PUBL0055: Introduction to Quantitative Methods

Lecture 1: Introduction

Michal Ovádek and Indraneel Sircar

Lecture Outline

Course Outline

Logistics

Quantitative Methods and Research Design

Introduction to Quantitative Data

Conclusion

Course Outline

What is this course?

- This is not a course on statistics
  - A statistics course would focus on the theory and derivation of statistical methods
  - We will discuss some theory at a basic level, but will not concern ourselves with the derivation
- This is a course on applied quantitative research methods
  - Focus on the developing intuition about quantitative methods
  - Focus on using these methods to answer social science questions
- This course is different to other similar courses
  - Stronger focus on causality, data visualisation, and application
  - Less focus on sampling, statistical inference and uncertainty

What is in this course?

1. Introduction
2. Causality I
3. Description & Measurement
4. Regression I (Prediction)
5. Regression II (Specification)
6. Regression III & Causality II
7. Causality III (Obs Data)
8. Uncertainty I (Sampling)
9. Uncertainty II (Hyp. Test)
10. Significance

Week 6 is reading week. There will be no lecture, but you will have a midterm assessment.

Why should you take research methods?

- This is a course on quantitative methods, not all research methods
- Many of you will also take a qualitative methods module – PUBL0010, PUBL0085, or PUBL0058
- The *science* of the 'social sciences' comes from the methodological rigour of the approaches you will learn in these courses
- These courses will...
  - ...provide you with the tools necessary to conduct social scientific research (relevant for writing your dissertations)
  - ...help you to better understand and evaluate quantitative claims (relevant for evaluating plausibility of current research)
  - ...help you to think more critically about evidence-based arguments made in the 'real world' (relevant for being a good human being)

Why should you take quantitative research methods?

You will learn...

- ...to apply a wide range of quantitative methods to answering your potential research questions
- ...the types of questions that can (and cannot) be answered using quantitative analysis
- ...to make more persuasive arguments using quantitative data
- ...to evaluate the quantitative evidence others present in their work
- ...some transferable skills

Logistics

Course Website

- The course website has several important resources for this module
  - Weekly class assignments and datasets
- The website can be found at https://uclspp.github.io/PUBL0055/

Moodle

- In addition to the course website, Moodle access is essential for this course
  - Lecture recordings
  - Links to student support and feedback hour signups
  - Assessments

Lecturers

Michal Ovádek

- E-mail: m.ovadek@ucl.ac.uk
- Student support and feedback hours, Mondays and Wednesdays 4-5pm.

Indraneel Sircar

- E-mail: i.sircar@ucl.ac.uk
- Student support and feedback hours, Thursdays 9:30-11:00 am.

Instructors

- César Burga Idrogo
- Irene Germani
- Guy Heilbrun
- Memta Jagtiani

Please check the Moodle page for their student support and feedback hour times and links to sign up.

Introduction or Intermediate?

We offer three quantitative methods modules at the MSc level:

	Introduction	Causal	Text
Term	One	Two	Two
Prior training? (methods)	No	Yes	Yes
Pre-requisites (R)	None	None	None
Focus	Intro	Causal inference	Text analysis

On most of the MSc programmes, it is possible for you to take a combination of these modules.

Which course should I take?

- This course has no pre-requisites: we will assume that you have no prior experience in either quantitative methods, or in coding
- The Causal Inference course requires you to have **at least one** prior course in quantitative methods/econometrics up the level that we cover on this course
- If you are unsure which course to take
    1. Take this quiz
    2. Book an student support and feedback hour appointment to speak to us

