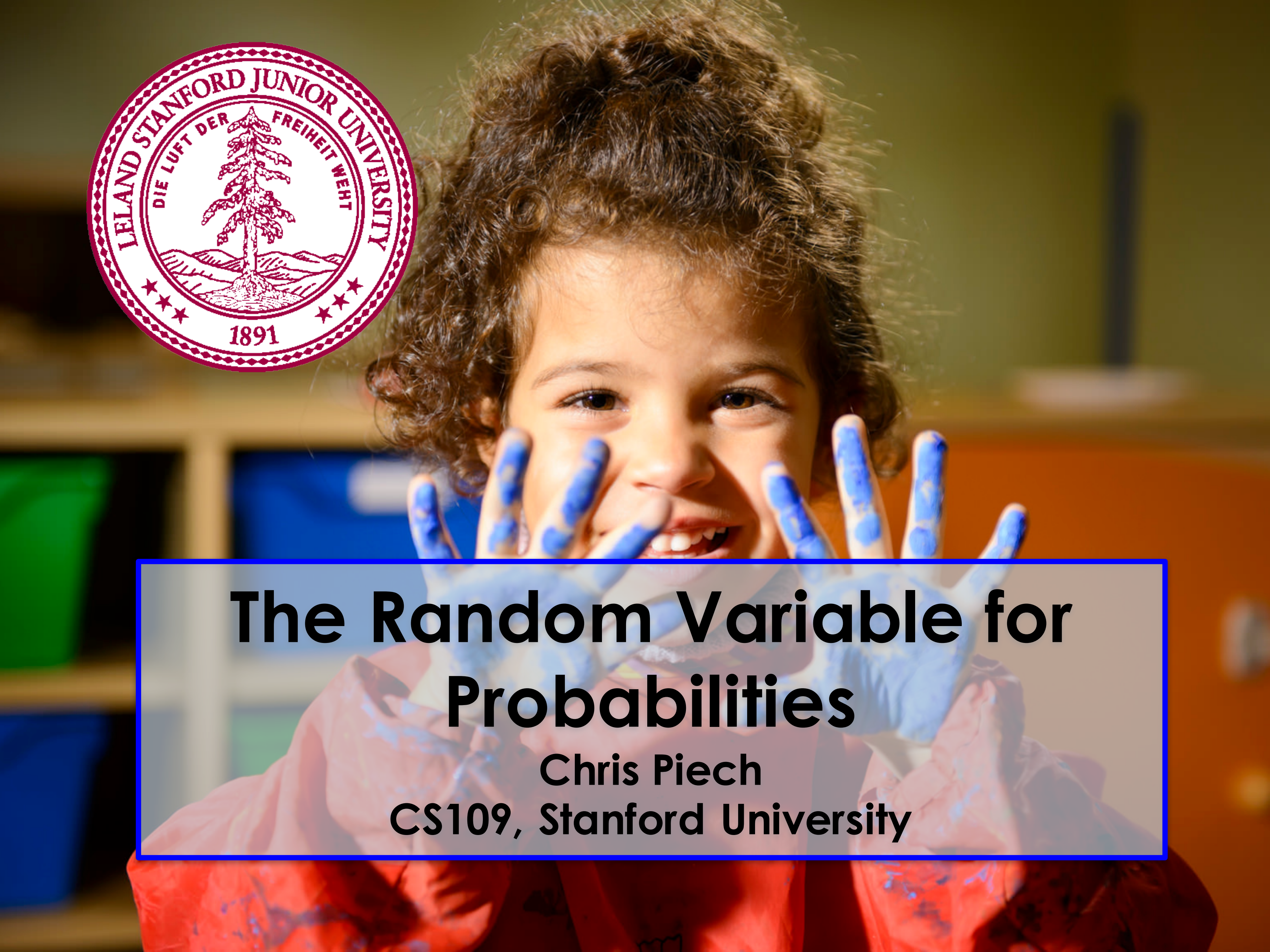
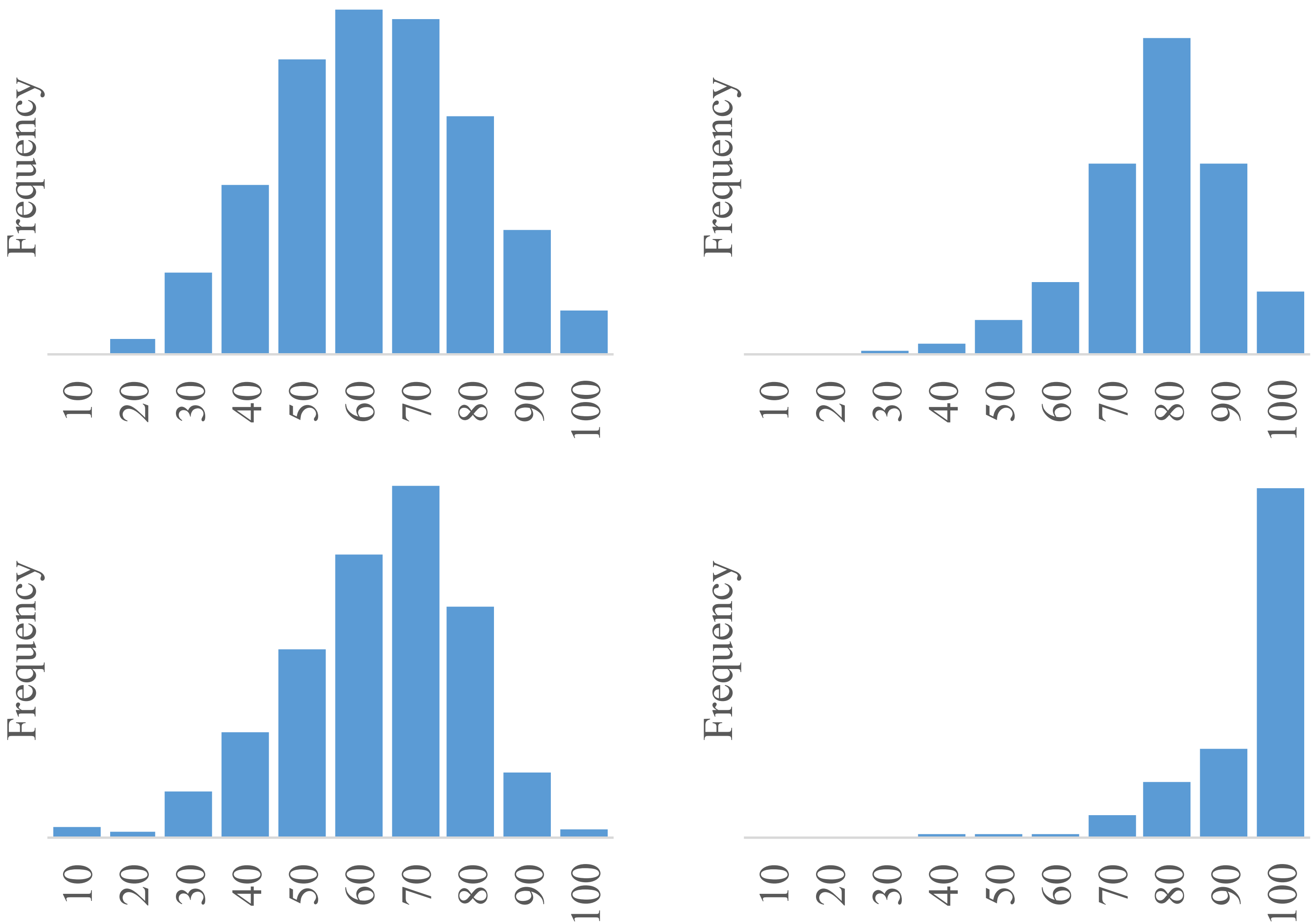


