What is statistics?

Probability theory computes probabilities of complex events given the underlying base probabilities.

Statistics takes us in the opposite direction.

We are given **data** that was generated by a **Stochastic process**

We **infer** properties of the underlying base probabilities.

Example: deciding whether a coin is biased.

In a previous video we discussed the distribution of the number of heads when flipping a fair coin many times.

Let's turn the question around: we flip a coin 1000 times and get 570 heads.

Can we conclude that the coin is biased (not fair) ?

What can we conclude if we got 507 heads?

The Logic of Statistical inference

The answer uses the following logic.

- Suppose that the coin is fair.
- Use **probability theory** to compute the probability of getting at least 570 (or 507) heads.
- If this probability is very small, then we can **reject** with confidence the hypothesis that the coin is fair.

Calculating the answer

Recall the simulations we did in the video "What is probability".

We used $x_i = -1$ for tails and $x_i = +1$ for heads.

We looked at the sum $S_k = \sum_{i=1}^{k} x_i$, here $k = 1000$.

If number of heads is 570 then $S_{1000} = 570 - 430 = 140$

It is very unlikely that $|S_{1000}| > 4\sqrt{k} \approx 126.5$

In [1]:
```
from math import sqrt
4*sqrt(1000)
```

Out[1]:
```
126.49110640673517
```

It is very unlikely that the coin is unbiased.

What about 507 heads?

507 heads = 493 tails $\Rightarrow S_n = 14, \quad 14 \ll 126.5$

We cannot conclude that coin is biased.

Conclusion

The probability that an unbiased coin would generate a sequence with 570 or more heads is extremely small. From which we can conclude, with high confidence, that the coin **is** biased.

On the other hand, $\left|S_{1000}\right| \geq 14$ is quite likely. So getting 507 heads does not provide evidence that the coin is biased.

Real-World examples

You might ask "why should I care whether a coin is biased?"

- This is a valid critique.
- We will give a few real-world cases in which we want to know whether a "coin" is biased or not.

Case I: Polls

- Suppose elections will take place in a few days and we want to know how people plan to vote.
- Suppose there are just two parties: **D** and **R**.

- We could try and ask **all** potential voters.

- That would be very expensive.

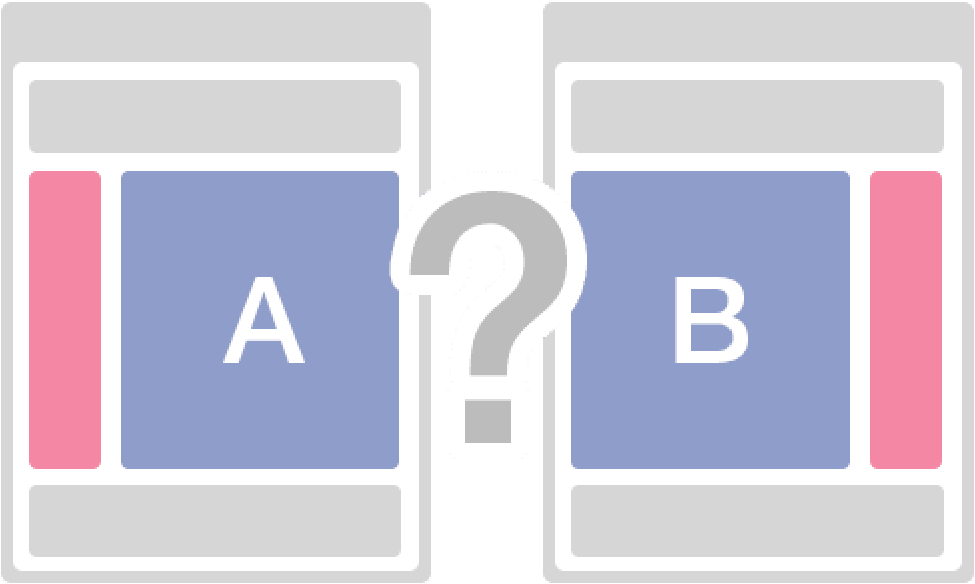
- Instead, we can use a poll: call up a small randomly selected set of people.

- Call n people at random and count the number of **D** votes.
- Can you say with confidence that there are more **D** votes, or more **R** votes?
- Mathematically equivalent to flipping a biased coin and
- asking whether you can say with confidence that it is biased towards "Heads" or towards "Tails"

Case 2: A/B testing

A common practice when optimizing a web page is to perform A/B tests.

- A/B refer to two alternative designs for the page.



- To see which design users prefer we randomly present design A or design B.

- We measure how long the user stayed on a page, or whether the user clicked on an advertisement.

- We want to decide, with confidence, which of the two designs is better.

- Again: similar to making a decision with confidence on whether "Heads" is more probably than "Tails" or vice versa.

Summary

Statistics is about analyzing real-world data and drawing conclusions.

Examples include:

- Using polls to estimate public opinion.
- performing A/B tests to design web pages
- Estimating the rate of global warming.
- Deciding whether a medical procedure is effective

The end!