Joint Distributions

Will Song
CS109, Stanford University

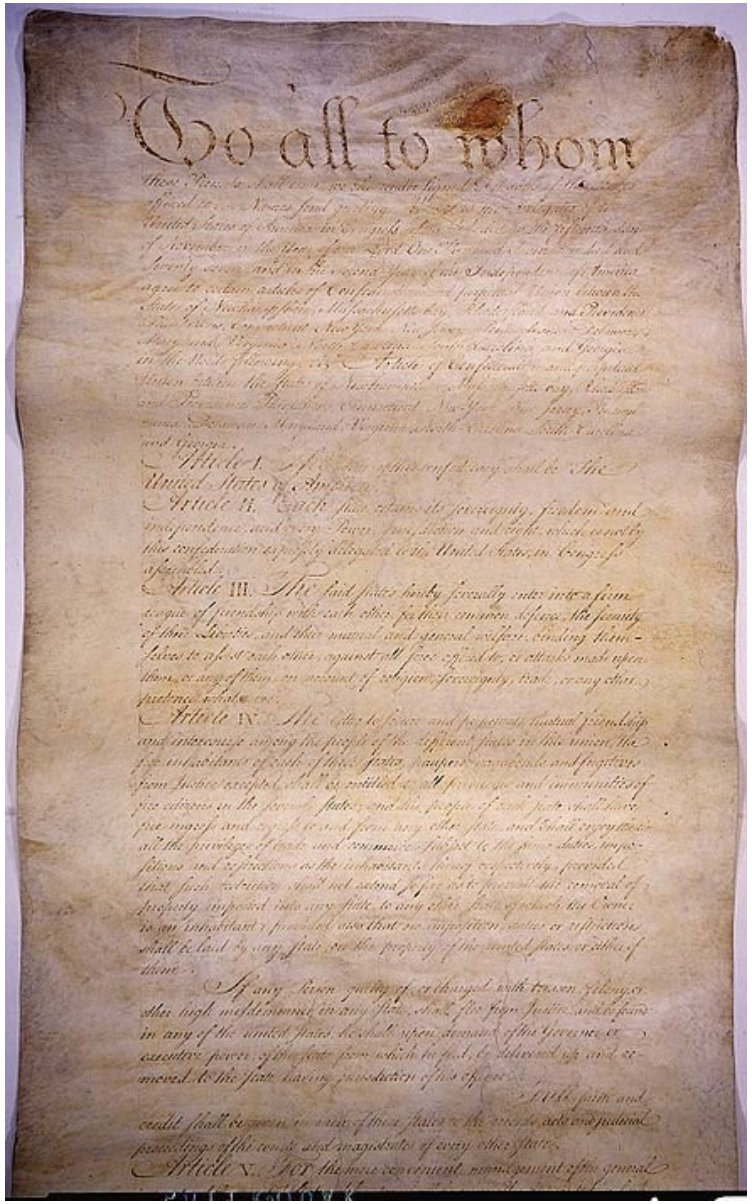
Announcements!

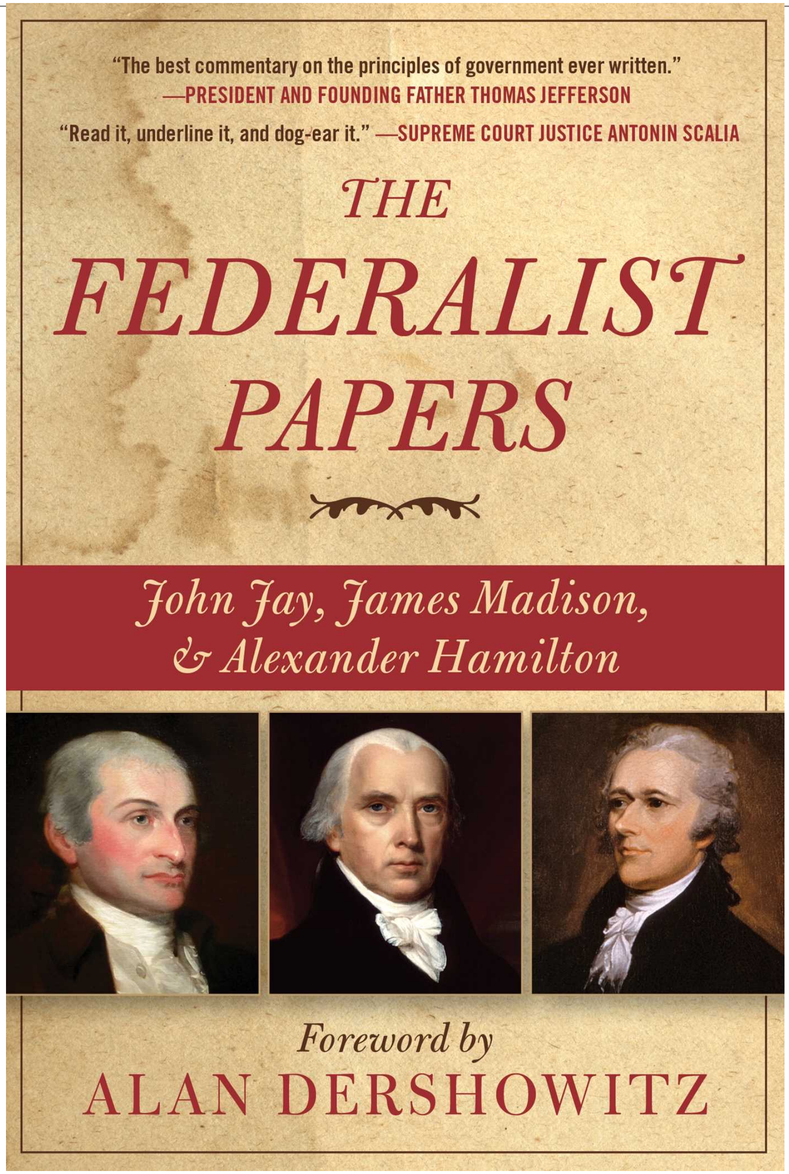
Midterm Next Tuesday 7p—9pm

- Location: 370-370

- Exam preparation going live soon!

Exciting Day!


