Lecture 4: Intervals and Predictive Distributions

Professor Alexander Franks

2023-01-24

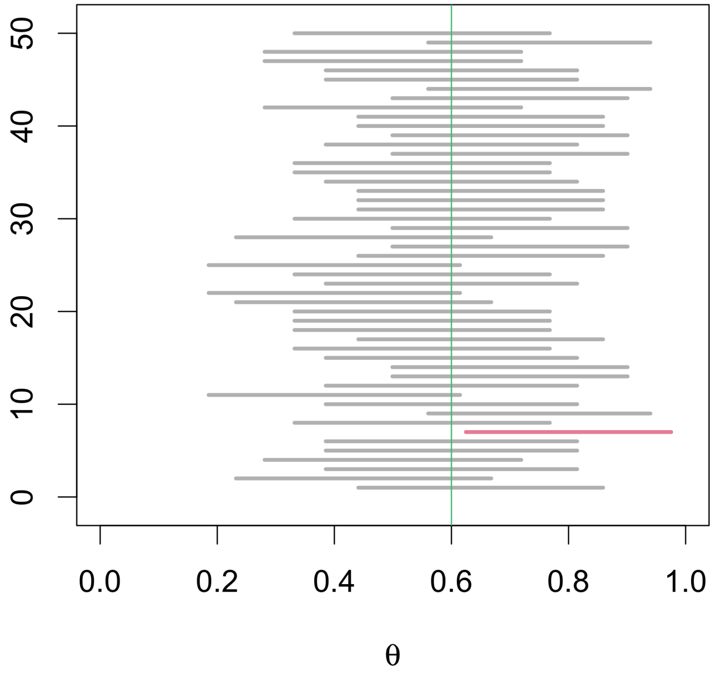
Announcements

- Reading: Chapter 8.1 (intervals), 8.3 (posterior prediction)
- Homework 2 due February 5, at 11:59pm

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 x 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 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 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 θ 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 θ is in the HPD region is $1 - \alpha$
2. If $\theta_a \in R(y),$ and $\theta_b \notin R(y),$ then $p\left(\theta_a | Y = y\right) > p\left(\theta_b | Y = y\right)$
    - All points in an HPD region have a higher posterior density than points out- side 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 $(1 - \alpha)\%$

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 political predictions

The best test of a probabilistic forecast is whether it's well calibrated. By that I mean: Out of all FiveThirtyEight forecasts that give candidates about a 75 percent shot of winning, do the candidates in fact win about 75 percent of the time over the long run? It's a problem if these candidates win only 55 percent of the time. But from a statistical standpoint, it's just as much of a problem if they win 95 percent of the time.

source: fivethirtyeight.com

Calibration of political predictions

Calibration for FiveThirtyEight "polls-plus" forecast

WIN PROBABILITY RANGE	NO. FORECASTS	EXPECTED NO. WINNERS	ACTUAL NO. WINNERS
95-100%	27	26.7	26
75-94%	15	13.1	14
50-74%	14	8.7	11
25-49%	13	4.8	3
5-24%	27	3.1	1
0-4%	88	0.8	1

source: https://fivethirtyeight.com/features/when-we-say-70-percent-it-really-means-70-percent/

