Lecture 2: One Parameter Models

Professor Alexander Franks

2023-01-19

Announcements

- Reading: Chapter 2 and 3, Bayes Rules
- Homework due: January 22 at midnight

Sunday

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
  $p(y, \theta) = \underbrace{p(y \mid \theta)}_{\text{Sampling Model / Likelihood}}\underbrace{p(\theta)}_{\text{Prior Distribution}}$

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
- 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

Sampling model / Likelihood

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

Prior

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

$$P(\Theta \mid y) \propto L(\Theta)P(\Theta)$$

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

- 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?"

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.

$P(\theta \mid y)$

Example: estimating the fraction of the Earth covered in water.

- Assume we sample the a point on the Earth and record whether it is land or water
- 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 $\frac{y}{n}$
- What would our estimates be if we use Bayesian inference?
    - What properties do we want for our prior distribution?

- Land anywhere (uniformly at random)

- Record water (1) or earth 0

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

# of times saw water. — # of times "sampling" — fraction of earth in water.

$$P(\theta) = \mathbb{1}[\theta \in [0,1]] \text{ (uniform)}$$

$$P(\theta \mid y) \propto \underbrace{\binom{n}{y}\theta^{y}(1-\theta)^{n-y}}_{L(\theta)} \times 1$$

6 waters
4 earths

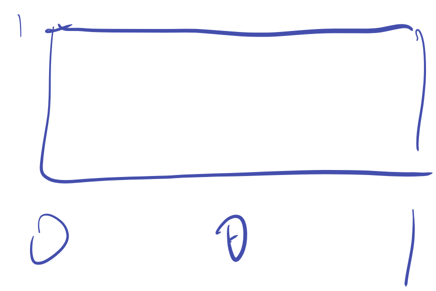
$$\propto \binom{10}{6}\theta^{6}(1-\theta)^{4}$$

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

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

$$P(\theta|y) \sim \underbrace{\theta^{y}(1-\theta)^{n-y}} \times \underbrace{\mathbb{1}[\theta\in[0,1]]}_{\text{Uniform}}$$

$$\rightarrow \text{Beta}(\overset{\alpha}{y+1}, \overset{\beta}{n-y+1})$$



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

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

$$p(\theta = a|y) \propto p(y|\theta = a)p(\theta = a) = 0$$

The Binomial Model

- The uniform prior: $p(\theta) = \text{Unif}(0,1) = \mathbf{1}\{\theta \in [0,1]\}$
  - 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.

$\text{Beta}(y+1, n-y+1)$

The Binomial Model

- The uniform prior: $p(\theta) = \text{Unif}(0,1) = \mathbf{1}\{\theta \in [0,1]\}$
    - 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.
- $p(\theta \mid y) \sim \text{Beta}(y+1, n-y+1) = \frac{\Gamma(n)}{\Gamma(n-y)\Gamma(y)} \theta^y (1-\theta)^{n-y}$

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?

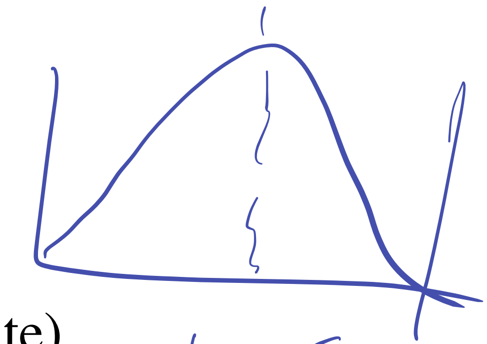
- *Point estimates:* posterior mean or mode:

  - $E[\theta \mid y] = \int_\Theta \theta p(\theta \mid y) d\theta$ (the posterior mean)

  - $\arg\max p(\theta \mid y)$ (*maximum a posteriori* estimate)

MAP

$E[\theta \mid y=5] = 1/2$



$y = 5$

Summarizing Posterior Results

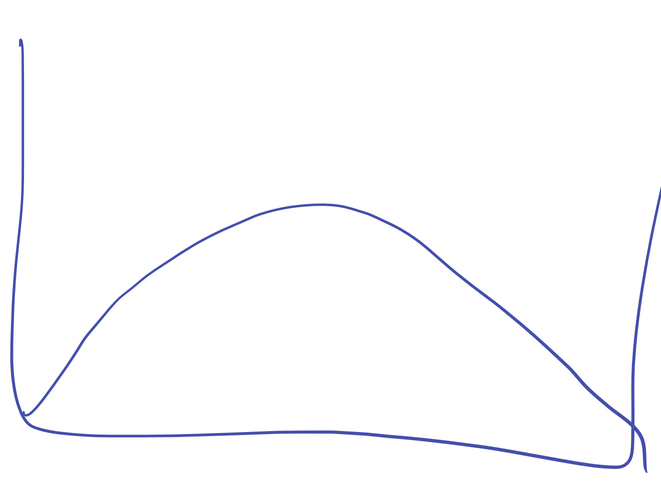
- An entire distribution describes our beliefs about the value for θ. How can we summarize these beliefs?

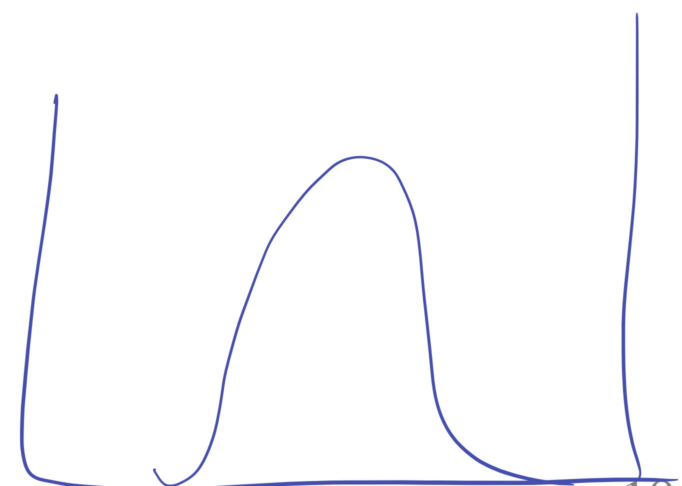
- *Point estimates:* posterior mean or mode:

    - $E[\theta \mid y] = \int_\Theta \theta p(\theta \mid y) d\theta$ (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?

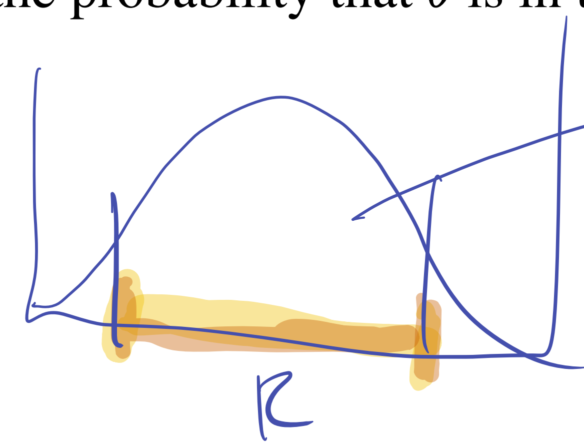
- *Point estimates:* posterior mean or mode:

  - $E[\theta \mid y] = \int_\Theta \theta p(\theta \mid y) d\theta$ (the posterior mean)

  - $\arg\max p(\theta \mid y)$ (*maximum a posteriori* estimate)

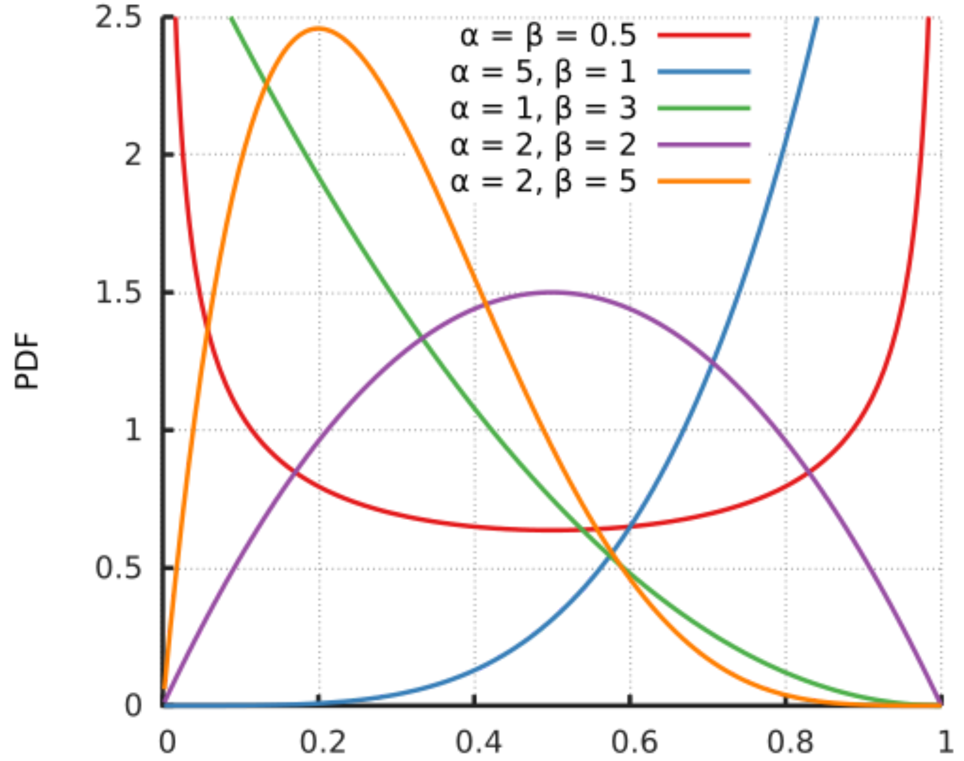
- Posterior variance: $\mathrm{Var}[\theta \mid y] = \int_\Theta (\theta - E[\theta \mid y])^2 p(\theta \mid y) d\theta$

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



95% prob. $\theta \in R$.

Beta Distributions



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

Summarizing Posterior Results

- $\text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- 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 $\frac{\alpha-1}{\alpha+\beta-2}$
- 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`

Example: estimating shooting skill in basketball

- 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
- 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

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

100

True shooting skill

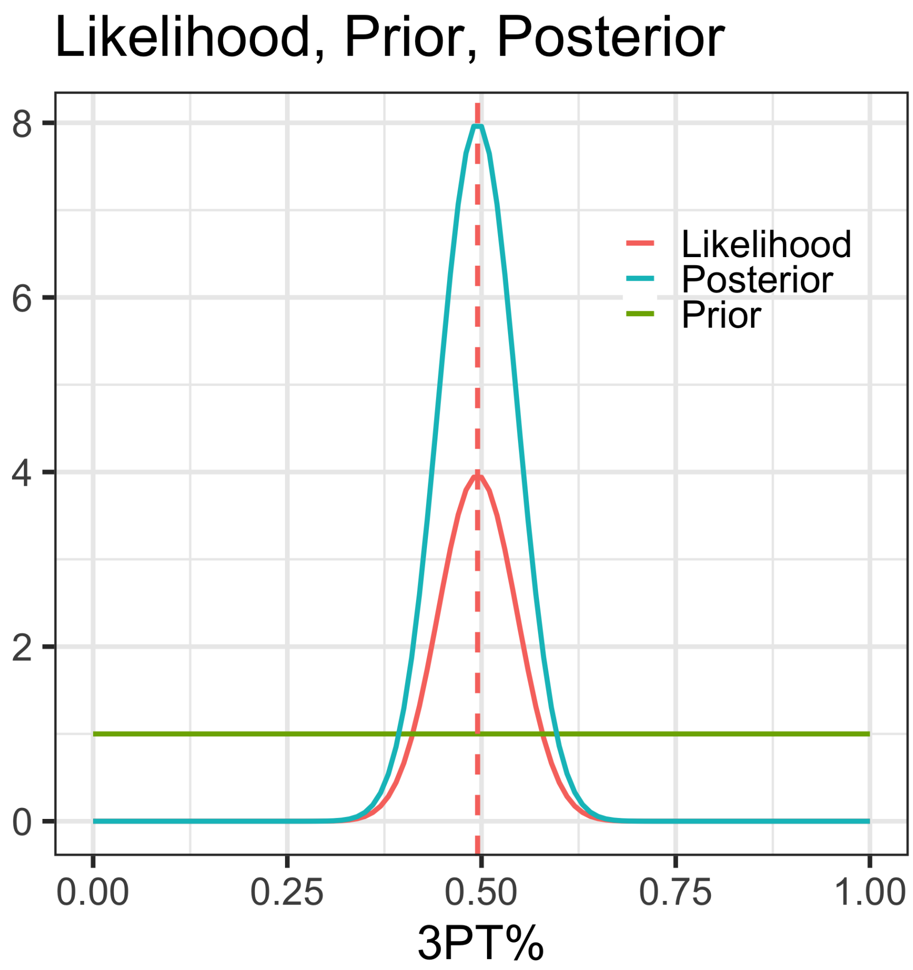
Example: estimating shooting skill in basketball

- 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 $\frac{y}{n} = 49/100 = 0.49$ MLE
- What would our estimates be if we use Bayesian inference?

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

$P(\theta|y) = \text{Beta}(50; 51)$

Example: estimating shooting skill in basketball



Posterior is proportional to the likelihood

Example: estimating shooting skill in basketball

- 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?
  - If our prior reflects "complete ignorance" about basketball?
  - What if we want to incorporate prior domain knowledge?

- Past data from Robert Covington
- Look at other players.

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:

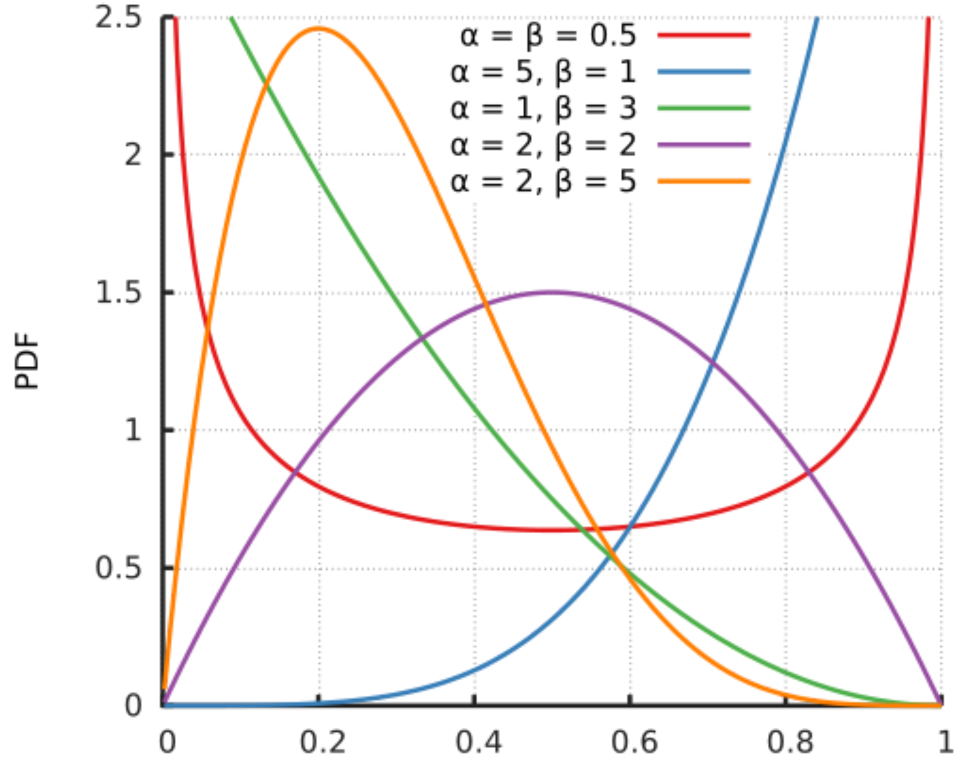
  - $E[\theta \mid y] = \int_\Theta \theta p(\theta \mid y) d\theta$ (the posterior mean)

  - $\arg\max p(\theta \mid y)$ (*maximum a posteriori* estimate)

- 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: $p(\theta \in R(y))$

Beta Distributions



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

Summarizing Posterior Results

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

- 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 $\frac{\alpha-1}{\alpha+\beta-2}$

- 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`

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 $p(\theta) \sim U(0,1)$ 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: $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$\text{Beta}(1,1) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \propto ?$$

$$p(\theta) \sim \text{Beta}(\alpha, \beta)$$

$$p(\theta \mid y) \propto \theta^{y}(1-\theta)^{n-y}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

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

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

α: prior or "pseudo"-made shots

β: prior or pseudo missed shots.

$\alpha+\beta$: prior sample size.

Mean of $\text{Beta}(\alpha, \beta) = \dfrac{\alpha}{\alpha+\beta}$

prior "guess" at fraction

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

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

Conjugate Prior Distributions

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

- The prior distribution and the posterior distribution are in the same family

Conjugate Prior Distributions

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

- The prior distribution and the posterior distribution are in the same family
- Conjugate priors are very convenient because they make calculations easy

Conjugate Prior Distributions

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

- 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

Conjugate Prior Distributions

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

- 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.

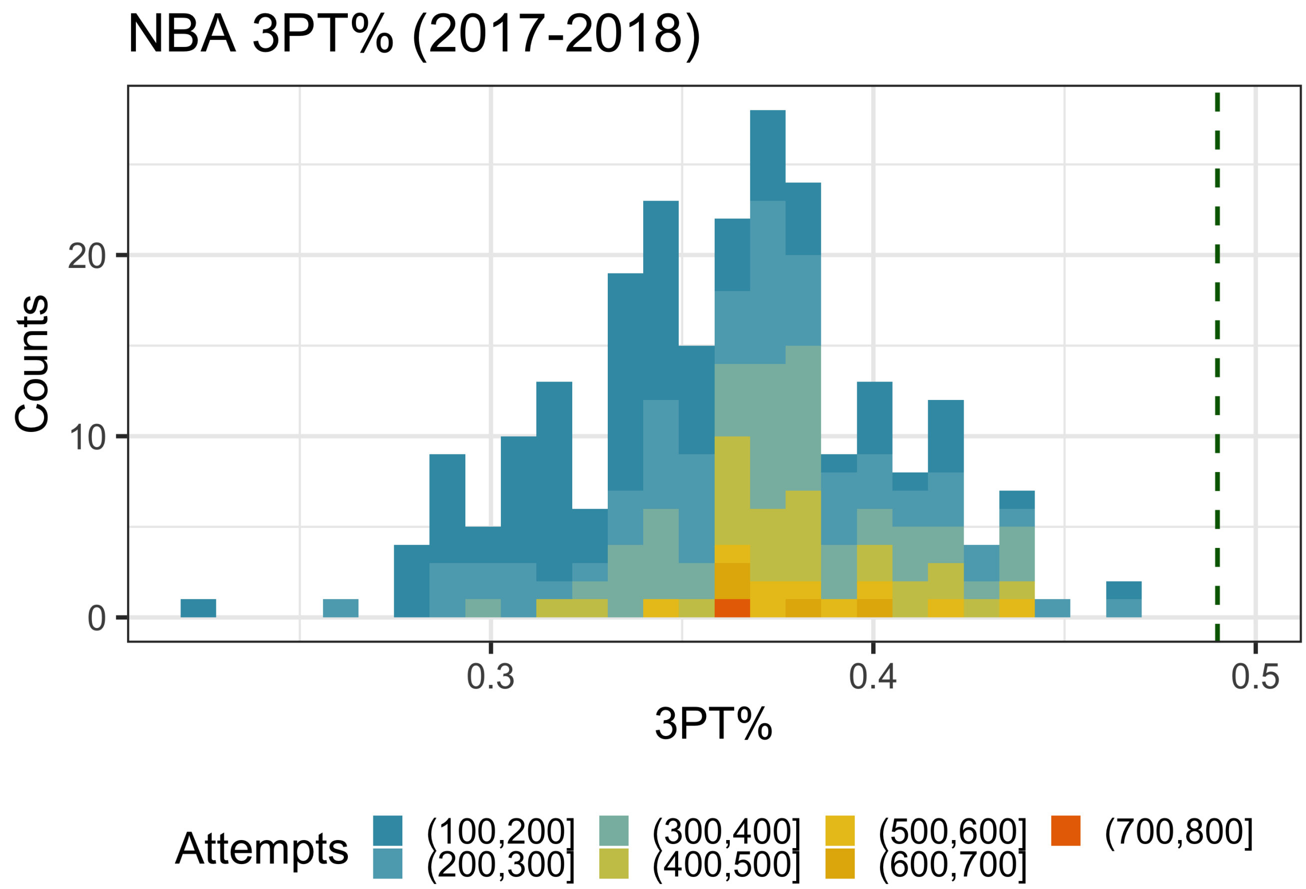
Pseudo-Counts Interpretation

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

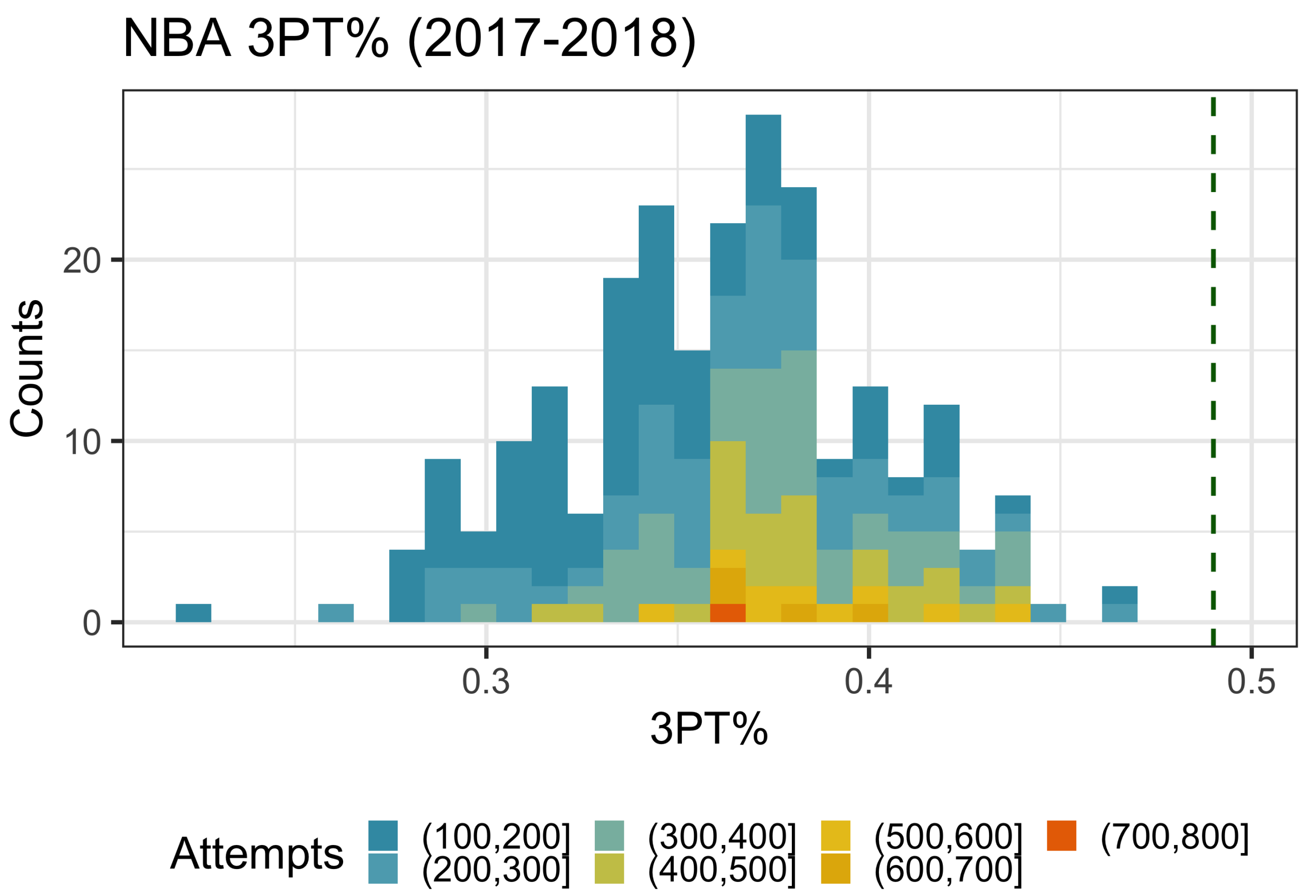
Example: estimating shooting skill in basketball

- 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 10 ten all time
- Prior knowledge tells us it is unlikely this will continue!
- How can we use Bayesian inference to better estimate his true skill?

Three point shooting in 2017-2018



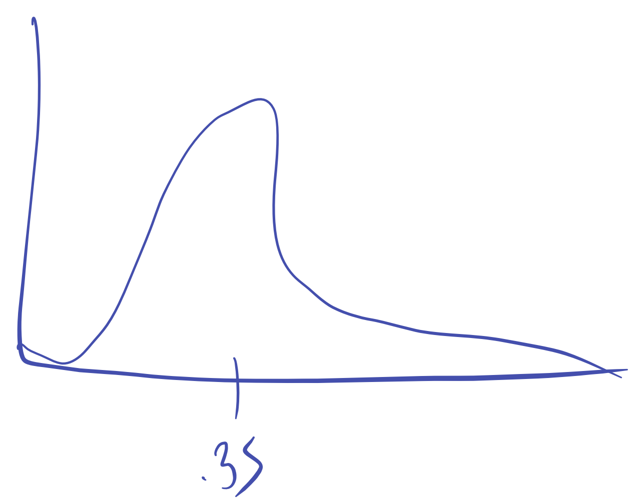
Three point shooting in 2017-2018



Regression Toward the Mean

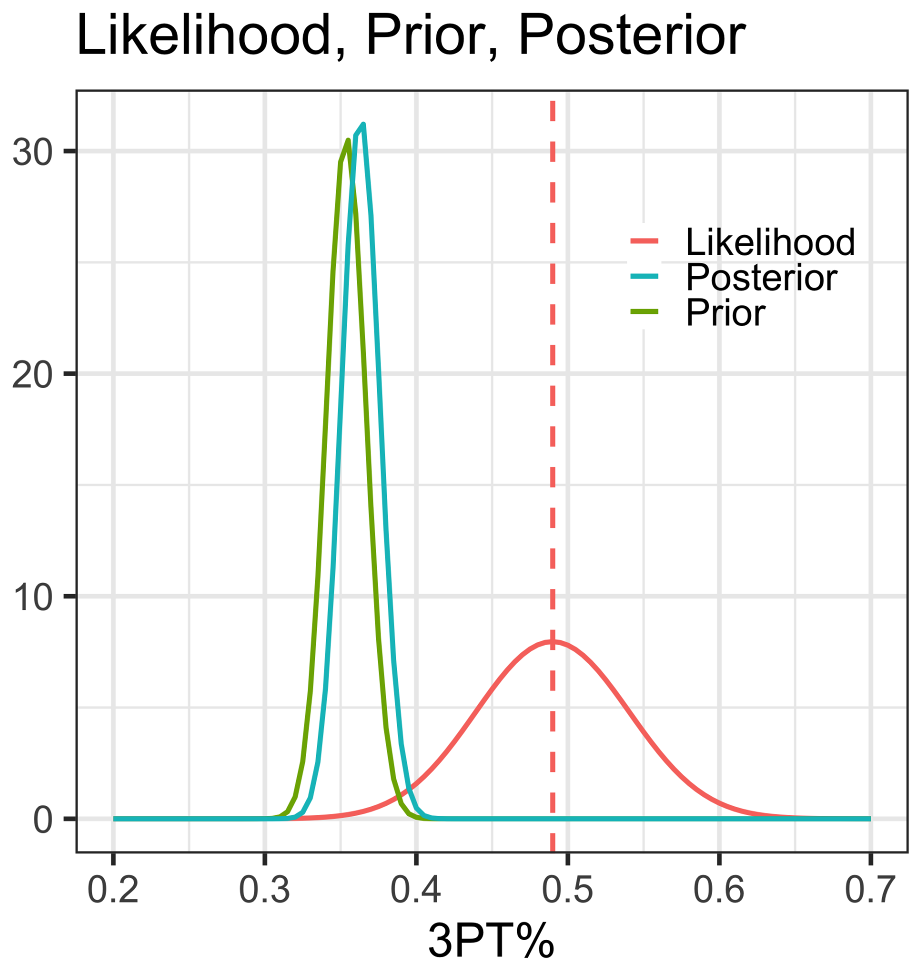
What is a reasonable model?

- If we believe that his skill doesn't change much year to year, use past data to inform prior
- In his first 4 seasons combined Robert Covington made a total of 478 out of 1351 three point shots (0.35%, just below average).
- Choose a Beta(478, 873) prior (pseudo-count interpretation)



Robert Covington 2017-2018 estimates

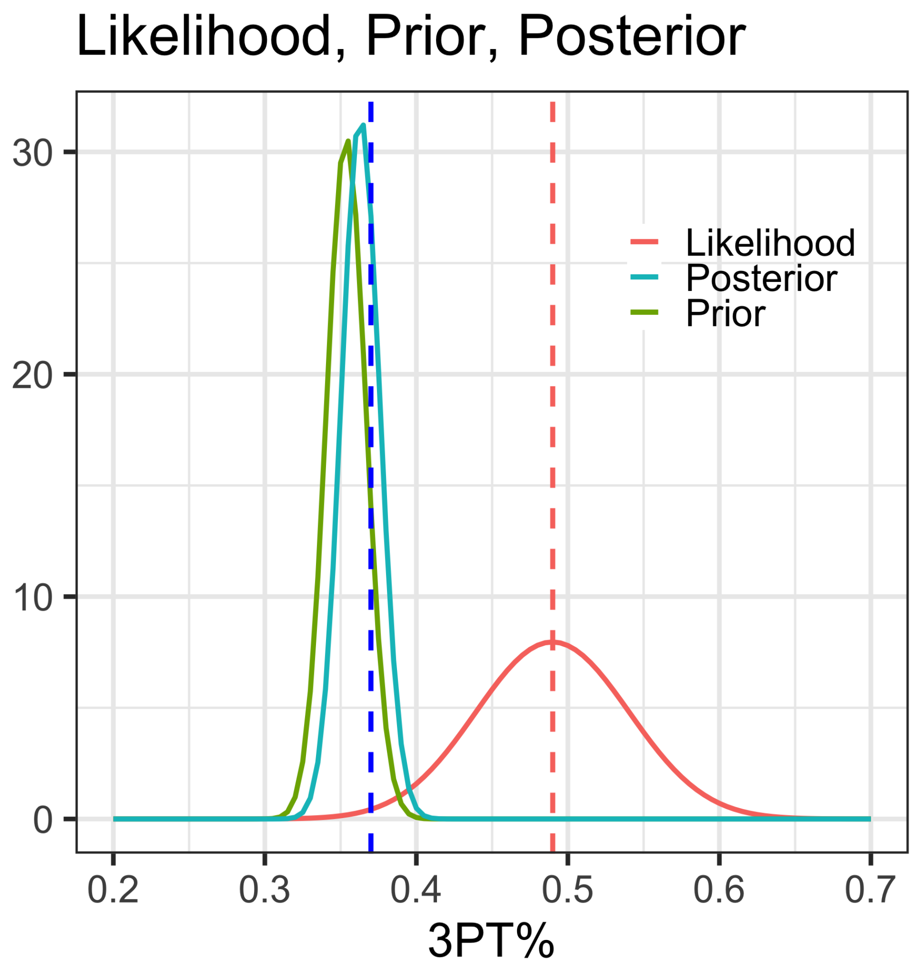
After 100 shots Robert Covington's 3PT% was 0.49



MLE = 0.49, posterior mean = 0.36

How did we do?

Robert Covington's end of season 3PT% was 0.37



MLE = 0.49, posterior mean = 0.36