

16: The Beta Random Variable

Jerry Cain
May 5, 2025

Starting Today!

Today we are going to learn something unintuitive,
beautiful, and incredibly powerful!

We are going to think of probabilities themselves as
random variables.

Enter: New Definition of Probability

Flip a coin $n + m$ times, produce n heads.

We **do not** know the **probability** θ that the coin comes up heads.

Note: We use θ instead of p , because p looks like a constant, and we're not talking about a constant here. We talking about a random variable that takes on some values between 0 and 1—as probabilities always do—with varying probability densities.



The world's first coin

Frequentist

θ is a single value p

$$p = \lim_{n+m \rightarrow \infty} \frac{n}{n+m} \approx \frac{n}{n+m}$$

Bayesian

θ is a **random variable**.

θ 's support: $(0, 1)$

Review: Bayesian Probability

Bayesian statistics: Probability represents our ever-evolving understanding of the world.

Mixing discrete and continuous random variables, combined with Bayes' Theorem, allows us to reason about **probabilities as random variables**.

Review: Bayes' Theorem in All Flavors

Let X, Y be **continuous** and M, N be **discrete** random variables.

Original Bayes:
$$p_{M|N}(m|n) = \frac{p_{N|M}(n|m)p_M(m)}{p_N(n)}$$

Mix Bayes #1:
$$f_{X|N}(x|n) = \frac{p_{N|X}(n|x)f_X(x)}{p_N(n)}$$

Mix Bayes #2:
$$p_{N|X}(n|x) = \frac{f_{X|N}(x|n)p_N(n)}{f_X(x)}$$

All continuous:
$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$$



Review: Mixing discrete and continuous

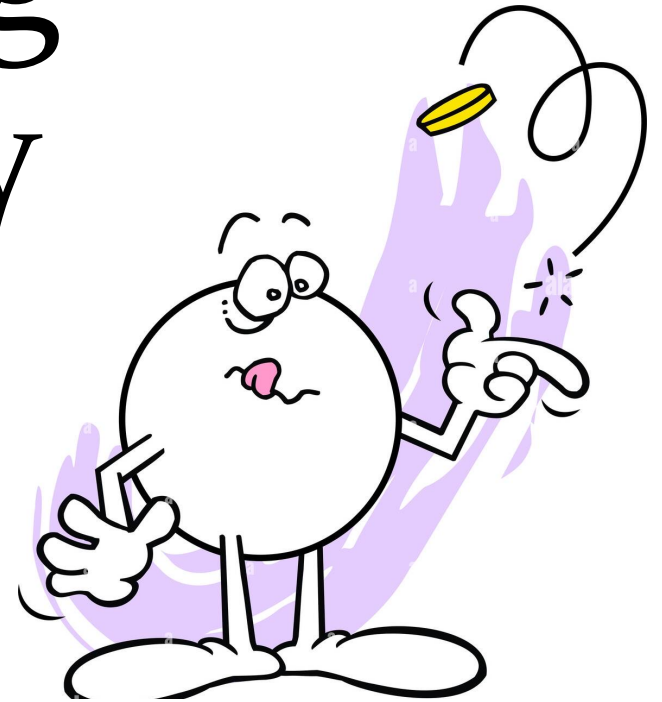
Let θ be a random variable for the probability your coin comes up heads and let N be the number of heads you observe in a series of coin flips.

$$\overset{\text{posterior}}{f_{\theta|N}(x|n)} = \frac{\overset{\text{likelihood}}{p_{N|\theta}(n|x)} \overset{\text{prior}}{f_{\theta}(x)}}{\underset{\text{normalization constant}}{p_N(n)}}$$

- **Prior** belief of parameter θ
- **Likelihood** of $N = n$ heads, given parameter $\theta = x$.
- **Posterior** updated belief of parameter θ .

$$\begin{aligned} f_{\theta}(x) \\ p_{N|\theta}(n|x) \\ f_{\theta|N}(x|n) \end{aligned}$$

Flipping Mystery Coins



Flipping Coins

Flip a coin $n + m$ times, observe n heads.

- Before flipping anything, assume θ (the probability that the coin comes up heads) is equally likely to be any probability in $(0, 1)$.
- Let N count the number of heads.
- For any $\theta = x$, assume coin flips are independent.

What is our updated belief of θ after we observe $N = n$?

What are reasonable distributions for each of the following?

1. θ Bayesian prior $\theta \sim \text{Uni}(0, 1)$
2. $N|\theta = x$ Likelihood $N|\theta = x \sim \text{Bin}(n + m, x)$
3. $\theta|N = n$ Bayesian posterior, using Bayes' Theorem.

We may not know what x is, but whatever it is, it's the probability any single Bernoulli trial comes up heads.



Flipping Coins

Flip a coin $n + m$ times, observe n heads.

- Before flipping anything, assume θ (the probability that the coin comes up heads) is equally likely to be any probability in $(0, 1)$.
- Let N count the number of heads.
- For any $\theta = x$, assume coin flips are independent.

Prior:
 $\theta \sim \text{Uni}(0,1)$

Likelihood:
 $N|\theta = x \sim \text{Bin}(n + m, x)$

What is our updated belief of θ after we observe $N = n$?