The age guessing game*

*Bayesian edition

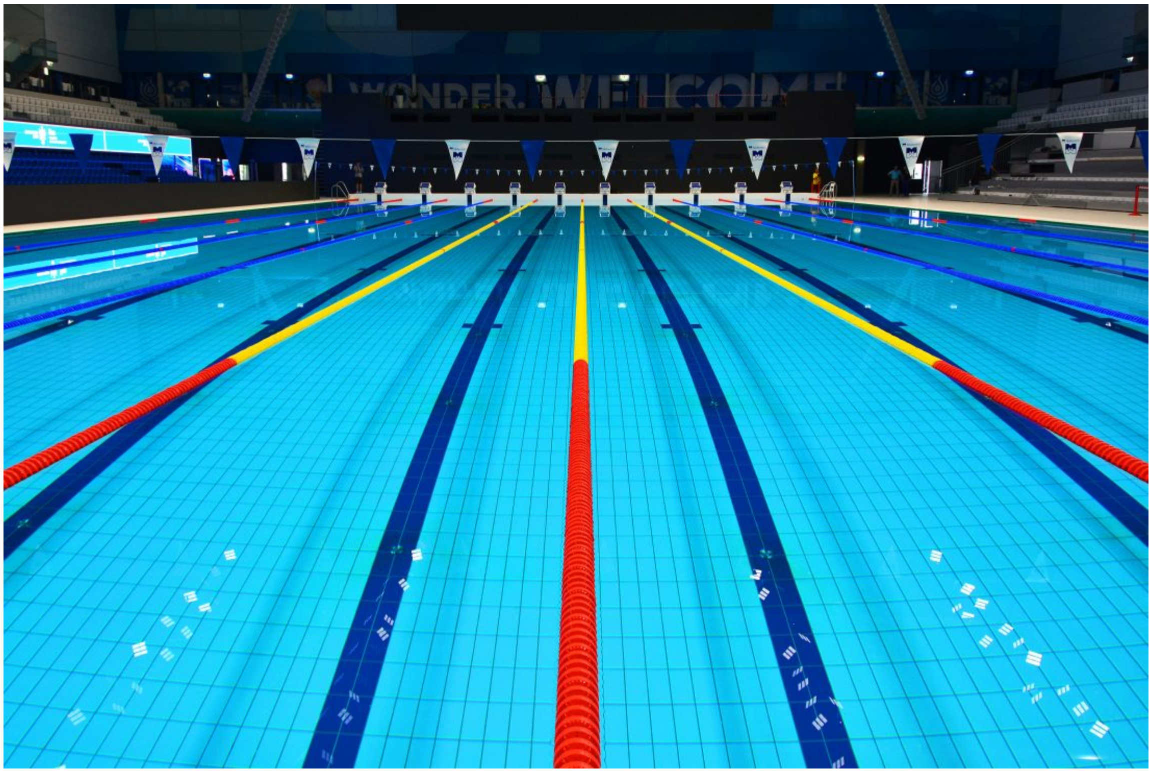
Interval Trivia

- I'm going to ask you ten questions about random facts
- For each, write down a 50% credible interval for *your* belief about the answer
- Goal:
  - Be well calibrated. 50% of your intervals should contain the true answer.

Percentage in California

What fraction of the US population is living in California?

Olympic Swimming Pool



How many gallons of water are there in an olympic swimming pool?

Household Income

What is the median household income in the US (dollars)?

Grooves on a quarter

How may grooves are there on the edge of quarter?

Gold in Fort Knox

How many pounds of gold are there Fort Knox?

Population of Australia

What is the population of Australia (in millions)?

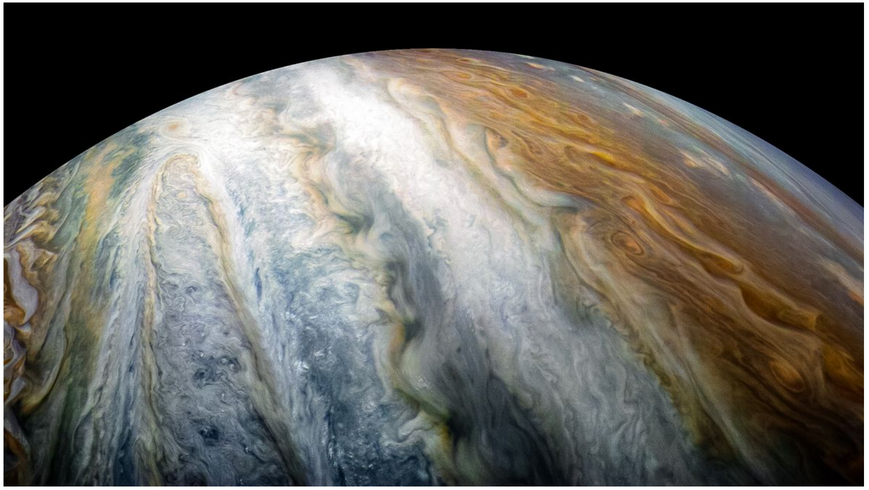
Tallest tsunami wave ever recorded

How tall was the tallest tsunami wave ever recorded (in feet)?

