Table of Contents

21: Bayesian Statistics and Beta

Jerry Cain

February 28, 2024

[Lecture Discussion on Ed]

brand new content

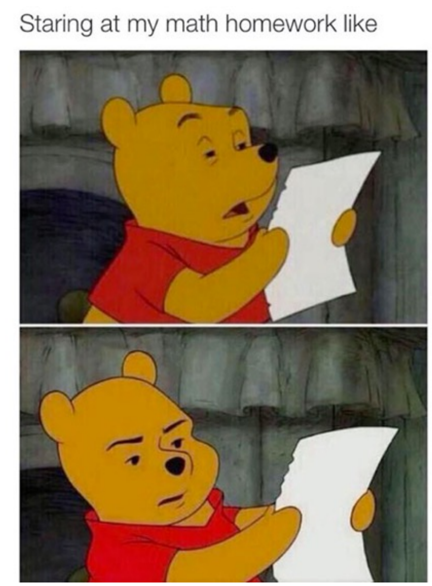
MLE: Multinomial

Okay, just one more MLE with the Multinomial

Consider a sample of n iid random variables where:

- Each element is drawn from one of m outcomes.
  $P(\text{outcome } i) = p_i$, where $\sum_{i=1}^{m} p_i = 1$
- X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

Let's give an example!



Okay, just one more MLE with the Multinomial

Consider a sample of n iid random variables where:

- Each element is drawn from one of m outcomes.
  $P(\text{outcome } i) = p_i$, where $\sum_{i=1}^{m} p_i = 1$
- X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

Example: Suppose each RV is outcome of 6-sided die. $m = 6, \sum_{i=1}^{6} p_i = 1$

- Roll the dice $n = 12$ times.
- Observe data: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

$$X_1 = 3, X_2 = 2, X_3 = 0,$$
$$X_4 = 3, X_5 = 1, X_6 = 3$$

Check: $X_1 + X_2 + \cdots + X_6 = 12$

Okay, just one more MLE with the Multinomial

Consider a sample of n iid random variables where:

- Each element is drawn from one of m outcomes.
  $P(\text{outcome } i) = p_i$, where $\sum_{i=1}^{m} p_i = 1$
- X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

1. What is the likelihood of observing the sample $(X_1, X_2, \dots, X_m)$, given the probabilities $p_1, p_2, \dots, p_m$?

A. $$\frac{n!}{X_1!\,X_2!\cdots X_m!}\,p_1^{X_1} p_2^{X_2} \cdots p_m^{X_m}$$

B. $$p_1^{X_1} p_2^{X_2} \cdots p_m^{X_m}$$

C. $$\frac{n!}{X_1!\,X_2!\cdots X_m!}\,X_1^{p_1} X_2^{p_2} \cdots X_m^{p_m}$$

Okay, just one more MLE with the Multinomial

Consider a sample of n iid random variables where:

- Each element is drawn from one of m outcomes.
  $P(\text{outcome } i) = p_i$, where $\sum_{i=1}^{m} p_i = 1$
- X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

1. What is the likelihood of observing the sample $(X_1, X_2, \ldots, X_m)$, given the probabilities $p_1, p_2, \ldots, p_m$?

$$L(\theta) = \frac{n!}{X_1!\,X_2!\cdots X_m!} p_1^{X_1} p_2^{X_2} \cdots p_m^{X_m}$$

2. What is θ_{MLE}?

$$LL(\theta) = \log(n!) - \sum_{i=1}^{m} \log(X_i!) + \sum_{i}^{m} X_i \log(p_i)\,,$$ such that $\sum_{i=1}^{m} p_i = 1$

Optimize with Lagrange multipliers in extra slides $\rightarrow$ θ_{MLE}: $p_i = \frac{X_i}{n}$

Intuitively, probability p_i = proportion of outcomes

When MLEs attack!

MLE for Multinomial: $p_i = \frac{X_i}{n}$

Consider a 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

What is θ_{MLE}?

When MLEs attack!

MLE for Multinomial: $p_i = \frac{X_i}{n}$

Consider a 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

θ_{MLE}:

$p_1 = 3/12$
$p_2 = 2/12$
$p_3 = 0/12$ ⚠
$p_4 = 3/12$
$p_5 = 1/12$
$p_6 = 3/12$

- MLE say you just never roll threes.
- Do you really believe that?

