Lecture 3: One Parameter Models

Professor Alexander Franks

1/22/24

Announcements

- Reading: Chapter 2 and 3, Bayes Rules
- Homework due: January 28, at midnight

Sunday.

- Quiz in Section.

Bayesian Inference

- In frequentist inference, θ is treated as a fixed unknown constant
- In Bayesian inference, θ is treated as a random variable
- Need to specify a model for the joint distribution

$$p(y, \theta) = \underbrace{p(y \mid \theta)}_{\text{Sampling Distribution. } L(\theta)}\underbrace{p(\theta)}_{\text{prior Distribution}} \longrightarrow \underbrace{P(\theta / y)}_{\text{posterior Distribution.}}$$

Setup

- The *sample space* $\mathcal{Y}$ is the set of all possible datasets. We observe one dataset y from which we hope to learn about the world.
    - Y is a random variable, y is a realization of that random variable

      data.
- The *parameter space* Θ is the set of all possible parameter values θ
    - θ encodes the population characteristics that we want to learn about!

Bayesian Inference in a Nutshell

1. The *prior distribution* $p(\theta)$ describes our belief about the true population characteristics, for each value of $\theta \in \Theta$.

2. Our *sampling model* $p(y \mid \theta)$ describes our belief about what data we are likely to observe when the true population parameter is θ.

3. Once we actually observe data, y, we update our beliefs about θ by computing *the posterior distribution* $p(\theta \mid y)$. We do this with Bayes' rule!

Bayes' Rule

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

- $P(A \mid B)$ is the conditional probability of A given B
- $P(B \mid A)$ is the conditional probability of B given A
- $P(A)$ and $P(B)$ are called the marginal probability of A and B (unconditional)

Bayes' Rule for Bayesian Statistics

$$P(\theta \mid y) = \frac{P(y \mid \theta)P(\theta)}{P(y)}$$

- $P(\theta \mid y)$ is the posterior distribution Belief after seeing data
- $L(\theta) \propto P(y \mid \theta)$ is the likelihood (Sampling model)
- $P(\theta)$ is the prior distribution (Belief before data)
- $P(y) = \int_\Theta p(y \mid \tilde{\theta})p(\tilde{\theta})d\tilde{\theta}$ is the model evidence

$$P(\theta \mid y) \propto L(\theta)\,P(\theta)\,\cancel{\frac{1}{P(y)}}$$

Computing the Posterior Distribution

$$
\begin{aligned}
P(\theta \mid y) &= \frac{P(y \mid \theta)P(\theta)}{P(y)} \\
&\propto P(y \mid \theta)P(\theta) \\
&\propto L(\theta)P(\theta)
\end{aligned}
$$

- Start with a subjective belief (prior)
- Update it with evidence from data (likelihood)
- Summarize what you learn (posterior)

The posterior is proportional to the likelihood times the prior!

Bayesian vs Frequentist

θ

- In frequentist inference, unknown parameters treated as constants
  - Estimators are random (due to sampling variability)
  - Asks: what would I expect to see if I repeated the experiment?"

("counterfactual world")

Bayesian vs Frequentist

- In frequentist inference, unknown parameters treated as constants
  - Estimators are random (due to sampling variability)
  - Asks: what would I expect to see if I repeated the experiment?"
- In Bayesian inference, unknown parameters are random variables.
  - Need to specify a prior distribution for θ (not easy)
  - Asks: "what do I *believe* are plausible values for the unknown parameters given the data?"
  - Who cares what might have happened, focus on what *did* happen by conditioning on observed data.