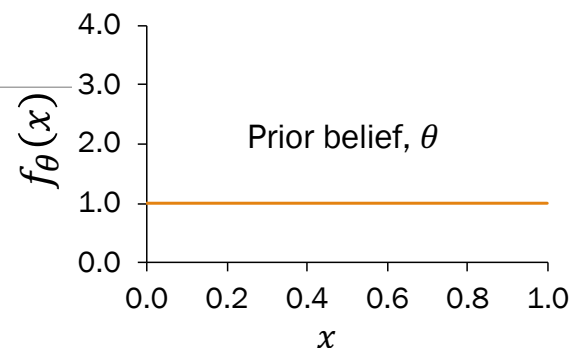
Posterior: $f_{\theta|N}(\theta|n)$

$$f_{\theta|N}(x|n) = \frac{p_{N|\theta}(n|x)f_{\theta}(x)}{p_N(n)} = \frac{\binom{n+m}{n} x^n (1-x)^m \cdot 1}{p_N(n)}$$
$$= \underbrace{\frac{\binom{n+m}{n}}{p_N(n)}}_{\text{constant with respect to } x} x^n (1-x)^m = \frac{1}{c} x^n (1-x)^m, \text{ where } c = \int_0^1 x^n (1-x)^m dx$$

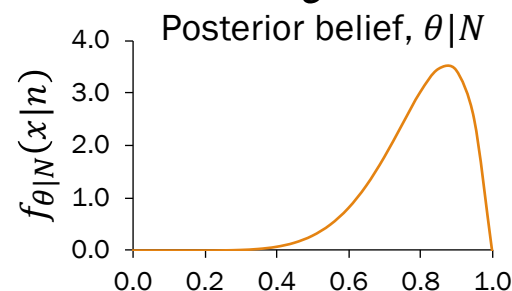
constant with respect to x ,
doesn't depend on x

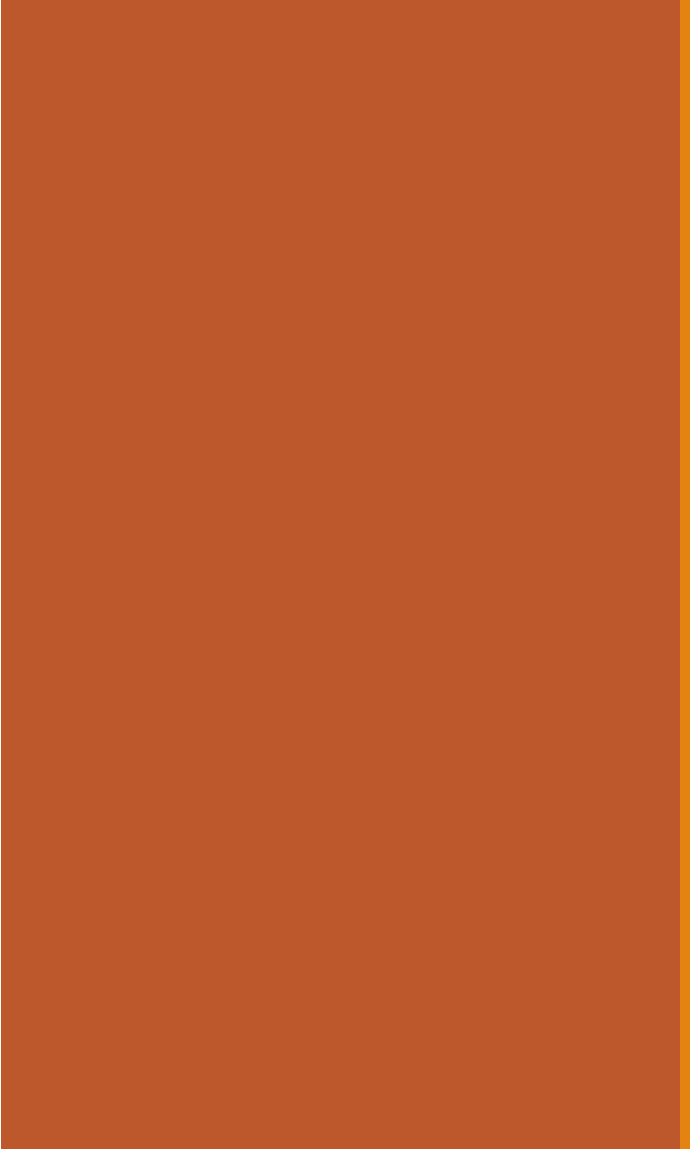
Flipping Coins

1. Start with a $\theta \sim \text{Uni}(0, 1)$ over probability that a coin lands heads.
2. Flip a coin 8 times. Observe $n = 7$ heads and $m = 1$ tail
3. What is our posterior belief of the probability θ ?



$$f_{\theta|N}(x|n) = \frac{1}{c} x^7 (1 - x)^1$$





Beta Random Variables

Introducing: Beta RV as $\text{Beta}(a, b)$

def A **Beta** random variable X is defined as follows:

$$\begin{aligned} X &\sim \text{Beta}(a, b) & \text{PDF } f(x) &= \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1} \\ a &> 0, b > 0 & \text{where } B(a, b) &= \int_0^1 x^{a-1} (1-x)^{b-1} dx, \text{ normalizing constant} \\ \text{Support of } X &: (0, 1) & \text{Cool fact: When } a \text{ and } b \text{ are integers,} & \\ E[X] &= \frac{a}{a+b} \quad \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)} & B(a, b) &= \frac{(a-1)!(b-1)!}{(a+b-1)!} \end{aligned}$$

Introducing: Beta RV as $\text{Beta}(a, b)$

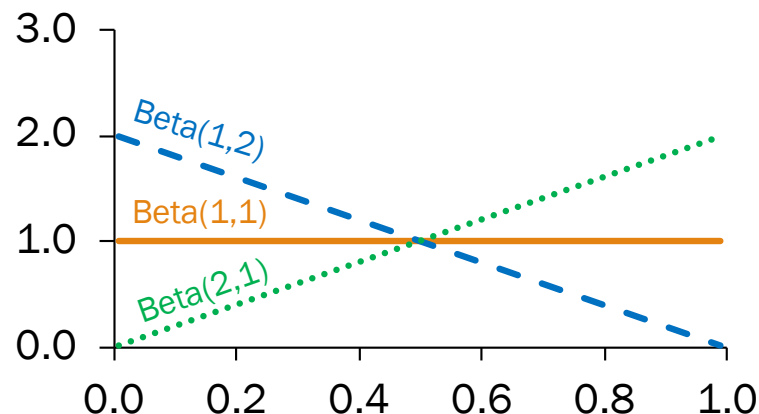
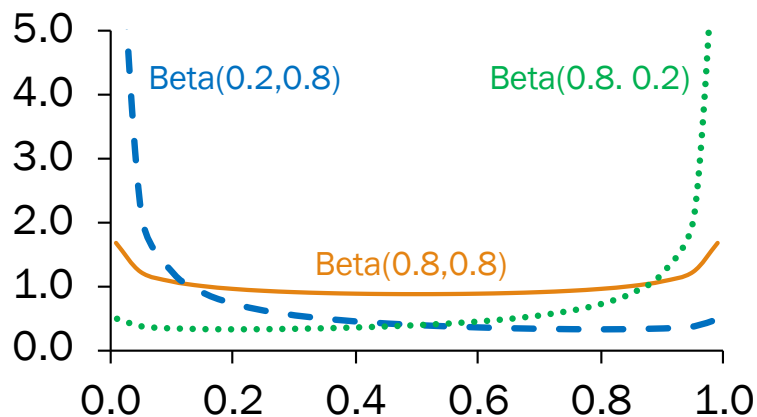
$$X \sim \text{Beta}(a, b)$$