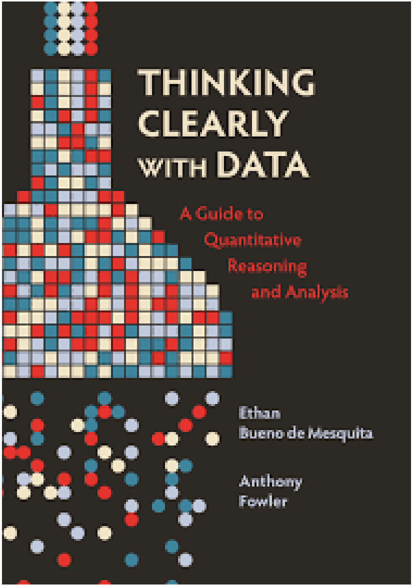
Learning objectives

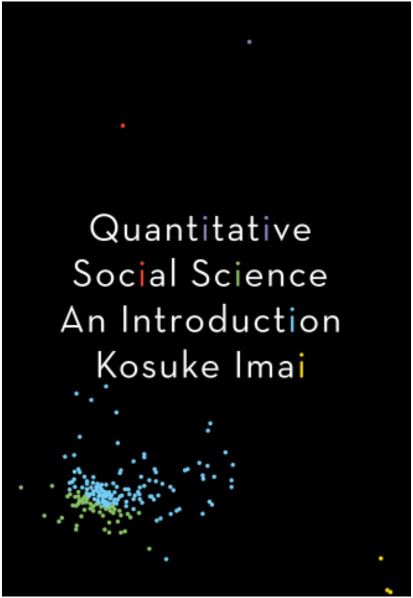
1. Understand the key tools used in modern quantitative methods
2. Understand which questions are and are not amenable to quantitative analysis
3. Improve ability to critically evaluate published work
4. Learn to implement key skills in R

Teaching philosophy

1. Building intuition is central to understanding statistical concepts
2. Examples and applications are central to building intuition
3. You cannot learn statistics or quantitative methods without analysing data on your own

Textbooks


THINKING CLEARLY WITH DATA
A Guide to Quantitative Reasoning and Analysis
Ethan Bueno de Mesquita
Anthony Fowler


Quantitative Social Science
An Introduction
Kosuke Imai

Advice on reading for this course

Statistical readings can be intimidating and on this course you should focus on an *in depth* reading of the textbook, rather than a broad and shallow reading of multiple sources.

1. Do the required reading before lecture
2. Do not expect to understand everything the first time
3. If overwhelmed, focus on the text, not the equations
4. After lecture, re-read to maximize understanding

R and Rstudio

- **R** is statistical programming language and software for data analysis
- **Rstudio** is software package that makes R more straightforward to use
- Why do we use R/Rstudio on this course? R is...
    - ...free!
    - ...more flexible than some alternatives – e.g. Excel, SPSS
    - ...widely used by researchers, companies, governments, non-profits, etc
    - ...also used on the Causal Inference and Text Analysis modules
- Learning to use R is essential to do well in this course
- You should install R and Rstudio on your personal computers
- Don't worry if you have trouble the first few weeks!

Lectures and classes

Lectures

- Mondays, 11:00 am - 12:50 pm, Christopher Ingold Building, XLG2 Auditorium
- Lecture recordings will be uploaded the middle of the week.

Seminars

- One hour seminar slots
  - Thursdays and Fridays
- Seminar attendance is mandatory in your assigned group

Homework

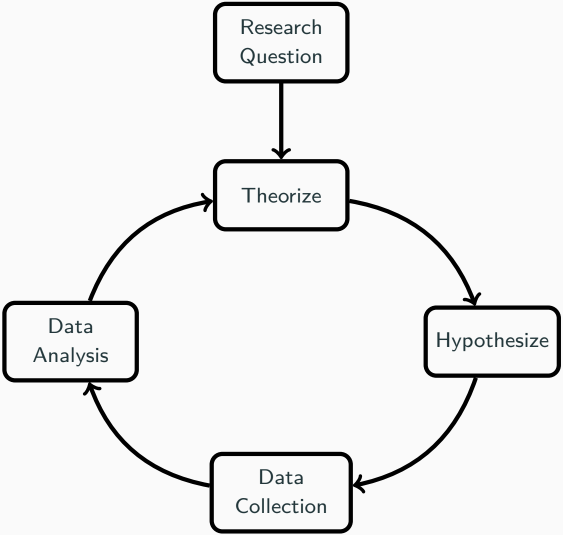
- The instructions and code you need for the seminars and homework will be available on the course website
- You should work through the exercises *before* your scheduled seminar time
- The site also includes useful information about the course, quantitative methods, and coding in R
- Each seminar includes a homework exercise which focusses on implementing the skills you have learned on new data
- Solutions will be posted on the Monday following the Friday class
- These homeworks are not assessed, but they will be very similar in style to the assessments!

