



Case Selection

QUANTITATIVE ANALYSIS
WEEK 4



Review

- Last week, we looked at the process of identifying a research question, developing a hypothesis, and then operationalising that hypothesis by finding concrete variables that can substitute for abstract concepts.
- We discussed the importance of the literature review – learning as much as you can about both the papers / books which are the foundation of the research in your field, and about the most up-to-date scholarship from the past few years.
- A thorough literature review is vital to making your research question important and original.



Review (2)

- We discussed three different kinds of question – descriptive, relational, and causal. Each of these has value, but most quantitative research focuses on causal questions – and their related hypotheses, which are a clear statement of cause and effect.
- A hypothesis always contains some abstract concepts. These may be obviously abstract – like “justice”, “freedom”, or “populism” – or they may seem deceptively concrete, but actually have a highly subjective aspect to them, like “wealth”, “intelligence”, or “unemployment”.



Review (3)

- To perform quantitative analysis, we must find an operational definition for each of these abstract concepts.
- These definitions must clearly and thoroughly state what the concept encapsulates in terms of concrete things that can be measured.
- Often, definitions are contested – you'll find different researchers arguing for different ways to operationalise a concept. You need to carefully consider which one to use, and justify why it fits your research better than the alternatives.

A Note on Variable Types

Continuous vs. Categorical



Variable Types

Categorical

- **Nominal**
 - Two or more categories, with no specific order – e.g. prefectures, ethnicities.
- **Ordinal**
 - Multiple categories with a ranked order – e.g. education level, degree of preference.
- **Dichotomous / Binary**
 - Categories with only two levels – e.g. Male / Female, Yes / No.

Continuous / Quantitative

- **Interval**
 - Measured quantities with a numeric value – e.g. temperature.
- **Ratio**
 - Measured numeric quantities with a fixed zero point – e.g. distance, vote counts.

Other Operationalisation Decisions

What is your unit / level of analysis?

- This might be individuals, groups, companies, towns, regions, nations...
- It's generally best practice to measure all your variables at the same **unit level**.

Measuring one variable with an individual-level survey, and another with national economic data, makes any comparison between them tricky (although not necessarily impossible).

Snapshot or Time Series?

- Snapshot data captures your variables at one point in time – e.g. election results, or a survey conducted all in one go.
- Time series data measures the same variables (usually in the same cases) repeatedly over time – e.g. economic trends, panel surveys conducted in multiple waves.
- The conclusions you can draw from each type of data are different.

Selection Bias

A hand is pointing at a bar chart displayed on a tablet screen. The chart consists of seven bars of varying heights, representing data points. The hand is positioned over the first bar, which is the shortest. The background is a blurred blue fabric. The text 'Selection Bias' is overlaid in the top left, and 'CHOOSING THE RIGHT CASES' is overlaid in the bottom left with a pink underline.

CHOOSING THE RIGHT CASES



Case Selection

- Every social science research project involves case selection.
- That might involve choosing which countries, events, political parties, or companies to study; or it might involve selecting which people to send an electoral survey to, or who to invite to participate in experimental research.
- Selection, or sampling, is always necessary because we can never observe the complete set of data that would be relevant to our hypothesis.
- Even when you can observe every event that happened, you cannot observe those that didn't happen – including those in the future, which your hypothesis should predict.



Large-n and Small-n

- "n" is the variable name commonly used to indicate the number of cases / samples being used in the research.
- From this we get the idea of "large-n" research – which uses a lot of samples, e.g. surveying thousands of people, or collecting data on hundreds of subjects like politicians or companies.
- Small-n research instead looks in more depth at a smaller number of cases – such as historical events like revolutions or wars, or case studies of specific politicians.



How large is “large-n”?

- Quantitative Analysis is mostly concerned with large-n research, for the simple reason that statistical methods are generally unreliable when you only have a small number of cases to observe.
- However, depending on your methodology, you can still use quantitative analysis techniques on a relatively small data set – this happens frequently in international relations, for example.
- 30 observations is often stated as a rough minimum guideline for quantitative research.
- However, the more variables you want to include in your model, the more observations you'll need.



Data Segmentation & Statistical Power

- Most forms of statistical analysis involve **segmenting** the data into smaller parts and re-confirming the existence of statistical relationships inside those segments.
 - For example, remember the example of the tobacco companies' claim that both lung cancer and smoking were caused by urbanisation? This could be tested by segmenting the data into "urban" and "rural" segments, then seeing if the relationship between smoking and cancer persisted in both segments.
- The more you need to segment your data, the more data you need to start with.
- 30 samples might be enough to confirm one **bivariate** relationship (i.e. a simple relationship between two variables) – but include just one other factor and you only have ~15 samples per segment.

