Lecture 2: One Parameter Models

Professor Alexander Franks

Announcements

- Reading: Chapter 2 (these lecture notes)
- Chapter3 (next lecture notes)

Example: estimating spam accounts on twitter

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

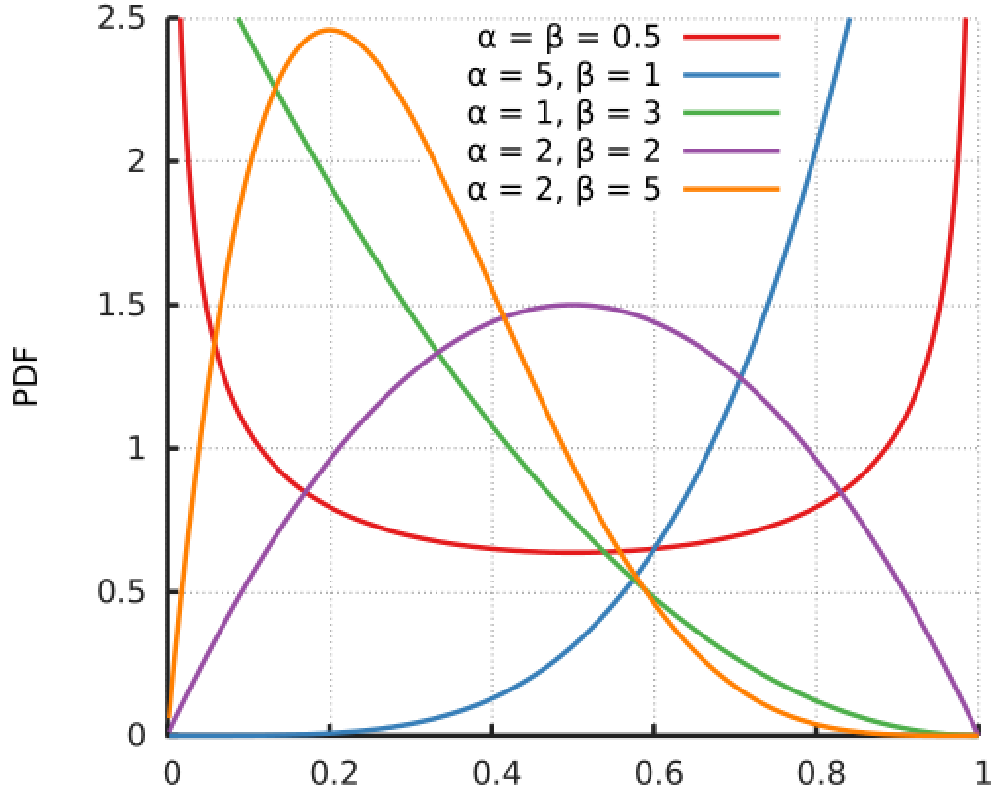
Example: estimating shooting skill in basketball

- Assume every shot is independent (reasonable) and identically distributed (less reasonble?)
- Let $Y \sim \text{Bin}(n, \theta)$ 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?

The Binomial Model

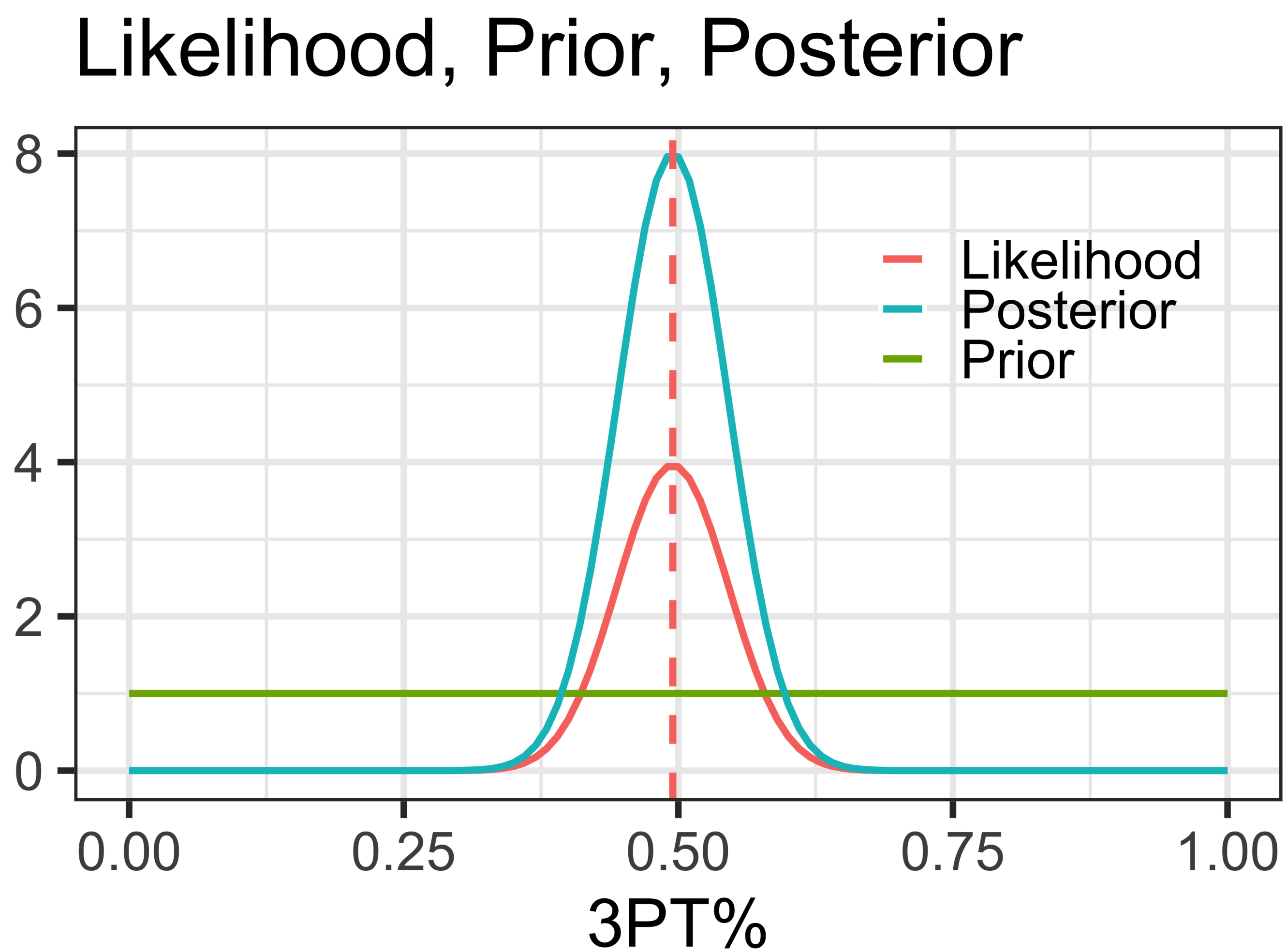
- The uniform prior: $p(\theta) = \mathrm{Unif}(0,1) = \mathbf{1}\{\theta \in [0,1]\}$
  - A “non-informative” prior
- Posterior: $p(\theta \mid y) \propto \underbrace{\theta^y(1-\theta)^{n-y}}_{\text{likelihood}} \times \underbrace{\mathbf{1}\{\theta \in [0,1]\}}_{\text{prior}}$
- The above posterior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