$$a > 0, b > 0$$

Support of X : $(0, 1)$

$$\text{PDF } f(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}$$

where $B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$, normalizing constant

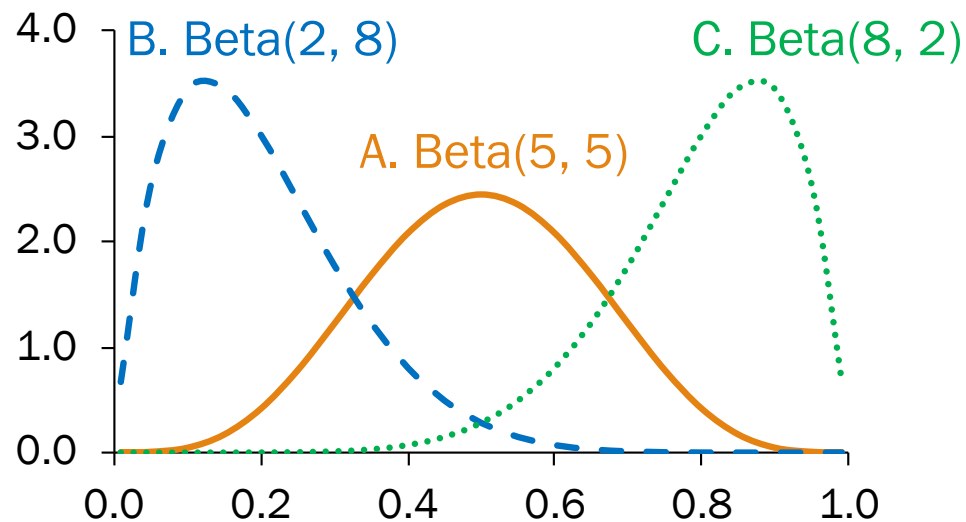


+ a third case
(next slide)

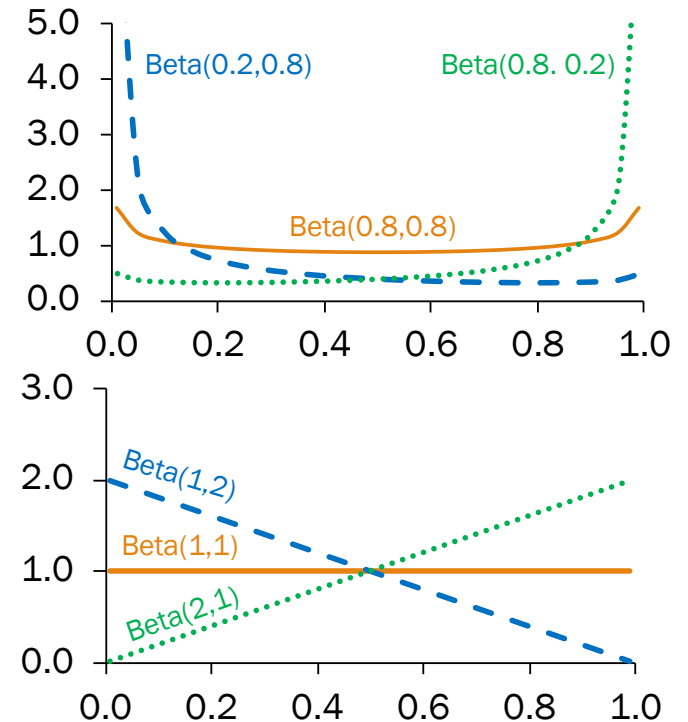
Note: PDF symmetric when $a = b$

Introducing: Beta RV as $\text{Beta}(a, b)$

Match PDF to distribution:



- A. Beta(5, 5)
- B. Beta(2, 8)
- C. Beta(8, 2)



In CS109, we focus on Beta functions where a, b are **both positive integers**.

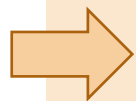


Introducing: Beta RV as PDF Modeling a Probability

def A **Beta** random variable X is defined as follows:

$$X \sim \text{Beta}(a, b)$$

$$a > 0, b > 0$$



Support of X : $(0, 1)$

$$\text{PDF } f(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}$$

where $B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$, normalizing constant

$$\text{Expectation } E[X] = \frac{a}{a+b}$$

$$\text{Variance } \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Beta can, in theory, be a distribution of probabilities.

Introducing: Beta RV as PDF Modeling a Probability

$$X \sim \text{Beta}(a, b)$$

$$a > 0, b > 0$$

Support of X : $(0, 1)$

PDF $f(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}$

where $B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$, normalizing constant



$$f_{\theta|N}(x|n) = \frac{1}{c} x^7 (1-x)^1$$

c normalizes to valid PDF

This is the PDF of Beta(8, 2)!

Flipping Coins: Evolution of θ ?

