Lecture 2: One Parameter Models

Professor Alexander Franks

Uncertainty Quantification

Posterior Credible Intervals

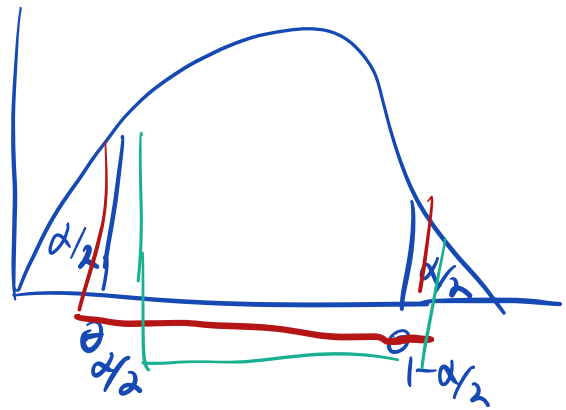
- Frequentist interval: $Pr(l(Y) < \theta < u(Y) \mid \theta) = 0.95$
  - Probability that the interval will cover the true value *before* the data are observed. "counterfactual world"
  - Interval is random since Y is random
- **Bayesian Interval**: $Pr(l(y) < \theta < u(y) \mid Y = y) = 0.95$
  - Information 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. $p(\theta > \theta_{1-\alpha/2} \mid Y = y) = \alpha/2$



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

$P(\Theta \in R \mid Y=y) = .95$

Interval for shooting skill in basketball

$y \overset{iid}{\sim} \text{Bin}(n, \theta)$, $y = 49$, $n = 100$

- The posterior distribution for Covington's shooting percentage is a $P(\theta \mid y) \propto \text{Beta}(y + \alpha, n - y + \beta)$

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

- 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!

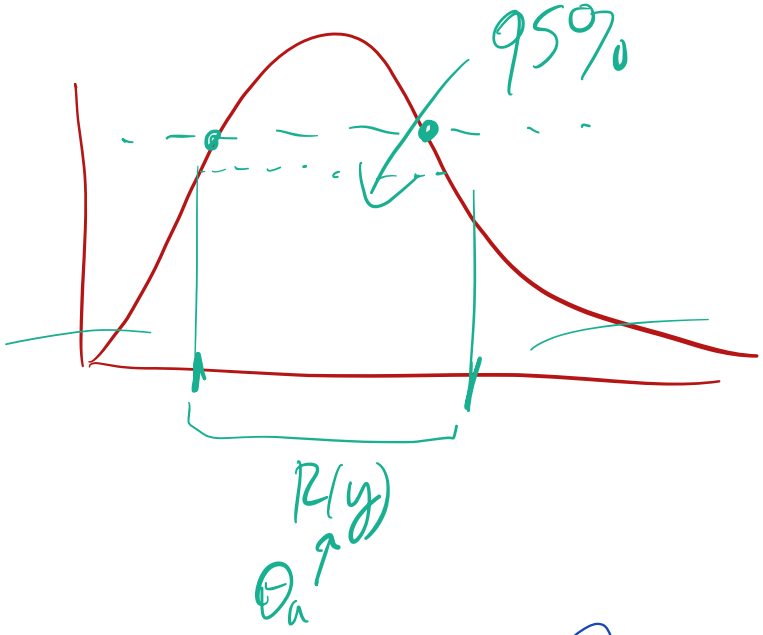
Highest Posterior Density (HPD) region

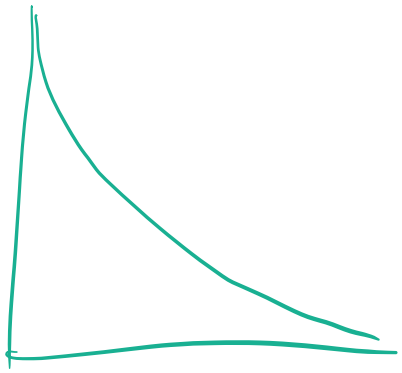
Definition: (HPD region) A $100 \times (1-\alpha)$ HPD region consists of a subset of the parameter space, $R(y) \in \Theta$ such that

