

# **Intervals and Predictive Distributions**

**Professor Laura Baracaldo**

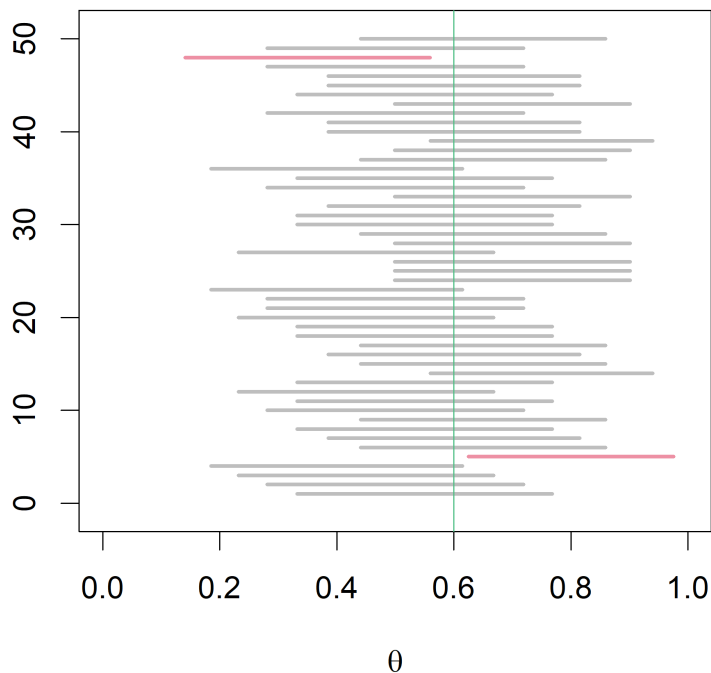
# Announcements

- Reading: Chapters 8.1 (intervals) and 8.3 (posterior prediction) Bayes Rules
-

# Reminder: Frequentist confidence interval

- 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

# Reminder: Frequentist confidence interval



We expect  $0.05 \times 50 = 2.5$  will *not* cover the true parameter 0.6

# 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

# 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  $\theta$  *after* observeing  $Y = y$ .
  - $\theta$  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 confidence 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:

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

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

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

- Use quantile functions in R, e.g. `qbeta`, `qpois`, `qnorm` etc.

## Example: 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)`
  - $[\theta_{\alpha/2}, \theta_{1-\alpha/2}] = [0.34, 0.39]$



# Example: 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)`
  - $[\theta_{\alpha/2}, \theta_{1-\alpha/2}] = [0.34, 0.39]$
- Compared to frequentist *confidence* interval without prior information:  $[0.39, 0.59]$
- 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  $\theta$  is in the HPD region is  $1 - \alpha$

# 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  $\theta$  is in the HPD region is  $1 - \alpha$

2. If  $\theta_a \in R(y)$ , and  $\theta_b \notin R(y)$ , then  $p(\theta_a | Y = y) > p(\theta_b | Y = y)$

- All points in an HPD region have a higher posterior density than points outside the region.

The HPD region can be discontinuous (hence "region")

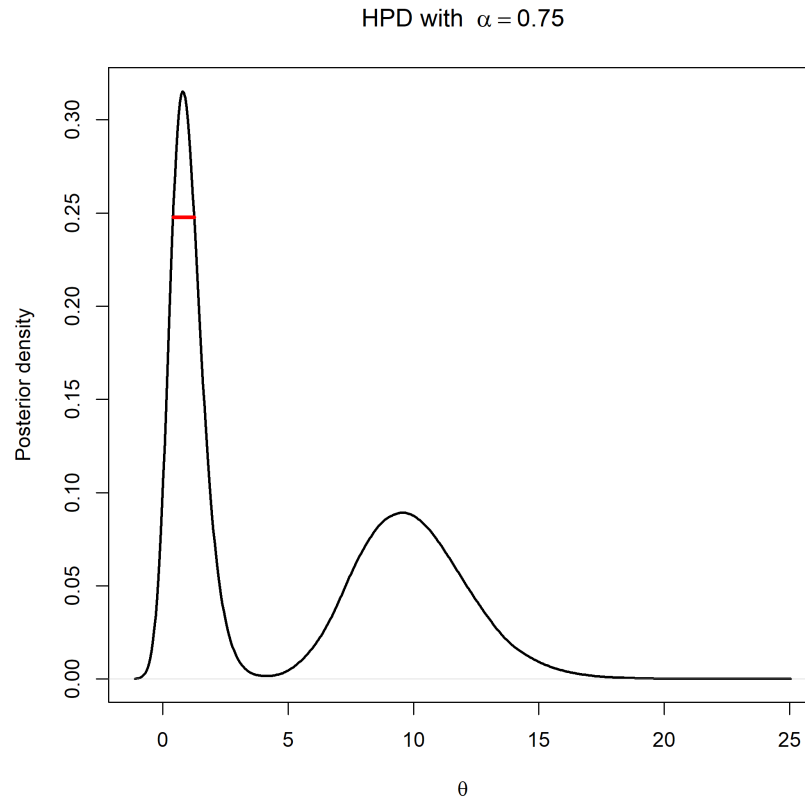
# Highest Posterior Density (HPD) region

