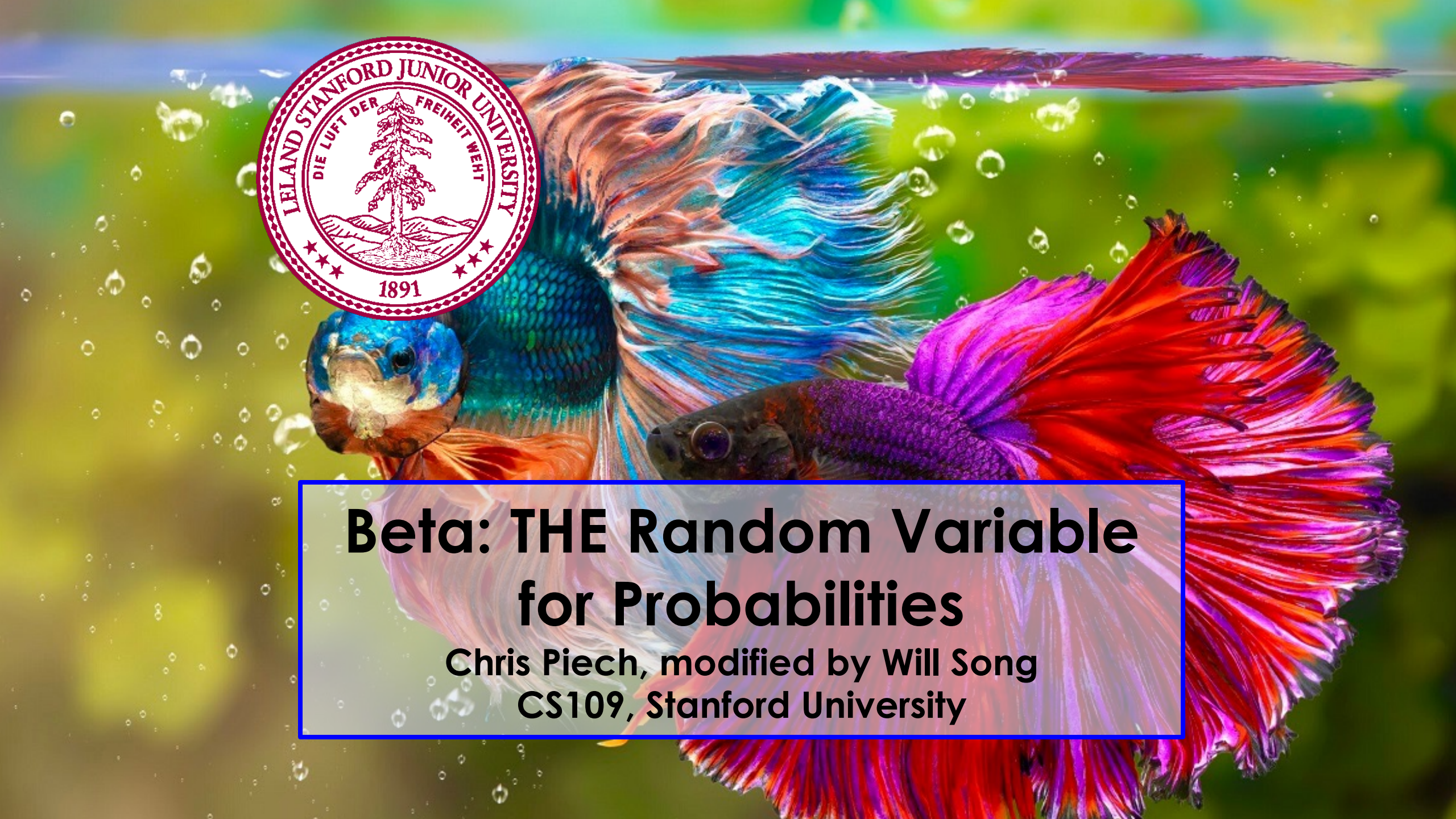


Beta: THE Random Variable for Probabilities

Chris Piech, modified by Will Song
CS109, Stanford University

Which mochi are you more likely to like?

10,000 50

10 0

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"

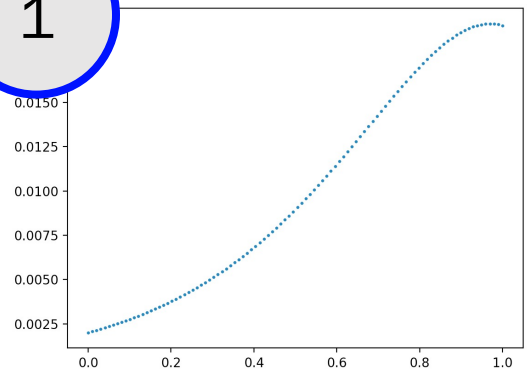
- CS109

Today we are going to learn how to quantify uncertainty.

Review

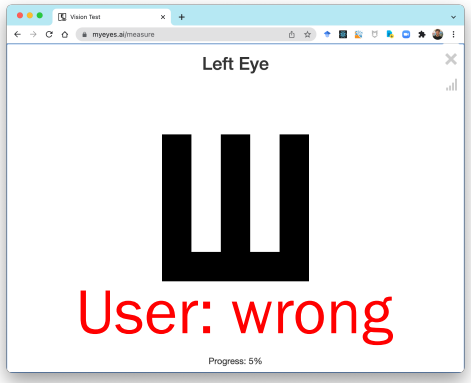
Inference on a non-bernoulli random variable

1



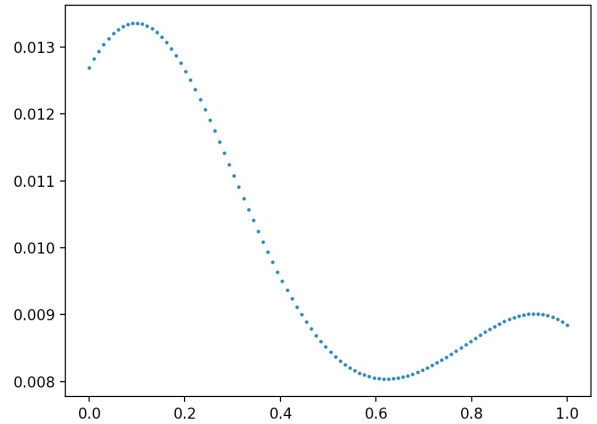
$P(A = a)$

2



Observation $Y = 0$

3



$P(A = a|Y = 0)$

We can perform **inference** when there are two random variables using Bayes!

Inference?

Updated Belief

$$P(A = a|Y = 0) = \frac{P(Y = 0|A = a)P(A = a)}{P(Y = 0)}$$

Likelihood Belief

Normalizer

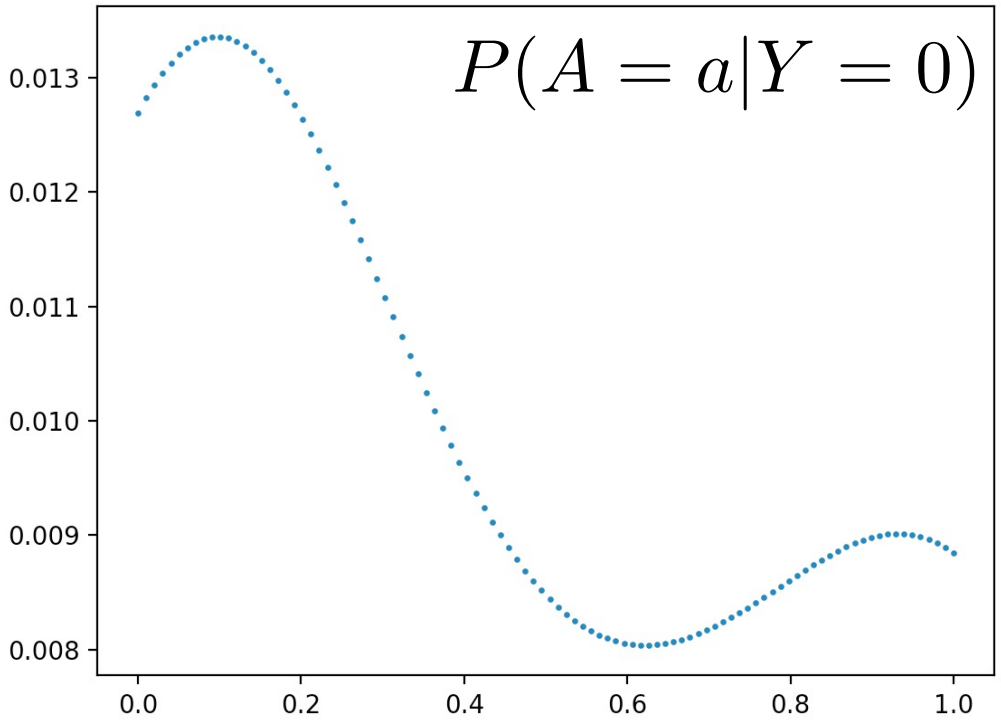
Belief(0.02) = 0.001

Value of a

P(A=a)

