# Lecture 1
# Introduction to Analytical Statistics

Dr Adam Mahdi

University of Oxford

October, 15, 2021

**This week:**

- Course overview

- Statistical thinking
- Evidence triangle
- Descriptive statistics

**Next week:**

- Probability (review)
- Inferential statistics
- Sampling distribution
- Central Limit Theorem

- Confidence intervals
- Hypothesis testing

# Course Overview

# Applied Analytical Statistics

**Instructors:**

- Lecturer: Dr Adam Mahdi [adam.mahdi@eng.ox.ac.uk]
- TA: Dr Alexander Zarebski [alexander.zarebski@zoo.ox.ac.uk]

**When:**

- Lectures: Friday, 13.00 - 15.00 [weeks 1-8]
- Tutorials: Thursday 9:00 - 10.00 [Group A]
- Tutorials: Thursday 10:00 - 10.30 [Group B]

# Applied Analytical Statistics

The a course about **Applied Analytical Statistics**

- **Statistics:** A body of tools and techniques for describing and analysing data.
- **Analytical Statistics:** These are statistics you can use to try and answer analytical questions about relations between cause and effect.
- **Applied Analytical Statistics:** The focus is on how these statistics are used to generate conclusions in the social sciences and other areas.

# How much information do we assume?

- This is a beginner level course; assumes you have no previous exposure to stats and only a kind of high school level maths.
- Aim to cover a few different techniques
- Develop the ability to implement, interpret and present the results from statistical models
- Statistical thinking

# Course overview

1. Introduction to analytical statistics
2. Sampling and statistical significance
3. Linear regression I
4. Linear regression II
5. Logistic regression
6. Multi-level modelling
7. Topics
8. Topics

Summative

# Deadlines

**Friday of Week 4, Michaelmas term**: please submit a short note, no more than one page, describing your idea for the summative project report.

- Include your research question and data sources
- We will provide feedback on the suitability of the proposal

**Friday of Week 0, Hilary term:** Please submit your final report via the Assignment Submission Site.

# Analytical Statistics

# Example: Average height of Londoners

Hannah was tasked with finding the average height of people living in London, UK in 2021. One way this could be done would be to measure the height of all Londoners and calculate this directly. Obviously this would be rather impossible since there are approximately nine million Londoners (in 2021). Instead, Hannah wandered around a park in South London, and asked people if they would mind being measured. Twenty people agree from those measurements Hannah calculated the average of their heights and reported back that the average height height is about 169 cm.

- **Population:** all the people living in London in 2021. A numerical summary of the population as a whole is a *parameter*.

- **Sample:** is the measured heights of the twenty people. The corresponding numerical summary of a sample is a *statistic*.

# Causality

What is causality?

- In short “the relationship between cause and effect” but can be tricky to define and verify precisely

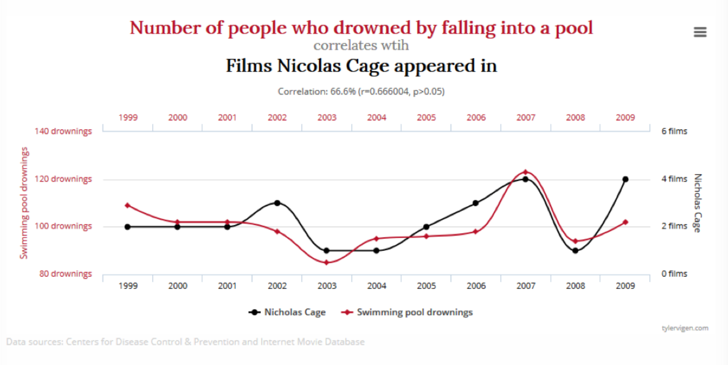
How is causality approached in quantitative social sciences?

- Define, as far as possible, a theory of the causes of a given phenomenon
- Test, using the best available data, whether the theory is supported
- Your conclusion will be that the data supports or undermines a causal theory.
- One study alone will never definitively prove a causal relationship. Point is to add to the evidence base

# Causality

Why is it difficult to make strong causal claims?

