L 5: Responsible Statistics

- Datasets
- Questions, objectives and uncertainties
- Statistics and the Law
- Semester 2, project 1

Population vs sample

- We use samples for inference about population parameters.
- Biases in these samples can invalidate this process.
- The result of any analysis is only as good as the data used as input. ‘Garbage in = Garbage out’.
- Need to be clear which data were used in any study, how they were collected, are there errors and whether there are any potential biases.
- From Huff (1954), ‘How to Lie With Statistics’:
  ‘Even if you can’t find a source of demonstrable bias, allow yourself some degree of scepticism about the result as long as there is a possibility of bias somewhere’
- From Reichman (1961), ‘The statistician must... rearrange the data that come to him and inspect them thoroughly in much the same way as a cabinet maker must ensure that he has the right timber and the right tools before commencing work’.

Definitions

- Lack of precision in definitions of populations and sub-populations can be problematic.
- The populations you define need to link to the question you want to ask.
- e.g developing a survey to assess what Londonders think of a particular political party.
- How to define London? By postal districts, by constituencies, City of London vs Greater London...
- A statistic for the number of deaths on train tracks is given as a worrying 4,712. It turns out this includes those that died in their cars on train tracks. Only 132 were train passengers. Huff (1954).

Example 1 (Reichmann 1961)

Women - One out of every Two gets Backache

Based on a dataset showing that 50 out of 100 women visiting a particular doctor admitted to backache.

Example 2

Building a model to assess risk of death from different levels of drug-taking

Based on a dataset of toxicology readings from postmortem blood samples taken from those whose primary cause of death (determined by coroner) was overdose.

The question

- From Reichmann (1961):
  'It may seem too obvious to iterate that... before attempting a survey one must be absolutely clear as to what information is required. But the obvious cannot be stated too often. If the researcher is not really clear as to exactly what information he requires, how can he expect others to know?'
- This quote is about survey questions but it is pertinent to all of statistical analysis. What is the question?!
- Before doing any analysis, first think about what the question is. The data that you need and the analysis that you do are both dependent on the question.
- Make sure you answer the question.
- Sometimes, there may not be an answer to the question (because there is not enough data, or the data are not relevant, etc.). Then you need to say so and explain why. Don't be tempted to answer a different question instead.

Uncertainty

- ‘One must never lose sight of the fact that statistics is concerned with reality’ (Reichmann 1961).
- In general, the less data you have, the more uncertain you are about your conclusions.
- It’s important to quantify uncertainty - difficult to make good decisions based on statistical output without knowing how unsure your conclusions are. Need to link to the real-world.
- Uncertainty can come from lots of different sources - have you considered all possible sources?
- E.g. in regression modelling we explicitly model uncertainty in the response variable conditional on the predictors. What about measurement error in the predictors or errors due to biases in the data?
- Do you have enough data for individual predictor variables? For example, you might have a lot of observations but only over a small number of years. Is it reasonable to fit a trend by year and extrapolate? And how can you model uncertainty due to limited years in the dataset?

Best Practice

- Compose a clear objective.
- Carefully define the population of interest with reference to this objective.
- Use random sampling (or some other experimental design, depending on the objective) to select samples.

In practice, much of the time we have to use existing datasets that were not collected with our objective in mind. So we must cast a critical eye on what the limitations of the dataset are and how these limitations will affect our conclusions.

Statistics and the law

- Much of forensic science consists of making statements about the extent to which data support a particular hypothesis. E.g., DNA, glass fragments, fibres, drug traces.
- This is a statistical problem.
- UK has adversarial legal system - prosecution vs defence.
- Two competing propositions, one from prosecution, one from defence.

Statistics and the law

How to assess the extent to which data support a particular hypothesis?

- E: evidence (e.g. chemical measurements of a glass fragment found on the suspect).
- H_p: prosecution proposition - what the prosecution is arguing regarding the evidence (e.g. the glass fragment is from a broken window at the crime scene).
- H_d: defence proposition - what the defence is arguing regarding the evidence (e.g. the glass fragment is from a broken beer bottle).