- Start with a $\theta \sim \text{Uni}(0,1)$ over probability
- Observe $n = 7$ successes and $m = 1$ failures
- Your new belief about the probability of θ is:

$$f_{\theta|N}(x|n) = \frac{1}{c} x^7 (1-x)^1, \text{ where } c = \int_0^1 x^7 (1-x)^1 dx$$

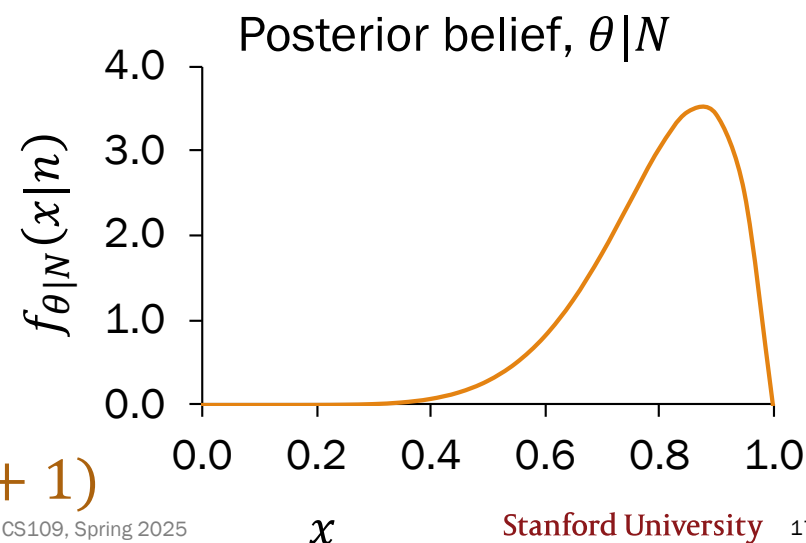
Posterior belief, $\theta|N$:

Beta($a = 8, b = 2$)

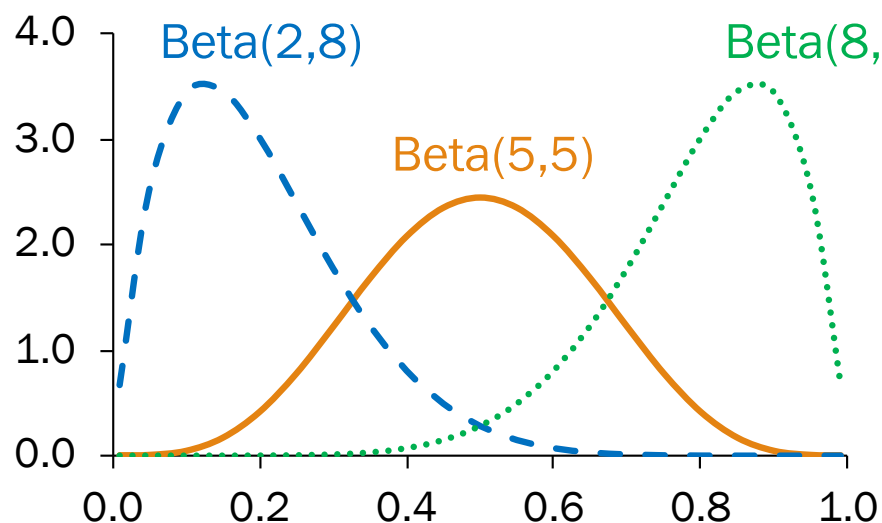
$$f_{\theta|N}(x|n) = \frac{1}{c} x^{8-1} (1-x)^{2-1}$$

More generally:

$$\theta|N \sim \text{Beta}(a = n + 1, b = m + 1)$$



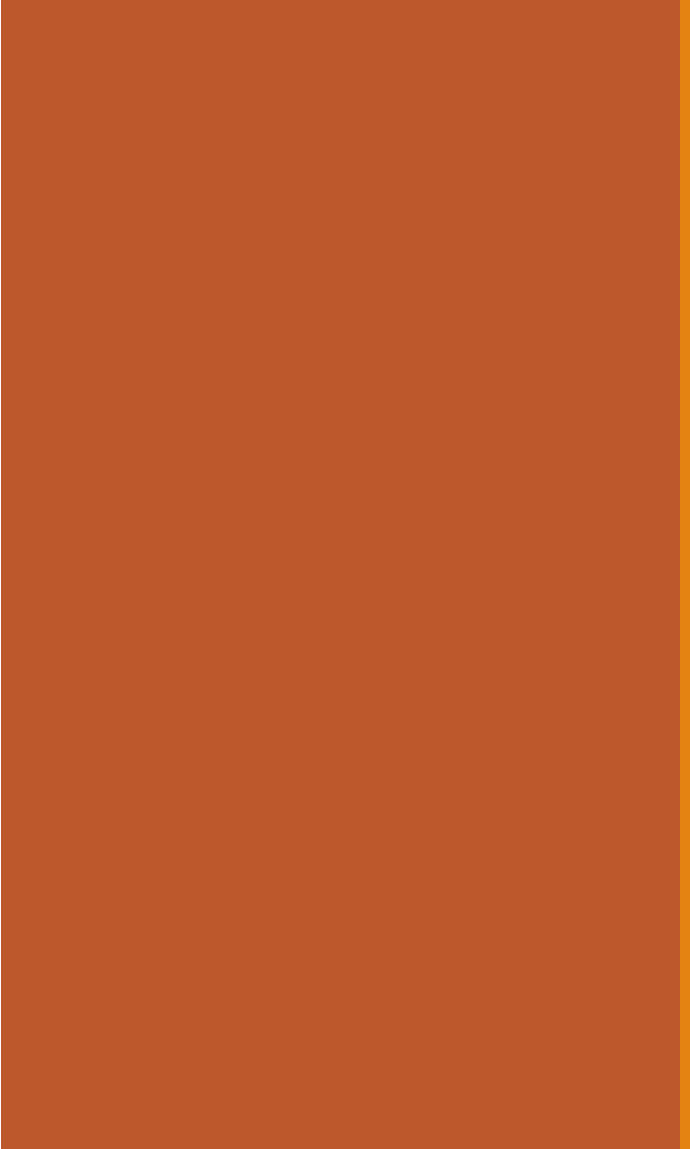
CS109 focus: Beta where a, b both positive integers



Beta parameters a, b are determined by the outcome of Bernoulli experiments.

$$a = \text{"successes"} + 1$$
$$b = \text{"failures"} + 1$$

- Beta can model the randomness of the probability of experiment success.
- Beta parameters depend on our data and our prior.

