21: Beta

Lisa Yan
May 22, 2020
Quick slide reference

3 MLE: Multinomial 21a_mle_multinomial

11 Bayesian statistics/Beta sneak peek 21b_bayesian

20 The Beta RV 21c_beta

37 Flipping a coin with unknown probability LIVE

* Extra: MLE: Multinomial Derivation 21e_extra
MLE: Multinomial
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 $\sum_{i=1}^{m} p_i = 1$
- $X_i = \# \text{ of trials with outcome } i$, 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 \( n \) 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 \( \sum_{i=1}^{m} X_i = n \)

Example: Suppose each RV is outcome of a 6-sided die.
- 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 \( n \) 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 \( (X_1, X_2, \ldots, X_m) \), given the probabilities \( p_1, p_2, \ldots, p_m \)?

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

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

C. \[ \frac{n!}{X_1! \cdots X_m!} X_1^{p_1} X_2^{p_2} \cdots X_m^{p_m} \]
Okay, just one more MLE with the Multinomial

Consider a sample of \( n \) i.i.d. random variables where

- Each element is drawn from one of \( m \) outcomes.
  \[ P(\text{outcome } i) = p_i, \text{ where } \sum_{i=1}^{m} p_i = 1 \]
- \( X_i \) = # 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 \( p_1, p_2, \ldots, p_m \)?

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 \( n \) i.i.d. random variables where

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

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 \( \theta_{MLE} \)?

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

   \( \theta_{MLE}: p_i = \frac{X_i}{n} \)

   Intuitively, probability \( p_i = \text{ proportion of outcomes} \)

Optimize with Lagrange multipliers in extra slides

Lisa Yan, CS109, 2020
When MLEs attack!

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 $\theta_{MLE}$?
When MLEs attack!

Consider a 6-sided die.

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

$\theta_{MLE}$:

\[ p_1 = \frac{3}{12} \]
\[ p_2 = \frac{2}{12} \]
\[ p_3 = 0/12 \]
\[ p_4 = \frac{3}{12} \]
\[ p_5 = \frac{1}{12} \]
\[ p_6 = \frac{3}{12} \]

• MLE: you’ll never...EVER... roll a three.
• Do you really believe that?

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

Today: A new definition of probability!
Bayesian Statistics
When MLEs attack!

Consider a 6-sided die.

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

$\theta_{MLE}$:

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

⚠️

- MLE: you’ll **never**...**EVER**... roll a three.
- Do you really believe that?

Roll more! Prob. = frequency in limit

But what if you cannot observe anymore rolls?
Today’s plan

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, come up with $n$ heads. We don’t know the probability $\theta$ that the coin comes up heads.

**Frequentist**

$\theta$ is a single value.

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

**Bayesian**

$\theta$ is a **random variable**.

$\theta$’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

wait hold up this situation is whack tho

Bayesian statistics: Update your prior beliefs of probability.
Bayesian statistics: Probability is a reasonable expectation representing a state of knowledge.

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:

$$P(X = x|N = n) = \frac{P(N = n|X = x)P(X = x)}{P(N = n)}$$

$$f_{X|N}(x|n)e_X = \frac{P(N = n|X = x)f_X(x)e_x}{P(N = n)} \quad \Rightarrow \quad f_{X|N}(x|n) = \frac{p_{N|X}(n|x)f_X(x)}{p_N(n)}$$
All your Bayes are belong to us

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

**OG 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 $\theta$ be a random variable for the probability your coin comes up heads, and $N$ be the number of heads you observe in an experiment.

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

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

More in live lecture!
Beta RV
Beta random variable

Def: A Beta random variable $X$ is defined as follows:

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

where $a > 0$, $b > 0$

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

Expectation

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

Variance

$$\text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$
Beta RV with different $a, b$

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

$a > 0, b > 0$

Support of $X$: $(0, 1)$

PDF

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

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

Note: PDF symmetric when $a = b$
Beta RV with different $a, b$

Match PDF to distribution:

A. $\text{Beta}(5, 5)$
B. $\text{Beta}(2, 8)$
C. $\text{Beta}(8, 2)$
Beta RV with different $a, b$

Match PDF to distribution:

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

In CS109, we focus on Betas where $a, b$ are both positive integers.
Beta random variable

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

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

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

Expectation

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

Variance

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

Beta can be a distribution of probabilities.
Beta can be a distribution of probabilities.

Beta parameters $a, b$ could come from an experiment...

