

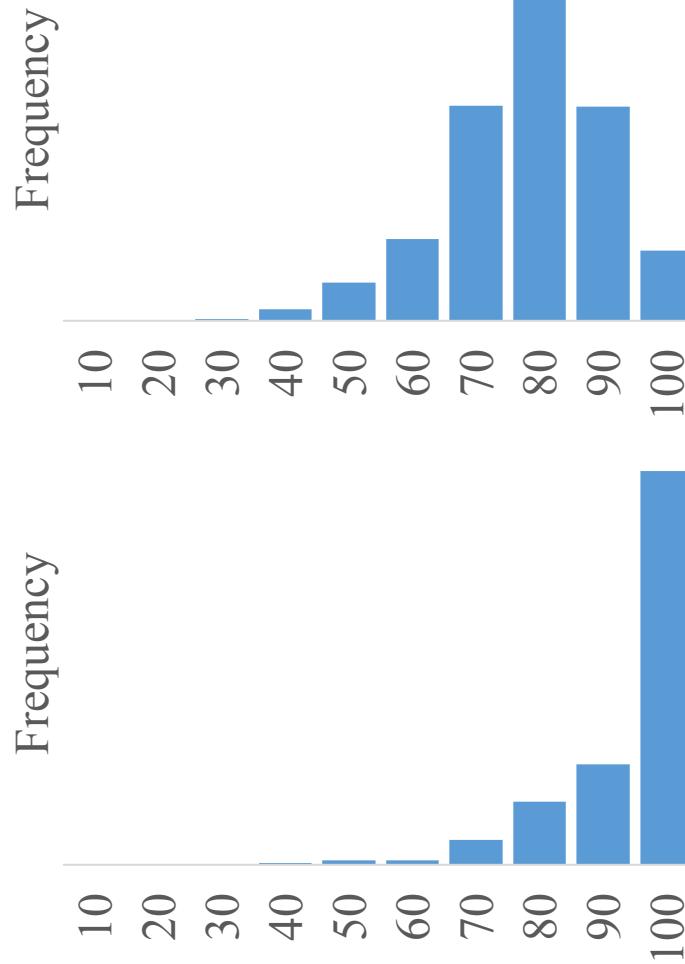
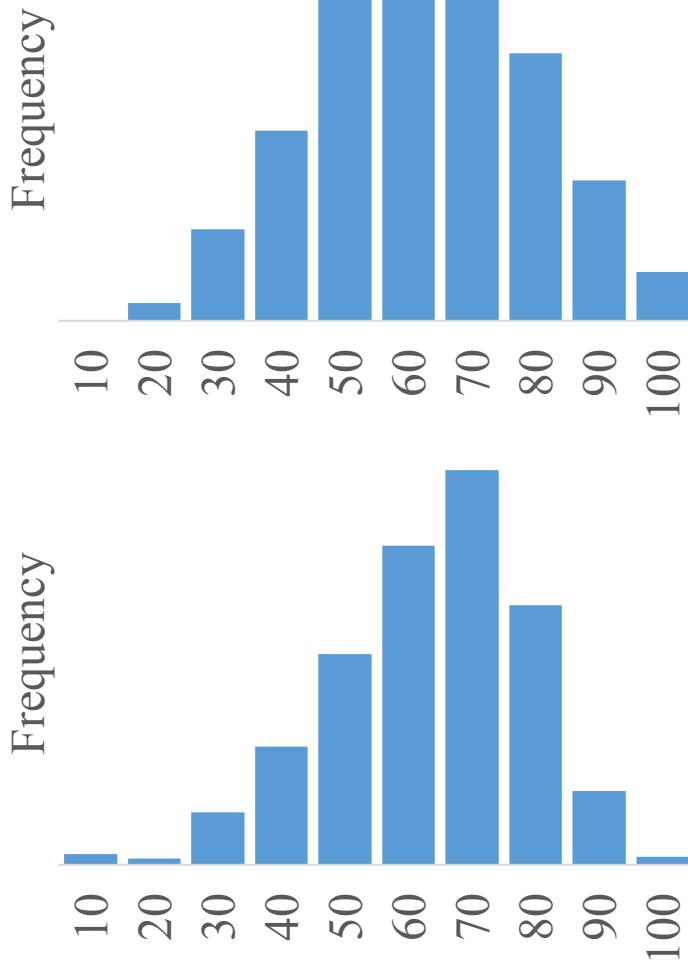


# The Random Variable for Probabilities

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 weird at first.  
But then it gets good.

Its like biking with a  
helmet...

# 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:

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

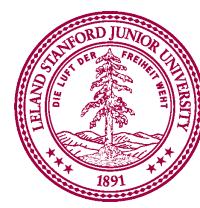
# Anomaly Detection

- 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?:

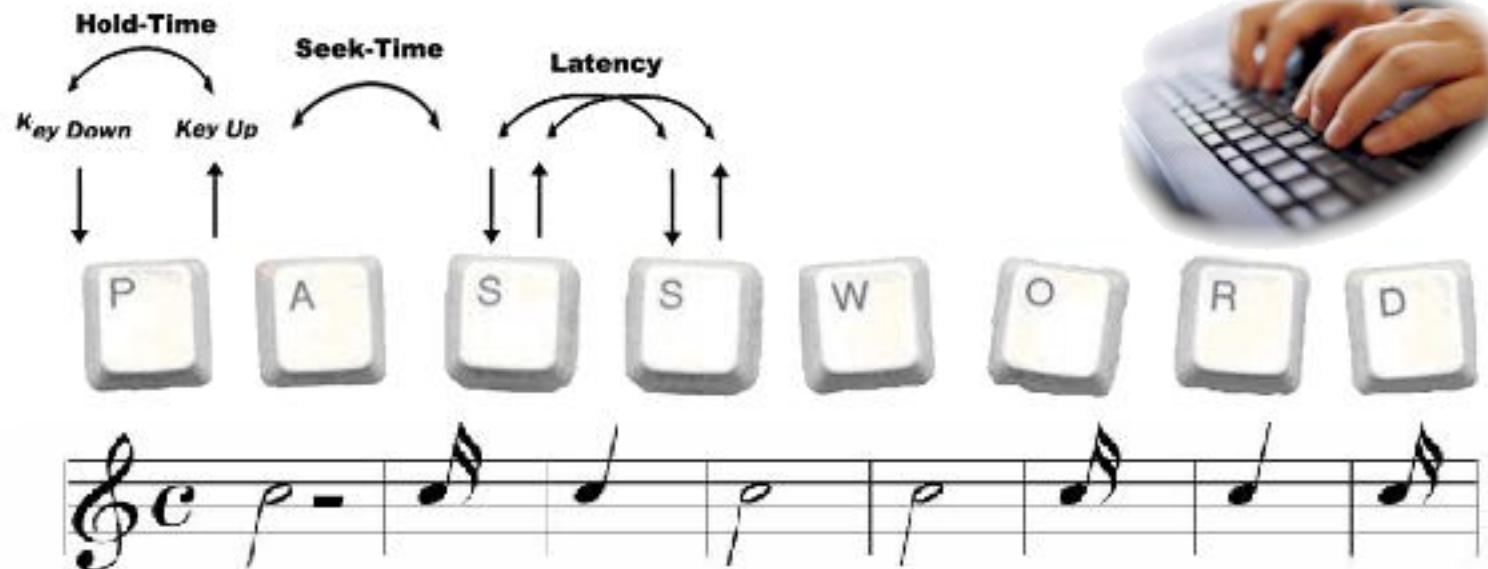
$$P(E|X = x) = \frac{f_X(x|E)P(E)}{f_X(x)}$$

Normal pdf                      Prior  
↓                                  ↙  
                                    ???