Beta Distributions



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

Example: estimating shooting skill in basketball



Summarizing Posterior Results

Summarizing Posterior Results

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

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

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

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

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

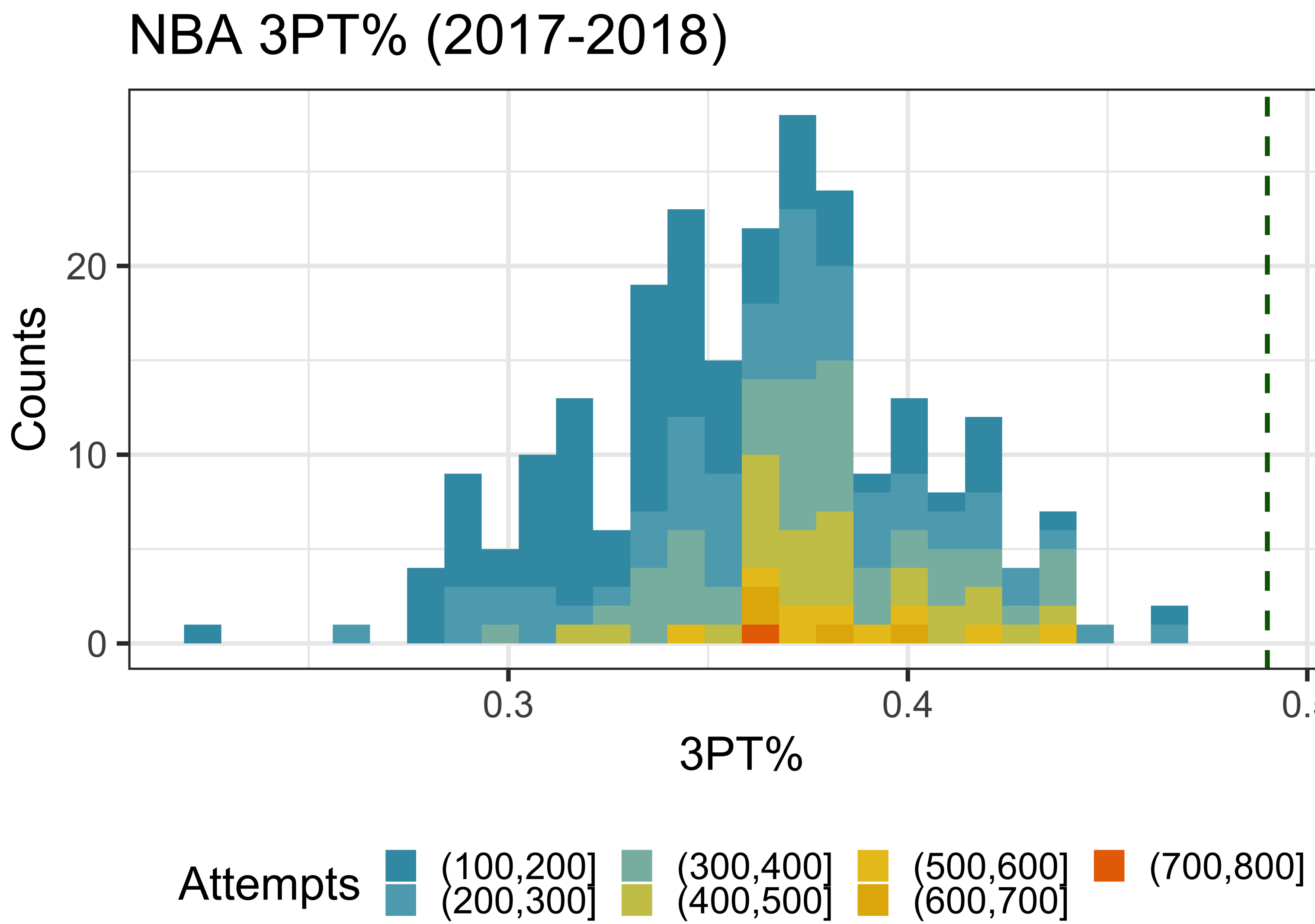
Pseudo-Counts Interpretation

- 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: 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