But which one?
Stay tuned...
21: Beta

Slides by Lisa Yan
August 3, 2020
Flipping a coin with unknown probability
A new definition of probability

Flip a coin $n + m$ times, comes up with $n$ heads. We don’t know the probability $\theta$ that the coin comes up heads.

Frequentist

$\theta$ is a single value.

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

Bayesian

$\theta$ is a random variable.

$\theta$’s continuous support: (0, 1)
Flip a coin with unknown probability

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

• Before our experiment, $\theta$ (the probability that the coin comes up heads) can be any probability.
• Let $N =$ number of heads.
• Given $\theta = x$, coin flips are independent.

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

What are reasonable distributions of the following?

1. $\theta$
2. $N|\theta = x$
3. $\theta|N$
Flip a coin with unknown probability

Flip a coin \( n + m \) times, observe \( n \) heads.

- Before our experiment, \( \theta \) (the probability that the coin comes up heads) can be any probability.
- Let \( N \) = number of heads.
- Given \( \theta = x \), coin flips are independent.

What is our updated belief of \( \theta \) after we observe \( N = n \)?

What are reasonable distributions of the following?

1. \( \theta \)  
   Bayesian prior \( \theta \sim \text{Uni}(0,1) \)

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

3. \( \theta | N \)  
   Bayesian posterior. Use Bayes’
Flip a coin with unknown probability

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

- Before our experiment, $X$ (the probability that the coin comes up heads) can be any probability.
- Let $N =$ number of heads.
- Given $X = x$, coin flips are independent.

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

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

$$= \frac{(n + m)^n}{p_N(n)} x^n (1 - x)^m$$

constant with respect to $x$, doesn’t depend on $x$

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

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

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

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

$c$ normalizes to valid PDF.
Beta RV with different $a, b$

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

$a > 0, b > 0$

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

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

is the PDF for 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 $\theta$?

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

$c$ normalizes to valid PDF

Beta(8,2)
3. What is our posterior belief of the probability $\theta$?

- 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 $\theta$ is:

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

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

If \(a, b\) are positive integers, Beta parameters \(a, b\) could come from 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 prior and our data.
Interlude for fun/announcements
Bayes’ on the waves

\[
P(I'm \text{ near the ocean} \mid I \text{ picked up a seashell}) = \frac{P(I \text{ picked up} \mid I'm \text{ near the ocean}) P(I'm \text{ near the ocean})}{P(I \text{ picked up a seashell})}
\]

Statistically speaking, if you pick up a seashell and don’t hold it to your ear, you can probably hear the ocean.
Announcements

Problem Set 6: **No late days or on-time bonus**
Interesting probability news

Why Rejection Sampling Is Useful in Cat Modeling

Note: Cat Modeling = Catastrophe Modeling (e.g., earthquakes, hurricanes, etc.)

Ethics in Probability: Frequentist vs. Bayesian

**Question:** What do these schools of thought have to do with ethics?

**Frequentist perspective.** There is a lot of room for ethical dilemmas when you infer a prior, because there are often different perspectives on prior belief.

**Bayesian perspective.** When some reasonable model for uncertainty is required, for example in clinical trials, a belief system like Bayesian statistics is more useful than a single answer (we see this with the Beta).

For a deeper dive:

**The False Dilemma: Bayesian vs. Frequentist**
Jordi Vallverdú
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

3. What is our posterior belief of the probability $\theta$?

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

$c$ normalizes to valid PDF

Beta(8,2)
Beta RV with different $a, b$

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

- $a > 0, b > 0$
- 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 $a = b$
A note about our prior

1. Start with a $\theta \sim \text{Uni}(0,1)$ over probability that a coin lands heads.
   
   \[ f(x) = \frac{1}{B(a, b)} x^{a-1} (1 - x)^{b-1} \]
   
   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 $\theta$?

   \[ \text{Beta}(8,2) \]

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

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

\[ = \frac{1}{\int_0^1 1 \, dx} \]

\[ = 1 \quad \text{where } 0 < x < 1 \]
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
3. Our posterior is also Beta.

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

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

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

\[
= C \cdot x^{n+a-1}(1-x)^{m+b-1} \quad \checkmark
\]
Beta is a conjugate distribution for Bernoulli

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 set the prior to reflect how biased you think the coin is apriori:

- $\theta \sim \text{Beta}(a, b)$: have seen $(a + b - 2)$ **imaginary trials**, where $(a - 1)$ are heads, $(b - 1)$ tails
- Then $\text{Beta}(1, 1) = \text{Uni}(0, 1)$ means we haven’t seen any imaginary trials

<table>
<thead>
<tr>
<th>Prior</th>
<th>$\text{Beta}(a = n_{imag} + 1, b = m_{imag} + 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment</strong></td>
<td>Observe $n$ successes and $m$ failures</td>
</tr>
<tr>
<td><strong>Posterior</strong></td>
<td>$\text{Beta}(a = n_{imag} + n + 1, b = m_{imag} + m + 1)$</td>
</tr>
</tbody>
</table>
The enchanted die

Let $\theta$ be the probability of rolling a 6 on Lisa’s die.

