Beta: The Random Variable for Probabilities

Chris Piech + Jerry Cain
CS109, Stanford University

Philosophical Ponderings:
You ask about the probability of rain tomorrow.

Person A: My leg itches when it rains and its kind of itchy…. Uh, p = .80

Person B: I have done complex calculations and have seen 10,451 days like tomorrow… p = 0.80

What is the difference between the two estimates?

"Those who are able to represent what they do not know make better decisions"

- CS109

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

Review

Conditioning with a continuous random variable is odd at first. But then it gets fun.

Its like snorkeling…

Continuous Conditional Distributions

Let X be continuous random variable

Let E be an event:

$$
\begin{aligned}
P(E|X = x) &= \frac{P(X = x, E)}{P(X = x)} \\
&= \frac{P(X = x|E)P(E)}{P(X = x)} \\
&= \frac{f_X(x|E)P(E)\epsilon_x}{f_X(x)\epsilon_x} \\
&= \frac{f_X(x|E)P(E)}{f_X(x)}
\end{aligned}
$$

Continuous Conditional Distributions

Let X be a measure of time to answer a question

Let E be the event that the user is a human:

$$
\begin{aligned}
P(E|X = x) &= \frac{P(X = x, E)}{P(X = x)} \\
&= \frac{P(X = x|E)P(E)}{P(X = x)} \\
&= \frac{f_X(x|E)P(E)\epsilon_x}{f_X(x)\epsilon_x} \\
&= \frac{f_X(x|E)P(E)}{f_X(x)}
\end{aligned}
$$

Biometric Keystrokes

Let X be a measure of time to answer a question

Let E be the event that the user is a human

What if you don't know normalization term?:

Normal pdf

Prior

$$P(E|X = x) = \frac{f_X(x|E)P(E)}{f_X(x)}$$

???

$$\frac{P(E|X = x)}{P(E^C|X = x)}$$

End Review

Let's play a game!

Roll a dice three times. If I roll a six twice (or more) I win $1 million. Otherwise you win $1 million. What should we charge to play?

$$P(W) = \left(\frac{5}{6}\right)^2 \approx 0.69$$

What if you don't know a probability?

What if you don't know a probability?

Pirate Supply Store, San Francisco

We are going to think of probabilities as random variables!!!

Flip a coin with unknown probability

Flip a coin (n + m) times, comes up with n heads

- We don't know probability X that coin comes up heads

Frequentist (never prior)

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

X is (often) a single value

Bayesian (prior is great)

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

X is a random variable. Leads to a belief distribution which captures confidence

What is your belief that you successfully roll a 6 on my die?

Flip a coin with unknown probability!

Flip a coin (n + m) times, comes up with n heads

- We don't know probability X that coin comes up heads
- Our belief before flipping coins is that: X ~ Uni(0, 1)
- Let N = number of heads
- Given X = x, coin flips independent: (N | X) ~ Bin(n + m, x)

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

Bayesian "posterior" probability distribution

Bayesian "prior" probability distribution

Flip a coin with unknown probability!

Flip a coin (n + m) times, comes up with n heads

- We don't know probability X that coin comes up heads
- Our belief before flipping coins is that: X ~ Uni(0, 1)
- Let N = number of heads
- Given X = x, coin flips independent: (N | X) ~ Bin(n + m, x)

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

Binomial 1

$$= \frac{\binom{n+m}{n} x^n (1-x)^m}{P(N = n)}$$

$$= \frac{\binom{n+m}{n}}{P(N = n)} x^n (1-x)^m$$

Move terms around

$$= \frac{1}{c} \cdot x^n (1-x)^m \qquad \text{where } c = \int_0^1 x^n (1-x)^m dx$$

Flip a coin with unknown probability!

If you start with a $X \sim \text{Uni}(0, 1)$ prior over probability, and observe:

n "successes" and
m "failures"…

Your new belief about the probability is:

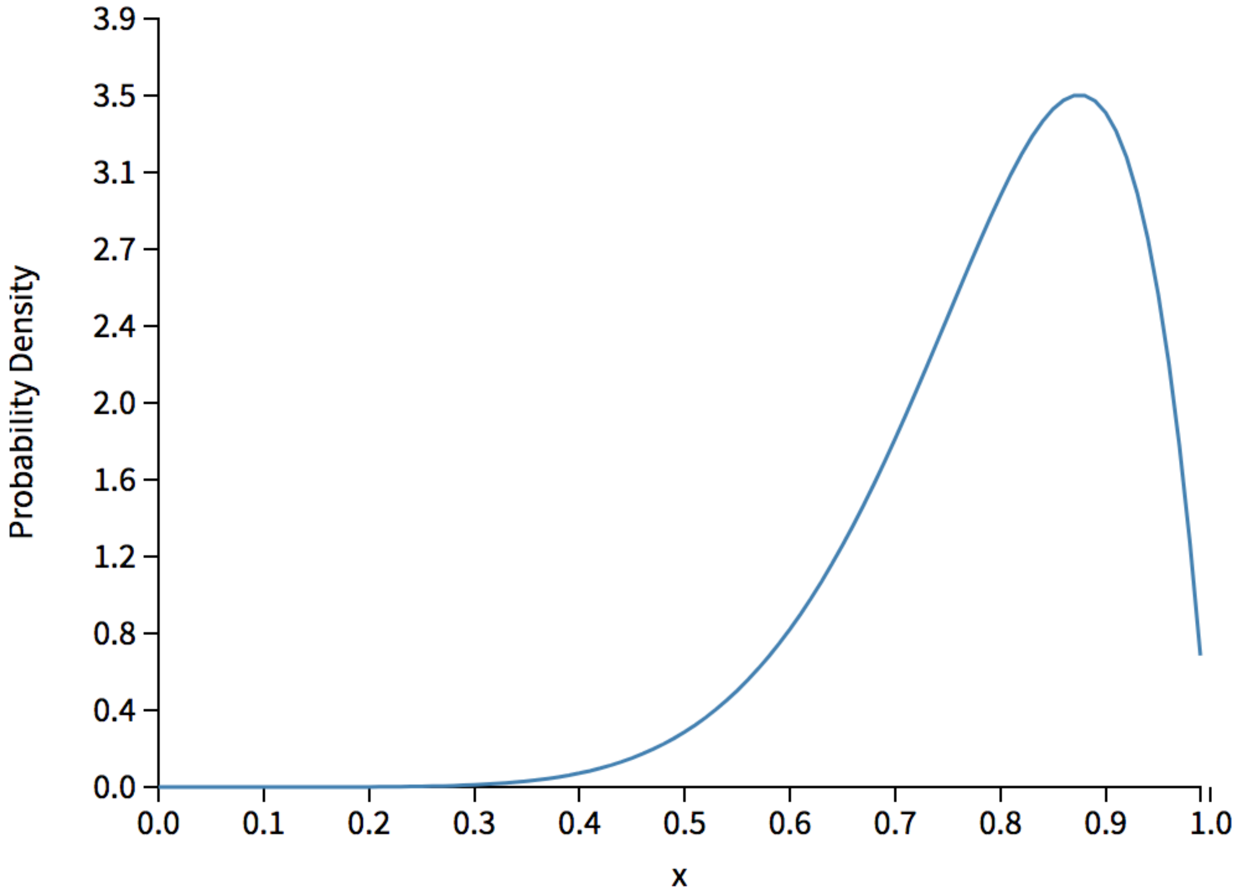
$$f_X(x) = \frac{1}{c} \cdot x^n (1-x)^m$$

where $c = \int_0^1 x^n (1-x)^m$

Belief after 7 success and 1 fail

$$f_X(x) = \frac{1}{c} \cdot x^n (1 - x)^m$$

$n = 7$ $m = 1$



Equivalently!

If you start with a $X \sim \text{Uni}(0, 1)$ prior over probability, and observe:

let a = num “successes” + 1

let b = num “failures” + 1

Your new belief about the probability is:

$$f_X(x) = \frac{1}{c} \cdot x^{a-1}(1-x)^{b-1}$$

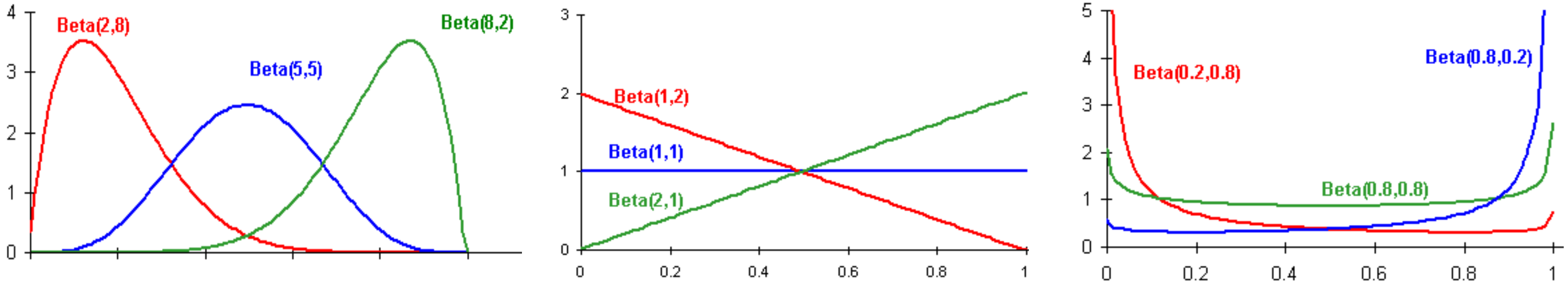
where $$c = \int_0^1 x^{a-1}(1-x)^{b-1}$$

Beta Random Variable

X is a **Beta Random Variable**: X ~ Beta(*a*, *b*)

- Probability Density Function (PDF): (where *a*, *b* > 0)

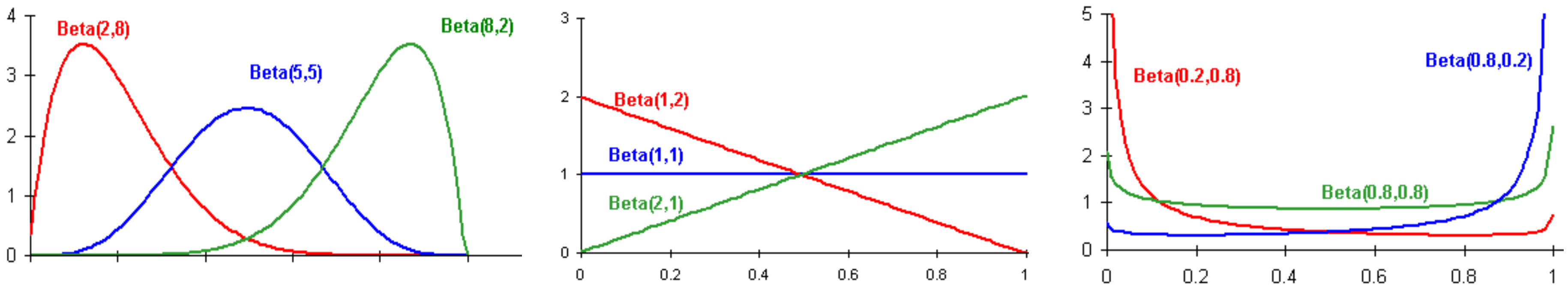
$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \qquad B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$