Assessment

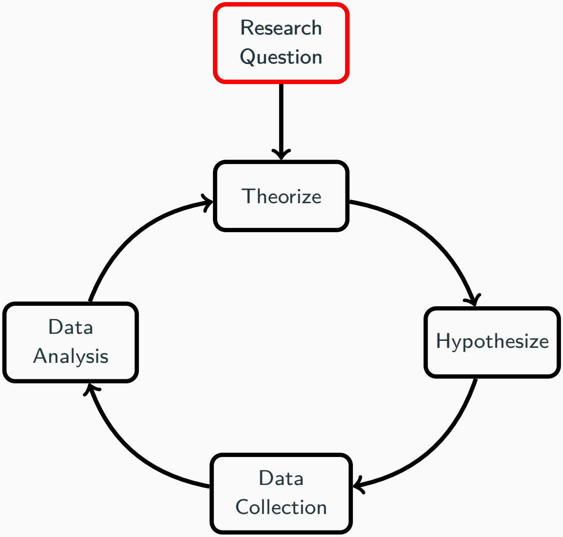
- 30% of the course mark is based on an online Moodle midterm (6 November 2024, details TBC)
- 70% of the course mark is based on a final take-home assessment (2000 words, due 15 January 2024, released 8 January 2024)
- The two courseworks will require you to:
  - understand the theoretical concepts
  - answer applied questions
  - work with R
- Details will follow during the term

Quantitative Methods and Research Design

Which part of the research process are we working on?

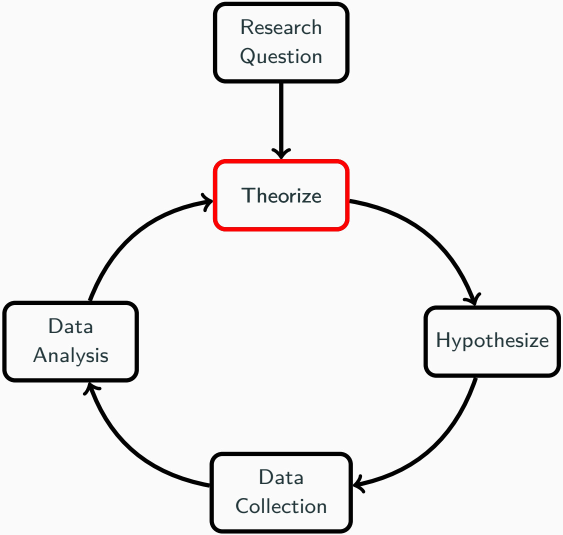


Which part of the research process are we working on?



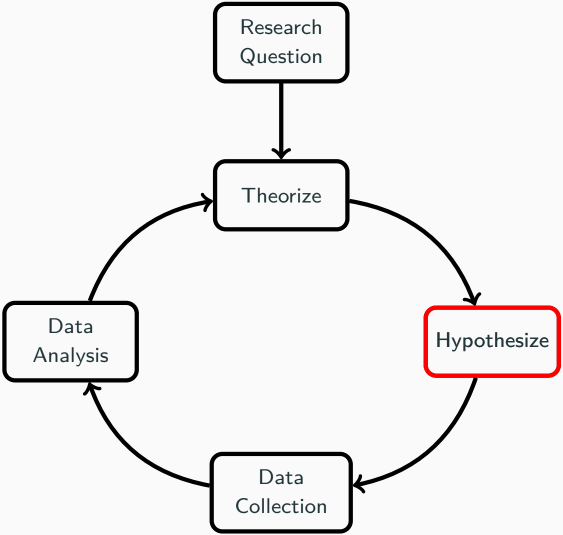
- A question that identifies the problem or puzzle one seeks to answer.
- E.g. Does economic development cause democratization?

Which part of the research process are we working on?



