DATA1001 Week 1

Abyan Majid

August 1, 2023

Contents

1 Lecture 1 Notes

1.1 Introduction to Data Science

Data Scientist: A person who is able to unlock insights of and storytell data.

Data science skills and tools sorted from most to least important:

1. Domain knowledge and soft skills
2. Communication and visualization skills
3. Ethics
4. Math and statistics
5. Programming language and databases

Domain knowledge: knowledge about the case/context being studied.
Evidence that makes a study/research reproducible: A reputable study of data will have all stages of the study (data collection, statistical methods that were applied, conclusions, etc.) documented, therefore allowing 3rd party verification of the conclusions via replication.

1.2 Controlled Experiment

A "controlled experiment" is a method of comparison to test the effect ("response") of a particular independent variable ("treatment"). It is called a "controlled" experiment because we have 2 groups of subjects, which are the:

- Experimental (or "Treatment") group: subjects who are affected by the independent variable (receives the treatment)
- Control group: subjects who are not affected by the independent variable (does not receive the treatment)

1.2.1 Confounding/lurking variables

Confounding (or "lurking") variables are variables that are hard to identify (hidden) - because of this, they may have a signfiicant influence on the conclusions without the data scientist's realization, therefore rendering the effect of the independent variable ("treatment") less accurate/less reliable.

1.2.2 Bias and types of biases

A "bias" is an error in how we collect or report data, which is a bad thing - because doing studies on biased data will produce skewed/inaccurate results. In the case of a controlled experiment, "biased data" produces inaccurate conclusion about the independent variable ("treatment").

As a way to visualize different types of biases, suppose we have the following scenario: We want to know whether or not caffeine improves academic performance. So, we conduct a controlled experiment where we have the:

- experimental ("treatment") group: students who consume caffeine at fixed times of the day
- control group: students who do not consume caffeine

Selection Bias: A bias created when a group (experimental or control) is comprised of subjects that specifically do not represent the whole sampled population.

> In our caffeine-academic performance example, one way a "selection bias" may occur is the control group (i.e. students who do not consume caffeine) being comprised of mostly highly contientious/dilligent students, meanwhile the experimental ("treatment") group is not.

Observer Bias: A bias created when the observer or the subjects know the identity of the 2 groups, so the subjects may intentionally report more or less favourably.

In our caffeine-academic performance example, one way an "observer bias" may occur is because the observer treats the experiment (treatment) group better than the control group (e.g. giving more encouragement, help, etc), thereby contributing to their academic success - rendering the role of caffeine smaller.

Consent Bias: A bias created when subjects can choose if they want to be in the experimental group or not.

In our caffeine-academic performance example, this would be a student (regardless of their character) refusing to consume caffeine.

Survivor Bias: A bias that occurs when we only consider the "survivors" or the "success stories" of a study such (usually contributing to rendering the hypothesis true).

In our caffeine-academic performance example, this bias could occur due to the observer's disregard for results that do not support their hypothesis, ie. does not report cases of students who consume caffeine but does not improve or does not do well academically.

Adherer Bias: A bias that occurs when some participants comply with the experiment (or take the treatment) more consistently than others. In other words, these people "adhere" to the experiment more than others.

In our caffeine-academic performance example, this bias could happen as a result of students in the experimental ("treatment") group skipping sessions of caffeine consumption, thereby consuming a total amount less than others

1.2.3 Placebo and placebo effect

- A "placebo" is a pretend treatment (ie. something that is not the treatment but is designed to mimic the treatment)
- A "placebo effect" is an effect on the subject caused by the subjects thinking they were given the treatment.

1.2.4 Randomised Controlled Trial (RCT)

A study where the assigment of subjects into either the experimental ("treatment") group or the control group is done randomly. RCTs are likely to spawn ethical questions, such as whether it is justifiable that students that do not like caffeine be forced to consume caffeine because they were randomly assigned to the experimental ("treatment") group.

1.2.5 Randomised Controlled Double-Blind Trial (gold standard!)

A study where both the subjects and the observers are not aware of the identity of the 2 groups. In such a study:

- there is control over the observers' evaluations and the subjects' responses
- there is usually a 3rd party that administers the treatment and placebo

- the placebo is designed to be as indistinguishable from the treatment as possible.

A double blind RCT is the "gold standard" because it minimizes biases! However, it's hard to achieve due to a lot of possible reasons! (e.g. subjects may be able to differentiate placebo from treatment, it still doesn't guarantee subjects' complete adherence to the protocol, etc.)

1.3 Observational studies

An "observational study" is a study in which the assignment of subjects into groups is not within the investigator's power and it cannot be done randomly.

1.3.1 Precautions

1. Observational studies can only establish association (that one thing is linked to another), not causation.
2. Observational studies may appear as a randomised trial due to confounding variable(s) when it fact it is not.
3. Observational studies may lead to simpson's paradox due to confounding variable(s).

1.3.2 Simpson's paradox

Simpson's paradox, in a nutshell, is the reversal of correlation between two variables A and B when subgroups defined by a confounding variable C are combined/aggregated together

Example: suppose you have the following data (which intuitively feels wrong!):

- People who drink alcohol on average earns $\$90,000$/yr
- People who do not drink alcohol on average earns $\$48,000$/yr

  The data suggests that drinking alcohol causes people to have a higher annual income! But actually, when you dig deeper into the data, you realize a confounding variable "age" which if taken into account, actually derives the conclusion that older people earn more money than younger people - and that more of alcohol drinkers are young! Wait, so that means alcohol consumption barely has any impact on income! What a 180 degrees turn of event!

The above example depicts a simpson's paradox which occured as a result of the investigator pooling together data of (1) alcohol consumption and (2) income by age, but completely disregarding the confounding variable "age".

1.3.3 Preventing Simpson's Paradox/dealing with confounders

When the conclusion suggested by a correlation in a data feels intuitively wrong, try to look for the confounding variables! Once you are confident of your suspicion toward a particular confounding variable, **try to split your subjects into subgroups with respect to the confounding variable!**

For example, to investigate the confounding variable "age" in the data:

- People who drink alcohol on average earns $\$90,000$/yr
- People who do not drink alcohol on average earns $\$48,000$/yr

We can try splitting the subjects to several "age subgroups", such as (A) ages 20-29, (B) ages 30-39, (C) ages 40-49, (D) ages 50-59, (E) ages 60-69, (F) ages 70-79, and (G) ages 80-89.

In this case, we say that we are "controlling for age".