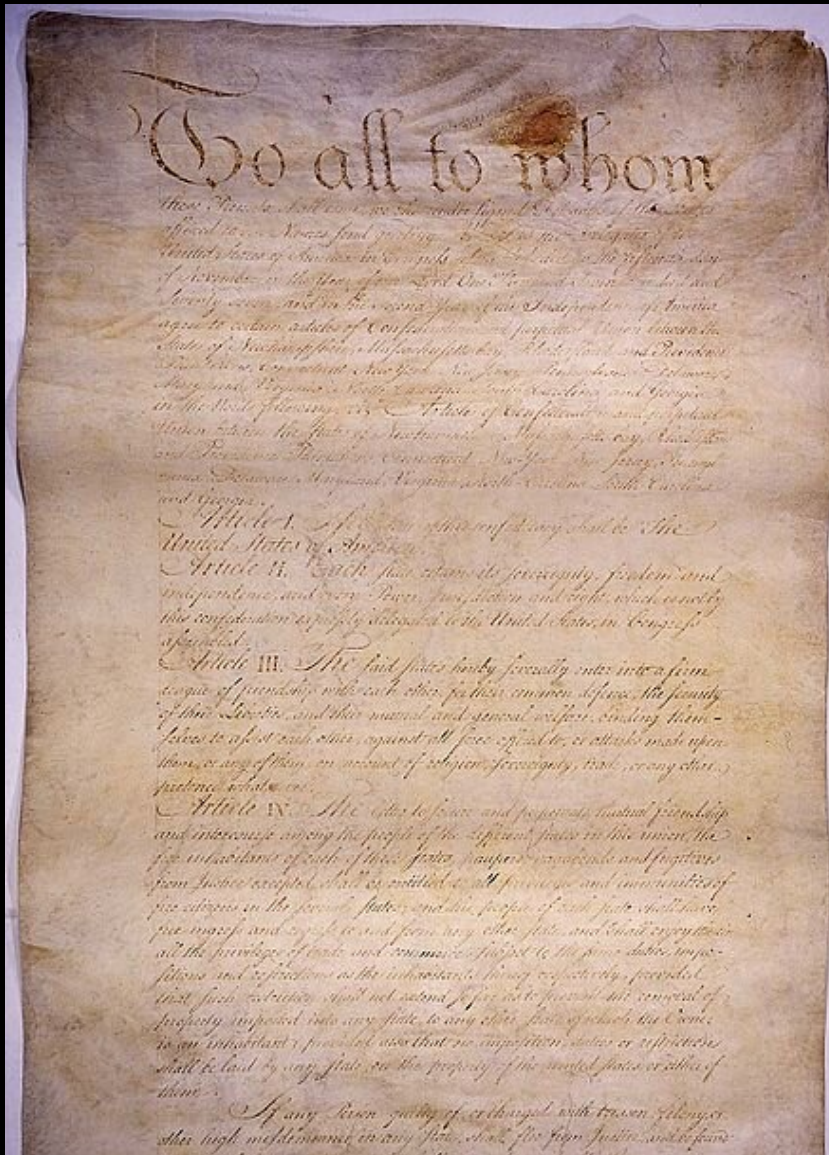




Probabilistic Models

CS109




"The best commentary on the principles of government ever written."
—PRESIDENT AND FOUNDING FATHER THOMAS JEFFERSON

"Read it, underline it, and dog-ear it." —SUPREME COURT JUSTICE ANTONIN SCALIA

THE FEDERALIST PAPERS

John Jay, James Madison,
& Alexander Hamilton



Foreword by
ALAN DERSHOWITZ

Who Wrote The Federalist Papers?

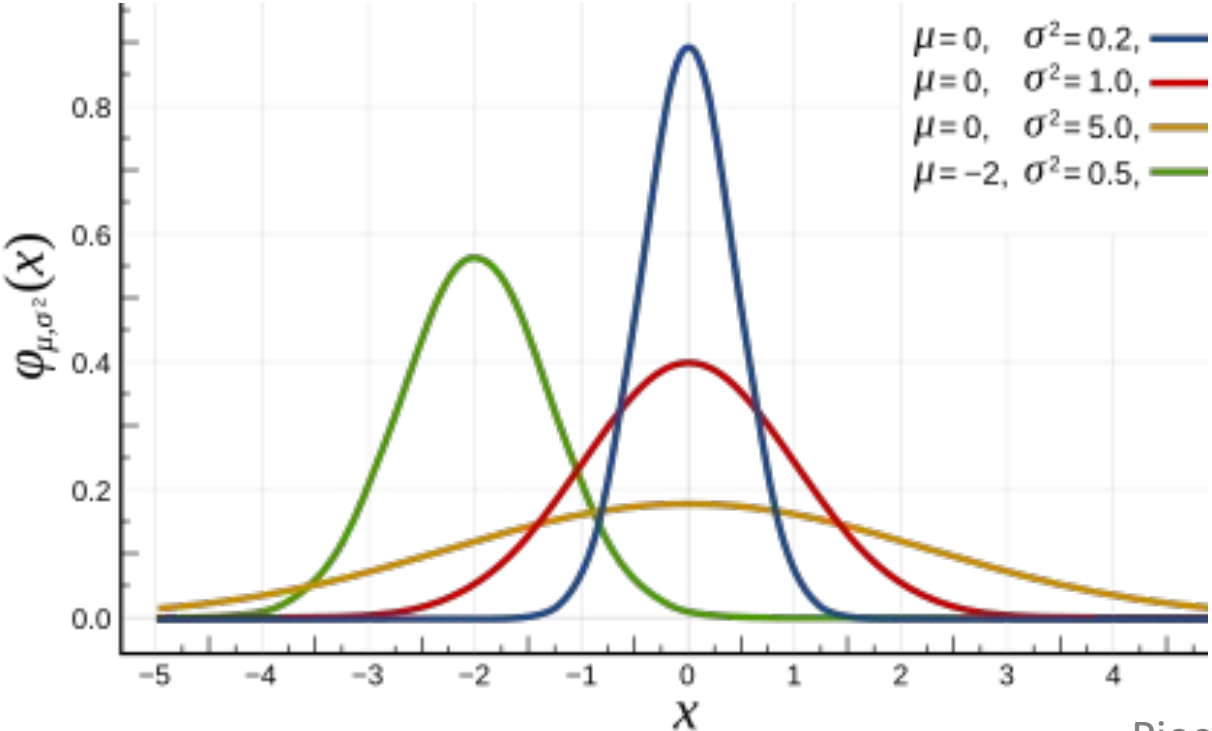
Review

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?

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ \longrightarrow $Y = aX + b$ \longrightarrow $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
is also Normal

What linear transform of X would get us to Z ?

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

$$Z \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

$$\sim \mathcal{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}\right)$$

$$\sim \mathcal{N}(0, 1)$$

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

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

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!!!

The course reader also has a calculator:

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

Norm CDF Calculator

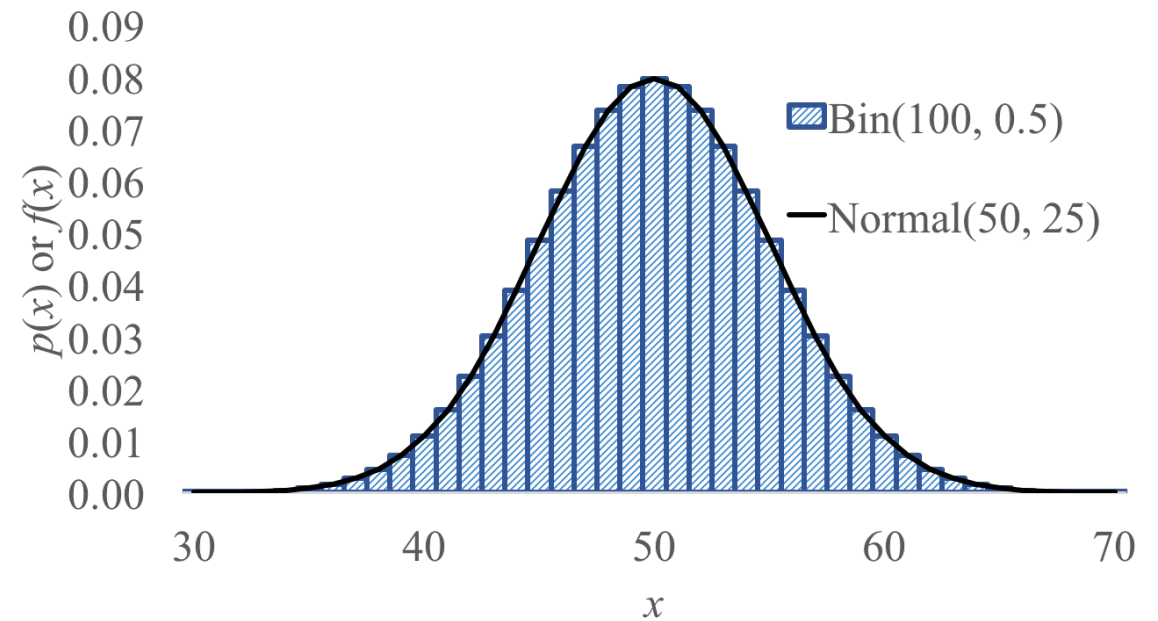
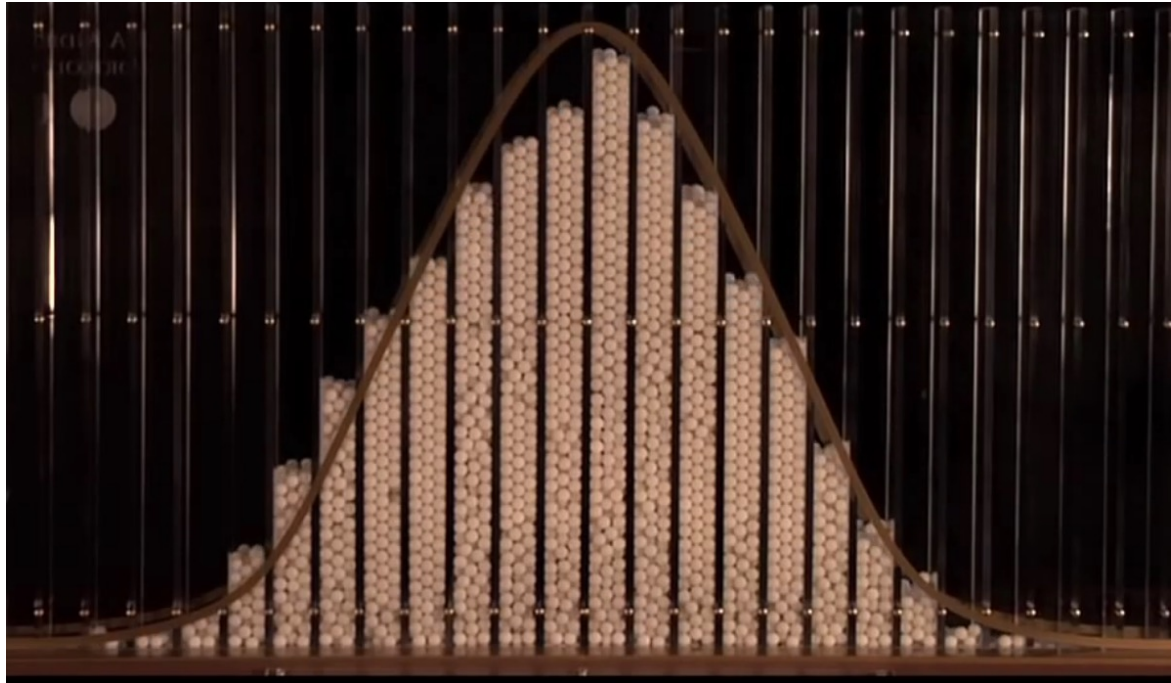
x

mu

std

`norm.cdf(x, mu, std)`

Normal Approximates Binomial, With Moderate p



The shapes are the same!

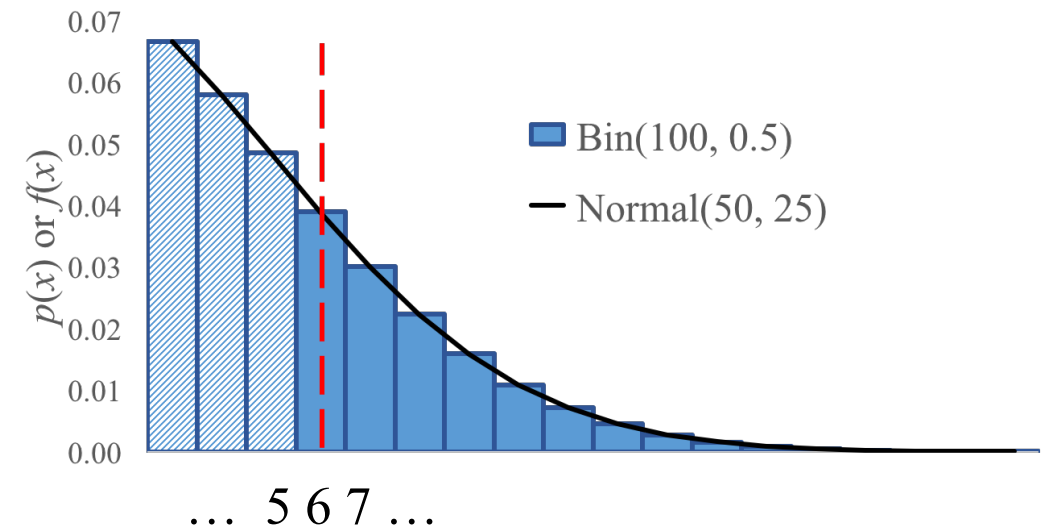
Just set the normal's μ, σ^2 to be the mean and variance of the binomial.

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: $P([\text{???}])$
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



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 $P(\text{no drop}) > 0.9999$?

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 $P(\text{no drop}) > 0.99999$?

Helpful fun fact -- phi has an inverse function:

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

End Review


Where are we in CS109?


Counting
Theory


Core
Probability

x_2
Random
Variables


Probabilistic
Models


Uncertainty
Theory


Machine
Learning



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?

WebMD Symptom Checker WITH BODY MAP

INFO SYMPTOMS QUESTIONS **CONDITIONS** DETAILS TREATMENT

Conditions that match your symptoms
[UNDERSTANDING YOUR RESULTS](#) ⓘ

- Gout**
Fair match
- Lyme Disease**
Fair match
- Osteoarthritis**
Fair match
- Pseudogout**
Fair match

Gender **Male** Age **50** [Edit](#)

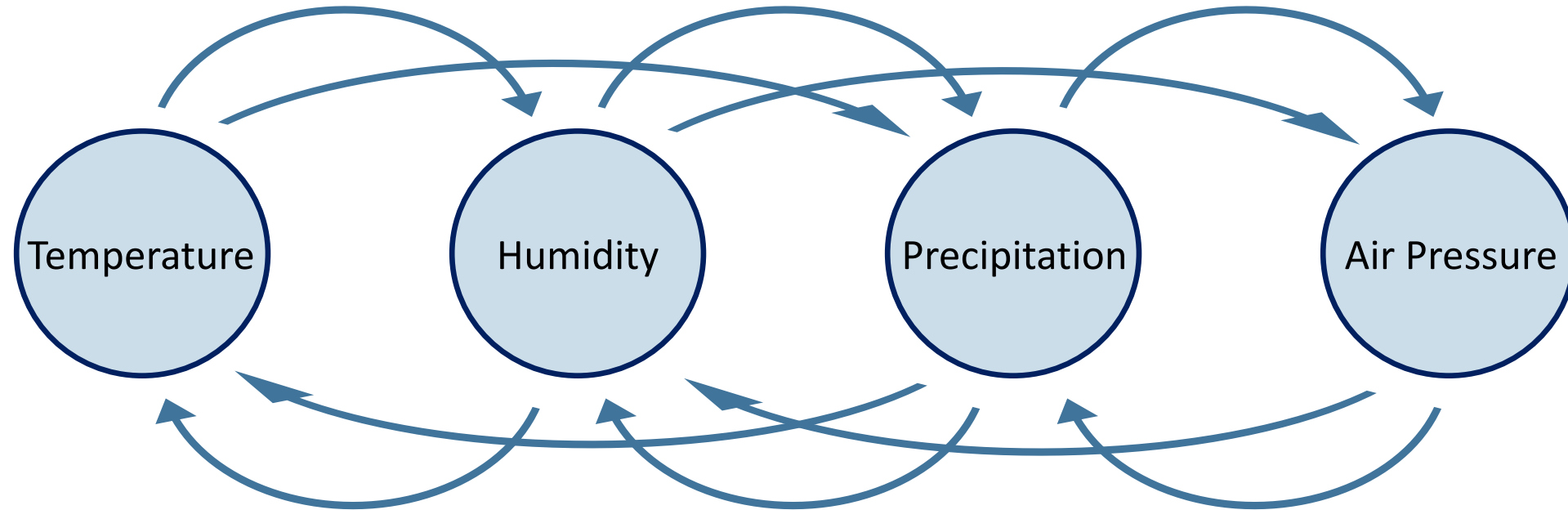
My Symptoms [Edit](#)
knee hurts , swelling , bruising

[Start Over](#)

<https://symptoms.webmd.com/>

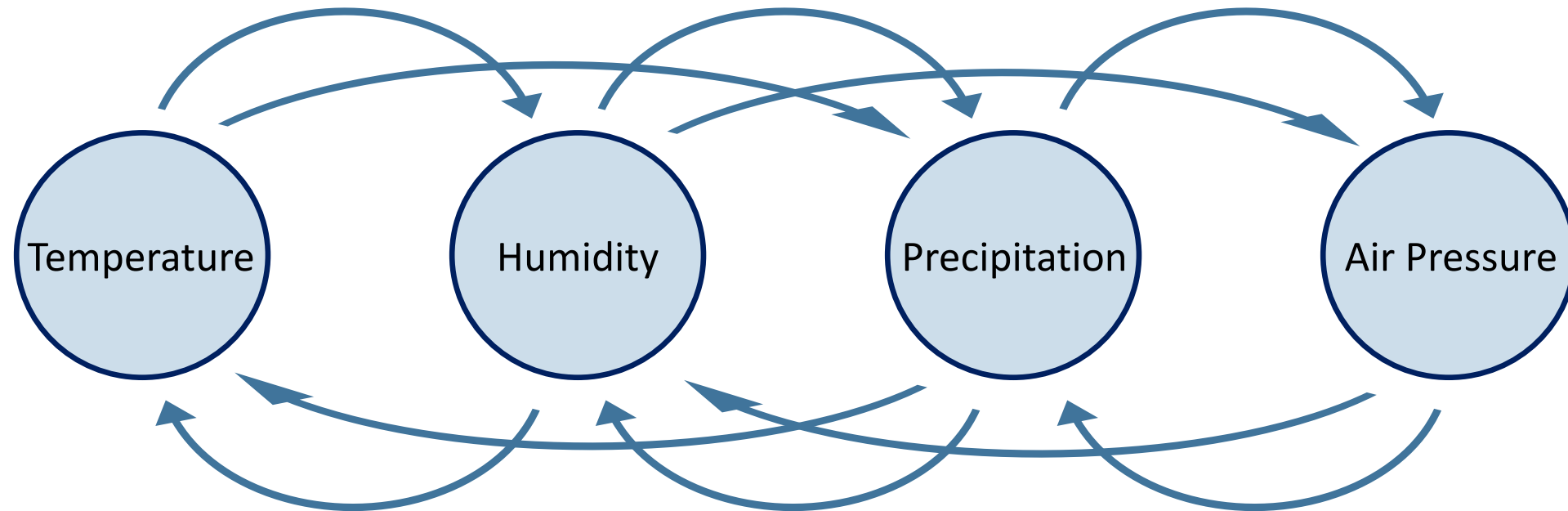
How Do We Model Multiple Random Variables Together?

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 $P(A \text{ and } B)$ 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**:

$$X = 2, Y = 4 \quad \nearrow \quad P(X = x, Y = y) \quad \nearrow \quad 0.5134 \dots$$

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

The same as $P(A \text{ and } B)$ for events A and B !

The “Joint” Distribution of Multiple Random Variables

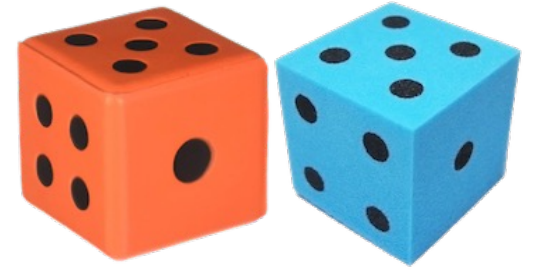
For *discrete* random variables X and Y , we have a **joint probability mass function**:

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

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

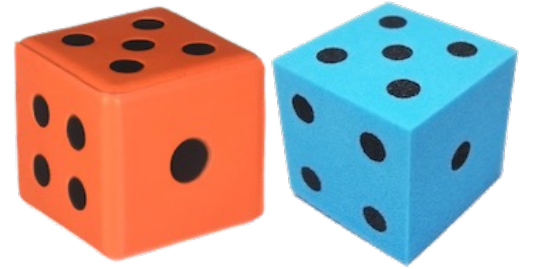
$$P(X = x, Y = y)$$

joint probability mass function

Example Joint PMF: Two Dice

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

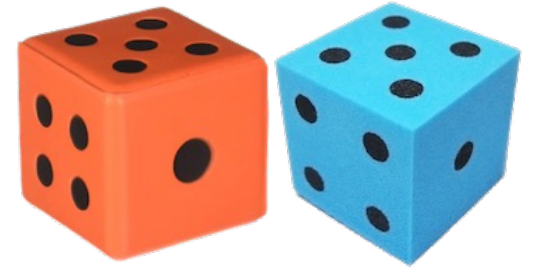
What is $P(X = x, Y = y)$?



Example Joint PMF: Two Dice

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

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



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

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

Example Joint PMF: Two Dice



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

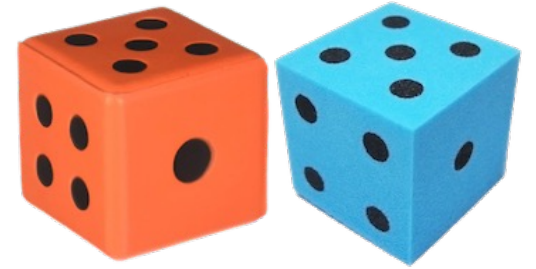
What is $P(X = x, Y = y)$?

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

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

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

Example Joint PMF: Two Dice



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

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

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

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

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

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

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

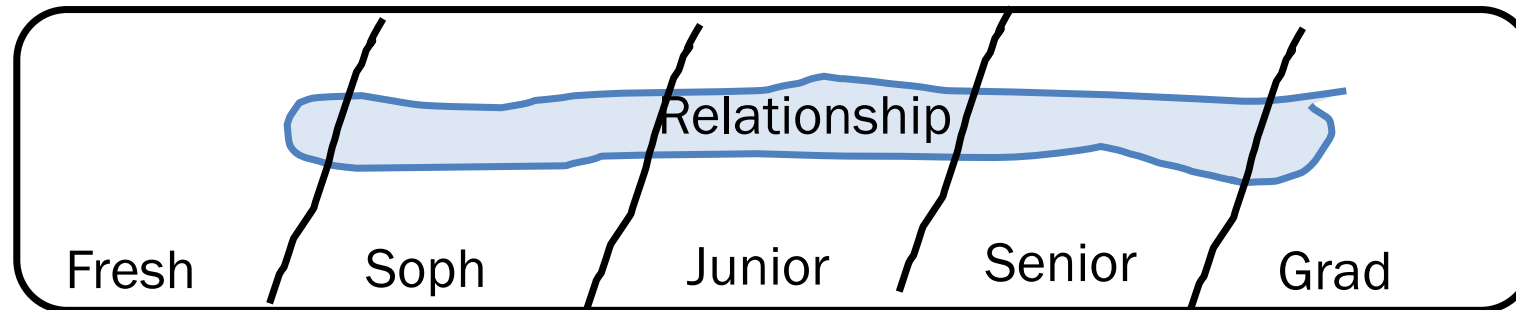
	Single	Relationship	Complicated
Freshman	0.13	0.08	0.02
Sophomore	0.17	0.11	0.02
Junior	0.09	0.10	0.02
Senior	0.02	0.07	0.01
Grad	0.06	0.09	0.04

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 $P(\text{relationship})$?



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

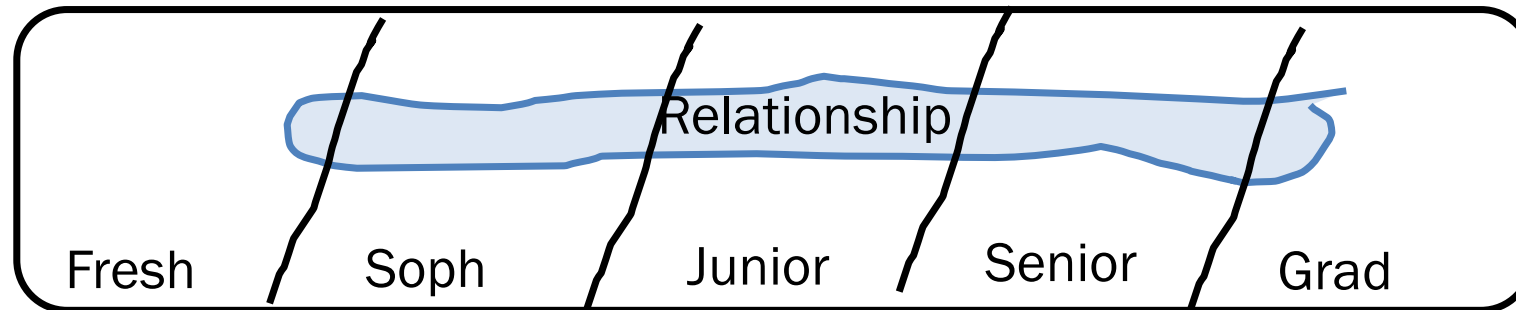
From the table, what is **P(relationship)**?

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

Concept
Check

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!

$$P(D = 1) = \sum_f \sum_s P(D = 1, F = f, S = s) = 0.051$$

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

$D = 0$

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

Concept
Check

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?

$$6^n$$

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

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

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.

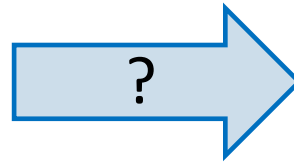


Key Limitation Of The Binomial: Only Binary Outcomes

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
- 0 threes
- 0 fives
- 1 two
- 2 fours
- 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
- 0 threes
- 0 fives
- 1 two
- 2 fours
- 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
- 0 threes
- 0 fives
- 1 two
- 2 fours
- 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

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

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

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?

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

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.?

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

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?

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

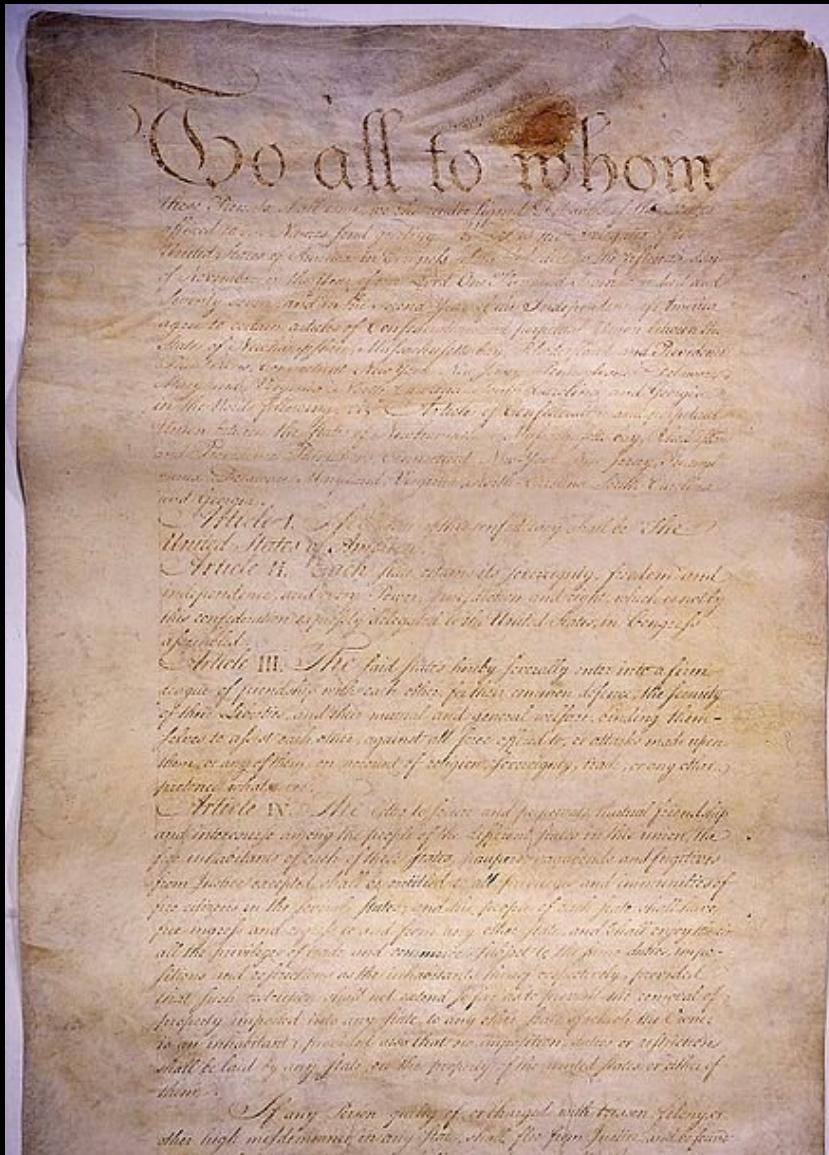
Multinomial coefficient:
ways to order the joint outcome

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




"The best commentary on the principles of government ever written."
—PRESIDENT AND FOUNDING FATHER THOMAS JEFFERSON

"Read it, underline it, and dog-ear it." —SUPREME COURT JUSTICE ANTONIN SCALIA

THE FEDERALIST PAPERS

John Jay, James Madison,
& Alexander Hamilton



Foreword by
ALAN DERSHOWITZ

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

Example text, with $n = 18$:

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

“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

Example text, with $n = 18$:

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

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

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

How We Model Text As A Multinomial: Spam

Example text, with $n = 18$:

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

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

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

Probability of seeing this text, if spam

How We Model Text As A Multinomial: Spam

Example text, with $n = 18$:

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

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

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

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

```
1 To the People of the State of New York:  
2  
3 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
```

hamilton_example.txt

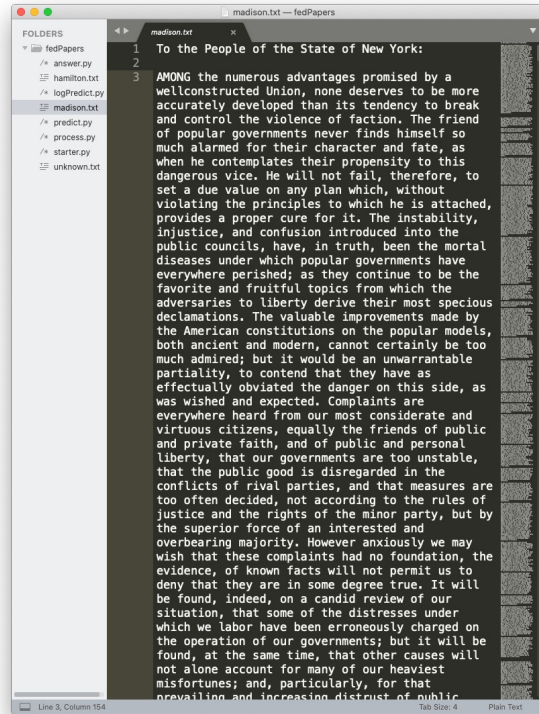
```
1 The Utility of the Union in Respect to Commercial  
Relations and a Navy  
Hamilton for the Independent Journal.  
2  
3  
4 To the People of the State of New York:  
5 THE importance of the Union, in a commercial  
light, is one of those points about which there  
is least room to entertain a 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.  
6  
7 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.  
8  
9 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
```

essay53.txt

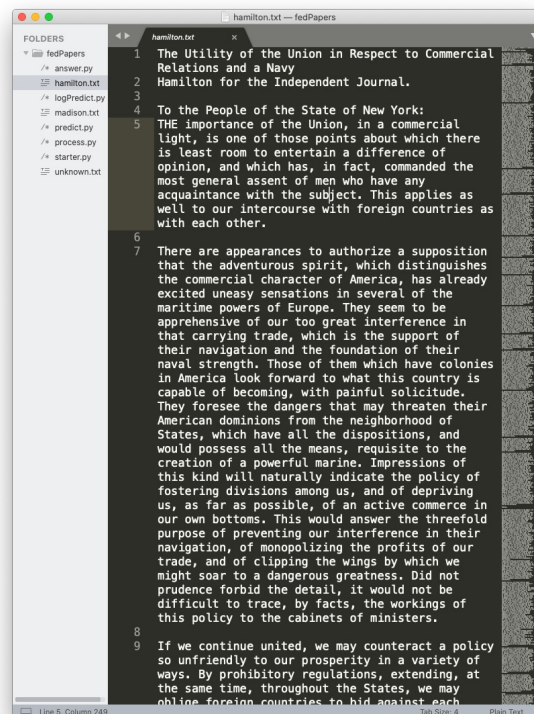
```
1 To the People of the State of New York:  
2 I SHALL here, perhaps, be reminded of a current  
observation, 'that where annual elections end,  
tyranny begins.' 'If it be true, as has often  
been remarked, that sayings which become  
proverbial are generally founded in reason, it is  
not less 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
```

Who wrote Federalist Paper 53?

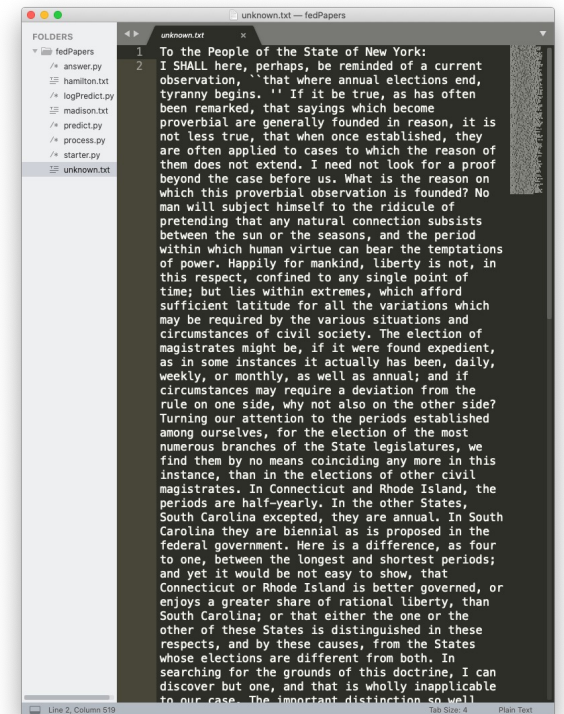
madison_example.txt



hamilton_example.txt



essay53.txt



We can fit a multinomial for Madison...

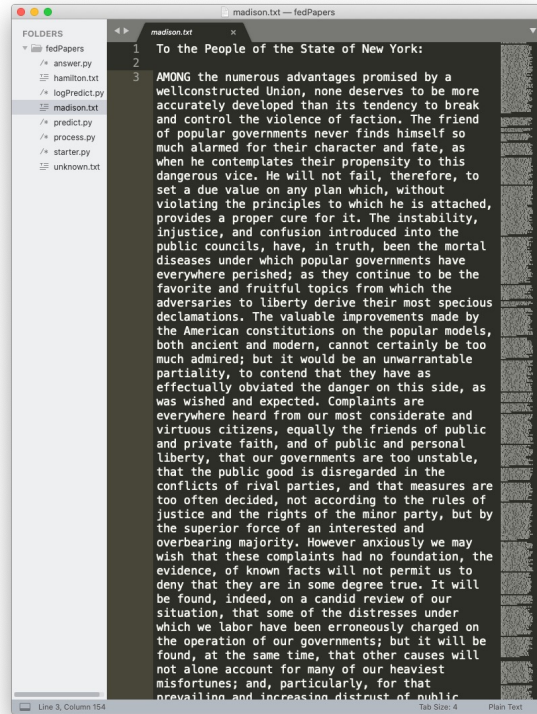
$$P(D|M)$$

...and fit another multinomial for Hamilton...

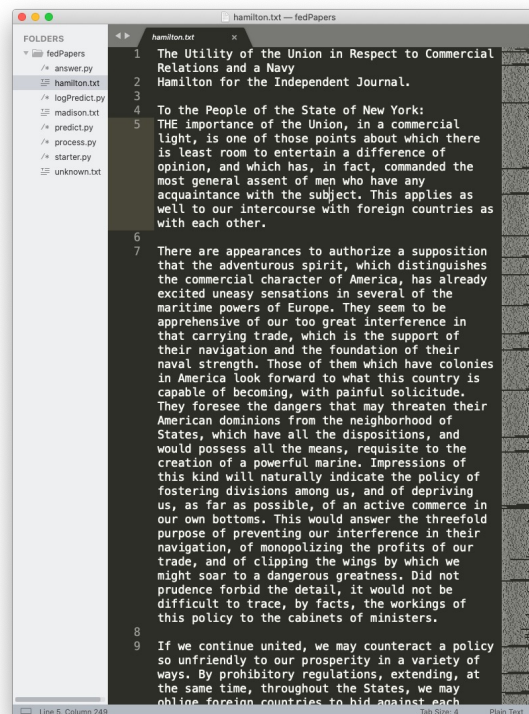
$$P(D|H)$$

Who wrote Federalist Paper 53?

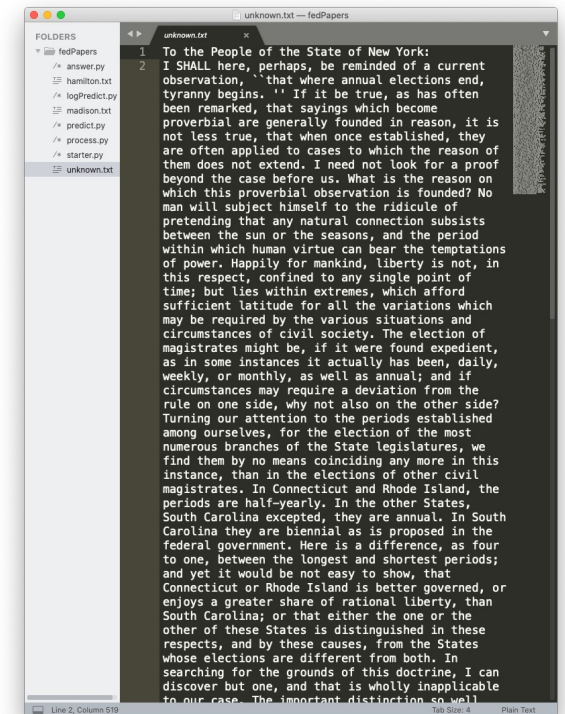
madison_example.txt



hamilton_example.txt



essay53.txt



We can fit a multinomial for Madison...

$$P(D|M)$$

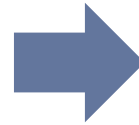
...and fit another multinomial for Hamilton...

$$P(D|H)$$

...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

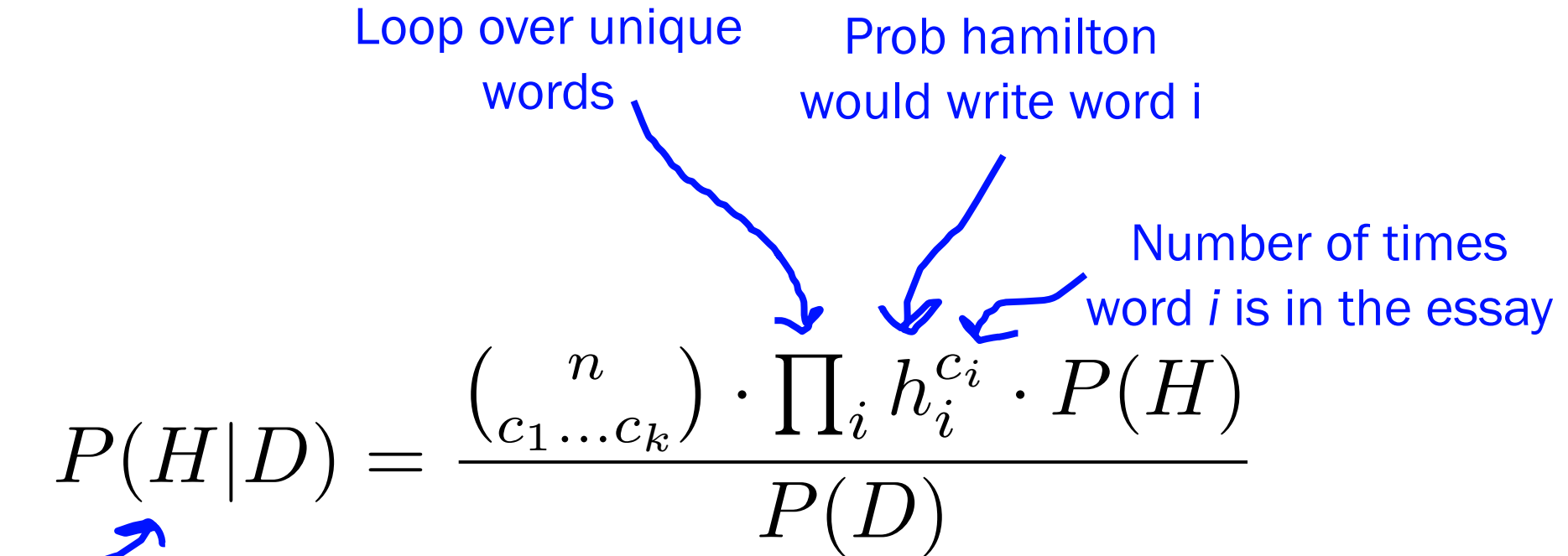
Prior belief it was
Hamilton

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

Prob Hamilton given
the text in essay 53

Prob of the essay??

Who wrote Federalist Paper 53? We Need Bayes

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


Loop over unique words

Prob hamilton would write word i

Number of times word i is in the essay

Prob Hamilton given the text in essay 53

Who wrote Federalist Paper 53? Comparing Multinomials

Probability that Hamilton wrote it:

$$\begin{aligned} P(H|D) &= \frac{P(D|H)P(H)}{P(D)} \\ &= \frac{P(H) \cdot \binom{n}{c_1 \dots c_m} \cdot \prod_i h_i^{c_i}}{P(D)} \end{aligned}$$

Probability that Madison wrote it:

$$\begin{aligned} P(M|D) &= \frac{P(D|M)P(M)}{P(D)} \\ &= \frac{P(M) \cdot \binom{n}{c_1 \dots c_m} \cdot \prod_i m_i^{c_i}}{P(D)} \end{aligned}$$

Who wrote Federalist Paper 53? Comparing Multinomials

Probability that Hamilton wrote it:

$$\begin{aligned} P(H|D) &= \frac{P(D|H)P(H)}{P(D)} \\ &= \frac{P(H) \cdot \binom{n}{c_1 \dots c_m} \cdot \prod_i h_i^{c_i}}{P(D)} \end{aligned}$$

Probability that Madison wrote it:

$$\begin{aligned} P(M|D) &= \frac{P(D|M)P(M)}{P(D)} \\ &= \frac{P(M) \cdot \binom{n}{c_1 \dots c_m} \cdot \prod_i m_i^{c_i}}{P(D)} \end{aligned}$$

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

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:

$$\begin{aligned} P(H|D) &= \frac{P(D|H)P(H)}{P(D)} \\ &= \frac{P(H) \cdot \binom{n}{c_1 \dots c_m} \cdot \prod_i h_i^{c_i}}{P(D)} \end{aligned}$$

Probability that Madison wrote it:

$$\begin{aligned} P(M|D) &= \frac{P(D|M)P(M)}{P(D)} \\ &= \frac{P(M) \cdot \binom{n}{c_1 \dots c_m} \cdot \prod_i m_i^{c_i}}{P(D)} \end{aligned}$$

$$\begin{aligned} \frac{P(H|D)}{P(M|D)} &= \frac{P(H) \cdot \binom{n}{c_1 \dots c_k} \cdot \prod_i h_i^{c_i}}{P(D)} \bigg/ \frac{P(M) \cdot \binom{n}{c_1 \dots c_k} \cdot \prod_i m_i^{c_i}}{P(D)} \\ &= \frac{\prod_i h_i^{c_i}}{\prod_i m_i^{c_i}} \end{aligned}$$

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!