1.  $p(\theta \in s(y) \mid Y = y) = 1 - \alpha$
2. If  $\theta_a \in s(y)$ , and  $\theta_b \notin s(y)$ , then  $p(\theta_a \mid Y = y) > p(\theta_b \mid Y = y)$ .
  - All points in an HPD region have a higher posterior density than points out- side the region.

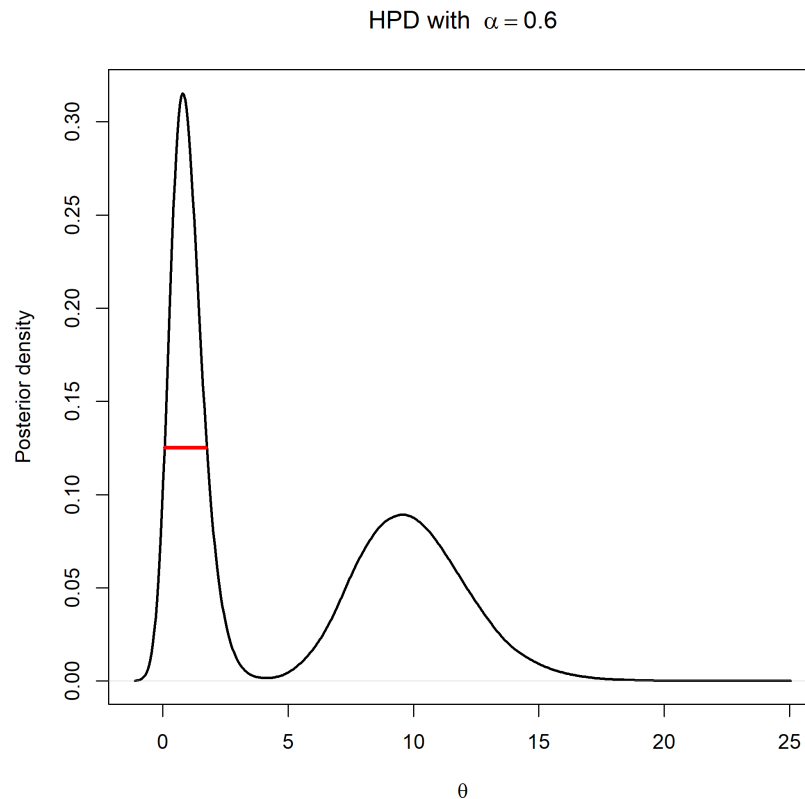
The HPD region is the *smallest* region with probability  $(1 - \alpha)\%$

# Highest Posterior Density (HPD) region

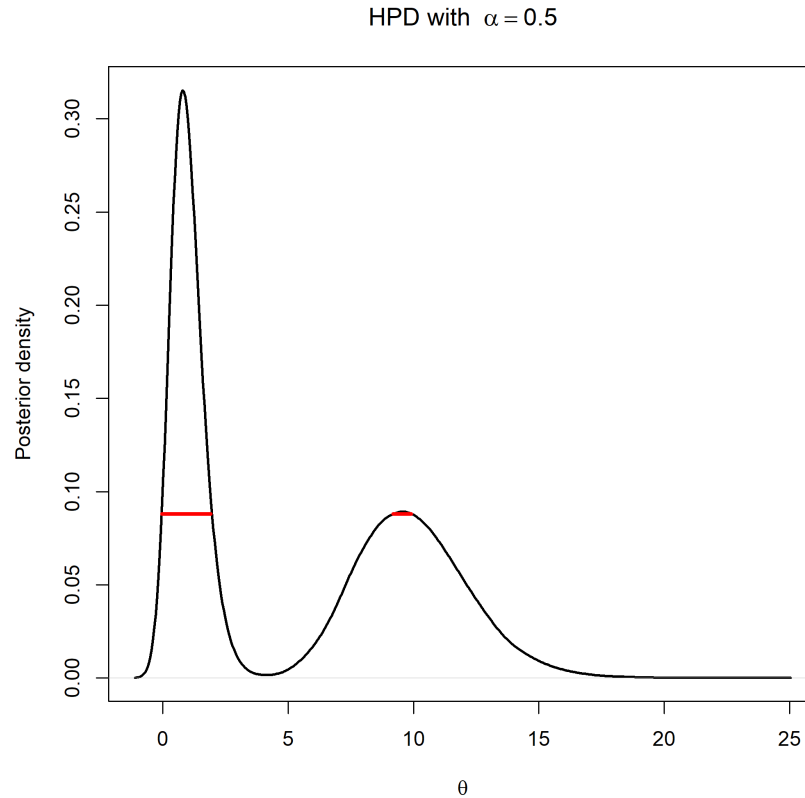
## Warning: package 'HDInterval' was built under R version 4.2.3