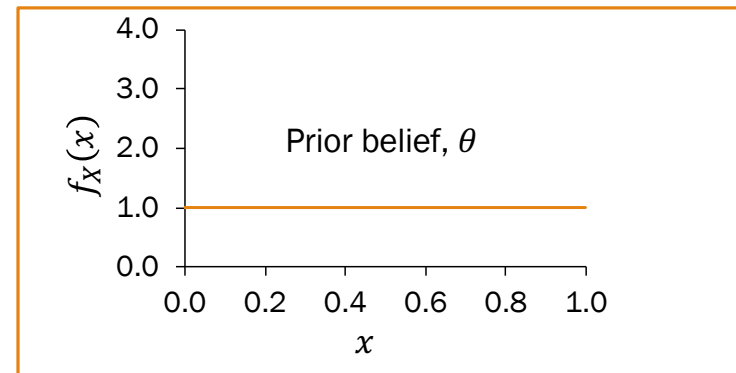
A large orange rectangle with a thin yellow border on the right side, positioned on the left side of the slide.

Conjugate distributions

The Uni(0, 1) Distribution

Wait a minute!

1. Start with a $\theta \sim \text{Uni}(0, 1)$ over probability that a coin lands heads.



okay 

2. Flip a coin 8 times. Observe $n = 7$ heads and $m = 1$ tail

3. What is our posterior belief of the probability θ ?

$$f_{\theta|N}(x|n) = \frac{1}{c} x^7 (1 - x)^1$$

c normalizes to valid PDF

Beta(8, 2)

The Uni(0, 1) Distribution

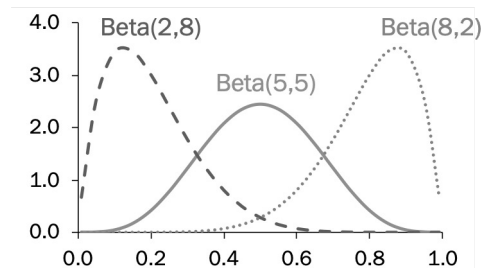
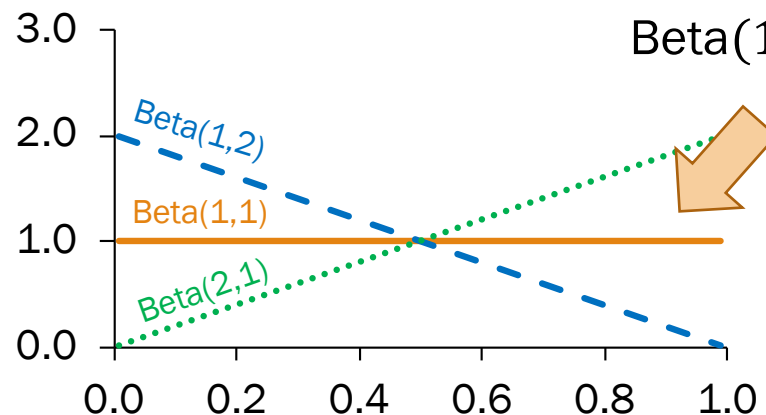
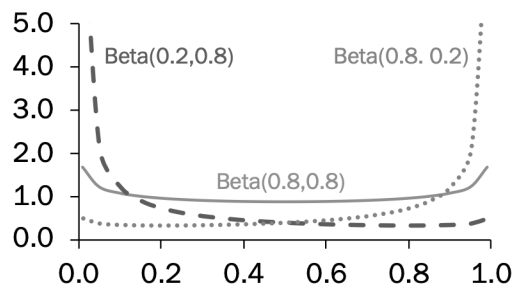
$X \sim \text{Beta}(a, b)$

$a > 0, b > 0$

Support of X : (0, 1)

PDF $f(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}$

where $B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$, normalizing constant



Note: PDF symmetric when $a = b$

The Uni(0, 1) Distribution is the Beta(1, 1) Distribution

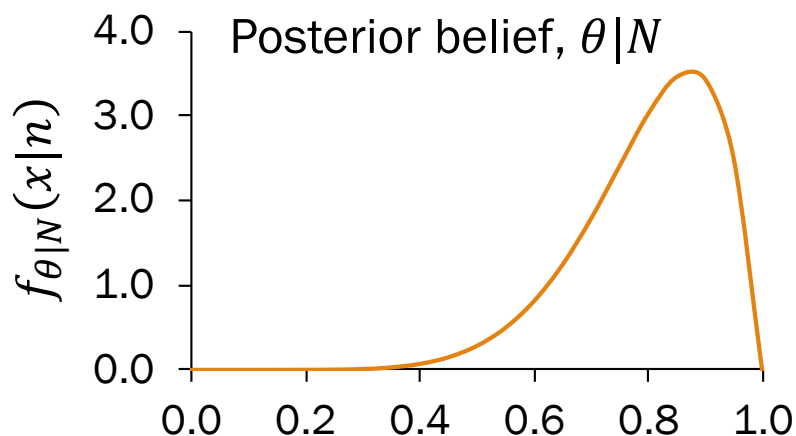
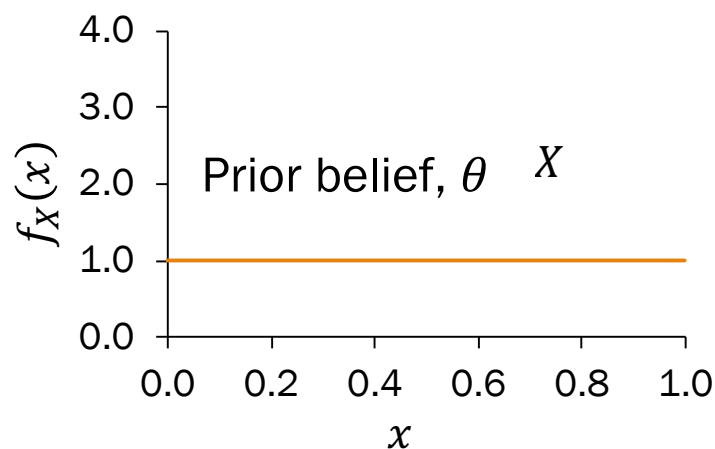
1. Start with a $\theta \sim \text{Uni}(0, 1)$ over probability that a coin lands heads.