The likelihood ratio

Best way - compute the likelihood ratio:

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)}$$

If $LR > 1$ the evidence supports H_p, if $LR < 1$, the evidence supports H_d. From odds form of Bayes theorem:

$$\frac{P(H_p \mid E)}{P(H_d \mid E)} = \frac{P(E \mid H_p)}{P(E \mid H_d)} \times \frac{P(H_p)}{P(H_d)}$$

The likelihood ratio - continuous

The likelihood ratio for continuous evidence $E = \mathbf{x}$:

$$LR = \frac{f(\mathbf{x} \mid H_p)}{f(\mathbf{x} \mid H_d)}$$

Where f is the probability density function of $\mathbf{x}$ conditional on H_p or H_d. As before, if $LR > 1$ the evidence supports H_p, if $LR < 1$, the evidence supports H_d.
Also from (continuous) odds form of Bayes theorem:

$$\frac{P(H_p \mid \mathbf{x})}{P(H_d \mid \mathbf{x})} = \frac{f(\mathbf{x} \mid H_p)}{f(\mathbf{x} \mid H_d)} \times \frac{P(H_p)}{P(H_d)}$$

Transposed conditional

Important not to transpose the conditional when interpreting the numbers:

$$P(E \mid H_d) \neq P(H_d \mid E)$$

Called the prosecutor's fallacy. Risks miscarriages of justice as it misses out the prior probabilities.

Example

- E: red jumper fibre found at crime scene
- H_p: fibre is from jumper A, with 80% red fibres and 20% blue fibres
- H_d: fibre is from jumper B, with 10% red fibres and 90% blue fibres

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)} = \frac{0.8}{0.1} = 8$$

Evidence supports prosecution proposition.

Other options

Sometimes you may not have the data to compute the LR.
Could try:

- If have data for one, compute $P(E \mid H_d)$ or $P(E \mid H_p)$ and compare to a feasible range for the one you don't have data for.
- Only compute $P(E \mid H_d)$ or $P(E \mid H_p)$ but be clear that this risks misinterpretation and why.
- Note that for the continuous case it wouldn't make sense to use $f(\mathbf{x} \mid H_d)$ or $f(\mathbf{x} \mid H_p)$ as a statistic unless taking a ratio using the LR - could use $P(X > x \mid H_d) = \int_x^\infty f(\mathbf{x} \mid H_d)dx$ (or similar, depending on application) instead.
- The Forensic Science Regulator sets out the LR as the way that forensic evidence should be presented and interpreted.

Plots can help you interpret the raw numbers.

Expert witnesses

The duty of an expert witness is to help the court to achieve the overriding objective by giving opinion which is objective and unbiased, in relation to matters within their expertise. This is a duty that is owed to the court and overrides any obligation to the party from whom the expert is receiving instructions.

Expert witnesses

Expert evidence is admissible where [1]:

- It will be of assistance to the court: ‘must be able to provide the court with information which is likely to be outside a judge’s or a jury’s knowledge and experience, [...] must also be evidence which gives the court the help it needs in forming its conclusions. The role of the expert is to give their opinion based on their analysis of the available evidence. The Bench or jury is not bound by that opinion but can take it into consideration’
- The expert has relevant expertise: ‘The individual claiming expertise must have acquired by study or experience sufficient knowledge of the relevant field to render their opinion of value.’
- The expert is impartial: ‘The expert must be able to provide impartial, unbiased, objective evidence on the matters within their field of expertise. ’
- The expert’s evidence is reliable: ‘There should be a sufficiently reliable scientific basis for the expert evidence, or it must be part of a body of knowledge or experience which is sufficiently organised or recognised to be accepted as a reliable body of knowledge or experience.’

[1] https://www.cps.gov.uk/legal-guidance/expert-evidence

Criminal Practice Directions 2023 at 7.1.2

‘Factors which the court may take into account in determining the reliability of expert opinion, and especially of expert scientific opinion, include:

- the extent and quality of the data on which the expert’s opinion is based, and the validity of the methods by which they were obtained
- the validity of the methodology employed by the expert
- if the expert’s opinion relies on an inference from any findings, whether the opinion properly explains how safe or unsafe the inference is (whether by reference to statistical significance or in other appropriate terms)
- if the expert’s opinion relies on the results of the use of any method (for instance, a test, measurement or survey), whether the opinion takes proper account of matters, such as the degree of precision or margin of uncertainty, affecting the accuracy or reliability of those results the extent to which any material upon which the expert’s opinion is based has been reviewed by others with relevant expertise (for instance, in peer-reviewed publications), and the views of those others on that material

Criminal Practice Directions 2023 at 7.1.2

...

- the extent to which the expert's opinion is based on material falling outside the expert's own field of expertise;
- the completeness of the information which was available to the expert, and whether the expert took account of all relevant information in arriving at the opinion (including information as to the context of any facts to which the opinion relates);
- if there is a range of expert opinion on the matter in question, where in the range the expert's own opinion lies and whether the expert's preference has been properly explained;
- whether the expert's methods followed established practice in the field and, if they did not, whether the reason for the divergence has been properly explained.'

Semester 2 - project 1

- Released on Learn.
- Self organise groups. Let me know if problems.
- Deadline 14 Feb 16:00.
- Next two classes on project work.