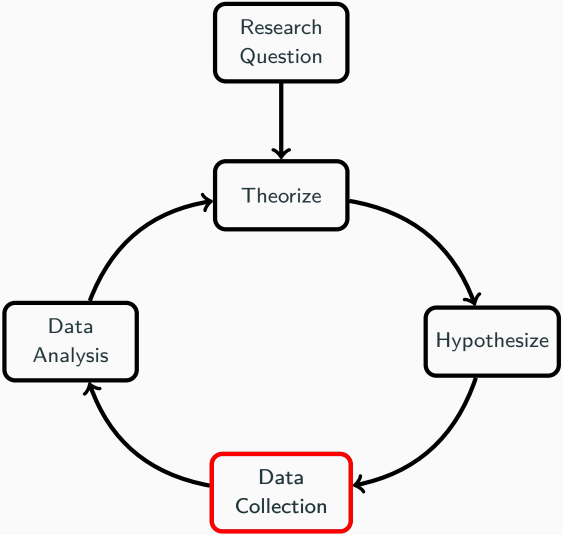
- An explanation of why or how something happens
- E.g. Economically developed countries are more likely to be democratic because they have a large middle-class that moderates political conflicts (Lipset 1959; Moore 1966)

Which part of the research process are we working on?



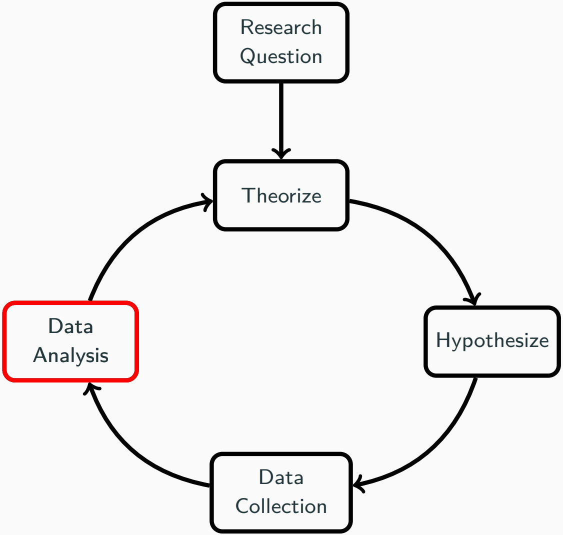
- A theory-based statement about a relationship we expect to observe
- E.g. Economically developed countries 1) are more likely to be democratic, 2) will have a large middle class, 3) will have more moderate political parties

Which part of the research process are we working on?



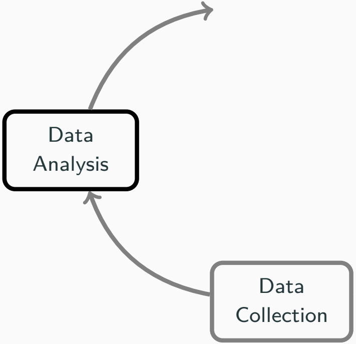
- Process of systematically gathering and measuring information on variables of interest
- E.g. For many countries, record the level of democracy; level of development; size of middle class; etc

Which part of the research process are we working on?



- Use the data you have collected to provide evidence either for or against your theory

Which part of the research process are we working on?



- We will only focus on the final two stages, with most emphasis on the analysis stage!
- PUBL0054 will focus on other parts of this process
- PUBL0010, PUBL0085, and PUBL0086 will introduce other types of data analysis

Description, prediction and causation

Within this scope, we will cover different types of research questions.

1. **Description**
   - Aims to describe differences in attributes across different units
   - E.g. Do men and women have different political preferences? Do politicans have the same priorities as their constituents?
2. **Prediction**
   - Aims to forecast likely outcomes of social processes
   - E.g. Who will win the next general election? What predicts civil war outbreaks?
3. **Causation**
   - Aims to establish the causal effects of one phenomenon on another
   - E.g. Did austerity cause Brexit? Does education increase income? What are the effects of immigration on employment?

Break

Introduction to Quantitative Data

Example

Who voted in the 2015 general election?

An important question in studies of representation is whether those who vote are similar to those who do not vote. This *descriptive* question can only be answered empirically: we need to look at data on the composition of voters and non-voters in an election.

We will use the 2015 British Election Study for this purpose.