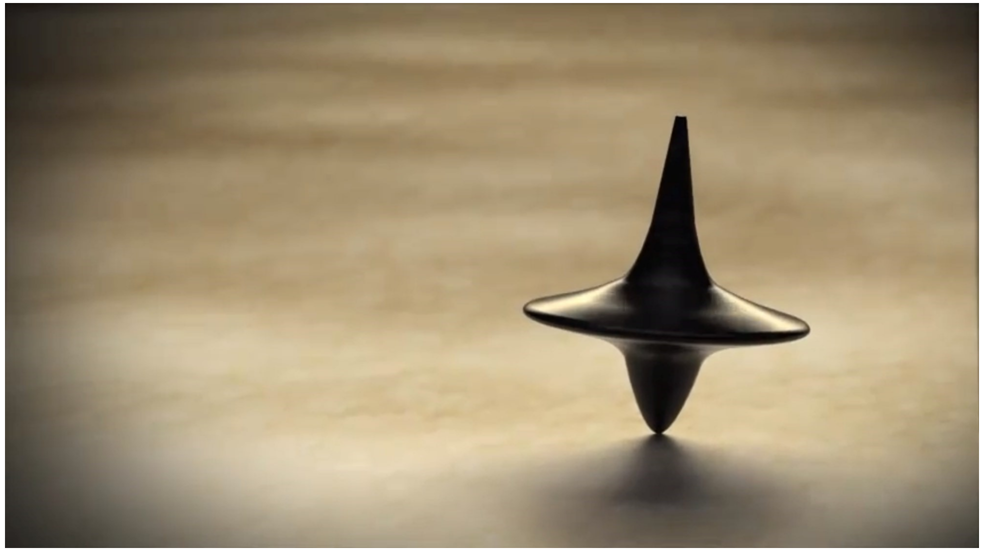
- Symmetric when *a* = *b*

$$E[X] = \frac{a}{a+b} \qquad Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Beta is the Random Variable for Probabilities



Used to represent a distributed belief of a probability



Philosophical Ponderings:
You ask about the probability of rain tomorrow.

Person A: My leg itches when it rains and its kind of itchy…. Uh, p = .80

Person B: I have done complex calculations and have seen 10,451 days like tomorrow… p = 0.80

What is the difference between the two estimates?

Beta is a distribution for probabilities. Its range is values between 0 and 1

Beta Parameters *can* come from experiments:

a = “successes” + 1

b = “failures” + 1

Back to Flipping Coins!

Flip a coin (n + m) times, comes up with n heads

- We don't know probability X that coin comes up heads
- Our belief before flipping coins is that: X ~ Uni(0, 1)
- Let N = number of heads
- Given X = x, coin flips independent: (N | X) ~ Bin(n + m, x)

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

$$= \frac{\binom{n+m}{n} x^n (1-x)^m}{P(N = n)}$$

$$= \frac{\binom{n+m}{n}}{P(N = n)} x^n (1-x)^m$$

$$= \frac{1}{c} \cdot x^n (1-x)^m \qquad \text{where } c = \int_0^1 x^n (1-x)^m dx$$

A beta understanding

X | (N = n, M = m) ~ Beta(a = n + 1, b = m + 1)

- Prior X ~ Uni(0, 1)

N successes

M failures

- Check this out, boss:
  - Beta(a = 1, b = 1) =?

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} = \frac{1}{B(a,b)} x^{0}(1-x)^{0}$$

$$= \frac{1}{\int_0^1 1\,dx} 1 = 1 \quad \text{where} \quad 0 < x < 1$$

  - Beta(a = 1, b = 1) = Uni(0, 1)

- So, prior X ~ Beta(a = 1, b = 1)

If the Prior was Beta?

X is our random variable for probability

If our **prior belief** about X was beta

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

What is our **posterior belief** about X after observing *n* heads (and *m* tails)?

$$f(X = x | N = n) = ???$$

If the Prior was Beta?

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

$$= \frac{\binom{n+m}{n}x^n(1-x)^m f(X=x)}{P(N=n)}$$

$$= \frac{\binom{n+m}{n}x^n(1-x)^m \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}}{P(N=n)}$$

$$= K_1 \cdot \binom{n+m}{n} x^n(1-x)^m \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

$$= K_3 \cdot x^n(1-x)^m x^{a-1}(1-x)^{b-1}$$

$$= K_3 \cdot x^{n+a-1}(1-x)^{m+b-1}$$

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

A beta understanding

- If "Prior" distribution of X (before seeing flips) is Beta

- Then "Posterior" distribution of X (after flips) is Beta

Beta is a **conjugate** distribution for Beta

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

A beta understanding

Can set X ~ Beta(*a*, *b*) as prior to reflect how biased you think coin is apriori

- This is a subjective probability (aka Bayesian)!
- Prior probability for X based on seeing (*a* + *b* – 2) "imaginary" trials, where

  (*a* – 1) of them were heads.

  (b – 1) of them were tails.
- Beta(1, 1) = Uni(0, 1) → we haven't seen any "imaginary trials", so apriori know nothing about coin

Update to get posterior probability

- X | (n heads and m tails) ~ Beta(a + n, b + m)

Enchanted Die

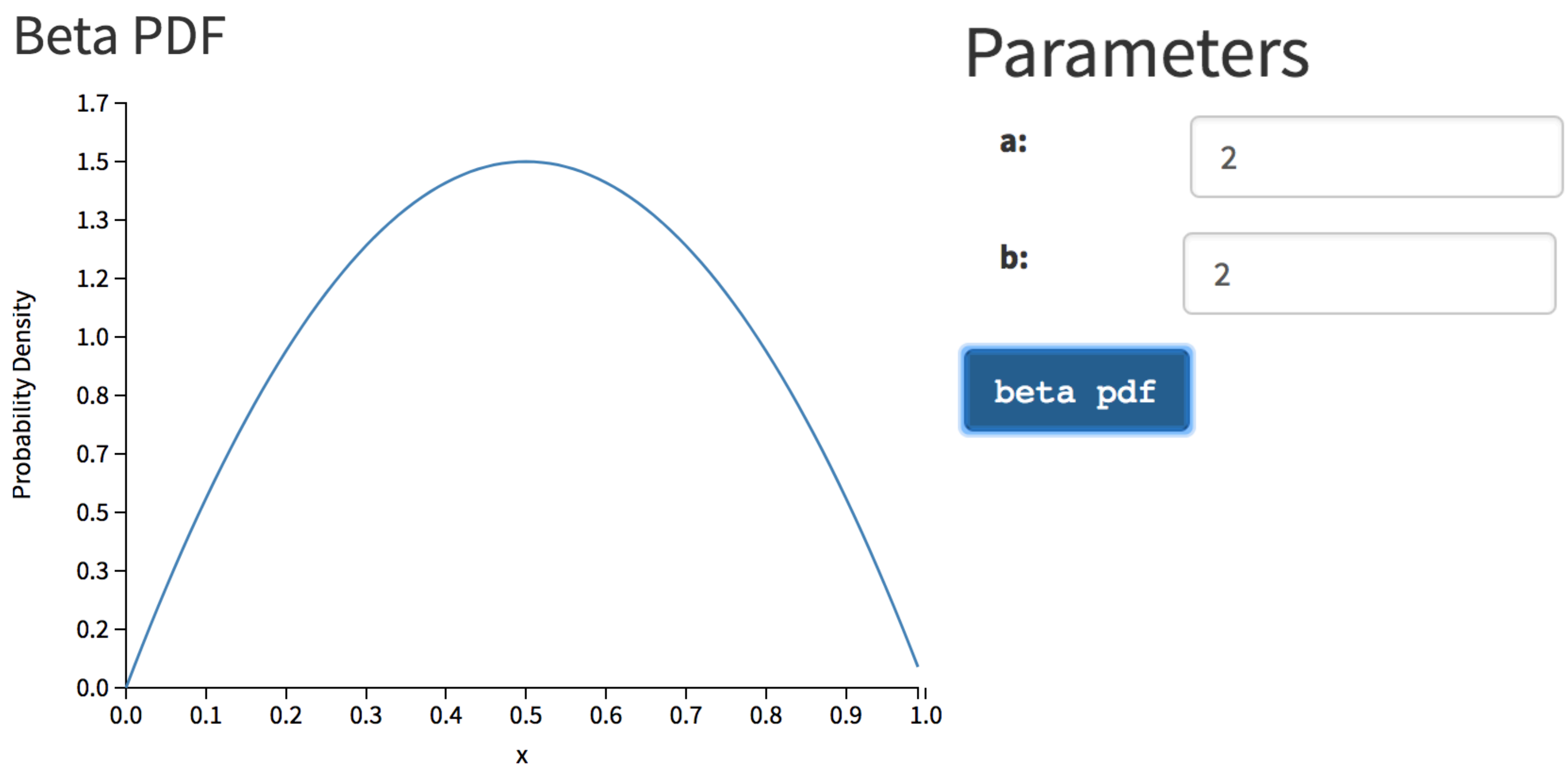
Let X be the probability of rolling a “6” on Chris’ die.

Prior: Imagine 5 die rolls where only showed up as a “6”

Observation: Roll it a few times…

What is the updated probability density function of X after our observations?

Check out the Demo!



Damn

A beta example

Before being tested, a medicine is believed to “work” about 80% of the time. The medicine is tried on 20 patients. It “works” for 14 and “doesn’t work” for 6. What is your new belief that the drug works?

Frequentist:

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

A beta example

Before being tested, a medicine is believed to "work" about 80% of the time. The medicine is tried on 20 patients. It "works" for 14 and "doesn't work" for 6. What is your new belief that the drug works?