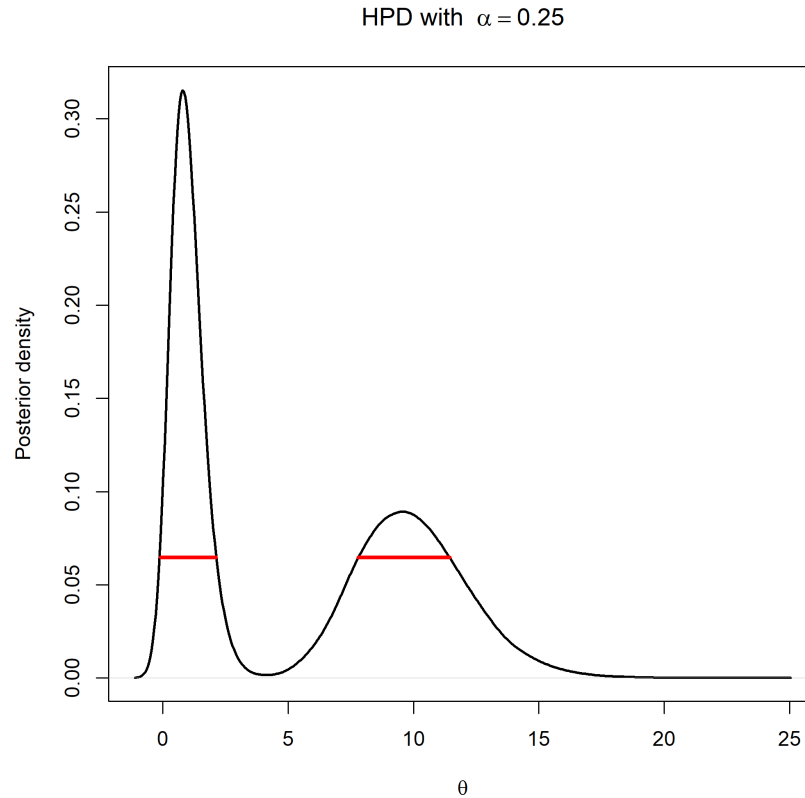
# Highest Posterior Density (HPD) region



