Foundations of Data Science: Randomness, sampling and simulation – Introduction to statistical inference
Where are we in the course?

I. About data: collection, representation, wrangling, exploration, visualisation and descriptive stats

II. Intro to Machine Learning

III. Linear models

IV. Statistical Inference

V. Regression and inference
Descriptive statistics
Inferential statistics

The process of drawing conclusions about quantities that are not observed.

E.g. "Manuscript on Deciphering Cryptographic Messages"
Al-Kindi, 9th Century, Baghdad

We observe the mean of a sample
We infer the mean of the population

We infer the meaning of the messages
Inferential statistics tasks

1. Estimation

2. Hypothesis testing

3. Comparing two samples (A/B testing)
Inferential statistics tasks: Estimation

\[ [30.4 g, 33.6 g] \leq \text{CI} \]

\[ \hat{\mu} = \bar{x} = 320 g \pm 16 g \quad n = 20 \]

|       | coef | std err | t     | P>|t| | [0.025] | [0.975] |
|-------|------|---------|-------|------|---------|---------|
| Intercept | -382.7372 | 108.680 | -3.522 | 0.001 | -604.692 | -160.783 |
| Length | 3.3515 | 0.503 | 6.661 | 0.000 | 2.324 | 4.379 |

Point estimates

Confidence intervals: how confident are we in the estimate?
Inferential statistics tasks: Hypothesis testing

Yes/no questions: E.g. 1: "Is Chocolate good for you"

E.g. 2: Swain versus Alabama (1965).
Is this jury selection procedure biased?

Population of Alabama
26% Black
74% Non-black

Jury panel of 100:
8 Black and 92 Non-black
Inferential statistics tasks:
Comparing two samples (A/B testing)

E.g. 1. Is a vaccine better than a placebo?
E.g. 2. Does providing recommendations lead to more sales?
Two approaches to statistical inference

1. Computational: "Statistical simulations"
   + Few assumptions $\Rightarrow$ can be applied to many situations
   + Little theory required
   + Hopefully intuitive
     - Can be compute-intensive

2. Mathematical: Statistical theory
   + Not compute-intensive
   + Standard in scientific literature
   - Can depend on assumptions that aren't true (e.g. normal distributions)
Plan for statistical inference

1. Randomness, sampling and simulations (S1 Week 10)
2. Estimation, including confidence intervals (S1 Week 11)
3. Hypothesis testing (S2 Week 1)
4. Logistic regression (S2 Week 1)
5. A/B testing (S2 Week 2)
How can we address these questions?

1. What is the mean and median age of the population of all 2p and 10p coins in circulation?
2. Are tosses of 2p and 10p coins biased, i.e. is the probability of heads or tails different from 1/2?
Let's get sampling!

1. Go to the form at the right
2. Record the
   - denomination (2p/10p)
   - style (old/new)
   - year
3. Toss the coin 8 times and record the results
4. Submit the form

https://forms.office.com/e/SKNgiQmB4N
Results

How certain are we that the mean year is what we compute?

Do we think that the coins are biased or not?