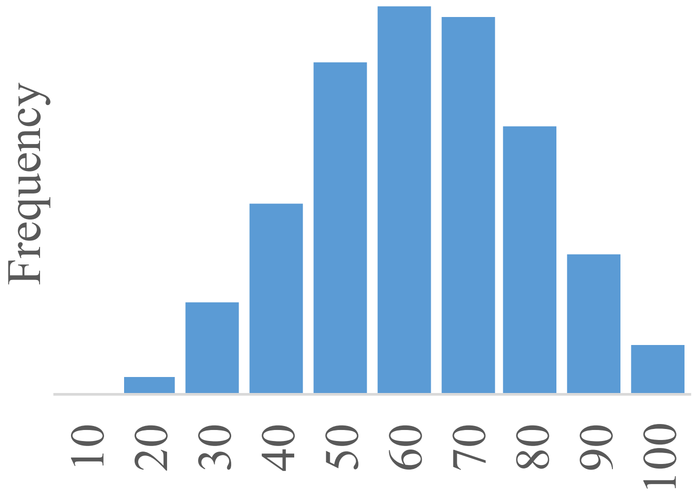
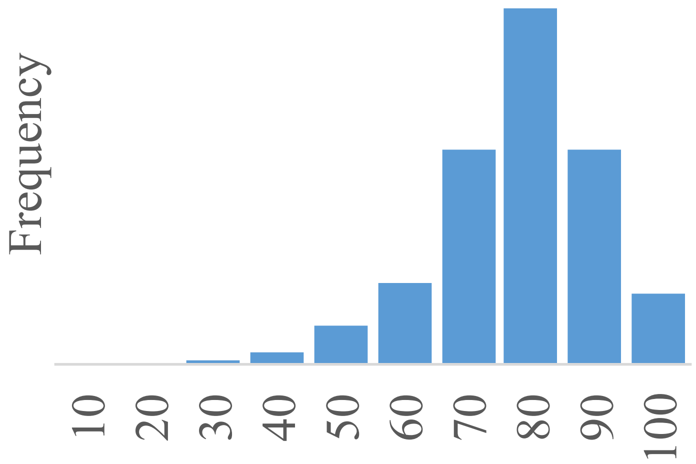
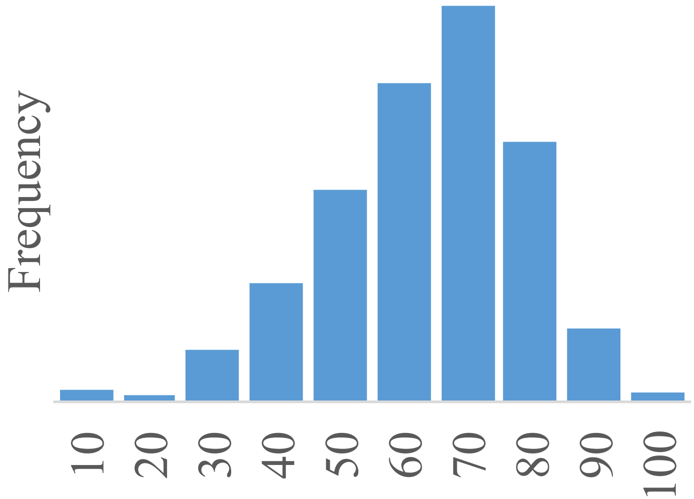
Meta Beta and the story of our prior beliefs

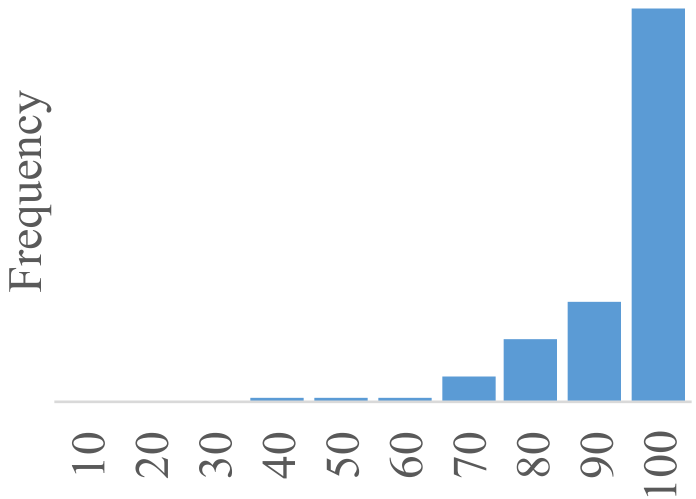
CS 109
Lecture 14
April 27th, 2016

Assignment Grades









We have 2055 assignment distributions from grade scope

Review

Two parts to last class:
Convolution and Conditionals with Random Variables

Convolution of Probability Distributions

We talked about sum of Binomial, Normal and Poisson…who's missing from this party?

Uniform.

Dance, Dance Convolution

- Let X and Y be independent random variables
  - Cumulative Distribution Function (CDF) of X + Y:

$$F_{X+Y}(a) = P(X+Y \le a)$$

$$= \iint_{x+y\le a} f_X(x) f_Y(y)\, dx\, dy = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{a-y} f_X(x)\, dx\, f_Y(y)\, dy$$