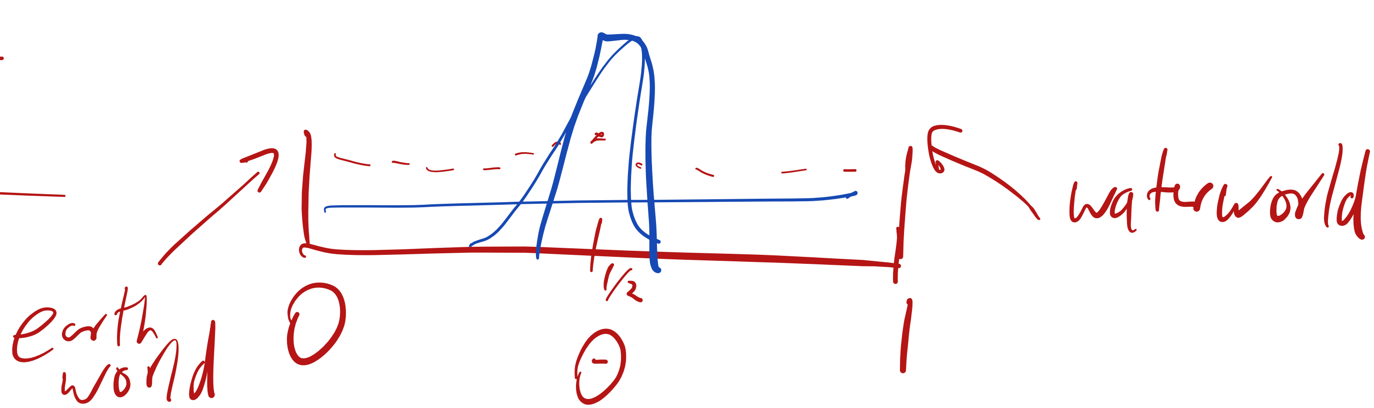
$P(\theta | y)$

"this happened"

Example

- Assume we sample the a point on the Earth and record whether it is land or water
- Let $Y \sim \mathrm{Bin}(n, \theta)$ where θ ~~corresponds to his true skill~~ is fraction of earth covered in water.
- Frequentist inference tells us that the maximum likelihood estimate is simply $\frac{y}{n}$ ← # of waters ← # of spins. MLE
- What would our estimates be if we use Bayesian inference?
  - What properties do we want for our prior distribution?

$P(\theta)$:



Cromwell's Rule

The use of priors placing a probability of 0 or 1 on events should be avoided except where those events are excluded by logical impossibility.

If a prior places probabilities of 0 or 1 on an event, then no amount of data can update that prior.

> I beseech you, in the bowels of Christ, think it possible that you may be mistaken.
> — Oliver Cromwell

$P(\theta | y) \propto L(\theta) P(\theta)$

Cromwell's Rule

> Leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million, but have it there since otherwise an army of astronauts returning with samples of the said cheese will leave you unmoved.
> — Dennis Lindley (1991)

If $p(\theta = a) = 0$ for a value of a, then the posterior distribution is always zero, regardless of what the data says

$$p(\theta = a|y) \propto p(y|\theta = a)\overbrace{p(\theta = a)}^{0} = 0$$

The Binomial Model

- The uniform prior:
  $p(\theta) = \mathrm{Unif}(0,1) = \mathbf{1}\{\theta \in [0,1]\}$
  - A "non-informative" prior
- Posterior: $p(\theta \mid y) \propto \underbrace{\theta^y(1-\theta)^{n-y}}_{\text{likelihood}} \times \underbrace{\mathbf{1}\{\theta \in [0,1]\}}_{\text{prior}}$
- The above posterior density is is a density over θ.