Roll more!
prob = frequency
in limit

But what if you cannot observe anymore rolls?

Frequentist

Bayesian Statistics

Starting Today!

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

We are going to think of probabilities as random variables.

A new definition of probability

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

We don't know the probability θ that the coin comes up heads.

The world's first coin

Frequentist

θ is a single value.

$$\theta = \lim_{n+m \to \infty} \frac{n}{n+m} \approx \frac{n}{n+m}$$

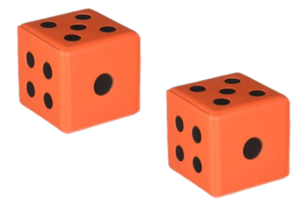
Bayesian

θ is a **random variable**.

θ's continuous support: (0, 1)

Let's play a game

Roll 2 dice. If neither roll is a 6,
you win (event W). Else, I win (event W^C).



- Before you play, what's the probability that you win?
- Play once. What's the probability that you win?
- Play three more times. What's the probability that you win?

Frequentist

$$P(W) = \left(\frac{5}{6}\right)^2$$

Bayesian

I am constantly re-evaluating the situation

Bayesian statistics: Constantly update your prior beliefs.

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.

Mixing discrete and continuous

Let X be a continuous random variable, and
N be a discrete random variable.

Bayes' Theorem:

$$f_{X|N}(x|n) = \frac{p_{N|X}(n|x) f_X(x)}{p_N(n)}$$

Intuition: $P(X = x|N = n) = \dfrac{P(N = n|X = x)P(X = x)}{P(N = n)}$

$$f_{X|N}(x|n)\varepsilon_X = \frac{P(N = n|X = x) f_X(x)\varepsilon_X}{P(N = n)} \Rightarrow f_{X|N}(x|n) = \frac{p_{N|X}(n|x) f_X(x)}{p_N(n)}$$

Bayes' Theorem: 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: $$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$$

Mixing discrete and continuous

Let θ be a random variable for the probability your coin comes up heads, and N be the number of heads you observe in an experiment.

likelihood prior

posterior

$$f_{\theta|N}(x|n) = \frac{p_{N|\theta}(n|x) f_\theta(x)}{p_N(n)}$$

normalization constant

- Prior belief of parameter θ — $f_\theta(x)$
- Likelihood of $N = n$ heads, given parameter $\theta = x$. — $p_{N|\theta}(n|x)$
- Posterior updated belief of parameter θ. — $f_{\theta|N}(x|n)$

