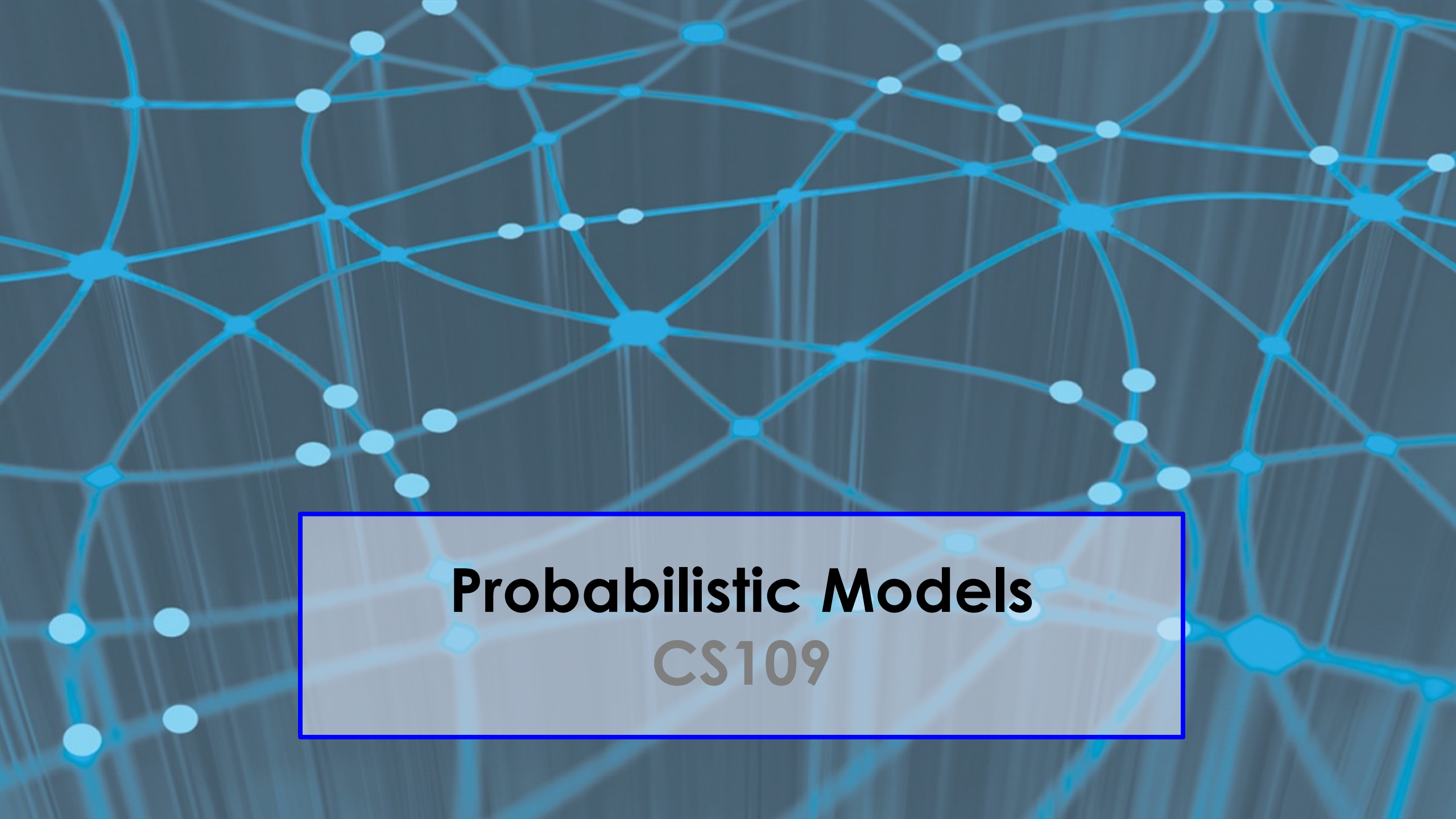
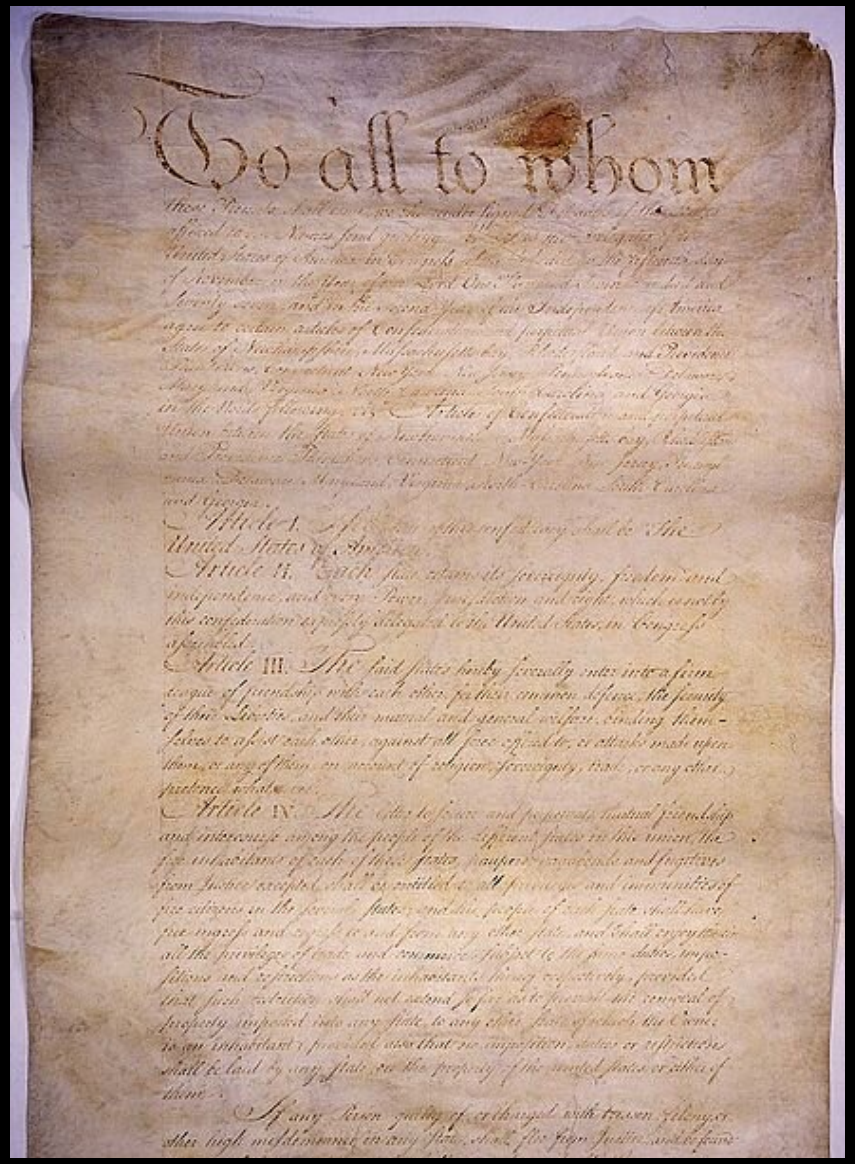
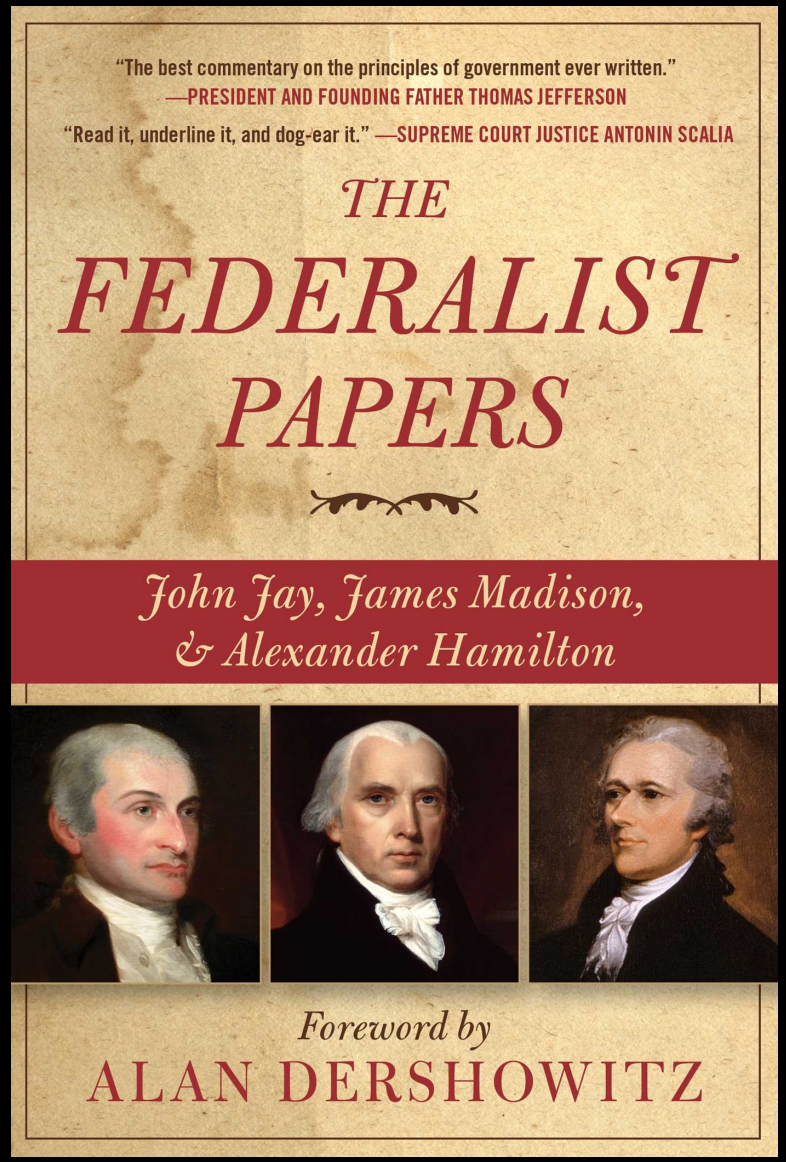


Probabilistic Models

CS109





Who Wrote The Federalist Papers?

Review

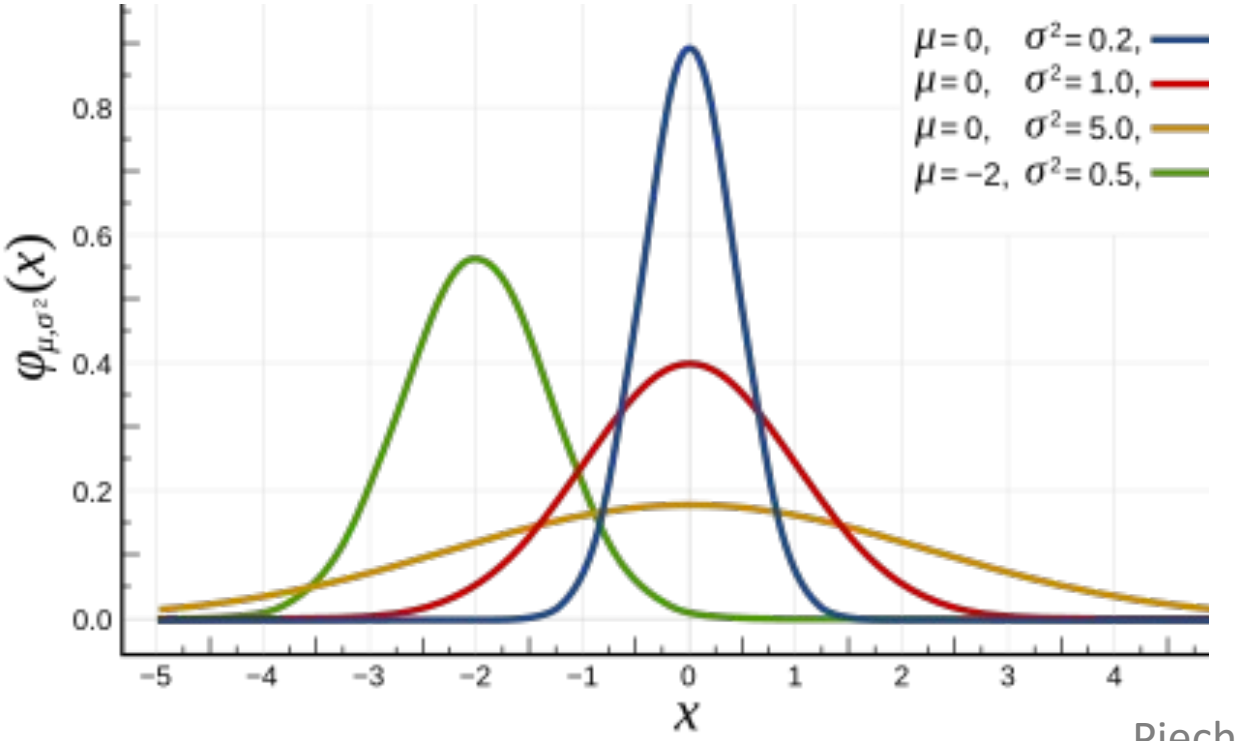
Normal (Gaussian) Random Variable

Support:
$(-\infty, \infty)$

mean

variance

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



Normal PDF & CDF

PDF:

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

CDF:

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

Where Does The Normal CDF Come From?

$$\text{Let } X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow Y = aX + b \Rightarrow Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

is also Normal

What linear transform of *X* would get us to *Z*?

$$Z = \frac{X - \mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma} \qquad a = \frac{1}{\sigma} \quad b = -\frac{\mu}{\sigma}$$

If we plug in these values for *a* and *b*, we get the standard normal:

$$\begin{aligned} Z &\sim \mathcal{N}(a\mu + b, a^2\sigma^2) \\ &\sim \mathcal{N}(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}) \\ &\sim \mathcal{N}(0, 1) \end{aligned}$$

$$Z = \frac{X - \mu}{\sigma}$$

We Are Computer Scientists! Compute The CDF With Code

Every modern programming language has phi stored in a library:

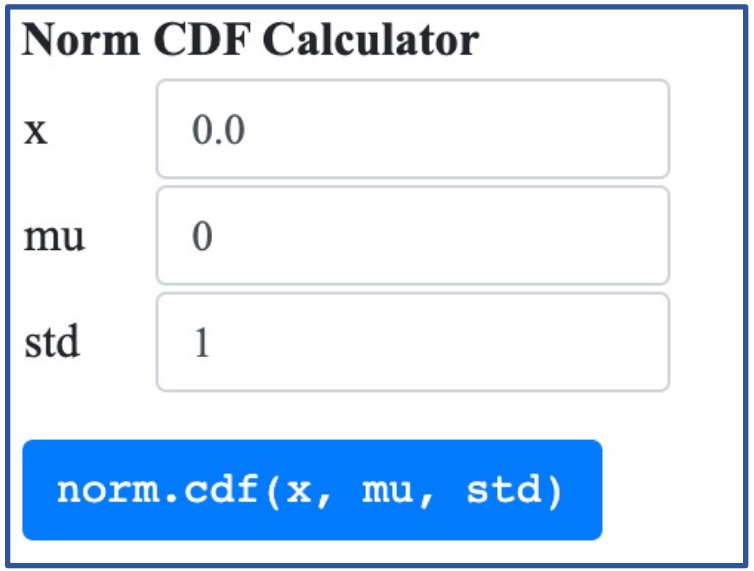
```
from scipy import stats

stats.norm.cdf(x, mean, std)
```

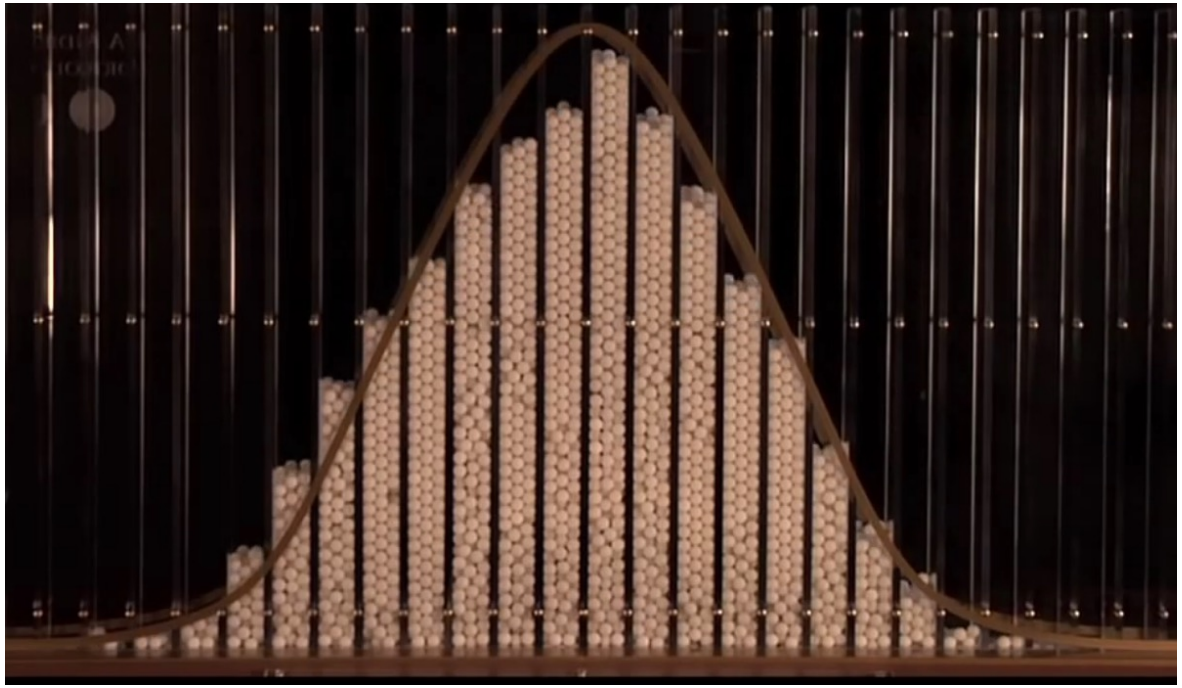
not variance!!!

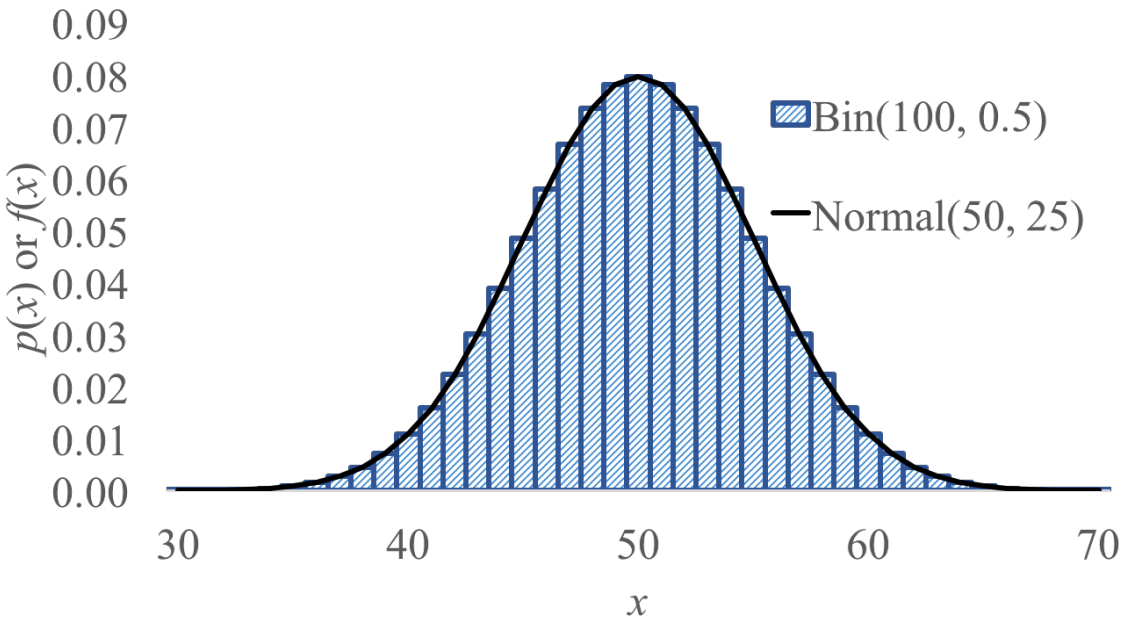
$$= P(X < x) \text{ where } X \sim \mathcal{N}(\mu, \sigma^2)$$

The course reader also has a calculator:



Normal Approximates Binomial, With Moderate p





The shapes are the same!

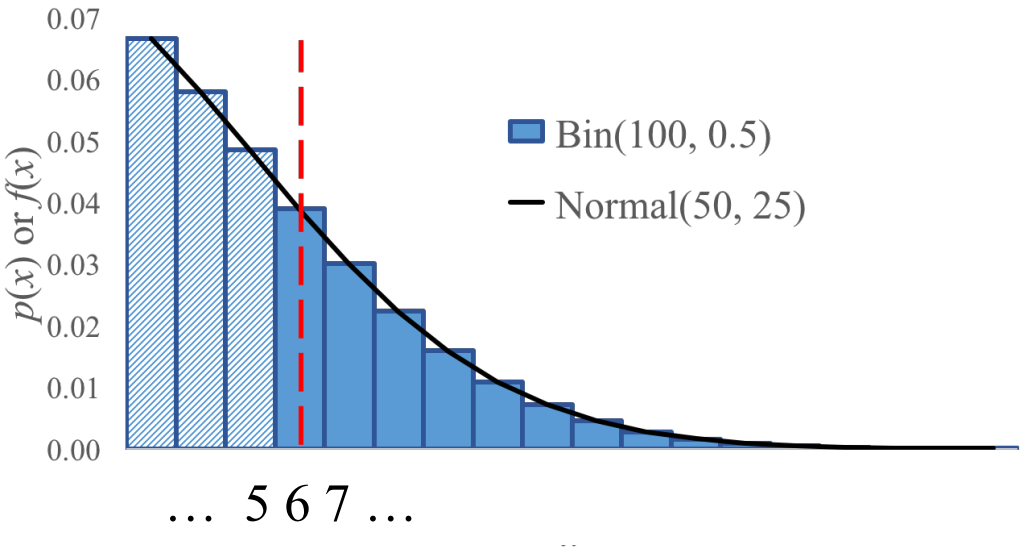
Just set the normal's μ, σ^2 to be the mean and variance of the binomial.

Continuity Correction

$Y \sim \mathcal{N}(np, np(1-p))$ approximates $X \sim \text{Bin}(n, p)$.

How do we approximate the following probabilities?

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



Guide To The Normal

1. The normal shows up in situations where you know mean and variance, and nothing else.

2. Treat the normal like any continuous RV: get your answer in terms of P($X < x$) so you can use the CDF.

3. Apply the CDF formula:

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

Guide To The Normal

1. The normal shows up in situations where you know mean and variance, and nothing else.

2. Treat the normal like any continuous RV: get your answer in terms of P($X < x$) so you can use the CDF.

3. Apply the CDF formula:

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

For approximating binomials:

1. Proceed as if it's a normal binomial problem -- find *n* and *p*, and decide what probability you want: P([???])

2. Then, fit your normal distribution to match your binomial:

$$Y \sim \mathcal{N}(np, np(1-p))$$

3. Translate your "probability you want" into terms of the normal distribution, using continuity correction.

Boss Battle Normal Distribution Problem


福
双11
11.11
光棍节
SINGLE'S DAY
SALE

How Many Servers Is Enough?

At the busiest minute of the shopping rush, your website receives R pings:

$$R \sim N(\mu = 10^6, \sigma = 10^4)$$

To anticipate the rush, you plan to buy *N* servers. Each server can handle 100 pings per minute, but if it receives any more, it will drop customers.

What is the smallest value of *N* such that P(no drop) > 0.9999?

How Many Servers Is Enough?

At the busiest minute of the shopping rush, your website receives R pings:

$$R \sim N(\mu = 10^6, \sigma = 10^4)$$

To anticipate the rush, you plan to buy *N* servers. Each server can handle 100 pings per minute, but if it receives any more, it will drop customers.

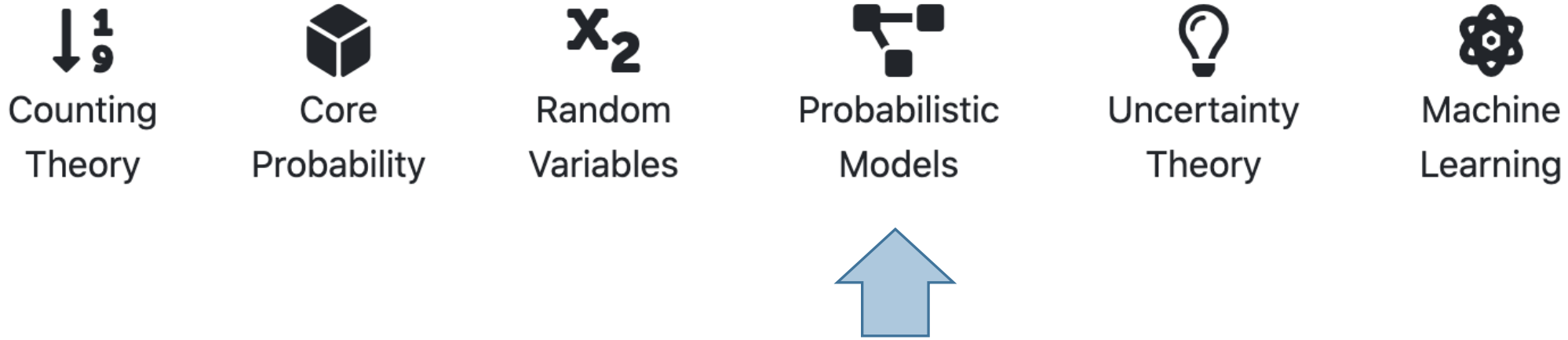
What is the smallest value of *N* such that P(no drop) > 0.9999?

Helpful fun fact -- phi has an inverse function:

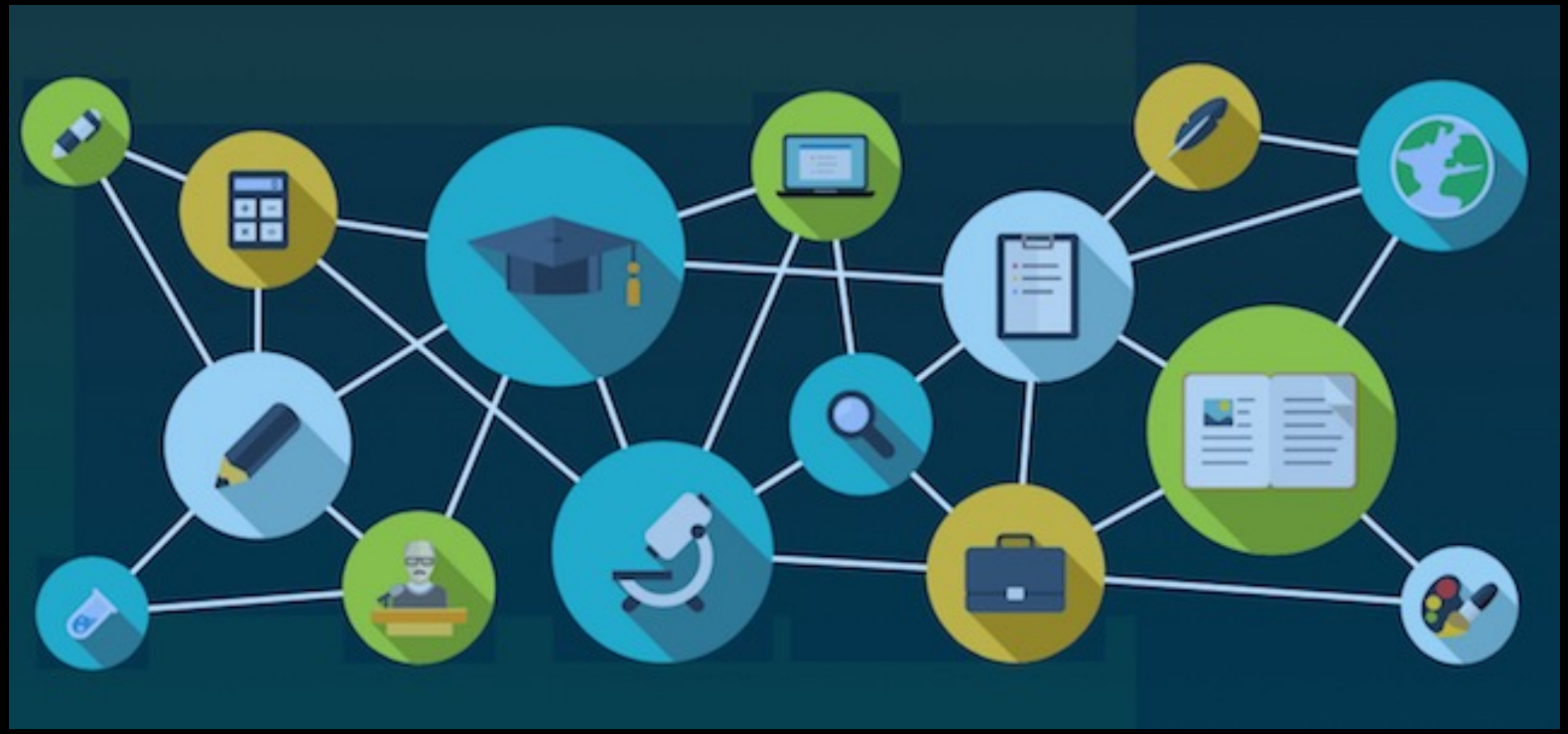
$$x = \phi^{-1}(\phi(x))$$

End Review

Where are we in CS109?

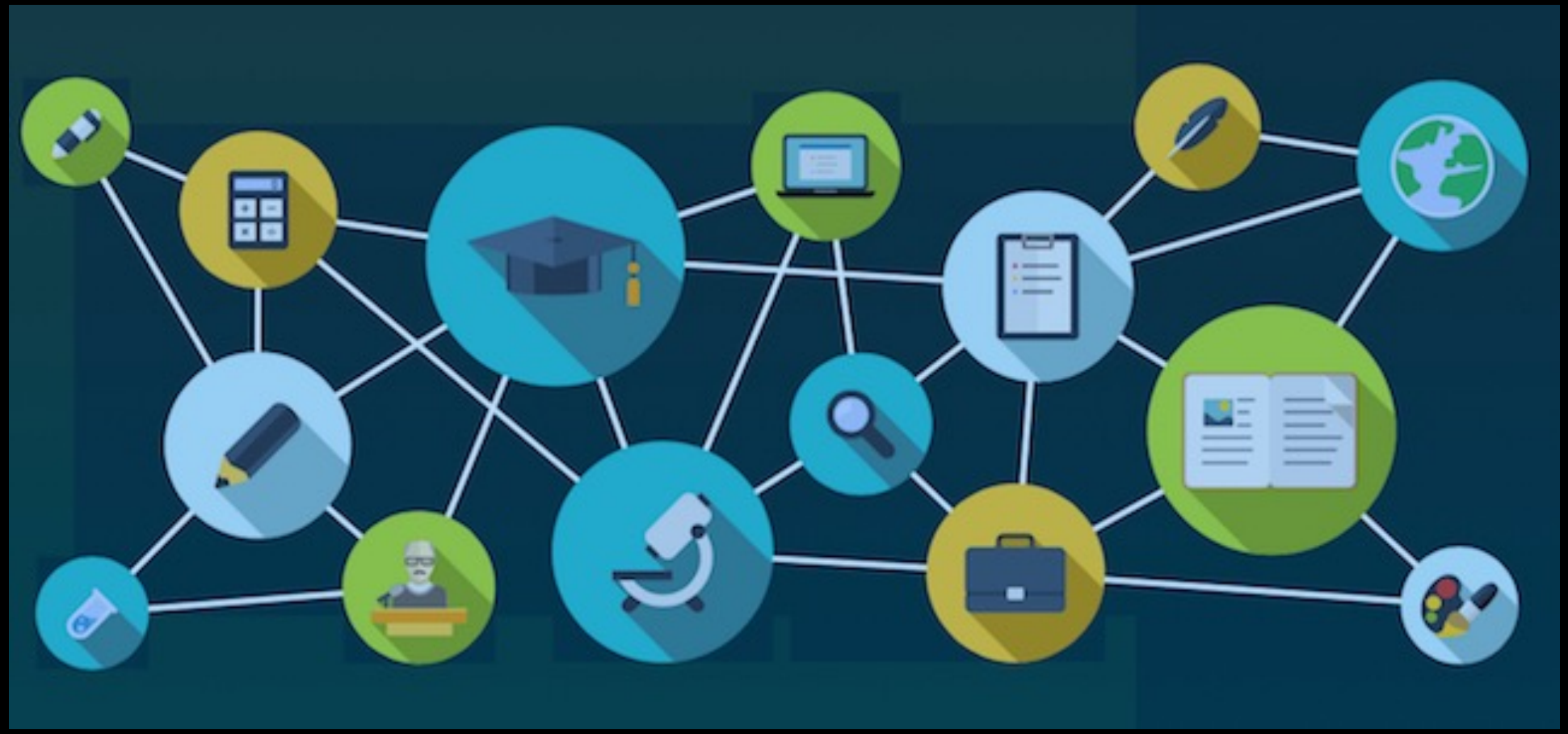


What Are We Missing?



The world is full of interesting probability problems…

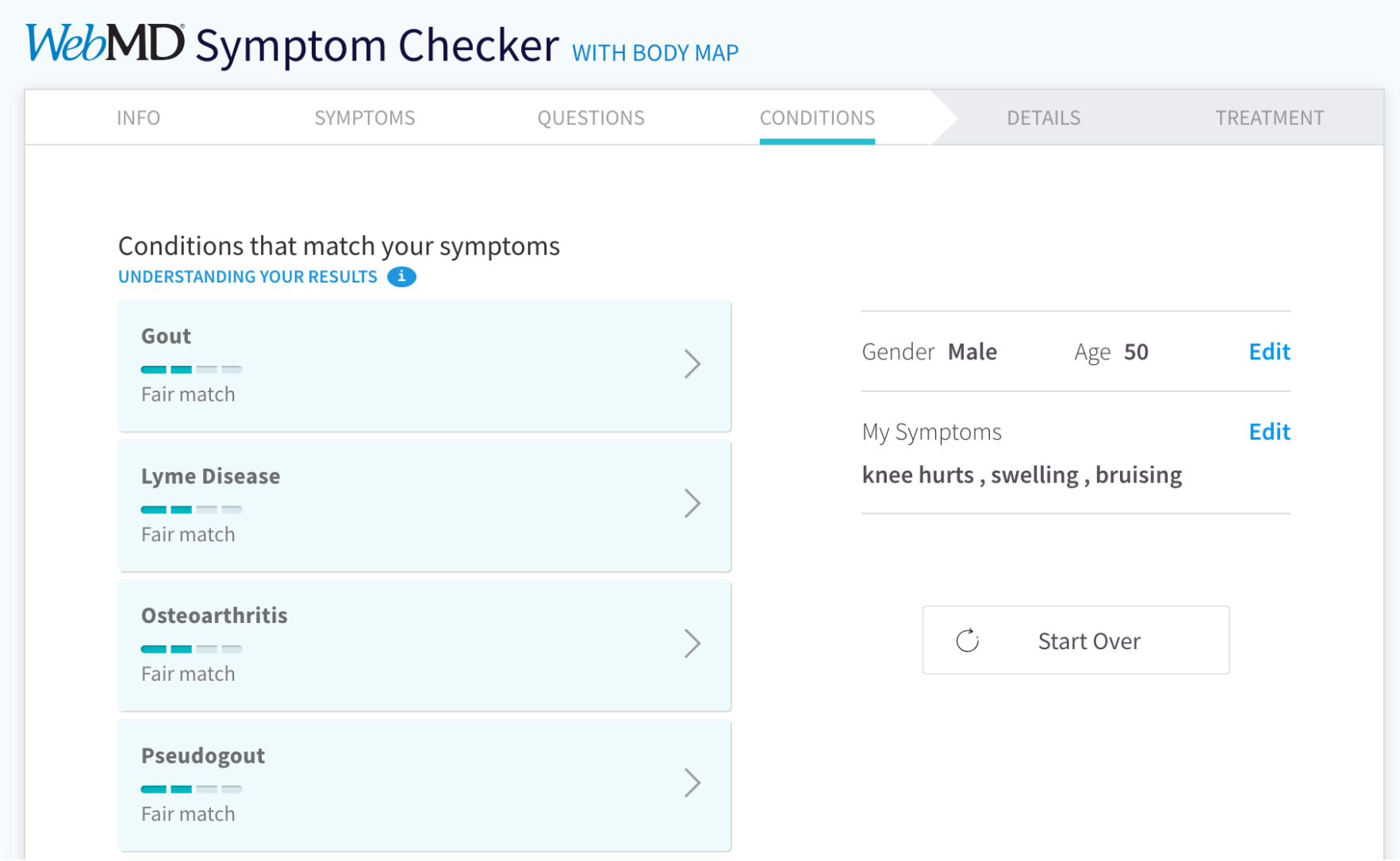
What Are We Missing?



The world is full of interesting probability problems…

…and many of them involve *multiple* random variables, being random *together*

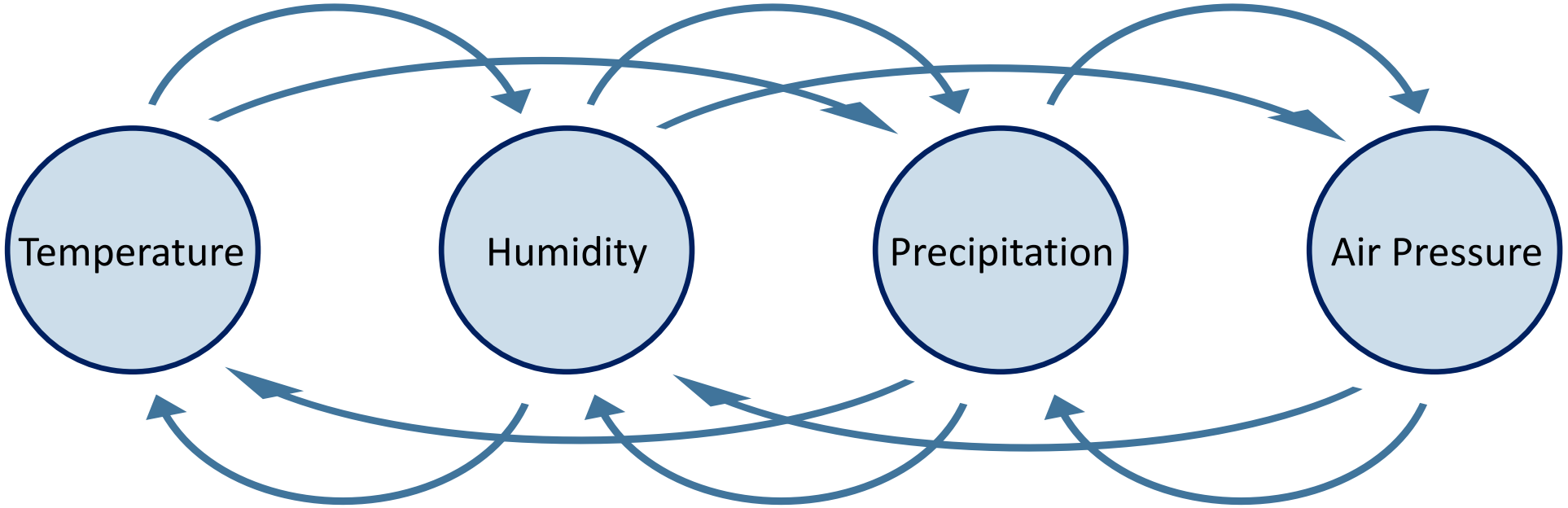
How Do We Model Multiple Random Variables Together?



https://symptoms.webmd.com/

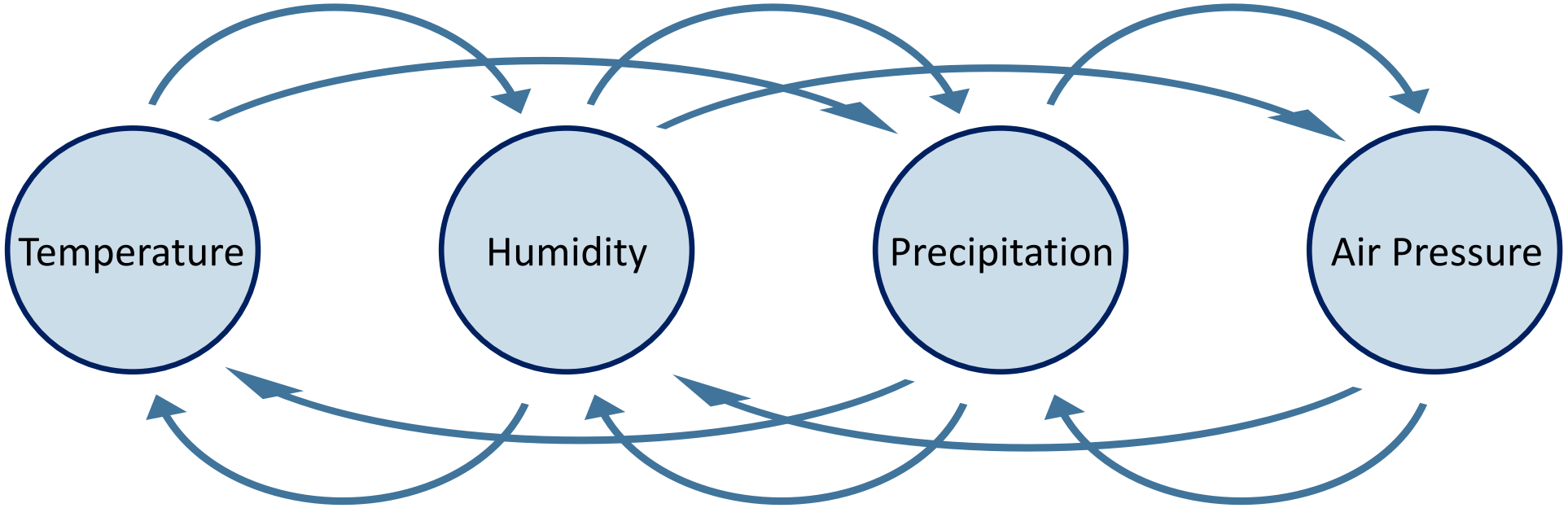
How Do We Model Multiple Random Variables Together?

Often, all the random variables involved are not independent of each other.



How Do We Model Multiple Random Variables Together?

Often, all the random variables involved are not independent of each other.



So we can't just have a single distribution for each random variable —
we need a way to talk about all the random variables at the same time.

The "Joint" Distribution of Multiple Random Variables

For *discrete* random variables X and Y, we have a **joint probability mass function**:

$$P(X = x, Y = y)$$

The joint is the "and" between an assignment to X, and an assignment to Y

The same as P(A and B) for events A and B!

The "Joint" Distribution of Multiple Random Variables

For *discrete* random variables X and Y, we have a **joint probability mass function**:

$$P(X = x, Y = y)$$

$X = 2, Y = 4$

0.5134 …

The joint is the "and" between an assignment to X, and an assignment to Y

The same as P(A and B) for events A and B!

The "Joint" Distribution of Multiple Random Variables

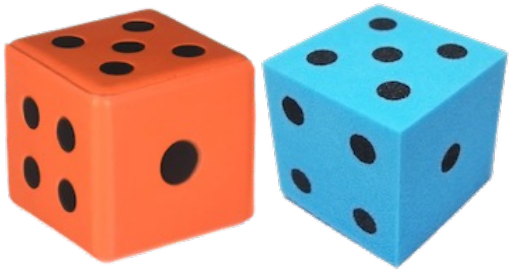
For *discrete* random variables X and Y, we have a **joint probability mass function**:

$$P(X = x, Y = y)$$

For *continuous* random variables, we have a **joint probability density function**:

$$f(X = x, Y = y)$$

Example Joint PMF: Two Dice



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

X

random variable

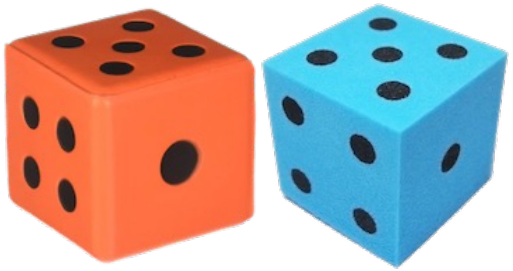
$P(X = 1)$

probability of an event

$P(X = k)$

probability mass function

Example Joint PMF: Two Dice



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

$P(X = 1, Y = 6)$

probability of the intersection of two events

$P(X = x, Y = y)$

joint probability mass function

Example Joint PMF: Two Dice

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

What is $P(X = x, Y = y)$?

Example Joint PMF: Two Dice

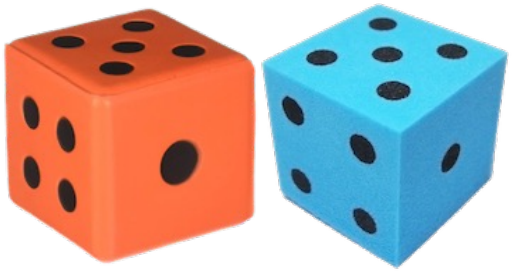
Roll two 6-sided dice, yielding values X and Y.

What is $P(X = x, Y = y)$?

$$P(X = x, Y = y) = \frac{1}{36}$$

$$(x, y) \in \{(1,1), \ldots, (6,6)\}$$

Example Joint PMF: Two Dice



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

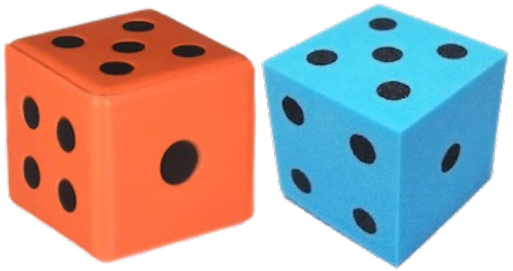
What is $P(X = x, Y = y)$?

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

$$P(X = x, Y = y) = \frac{1}{36}$$

$$(x, y) \in \{(1,1), \ldots, (6,6)\}$$

Example Joint PMF: Two Dice



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

What is $P(X = x, Y = y)$?

$$P(X = x, Y = y) = \frac{1}{36}$$

$$(x, y) \in \{(1,1), \ldots, (6,6)\}$$

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

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

This is a **joint probability table**: it contains the probabilities of all possible outcomes for a set of discrete random variables

More Interesting Data: Dating at Stanford

	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
Grad Student	0.06	0.09	0.04

Fun Facts About Joint Tables

	Single	Relationship	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
Grad	0.06	0.09	0.04

Fact 1: Each cell in the table is one outcome; all cells are mutually exclusive.

Fun Facts About Joint Tables

	Single	Relationship	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
Grad	0.06	0.09	0.04

Fact 1: Each cell in the table is one outcome; all cells are mutually exclusive.

Fact 2: The sum over the whole table is 1.

Let X be dating status, and Y be year.

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

Fun Facts About Joint Tables

	Single	Relationship	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
Grad	0.06	0.09	0.04

Fact 1: Each cell in the table is one outcome; all cells are mutually exclusive.

Fact 2: The sum over the whole table is 1.

Let X be dating status, and Y be year.

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

Fact 3: A joint distribution is *complete information*.
It can be used to answer *any* probability question about the RVs.

The Joint Is Complete Information

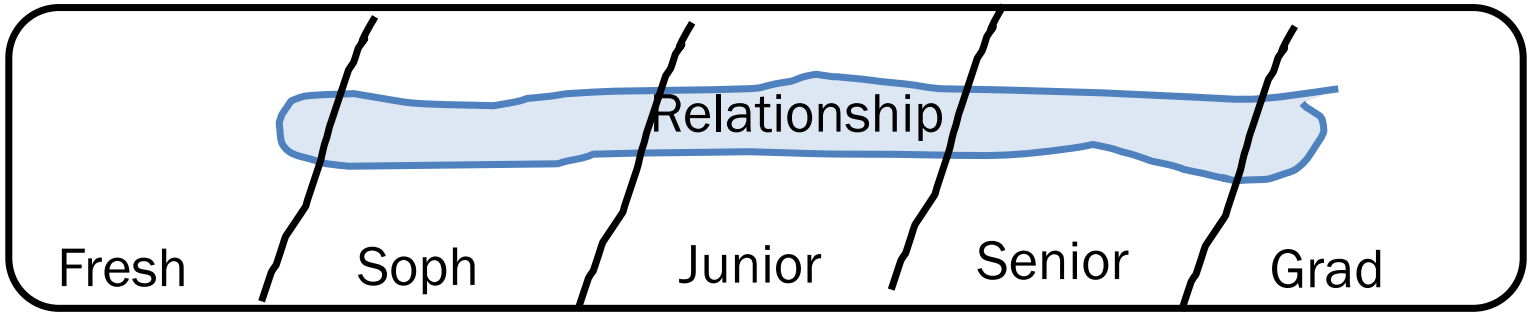
	Single	Relationship	Complicated
Freshman	0.13	0.08	0.02
Sophomore	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
Grad	0.06	0.09	0.04

From the table, what is **P(relationship)?**

The Joint Is Complete Information

	Single	Relationship	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
Grad	0.06	0.09	0.04

From the table, what is **P(relationship)?**



The Joint Is Complete Information

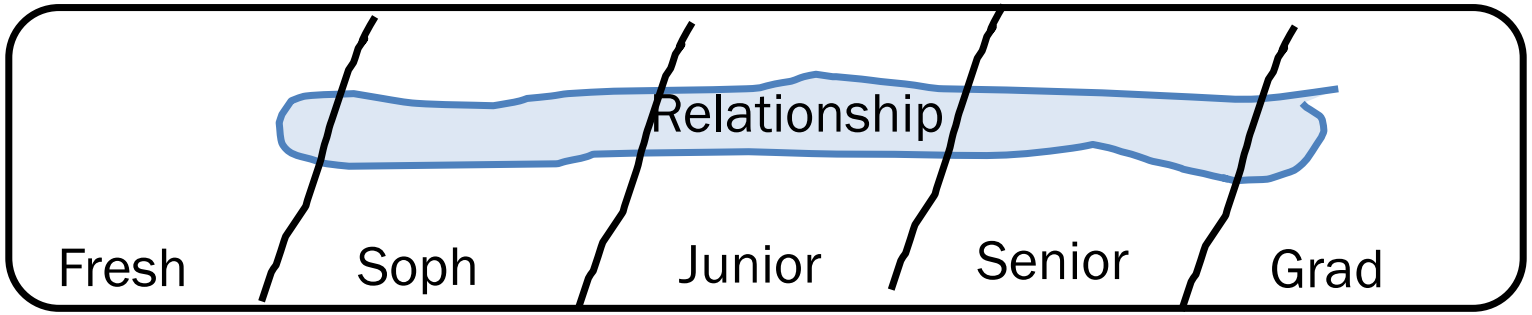
	Single	Relationship	Complicated
Freshman	0.13	0.08	0.02
Sophomore	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
Grad	0.06	0.09	0.04

From the table, what is **P(relationship)?**

Let X be dating status, and Y be year.

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

Law of Total Probability!



The Joint Is Complete Information

	Single	Relationship	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
Grad	0.06	0.09	0.04

From the table, what is **P(relationship)?**

Let X be dating status, and Y be year.

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

Law of Total Probability!

Summing probabilities in a joint distribution across all outcomes of one RV has a name.

We call this the **marginal probability** of X: $P(X = a) = \sum_{y} P(X = a, Y = y)$

Another Example Joint Table

What if we have 3 random variables?

D: do you have covid

S: can you smell

F: do you have a fever

Now the joint table is 3D!

$D = 0$

	$S = 0$	$S = 1$
$F = \text{none}$	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

Another Example Joint Table

What if we have 3 random variables?

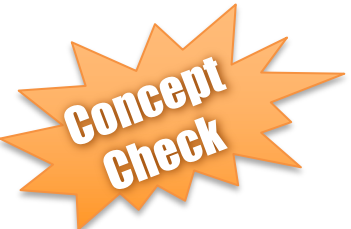
D: do you have covid

S: can you smell

F: do you have a fever

Now the joint table is 3D!

$$P(D = 1) = \ ?$$



$D = 0$

	$S = 0$	$S = 1$
$F = \text{none}$	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

Another Example Joint Table

What if we have 3 random variables?

D: do you have covid

S: can you smell

F: do you have a fever

Now the joint table is 3D!

$D = 0$

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

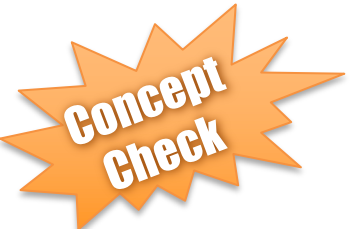
$D = 1$

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

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

$$= 0.051$$

With more RVs, calculate marginal probabilities by summing over all RVs except one.



We can do anything with a joint table!

But *should* we do *everything* with a joint table?

Joint Tables Get Bigger And Bigger

The joint table has to contain probabilities for *every possible joint outcome* across all the RVs.

For 2 dice, the size of the joint table was:

6 outcomes for X * 6 outcomes for Y = 36

What is the size of the joint table for n dice?

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

Joint Tables Get Bigger And Bigger

The joint table has to contain probabilities for *every possible joint outcome* across all the RVs.

For 2 dice, the size of the joint table was:

6 outcomes for X * 6 outcomes for Y = 36

What is the size of the joint table for n dice?

6^n

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

The joint table grows *exponentially* with the number of random variables…

Joint Tables Don't Scale

Joint tables grow in size exponentially with more random variables – yikes.

So, we prefer to represent joint distributions more efficiently: with equations!

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

$$P(X = x, Y = y) = \frac{1}{36}$$

$$(x, y) \in \{(1,1), \dots, (6,6)\}$$

Multinomial Distribution

Key Limitation Of The Binomial: Only Binary Outcomes

The binomial models the number of successes seen in a set of *n* trials.

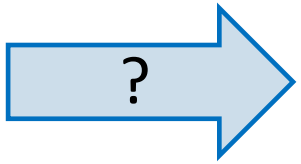
Every trial can only end in success or failure.

Key Limitation Of The Binomial: Only Binary Outcomes

The binomial models the number of successes seen in a set of *n* trials.

What about a set of *n* trials, where outcomes are not binary?

Every trial can only end in success or failure.



Imagine we 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?

Imagine we 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 would the joint table be?)

We Can Roll Dice Many Times Now!!

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

We Can Roll Dice Many Times Now!!

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 we got a 6

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

We Can Roll Dice Many Times Now!!

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 we got a 6

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

Can you figure out the PMF from first principles?

Remember how we found the Binomial PMF:

1) Any specific ordered outcome had probability $p^k(1-p)^{n-k}$
2) Then we had to multiply by the number of possible orderings

We Can Roll Dice Many Times Now!!

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 we got a 6

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

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

Need to account for possible orderings!

probability of rolling a six, 3 times

Let's Generalize The Binomial

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

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

Binomial coefficient: # ways to order the outcomes

Probability of each ordering of k successes

Let's Generalize The Binomial

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

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

Binomial coefficient: # ways to order the outcomes

Probability of each ordering of k successes

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

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

Multinomial coefficient: # ways to order the outcomes

Probability of each ordering

Binomial vs. Multinomial Coefficient

Binomial coefficient

How many ways are there to order *n* objects, such that *k* are indistinct, and (*n*-*k*) are indistinct?

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

Multinomial coefficient

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

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

MISSISSIPPI

Multinomial: Generalized Binomial

For experiments with n trials, where each trial results in one of m outcomes, let $P(\text{outcome } i) = p_i$, and let X_i be the number of trials with outcome i.

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

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

Multinomial coefficient: # ways to order the joint outcome

Probability of one joint outcome, ordered

Multinomial: Generalized Binomial

For experiments with *n* trials, where each trial results in one of m outcomes, let $P(\text{outcome } i) = p_i$, and let X_i be the number of trials with outcome i.

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

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

Multinomial coefficient: # ways to order the joint outcome

Probability of one joint outcome, ordered

The counts of each of the possible outcomes sum to *n*: $\sum_{i=1}^{m} c_i = n$

Multinomial: Generalized Binomial

For experiments with *n* trials, where each trial results in one of m outcomes, let $P(\text{outcome } i) = p_i$, and let X_i be the number of trials with outcome i.

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

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

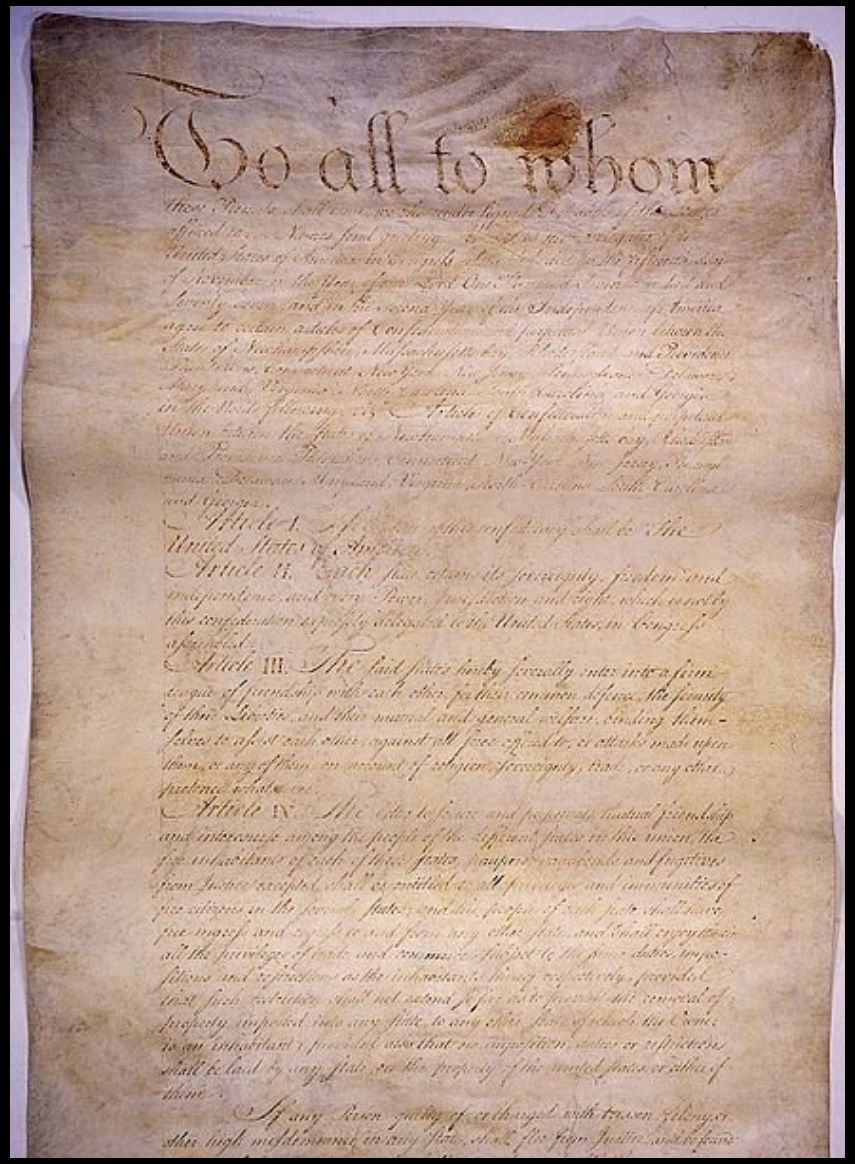
Multinomial coefficient: # ways to order the joint outcome

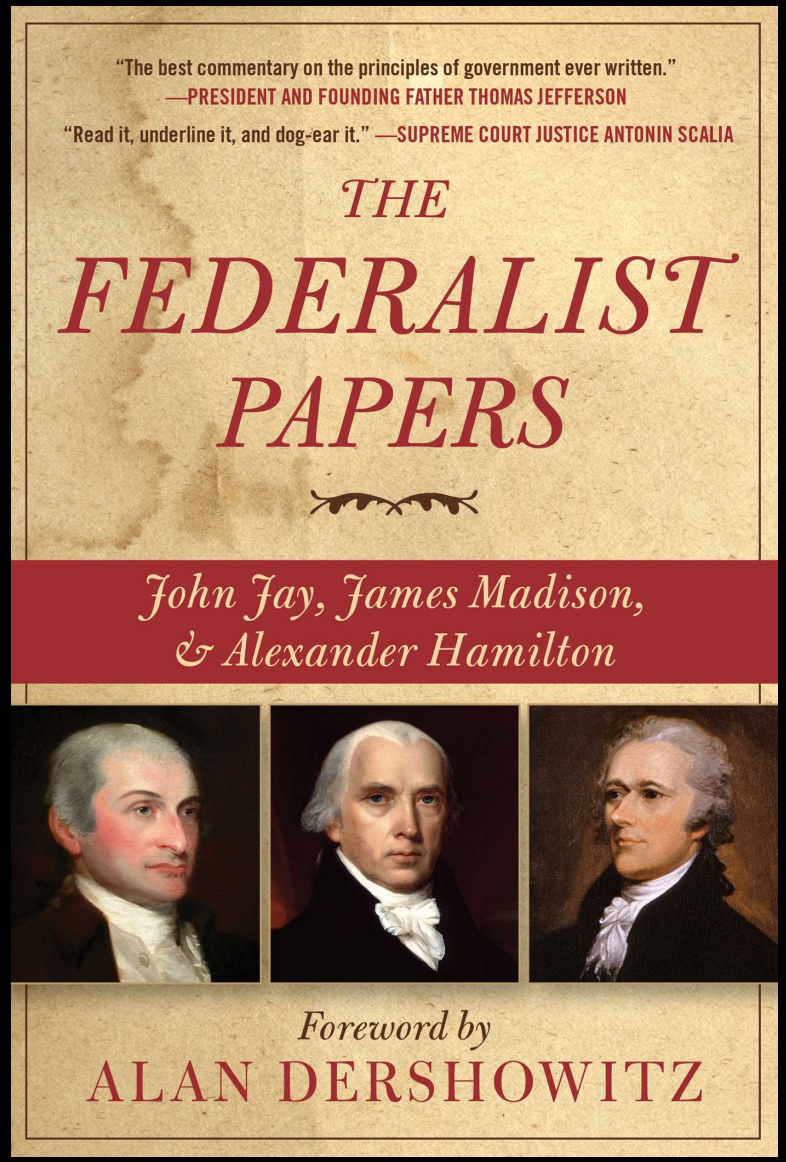
Probability of one joint outcome, ordered

The counts of each of the possible outcomes sum to *n*: $\sum_{i=1}^{m} c_i = n$

The probabilities of each of the possible outcomes sum to 1: $\sum_{i=1}^{m} p_i = 1$

The multinomial is *one* example of a joint distribution represented as an equation, instead of a table





Who Wrote The Federalist Papers?

The Federalist Papers

- 85 essays from the 1780s, advocating for ratification of the US constitution
- Written under the pseudonym Publius (really, a combo of Alexander **Hamilton**, James **Madison**, and John Jay)
- Years later, Hamilton and Madison *both* claimed to write essay #53…

The Federalist Papers

- 85 essays from the 1780s, advocating for ratification of the US constitution
- Written under the pseudonym Publius (really, a combo of Alexander **Hamilton**, James **Madison**, and John Jay)
- Years later, Hamilton and Madison *both* claimed to write essay #53…

Question: Who wrote essay 53?

The Federalist Papers

- 85 essays from the 1780s, advocating for ratification of the US constitution
- Written under the pseudonym Publius (really, a combo of Alexander **Hamilton**, James **Madison**, and John Jay)
- Years later, Hamilton and Madison *both* claimed to write essay #53…

Question: Who wrote essay 53?

Strategy: model the probabilities of words in the essay, and compare to models of probabilities from known writings of each author

Crash Course in Natural Language Processing

Probabilistic Text Analysis

Ignoring the order of words, documents are collections of words.

- Some words are more common than others: $P(\text{“the”}) > P(\text{“pokemon”})$
- Different people use words at different frequencies

Probabilistic Text Analysis

Ignoring the order of words, documents are collections of words.

- Some words are more common than others: $P(\text{“the”}) > P(\text{“pokemon”})$
- Different people use words at different frequencies

Idea: Probabilities of *counts* of words = Multinomial distribution

We will think of essays as a collection of outcomes when we roll a very large word-dice, representing many possible outcomes for which word is at each position.

How We Model Text As A Multinomial: Spam

(this way of modeling text was first used to detect spam emails)

Example text, with n = 18:

"Pay for Viagra with a credit-card. Viagra is great. So are credit-cards. Risk free Viagra. Click for free."

How We Model Text As A Multinomial: Spam

(this way of modeling text was first used to detect spam emails)

Example text, with n = 18:

"Pay for Viagra with a credit-card. Viagra is great. So are credit-cards. Risk free Viagra. Click for free."

$$P\left(\text{this text} \mid \text{spam}\right)$$

How We Model Text As A Multinomial: Spam

(this way of modeling text was first used to detect spam emails)

Example text, with n = 18:

"Pay for Viagra with a credit-card. Viagra is great. So are credit-cards. Risk free Viagra. Click for free."

$$P\left(\begin{array}{l}\text{Viagra = 2}\\ \text{Free = 2}\\ \text{Risk = 1}\\ \text{Credit-card: 2}\\ \ldots\\ \text{For = 2}\end{array}\right)$$

Probability of seeing this text, if spam

How We Model Text As A Multinomial: Spam

(this way of modeling text was first used to detect spam emails)

Example text, with n = 18:

"Pay for Viagra with a credit-card. Viagra is great. So are credit-cards. Risk free Viagra. Click for free."

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

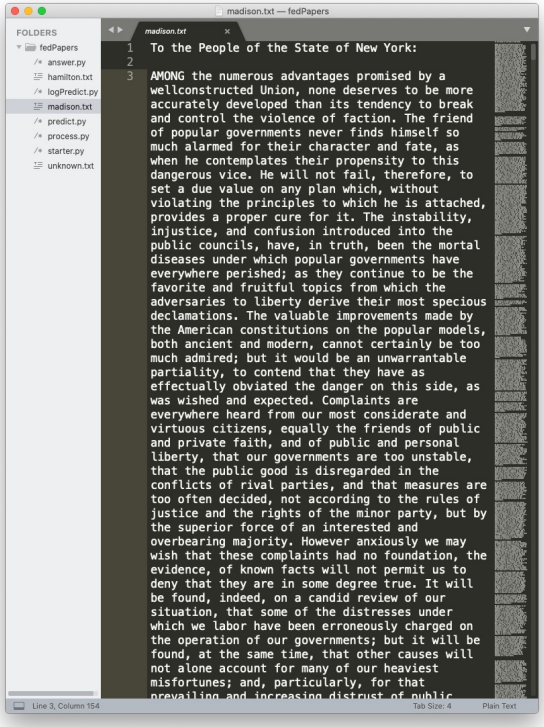
Multinomial PMF:

Probability of seeing this text, if spam

The probability of a word in spam email being viagra

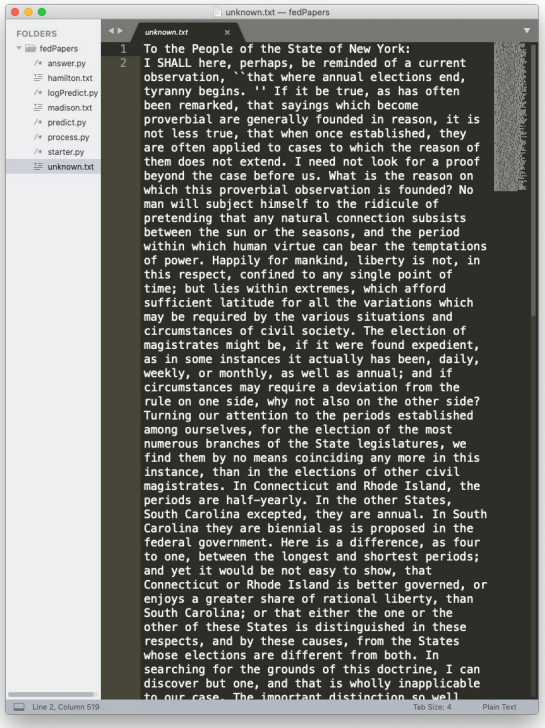
Who wrote Federalist Paper 53?

madison_example.txt



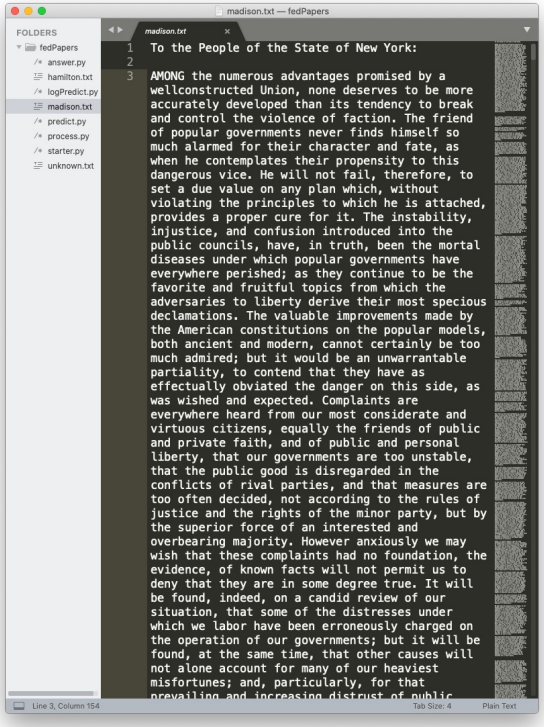
hamilton_example.txt

essay53.txt



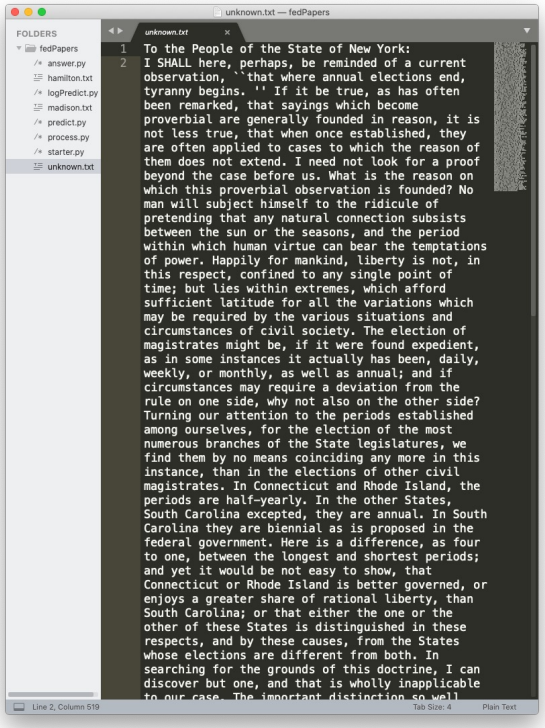
Who wrote Federalist Paper 53?

madison_example.txt



hamilton_example.txt

essay53.txt



We can fit a multinomial for Madison…

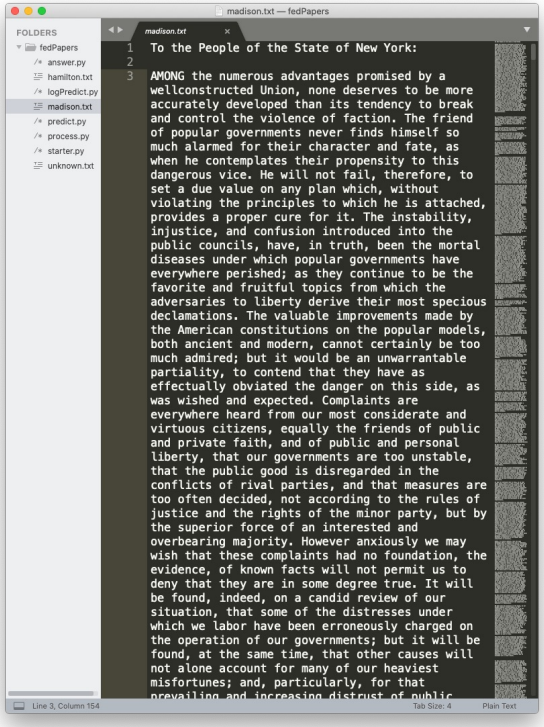
$$P(D|M)$$

…and fit another multinomial for Hamilton…

$$P(D|H)$$

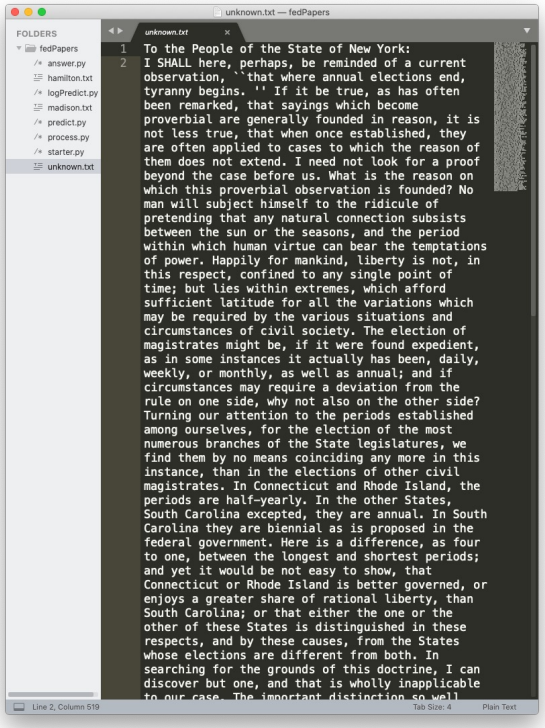
Who wrote Federalist Paper 53?

madison_example.txt



hamilton_example.txt

essay53.txt



We can fit a multinomial for Madison…

$$P(D|M)$$

…and fit another multinomial for Hamilton…

$$P(D|H)$$

…we could see which multinomial essay 53 is more likely to be from!

Guess Who's Back!

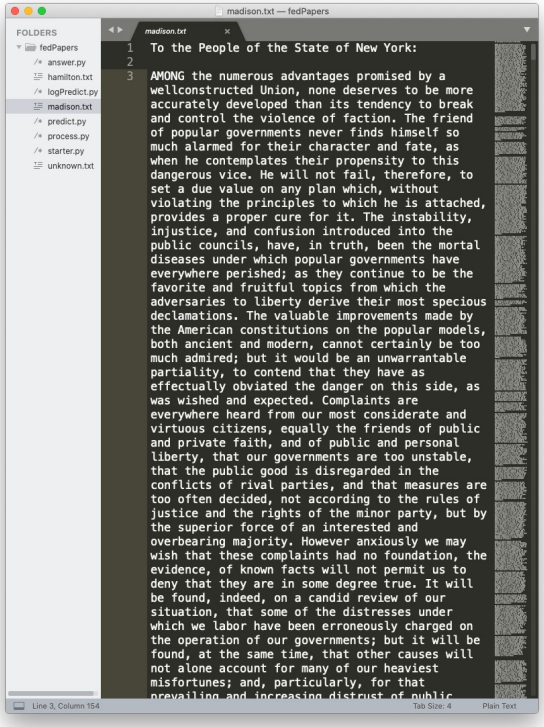
What we can calculate:
probabilities of words,
given their author (Madison/Hamilton)

But what we want to know is:
probability of author,
given the words in essay 53

Well hello again…

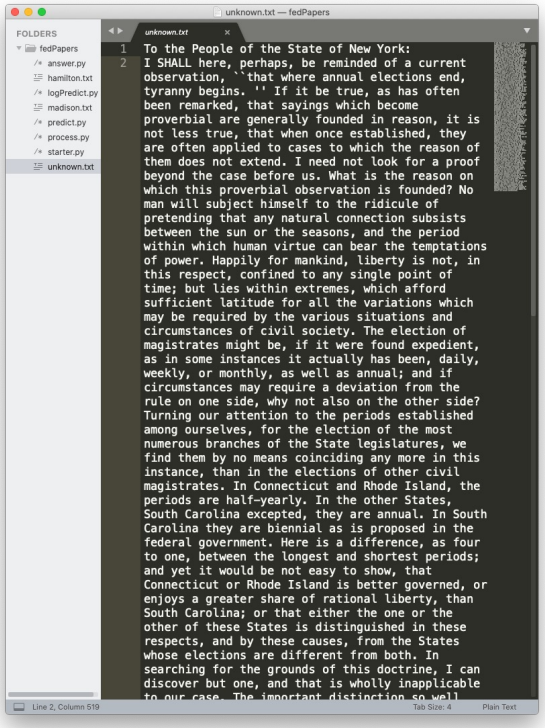
Who wrote Federalist Paper 53?

madison_example.txt



hamilton_example.txt

essay53.txt



We can fit a multinomial for Madison…

$$P(D|M)$$

…and fit another multinomial for Hamilton…

$$P(D|H)$$

…we could see which multinomial essay 53 is more likely to be from!

$$P(H|D) > P(M|D)\ ?$$

Who wrote Federalist Paper 53? We Need Bayes

Model essay as a multinomial where we care about counts of words

Prior belief it was Hamilton

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

Prob Hamilton given the text in essay 53

Prob of the essay??

Who wrote Federalist Paper 53? We Need Bayes

Loop over unique words

Prob hamilton would write word i

Number of times word *i* is in the essay

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

Prob Hamilton given the text in essay 53

Who wrote Federalist Paper 53? Comparing Multinomials

Probability that Hamilton wrote it:

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

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

Probability that Madison wrote it:

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

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

Who wrote Federalist Paper 53? Comparing Multinomials

Probability that Hamilton wrote it:

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

Probability that Madison wrote it:

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

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

This ratio will tell us which probability is bigger.

If this ratio > 1, Hamilton wrote essay 53!

Who wrote Federalist Paper 53? Comparing Multinomials

Probability that Hamilton wrote it:

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

Probability that Madison wrote it:

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

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

So much cancels out!

But If You Computed This As-Is:

Tip: Use Logs When Probabilities Become Too Small

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

If you calculated this literally, Python would tell you that the numerator and denominator are both zero.

Why?

- Each probability h_i or m_i is really small
- Multiplying lots of very small numbers together means REALLY small results
- Python eventually ”rounds down” to 0 due to finite precision

Tip: Use Logs When Probabilities Become Too Small

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

If you calculated this literally, Python would tell you that the numerator and denominator are both zero.

Why?

- Each probability h_i or m_i is really small
- Multiplying lots of very small numbers together means REALLY small results
- Python eventually ”rounds down” to 0 due to finite precision

Solution: Take the log!

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

Thanks for being awesome!