$P(\theta | y) \propto \theta^y (1-\theta)^{n-y}$

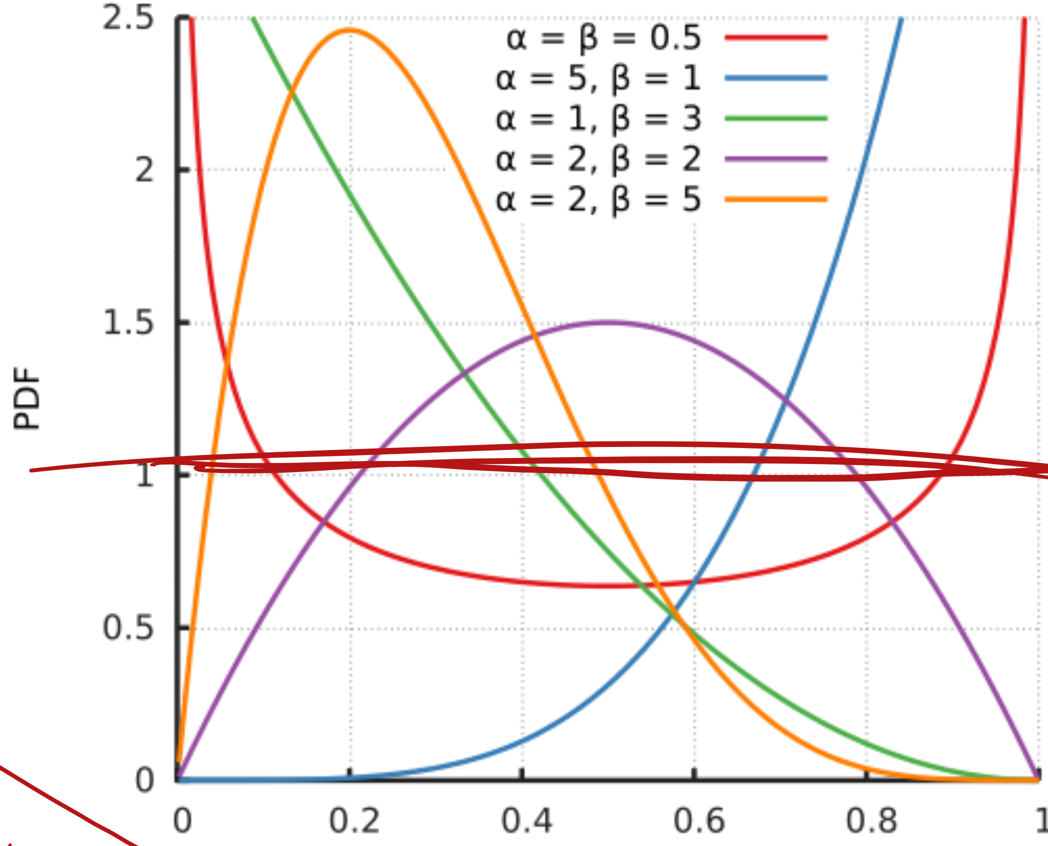
The Binomial Model

- The uniform prior:
  $p(\theta) = \mathrm{Unif}(0,1) = \mathbf{1}\{\theta \in [0,1]\}$
  - A "non-informative" prior
- Posterior: $p(\theta \mid y) \propto \underbrace{\theta^y (1-\theta)^{n-y}}_{\text{likelihood}} \times \underbrace{\mathbf{1}\{\theta \in [0,1]\}}_{\text{prior}}$
- The above posterior density is is a density over θ.

$$p(\theta \mid y) \sim \mathrm{Beta}(y+1, n-y+1)$$
$$= \frac{\Gamma(n)}{\Gamma(n-y)\Gamma(y)} \theta^y (1-\theta)^{n-y}$$

normalizing constant

Beta Distributions



Gamma Function, generalization of factorial(!)

Beta(1, 1) is Unif.

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

Sequential Bayesian Updating

new prior for the next obs.

$$\underbrace{p(\theta)}_{\text{prior}} \rightarrow \underbrace{p(\theta \mid y_1)}_{\text{reveal the first observation}} \rightarrow \underbrace{p(\theta \mid y_1, y_2)}_{\text{reveal the second observation}}$$

(one spin) (second spin)

Unif.

$$\rightarrow \cdots \rightarrow \underbrace{p(\theta \mid y_1, \ldots, y_n)}_{\text{reveal all } n \text{ observations}}$$

When data are i.i.d., final posterior is the same, regardless of whether we analyze data sequentially or as a single batch.

$$L(\theta) \propto P(y_1, \ldots y_n \mid \theta) \propto \prod P(y_i \mid \theta)$$

$$P(\theta \mid y) \propto P(y_3 \mid \theta)\, P(y_2 \mid \theta)\, \underbrace{P(y_1 \mid \theta)\, P(\theta)}_{P_1(\theta)}$$

Demo

$P(\theta \mid y_1, y_2) = \tilde{P}_2(\theta)$

Summarizing Posterior Results

- An entire distribution describes our beliefs about the value for θ. How can we summarize these beliefs?

Summarizing Posterior Results

- An entire distribution describes our beliefs about the value for θ. How can we summarize these beliefs?