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



# Biometric Keystroke



End Review

# Flip a Coin With Unknown Probability



Demo

# Roll a Die 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

$$\begin{aligned} X &= \lim_{n+m \rightarrow \infty} \frac{n}{n+m} \\ &\approx \frac{n}{n+m} \end{aligned}$$

$X$  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:  $X \sim \text{Uni}(0, 1)$
  - Let  $N = \text{number of heads}$
  - Given  $X = x$ , coin flips independent:  $(N | X) \sim \text{Bin}(n + m, x)$

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

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:  $X \sim \text{Uni}(0, 1)$
  - Let  $N = \text{number of heads}$
  - Given  $X = x$ , coin flips independent:  $(N | X) \sim \text{Bin}(n + m, x)$

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

Binomial

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

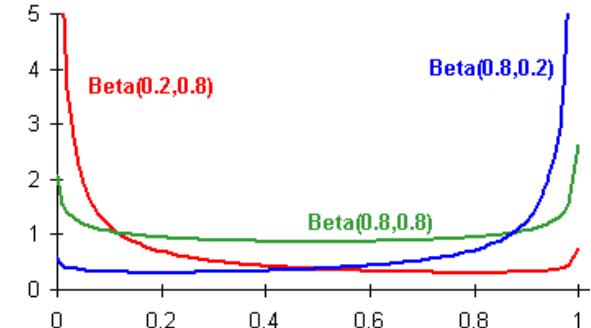
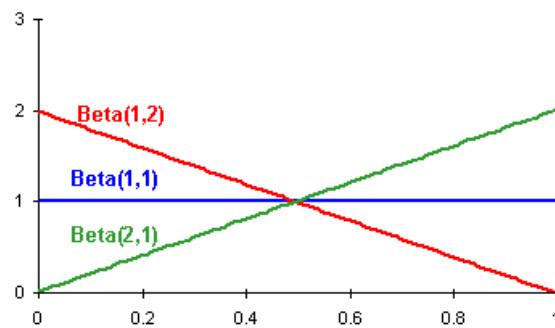
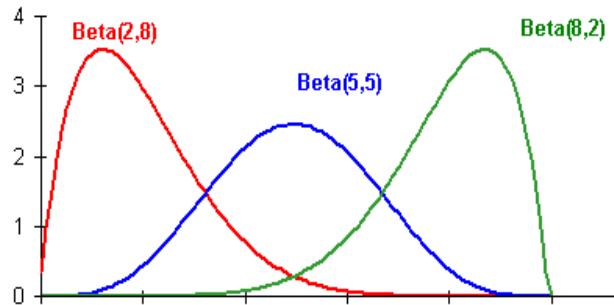
$$= \frac{1}{c} \cdot x^n (1-x)^m \quad \text{where } c = \int_0^1 x^n (1-x)^m dx$$

Move terms  
around

# Beta Random Variable

- $X$  is a Beta Random Variable:  $X \sim \text{Beta}(a, b)$ 
  - Probability Density Function (PDF): (where  $a, b > 0$ )

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

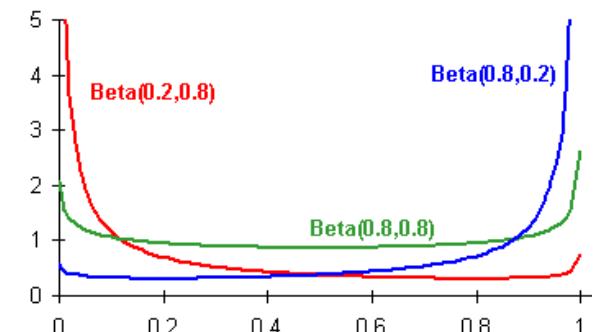
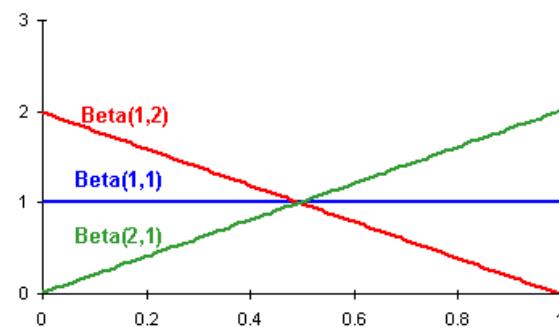
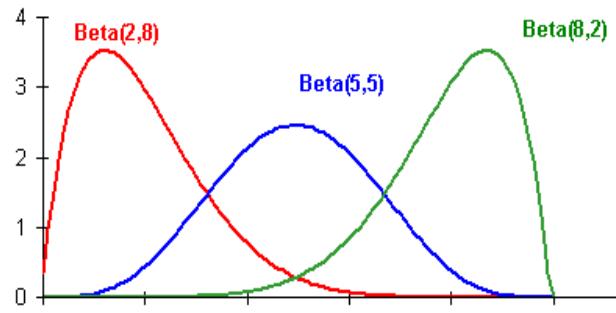


- Symmetric when  $a = b$

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

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

# Meta Beta



Used to represent a distributed belief of a probability



Beta is a distribution for  
probabilities



# 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:  $X \sim \text{Uni}(0, 1)$
  - Let  $N = \text{number of heads}$
  - Given  $X = x$ , coin flips independent:  $(N | X) \sim \text{Bin}(n + m, x)$

$$\begin{aligned}f_{X|N}(x|n) &= \frac{P(N = n | X = x) f_X(x)}{P(N = n)} \\&= \frac{\binom{n+m}{n} x^n (1-x)^m}{P(N = n)} \\&= \frac{\binom{n+m}{n}}{P(N = n)} x^n (1-x)^m \\&= \frac{1}{c} \cdot x^n (1-x)^m \quad \text{where } c = \int_0^1 x^n (1-x)^m dx\end{aligned}$$

# Dude, Where's My Beta?

- Flip a coin  $(n + m)$  times, comes up with  $n$  heads
  - Conditional density of  $X$  given  $N = n$

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

- Note:  $0 < x < 1$ , so  $f_{X|N}(x | n) = 0$  otherwise
- Recall Beta distribution:

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

- Hey, that looks more familiar now...
- $X | (N = n, n + m \text{ trials}) \sim \text{Beta}(n + 1, m + 1)$

# Understanding Beta

- $X | (N = n, m + n \text{ trials}) \sim \text{Beta}(n + 1, m + 1)$ 
  - $X \sim \text{Uni}(0, 1)$
  - Check this out, boss:
    - $\text{Beta}(1, 1) = ?$ 
$$f(x) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} = \frac{1}{B(a,b)} x^0 (1-x)^0$$
$$= \frac{1}{\int_0^1 1 dx} 1 = 1 \quad \text{where } 0 < x < 1$$
    - $\text{Beta}(1, 1) = \text{Uni}(0, 1)$
  - So,  $X \sim \text{Beta}(1, 1)$

# If the Prior was a Beta...

If our belief about X (that random variable for probability) was beta

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

What is our belief about X after observing N heads?