Selecting on the Dependent Variable

- One of the most dangerous types of selection bias involves selecting cases based on the dependent variable – in other words, choosing your cases based on the outcome you wish to study.
 - For example, you might try to examine the factors that lead people to vote for populist politicians by surveying / interviewing voters for populists.
 - It seems logical – but voting for a populist is the outcome you are trying to observe, and this method means your outcome variable is always 1, never 0.
-



Restricting the Dependent Variable

- Perhaps your research design does allow for the outcome variable to differ – but might some aspect of it restrict the potential variation in the outcome?
- For example, you might want to research the benefits of doing a Masters degree by observing its effects on salaries – comparing the salaries of people with BA and MA degrees from the same university after ten years in employment.
- But if your data does not include people who were unemployed, or who dropped out of the workforce, or who have no fixed income (freelance / part time workers), then you are restricting the possible values of the outcome variable.



Survivorship Bias

- This is a kind of bias that occurs when we select cases only from a group that has passed some form of filtering process that is relevant to our research.
- For example: researching election campaigns by studying elected officials; researching the factors for business success by surveying top businesspeople.
- Malcolm Gladwell's "10,000 Hours" claim.
 - Gladwell found that 10,000 hours was the amount of time learning and practicing required to reach the top of many fields of human achievement.
 - But he only interviewed top athletes/artists/researchers etc. for his study. Many people may spend 10,000 hours on something and never become world-class at it!




Self-Selection Bias

- Also called volunteer bias, this is a kind of bias that emerges when your research participants can put themselves forward.
- In observational studies, this happens when the research subject chooses the treatment of their own volition – e.g., a legislature choosing to adopt a certain policy.
- The people who choose to take surveys or participate in interviews and focus groups may be doing so for specific reasons – e.g. being interested in the field you are studying.
- A related problem in survey research is nonresponse bias. You can't force people to take a survey, and people may decline to respond for specific, systematic reasons.