- Prior: Imagine 1 out of 6 die rolls where only 6 showed up
- Observation: roll it a few times...

What is the updated distribution of $\theta$ after our observation?

Check out the demo!
Medicinal Beta

• Before being tested, a medicine is believed to “work” 80% of the time.
• The medicine is tried on 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 tried on 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 $\theta$ be the probability your drug works.

$\theta$ is a random variable.
Medicinal Beta

• Before being tested, a medicine is believed to “work” 80% of the time.
• The medicine is tried on 20 patients.
• It “works” for 14, “doesn’t work” for 6.

What is your new belief that the drug “works”?  

What is the prior distribution of $\theta$? (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

- Before being tested, a medicine is believed to “work” 80% of the time.
- The medicine is tried on 20 patients.
- It “works” for 14, “doesn’t work” for 6.

What is your new belief that the drug “works”?

What is the prior distribution of $\theta$? (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 strong your belief is (an engineering choice). We choose E on next slide)
Medicinal Beta

• Before being tested, a medicine is believed to “work” 80% of the time.
• The medicine is tried on 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: \( \theta \sim \text{Beta}(a = 5 + 14, b = 2 + 6) \sim \text{Beta}(a = 19, b = 8) \)

(Bayesian interpretation)
Medicinal Beta

• Before being tested, a medicine is believed to “work” 80% of the time.
• The medicine is tried on 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 to pharmacists?
A. Expectation of posterior
B. Mode of posterior
C. Distribution of posterior
D. Nothing
Medicinal Beta

- Before being tested, a medicine is believed to “work” 80% of the time.
- The medicine is tried on 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 to pharmacists?

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

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.
In this lecture: If we don’t know the parameter $p$, Bayesian statisticians will:

- Treat the parameter as a random variable $\theta$ with a Beta prior distribution
- Perform an experiment
- Based on experiment outcomes, update the posterior distribution of $\theta$

Food for thought: Any parameter for a “parameterized” random variable can be thought of as a random variable.
Estimating our parameter directly

(our focus so far)

Maximum Likelihood Estimator (MLE)

What is the parameter \( \theta \) that maximizes the likelihood of our observed data \((x_1, x_2, \ldots, x_n)\)?

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

\[
\theta_{MLE} = \arg \max_{\theta} f(X_1, X_2, \ldots, X_n | \theta)
\]

Observations:

- MLE maximizes probability of observing data given a parameter \( \theta \).
- If we are estimating \( \theta \), shouldn’t we maximize the probability of \( \theta \) directly?

See you next time!
Next: Maximum A Posteriori (MAP) Estimation

$$\theta_{MAP} = \arg \max_{\theta} ??$$
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.
  
  \[ P(\text{outcome } i) = p_i, \text{ where } \sum_{i=1}^{m} p_i = 1 \]
  
- $X_i = \# \text{ of trials with outcome } i$, 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 $\theta_{MLE}$?

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

   \[ \theta_{MLE}: \quad p_i = \frac{X_i}{n} \quad \text{Intuitively, probability} \quad p_i = \text{proportion of outcomes} \]

   Optimize with Lagrange multipliers in extra slides
Optimizing MLE for Multinomial

\[ \theta = (p_1, p_2, \ldots, p_m) \]

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

Use Lagrange multipliers to account for constraint

Lagrange multipliers:

\[
A(\theta) = LL(\theta) + \lambda \left( \sum_{i=1}^{m} p_i - 1 \right) = \sum_{i} X_i \log(p_i) + \lambda \left( \sum_{i=1}^{m} p_i - 1 \right) \quad \text{(drop non-} p_i \text{ terms)}
\]

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 \( \lambda \), noting

\[
\sum_{i=1}^{m} X_i = n, \sum_{i=1}^{m} p_i = 1:
\]

\[
\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} -\frac{X_i}{\lambda} = 1 \Rightarrow 1 = -\frac{n}{\lambda} \quad \Rightarrow \lambda = -n
\]

Substitute \( \lambda \) into \( p_i \)

\[
p_i = \frac{X_i}{n}
\]