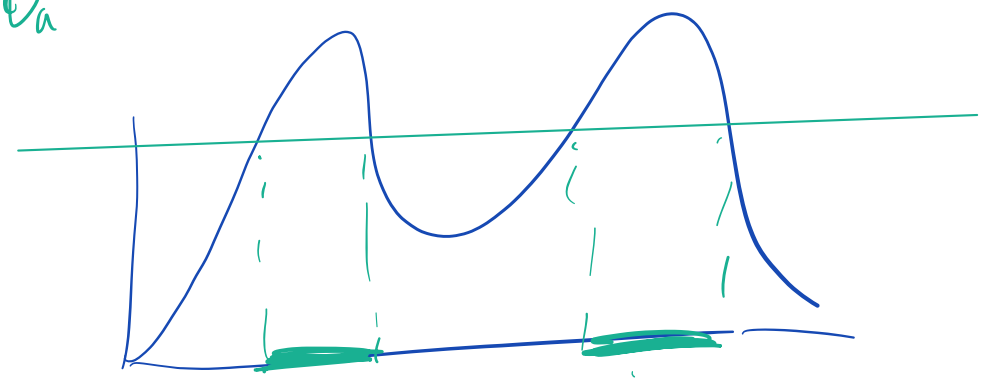
1. $\Pr(\theta \in R(y)|Y = y) = 1 - \alpha$
   - The probability that θ is in the HPD region is $1 - \alpha$
2. If $\theta_a \in R(y),$ and $\theta_b \notin R(y),$ then $p(\theta_a|Y = y) > p(\theta_b|Y$
   - All points in an HPD region have a higher posterior density tha points out- side the region.

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

$\rightarrow P(\Theta_a|y) > P(\Theta_b|y)$


95%
R(y)





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

Ch. 4 (Stein)

$$\log P(\theta|y) \approx \log P(\hat{\theta}|y) + \underbrace{0}_{\text{linear}} + \frac{1}{2}(\hat{\theta}_{mode} - \theta)\left(\frac{d^2 \log P(\theta|y)}{d\theta}\right)(\hat{\theta}_{mode} - \theta) + \dots$$

(expand about $\hat{\theta}_{mode}$)

$$-I(\hat{\theta})$$

Bayes $P(\theta|y_1, \dots y_n) \xrightarrow{d} N(\hat{\theta}_{mode}, I(\hat{\theta})^{-1})$

Classical Freq: $\hat{\theta}_{MLE} \approx N(\theta, I(\theta)^{-1})$

Bayes $I(\theta)^{1/2}(\theta - \hat{\theta})|y \sim N(0, I)$ ← identity

Freq $I(\theta)^{1/2}(\hat{\theta} - \theta)|\theta \sim N(0, I)$

Frequentist-Bayes Unification

Under regularity conditions:

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

Classical result:

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

Sequential Bayesian Updating

New Prior $\tilde{P}_1(\theta)$

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

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

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

$$L(\theta) \propto \prod_{i=1}^{n} P(y_i \mid \theta)$$

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

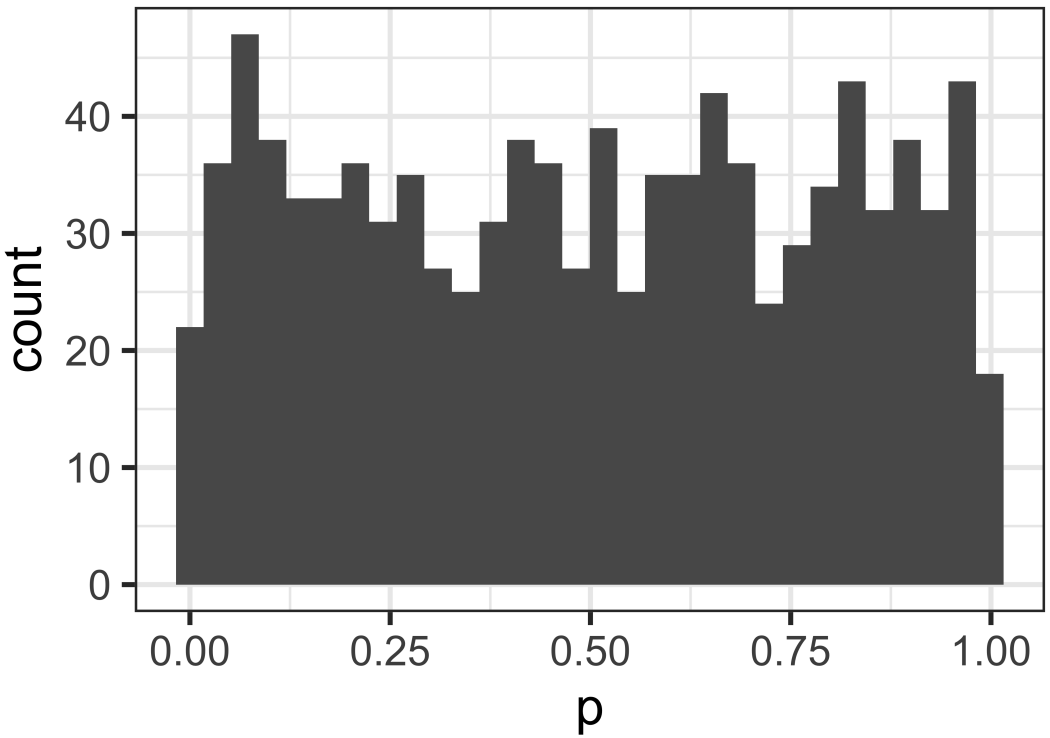
Objective Bayes ("Obayes")

- 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 (Jeffreys)