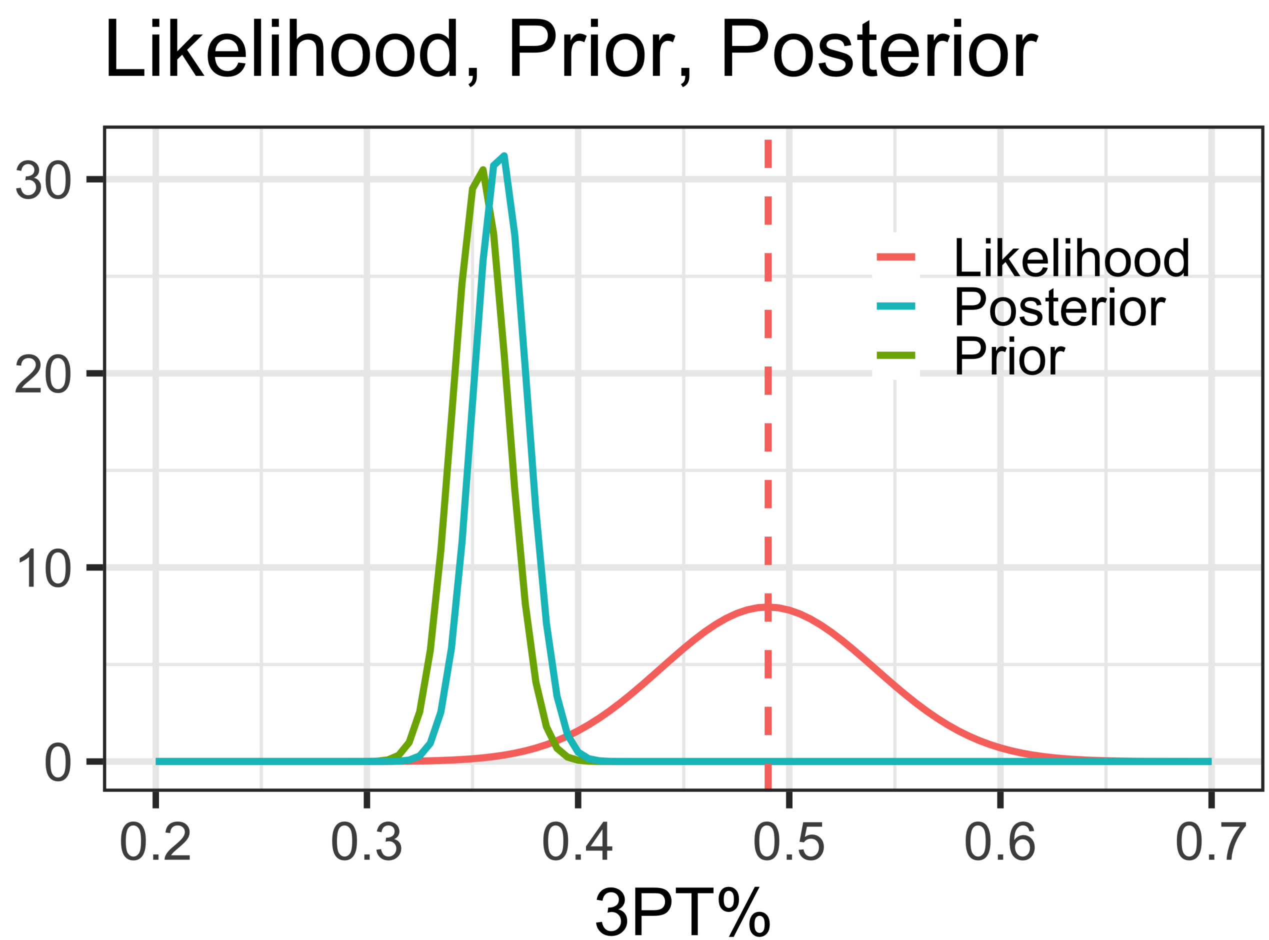
Regression Toward the Mean

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)

Robert Covington 2017-2018 estimates

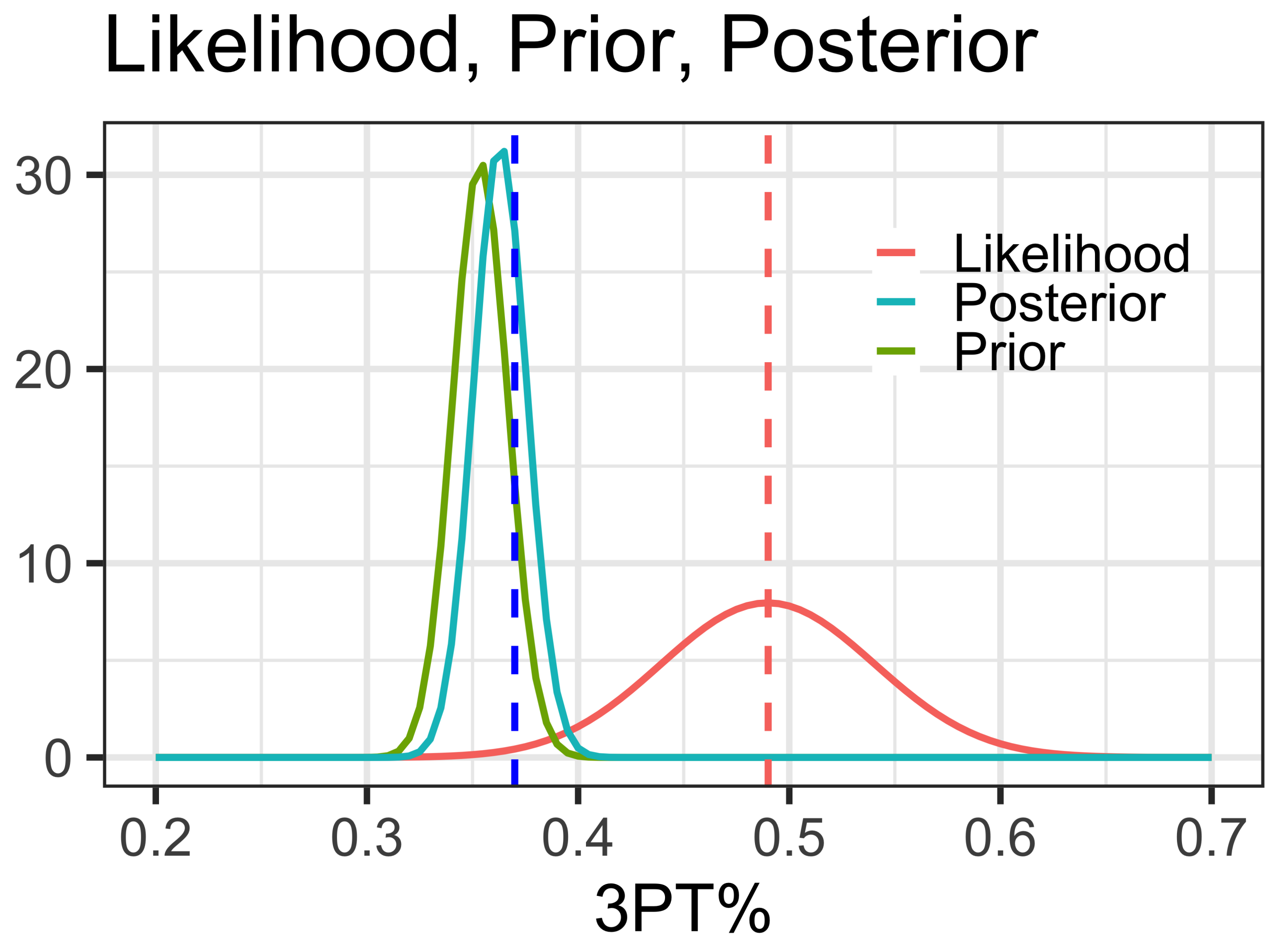
After 100 shots Robert Covington's 3PT% was 0.49