- *Point estimates:* posterior mean or mode:
  - $E[\theta \mid y] = \int_{\Theta} \theta p(\theta \mid y) d\theta$ (the posterior mean)
  - $\arg\max p(\theta \mid y)$ (*maximum a posteriori* estimate)

Summarizing Posterior Results

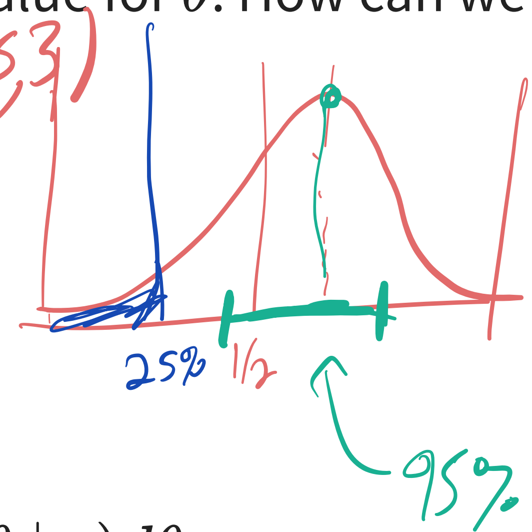
- An entire distribution describes our beliefs about the value for θ. How can we summarize these beliefs?

- *Point estimates:* posterior mean or mode:
  - $E[\theta \mid y] = \int_\Theta \theta p(\theta \mid y) d\theta$ (the posterior mean)
  - $\arg\max p(\theta \mid y)$ (*maximum a posteriori* estimate)

- Posterior variance: $\mathrm{Var}[\theta \mid y] = \int_\Theta (\theta - E[\theta \mid y])^2 p(\theta \mid y) d\theta$

Summarizing Posterior Results

- An entire distribution describes our beliefs about the value for θ. How can we summarize these beliefs?

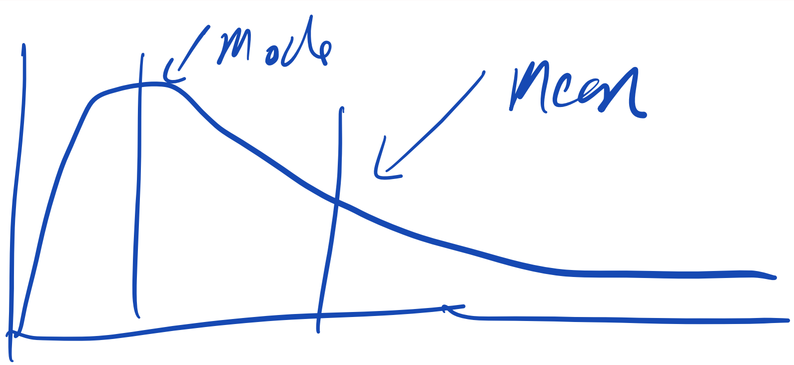


- *Point estimates:* posterior mean or mode:
  - $E[\theta \mid y] = \int_\Theta \theta p(\theta \mid y) d\theta$ (the posterior mean)
  - $\arg\max p(\theta \mid y)$ (*maximum a posteriori* estimate)

(MAP)

- Posterior variance: $\mathrm{Var}[\theta \mid y] = \int_\Theta (\theta - E[\theta \mid y])^2 p(\theta \mid y) d\theta$

- Posterior credible intervals: for any region $R(y)$ of the parameter space compute the probability that θ is in that region: $p(\theta \in R(y))$



Summarizing Posterior Results

Memorize

- $\text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The mean of a $\text{Beta}(\alpha, \beta)$ distribution r.v. $\frac{\alpha}{\alpha+\beta}$
- The mode of a $\text{Beta}(\alpha, \beta)$ distributed r.v. is $\frac{\alpha-1}{\alpha+\beta-2}$ (MAP)
- The variance of a $\text{Beta}(\alpha, \beta)$ r.v. is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- In R: `dbeta`, `rbeta`, `pbeta`, `qbeta`

Beta(5,3)

Mean: 5/8

Example

- On November 18, 2017, an NBA basketball player, Robert Covington, had made 49 out of 100 three point shot attempts.
- At that time, his three point field goal percentage, 0.49, was the best in the league and would have ranked in the top ten all time $\parallel$ $\hat{\theta}_{MLE} = 49\%$
- How can we estimate his true shooting skill?
  - Think of “true shooting skill” as the fraction he would make if he took infinitely many shots