- Survey conducted at each general election in the UK
- Face-to-face interviews of a representative sample of the population

Units and variables

There are 2 organising features of any data that we study

1. **Units ($i \in 1, ..., N$)**
    - The objects that we are studying
    - Usually these are the rows of the dataset
    - E.g. individuals; countries; companies; Members of Parliament; etc
    - We usually use i to indicate a unit, and N to mean the total number of units
2. **Variables**
    - Measurements of characteristics that vary across units
    - Usually these are the columns of the dataset
    - E.g. age; income; vote choice; profit/loss; GDP; etc

The first question we should ask when given data is "what are the units and variables in this data?"

Dependent and independent variables

An important conceptual distinction between types of variable:

- **Dependent variable (Y)**
  - Variable to be explained
  - Also called the outcome or response variable
- **Independent variables (X)**
  - Determinant(s) of the dependent variable
  - Also called the explanatory or predictor variables
  - Sometimes (somewhat confusingly) expressed as T or D

Units and variables (example)

In our British Election Study, the *units* are 1669 individuals who responded to the survey, and the *variables* are listed in the table below.

Variable	Description
`turnout`	1 if voted in 2015, 0 otherwise
`age`	Age in years
`gender`	Female/Male
`left_right`	Self-placement on left (0) to right (10) scale
`education`	Highest level of education achieved

Question: Which are the dependent and independent variables?

Looking at our data

We can load this data using:

```r
bes <- read.csv("data/bes.csv")
```

where

- `read.csv` tells R we want to read data from a `.csv` file
- `"data/bes.csv"` is the location in which our file is saved
- `<-` is the "assignment operator" which tells R that we want to save our data in memory
- `bes` is the name of the object we have saved (we can choose any name for objects)

Looking at our data

We can load this data using:

```r
bes <- read.csv("data/bes.csv")
```

The `head()` function shows the top 6 rows (units) in our data:

```r
head(bes)
```

```
##        turnout age gender left_right education
## 1        Voted  67 Female          5      GCSE
## 2        Voted  65 Female          5    Degree
## 3        Voted  65   Male          3    Degree
## 4        Voted  83   Male          5      None
## 5        Voted  56 Female          3      GCSE
## 6 Did not vote  40 Female          5      GCSE
```

Looking at our data

We can load this data using:

```r
bes <- read.csv("data/bes.csv")
```

The `dim()` function shows the number of units and columns in our data:

```r
dim(bes)
```

```
## [1] 1669    5
```

We have 1669 units, and 5 variables.

Looking at our data

We can load this data using:

```
bes <- read.csv("data/bes.csv")
```

The `str()` function gives information on the *structure* of our data:

```
str(bes)
```

```
## 'data.frame':    1669 obs. of  5 variables:
##  $ turnout   : Factor w/ 2 levels "Did not vote",..: 2 2
##  $ age       : int  67 65 65 83 56 40 44 39 30 68 ...
##  $ gender    : Factor w/ 2 levels "Female","Male": 1 1 2
##  $ left_right: num  5 5 3 5 3 5 5 5 5 1 ...
##  $ education : Factor w/ 4 levels "None","GCSE",..: 2 4
```

Looking at our data

We can load this data using:

```r
bes <- read.csv("data/bes.csv")
```

As the `str()` function revealed, R calls this `bes` object a “data frame”

A data frame is a data set with any number of variables (columns) measured for each of any number of units (rows)

Levels of measurement

- **Continuous/Interval**
  - Values indicate precise differences between categories
  - Differences (intervals) have the same meaning anywhere on the scale
  - E.g age
- **Categorical/Nominal**
  - Values indicate different, mutually exclusive categories
  - No relative information in the categories
  - E.g gender
- **Ordinal**
  - Values indicate relative differences between categories
  - Imply a ranking, but difference between categories may be unknown
  - E.g. educational achievement

Determining the correct level of measurement is important for making decisions about how to analyse your data.

Sums and Sigma notation

- N is the number of units or the **sample size**
- If $N = 100$, we have 100 measurements of each variable $(Y_1, Y_2, Y_3, ..., Y_N)$
- We will often want to refer to the **sum** of a variable:

$$Y_1 + Y_2 + Y_3 + ... + Y_N$$

  But this gets cumbersome if N is large!
- Instead, we will often use **Sigma notation**:

$$\sum_{i=1}^{N} Y_i = Y_1 + Y_2 + Y_3 + ... + Y_N$$

  where $\sum_{i=1}^{N} Y_i$ means "sum up all instances of Y starting from 1 and ending at N".