CDF of X

$$= \int_{y=-\infty}^{\infty} F_X(a-y)\, f_Y(y)\, dy$$

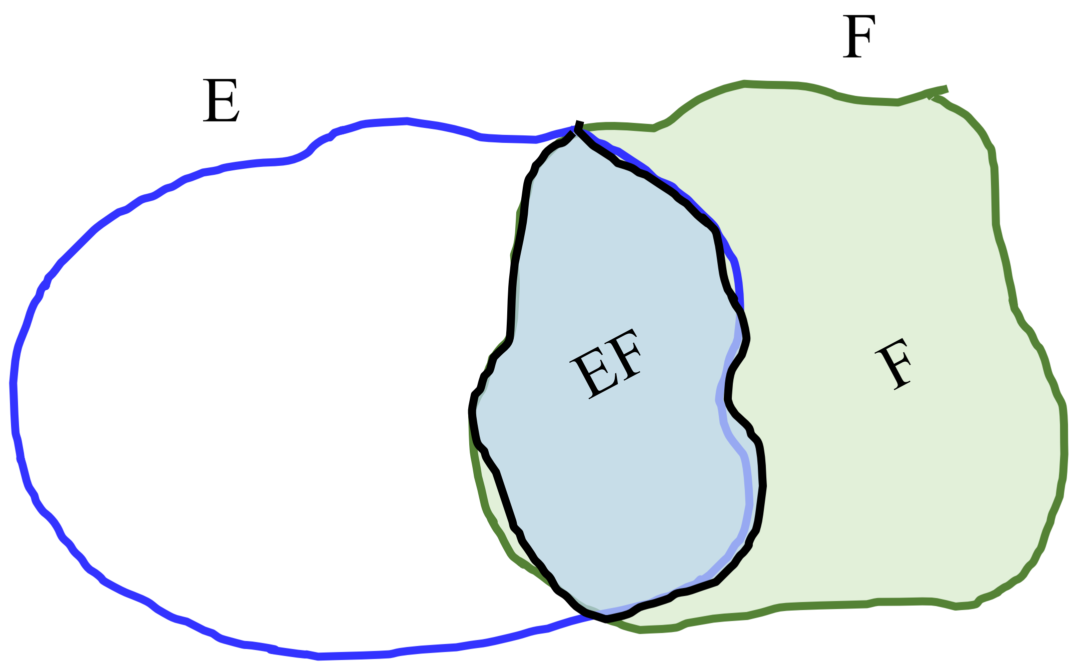
PDF of Y

  - F_{X+Y} is called **convolution** of F_X and F_Y

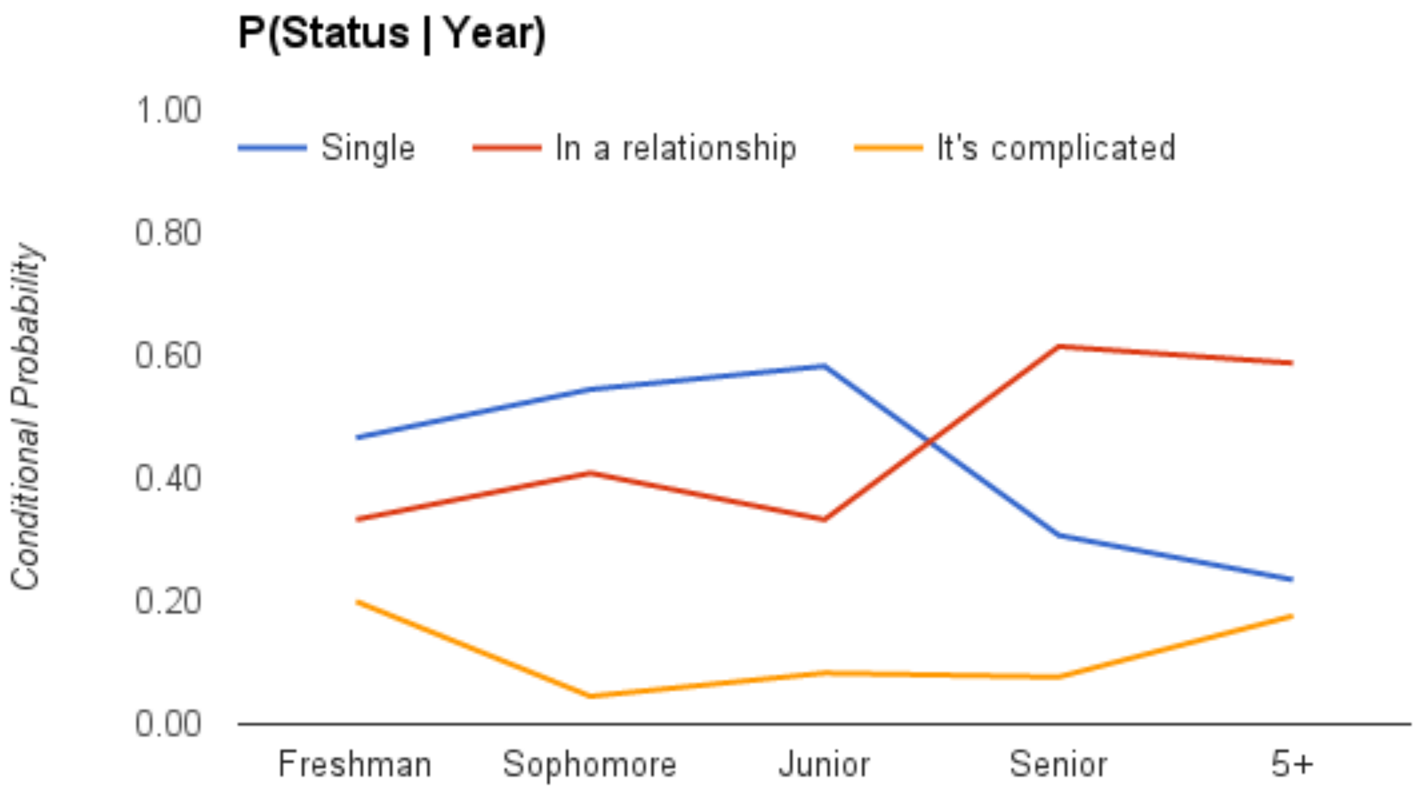
Discrete Conditional Distribution

- Recall that for *events* E and F:

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



Relationship Status



Web Server Requests Redux

- Requests received at web server in a day
  - X = # requests from humans/day $X \sim \text{Poi}(\lambda_1)$
  - Y = # requests from bots/day $Y \sim \text{Poi}(\lambda_2)$
  - X and Y are independent → $X + Y \sim \text{Poi}(\lambda_1 + \lambda_2)$
  - What is P(X = *k* | X + Y = *n*)?

$$P(X = k \mid X + Y = n) = \frac{P(X = k, Y = n - k)}{P(X + Y = n)} = \frac{P(X = k)P(Y = n - k)}{P(X + Y = n)}$$

$$= \frac{e^{-\lambda_1}\lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!} \cdot \frac{n!}{e^{-(\lambda_1+\lambda_2)}(\lambda_1 + \lambda_2)^n} = \frac{n!}{k!(n-k)!} \cdot \frac{\lambda_1^k \lambda_2^{n-k}}{(\lambda_1 + \lambda_2)^n}$$

$$= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n-k}$$

$$(X|X + Y = n) \sim Bin\left(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)$$

End Review

Let’s talk a little more about conditional probabilities with RVs

Continuous Conditional Distributions

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

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

Let's Do an Example

- X and Y are continuous RVs with PDF:

$$f(x,y) = \begin{cases} \frac{12}{5}x(2-x-y) & \text{where } 0 < x,y < 1 \\ 0 & \text{otherwise} \end{cases}$$

  - Compute conditional density: $f_{X|Y}(x \mid y)$

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{\int_0^1 f_{X,Y}(x,y)\,dx}$$

$$= \frac{\frac{12}{5}x(2-x-y)}{\int_0^1 \frac{12}{5}x(2-x-y)\,dx} = \frac{x(2-x-y)}{\int_0^1 x(2-x-y)\,dx} = \frac{x(2-x-y)}{\left[x^2 - \frac{x^3}{3} - \frac{x^2 y}{2}\right]_0^1}$$

$$= \frac{x(2-x-y)}{\frac{2}{3}-\frac{y}{2}} = \frac{6x(2-x-y)}{4-3y}$$

Independence and Conditioning

- If X and Y are independent discrete RVs:

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

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} = \frac{p_X(x)p_Y(y)}{p_Y(y)} = p_X(x)$$

Sanity check: what do these different notations mean?

