(\theta) = \mathbf{N}(\theta|\mu_0, \tau_0^2)$
- Posterior: $\pi(\theta|X) = \mathbf{N}(\theta|\mu_n, \tau_n^2)$
- Posterior is in the same family of distribution as prior, with information updated by data:

Normal-Normal Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \mathbf{N}(X|\theta, \sigma^2)$
  - $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta) = \prod_{i=1}^{n} \mathbf{N}(X_i|\theta, \sigma^2)$
  - Assume θ is the parameter of interest, and σ^2 is known
  - Remember $\hat{\theta}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$
- Prior distribution: $\pi(\theta) = \mathbf{N}(\theta|\mu_0, \tau_0^2)$
- Posterior: $\pi(\theta|X) = \mathbf{N}(\theta|\mu_n, \tau_n^2)$
- Posterior is in the same family of distribution as prior, with information updated by data:
  - $\mu_n = \frac{\mu_0/\tau_0^2 + n\bar{X}/\sigma^2}{1/\tau_0^2 + n/\sigma^2}$
  - $\tau_n^2 = \frac{1}{1/\tau_0^2 + n/\sigma^2}$

Normal-Normal Model

- Likelihood: $X_1, X_2, \cdots X_n \overset{iid}{\sim} \mathbf{N}(X|\theta, \sigma^2)$
  - $p(X|\theta) = \prod_{i=1}^n p(X_i|\theta) = \prod_{i=1}^n \mathbf{N}(X_i|\theta, \sigma^2)$
  - Assume θ is the parameter of interest, and σ^2 is known
  - Remember $\hat{\theta}^{MLE}$ is the sample mean $\bar{X} = \sum_{i=1}^n X_i/n$
- Prior distribution: $\pi(\theta) = \mathbf{N}(\theta|\mu_0, \tau_0^2)$
- Posterior: $\pi(\theta|X) = \mathbf{N}(\theta|\mu_n, \tau_n^2)$
- Posterior is in the same family of distribution as prior, with information updated by data:
  - $\mu_n = \frac{\mu_0/\tau_0^2 + n\bar{X}/\sigma^2}{1/\tau_0^2 + n/\sigma^2}$
  - $\tau_n^2 = \frac{1}{1/\tau_0^2 + n/\sigma^2}$
  - Posterior mean is a weighted average of prior mean and sample mean
  - $\mu_n = \frac{1/\tau_0^2}{1/\tau_0^2 + n/\sigma^2}\mu_0 + \frac{n/\sigma^2}{1/\tau_0^2 + n/\sigma^2}\bar{X} = \omega_n\mu_0 + (1-\omega_n)\bar{X}$

Outline

Posterior Distribution

- The posterior distribution tells the whole story!

Posterior Distribution

- The posterior distribution tells the whole story!
- Point estimate of the unknown parameter:
  - posterior mean
  - posterior mode

Posterior Distribution

- The posterior distribution tells the whole story!
- Point estimate of the unknown parameter:
  - posterior mean
  - posterior mode
- Uncertainty quantification of the unknown parameter:
  - standard error: posterior standard deviation
  - confidence interval: the "middle chunk" of posterior distribution
  - in Bayesian this is called credible interval $[\theta_l, \theta_u]$:

$$\mathbb{P}(\theta \in [\theta_l, \theta_u]) = 95\%$$

the Monte Carlo Method

When things are not so conjugate:

the Monte Carlo Method

When things are not so conjugate:

- Any posterior distribution $p(\theta|X)$ may be approximated by taking a very large random sample of realizations of $\{\theta^1, \theta^2, \cdots, \theta^M\}$ from $p(\theta|X)$.

the Monte Carlo Method

When things are not so conjugate:

- Any posterior distribution $p(\theta|X)$ may be approximated by taking a very large random sample of realizations of $\{\theta^1, \theta^2, \cdots, \theta^M\}$ from $p(\theta|X)$.
  - Theory: approximate the true posterior distribution by empirical distribution of the random sample draw from the posterior.
  - Practice: approximate mean/variance/quantile by sample mean/variance/quantile.
  - For arbitrary function of θ: $\hat{g}(\theta) = \frac{1}{M}\sum_{m=1}^{M} g(\theta^m)$

the Monte Carlo Method

When things are not so conjugate:

- Any posterior distribution $p(\theta|X)$ may be approximated by taking a very large random sample of realizations of $\{\theta^1, \theta^2, \cdots, \theta^M\}$ from $p(\theta|X)$.
  - Theory: approximate the true posterior distribution by empirical distribution of the random sample draw from the posterior.
  - Practice: approximate mean/variance/quantile by sample mean/variance/quantile.
  - For arbitrary function of θ: $\hat{g}(\theta) = \frac{1}{M}\sum_{m=1}^{M} g(\theta^m)$
- Samples from the posterior can be generated in several ways, without exact knowledge of $p(\theta|X)$.