Measures of central tendency

To compare voters to non-voters, we need some way of summarising their characteristics. The most common summaries for most variables are those that measure the **central tendency** of the variable.

Central Tendency

The value of a “typical” observation, or the value of the observation at the center of a variable’s distribution.

We will consider three measures of central tendency:

1. Mean
2. Median
3. Mode

Mean

The mean is the “average” or expected value of a variable

It is denoted $\bar{Y}$ or $\bar{X}$, which can be read as “Y bar” or “X bar”

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

I.e. we add up the values of Y and divide by the sample size.

Median

The median is the value of a variable that divides the data into two groups such that there are an equal number above and below.

$$Median = \begin{cases} x_{((N+1)/2)} & \text{when N is odd} \\ \frac{1}{2}\left(x_{(N/2)} + x_{(N/2+1)}\right) & \text{when N is even} \end{cases}$$

where x_i is the ith smallest value of variable x.

I.e. the median is the middle value when the total number of observations is odd, and the average of the two middle values when the total number of observations is even

Mode

The mode is simply the most common value of a variable.

For example, in the BES data:

- 1282 respondents **voted**
- 387 respondents **did not vote**

The modal outcome for this variable is **voted**.

Mean, Median or Mode?

Which measure we use depends on the level of measurement:

- The **mean** is most appropriate for *continuous* variables
- The **median** is most appropriate for *ordinal* variables
- The **mode** is most appropriate for *categorical* variables

We will see examples throughout the course of many of these.

Implementing in R

Fortunately, all of these are easily implemented in R for a given variable.

- `mean()` is a function that calculates the mean
- the $ sign allows us to select a variable from our data

```r
## Mean
mean(bes$age)
```

```
## [1] 53.54763
```

Implementing in R

Fortunately, all of these are easily implemented in R for a given variable.

- `median()` is a function that calculates the median
- here we use `bes$left_right` to select the `left_right` variable

```
## Median
median(bes$left_right)
```

```
## [1] 5
```

Implementing in R

Fortunately, all of these are easily implemented in R for a given variable.

- `table()` counts the number of times each variable value appears
- The modal value here is “Degree”

```r
## Mode
table(bes$education)
```

```
##
##   None   GCSE Alevel Degree
##    383    392    339    555
```

Subsetting data

We will frequently want to **subset** our data in order to make statements about different groups of observations (e.g. voters v non-voters).

We can denote subsets of a variable using subscripts. For instance:

$$\bar{Y}_{X=1}$$

means “the average value of Y when X is equal to 1.”

We can then compare the average of Y in this subset to the average of Y in another subset (i.e. $\bar{Y}_{X=0}$).

Subsetting data

We can subset our data in R using the [,] brackets, which allow us to select certain rows and columns from the data.

To select **rows** use the space **before the comma**

```
bes[1:3,]
```

```
##   turnout age gender left_right education
## 1   Voted  67 Female          5      GCSE
## 2   Voted  65 Female          5    Degree
## 3   Voted  65   Male          3    Degree
```

Subsetting data

We can subset our data in R using the `[,]` brackets, which allow us to select certain rows and columns from the data.

To select **columns** use the space **after the comma**

```
bes[1:3,1:3]
```

```
##   turnout age gender
## 1   Voted  67 Female
## 2   Voted  65 Female
## 3   Voted  65   Male
```

Brackets and braces and parentheses

R makes different use of `( )`, `[ ]`, and `{ }` characters, and many new user errors arise from confusing these.

- *Parentheses* `( )` are used when calling a named function to do something to some objects.
  - As in `mean(bes$age)`, where we are using the `mean()` function on the data `bes$age`.
- *Brackets* `[ ]` are used to access a subset of an object.
  - As in `bes[1,]`, where we are accessing the first row (unit) in `bes`.
- *Braces* `{ }` are used for grouping multiple lines of code so that they act like a single line of code.
  - We will see these later in the module.

