Introduction to Bayesian Methods

Jessi Cisewski
Department of Statistics
Yale University

Sagan Summer Workshop 2016

Our goal: introduction to Bayesian methods

- Likelihoods
- Priors: conjugate priors, "non-informative" priors
- Posteriors

Related topics covered this week

- Markov chain Monte Carlo (MCMC)
- Selecting priors
- Bayesian modeling comparison
- Hierarchical Bayesian modeling

Some material is from Tom Loredo, Sayan Mukherjee, Beka Steorts

Likelihood Principle

All of the information in a sample is contained in the likelihood function, a density or distribution function.

Likelihood Principle

All of the information in a sample is contained in the likelihood function, a density or distribution function.

- The data are modeled by a likelihood function.

Likelihood Principle

All of the information in a sample is contained in the likelihood function, a density or distribution function.

- The data are modeled by a likelihood function.
- Not all statistical paradigms agree with this principle.

Likelihood functions

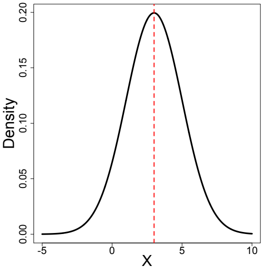
Consider a random sample of size $n = 1$ from a Normal($\mu = 3$, $\sigma = 2$): $X_1 \sim N(3, 2)$

- **Probability density function (pdf)**

  $\longrightarrow$ the function $f(x, \theta)$, where θ is fixed and x is variable

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$= \frac{1}{\sqrt{2\pi 2^2}} e^{-\frac{(x-3)^2}{(2)(2^2)}}$$



The data are drawn from this

- **Likelihood**

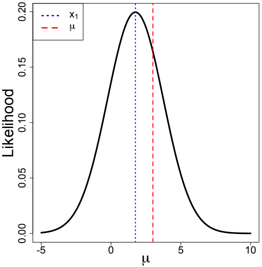
  $\longrightarrow$ the function $f(x, \theta)$, where θ is variable and x is fixed

Likelihood functions

Consider a random sample of size $n = 1$ from a Normal($\mu = 3$, $\sigma = 2$): $X_1 \sim N(3, 2)$

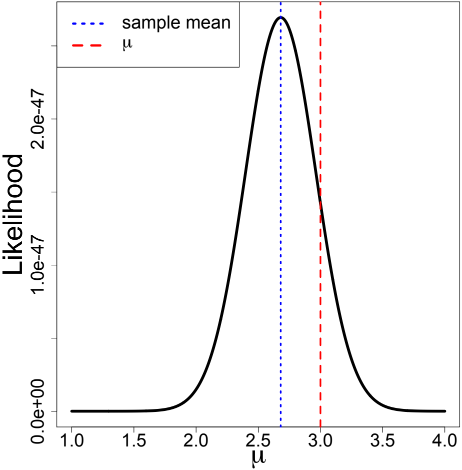
- **Probability density function (pdf)**
  $\longrightarrow$ the function $f(x, \theta)$, where θ is fixed and x is variable
- **Likelihood**
  $\longrightarrow$ the function $f(x, \theta)$, where θ is variable and x is fixed

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(1.747-\mu)^2}{2\sigma^2}}$$



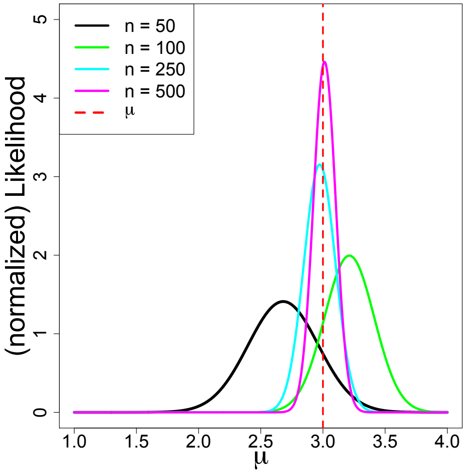
- Consider a random sample of size $n = 50$ (assuming independence, and a known σ): $X_1, \ldots, X_{50} \sim N(3, 2)$

$$f(x, \mu, \sigma) = f(x_1, \ldots, x_{50}, \mu, \sigma) = \prod_{i=1}^{50} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$



- Consider a random sample of size $n = 50$ (assuming independence, and a known σ): $X_1, \ldots, X_{50} \sim N(3, 2)$

$$f(x, \mu, \sigma) = f(x_1, \ldots, x_{50}, \mu, \sigma) = \prod_{i=1}^{50} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$



Likelihood Principle

All of the information in a sample is contained in the likelihood function, a density or distribution function.