Markov Chain Monte Carlo (MCMC)

- Realizations from the posterior used in Monte Carlo methods need not be independent, or generated directly.

Markov Chain Monte Carlo (MCMC)

- Realizations from the posterior used in Monte Carlo methods need not be independent, or generated directly.
- Under conditional conjugacy: Gibbs Sampling.

Markov Chain Monte Carlo (MCMC)

- Realizations from the posterior used in Monte Carlo methods need not be independent, or generated directly.
- Under conditional conjugacy: Gibbs Sampling.
  - generates one parameter at a time;
  - sequentially updates each parameter, the entire parameter space is explored;
  - in the long-run, the "chains" of realizations produced will reflect the posterior of interest.

Markov Chain Monte Carlo (MCMC)

- Realizations from the posterior used in Monte Carlo methods need not be independent, or generated directly.
- Under conditional conjugacy: Gibbs Sampling.
  - generates one parameter at a time;
  - sequentially updates each parameter, the entire parameter space is explored;
  - in the long-run, the "chains" of realizations produced will reflect the posterior of interest.
- More general:
  - Metropolis-Hastings algorithm;
  - Rejection sampling;
  - Importance sampling;

Markov Chain Monte Carlo (MCMC)

- Realizations from the posterior used in Monte Carlo methods need not be independent, or generated directly.
- Under conditional conjugacy: Gibbs Sampling.
  - generates one parameter at a time;
  - sequentially updates each parameter, the entire parameter space is explored;
  - in the long-run, the "chains" of realizations produced will reflect the posterior of interest.
- More general:
  - Metropolis-Hastings algorithm;
  - Rejection sampling;
  - Importance sampling;
- Developing practical algorithms to approximate posterior distributions for complex problems remains an active area of research.

Gibbs Sampling

- Normal-normal model with joint inference for the mean and variance.

Gibbs Sampling

- Normal-normal model with joint inference for the mean and variance.
- Data: $y_1, y_2, \cdots y_n \overset{iid}{\sim} \mathbf{N}(\mu, \sigma^2)$ with μ and σ^2 unknown, use parameterization $\mathbf{N}(\mu, \tau = 1/\sigma^2)$

Gibbs Sampling

- Normal-normal model with joint inference for the mean and variance.
- Data: $y_1, y_2, \cdots y_n \overset{iid}{\sim} \mathbf{N}(\mu, \sigma^2)$ with μ and σ^2 unknown, use parameterization $\mathbf{N}(\mu, \tau = 1/\sigma^2)$
- Likelihood: $p(y_1, y_2, \cdots y_n | \mu, \tau) = \prod_{i=1}^{n} \mathbf{N}(y_i | \mu, \tau)$

Gibbs Sampling

- Normal-normal model with joint inference for the mean and variance.
- Data: $y_1, y_2, \cdots y_n \overset{iid}{\sim} \mathbf{N}(\mu, \sigma^2)$ with μ and σ^2 unknown, use parameterization $\mathbf{N}(\mu, \tau = 1/\sigma^2)$
- Likelihood: $p(y_1, y_2, \cdots y_n|\mu, \tau) = \prod_{i=1}^{n} \mathbf{N}(y_i|\mu, \tau)$
- Prior: $p(\mu, \tau) =???$
- Posterior: $\pi(\mu, \tau|y_1, y_2, \cdots y_n) \propto p(y_1, y_2, \cdots y_n|\mu, \tau)p(\mu, \tau)$

Gibbs Sampling

- To have conditional conjugacy, we assume $p(\mu, \tau) = p(\mu)p(\tau)$

Gibbs Sampling

- To have conditional conjugacy, we assume $p(\mu, \tau) = p(\mu)p(\tau)$
- Choose $p(\mu) = \mathbf{N}(\mu|0, v^2)$ and $p(\tau) = \mathbf{Gamma}(\tau|\alpha, \beta)$, where v^2, α and β are hyper-parameters to be determined.