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:

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

- 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

The Gamma distribution

We use the shape-rate parameterization of the Gamma

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

The Gamma distribution

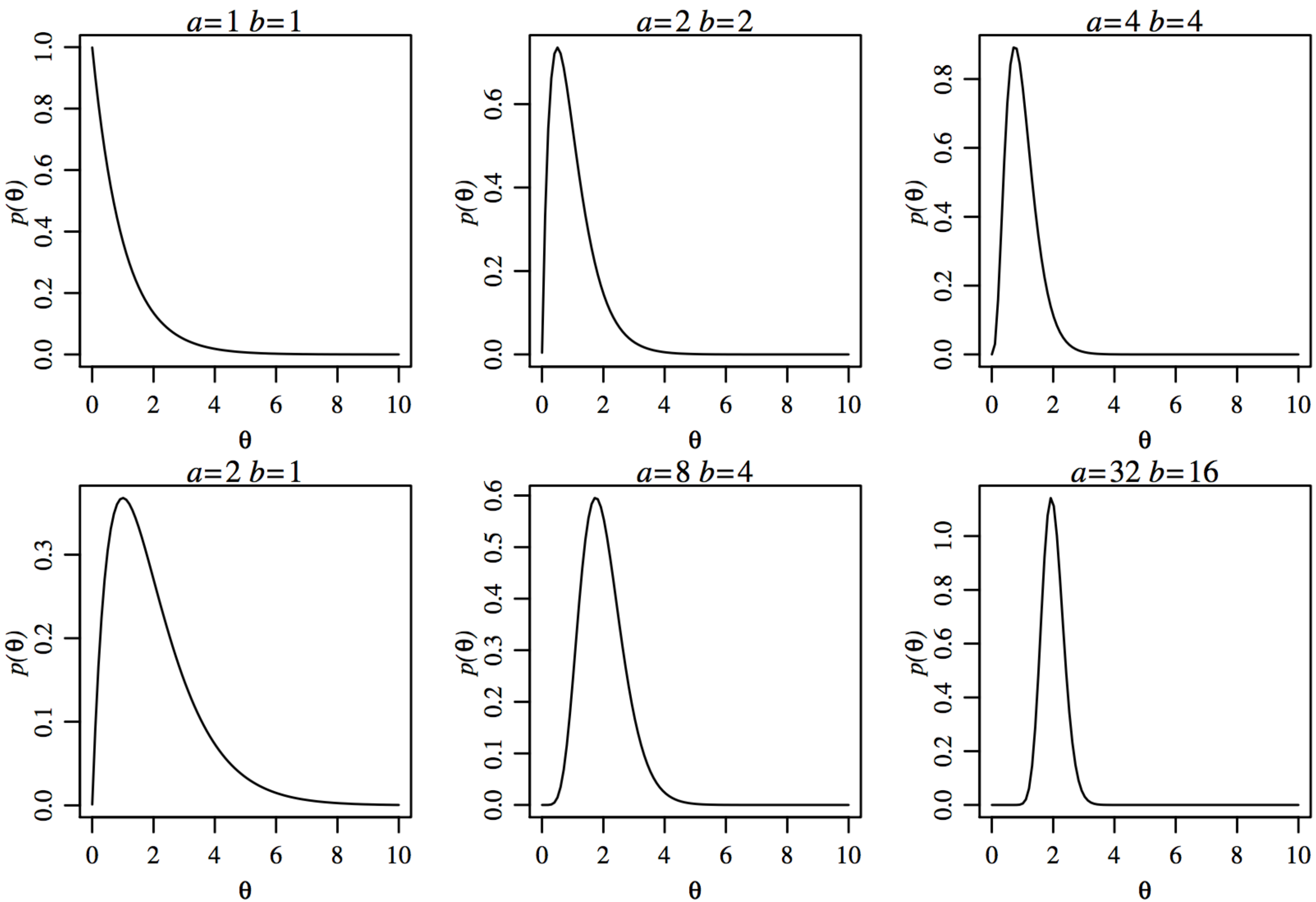


Fig. 3.8. Gamma densities.

The posterior mean

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.

Uncertainty Quantification

Posterior Credible Intervals

- Frequentist interval: $Pr(l(Y) < \theta < u(Y) \mid \theta) = 0.95$
  - Probability that the interval will cover the true value *before* the data are observed.
  - Interval is random since Y is random
- **Bayesian Interval**: $Pr(l(y) < \theta < u(y) \mid Y = y) = 0.95$
  - Information about the the true value of θ *after* observeing $Y = y$.
  - θ is random (because we include a prior), y is observed so interval is non-random.

Posterior Credible Intervals (Quantile-based)

- The easiest way to obtain a credible interval is to use the quantiles of the posterior distribution.