Inference on a non-bernoulli random variable

In plain English: run bayes for each value of a




```
# RV bayes as code
def update(belief, obs):
    for a in support:
      prior_a = belief[a]
      likelihood = calc_likelihood(a, obs)
      belief[a] = prior_a * likelihood
    normalize(belief)
```

likelihood

$$P(A = a|Y = 0) = \frac{P(Y = 0|A = a)P(A = a)}{P(Y = 0)}$$

Normalize???

```
# RV bayes as code
def update(belief, obs):
    for a in support:
      prior_a = belief[a]
      likelihood = calc_likelihood(a, obs)
      belief[a] = prior_a * likelihood
    normalize(belief)
```

In plain English: this is the sum of all the things in belief

$$P(A=a|Y=0) = \frac{P(Y=0|A=a)P(A=a)}{P(Y=0)}$$

$$= \frac{P(Y=0|A=a)P(A=a)}{\sum_a P(Y=0, A=a)}$$

$$= \frac{P(Y=0|A=a)P(A=a)}{\sum_a P(Y=0|A=a)P(A=a)}$$

End Review

Where are we in CS109?

Let's play a game!

Flip a plate 5 times. If you get heads 3 times you win

Credit: Rembrandt via Dall E

$$P(X = 3) = \binom{5}{3} \cdot \frac{1}{2}^3 \cdot \frac{1}{2}^2$$
$$= 0.3125$$

What if you don't know a probability?

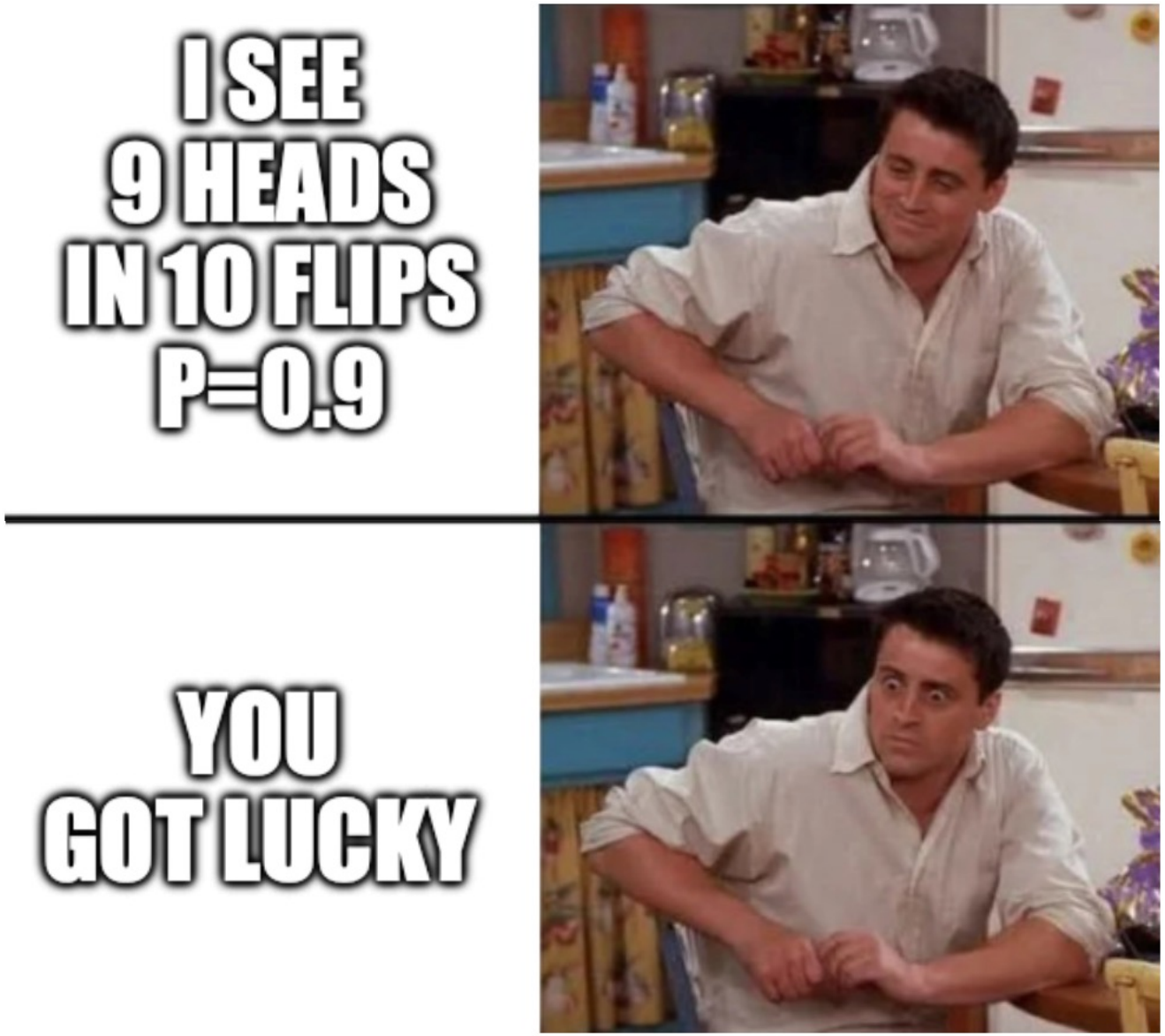
What is your belief that you flip a heads on my coin?

The parameter p to a binomial can be a random variable

9 Heads out of 10 Flips. What is your Belief in p?

$$p = \frac{9}{10}$$

Great! We're done



Uncertainty

9 Heads out of 10 Flips. What is your Belief in p?

Let X be our belief about the probability of heads:

$$P(X = x|H = 9, T = 1)$$

$$= \frac{P(H = 9, T = 1|X = x) f(X = x)}{P(H = 9, T = 1)}$$

Binomial

Uniform

9 Heads out of 10 Flips. What is your Belief in p?

Let X be our belief about the probability of heads:

$$P(X = x|H = 9, T = 1)$$

$$= \frac{P(H = 9, T = 1|X = x)f(X = x)}{P(H = 9, T = 1)}$$

Binomial

Uniform

$$= \frac{\binom{10}{9}x^9(1-x)^1}{P(H = 9, T = 1)}$$

9 Heads out of 10 Flips. What is your Belief in p?

Let X be our belief about the probability of heads:

$$P(X = x|H = 9, T = 1)$$

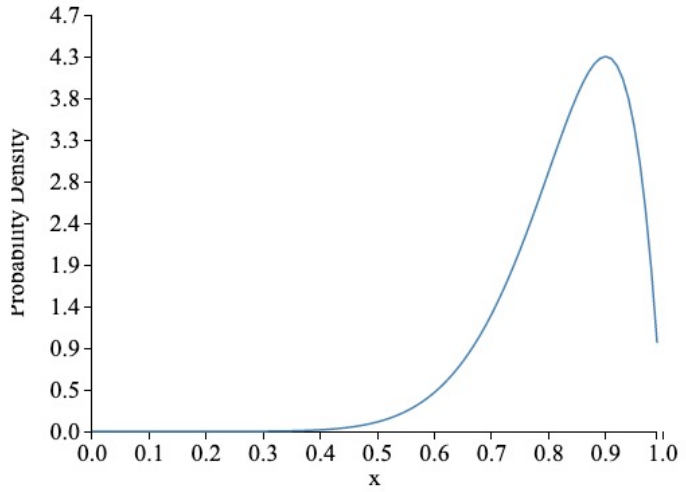
Binomial

$$= \frac{P(H = 9, T = 1|X = x) f(X = x)}{P(H = 9, T = 1)}$$

Uniform

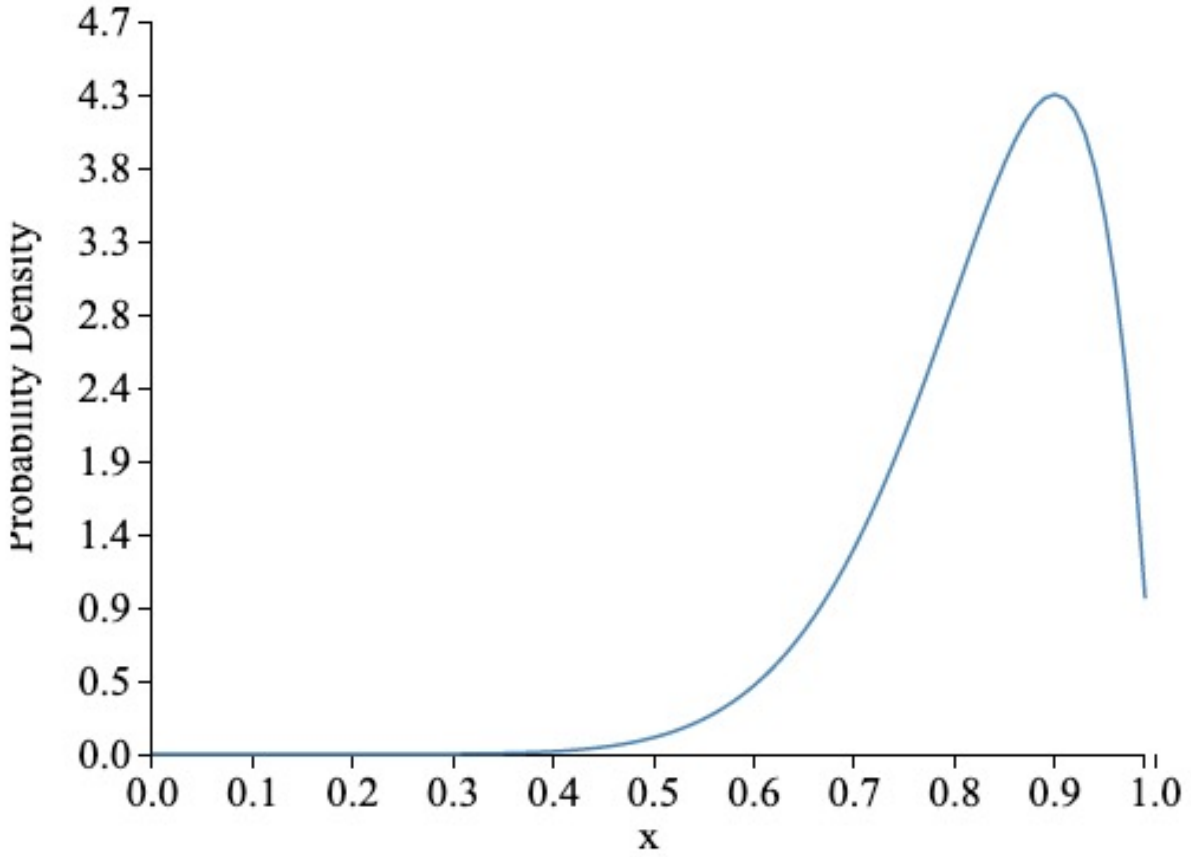
$$= \frac{\binom{10}{9} x^9 (1-x)^1}{P(H = 9, T = 1)}$$

$$= K \cdot x^9 (1-x)^1$$



9 Heads out of 10 Flips. What is your Belief in p?

$$P(X = x | H = 9, T = 1)$$



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

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)

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

1

Binomial

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

Move terms around

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

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

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:

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

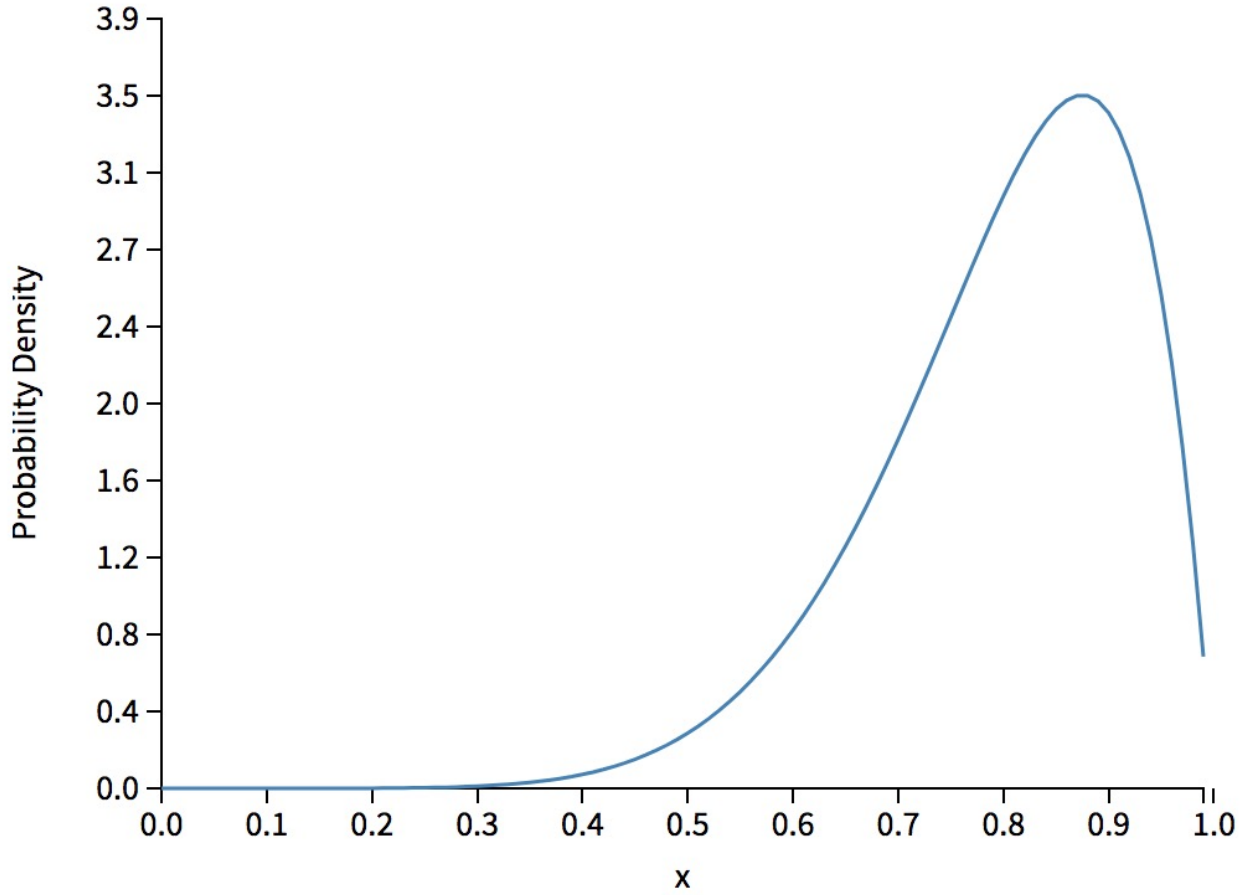
where $c = \int_0^1 x^n(1-x)^m$

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

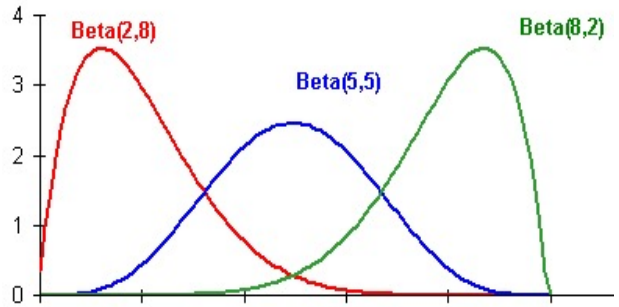
Beta Random Variable

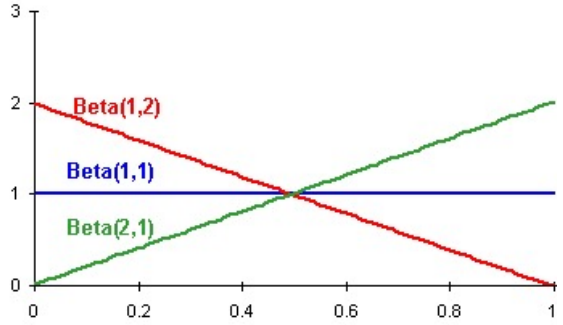
X is a **Beta Random Variable**: X ~ Beta(a, b)

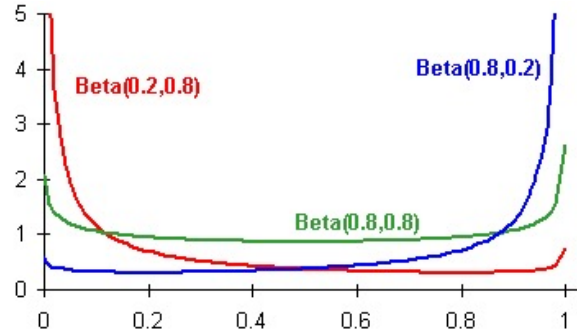
- Probability Density Function (PDF): (where a, $b > 0$)

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

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





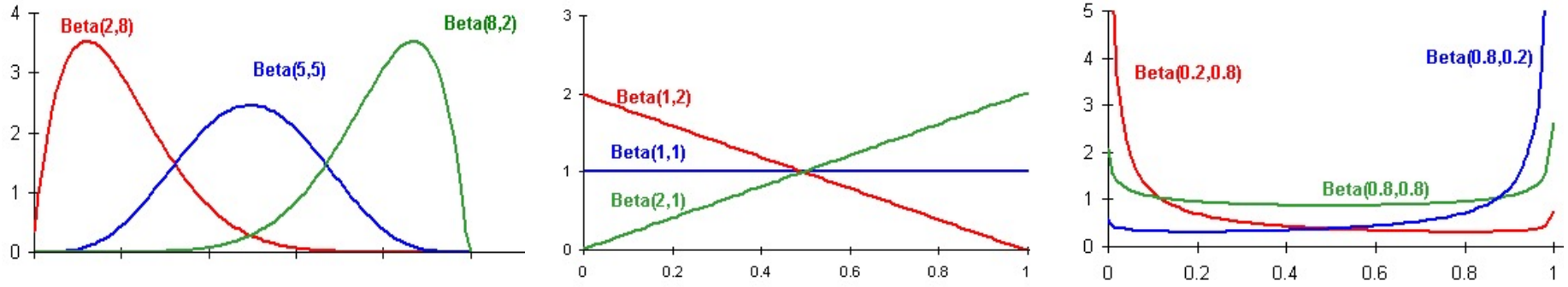


- Symmetric when $a = b$

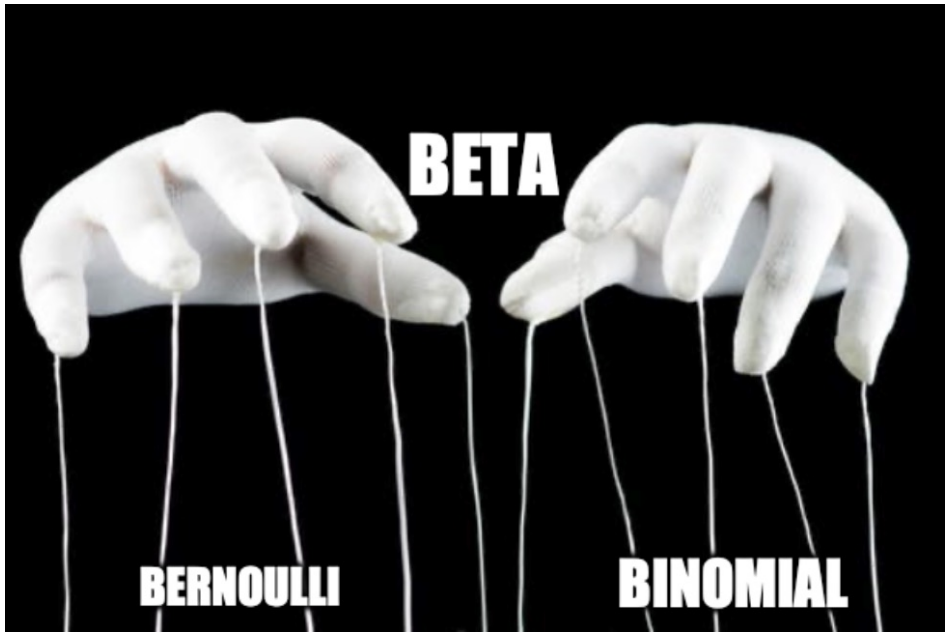
$$E[X] = \frac{a}{a+b} \qquad Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

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

Beta is the Random Variable for Probabilities



Used to represent a distributed belief of a probability



(Still an R.V tho :P)

Beta Parameters *can* come from experiments:

a = “successes” + 1

b = “failures” + 1

Eh? +1

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

- Uniform prior means both 0 and 1 is possible.
- How this is encoded is we hallucinate seeing a 0 and a 1 once each.

Concretely:

$$\text{Uni}(0, 1) = \text{Beta}(1, 1)$$

Proof:

$$X \sim \text{Uni}(0, 1)$$
$$Y \sim \text{Beta}(1, 1)$$

Set up R.Vs.

$$B(1, 1) = \int_0^1 y^{1-1}(1-y)^{1-1} dy = 1$$

Solve beta constant

$$f(X = a) = \frac{1}{1-0} = \frac{1}{B(1,1)} \cdot a^{1-1} \cdot (1-a)^{1-1} = f(Y = a)$$

PDF equality to show X=Y

Laplace Smoothing

One imagined heads

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

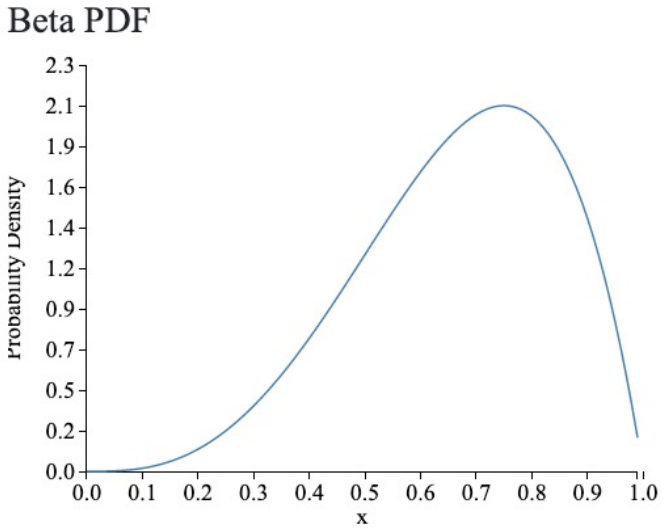
One imagined tail

Fancy name. Simple prior

Think about the difference between a **point estimate** and a **distribution**

$p = 0.75$

$p =$



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?

n heads, m tails, x is param

$$
\begin{aligned}
f(X = x|N = n) &= \frac{P(N = n|X = x)f(X = x)}{P(N = n)} \\
&= \frac{\binom{n+m}{n} x^n (1-x)^m \, f(X = x)}{P(N = n)} \\
&= \frac{\binom{n+m}{n} x^n (1-x)^m \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}}{P(N = n)} \\
&= K_1 \cdot \binom{n+m}{n} x^n (1-x)^m \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \\
&= K_3 \cdot x^n (1-x)^m x^{a-1}(1-x)^{b-1} \\
&= K_3 \cdot x^{n+a-1}(1-x)^{m+b-1} \\
X|N &\sim \text{Beta}(n+a, m+b)
\end{aligned}
$$

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 Bernoulli and Binomial

- 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.

