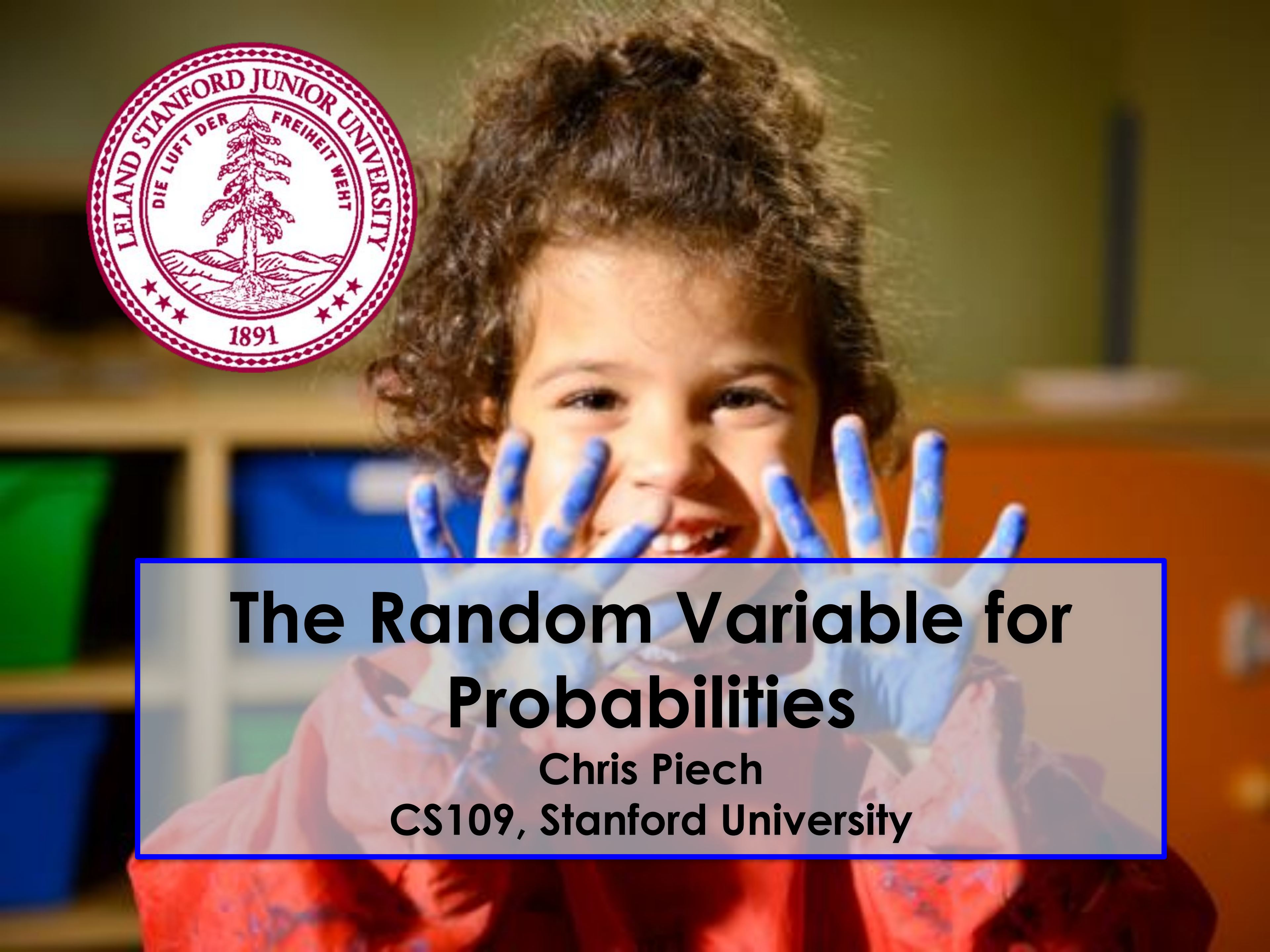
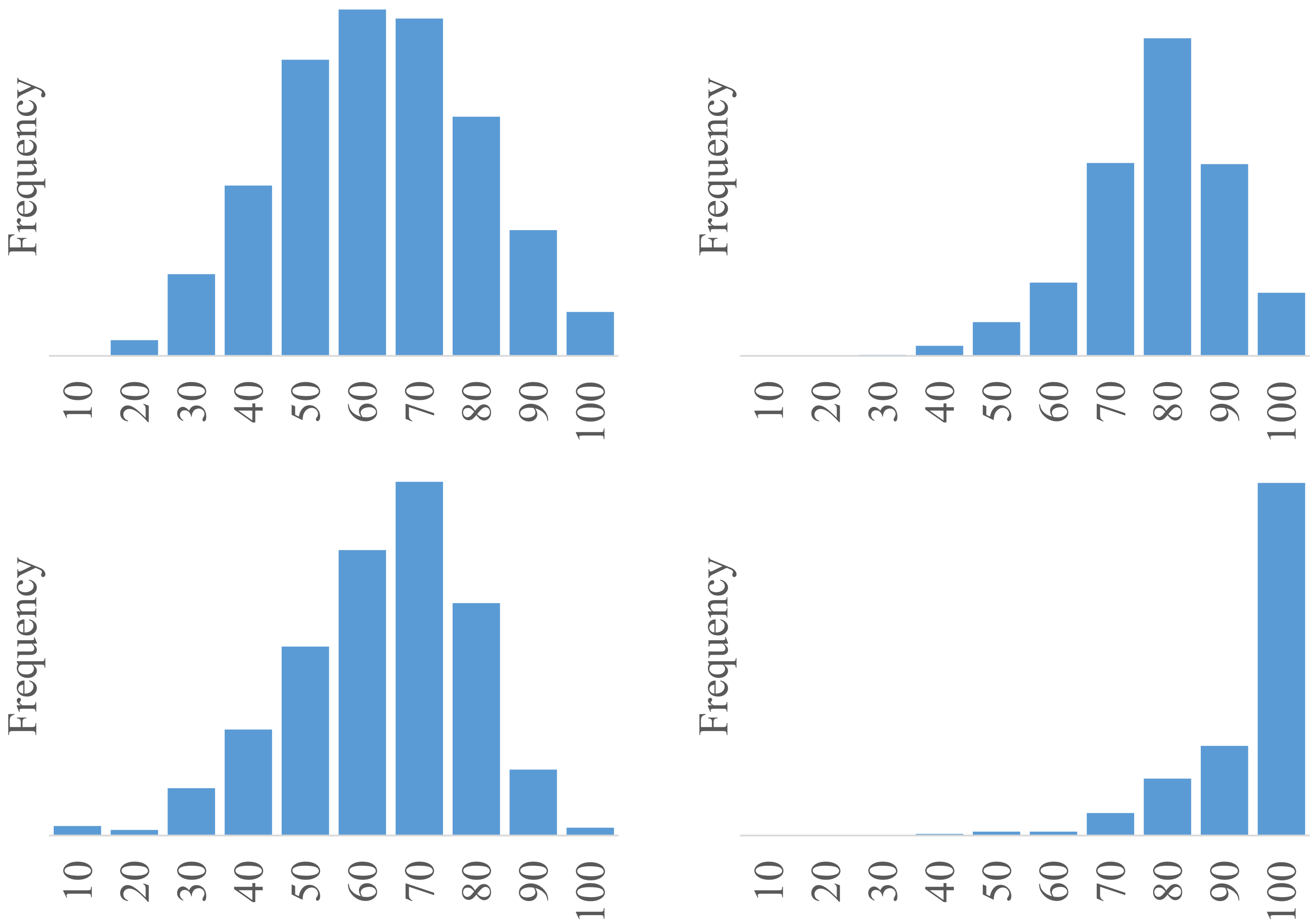


The Random Variable for Probabilities

Chris Piech
CS109, Stanford University

Assignment Grades



We have 2055 assignment distributions from grade scope

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

Review

Conditioning with a continuous random variable feels 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 Keystroke

- 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

Lets play a game

Roll a dice twice. If either time you roll a 6, I win.
Otherwise you win.