The Random Variable for Probabilities

Chris Piech
CS109, Stanford University

Assignment Grades



We have 2055 assignment distributions from grade scope

Flip a Coin With Unknown Probability

Demo

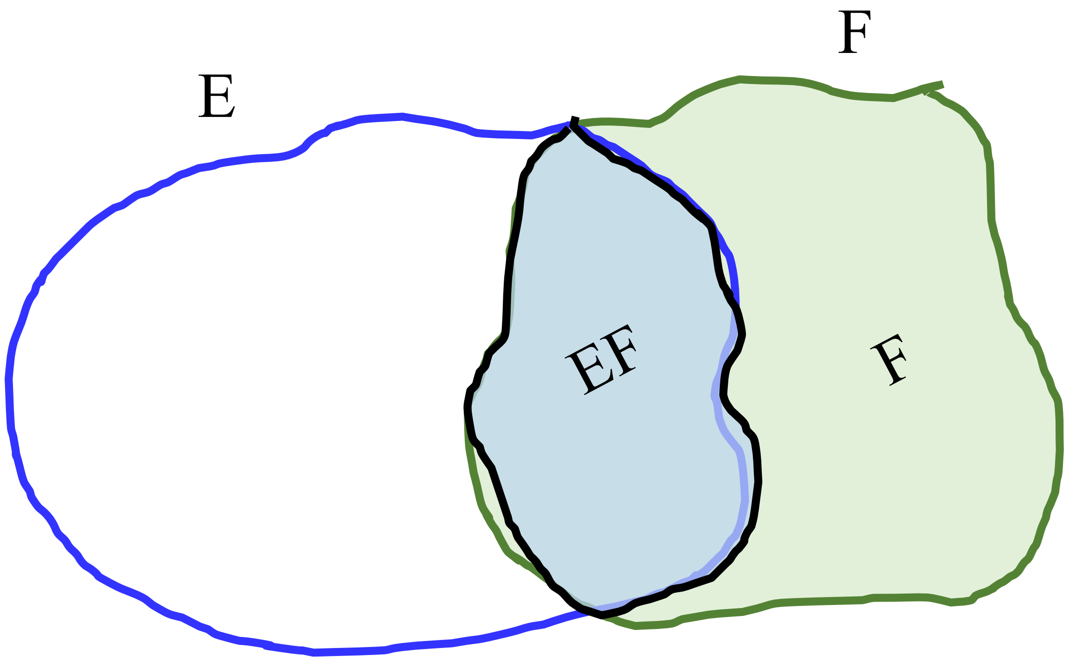
Today we are going to learn something unintuitive, beautiful and useful

Review

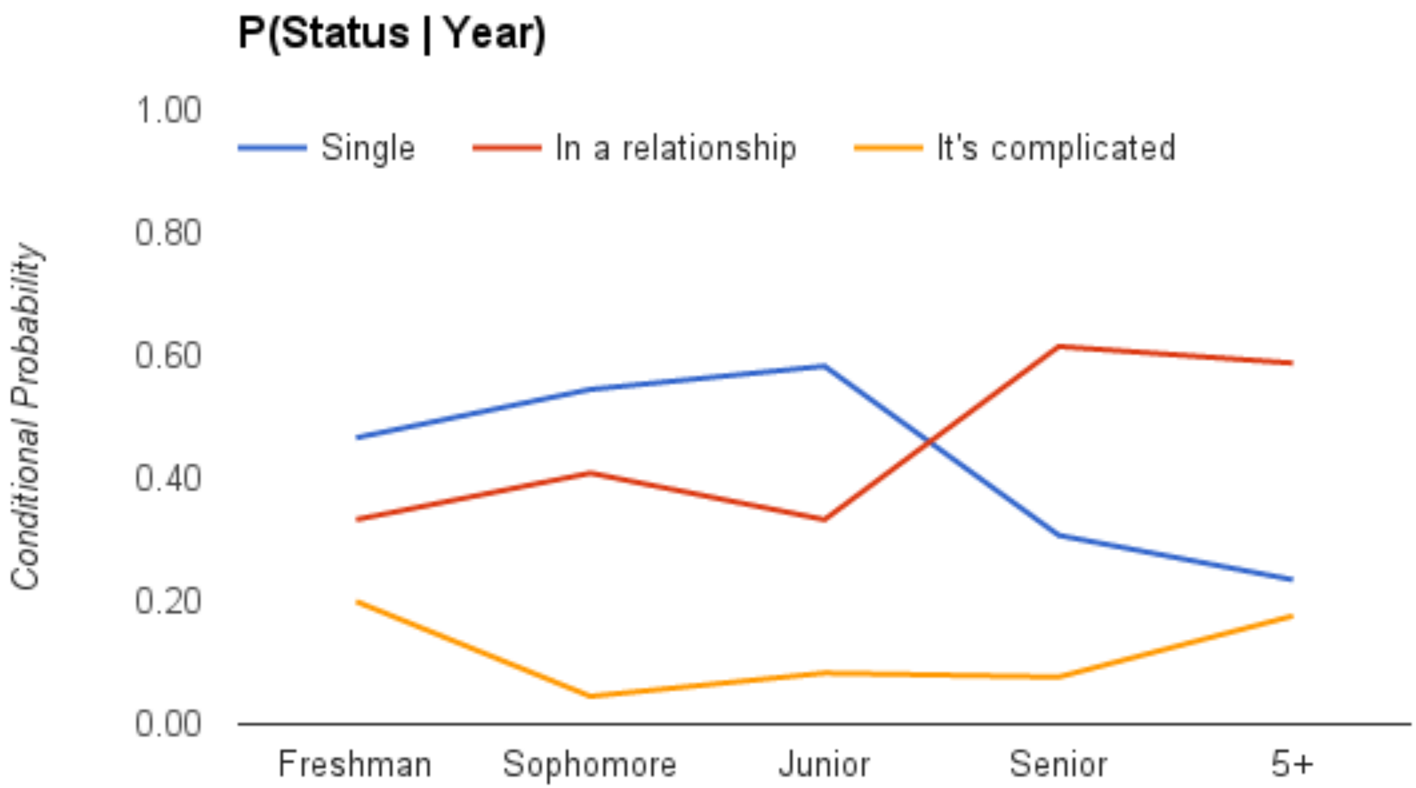
Conditional Events

- Recall that for *events* E and F:

$$P(E \mid F) = \frac{P(EF)}{P(F)} \quad \text{where} \quad P(F) > 0$$



Discrete Conditional Distributions



Continuous Conditional Distributions

- Let X and Y be continuous random variables
  - Conditional PDF of X given Y (where $f_Y(y) > 0$):

$$P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

$$f_{X|Y}(x, y)\partial x = \frac{f_{X,Y}(x, y)\partial x \partial y}{f_{\cancel{X}|Y}(y)\partial y}$$

$$f_{X|Y}(x, y)\partial x = \frac{f_{X,Y}(x, y)\partial x}{f_{\cancel{X}|Y}(y)}$$

$$f_{X|Y}(x, y) = \frac{f_{X,Y}(x, y)}{f_{\cancel{X}|Y}(y)}$$

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:

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

Anomaly Detection

- Let X be a measure of time to answer a question
- Let E be the event that the user is a human:

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

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 unconditional?:

Normal pdf

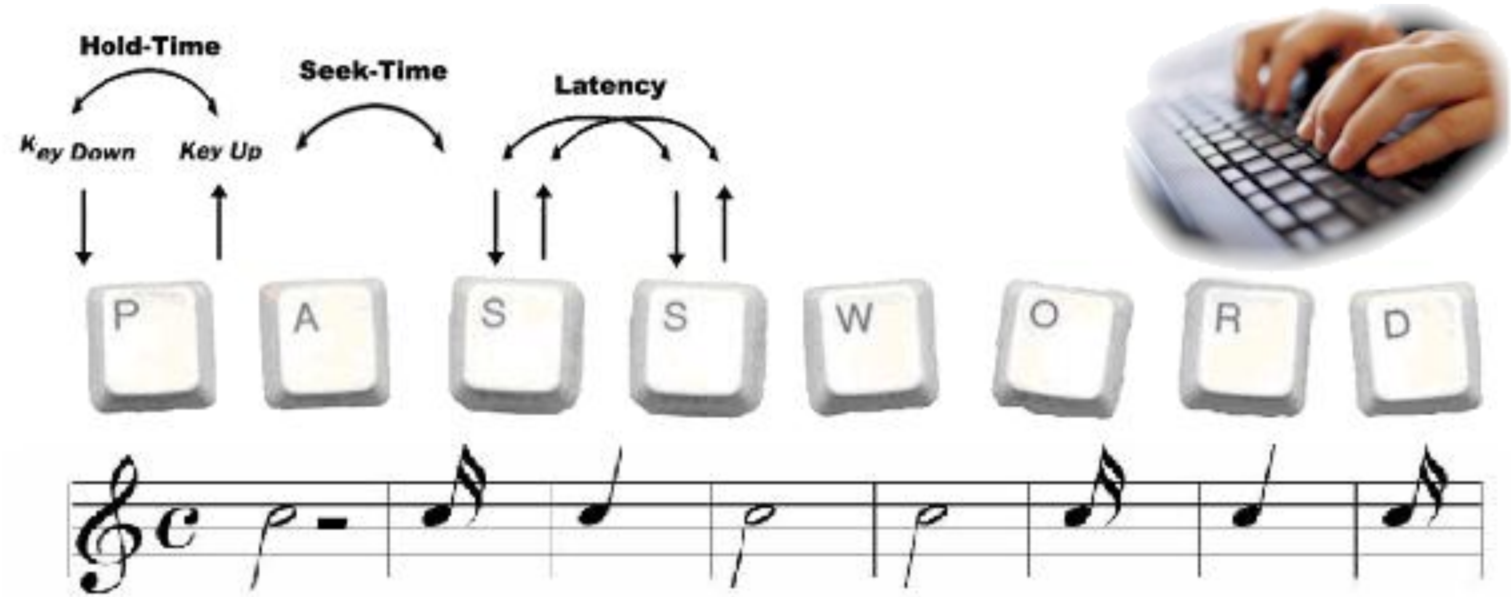
Prior

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

???

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

Biometric Keystroke



Mixing Discrete and Continuous

- Let X be a <u>continuous</u> random variable
- Let N be a <u>discrete</u> random variable
  - Conditional PDF of X given N:

$$f_{X|N}(x \mid n) = \frac{p_{N|X}(n \mid x) f_X(x)}{p_N(n)}$$

  - Conditional PMF of N given X:

$$p_{N|X}(n \mid x) = \frac{f_{X|N}(x \mid n) p_N(n)}{f_X(x)}$$

  - If X and N are independent, then:

$$f_{X|N}(x \mid n) = f_X(x) \qquad p_{N|X}(n \mid x) = p_N(n)$$

End Review


CITY

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 | 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)

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

1

Binomial

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

Move terms around

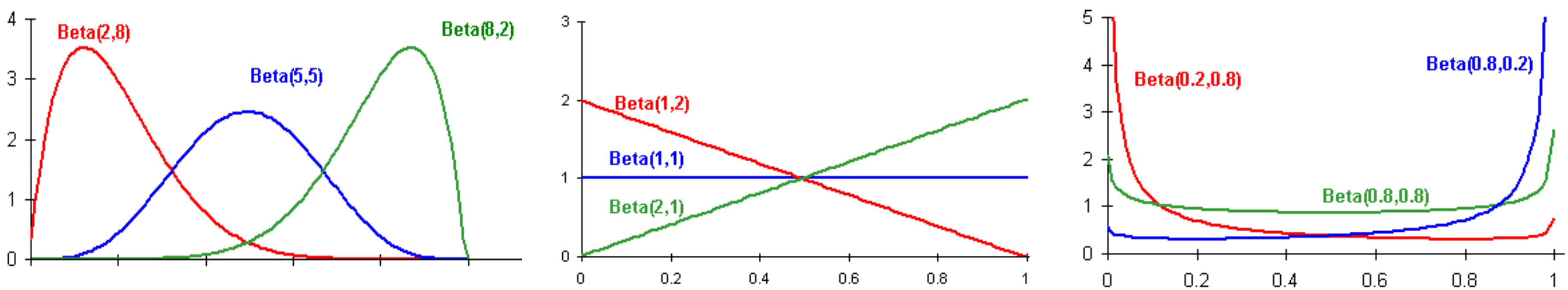
$$= \frac{\binom{n+m}{n}}{P(N=n)} x^n (1-x)^m$$

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

Beta Random Variable

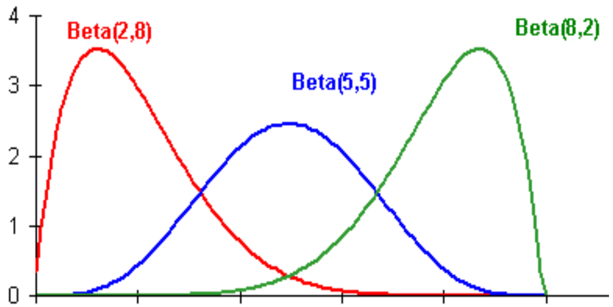
- X is a **Beta Random Variable**: X ~ Beta(*a*, *b*)
  - Probability Density Function (PDF): (where *a*, *b* > 0)

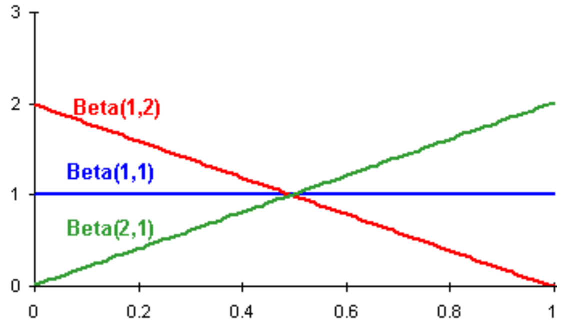
$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{where} \quad B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$

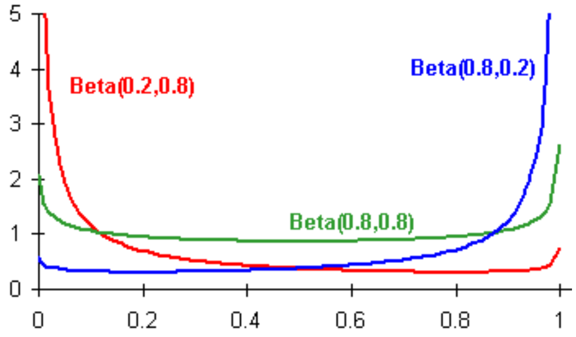


  - Symmetric when *a* = *b*
  - $E[X] = \dfrac{a}{a+b}$ $\quad Var(X) = \dfrac{ab}{(a+b)^2(a+b+1)}$

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 ~ Uni(0, 1)
  - Let N = number of heads
  - Given X = x, coin flips independent: (N | X) ~ Bin(n + m, x)

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

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

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

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

Dude, Where's My Beta?

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

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

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

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

  - Hey, that looks more familiar now...
  - X | (N = n, n + m trials) ~ Beta(n + 1, m + 1)

Understanding Beta

- X | (N = n, m + n trials) ~ Beta(n + 1, m + 1)
  - X ~ Uni(0, 1)
  - Check this out, boss:
    - Beta(1, 1) =?

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

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

    - Beta(1, 1) = Uni(0, 1)
  - So, X ~ 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?

$$f_{X|N}(x, n) = ???$$

If the Prior was a Beta...