To all to whom

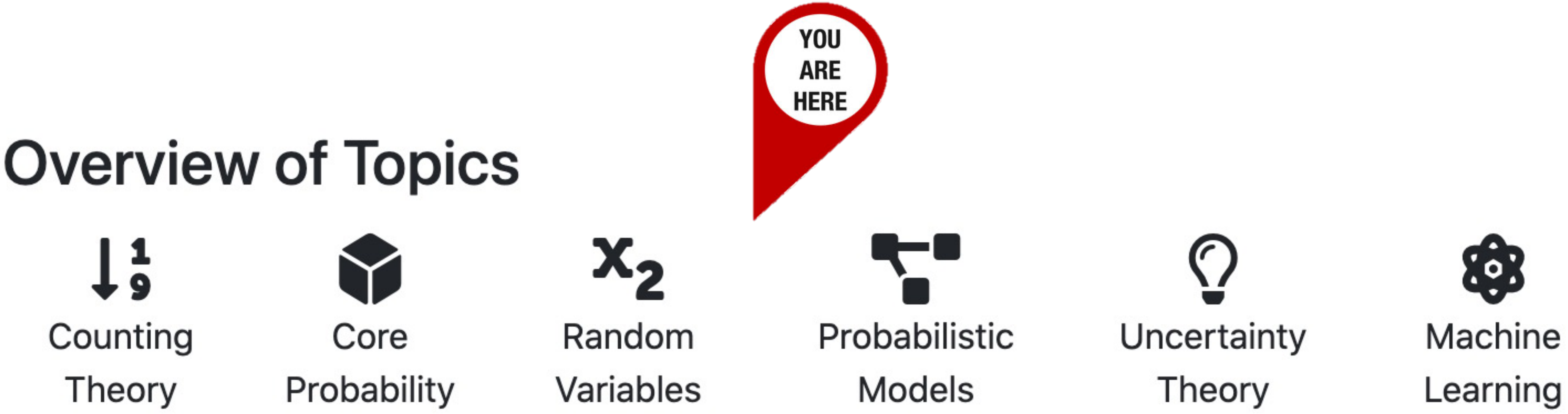


Learning Goals

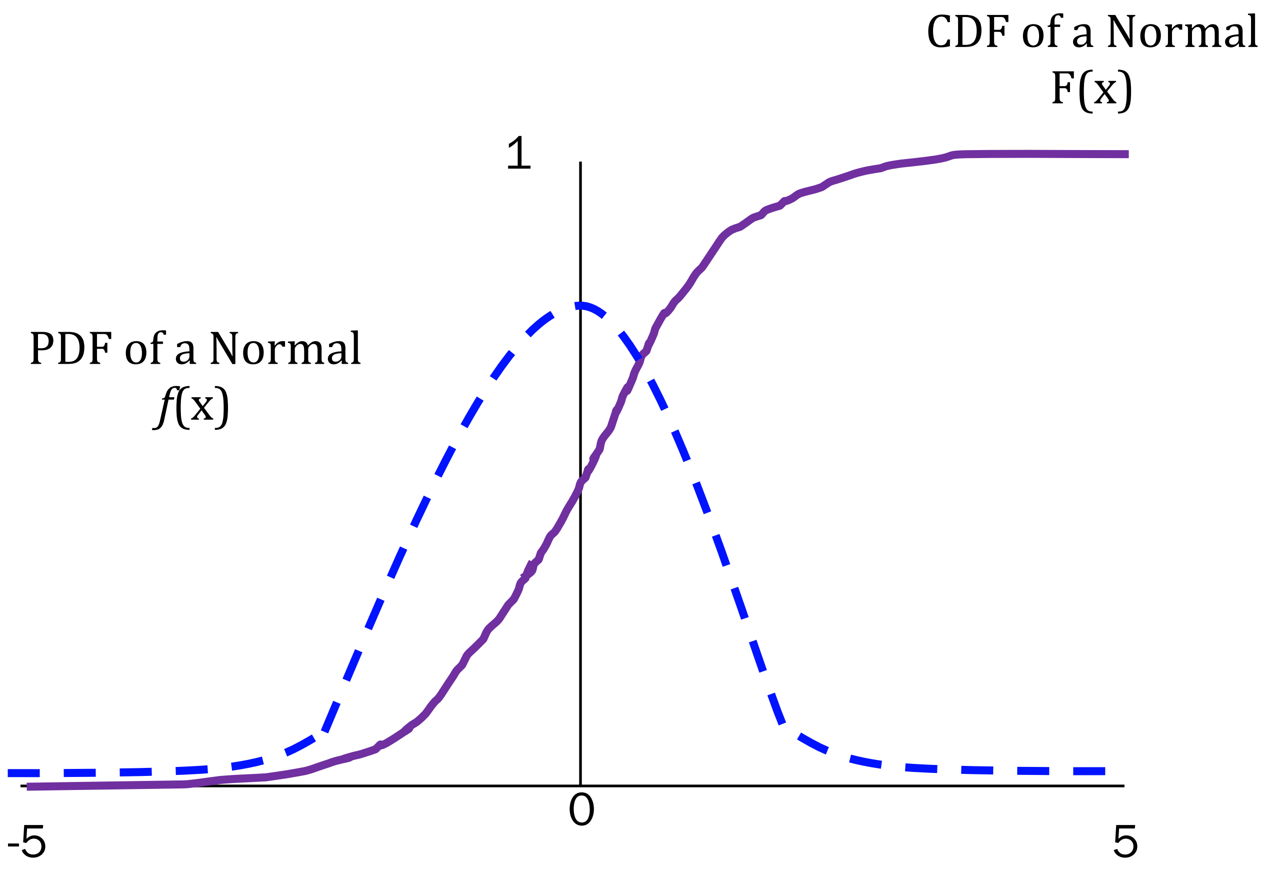
1. Be able to Approximate Binomials with a Normal!
2. Be able to reason about joint distributions.
3. Be able to recognize applications of multinomials.

First, some review

Where are we in CS109?



Normal Distribution



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?