- **“correlation is not causation”**
- associations between variables could be driven by hidden third factors, or by random chance



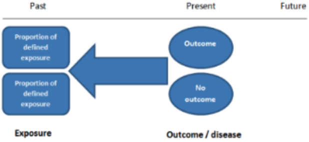
# Type of studies

- **Descriptive** (or nonanalytical) studies, aim to describe the data without trying to establish relationships between variables.
  - Case study
  - Case series study

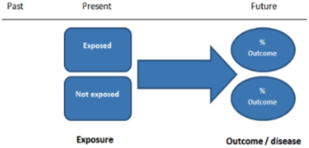
- **Observational** studies aim to draw inferences where the independent variable of interest is out of control of the researcher.
  - Case-control studies
  - Cohort studies
  - Cross-sectional studies

- **Experimental** studies are ones where researchers introduce an intervention and study the effects. Experimental studies are usually randomized, meaning the subjects are grouped by chance.
  - Randomized controlled trial
  - Non-RCT

# Case-control vs Cohort studies


Past
Present
Future
Proportion of defined exposure
Proportion of defined exposure
Outcome
No outcome
Exposure
Outcome / disease

Case-control studies

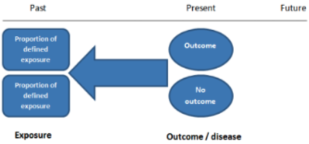


Cohort studies

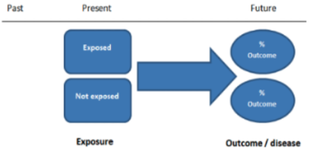
**Source:** https://s4be.cochrane.org/blog/2017/12/06/case-control-and-cohort-studies-overview/.

- **Case-control studies.** At the start of the study we have clearly defined two groups: one with the outcome (e.g. disease) or some condition of interest (referred to as cases) and one without the outcome (referred to as controls or control group). In case-control study we look back in time to learn which subjects in each group had the exposure by comparing the frequency of the exposure in the case of group to the control group.

# Case-control vs Cohort studies


Past
Present
Future
Proportion of defined exposure
Proportion of defined exposure
Outcome
No outcome
Exposure
Outcome / disease

Case-control studies



Cohort studies

**Source:** https://s4be.cochrane.org/blog/2017/12/06/case-control-and-cohort-studies-overview/.

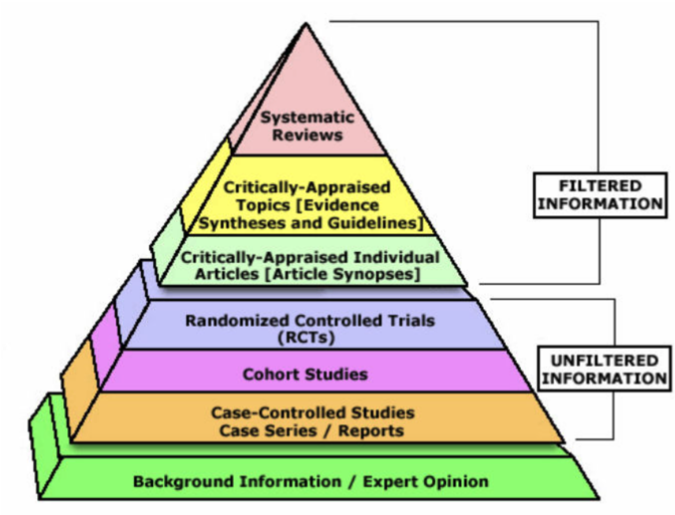
- **Cohort studies.** The participants do not have the outcome of interest to begin with. They are selected based on the exposure status of the individual. They are then followed over time to evaluate for the occurrence of the outcome of interest.

# Cross-sectional studies, RCTs

- **Cross-sectional study.** The investigator measures the outcome and the exposures in the study participants at the same time. Unlike in case-control studies (participants selected based on the outcome status) or cohort studies (participants selected based on the exposure status), the participants in a cross-sectional study are just selected based on the inclusion and exclusion criteria set for the study. Once the participants have been selected for the study, the investigator follows the study to assess the exposure and the outcomes.

- **Randomized controlled trial (RCT).** Eligible people are randomly assigned to one of two or more groups. One group receives the intervention (e.g. new drug) while the control group receives nothing or an inactive placebo. The researchers then study what happens to people in each group. Any difference in outcomes can then be linked to the intervention.

# Some types of studies are stronger than other


Systematic Reviews
Critically-Appraised Topics [Evidence Syntheses and Guidelines]
Critically-Appraised Individual Articles [Article Synopses]
Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies Case Series / Reports
Background Information / Expert Opinion
FILTERED INFORMATION
UNFILTERED INFORMATION

**Source:** https://s4be.cochrane.org/blog/2014/04/29/the-evidence-based-medicine-pyramid/

# Qualitative vs Quantitative Paradigms

**Quantitative:**

- Who does what? When?
- Characterised by having so many observations that in-depth knowledge of any particular observation is not feasible

**Qualitative:**

- Why? How?
- Focus on in-depth exploration of a small amount of cases.

# Qualitative vs Quantitative Paradigms

Example: *What drives people to post hate speech on social media?*

**Qualitative**. We could find and interview people who have produced hate speech, and ask them why they did it! This will uncover the extent to which people were conscious of what they were doing, and what their motivations were.