Beta RV

Beta random variable

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

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

$a > 0, b > 0$

Support of X: $(0, 1)$

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

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

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

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

Beta RV with different a, b

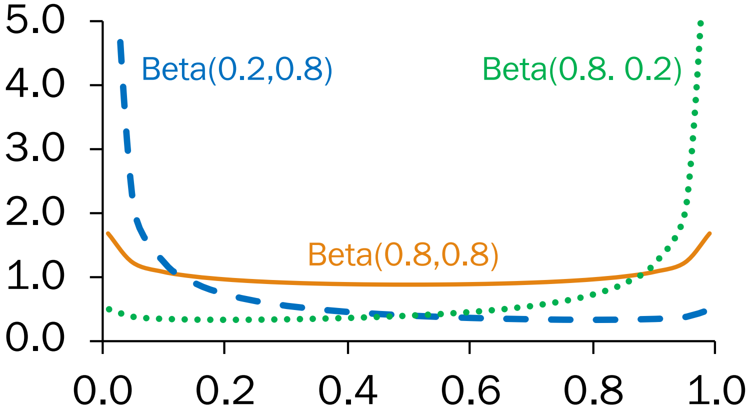
$$X \sim \text{Beta}(a, b)$$

$$a > 0, b > 0$$

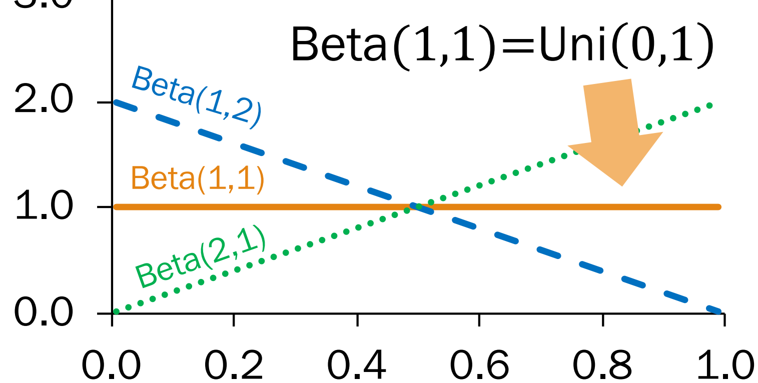
Support of X: $(0, 1)$

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

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



SAT word: *adumbrate*



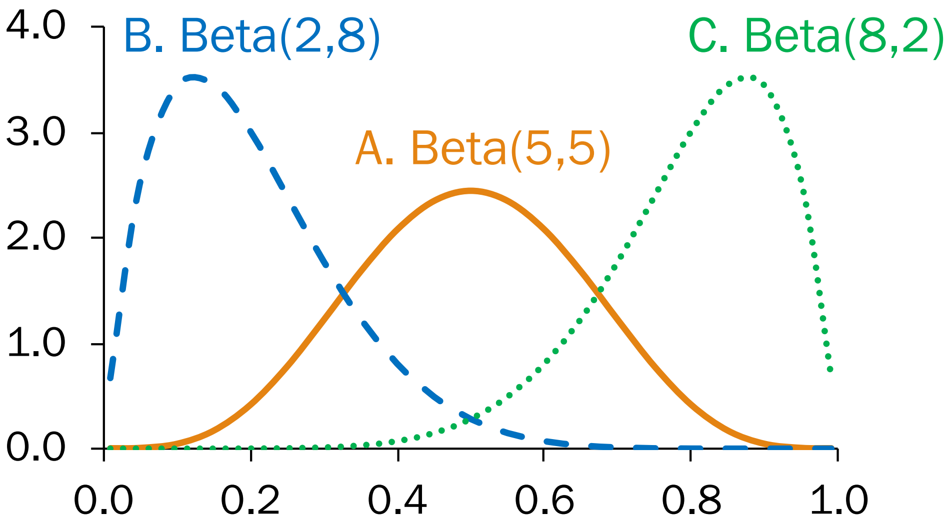
+ a third case (next slide)

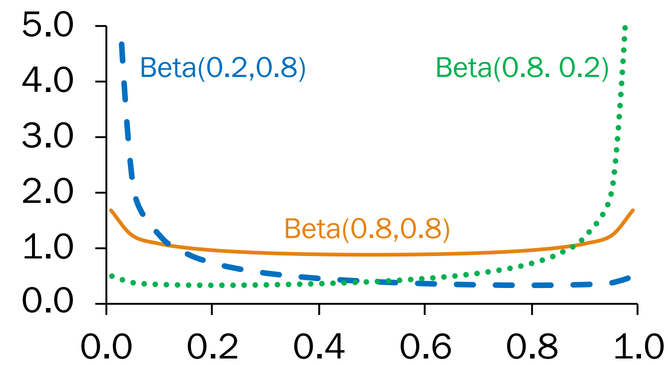
Note: PDF symmetric when $a = b$

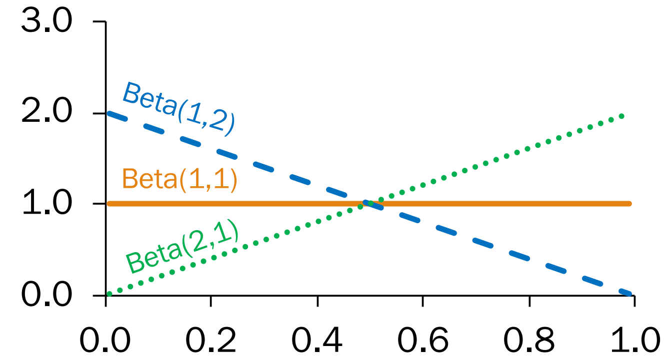
Beta RV with different a, b

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

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.

