Introduction
ooo

Causality
ooooooooo

Statistics: The Basics
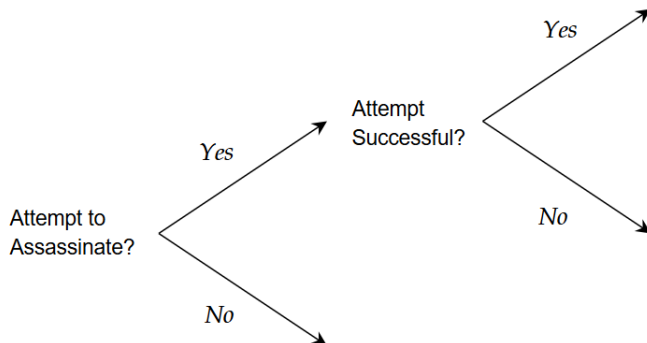oooooooo

Data Analysis in R
Causality & The Basics of Statistics

Ken Stiller

16th January 2026

Introduction
●○○

Causality
○○○○○○○○○

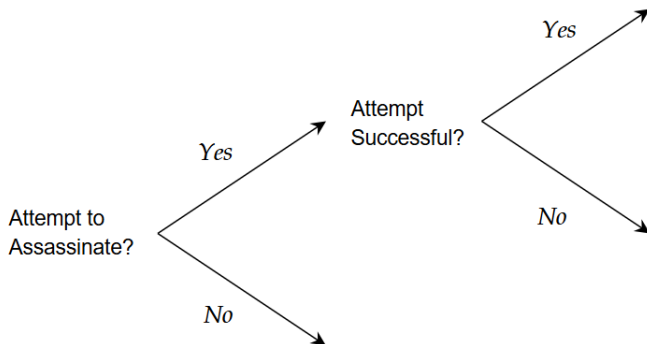Statistics: The Basics
○○○○○○○○

Syllabus: Data Analysis in R

1. Introduction
2. **Causality & Basics of Statistics**
3. Visualisations & Sampling
4. Prediction
5. Multivariate Regression
6. Probability & Uncertainty
7. Hypothesis Testing
8. Assumptions & Limits of OLS
9. Interactions & Non-Linear Effects

Introduction
○●○

Causality
○○○○○○○○○

Statistics: The Basics
○○○○○○○○

So About Those Assassinations...

Introduction
○○●

Causality
○○○○○○○○○

Statistics: The Basics
○○○○○○○○

So About Those Assassinations...

- ▶ 2. Step: **As-if random** *[Think Trump]*
- ▶ 1. Step: **Non-random** *[Why?]*

Introduction
ooo

Causality
●oooooooo

Statistics: The Basics
ooooooooo

Table of Contents

Introduction
000

Causality
○●○○○○○○○○

Statistics: The Basics
○○○○○○○○

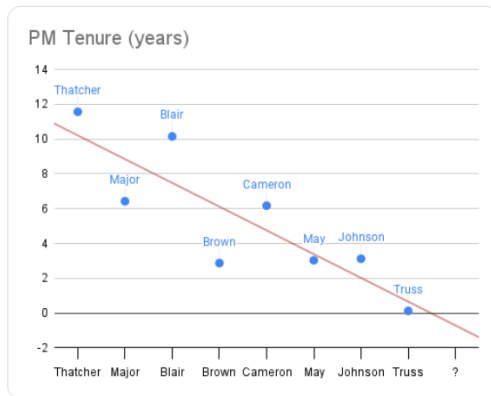# Why Do We Analyse Data?

Rob Sansom
@Sansom_Rob

Following current trends, the next PM will be in office for approximately minus 200 days

PM Tenure (years)

6:59 pm · 20 Oct 2022 · Twitter Web App

Introduction
000

Causality
00●000000

Statistics: The Basics
00000000

Definitions

Causality

Refers to the relationship between events where one set of events (the effects) is a direct consequence of another set of events (the causes). (Hidalgo & Sekhon 2012)

Data are Key

The process by which one can use data to make claims about causal relationships. (Hidalgo & Sekhon 2012)

Inferring causal relationships is a central task of science.