If we want $100 \times (1 - \alpha)$ interval, we find numbers $\theta_{\alpha/2}$ and $\theta_{1-\alpha/2}$ such that:

1. $p(\theta < \theta_{\alpha/2} \mid Y = y) = \alpha/2$
2. $p(\theta > \theta_{1-\alpha/2} \mid Y = y) = \alpha/2$

$$p(\theta \in [\theta_{\alpha/2}, \theta_{1-\alpha/2}] \mid Y = y) = 1 - \alpha$$

Interval for shooting skill in basketball

- The posterior distribution for Covington's shooting percentage is a

$$\text{Beta}(49 + 478, 50 + 873) = \text{Beta}(528, 924)$$

- For a 95% *credible* interval, $\alpha = 0.05$
  - Lower endpoint: `qbeta(0.025, 528, 924)`
  - Upper endpoint: `qbeta(0.975, 528, 924)`
  - $[\theta_{\alpha/2}, \theta_{1-\alpha/2}]$ = [0.34, 0.39]

Interval for shooting skill in basketball

- Bayes credible interval: $[\theta_{\alpha/2}, \theta_{1-\alpha/2}]$ = [0.34, 0.39]
- Frequentist *confidence* interval: [0.39, 0.59]
- End-of-season percentage was 0.37
- Credible intervals and confidence intervals have different meanings!

Highest Posterior Density (HPD) region

Definition: (HPD region) A $100 \times (1-\alpha)$ HPD region consists of a subset of the parameter space, $R(y) \in \Theta$ such that

1. $\Pr(\theta \in R(y) | Y = y) = 1 - \alpha$
   - The probability that θ is in the HPD region is $1 - \alpha$
2. $\text{If } \theta_a \in R(y), \text{ and } \theta_b \notin R(y), \text{ then } p\left(\theta_a | Y = y\right) > p\left(\theta_b | Y\right.$
   - All points in an HPD region have a higher posterior density tha points out- side the region.

The HPD region is the *smallest* region with probability $(1 - \alpha)$

Frequentist behavior of Bayesian intervals

- Bayesian credible intervals usually won't have exactly correct frequentist coverage
- If our prior was well-calibrated and the sampling model was correct, we'd have well-calibrated credible intervals
- And: asymptotically, a central posterior interval will cover the true value 95% of the time under repeated sampling

Frequentist-Bayes Unification

Under regularity conditions:

$$p(\theta \mid y_1, \ldots, y_n) \approx \mathrm{N}\left(\hat{\theta}, [I(\hat{\theta})]^{-1}\right)$$

Classical result:

$$\hat{\theta} \approx N(\theta, [I(\theta)]^{-1})$$

Sequential Bayesian Updating

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

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

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

Improper prior distributions

- For the Beta distribution we chose a uniform prior (e.g. $p(\theta) \propto \text{const}$). This was ok because
  - $\int_0^1 p(\theta) d\theta = \text{const} < \infty$
  - We say this prior distribution is *proper* because it is integrable
- For the Poisson distribution, try the same thing: $p(\lambda) \propto \text{const}$
  - $\int_0^\infty p(\lambda) d\lambda = \infty$
  - In this case we say $p(\lambda)$ is an *improper* prior

Improper prior distributions

- Sometimes there is an absence of precise prior information
- The prior distribution does not have to be proper but the posterior does!
  - A proper distribution is one with an integrable density
  - If you use an improper prior distribution, you need to check that the posterior distribution is also proper

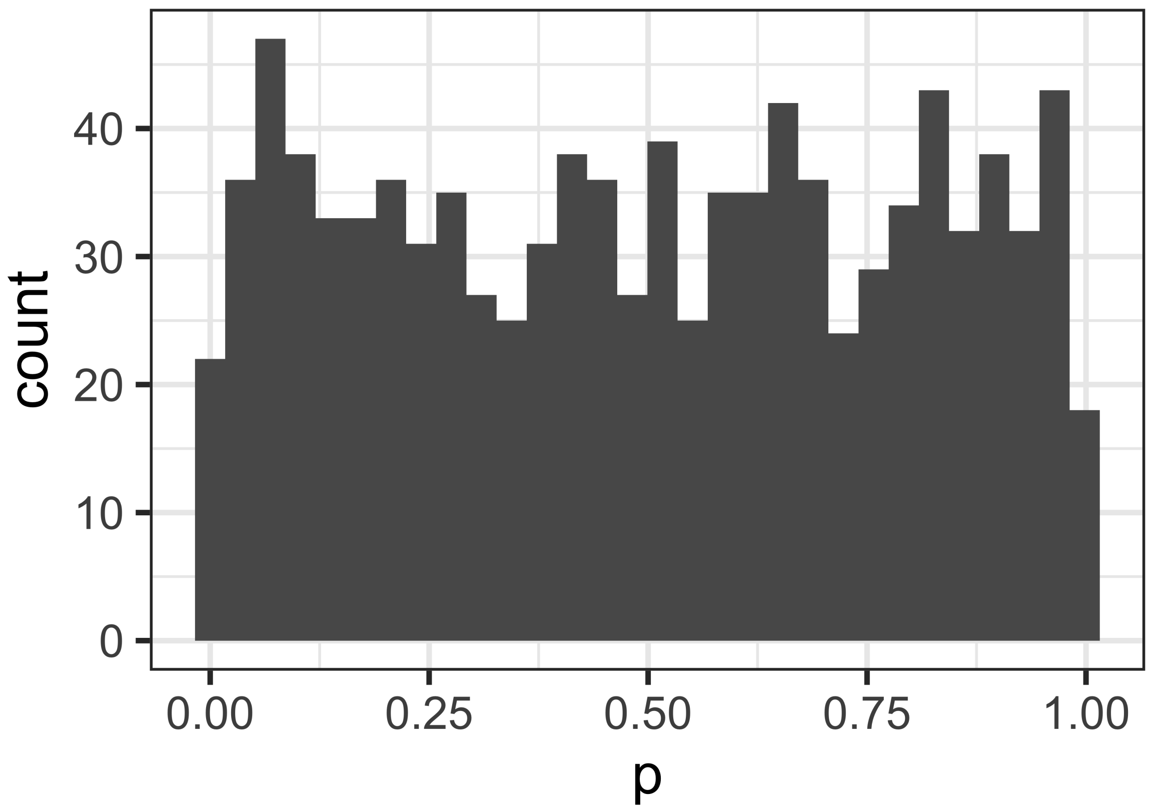
Objective Bayes

- Also called “default”, “reference”, “non-informative” prior distributions
- Laplace’s principle of insufficient reason
- Principle of maximum entropy (MAXENT).
- Matching prior distributions
  - Find prior distributions which lead to posterior intervals with approximate frequentist coverage
- Invariant priors