# Highest Posterior Density (HPD) region

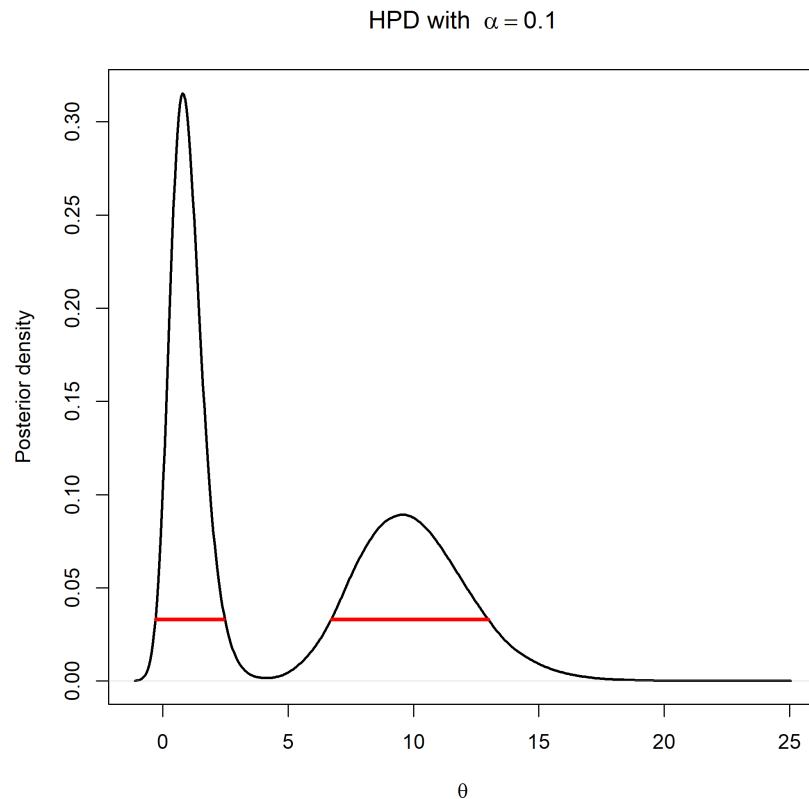


# Highest Posterior Density (HPD) region

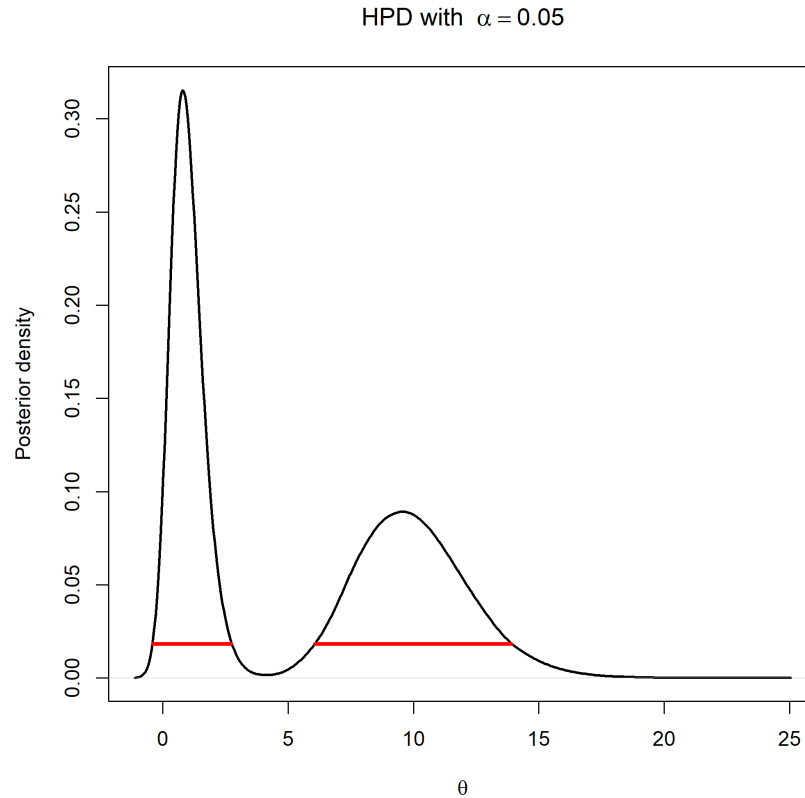




# Highest Posterior Density (HPD) region



# Highest Posterior Density (HPD) region



# Calibration: Frequentist Behavior of Bayesian Intervals

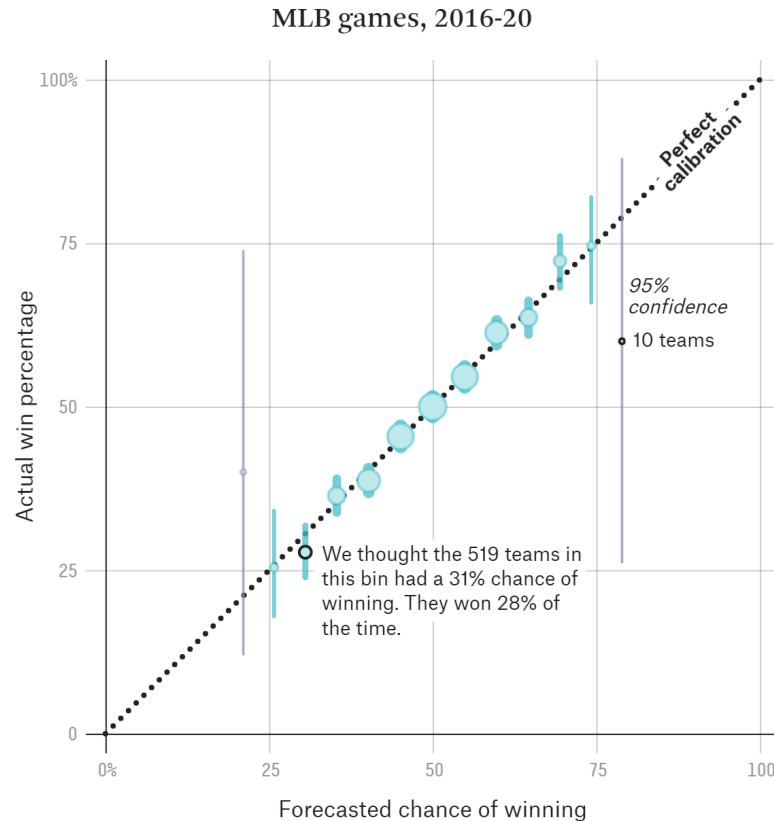
- A credible interval is calibrated if it has the right frequentist coverage
- Bayesian credible intervals usually won't have correct coverage
- If our prior was well-calibrated and the sampling model was correct, we'd have well-calibrated credible intervals
- Specifying *nearly* calibrated prior distributions is hard!