Example

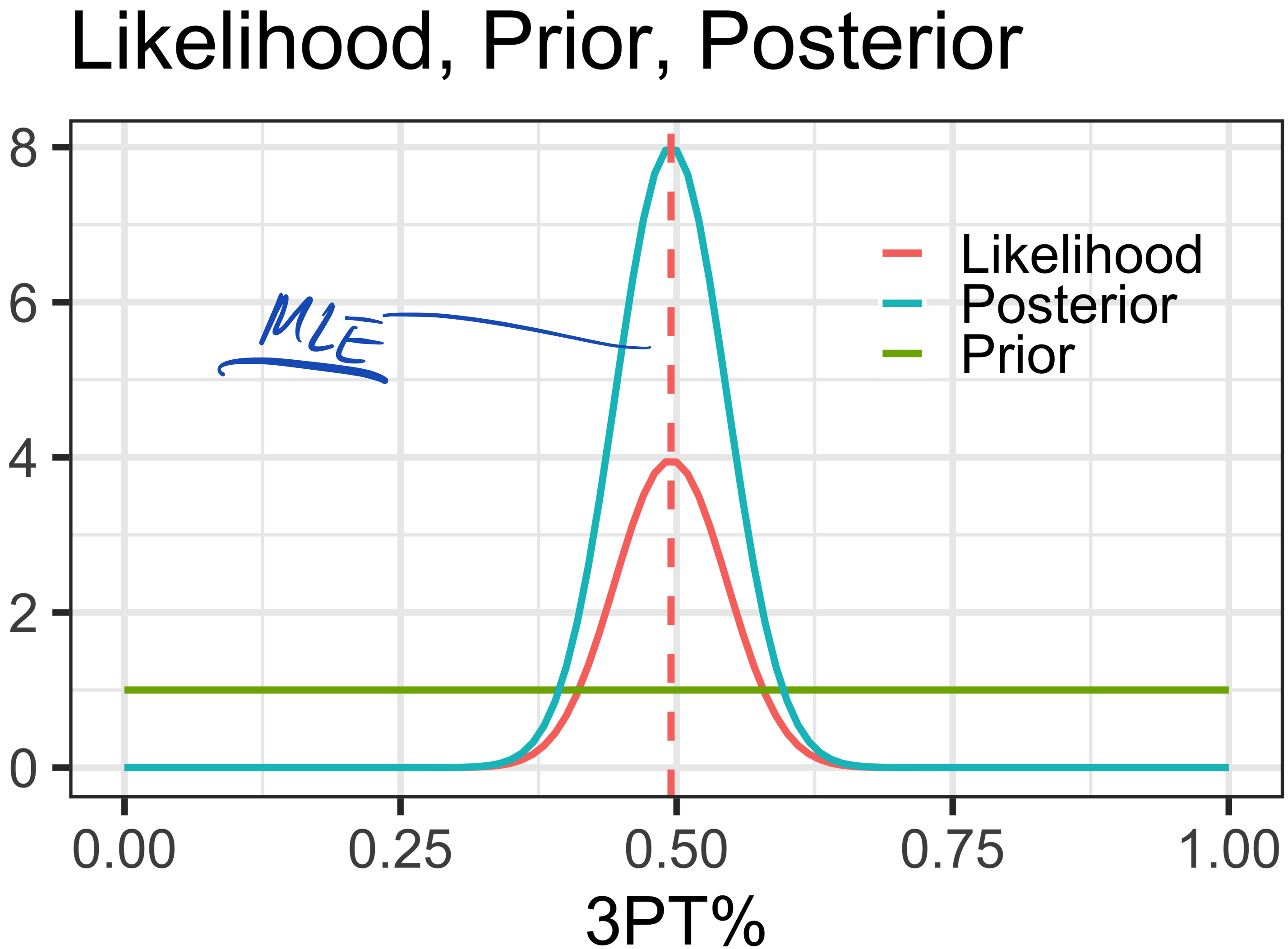
- Assume every shot is independent (reasonable) and identically distributed (less reasonble?)
- Let $Y \sim \text{Bin}(n, \theta)$ where θ corresponds to his true skill
- Frequentist inference tells us that the maximum likelihood estimate is simply $\frac{y}{n} = 49/100 = 0.49$
- What would our estimates be if we use Bayesian inference?

$$y \sim \text{Bin}(100, \theta)$$

$$P(\theta) \sim \text{Unif}[0, 1]$$

$$P(\theta | y) \propto \theta^{y}(1-\theta)^{n-y} \times 1 \rightarrow \text{Beta}(y+1, n-y+1)$$

Example



Posterior is proportional to the likelihood

Example

- Assume every shot is independent (reasonable) and identically distributed (less reasonable?)
- Let $Y \sim \text{Bin}(n, \theta)$ where θ corresponds to his true skill
- Frequentist inference tells us that the maximum likelihood estimate is simply $\frac{y}{n} = 49/100 = 0.49$
- What would our estimates be if we use Bayesian inference?
  - If our prior reflects "complete ignorance" about basketball?
  - What if we want to incorporate prior domain knowledge?

