

# STA 702: Lecture 1

## Basics of Bayesian Statistics

Merlise Clyde



# Ingredients

(1) **Prior Distribution**  $\pi(\theta)$  for unknown  $\theta$

(2) **Likelihood Function**  $\mathcal{L}(\theta | y) \propto p(y | \theta)$  (sampling model)

(3) **Posterior Distribution**

$$\pi(\theta|y) = \frac{\pi(\theta)p(y | \theta)}{\int_{\Theta} \pi(\theta)p(y | \theta)d\theta} = \frac{\pi(\theta)p(y | \theta)}{p(y)}$$

(4) **Loss Function** Depends on what you want to report; estimate of  $\theta$ , predict future  $Y_{n+1}$ , etc



# Posterior Depends on Likelihoods

- Likelihood is defined up to a constant

$$c \mathcal{L}(\theta | Y) = p(y | \theta)$$

- Bayes' Rule

$$\pi(\theta|y) = \frac{\pi(\theta)p(y | \theta)}{\int_{\Theta} \pi(\theta)p(y | \theta)d\theta} = \frac{\pi(\theta)c\mathcal{L}(\theta | y)}{\int_{\Theta} \pi(\theta)c\mathcal{L}(\theta | y)d\theta} = \frac{\pi(\theta)\mathcal{L}(\theta | y)}{m(y)}$$

- $m(y)$  is proportional to the marginal distribution of data

$$m(y) = \int_{\Theta} \pi(\theta)\mathcal{L}(\theta | y)d\theta$$

- marginal likelihood of this model or "evidence"

**Note:** the marginal likelihood and maximized likelihood are *very* different!



# Binomial Example

$$Y \mid n, \theta \sim \text{Binomial}(n, \theta)$$

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

$$\mathcal{L}(\theta \mid y) = \theta^y (1 - \theta)^{n-y}$$

MLE  $\hat{\theta}$  of Binomial is  $\bar{y} = y/n$  proportion of successes

Recall Derivation:



# Marginal Likelihood

$$m(y) = \int_{\Theta} \mathcal{L}(\theta | y) \pi(\theta) d\theta = \int_{\Theta} \theta^y (1 - \theta)^{n-y} \pi(\theta) d\theta$$

"Averaging" likelihood over prior



# Binomial Example

- **Prior**  $\theta \sim U(0, 1)$  or  $\pi(\theta) = 1$ , for  $\theta \in (0, 1)$

- **Marginal**

$$m(y) = \int_0^1 \theta^y (1 - \theta)^{n-y} 1 \, d\theta$$

$$m(y) = \int_0^1 \theta^{(y+1)-1} (1 - \theta)^{(n-y+1)-1} 1 \, d\theta = B(y + 1, n - y + 1)$$

- Special function known as the **beta function** (see Rudin)

$$B(a, b) = \int_0^1 \theta^{a-1} (1 - \theta)^{b-1} \, d\theta$$

## Posterior Distribution

$$\pi(\theta \mid y) = \frac{1}{B(y + 1, n - y + 1)} \theta^{(y+1)-1} (1 - \theta)^{(n-y+1)-1} \quad \theta \mid y \sim \text{Beta}((y + 1, n - y + 1))$$



# Beta Prior Distributions

**Beta(a,b)** is a probability density function (pdf) on (0,1),

$$\pi(\theta) = \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}$$

Use the "**kernel**" trick

$$\pi(\theta \mid y) \propto \mathcal{L}(\theta \mid y) \pi(\theta)$$



# Prior to Posterior Updating

- **Prior**  $\text{Beta}(a, b)$
- **Posterior**  $\text{Beta}(a + y, b + n - y)$
- **Conjugate** prior & posterior distribution are in the same family of distributions, (Beta)
- Simple updating of information from the prior to posterior
  - $a + b$  "prior sample size" (number of trials in a hypothetical experiment)
  - $a$  "number of successes"
  - $b$  "number of failures"
- Should be easy to do "prior elicitation " (process of choosing the prior hyperparameters)