Identify the Sources of Bias

IN-CLASS EXERCISE



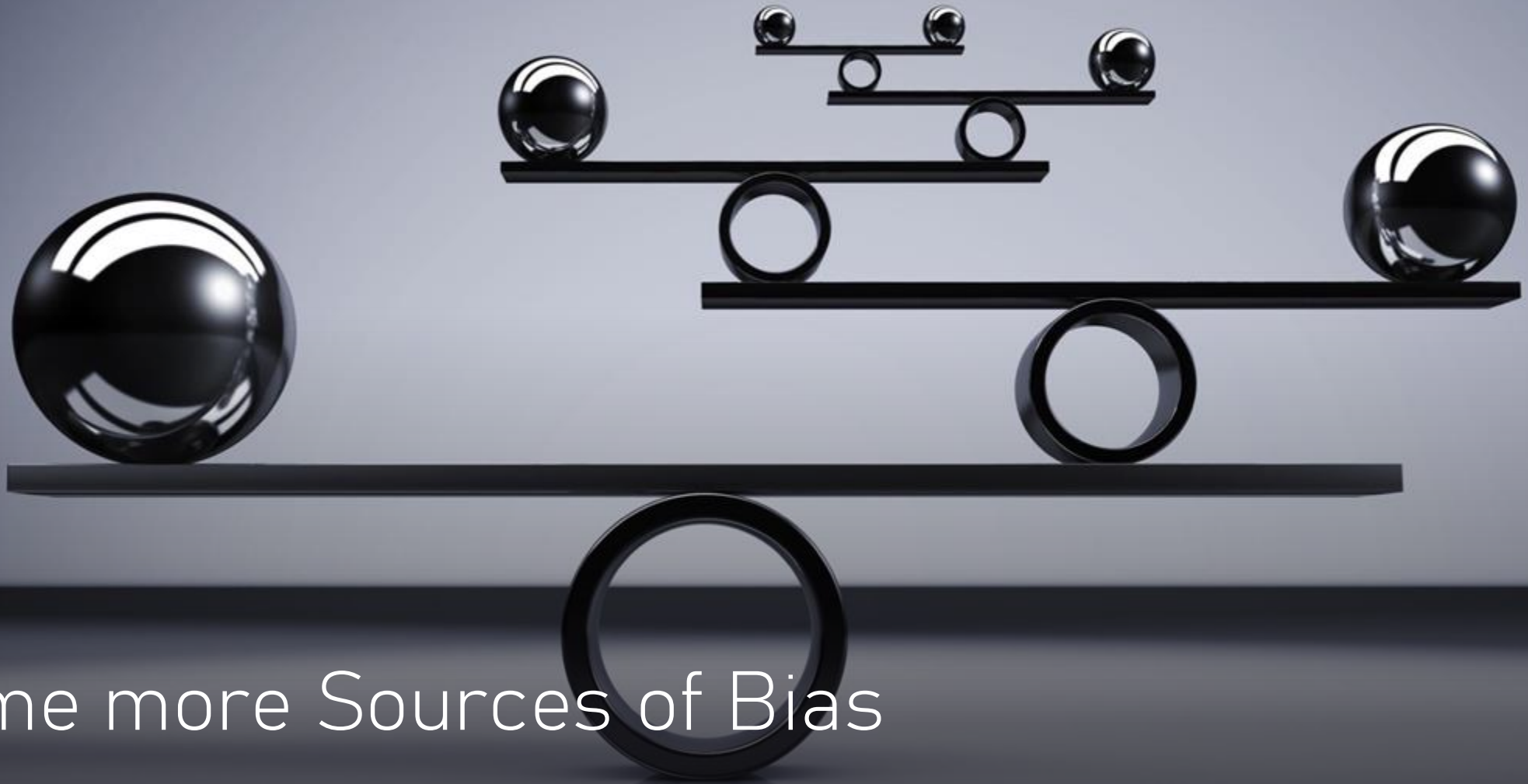


A university admissions office wants to know whether a student's grades in high school can predict their academic performance at university. To do this, they examine a randomly selected sample of the university's students, using quantitative analysis to see if their high school grades are correlated with their present academic performance. To their surprise, better grades in high school are actually negatively correlated with academic performance.

The education ministry is looking for ways to prevent pupils dropping out of school. It conducts interviews with hundreds of school dropouts to find out what their reasons for leaving education were. Many of them report that the lessons were boring and irrelevant to them, leading the ministry to propose overhauling the school curriculum to tackle these problems.

A doctor believes that a certain medical procedure may be harmful to children in the long term. Working with a campaign group of parents who are also concerned about this possible link, they conduct interviews and tests on parents and children recruited from the group, and conclude that while a causal link is difficult to prove, their data raises sufficient doubts about the safety of the procedure to recommend that it should be discontinued.

Anti-immigration politicians are winning an increasing number of votes in a certain city. Researchers at a major university in that city recruit local students for an experiment in which they are questioned about their views on immigration, shown a short video about the positive benefits of immigration to the local economy, and then questioned again about their views to see if they have changed. The researchers find that the video does not change people's views significantly.



Some more Sources of Bias



Attrition Bias

- Just like you cannot force people to take a survey – leading to nonresponse bias – you also can't force someone to continue with a survey or experiment if they want to stop.
- This leads to attrition – participants dropping out midway through your study – which may also create systematic bias in your data.
- A common example is in panel surveys, which question the same group of people at different points in time. These usually have a significant drop-out rate between waves.
- In observational studies, this may happen when certain subjects stop reporting data during the time period being observed. For example, small companies may drop below certain thresholds for public reporting of tax affairs.



Recall Bias

- Humans are terrible at remembering things.
- Even memories that seem very clear to us can actually be reconstructed over time – our memory of an event changes to fit with knowledge we learned later.
- In post-election surveys, people are more likely to claim that they voted, and more likely to claim that they voted for the winner – creating biased data for turnout and vote shares.
 - They're not necessarily lying – their memory may seem genuine, but it has been altered post-facto.
- This form of bias makes asking about memories of prior events somewhat unreliable.



Social Desirability Bias

- Another common problem in survey and interview research is that many respondents will give you answers they think you want to hear.
- People who hold views they think are unpopular – especially views that may be discriminatory, like racism or homophobia – may deliberately hide those views in an interview or survey.
 - This is an example of the spiral of silence theory at work.
- This may also explain systemic differences in Japanese election surveys!
 - People being surveyed by the Asahi Shimbun may be more likely to state preferences for the opposition parties; people being surveyed by the Yomiuri Shimbun may be more likely to express support for the LDP.



Next week... Statistical Tests in R

- Your task for next week: make sure your computer has a working installation of R and RStudio.
- There are instructions on Moodle for where you can download these.
- **Ensure that you have the tidyverse packages installed before the class.** Again, instructions are on Moodle.
- If you have not used R before (or you're just a bit rusty, or unfamiliar with the Tidyverse packages in R), then you should follow through the beginners' tutorial that was posted a few weeks ago.