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

Flip a Coin With Unknown Probability

Demo

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

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

X is a single value

Bayesian

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

X is a random variable

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: $(N \mid X) \sim \text{Bin}(n + m, x)$

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

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)

$$f_{X|N}(x|n) = \frac{\boxed{P(N = n|X = x)}\,\boxed{f_X(x)}}{P(N = n)} \quad 1$$

Binomial

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

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

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

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:

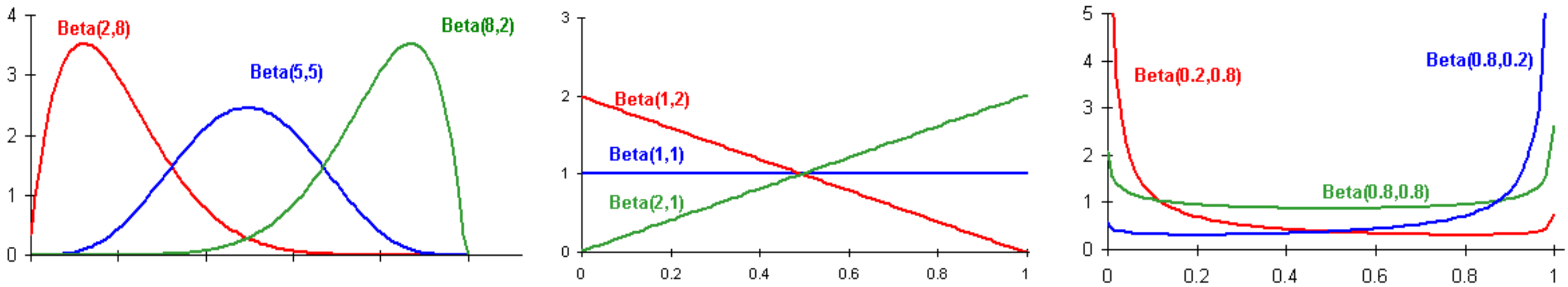
$$f_X(x) = \frac{1}{c} \cdot x^{a-1}(1-x)^{b-1}$$

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

Beta Random Variable

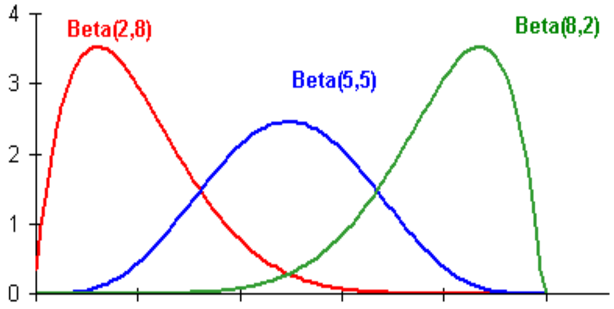
- X is a **Beta Random Variable**: X ~ Beta(*a*, *b*)
  - Probability Density Function (PDF): (where *a*, *b* > 0)

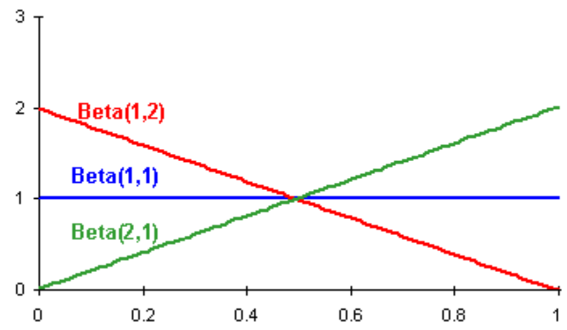
$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{where} \quad B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$

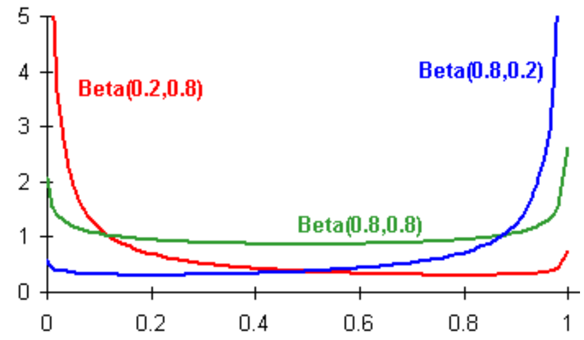


  - Symmetric when *a* = *b*
  - $E[X] = \frac{a}{a+b}$ $\quad Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$

Meta Beta







Used to represent a
distributed belief of a probability

Beta is a distribution for probabilities

Beta Parameters:

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: $(N \mid X) \sim \mathrm{Bin}(n+m, x)$

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

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

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

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

Understanding Beta

- $X \mid (N = n, M = m) \sim$ Beta$(a = n + 1, b = m + 1)$
  - Prior $X \sim$ Uni$(0, 1)$
  - 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 $X \sim$ Beta$(a = 1, b = 1)$

N successes

M failures

If the Prior was a 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 a Beta...

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

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

Understanding Beta

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

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

- Beta is a **<u>conjugate</u>** 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