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

Understanding Beta

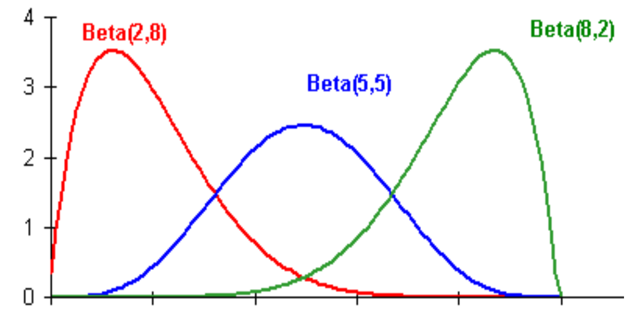
- If "Prior" distribution of X (before seeing flips) is Beta

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

- Beta is a <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!
  - Then observe n + m trials, where n of trials are heads



- Update to get posterior probability
  - X | (n heads in n + m trials) ~ 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.
  - Beta(1, 1) ~ Uni(0, 1) → we haven't seen any "imaginary trials", so apriori know nothing about coin

Check out Demo!

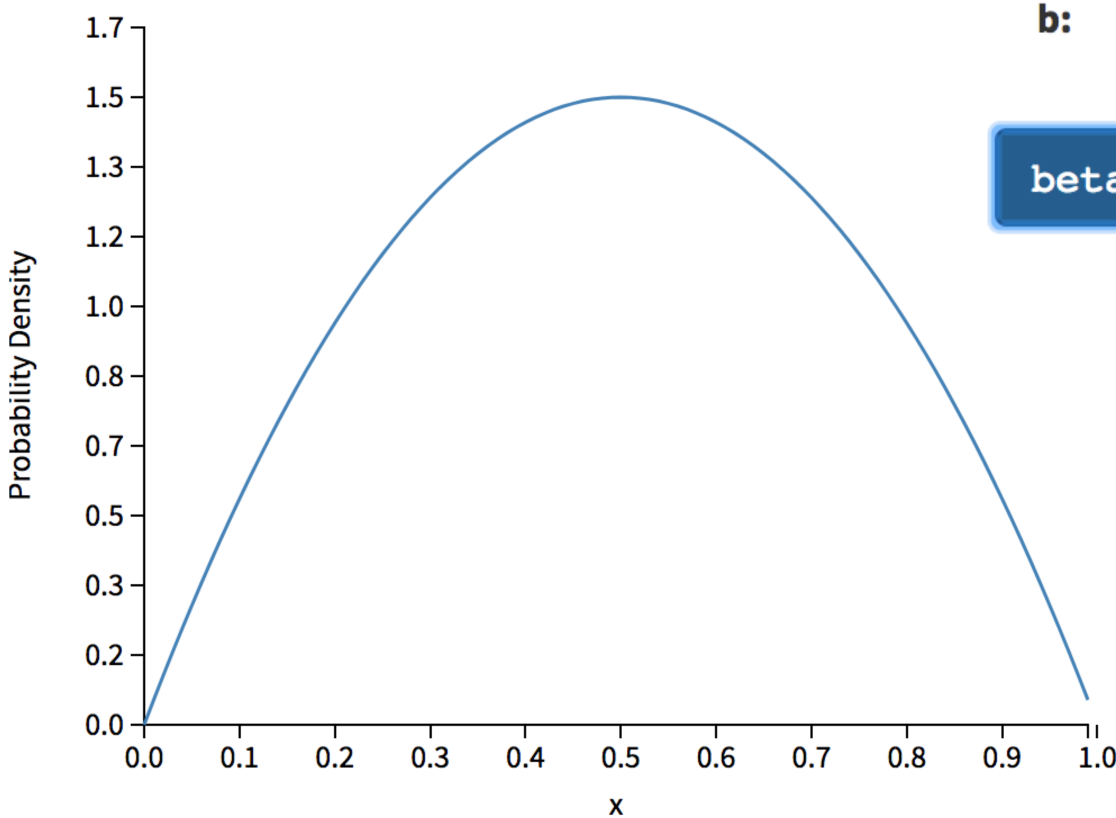
Parameters

a: 2

b: 2

beta pdf

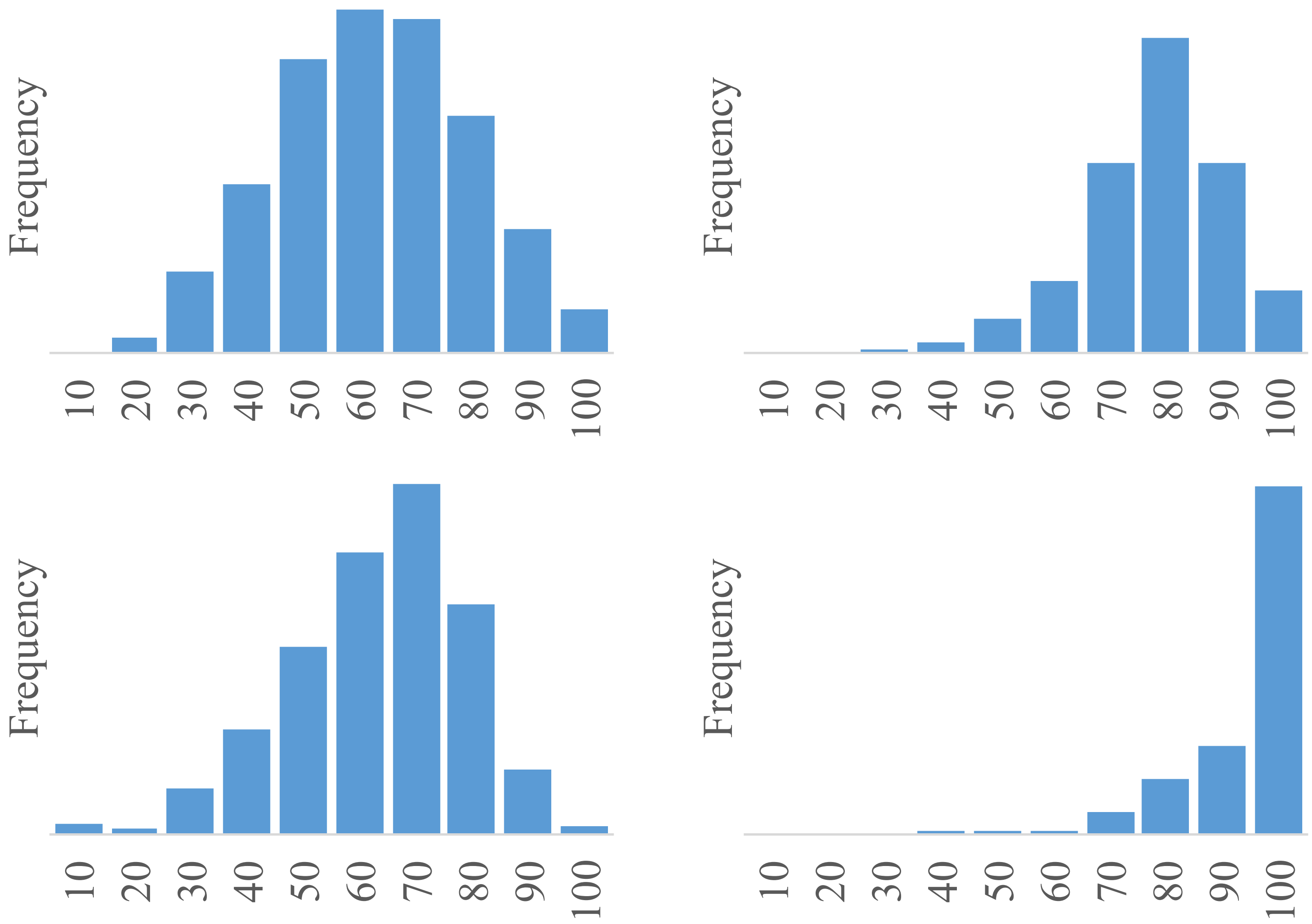
Beta PDF



Damn

Next level?