# Calibration of MLB predictions

Indeed, single predictions are hard to judge on their own. So let's group every MLB game prediction (not just those from September 2018) into bins — for example, we'll throw every prediction that gave a team between a 37.5 percent and 42.5 percent chance of winning into the same “40 percent” group — and then plot the averages of each bin's forecasted chances of winning against their actual win percentage. If our forecast is **well-calibrated** — that is, if events happened roughly as often as we predicted over the long run — then all the bins on the calibration plot will be close to the 45-degree line; if our forecast was **poorly calibrated**, the bins will be further away.

source: <https://projects.fivethirtyeight.com/checking-our-work/>

# Calibration of MLB predictions



source: <https://projects.fivethirtyeight.com/checking-our-work/>

# Calibrated probability intervals

- Calibration is important but only part of the story!
- Want well calibrated but *small* intervals (big intervals tell us nothing)

# Calibrated probability intervals

- Calibration is important but only part of the story!
- Want well calibrated but *small* intervals (big intervals tell us nothing)
- Domain expertise helps us develop smaller prior distributions (calibration?)
  - Usually at the cost of calibration
  - My experience: people tend to be overconfident
  - Alternatives to domain expertise?

# Subjective Bayesianism

- So far we have focused on defining priors using domain expertise
- "Subjective" Bayes
  - Essentially what we have discussed so far
  - Priors usually represent subjective judgements can't always be rigorously justified
- Alternative: "objective" Bayes



# Objective Bayesianism

- Is there a way to define "objective" prior distributions?
  - Good default prior distributions for some problems?
  - "Non-informative" prior distributions?
- Also called "reference" or "default" priors

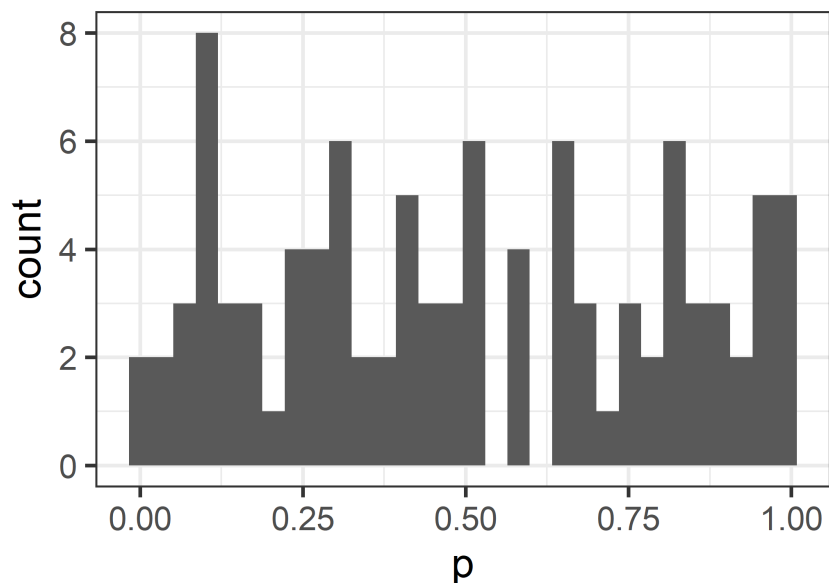
# Objective Bayesianism

- Is there a way to define "objective" prior distributions?
  - Good default prior distributions for some problems?
  - "Non-informative" prior distributions?
- Also called "reference" or "default" priors
- Can we find prior distributions that lead to (approximately) correct frequentist calibration?
- Can we find prior distributions which minimize the amount of information contained in the distribution?
  - Principle of maximum entropy (MAXENT).