Laplace's principle of insufficient reason

Uniform distribution for p

```
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

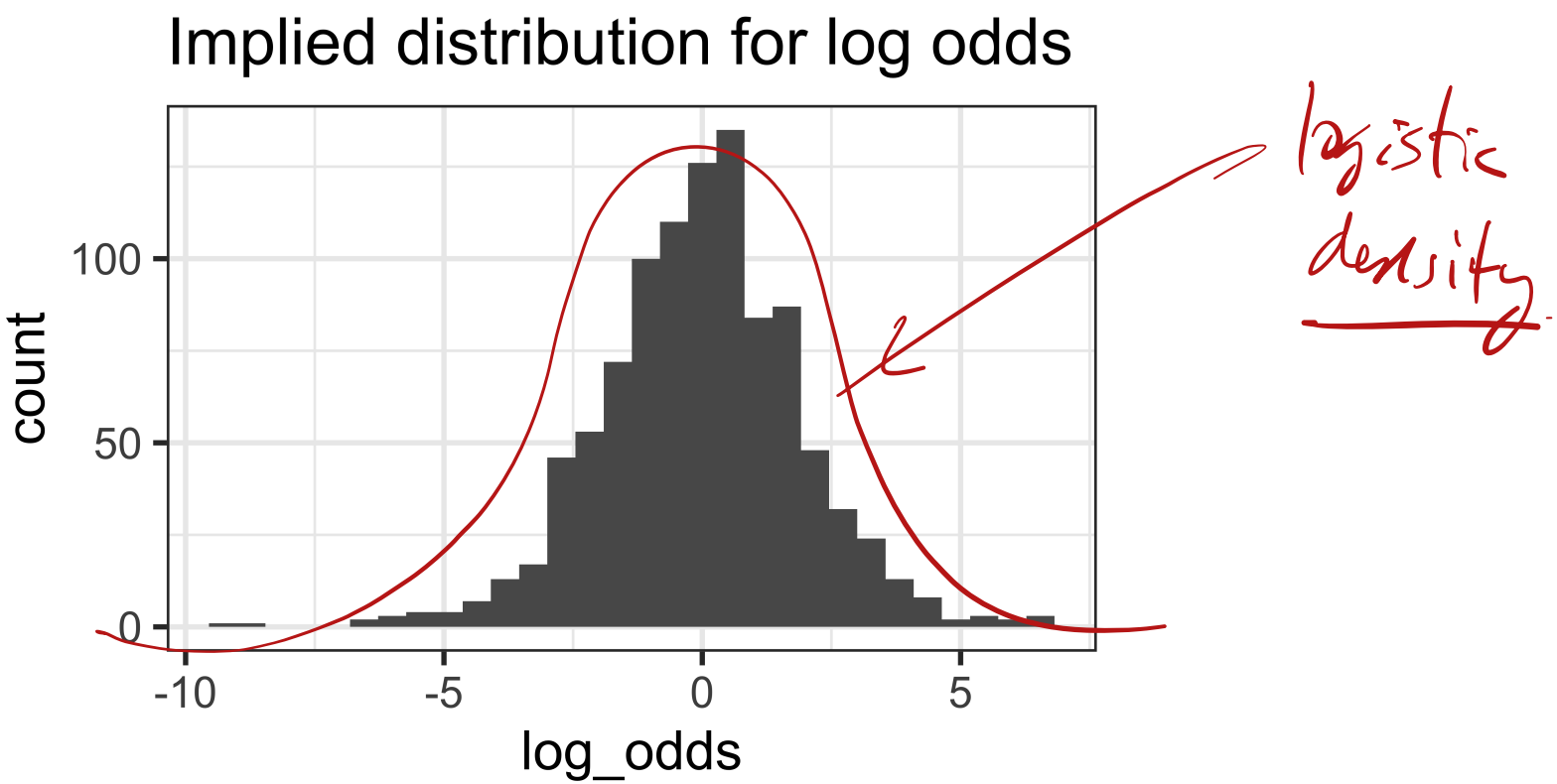
- Assume that $Y \sim \text{Bin}(n, \theta)$ but that we're most interested the log odds:

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

- 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, $\gamma = g(\theta)$?

1-1

g^{-1}

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

$\gamma = \log\left(\frac{\theta}{1-\theta}\right)$

$\Rightarrow \theta = \frac{e^\gamma}{1+e^\gamma}$

$P_\theta(\theta) \sim \text{Unif}$

$P_\gamma(\gamma) = \left|\frac{d\, e^\gamma/1+e^\gamma}{d\gamma}\right| \times 1 = \frac{e^\gamma}{(1+e^\gamma)^2}$

Jeffreys prior

$$I(\theta) = -E\left[\frac{d^2 \ell(\theta)}{d\theta}\right]$$

- Idea: find a parameterization that is invariant under transformations $p(\theta) \overset{?}{\rightarrow} p_\phi(\phi), \quad \phi = g(\theta)$
- Derivation:

$$p_\theta(\theta) \propto \sqrt{I(\theta)} \quad \text{(Jeffreys prior)}$$

Fact: $I_\phi(\phi) = I_\theta(\theta)\left(\frac{d\theta}{d\phi}\right)^2$

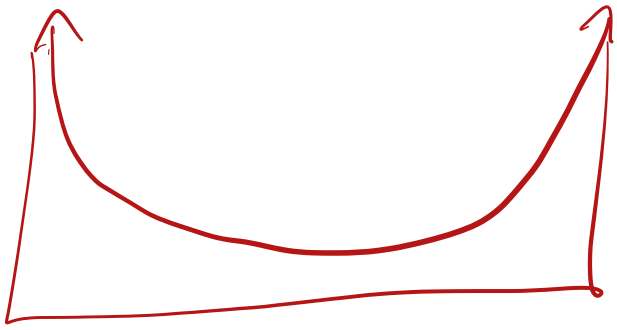