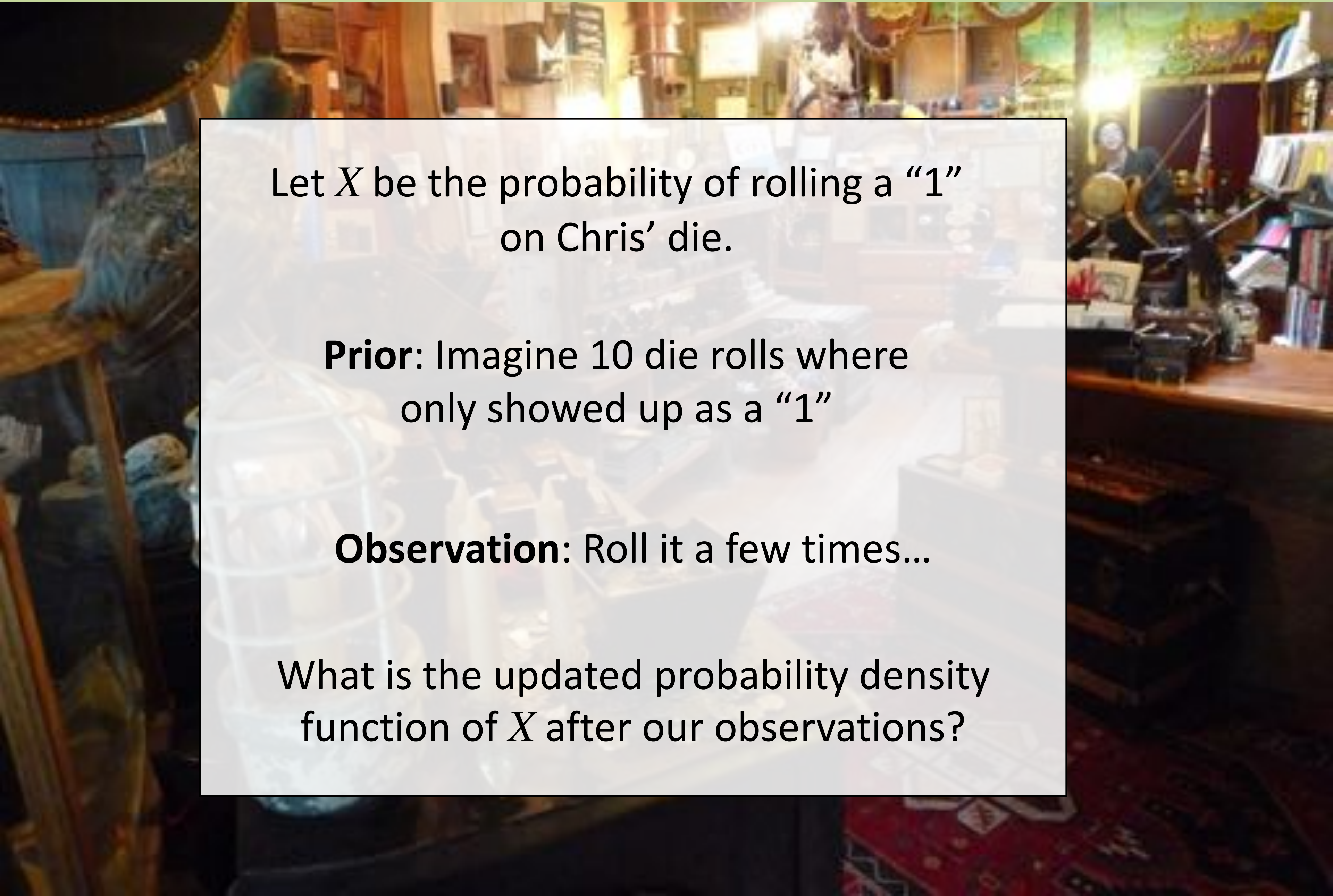
Further Understanding Beta

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



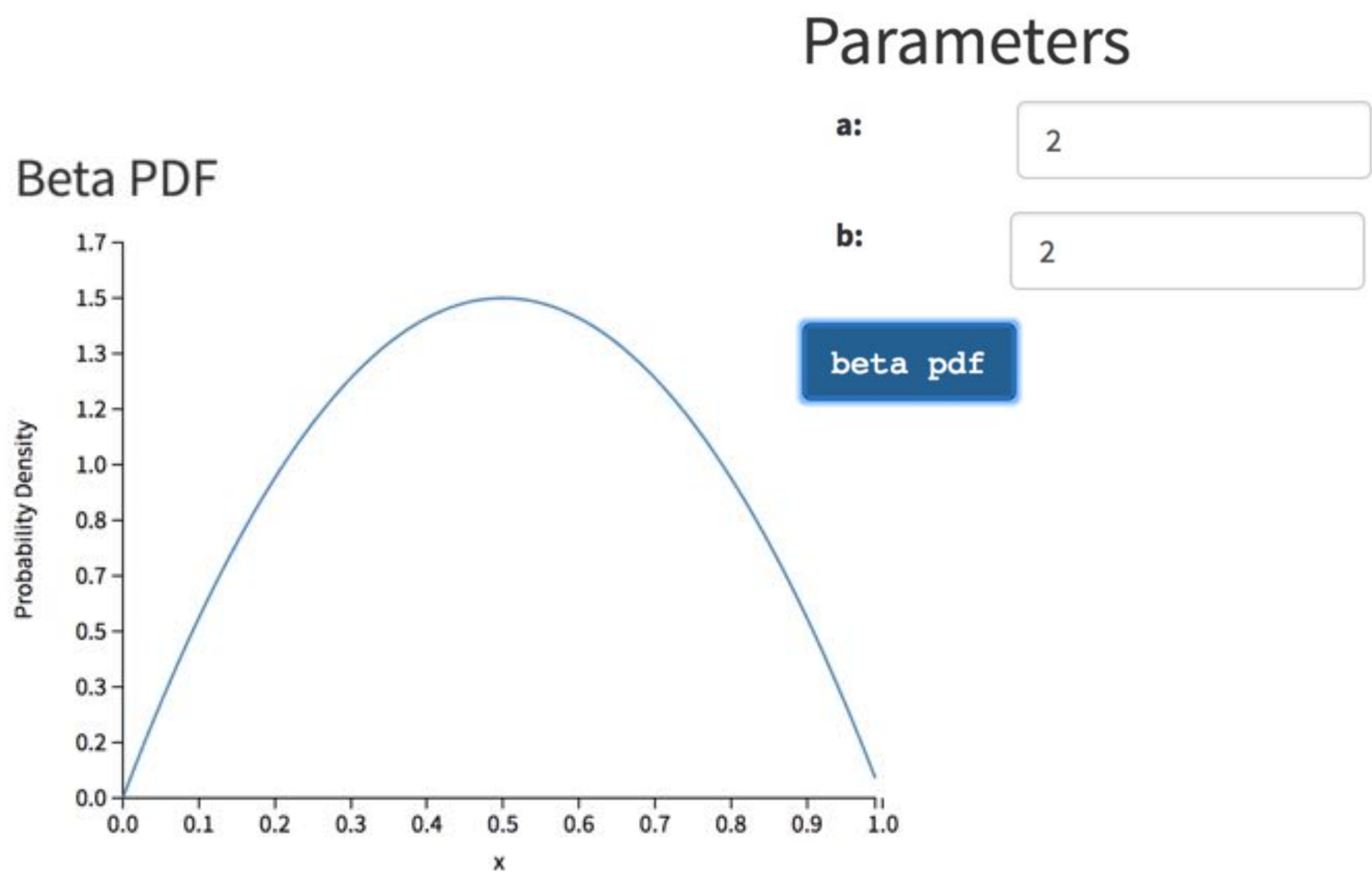
Let X be the probability of rolling a "1" on Chris' die.

Prior: Imagine 10 die rolls where only showed up as a "1"

Observation: Roll it a few times…

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

Check out Demo!



Damn

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:

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

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

Beta Example

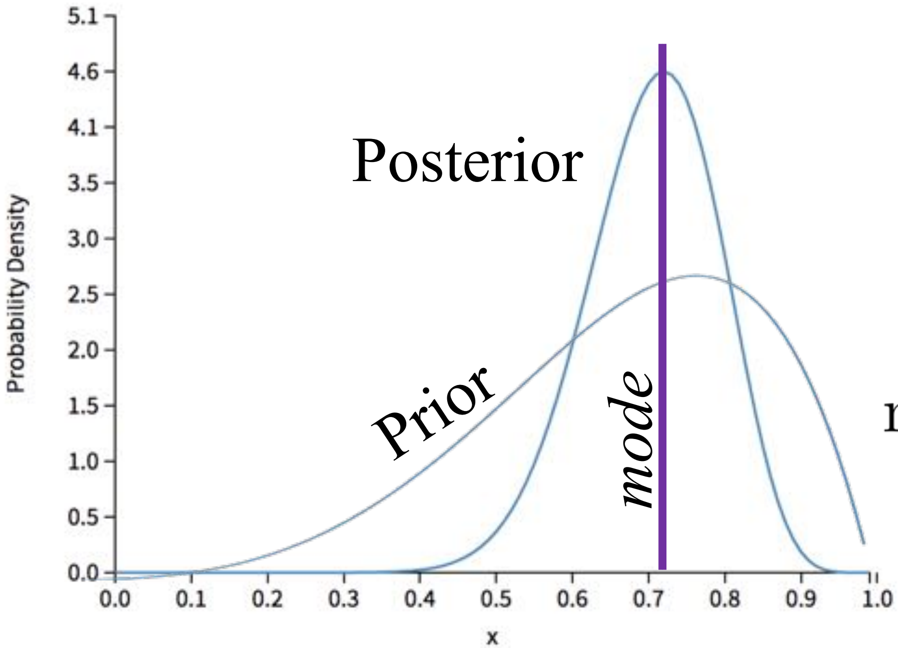
Before being tested, a medicine is believed to "work" about 80% of the time. The medicine is tried on 20 patients. It "works" for 14 and "doesn't work" for 6. What is your new belief that the drug works?