Gibbs Sampling

- To have conditional conjugacy, we assume $p(\mu, \tau) = p(\mu)p(\tau)$
- Choose $p(\mu) = \mathbf{N}(\mu|0, v^2)$ and $p(\tau) = \mathbf{Gamma}(\tau|\alpha, \beta)$, where v^2, α and β are hyper-parameters to be determined.
- Notice here we still don't know how to calculate the joint posterior or sample from it directly. However, conditional conjugacy allows us to sample μ and τ iteratively from the conditional posterior:

Gibbs Sampling

- To have conditional conjugacy, we assume $p(\mu, \tau) = p(\mu)p(\tau)$
- Choose $p(\mu) = \mathbf{N}(\mu|0, v^2)$ and $p(\tau) = \mathbf{Gamma}(\tau|\alpha, \beta)$, where v^2, α and β are hyper-parameters to be determined.
- Notice here we still don't know how to calculate the joint posterior or sample from it directly. However, conditional conjugacy allows us to sample μ and τ iteratively from the conditional posterior:
  - Randomly set initial value $\mu^{(0)}$ and $\tau^{(0)}$, for $m = 1, 2, 3, \cdots M$.
  - Sample $\mu^{(m)}$ from $\pi(\mu|\tau, y_1, y_2, \cdots y_n) = \mathbf{N}\left(\mu|\frac{\bar{y}n\tau}{n\tau+1/v^2}, \frac{1}{n\tau+1/v^2}\right)$ where we use $\tau = \tau^{(m-1)}$.
  - Sample $\tau^{(m)}$ from $\pi(\tau|\mu, y_1, y_2, \cdots y_n) = \mathbf{Gamma}\left(\tau|\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}\right)$ where we use $\mu = \mu^{(m)}$.

Posterior Inference for Parameters

- For M large enough, we assume the chain has converged, and $(\mu^{(m)}, \tau^{(m)})$ can be treated as samples from the joint posterior $\pi(\mu, \tau | y_1, y_2, \cdots y_n)$.

Posterior Inference for Parameters

- For M large enough, we assume the chain has converged, and $(\mu^{(m)}, \tau^{(m)})$ can be treated as samples from the joint posterior $\pi(\mu, \tau | y_1, y_2, \cdots y_n)$.
- Use Monte Carlo approximation to get: point estimator, uncertainty measure, credit interval, any function of μ and τ etc.

Posterior Inference for Parameters

- For M large enough, we assume the chain has converged, and $(\mu^{(m)}, \tau^{(m)})$ can be treated as samples from the joint posterior $\pi(\mu, \tau | y_1, y_2, \cdots y_n)$.
- Use Monte Carlo approximation to get: point estimator, uncertainty measure, credit interval, any function of μ and τ etc.
- In practice, usually discard the first $x\%$ samples (burn-in) from the chain $(\mu^{(m)}, \tau^{(m)})$ and use the rest for posterior inference.

Predictive Distribution

- Prior distribution: $\pi(\theta)$
- Likelihood: $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$
- Posterior: $\pi(\theta|X)$

Predictive Distribution

- Prior distribution: $\pi(\theta)$
- Likelihood: $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$
- Posterior: $\pi(\theta|X)$
- Predictive density: $p(X^{pred}|\theta) = \int p(X^{pred}|\theta)\pi(\theta|X)d\theta$

Predictive Distribution

- Prior distribution: $\pi(\theta)$
- Likelihood: $p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$
- Posterior: $\pi(\theta|X)$
- Predictive density: $p(X^{pred}|\theta) = \int p(X^{pred}|\theta)\pi(\theta|X)d\theta$
  - to predict the next data point.
  - in practice, can sample X^{pred} using samples of θ from Gibbs sampling, then use Monte Carlo approximation again.
  - after all, we can view the predictive density as a function of θ.

Posterior Consistency

As we get more and more data, with appropriate prior information, can we recover the "truth"?

Posterior Consistency

As we get more and more data, with appropriate prior information, can we recover the "truth"?

- Posterior consistency is a freqeuntist justification of Bayesian methods.

Posterior Consistency

As we get more and more data, with appropriate prior information, can we recover the "truth"?

- Posterior consistency is a freqeuntist justification of Bayesian methods.
- It is frequentist because we assume there is a "true" parameter out there.

Posterior Consistency

As we get more and more data, with appropriate prior information, can we recover the "truth"?

- Posterior consistency is a freqeuntist justification of Bayesian methods.
- It is frequentist because we assume there is a "true" parameter out there.
- Does the posterior distribution converge to the point mass at the "true" parameter?

Posterior Consistency

As we get more and more data, with appropriate prior information, can we recover the "truth"?

- Posterior consistency is a freqeuntist justification of Bayesian methods.
- It is frequentist because we assume there is a "true" parameter out there.
- Does the posterior distribution converge to the point mass at the "true" parameter?
  - Converge in what sense?
  - How fast is the convergence rate?
  - How does the Bayesian estimate compare to MLE?

Why Bayesian?

Almost every statistical learning method has the "Bayesian" version!

Why Bayesian?

Almost every statistical learning method has the "Bayesian" version!

- To include quantitative prior judgments due to lack of data.

Why Bayesian?