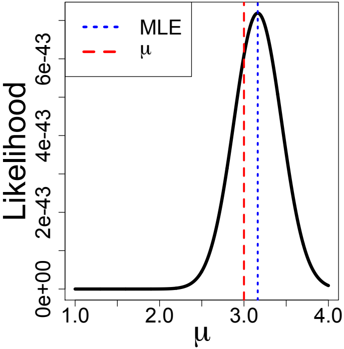
- The data are modeled by a likelihood function.
- How do we infer θ?

Maximum likelihood estimation

The parameter value, θ, that maximizes the likelihood:

$$\hat{\theta} = \max_{\theta} f(x_1, \ldots, x_n, \theta)$$

"Minimizing χ^2 statistic" (under the Gaussian assumption)



$$\max_\mu f(x_1, \ldots, x_n, \mu, \sigma) =$$
$$\max_\mu \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Hence, $\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$

Bayesian framework

- Classical or Frequentist methods for inference consider θ to be fixed and unknown
  $\longrightarrow$ performance of methods evaluated by repeated sampling
  $\longrightarrow$ consider all possible data sets

- Bayesian methods consider θ to be random
  $\longrightarrow$ only considers observed data set and prior information

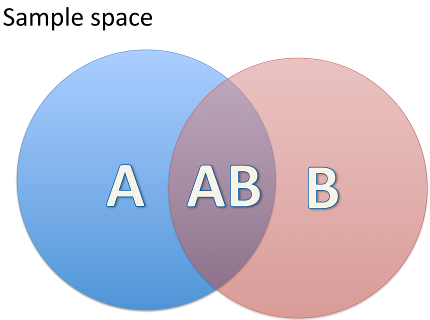
Bayes' Rule

Let A and B be two events in the sample space. Then

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

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

$\longrightarrow$ It is really just about conditional probabilities.



Posterior distribution

$$\pi(\theta \mid \overbrace{x}^{\text{Data}}) = \frac{\overbrace{f(x \mid \theta)}^{\text{Likelihood}} \cdot \overbrace{\pi(\theta)}^{\text{Prior}}}{f(x)} = \frac{f(x \mid \theta)\pi(\theta)}{\int_{\Theta} d\theta f(x \mid \theta)\pi(\theta)} \propto f(x \mid \theta)\pi(\theta)$$

- The prior distribution allows you to "easily" incorporate your beliefs about the parameter(s) of interest
- Posterior is a distribution on the parameter space given the observed data

Gaussian example

Consider $y_{1:n} = y_1, \ldots, y_n$ drawn from a Gaussian(μ, σ), μ unknown

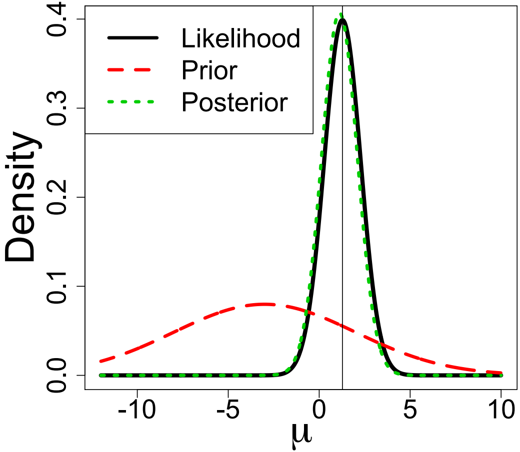
- Likelihood: $f(y_{1:n} \mid \mu) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\mu)^2}{2\sigma^2}} \right)$
- Prior: $\pi(\mu) \sim N(\mu_0, \sigma_0)$
- Posterior:

$$
\begin{aligned}
\pi(\mu \mid Y_{1:n}) &\propto f(Y_{1:n} \mid \mu)\pi(\mu) \\
&= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\mu)^2}{2\sigma^2}} \right) \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}} \right) \\
&\sim N(\mu_1, \sigma_1)
\end{aligned}
$$