Disney world

In what year did Disneyland Park open to the public?

Jupiter



How many times larger in volume is Jupiter than Earth?

Netflix

In what year was Netflix founded?

All questions

N	Question	lower	upper
Q1	What fraction of the US population is living in California?	______	______
Q2	How many gallons of water are there in an olympic swimming pool?	______	______
Q3	What is the median household income in the US (dollars)?	______	______
Q4	How may grooves are there on the edge of quarter?	______	______
Q5	How many pounds of gold are there fort knox?	______	______
Q6	What is the population of australia (in millions)?	______	______
Q7	How tall was the tallest tsunami wave ever recorded (in feet)?	______	______
Q8	In what year did Disneyland open to the public?	______	______
Q9	How many times larger in volume, is Jupiter than Earth?	______	______
Q10	In what year was Netflix founded?	______	______

Answers

N	Question	Answers
Q1	What fraction of the US population is living in California?	12%
Q2	How many gallons of water are there in an olympic swimming pool?	660,000
Q3	What is the median household income in the US?	$59,039
Q4	How may grooves are there on the edge of quarter?	119
Q5	How many pounds of gold are there fort knox?	9,206,250
Q6	What is the population of australia in millions?	24.13 mil
Q7	How tall was the tallest tsunami wave ever recorded?	1720 feet
Q8	In what year did Disneyland open to the public?	1955
Q9	How many times larger in volume, is Jupiter than Earth?	1,321.33
Q10	In what year was Netflix founded?	1997

Calibrated probability intervals

- Calibration is important but only part of the story!
- Want well calibrated but *small* intervals (big intervals tell us nothing)
- How you could have gotten a perfect score on the quiz:
    - For 5 of the answers select [-1 Trillion, +1 Trill] (ensures it will cover)
    - For the other 5 answers select[-0.01, + 0.01] (ensures it won't cover)

Calibrated probability intervals

- Calibration is important but only part of the story!
- Want well calibrated but *small* intervals (big intervals tell us nothing)
- How you could have gotten a perfect score on the quiz:
  - For 5 of the answers select [-1 Trillion, +1 Trill] (ensures it will cover)
  - For the other 5 answers select[-0.01, + 0.01] (ensures it won't cover)
- 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