Bayesian: $X \sim \text{Beta}$

Prior:	Interpretation:
$X \sim \text{Beta}(a = 81, b = 21)$	80 successes / 100 trials
$X \sim \text{Beta}(a = 9, b = 3)$	8 successes / 10 trials
$X \sim \text{Beta}(a = 5, b = 2)$	4 successes / 5 trials

A beta example

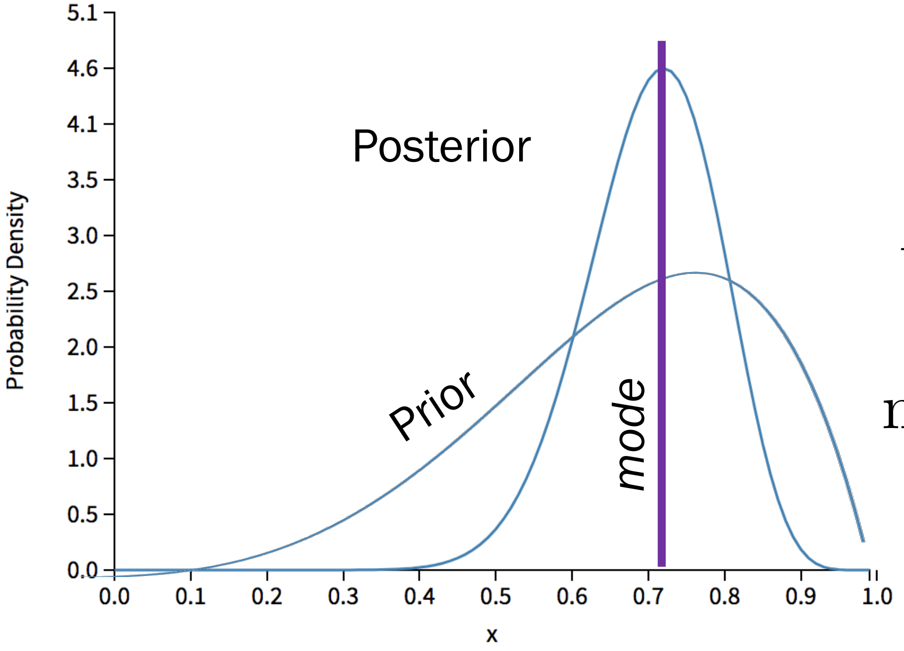
Before being tested, a medicine is believed to "work" about 80% of the time. The medicine is tried on 20 patients. It "works" for 14 and "doesn't work" for 6. What is your new belief that the drug works?

Bayesian: $X \sim \text{Beta}$

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

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

$$\sim \text{Beta}(a = 19, b = 8)$$



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

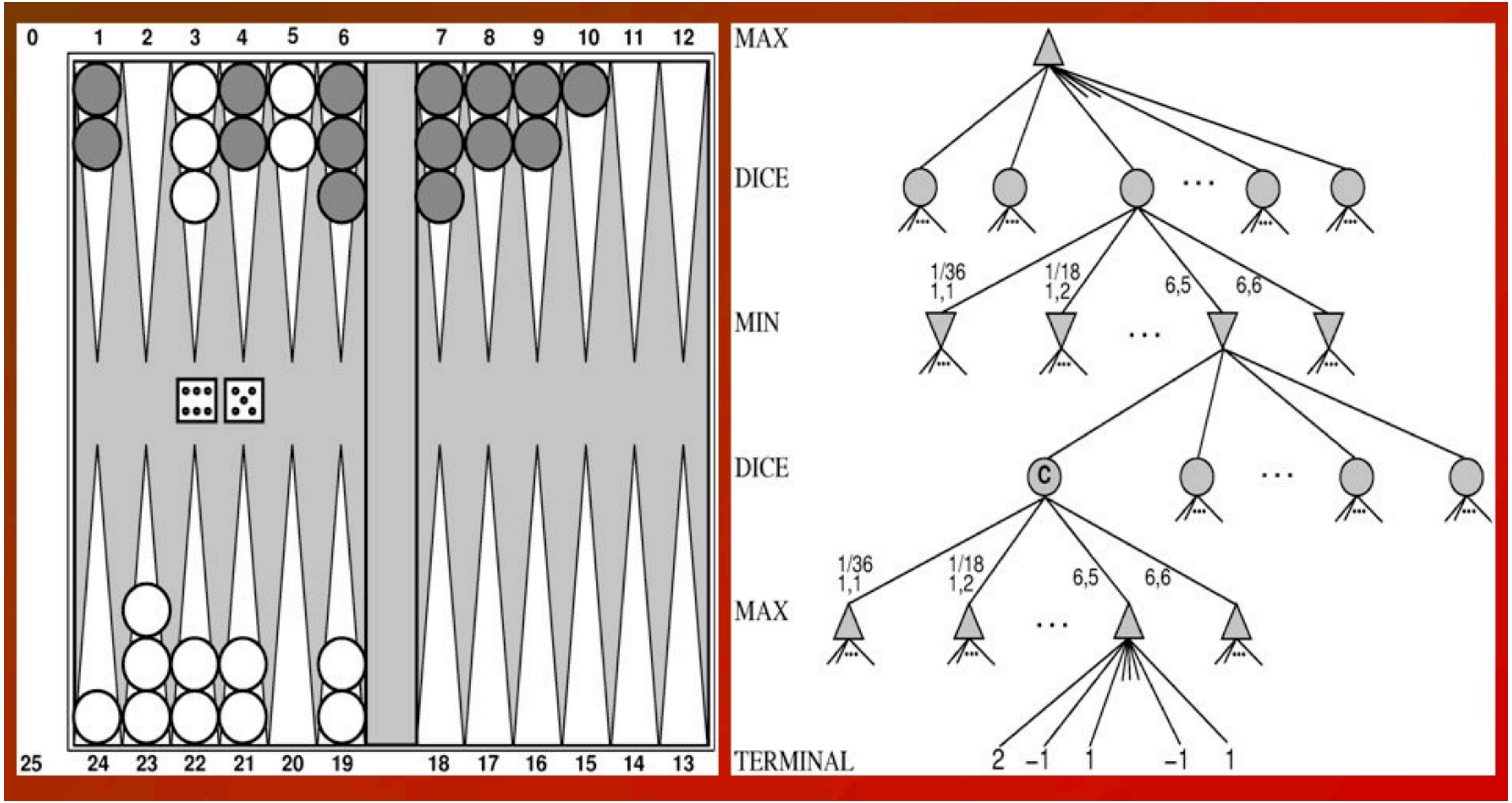
$$\text{mode}(X) = \frac{a-1}{a+b-2}$$

$$= \frac{19}{18+7} \approx 0.72$$

Next level?

Alpha GO mixed deep learning and core reasoning under uncertainty

Multi Armed Bandit



Multi Armed Bandit

Drug A

Drug B

Which one do you give to a patient?

Lets Play!

Drug A

Drug B

Which one do you give to a patient?

Lets Play!

sim.py

```python
import pickle
import random

def main():
  X1, X2  = pickle.load(open('probs.pkl', 'rb'))

  print("Welcome to the drug simulator. There are two drugs")

  while True:
    choice = getChoice()
    prob = X1 if choice == "a" else X2
    success = bernoulli(prob)
    if success:
      print('Success. Patient lives!')
    else:
      print('Failure. Patient dies!')
    print('')
```

Optimal Decision Making

You try drug B, 5 times. It is successful 2 times.

If you had a uniform prior, what is your posterior belief about the likelihood of success?

---

2 successes

3 failures

$$X \sim \text{Beta}(a = 3, b = 4)$$

Optimal Decision Making

You try drug B, 5 times. It is successful 2 times.
X is the probability of success.

$$X \sim \text{Beta}(a = 3, b = 4)$$

What is expectation of X?

$$E[X] = \frac{a}{a+b} = \frac{3}{3+4} \approx 0.43$$

Optimal Decision Making

You try drug B, 5 times. It is successful 2 times.
X is the probability of success.

$$X \sim \text{Beta}(a = 3, b = 4)$$

What is the probability that $X > 0.6$

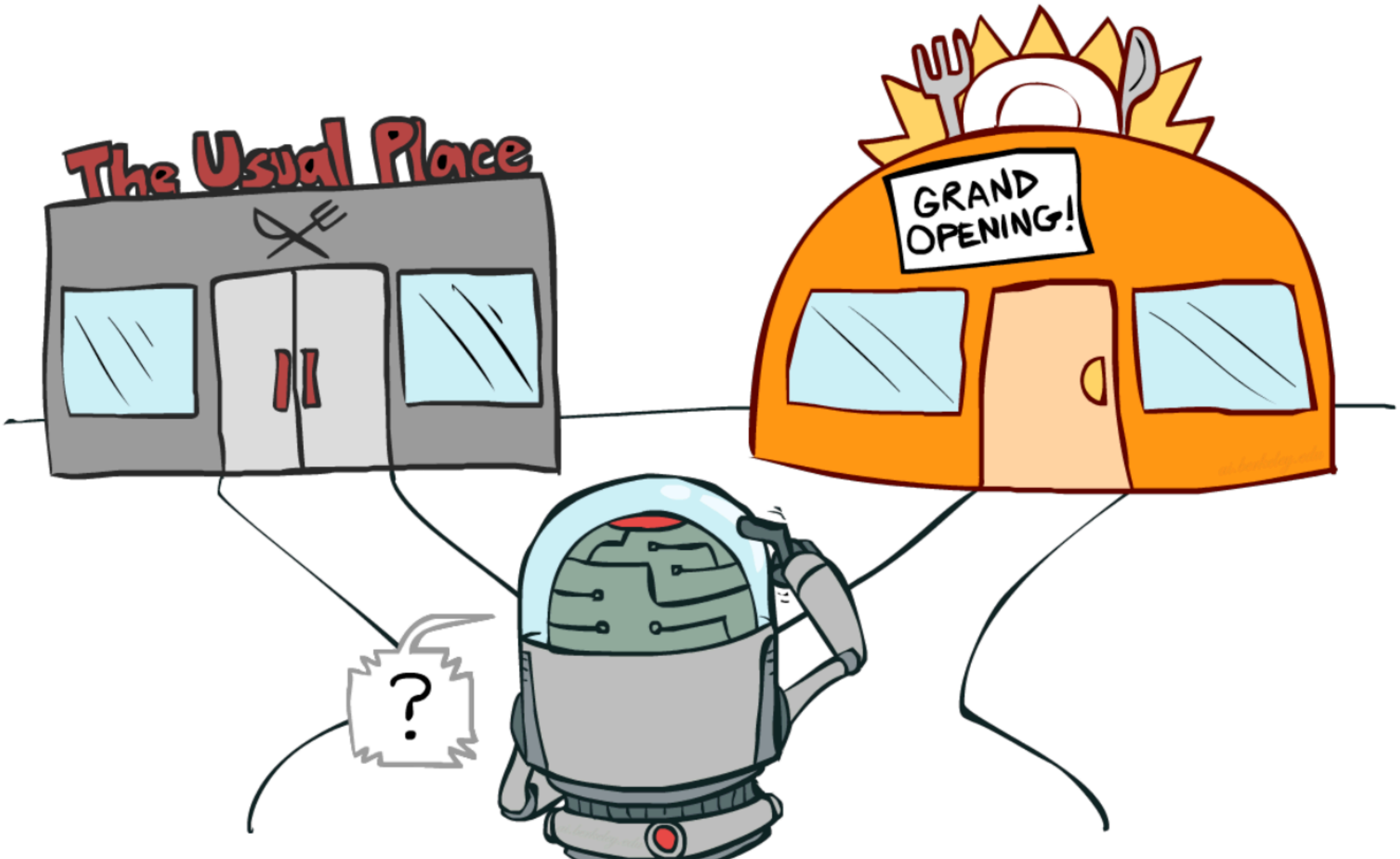
$$P(X > 0.6) = 1 - P(X < 0.6) = 1 - F_X(0.6)$$

Wait what? Chris are you holding out on me?