Update to get posterior probability

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

Examples!

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

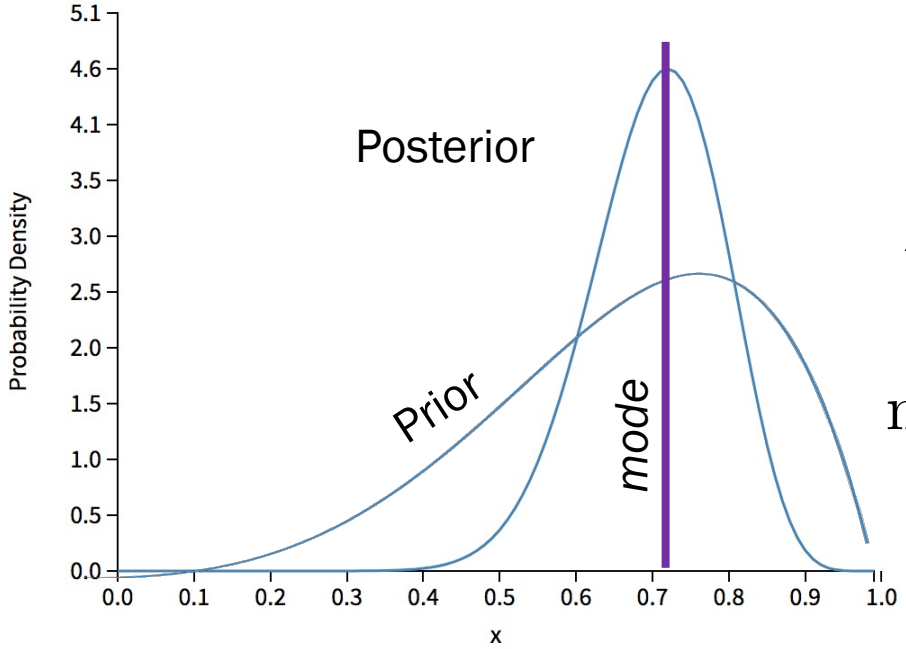
Before being tested, a medicine is believed to "work" about 80% of the time. The medicine is tried on 20 patients. It "works" for 14 and "doesn't work" for 6. What is your new belief that the drug works?

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

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

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

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



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

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

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

Which video are you more likely to like?

10,000 50

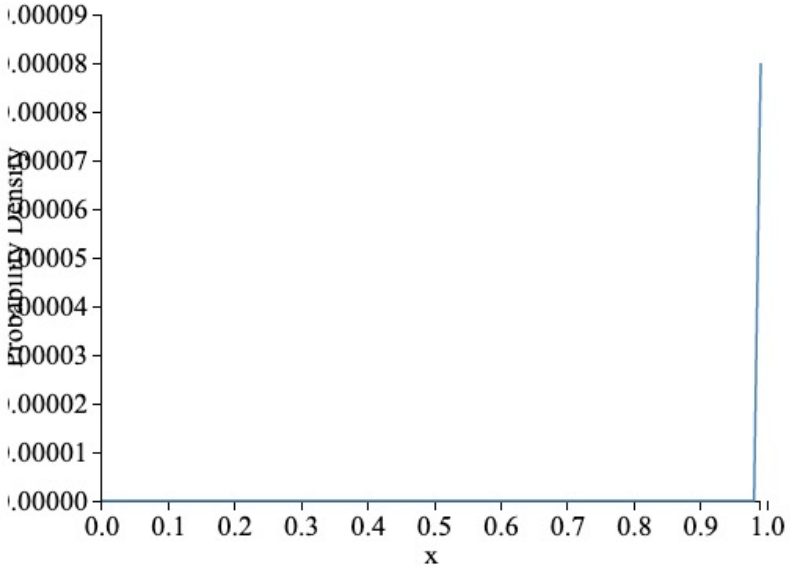
10 0

Which mochi are you more likely to like?

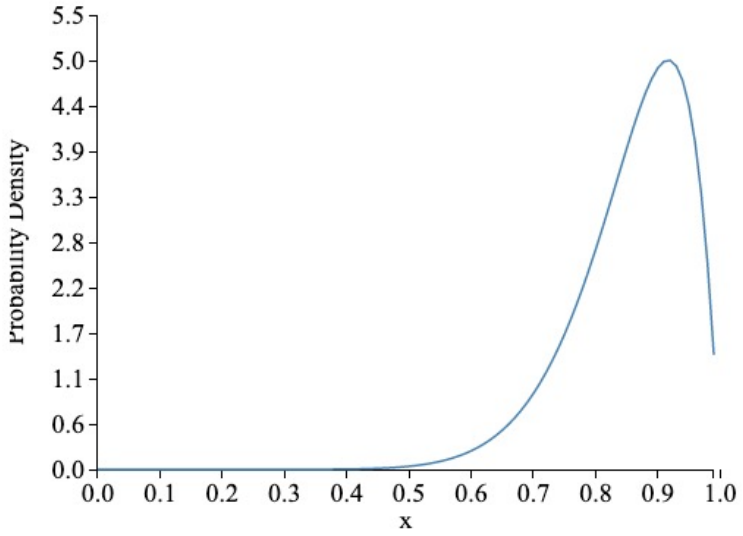
10,000 50

10 0

Beta PDF (Using Beta(2,2) prior)

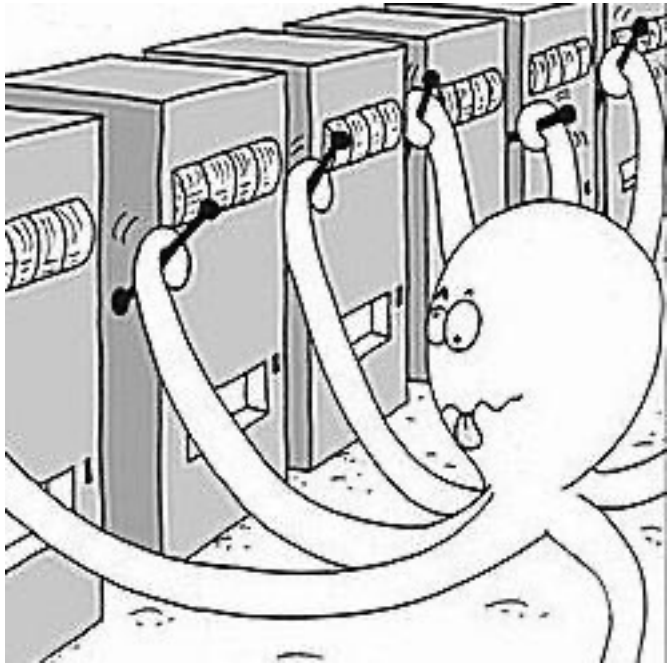


Beta PDF (Using Beta(2,2) prior)



Next level?

Multi Armed Bandit



What if you don't know a probability?

Multi Armed Bandit

Drug A

Drug B

Which one do you approve to give to a pharmacy?

Lets Play!

Drug A

X_A

Drug B

X_B

Which one do you approve to give to a pharmacy?

Optimal Decision Making

You try drug B, 10 times. It is successful 7 times.

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

Drug A

Drug B

7 successes

3 failures

What's this R.V?

$$X_B \sim Beta(7+1, 3+1)$$

Optimal Decision Making

You try drug A, 2 times. It is successful 2 times.