Beta(1, 1)

2. Flip a coin 8 times. Observe $n = 7$ heads and $m = 1$ tail

3. What is our posterior belief of the probability θ ?

Beta(8, 2)



Beta ❤️ Binomial: Conjugate Distributions

Beta is a **conjugate distribution** for Binomial, meaning:

1. If our prior belief of the parameter is a Beta, and
2. Our experiment is a Binomial (assuming n successes, m failures), then
3. Our posterior is also Beta.

Proof: $\theta \sim \text{Beta}(a, b)$ $N | \theta = x \sim \text{Bin}(n + m, x)$

$$\begin{aligned} f_{\theta|N}(x|n) &= \frac{p_{N|\theta}(n|x)f_{\theta}(x)}{p_N(n)} = \frac{\binom{n+m}{n} x^n (1-x)^m \cdot \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1}}{p_N(n)} \\ &= \text{constants that don't depend on } x \cdot C \cdot x^n (1-x)^m \cdot x^{a-1} (1-x)^{b-1} \\ &= C \cdot x^{n+a-1} (1-x)^{m+b-1} \quad \checkmark \end{aligned}$$

Beta ♥ Binomial : Conjugate Distributions

Beta is a **conjugate distribution** for Binomial, meaning:

- Prior and posterior parametric forms are the same
- Practically speaking, "conjugate" means easy update to our belief distributions:
Add number of "heads" and "tails" seen to Beta parameters.

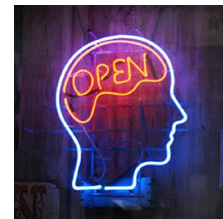
You **invent a prior** to express how biased you believe the coin is **a priori**:

- $\theta \sim \text{Beta}(a, b)$: pretend you've conducted $(a + b - 2)$ **imaginary trials**, where $(a - 1)$ trials produced a head and $(b - 1)$ produced a tail
- Choosing $\text{Beta}(1, 1) = \text{Uni}(0, 1)$ means you don't hold any prior beliefs

Prior $\text{Beta}(a = n_{\text{imag}} + 1, b = m_{\text{imag}} + 1)$

Experiment Observe n successes and m failures

Posterior $\text{Beta}(a = n_{\text{imag}} + n + 1, b = m_{\text{imag}} + m + 1)$



Medicinal Beta

- Before being tested, a medicine is believed to "work" 80% of the time.
- The medicine is administered to 20 patients.
- It "works" for 14, "doesn't work" for 6.

What is your new belief that the drug "works"?

(Bayesian interpretation)

What is the prior distribution of θ ? (select all that apply)

- A. $\theta \sim \text{Beta}(1, 1) = \text{Uni}(0, 1)$
- B. $\theta \sim \text{Beta}(81, 101)$
- C. $\theta \sim \text{Beta}(80, 20)$
- ☒ D. $\theta \sim \text{Beta}(81, 21)$ Interpretation: 80 successes / 100 imaginary trials
- ☒ E. $\theta \sim \text{Beta}(5, 2)$

(you can choose either based on how strongly you believe in prior data.

We choose E on next slide)



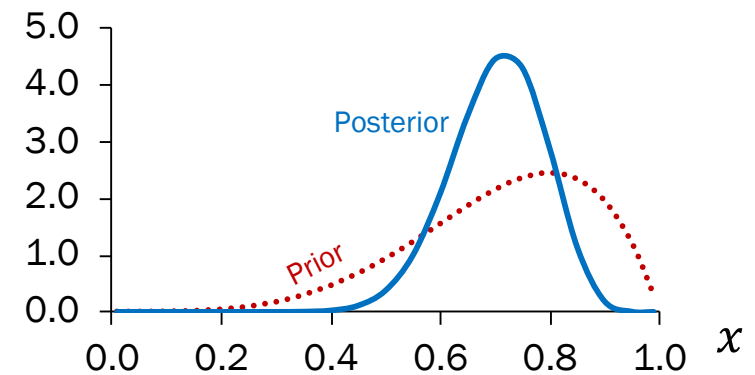
Medicinal Beta

- Before being tested, a medicine is believed to "work" 80% of the time.
- The medicine is administered to 20 patients.
- It "works" for 14, "doesn't work" for 6.

What is your new belief that the drug "works"?

Prior: $\theta \sim \text{Beta}(a = 5, b = 2)$

Posterior: $\theta \sim \text{Beta}(a = 5 + 14, b = 2 + 6)$
 $\sim \text{Beta}(a = 19, b = 8)$



Medicinal Beta

- Before being tested, a medicine is believed to "work" 80% of the time.
- The medicine is administered to 20 patients.
- It "works" for 14, "doesn't work" for 6.

What is your new belief that the drug "works"?

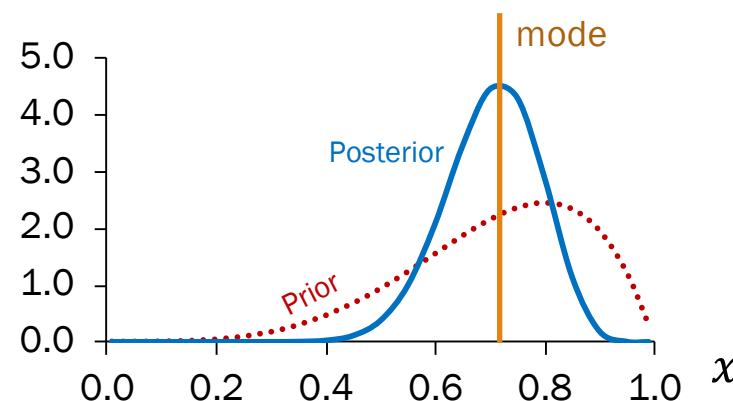
Prior: $\theta \sim \text{Beta}(a = 5, b = 2)$

Posterior: $\theta \sim \text{Beta}(a = 5 + 14, b = 2 + 6)$
 $\sim \text{Beta}(a = 19, b = 8)$

What do you report?