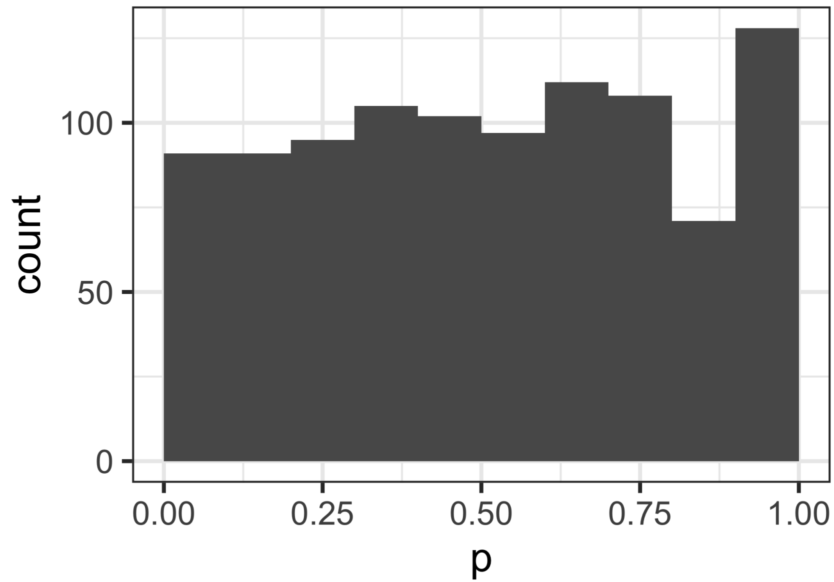
Objective Bayesianism

- 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

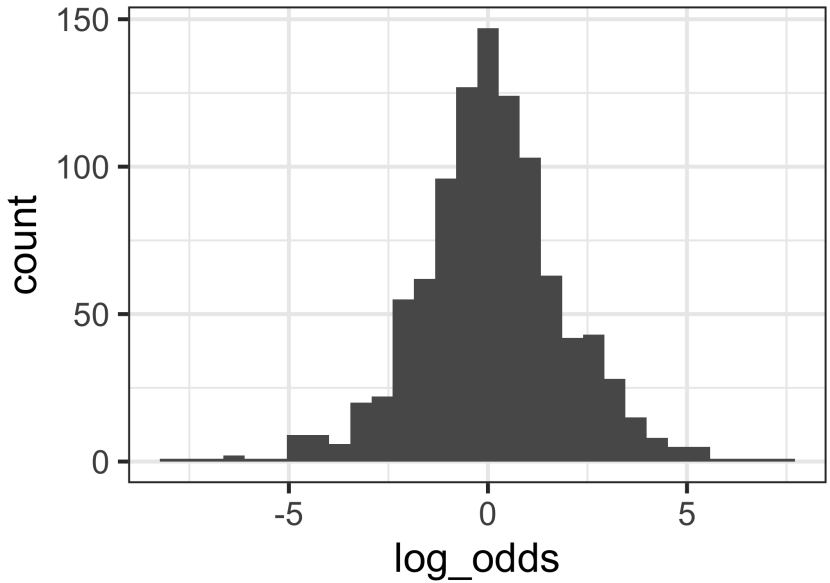
```
p <- runif(1000)
tibble(p=p) %>% ggplot() +
  geom_histogram(aes(x=p), boundary=0.5, binwidth=0.1) +
  theme_bw(base_size=24)
```



Difficulties with non-informative priors

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

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



Difficulties with non-informative priors

Improper prior distributions

- For the Beta distribution we chose a uniform prior, where $p(\theta) \propto \text{const}$. 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

