

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
 - No lab sessions – ask on Piazza
- **Workshops:** in weeks 2, 4 and 6
 - May be in a new group or at a different time
 - To change group, use Group Change in Timetabler; turn up if change not actioned.
- **CW1 – 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



THE UNIVERSITY of EDINBURGH
informatics
Open Course Materials

Resources

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

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.

Informatics 2 - Foundations of Data Science

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

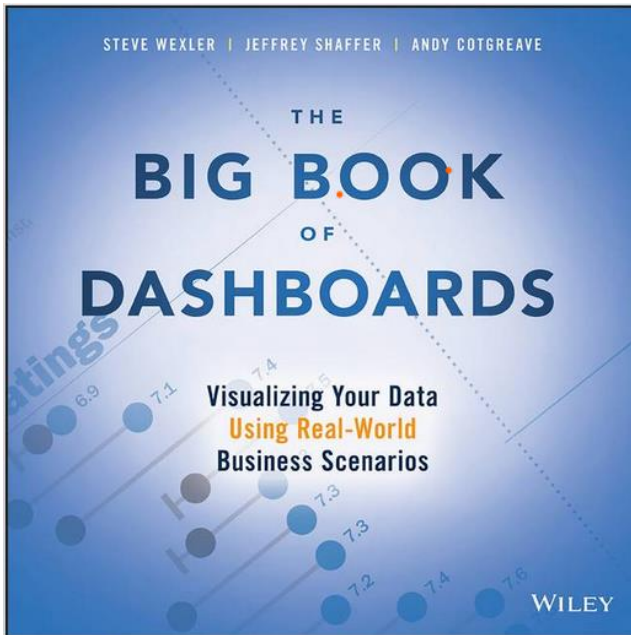
Part III Comprehension Questions: Introduction to machine learning

Part IV Comprehension Questions: Inferential statistics

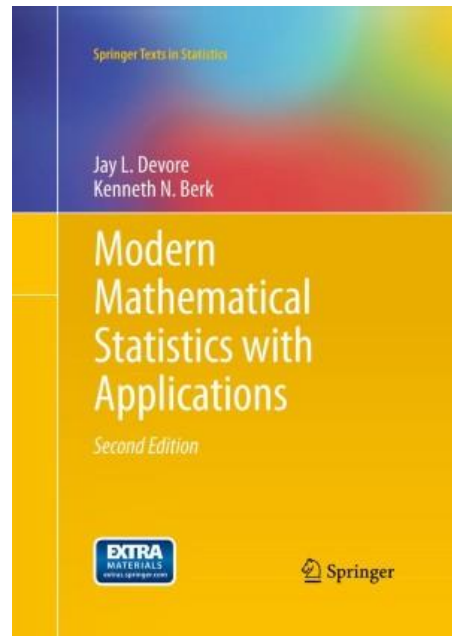
Coursework Planner
Overview of all coursework including submission and feedback timing.

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 Monday, 4pm in AT Cafe



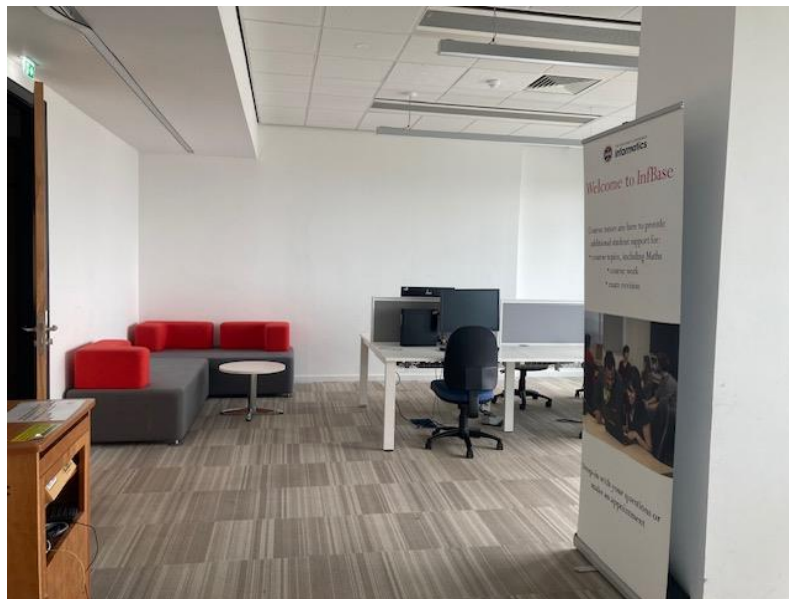
THE UNIVERSITY of EDINBURGH
informatics

InfBase

student learning support

Appleton Tower 7.03
(7th floor)

Monday to Thursday
11:10-13:00



InfBase Semester 2

- Starting Wednesday 14th
- Drop-in support space for year 1 & 2 students
- Run by senior peers
- 1-on-1 support and advice to help work through problems with lab exercises
- Also just a nice place to work together just in case you need assistance
- Semester 2 courses covered:
 - Informatics 1 - Object Oriented Programming
 - Informatics 2 - Software Engineering and Professional Practice
 - Informatics 2D - Reasoning and Agents
 - Informatics 2 - Foundations of Data Science
 - Informatics 2 - Introduction to Algorithms and Data Structures

<https://infbase-learning-support.github.io/>

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 (50%) 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**

Week 1 task / Week 2 workshop

- Critical reading of a data science study, which we will base at least some exam questions on
- Task for now: find and read the paper on OpenCourse
- You may want to read the associated media article too

Original research

Body dissatisfaction predicts the onset of depression among adolescent females and males: a prospective study

Anna Bornioli , Helena Lewis-Smith, Amy Slater, Isabelle Bray

► Supplemental material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/jech-2019-213033>).

University of the West of England Bristol, Bristol, UK

ABSTRACT

Rationale Body dissatisfaction is prevalent in mid-adolescence and may be associated with the onset of depression.

Objective The study assessed the influence of body dissatisfaction on the occurrence of later depressive episodes in a population-based sample of British adolescents.

extent to which a person cognitively 'buys into' socially determined ideals of beauty) and appearance comparisons (ie, the extent to which a person compares their own appearance with that of others). This model has received substantial support^{7 8}; and scholars have also highlighted the prominent impact of the media in body dissatisfaction processes.^{9 10}

Protected by copyright

J Epidemiol Community Health: first published as 1



The image shows a screenshot of the Daily Mail website's Science section. At the top, the 'Daily Mail' logo is visible, followed by a navigation bar with links to Home, News, Royals, U.S., Sport, TV, Showbiz, Lifestyle, Health, Science, and Money. Below this is a sub-navigation bar with links to Latest Headlines, Blue Origin, SpaceX, NASA, Apple, Google, Twitter, and Microsoft. The main content area features a large advertisement for 'Daily Mail Your Health' with a green background and images of pills. The ad text says 'Transform your health every Tuesday' and 'guides, real-life stories and advice from'. Below the ad, a headline reads 'Teenagers who dislike their own body are THREE TIMES more likely to be depressed as adults, study finds'.

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?

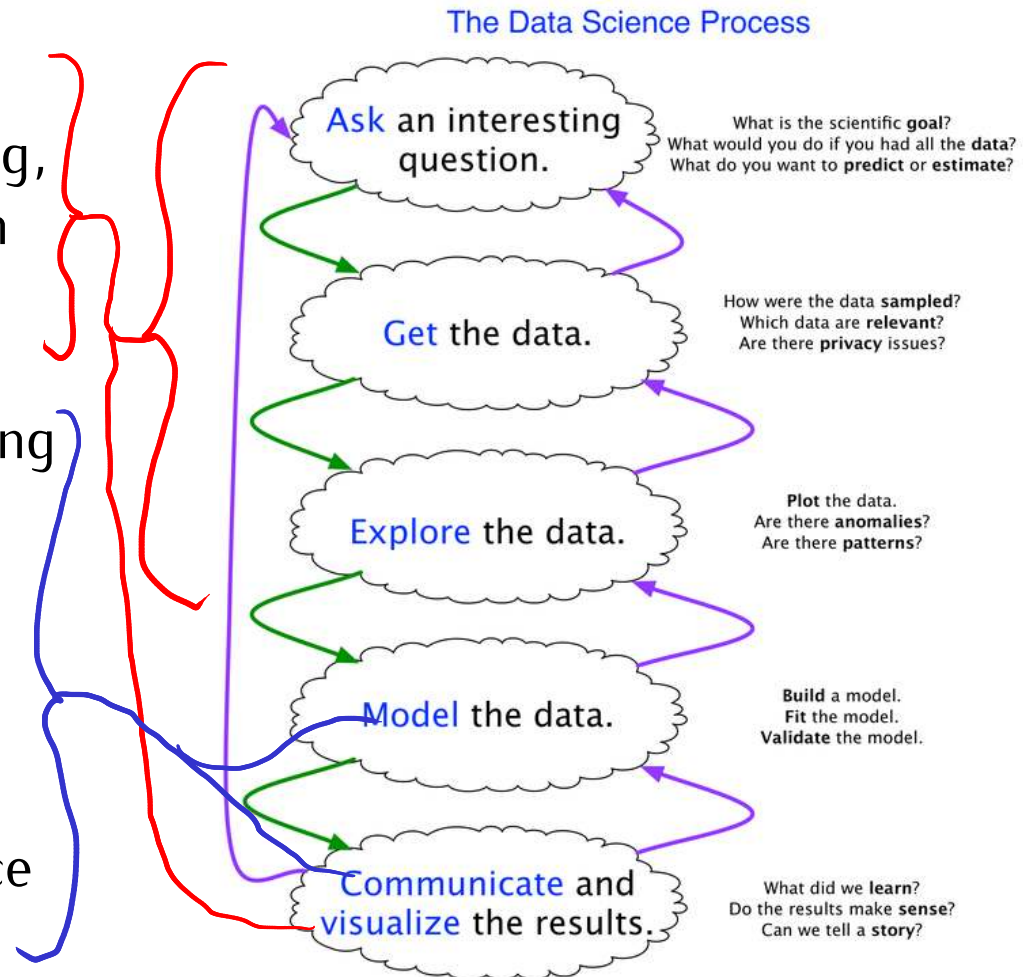
I. Data: ethics, collection, representation, wrangling, exploration, visualisation and descriptive stats

II. Intro to Machine Learning

III. Linear models

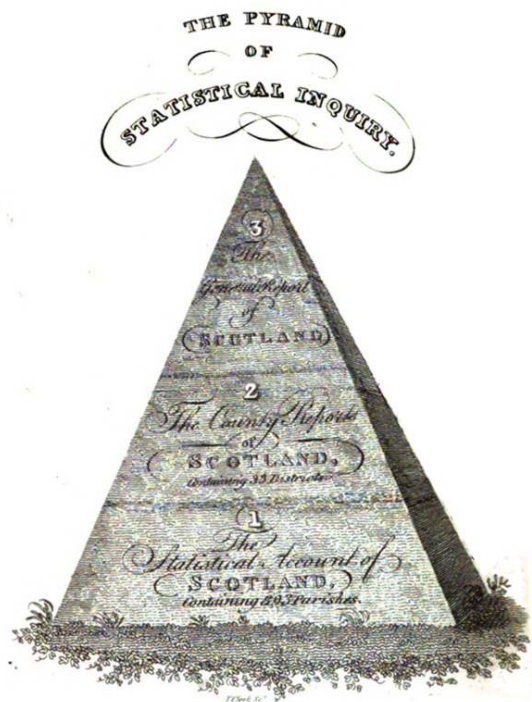
IV. Statistical inference

V. Regression and inference



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

Descriptive statistics



1 The Statistical account of Scotland, commenced in May 1790, and was completed in 1800.
 2 The Publication of the corrected County Reports, commenced in June 1795 and was completed in 1804.
 3 The General Report of Scotland, commenced in 1801, and was completed in 1804.

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 shall further say on
 Lime, is, that by it we can produce good crops of rough Beer & pease, without
 it, these last will not grow in this country, the crop is indeed precarious,
 but often productive & if well got, the straw is excellent winter fodder for
 the sheep. would the farmers consult their own Interest, & after lime sow
 their lands with grass the second or third crop, the Hay & pasture would
 more than compensate them, besides leaving the land in good heart for after
 crops, but men seldom forgoe a present profit for future gain!

12th Plough: gates in the County of M^d Lothian by which the statute labor
 is determined at 15 35 $\frac{1}{2}$ ----- L26..11..3

Do Selkirk shire -----	20
Horses M ^d Lothian -----	182
Do Selkirk shire -----	60
Black Cattle M ^d Lothian -----	1290
Do Selkirk shire -----	199
Carts M ^d Lothian -----	91
Do Selkirk shire -----	30

13th The next Article to be
 considered is Sheep, for which nature seems chiefly to have designed
 this part of the country, If we examine the sheep in the higher parts of
 the parish, they seem originally to have been of the black faced kind

Inferential statistics

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

E.g. "Manuscript on Deciphering Cryptographic Messages"

Al-Kindi, 9th Century, Baghdad

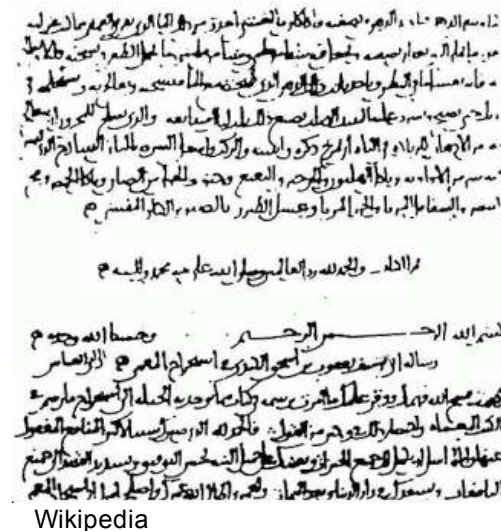
E.g. Wildcats



Wikipedia, Peter Trimming, CC BY 2.0

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

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

$[320g \pm 16g]$ contains μ

Report as : $\mu = 320 \pm 16g$, $n = 20$

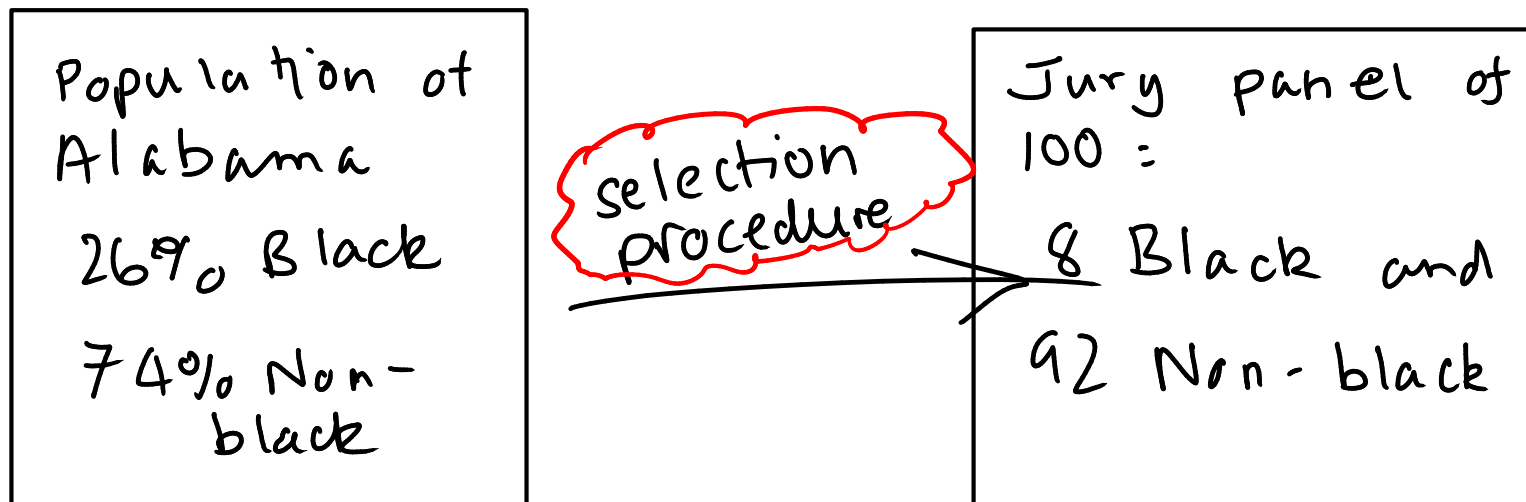
E.g. 2: Opinion polling

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?



Question: what if

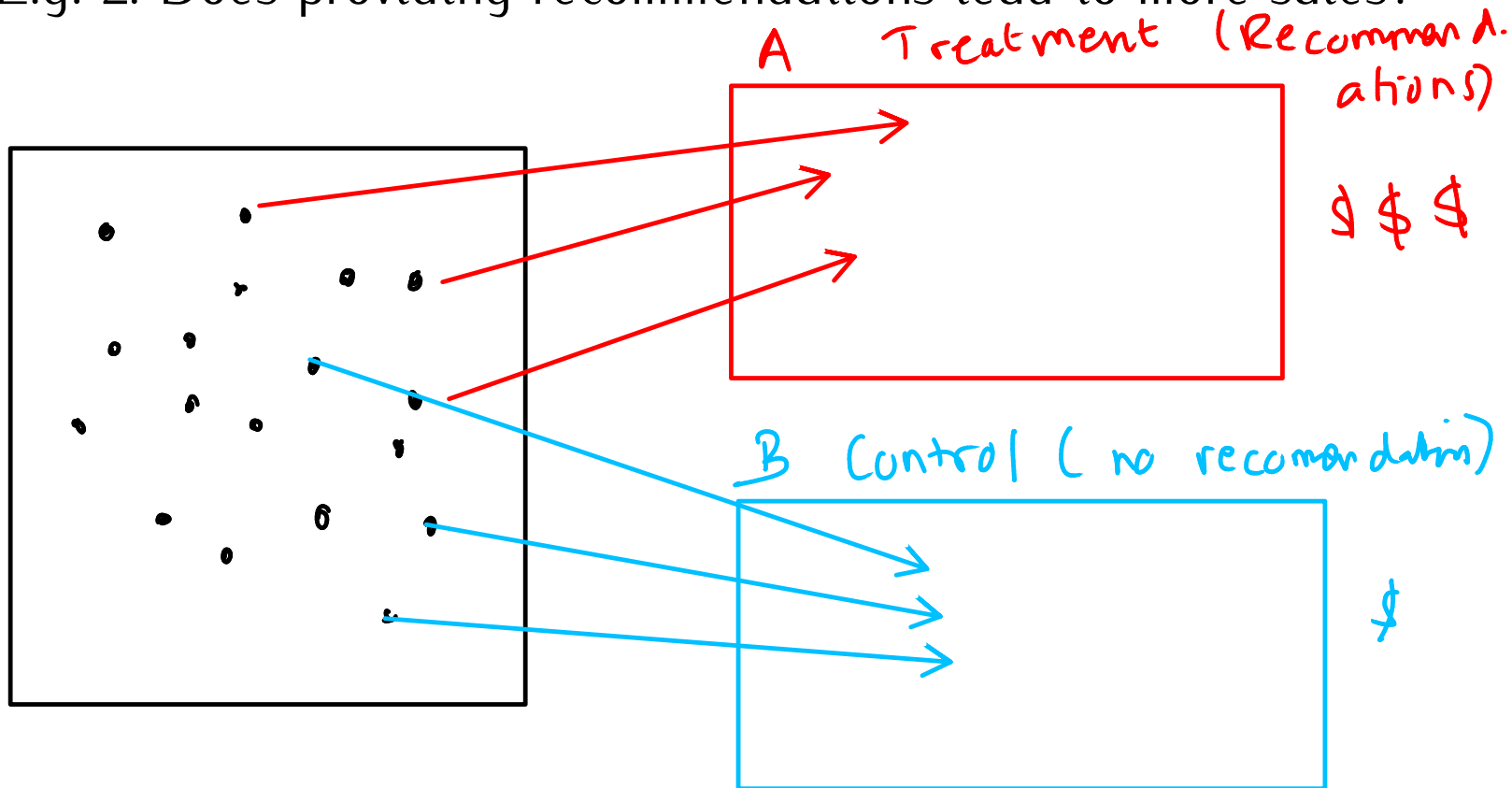
- (a) there had been 26 black and 74 non-black?
- (b) there had been 20 black and 80 non-black?

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 $\Rightarrow$ 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 (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?