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

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

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

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



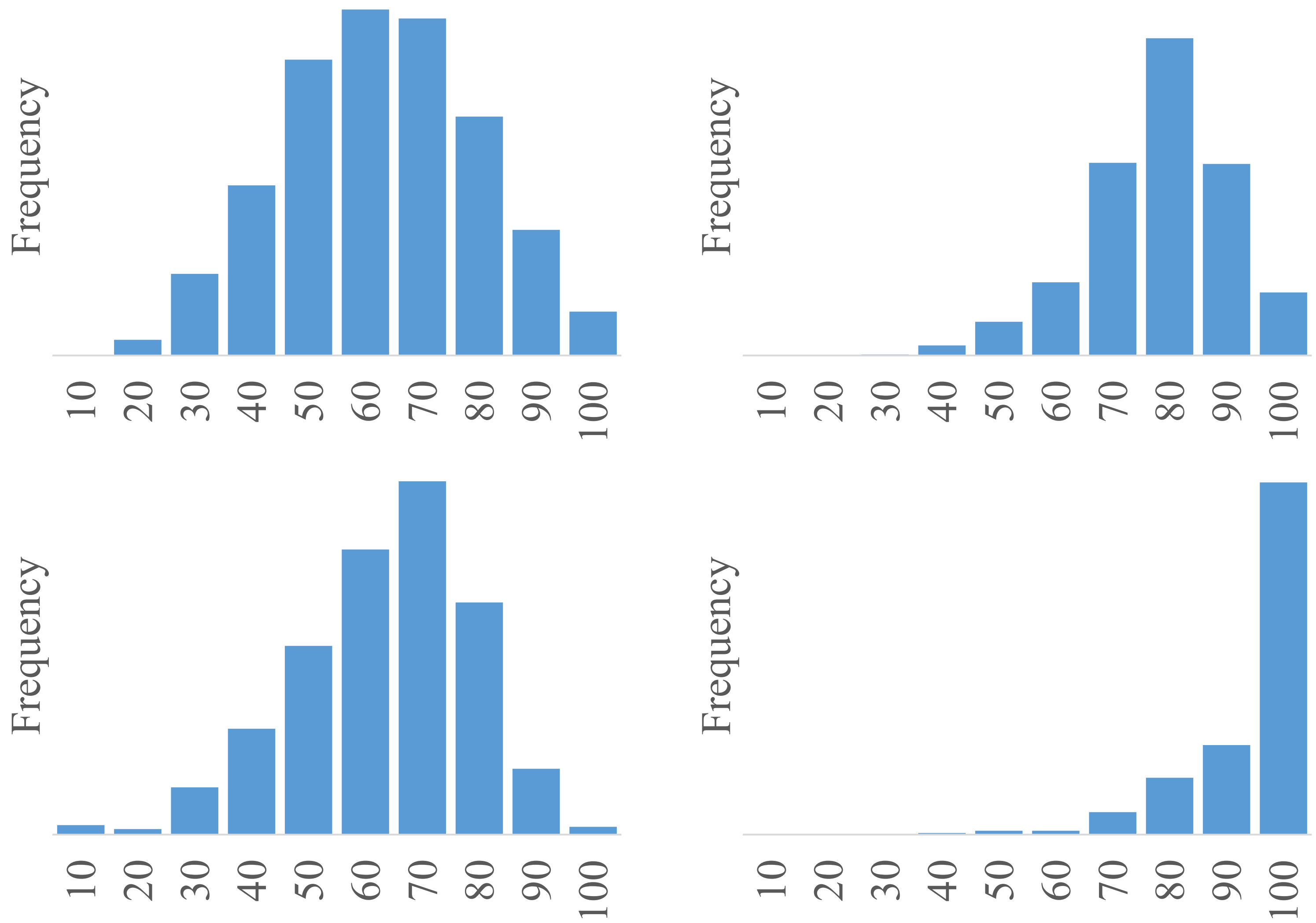
$$E[X] = \frac{a}{a+b} = \frac{19}{19+8} \approx 0.70$$

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

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

Next level?

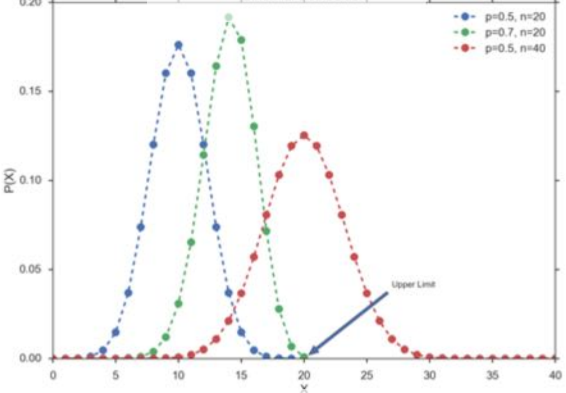
Assignment Grades



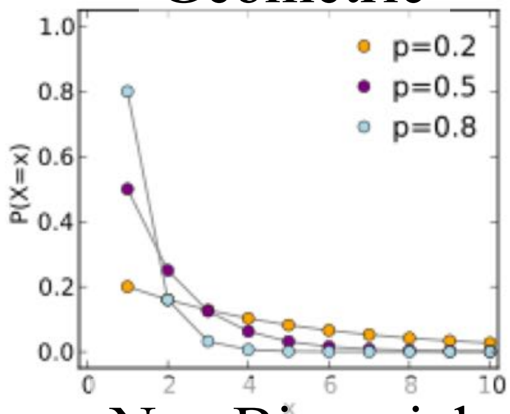
We have 2055 assignment distributions from gradescope