Cumulative Density Function

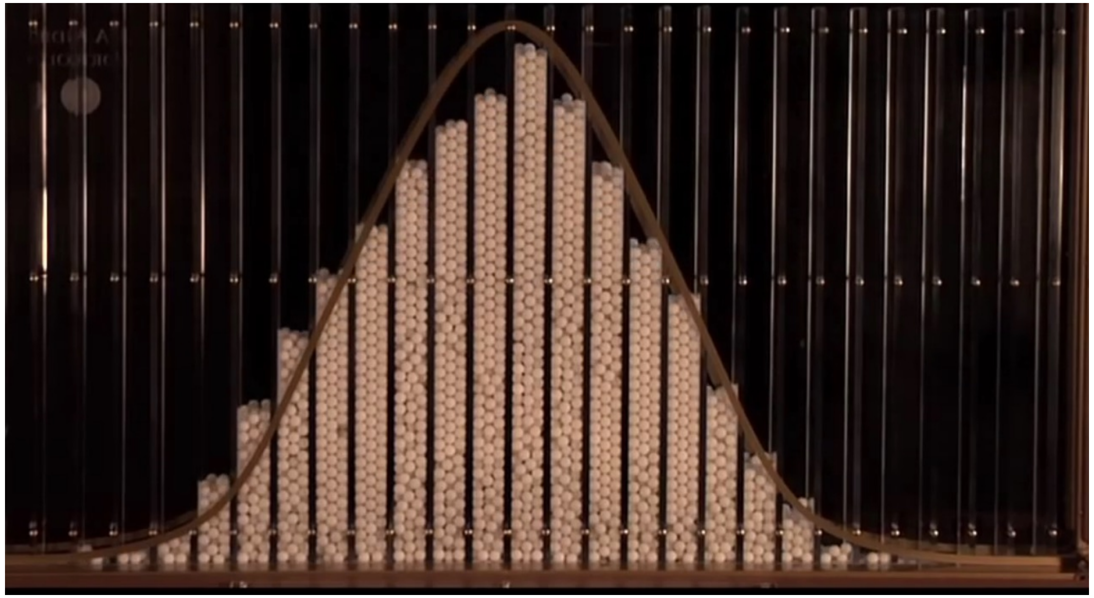
$$\mathcal{N}(\mu, \sigma^2)$$

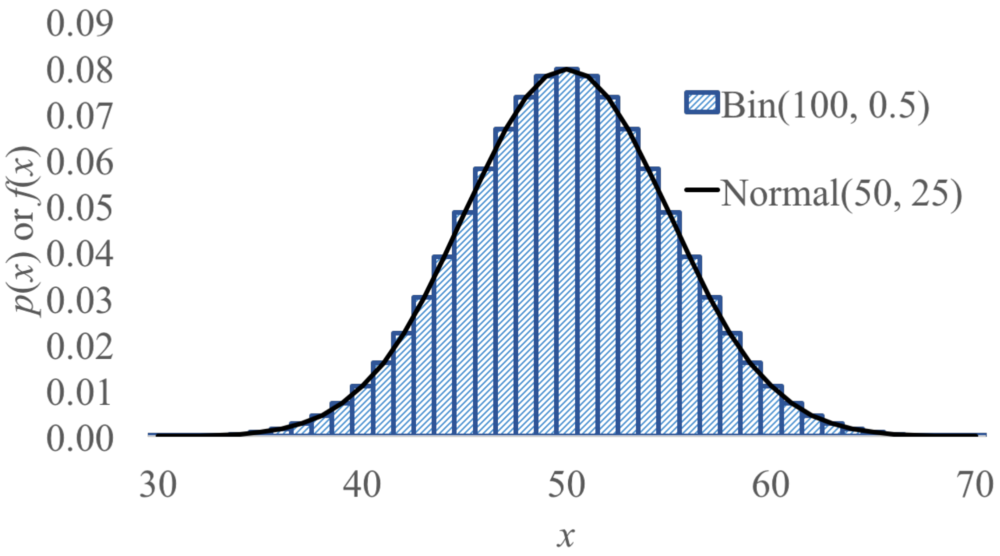
CDF of Standard Normal: A function that has been solved for numerically

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

The cumulative density function (CDF) of any normal

Don’t worry, Normal approximates Binomial





Galton Board

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(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

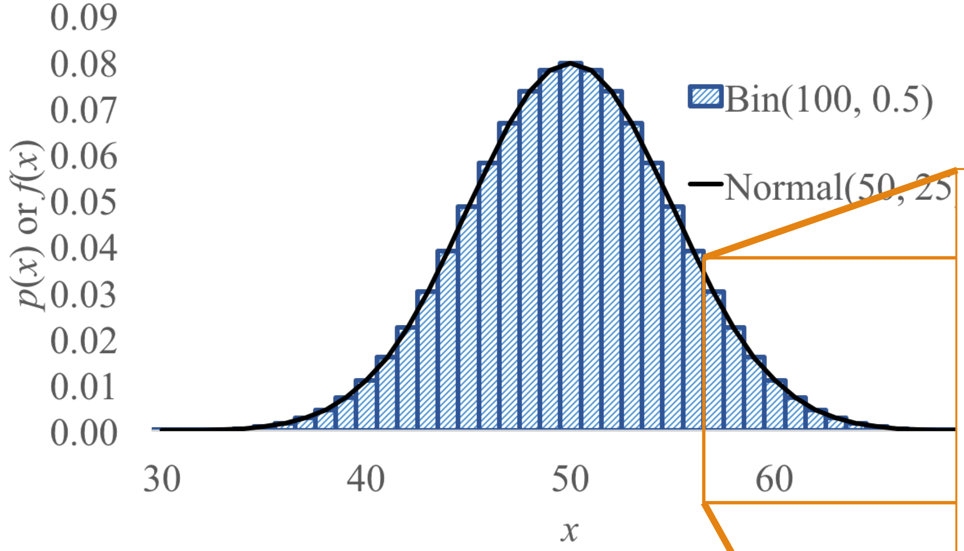
$$P(X \geq 65) \approx P(Y \geq 65) = 1 - F_Y(65)$$

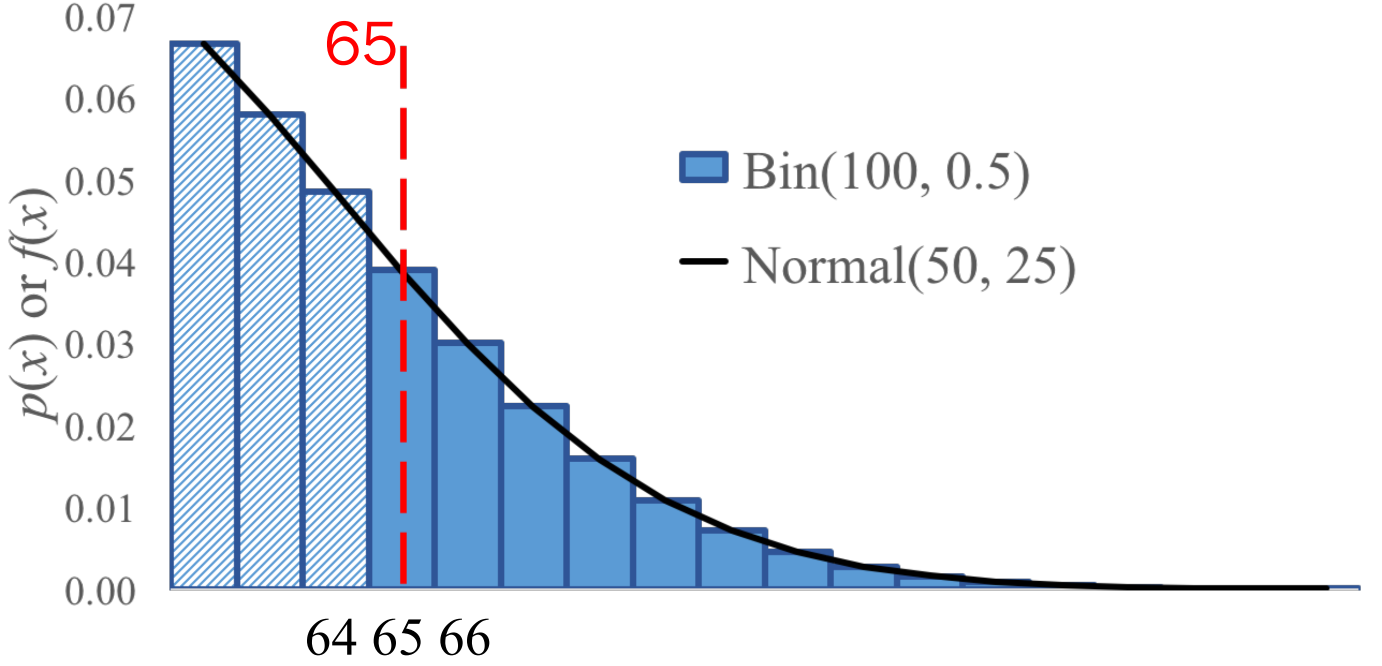
$$= 1 - \Phi\left(\frac{65-50}{5}\right) = 1 - \Phi(3) \approx 0.0013 \text{ ?}$$

⚠️ ⚠️ (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
$P(X = 6)$	$P(5.5 \leq Y \leq 6.5)$
$P(X \geq 6)$	$P(Y \geq 5.5)$
$P(X > 6)$	$P(Y \geq 6.5)$
$P(X < 6)$	$P(Y \leq 5.5)$
$P(X \leq 6)$	$P(Y \leq 6.5)$

$$P(5.5 \leq Y \leq 6.5)$$

… 5 6 7 …

Who gets to approximate?

