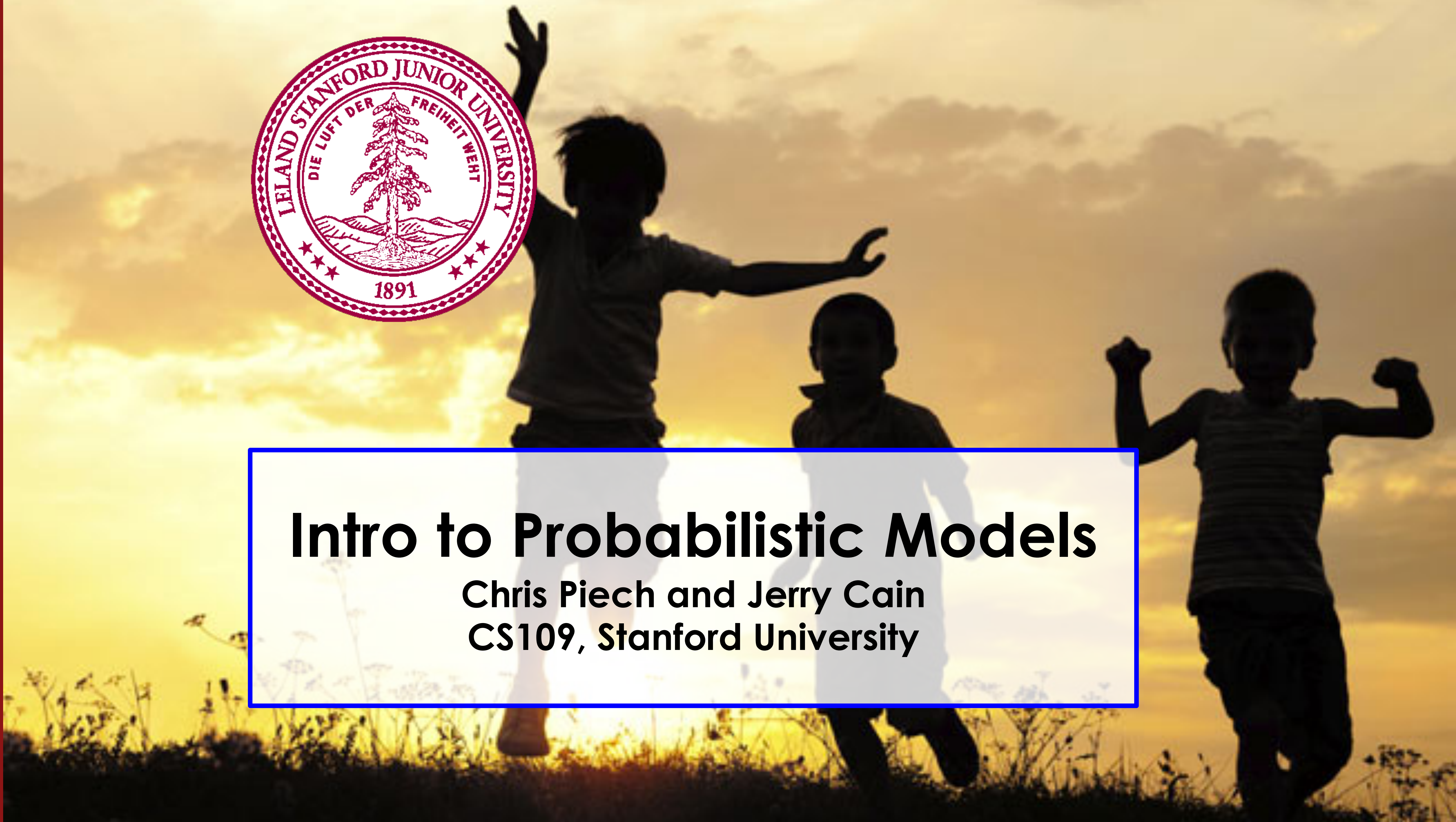
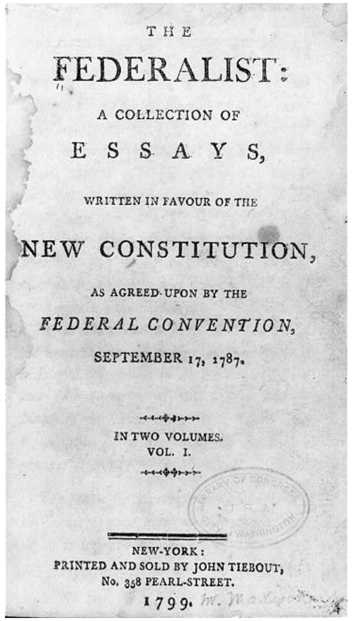


Intro to Probabilistic Models

Chris Piech and Jerry Cain
CS109, Stanford University

Terribly exciting day in CS109

Exciting Day



THE

FEDERALIST:

A COLLECTION OF

ESSAYS,

WRITTEN IN FAVOUR OF THE

NEW CONSTITUTION,

AS AGREED UPON BY THE

FEDERAL CONVENTION,

SEPTEMBER 17, 1787.

IN TWO VOLUMES.

VOL. I.

NEW-YORK:

PRINTED AND SOLD BY JOHN TIEBOUT,

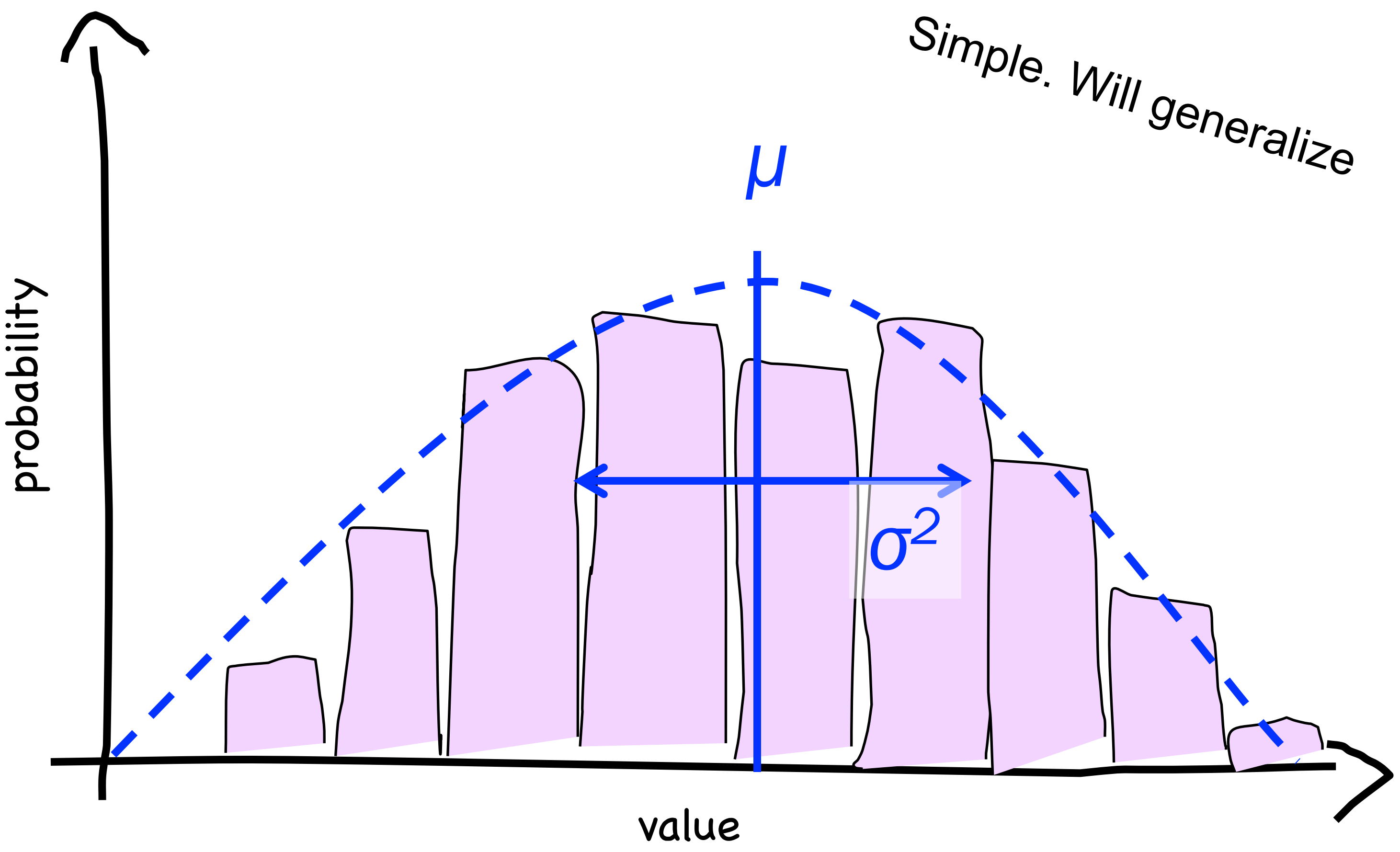
No. 358 PEARL-STREET.

1799.

Quick slide reference

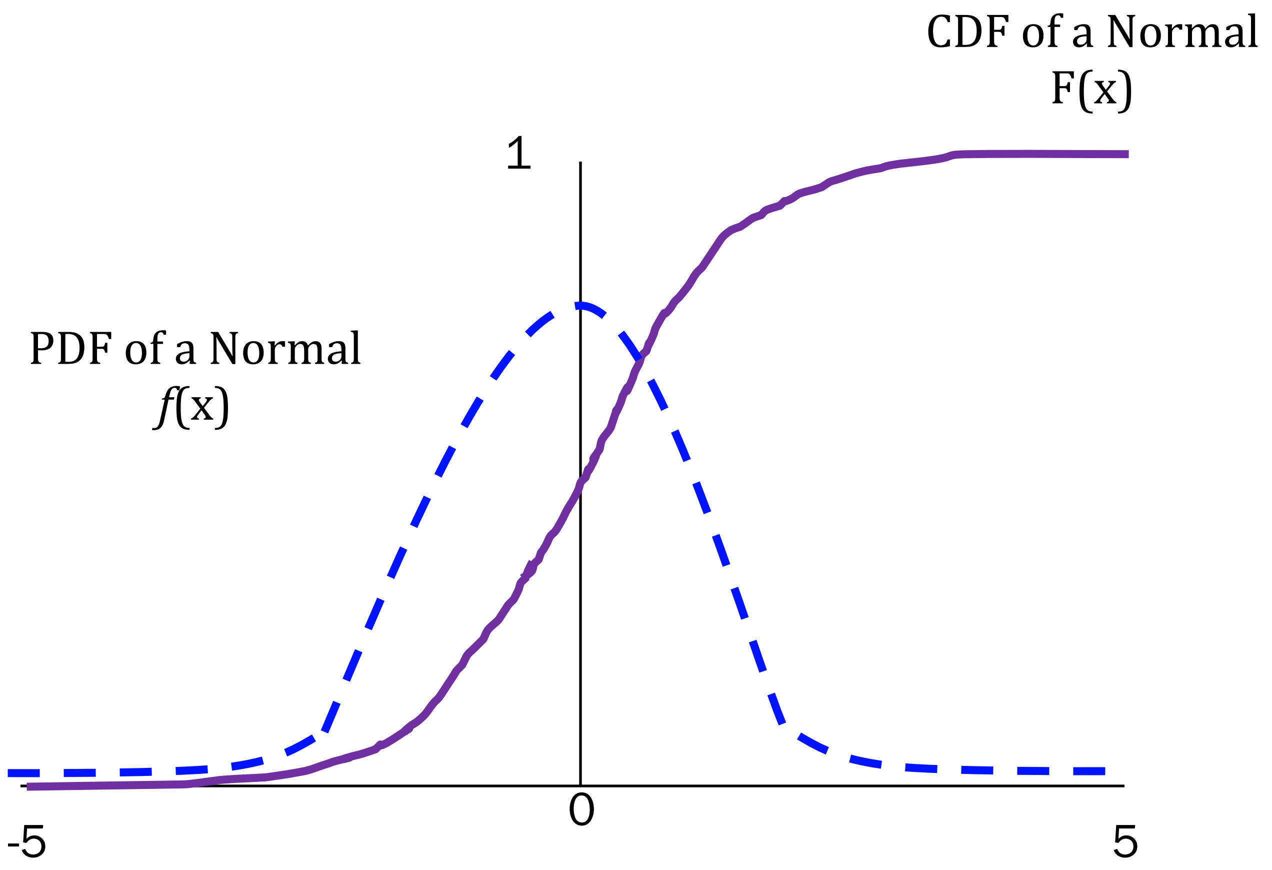
First, some review

Simplicity is Humble



* A Gaussian maximizes entropy for a given mean and variance

Density vs Cumulative



$f(x)$ = derivative of probability

$F(x) = P(X < x)$

Probability Density Function

$$\mathcal{N}(\mu, \sigma^2)$$

"exponential"

the distance to the mean

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

probability density at x

a constant

sigma shows up twice

Does it look less scary like this?

