Topic 1: Design of Experiments

Data science intro

Data science: using data to solve problems and find patterns
Data scientist: ask questions, gather and clean data, use tools, and explain findings → e.g using data to understand road accidents

Ethics and privacy

Handle personal data carefully → do not share private info
Data must be non-identifiable (no names, no private info)
Special care is needed for indigenous or vulnerable groups (e.g children)

Big data

Massive datasets
Big data has many "V's" - volume, variety, velocity etc.

Randomised controlled trials (RCT's)

What is a Controlled trial?

- ➔ You want to know if a treatment works
- ➔ You create 2 groups:
  - **Treatment group** gets the drug
  - **Control group** doesn't
- ➔ You compare the outcomes

Randomisation

- ➔ Randomly assign people to groups so it is fair
- ➔ Prevents selection bias (e.g only putting healthy people in one group)

Blinding

Single-blind: participants don't know what group they're in
Double-blind: Participants and experimenters don't know what group participants are in
Prevents *observer bias* (researcher accidentally influences the results because they know who got the real treatment e.g treating treatment group differently) and *placebo effect* (someone feels better just because they believe they got a real treatment - even if they didn't)

Common biases

Selection bias: Picking only certain types of people/participants non-randomly chosen e.g hospital selects healthier subjects for surgery
Observer bias: Experimenter expectations influence results.
Consent bias: Only certain people agree to join the study.
Survivor bias: Only those who finish are counted.

Observational studies

Whats an observational study
You don't assign treatments - you just observe real life e.g you observes smokers and non-smokers, but don't make them smoke

Key precautions

- ➔ Association ≠ Causation → just because two things are linked doesn't mean one causes the other
- ➔ Confounding variables
  - A confounding variable is a hidden factor that affects both the thing you're studying (e.g., smoking), and the result you're measuring (e.g., liver cancer)
  - This can confuse the results and make it look like one thing is causing another, when it's actually the hidden third thing
  - **E.g** You're studying 'does smoking cause liver cancer?' → but people who smoke may also drink more alcohol, and we know that alcohol can cause liver cancer → So maybe it's not smoking causing the cancer, maybe it's the alcohol, and smoking just happens to be linked with drinking
- ➔ **Simpson's Paradox** → when a pattern in individual groups changes or reverses when the groups are combined
  - **Example**:
  - Comparing two hospitals (Hospital A and B) and looking at surgery success rates.
  - Hospital A:
    - Easy surgeries: 90/100 succeeded (90%)
    - Hard surgeries: 10/20 succeeded (50%)
  - Hospital B:
    - Easy surgeries: 80/100 succeeded (80%)
    - Hard surgeries: 40/50 succeeded (80%)
  - Within each group, Hospital A seems better at easy surgeries (90% vs 80%) and worse at hard surgeries (50% vs 80%).
  - But look what happens when you combine:
    - Hospital A total: 100/120 = 83%
    - Hospital B total: 120/150 = 80%
  - Now Hospital A looks better overall, even though it was worse for hard surgeries, and only slightly better for easy surgeries

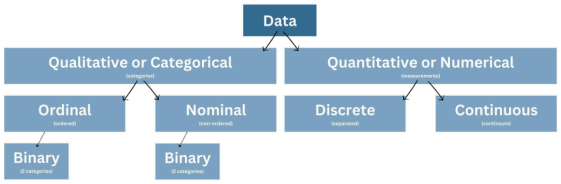
Topic 2: Data Graphical Summaries

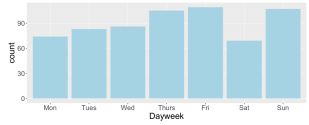
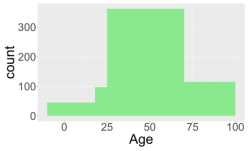
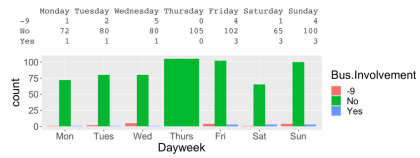
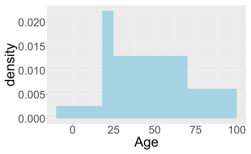
Data basics

Data: information about things we're studying → usually we work with a sample (part of population)
Initial Data analysis (IDA): first step in any data project → Assess data's ability to answer research questions, Identify potential new research questions, Understand data's main qualities, Determine population source
IDA Components:

- ➔ **Data Background** → checking quality and integrity of data
- ➔ **Data Structure** → what information has been collected?
- ➔ **Data Wrangling** → Cleaning, Tidying, Reshaping
- ➔ **Data Summaries** → Graphical representations, Numerical analysis

Types of variables



Graphical Summaries for Qualitative Data	Graphical Summaries for Quantitative Data
Simple Barplot Purpose: Summarise 1 qualitative variable 	**Simple Histogram** Used for quantitative data - to see how a variable is distributed across different class intervals X-axis = variable Each block = class interval Y-axis represents count of subjects 
Double barplot Purpose: summarises 2 qualitative variables Visualisation: Secondary variable represented by color/shading 	**Density/Probability Histogram** Area of each block represents percentage of subjects in particular class interval Adjusts for varying interval sizes 
Sliced histogram Adding a qualitative variable to a histogram by slicing each class interval by colour - we can see how each variable is distributed within each class interval	