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

Drug A

Drug B

2 successes

0 failures

What's this R.V?

$$X_A \sim Beta(2+1, 0+1)$$

Optimal Decision Making

You try drug B, 10 times. It is successful 7 times.
You try drug A, 2 times. It is successful 2 times.

$$X_B \sim Beta(8, 4)$$

$$X_A \sim Beta(3, 1)$$

What is expectation of the outcome of drug B?

$$E[X_B] = \frac{a}{a+b} = \frac{8}{8+4} \approx 0.66$$

What is expectation of the outcome of drug A?

$$E[X_A] = \frac{3}{3+1} = 0.75$$

Optimal Decision Making

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

$$X_B \sim Beta(8, 4)$$
$$X_A \sim Beta(3, 1)$$

What is the probability that $X_B > 0.5$

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

$$P(X_B > 0.5) = 1 - P(X_B < 0.5) = 1 - F_{X_B}(0.5) = 0.886$$

What is the probability that $X_A > 0.5$

$$P(X_A > 0.5) = 1 - P(X_A < 0.5) = 1 - F_{X_A}(0.5) = 0.875$$

Wait... so, $P(X_B > 0.5)$ is better but $E[X_A]$ is better?? What?

$$P(X_A > 0.5) > P(X_B > 0.5)$$
$$E[X_A] < E[X_B]$$

Confidence

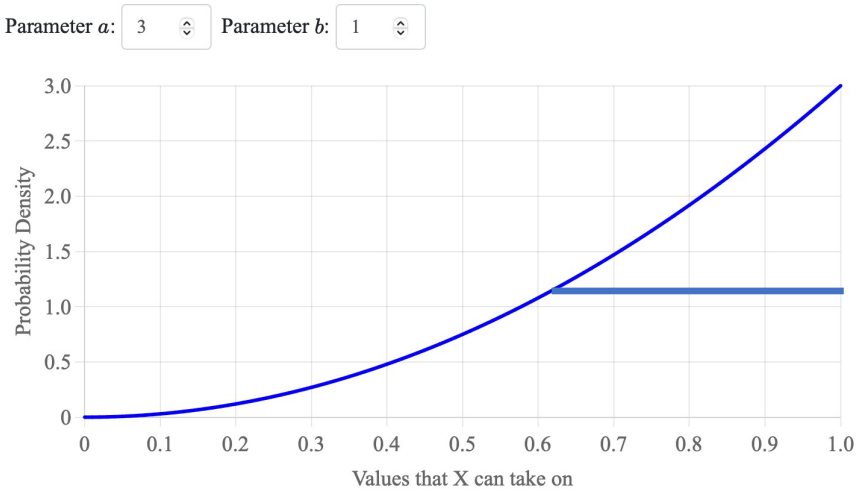
So Why?

$$P(X_A > 0.5) > P(X_B > 0.5)$$

$$E[X_A] < E[X_B]$$

Drug A

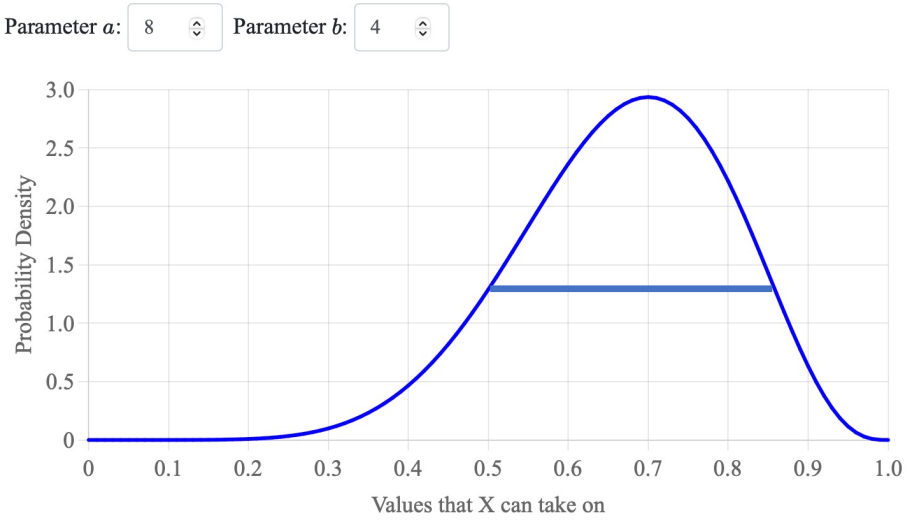
$X_A \sim Beta(3,1)$



We are more confident about drug B

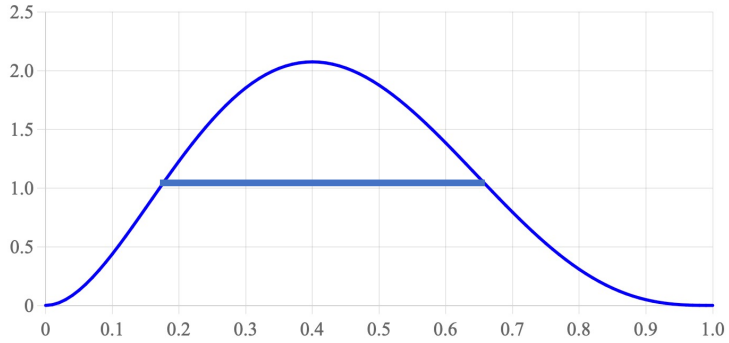
Drug B

$X_B \sim Beta(8,4)$

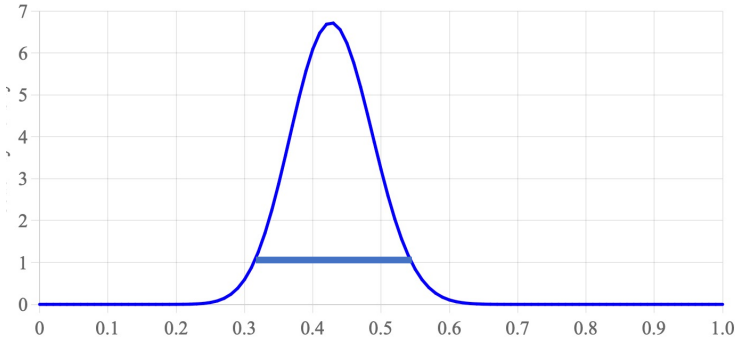


Confidence

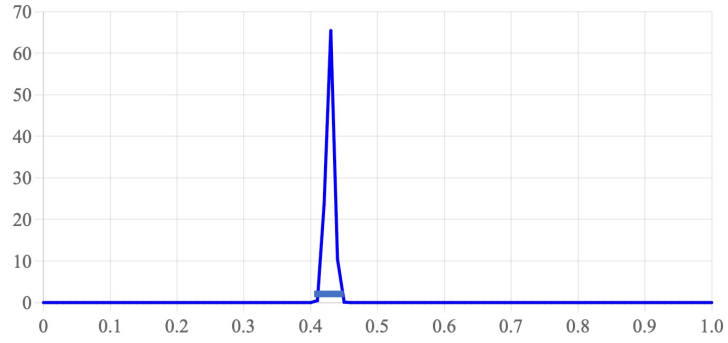
$X_B \sim Beta(3,4)$



$X_B \sim Beta(30,40)$



$X_B \sim Beta(3000,4000)$



As we see more and more evidence, we start becoming more and more confident!!