# Summaries & Properties

Recall that for  $\theta \sim \text{Beta}(a, b)$   $a + b = n_0$

$$\mathbb{E}[\theta] = \frac{a}{a + b} \equiv \theta_0$$

Posterior mean

$$\mathbb{E}[\theta \mid y] = \frac{a + y}{a + b + n} \equiv \tilde{\theta}$$

Rewrite with MLE  $\hat{\theta} = \bar{y} = \frac{y}{n}$  and prior mean

$$\mathbb{E}[\theta \mid y] = \frac{a + y}{a + b + n} = \frac{n_0}{n_0 + n} \theta_0 + \frac{n}{n_0 + n} \hat{\theta}$$

Weighted average of prior mean and MLE where weight for  $\theta_0 \propto n_0$  and weight for  $\hat{\theta} \propto n$



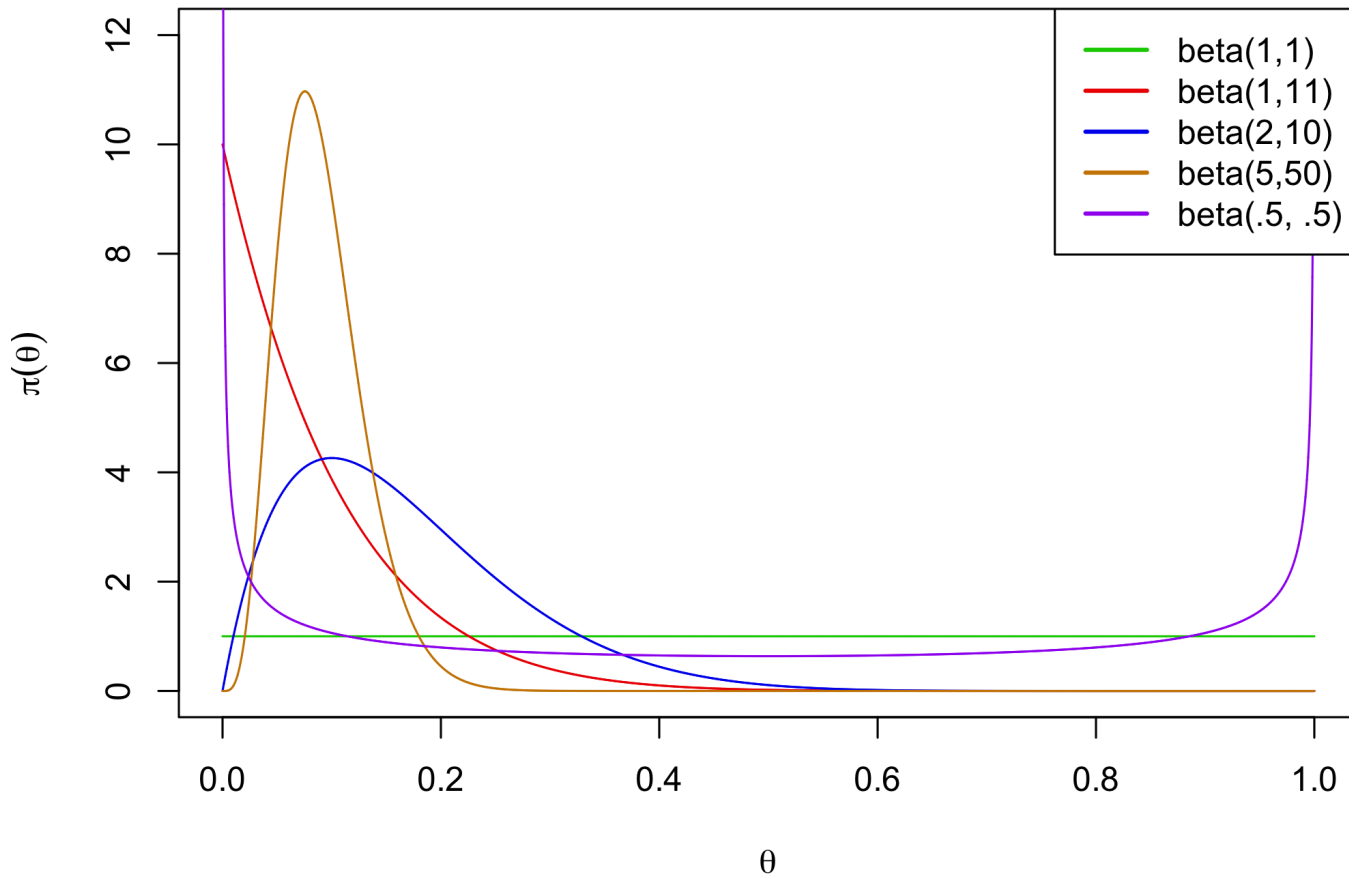
# Properties

$$\tilde{\theta} = \frac{n_0}{n_0 + n} \theta_0 + \frac{n}{n_0 + n} \hat{\theta}$$

- in finite samples we get **shrinkage**: posterior mean pulls the MLE toward the prior mean; amount depends on prior sample size  $n_0$  and data sample size  $n$
- **regularization** effect to reduce Mean Squared Error for estimation with small sample sizes and noisy data
  - introduces some bias (in the frequentist sense) due to prior mean  $\theta_0$
  - reduces variance (bias-variance trade-off)
- helpful in the Binomial case, when sample size is small or  $\theta_{\text{true}} \approx 0$  (rare events) and  $\hat{\theta} = 0$  (inbalanced categorical data)
- as we get more information from the data  $n \rightarrow \infty$  we have  $\tilde{\theta} \rightarrow \hat{\theta}$  and **consistency** ! As  $n \rightarrow \infty, E[\tilde{\theta}] \rightarrow \theta_{\text{true}}$



# Some possible prior densities



# Prior Choice

- Is the uniform prior  $\text{Beta}(1, 1)$  non-informative?
  - No- if  $y = 0$  (or  $n$ ) sparse/rare events saying that we have a prior "historical" sample with 1 success and 1 failure ( $a = 1$  and  $b = 1$ ) can be very informative