- Analogously, for independent continuous RVs:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$$

Mixing Discrete and Continuous

- Let X be a continuous random variable
- Let N be a discrete 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:

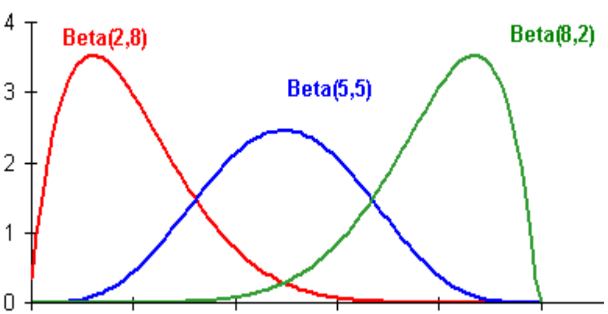
$$f_{X|N}(x \mid n) = f_X(x) \qquad p_{N|X}(n \mid x) = p_N(n)$$

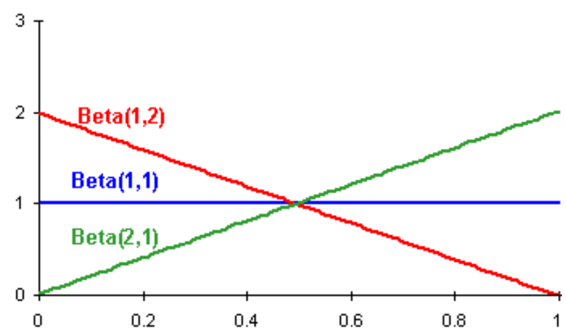
We will use that shortly

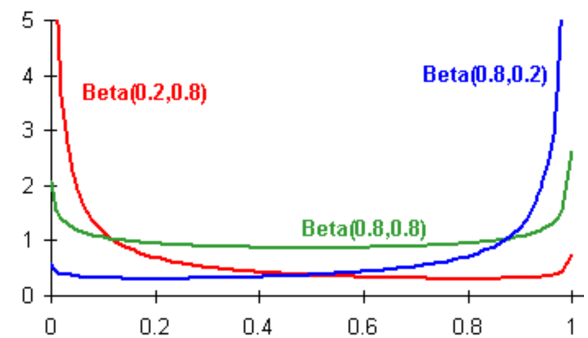
Beta Random Variable

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

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

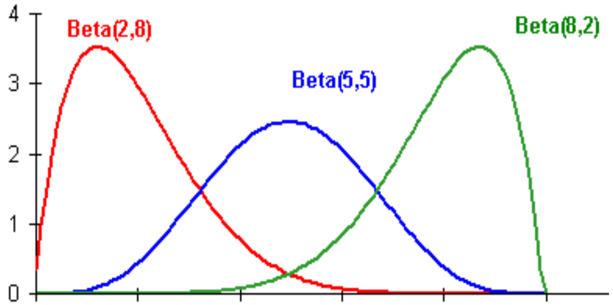


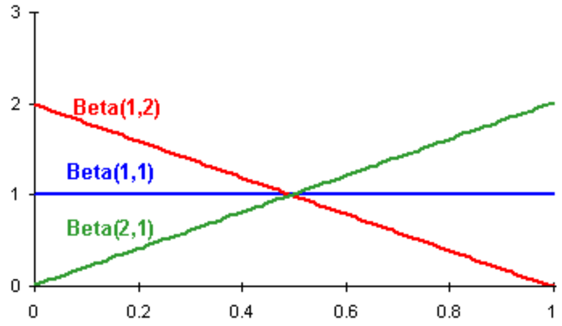


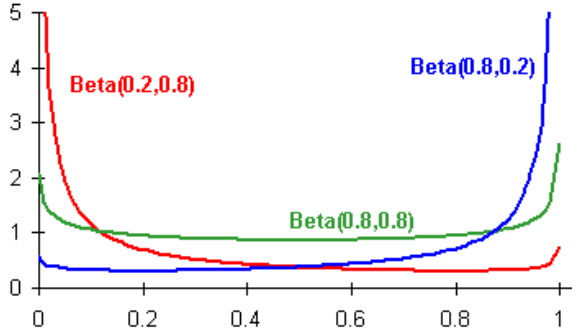


  - Symmetric when *a* = *b*
  - $E[X] = \frac{a}{a+b}$ $\qquad Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$

Meta Beta







Used to represent a
distributed belief of a probability

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)

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

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

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

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

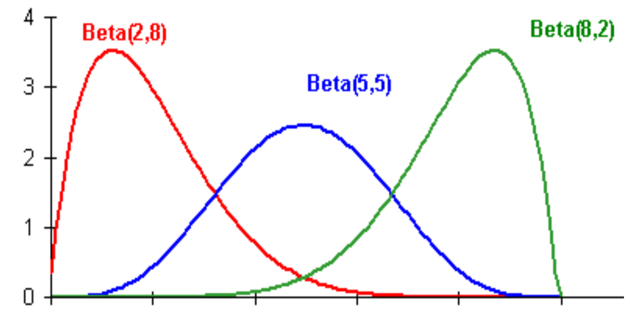
- X | (N = n, m + n trials) ~ Beta(n + 1, m + 1)
  - X ~ Uni(0, 1)
  - Check this out, boss:
    - Beta(1, 1) = Uni(0, 1)
  - So, X ~ Beta(1, 1)
  - "Prior" distribution of X (before seeing any flips) is Beta
  - "Posterior" distribution of X (after seeing flips) is Beta

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

- Beta is a **<u>conjugate</u>** distribution for Beta
  - Prior and posterior parametric forms are the same!
  - Beta is also conjugate for Bernoulli and Binomial
  - 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

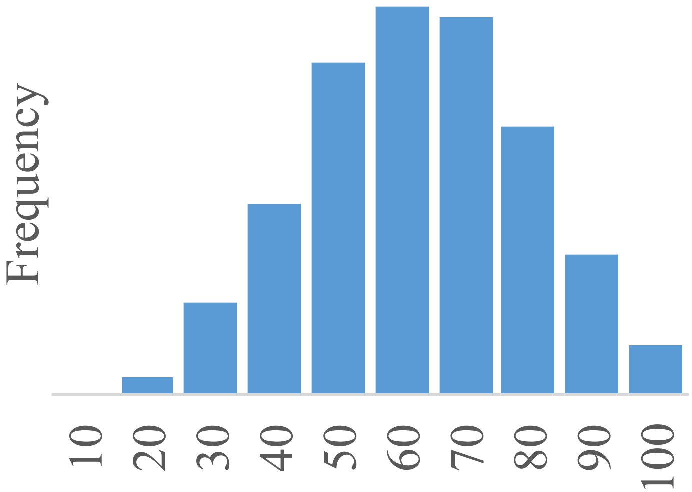


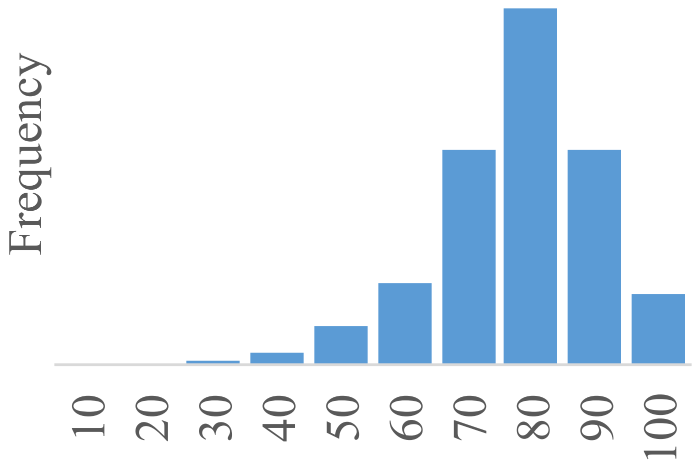
- Update to get posterior probability
  - X | (n heads in n + m trials) ~ Beta(a + n, b + m)
  - Sometimes call *a* and *b* the "equivalent sample size"
  - Prior probability for X based on seeing (*a* + *b* – 2) "imaginary" trials, where (*a* – 1) of them were heads.
  - Beta(1, 1) ~ Uni(0, 1) → we haven't seen any "imaginary trials", so apriori know nothing about coin