So, which drug??

$$P(X_A > 0.5) > P(X_B > 0.5)$$
$$E[X_A] < E[X_B]$$

Drug A

$X_A \sim Beta(3,1)$

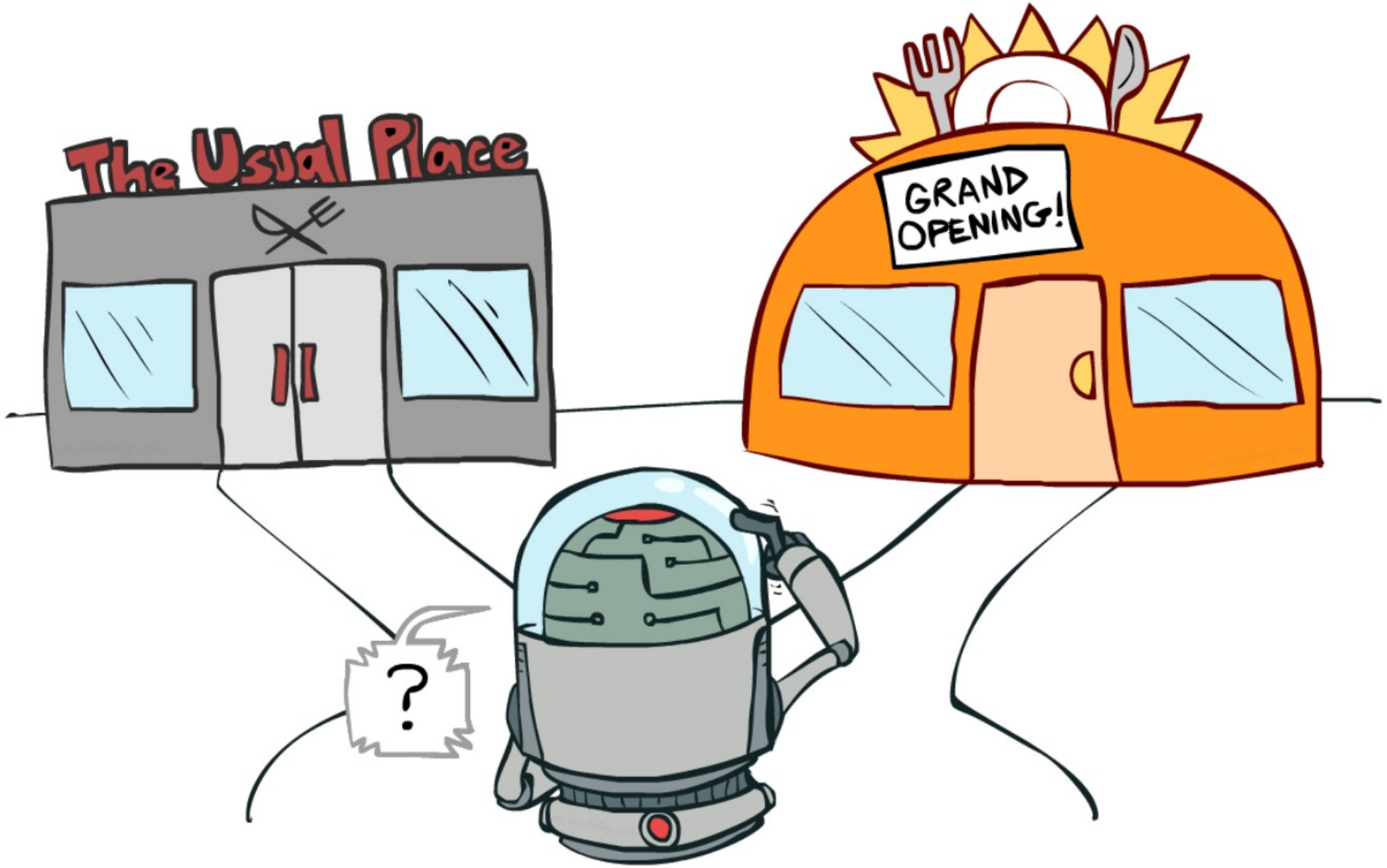
Drug B

$X_B \sim Beta(8,4)$

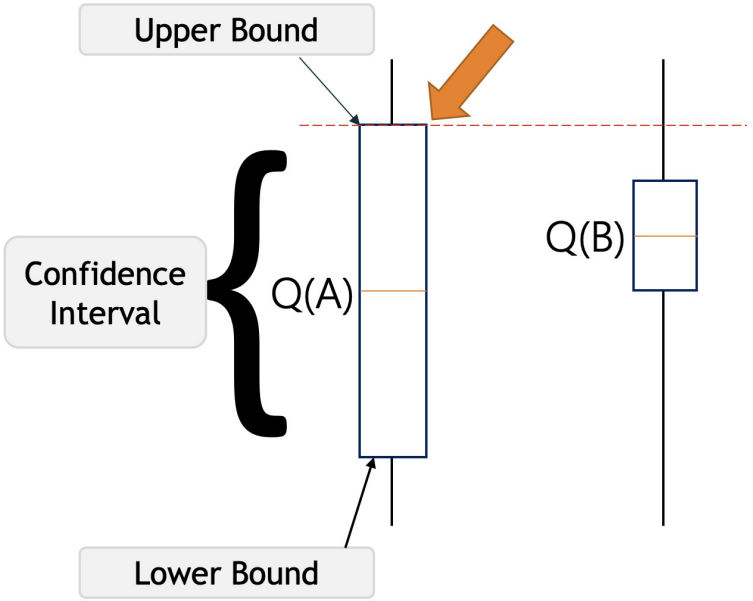
Drug C

Test Drugs A and B some more!!

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



Extension: Upper Confidence Bound



Beta Random Variable gives information on the confidence of the random variable.

$$\text{Treatment} = \arg\max_{d \in \{a,b\}} \left(p_d + c \sqrt{\frac{a+b}{\ln(d)}} \right)$$

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

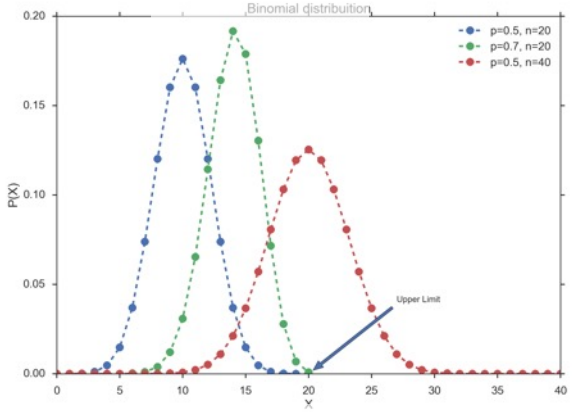
Your new belief about the probability is:

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

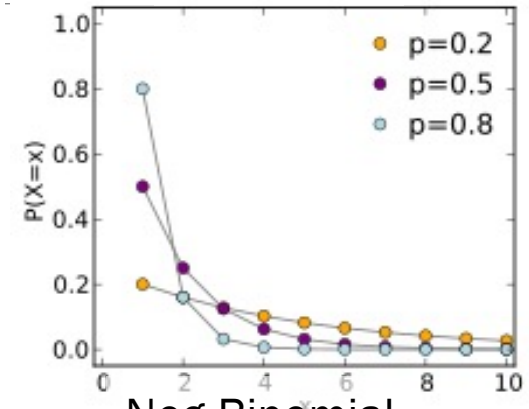
$$\text{where} \quad c = \int_0^1 x^{a-1}(1-x)^{b-1}$$