Beta random variable

<u>def</u> A **Beta** random variable X is defined as follows:

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

$a > 0, b > 0$

Support of X: $(0, 1)$

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

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

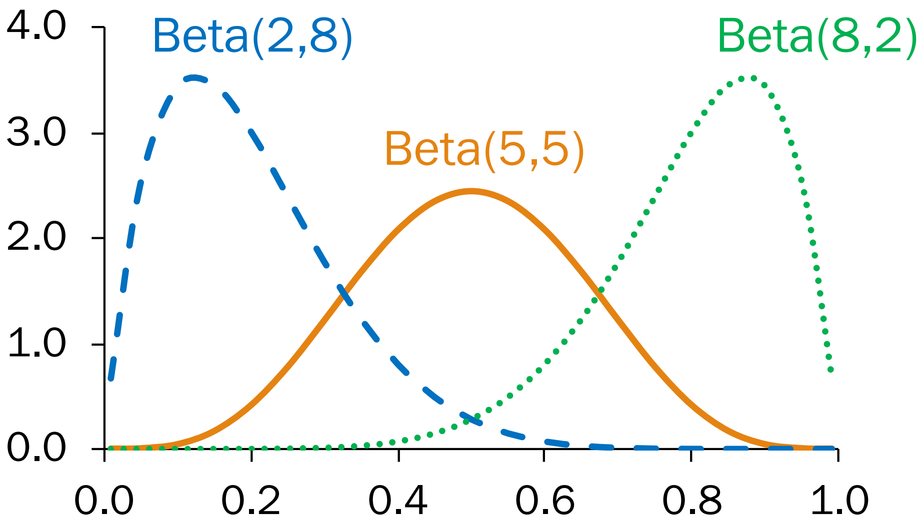
Expectation $\quad E[X] = \frac{a}{a+b}$

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

Beta can be a distribution of probabilities.

Beta can be a distribution of probabilities.

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



Beta parameters a, b are determined by the outcome of an experiment.

But which experiment?

Flipping a coin with unknown probability

Flip a coin with unknown probability

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

- Before our experiment, θ (the probability that the coin comes up heads) is equally like to be any probability in (0, 1).
- Let N = number of heads.
- Given $\theta = x$, coin flips are independent.

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

What are reasonable distributions 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. Use Bayes'!

Flip a coin with unknown probability

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

- Before our experiment, θ (the probability that the coin comes up heads) is equally like to be any probability in (0, 1).
- Let $N =$ number of heads.
- Given $\theta = x$, coin flips are independent.

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

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

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

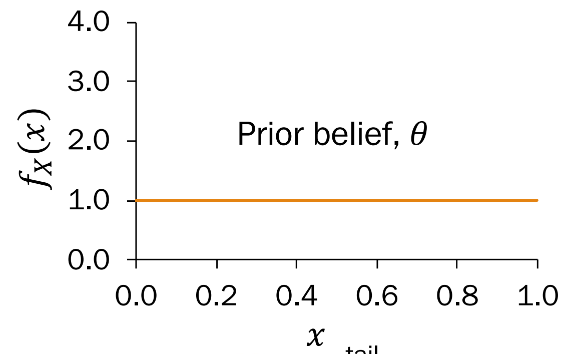
Posterior: $f_{\theta|N}(\theta|n)$

$$f_{\theta|N}(x|n) = \frac{p_{N|\theta}(n|x) f_\theta(x)}{p_N(n)} = \frac{\binom{n+m}{n} x^n (1-x)^m \cdot 1}{p_N(n)}$$

$$= \frac{\binom{n+m}{n}}{\underbrace{p_N(n)}_{\text{constant with respect to } x\text{, doesn't depend on } x}} x^n (1-x)^m = \frac{1}{c} x^n (1-x)^m, \text{ where } c = \int_0^1 x^n (1-x)^m dx$$

Let's try it out

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?

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

c normalizes to valid PDF

Wait a minute! #looksbetalike

Beta RV with different a, b

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

$a > 0, b > 0$

Support of X: $(0, 1)$

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

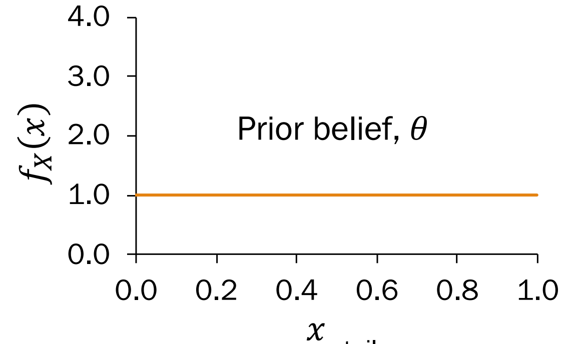
where $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$, normalizing constant

🌟 $\quad f_{\theta|N}(x|n) = \frac{1}{c}\, x^7(1-x)^1 \quad$ is the PDF for $\text{Beta}(8, 2)$!

c normalizes to valid PDF

Let's try it out

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?