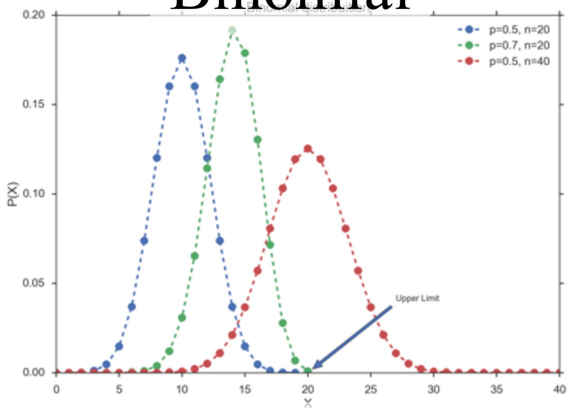
Assignment Grades



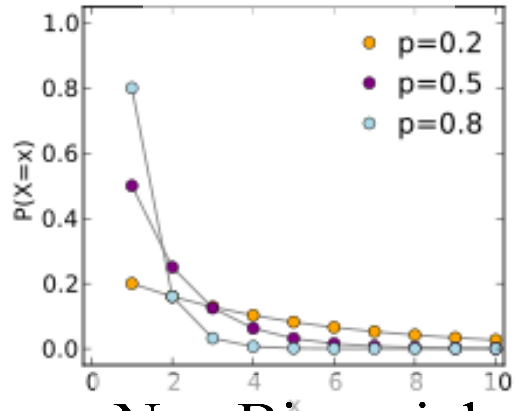
We have 2055 assignment distributions from gradescope