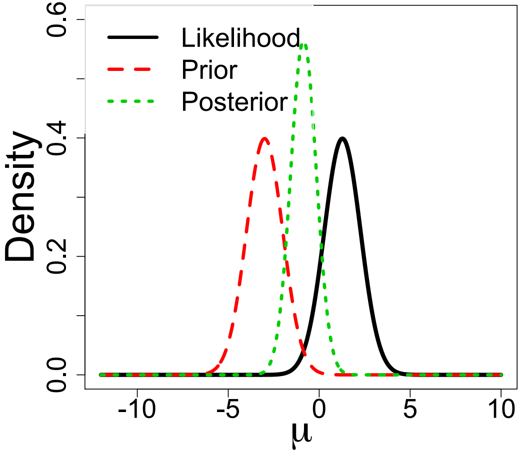
where

$$
\mu_1 = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{\sum y_i}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \qquad \sigma_1 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1/2}
$$

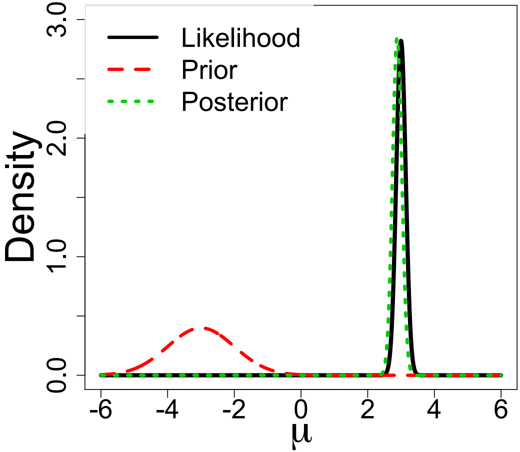
- Data: $y_1, \ldots, y_4 \sim N(\mu = 3, \sigma = 2)$, $\bar{y} = 1.278$
- Prior: $N(\mu_0 = -3, \sigma_0 = 5)$
- Posterior: $N(\mu_1 = 1.114, \sigma_1 = 0.981)$



- Data: $y_1, \ldots, y_4 \sim N(\mu = 3, \sigma = 2)$, $\bar{y} = 1.278$
- Prior: $N(\mu_0 = -3, \sigma_0 = 1)$
- Posterior: $N(\mu_1 = -0.861, \sigma_1 = 0.707)$



- Data: $y_1, \ldots, y_{200} \sim N(\mu = 3, \sigma = 2)$, $\bar{y} = 2.999$
- Prior: $N(\mu_0 = -3, \sigma_0 = 1)$
- Posterior: $N(\mu_1 = 2.881, \sigma_1 = 0.140)$



Example 2 - on your own

Consider the following model:

$$Y \mid \theta \sim U(0, \theta)$$
$$\theta \sim Pareto(\alpha, \beta)$$

- $\pi(\theta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}} \mathbf{I}_{(\beta,\infty)}(\theta)$
  where $\mathbf{I}_{(a,b)}(x) = 1$ if $a < x < b$ and 0 otherwise
- Find the posterior distribution of $\theta \mid y$

$$\begin{aligned}
\pi(\theta \mid y) &\propto \frac{1}{\theta} \mathbf{I}_{(0,\theta)}(y) \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}} \mathbf{I}_{(\beta,\infty)}(\theta) \\
&\propto \frac{1}{\theta} \mathbf{I}_{(y,\infty)}(\theta) \frac{1}{\theta^{\alpha+1}} \mathbf{I}_{(\beta,\infty)}(\theta) \\
&\propto \frac{1}{\theta^{\alpha+2}} \mathbf{I}_{(\max\{y,\beta\},\infty)}(\theta) \\
&\implies Pareto(\alpha+1, \max\{y, \beta\})
\end{aligned}$$

Prior distribution

- The prior distribution allows you to “easily” incorporate your beliefs about the parameter(s) of interest
- If one has a specific prior in mind, then it fits nicely into the definition of the posterior
- But how do you go from prior information to a prior distribution?
- And what if you don’t actually have prior information?