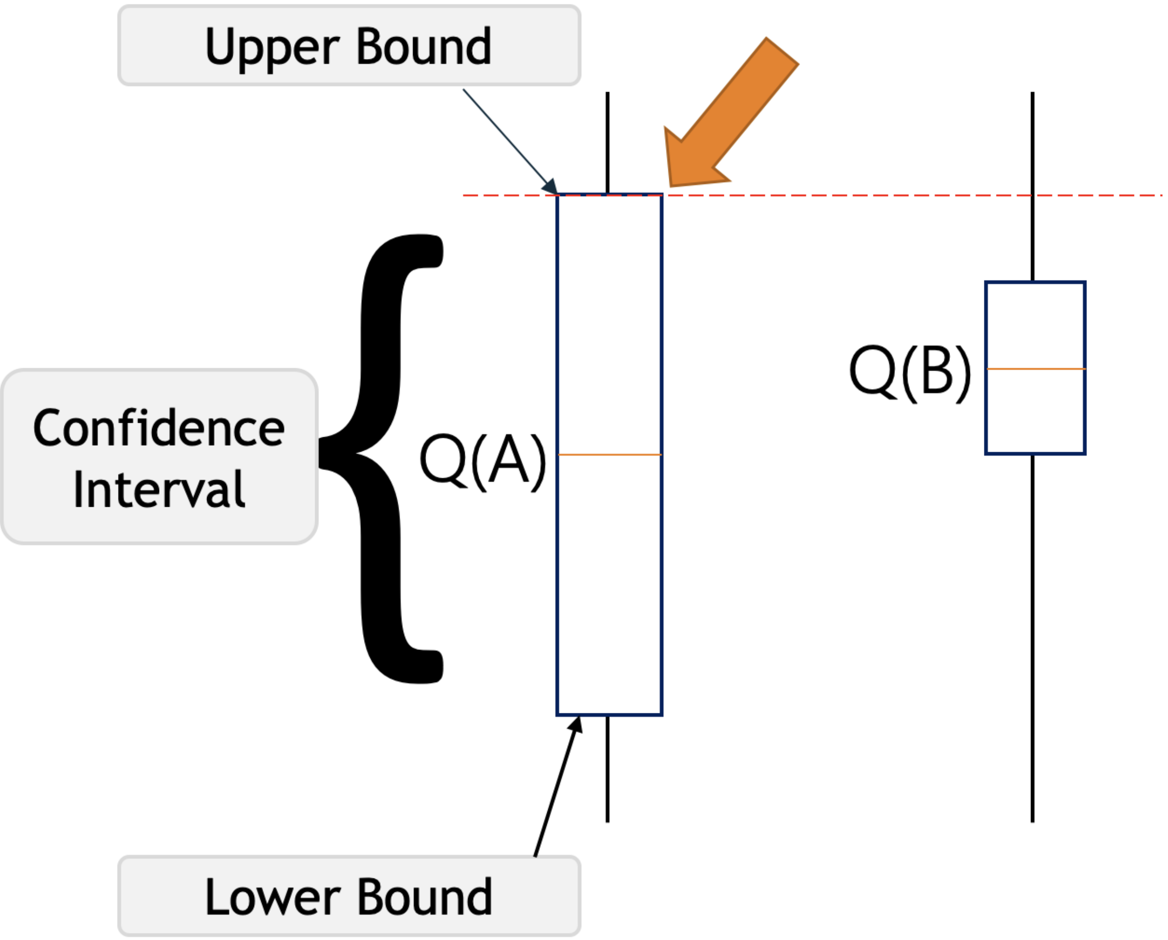
```
stats.beta.cdf(x, a, b)
```

$$P(X > 0.6) = 1 - F_X(0.6) = 0.1792$$

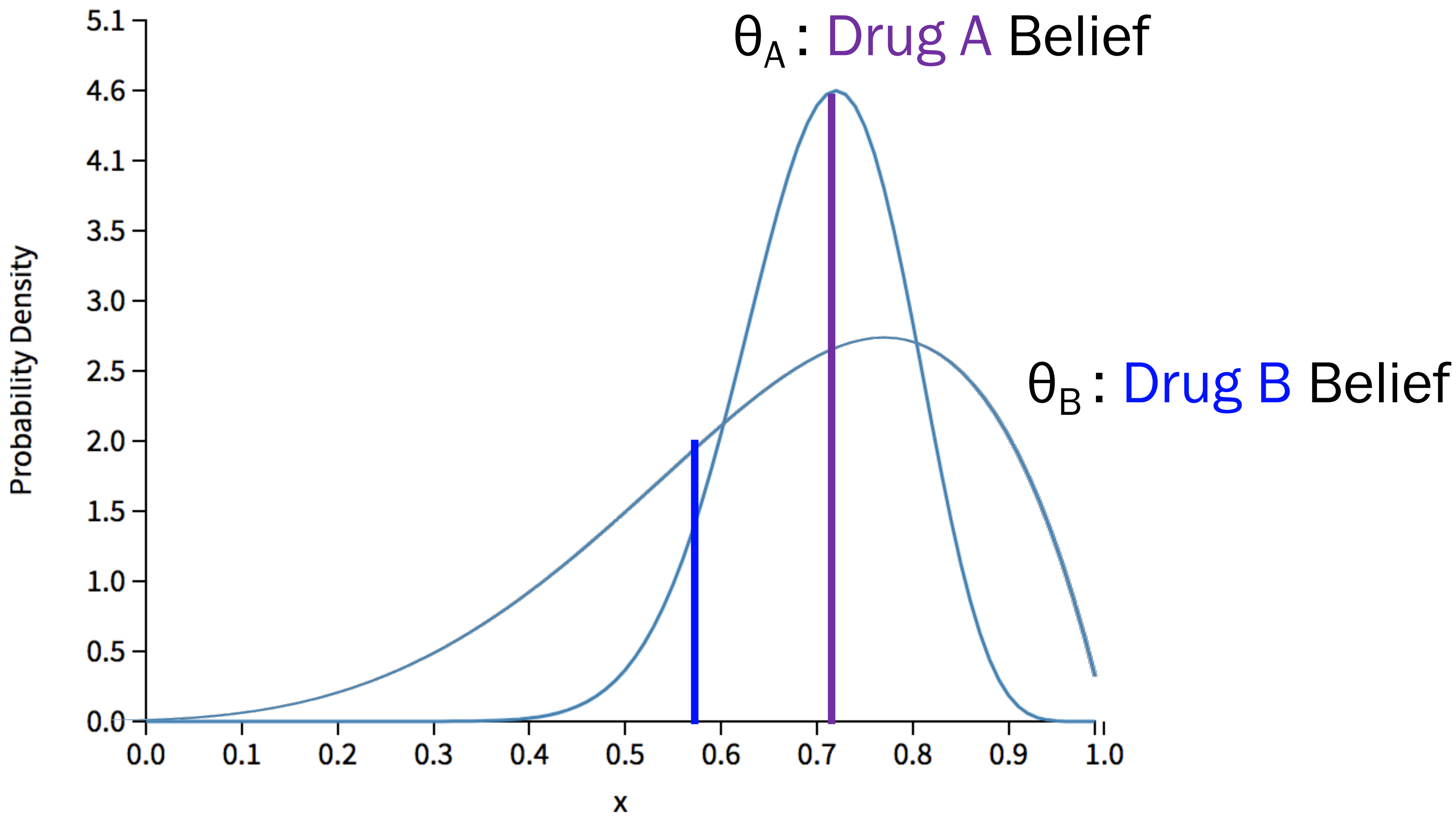
Explore something new? Or go for what looks good now?



