Significance tests

Part 1: Critical appraisal

By: Noureddin Sadawi, PhD

University of London

Critical appraisal

- This refers to the process of **assessment of evidence**.
- This is done by systematically reviewing its:
  - relevance
  - validity
  - results.

How can we systematically judge the trustworthiness of results of a research study?

Problem formulation is key

‘**The formulation** of a **problem** is often **more** essential than its solution, which may be merely a matter of mathematical or experimental skill. To raise new **questions**, new possibilities, to regard old **problems** from a new angle requires creative imagination and marks real advances in science.’

Albert Einstein

PICO questions

These questions are highly useful when conducting or analysing studies:

P: What is the Population? Problem? Patient?

I: What is the Intervention being considered? test ? exposure?

C: Do we have a Comparative group (e.g. healthy controls)?

O: What are we measuring? What is the Outcome?

Example

P: Obese men with sleeping problems.

I: 40 minute walk and low energy diet every day for six consecutive weeks.

C: Normal lifestyle and diet.

O: Extent of sleeping problems.

Reliability and validity of outcome

How many outcomes does the study have?

Reliability:

If the conditions are different, are the measurement results consistently the same?

Validity:

How accurate is the measurement? Does it give a true assessment?

Type of outcome

- We have seen before that different types of data require different summary measures.
- The same applies for statistical analysis.
- **Numeric:**

  The outcome measurement is made on a number scale. (e.g. blood pressure)
- **Categorical:**

  The outcome is one of two or more categories or classes (e.g. severity of asthma: critical, mild or low).

Sample size and confidence interval (CI)

- Generally, the larger the study the more accurate the results.
- Confidence intervals should always be used when reporting estimates (CI is a measure of precision).
- Often we use 95% CI which gives the approximate range of population values (based on the sample(s)).
- Asthmatic children given seretide have 12 (95% CI (7,16)) less asthma attacks.
- Power calculation is useful (a sample size calculation that is conducted before the study).

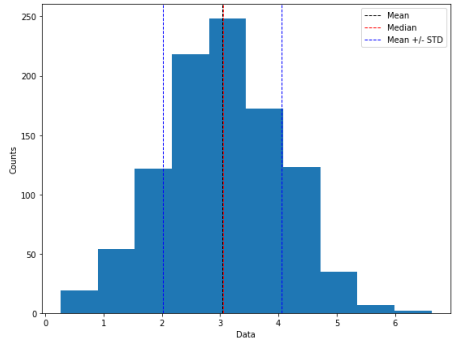
Summarising measurements

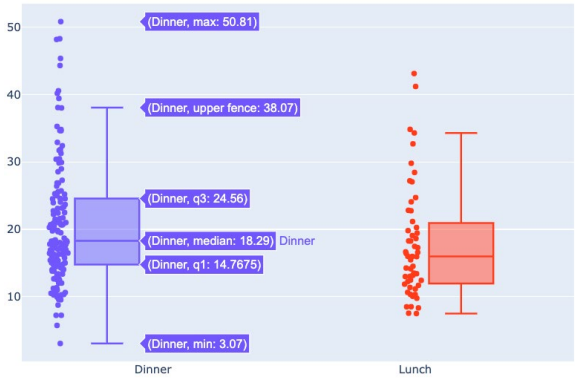
- **Numeric:** depends on the data distribution:

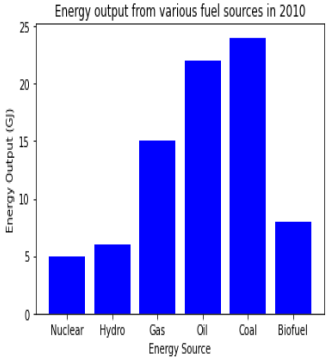
  Symmetric: mean and STD.

  Skewed: median, range and IQR.
- **Categorical:**

  mode, bar charts, pie charts and so on.







Summarising differences between two groups

Depends on the outcome values being compared (i.e. numeric or categorical).

Numeric:

The difference between means or medians (if the distribution of the outcome is skew).

Categorical:

Binary: some example methods are:

a) odds ratio (OR), more on it in logistic regression (Topic 6)
b) telative risk (RR), sometimes called risk ratio
c) the number needed to treat (NNT)

Statistical significance tests

- The correct test depends on:
    1. the type of outcome
    2. the number of groups
    3. if the groups are paired.
- The null hypothesis is usually: there is no difference between groups (e.g. treatment has no effect).
- The p-value represents the probability of the hypothesis being true based on the available data.
- If the p-value is close to zero, this is an indication that the null hypothesis is unlikely to be true.
- If the p-value is close to one, this is an indication that the null hypothesis is true (sample data is compatible with H_0).

Interpretation of p-value

Results are said to be significant if p-value is close to zero.

Results are said to be non-significant if p-value is away from zero (usually p-value > 0.05 means non-significance).

Two rules:

- The 95% CI will not contain 0 if p-value < 0.05.
- The 95% CI will contain 0 if p-value > 0.05.

If these two rules are not satisfied then the analysis has a problem and results are not to be trusted.