- A. Expectation of posterior
- B. Mode of posterior
- C. Distribution of posterior
- D. Nothing

(Bayesian interpretation)



Medicinal Beta

- Before being tested, a medicine is believed to "work" 80% of the time.
- The medicine is administered to 20 patients.
- It "works" for 14, "doesn't work" for 6.

What is your new belief that the drug "works"?

Prior: $\theta \sim \text{Beta}(a = 5, b = 2)$

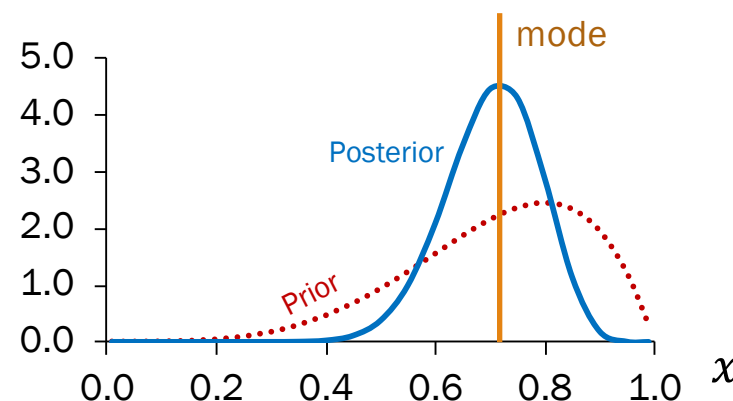
Posterior: $\theta \sim \text{Beta}(a = 5 + 14, b = 2 + 6)$
 $\sim \text{Beta}(a = 19, b = 8)$

What do you report?

$$E[\theta] = \frac{a}{a+b} = \frac{19}{19+8} \approx 0.70$$

$$\text{mode}(\theta) = \frac{a-1}{a+b-2} = \frac{18}{18+7} \approx 0.72$$

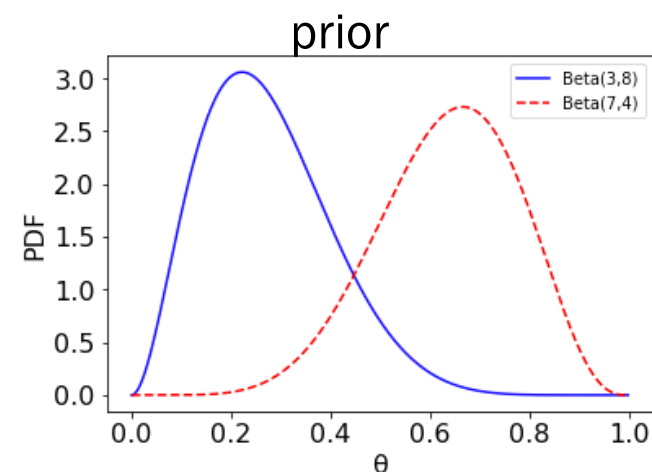
(Bayesian interpretation)



In CS109, we report the **mode**: The "most likely" parameter given the data.

Priors Should Become Irrelevant

- Let θ be the probability a coin turns up heads.
- Model θ with 2 different priors:
 - Prior 1: **Beta(3, 8)**: 2 imaginary heads, 7 imaginary tails mode: $\frac{2}{9}$
 - Prior 2: **Beta(7, 4)**: 6 imaginary heads, 3 imaginary tails mode: $\frac{6}{9}$



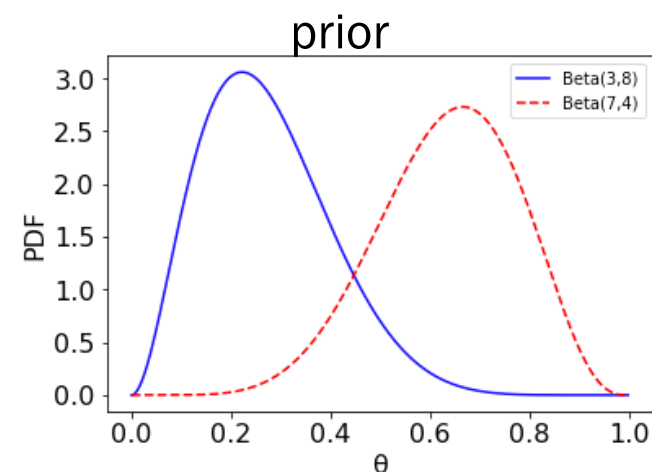
Now flip 100 coins and get 58 heads and 42 tails.

1. What are the two posterior distributions?
2. What are the modes of the two posterior distributions?



Priors Should Become Irrelevant

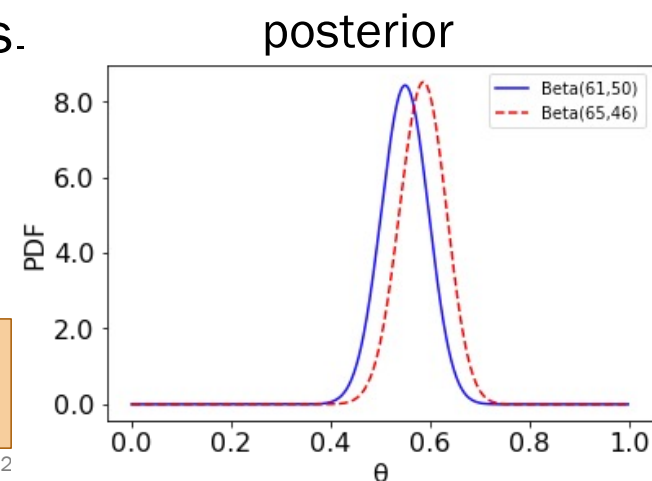
- Let θ be the probability a coin turns up heads.
- Model θ with 2 different priors:
 - Prior 1: **Beta(3, 8)**: 2 imaginary heads, 7 imaginary tails mode: $\frac{2}{9}$
 - Prior 2: **Beta(7, 4)**: 6 imaginary heads, 3 imaginary tails mode: $\frac{6}{9}$



Now flip 100 coins and get 58 heads and 42 tails.