One option: Upper Confidence Bound



Amazing option: Thompson Sampling

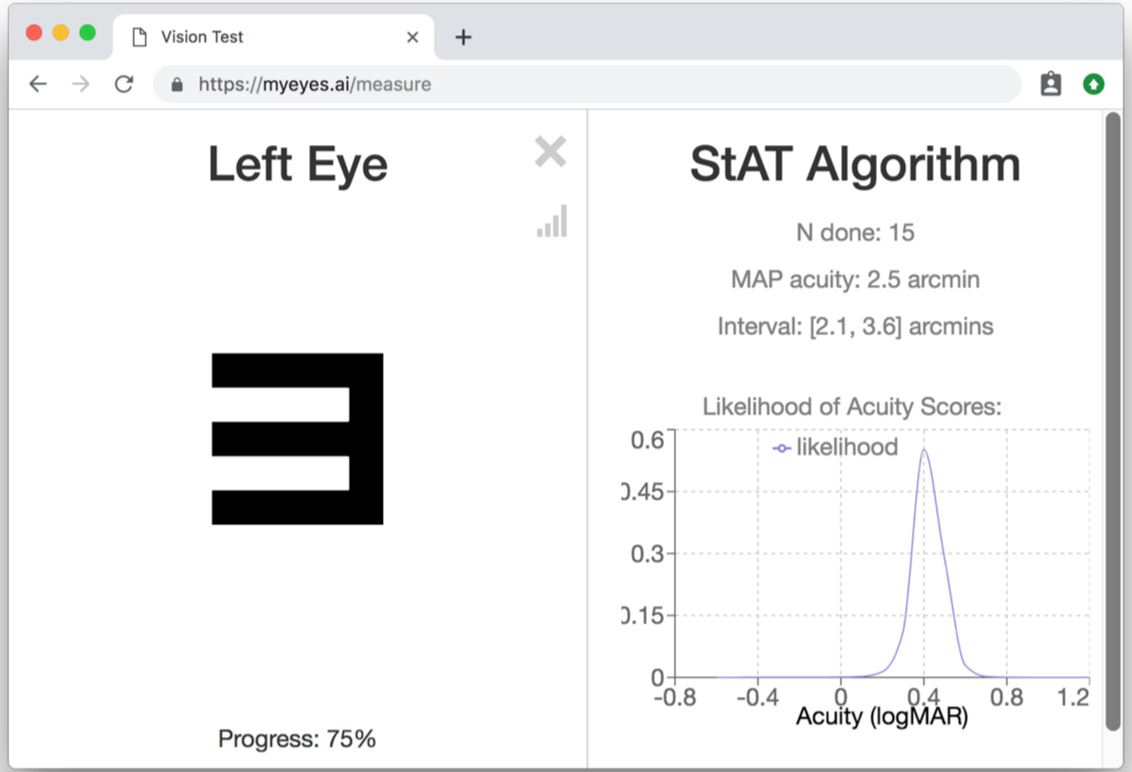


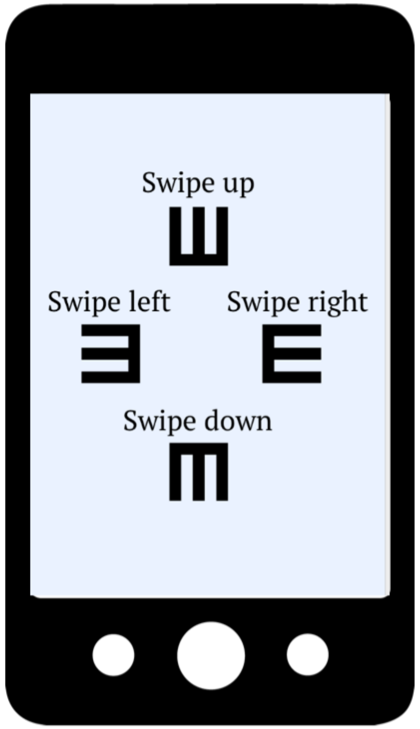
Probability that you chose drug A? Make it $\Pr(\theta_a > \theta_b)$

Stanford Acuity Test

1 Take an eye exam on this website

2 Connect your phone



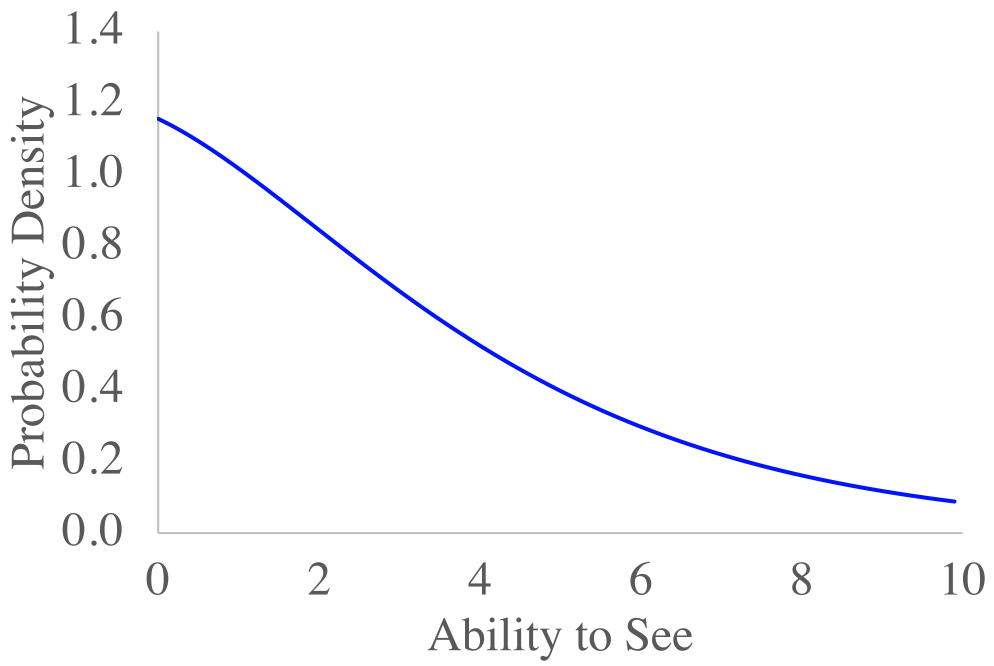


3 Visualize the math

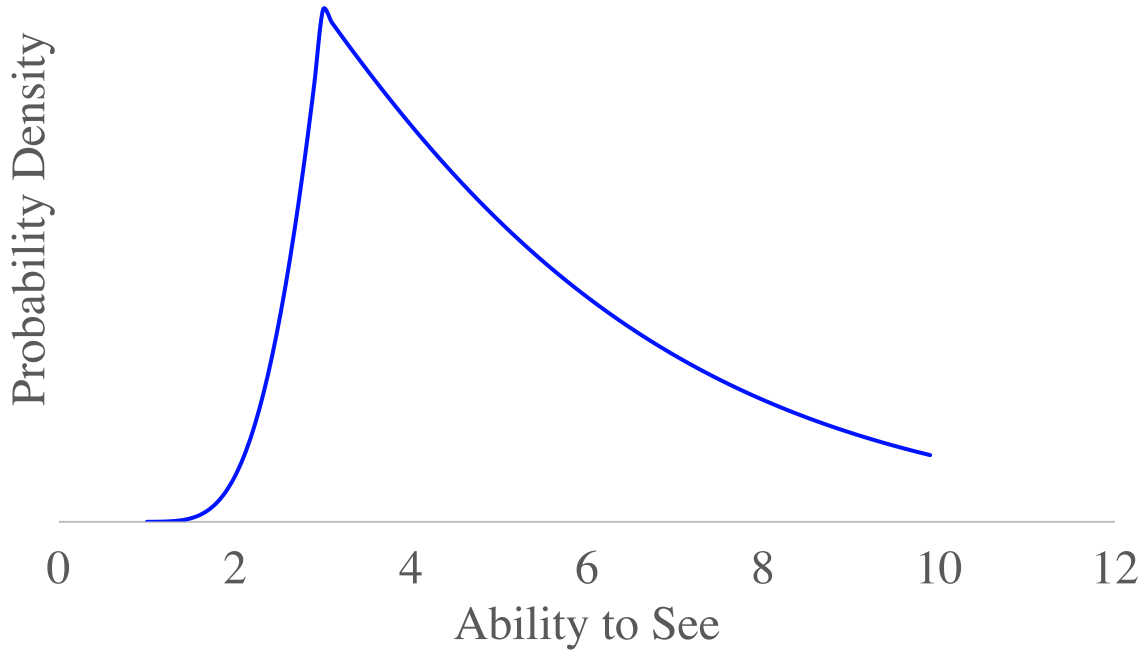
An Updated Belief

A user is shown a letter at **font size 3** and gets it **wrong**.
What is your new belief that their **visual ability is 3**?

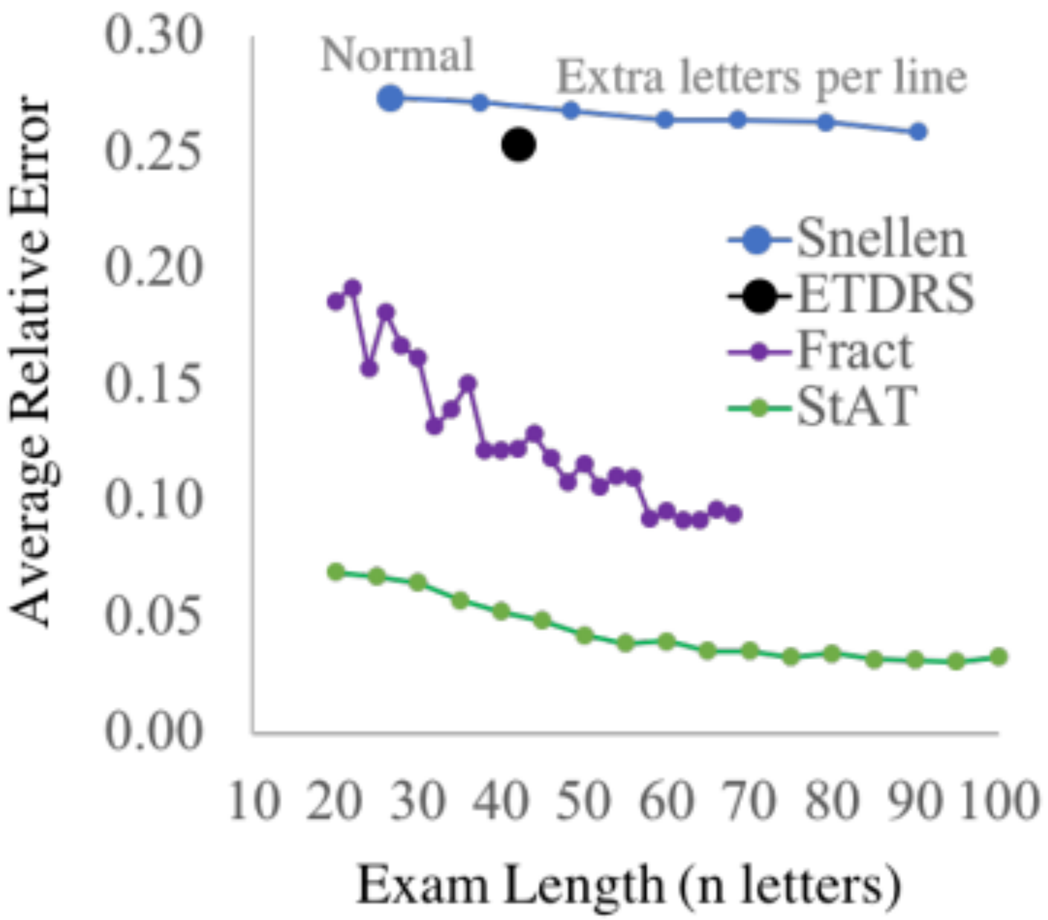
Prior



Posterior



Thompson Sampling belongs to a family called Optimistic



Actual model also included
+ a probability of "slip"
+ an intelligent algorithm for choosing the next letter size