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

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

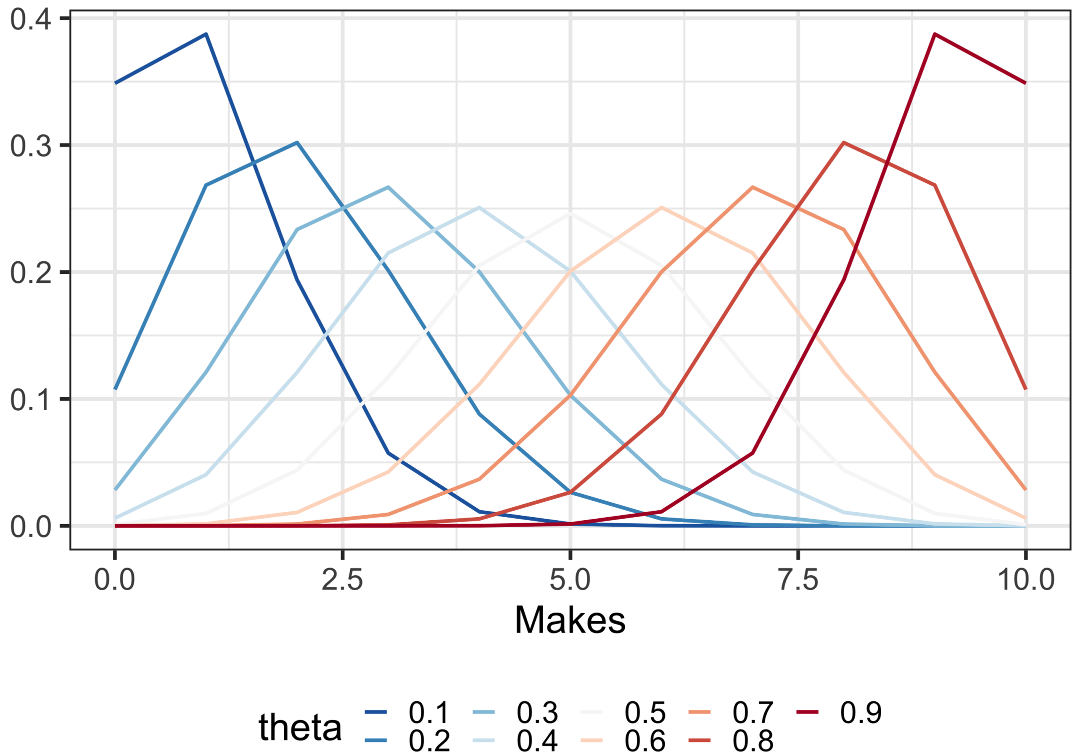
Another Basketball Example

- 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 θ, then we know the distribution over future attempts:

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

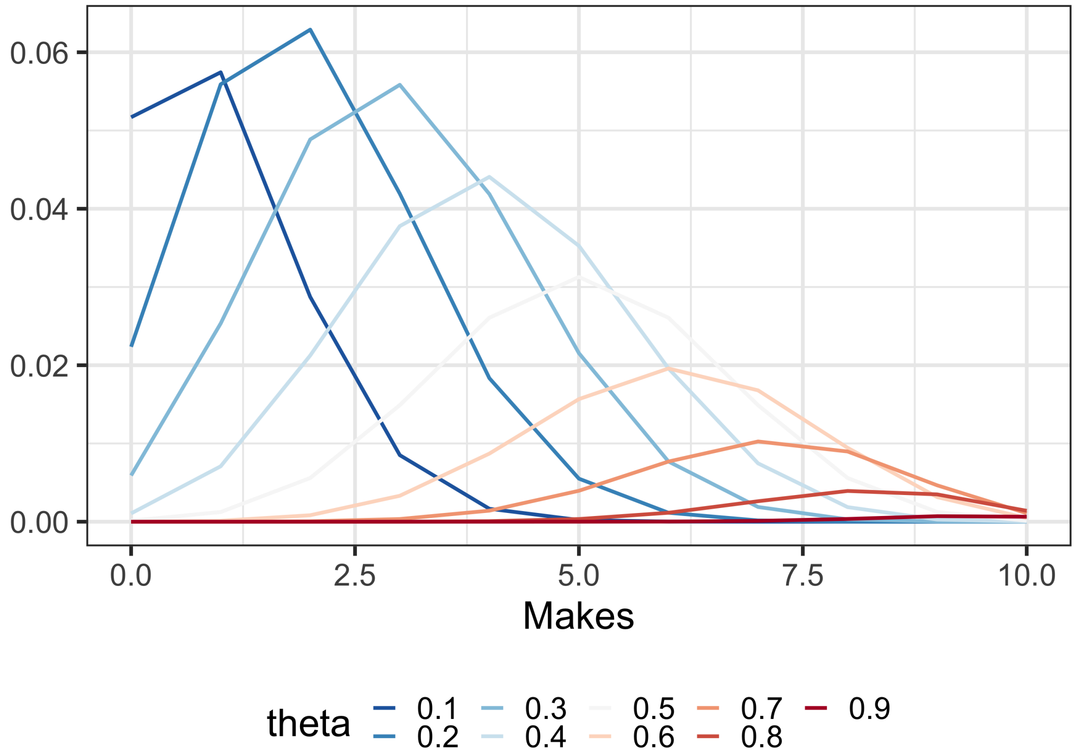


Posterior Prediction

- We already observed 1 make out of 2 tries.
- Assume a $\text{Beta}(1, 3)$ 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