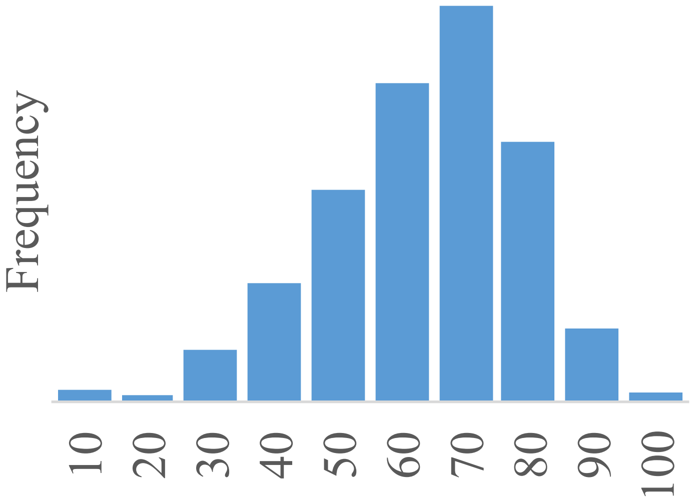
Flip a Coin With Unknown Probability

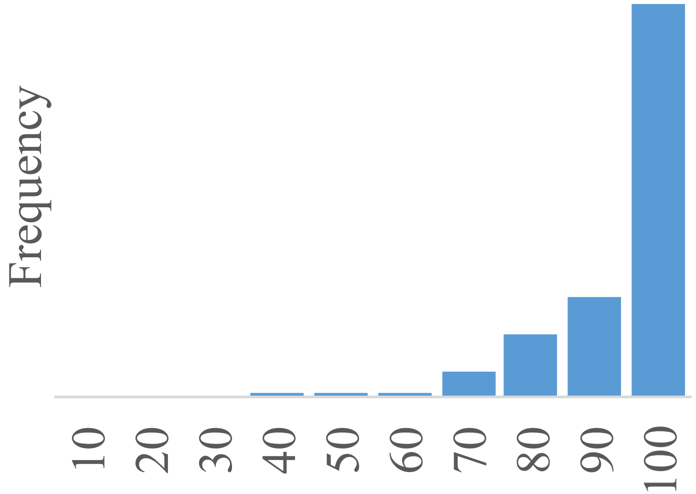
Demo

Assignment Grades



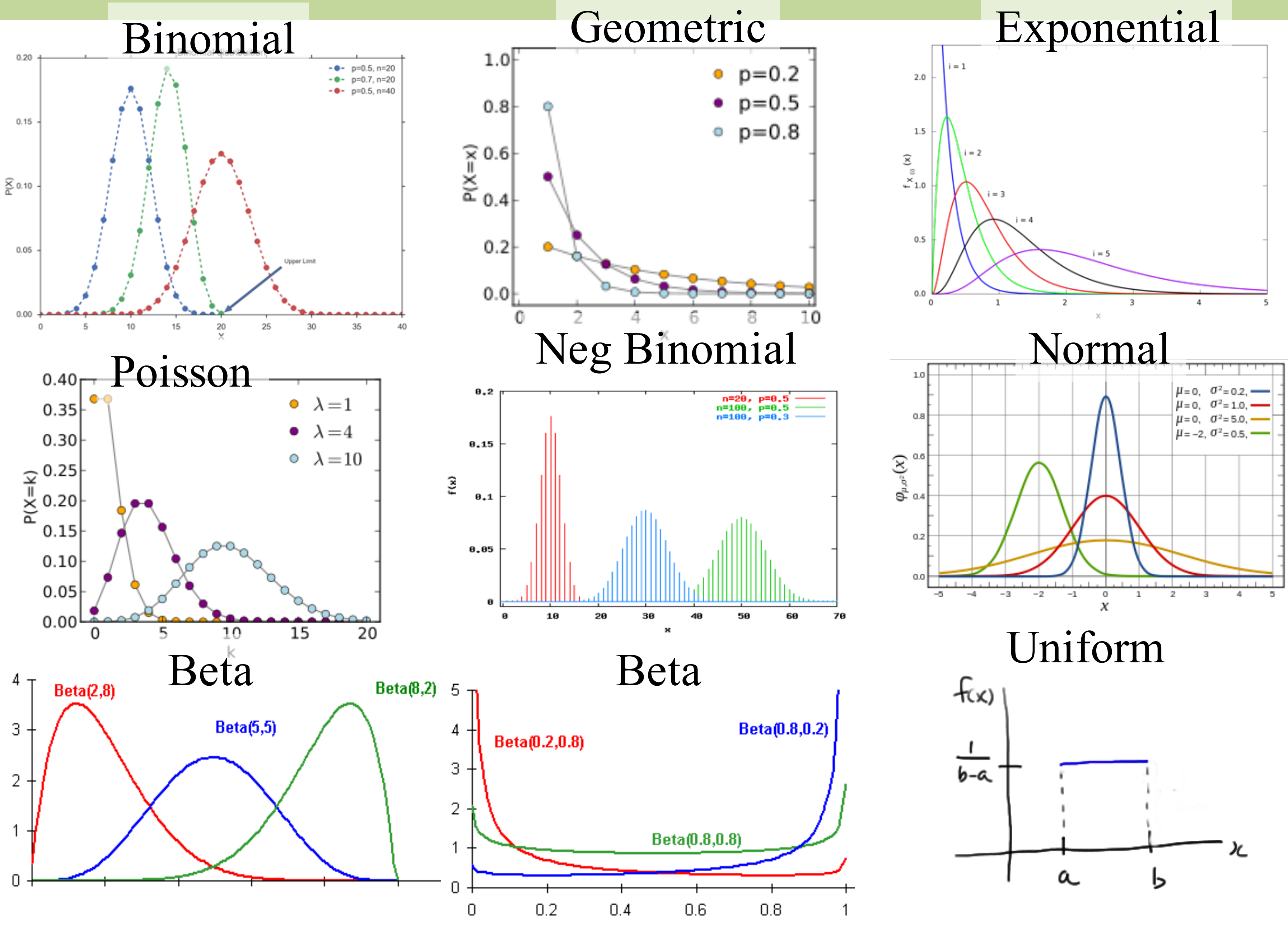






We have 2055 assignment distributions from gradescope

Distributions