Choosing a prior

- Informative/Subjective prior: choose a prior that reflects our belief/uncertainty about the unknown parameter
  - Based on experience of the researcher from previous studies, scientific or physical considerations, other sources of information
  - ⋆ Example: For a prior on the mass of a star in a Milky Way-type galaxy, you likely would not use an infinite interval

- Objective, non-informative, vague, default priors

- Hierarchical models: put a prior on the prior

- Conjugate priors: priors selected for convenience

Conjugate priors

- The posterior distribution is from the same family of distributions as the prior

  We saw this with a Gaussian prior on μ resulted in a Gaussian posterior $\mu \mid Y_{1:n}$

  (Gaussian priors are conjugate with Gaussian likelihoods resulting in a Gaussian posterior)

Some conjugate priors

- **Normal - normal**: normal priors are conjugate with normal likelihoods
- **Beta - binomial**: beta priors are conjugate with binomial likelihoods
- **Gamma - Poisson**: gamma priors are conjugate with Poisson likelihoods
- **Dirichlet - multinomial**: Dirichlet priors are conjugate with multinomial likelihoods

Beta-Binomial

- Suppose we have an iid sample, $x_1, \ldots, x_n$, from a Bernoulli(θ)

$$\begin{aligned} X = 1, &\qquad \text{with probability } \theta \\ X = 0, &\qquad \text{with probability } 1 - \theta \end{aligned}$$

Let $y = \sum_{i=1}^{n} x_i \implies y$ is a draw from a Binomial(n, θ)

$$p(Y = k) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

We want the posterior distribution for θ

Beta-Binomial

- We have a binomial likelihood, and need to specify a prior on θ
  Note that $\theta \in [0, 1]$

- If prior $\pi(\theta) \sim \text{Beta}(\alpha,\ \beta)$, then posterior

$$\pi(\theta \mid y) \sim \text{Beta}(y + \alpha, n - y + \beta)$$

- The beta distribution is the conjugate prior for binomial likelihoods

Beta - Binomial posterior derivation

$$f(y,\theta) = \left\{ \binom{n}{y} \theta^{y}(1-\theta)^{n-y} \right\} \left\{ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \right\}$$

$$= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$$

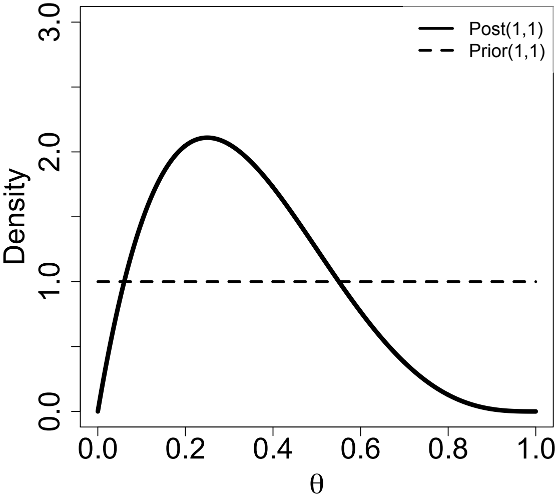
$$f(y) = \int_0^1 f(y,\theta)d\theta = \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)} \right)$$

$$\pi(\theta \mid y) = \frac{f(y,\theta)}{f(y)} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$$

$$\sim \text{Beta}(y+\alpha, n-y+\beta)$$

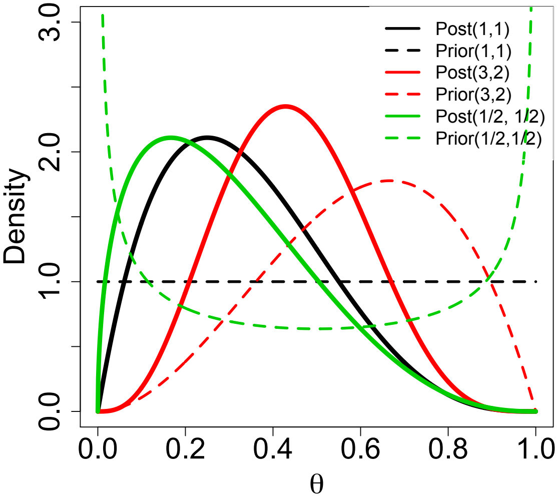
Beta priors and posteriors

$$\pi(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$



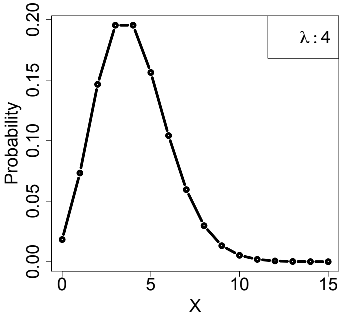
Beta priors and posteriors

$$\pi(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$



Poisson distribution

$$Y \sim \text{Poisson}(\lambda) \implies P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$$



- Mean = Variance = λ
- Bayesian inference on λ:

$$Y \mid \lambda \sim \text{Poisson}(\lambda)$$

What prior to use for λ? ($\lambda > 0$)