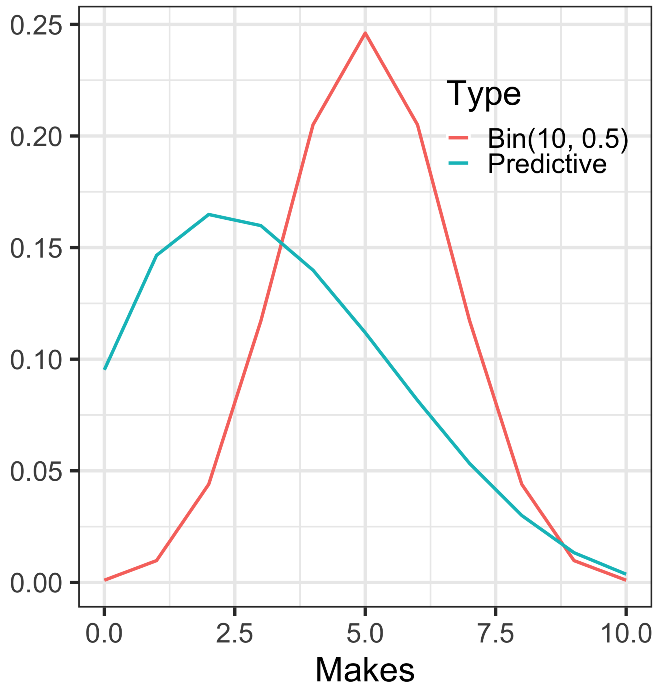
If I take 10 more shots how many will I make?



Posterior predictive distribution

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 1 of 2".

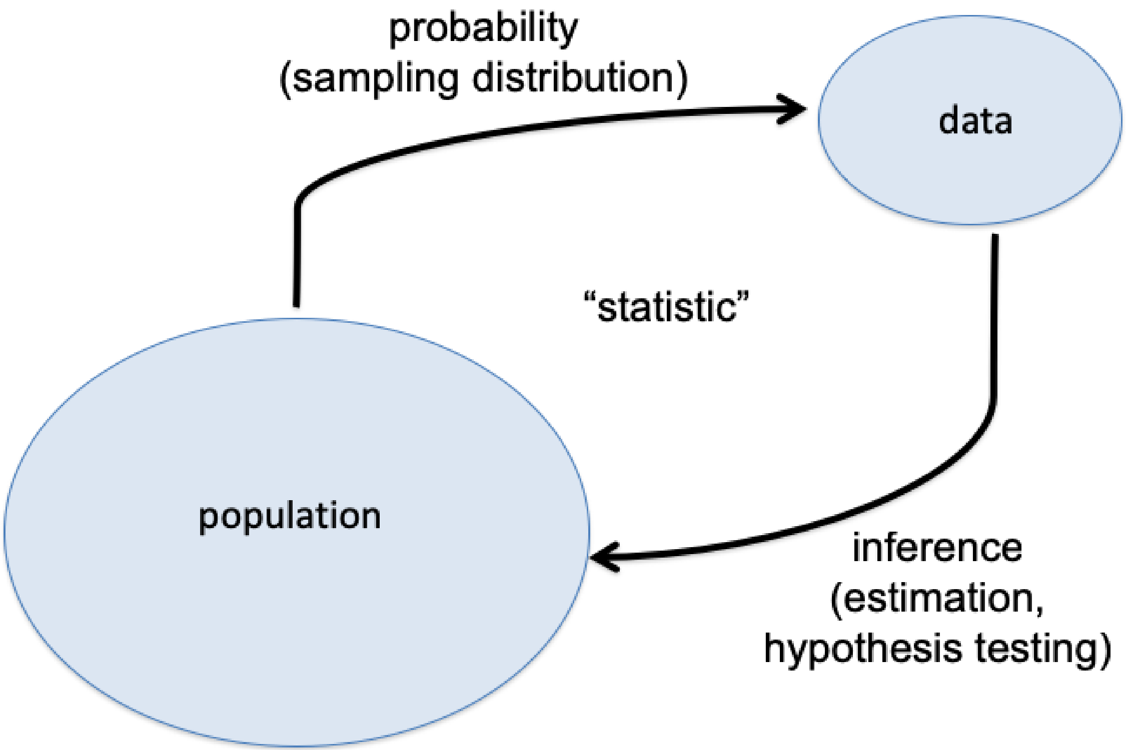
The posterior predictive distribution

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

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

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 Extra Credit

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

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