Introduction to experimental design

Yao-ban Chan & Kim-Anh Lê Cao

The University of Melbourne

Objectives of this workshop

- To understand that statistical experimental design is the entire process leading to a dataset.
- To be able to identify and control sources of variation that influence your results.
- To understand how we can obtain data that provide 'honest' and accurate answers to your questions.

"There's a flaw in your experimental design. All the mice are scorpios."

Samples and populations

Population

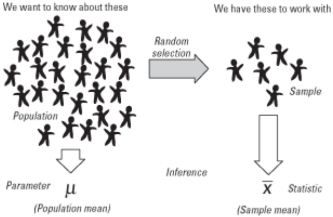
Complete set of items (also referred to as ‘subjects’, ‘individuals’, ‘units’), that we would like to make inferences about.

Sample

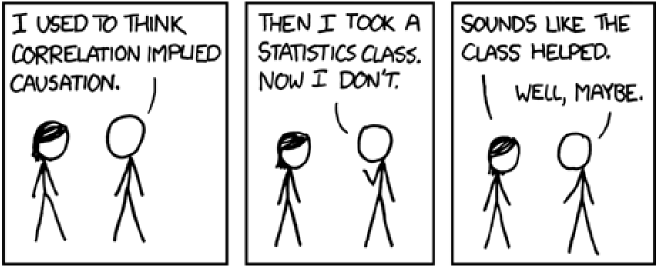
Subset (subgroup) drawn from the population.

Statistics

Uses the sample to make inferences about the population.



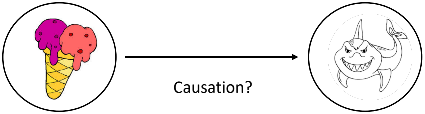
Correlation vs. causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Source: `xkcd.com`

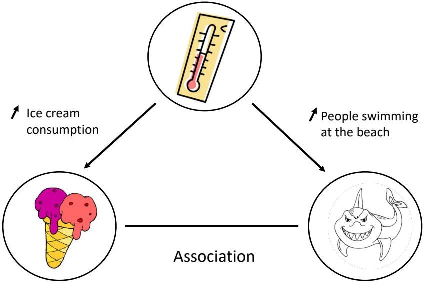
Correlation vs. causation

'An increase of shark attacks on Bondi beach is highly correlated with ice cream selling!'



Correlation vs. causation

'An increase of shark attacks on Bondi beach is highly correlated with ice cream selling!'



Separating correlation and causation is best done by experiment rather than observational study.

Confounding

A study on a neurological disorder wants to assess whether the expression of a particular gene increases for a specific treatment. Scientists recruit a number of patients, give them the treatment, and sequence their RNA.

What could go wrong?

Confounding

A study on a neurological disorder wants to assess whether the expression of a particular gene increases for a specific treatment. Scientists recruit a number of patients, give them the treatment, and sequence their RNA.

What could go wrong?

There is no control group to compare against.

Confounding

A bit later, the scientists decide to recruit more patients, give them a placebo, and sequence their RNA.

They find that the expression of the gene increases less for the treatment than the placebo! What happened?

Confounding

A bit later, the scientists decide to recruit more patients, give them a placebo, and sequence their RNA.

They find that the expression of the gene increases less for the treatment than the placebo! What happened?

The treatment group was sequenced on one plate, while the placebo group was sequenced on a different plate. Differences in experimental conditions were incorrectly attributed to the treatment.

The plate is a confounding variable for the experiment.

Confounding

The scientists start again: they recruit some patients, give half the treatment and half the placebo, and sequence their RNA, taking care to sequence all samples on the same plate.

They still find that the expression of the gene increases less for the treatment than the placebo! What happened?

Confounding

The scientists start again: they recruit some patients, give half the treatment and half the placebo, and sequence their RNA, taking care to sequence all samples on the same plate.

They still find that the expression of the gene increases less for the treatment than the placebo! What happened?