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

c normalizes to valid PDF

Beta(8,2)

3. What is our posterior belief of the probability 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:

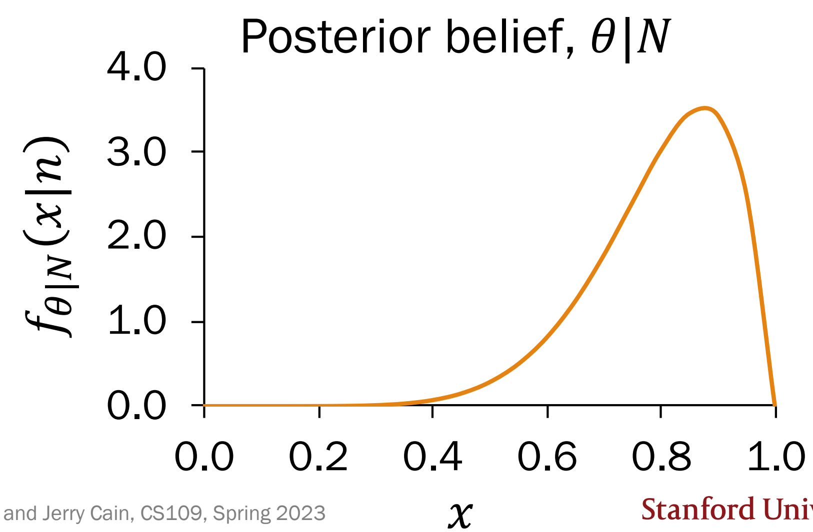
$$f_{\theta|N}(x|n) = \frac{1}{c}\, x^{7}(1-x)^{1}, \text{where } c = \int_0^1 x^{7}(1-x)^{1} dx$$

Posterior belief, $\theta|N$:

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

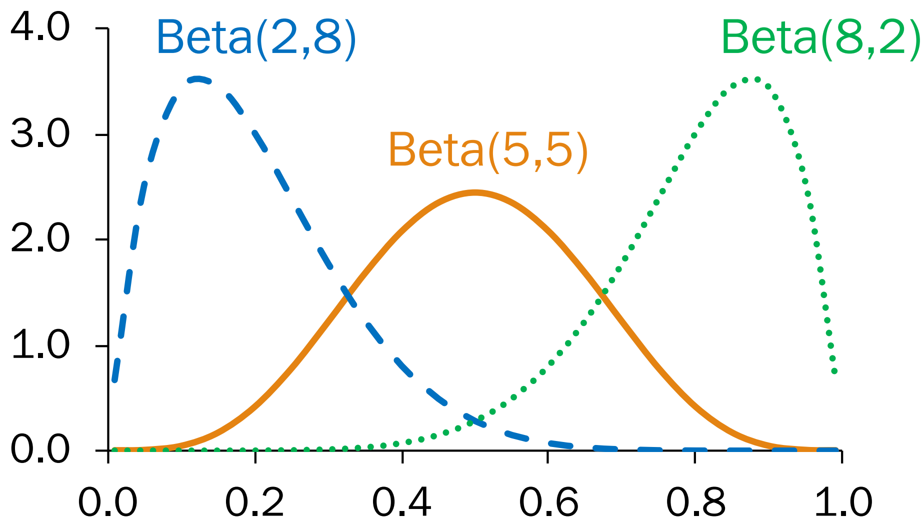
$$f_{\theta|N}(x|n) = \frac{1}{c}\, x^{8-1}(1-x)^{2-1}$$

Beta$(a = n + 1, b = m + 1)$



CS109 focus: Beta where a, b both positive integers

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



Beta parameters a, b are determined by the outcome of an experiment.