Grades must be bounded

Normal: No

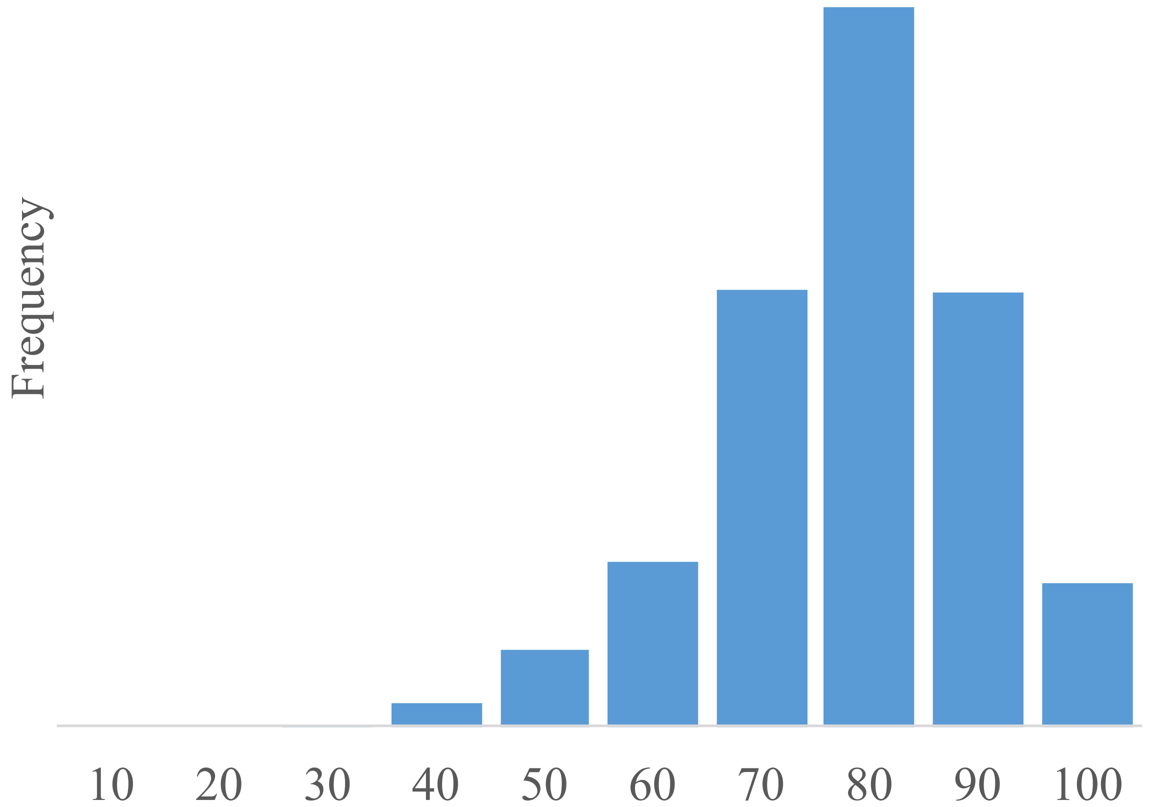
Poisson: No

Exponential: No

Beta: Yes

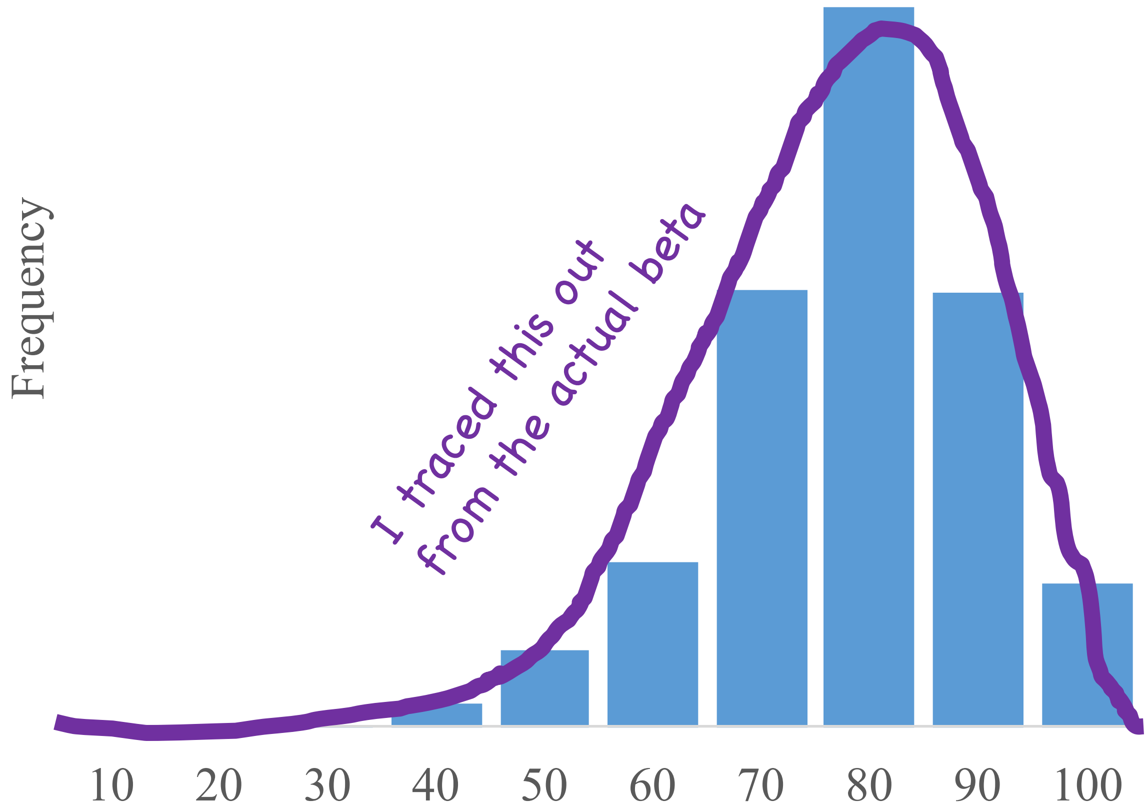
Assignment Grades Demo

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