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

This means "e to the power of" and is common function in code math libraries

$$f(x) \propto \frac{1}{\sigma} \cdot \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

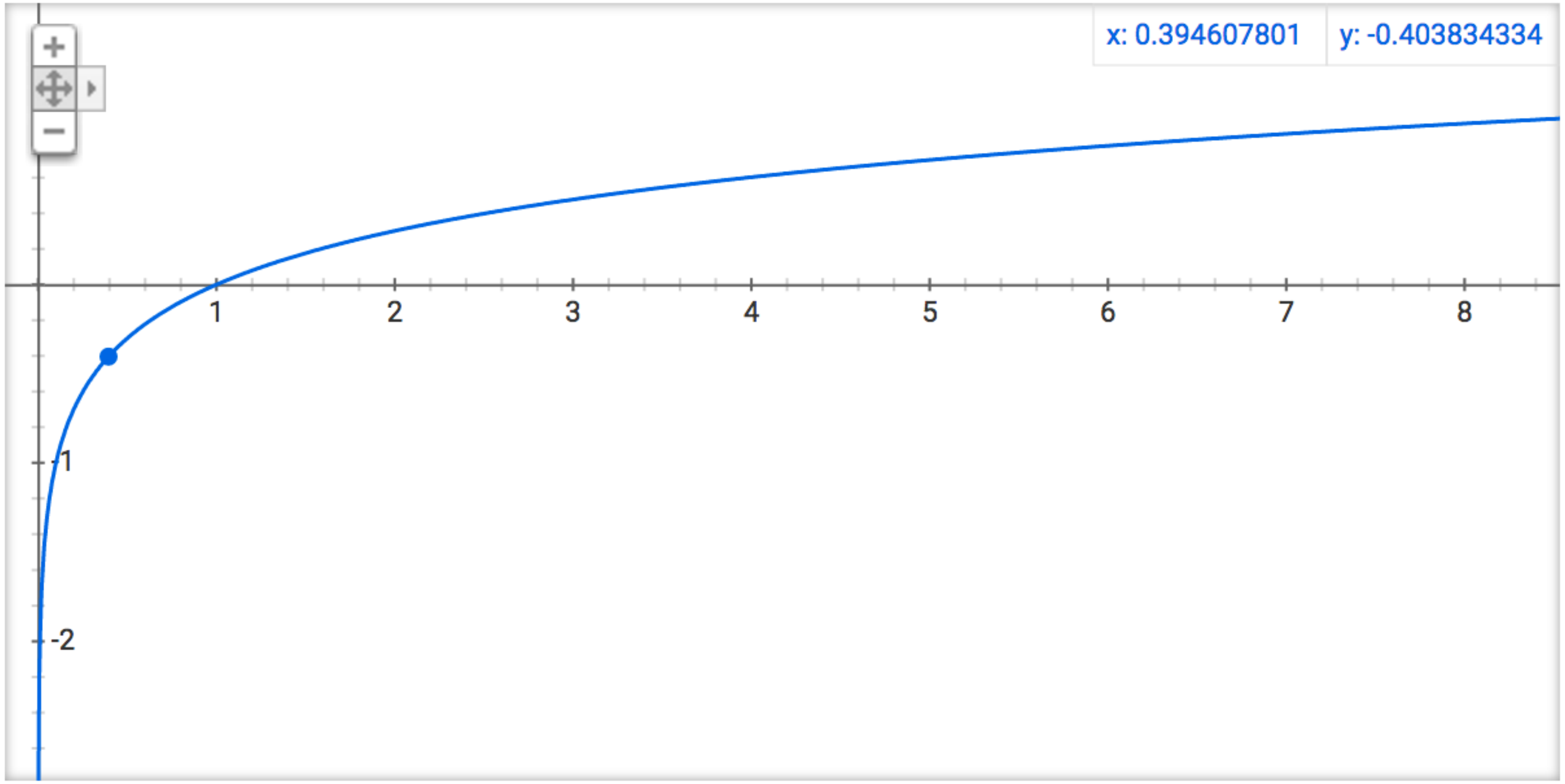
This means "proportional to". There is a constant but there are many cases where we don't care what it is!

What if you had to take the log of this function?

Log Review

$$e^y = x \qquad \log(x) = y$$

Graph for log(x)



More info

Log Identities

$$\log(a \cdot b) = \log(a) + \log(b)$$

$$\log(a/b) = \log(a) - \log(b)$$

$$\log(a^n) = n \cdot \log(a)$$

Products become sums!

$$\log(a \cdot b) = \log(a) + \log(b)$$

$$\log(\prod_i a_i) = \sum_i \log(a_i)$$

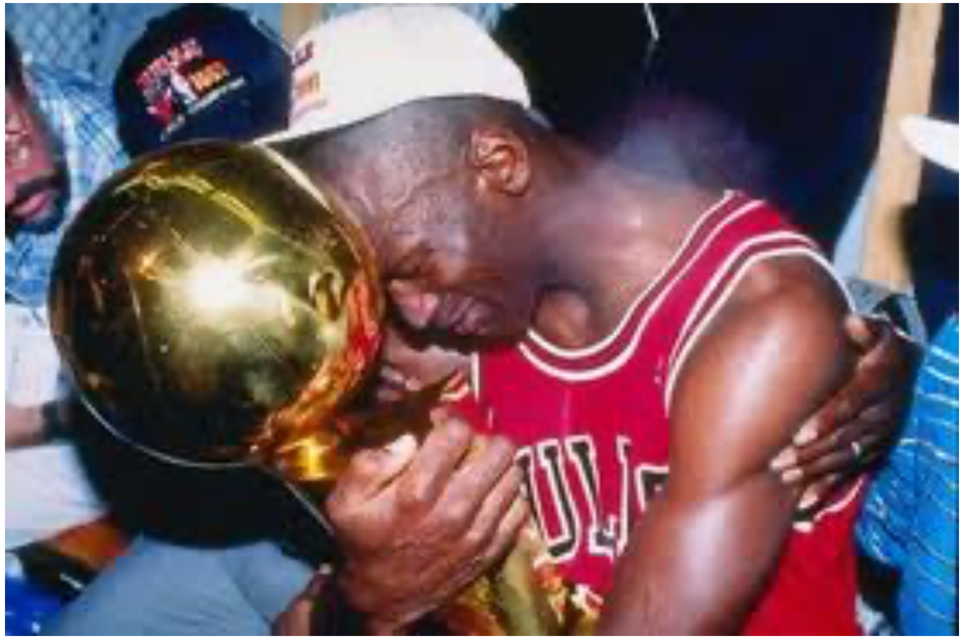
* Spoiler alert: This is important because the product of many small numbers gets hard for computers to represent.

Log for normal pdf

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

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$\log(f(x)) = -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{(x-\mu)^2}{2\sigma^2}$$



(happy tears)

Cumulative Density Function

$$\mathcal{N}(\mu, \sigma^2)$$

CDF of Standard Normal: A function that has been solved for numerically

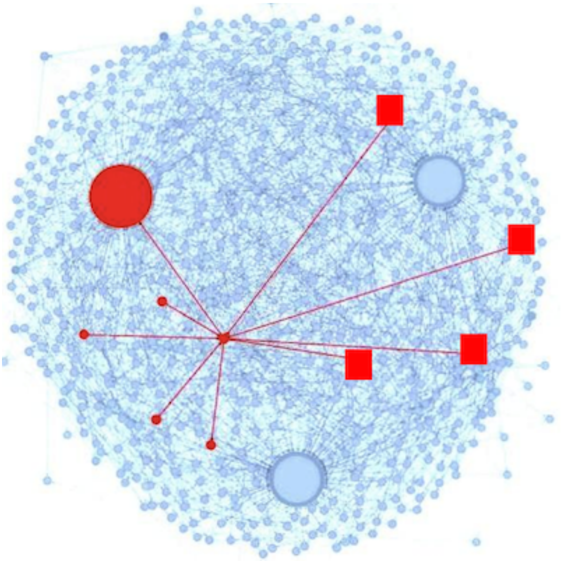
$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

The cumulative density function (CDF) of any normal

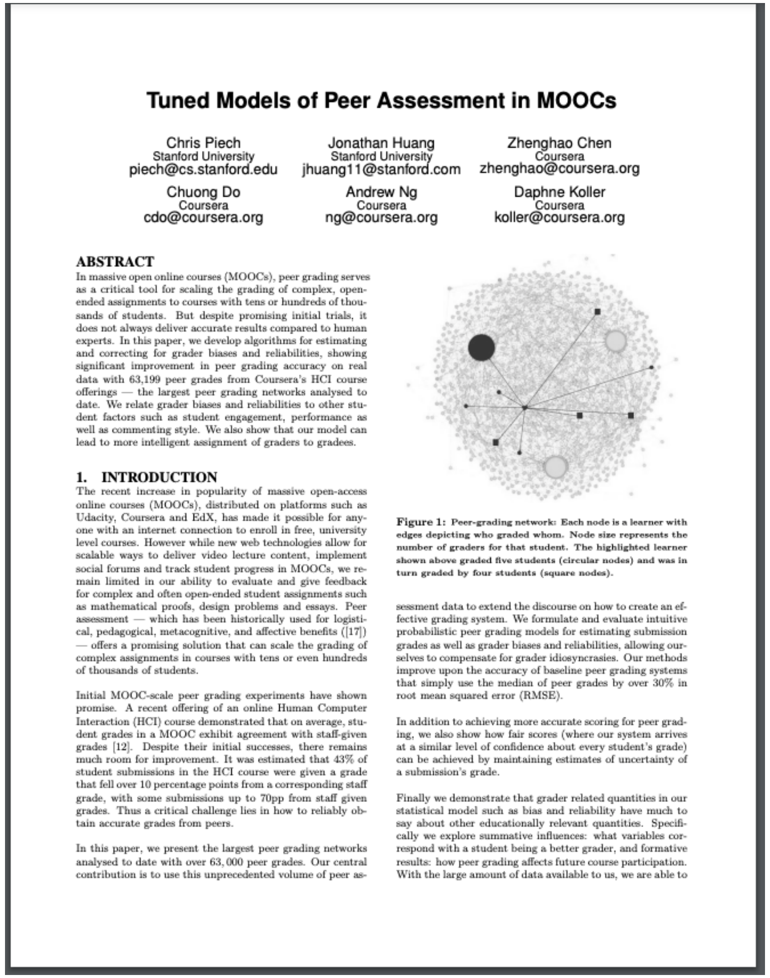
Table of $\Phi(Z)$ values in textbook, p. 201 and handout

End of review

My first paper as a PhD student was working with normals



You have 70k peer grades. Jointly figure out each student's true grade, and how good each person is at grading.



Tuned Models of Peer Assessment in MOOCs

Chris Piech
Stanford University
piech@cs.stanford.edu

Jonathan Huang
Stanford University
jhuang11@stanford.com

Zhenghao Chen
Coursera
zhenghao@coursera.org

Chuong Do
Coursera
cdo@coursera.org

Andrew Ng
Coursera
ng@coursera.org

Daphne Koller
Coursera
koller@coursera.org

ABSTRACT
In massive open online courses (MOOCs), peer grading serves as a critical tool for scaling the grading of complex, open-ended assignments to courses with tens or hundreds of thousands of students. But despite promising initial trials, it does not always deliver accurate results compared to human experts. In this paper, we develop algorithms for estimating and correcting for grader biases and reliabilities, showing significant improvement in peer grading accuracy on real data with 63,199 peer grades from Coursera's HCI course offerings — the largest peer grading networks analysed to date. We relate grader biases and reliabilities to other student factors such as student engagement, performance as well as commenting style. We also show that our model can lead to more intelligent assignment of graders to gradees.

1. INTRODUCTION
The recent increase in popularity of massive open-access online courses (MOOCs), distributed on platforms such as Udacity, Coursera and EdX, has made it possible for anyone with an internet connection to enroll in free, university level courses. However while new web technologies allow for scalable ways to deliver video lecture content, implement social forums and track student progress in MOOCs, we remain limited in our ability to evaluate and give feedback for complex and often open-ended student assignments such as mathematical proofs, design problems and essays. Peer assessment — which has been historically used for logistical, pedagogical, metacognitive, and affective benefits ([17]) — offers a promising solution that can scale the grading of complex assignments in courses with tens or even hundreds of thousands of students.

Initial MOOC-scale peer grading experiments have shown promise. A recent offering of an online Human Computer Interaction (HCI) course demonstrated that on average, student grades in a MOOC exhibit agreement with staff-given grades [12]. Despite their initial successes, there remains much room for improvement. It was estimated that 43% of student submissions in the HCI course were given a grade that fell over 10 percentage points from a corresponding staff grade, with some submissions up to 70pp from staff given grades. Thus a critical challenge lies in how to reliably obtain accurate grades from peers.

In this paper, we present the largest peer grading networks analysed to date with over 63,000 peer grades. Our central contribution is to use this unprecedented volume of peer assessment data to extend the discourse on how to create an effective grading system. We formulate and evaluate intuitive probabilistic peer grading models for estimating submission grades as well as grader biases and reliabilities, allowing ourselves to compensate for grader idiosyncrasies. Our methods improve upon the accuracy of baseline peer grading systems that simply use the median of peer grades by over 30% in root mean squared error (RMSE).

In addition to achieving more accurate scoring for peer grading, we also show how fair scores (where our system arrives at a similar level of confidence about every student's grade) can be achieved by maintaining estimates of uncertainty of a submission's grade.

Finally we demonstrate that grader related quantities in our statistical model such as bias and reliability have much to say about other educationally relevant quantities. Specifically we explore summative influences: what variables correspond with a student being a better grader, and formative results: how peer grading affects future course participation. With the large amount of data available to us, we are able to

Figure 1: Peer-grading network: Each node is a learner with edges depicting who graded whom. Node size represents the number of graders for that student. The highlighted learner shown above graded five students (circular nodes) and was in turn graded by four students (square nodes).

1. Gibbs sampling for Model $\mathbf{PG}_1$

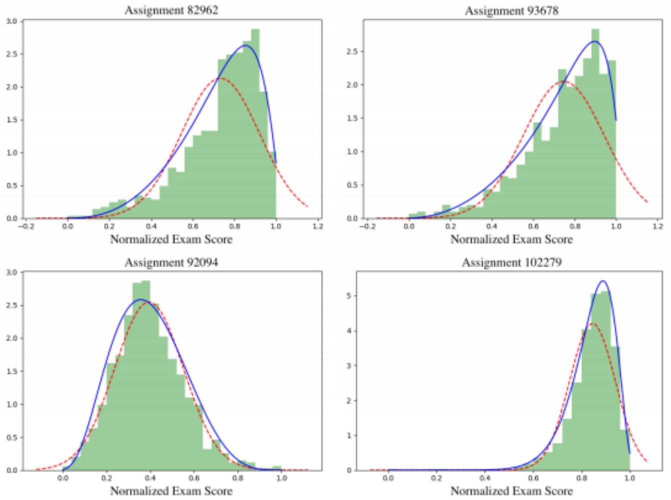
Model $\mathbf{PG}_1$ is given as follows:

$$\begin{aligned}
\text{(Reliability) } \tau_v &\sim \mathcal{G}(\alpha_0, \beta_0) \text{ for every grader } v,\\
\text{(Bias) } b_v &\sim \mathcal{N}(0, 1/\eta_0) \text{ for every grader } v,\\
\text{(True score) } s_u &\sim \mathcal{N}(\mu_0, 1/\gamma_0) \text{ for every user } u, \text{ and}\\
\text{(Observed score) } z_u^v &\sim \mathcal{N}(s_u + b_v, 1/\tau_v),\\
&\quad \text{for every observed peer grade.}
\end{aligned}$$

The joint posterior distribution is:

$$P(Z|\{s_u\}_{u\in U}, \{b_v\}_{v\in G}, \{\tau_v\}_{v\in G}) = \prod_u P(s_u|\mu_0, \gamma_0) \cdot \prod_v P(b_v|\eta_0) \cdot P(\tau_v|\alpha_0, \beta_0) \prod_{z_u^v} P(z_u^v|s_u, b_v, \tau_v).$$

But grades are not normal...



http://stanford.edu/~cpiech/bio/papers/appendices/edm13_appendix.pdf

Great questions!
Great thinkers start with great questions. Ask away!!!

How does python sample from a Gaussian?

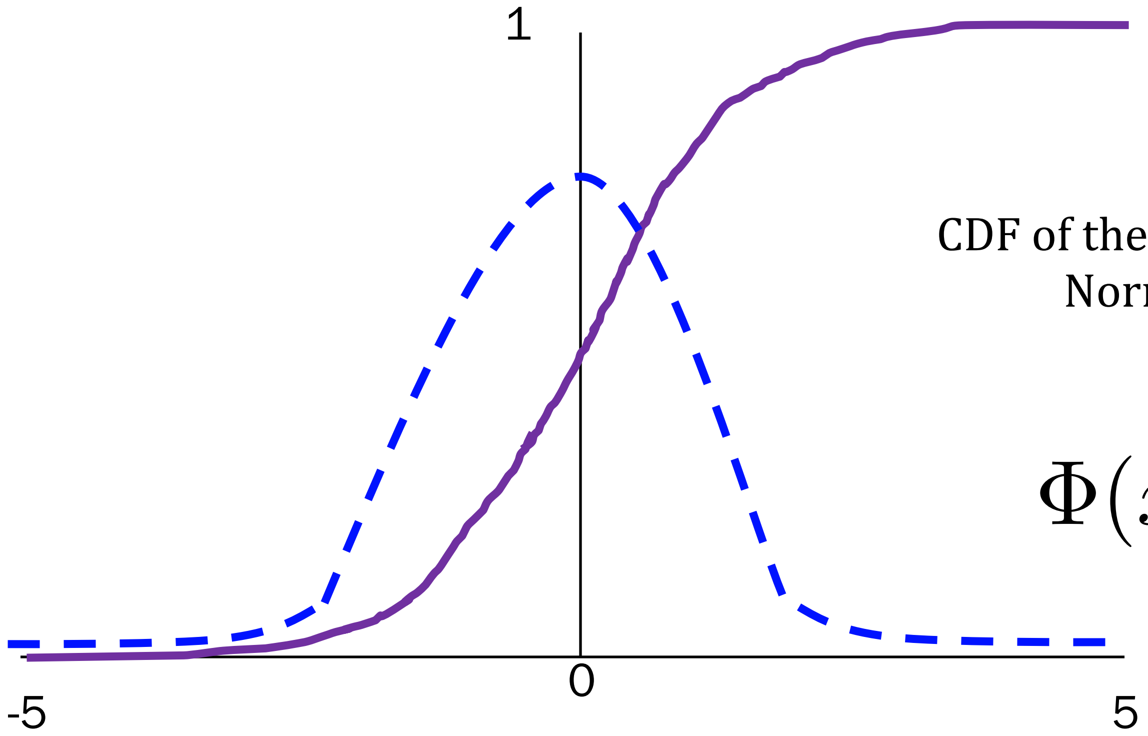
```python
from random import *

for i in range(10):
    mean = 5
    std = 1
    sample = gauss(mean, std)
    print sample
```

How does this work?

```
3.79317794179
5.19104589315
4.209360629
5.39633891584
7.10044176511
6.72655475942
5.51485158841
4.94570606131
6.14724644482
4.73774184354
```

How Does a Computer Sample a Normal?



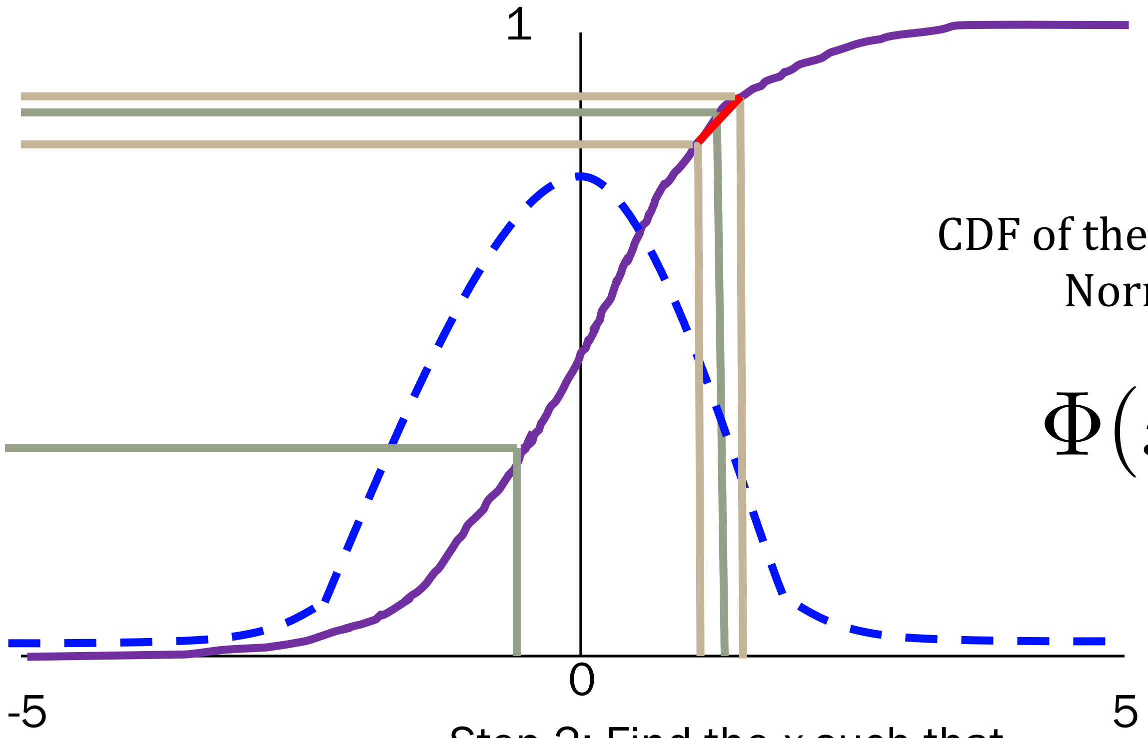
CDF of the Standard Normal

$$\Phi(x)$$

How Does a Computer Sample a Normal?

Inverse Transform Sampling

Step 1: pick a uniform number y between 0,1



CDF of the Standard Normal

$$\Phi(x)$$

Step 2: Find the x such that

$$\Phi(x) = y$$

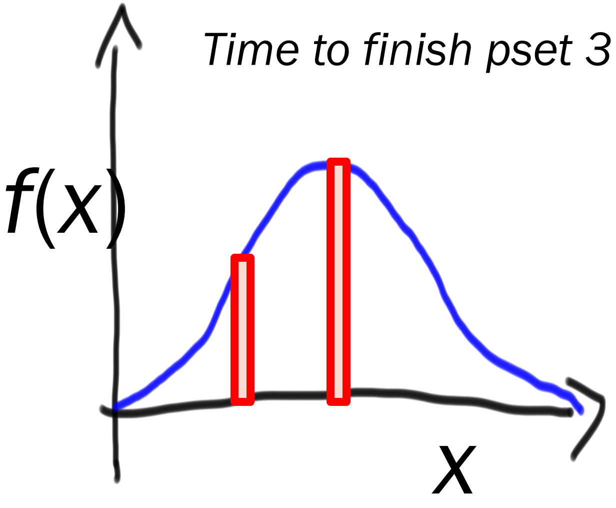
$$x = \Phi^{-1}(y)$$

Further reading: Box–Muller transform

Relative Probability of Continuous Variables

X = time to finish pset 3

X ~ N(10, 2)



How much more likely are you to complete in 10 hours than in 5?

$$\frac{P(X=10)}{P(X=5)} = \frac{\varepsilon f(X=10)}{\varepsilon f(X=5)}$$

$$= \frac{f(X=10)}{f(X=5)}$$

$$= \frac{\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(10-\mu)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(5-\mu)^2}{2\sigma^2}}}$$

$$= \frac{\frac{1}{\sqrt{4\pi}} e^{-\frac{(10-10)^2}{4}}}{\frac{1}{\sqrt{4\pi}} e^{-\frac{(5-10)^2}{4}}}$$

$$= \frac{e^0}{e^{-\frac{25}{4}}} = 518$$

Imagine you are taking a quiz…
With no computer!!!

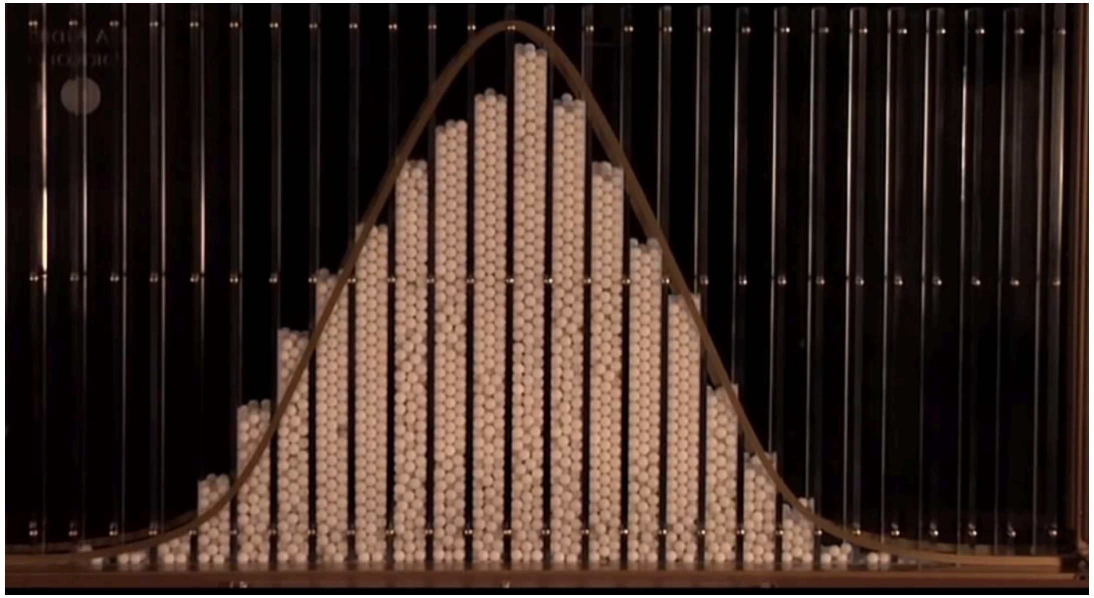
Website Testing

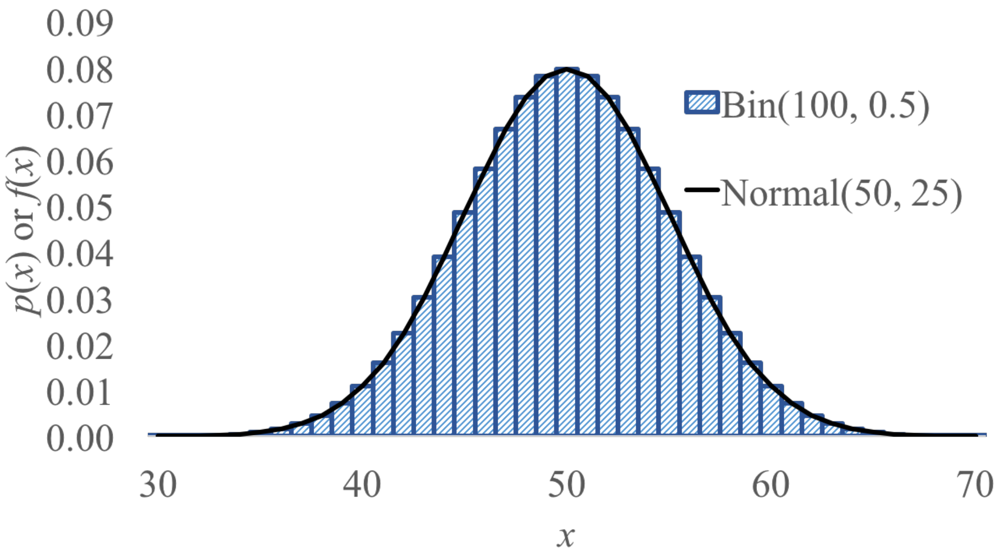
100 people are given a new website design

- X = # people whose time on site increases
- CEO will endorse new design if X ≥ 65 What is P(CEO endorses change| it has no effect)?
- X ~ Bin(100, 0.5). Want to calculate P(X ≥ 65)
- Give a numerical answer…

$$P(X \geq 65) = \sum_{i=65}^{100} \binom{100}{i} (0.5)^i (1 - 0.5)^{100-i}$$

Don’t worry, Normal approximates Binomial





Galton Board

(We’ll explain *why* in 2 weeks’ time)

Website testing

- 100 people are given a new website design.
- X = # people whose time on site increases
- The design actually has no effect, so P(time on site increases) = 0.5 independently.
- CEO will endorse the new design if $X \geq 65$.

What is $P(\text{CEO endorses change})$? *Give a numerical approximation*.

Approach 1: Binomial

Define

$X \sim \text{Bin}(n = 100, p = 0.5)$

Want: $P(X \geq 65)$

Solve

$P(X \geq 65) \approx 0.0018$

Approach 2: approximate with Normal

Define

$Y \sim \mathcal{N}(\mu, \sigma^2)$

$\mu = np = 50$

$\sigma^2 = np(1-p) = 25$

$\sigma = \sqrt{25} = 5$

Solve

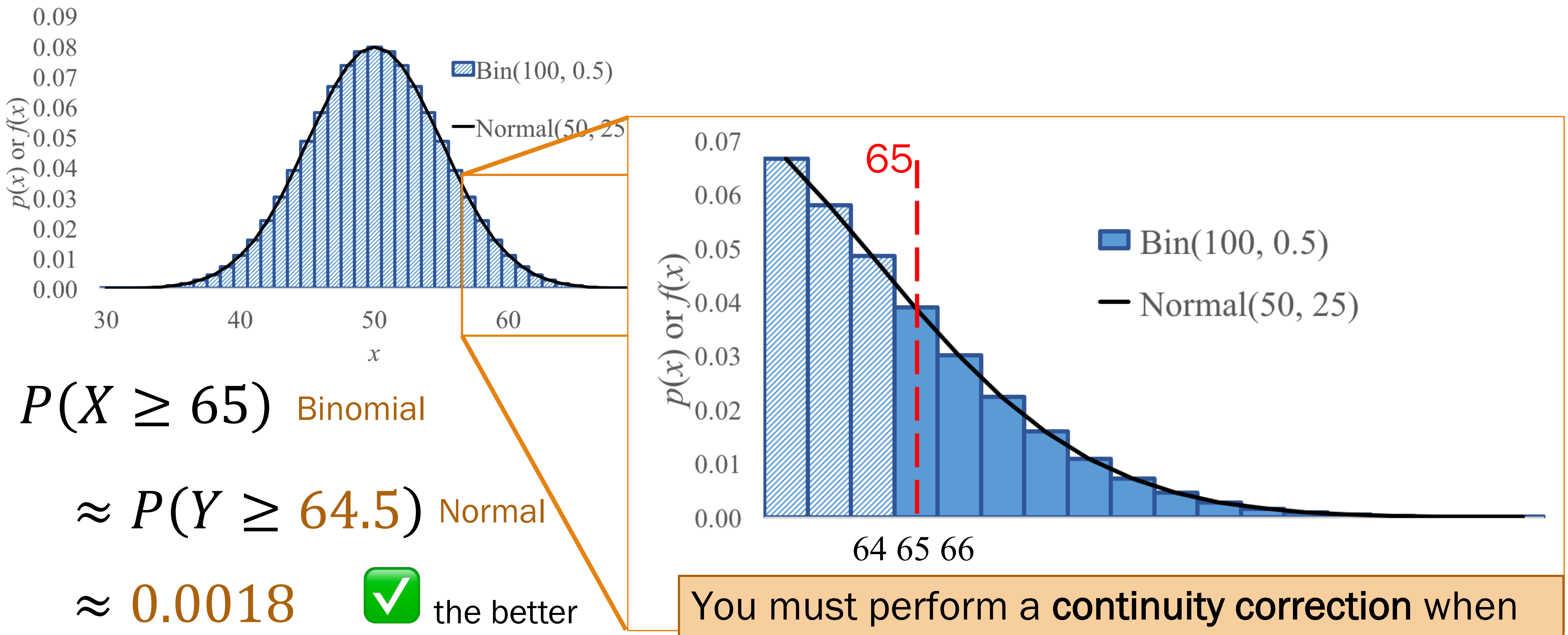
$$P(X \geq 65) \approx P(Y \geq 65) = 1 - F_Y(65)$$

$$= 1 - \Phi\left(\frac{65-50}{5}\right) = 1 - \Phi(3) \approx 0.0013 \text{ ?}$$

⚠️ ⚠️ (this approach is missing something important)

Website testing (with continuity correction)

In our website testing, $Y \sim \mathcal{N}(50, 25)$ approximates $X \sim \text{Bin}(100, 0.5)$.



$P(X \geq 65)$ Binomial

$\approx P(Y \geq 64.5)$ Normal

≈ 0.0018 ✅ the better

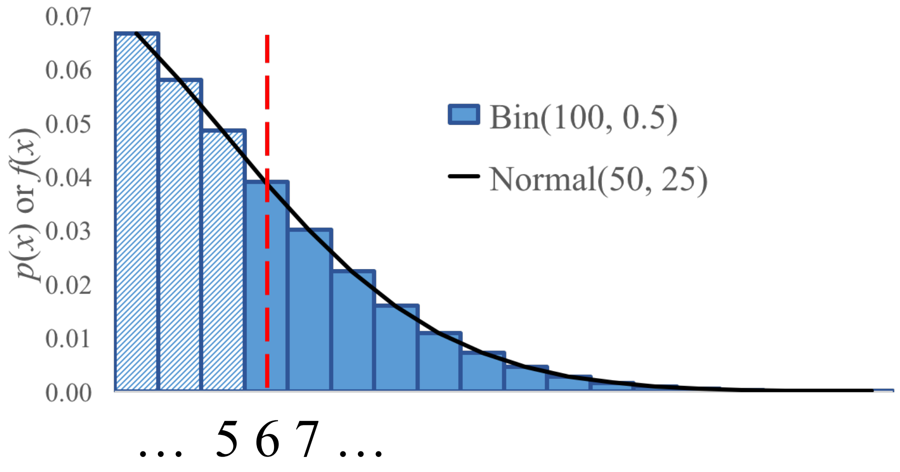
Approach 2

You must perform a **continuity correction** when approximating a Binomial RV with a Normal RV.

Continuity correction

If $Y \sim \mathcal{N}(np, np(1-p))$ approximates $X \sim \text{Bin}(n, p)$, how do we approximate the following probabilities?

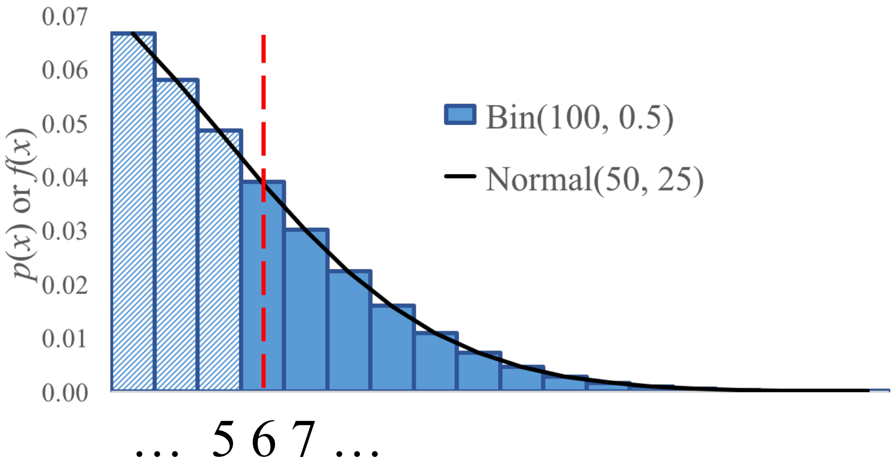
Discrete (e.g., Binomial) probability question	Continuous (Normal) probability question
$P(X = 6)$	
$P(X \geq 6)$	
$P(X > 6)$	
$P(X < 6)$	
$P(X \leq 6)$	



Continuity correction

If $Y \sim \mathcal{N}(np, np(1-p))$ approximates $X \sim \text{Bin}(n, p)$, how do we approximate the following probabilities?

Discrete (e.g., Binomial) probability question	Continuous (Normal) probability question
$P(X = 6)$	$P(5.5 \le Y \le 6.5)$
$P(X \ge 6)$	$P(Y \ge 5.5)$
$P(X > 6)$	$P(Y \ge 6.5)$
$P(X < 6)$	$P(Y \le 5.5)$
$P(X \le 6)$	$P(Y \le 6.5)$



Normal Approximation

Who gets to approximate?

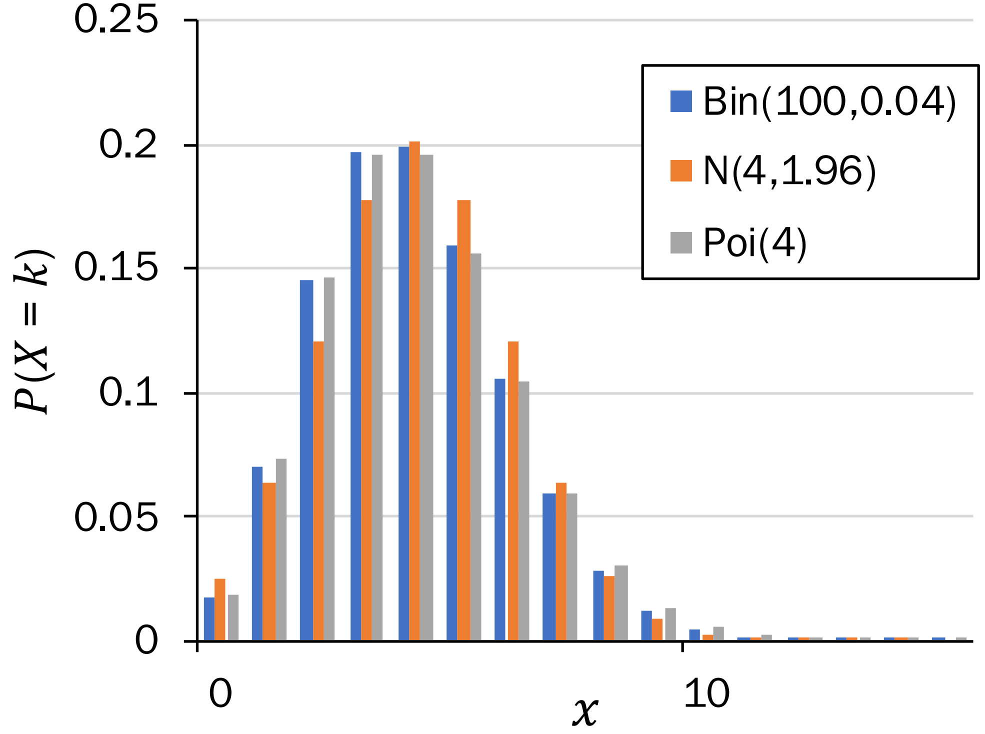
$$X \sim \text{Bin}(n, p)$$
$$E[X] = np$$
$$\text{Var}(X) = np(1-p)$$

$$Y \sim \text{Poi}(\lambda)$$
$$\lambda = np$$

?

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$
$$\mu = np$$
$$\sigma^2 = np(1-p)$$

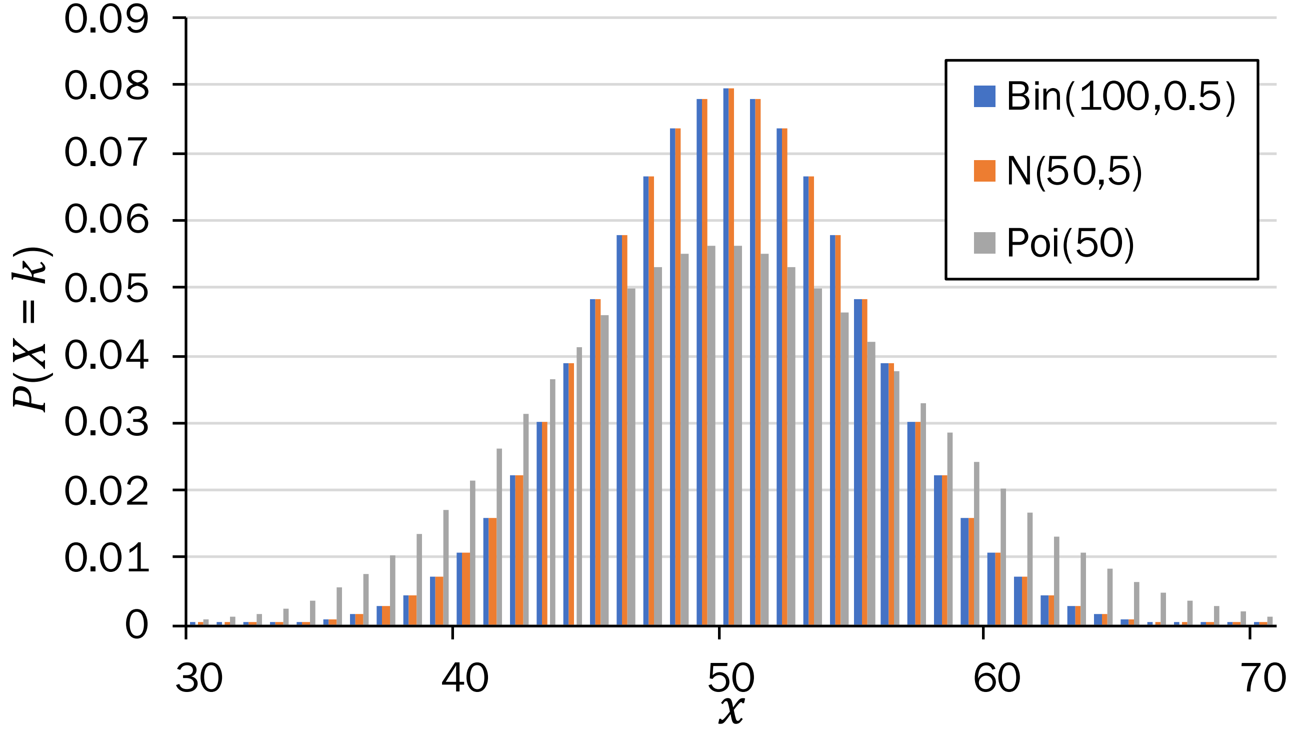
Who gets to approximate?



Poisson approximation

n large (> 20), p small (< 0.05)

slight dependence okay



Normal approximation

n large (> 20), p mid-ranged ($np(1-p) > 10$)

independence

1. If there is a choice, use Normal to approximate.
2. When using Normal to approximate a discrete RV, use a continuity correction.

Stanford Admissions (a while back)

Stanford accepts 2480 students.

- Each admitted student matriculates w.p. 0.68 (independent trials)
- Let X = # of students who will attend

What is $P(X > 1745)$? *Give a numerical approximation*.

Strategy:

A. Just Binomial
B. Poisson
C. Normal
D. None/other

(by yourself)

Stanford Admissions

Stanford accepts 2480 students.

- Each admitted student matriculates w.p. 0.68 (independent trials)
- Let X = # of students who will attend

What is $P(X > 1745)$? *Give a numerical approximation*.

Strategy:

A. Just Binomial — not an approximation (also computationally expensive)
B. Poisson — $p = 0.68$, not small enough
C. Normal ✅ Variance $np(1-p) = 540 > 10$
D. None/other

Define an approximation

Let $Y \sim \mathcal{N}(E[X], \text{Var}(X))$

$E[X] = np = 1686$

$\text{Var}(X) = np(1-p) \approx 540 \to \sigma = 23.3$

$P(X > 1745) \approx P(Y \geq 1745.5)$ ⚠️ Continuity correction

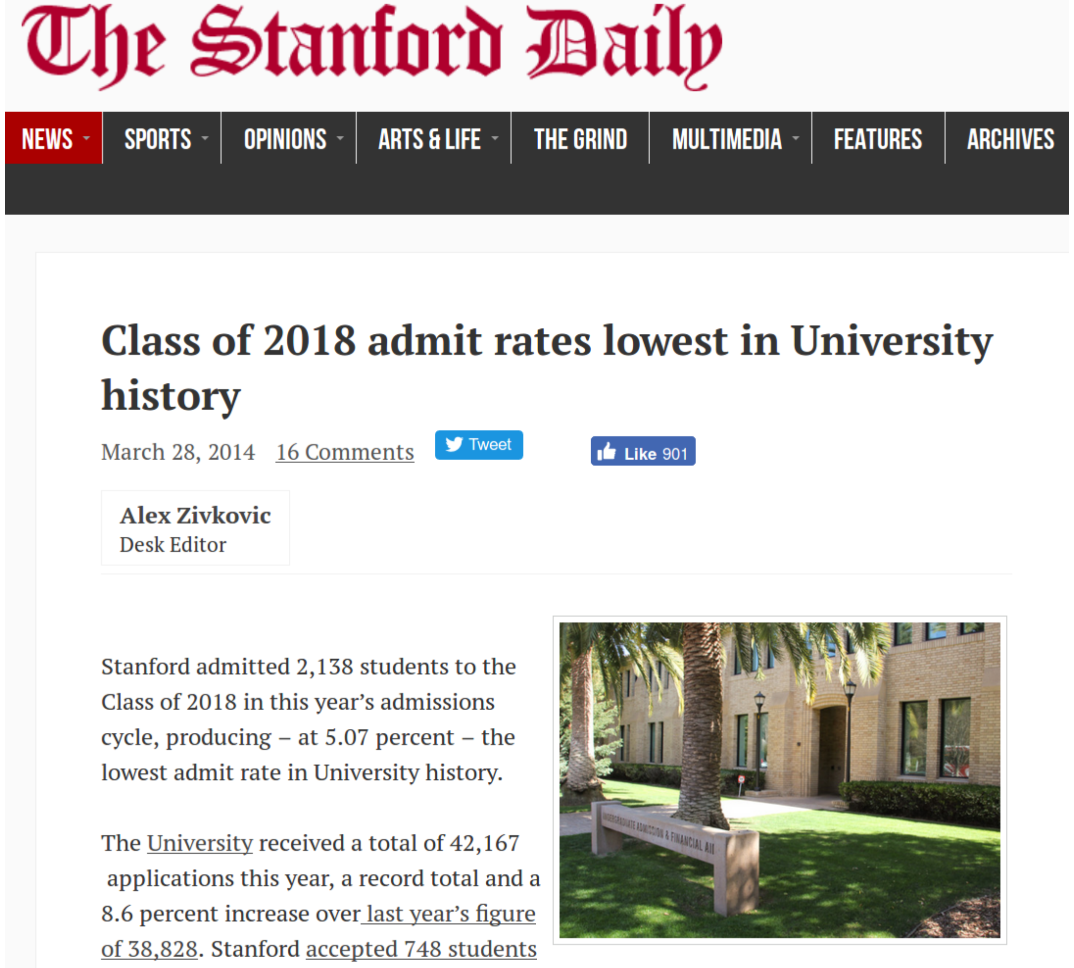
Solve

$$P(Y \geq 1745.5) = 1 - F(1745.5)$$

$$= 1 - \Phi\left(\frac{1745.5 - 1686}{23.3}\right)$$

$$= 1 - \Phi(2.54) \approx 0.0055$$

How many students should Stanford admit?



The Stanford Daily

NEWS | SPORTS | OPINIONS | ARTS & LIFE | THE GRIND | MULTIMEDIA | FEATURES | ARCHIVES

Class of 2018 admit rates lowest in University history

March 28, 2014 16 Comments Tweet Like 901

Alex Zivkovic
Desk Editor

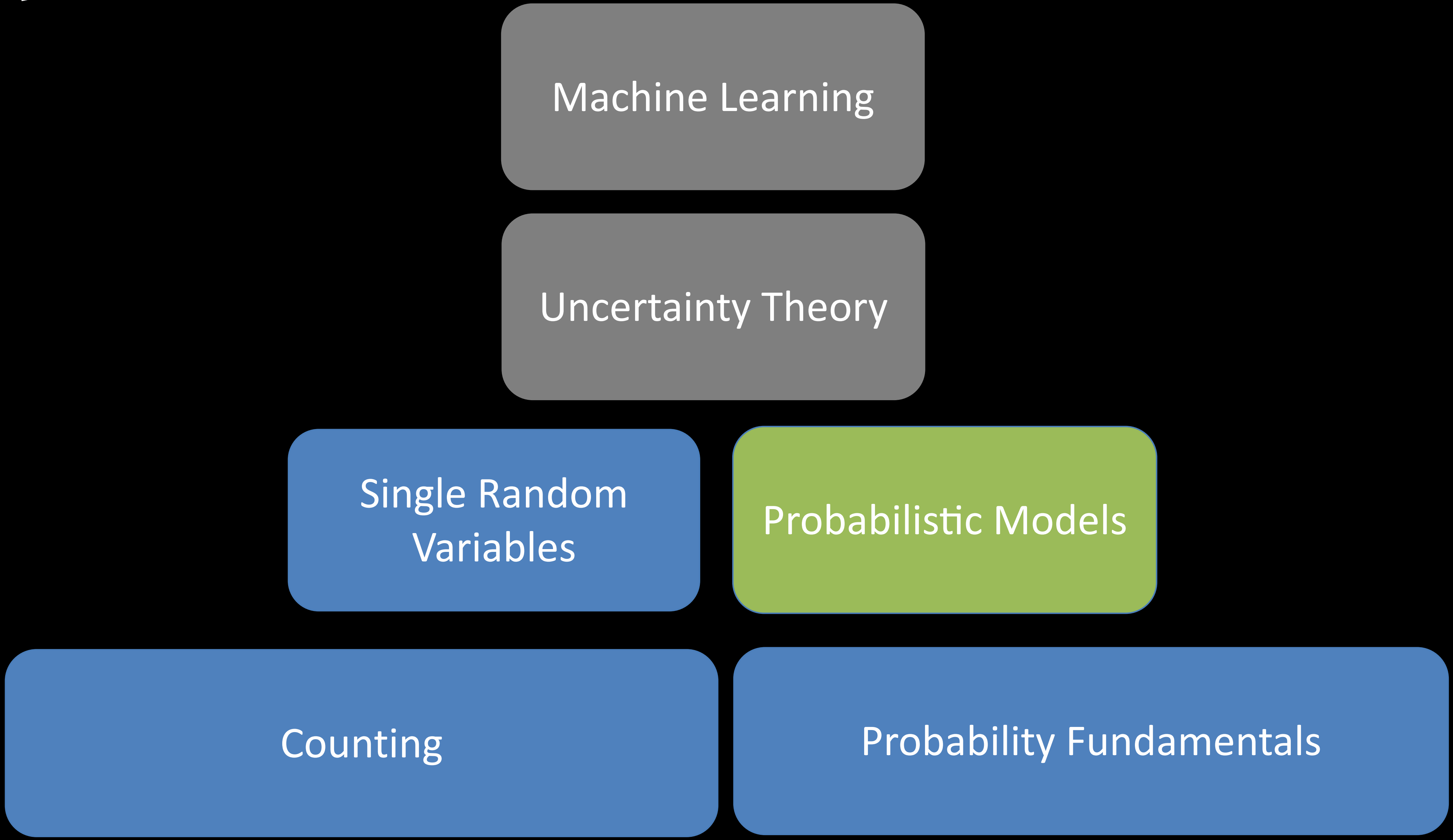
Stanford admitted 2,138 students to the Class of 2018 in this year's admissions cycle, producing – at 5.07 percent – the lowest admit rate in University history.

The University received a total of 42,167 applications this year, a record total and a 8.6 percent increase over last year's figure of 38,828. Stanford accepted 748 students

Admit rate: 4.3%

Yield rate: 81.9%

CS109


Machine Learning
Uncertainty Theory
Single Random Variables
Probabilistic Models
Counting
Probability Fundamentals

[suspense]

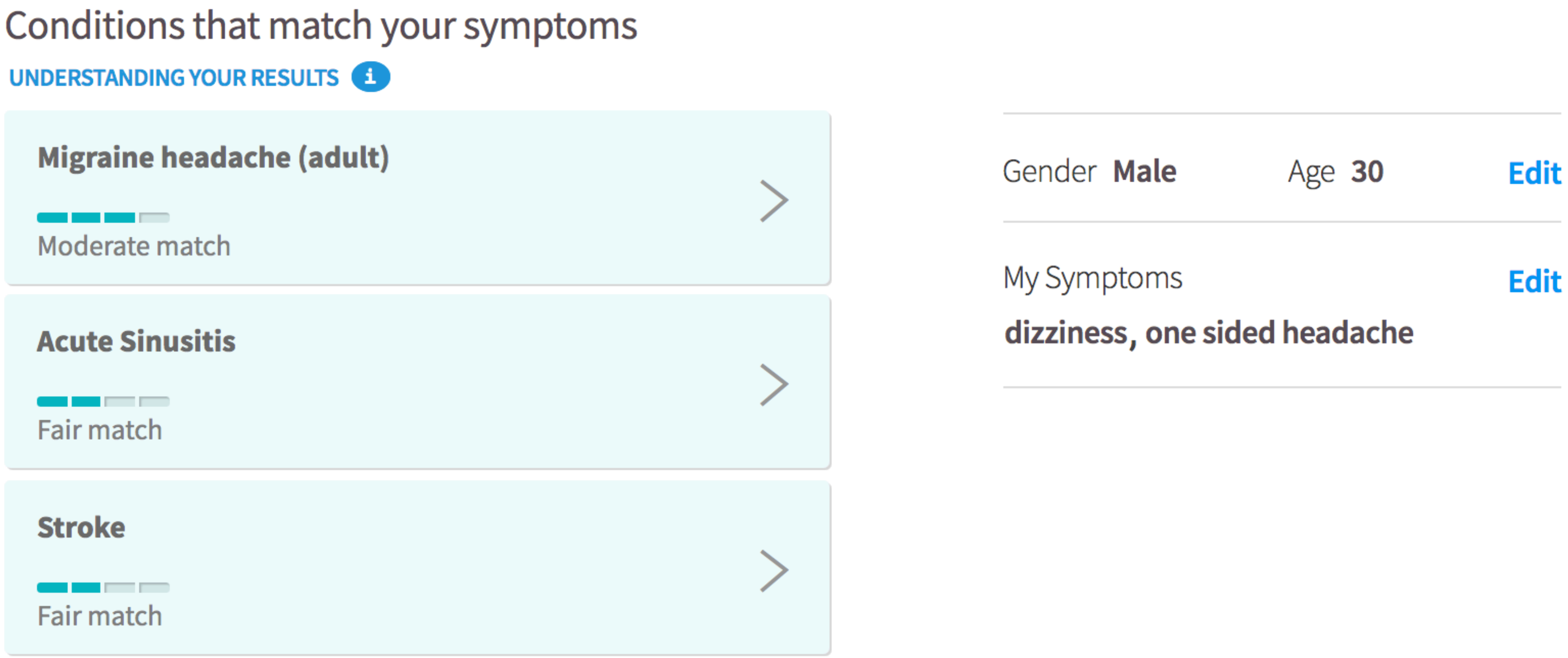
Discrete Probabilistic Models

The world is full of interesting probability problems

Have multiple random variables interacting with one another

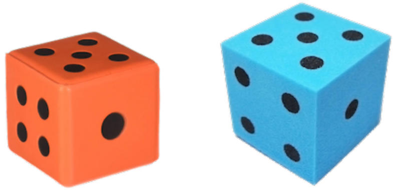
Multiple Random Variables. Start of Digital Revolution

Multiple Random Variables. Start of Digital Revolution



Joint probability mass functions

Roll two 6-sided dice, yielding values X and Y.



$$X$$

random variable

$$P(X = 1)$$

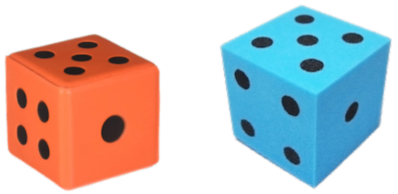
probability of an event

$$P(X = k)$$

probability mass function

Joint probability mass functions

Roll two 6-sided dice, yielding values X and Y.



X random variable	$P(X = 1)$ probability of an event	$P(X = k)$ probability mass function
X, Y random variables	$P(X = 1 \cap Y = 6)$ $P(X = 1, Y = 6)$ new notation: the comma probability of the intersection of two events	$P(X = a, Y = b)$ joint probability mass function

Discrete joint distributions

For two discrete joint random variables X and Y,
the joint probability mass function is defined as:

$$p_{X,Y}(a,b) = P(X = a, Y = b)$$

The marginal distributions of the joint PMF are defined as:

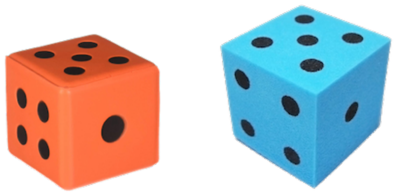
$$p_X(a) = P(X = a) = \sum_y p_{X,Y}(a,y)$$

$$p_Y(b) = P(Y = b) = \sum_x p_{X,Y}(x,b)$$

Use marginal distributions to get a 1-D RV from a joint PMF.

Two dice

Roll two 6-sided dice, yielding values X and Y.



1. What is the joint PMF of X and Y?

$$p_{X,Y}(a,b) = 1/36 \qquad (a,b) \in \{(1,1), \dots, (6,6)\}$$

		X					
		1	2	3	4	5	6
Y	1	1/36	…	…	…	…	1/36
	2	…	…	…	…		
	3	…	…	…	…	…	…
	4	…	…	…	…	…	…
	5	…	…	…	…	…	…
	6	1/36	…	…	…	…	1/36

$P(X = 4, Y = 2)$

Probability table

- All possible outcomes for several discrete RVs
- Not parametric (e.g., parameter p in Ber(p))

Marriage Pact in CS109. Data from a few years ago

	Single	In a relationship	It's complicated
Freshman	0.13	0.08	0.02
Sophomore	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.76
5+	0.06	0.09	0.04

Joint is Complete Information!

	Single	Relationship	Complicated
Frosh	0.13	0.08	0.02
Soph	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
5+	0.06	0.09	0.04

A joint distribution is complete information. It can be used to answer any probability question.

Joint table: mutually exclusive and covers sample space.

	Single	Relationship	Complicated
Frosh	0.13	0.08	0.02
Soph	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
5+	0.06	0.09	0.04

Each combination is mutually exclusive, and they span the sample space

$$\sum_{x \in X} \sum_{y \in Y} P(x, y) = 1$$

X is dating status.
Y is year.

Joint table: mutually exclusive and covers sample space.

	Single	Relationship	Complicated
Frosh	0.13	0.08	0.02
Soph	0.17	0.11	0.02
Junior	0.09	?	0.02
Senior	0.02	0.07	0.01
5+	0.06	0.09	0.04

Each combination is mutually exclusive, and they span the sample space

$$\sum_{x \in X} \sum_{y \in Y} P(x, y) = 1$$

X is dating status.
Y is year.

Joint table: mutually exclusive and covers sample space.

	Single	Relationship	Complicated
Frosh	0.13	0.08	0.02
Soph	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
5+	0.06	0.09	0.04

Each combination is mutually exclusive, and they span the sample space

$$\sum_{x \in X} \sum_{y \in Y} P(x, y) = 1$$

X is dating status.
Y is year.

What is the probability someone is in a relationship?

	Single	Relationship	Complicated
Frosh	0.13	0.08	0.02
Soph	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
5+	0.06	0.09	0.04

We can use the law of total probability!

X is dating status. Y is year.

$$P(X = \text{single}) = \sum_{y \in Y} P(X = \text{single}, Y = y)$$

$$P(X = \text{relation}) = \sum_{y \in Y} P(X = \text{relation}, Y = y)$$

$$P(Y = \text{frosh}) = \sum_{x \in X} P(X = x, Y = \text{frosh})$$

$$P(Y = \text{soph}) = \sum_{x \in X} P(X = x, Y = \text{soph})$$

Why is that called the marginal?

Mini WebMd

Variable	Symbol	Type
Has Determinitis	D	Bernoulli (1 indicates has Determinitis)
Fever	F	Categorical (none, low, high)
Can Smell	S	Bernoulli (1 indicates can smell)

$D = 0$

	$S = 0$	$S = 1$
$F =$ none	0.024	0.783
$F =$ low	0.003	0.092
$F =$ high	0.001	0.046

$D = 1$

	$S = 0$	$S = 1$
$F =$ none	0.006	0.014
$F =$ low	0.005	0.011
$F =$ high	0.004	0.011

Mini WebMd: What is P(D = 1 | F = low, S = 1)

Variable	Symbol
Has Determinitis	D
Fever	F
Can Smell	S

$$\Pr(D=1|F=\text{low},S=1) = \frac{\Pr(D=1,F=\text{low},S=1)}{\Pr(F=\text{low},S=1)} = \frac{0.011}{0.011+0.092} = 0.107$$

$D = 0$

	$S = 0$	$S = 1$
$F = \text{none}$	0.024	0.783
$F = \text{low}$	0.003	0.092
$F = \text{high}$	0.001	0.046

$D = 1$

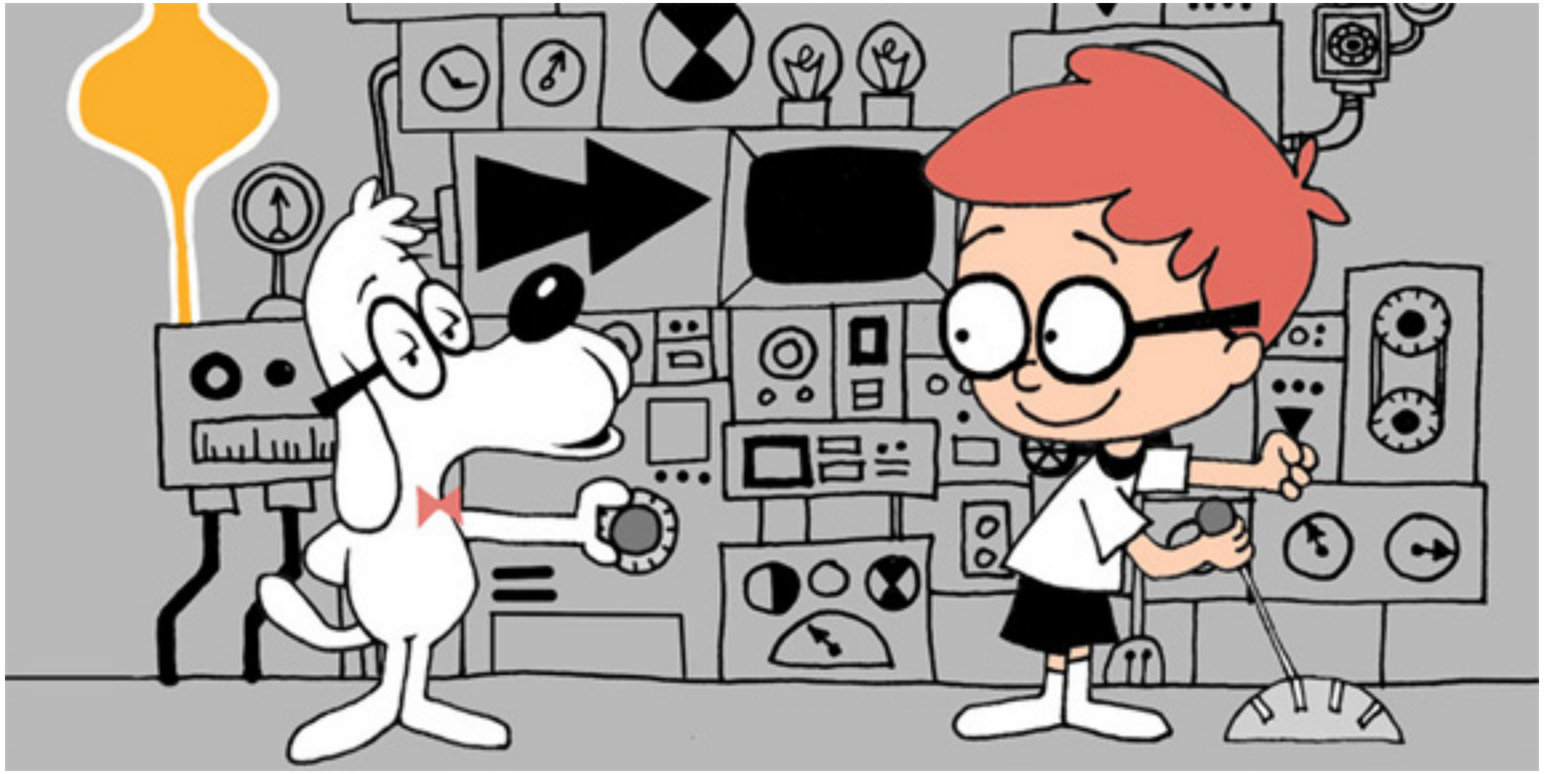
	$S = 0$	$S = 1$
$F = \text{none}$	0.006	0.014
$F = \text{low}$	0.005	0.011
$F = \text{high}$	0.004	0.011

Timnit + Joy: Gender Shades

https://chrispiech.github.io/probabilityForComputerScientists/en/examples/fairness/

Multinomial RV

Recall the good times



Permutations

$n!$

How many ways are there to order n objects?

Ways to put elements into fixed size containers

How many ways are there to put *n* objects into r buckets such that:

n_1 go into bucket 1
n_2 go into bucket 2
...
n_r go into bucket r?

$$\frac{n!}{n_1!n_2!\dots n_r!} = \binom{n}{n_1, n_2, \dots, n_r}$$

Note: Multinomial > Binomial

Counting unordered objects

Binomial coefficient

How many ways are there to group n objects into two groups of size k and $n-k$, respectively?

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Called the binomial coefficient because of something from Algebra

Multinomial coefficient

How many ways are there to group n objects into r groups of sizes $n_1, n_2, \ldots, n_r$ respectively?

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1!\,n_2!\cdots n_r!}$$

Multinomials generalize Binomials for counting.

Probability

Binomial RV

What is the probability of getting k successes and $n - k$ failures in n trials?

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Binomial # of ways of ordering the successes

Probability of each ordering of k successes is equal + mutually exclusive

Multinomial RV

What is the probability of getting c_1 of outcome 1, c_2 of outcome 2, …, and c_m of outcome m in n trials?

Multinomial RVs also generalize Binomial RVs for probability!

Multinomial Random Variable

Consider an experiment of n independent trials:

- Each trial results in one of m outcomes. $P(\text{outcome } i) = p_i, \sum_{i=1}^{m} p_i = 1$
- Let X_i = # trials with outcome i

Joint PMF

$$P(X_1 = c_1, X_2 = c_2, \dots, X_m = c_m) = \binom{n}{c_1, c_2, \dots, c_m} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$$

where $\sum_{i=1}^{m} c_i = n$ and $\sum_{i=1}^{m} p_i = 1$

Multinomial # of ways of ordering the outcomes

Probability of each ordering is equal + mutually exclusive

Hello dice rolls, my old friends

A 6-sided die is rolled 7 times.

What is the probability of getting:

- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes

Hello dice rolls, my old friends

A 6-sided die is rolled 7 times.

What is the probability of getting:

- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes

$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 2, X_5 = 0, X_6 = 3)$$

$$= \binom{7}{1,1,0,2,0,3} \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^3 = 420 \left(\frac{1}{6}\right)^7$$

Hello dice rolls, my old friends

A 6-sided die is rolled 7 times.

What is the probability of getting:

- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes

# of times a six appears

$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 2, X_5 = 0, X_6 = 3)$$

$$= \binom{7}{1,1,0,2,0,3} \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^3 = 420 \left(\frac{1}{6}\right)^7$$

choose where the sixes appear

probability of rolling a six

this many times

Multinomial Random Variable

Review

Consider an experiment of n independent trials:

- Each trial results in one of m outcomes. $P(\text{outcome } i) = p_i$, $\sum_{i=1}^{m} p_i = 1$
- Let X_i = # trials with outcome i

Joint PMF

$$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m) = \binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$$

where $\sum_{i=1}^{m} c_i = n$ and $\sum_{i=1}^{m} p_i = 1$

Example:

- Rolling 2 twos, 3 threes, and 5 fives on 10 rolls of a fair-sided die
- Generating a random 5-word phrase with 1 “the”, 2 “bacon”, 1 “put”, 1 “on”

Hello dice rolls, my old friends

Review

A 6-sided die is rolled 7 times.

What is the probability of getting:

- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes

# of times a six appears

$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 2, X_5 = 0, X_6 = 3)$$

$$= \binom{7}{1,1,0,2,0,3} \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^3 = 420 \left(\frac{1}{6}\right)^7$$

choose where the sixes appear

probability of rolling a six

this many times

Parameters of a Multinomial RV?

$X \sim \text{Bin}(n, p)$ has parameters n, p...

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

p: probability of success outcome on a single trial

A Multinomial RV has parameters $n, p_1, p_2, \ldots, p_m$ (Note $p_m = 1 - \sum_{i=1}^{m-1} p_i$)

$$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m) = \binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$$

p_i: probability of outcome i on a single trial

Where do we get p_i from?

Pedagogic pause

The Federalist Papers

Intro to Natural Language Processing

Probabilistic text analysis

Ignoring the order of words…

What is the probability of any given word that you write in English?

- $P(\text{word = "the"}) > P(\text{word = "pokemon"})$
- $P(\text{word = "Stanford"}) > P(\text{word = "Cal"})$

Probabilities of *counts* of words = Multinomial distribution

A document is a large multinomial.

(according to the Global Language Monitor, there are 988,968 words in the English language used on the internet.)

Model text as a multinomial

Example document:
"Pay for Viagra with a credit-card. Viagra is great. So are credit-cards. Risk free Viagra. Click for free."
n = 18

$$P\left(\begin{array}{l}\text{Viagra} = 2 \\ \text{Free} = 2 \\ \text{Risk} = 1 \\ \text{Credit-card: } 2 \\ \ldots \\ \text{For} = 2\end{array} \middle| \text{spam}\right) = \frac{n!}{2!2!\ldots2!} p_{\text{viagra}}^2 p_{\text{free}}^2 \ldots p_{\text{for}}^2$$

It's a Multinomial!

Probability of seeing this document | spam

The probability of a word in spam email being viagra

Who wrote the federalist papers?

Old and New Analysis

Authorship of the Federalist Papers

- 85 essays advocating ratification of the US constitution
- Written under the pseudonym “Publius” (really, Alexander Hamilton, James Madison, John Jay)

Who wrote which essays?

- Analyze probability of words in each essay and compare against word distributions from known writings of three authors

Who wrote Federalist Paper 53?

madison.txt

To the People of the State of New York:

AMONG the numerous advantages promised by a wellconstructed Union, none deserves to be more accurately developed than its tendency to break and control the violence of faction. The friend of popular governments never finds himself so much alarmed for their character and fate, as when he contemplates their propensity to this dangerous vice. He will not fail, therefore, to set a due value on any plan which, without violating the principles to which he is attached, provides a proper cure for it. The instability, injustice, and confusion introduced into the public councils, have, in truth, been the mortal diseases under which popular governments have everywhere perished; as they continue to be the favorite and fruitful topics from which the adversaries to liberty derive their most specious declamations. The valuable improvements made by the American constitutions on the popular models, both ancient and modern, cannot certainly be too much admired; but it would be an unwarrantable partiality, to contend that they have as effectually obviated the danger on this side, as was wished and expected. Complaints are everywhere heard from our most considerate and virtuous citizens, equally the friends of public and private faith, and of public and personal liberty, that our governments are too unstable, that the public good is disregarded in the conflicts of rival parties, and that measures are too often decided, not according to the rules of justice and the rights of the minor party, but by the superior force of an interested and overbearing majority. However anxiously we may wish that these complaints had no foundation, the evidence, of known facts will not permit us to deny that they are in some degree true. It will be found, indeed, on a candid review of our situation, that some of the distresses under which we labor have been erroneously charged on the operation of our governments; but it will be found, at the same time, that other causes will not alone account for many of our heaviest misfortunes; and, particularly, for that prevailing and increasing distrust of public

Line 3, Column 154 · Tab Size: 4 · Plain Text

hamilton.txt

The Utility of the Union in Respect to Commercial Relations and a Navy
Hamilton for the Independent Journal.

To the People of the State of New York:
THE importance of the Union, in a commercial light, is one of those points about which there is least room to entertain a difference of opinion, and which has, in fact, commanded the most general assent of men who have any acquaintance with the subject. This applies as well to our intercourse with foreign countries as with each other.

There are appearances to authorize a supposition that the adventurous spirit, which distinguishes the commercial character of America, has already excited uneasy sensations in several of the maritime powers of Europe. They seem to be apprehensive of our too great interference in that carrying trade, which is the support of their navigation and the foundation of their naval strength. Those of them which have colonies in America look forward to what this country is capable of becoming, with painful solicitude. They foresee the dangers that may threaten their American dominions from the neighborhood of States, which have all the dispositions, and would possess all the means, requisite to the creation of a powerful marine. Impressions of this kind will naturally indicate the policy of fostering divisions among us, and of depriving us, as far as possible, of an active commerce in our own bottoms. This would answer the threefold purpose of preventing our interference in their navigation, of monopolizing the profits of our trade, and of clipping the wings by which we might soar to a dangerous greatness. Did not prudence forbid the detail, it would not be difficult to trace, by facts, the workings of this policy to the cabinets of ministers.

If we continue united, we may counteract a policy so unfriendly to our prosperity in a variety of ways. By prohibitory regulations, extending, at the same time, throughout the States, we may oblige foreign countries to bid against each

Line 5, Column 249 · Tab Size: 4 · Plain Text

unknown.txt

To the People of the State of New York:
I SHALL here, perhaps, be reminded of a current observation, ``that where annual elections end, tyranny begins. '' If it be true, as has often been remarked, that sayings which become proverbial are generally founded in reason, it is not less true, that when once established, they are often applied to cases to which the reason of them does not extend. I need not look for a proof beyond the case before us. What is the reason on which this proverbial observation is founded? No man will subject himself to the ridicule of pretending that any natural connection subsists between the sun or the seasons, and the period within which human virtue can bear the temptations of power. Happily for mankind, liberty is not, in this respect, confined to any single point of time; but lies within extremes, which afford sufficient latitude for all the variations which may be required by the various situations and circumstances of civil society. The election of magistrates might be, if it were found expedient, as in some instances it actually has been, daily, weekly, or monthly, as well as annual; and if circumstances may require a deviation from the rule on one side, why not also on the other side? Turning our attention to the periods established among ourselves, for the election of the most numerous branches of the State legislatures, we find them by no means coinciding any more in this instance, than in the elections of other civil magistrates. In Connecticut and Rhode Island, the periods are half-yearly. In the other States, South Carolina excepted, they are annual. In South Carolina they are biennial as is proposed in the federal government. Here is a difference, as four to one, between the longest and shortest periods; and yet it would be not easy to show, that Connecticut or Rhode Island is better governed, or enjoys a greater share of rational liberty, than South Carolina; or that either the one or the other of these States is distinguished in these respects, and by these causes, from the States whose elections are different from both. In searching for the grounds of this doctrine, I can discover but one, and that is wholly inapplicable to our case. The important distinction so well

Line 2, Column 519 · Tab Size: 4 · Plain Text

Where to start?

We have words, we want to know probability of authorship. We also know probability of words given author…

Well hello again…

Who wrote Federalist Paper 53?

Prob Document given Hamilton

Prior belief it was Hamilton

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Prob Hamilton given Document

Prob of the document???

Who wrote Federalist Paper 53?

Model document as a multinomial where we care about count of words

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Who wrote Federalist Paper 53?

Loop over unique words

Prob hamilton would write word i

Prior belief it was Hamilton

Number of times word i is in the doc

$$P(H|D) = \frac{\binom{n}{c_1 \ldots c_k} \cdot \prod_i h_i^{c_i} \cdot P(H)}{P(D)}$$

Prob Hamilton given Document

Prob of the document???

Who wrote Federalist Paper 53?

Prob that Hamilton wrote it

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

$$= \frac{P(H) \cdot \binom{n}{c_1 \ldots c_m} \cdot \prod_i h_i^{c_i}}{P(D)}$$

Prob that Madison wrote it

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

$$= \frac{P(M) \cdot \binom{n}{c_1 \ldots c_m} \cdot \prod_i m_i^{c_i}}{P(D)}$$

$$\frac{P(H|D)}{P(M|D)} = \frac{P(M) \cdot \binom{n}{c_1 \ldots c_k} \cdot \prod_i h_i^{c_i}}{P(D)} \Big/ \frac{P(H) \cdot \binom{n}{c_1 \ldots c_k} \cdot \prod_i m_i^{c_i}}{P(D)}$$

$$= \frac{\prod_i h_i^{c_i}}{\prod_i m_i^{c_i}}$$

Who wrote Federalist Paper 53?

Probability Hamilton wrote the document

For each word i

Prob that hamilton writes word i

Times word i shows up in the doc

$$\frac{P(H|D)}{P(M|D)} = \frac{\prod_i h_i^{c_i}}{\prod_i m_i^{c_i}}$$

Probability Madison wrote the document

Prob that Madison writes word i

To the code

What happened?

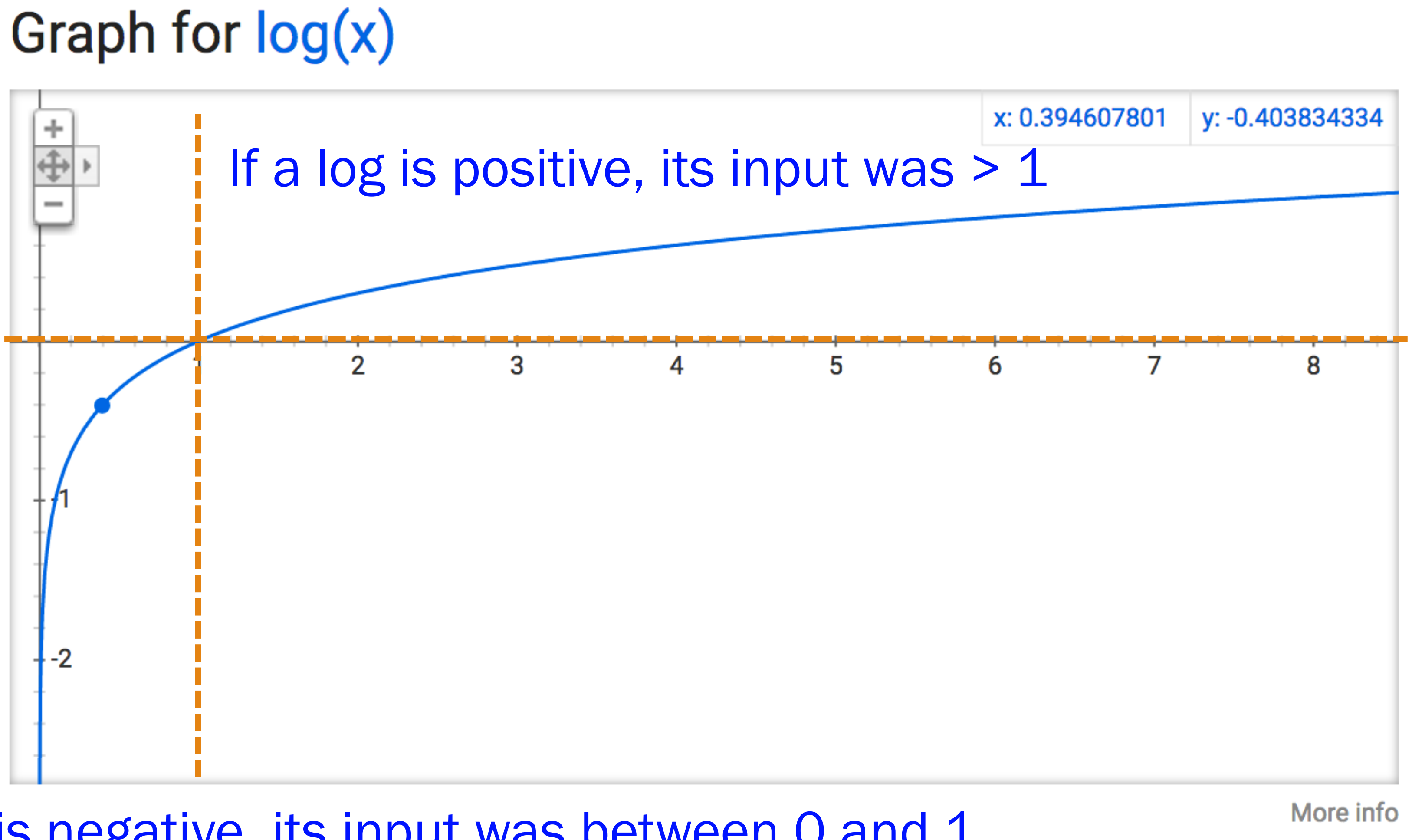
All our probabilities are zero…

Use logs when probabilities become too small!

$$\frac{P(H|D)}{P(M|D)} = \frac{\prod_i m_i^{c_i}}{\prod_i h_i^{c_i}}$$

$$\log \frac{P(H|D)}{P(M|D)} = \log \frac{\prod_i h_i^{c_i}}{\prod_i m_i^{c_i}}$$
$$= \sum_i \log h_i^{c_i} - \sum_i \log m_i^{c_i}$$
$$= \sum_i c_i \cdot \log h_i - \sum_i c_i \log m_i$$

What does it mean if a log value is positive / negative



If a log is negative, its input was between 0 and 1

To be continued…