Examples

- ▶ What is the effect of peace-keeping missions on peace?
- ▶ What is the effect of church attendance on social capital?
- ▶ What is the effect of minimum wage on employment?

Introduction
000

Causality
000●00000

Statistics: The Basics
00000000

A Counterfactual Logic

Counterfactual Logic

If X had/had not been the case, Y would/would not have happened

Example:*Does college education increase earnings?*

▶ If high school grads had instead obtained a college degree, how much would their income change?

▶ If college grads had only obtained a high school diploma, how much would their income change?

Introduction
ooo

Causality
ooooo●oooo

Statistics: The Basics
ooooooooo

A hypothetical example

Imagine two students who are interested in getting a very high score on their thesis. They are considering the courses they should take and they are undecided between *Data Analysis in R* or sticking with *SPSS*.

Y_i : Thesis score is the outcome variable of interest for unit i.

$$D_i = \left\{ \begin{array}{ll} 1 & \text{if unit } i \text{ received the treatment (taking Data Analysis in R)} \\ 0 & \text{otherwise.} \end{array} \right.$$

$$Y_{di} = \left\{ \begin{array}{ll} Y_{1i} & \text{Potential thesis score for student } i \text{ with Data Analysis in R} \\ Y_{0i} & \text{Potential thesis score for student } i \text{ without Data Analysis in R} \end{array} \right.$$

Q: What is the effect of taking Data Analysis in R on your thesis score?

Introduction
000

Causality
00000●000

Statistics: The Basics
00000000

Defining the Potential Outcomes

Definition: Treatment

D_i : Indicator of treatment status for unit i

$$D_i = \left\{ \begin{array}{ll} 1 & \text{if unit } i \text{ received the treatment} \\ 0 & \text{otherwise.} \end{array} \right.$$

Definition: Observed Outcome

Y_i : Observed outcome variable of interest for unit i. (Realized after the treatment has been assigned)

Introduction
ooo

Causality
ooooooo●oo

Statistics: The Basics
ooooooooo

Defining the Potential Outcomes

Definition: Potential Outcomes

Y_{0i} and Y_{1i}: Potential Outcomes for unit i

$$Y_{di} = \left\{ \begin{array}{ll} Y_{1i} & \text{Potential outcome for unit } i \text{ with treatment} \\ Y_{0i} & \text{Potential outcome for unit } i \text{ without treatment} \end{array} \right.$$

Introduction
○○○

Causality
○○○○○○○●○

Statistics: The Basics
○○○○○○○○

The Fundamental Problem of Causal Inference

The Fundamental Problem of Causal Inference

It is impossible to observe for the same unit i the values $D_i = 1$ and $D_i = 0$ as well as the values Y_{1i} and Y_{0i} and, therefore, it is impossible to observe the effect of D on Y for unit i.

This is why we call this a missing data problem. We cannot observe both potential outcomes, hence we cannot estimate:

$$\tau_i = Y_{1i} - Y_{0i}$$

		Y_{1i}	Y_{0i}
Person 1	Treatment Group ($D = 1$)	Observable as Y	Counterfactual
Person 2	Control Group ($D = 0$)	Counterfactual	Observable as Y

Dealing with this is core challenge of social science research!

Introduction
ooo

Causality
oooooooo●

Statistics: The Basics
oooooooo

Causal Identification & Internal Validity

► **Association is not causation.**

► *Internal validity* refers to the concern that the difference in outcomes we observe between treated and untreated units are truly caused by the treatment.

► Some threats to internal validity are:
  ► Omitted variables
  ► Selection bias: Non-random selection into the treatment group
  ► Endogeneity and reverse causality

► **Randomised experiments v observational studies**
  ► But we have tools to make use of observations

Introduction
000

Causality
000000000

Statistics: The Basics
●0000000

Statistics: The Basics

Now, we'll briefly discuss the very basics of descriptive statistics:

- ▶ Types of variables
- ▶ Measures of central tendency
- ▶ Quantiles
- ▶ Standard Deviation

Introduction
ooo

Causality
ooooooooo

Statistics: The Basics
o●oooooo

Types of Variables: Discrete Variables

A **variable** is a measurement of a characteristic of a *unit of analysis* that (usually) varies across unit in a population of units.

