CSE103: Introduction to Probability and Statistics

Prof. Yoav Freund

Why should you care about prob&stat?

- Learn to reason under uncertainty.
- Navigation software:
  - **Certainty**: Find the shortest route from UCSD to UCLA.
  - **Uncertainty**: Find the fastest route from UCSD to UCLA.
- Search Engine:
  - **Certainty**: Find all web pages that contain the words "Trump","Hillary" and "debate"
  - **Uncertainty**: Find the 10 most relevant pages for the query "Trump, Hillary debate"
- Insurance Company:
  - **Certainty**: If a person with life insurance dies, the insurance company has to pay the family $X
  - **Uncertainty**: What is the minimal life insurance premium such that the probability that the life insurance company will be bankrupt in 10 years is smaller than 1% ?

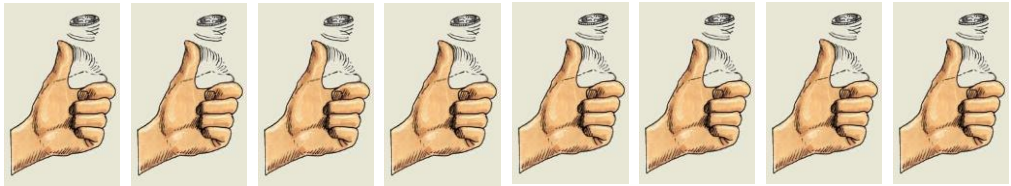
What you will learn

- The navigation and search engine problems are advanced, in this class you will learn the conceptual foundation of P&S and solve simple problems of reasoning under uncertainty.
- Examples:
  - If you flip a coin 100 times, what is the probability of getting at most 10 "heads" ?
  - What is the probability of getting a "4 of a kind" hand in poker.
  - If you want to hash 1,000,000 elements and can allow more than 5 indirections for only 10 elements, how big does the table need to be?
  - Suppose that the expected time between failures for a router is one year. What is the probability that the router will fail during the first month?

Flipping a coin

0+ 1+ 0 +0 +0+ 0 +0 +1=2 heads

T H T T T T T H



Suppose that we flip the coin 100 times, and count the number of heads

Flipping two dice

- 1 dice-What is the probability that it will land on 6 ? Or on 5 ?
  - 1/6
- G,R dice. What is the probability of green=6 and red=5
  - 1/6 X 1/6 = 1/36
- R,R dice. What is the probability of red=5 and red=6?
  - 2 X 1/36 = 1/18
- R,R dice. What is the probability of red=5 and red=5 ?
  - 1/36
- Dice of the same color are

indistinguishable or interchangeable

Indistinguishability / Exchangeability

- Two object are indistinguishable if exchanging them makes not difference.
- What makes two objects indistinguishable?
- Mathematical objects (points, lines) are indistinguishable
  - Physical objects are distinguishable (we can mark them)
- Poker cards are indistinguishable, unless marked (illegal)
- iPhones (of the same model) are indistinguishable
  - iPhone covers make the iPhones distinguishable.
- Dollars are indistinguishable – that is what makes the economy work (compare that to bartering).
- Are fruit indistinguishable? (same DNA)
- Are animals of a species distinguishable?
- Are people distinguishable?

Probabilities regarding people

- Which of the following is more correct?
  1. Each of us is unique, we have our own free will.
  2. We belong to groups, our opinions are the opinions of the group.
- When the number of people is large, a very effective way to reason is to think of people as interchangeable:
  1. How many children in this district have special needs?
  2. How many voters in San Diego county will vote republican?
  3. Does using seat belts save lifes?
- Are we all the same or are we all different?

We are individuals – The life of Brian

Concepts, not formulas!

- The most important skill that you will learn in this class is the ability to analyze problems that involve uncertainty and express your understanding in a precise mathematical way.
- Computing the correct final answer is much less important!
- Different formulas are relevant for Machine Learning, Search engines, route planning, medical research, computer vision
  - You can learn those later if you understand the fundamental concepts.
- The fundamental concepts remain the same: Probability, expectation, central limit theorem …
- I supply a cheat sheet with all of the relevant formulas,
  - You don't need to memorize formulas.
  - You need to understand the concepts!

What you should expect from this class

- This is a math Class!
  - This is a hard class, requires work, requires communication.
  - No Programming.
  - When solving problems, you need to find a short expression, the final numerical answer is less important.
  - A computer can easily compute the value of an expression, the computer cannot translate the problem into an expression.
- Concepts, not formulas:
  - Probability distribution
  - Random variables
  - Dependence
  - Expectation and Variance
  - …
- The examples I give in class and the examples in your homework will involve different numbers and sometimes different formulas, but they will use the same concepts.
- Concepts are harder than recipes: they require understanding, not just plugging in.

Tips on Learning concepts

- Words in math have precise meaning, be sure you understand that meaning.
- Don't let new words/concepts wash over you.
  To do the HW, you will need to understand those concepts.
- **Ask questions!**
  - Ask questions during Class.
  - Ask questions during Discussion.
  - Form study groups and discuss concepts.
  - Post questions on Piazza.
- Asking good questions will earn you points!

At Your Service

Our job is to answer your questions!

Instructor Yoav Freund

TAs

TA Zhen Zhai

TA Akanksha Maurya

TA Jigarkumar Pankajbhai Patel

TA Shen Ting Ang

TA Sunil Ramnik Raiyani

Tutors

Tutor Chang Qiu

Tutor Geovonni Allen Najera

Tutor Han Liu

Tutor Tony Wijaya Salim

Tutor Vrushali Mahesh Samant

How to earn points for asking questions

- Ask a question during class/section
  - Instructor/TA decides if this is a good question "POP=Post On Piazza"
    - Example of a good question: "Why do you take the product of the number of possibilities?"
    - Example of a weak question: "Why is 6 not the right answer?"
  - If POP:
    - Instructor writes down question and name.
    - Student posts question and answer on Piazza, with "POP" in title.
    - Instructor marks the question as "good question"
    - Instructor adds 5 points to student's class participation score.

Class Mechanics

- Final Grade:
  - 50% HW assignments. Bottom 2/9 HW grades dropped.
  - 40% Final
  - 20% Classroom participation (asking good questions).
- HW:
  - Delivered using the open-edx platform.
  - Assignments will open and close at noon on Friday.
  - You can attempt each problem several times:
    - Initially: unlimited
    - Close to final: limited.
- Hint system:
  - Will sometimes send you a hint to help you towards the answer.
  - The most common hint: "Please write an expression, not just the final result".
  - Part of a research project: **we will ask for your consent to use results in publication**.
  - Published results are anonymized.

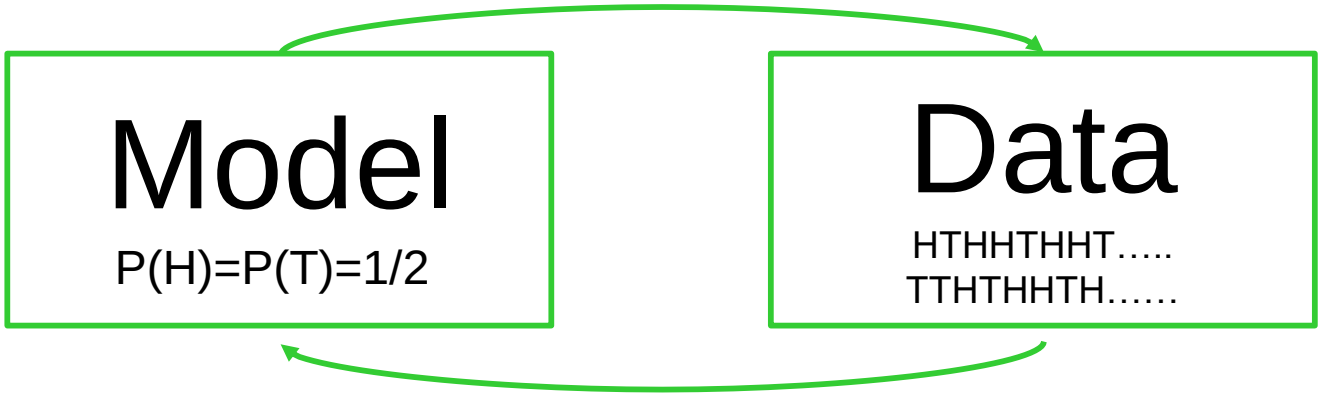
The mechanics (continued)

- Final:
  - similar questions to the HWs.
  - But you get only one attempt.
  - 3 hours, most finish in < 2hours.
- Classroom participation:
  - Each good question earns you 5 points (to a max of 20 points)
  - If you asked a good question in class/section, post the question and the answer to it on Piazza.

Probability vs. statistics

Probability

What is the probability that 60 out of 100 flips of fair coin are heads?

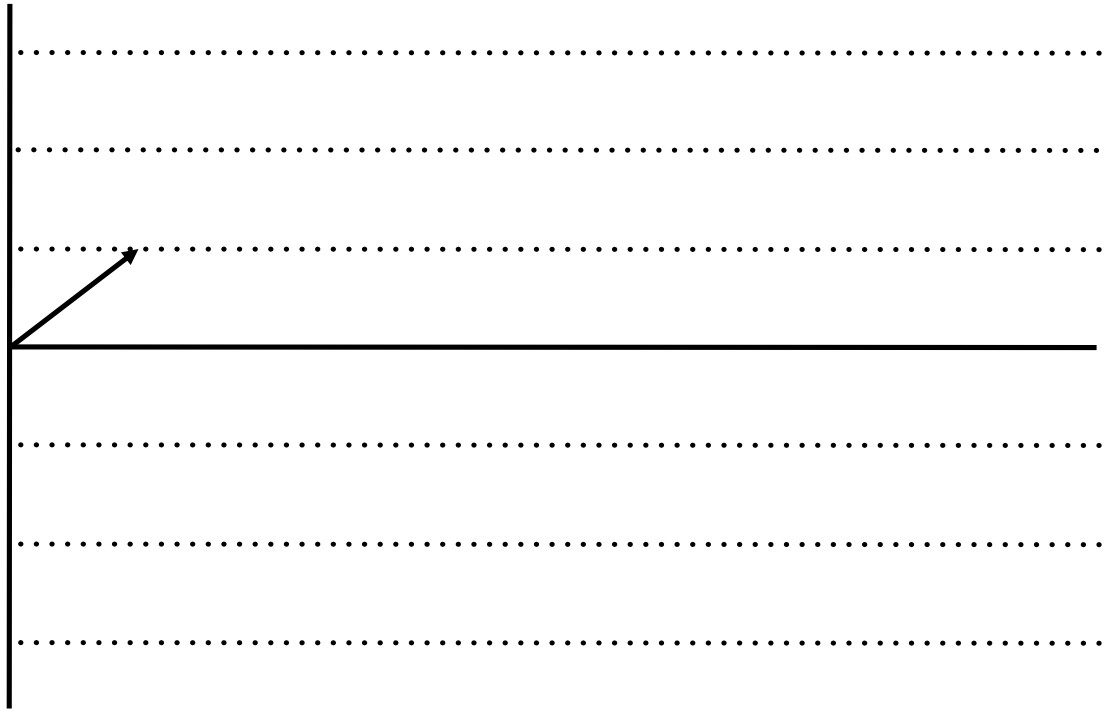


Statistics

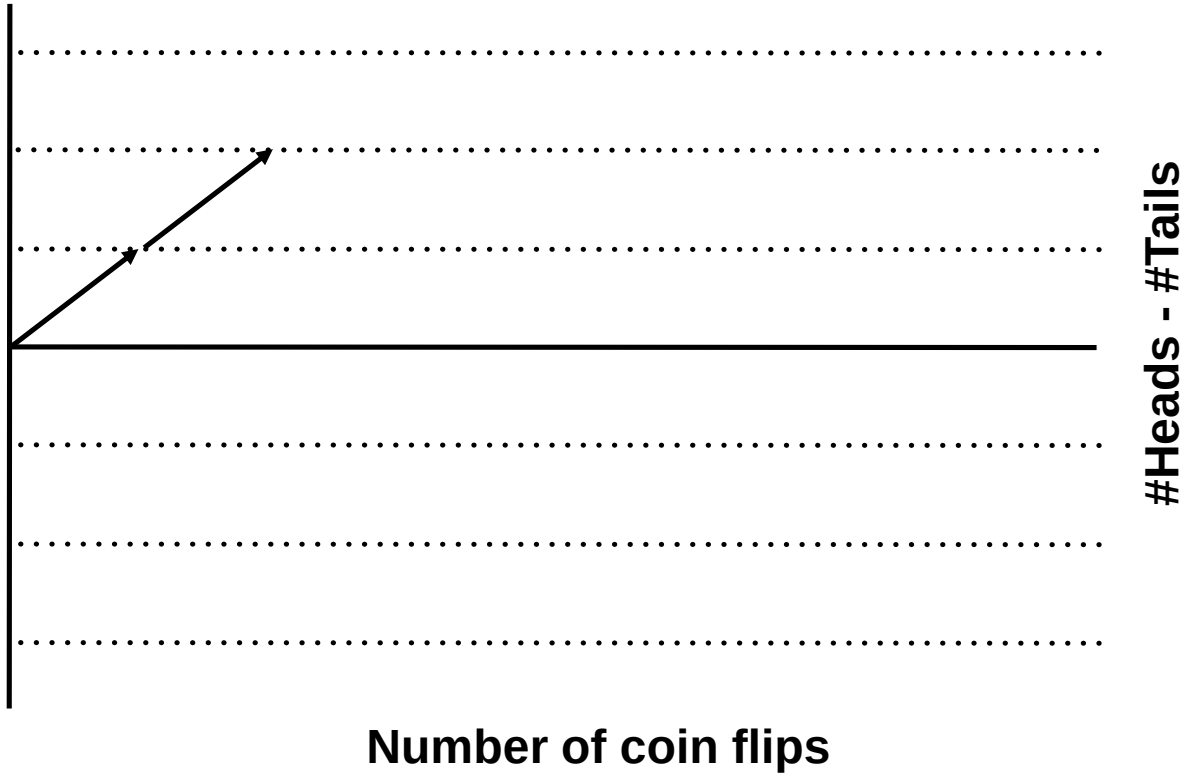
Given that 60 out of 100 coin flips are "heads", how sure can we be that the coin is not fair?

PROBABILITY

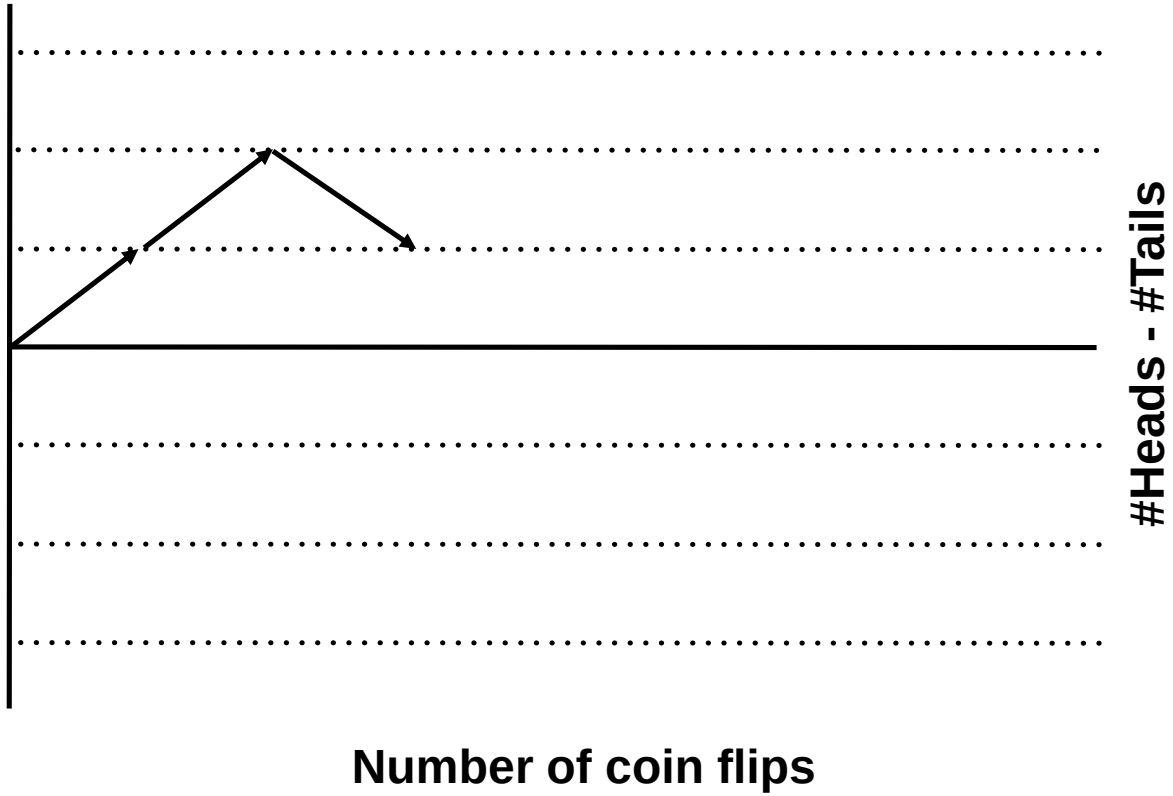
A Basic Problem: Is a Coin Fair?



A Basic Problem: Is a Coin Fair?



A Basic Problem: Is a Coin Fair?



A Basic Problem: Is a Coin Fair?

Null hypothesis H_0: $h = 0.5$ (coin is fair)

Alternative hypothesis H_1: $h \neq 0.5$ (coin is biased)

The law of large numbers

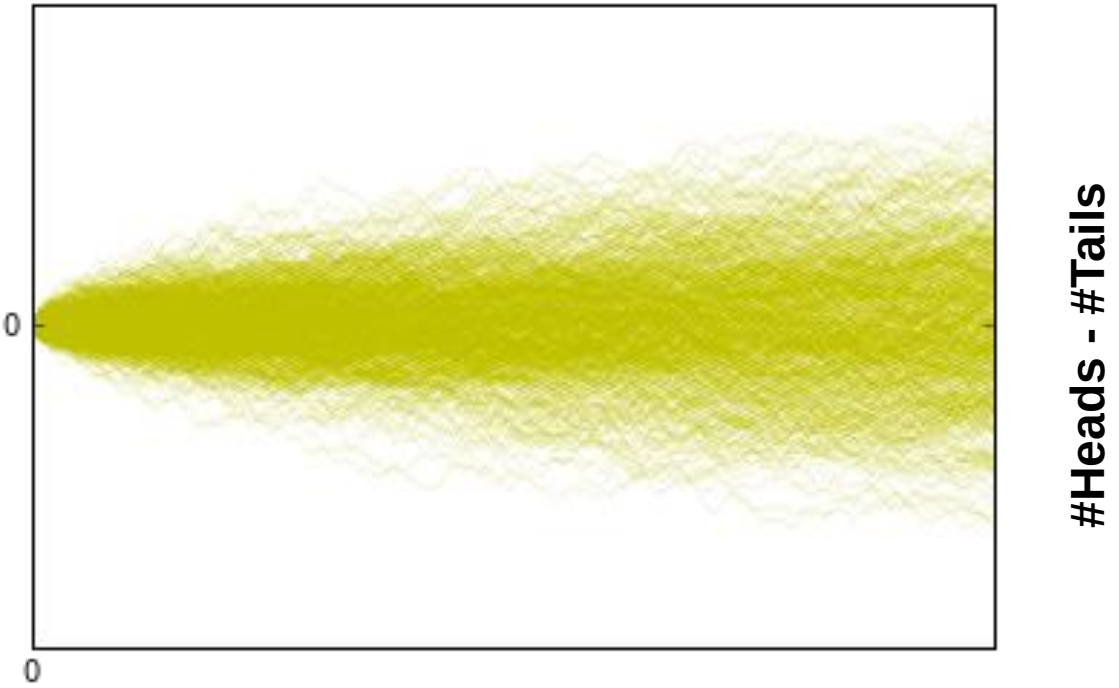
- If we flip forever, the fraction of times that we get "heads" converges to the true bias of the coin.
- This is called "the law of large numbers"
- If bias =1/2 then the coin is fair.
- But we cannot wait forever!
- How many times do we need to flip the coin in order to know whether the true bias is between 0.45 and 0.55 ?
- We will give increasingly more accurate answers to this question throughout the quarter.
- For now, consider some examples.

Statistical terminology and notation

- h = the true bias of the coin = probability of Heads.
- Null hypothesis
  - = our default assumption
  - = the coin is unbiased
  - = the bias of the coin is between 0.45 and 0.55
- Alternative hypothesis
  - = the coin is biased
  - = the bias of the coin is outside 0.45 and 0.55

Fair Coin (Null Hypothesis)

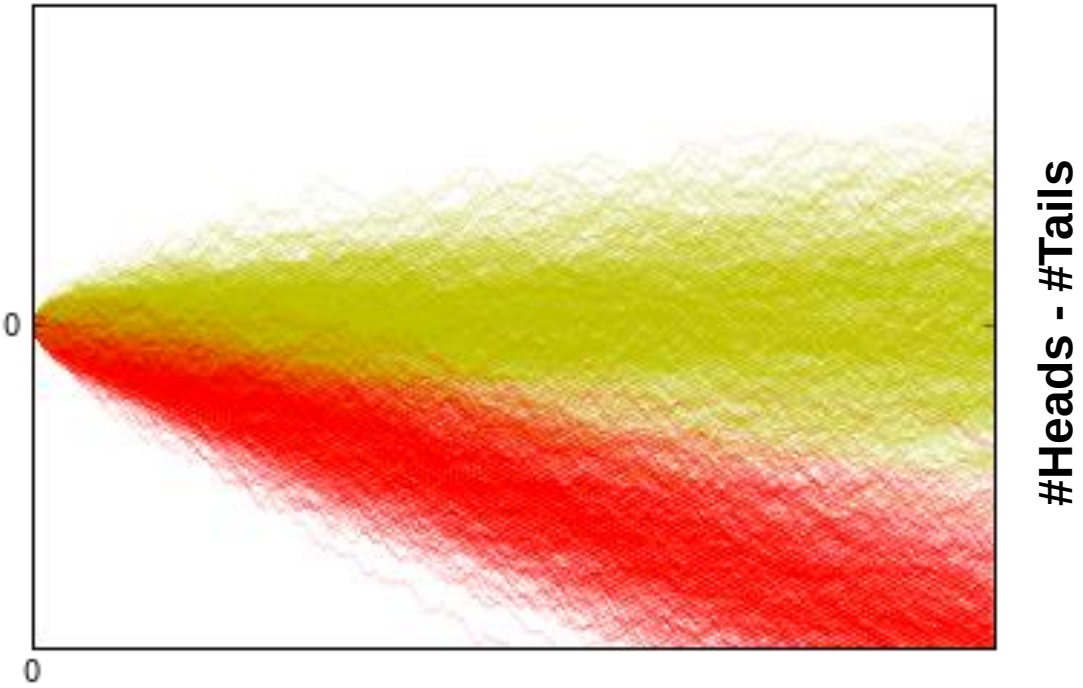
h = 0.5



Biased Coin (Alternate Hypothesis)

h = 0.5

h = 0.4

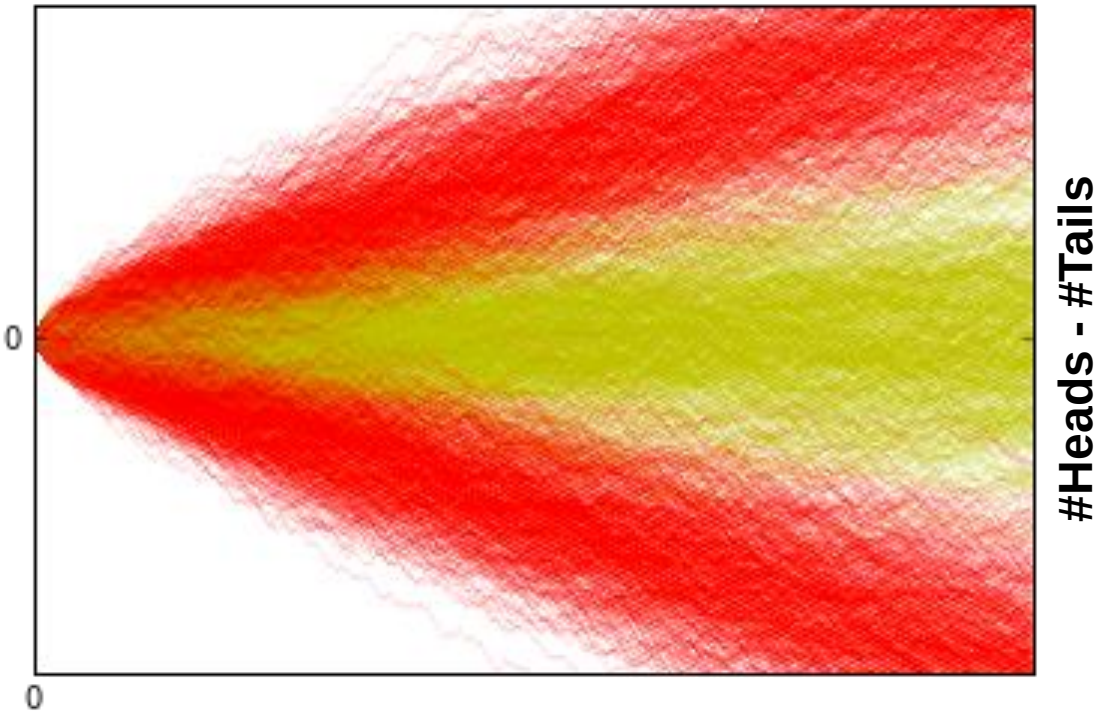


Biased Coin (Alternate Hypothesis)

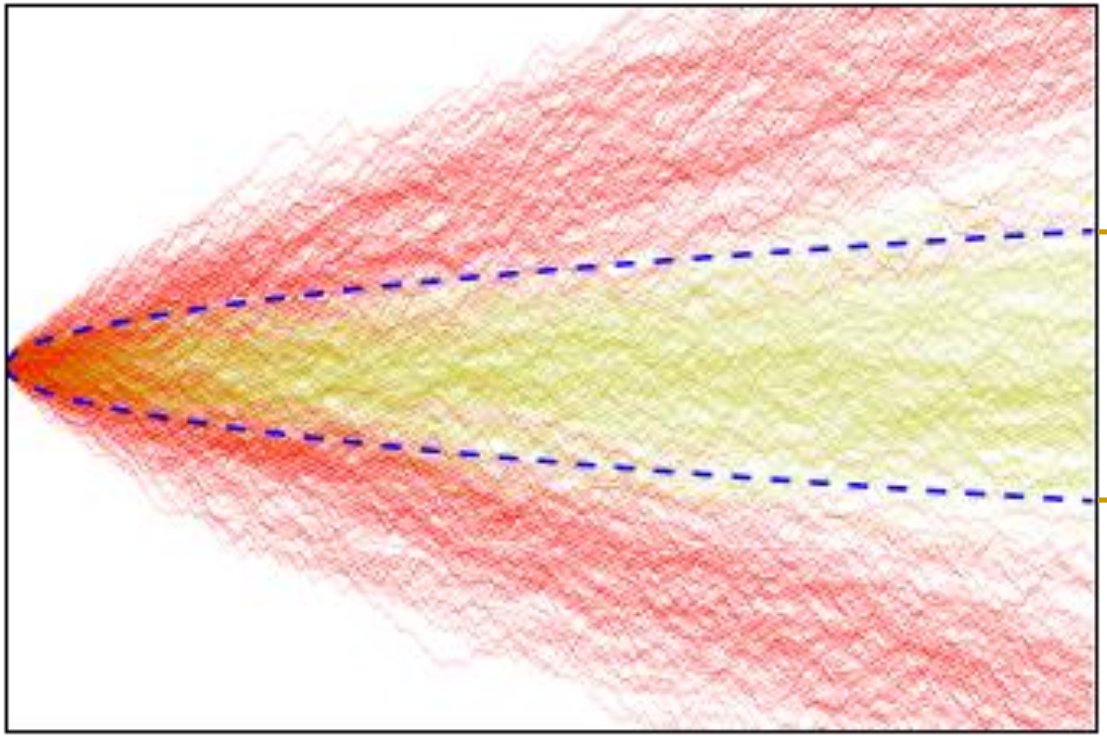
h = 0.6

h = 0.5

h = 0.4



Testing a Coin



Reject null
(i.e. coin is unfair)

Can't reject null
Don't know whether
or not coin is fair

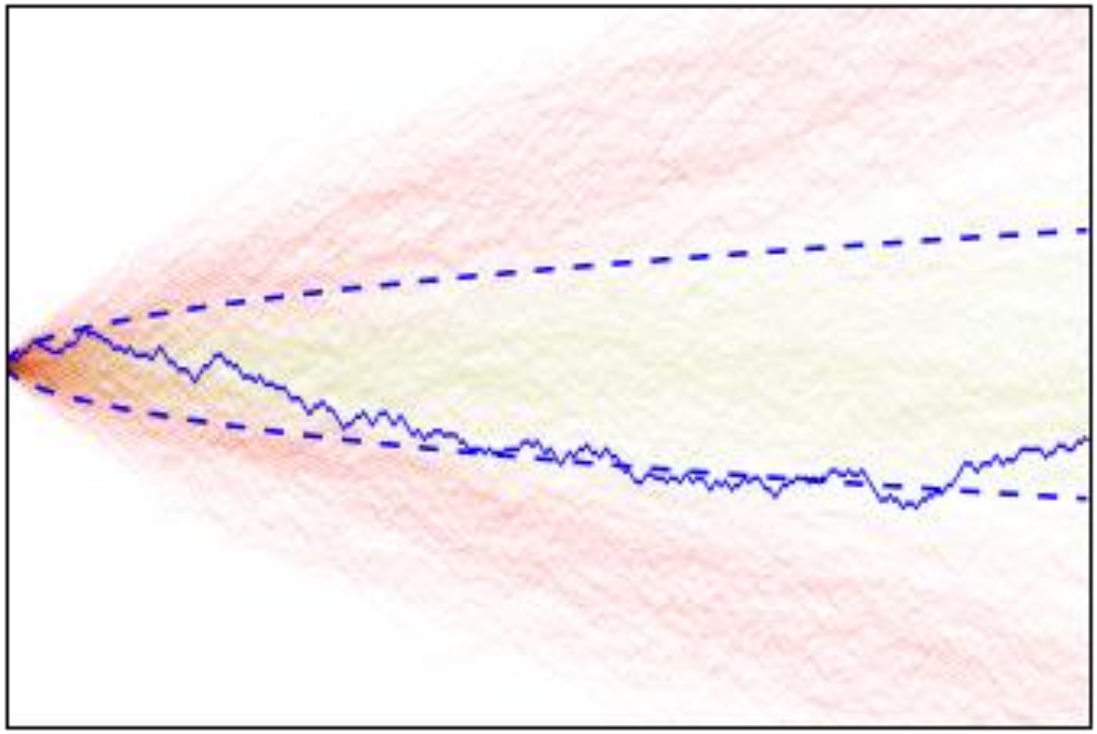
Reject null
(i.e. coin is
unfair)

Number of coin flips

N flips

Testing a Coin

Remember that our test is **a single** random path



Reject null (i.e. coin is unfair)

Reject null (i.e. coin is unfair)

Number of coin flips

N

STATISTICS

Who cares about coin flips?

- Gamblers, sport referees, …
- Coin flips, especially biased coin flips, can be used to represent many types of random events:
  1. Car insurance: T=day without accident, H=day with accident.
  2. Polls: T=polled leans republican, H= polled leans democratic
  3. Poker: T=your hand is the highest, H=your hand is not the highest
  4. Internet: H=the sent packet will arrive within 1 second, T= it will not arrive within 1 second.
- Each of these examples can be seen as a flip of a biased coin.
  - 1,2,4 are case in which you want to estimate the bias from data (Statistics)
  - 3: Poker is a case in which you want to calculate the bias from first principles (Probability)

Stents for preventing heart attacks

Wingspan Stents

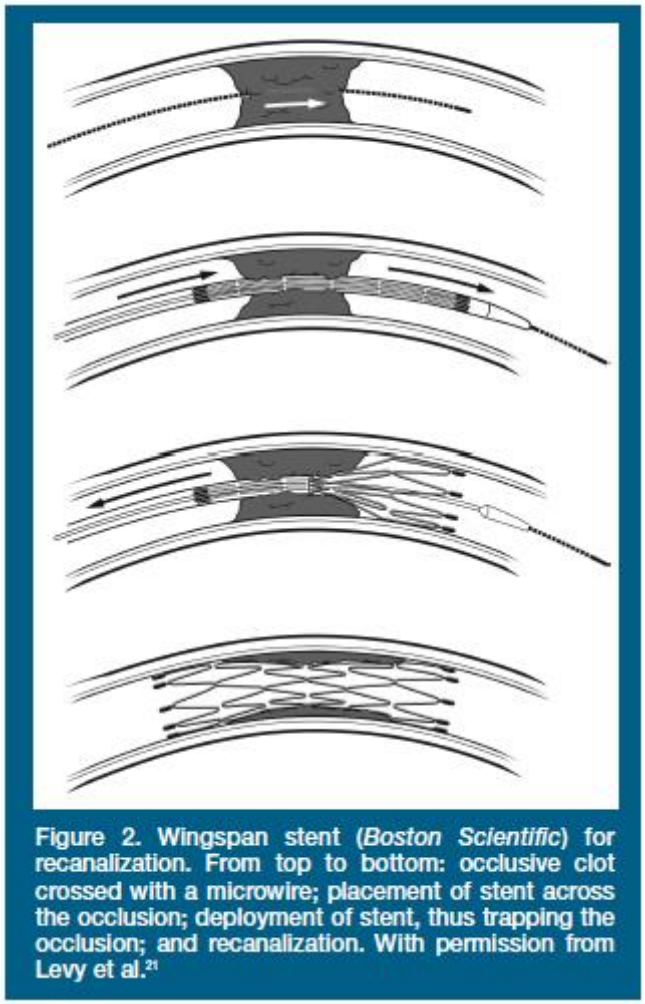


Figure 2. Wingspan stent (*Boston Scientific*) for recanalization. From top to bottom: occlusive clot crossed with a microwire; placement of stent across the occlusion; deployment of stent, thus trapping the occlusion; and recanalization. With permission from Levy et al.[21]

- A stent is a medical device that is inserted into a blood vessel with a clot and expands to remove the clot.
- Stents have been shown to help patients after a heart attack by reducing the chance of additional heart attacks.
- Doctors hoped that stents can be used to prevent heart attacks (before an attack has occurred).
- A study was conducted, involving 451 patients at high risk for heart attack.

The study

- Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. New England Journal of Medicine 365:993- 1003.
  - More details in the book "Open Intro to Statistics"
- 451 patients with high risk of heart attack.
  - Divided randomly to "Treatment" and "Control"
  - Treatment received standard care and a Stent
  - Control received standard care but no Stent

Patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
⋮	⋮	⋮	
450	control	no event	no event
451	control	no event	no event

Table 1.1: Results for five patients from the stent study.

The results of the study

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Table 1.2: Descriptive statistics for the stent study.

- Within 30 days:
  - proportion of stroke in treated patients: 33/224=0.15=15%
  - Proportion of stroke in control patients: 13/227=0.06=6%
- Within 1 year:
  - proportion of stroke in treated patients: 45/224=0.2=20%
  - Proportion of stroke in control patients: 28/227=0.12=12%
- Clearly, we cannot say that using the stent helps.
- Can we say that using a stent hurts?
  - Not necessarily.
  - The numbers might be a random fluctuation.
  - There can be other factors, such as the procedure for inserting the stent.
  - The chosen patients might not be representative.
  - …

For next class

- Make sure you have accounts on:
  - Open-edx (ID@ucsd.edu, password=PID)
  - Piazza
- Read chapter 1 of the class notes, comment in places that you don't understand.
- Post your questions on Piazza
- See you on wed!