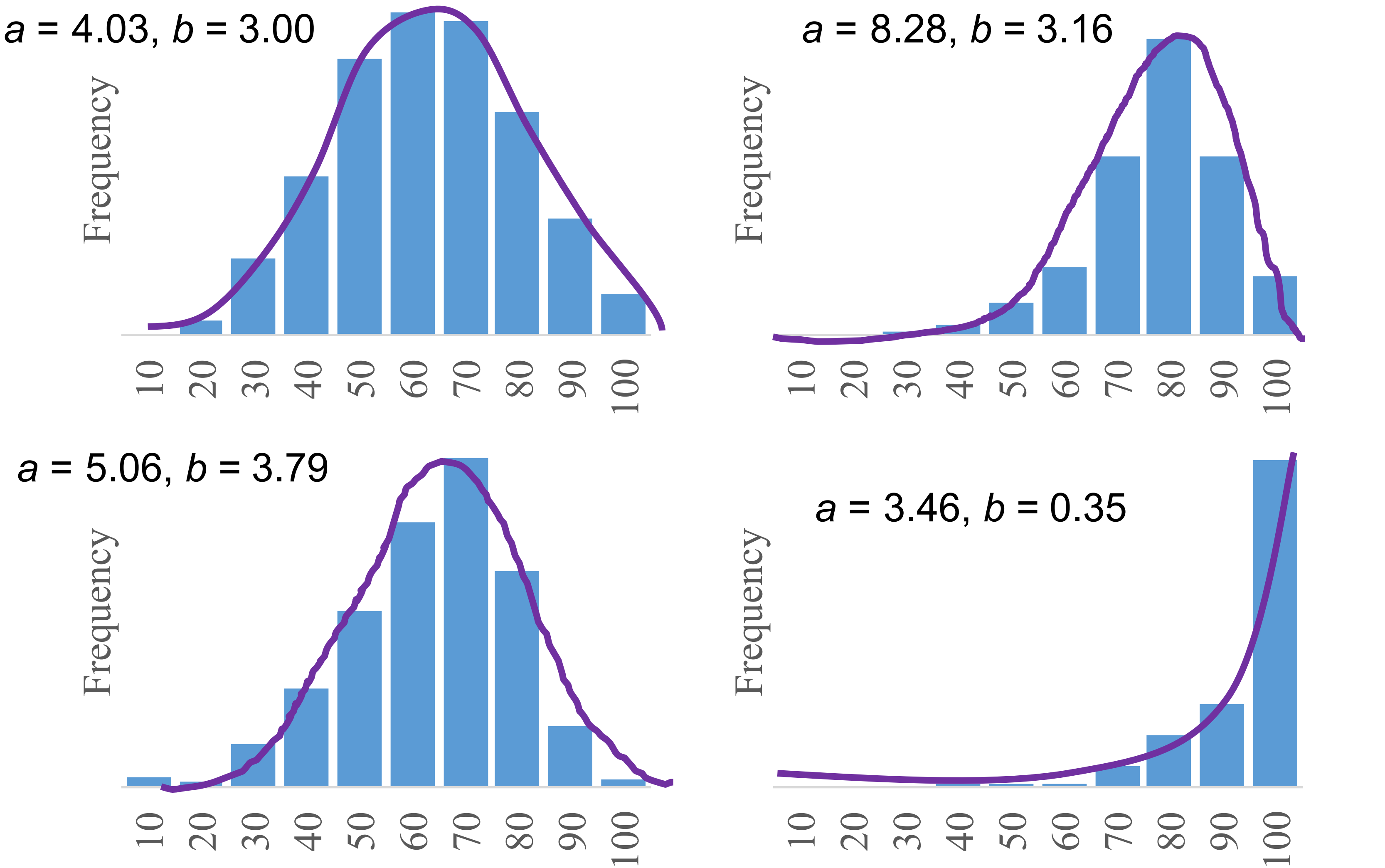
Assignment Grades Demo

`Assignment id = '1613'`



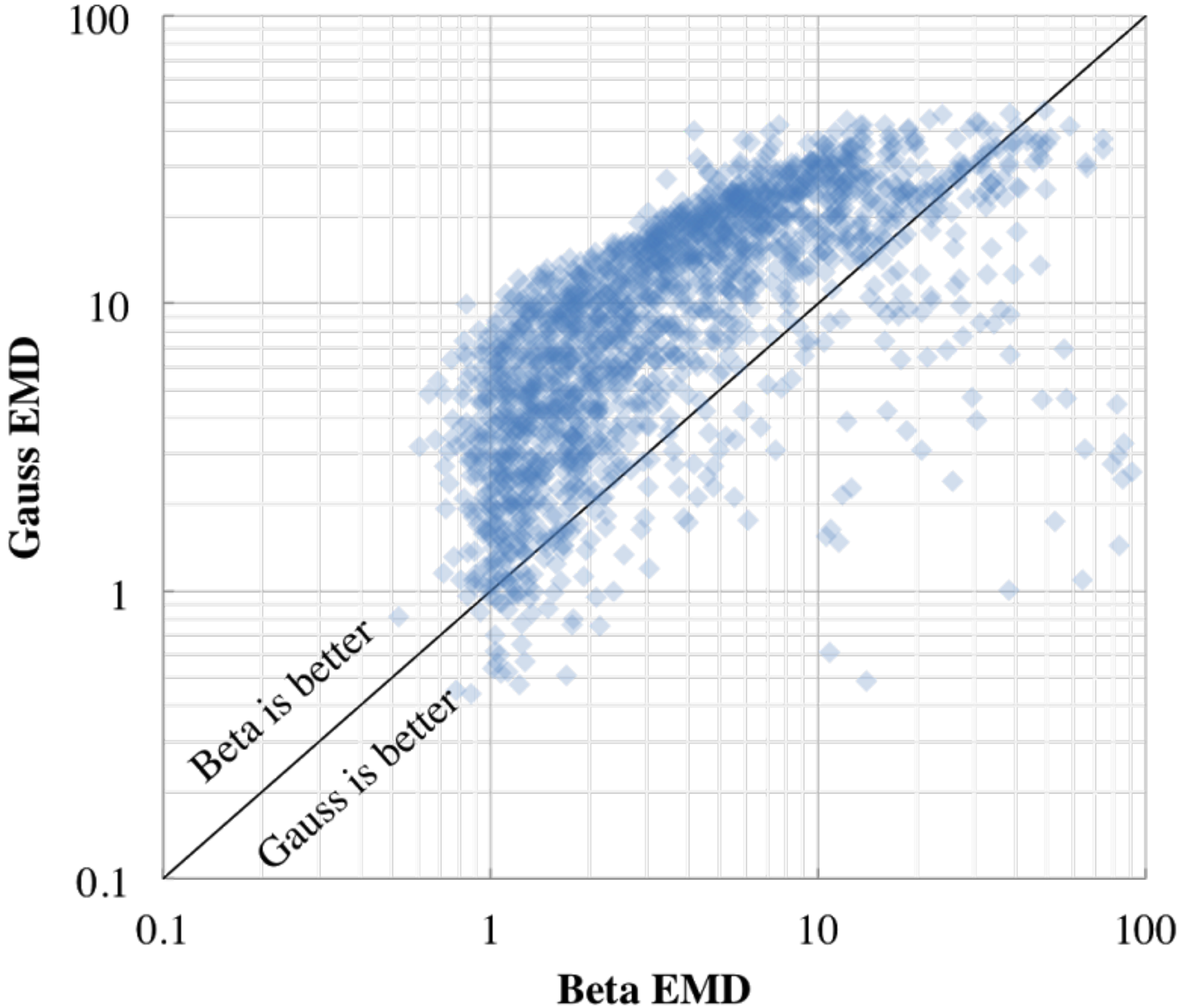
$$X \sim Beta(a = 8.28, b = 3.16)$$

Assignment Grades



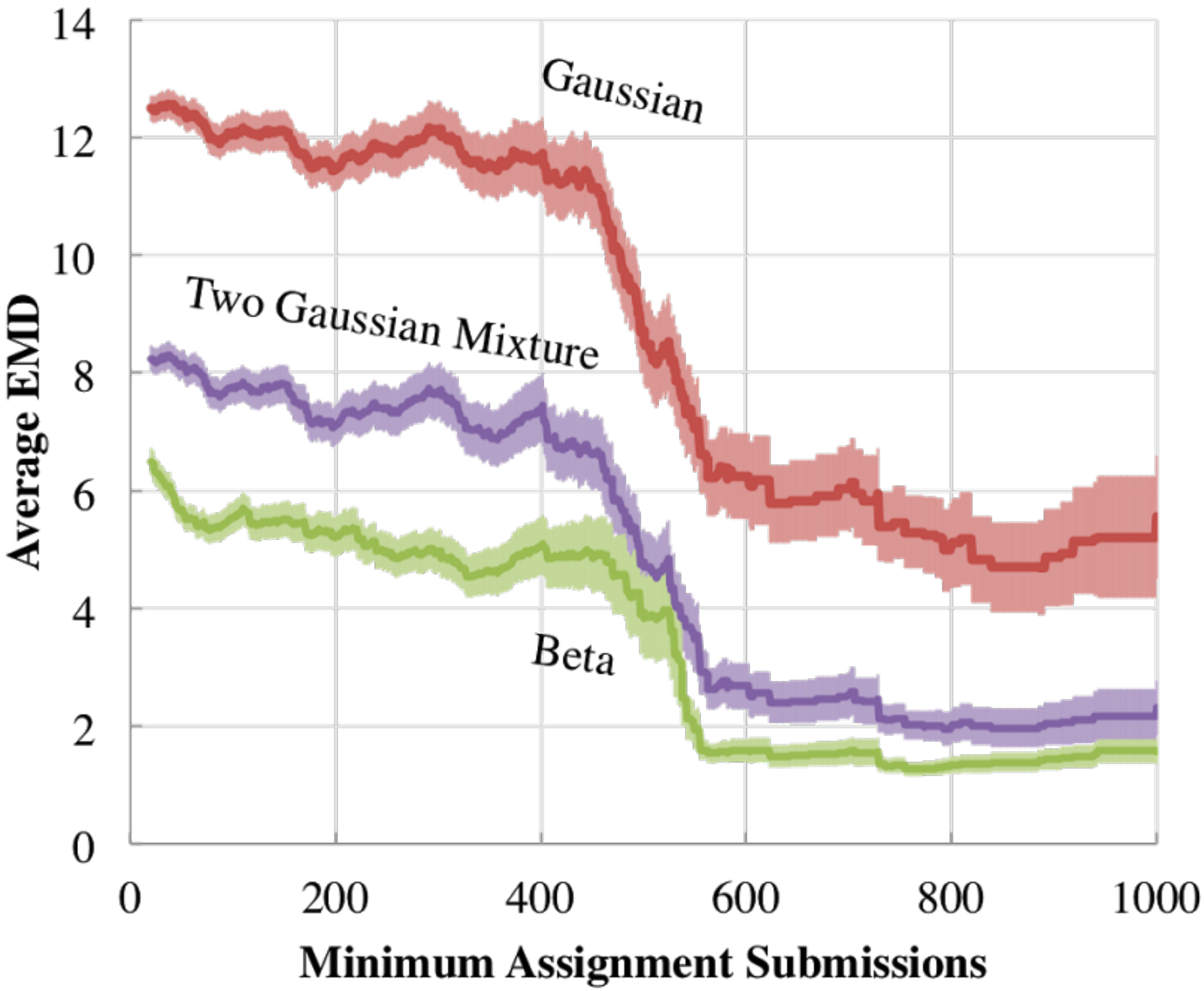
We have 2055 assignment distributions from grade scope