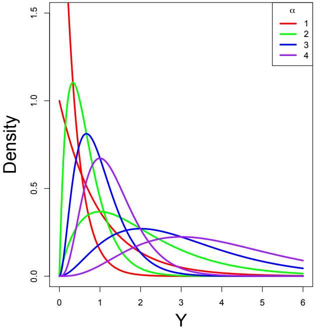
Astronomical example

⋆ Photons from distant quasars, cosmic rays

For more details see Feigelson and Babu (2012), Section 4.2

Gamma density

$$f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, y > 0$$



- Often written as $Y \sim \Gamma(\alpha, \beta)$
- $\alpha > 0$ (shape parameter), $\beta > 0$ (rate parameter)

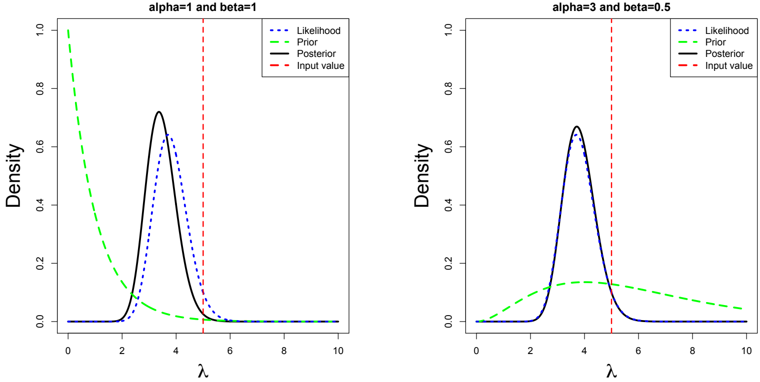
  Note: sometimes $\theta = 1/\beta$ is used instead
- Mean $= \alpha/\beta$, Variance $= \alpha/\beta^2$
- $\Gamma(1, \beta) \sim$ Exponential(β), $\Gamma(d/2, 1/2) \sim \chi^2_d$

Poisson - Gamma Posterior

$$\begin{cases} f(y_{1:n} \mid \lambda) & = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n}(y_i!)} \text{ (Likelihood)} \\ \pi(\lambda) & = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \text{ (Prior)} \\ & \\ \pi(\lambda \mid y_{1:n}) & \propto \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n}(y_i!)} \times \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \\ & \propto e^{-n\lambda}\lambda^{\sum_{i=1}^{n} y_i}\lambda^{\alpha-1}e^{-\beta\lambda} \\ & = e^{-\lambda(n+\beta)}\lambda^{\sum_{i=1}^{n} y_i+\alpha-1} \\ & \sim \Gamma\left(\sum y_i + \alpha, n + \beta\right) \end{cases}$$

- The gamma distribution is the conjugate prior for Poisson likelihoods

Poisson - Gamma Posterior illustrations



Same dataset

Hierarchical priors

A prior is put on the parameters of the prior distribution $\Longrightarrow$ the prior on the parameter of interest, θ, has additional parameters

$$
\begin{aligned}
Y \mid \theta, \gamma &\sim f(y \mid \theta, \gamma) \text{ (Likelihood)} \\
\Theta \mid \gamma &\sim \pi(\theta \mid \gamma) \text{ (Prior)} \\
\Gamma &\sim \phi(\gamma) \text{ (Hyper prior)}
\end{aligned}
$$

- It is assumed that $\phi(\gamma)$ is fully known, and γ is called a *hyper parameter*
- More layers can be added, but of course that makes the model more complex $\longrightarrow$ posterior may require computational techniques (e.g. MCMC)

Simple illustration

$$Y \mid (\mu, \phi) \sim N(\mu, 1) \text{ Likelihood}$$
$$\mu \mid \phi \sim N(\phi, 2) \text{ Prior}$$
$$\phi \sim N(0, 1) \text{ Hyperprior}$$

Maybe we want to put a hyperhyperprior on ϕ?

Posterior

$$\mu \mid Y \propto f(y \mid \mu, \phi)\pi_1(\mu \mid \phi)\pi_2(\phi)$$

Non-informative priors

What to do if we don't have relevant prior information? What if our model is too complex to know what reasonable priors are?