Distributions

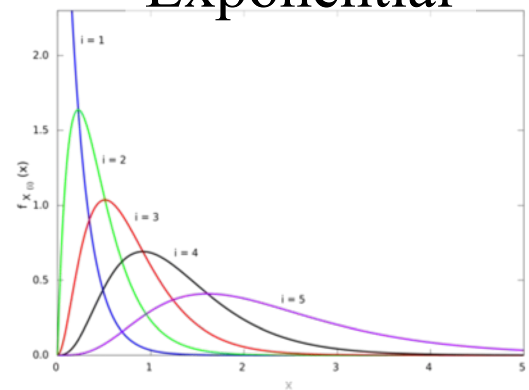
Binomial



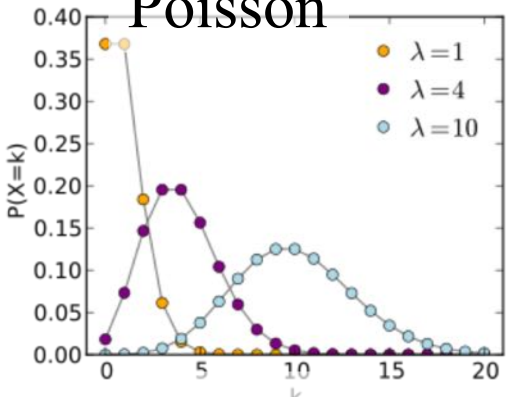
Geometric



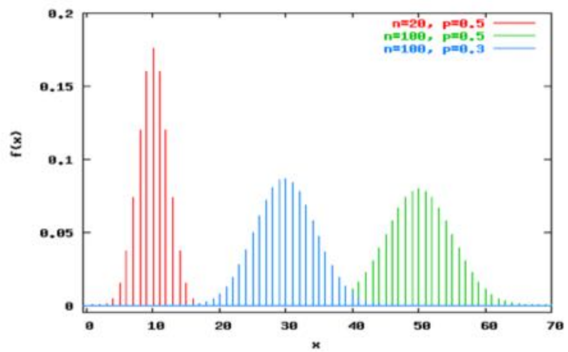
Exponential



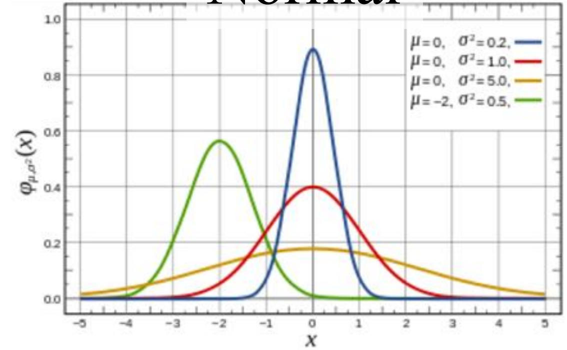
Poisson



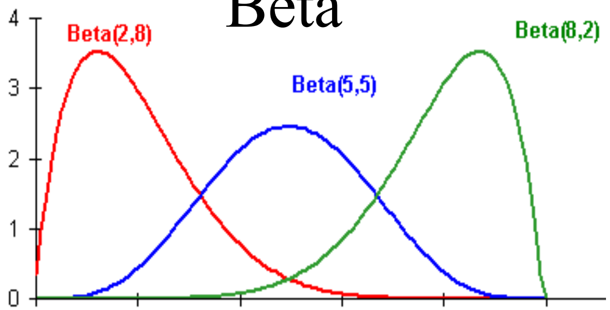
Neg Binomial



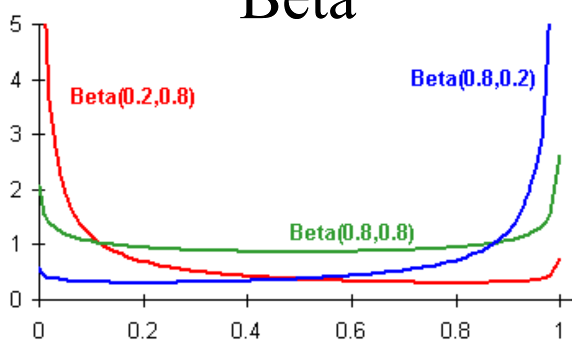
Normal



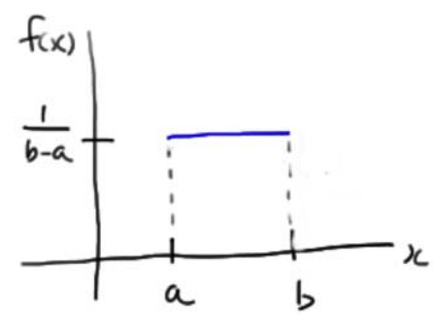
Beta



Beta



Uniform



Grades must be bounded

Normal: No

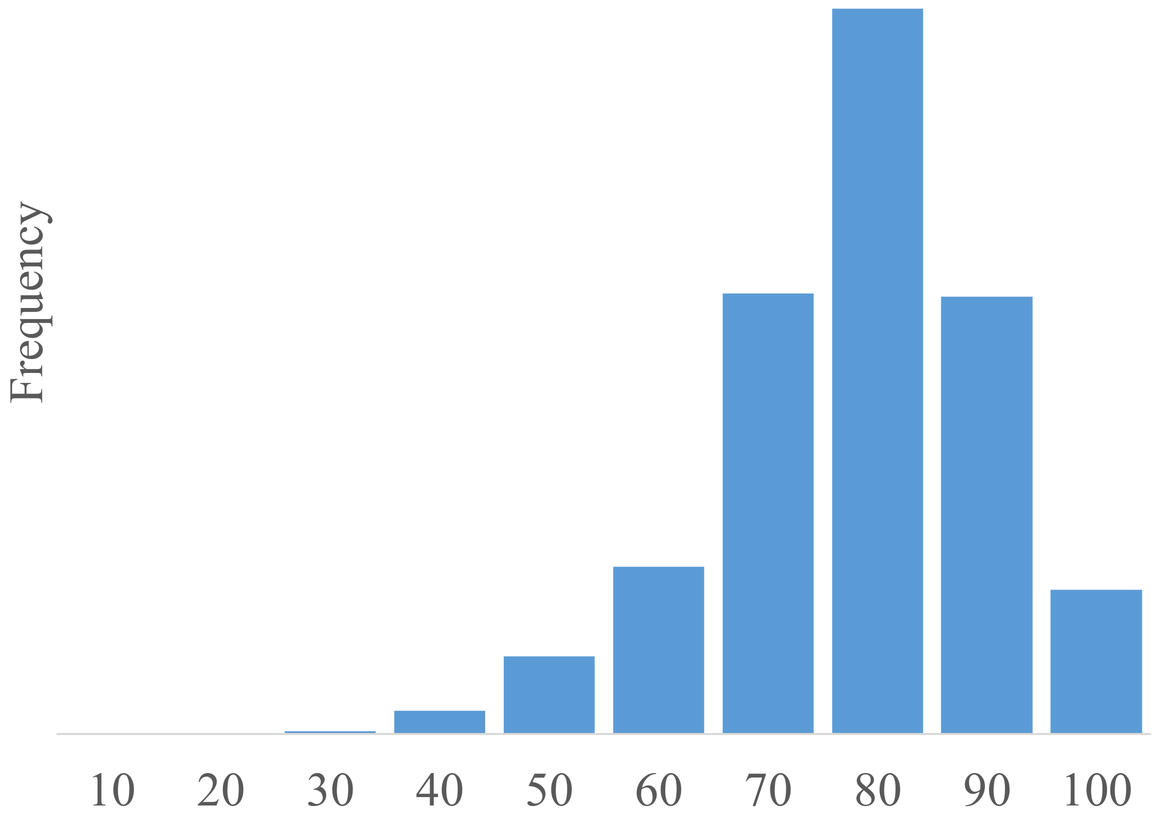
Poisson: No

Exponential: No

Beta: Looks Good!