There are different levels of measurement:

- ▶ **Nominal**: categorical measure, with no ordering
    - ▶ e.g . employed/unemployed; single/married/divorced
- ▶ **Ordinal**: ordered categorical measure
    - ▶ The distance between each category is unknown (strongly agree v agree)
    - ▶ e.g., many survey questions

Introduction
000

Causality
000000000

Statistics: The Basics
00●00000

Types of Variables: Continuous Variables

- ▶ **Interval**: numbers represent a quantitative variable - where we can quantify distances
  - ▶ The distance between each level is known and uniform
  - ▶ e.g . temperatures, voting cohesion, HDI, measures of democratisation? etc.
  - ▶ We can say that it's 10°C more than yesterday
- ▶ **Ratio**: There is a meaningful zero mark - which marks complete absence of the measure
  - ▶ We can divide measures and express them as multiples
  - ▶ e.g. age: someone might be twice as old as you are whereas this is not the case for temperature, or is it? (human development?)

Introduction
ooo

Causality
ooooooooo

Statistics: The Basics
oooo●oooo

Descriptive Statistics

▶ **Descriptive statistics** are simply that: they describe a large
amount of data by summarising it
  ▶ Think of all the values of a variable, which is not very informative -
    but we somehow want to make sense of them

▶ Why descriptive stats?
  ▶ Because we're often interested in what a typical unit (e.g
    person/country/district etc.) looks like
  ▶ Because it's useful to reduce many measurements to key indicators
    - either we're interested in them or as a preparatory step

▶ **Descriptive statistics $\neq$ inferential statistics**

Introduction
ooo

Causality
ooooooooo

Statistics: The Basics
ooooo●ooo

Measures of Central Tendency

- ▶ Measuring the *centre* of data - but which one?
  - ▶ **Mean:** most common, also referred to as the *average*
    - ▶ Sum of measures divided by number of observations

$$\text{mean} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- ▶ **Median:** More robust to *outliers*
  - ▶ Value at 50% mark of all observed values.

$$\text{median} = \begin{cases} \text{middle value} & \text{if number of entries is odd} \\ \frac{\text{sum of two middle values}}{2} & \text{if number of entries is even} \end{cases}$$

Example: data $= \{0, 1, 2, 3, 100\}$, mean $= 21.2$, median $= 2$

Range & Quantiles

- ▶ Measuring the **spread** or **dispersion** of data
  - ▶ **Range:** $[\min(x), \max(x)]$
  - ▶ **Quantile:** 'Portions' of the sorted data: quartile, quantile, percentile, etc.:
    - ▶ 25th percentile = lower quartile
    - ▶ 50th percentile = median
    - ▶ 75th percentile = upper quartile
  - ▶ **Interquartile Range (IQR):** Measure of variability and dispersion of the overall variable
    - ▶ One definition of *outliers*: over 1.5 IQR above upper quartile or below lower quartile

Example:

0%	25%	50%	75%	100%
9.9	16.2	29.2	42.3	75.2

Introduction
ooo

Causality
ooooooooo

Statistics: The Basics
ooooooo●o

Standard Deviation

- On average, how far away are data points from their mean?

$$\text{standard deviation} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

- Root-Mean-Square (RMS) of deviation from average
- For populations, it's divided by n instead of $n-1$
- Variance = standard deviation^2

Introduction
000

Causality
000000000

Statistics: The Basics
0000000●

Wrap Up

- ▶ Key points from this:
    - ▶ We're interested in causal effects, but can't observe counterfactuals - the fundamental problem of causal inference
    - ▶ Randomization is ideal, but possibilities with observational data - need a carefuly drafted identification strategy though
    - ▶ Description also important; there are different ways to describe data
- ▶ Next, we'll be talking about:
    - ▶ Visualisation, associations and sampling
    - ▶ Summarizing relationships between two variables
    - ▶ Visualizing data and distributions