**Quantitative**. Observe the behavior of a large sample of users on social media, and look at how the volume of hate speech they post correlates with things that, theoretically, we expect to be related to posting hate speech.

# Qualitative vs Quantitative

- Part of SDS is about learning how to run analytical research projects
- These are projects where we try and understand socially important relations of cause and effect
- They allow us to evaluate whether a given dataset supports a given causal theory

# Descriptive Statistics

# Descriptive statistics

**Statistics:** collection, organisation and interpretation of data.

- Descriptive Statistics: Presenting, organising and summarising the data (usually a sample). These are characteristics of a sample

- Interential Statistics: Drawing conclusions about a population based on data observed in a sample.

# Descriptive statistics: types (some)

**Measures of frequency**

- Count, Percent, Frequency
- Use to show how often a response is given

**Measures of central tendency**

- Mean, Median, Mode
- Use to show how often a response is given

**Measures of dispersion**

- Range, Variance, Standard Deviation
- Use this when you want to show how "spread out" the data are. It is helpful to know when your data are so spread out that it affects the mean

**Measures of position**

- Mean, Median, Mode
- Use this when you need to compare scores to a normalised score

# Descriptive statistics: Table 1

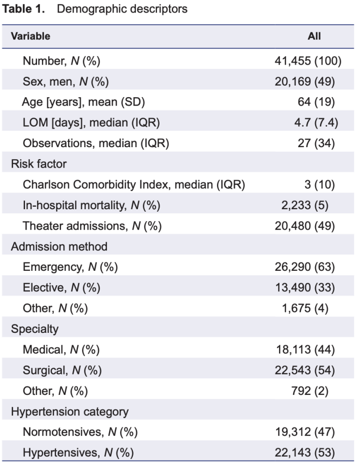
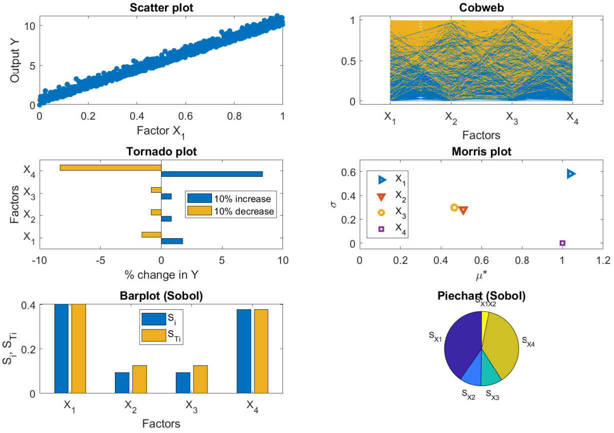


**Table 1.** Demographic descriptors

| Variable | All |
|---|---|
| Number, *N* (%) | 41,455 (100) |
| Sex, men, *N* (%) | 20,169 (49) |
| Age [years], mean (SD) | 64 (19) |
| LOM [days], median (IQR) | 4.7 (7.4) |
| Observations, median (IQR) | 27 (34) |
| Risk factor | |
| Charlson Comorbidity Index, median (IQR) | 3 (10) |
| In-hospital mortality, *N* (%) | 2,233 (5) |
| Theater admissions, *N* (%) | 20,480 (49) |
| Admission method | |
| Emergency, *N* (%) | 26,290 (63) |
| Elective, *N* (%) | 13,490 (33) |
| Other, *N* (%) | 1,675 (4) |
| Specialty | |
| Medical, *N* (%) | 18,113 (44) |
| Surgical, *N* (%) | 22,543 (54) |
| Other, *N* (%) | 792 (2) |
| Hypertension category | |
| Normotensives, *N* (%) | 19,312 (47) |
| Hypertensives, *N* (%) | 22,143 (53) |

**Source:** Mahdi et al. American J of Hyper, 32 (2019), 1154-1161.

# Descriptive statistics: data summary



G. Qian, A. Mahdi, Sensitivity analysis methods in the biomedical sciences, *Mathematical Biosciences* **323** (2020), 108306

# Summatives

**Aim:** To show that you can apply the skills you have built up in class.

A maximum 5,000 word project report, should include:

- **Question:** Describe an analytical research question with reference to existing literature on the subject. We do not expect an extensive literature review but students will be rewarded for clearly linking their question to a body of theory to be tested.
- **Data:** Describe the data, how it was collected including descriptive statistics.
- **Analysis:** Report on the analysis they have conducted.
- **Discussion:** Draw conclusions and thus answer the research question. Discuss the limitations of the analysis.
- **Code:** Code for the project can be submitted as an appendix which does not count towards the maximum 5,000 word limit.