- Desire is for a prior that does not favor any particular value on the parameter space
- ⋆ Side note: some may have philosophical issues with this (e.g. R.A. Fisher, which lead to fiducial inference)

- We will discuss some methods for finding "non-informative priors." It turns out these priors can be improper (i.e. they integrate to ∞ rather than 1), so you need to verify that the resulting posterior distribution is proper

- Example of improper prior with proper posterior:
  Data: $x_1, \ldots, x_n \sim N(\theta, 1)$
  (Improper) prior: $\pi(\theta) \propto 1$
  (Proper) posterior: $\pi(\theta \mid x_{1:n}) \sim N(\bar{y}, n^{-1/2})$

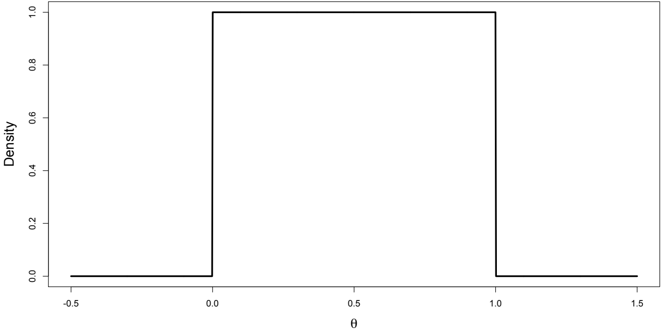
- Example of improper prior with improper posterior
  Data: $x_1, \ldots, x_n \sim Bernouilli(\theta)$, $y = \sum_{i=1}^{n} x_i \sim Binomial(n, \theta)$
  (Improper) prior: $\pi(\theta) \sim Beta(-1, -1)$
  (Improper) posterior: $\pi(\theta \mid x_{1:n}) \propto \theta^{y-1}(1-\theta)^{n-y-1}$
  This is improper for $y = 0$ or n

If you use improper priors, you have to check that the posterior is proper

Uniform prior

This is what many astronomers use for non-informative priors, and is often what comes to mind when we think of "flat" priors

$$\theta \sim U(0, 1)$$

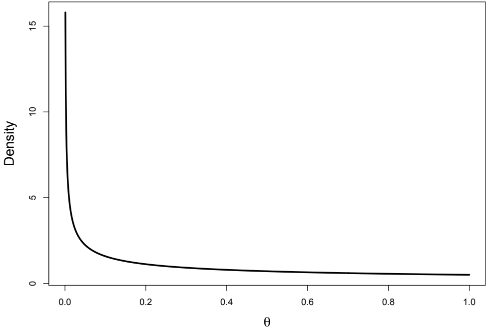


What if we consider a transformation of θ, such as θ^2?

Uniform prior

$$\theta \sim U(0, 1)$$

Prior for θ^2:



$\longrightarrow$ Notice that the above is not Uniform - the prior on θ^2 is *informative*. This is an undesirable property of Uniform priors. We would like the "un-informativeness" to be invariant under transformations.

There are a number of reasons why you may not have prior information:

1. your work may be the first of its kind
2. you are skeptical about previous results that would have informed your priors
3. the parameter space is too high dimensional to understand how your informative priors work together
4. . . .

If this is the case, then you may like the priors to have little effect on the resulting posterior

Objective priors

- Jeffreys' prior

  Uses Fisher information

- Reference priors

  Select priors that maximize some measure of divergence between the posterior and prior (hence minimizing the impact a prior has on the posterior)

  "The Formal Definition of Reference Priors" by Berger et al. (2009)

More about selecting priors can be found here: Kass and Wasserman (1996)

Jeffrey's Prior

$$\pi_J(\theta) \propto \sqrt{|I(\theta)|}$$

where $\mathbb{I}(\theta)$ is the Fisher information

$$\begin{aligned}\mathbb{I}(\theta) &= E\left(\frac{d}{d\theta}\log L(\theta \mid Y)\right)^2 \\ &= -E\left(\frac{d^2}{d\theta^2}\log L(\theta \mid Y)\right) \text{ (for exponential family)}\end{aligned}$$

Some intuition[1]

- $\mathbb{I}(\theta)$ is understood to be a proxy for the information content in the model about $\theta \rightarrow$ high values of $\mathbb{I}(\theta)$ correspond with likely values of θ. This reduces the effect of the prior on the posterior
- Most useful in single-parameter setting; not recommended with multiple parameters

[1] For more details see Robert (2007): "The Bayesian Choice"

Exponential example

$$f(y \mid \theta) = \theta e^{-\theta y}$$

Calculate the Fisher Information:

$\log(f(y \mid \theta)) = \log(\theta) - \theta y$

$\frac{d}{d\theta} \log(f(y \mid \theta)) = \frac{1}{\theta} - y$

$\frac{d^2}{d\theta^2} \log(f(y \mid \theta)) = -\frac{1}{\theta^2}$

$-E \frac{d^2}{d\theta^2} \log(f(y \mid \theta)) = \frac{1}{\theta^2}$

Hence,

$$\pi_J(\theta) \propto \sqrt{\frac{1}{\theta^2}} = \frac{1}{\theta}$$

Exponential example, continued

$$\pi_J(\theta) \propto \frac{1}{\theta}$$

- Suppose we consider $\phi = f(\theta) = \theta^2 \implies \theta = \sqrt{\phi}$
- $\frac{d\theta}{d\phi} = -\frac{1}{2\sqrt{\phi}}$
- Hence, $\pi'_J(\phi) = \pi_J(\sqrt{\phi}) \left| \frac{d\theta}{d\phi} \right| = \frac{1}{\sqrt{\phi}} \frac{1}{2\sqrt{\phi}} \propto \frac{1}{\phi}$

$$\pi'_J(\phi) \propto \frac{1}{\phi}$$

We see here that Jeffreys prior is invariant to the transformation $f(\theta) = \theta^2$

Binomial example

$$f(Y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Calculate the Fisher Information:

$\log(f(Y \mid \theta)) = \log(\binom{n}{y}) + y \log(\theta) + (n-y) \log(1-\theta)$

$\frac{d}{d\theta} \log(f(Y \mid \theta)) = \frac{y}{\theta} - \frac{(n-y)}{1-\theta}$

$\frac{d^2}{d\theta^2} \log(f(Y \mid \theta)) = -\frac{y}{\theta^2} - \frac{(n-y)}{(1-\theta)^2}$

$\rightarrow$ Note that $E(y) = n\theta$

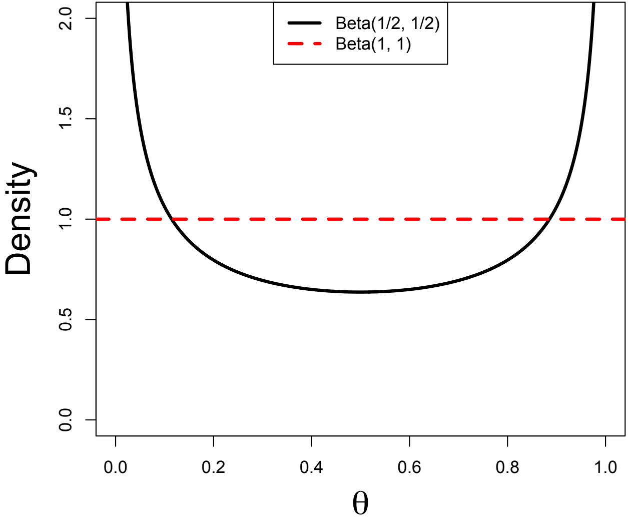
$-E \frac{d^2}{d\theta^2} \log(f(Y \mid \theta)) = \frac{n\theta}{\theta^2} + \frac{(n-n\theta)}{(1-\theta)^2} = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)}$

Hence,

$$\pi_J(\theta) \propto \sqrt{\frac{n}{\theta(1-\theta)}} \propto \theta^{-\frac{1}{2}} (1-\theta)^{-\frac{1}{2}}$$

Beta$(\frac{1}{2}, \frac{1}{2})$

$$\pi_J(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$$



If we use Jeffreys' prior $\pi_J(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$, what is the posterior for $\theta \mid Y$?

$$\begin{aligned}
\pi(\theta \mid Y) &\propto \binom{n}{y}\theta^y(1-\theta)^{n-y}\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \\
&\propto \theta^y(1-\theta)^{n-y}\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \\
&\propto \theta^{y-1/2}(1-\theta)^{n-y-1/2} \\
&\sim \text{Beta}(y+1/2, n-y+1/2)
\end{aligned}$$

The posterior distribution is proper (which we knew would be the case since the prior is proper)

It is just a coincidence that the Jeffreys prior is conjugate.