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

# If the Prior was a Beta...

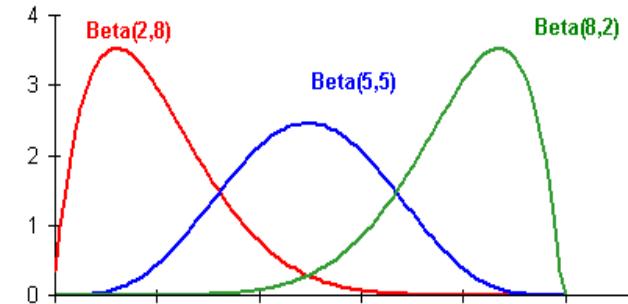
$$\begin{aligned} f(X = x | N = n) &= \frac{P(N = n | X = x) f(X = x)}{P(N = n)} \\ &= \frac{\binom{n+m}{n} x^n (1-x)^m f(X = x)}{P(N = n)} \\ &= \frac{\binom{n+m}{n} x^n (1-x)^m \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1}}{P(N = n)} \\ &= K_1 \cdot \binom{n+m}{n} x^n (1-x)^m \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} \\ &= K_2 \cdot x^n (1-x)^m \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} \\ &= K_3 \cdot x^n (1-x)^m x^{a-1} (1-x)^{b-1} \\ &= K_3 \cdot x^{n+a-1} (1-x)^{m+b-1} \\ X | N &\sim \text{Beta}(n + a, m + b) \end{aligned}$$

# Understanding Beta

- 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

# Further Understanding Beta

- Can set  $X \sim \text{Beta}(a, b)$  as prior to reflect how biased you think coin is apriori
  - This is a subjective probability!
  - Then observe  $n + m$  trials, where  $n$  of trials are heads
- Update to get posterior probability
  - $X | (n \text{ heads in } n + m \text{ trials}) \sim \text{Beta}(a + n, b + m)$
  - Sometimes call  $a$  and  $b$  the “equivalent sample size”
  - Prior probability for  $X$  based on seeing  $(a + b - 2)$  “imaginary” trials, where  $(a - 1)$  of them were heads.
  - $\text{Beta}(1, 1) \sim \text{Uni}(0, 1) \rightarrow$  we haven’t seen any “imaginary trials”, so apriori know nothing about coin



# 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!

## Parameters

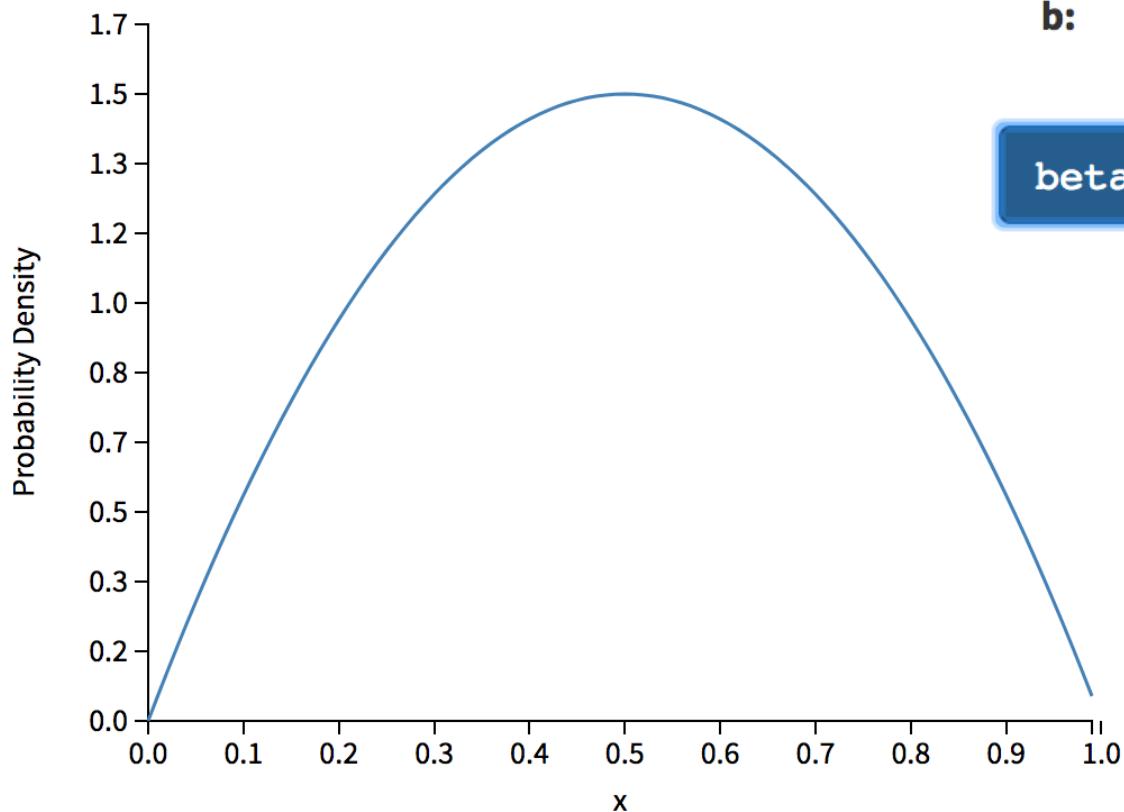
a:

2

b:

2

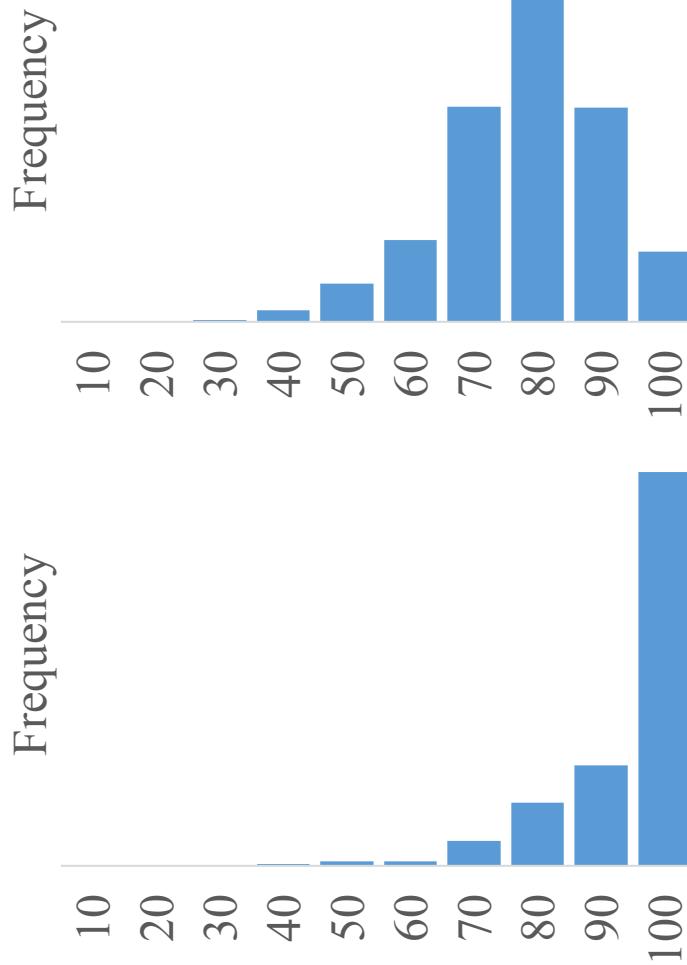
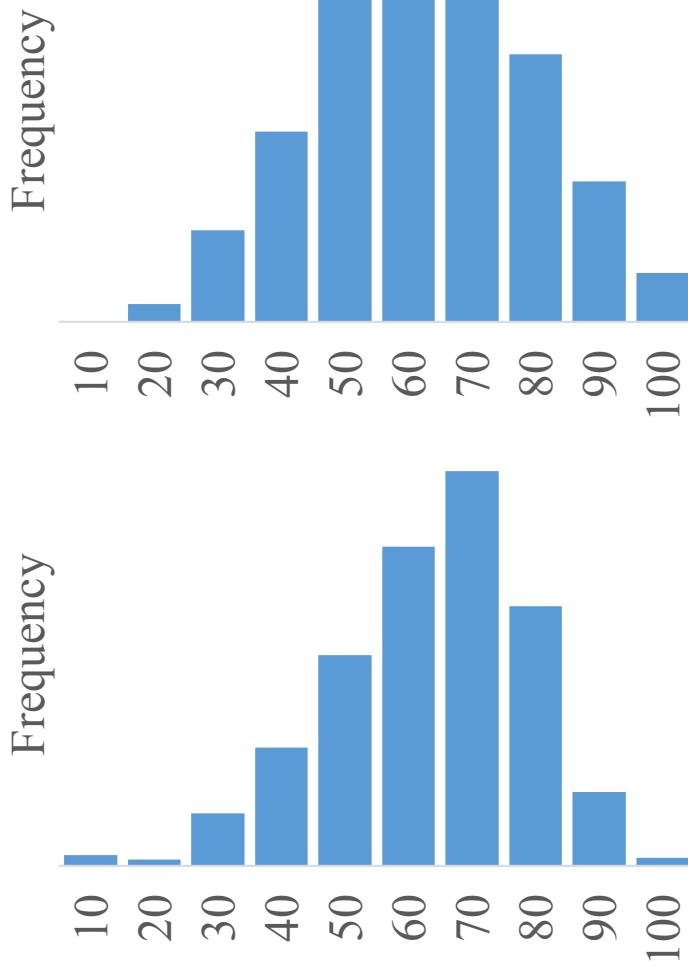
### Beta PDF



Damn

Next level?

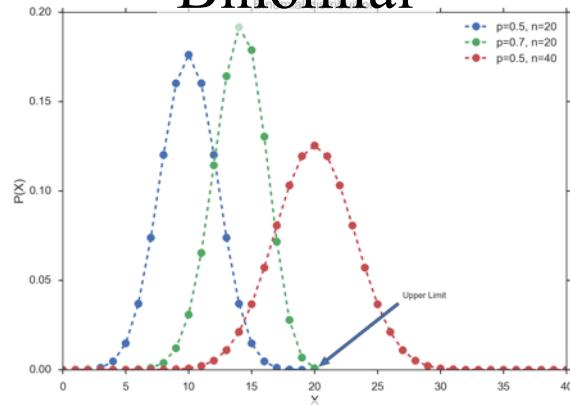
# Assignment Grades



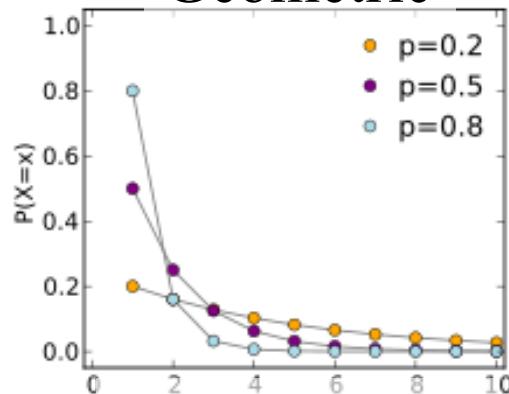
We have 2055 assignment distributions from gradescope