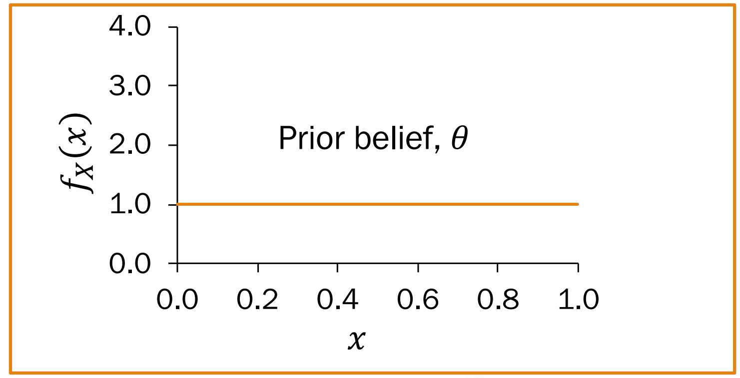
$$a = \text{"successes"} + 1$$
$$b = \text{"failures"} + 1$$

- Beta (in CS109) models the randomness of the probability of experiment success.
- Beta parameters depend on our data and our prior.

Conjugate distributions

A note about our prior

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

okay

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)

Wait another minute!

Beta RV with different a, b

Review

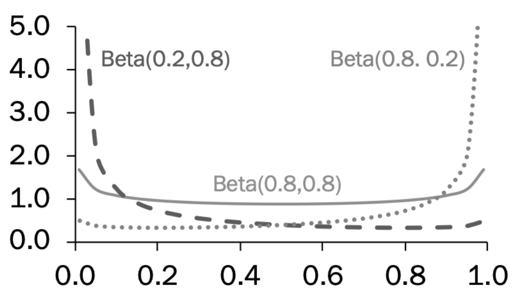
$$X \sim \text{Beta}(a, b)$$

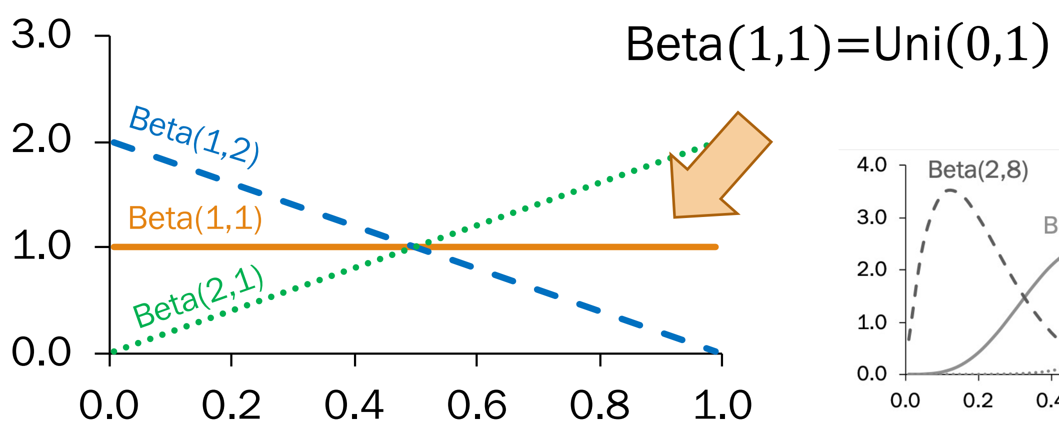
$$a > 0, b > 0$$

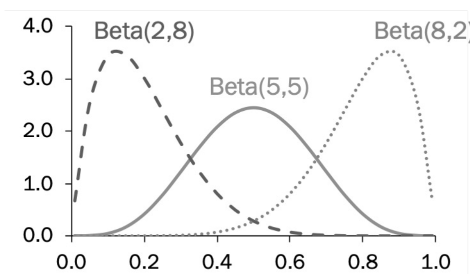
Support of X: $(0, 1)$

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

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







Note: PDF symmetric when $a = b$

A note about our prior

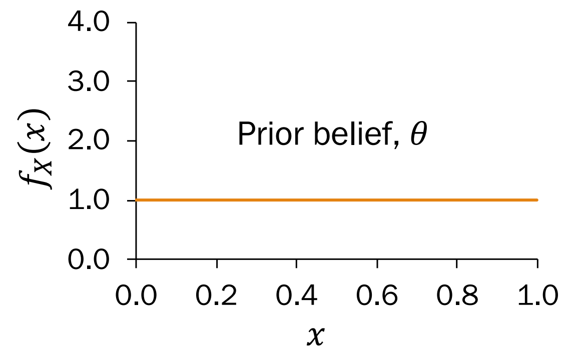
1. Start with a $\theta \sim \text{Uni}(0,1)$ over probability that a coin lands heads.