Laplace’s principle of insufficient reason

Uniform distribution for p

```r
p <- runif(1000)
tibble(p=p) %>% ggplot() +
  geom_histogram(aes(x=p), bins=30) +
  theme_bw(base_size=24)
```



Laplace's principle of insufficient reason

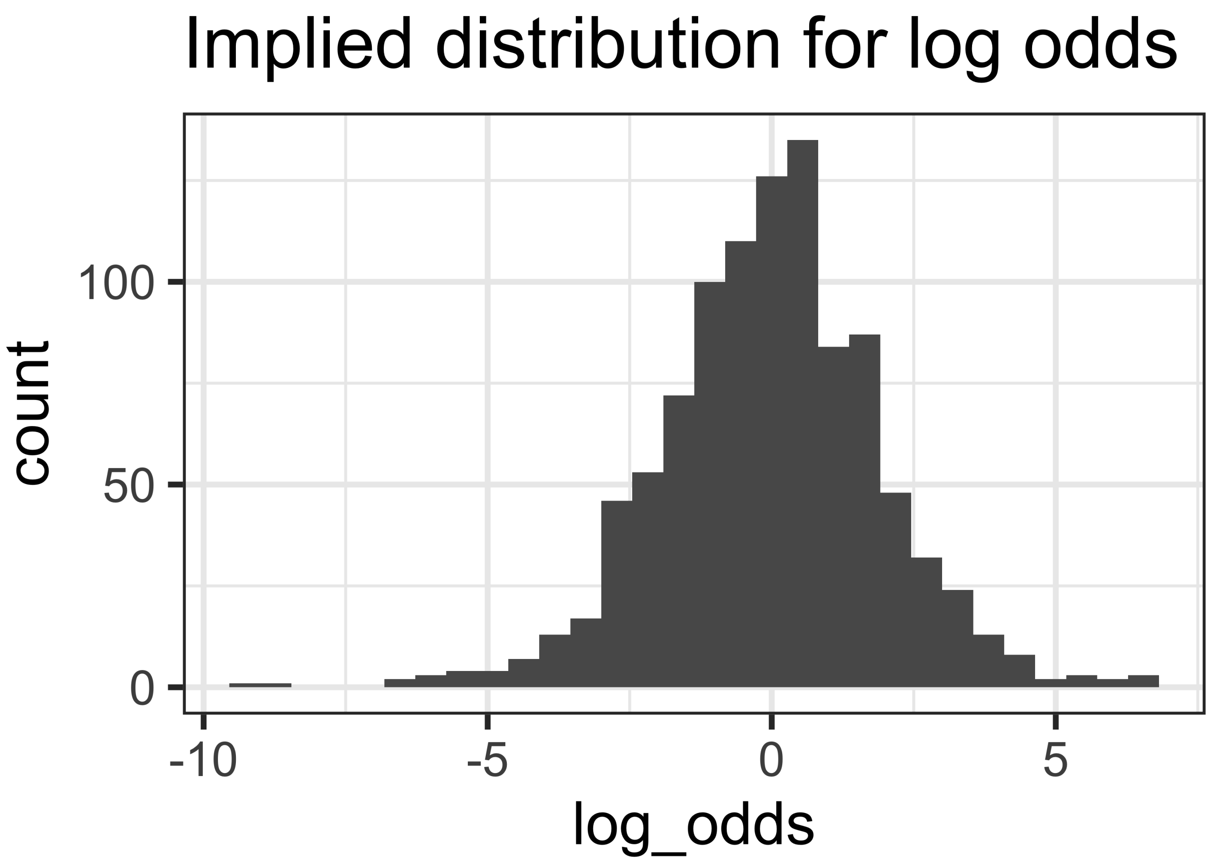
- Assume that $Y \sim \mathrm{Bin}(n, \theta)$ but that we're most interested the log odds:

$$\gamma = \log \mathrm{odds}(\theta) = \log \frac{\theta}{1 - \theta}$$

- What prior should we use if we want to be "noninformative"?

Difficulties with non-informative priors

```
log_odds <- log(p/(1-p))
tibble(log_odds=log_odds) %>% ggplot() +
  geom_histogram(aes(x=log_odds)) +
  theme_bw(base_size=24) +
  ggtitle("Implied distribution for log odds")
```



Method of transformations

Assume the prior density, $p(\theta)$. What is the implied prior density for the transformed parameter, $\gamma = g(\theta)$?

1. Find the inverse, $\theta = g^{-1}(\gamma)$
2. Compute $\frac{dg^{-1}(\gamma)}{d\gamma}$
3. Find $p_\gamma(\gamma) = \left| \frac{dg^{-1}(\gamma)}{d\gamma} \right| \times p_\theta(g^{-1}(\gamma))$

Jeffreys prior

- Idea: find a parameterization that is invariant under transformations
- Derivation:

Weakly Informative Priors

- A proper prior distribution, but intentionally include less information than is actually available a priori
- Construction:
  - Start with a noninformative prior and add then add enough information to constrain inferences to be more reasonable
  - Or: start with an informative prior and remove information
- Example: coefficients in a logistic regression should have a magnitude less than 10, in general

Prediction

Posterior predictive distribution

- An important feature of Bayesian inference is the existence of a predictive distribution for new observations.
  - Let $\tilde{y}$ be a new (unseen) observation, and $y_1, \ldots y_n$ the observed data.
  - The Posterior predictive distribution is $p(\tilde{y} \mid y_1, \ldots y_n)$

Posterior predictive distribution

- The predictive distribution does not depend on unknown parameters
- The predictive distribution only depends on observed data
- Asks: what is the probability distribution for new data given observations of old data?

Another Basketball Example

- I take free throw shots and make 1 out of 2. How many do you think I will make if I take 10 more?
- If my true "skill" was 50%, then $\tilde{Y} \sim \mathrm{Bin}(10, 0.50)$
- Is this the correct way to calculate the predictive distribution?