Distributions

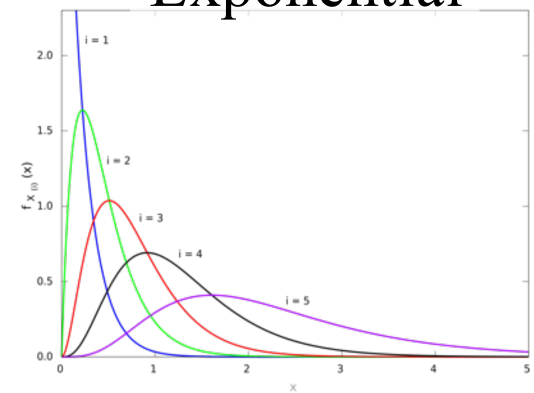
Binomial



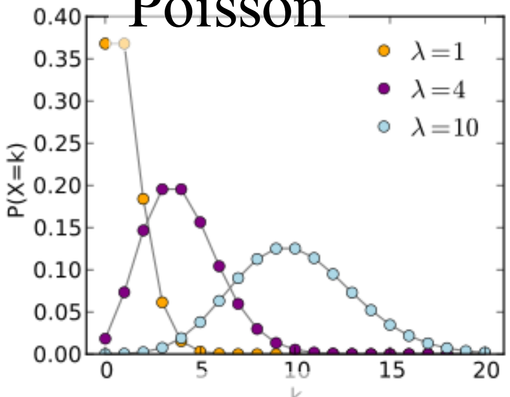
Geometric



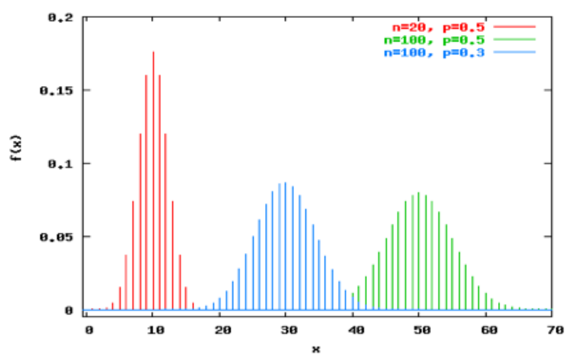
Exponential



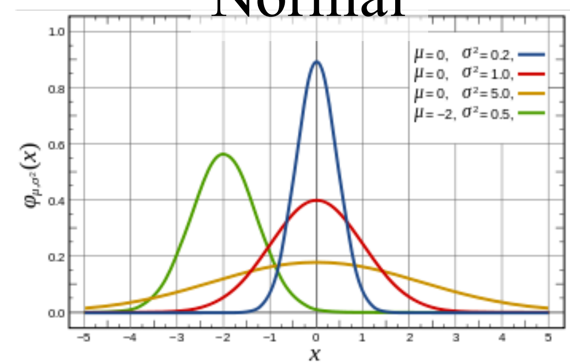
Poisson



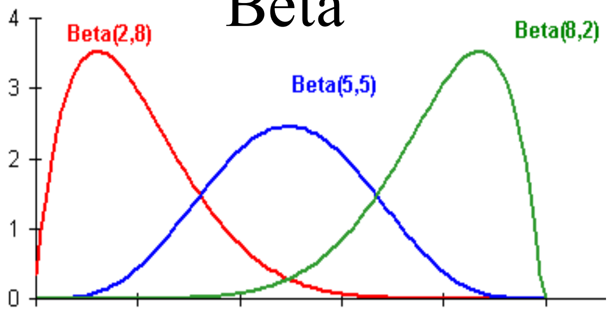
Neg Binomial



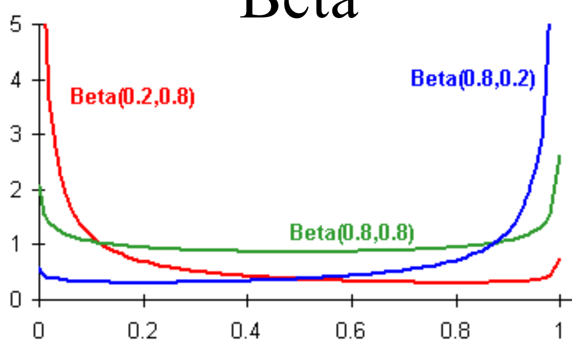
Normal



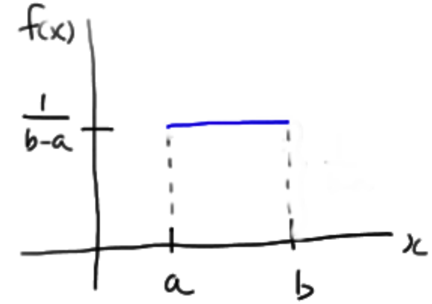
Beta



Beta



Uniform



Grades must be bounded

Normal: No

