Inf2 – Foundations of Data Science
S2 Week 1: Semester 2 Logistics

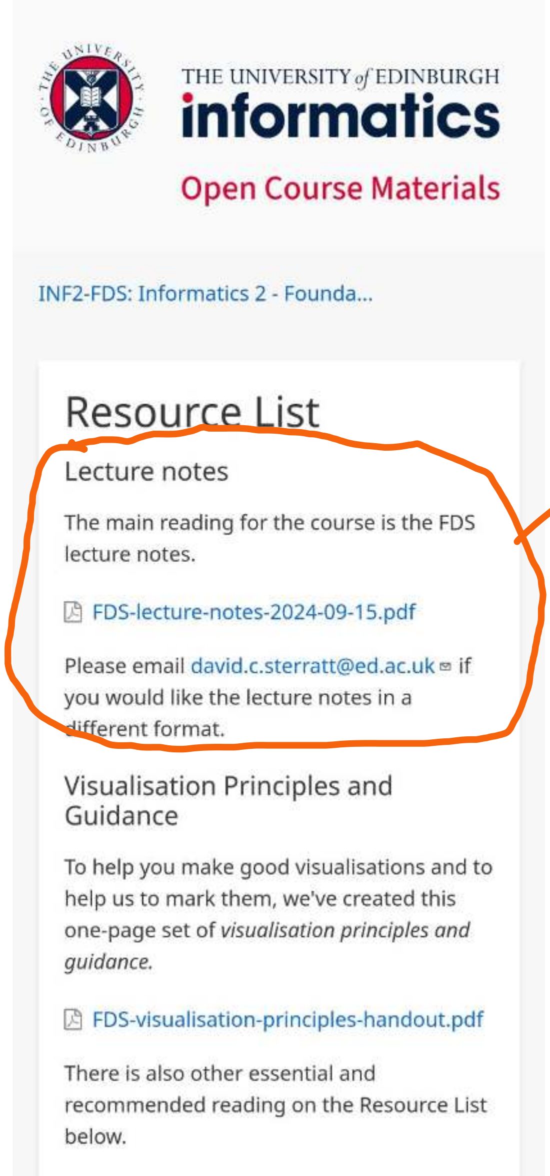
THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Semester 2 logistics

- **Lectures until week 6**
  - On **Statistical Inference**, **Maximum Likelihood and Regression**, **Ethics** and **Software Engineering**
  - Accompanying **Comprehension Questions** in Learn
- **Labs:** Lab notebooks for weeks 1, 2 and 4, and one on web-scraping
  - No lab sessions – ask on Piazza
- **Workshops:** in weeks **3**, 4 and 6
  - May be in a new group or at a different time
  - To change group, use Group Change Request Form; turn up if change not actioned.
- **CW2 – Project** from week 5 to week 10
  - **including opportunities to present in workshop sessions in weeks 8 and 10**
  - In response to feedback, earlier release than previous years

Resources



THE UNIVERSITY *of* EDINBURGH
informatics
Open Course Materials

INF2-FDS: Informatics 2 - Founda...

Resource List

Lecture notes

The main reading for the course is the FDS lecture notes.

FDS-lecture-notes-2024-09-15.pdf

Please email david.c.sterratt@ed.ac.uk if you would like the lecture notes in a different format.

Visualisation Principles and Guidance

To help you make good visualisations and to help us to mark them, we've created this one-page set of *visualisation principles and guidance.*

FDS-visualisation-principles-handout.pdf

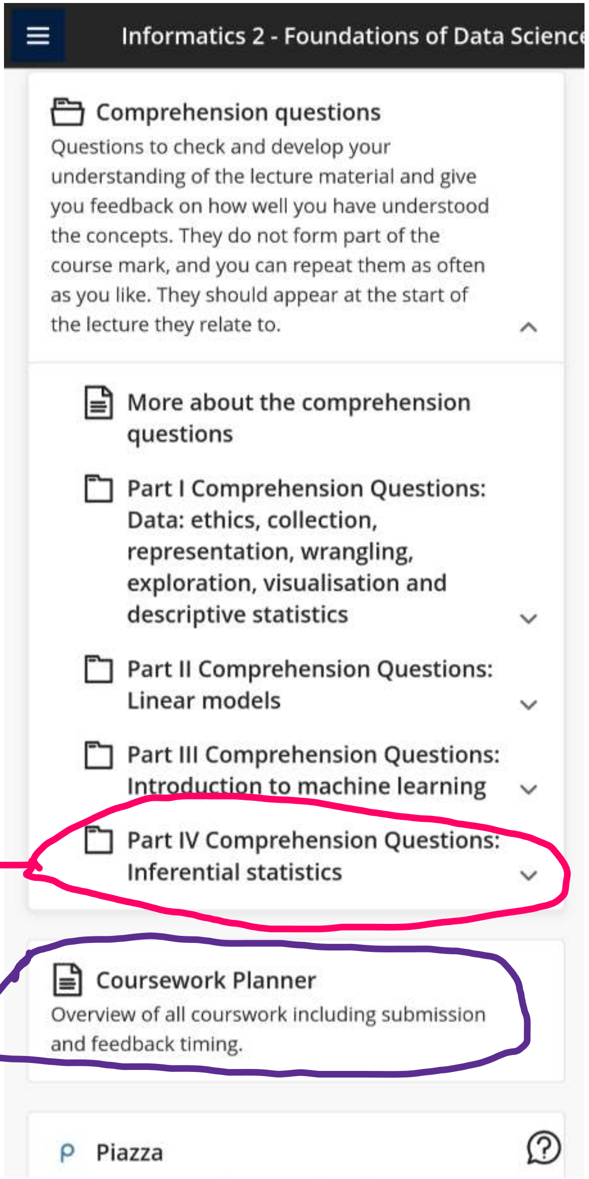
There is also other essential and recommended reading on the Resource List below.

Lecture notes:

- Read before each lecture
- Ask David for different formats
- Updated in response to queries!

Comprehension questions: Released each week, should all be do-able after Wednesday lectures

Coursework planner: includes link to show CW deadlines in your Outlook calendar



Informatics 2 - Foundations of Data Scienc

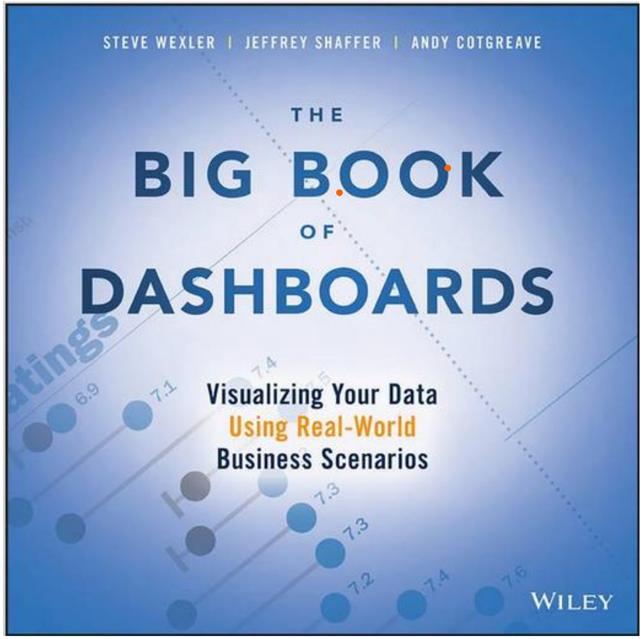
Comprehension questions
Questions to check and develop your understanding of the lecture material and give you feedback on how well you have understood the concepts. They do not form part of the course mark, and you can repeat them as often as you like. They should appear at the start of the lecture they relate to.

- More about the comprehension questions
- Part I Comprehension Questions: Data: ethics, collection, representation, wrangling, exploration, visualisation and descriptive statistics
- Part II Comprehension Questions: Linear models
- Part III Comprehension Questions: Introduction to machine learning
- Part IV Comprehension Questions: Inferential statistics

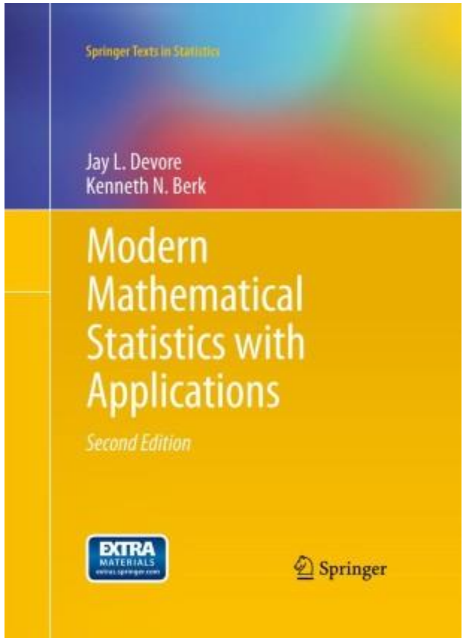
Coursework Planner
Overview of all coursework including submission and feedback timing.

Piazza

Other recommended reading



The Big Book of Dashboards
Ch. 1: Data Visualization: A Primer
Wexler & al. (2017)
Essential reading – in Resource List



Modern Mathematical Statistics with Applications
Devore & Berk
Buy softcover version via University Library for £25

- See **Resource List** and **Schedule** for other essential and recommended reading, including
- Shannon Vallor's **An Introduction to Data Ethics**
- Berkley Data 8 **Inferential Thinking**
- Some recommended reading each week

Support

- Piazza
    - Please try to answer each other's content questions - it helps you all to learn
    - We will try to get to logistics questions and urgent questions by the next working day (not Saturdays or Sundays)
- Office hour
    - Now every Wednesday, 10am in AT8.15
    - Please ring the bell to get in to Level 8, or ping David Sterratt on Teams

Dataset suggestions for final project

- Project:
  - Choose from one of three datasets to explore/analyse
  - Answer a few “seed” questions
  - Share the results in a written data science report
- Examples of datasets used in previous years:
  - Performance of Scottish A&E services
  - Worldwide trends in music streaming according to Spotify
  - Student learning on the EEdi educational platform
- Please add suggestions to the pinned Piazza post "**Request for Dataset suggestions for final project**"
- Deadline: end of week 2. We will then finalise the choices

What the exam (40%) will cover and how to do well

- Your knowledge of good practices for storing, manipulating, summarising and visualising data (Learning Outcome 1)
    - **Revise Semester 1 material, including comprehension questions**
- How well you can apply basic techniques from descriptive and inferential statistics and machine learning and interpret and describe the output from such analyses (Learning Outcome 3)
    - **Do statistical problems tasks and workshops this semester**
    - **Do comprehension questions**
    - **Do labs**
- How well you can evaluate claims made in case study and your understanding of ethical issues (Learning Outcome 4)
    - **Read target paper and attend workshop in which we'll get to grips with it**

Questions?

Inf2 - Foundations of Data Science: Introduction to statistical inference

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Where are we in the course?

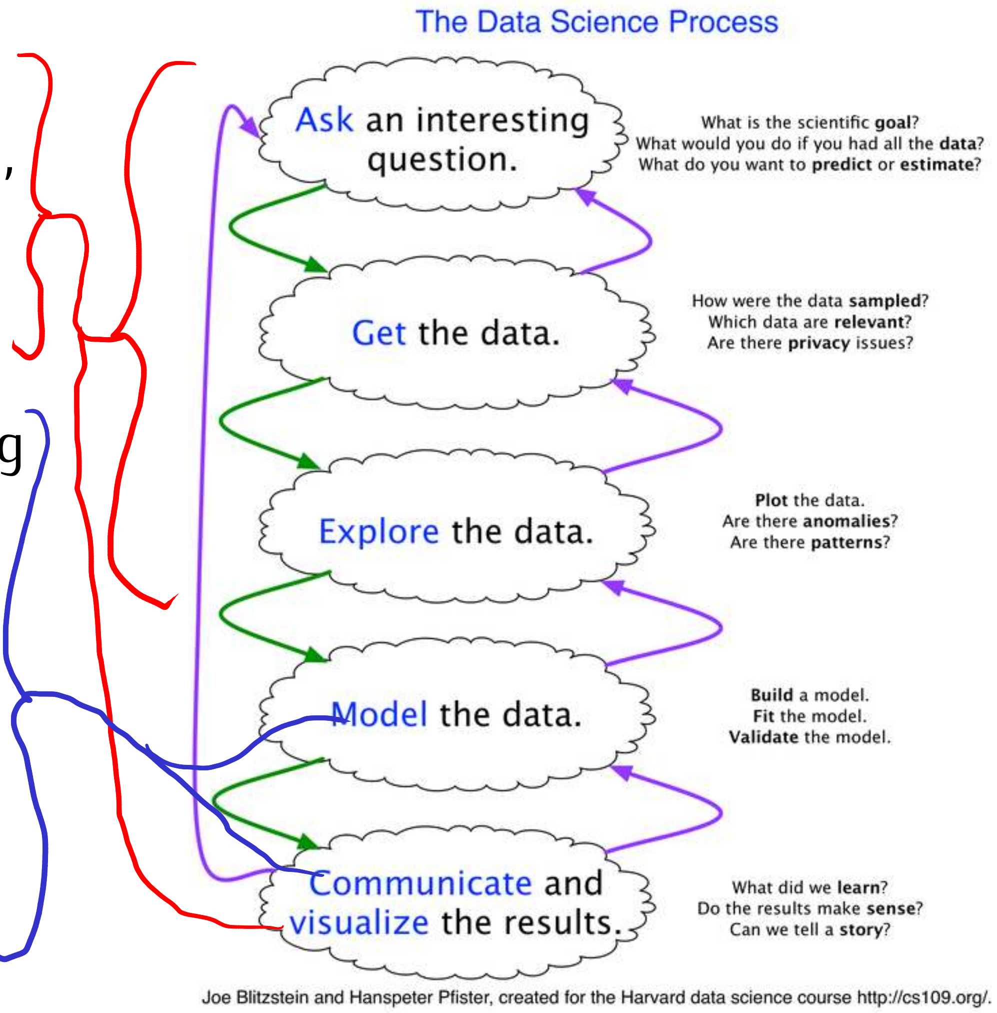
I. Data: ethics, collection, representation, wrangling, exploration, visualisation and descriptive stats

II. Intro to Machine Learning

III. Linear models

IV. Statistical inference

V. Regression and inference

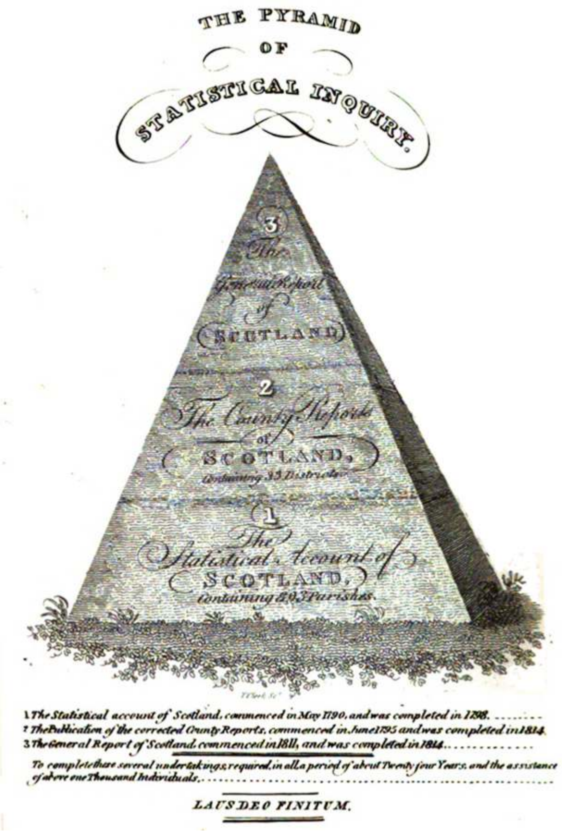


The Data Science Process

Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

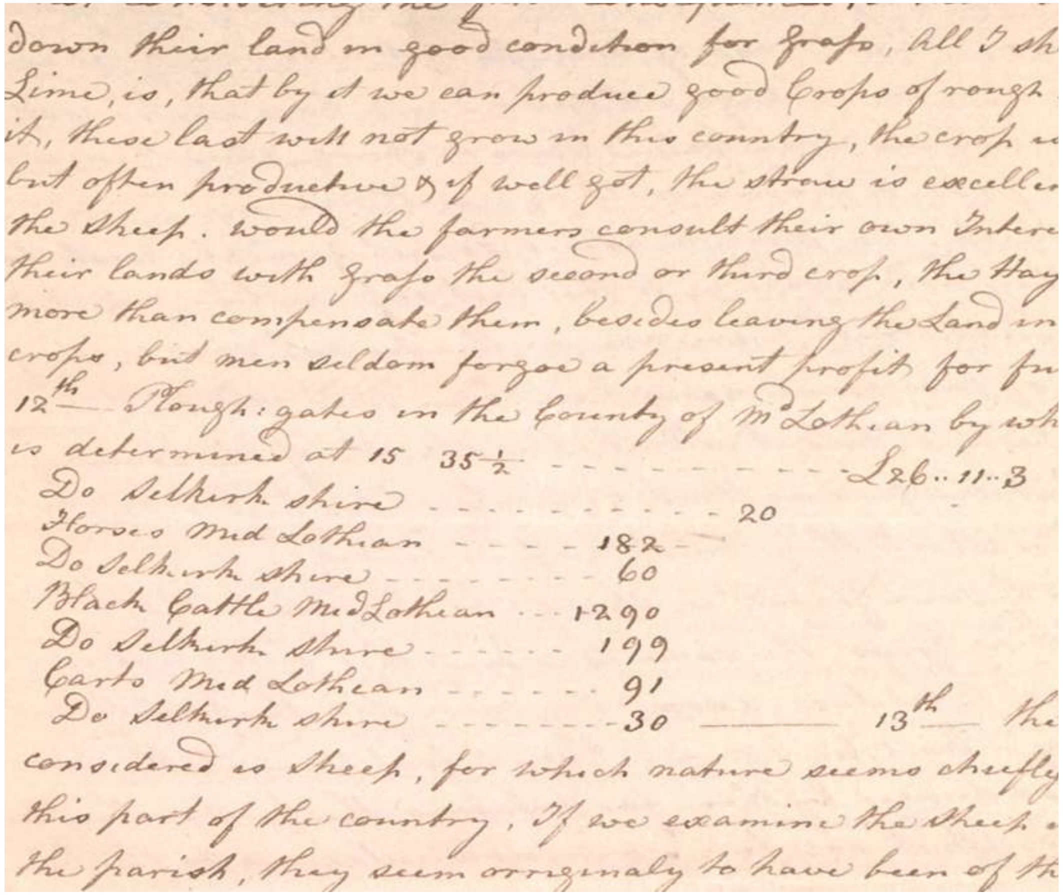
Descriptive statistics


THE PYRAMID OF STATISTICAL INQUIRY.
3
The General Report of SCOTLAND
2
The County Reports of SCOTLAND,
Containing 33 Districts
1
The Statistical Account of SCOTLAND,
Containing 893 Parishes.

1 The Statistical account of Scotland, commenced in May 1790, and was completed in 1798. ....
2 The Publication of the corrected County Reports, commenced in June 1795 and was completed in 1814.
3 The General Report of Scotland, commenced in 1811, and was completed in 1814. ....

To complete these several undertakings, required, in all, a period of about Twenty four Years, and the assistance of above one Thousand Individuals, ...

LAUS DEO FINITUM.



down their land in good condition for graſs, All I sh
Lime, is, that by it we can produce good Crops of rough
it, these last will not grow in this country, the crop
but often productive & if well got, the straw is excelle
the Sheep. would the farmers consult their own Intere
their lands with graſs the second or third crop, the Hay
more than compensate them, besides leaving the Land in
crops, but men seldom forgoe a present profit for fu
12th — Plough: gates in the County of Md Lothian by wh
is determined at 15 35½ ------------------ £26..11..3
Do Selkirk shire ------------------- 20
Horses Mid Lothian ------ 182
Do Selkirk shire ---------- 60
Black Cattle Mid Lothian ... 1290
Do Selkirk Shire -------- 199
Carts Mid Lothian ------- 91
Do Selkirk shire ---------- 30 ——— 13th — the
considered is Sheep, for which nature seems chiefly
this part of the country, If we examine the Sheep
the parish, they seem orriginaly to have been of th

https://stataccscot.edina.ac.uk

Inferential statistics

Statistical inference is the process of drawing conclusions about quanties that are not observed

E.g. Wildcats

Wikipedia, Peter Trimming, CC BY 2.0

We observe the mean of a sample

We infer the mean of the population

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

Wikipedia

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

How big is a quantity, and how certain are we about our answer?

E.g. weight of a population of squirrels from sample of 20

Peter Trimming, CC BY 2.0, Wikipedia

Point estimates

$\hat{\mu} = \bar{x} = 320g$

Confidence intervals: how confident are we in the estimate?

95% confident that

$[304g, 336g]$

contains μ

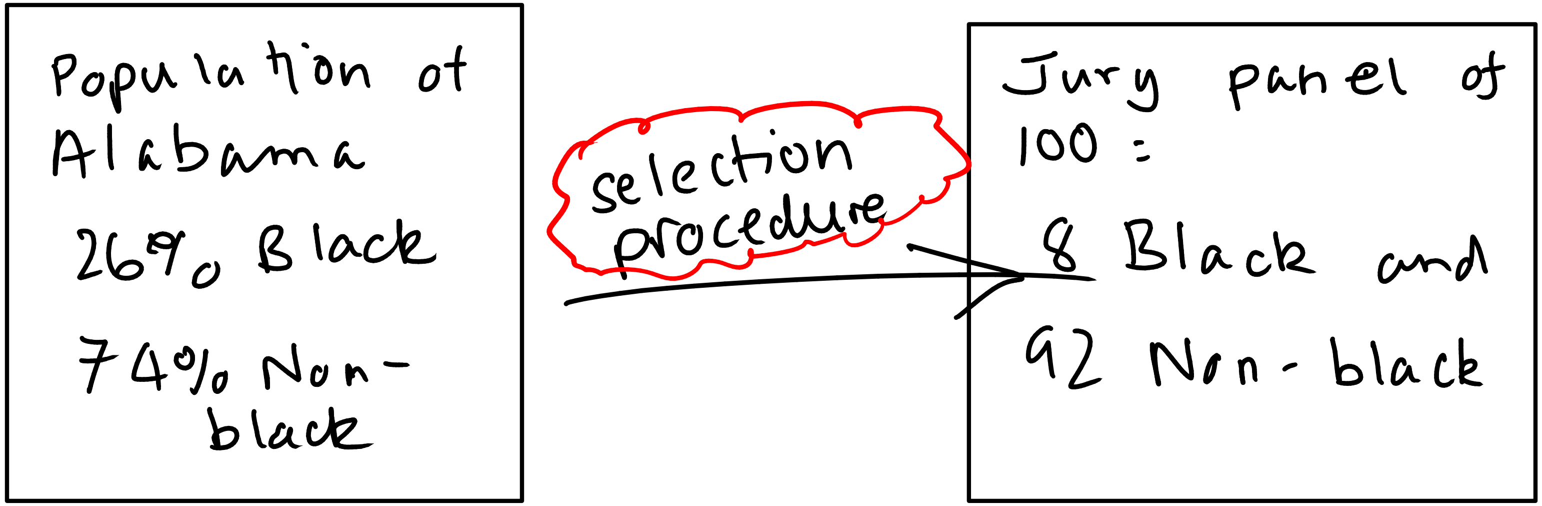
$\mu = 320g \pm 16g, \; n = 20$

E.g. 2: Opinion polling

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

selection procedure

Jury panel of 100 :

8 Black and

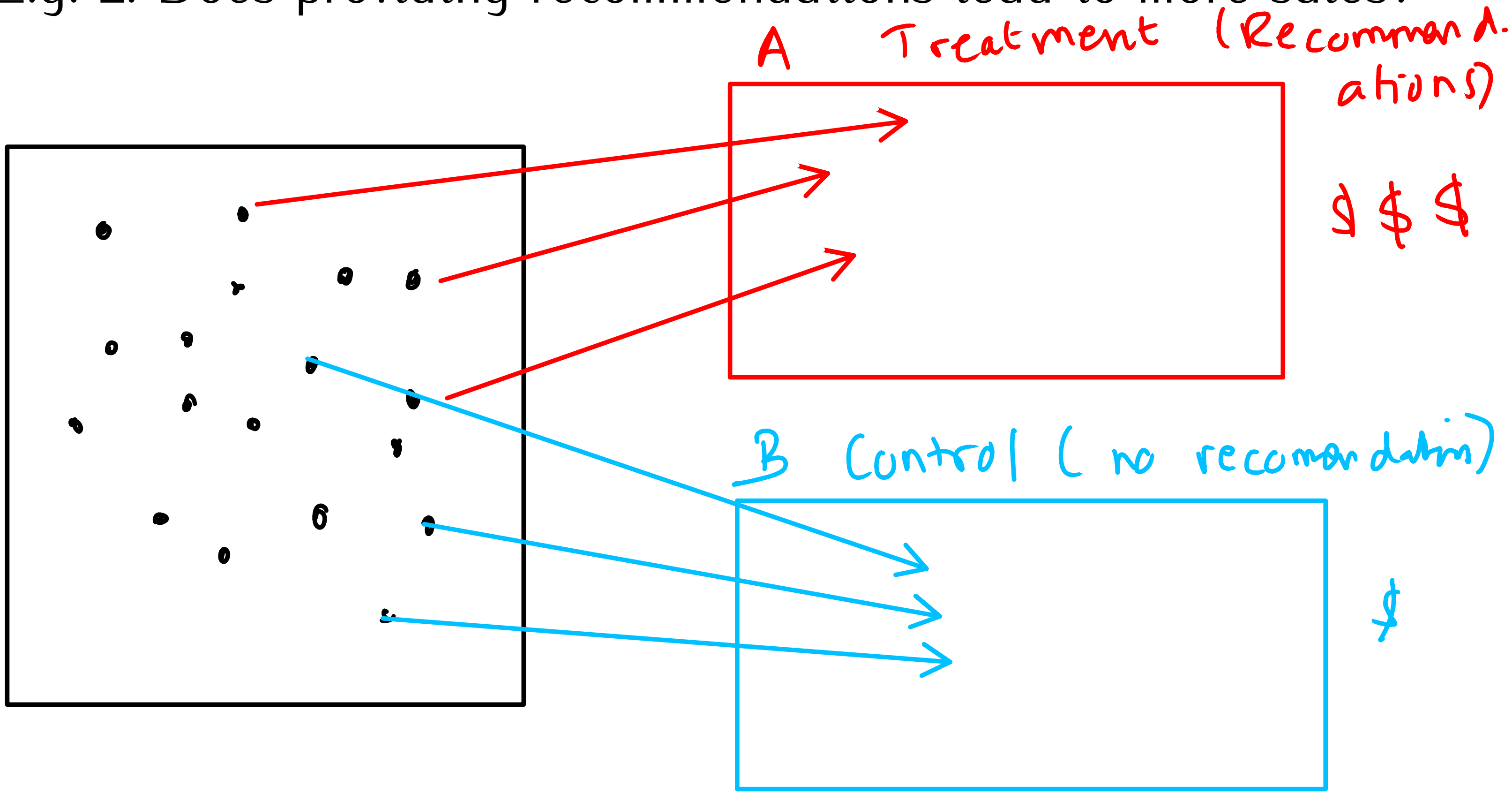
92 Non-black

Question: what if
(a) there had been 26 black and 74 non-black?
(b) there had been 20 black and 80 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 => 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 (S2 Week 1)
2. Estimation, including confidence intervals (S2 Week 2)
3. Hypothesis testing (S2 Week 3)
4. A/B testing (S2 Week 3)

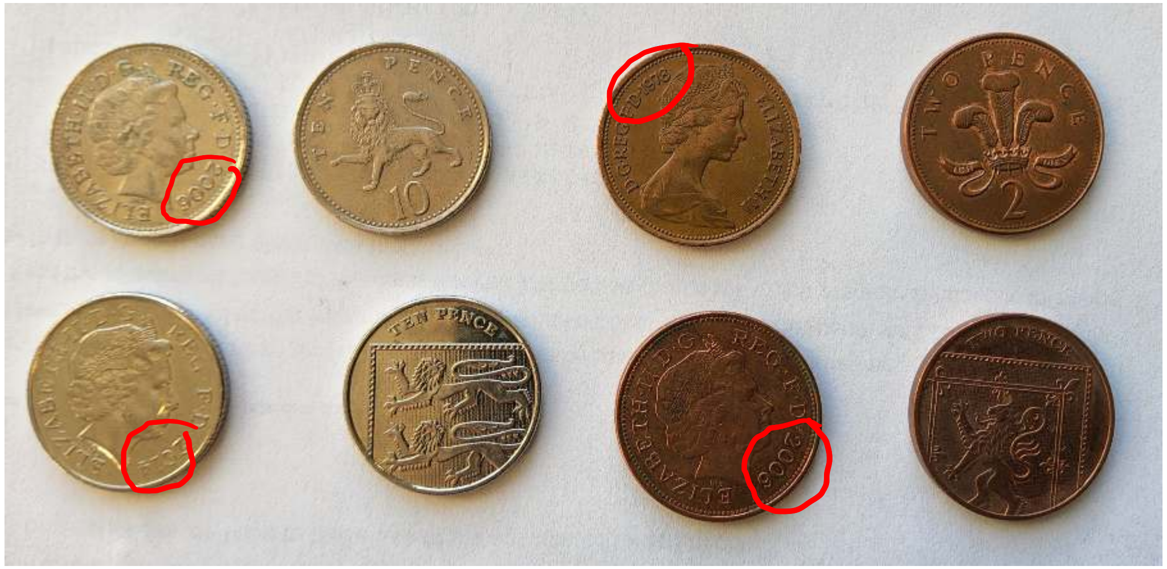
How can we address these questions?

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 ?

Head Tail Head Tail

Old style

New style



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

Coin tossing data

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?