Posterior Prediction

If you know θ, then we know the distribution over future attempts:

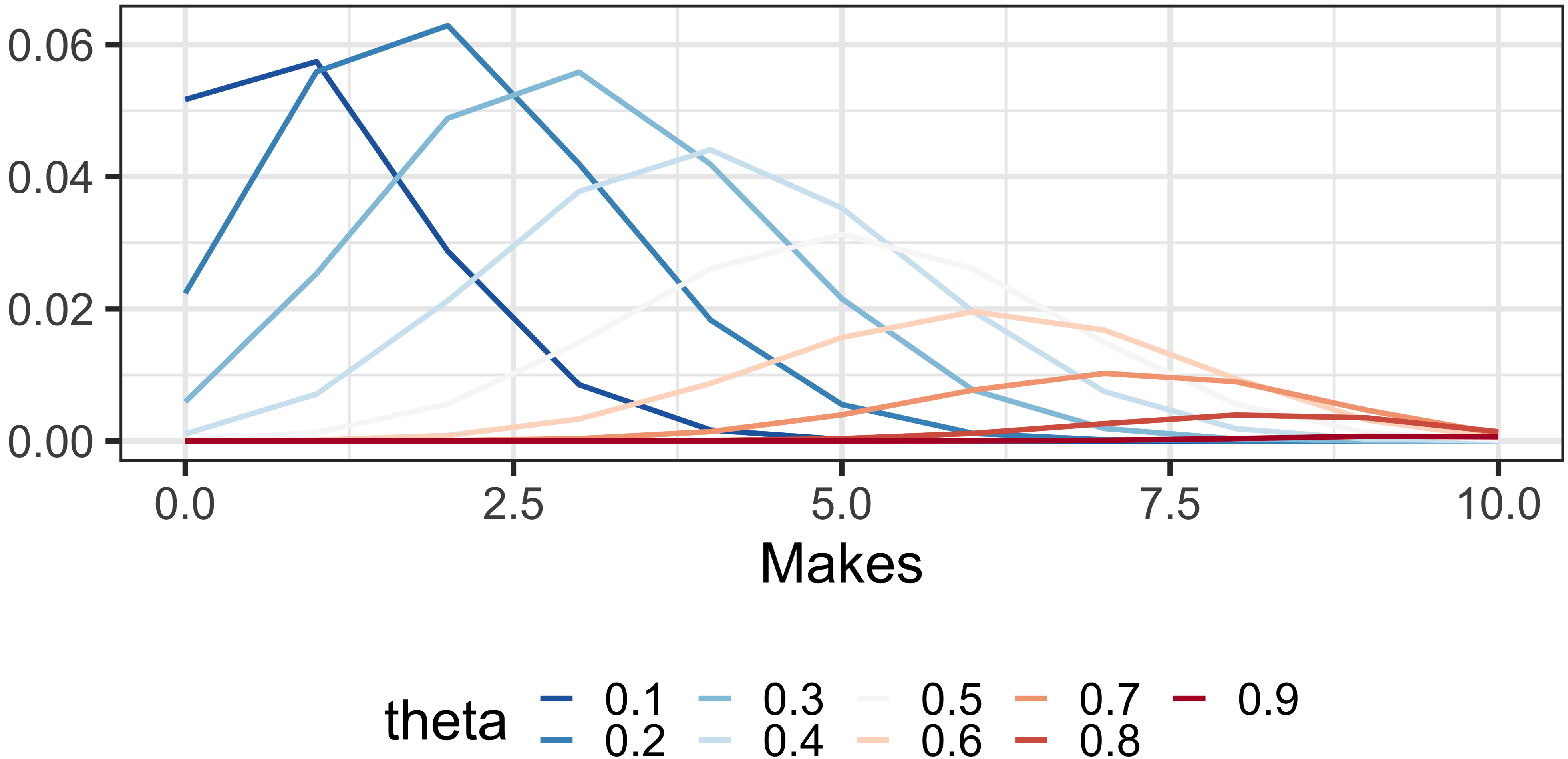
$$\tilde{Y} \sim \mathrm{Bin}(10, \theta)$$

Posterior Prediction

- We already observed 1 make out of 2 tries.
- Assume a Beta(1, 3) prior distribution
  - e.g. a priori you think I'm more likely to make 25% of my shots
- Then $p(\theta \mid Y = 1, n = 2)$ is a $\mathrm{Beta}(2, 4)$
- Intuition: weight $\tilde{Y} \sim \mathrm{Bin}(10, \theta)$ by $p(\theta \mid Y = 1, n = 2)$

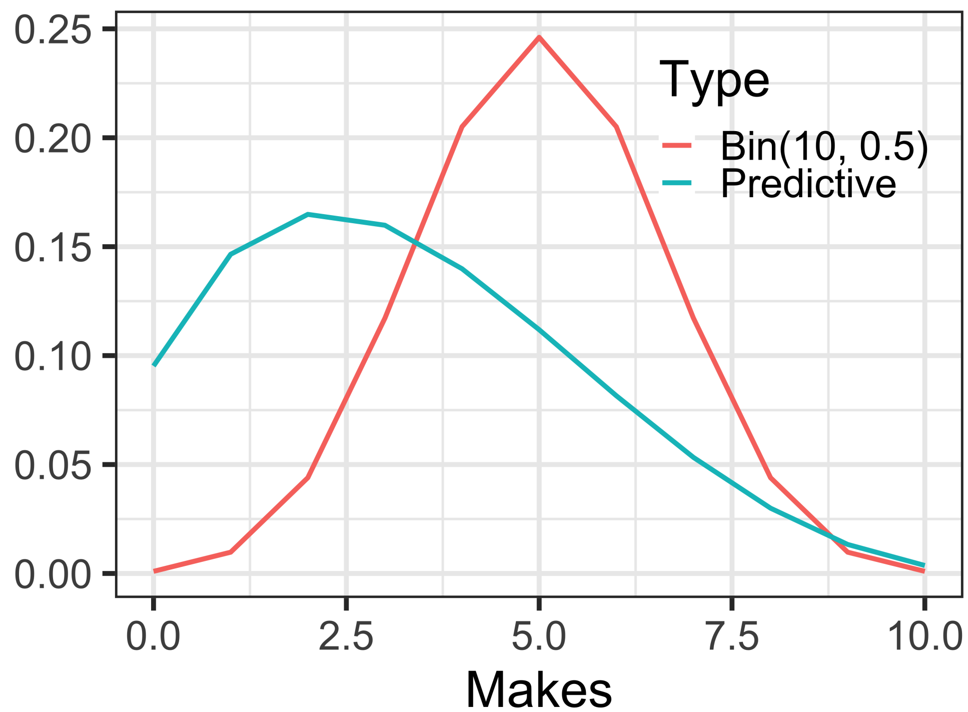
Posterior Prediction

If I take 10 more shots how many will I make?