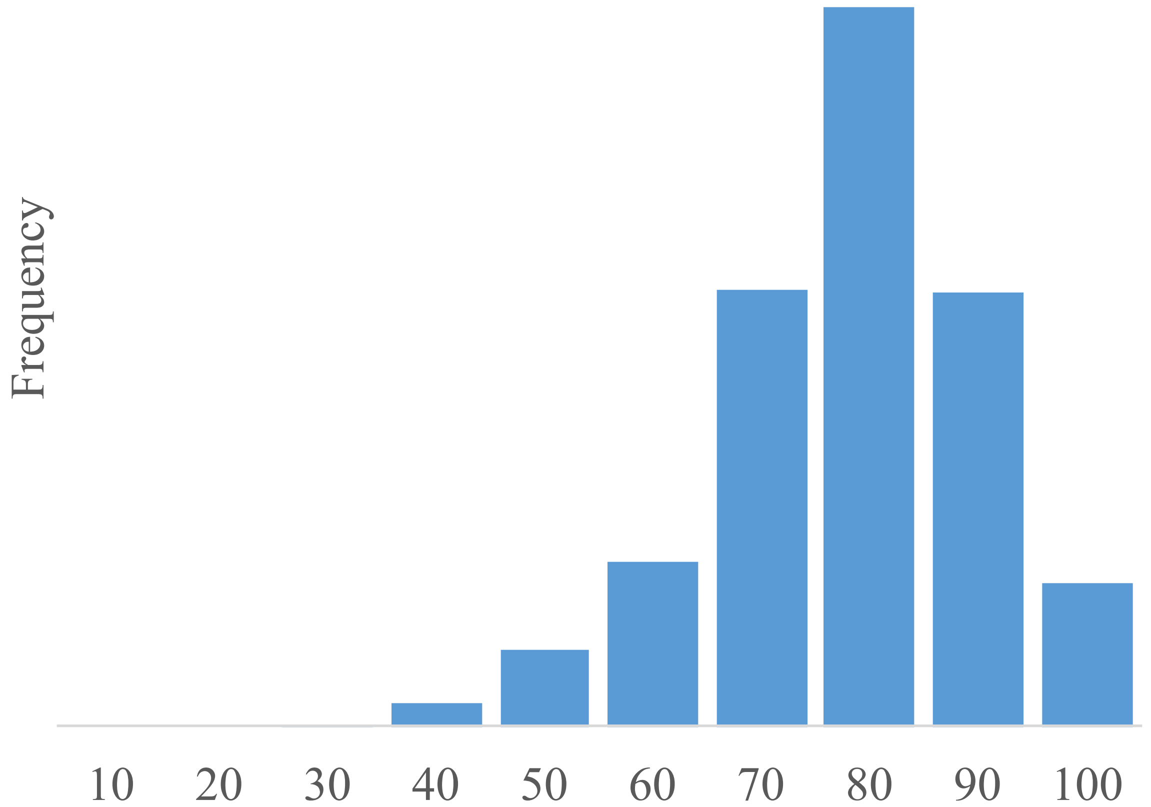
Poisson: No

Exponential: No

Beta: Yes

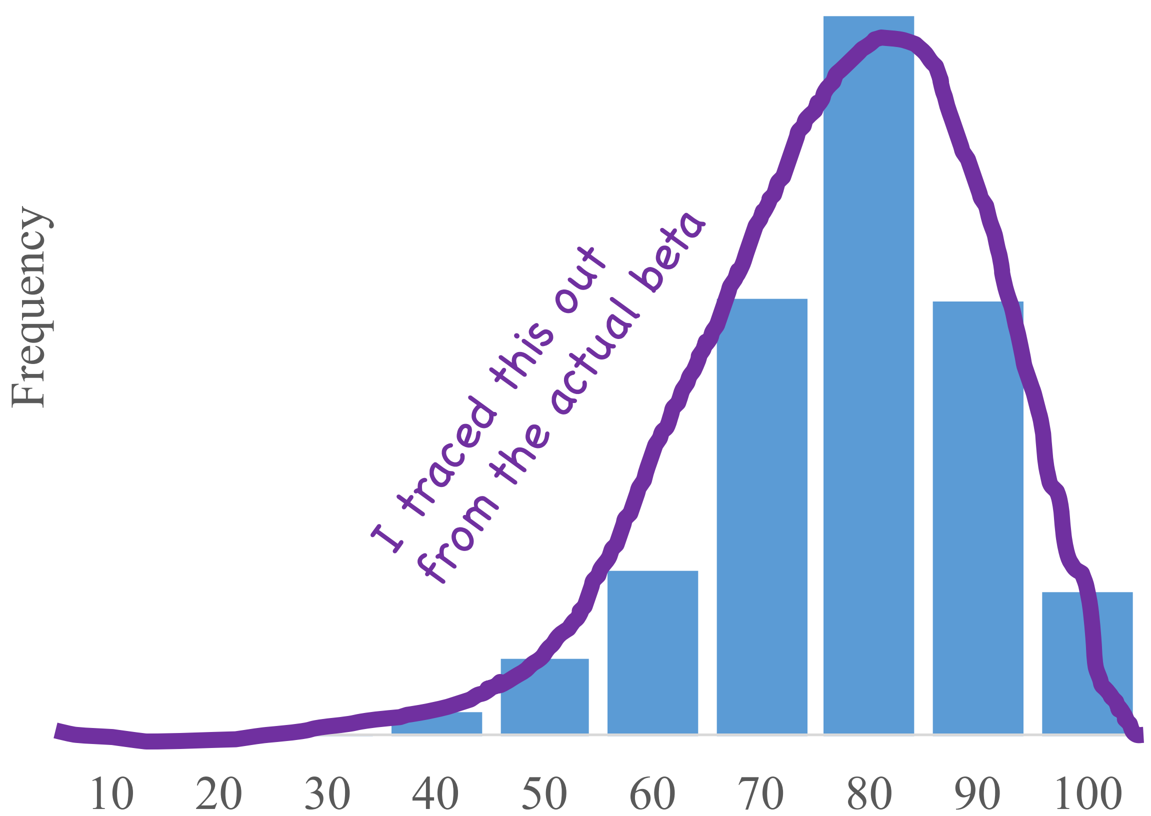
Assignment Grades Demo

`Assignment id = '1613'`



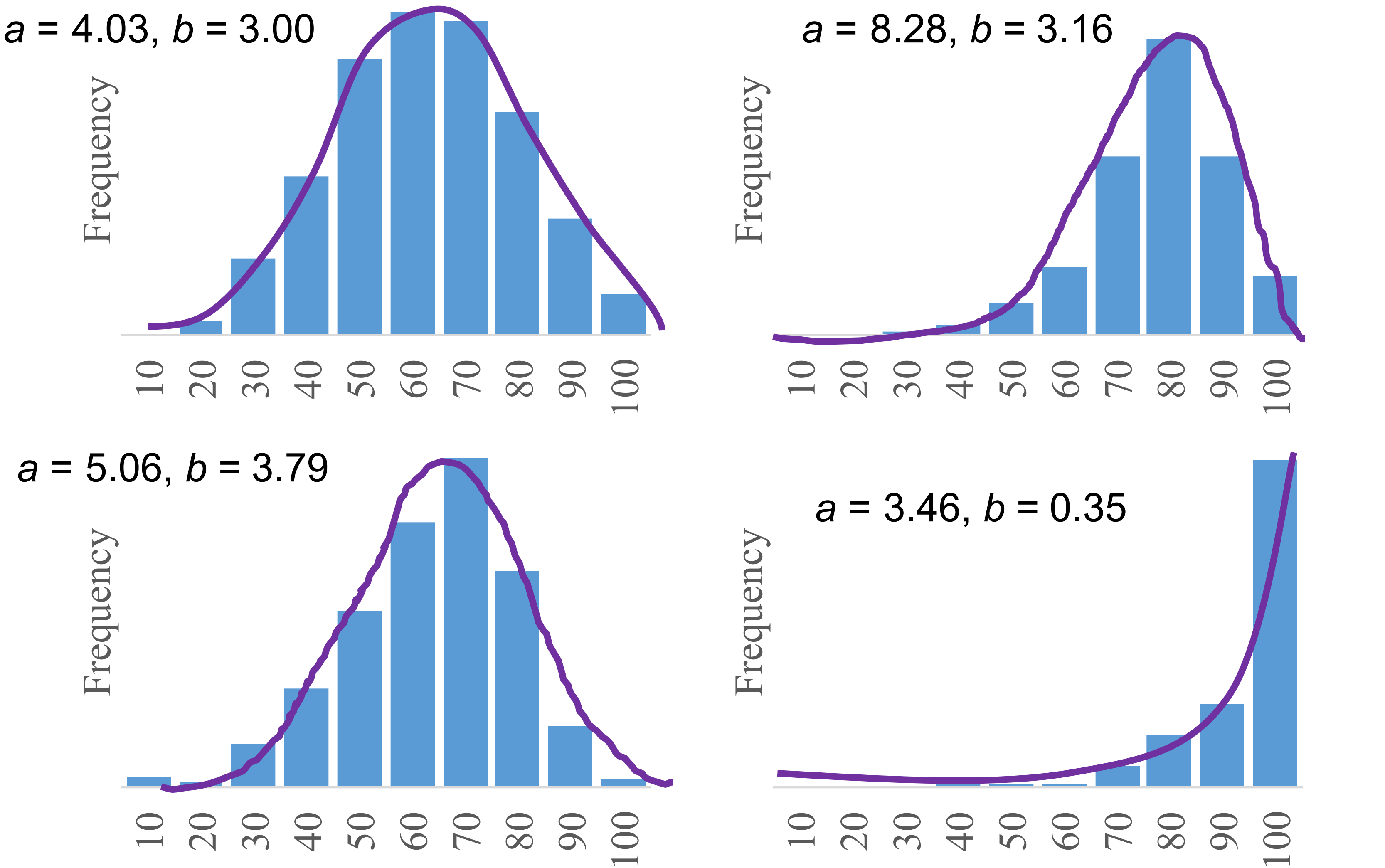
Assignment Grades Demo

`Assignment id = '1613'`



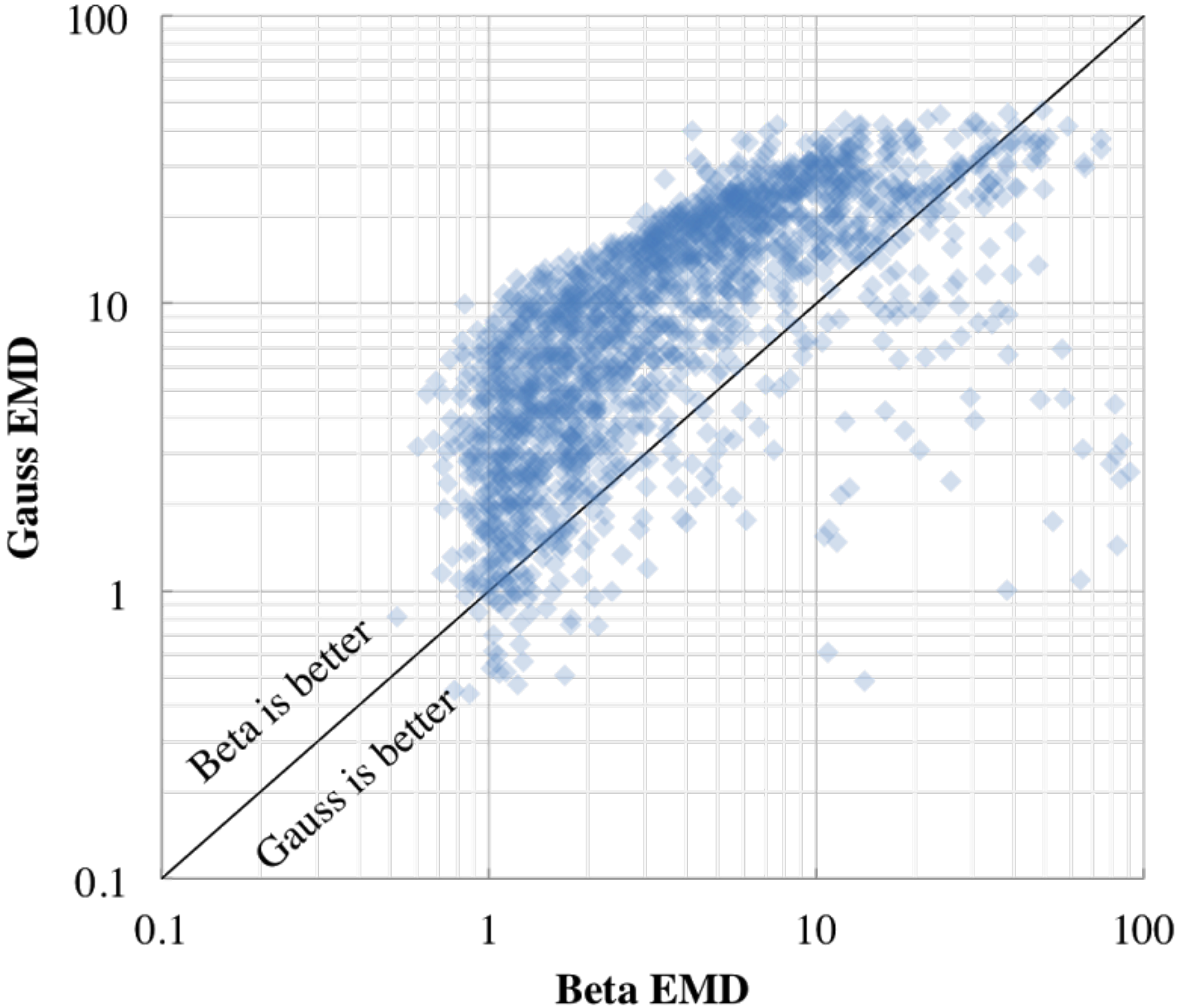
$$X \sim Beta(a = 8.28, b = 3.16)$$

Assignment Grades



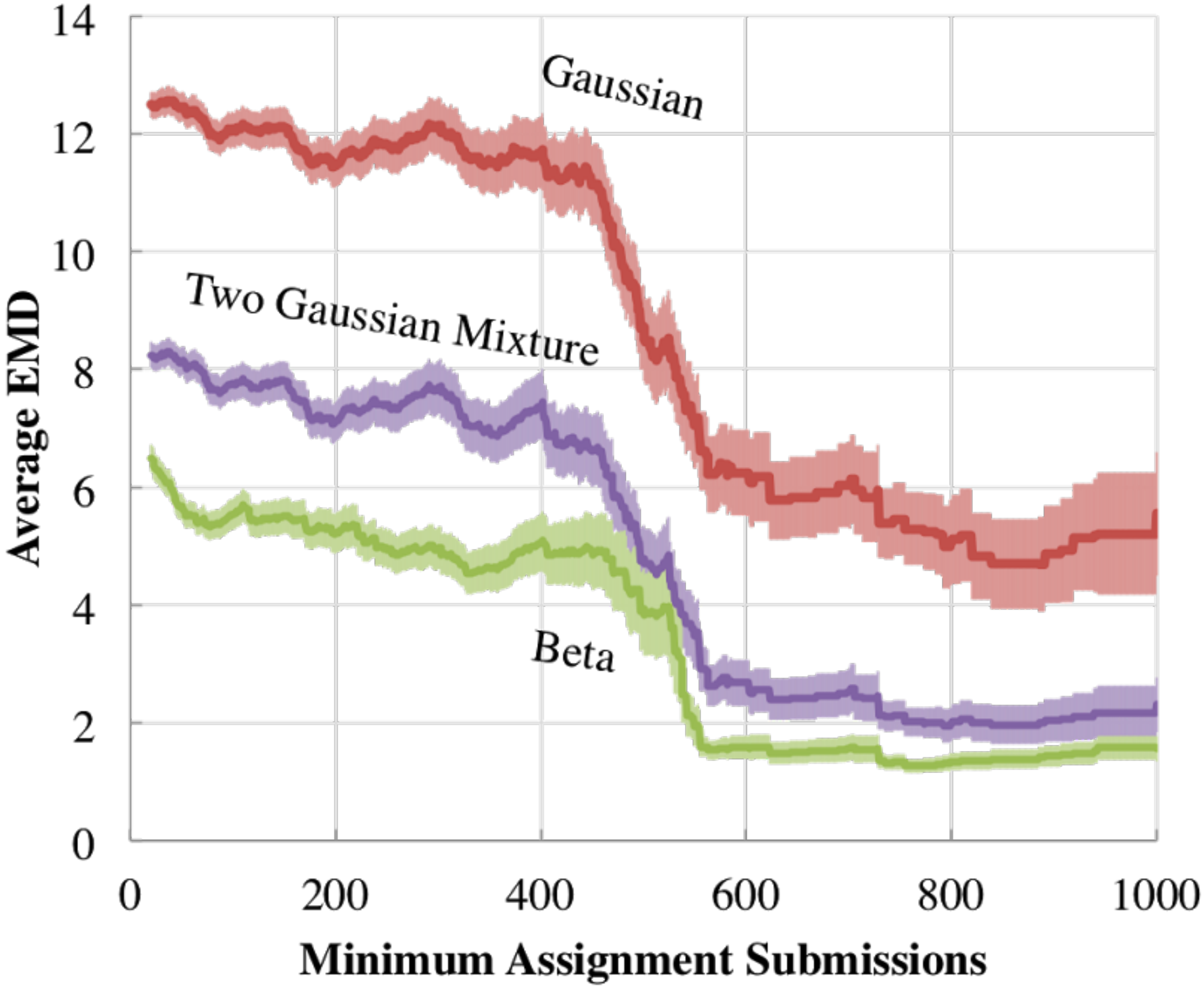
We have 2055 assignment distributions from grade scope