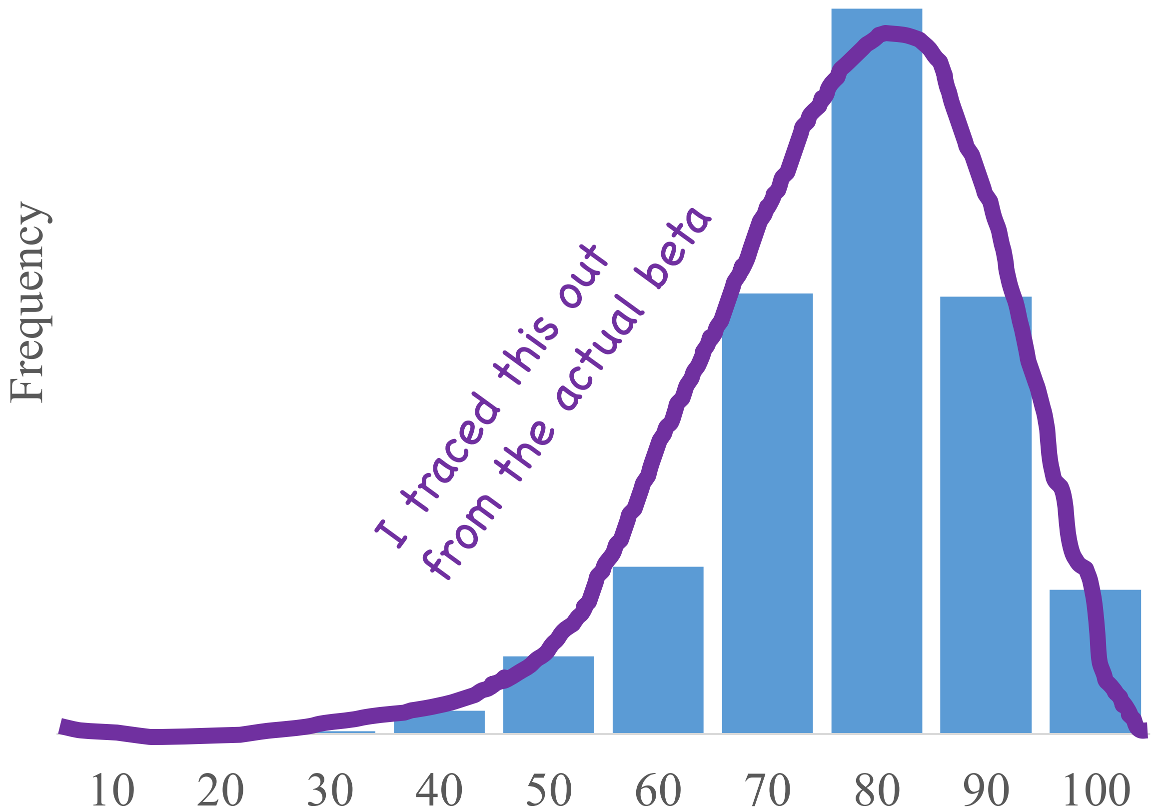
Assignment Grades Demo

`Assignment id = '1613'`



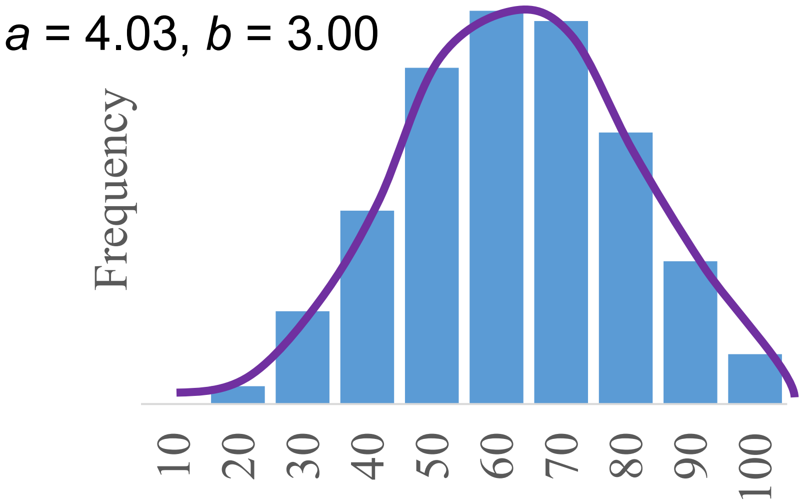
Assignment Grades Demo

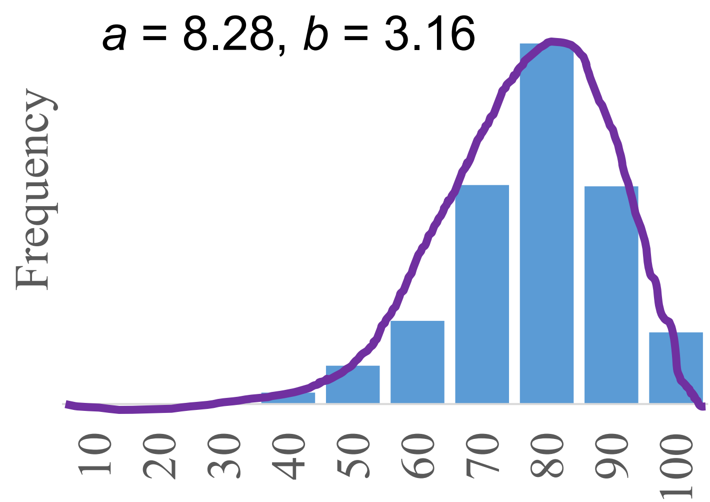
`Assignment id = '1613'`

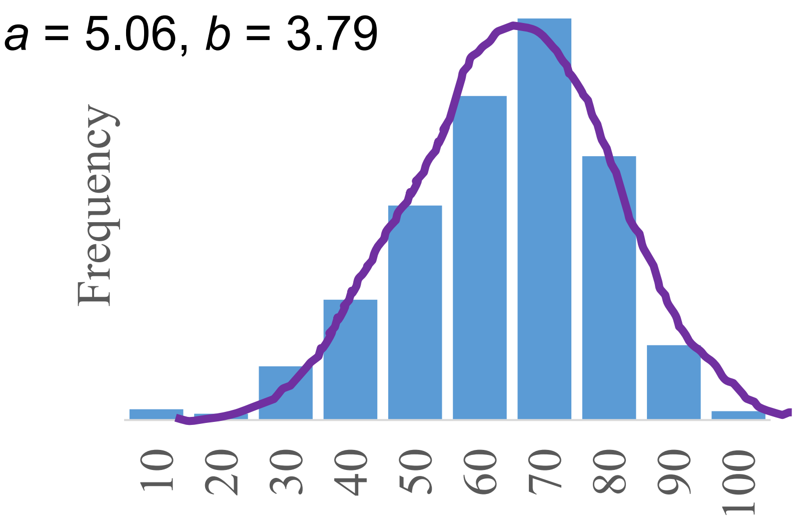


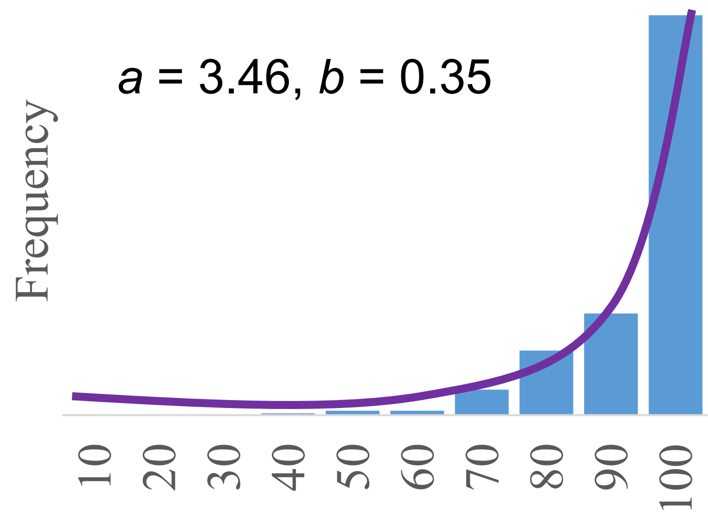
$$X \sim Beta(a = 8.28, b = 3.16)$$

Assignment Grades