Posterior 1: **Beta(61, 50)** with mode: $\frac{60}{109}$

Posterior 2: **Beta(65, 46)** with mode: $\frac{64}{109}$



Provided we collect enough data, posteriors will converge to the true value and choices of prior will matter less and less.

Beta: Recap



In this lecture:

$$X \sim \text{Ber}(p)$$

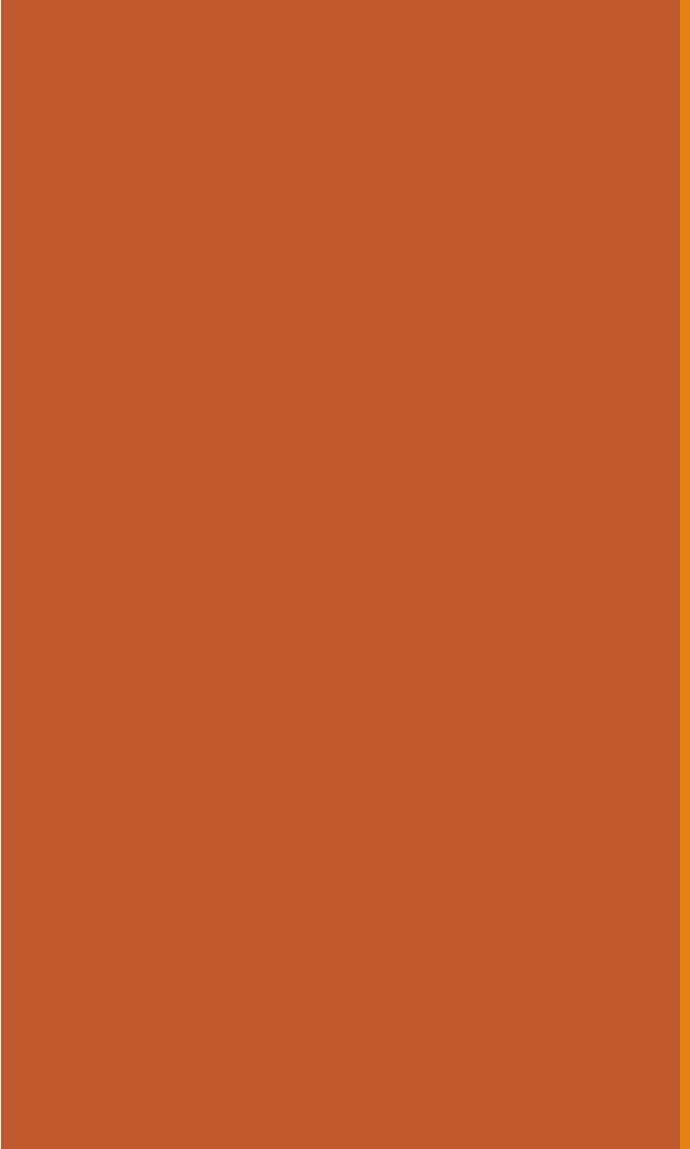
If nothing is known about the **parameter** p , Bayesian statisticians will:

- Treat the parameter as a random variable θ with a Beta prior distribution
- Conduct experiments
- Based on the outcomes of those experiments, update the posterior distribution of θ

Food for thought:

Any parameter for a "parameterized" random variable can be thought of as a random variable.

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$



Extra: Other Conjugates

only if we have time

Full Family: Conjugate Distributions

Extra and Optional

Distribution parameter	Conjugate distribution
Bernoulli p	Beta
Binomial p	Beta
Multinomial p_i	Dirichlet
Poisson λ	Gamma
Exponential λ	Gamma
Normal μ	Normal
Normal σ^2	Inverse Gamma

CS109: We'll only focus on conjugates for Bernoulli/Binomial p

Gamma Gamma Gamma

Extra and Optional

Gamma(α, β) is a conjugate for Poisson.

- Also conjugate for Exponential, but we won't delve into that
- Mode of gamma: $(\alpha - 1)/\beta$

Prior

$$\theta \sim \text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$$

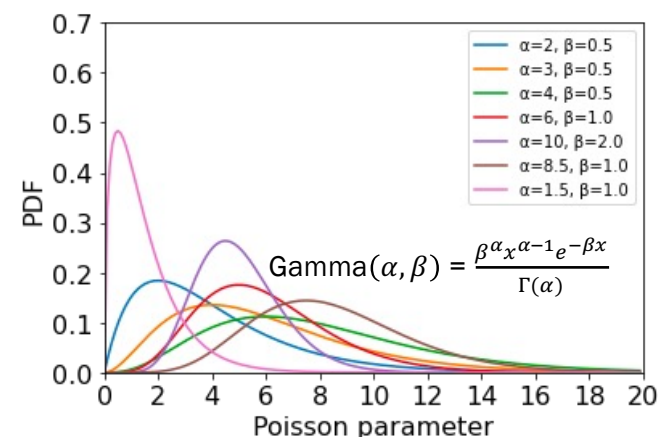
See $\alpha - 1$ total events across β prior time periods

Experiment Observe n events during next k additional time periods

Posterior $(\theta | n \text{ events in } k \text{ periods}) \sim \text{Gamma}(\alpha + n, \beta + k)$

Mode:

$$\theta_{\text{mode}} = \frac{\alpha + n - 1}{\beta + k}$$



Gamma Gamma Gamma

Extra and Optional

Let λ be the average # of successes in a time period.

1. What does it mean to have a prior of $\theta \sim \text{Gamma}(11, 5)$?

Observe 10 imaginary events in 5 time periods, i.e., observe at Poisson rate = 2

Now perform the experiment and see 11 events in next 2 time periods.

2. Given your prior, what is the posterior distribution?

$(\theta | n \text{ events in } k \text{ periods}) \sim \text{Gamma}(22, 7)$

3. What is θ_{mode} ?

$\theta_{\text{mode}} = 3$, updated Poisson rate