# Distributions

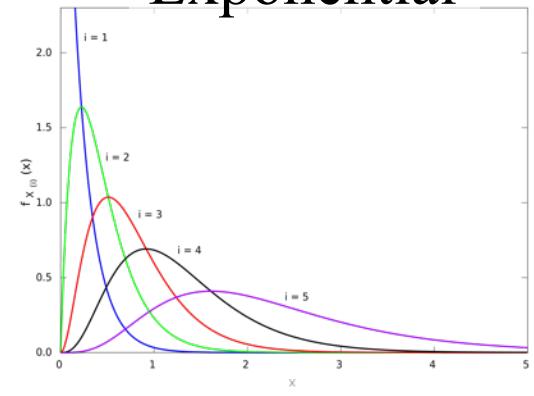
## Binomial



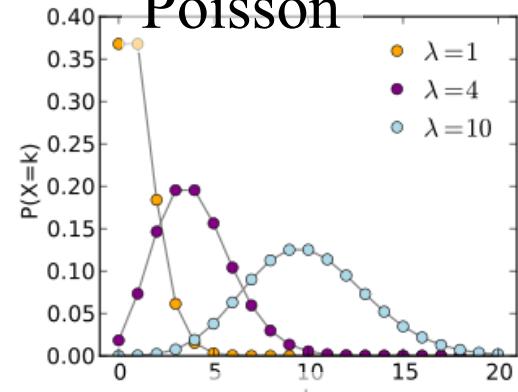
## Geometric



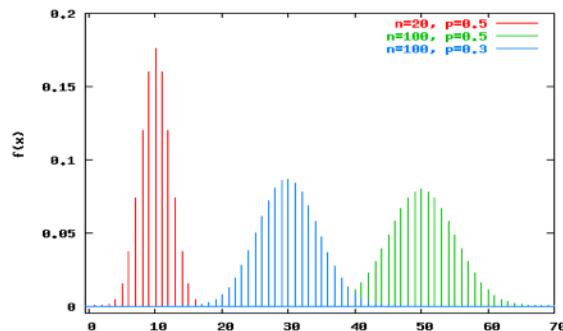
## Exponential



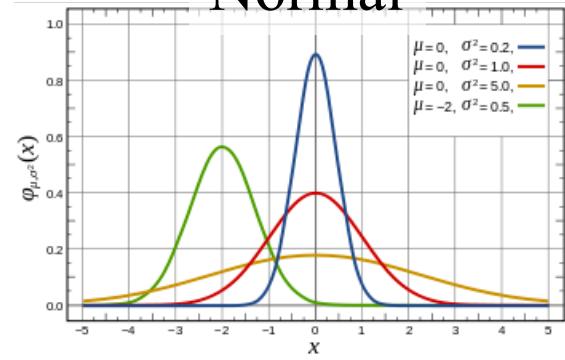
## Poisson



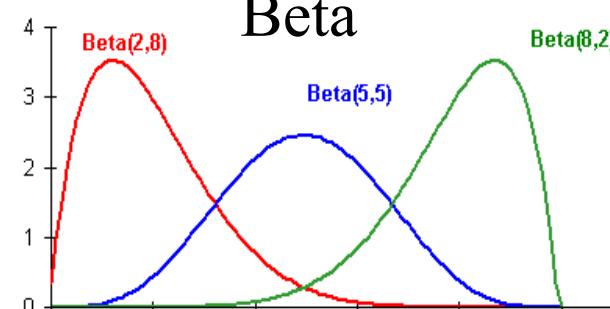
## Neg Binomial



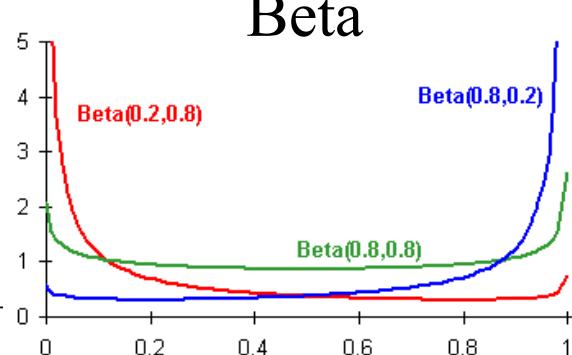
## Normal



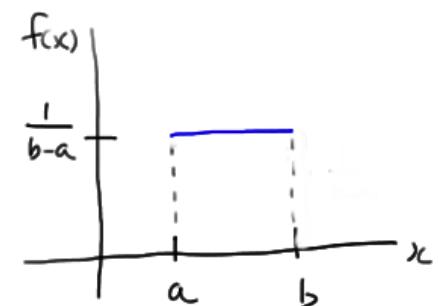
## Beta



## Beta



## Uniform



Grades must be bounded

Normal: No

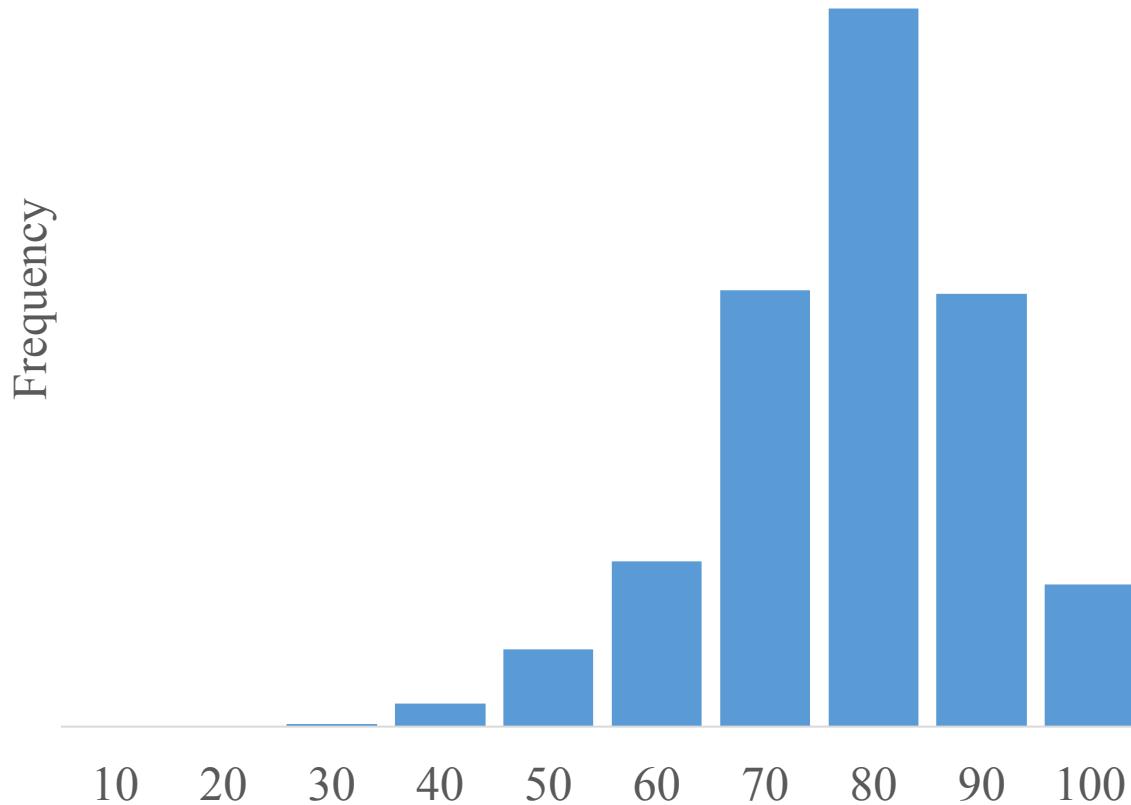
Poisson: No

Exponential: No

Beta: Yes

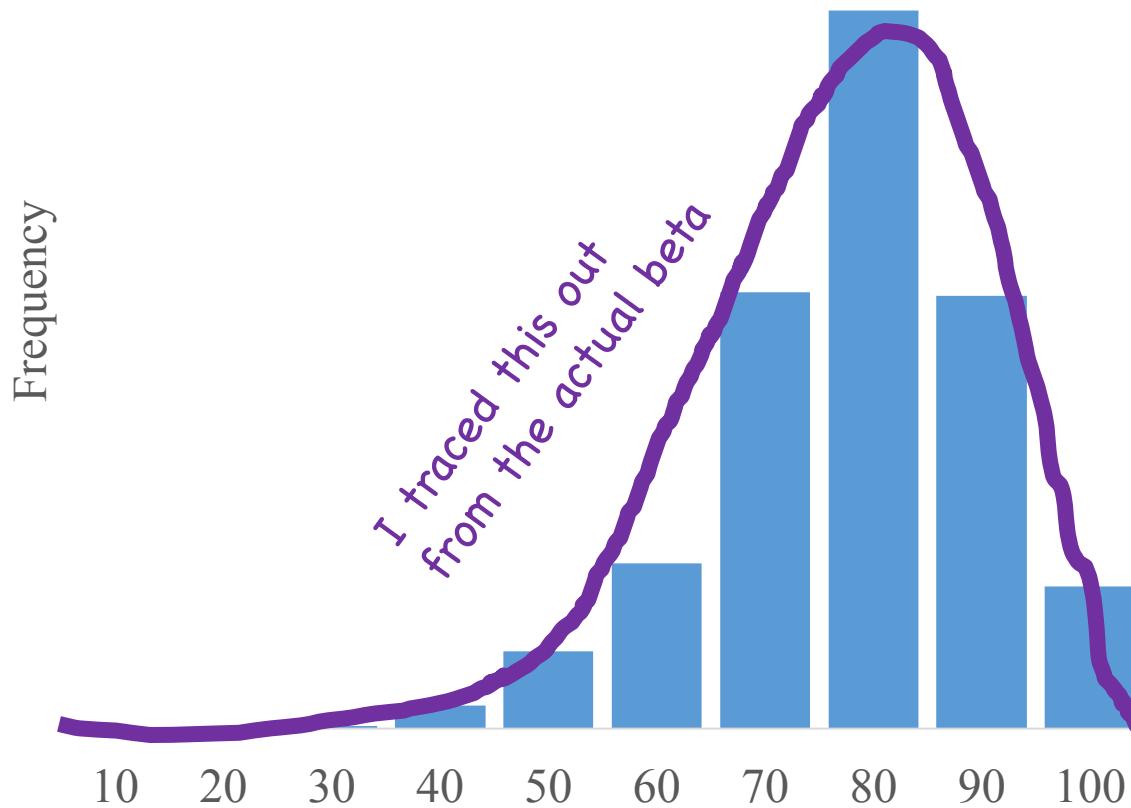
# Assignment Grades Demo

Assignment id = '1613'