Posterior predictive distribution

$$p(\theta) = \text{Beta}(1, 3), p(\theta \mid y) = \text{Beta}(2, 4)$$



The predictive density, $p(\tilde{y} \mid y)$, answers the question "if I take 10 more shots how many will I make, given that I already made

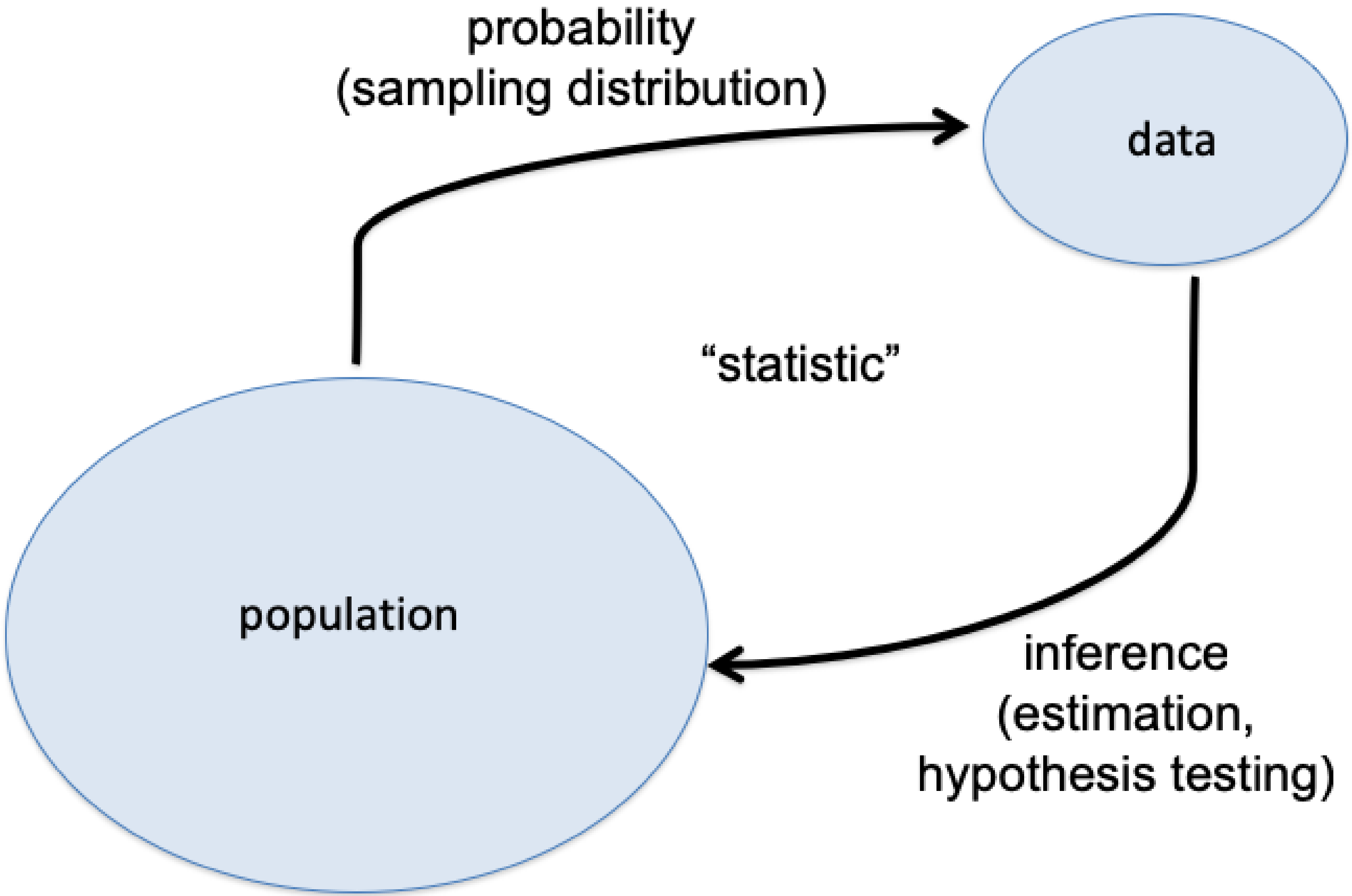
The posterior predictive distribution

$$\begin{aligned} p(\tilde{y} \mid y_1, \ldots y_n) &= \int p(\tilde{y}, \theta \mid y_1, \ldots y_n) d\theta \\ &= \int p(\tilde{y} \mid \theta) p(\theta \mid y_1, \ldots y_n) d\theta \end{aligned}$$

- The posterior predictive distribution describes our uncertainty about a new observation after seeing n observations
- It incorporates uncertainty due to the sampling in a model $p(\tilde{y} \mid \theta)$ *and* our posterior uncertainty about the data generating parameter, $p(\theta \mid y_1, \ldots y_n)$

The Beta-Binomial Distribution

Posterior predictive density



The prior predictive distribution

$$\begin{aligned} p(\tilde{y}) &= \int p(\tilde{y}, \theta) d\theta \\ &= \int p(\tilde{y} \mid \theta) p(\theta) d\theta \end{aligned}$$

- The prior predictive distribution describes our uncertainty about a new observation before seeing data
- It incorporates uncertainty due to the sampling in a model $p(\tilde{y} \mid \theta)$ *and* our prior uncertainty about the data generating parameter, $p(\theta)$

Summary

- Conjugate priors
- Credible intervals
- Noninformative priors
- Posterior (or prior) predictive distribution