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

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 $p(y, \theta) = p(y \mid \theta)p(\theta)$

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

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

- $P(\theta \mid y)$ 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

Computing the Posterior Distribution

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

- 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.

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

- 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?

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:
$$p(\theta) = \mathrm{Unif}(0, 1) = \mathbf{1}\{\theta \in [0, 1]\}$$
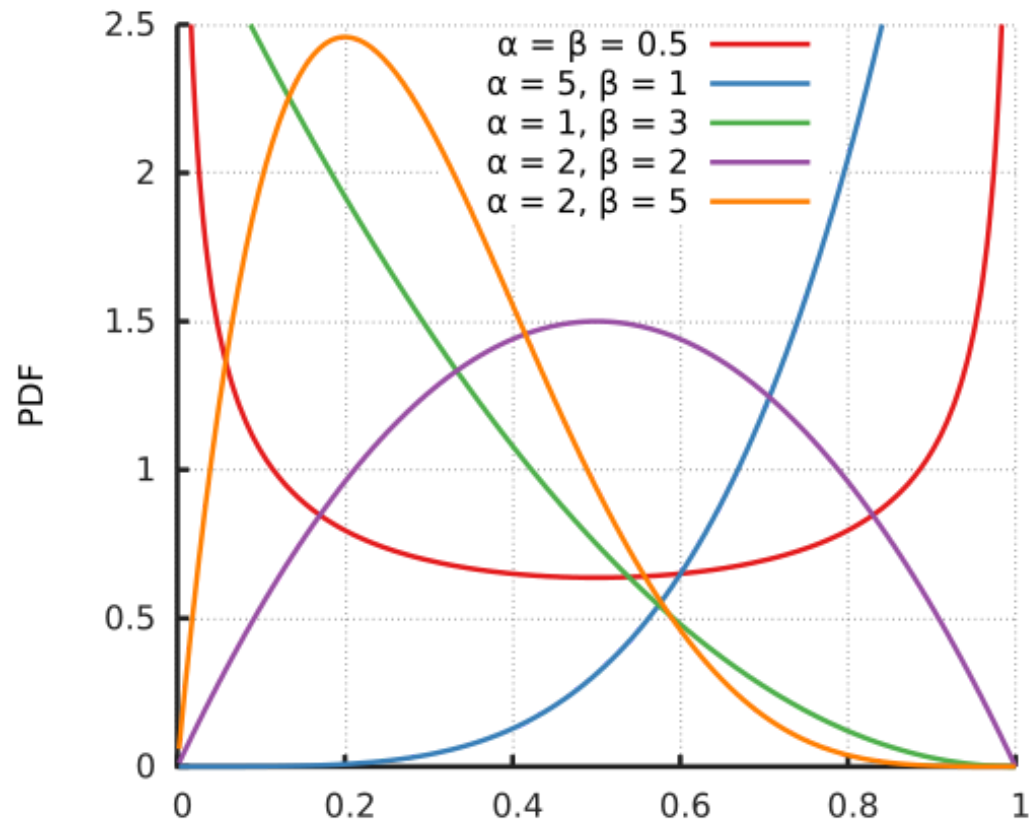
  - A "non-informative" prior

- 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.

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.

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

Beta Distributions

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

Sequential Bayesian Updating

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

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

$\underbrace{}_{\text{prior}}$

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

Demo

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

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

- 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

Example

- Assume every shot is independent (reasonable) and identically distributed (less reasonble?)

- Let $Y \sim \mathrm{Bin}(n, \theta)$ 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?

Example

Likelihood, Prior, Posterior
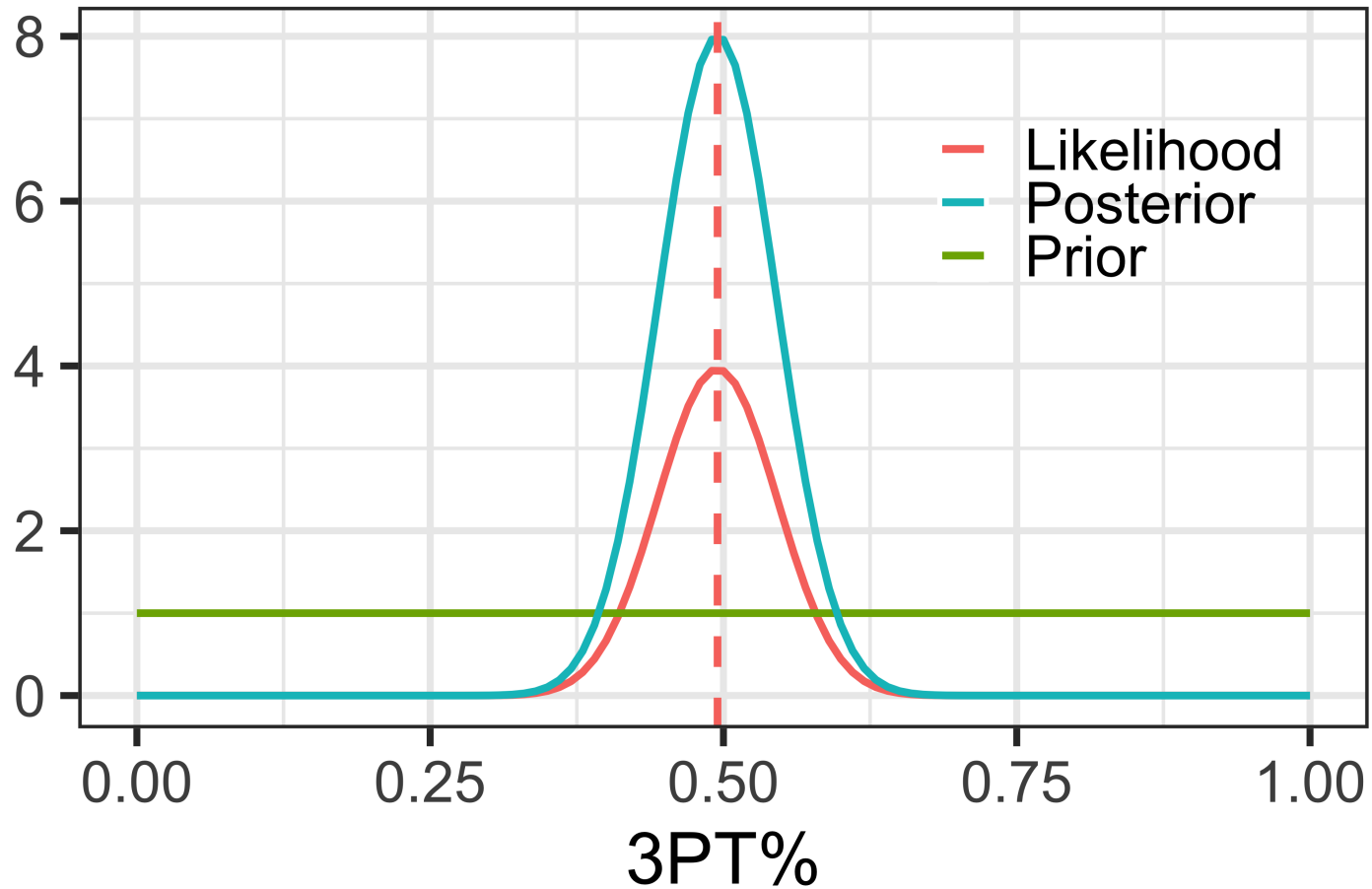
Posterior is proportional to the likelihood

Example

- Assume every shot is independent (reasonable) and identically distributed (less reasonble?)

- Let $Y \sim \mathrm{Bin}(n, \theta)$ 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?

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: $p(\theta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1}$

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

- 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 $p(\theta) \in \mathcal{P} \implies p(\theta|y) \in \mathcal{P}$

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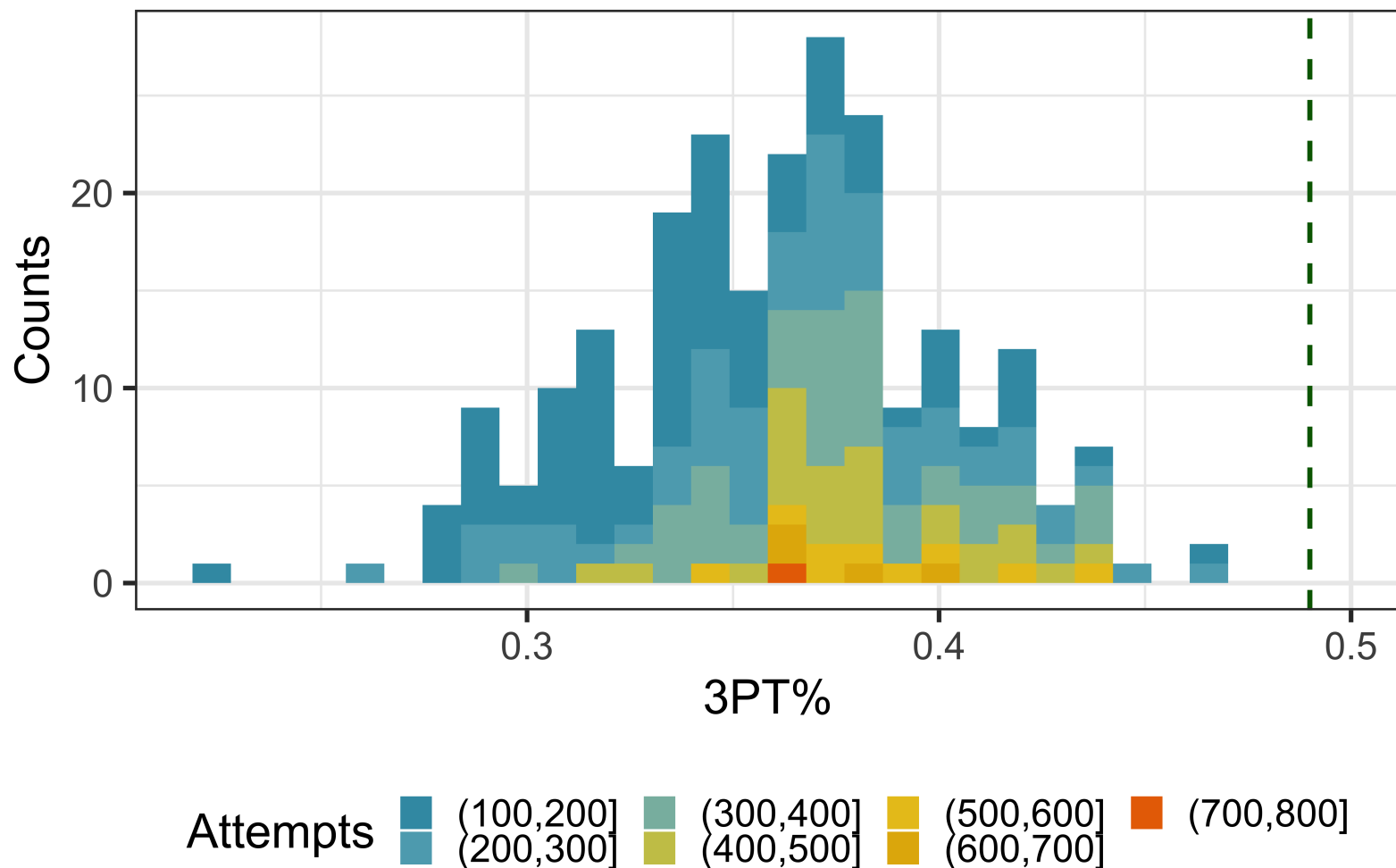
- What is $E[\theta \mid y]$?

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 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

NBA 3PT% (2017-2018)

Three point shooting in 2017-2018

NBA 3PT% (2017-2018)
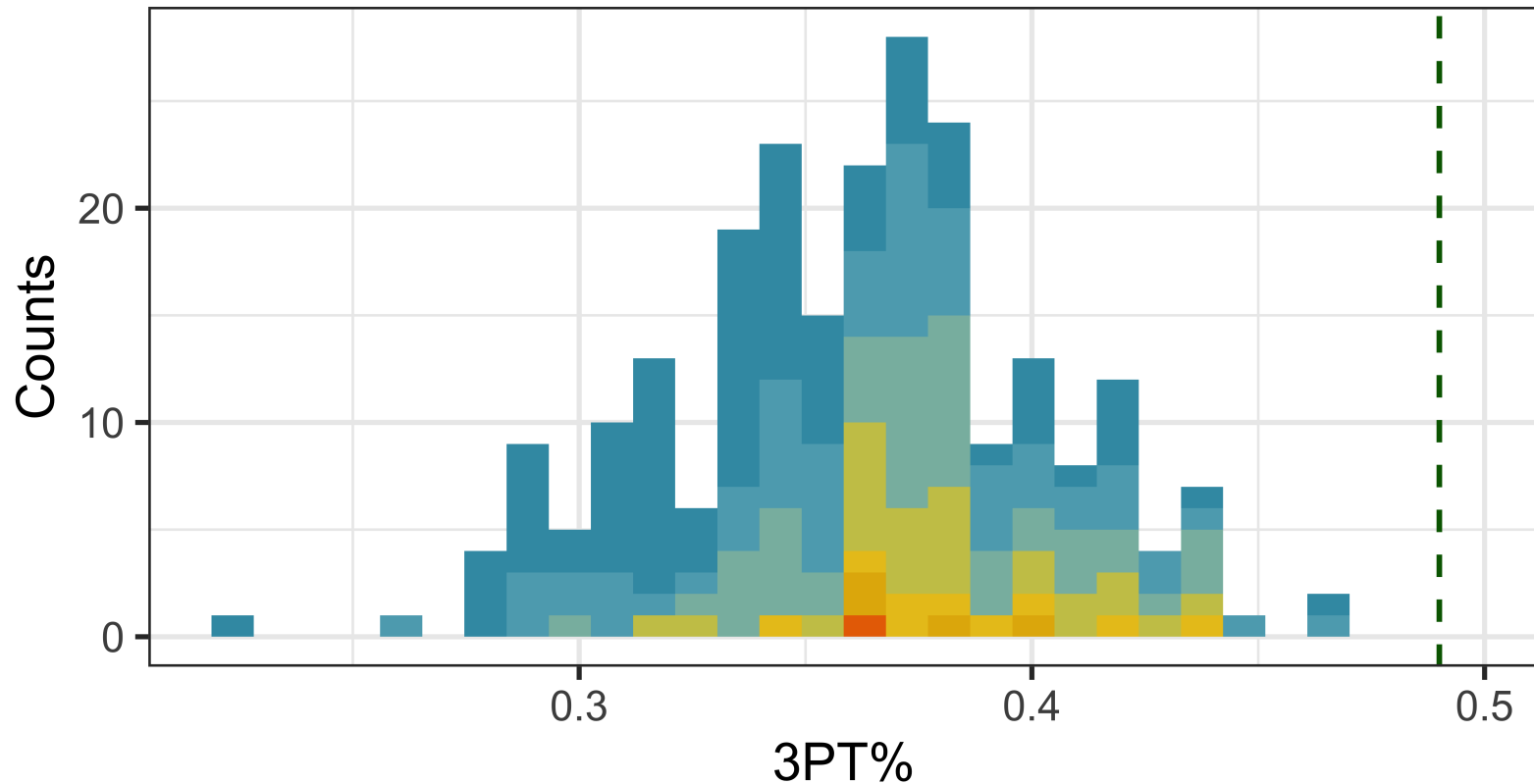
Regression Toward the Mean

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

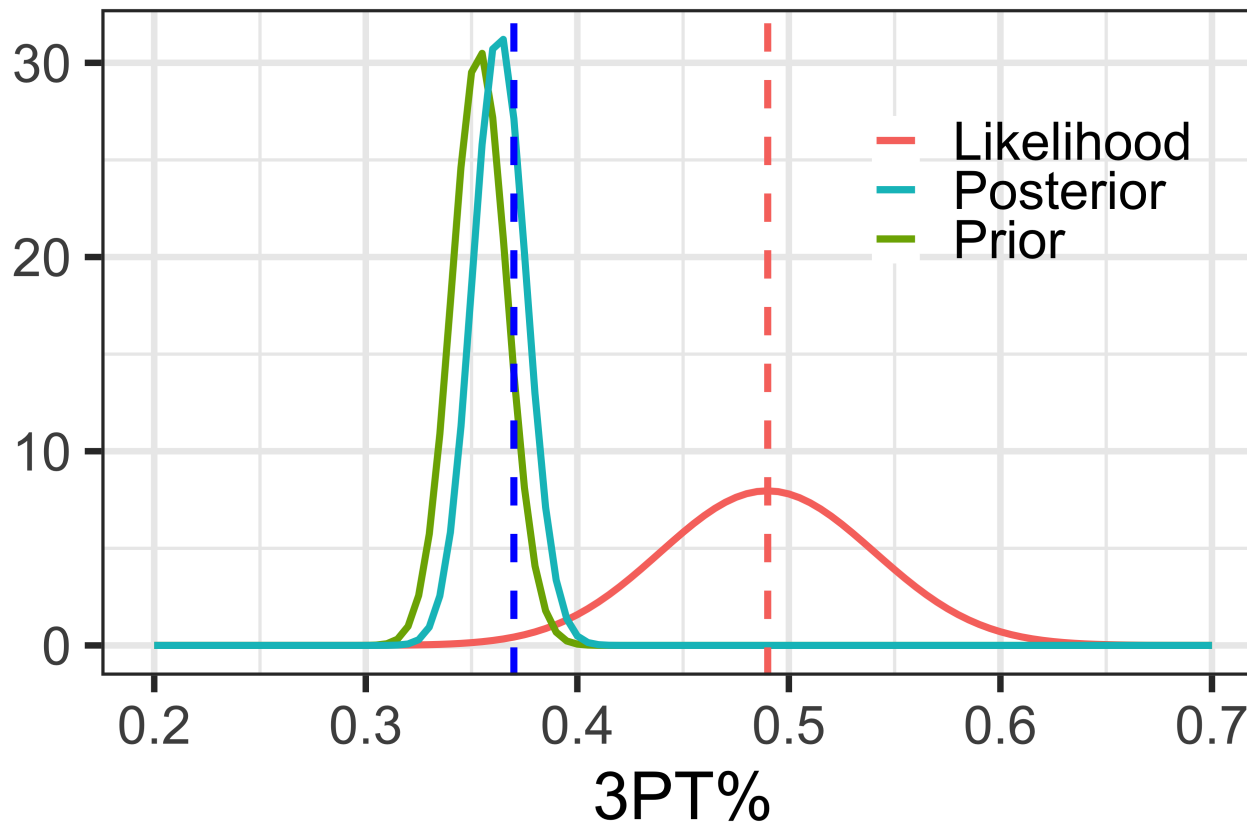

Likelihood, Prior, Posterior
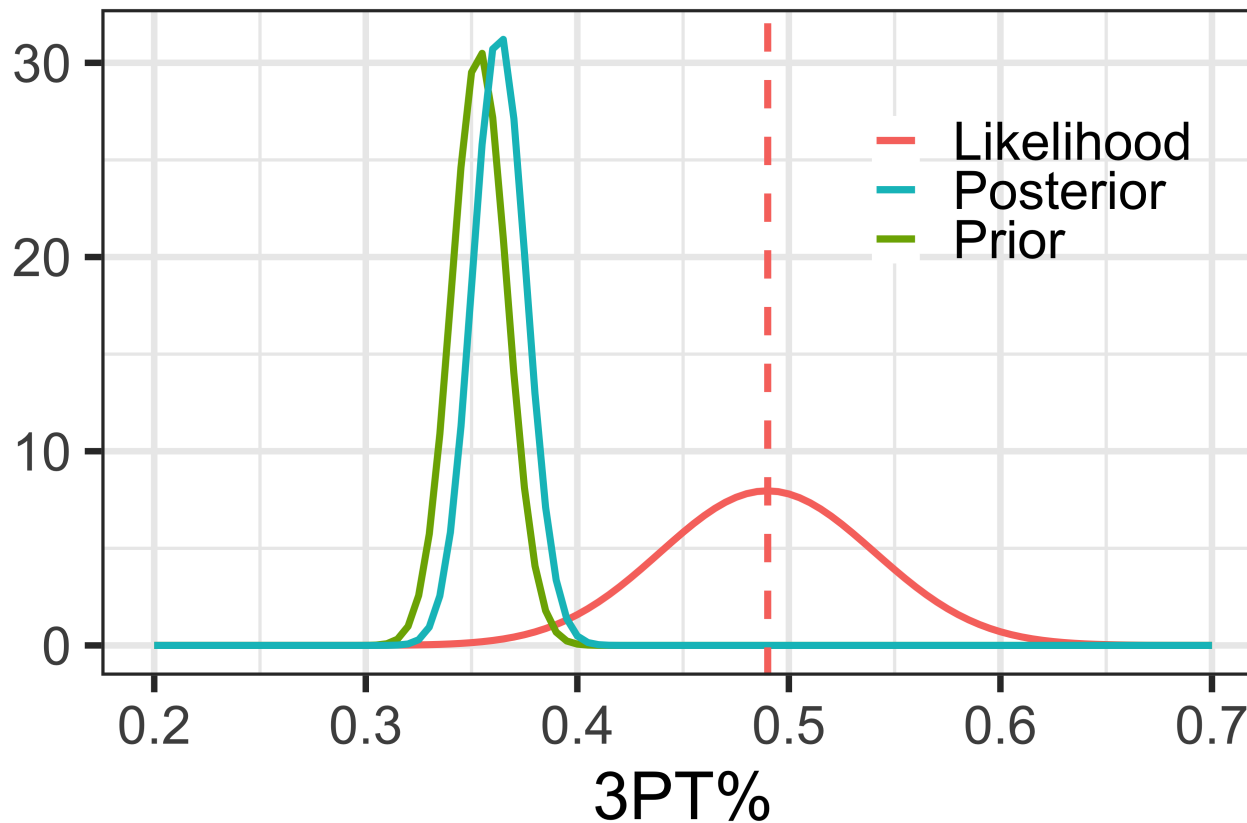
MLE = 0.49, posterior mean = 0.36

How did we do?

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

Likelihood, Prior, Posterior

MLE = 0.49, posterior mean = 0.36

The Poisson Distribution

Poisson model

Assume $Y_1, \ldots, Y_n$ are n i.i.d. $\mathrm{Pois}(\lambda)$

Poisson model with exposure

- Often times we include an "exposure" term in the Poisson model:

$$p(y_i \mid \nu_i \lambda) = (\nu_i \lambda)^y e^{\nu_i \lambda} / y_i!$$

- How many cars do we expect to pass an intersection in one hour? How many in two hours?
  - If we model the distribution as Poisson, we expect twice as many in two hours as in one hours.

- Homework: exposure is the length of the chapter

Poisson model example

- In a particular county 3 people out of a population of 100,000 died of asthma

- Assume a Poisson sampling model with rate λ (units are rate of deaths per 100,000 people)

- How do we specify a prior distribution for λ?

- How would our Bayesian estimate for λ differ?

Conjugate Prior for the Poisson

Assume n i.i.d observations of a $\mathrm{Poisson}(\lambda)$

$$p(\lambda \mid y_1, \ldots y_n) \propto L(\lambda) \times p(\lambda)$$
$$\propto \lambda^{\sum y_i} e^{-n\lambda} \times p(\lambda)$$

- A prior distribution for λ should have support on $\mathbb{R}^+$, the positive real line
- Bayesian definition of sufficiency: $p(\lambda \mid s, y_1, \ldots y_n) = p(\lambda \mid s)$
  - For the Poisson, $\sum y_i$ is sufficient
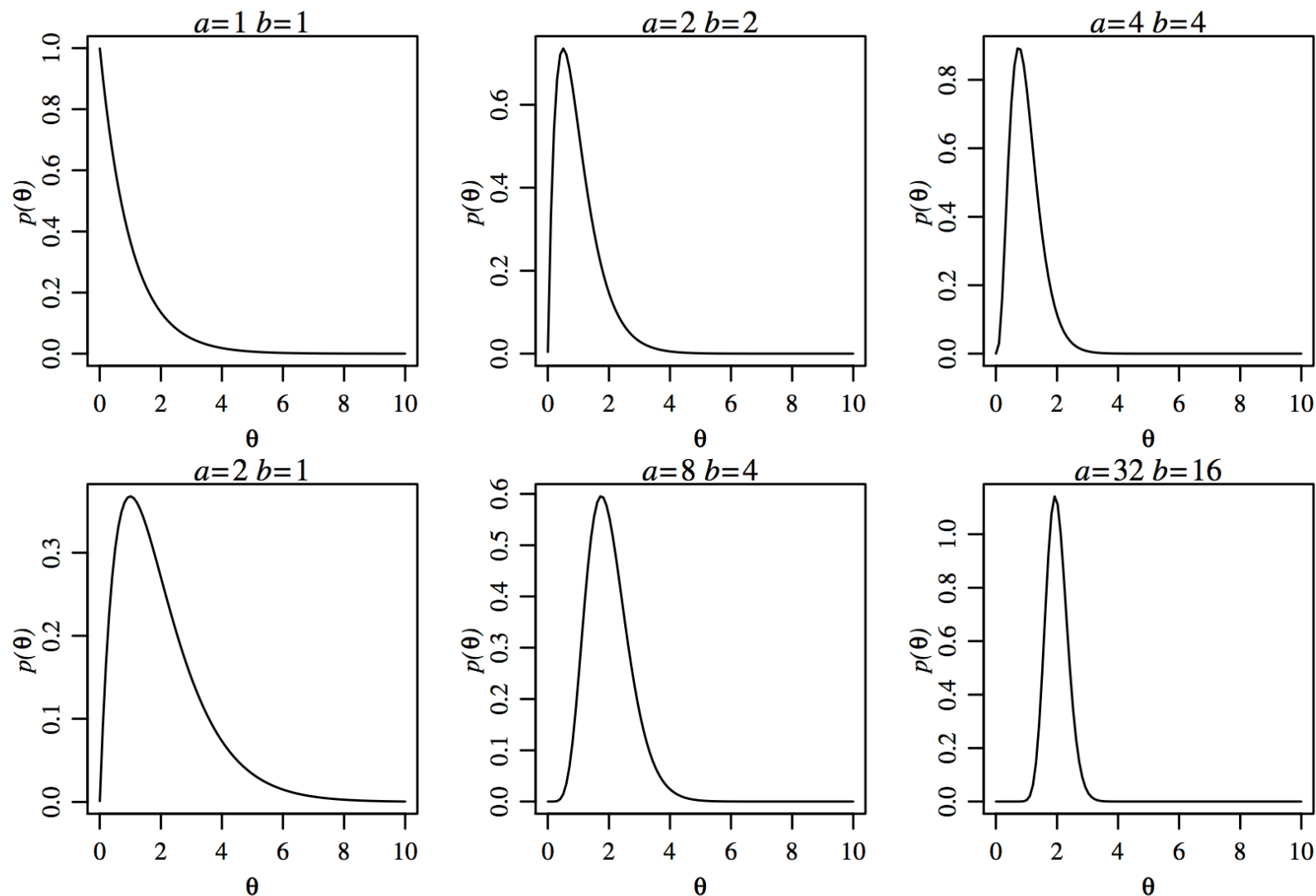- Can we find a density of the form $p(\lambda) \propto \lambda^{k_1} e^{k_2 \lambda}$?

Conjugate Prior for the Poisson

Assume n i.i.d observations of a $\mathrm{Poisson}(\lambda)$

$$p(\lambda \mid y_1, \ldots y_n) \propto L(\lambda) \times p(\lambda)$$
$$\propto \lambda^{\sum y_i} e^{-n\lambda} \times p(\lambda)$$

- A prior distribution for λ should have support on $\mathbb{R}^+$, the positive real line
- Bayesian definition of sufficiency: $p(\lambda \mid s, y_1, \ldots y_n) = p(\lambda \mid s)$
  - For the Poisson, $\sum y_i$ is sufficient
- Can we find a density of the form $p(\lambda) \propto \lambda^{k_1} e^{k_2 \lambda}$?

- $\mathrm{Gamma}(a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$

The Gamma distribution

Fig. 3.8. Gamma densities.

The Gamma distribution

Useful properties of the Gamma distribution:

- $E[\lambda] = a/b$

- $\mathrm{Var}[\lambda] = a/b^2$

- $\mathrm{mode}[\lambda] = (a-1)/b$ if $a > 1$, 0 otherwise

- In R: dgamma, rgamma, pgamma, qgamma

The posterior in the Poisson-Gamma model

Assume one observation with $y_i \sim \mathrm{Pois}(\lambda\nu_i)$ where ν_i is the exposure

$$p(\lambda \mid y_i) \propto L(\lambda) \times p(\lambda)$$

$$\propto (\lambda\nu_i)^{y_i}e^{-\lambda\nu_i} \times \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}$$

$$\propto (\lambda)^{y_i+a-1}e^{-(b+\nu_i)\lambda}$$

$$p(\lambda \mid y, a, b) = \mathrm{Gamma}(y_i + a, b + \nu_i)$$

Poisson model example

- In a particular county 3 people out of a population of 100,000 died of asthma

- Assume a Poisson sampling model with rate λ

  - Units are rate of deaths per 100,000 people/year

- Experts know that typical rates of asthma mortality in the US are closer to 0.6 per 100,000

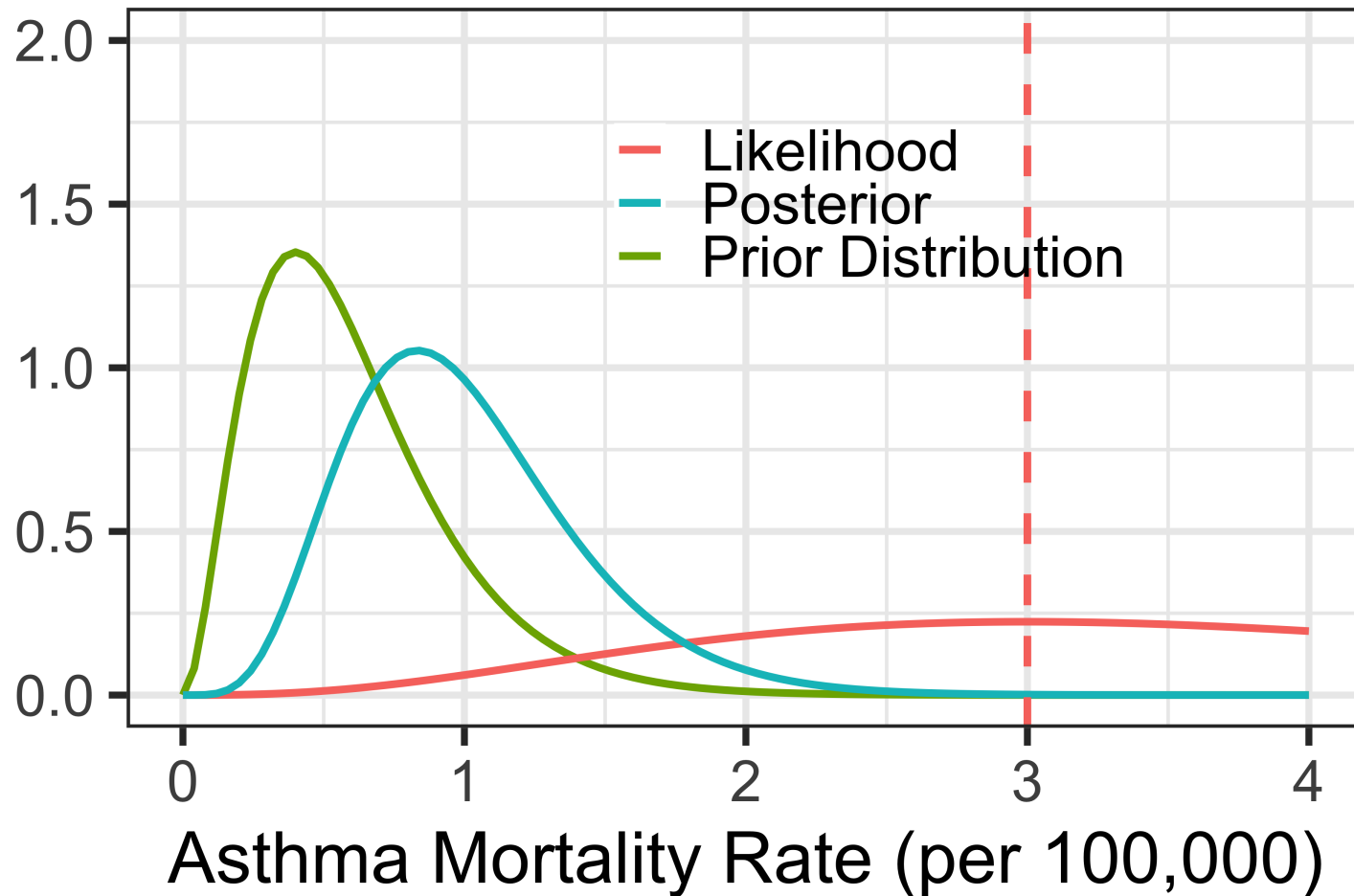- Let's choose a Gamma distribution with a mean of 0.6 and appropriate uncertainty.

Possible Gamma prior distributions
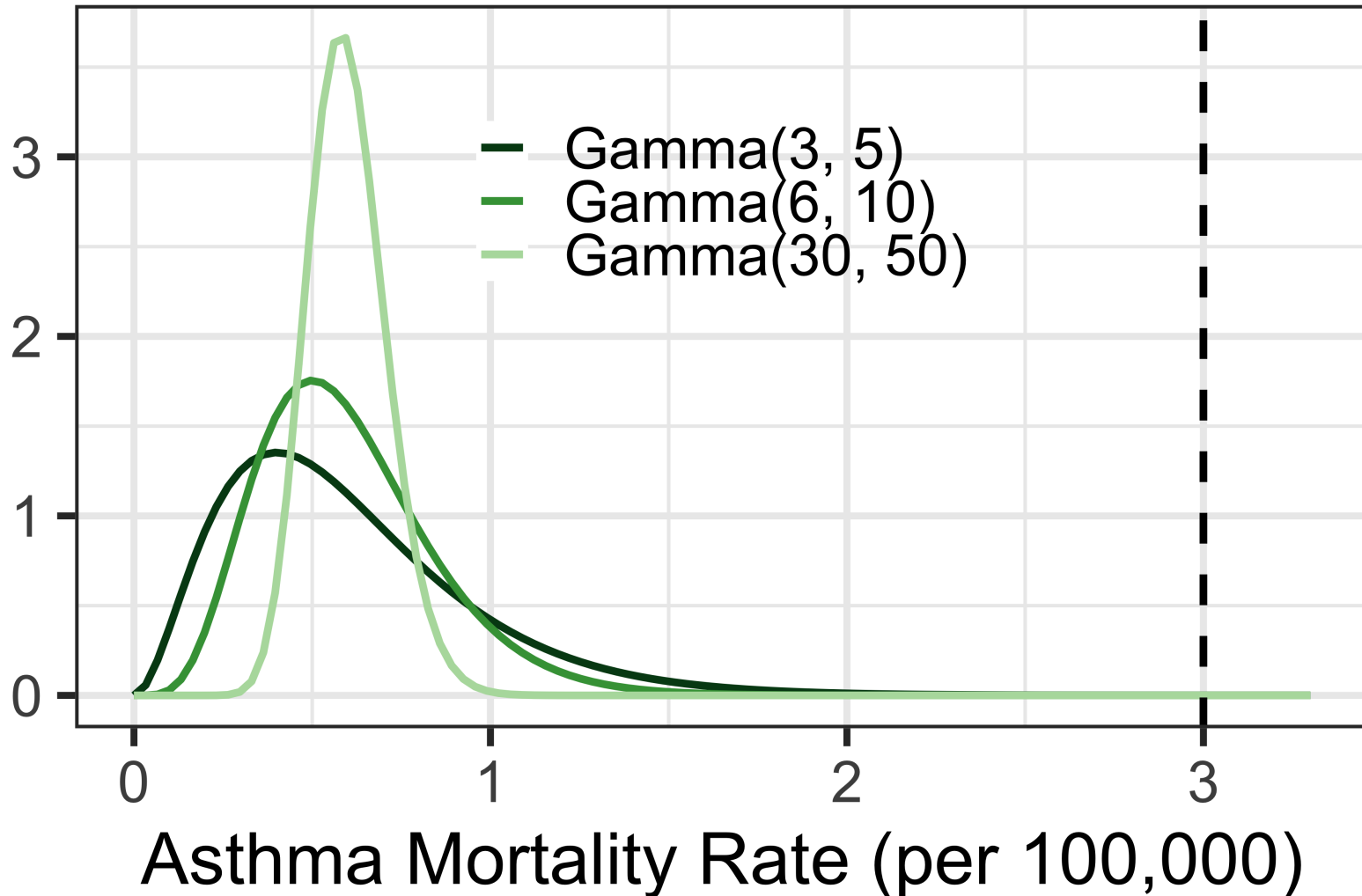
Asthma Mortality

Likelihood, Prior and Posterior

Using Gamma(3, 5) prior distribution

The posterior mean

$$E[\lambda \mid y_1, \dots y_n] = \frac{a + \sum y_i}{b + n}$$

$$= \frac{b}{b + n} \frac{a}{b} + \frac{n}{b + n} \frac{\sum y_i}{n}$$

$$= (1 - w)\frac{a}{b} + w\hat{\lambda}_{\mathrm{MLE}}$$

- $w \to 1$ as $n \to \infty$ (data dominates prior)
- b can be interpreted as the number of *prior* observations
  - Analogous to n or total prior exposure
- a can be interpreted as the sum of the counts from prior total exposure of b
  - Analogous to $\sum_i y_i$

Summary

- The Beta distribution
  - Conjugate prior for Binomial likelihood
- The Gamma distribution
  - Conjugate prior for the Poisson likelihood
- Pseudo-counts interpretations of conjugate prior distributions