Almost every statistical learning method has the "Bayesian" version!

- To include quantitative prior judgments due to lack of data.
- To construct hierarchical models on the assumption of shared prior distributions whose parameters can be estimated from the data.

Why Bayesian?

Almost every statistical learning method has the "Bayesian" version!

- To include quantitative prior judgments due to lack of data.
- To construct hierarchical models on the assumption of shared prior distributions whose parameters can be estimated from the data.
- To make inferences on a huge joint probability model where there are possibly thousands of observations and parameters.

Why Bayesian?

Almost every statistical learning method has the "Bayesian" version!

- To include quantitative prior judgments due to lack of data.
- To construct hierarchical models on the assumption of shared prior distributions whose parameters can be estimated from the data.
- To make inferences on a huge joint probability model where there are possibly thousands of observations and parameters.
- To use Bayesian ideas to quantify uncertainty of parameters.

Why Bayesian?

Almost every statistical learning method has the "Bayesian" version!

- To include quantitative prior judgments due to lack of data.
- To construct hierarchical models on the assumption of shared prior distributions whose parameters can be estimated from the data.
- To make inferences on a huge joint probability model where there are possibly thousands of observations and parameters.
- To use Bayesian ideas to quantify uncertainty of parameters.
- The "updating" inherent in the Bayesian approach is suitable in machine-learning.

Summary

- Classical likelihood-based inference closely resembles Bayesian inference using a flat prior. In many cases, estimates, intervals, and other decisions will be extremely similar for Bayesian and frequentist analyses.

Summary

- Classical likelihood-based inference closely resembles Bayesian inference using a flat prior. In many cases, estimates, intervals, and other decisions will be extremely similar for Bayesian and frequentist analyses.
- There is deep philosophical differences between Bayesian and frequentist inference.

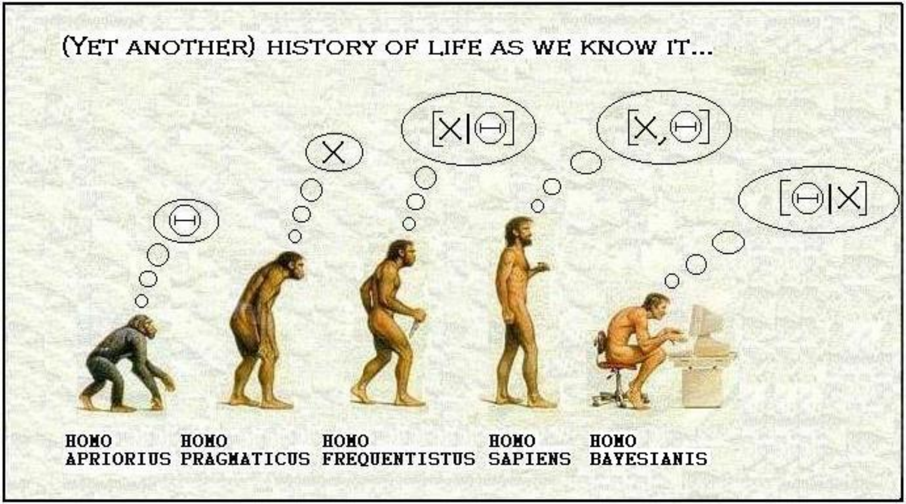
Summary

- Classical likelihood-based inference closely resembles Bayesian inference using a flat prior. In many cases, estimates, intervals, and other decisions will be extremely similar for Bayesian and frequentist analyses.
- There is deep philosophical differences between Bayesian and frequentist inference.
- Bayesian make statements about the relative evidence for parameter values given a dataset, while frequentists compare the relative chance of datasets given a parameter value.

Summary

- Classical likelihood-based inference closely resembles Bayesian inference using a flat prior. In many cases, estimates, intervals, and other decisions will be extremely similar for Bayesian and frequentist analyses.
- There is deep philosophical differences between Bayesian and frequentist inference.
- Bayesian make statements about the relative evidence for parameter values given a dataset, while frequentists compare the relative chance of datasets given a parameter value.
- Bayesian statistics is getting more and more popular due to its advantages and increase of computation power.

I'm a homo Bayesian



Outline

1. The Big Picture
2. The Bayesian "Recipe"
3. Conjugate Family
4. Posterior Inference
5. Code Demo

Recommended References

- Textbook: **A first course in Bayesian Statistic method, by Peter Hoff.**
- Lecture notes: **Introduction to Bayesian Statistics, by Brendon Brewer.**
- Talk slides: **Bayesian Statistics, a very brief introduction, by Ken Rice.**
- Websites:
  - **Scholarpedia entry on Bayesian statistics.**
  - **Bayesian statistics for beginners in simple English.**