$$p_\phi(\phi) = p_\theta(\theta)\left|\frac{d\theta}{d\phi}\right| \propto \sqrt{I(\theta)}\left|\frac{d\theta}{d\phi}\right| = \sqrt{I(\phi)\left|\frac{d\phi}{d\theta}\right|^2}\left|\frac{d\theta}{d\phi}\right|$$

$= \sqrt{I(\phi)}$

$Y \sim Bin(n, \theta)$

$P(\theta) \propto \sqrt{I(\theta)} = \theta^{-1/2}(1-\theta)^{-1/2}$

$Beta(1/2, 1/2)$



$Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$ ~~σ^2~~ known $\quad I(\mu) = n/\sigma^2$

$P_N(\mu) \propto$ const. no μ!

Improper!

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 $p(\tilde{y} \mid y_1, \ldots y_n)$

new obs

previous

$$P(\tilde{y} \mid y_1, \ldots y_n) = \int P(\tilde{y} \mid \theta) P(\theta \mid y_1, \ldots y_n) d\theta$$

obs data

No θ!

$P(\tilde{y} \mid y_1, \ldots y_n)$ vs $P(\theta \mid y_1, \ldots y_n)$

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 $\tilde{Y} \sim \text{Bin}(10, 0.50)$
- Is this the correct way to calculate the predictive distribution?

Posterior Prediction

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

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

Posterior Prediction

- We already observed 1 make out of 2 tries.
- Assume a Beta(1, 3) prior distribution → $1/(1+3) = .25$
  - 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 $p(\theta \mid Y = 1, n = 2)$