How to operationalize?

+ Other players
+ R.C. past play.
+ R.C. playing style / kind of shots.

Informative prior distributions

- At that time, his three point field goal percentage, 0.49, was the best in the league and would have ranked in the top ten all time
- It seems very unlikely that this level of skill would continue for an entire season of play.
- A uniform prior distribution doesn't reflect our known beliefs. We need to choose a more *informative* prior distribution

Informative prior distributions

- When $p(\theta) \sim \underbrace{U(0,1)}_{}$ then the posterior was a Beta distribution

  Beta(1,1)

- Remember: the binomial likelihood is $L(\theta) \propto \theta^y(1-\theta)^{n-y}$
- Choose a prior with a similar looking form: $p(\theta) \propto \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{Beta}(\alpha,\beta)}$

$$P(\theta|y) \propto \underbrace{\theta^{y}(1-\theta)^{n-y}}_{L(\theta)}\ \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{P(\theta)}$$

$$\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \rightarrow$$

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

Informative prior distributions

- When $p(\theta) \sim U(0,1)$ then the posterior was a Beta distribution
- Remember: the binomial likelihood is $L(\theta) \propto \theta^y(1-\theta)^{n-y}$
- Choose a prior with a similar looking form: $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- Then $p(\theta \mid y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$ is a $\mathrm{Beta}(y+\alpha, n-y+\beta)$
- For the binomial model, a beta prior distribution implies a beta posterior distribution!
- The family of Beta distributions is called a **conjugate prior** distribution for the binomial likelihood.

Conjugate Prior Distributions

e.g. Beta's

Definition: A class of prior distributions, $\mathcal{P}$ for θ is called *conjugate* for a sampling model $p(Y|\theta)$ if $p(\theta) \in \mathcal{P} \implies p(\theta|y) \in \mathcal{P}$

(e.g Bin)

- The prior distribution and the posterior distribution are in the same family
- Conjugate priors are very convenient because they make calculations easy
- The parameters for conjugate prior distribution have nice interpretations . . .
- Note: convenience is not correctness. Best to choose prior distributions that reflect your true knowledge / experience, not convenience. We'll return to this later.

Prior: $\text{Beta}(\alpha, \beta)$

$L(\theta)$: $y \sim \text{Bin}(n, \theta)$

$\Rightarrow$ $P(\theta | y)$: $\text{Beta}(\underbrace{y}_{\text{Actual makes (49)}} + \overbrace{\alpha}^{\text{"Imagined" "pseudo" made shots}}, \underbrace{n-y}_{\text{Actual misses (51)}} + \beta)$

"pseudo" misses.

Mean is

$\alpha / \alpha+\beta$

Imagine $\alpha = 35$
$\beta = 65$

$\Rightarrow \alpha / \alpha+\beta = .35$

Pseudo-Counts Interpretation

- Observe y successes, $n - y$ failures
- If $p(\theta) \sim \text{Beta}(\alpha, \beta)$ then $p(\theta \mid y) = \text{Beta}(\underline{y + \alpha}, n - y + \beta)$
- What is $E[\theta \mid y]$? ("posterior mean")

$$E[\theta|y] = \frac{y+\alpha}{y+\alpha+n-y+\beta} = \frac{y+\alpha}{n+\alpha+\beta}$$

$$= \left(\frac{n}{n}\right)\frac{y}{n+\alpha+\beta} + \frac{\alpha}{n+\alpha+\beta}\left(\frac{\alpha+\beta}{\alpha+\beta}\right)$$

$$= \frac{n}{n+\alpha+\beta}\frac{y}{n} + \frac{\alpha+\beta}{n+\alpha+\beta}\frac{\alpha}{\alpha+\beta}$$

$$= w \frac{y}{n} + (1-w)\frac{\alpha}{\alpha+\beta} \qquad w = \frac{n}{n+\alpha+\beta}$$

$$w\,\hat{\theta}_{MLE} + (1-w)\,\hat{\theta}_{\text{prior mean}}$$

.49

(maybe closer to .35)

$\alpha+\beta$ "prior attempts"