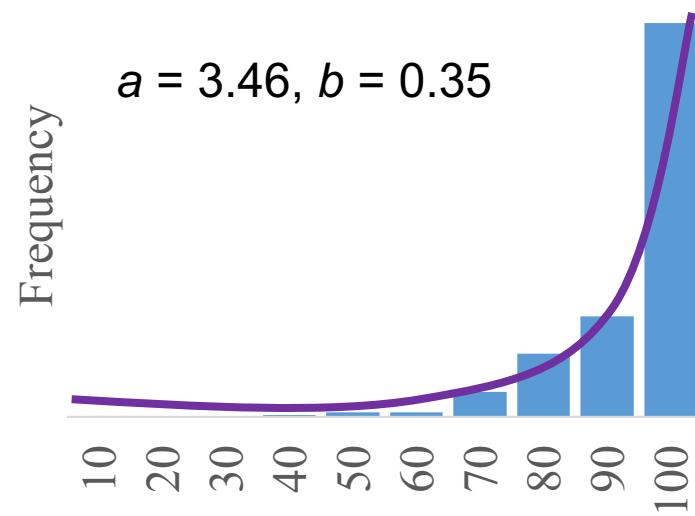
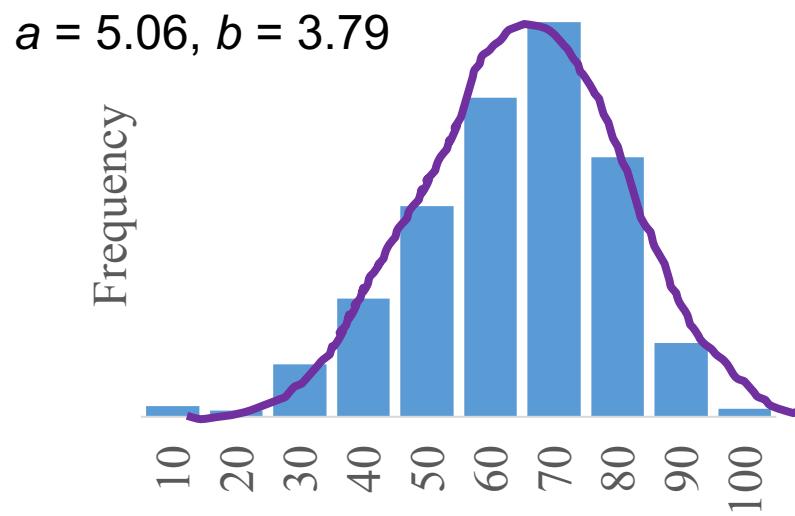
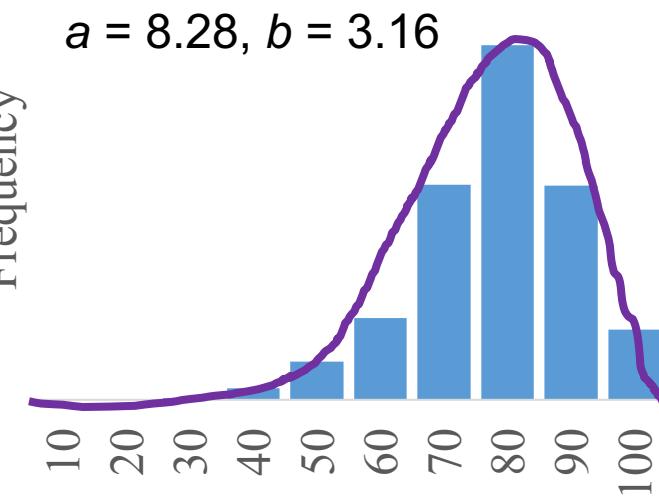
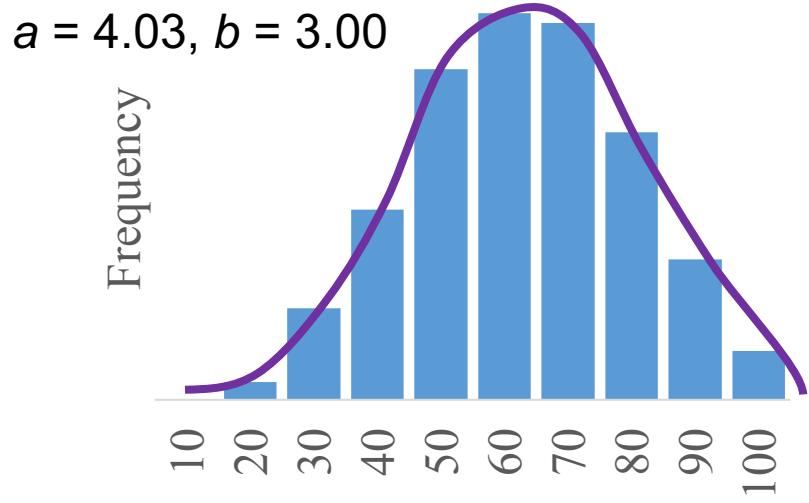
# Assignment Grades Demo

Assignment id = '1613'