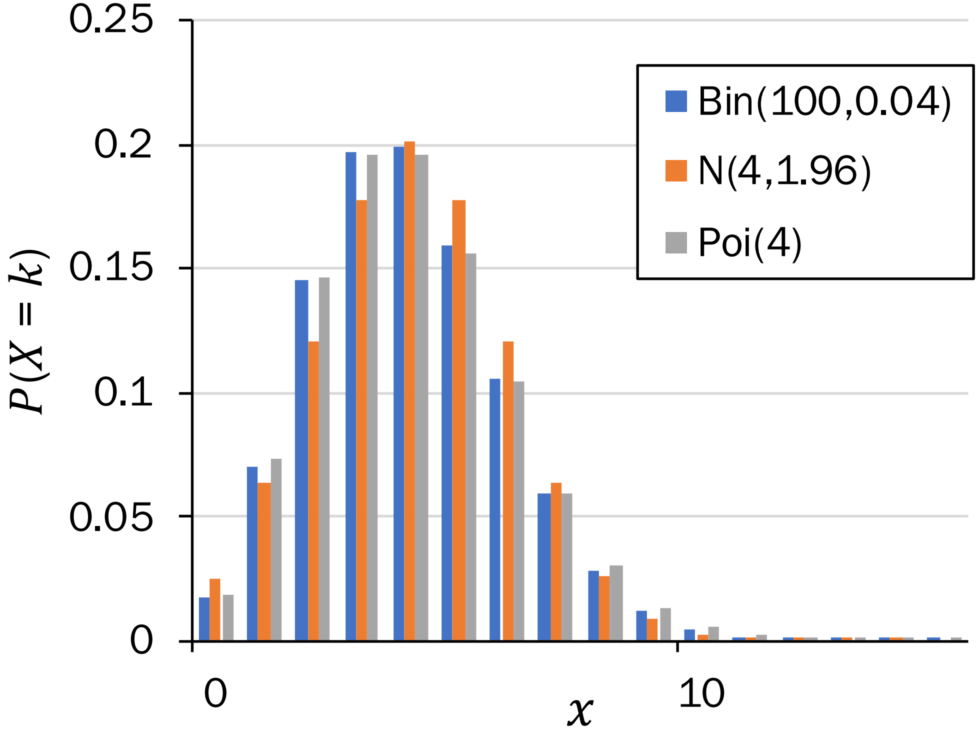
$$X \sim \text{Bin}(n, p)$$
$$E[X] = np$$
$$\text{Var}(X) = np(1-p)$$

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

?

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

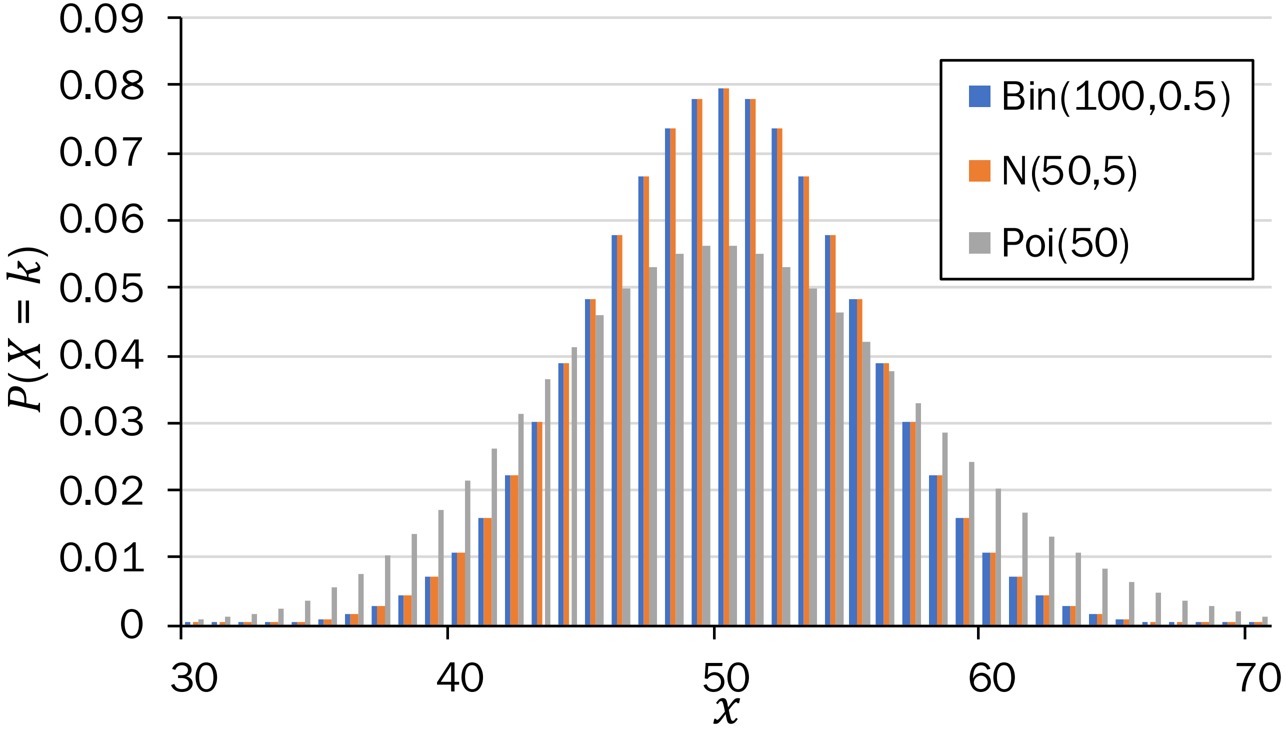
Who gets to approximate?