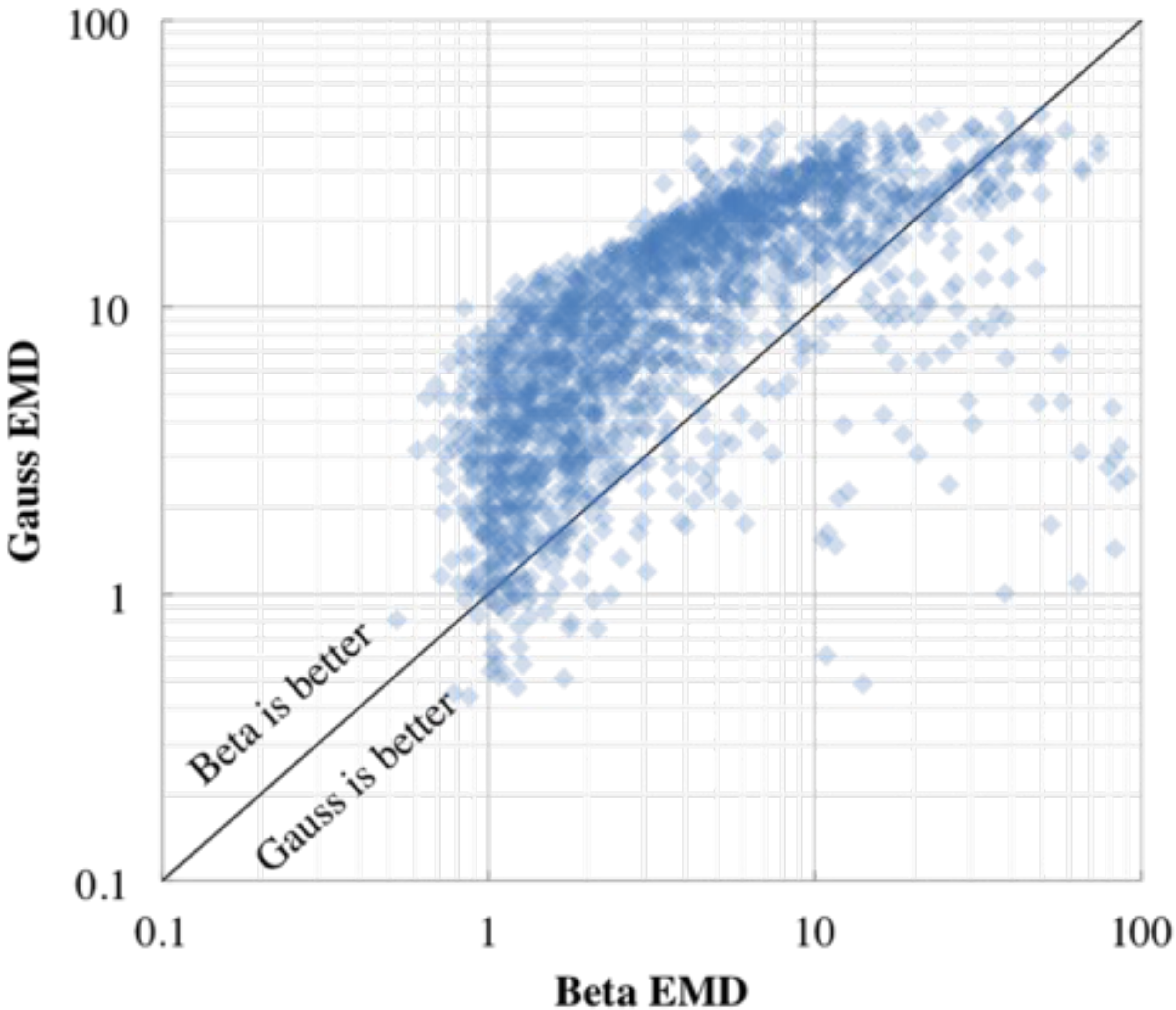






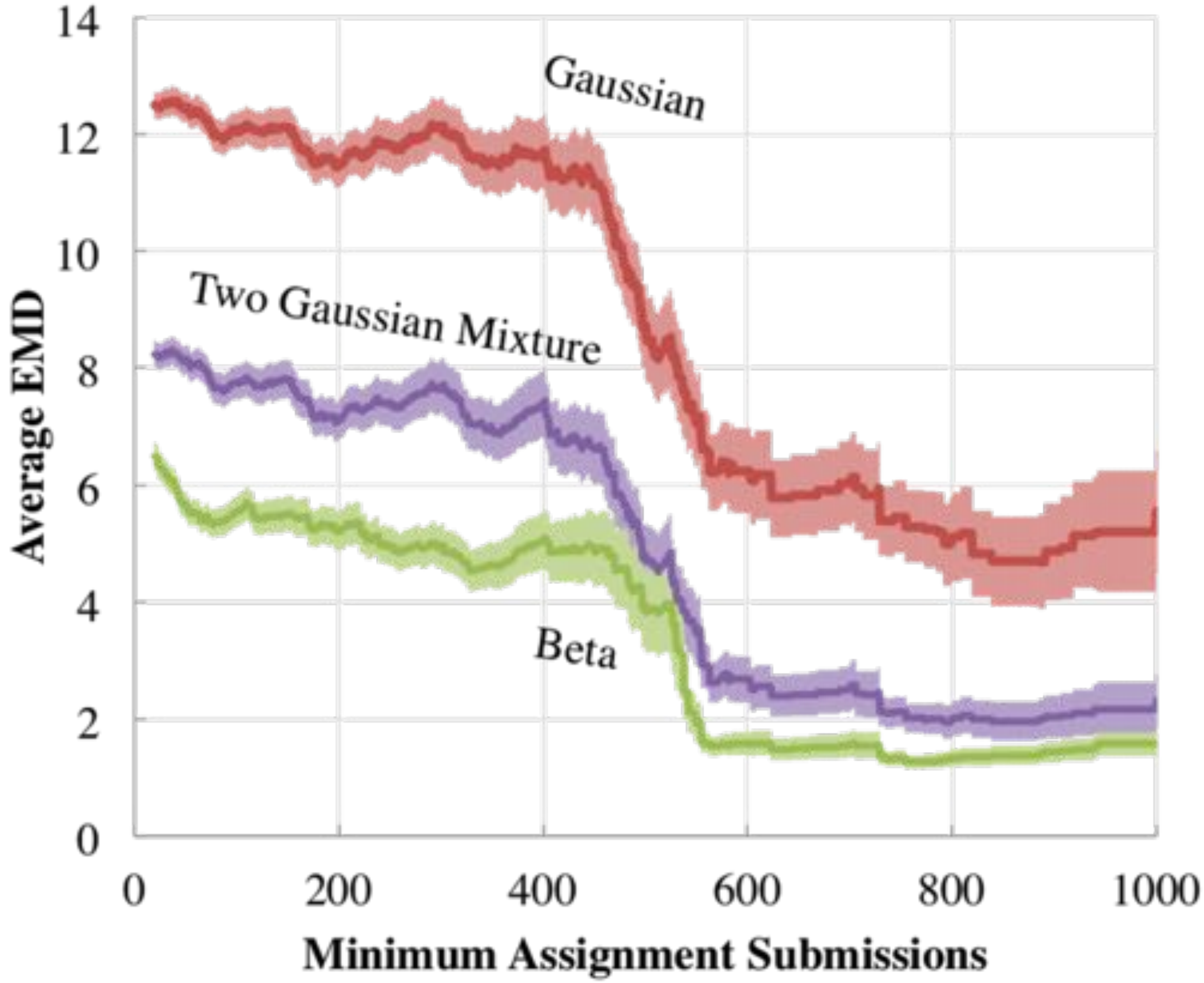
We have 2055 assignment distributions from grade scope

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

Binomial Interpretation

> Each student has **the same** probability of getting each point. Generate grades by flipping a coin 100 times for each student. The resulting distribution is binomial.