Distributions

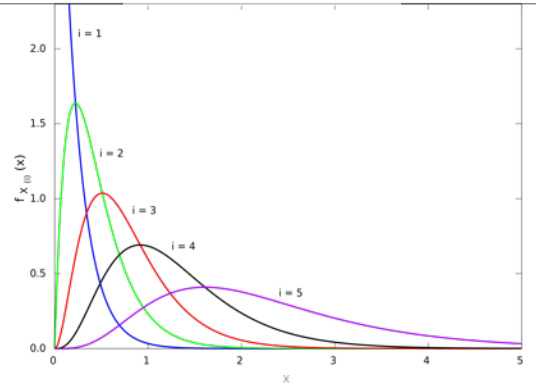
Binomial



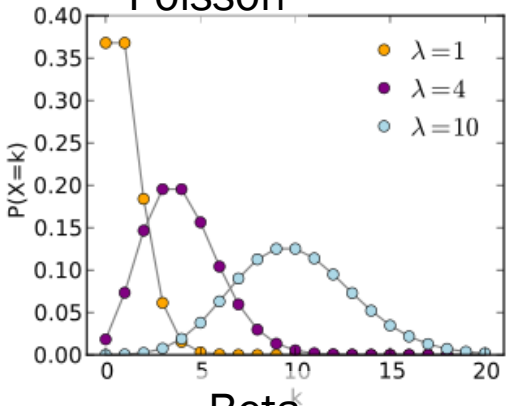
Geometric



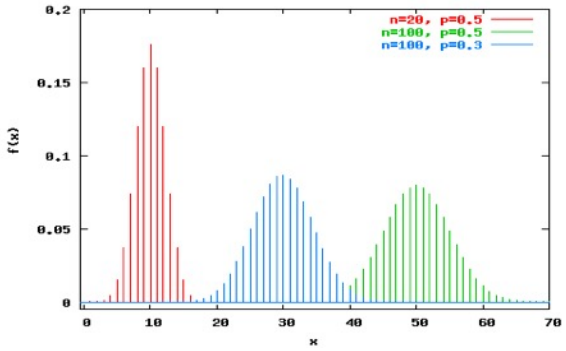
Exponential



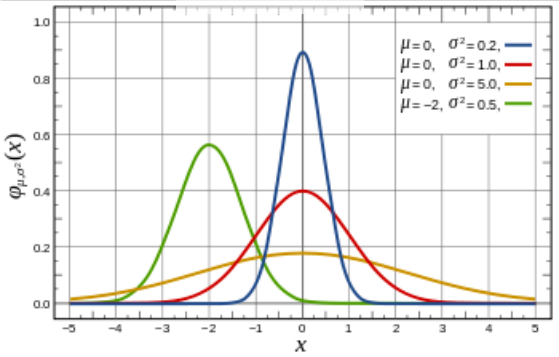
Poisson



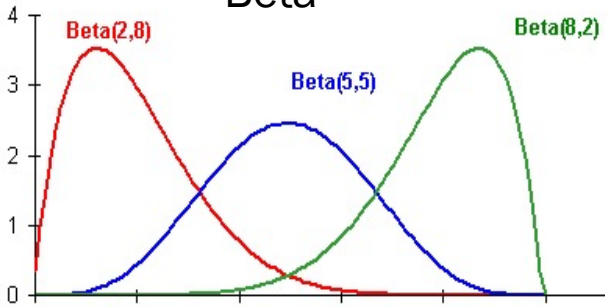
Neg Binomial



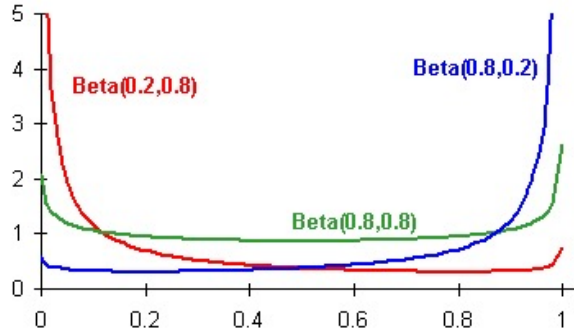
Normal



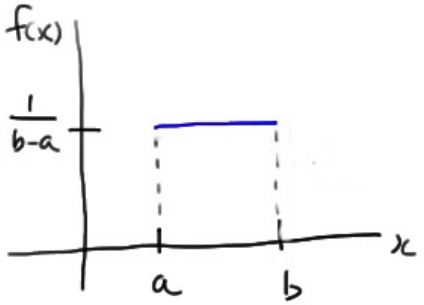
Beta



Beta

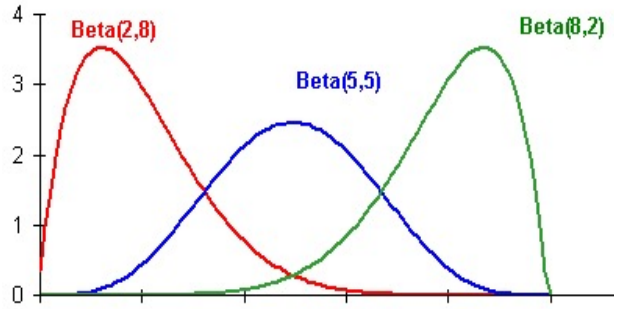


Uniform



Weights

Beta



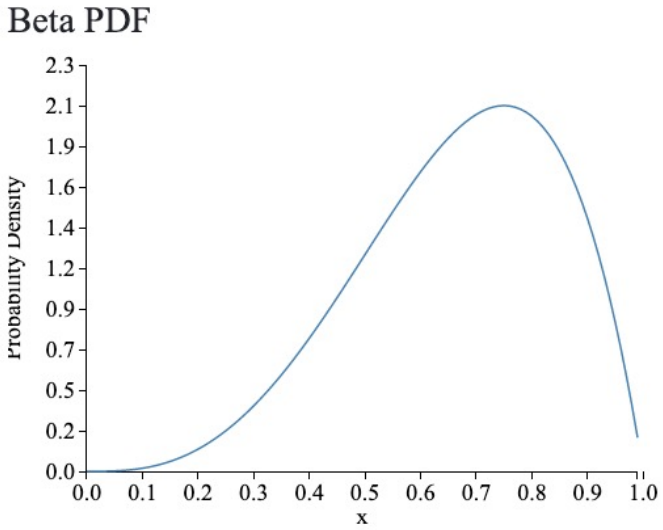
$\text{Binomial}(n, p)$ $\text{Bernoulli}(p)$

p sampled from a Beta represents a tuneable parameter!

Think about the difference between a **point estimate** and a **distribution**

$p = 0.75$

$p =$



Problem with a point estimate:

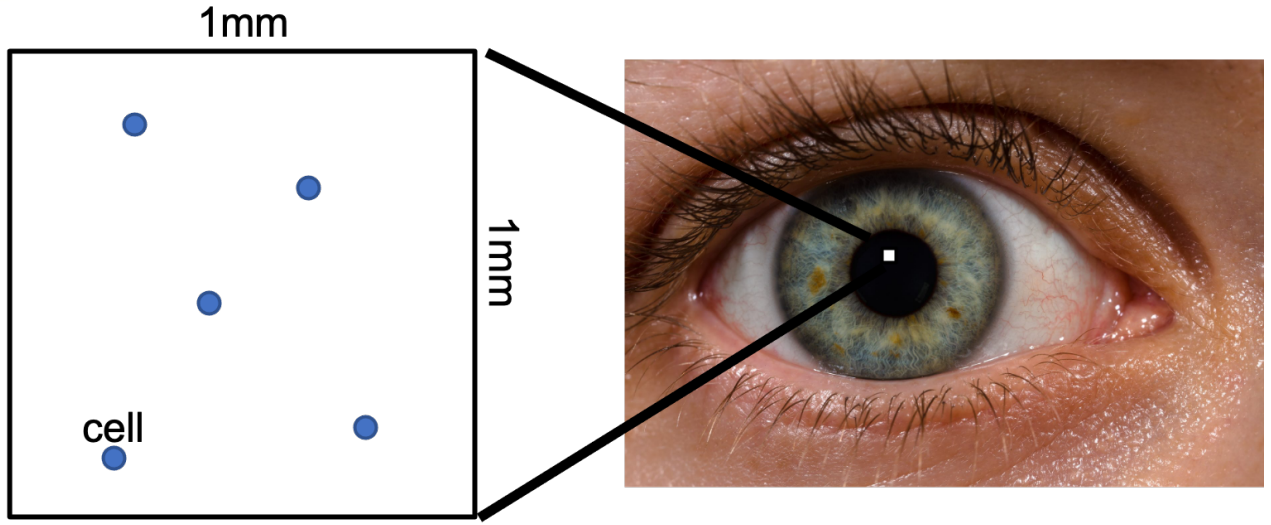
Person A: My leg itches when it rains and its kind of itchy…. Uh, p = .80

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

Give me the uncertainty!!!

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

Eg:



$$P(\Lambda = \lambda | N = 5)$$