Logical values and operators

We can also use **logical values** and **logical operators** to select rows/columns of interest.

For instance, we can ask R to return all rows in our data where the respondent's value for `turnout` is "Voted":

```
bes$turnout == "Voted"
```

Where

- the $ says that we would like to access the `turnout` variable from the `bes` data
- the == says we would like the elements of that variable that **are equal to** the value "Voted"

We will learn more logical operators (such as <, >, >=) in the seminar.

Logical values and operators

We can combine == and [ ] to select **rows that match a criterion**:

```r
bes_voters <- bes[bes$turnout == "Voted",]
head(bes_voters)
```

```
##    turnout age gender left_right education
## 1    Voted  67 Female          5      GCSE
## 2    Voted  65 Female          5    Degree
## 3    Voted  65   Male          3    Degree
## 4    Voted  83   Male          5      None
## 5    Voted  56 Female          3      GCSE
## 11   Voted  33 Female          5    Alevel
```

Logical values and operators

We can combine == and [ ] to select **rows that match a criterion**:

```
bes_non_voters <- bes[bes$turnout == "Did not vote",]
head(bes_non_voters)
```

```
##         turnout age gender left_right education
## 6  Did not vote  40 Female          5      GCSE
## 7  Did not vote  44 Female          5      GCSE
## 8  Did not vote  39   Male          5    Alevel
## 9  Did not vote  30 Female          5      GCSE
## 10 Did not vote  68   Male          1      GCSE
## 19 Did not vote  36   Male          5      GCSE
```

Logical values and operators

We can combine == and `[ ]` to select **rows that match a criterion**:

```r
bes_voters <- bes[bes$turnout == "Voted",]
bes_non_voters <- bes[bes$turnout == "Did not vote",]
```

- `bes_voters` includes units who voted
- `bes_non_voters` includes units who did not vote

We can use these new datasets to characterise the central tendency of voters and non-voters for different variables.

Subsetting data

```r
## Age
mean(bes_non_voters$age)
```

```
## [1] 47.86563
```

```r
mean(bes_voters$age)
```

```
## [1] 55.26287
```

→ Voters are on average 7 years older than non-voters

Subsetting data

```r
## Education
table(bes_voters$education)
```

```
##
##   None   GCSE Alevel Degree
##    276    260    258    488
```

```r
table(bes_non_voters$education)
```

```
##
##   None   GCSE Alevel Degree
##    107    132     81     67
```

→ The modal qualification for voters is a degree, for non-voters it is GCSE

Subsetting data

```r
## Left-right placement
median(bes_voters$left_right)
```

```
## [1] 5
```

```r
median(bes_non_voters$left_right)
```

```
## [1] 5
```

→ Voters and non-voters are similar in terms of left-right placement

Example summary

Who voted in the 2015 general election?

Using data on 1669 individuals from the BES, we used measures of the **mean**, **median** and **mode** to investigate differences between voters and non-voters.

1. Voters are older, on average, than non-voters
2. Voters are more educated, on average, than non-voters
3. Voters and non-voters are similar in terms of ideology

Conclusion

What have we covered?

- Quantitative methods are a collection of tools we can use to investigate research questions and theories
- Quantitative data is a collection of information structured in terms of units and variables
- We can summarise variables by examining measures of central tendency
- We can compare groups of observations using these measures of central tendency

Recap of functions and notation

Code:

- `read.csv()` – load data into R from a `.csv` file
- `head()` – look at the first 6 rows of the data
- `mean()`, `median()` and `table()`
- `data_object[row_indexes, column_indexes]` – subsetting data
- `data_object$variable_name` – selecting variables from the data

Notation:

- i – a given unit
- N – the total number of units (sample size)
- $\sum_{i=1}^{N} X_i$ – add up all the numbers in X, from the first to the Nth
- $\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$ – the mean, or expected value, of Y

Seminar

In seminars this week, you will learn about ...

1. ... the Rstudio interface to R
2. ... objects and assignment
3. ... vectors and `data.frames`
4. ... subsetting

- Before coming to the seminar, install R and then Rstudio on your computer.
  - https://cran.r-project.org
  - https://rstudio.com/products/rstudio/download/#download