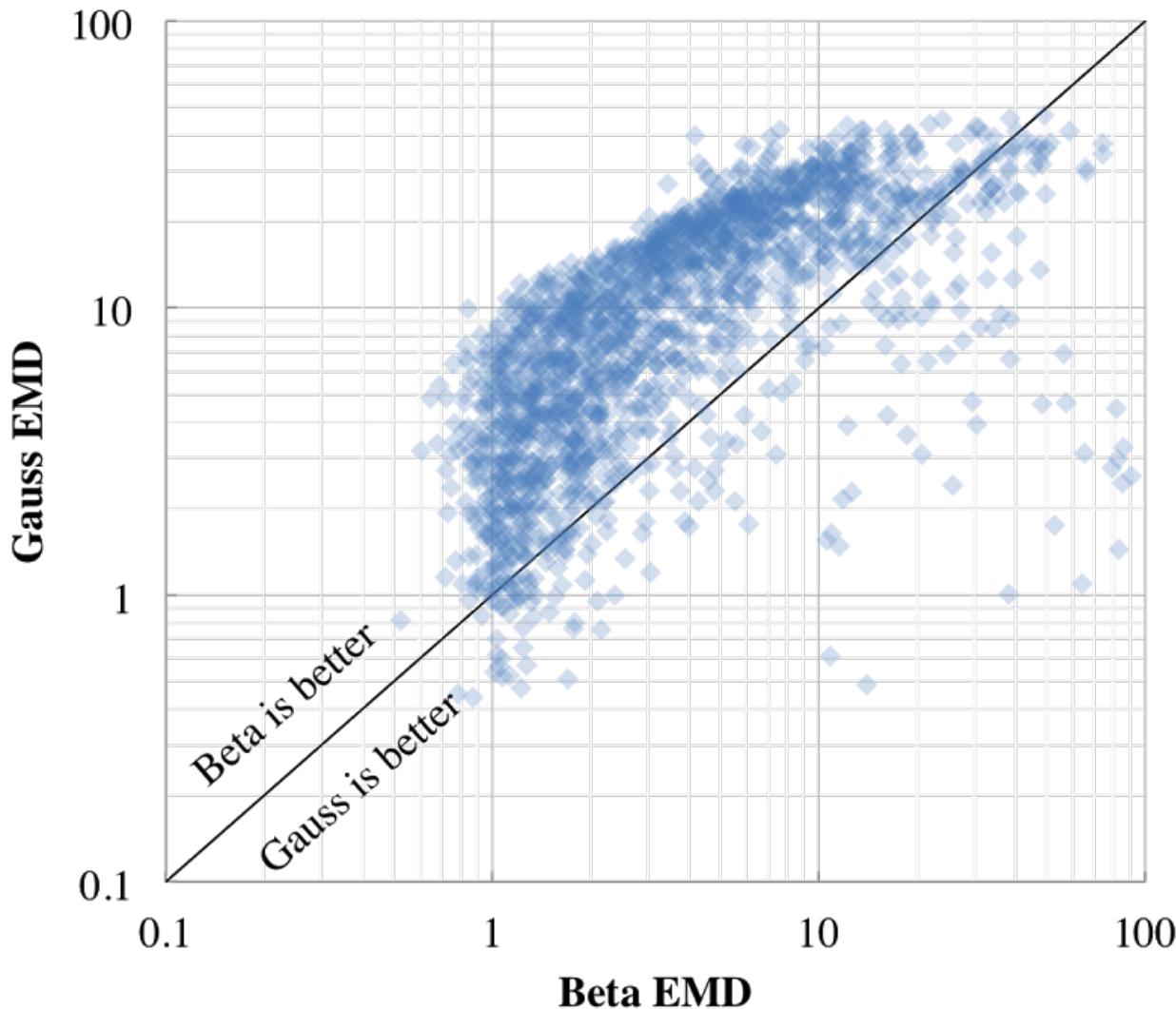
$$X \sim Beta(a = 8.28, b = 3.16)$$

# Assignment Grades



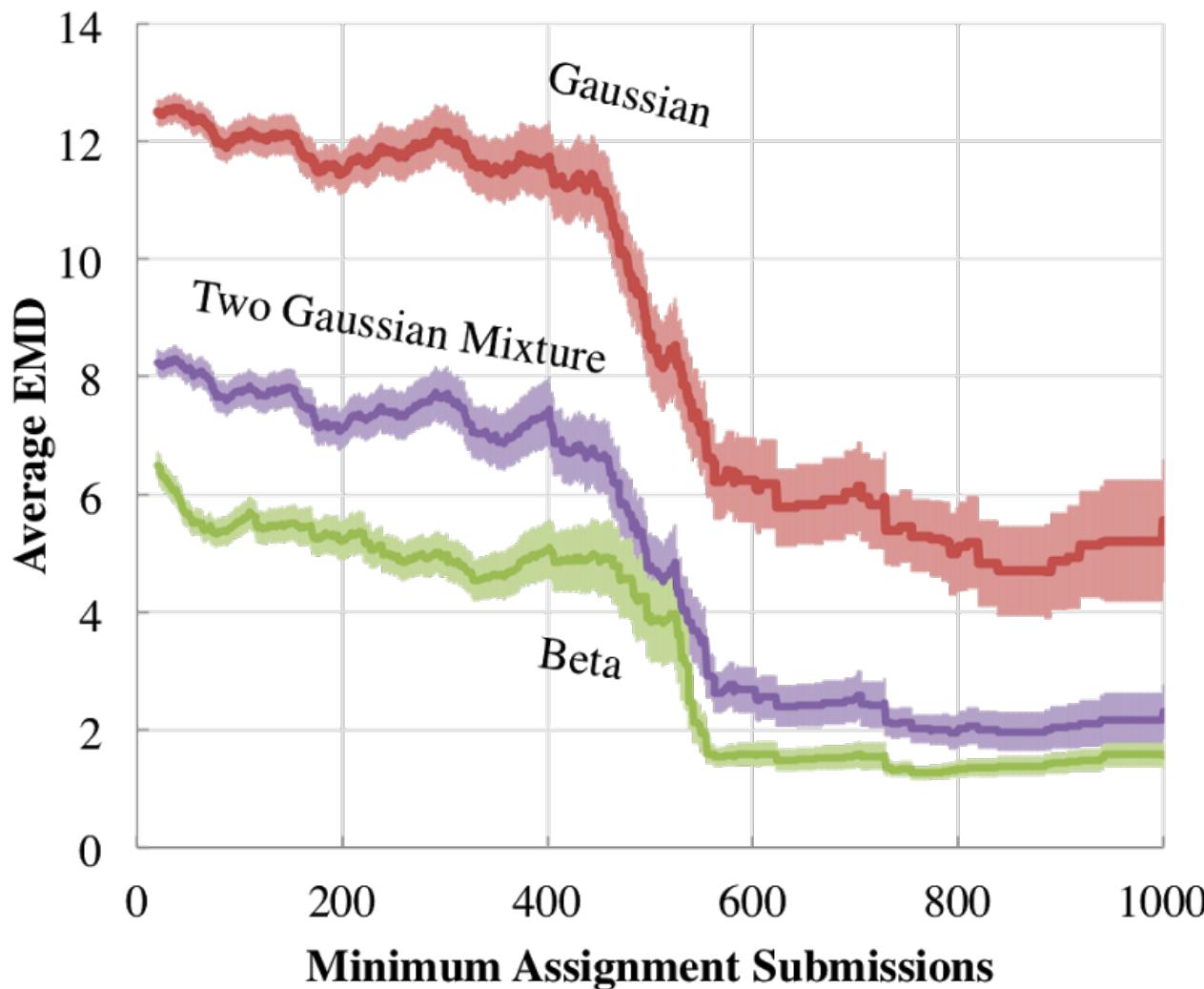
We have 2055 assignment distributions from grade scope