- Binomial

Normal Interpretation

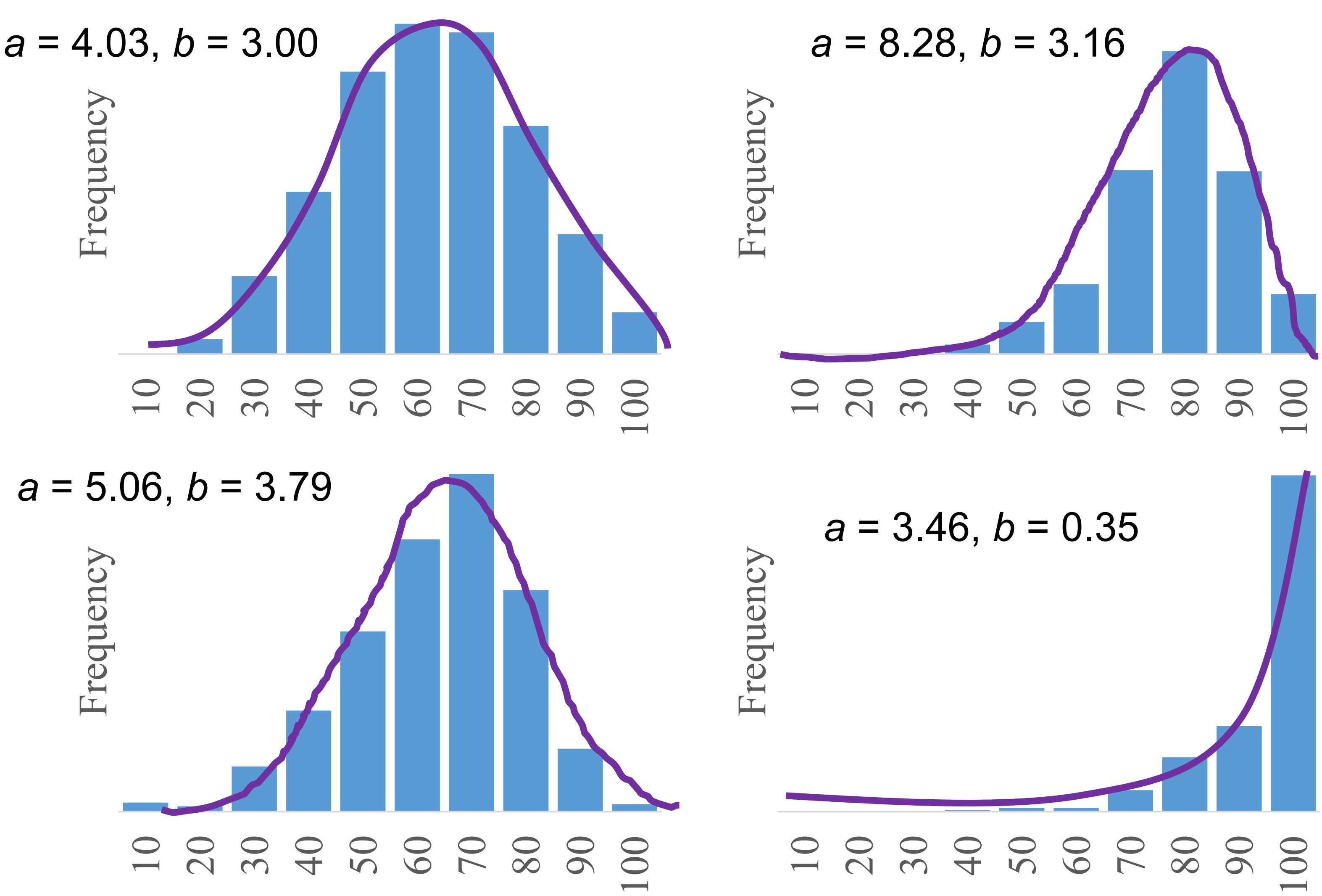
> What the Binomial said, but approximated.
>
> \- Normal

Beta Interpretation

> Each student's ability is represented as a probability – perhaps their probability of getting a generic point. Each student has their **own** probability, however, the distribution of probabilities in a class is a Beta distribution.
>
> \- Beta

* This is an opinion. It is open for debate

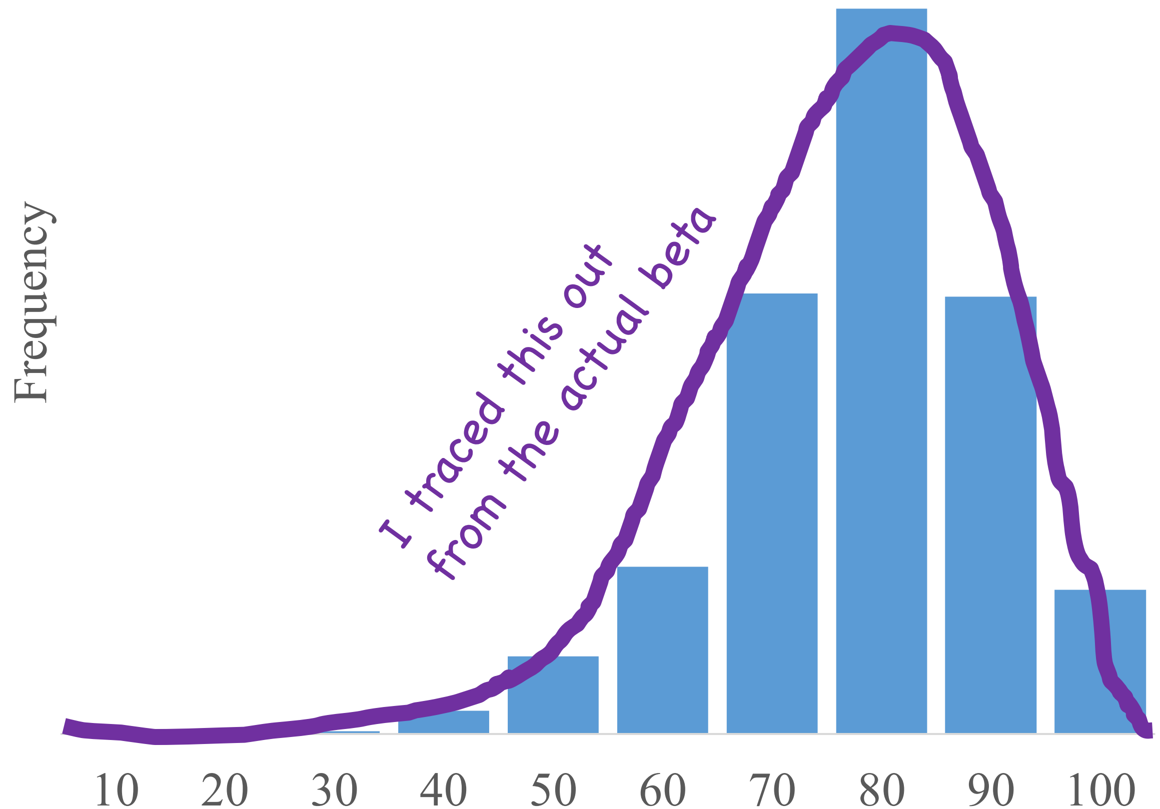
Assignment Grades



These are the distribution of student *point probabilitities*

Assignment Grades Demo

What is the semantics of E[X]?



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

Assignment Grades

What is the probability that a student is bellow the mean?

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

$$E[X] = \frac{a}{a+b} = \frac{8.28}{8.28 + 3.16} \approx 0.7238$$

$$P(X < 0.7238) = F_X(0.7238)$$

Wait what? Chris are you holding out on me?

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

$$P(X < E[X]) = 0.46$$

Implications

- Will be combined with Item Response Theory which models how assignment difficulty and student ability combine to give *point probabilities*.
- Machine learning on education data will be more accurate.
- Analysis of “mixture” distributions can be better.
- Better understand how variance impacts weighting.

Beta:
The probability density
for probabilities

Beta is a distribution for probabilities

Beta Distribution

If you start with a $X \sim \mathrm{Uni}(0, 1)$ 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 $$c = \int_0^1 x^{a-1}(1-x)^{b-1}$$

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

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