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$(8,2)$



Check this out. Beta$(a = 1, b = 1)$:

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

$$= \frac{1}{\int_0^1 1 dx}$$

$$= 1 \qquad \text{where } 0 < x < 1$$

Beta is a conjugate distribution for Bernoulli

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

- Prior and posterior parametric forms are the same

(proof on next slide)

Beta is a conjugate distribution for Bernoulli

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

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

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

$$f_{\theta|N}(x|n) = \frac{p_{N|\theta}(n|x) f_\theta(x)}{p_N(n)} = \frac{\binom{n+m}{m} x^n (1-x)^m \cdot \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}}{p_N(n)}$$

constants that don't depend on x

$$= C \cdot x^n (1-x)^m \cdot x^{a-1}(1-x)^{b-1}$$

$$= C \cdot x^{n+a-1}(1-x)^{m+b-1} \quad ✅$$

Beta is a conjugate distribution for Bernoulli

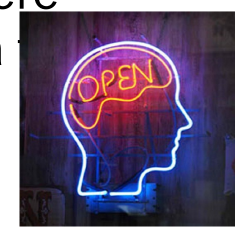
This is the main takeaway of Beta.

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

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

You can **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
- Choosing $\text{Beta}(1, 1) = \text{Uni}(0, 1)$ means you don't hold any prior beliefs



Prior	$\text{Beta}(a = n_{imag} + 1, b = m_{imag} + 1)$
Experiment	Observe n successes and m failures
Posterior	$\text{Beta}\left(a = n_{imag} + n + 1, b = m_{imag} + m + 1\right)$

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"?

Frequentist

Let p be the probability your drug works.

$$p \approx \frac{14}{20} = 0.7$$

Bayesian

A frequentist view will not incorporate prior/expert belief about probability.

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"?

Frequentist

Let p be the probability your drug works.

$$p \approx \frac{14}{20} = 0.7$$

Bayesian

Let θ be the probability your drug works.

θ is a random variable.

Medicinal Beta

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

- 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. $\theta \sim \text{Beta}(81, 21)$

E. $\theta \sim \text{Beta}(5, 2)$

Medicinal Beta

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

- 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

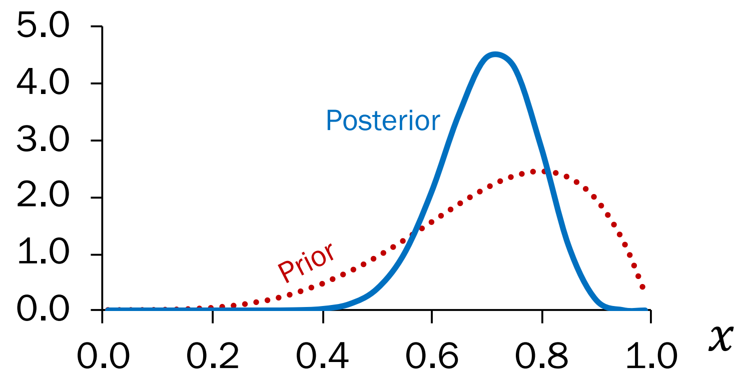
Prior	$\text{Beta}(a = n_{imag} + 1, b = m_{imag} + 1)$
Posterior	$\text{Beta}(a = n_{imag} + n + 1, b = m_{imag} + m + 1)$

- 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)

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

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

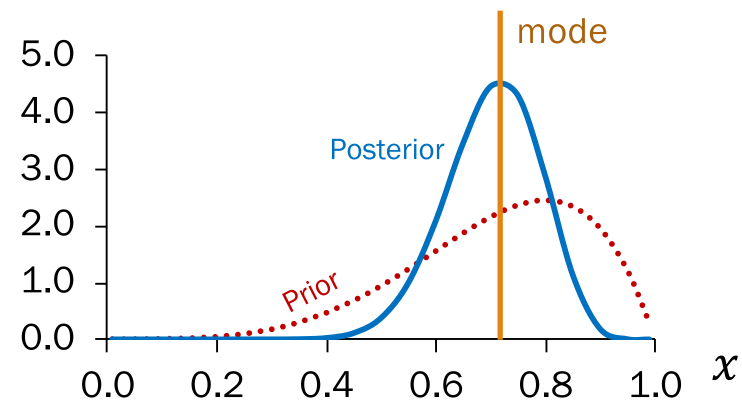
- Before being tested, a medicine is believed to "work" 80% of the time.
- The medicine is administered to 20 patients.
- It "works" for 14, "doesn't work" for 6.