$$p(\tilde{y} \mid y_1, \ldots y_n) = \int \theta^{\tilde{y}} (1-\theta)^{10-\tilde{y}} p(\theta \mid y)\, d\theta$$

$$P(\tilde{y} \mid y=1, n=2) = \int_0^1 P(\tilde{y} \mid \theta) P(\theta \mid y)\, d\theta$$

$$= \int_0^1 \binom{10}{\tilde{y}} \theta^{\tilde{y}} (1-\theta)^{10-\tilde{y}} \theta^{2-1} (1-\theta)^{4-1} \frac{\Gamma(6)}{\Gamma(2)\Gamma(4)}\, d\theta$$

$$= \binom{10}{\tilde{y}} \frac{\Gamma(6)}{\Gamma(4)\Gamma(2)} \int_0^1 \theta^{\tilde{y}+1} (1-\theta)^{13-\tilde{y}}\, d\theta$$

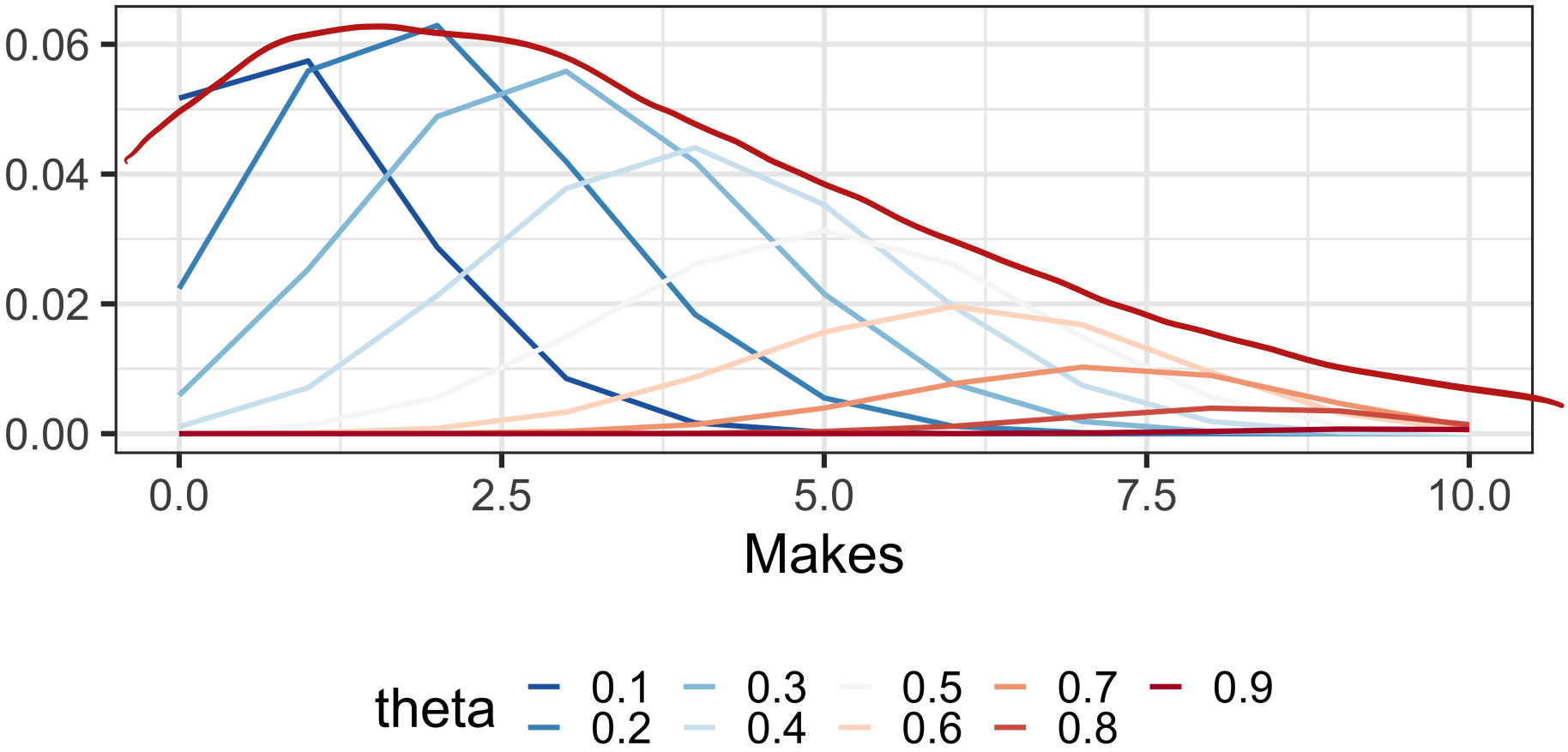
Unnormalized Beta $\rightarrow$ Beta$(\tilde{y}+2, 14)$

$$= \binom{10}{\tilde{y}} \frac{\Gamma(6)}{\Gamma(4)\Gamma(2)} \frac{\Gamma(\tilde{y}+2)\Gamma(14-\tilde{y})}{\Gamma(16)}$$

Beta-Binomial

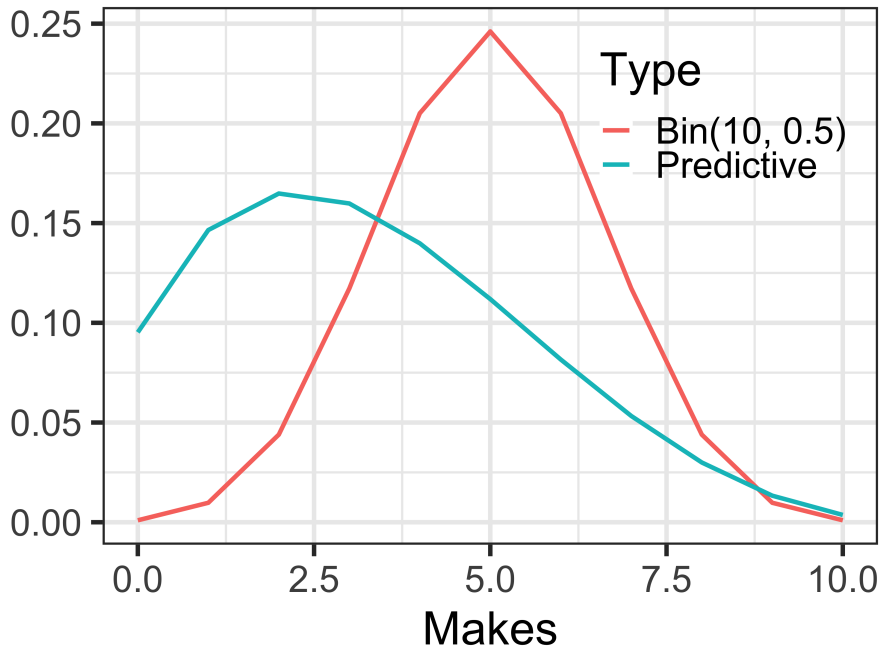
Posterior Prediction

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

$Y_i \sim \text{Bin}(n_i, \theta_i)$ (makes by player i)

$P(\theta_i) \overset{iid}{\sim} \text{Beta}(\alpha, B)$

Idea: Learn α, B from data.

$$P(Y_1, \ldots Y_n \mid \alpha, B) = \int \prod P(Y_i \mid \theta_i) P(\theta_i) \, d\theta$$

"prior" predictive Distribution.

and maximize over α, B

Then do Bayes

with $\theta_i \sim \text{Beta}(\hat{\alpha}, \hat{B})$

$$P(\theta_1 \mid y_1, \hat{\alpha}, \hat{B})$$