80% of the treatment group were women, while 15% of the placebo group were women. It happens that women have a lower expression of the gene than men.

Gender is a confounding variable for the experiment.

Confounding

Confounding

Confounding occurs when observed effects can be explained by more than one variable, and the effects of the variables cannot be separated.

Confounding variable

A confounding variable is a variable that affects the relationship between the variables in question.

Confounding

A variable is a confounding variable, if it is:

- a causal factor for the outcome (e.g., women have lower expression for the gene);
- associated with (related to) the exposure (e.g., there were more women in the treatment group);
- not caused by the exposure (...).

A lurking variable is a confounding variable which is not observed.

Dealing with confounding

Confounding can occur in many biological studies.

How can we deal with it?

- Stratification (blocking)
- Statistical correction
- Randomisation

→ These concepts form the basis of many of the key concepts involved in designing experiments (more to come next).

Stratification or blocking

We can *stratify* or *block* the study group.

We chop the population up into strata where all individuals have same ‘level’ of confounder.

This requires us to know what the confounding variables are!

It is standard to stratify by age and/or gender because they are common confounding variables.

Stratification or blocking

At first, we can consider just one stratum: for example, women aged 35–44 years.

Within one stratum there is not much confounding, as all individuals are the same in terms of age and gender.

However, results obtained apply only to that strata.

We will also need to consider how to deal with results for a population.

We can either combine the results for the strata appropriately to obtain a population value; or report results for different strata.

Statistical correction

We can directly account for confounding variables in our statistical model:

$$Response = Exposure + \text{effect of other variables} + Error.$$

This can be quite complex, and we will discuss this in the next workshop 'linear models' in May.

Randomisation

We can deal with confounding by randomly assigning the exposure variable in a controlled study.

If the exposure is randomly assigned, it cannot be ‘associated’ with any other cause — the groups with different exposure are all roughly the same.

Thus there cannot be any other confounding variables, whether we know them or not!

This is a very powerful approach for dealing with confounding.

Randomisation

Randomisation isn't always possible, if we don't control the exposure.

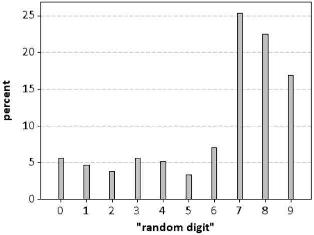
For the gene expression example, we must be able to assign patients to the treatment or the placebo.

In a study where smoking is the exposure, it can be problematic to assign subjects to smoke or not smoke!

Randomisation

How do we randomise?

Humans are notoriously bad at randomisation.



Any systematic method of allocating exposures (e.g. assign treatments in order of arrival, or lexicographical order) can fail.

Computers are much better (though still not foolproof).

Principles of experimental design

Definition

Experiment. An investigation in which the investigator applies some treatments to experimental units and then observes the effect of the treatments on the experimental units by measuring one or more response variables.

Definition

Treatment / Exposure. A condition or set of conditions applied to experimental units in an experiment.

Principles of experimental design

Definition

Experimental unit. The physical entity to which a treatment is randomly assigned and independently applied.

Definition

Observational unit. The unit on which a response variable is measured. There is often a one-to-one correspondence between experimental units and observational units, but that is not always true.

Principles of experimental design

Exercise

Soil moisture.

The exercise hand-out is available on our gitHub page.

Brainstorm in group of 3-4 for 10 min, we will ask some groups to debrief or comment.

Principles of experimental design

When we design an experiment, we need to think about:

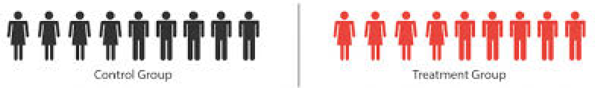
Validity — are we drawing the right conclusion?

Precision — how accurate are the results?

To ensure these, we apply certain general principles:

1. Control
2. Randomisation
3. Blocking
4. Replication
5. Blinding
6. