Lecture 2: One Parameter Models

Professor Alexander Franks

Announcements

- Reading: Chapter 2 (these lecture notes)
- Chapter3 (next lecture notes)
- HW1 due the 28th.
- OH: Fri 2-3 on Zoom.

$$Y \sim N(\mu, \sigma^2) \rightarrow Y = \mu + \sigma Z$$

$$Z \sim N(0,1)$$

$$\sigma \sim Exp(1/2)$$

$$\frac{Z_1}{Z_2}, \qquad Z_i \sim N(0,1)$$

Cauchy

$$\sigma = \frac{1}{|Z_2|}, \quad Z_1$$

$$Y = \sigma Z_1$$

Example: estimating spam accounts on twitter

θ: Frac of bots

$P(Y \mid \theta)$

$X_t \sim Bin(100, \theta_t)$

$P(\theta_t)$

$Unif(0,1)$

David Sacks
@DavidSacks

Twitter is toast.

4. On April 28, just three days after signing the Agreement, Twitter restated three years of its mDAU numbers, despite never disclosing the issue to Defendants pre-signing. Post-signing, Defendants promptly sought to understand Twitter's process for identifying false or spam accounts. In a May 6 meeting with Twitter executives, Musk was flabbergasted to learn just how meager Twitter's process was. Human reviewers randomly sampled 100 accounts per day (less than 0.00005% of daily users) and applied unidentified standards to somehow conclude every quarter for nearly three years that fewer than 5% of Twitter users were false or spam. That's it. No automation, no AI, no machine learning.

Elon Musk

9:02 AM · Jul 16, 2022 · Twitter for iPhone

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.

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

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

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

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

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

Example: estimating shooting skill in basketball

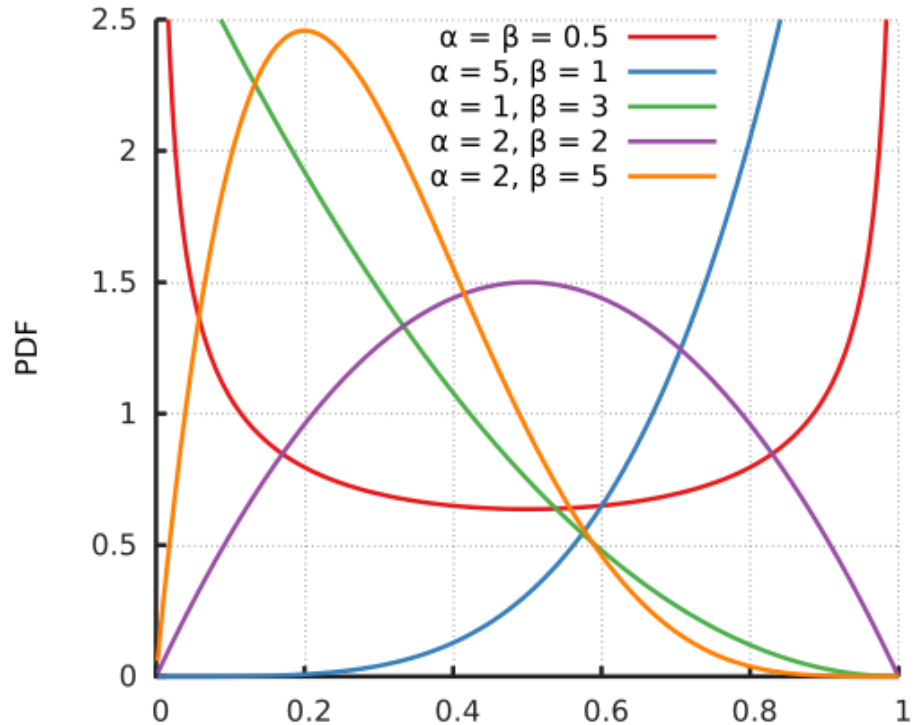
- Assume every shot is independent (reasonable) and identically distributed (less reasonble?)
- Let $Y \sim \mathrm{Bin}(n, \theta)$ where θ corresponds to his true skill
- Frequentist inference tells us that the maximum likelihood estimate is simply $\frac{y}{n} = 49/100 = 0.49$
- What would our estimates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

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.

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

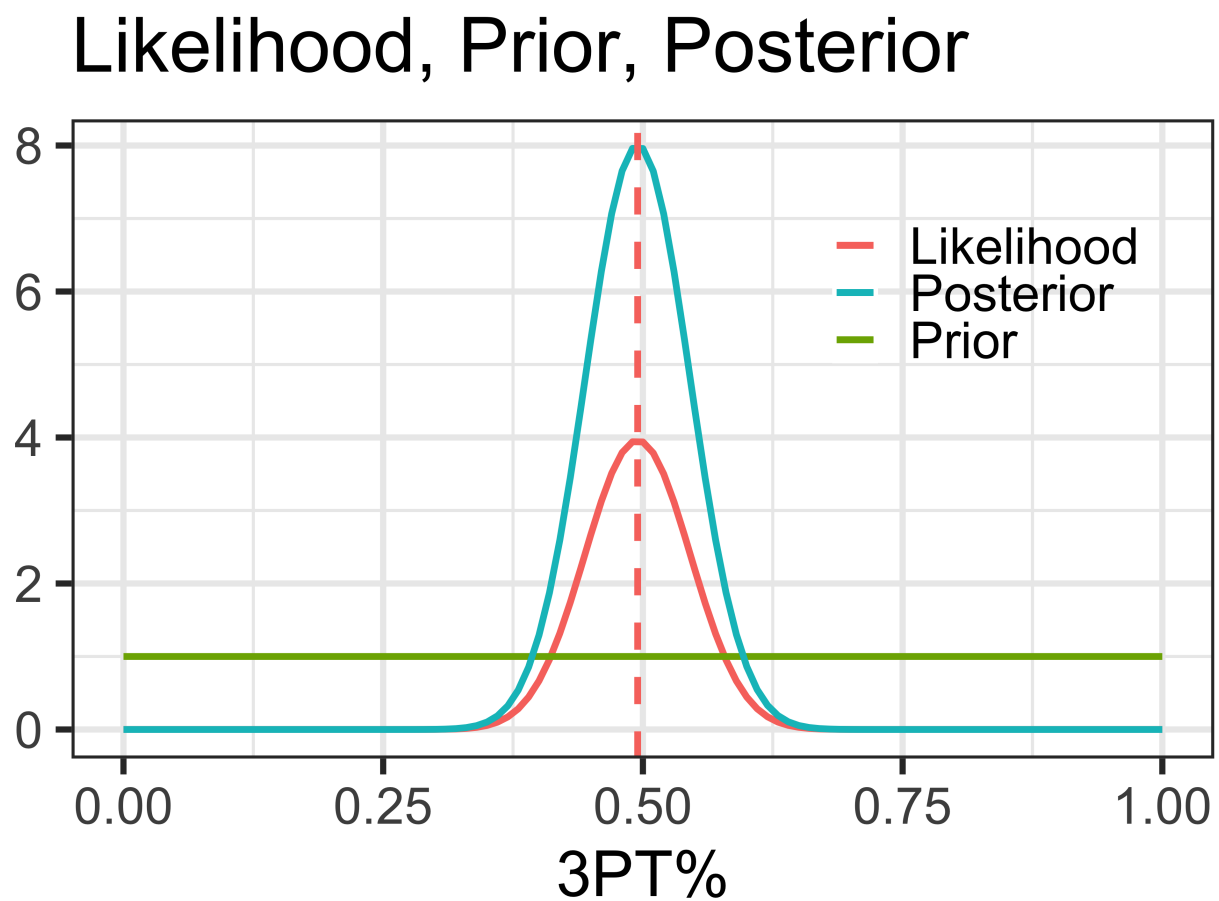
Beta Distributions



$\alpha, \beta > 0$

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

Example: estimating shooting skill in basketball



Summarizing Posterior Results

$$E[\theta|y] = \int_{\Theta} \theta \, p(\theta|y) \, d\theta \qquad \text{p.m.}$$

$$\arg\max_{\theta} \; p(\theta|y) \qquad \text{MAP}$$

$$Var(\theta|y) = \int (\theta - E[\theta|y])^2 \, p(\theta|y) \, d\theta$$

"Credible Intervals", uncertainty Quantif.

Summarizing Posterior Results

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

$Y \sim \text{Gam}(\alpha, 1)$

$Z \sim \text{Gam}(\beta, 1)$

$\frac{Y}{Y+Z} \sim \text{Beta}(\alpha, \beta)$

The Bayesian Advantage

- Particularly effective with small samples
  - With a true shooting skill of 0.35, there is about a 1/4 chance of making 5 or more out of the first 10 shots
- Ability to incorporate real prior knowledge
- Can easily "share information" across related observations

$Y|X \sim Bin(\ \ , \theta(x))$

Informative prior distributions

- At that point in November, Covington's 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

+ Similar players / Average Players

+ R.C. past performance.

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

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

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

"pseudo-counts"

$$\frac{\alpha}{\alpha+\beta} \qquad \mathrm{Beta}(y+\alpha, n-y+\beta)$$

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

Pseudo-Counts Interpretation

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

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

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

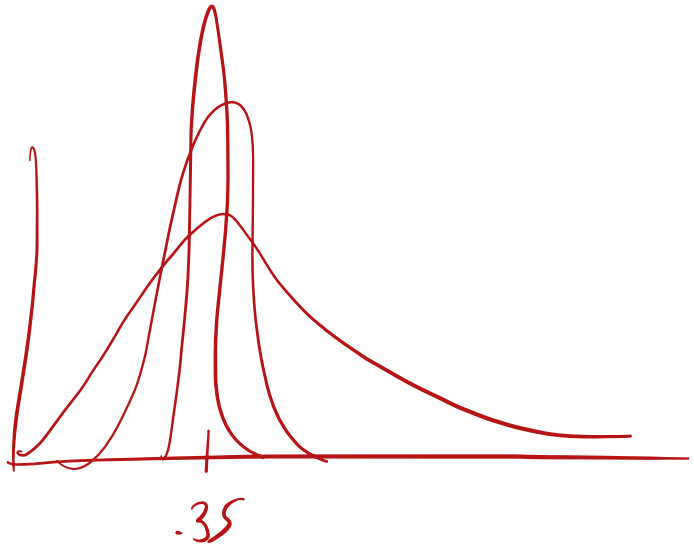
$$= \frac{n}{n + \alpha + \beta}\,\frac{y}{n} + \frac{\alpha + \beta}{n + \alpha + \beta}\,\frac{\alpha}{\alpha + \beta}$$

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

$$w\,\underbrace{\bar{y}}_{= \hat{\theta}_{MLE}} + (1 - w)\,E[\theta]$$

$$\frac{\alpha}{\alpha+\beta} \doteq .35$$

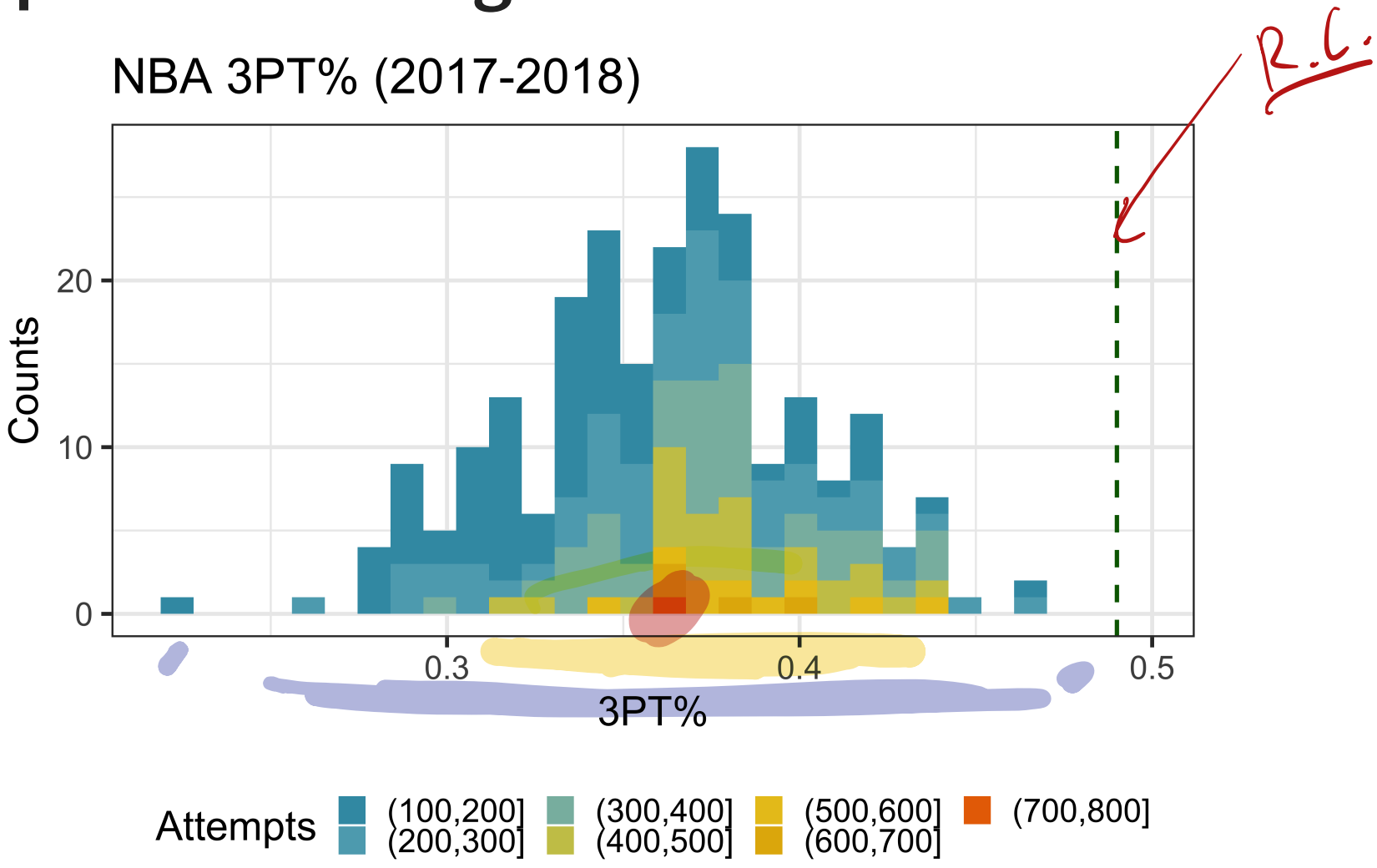
Calibrate $\alpha+\beta$



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



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)

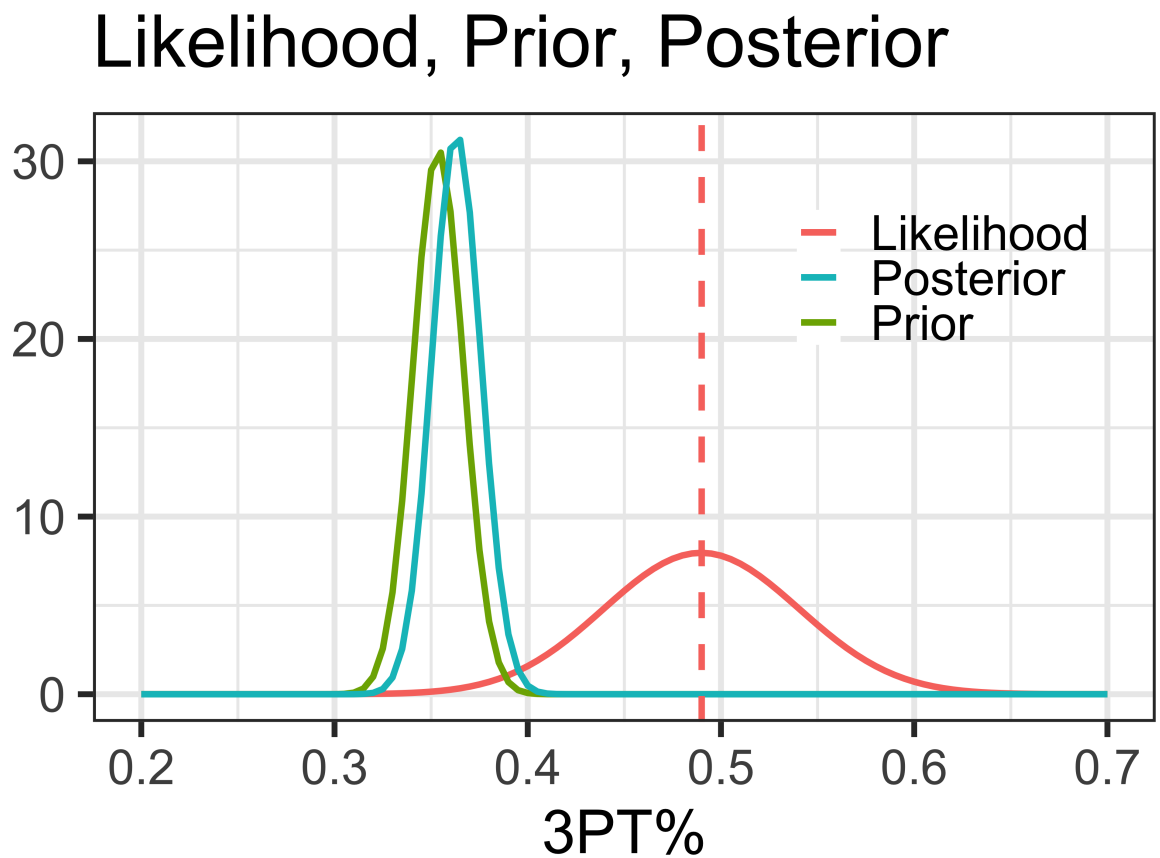
$\frac{\alpha}{\alpha+\beta} = .35$

$\alpha + \beta < 1300$

2013 . . . 2017

Robert Covington 2017-2018 estimates

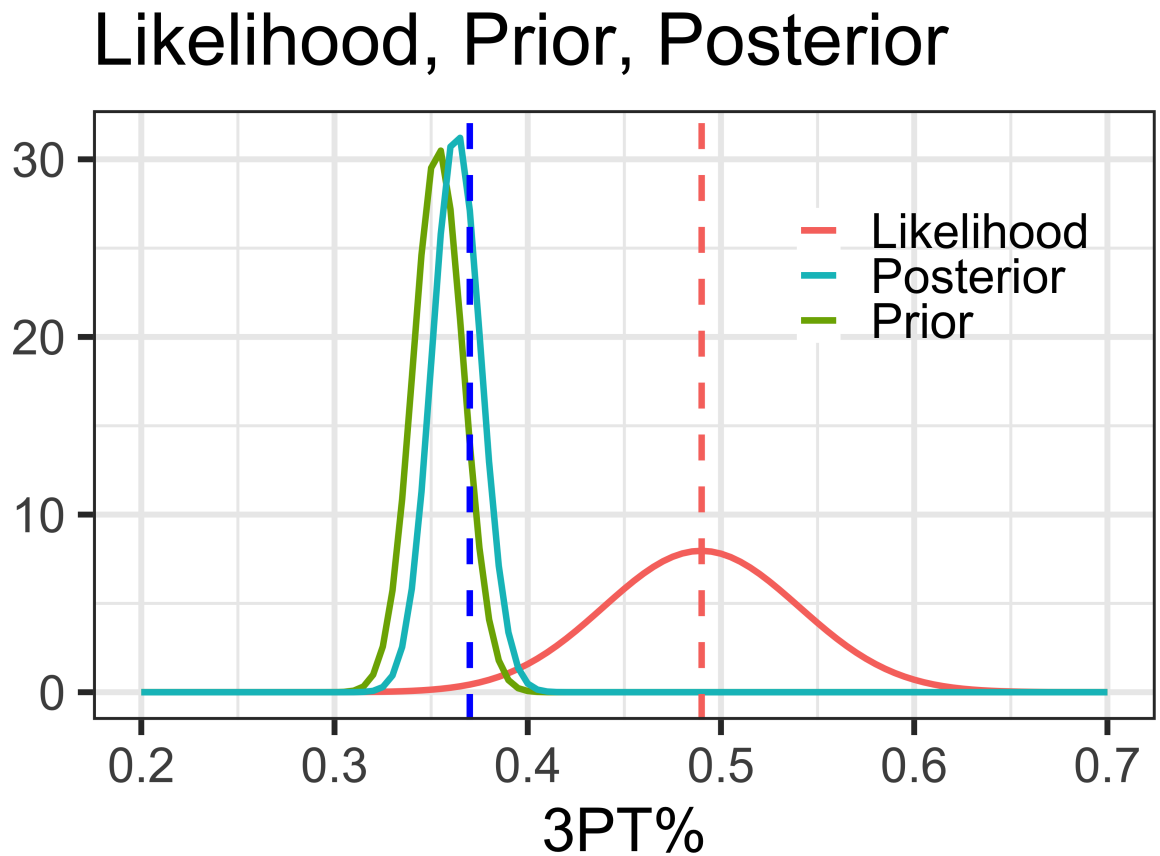
After 100 shots Robert Covington's 3PT% was 0.49



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

The Poisson Distribution

- A useful model for count data
- Events occur independently at some rate λ
- Mean = variance = λ.
- Example applications:
  - Epidemiology (disease incidence)
  - Astronomy (e.g. the number of meteorites entering the solar system each year)
  - The number of patients entering the emergency room
  - The number of times a neuron in the brain “fires”

Poisson model with exposure

- Assume y_i is Poisson with rate λ and exposure ν_i:

$$\frac{\lambda^y e^{\lambda}}{y!}$$

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

- 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.

Conjugate Prior for the Poisson Distribution

$Y_1, \ldots Y_n \sim Pois(v_i \lambda)$

$$L(\lambda) \propto \prod_i \frac{e^{-v_i\lambda} (v_i\lambda)^{y_i}}{\cancel{y_i!}} \propto e^{-\sum v_i \lambda} \lambda^{\sum y_i}$$

$$\Longrightarrow \hat{\lambda}_{MLE} = \frac{\sum y_i}{\sum v_i}$$

$$P(\lambda) \propto e^{-\beta\lambda} \lambda^{\alpha-1}$$

$Gam(\alpha, \beta)$

"pseudo-counts"

$$P(\lambda \mid y_1, \ldots y_n, v's) \propto \lambda^{\sum y_i + \alpha - 1} e^{-(\beta + \sum v_i)\lambda}$$

$$\longrightarrow Gam(\sum y_i + \alpha, \sum v_i + \beta)$$

The Gamma distribution

We use the shape-rate parameterization of the Gamma

- $\text{Gamma}(a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$
- $E[\lambda] = a/b$ and $\text{Var}[\lambda] = a/b^2$ Memorize.
- $\text{mode}[\lambda] = (a-1)/b$ if $a > 1, 0$ otherwise
- In R: `dgamma`, `rgamma`, `pgamma`, `qgamma`

a: pseudo totals

b: pseudo exposure.

$\frac{a}{b} \sim$ totals/time

The Gamma distribution

Mean = 1

Mean = 2

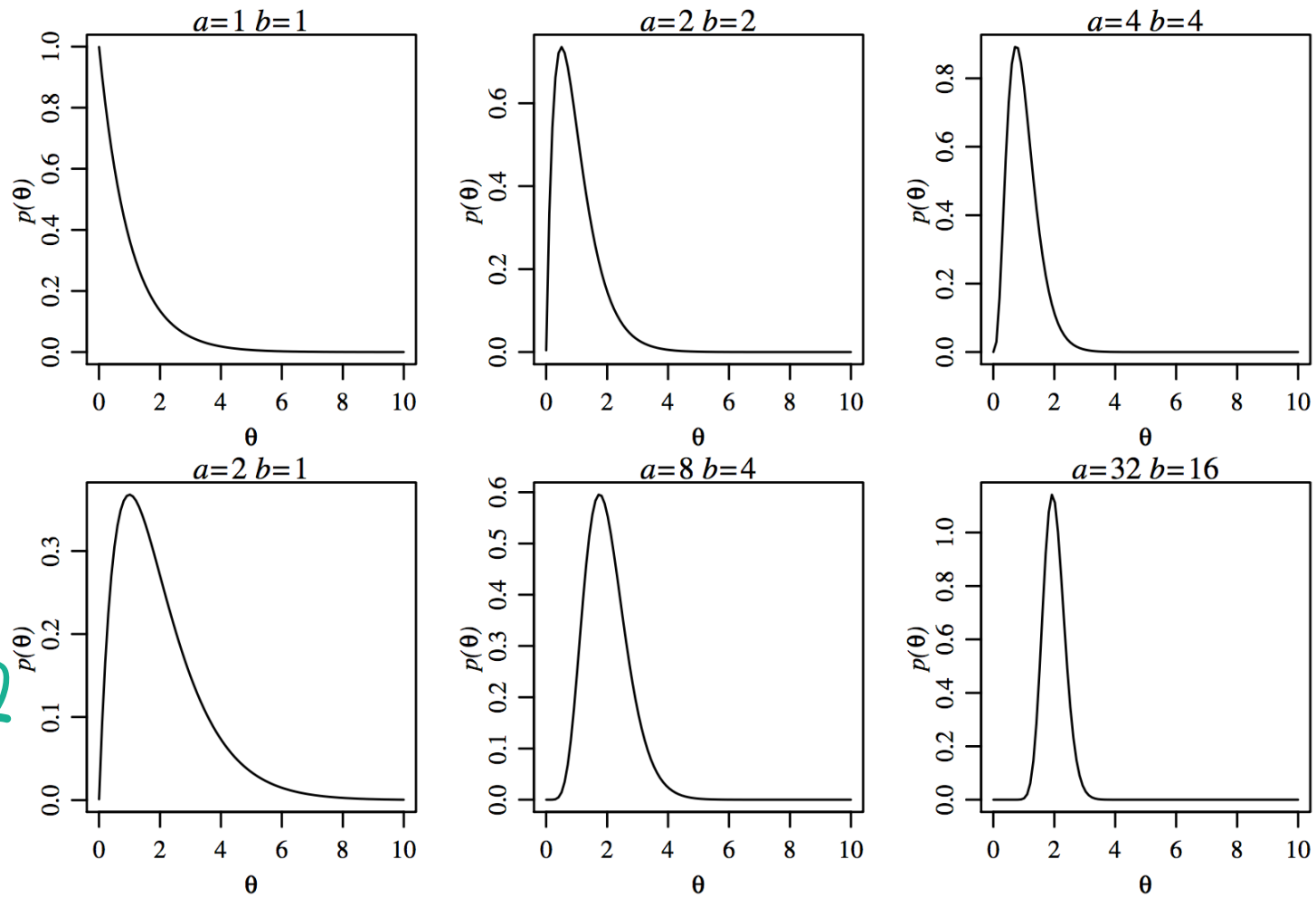


Fig. 3.8. Gamma densities.

The posterior mean

$$\lambda \mid y_1, \ldots y_n, v_1, \ldots v_n \sim \text{Gam}\left(\sum y_i + \alpha, \sum v_i + B\right)$$

$$E[\lambda \mid \ldots] = \frac{\sum y_i + \alpha}{\sum v_i + B}$$

$$= w \hat{\lambda}_{MLE} + (1-w)\lambda_{prior}$$

Prove this

$$= w \frac{\sum y_i}{\sum v_i} + (1-w)\frac{\alpha}{B} \quad \text{where} \quad w = \frac{\sum v_i}{\sum v_i + B}$$

Conjugate Priors and Exponential Families

- An k-dimensional exponential family distribution can be written as $p(y \mid \theta) = h(y)exp(\eta(\theta)'T(y) - A(\theta))$
- For any k-parameter exponential family there exists a (k + 1)-parameter conjugate prior.

K-parameters

$$L(\theta) \propto \exp\left[\sum^{K} \eta_i(\theta) T_i(y) - A(\theta)\right]$$

$$P_\alpha(\theta) \propto \exp\left[\sum^{K} \eta_i(\theta)\alpha_i - \alpha_{k+1} A(\theta)\right]$$

$$\Longrightarrow P(\theta \mid \cdots) \propto \exp\left[\eta_i(\theta)[\alpha_i + T_i(y)] + (1+\alpha_{k+1})A(\theta)\right]$$