Gaussian with unknown μ - on your own

$$f(Y \mid \mu, \sigma^2) \propto e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Calculate the Fisher Information:

$\log(f(Y \mid \theta)) = -\frac{(y-\mu)^2}{2\sigma^2}$

$\frac{d}{d\theta}\log(f(Y \mid \theta)) = 2\frac{(y-\mu)}{2\sigma^2} = \frac{(y-\mu)}{\sigma^2}$

$\frac{d^2}{d\theta^2}\log(f(Y \mid \theta)) = -\frac{1}{\sigma^2}$

$-E\frac{d^2}{d\theta^2}\log(f(Y \mid \theta)) = \frac{1}{\sigma^2}$

Hence,

$$\pi_J(\mu) \propto 1$$

Inference with a posterior

Now that we have a posterior, what do we want to do with it?

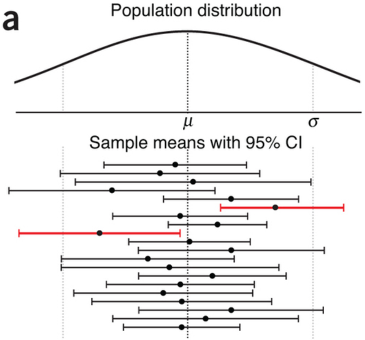
- Point estimation:
  - posterior mean: $<\theta>= \int_{\Theta} d\theta p(\theta \mid Y, model)$
  - posterior mode (MAP = maximum a posteriori)
- Credible regions: posterior probability p that θ falls in regions R $p = P(\theta \in R \mid Y, model) = \int_{R} d\theta p(\theta \mid Y, model)$ highest posterior density (HPD) region
- Posterior predictive distributions: predict new $\tilde{y}$ given data y

- Posterior predictive distributions: predict new $\tilde{y}$ given data y

$$
\begin{aligned}
f(\tilde{y} \mid y) &= \frac{f(\tilde{y}, y)}{f(y)} = \frac{\int f(\tilde{y}, y, \theta) d\theta}{f(y)} = \frac{\int f(\tilde{y} \mid y, \theta) f(y, \theta) d\theta}{f(y)} \\
&= \int f(\tilde{y} \mid y, \theta) \pi(\theta \mid y) d\theta \\
&= \int f(\tilde{y} \mid \theta) \pi(\theta \mid y) d\theta \text{ (if y and } \tilde{y} \text{ are independent)}
\end{aligned}
$$

Confidence intervals $\neq$ credible intervals

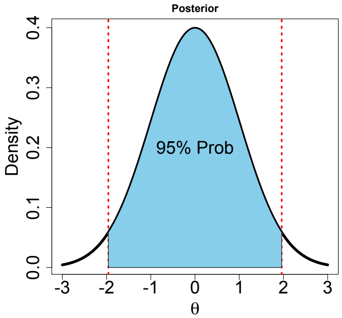
A 95% confidence interval is based on repeated sampling of datasets - about 95% of the confidence intervals will capture the true parameter value



Parameters are **not** random

http://www.nature.com

A 95% credible interval is defined using the posterior distribution



Parameters are random

Summary

- We discussed some basics of Bayesian methods
- Bayesian inference relies on the posterior distribution

$$\pi(\theta \mid \overbrace{x}^{\text{Data}}) = \frac{\overbrace{f(x \mid \theta)}^{\text{Likelihood}} \cdot \overbrace{\pi(\theta)}^{\text{Prior}}}{f(x)}$$

- There are different ways to select priors: subjective, conjugate, “non-informative”
- Credible intervals and confidence intervals have different interpretations
- We’ll be hearing a lot more about Bayesian methods throughout the week.

Summary

- We discussed some basics of Bayesian methods
- Bayesian inference relies on the posterior distribution

$$\pi(\theta \mid \overbrace{x}^{\text{Data}}) = \frac{\overbrace{f(x \mid \theta)}^{\text{Likelihood}} \cdot \overbrace{\pi(\theta)}^{\text{Prior}}}{f(x)}$$

- There are different ways to select priors: subjective, conjugate, "non-informative"
- Credible intervals and confidence intervals have different interpretations
- We'll be hearing a lot more about Bayesian methods throughout the week.

Thank you!

Bibliography I

Berger, J. O., Bernardo, J. M., and Sun, D. (2009), "The formal definition of reference priors," *The Annals of Statistics*, 905–938.

Feigelson, E. D. and Babu, G. J. (2012), *Modern Statistical Methods for Astronomy: With R Applications*, Cambridge University Press.

Kass, R. E. and Wasserman, L. (1996), "The selection of prior distributions by formal rules," *Journal of the American Statistical Association*, 91, 1343–1370.

Robert, C. (2007), *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Springer Science & Business Media.