Poisson approximation

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

slight dependence okay



Normal approximation

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

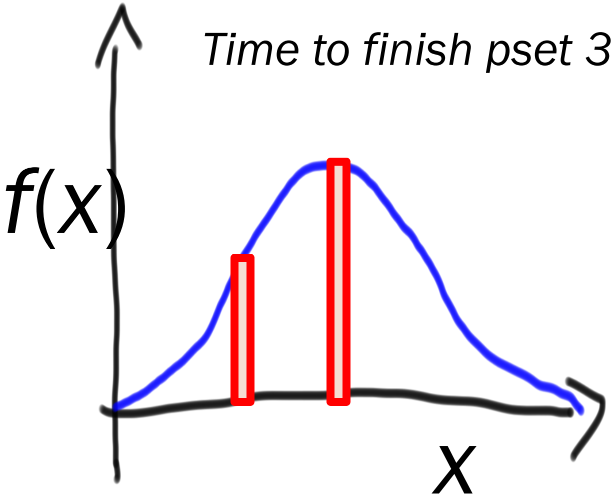
independence

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

Relative Probability of Continuous Variables

X = time to finish pset 3

$X \sim N(\mu = 10, \sigma^2 = 2)$



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

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

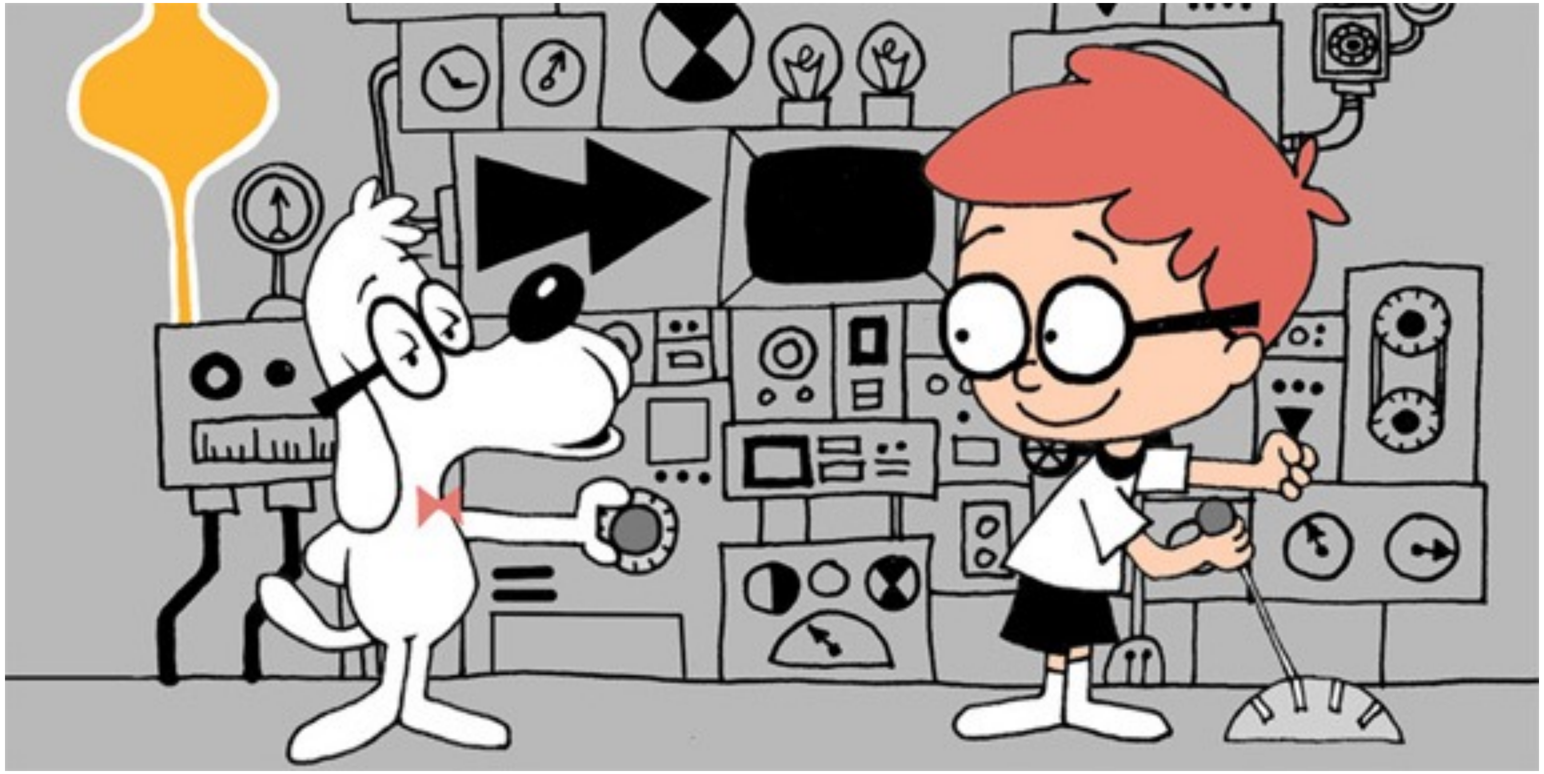
$$= \frac{f(X=10)}{f(X=5)}$$

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

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

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

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 order n objects such that k are indistinct and (n-k) are indistinct

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

Called the binomial coefficient because of something from Algebra

Multinomial coefficient

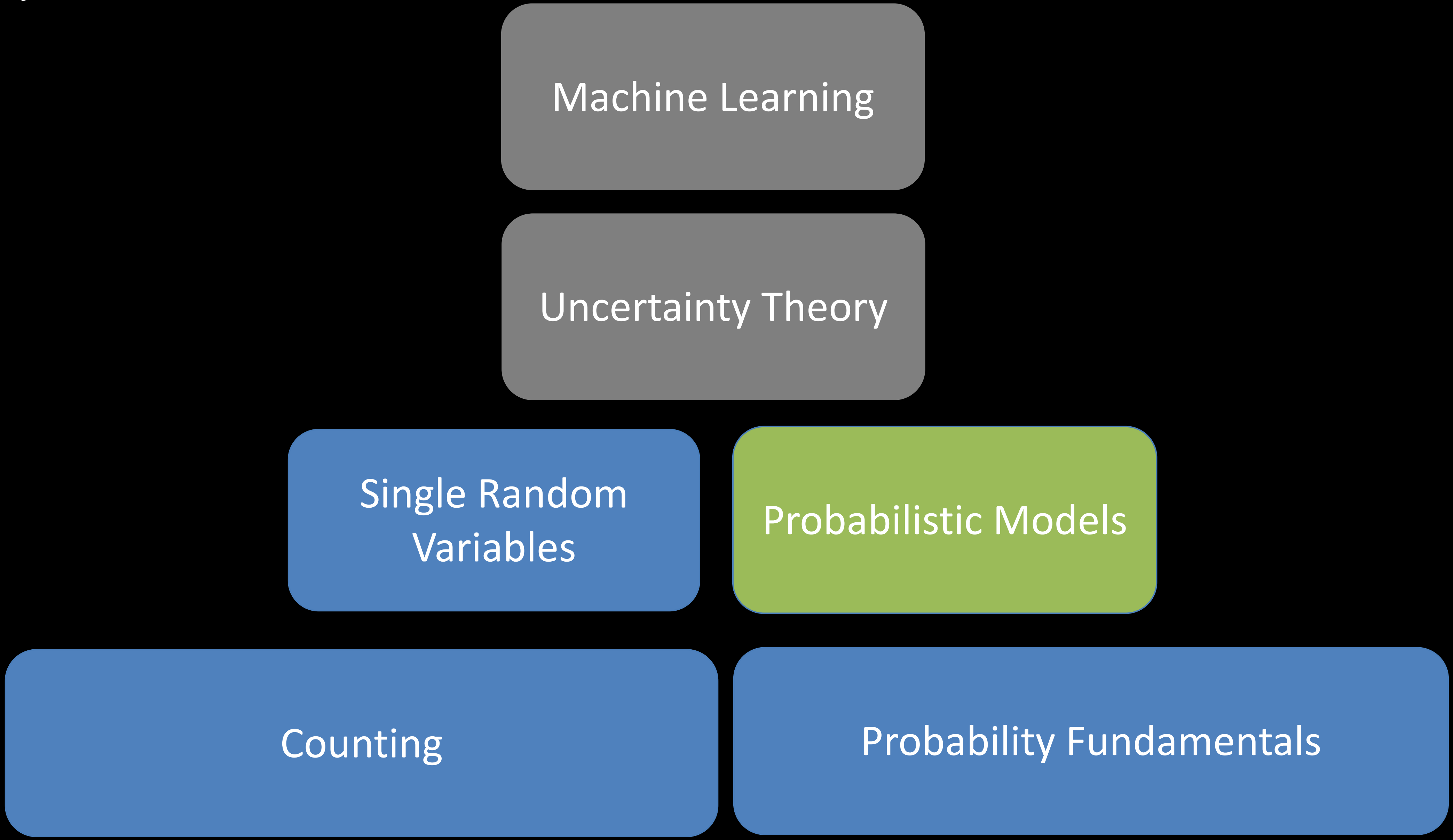
How many ways are there to order n objects such that n_1 are indistinct, n_2 are indistinct etc.

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

Multinomials generalize Binomials for counting.

End of review

CS109


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

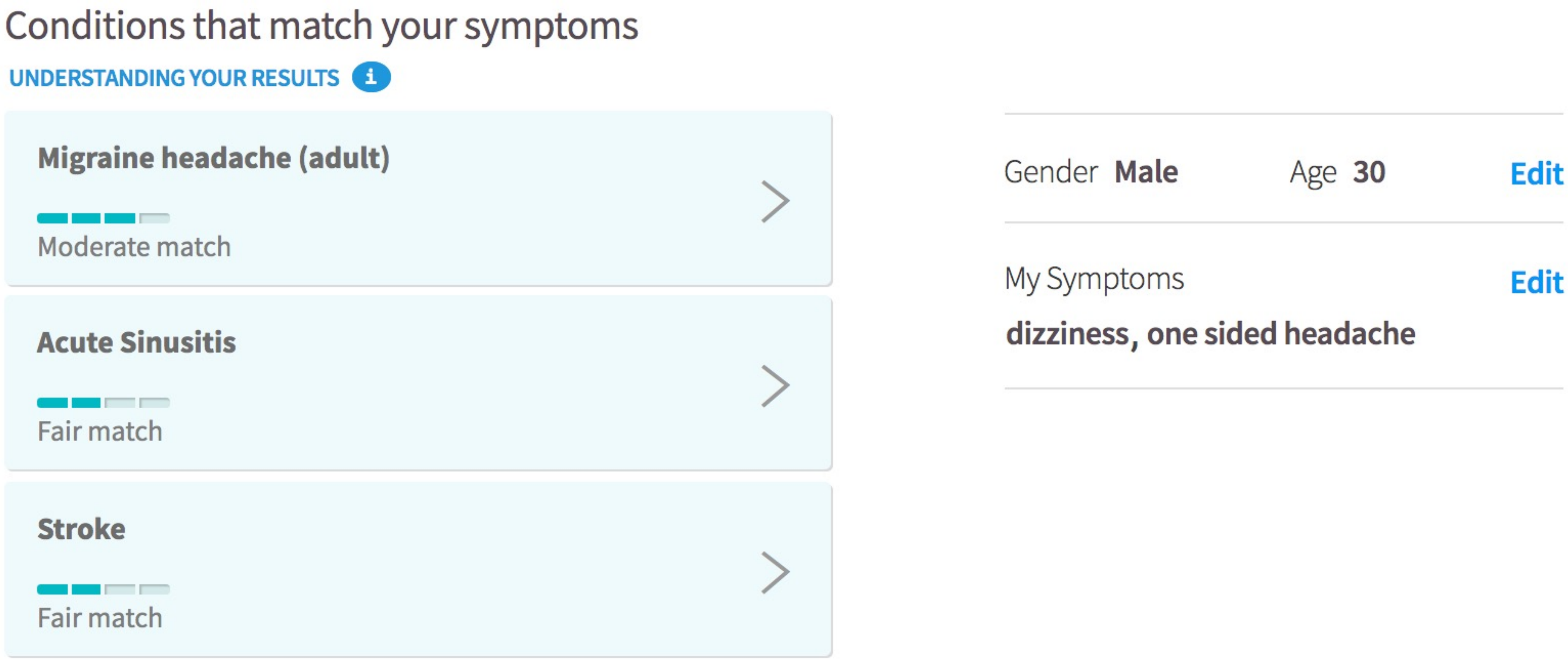
Discrete Probabilistic Models

The world is full of interesting probability problems

Have multiple random variables interacting with one another

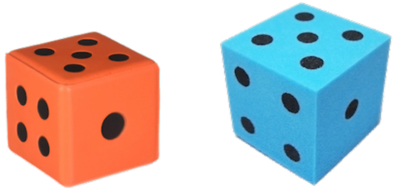
Multiple Random Variables. Start of Digital Revolution

Multiple Random Variables. Start of Digital Revolution



Joint probability mass functions

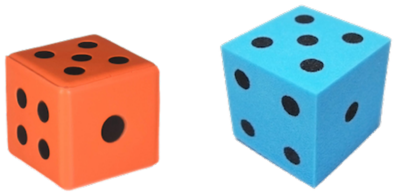
Roll two 6-sided dice, yielding values X and Y.



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.

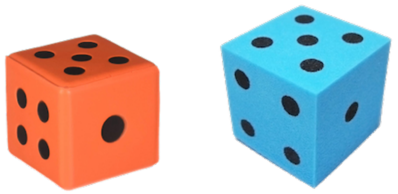


X	$P(X = 1)$	$P(X = k)$
random variable	probability of an event	probability mass function

X, Y	$P(X = 1 \text{ and } Y = 6)$ $P(X = 1, Y = 6)$ recall: the comma	$P(X = a, Y = b)$
random variables	probability of the intersection of two events	joint probability mass function

Two dice

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



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

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

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

$P(X = 4, Y = 3)$

Probability table

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

Dating at Stanford. 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	0.01
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.

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

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

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

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

Welcome the marginal

Marginal Distribution

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

$$P(X = a, Y = b)$$

The marginal distributions of the joint PMF are defined as:

$$P(X = a) = \sum_y P(X = a, Y = y)$$

$$P(Y = b) = \sum_x P(X = x, Y = b)$$

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

What about 3 Random Variables?

D is disease, ***S*** is can smell, ***F*** is fever status

$D = 0$

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

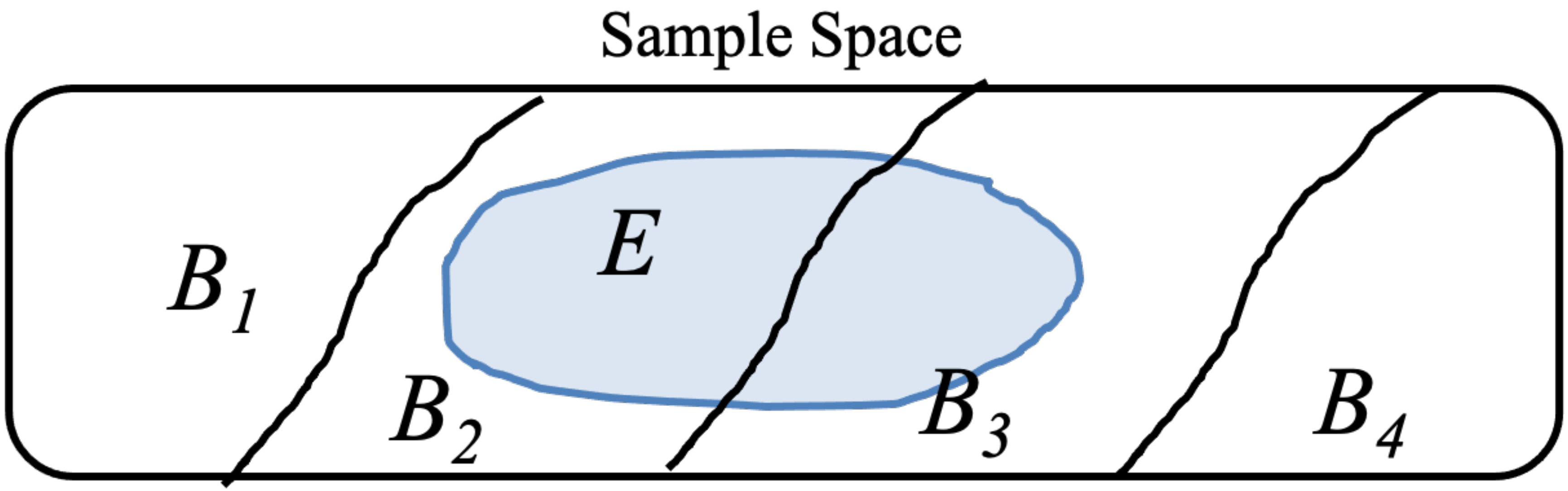
$D = 1$

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

$$P(D = 1) = \sum_f \sum_s P(D = 1, F = f, S = s)$$

Marginal Distribution. Law of Total Probability for RVs

$$P(X = a) = \sum_y P(X = a, Y = y)$$



Why is that called the marginal?

Key limitation of the joint: it is too big

What about 10 Random Variables?

Imagine you have **10 discrete** RVs which can each take on **5 values**

unique assignments $= 5^{10}$

10 million entries in your joint table.

So, we are going to need models …

… **probabilistic models** …

Sometimes the structure of the variables suggests a more efficient representation

Roll 100 dice.

X_1 = How many 1s?

X_2 = How many 2s?

X_3 = How many 3s?

X_4 = How many 4s?

X_5 = How many 5s?

X_6 = How many 6s?

How big is the joint table?

Multinomial RV

Making Binomial Bigger.

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

Binomial # of ways of ordering the successes

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

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

Binomial # of ways of ordering the successes

# of ways of ordering the fails

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

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

Case 1 # of ways and probabilities.

Case 2 # of ways and probabilities

Generalized Binomial

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

Case 1 # of ways and probabilities.

Case 2 # of ways and probabilities

$$\binom{n}{k_1} p_1^{k_1} \binom{n-k_1}{k_2} p_2^{k_2} \binom{n-k_2-k_1}{k_3} p_3^{k_3}$$

Case 1 # of ways and probabilities.

Case 2 # of ways and probabilities

Case 3 # of ways and probabilities

Notation Detail:

$$\binom{n}{k_1}\binom{n-k_1}{k_2}\binom{n-k_2-k_1}{k_3} = \binom{n}{k_1, k_2, k_3}$$

Fine Print

$$k_1 + k_2 + k_3 = n$$
$$p_1 + p_2 + p_3 = 1$$

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?

$$\binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$$

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

$$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m) =$$

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

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

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

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

Probability of each ordering is equal + mutually exclusive

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

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

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

Multinomial # of ways of ordering the outcomes

Probability of each ordering is equal + mutually exclusive

Hello dice rolls, my old friends

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

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

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

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

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

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(word = "the") $>$ P(word = "pokemon")
- P(word = "Stanford") $>$ P(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

It's a Multinomial!

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

Probability of seeing this document | spam

The probability of a word in spam email being viagra

Stanford University

Who wrote the federalist papers?

Old and New Analysis

Authorship of the Federalist Papers

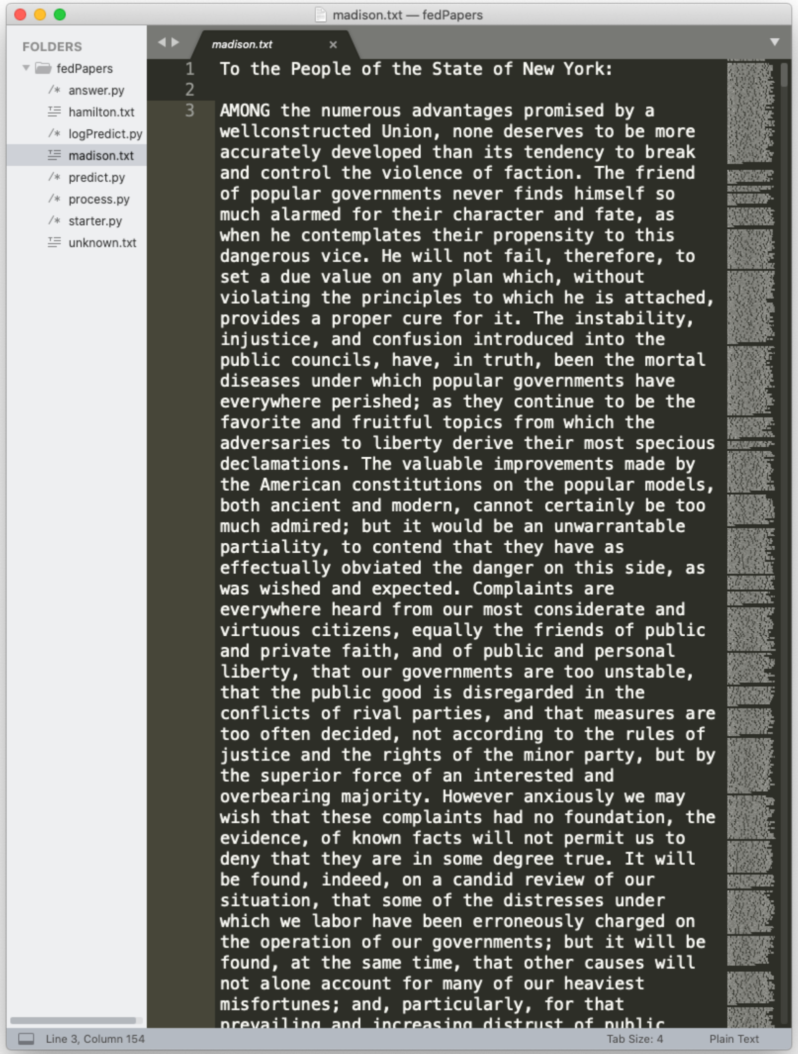
- 85 essays advocating ratification of the US constitution
- Written under the pseudonym “Publius” (really, Alexander Hamilton, James Madison, John Jay)

Who wrote which essays?

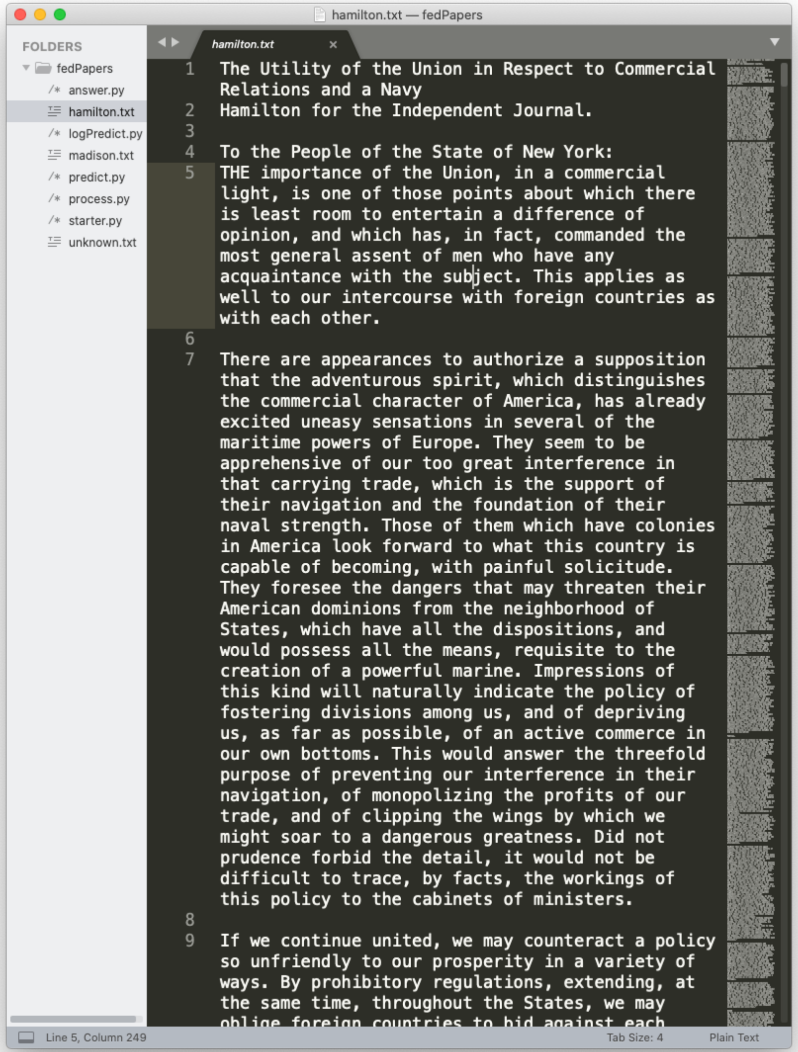
- Analyze probability of words in each essay and compare against word distributions from known writings of three authors

Who wrote Federalist Paper 53?

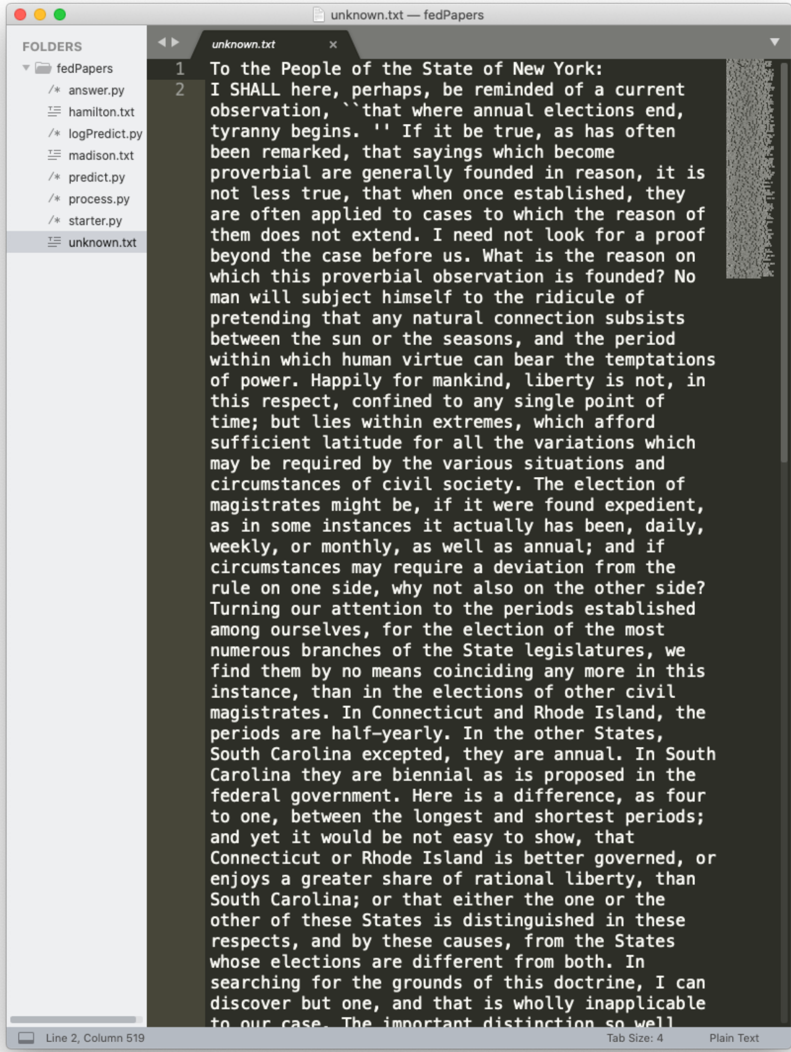
madison.txt



hamilton.txt



unknown.txt



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

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

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

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

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

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

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

(Detail) The power of Log

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

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

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

Numerically unstable.

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

$h_i \leq 1$

$m_i \leq 1$

The numerator and denominator will almost vanish!

Instead, take a log!

$$\sum_i c_i \log(h_i) - c_i \log(m_i) = -1353.6.$$ Madison?

If positive, then ratio on left is greater than 1.

Fin!