10. Probability and statistics
LING 471

Learning outcomes

- Describe and calculate (frequentist) **probability**, **conditional probability**, **joint probability**, **sequence probability**, and **marginal probability**
- Define **independent and mutually exclusive events**
- Simulate **binary outcomes/experiments** in Python (like coin flipping) using the `random` module
- Describe what a **parameter** is, especially in a binary outcome scenario
- Describe why **logarithms** are important when calculating probability
- Describe **maximum likelihood estimation**

Probability theory

Data science and statistics

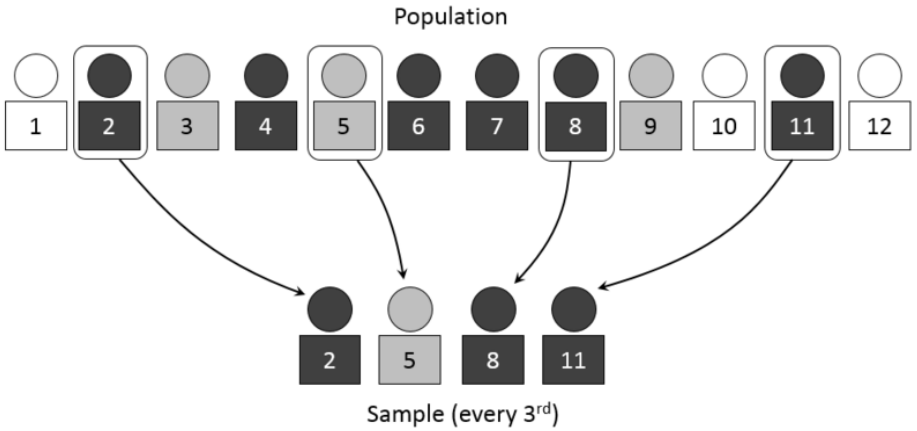


Image from Mathprofdk under CC BY-SA 4.0

- There is a lot of **randomness** and **uncertainty** in the world
- Many processes in our lives generate data
  - How many times you click on something, how many messages we send and receive, etc.
- Statistics tries to make sense of the world by **sampling** data
  - What is true of your sample should be true of your population (if the sample is truly **random** and sufficiently **large**)

Probability theory and statistics

PROBABILITY THEORY

- Formally, estimate the **likelihood** of an outcome
- Informally, predict **future** events
  - Given what I know about the population, what sample can I draw?
- Relies on the notion of the probability distribution
  - How are probabilities of all possible outcome distributed?

STATISTICS

- Use probability distributions to make sense of large data formally
- Informally, analyze past events
  - Given the samples I drew, what I can I say about the population?
- No distribution → no statistics!

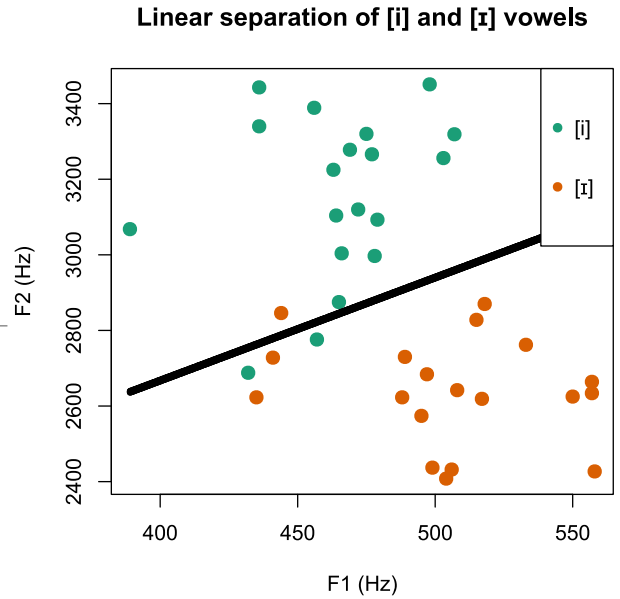
Probability theory and statistics

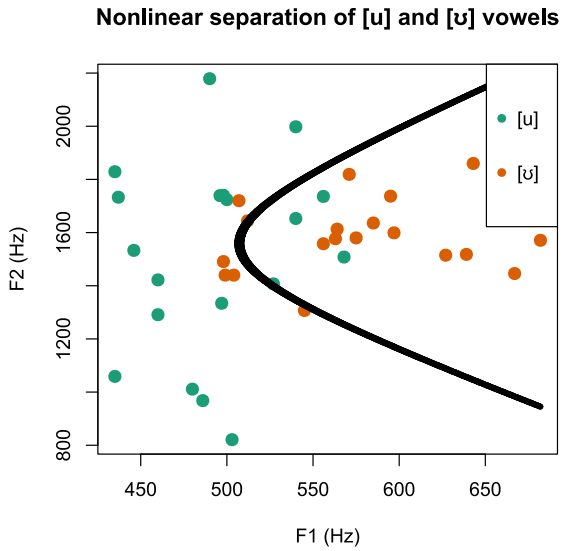
Data science

- Probability + statistics
- Analyze past events **and** predict future events, **at scale**, in real world

Prediction and probabilities: Classification problems

- Predictions in ML need to be quantified
- To predict whether a review is good or bad:
  - E.g., compute the **probability** of the review being good
  - Predict "good" if probability is **high**
  - Predict "bad" otherwise





Conditional probability

- Conditional probability: P(Y|X)
  - Read "The probability of Y given X"
  - That is, "what is the probability of Y if we know X"?
- In this case, Y is the label and X is the observation
  - E.g., Y = "good" and X = "This is a good movie!"
  - That is, "What is the probability the review is 'good' if we know the review said 'This is a good movie!'?"
- How do we learn P(Y|X)?
  - There are mathematical functions we use
  - We'll cover more on some of these techniques later

Probability theory

- It is notoriously unintuitive and hard! ☹
- Our goal is to be familiar with some basic concepts
  - Not necessarily in the most formal, rigorous way
- …Such that we can experiment with some data science models in Assignments 4 and 5

Probability theory for today

- Definitions: events, outcomes, sample space, random variables
- Mutually exclusive events
- Sequences and independent events
- Joint probabilities
- Conditional probability
- Marginalizing joint probabilities
- Bonus: Maximum likelihood estimation

Probability: Basic intuitions

- How likely is something to happen?
  - Well, we don't know for sure
  - But, we can **estimate** based on **prior observations** or what we **assume** about a situation
- **Out of n** experiments (the "sample space"), how many resulted in a specified outcome?
  - This is the **frequentist** interpretation of probability

Frequentist probability

- Based on the **frequency** of an event occurring
- **Probability** is the ratio between the frequency of a specific event to all observed events
- Understanding the sample space exactly is crucial
- The probability will be different based on what the sample space is

Coin flipping: The classic example

- Sample space: {T, H} (heads or tails)
- Experiment: one coin flip
- Outcome: either T or H
- Extra assumption: A coin will not land on an edge

Flipping a coin in Python

- We can randomly generate numbers in Python with the `random` module
- One way is using the `random.choices` function, which will randomly choose an item from a list *k* times
- Assume 0 = T and 1 = H

```
import random as rand
rand.seed(20260430)
flip = rand.choices([0, 1], k=1)
# flip is [0], i.e., T
```

Coin flip series

- Sample space
  - Depends on number of flips, i.e., the experiment
  - For 2: {TT, HT, TH, HH}
- Experiment: A number of flips
- Outcome: A sequence made up of H and T
- Statistically, P(H) is estimated by a large number of experiments
  - Toss the coin 1 million times
  - Count how many times you got H
  - N / 1 million is the empirical (frequentist) estimate of P(H)
    - This can be proved formally with maximum likelihood estimation (in a few slides from now)

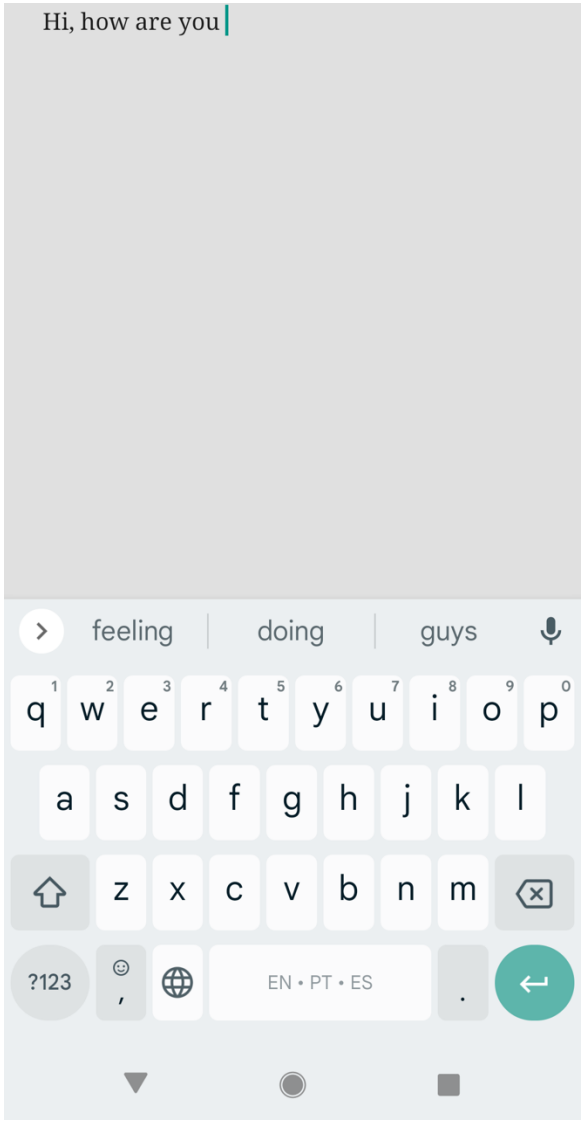
Coin flip series in Python

- It is easy to get more coin flips in Python
- Just change the value of *k* in `random.choices`
- This will take a moment to run, though…
  - And, my computer runs out of memory when doing this, so we need to use a generator instead
- To get a sense of how much larger 1 billion is than 1 million, try increasing k to 1 billion (takes over 20 minutes on my computer)

```
import random as rand
rand.seed(20260430)
n_H = sum(rand.choices([0, 1])[0] == 1 for f in range(1000000))
# n_H = 499216
p_H = n_H / 1000000
# p_H = 0.499216
```

A fair coin

- A fair coin is one such that P(H) = 0.5
- In other words, you can toss it a million (or a billion) times and expect H to come up ≈500K times
  - What if you instead got only 499216 heads? (Like we just did)
- The probability is 0.499612
  - For all practical purposes, that's still 0.5
- In a strict sense, the theoretical probability is only true for very large numbers (i.e., as you approach infinity)
  - Law of large numbers



Sequence probability in NLP

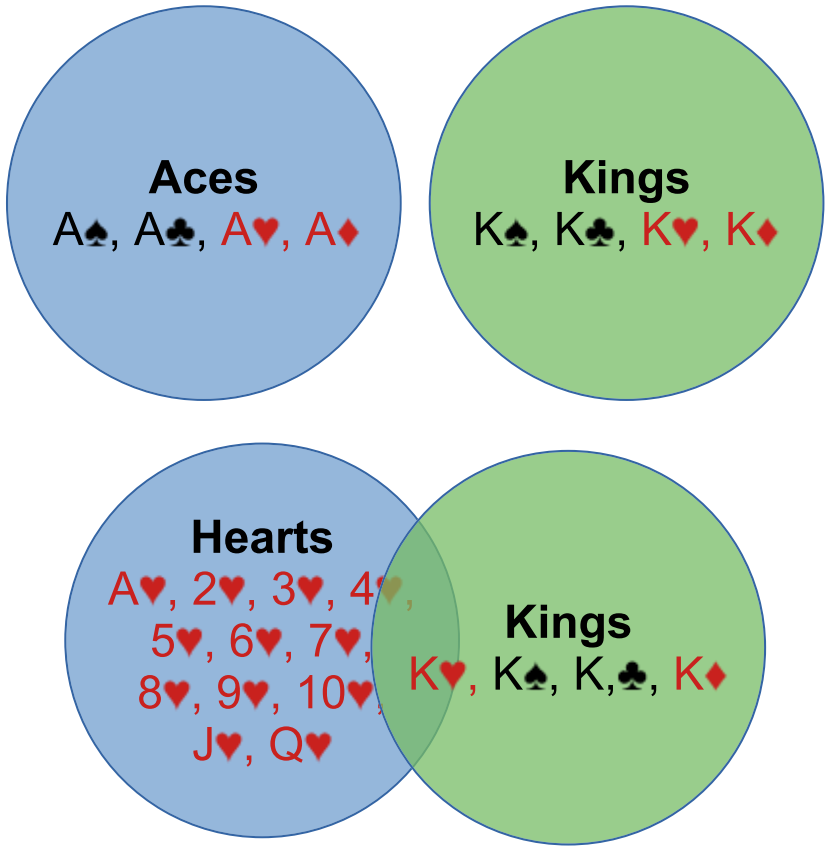
- Why are we flipping coins when we want to work with language data??!!!
  - Coins give us a simple example of sequences and sequence probabilities
  - And, **texts** are **sequences** of words, characters, symbols, sentences, paragraphs, etc.
- Language modeling:
  - Estimating probabilities of textual sequences
  - Given what we've seen before, what is the most likely continuation?
    - This is how word suggestions work! (See left)

Probability and frequency

- How probable is some outcome?
  - E.g., H or T
- How frequent is some outcome?
  - E.g., H or T
- What's the difference?
  - Frequency is observed
  - Probability is estimated (or maybe derived)
  - Probability in this sense is also sometimes called "relative frequency"

Probabilities sum to 1 for mutually exclusive events

- Refers not to just any set of probabilities
  - Rather, refer to only the sets of probabilities that account for all possible outcomes in a specific setting
  - Really, just a convention/definition; 1 = 100%
- Example: coin flip
  - When you flip the coin, you **must** get T or H
  - That is, there is a 100% chance you will get one of T or H
- Notation: P(T) + P(H) = 1



Mutually exclusive events

- T and H in coin flip
  - P(H **and** T) = 0
    - For one coin toss
- Playing cards
  - P(King **and** Ace) = 0
    - If drawing **one** card
  - **But**, P(King **and** Hearts) > 0
    - **Not** mutually exclusive

Sequence probabilities for independent events

- Suppose you flip a fair coin twice
- Sample space: {HH, HT, TT, TH}
- What's P(HH)? **0.25**
  - This is P(H) * P(H) (that is, 0.5 * 0.5)
  - Probability of a sequence is a **product**
- What's P(HT) (in that order)? **0.25**
- What's P of getting one H and one T, any order? **0.5**
  - This is P(HT) + P(TH)
  - Probability of a disjunction (logical or) is a sum
  - P(HT **or** TH) = 0.25 + 0.25 = 0.5

Random variables

- A random variable is the set of possible values from a probabilistic experiment
  - E.g., {T, H}
  - Let's define T=0 and H=1
- A random variable is often denoted with a capital letter
- For our coin flip, let's use X as our random variable
- So, when we have our random variable X, we are saying
  - "It could take on any of these possible values: {0, 1}"
- We don't know what the values are until we perform our experiment (i.e., flip our coin)

Potentially confusing notation with random variables

- What do people mean when saying P(X) or P(A), etc.?
- Usually, if A is a random variable and the values are, e.g., {1,2,3,4,5,6}
- Then P(A) may refer specifically to P(A=1), P(A=5), etc.
- We will try to be more explicit and actually say P(A=1), P(A=5), etc. instead of P(A)

Independent events

- Two events are independent if they do not affect each other
  - E.g., a coin flip, rolling a die, etc.
- P(A **and** B) = P(A) * P(B) if and **only** if A and B are independent from each other
- When you flip a coin, the result does not depend on any previous result

Independent events

- Two events are independent if they do not affect each other
  - E.g., a coin flip, rolling a die, etc.
- P(A **and** B) = P(A) * P(B) if and **only** if A and B are independent from each other
- When you flip a coin, the result does not depend on any previous result

- Let's flip a coin 10 times and then flip it 10 more times
  - Assume we got "HHHHHHHHHH" (10 heads in a row) for the first sequence, S1
  - Assume we got "HTTHTHTHHT" for the second sequence, S2
- P(S1) = 1/1024
- P(S2) = 1/1024
  - This is unintuitive because S2 seems much more balanced

Conditional probability

- What is the probability of A given B?
  - Notation: P(A|B)
- E.g., if it is very sunny right now, is it more or less likely that it will rain in 30 minutes?
  - Compared to when it is not sunny
  - P(rain in 30 mins | it is very sunny)
- E.g., If you see lightning, is it more or less likely that you hear thunder in a few seconds?
  - P(hear thunder | just saw lightning)
- Always have to consider the sample space!

Definition of conditional probability

- P(thunder | lightning) = P(L **and** T) / P(L)
  - P(L **and** T): Estimated by counting all occurrences were **BOTH** thunder and lightning occurred
  - P(L): Estimated by counting all occurrences when lightning occurred
- Conditional probability is crucial in the **Bayes Theorem**
  - And in Naïve Bayes classifiers
  - And in many data science techniques
  - And in Assignment 4!
- Sometimes easier to just remember this definition than try to develop an intuition…

Marginal probabilities

Joint Probability Table				
	Single	In a relationship	It's complicated	**Marginal Year**
Freshman	0.13	0.09	0.02	0.24
Sophomore	0.16	0.10	0.02	0.28
Junior	0.12	0.10	0.02	0.23
Senior	0.01	0.09	0.00	0.10
5+	0.03	0.12	0.01	0.15
Marginal Status	0.45	0.48	0.07	

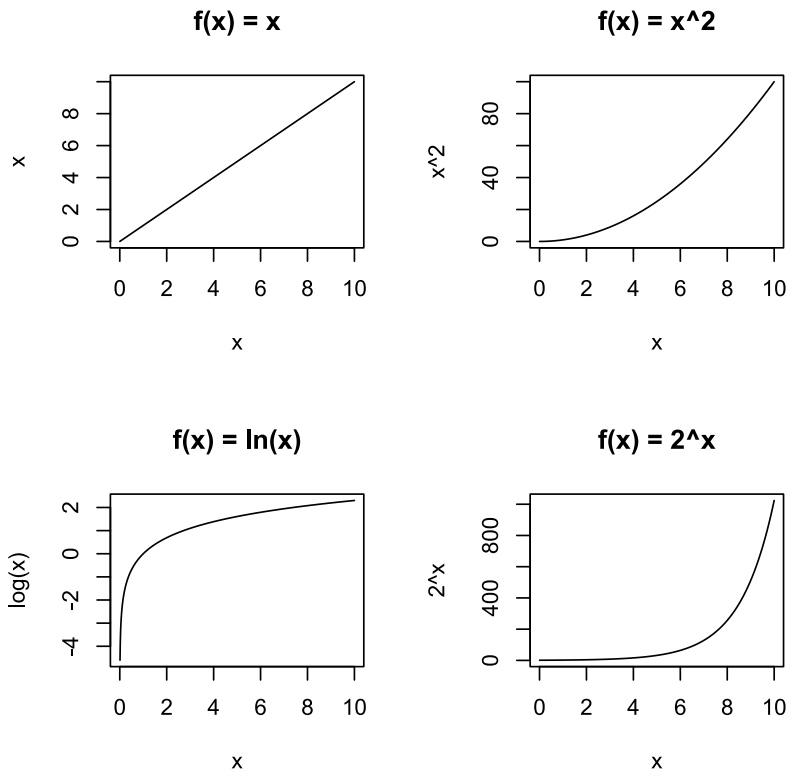
https://web.stanford.edu/class/archive/cs/cs109/cs109.1176/lectures/12-Continuous.Joint.pdf

- Put conditional probabilities in tables
- The table has joint probabilities in it, i.e., probabilities of two events
  - Notation: P(A, B) (equivalent to our P(A **and** B))
- To marginalize the probability of A, you compute P(A) by removing any dependencies on other events
  - By summing along a row or column
- The marginals should sum up to 1

Programming activities

- Using the `random` module, flip some coins
  - Assume 0 is tails and 1 is heads
- Flip 1,000 coins and count how many heads or tails you get
- Flip 1,000 more and compare your results
- For a challenge: create an unfair coin, i.e., one where heads and tails are not equally probable
  - Hint: Give more options to `random.choices` and rethink what numbers correspond to heads and tails

Statistics

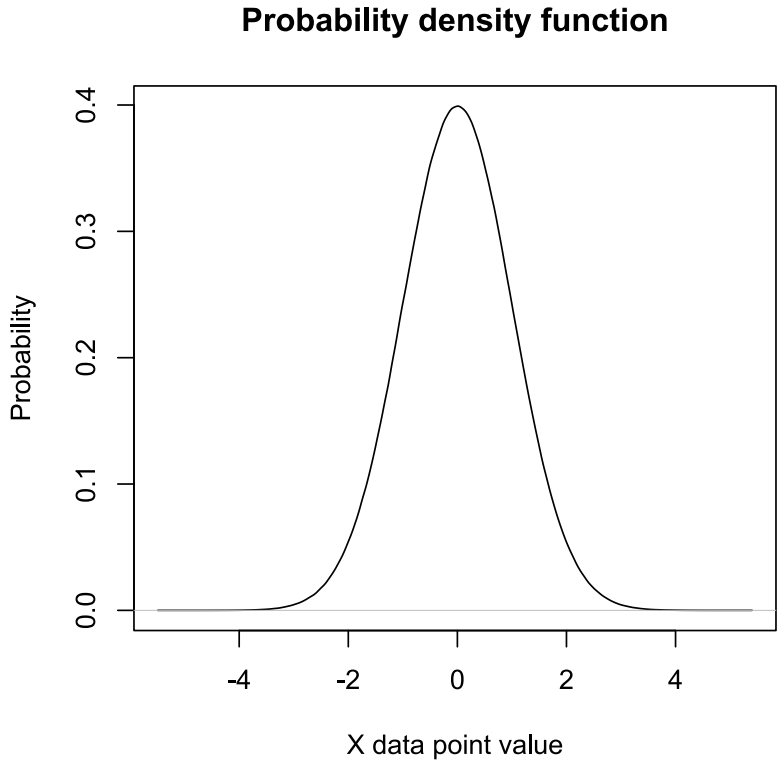


Functions

- Functions are the bread and butter of statistics
- Definition:
  - Input → output (map an input to an output)
  - Given x, what is the value of y?
  - f(x), e.g., y = f(x) = 2x
- Functions of one variable can be visualized as lines and curves (in 2D)

Probabilities and functions

- We can treat probabilities as a function
- What is the **probability** of observing **data point** x (maybe in some range)?
  - Need to know how the data points are **distributed**
- Probability functions describe these **distributions**
- Many distributions are described by parameters, but how do we find them?
  - E.g., mean, standard deviation



Maximum likelihood estimation: High level

GOAL

- Represent probabilities abstractly, as formulas
- Probability of each outcome is a parameter
- Parameters can be unknown; we want to estimate their values

EXAMPLE: UNFAIR COIN TOSS

- What's P(H) for an unfair coin?
  - We don't know, so we will use an abstract parameter θ, i.e., P(H) = θ
- Then, P(T) = $1 - \theta$
- Also, P(HT) = $\theta * (1 - \theta)$
- And, P(HHHTT) = $\theta^3 * (1 - \theta)^2$
- What is the value of θ?

Probability from events

- Suppose we tossed our unfair coin 5 times
  - Result: HHHTT
- What is P(H)?
  - 3/5 = 0.6
- This is from our frequentist definition previously
  - Means this is theoretical
- Can we get some other kind of evidence that this is the probability?

Maximum likelihood estimation I

- We can get other evidence!
- We know that P(H) exists
  - Let's call it θ
- What do we know about P(T)?
  - It must be $1 - \theta$
- Let's assign $D = \{HHHTT\}$
  - What is P(D)?
  - P(D) is the **product** of the probabilities

Maximum likelihood estimation II

- $P(D) = \theta^3 * (1-\theta)^2$
- What exactly do we want?
  - We want to know the actual value of θ
- But we also want a value for θ such that P(D) is maximized
  - Why?
  - We want to know the probabilities that are **most likely** for our given sequence
    - In theory, a fair coin could have given HHHTT, right?

Maximizing P(D)

- How do we find the maximum point of a function?
  - Well, calculus! (But that wasn't a pre-req…)
    - We're going through this to demonstrate a concept, not to have technical mastery!
- Generally,
  - Think of your function as a curve
  - A curve becomes flat at its maximum (its slope = 0)
  - A curve's slope at a given point is known as its derivative
  - You can often directly compute this value; many functions have well known derivative you can look up
- We need to find the maximum point of our P(D) function

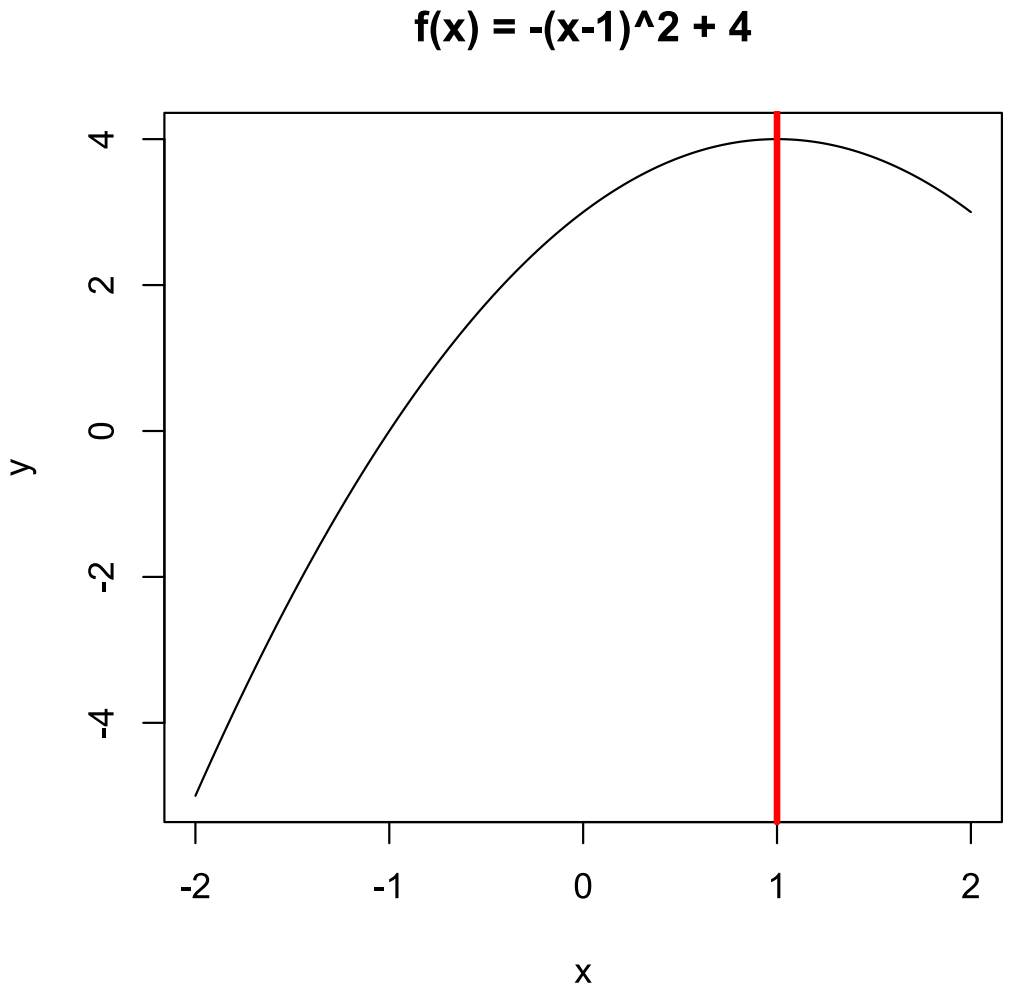
Programming activity

- Numerically estimate the value of θ that maximizes $P(D)$
- Basic strategy: try different numbers and see which one gives you the maximum value
  - Try using the step keyword in the range function, or randomly generate values and check each one
  - Use a for loop to find which value in the range gives you the highest P(D)
  - You will need to keep track of your current maximum value
- Feel free to come up with your own numerical approximation method
  - Don't solve analytically for it if you happen to know how… ;)

Excursus: arg max and logarithms

Arg max

- Functions look like curves in 2D
- Those curves have **maxima** along the y-axis
- The point on the x-axis where y is maximum…
  - …is defined as the arg max
- Why is this important?
  - We want to find the parameters for probability functions, given our observations
  - If the function has θ as a parameter, we want to know what value of θ results in the maximum probability for the observed sequence/data
- So, we want to find: $\arg\max P(\theta)$



Arg max visual

- This parabola has a maximum value of 4
  - Largest y value is 4
- Where does this maximum occur?
  - When x = 1
- $\arg\max(-(x-1)^2+4) = 1$

Small numbers and computers

- Probabilities range from 0 to 1
- Suppose you have a very long sequence of events
- What happens when you multiply many, many numbers, each ranging between 0 and 1?
  - Your number becomes so small your computer cannot represent it!

```
0.1 ** 100
# 1.0000000000000056e-100
0.1 ** 1000
# 0.0
```

Logarithms to the rescue!

- We're doing lots of products, and logs have a convenient property:
  - $\log(xy) = \log(x) + \log(y)$
- Can use $\log(P(A))$ in place of P(A) for likelihood estimation
  - arg max of P(D) will be where arg max is for log(P(D))
- Can use sum of logs instead of product

```
from math import log

# log in programming defaults to
natural log,

# not log10

sum(log(0.1) for x in range(1000))

# -2302.585092994075

# this is a log probability and
very useful!
```

Returning to our main event…

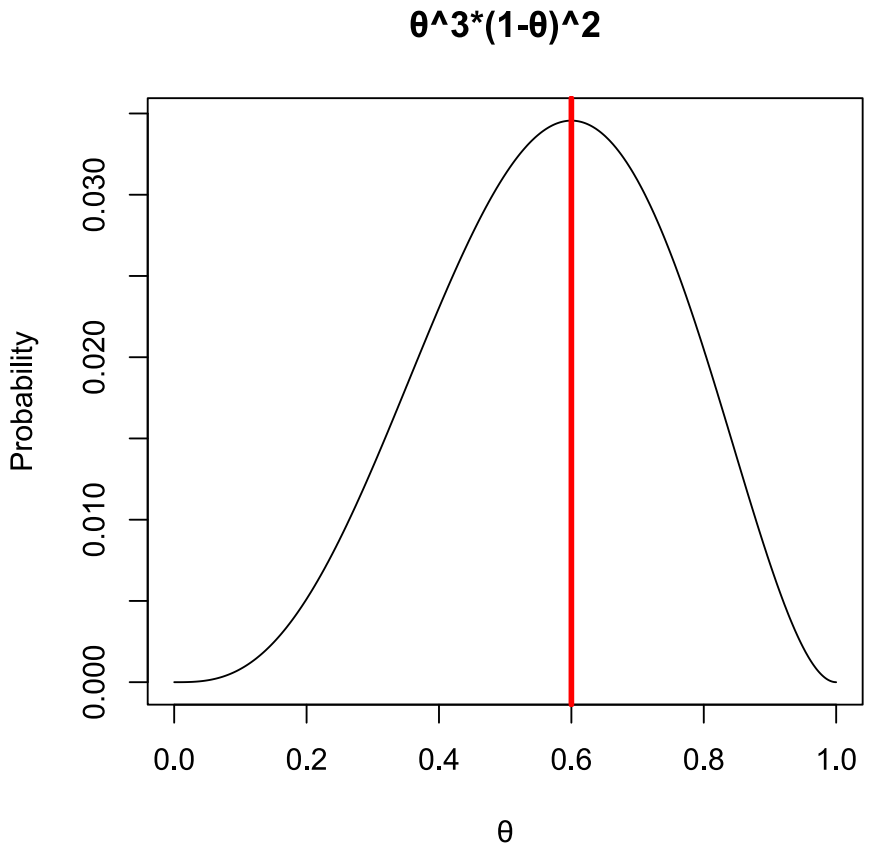
Maximum likelihood estimation III

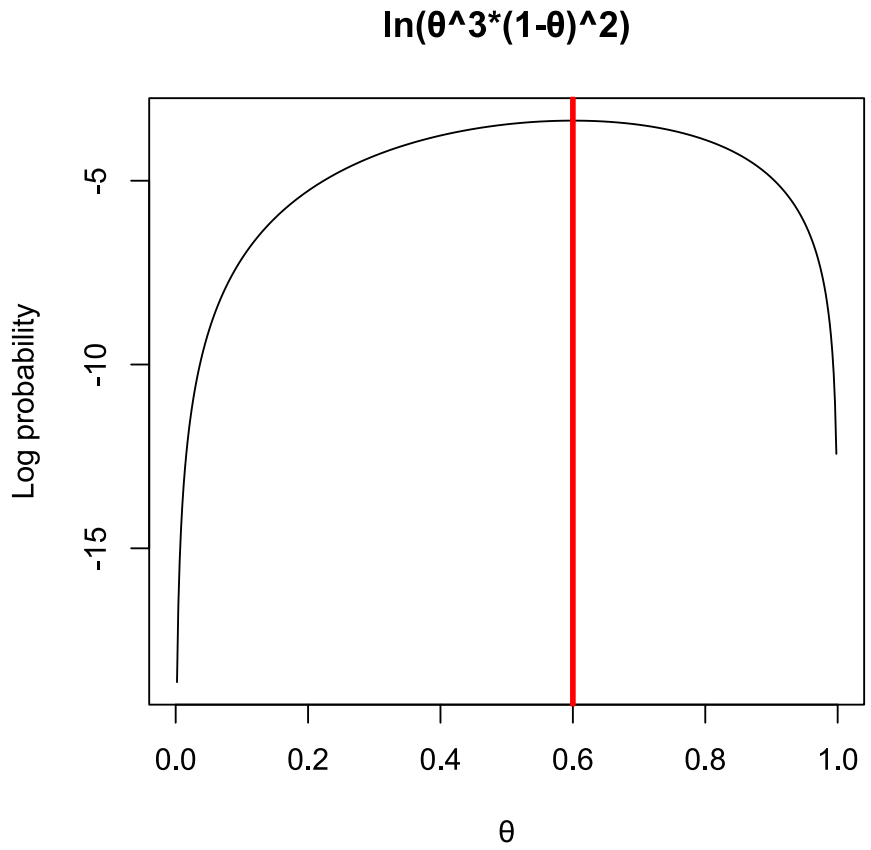
Want to find the value of θ that maximizes $\theta^3 * (1-\theta)^2$. So, we must solve for θ when $\frac{d}{d\theta}\theta^3 * (1-\theta)^2 = 0$. Because ln is monotonically increasing, we can instead solve $\ln(\frac{d}{d\theta}\theta^3 * (1-\theta)^2) = 0$ for θ.

$$
\begin{aligned}
0 &= \frac{d}{d\theta}\ln(\theta^3 * (1-\theta)^2) && < \text{problem statement} > && (1)\\
&= \frac{d}{d\theta}\ln(\theta^3) + \ln((1-\theta)^2)) && < \ln(ab) = \ln a + \ln b > && (2)\\
&= \frac{d}{d\theta}3\ln(\theta) + 2\ln(1-\theta) && < \ln a^b = b\ln a > && (3)\\
&= \frac{3}{\theta} + \frac{2}{1-\theta} * -1 && < \ln \text{rule, chain rule, power rule} > && (4)\\
\frac{2}{1-\theta} &= \frac{3}{\theta} && < \text{algebra} > && (5)\\
3 - 3\theta &= 2\theta && < \text{algebra} > && (6)\\
3 &= 5\theta && < \text{algebra} > && (7)\\
\theta &= \frac{3}{5} && < \text{algebra} > && (8)
\end{aligned}
$$

- So, can we solve now? Yes! See left
- Do you need to ever do this in this class? No!!!

Visual check: 3/5 is it!





The larger picture

- That was very complicated for a simple answer
  - So, why the #!@#% did I show it to you?
- This is (conceptually) how the statistical/ML systems we will be working with calculate things for us
  - They are trying to choose numbers that maximize the probability of **the events we've seen**
- Please note: I will **not** ask you to manually calculate MLE!!!!
  - That's what other people's software is for! ☺

Linguistic relevance?

- What, exactly, does coin flipping have to do with linguistics or language data?
- In fact, a lot! Many linguistic phenomena are binary in nature (or semi-reasonably treated as such)
  - Stop aspiration in English
  - Using "that" in English ("The thing (that?) you want")
- Many classifications are binary in nature (or coerced to be as such)
  - Is this text in Chinese, or not?
  - Is this review good, or bad?