# Difficulties with non-informative priors

Uniform distribution for  $p$

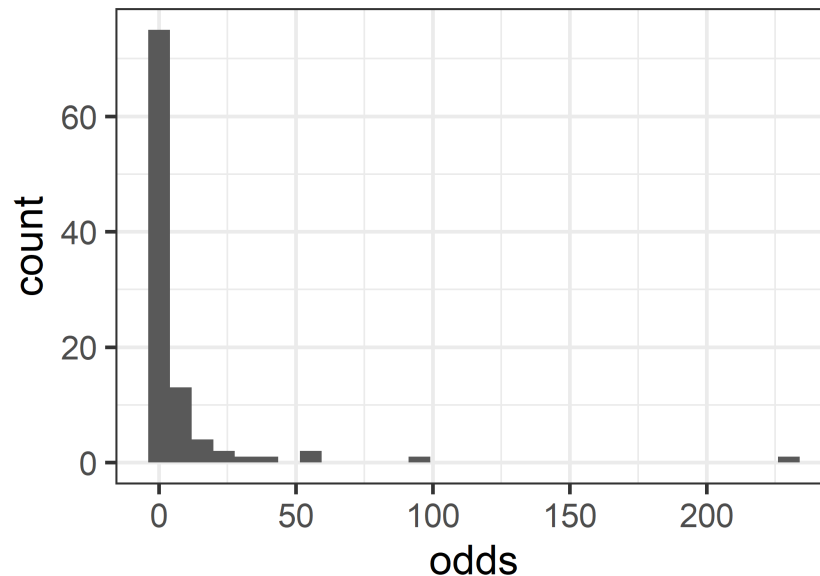
```
p <- runif(100)
tibble(p=p) %>% ggplot() +
  geom_histogram(aes(x=p), bins=30) +
  theme_bw(base_size=24)
```



# Difficulties with non-informative priors

Implied distribution for odds =  $p/(1-p)$

```
odds <- p/(1-p)
tibble(odds=odds) %>% ggplot() +
  geom_histogram(aes(x=odds)) +
  theme_bw(base_size=24)
```



# 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

- In the Poisson case, let us pretend the prior for  $\lambda$  is proper...

$$p(\lambda \mid y) \propto$$

# 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

# Posterior Predictive Distributions



# 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  $y_1, \dots, y_n$  the observed data.
  - The posterior predictive distribution is  $p(\tilde{y} \mid y_1, \dots, y_n)$

# 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  $y_1, \dots, y_n$  the observed data.
  - The posterior predictive distribution is  $p(\tilde{y} \mid y_1, \dots, y_n)$
- 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?

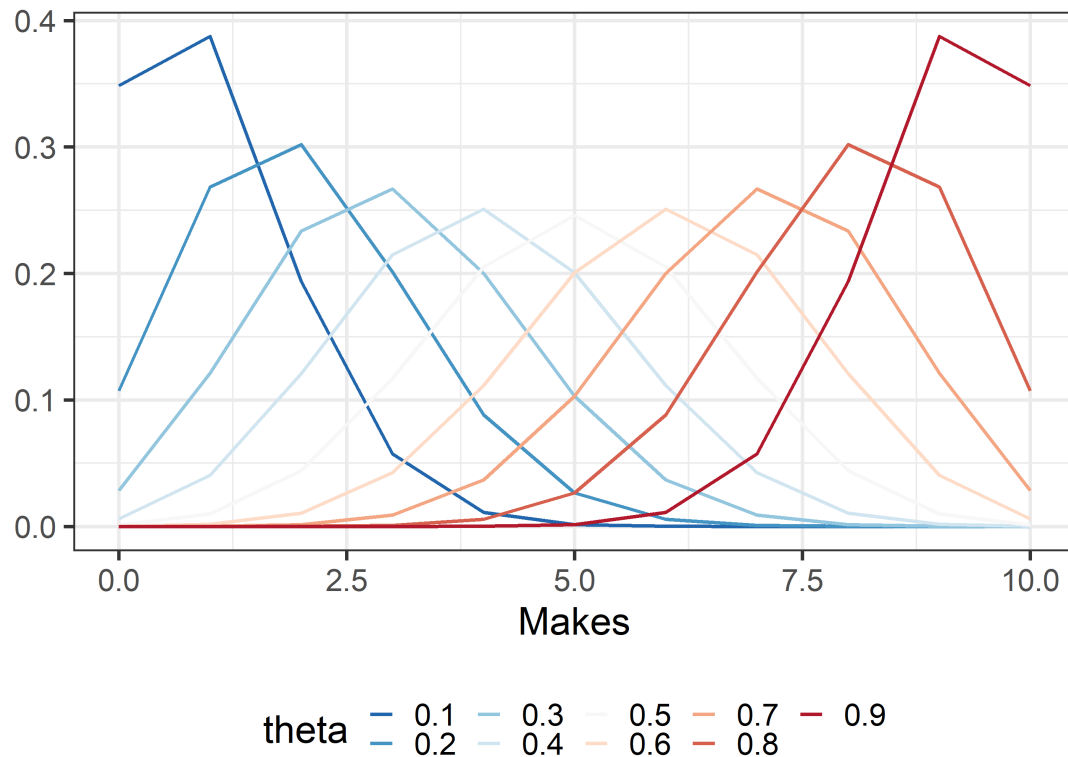
## Another Basketball Example

- 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  $\theta$ , then we know the distribution over future attempts:

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

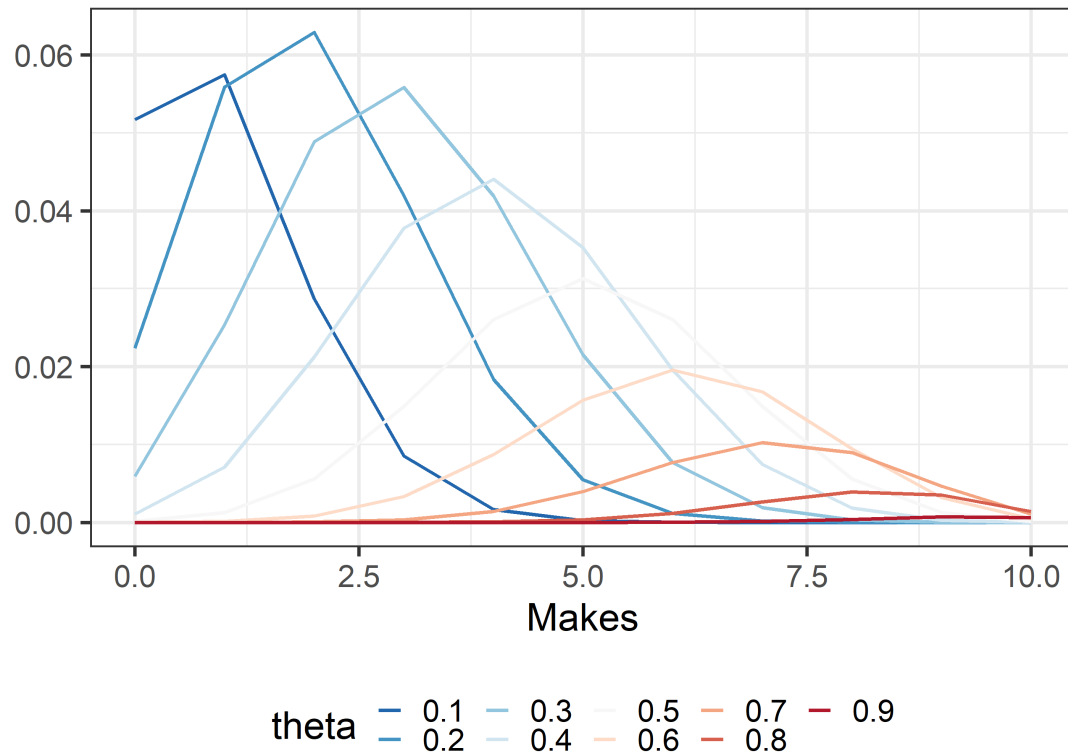


# Posterior Prediction

- We already observed 1 make out of 2 tries.
- Assume a  $\text{Beta}(1, 3)$  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  $\text{Beta}(2, 4)$
- Intuition: weight  $\tilde{Y} \sim \text{Bin}(10, \theta)$  by  $p(\theta \mid Y = 1, n = 2)$

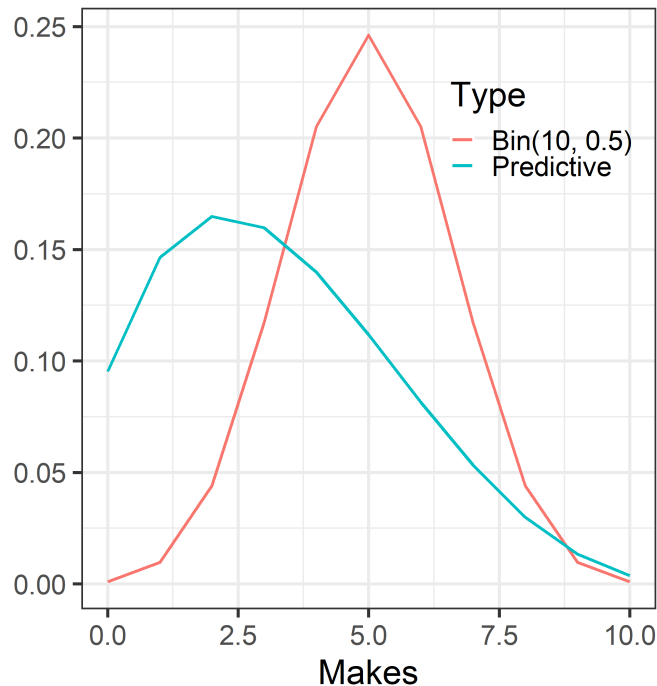
# Posterior Prediction

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



# Posterior Predictive Distribution

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



The predictive density,  $p(\tilde{y} | y)$ , answers the question "if I take 10 more shots how many will I make, given that I already made 1 of 2".