# 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

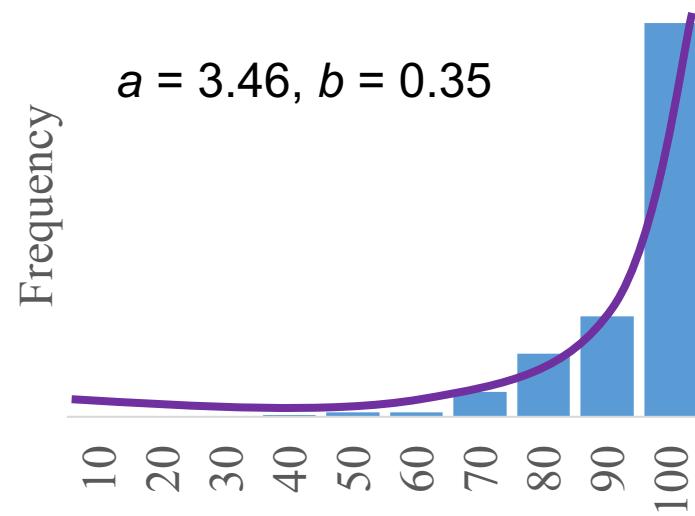
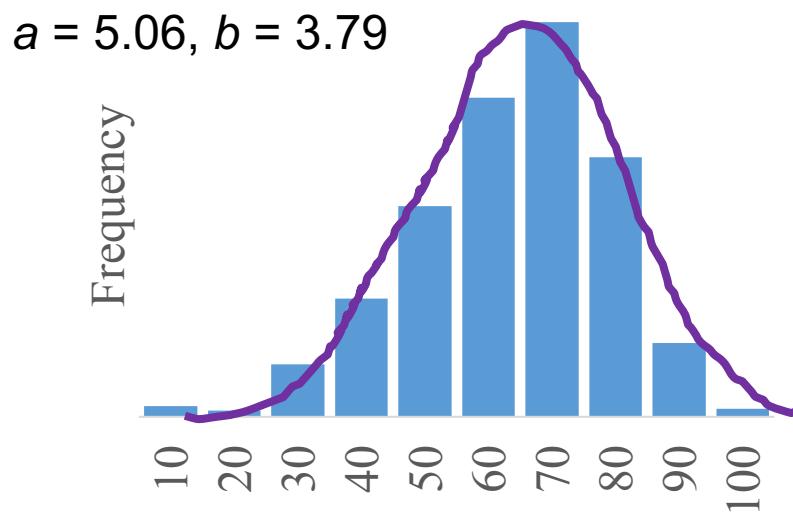
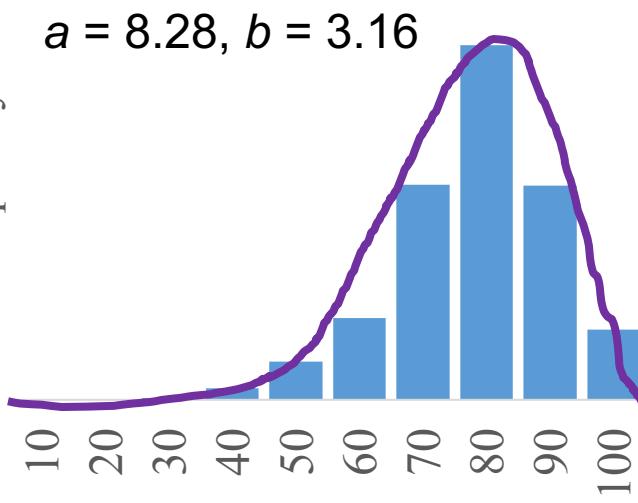
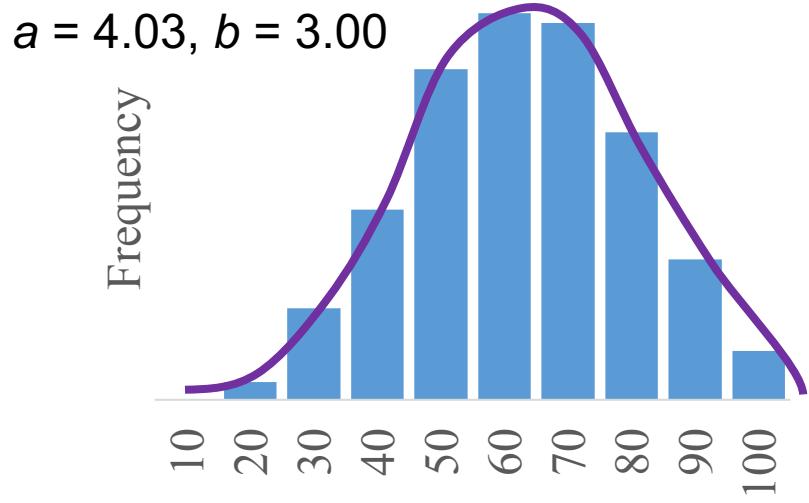
Each student's ability is represented as a probability – perhaps their probability of getting a generic point.

Each student has a **different** probability, however, the distribution of probabilities in a class is a Beta distribution.

- Beta

- This is Chris Piech's opinion. It is open for debate

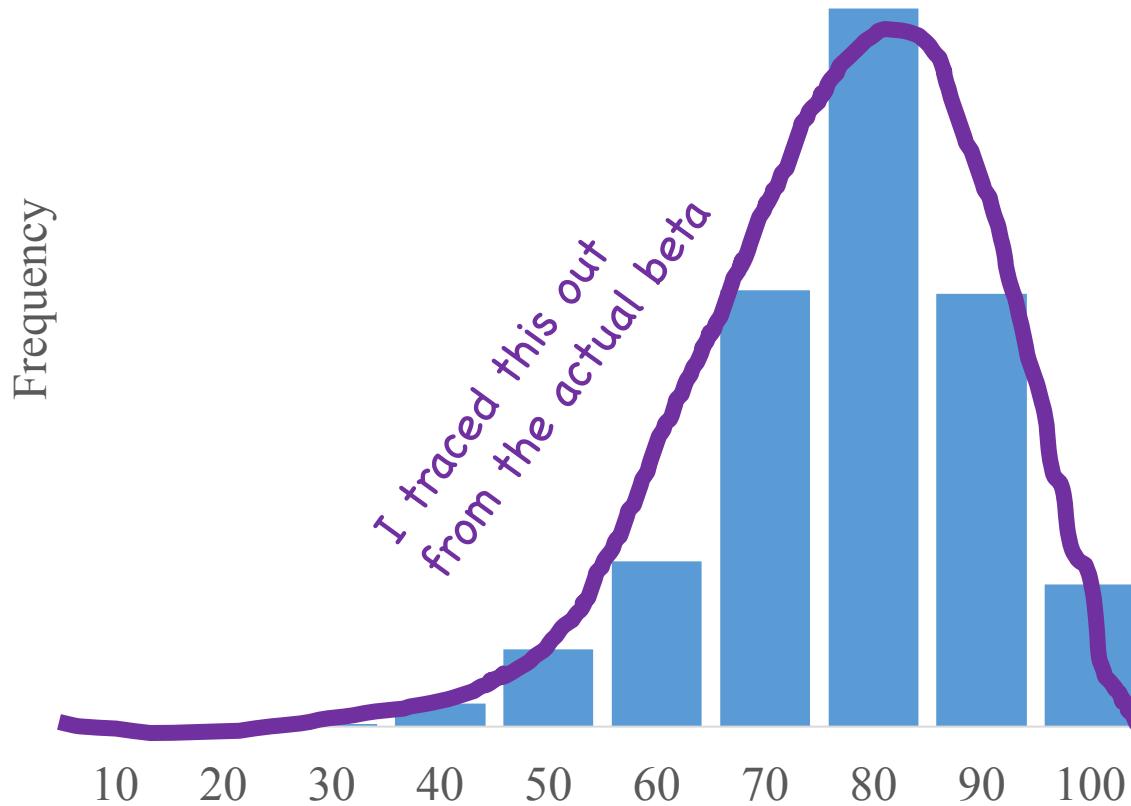
# Assignment Grades



These are the distribution of student *point probabilities*

# 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 below 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)
```

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

As far as I know, this is an  
unpublished result

# 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 fixed.

Will you use this on us?

Not yet ☺

Beta:  
The probability density  
for probabilities



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

# Course Mean

$E[CS109]$

*This is actual midpoint of course  
(Just wanted you to know)*