Beta:
The probability density for probabilities

Beta is a distribution for probabilities

Beta Distribution

If you start with a $X \sim \text{Uni}(0, 1)$ prior over probability, and observe:

let a = num "successes" + 1

let b = num "failures" + 1

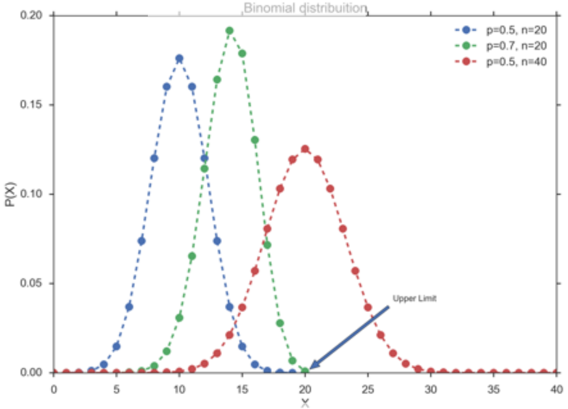
Your new belief about the probability is:

$$f_X(x) = \frac{1}{c} \cdot x^{a-1}(1-x)^{b-1}$$

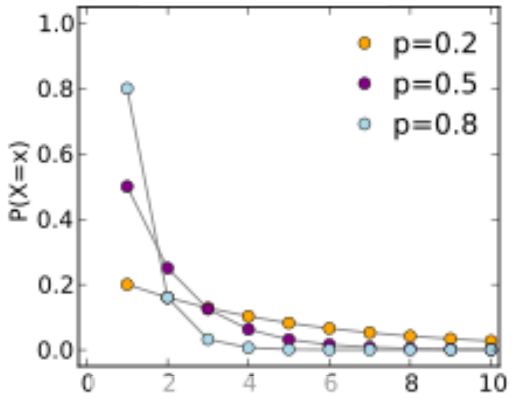
$$\text{where} \quad c = \int_0^1 x^{a-1}(1-x)^{b-1}$$

Distributions

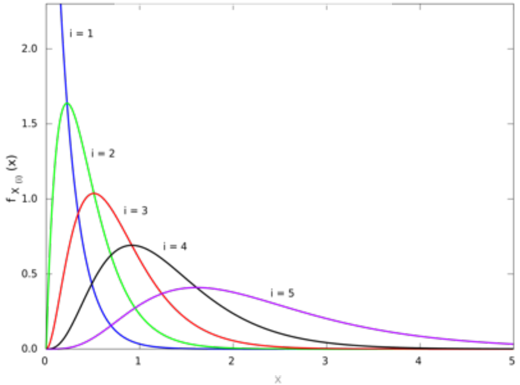
Binomial



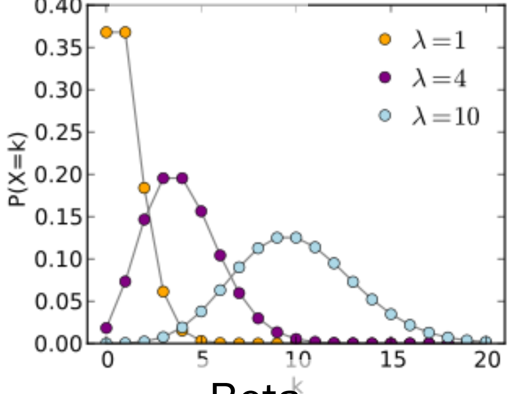
Geometric



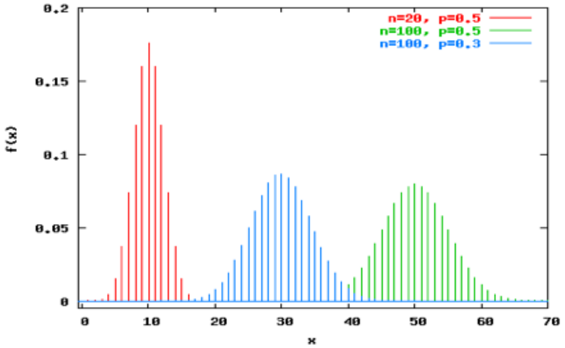
Exponential



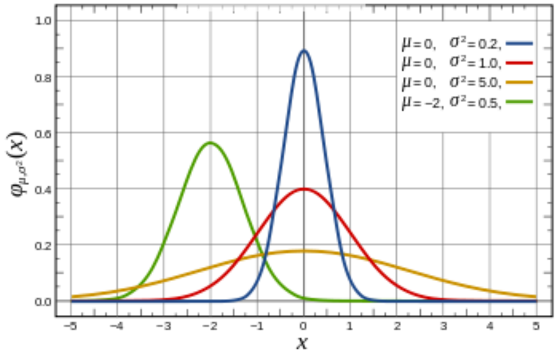
Poisson



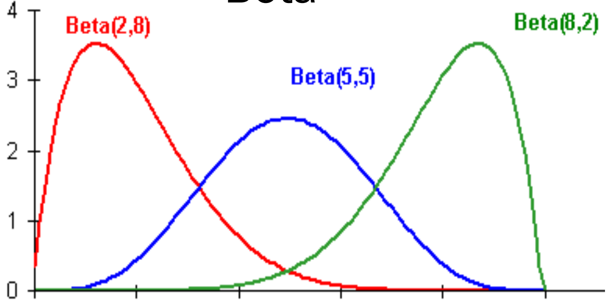
Neg Binomial



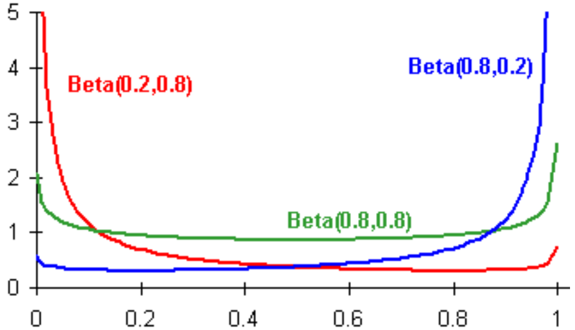
Normal



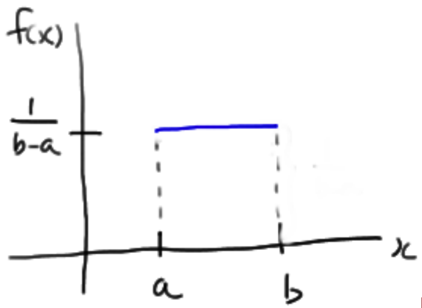
Beta



Beta



Uniform



Grades must be bounded

Normal: No

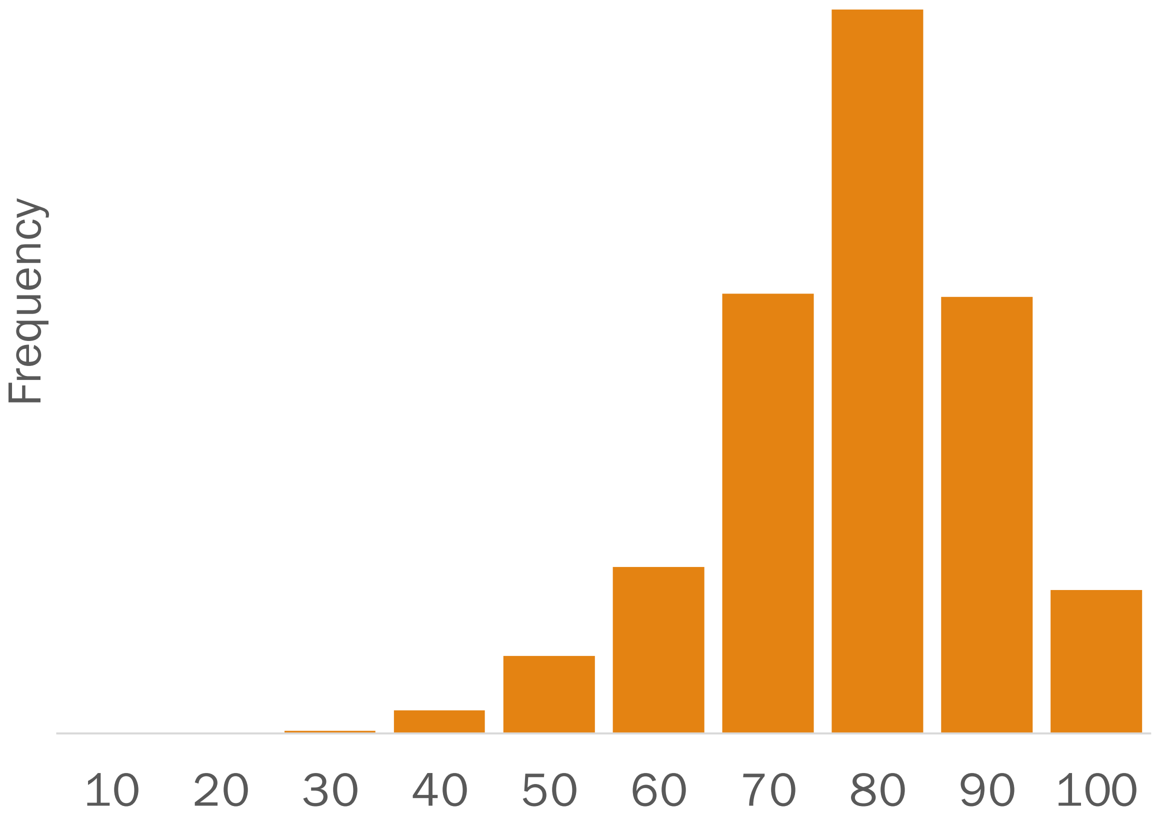
Poisson: No

Exponential: No

Beta: Looks Good!