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

(Bayesian interpretation)

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 to pharmacists?

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

Medicinal Beta

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

- 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.

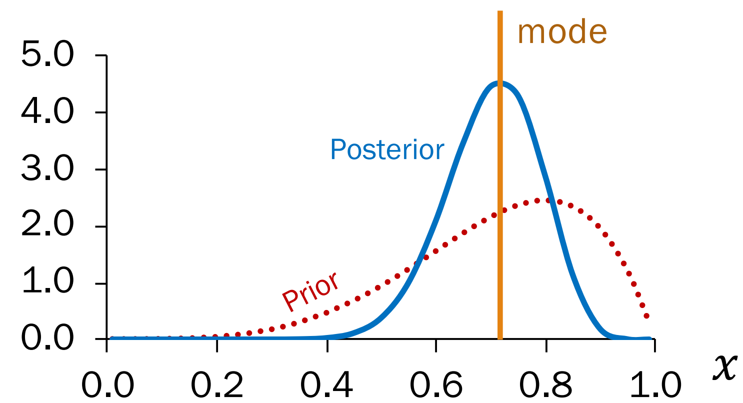
What is your new belief that the drug "works"?

(Bayesian interpretation)

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 to pharmacists?

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

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

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

Food for thought

👉 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.

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

Estimating our parameter directly

(our focus so far)

Maximum Likelihood Estimator (MLE)

What is the parameter θ that **maximizes the likelihood** of our observed data $(x_1, x_2, \dots, x_n)$?

$$L(\theta) = f(X_1, X_2, \dots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

$$\theta_{MLE} = \arg\max_{\theta} \underbrace{f(X_1, X_2, \dots, X_n|\theta)}_{\text{likelihood of data}}$$

Observations:

- MLE maximizes probability of observing data given a parameter θ. It's fitting the curve to match the data.
- If we are estimating θ, shouldn't we maximize the probability of θ directly? SAT word: *adumbrate*

Extra: MLE: Multinomial derivation

Okay, just one more MLE with the Multinomial

Consider a sample of n i.i.d. random variables where

- Each element is drawn from one of m outcomes.
  $P(\text{outcome } i) = p_i$, where $\sum_{i=1}^{m} p_i = 1$
- X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

1. What is the likelihood of observing the sample $(X_1, X_2, \dots, X_m)$, given the probabilities $p_1, p_2, \dots, p_m$?

$$L(\theta) = \frac{n!}{X_1!\,X_2! \cdots X_m!} p_1^{X_1} p_2^{X_2} \cdots p_m^{X_m}$$

2. What is θ_{MLE}?

$$LL(\theta) = \log(n!) - \sum_{i=1}^{m} \log(X_i!) + \sum_{i=1}^{m} X_i \log(p_i), \quad \text{such that } \sum_{i=1}^{m} p_i = 1$$

$$\theta_{MLE}\colon \ p_i = \frac{X_i}{n}$$

Intuitively, probability p_i = proportion of outcomes

Optimizing MLE for Multinomial

$\theta = (p_1, p_2, \ldots, p_m)$

$\theta_{MLE} = \arg\max_{\theta} LL(\theta)$, where $\sum_{i=1}^{m} p_i = 1$

Use Lagrange multipliers to account for constraint

Lagrange multipliers: $A(\theta) = LL(\theta) + \lambda\left(\sum_{i=1}^{m} p_i - 1\right) = \sum_{i=1}^{m} X_i \log(p_i) + \lambda\left(\sum_{i=1}^{m} p_i - 1\right)$ (drop non-p_i terms)

Differentiate w.r.t. each p_i, in turn: $\frac{\partial A(\theta)}{\partial p_i} = X_i \frac{1}{p_i} + \lambda = 0 \Rightarrow p_i = -\frac{X_i}{\lambda}$

Solve for λ, noting $\sum_{i=1}^{m} X_i = n, \sum_{i=1}^{m} p_i = 1$: $\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} -\frac{X_i}{\lambda} = 1 \Rightarrow 1 = -\frac{n}{\lambda} \Rightarrow \lambda = -n$

Substitute λ into p_i: $p_i = \frac{X_i}{n}$