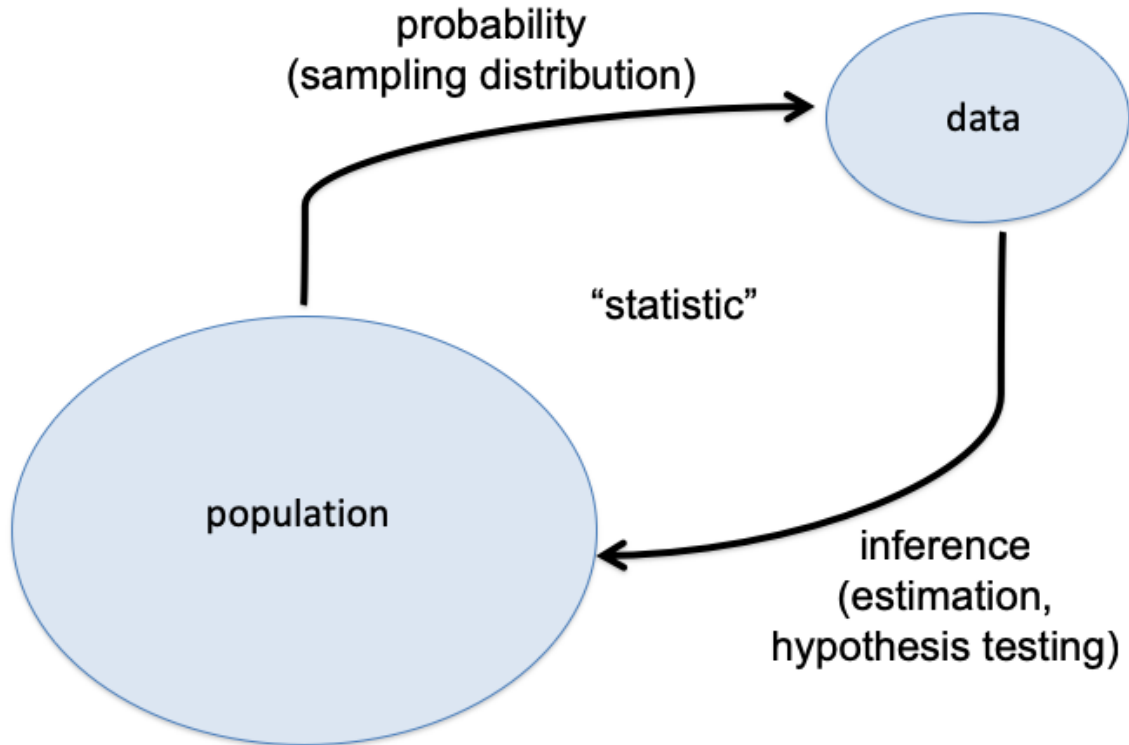
# The posterior predictive distribution

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

- 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,  $p(\theta \mid y_1, \dots, y_n)$



# 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

# 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,  $p(\theta)$

# Homework 1

- $\lambda \sim \text{Gamma}(\alpha, \beta)$
- $\tilde{Y} \sim \text{Pois}(\lambda)$
- $p(\tilde{y}) = \int p(\tilde{y} \mid \lambda)p(\lambda)d\lambda$  is a prior predictive distribution!
- "A Gamma-Poisson mixture is a Negative-Binomial Distribution"

# Homework 1

$$\begin{aligned} p(\tilde{y}) &= \int p(\tilde{y} \mid \lambda) p(\lambda) d\lambda \\ &= \int \left( \frac{\lambda^{\tilde{y}}}{y!} e^{-\lambda} \right) \left( \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{(\alpha-1)} e^{-\beta\lambda} \right) d\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha) y!} \int (\lambda^{(\alpha+y-1)} e^{-(\beta+1)\lambda}) d\lambda \end{aligned}$$

$\int (\lambda^{(\alpha+y-1)} e^{-(\beta+1)\lambda}) d\lambda$  looks like an unnormalized Gamma( $\alpha + y, \beta + 1$ )

# Summary

- Bayesian credible intervals
  - Posterior probability that the value falls in the interval
  - Still strive for well-calibrated intervals (in the frequentist sense)
- Non-informative prior distributions
- Posterior predictive distributions
  - Estimated distribution for new data our uncertainty about the parameters