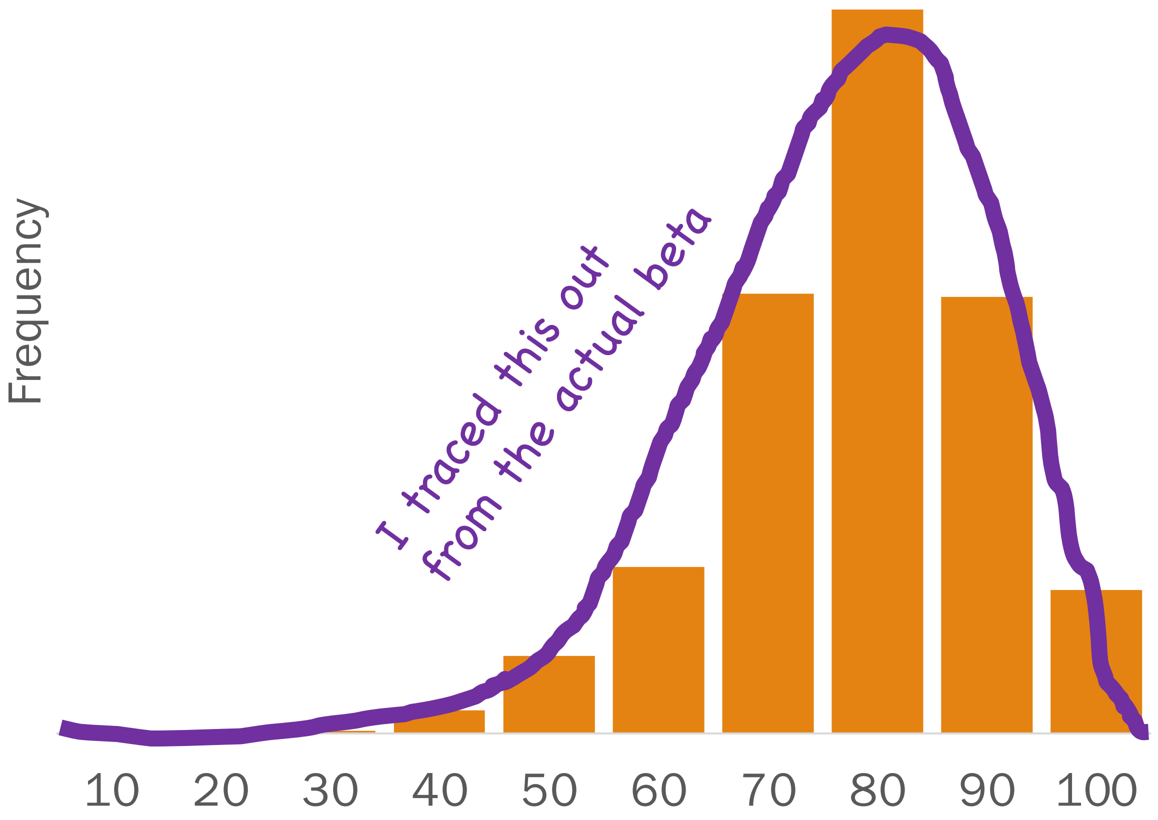
Assignment Grades Demo

```
Assignment id = '1613'
```



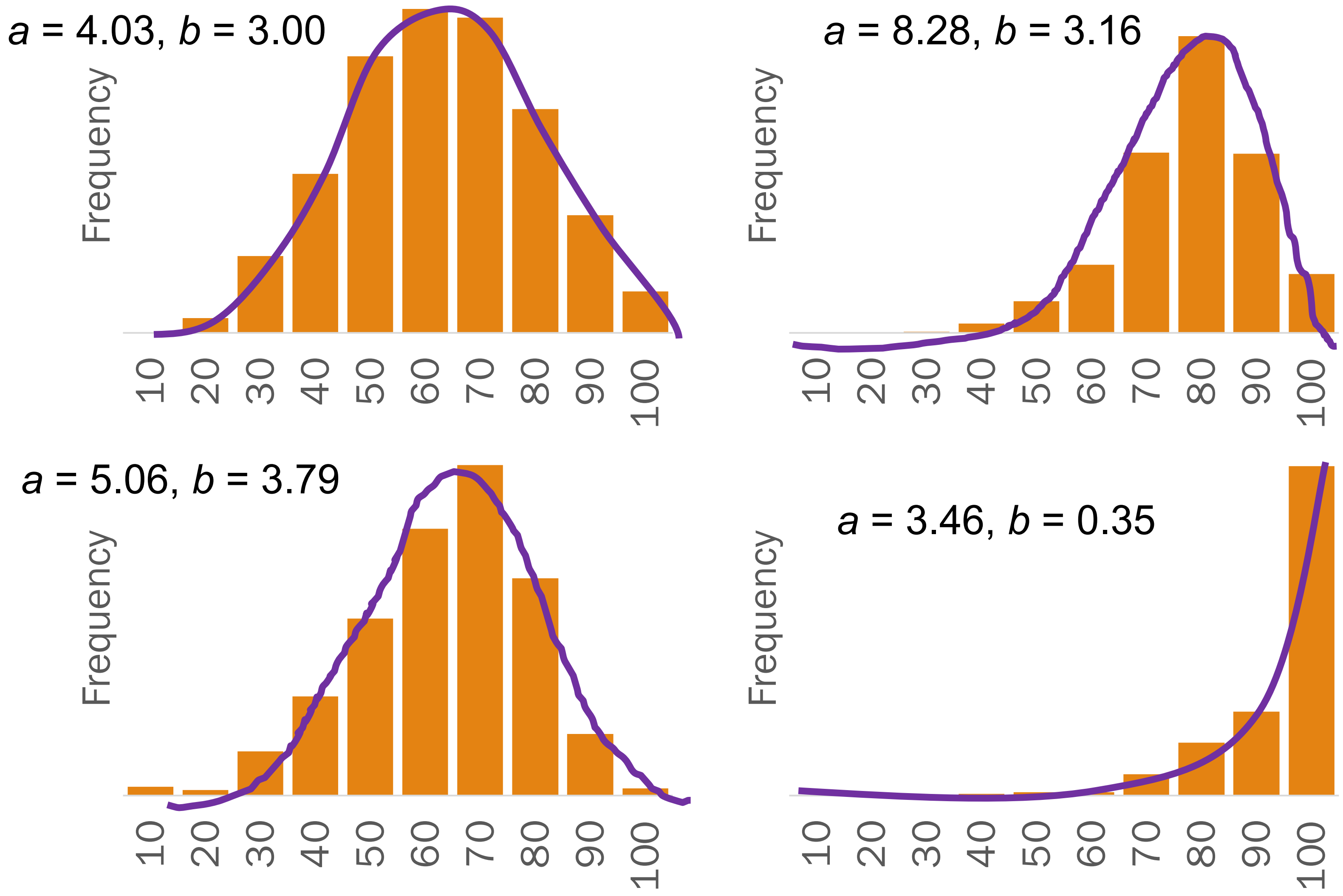
Assignment Grades Demo

```
Assignment id = '1613'
```