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

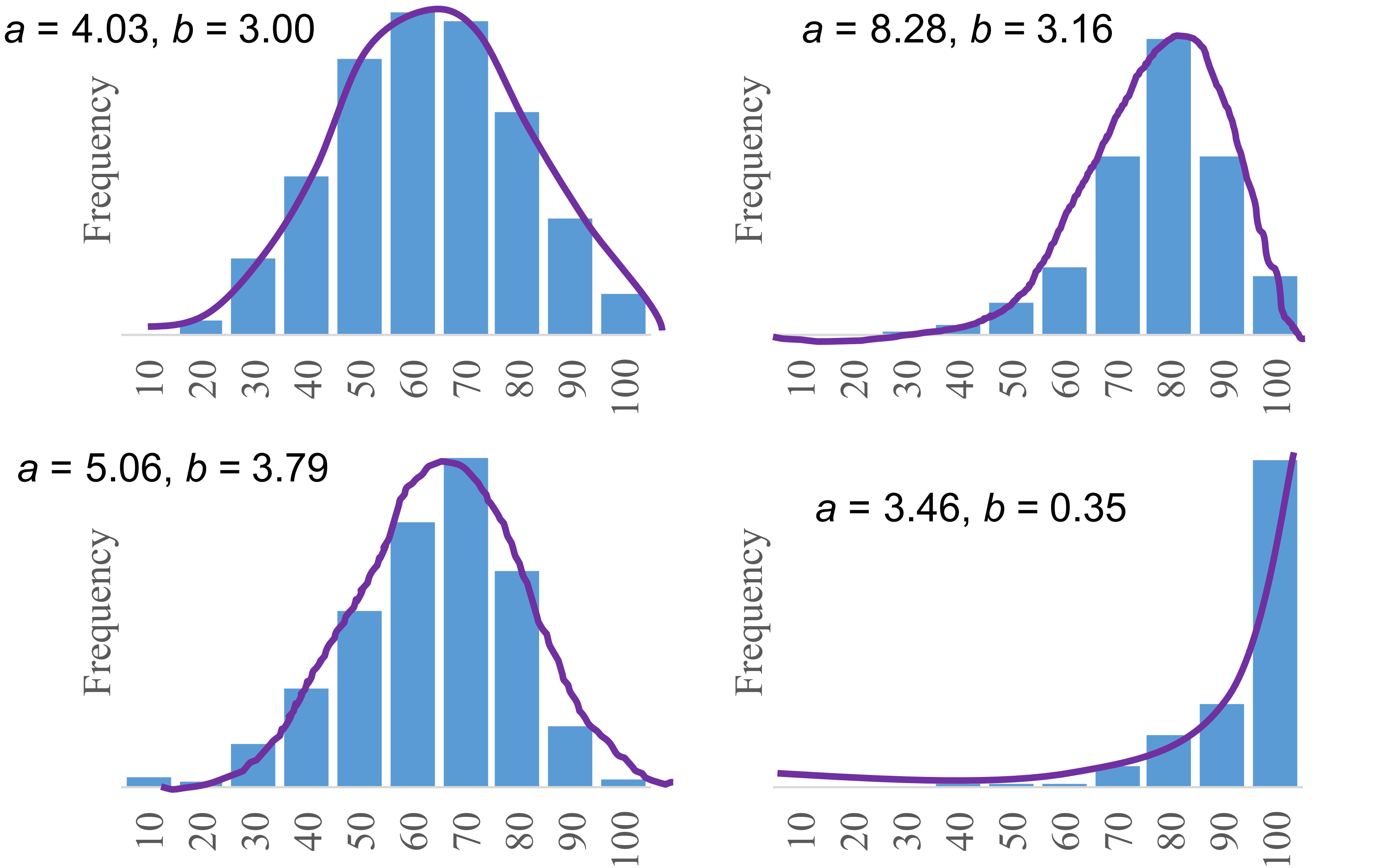
> The sum of equally weighted independent random variables will produce a normal. Each point is an **equally weighted**, **independent** random variable.