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

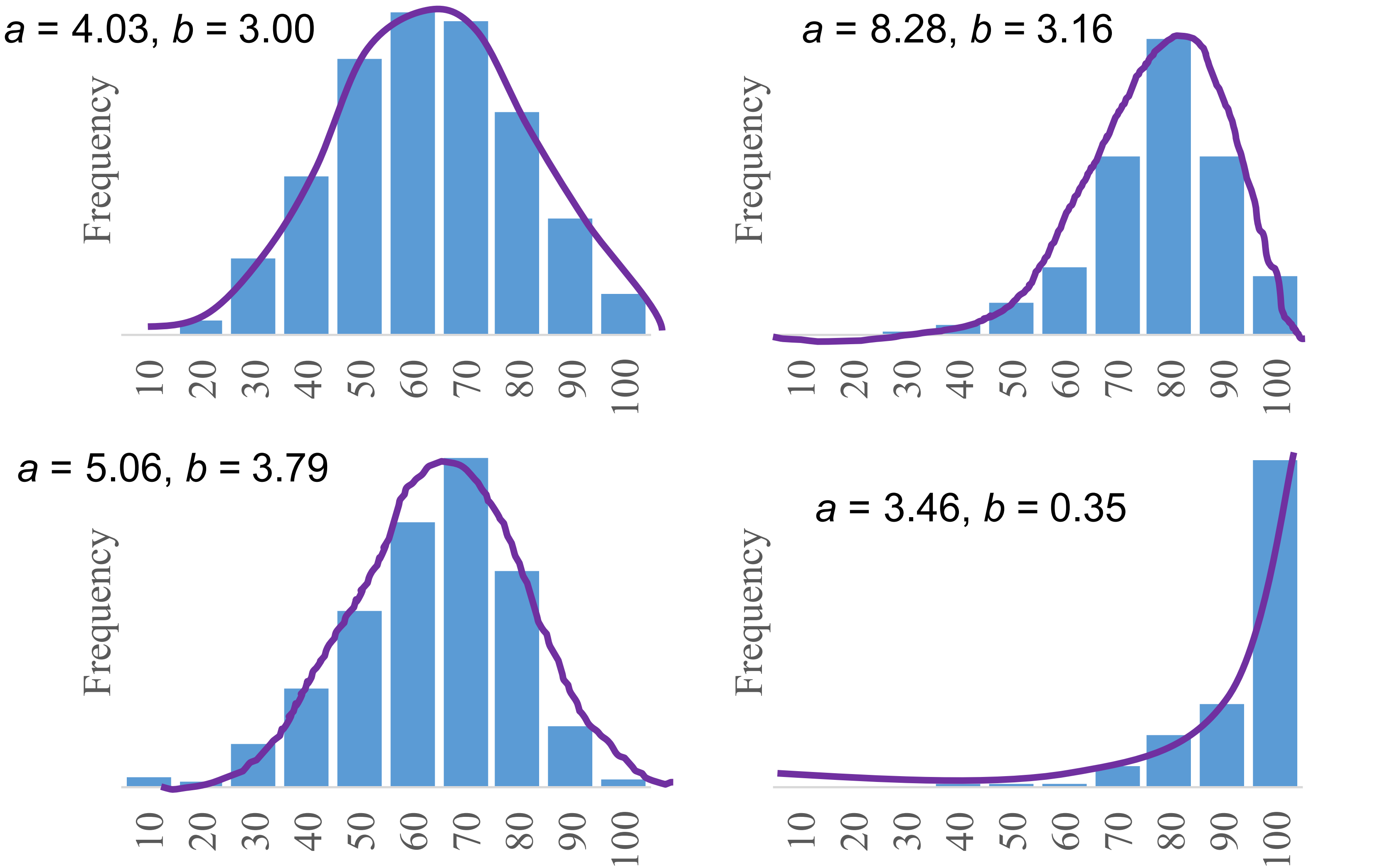
> What the Binomial said, but approximated.
>
> \- Normal

Beta Interpretation

> Each student's ability is represented as a probability. The distribution of probabilities is a Beta distribution. Each student has **a different** probability of getting points, and that probability is sampled from a Beta distribution.
>
> \- Beta

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

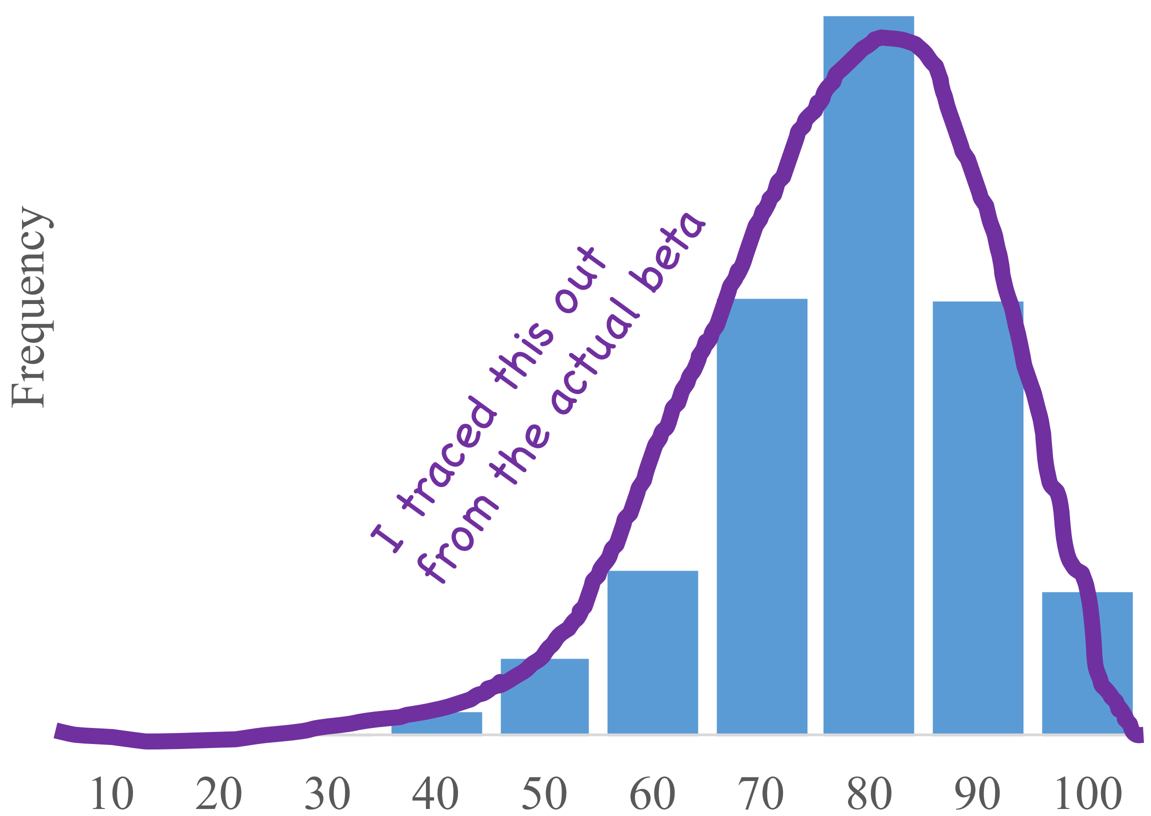
Assignment Grades



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, alpha, beta)
```

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

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*.
- Suggests a way to calculate final grades as a probabilistic most likely estimate of “ability”.
- 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

$$E[\mathrm{CS109}]$$

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