$$X \sim Beta(a = 8.28, b = 3.16)$$

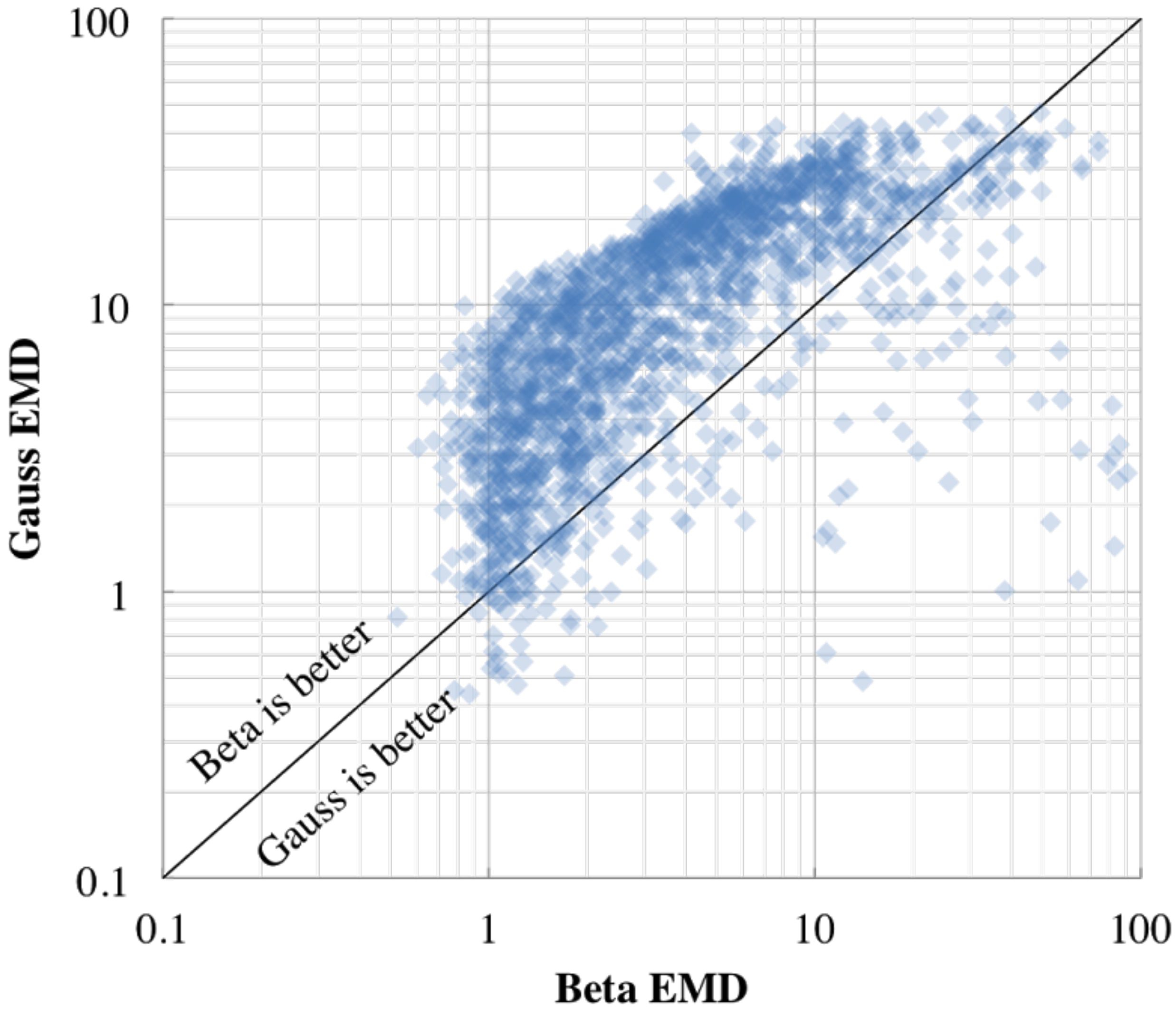
Assignment Grades



We have 2055 assignment distributions from grade scope

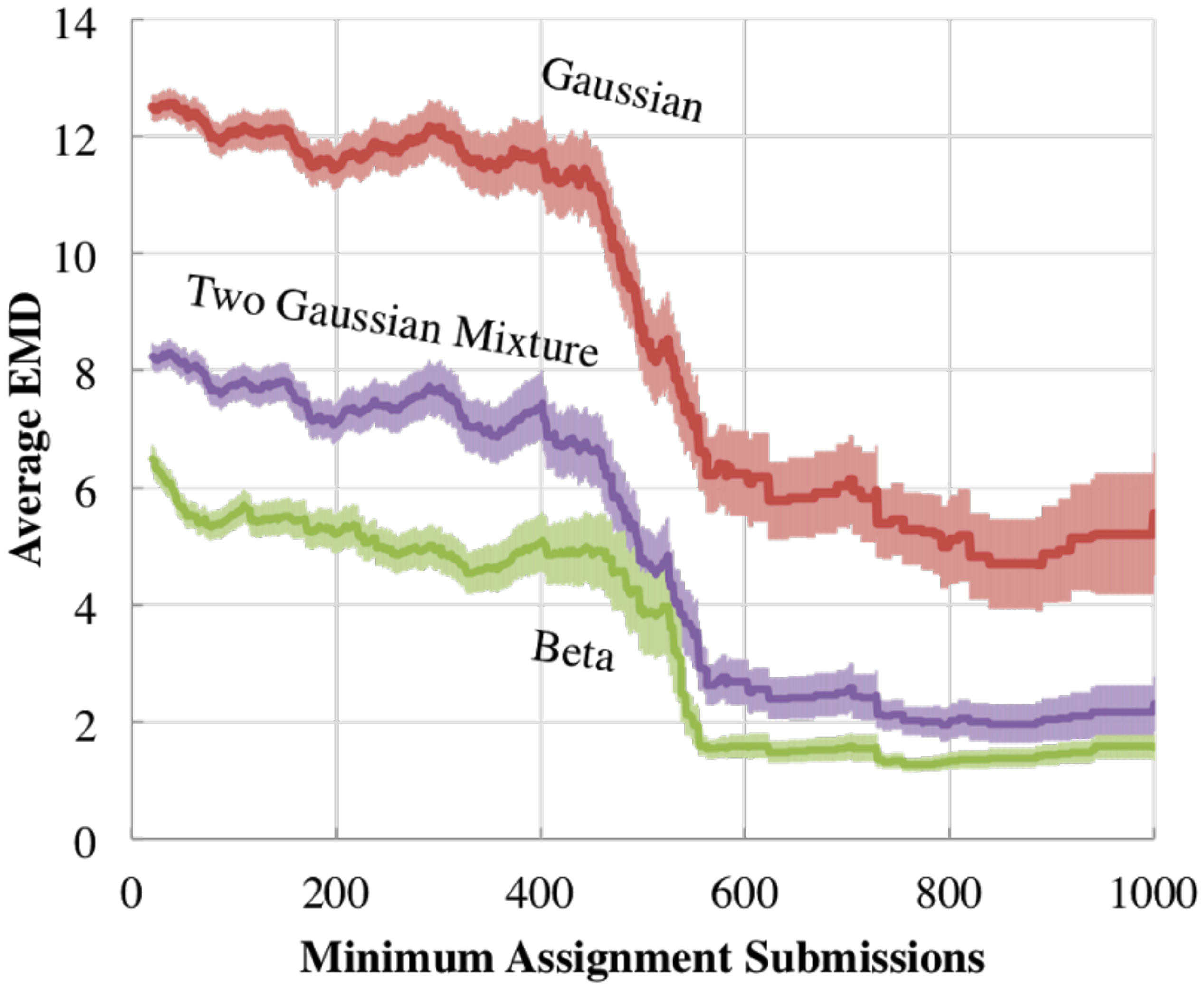
Stanford University

Beta is a Better Fit



Unpublished results. Based on Gradescope data

Beta is a Better Fit For All Class Sizes

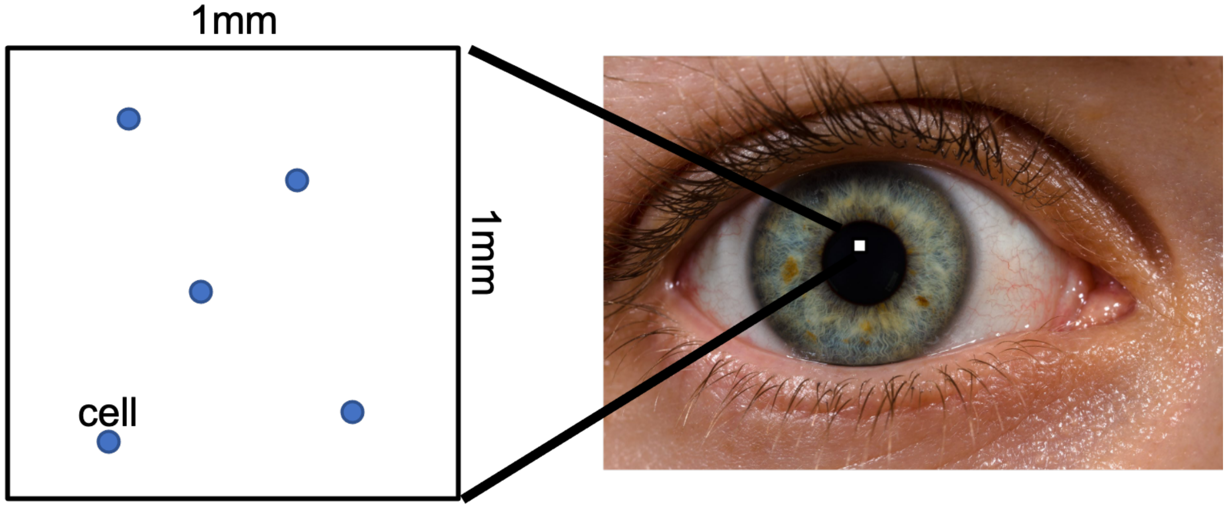


Unpublished results. Based on Gradescope data

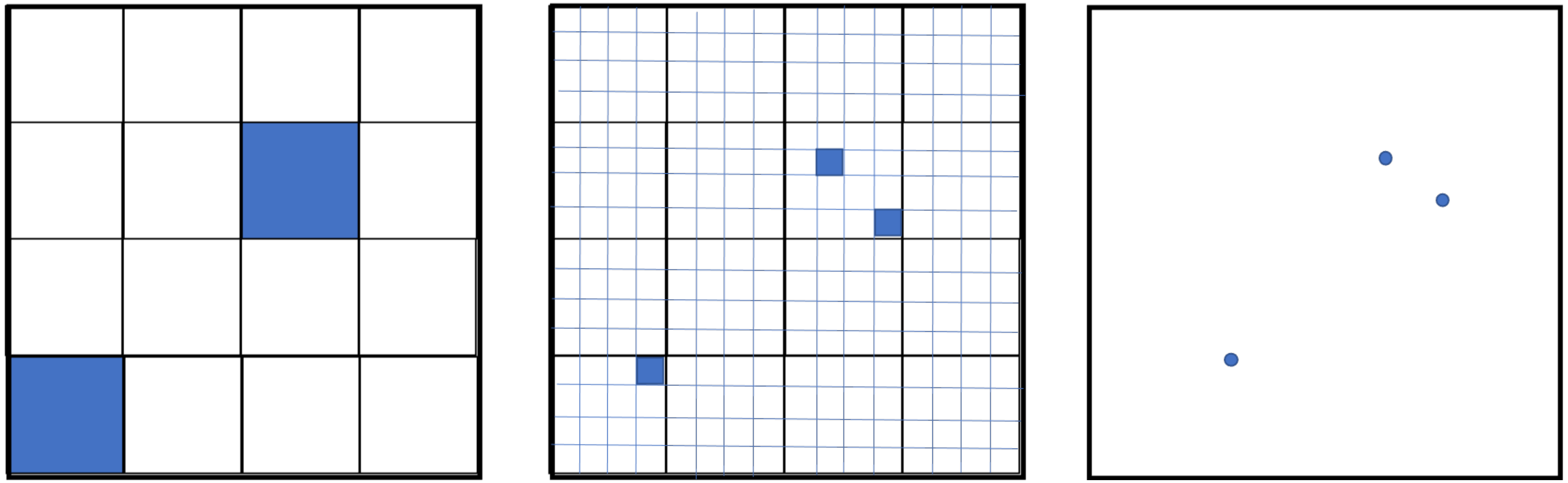
Any parameter for a "parameterized" random variable can be thought of as a random variable.

Eg: $X \sim N(\mu, \sigma^2)$

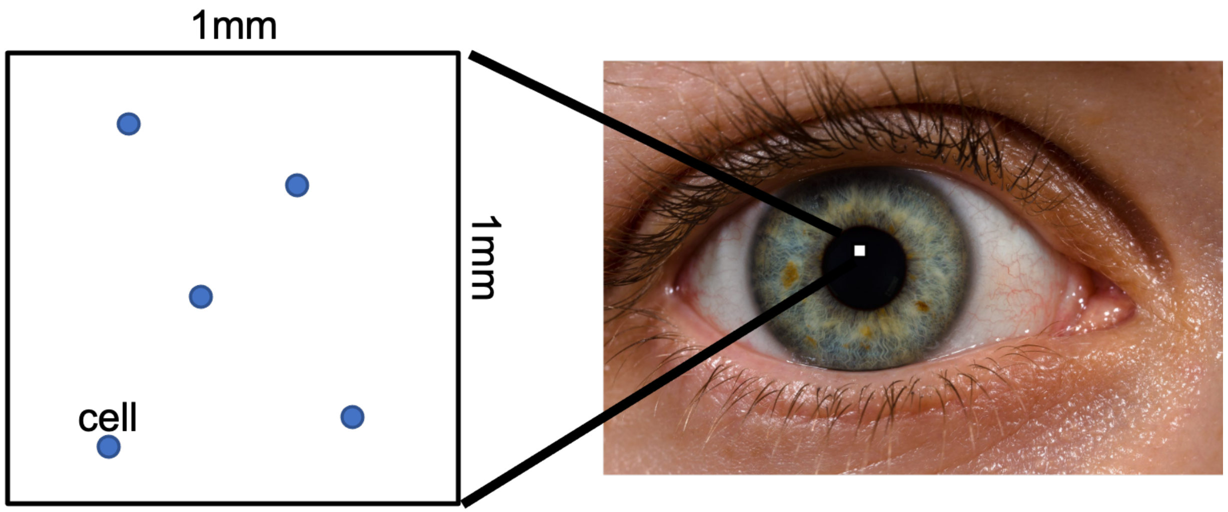
Better Measure for Eye Disease: Counting Cells in Space



$X \sim \text{Bin}(n = 16, p = \lambda/16)$ $X \sim \text{Bin}(n = 256, p = \lambda/256)$ $X \sim \lim_{n\to\infty} \text{Bin}(n, p = \lambda/n)$



Better Measure for Eye Disease: Counting Cells in Space



On the exam: True lambda is 5, what is the probability of observing 4 cells?

Next level: You observe 4 cells, what is the distribution of belief over the true average?

Wow level: One day you observe 4 cells, two days later you observe 5. What is your belief that the patient actually got worse?

Random Variables for Parameters

Parameter	Chosen Distribution
Bernoulli p	Beta
Poisson λ	Gamma
Normal μ	Normal
Normal σ^2	Gamma
Beta α	Gone too far…