- Normal

Beta Interpretation

> Each student has **a different** probability of getting points. A student's grade is a sum of coin flips based on their personal probability. The distribution of individual probabilities is a Beta distribution.
>
> \- Beta

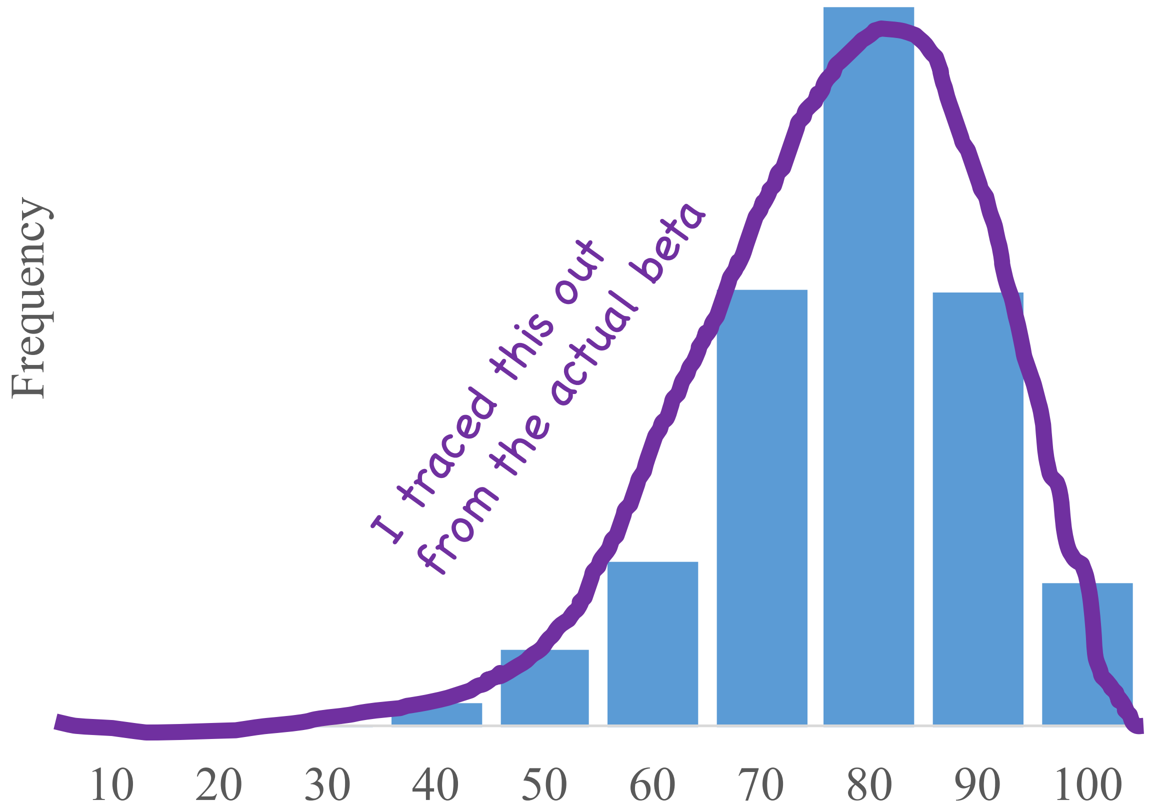
Assignment Grades



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?

```
jStat.beta.cdf(x, alpha, beta)
```

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

Implications

- Will be combined with Item Response Theory which models how assignment difficulty and student ability combine to give *point probabilities*.
- 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 in CS109?

No

Beta:
The probability density for probabilities

Recall:
Expectation with Multiple Variables

Joint Expectation

$$E[X] = \sum_x x p(x)$$

- Expectation over a joint isn't nicely defined because it is not clear how to compose the multiple variables:
  - Add them? Multiply them?

- Lemma: For a function g(X,Y) we can calculate the expectation of that function:

$$E[g(X,Y)] = \sum_{x,y} g(x,y) p(x,y)$$

- By the way, this also holds for single random variables:

$$E[g(X)] = \sum_x g(x) p(x)$$

Expected Values of Sums

Big deal lemma: first stated without proof

E[X + Y] = E[X] + E[Y]

Generalized: $E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$

Holds regardless of dependency between X_i's

Skeptical Chris Wants a Proof!

Let g(X,Y) = [X + Y]

$$E[X+Y] = E[g(X,Y)] = \sum_{x,y} g(x,y)p(x,y)$$

What a useful lemma

$$= \sum_{x,y} [x+y]p(x,y)$$

By the definition of g(x,y)

Break that sum into parts!

$$= \sum_{x,y} xp(x,y) + \sum_{x,y} yp(x,y)$$

Change the sum of (x,y) into separate sums

$$= \sum_{x} x \sum_{y} p(x,y) + \sum_{y} y \sum_{x} p(x,y)$$

That is the definition of marginal probability

$$= \sum_{x} xp(x) + \sum_{y} yp(y)$$

That is the definition of expectation

$$= E[X] + E[Y]$$

Hash Tables (aka Toy Collecting)

- Consider a hash table with *n* buckets
  - Each string equally likely to get hashed into any bucket
  - Let X = # strings to hash until each bucket ≥ 1 string
  - What is E[X]?
  - Let X_i = # of trials to get success after *i*-th success
    - where "success" is hashing string to previously empty bucket
    - After *i* buckets have ≥ 1 string, probability of hashing a string to an empty bucket is $p = (n - i) / n$
    - $P(X_i = k) = \frac{n-i}{n}\left(\frac{i}{n}\right)^{k-1}$ equivalently: $X_i \sim \text{Geo}((n - i) / n)$
    - $E[X_i] = 1 / p = n / (n - i)$
  - $X = X_0 + X_1 + \ldots + X_{n-1} \Rightarrow E[X] = E[X_0] + E[X_1] + \ldots + E[X_{n-1}]$

$$E[X] = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \ldots + \frac{n}{1} = n\left[\frac{1}{n} + \frac{1}{n-1} + \ldots + 1\right] = O(n \log n)$$

Course Mean

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

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