- What about a uniform prior on the log odds?  $\eta \equiv \log\left(\frac{\theta}{1-\theta}\right)$ ?

$$\pi(\eta) \propto 1, \quad \eta \in \mathbb{R}$$

- Is this a **proper** prior distribution?
- what would be induced measure for  $\theta$ ?
- Find Jacobian

$$\pi(\theta) \propto \theta^{-1}(1 - \theta)^{-1}, \quad \theta \in (0, 1)$$

- limiting case of a Beta  $a \rightarrow 0$  and  $b \rightarrow 0$  (Haldane's prior)



# Formal Bayes

- use of improper prior and turn the Bayesian crank
- calculate  $m(y)$  and renormalize likelihood times "improper prior" if  $m(y)$  is finite
- formal posterior is  $\text{Beta}(y, n - y)$  and reasonable only if  $y \neq 0$  or  $y \neq n$  as  $B(0, -)$  and  $B(-, 0)$  (normalizing constant) are undefined!
- no shrinkage  $E[\theta \mid y] = \frac{y}{n} = \tilde{\theta} = \hat{\theta}$



# Invariance

Jeffreys argues that priors should be invariant to transformations to be non-informative

i.e. if we reparameterize with  $\theta = h(\rho)$  then the rule should be that

$$\pi_{\theta}(\theta) = \left| \frac{d\rho}{d\theta} \right| \pi_{\rho}(h^{-1}(\theta))$$

Jefferys' rule is to pick  $\pi(\rho) \propto (I(\rho))^{1/2}$

Expected Fisher Information for  $\rho$

$$I(\rho) = -\mathbb{E} \left[ \frac{d^2 \log(\mathcal{L}(\rho))}{d^2 \rho} \right]$$

For the Binomial example  $\pi(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$

Thus Jefferys' prior is a Beta(1/2, 1/2)



# Why ?

Chain Rule!

Find Jefferys' prior for  $\theta$

Find information matrix for  $\rho$  from  $I(\theta)$

Show that the prior satisfies the invariance property!

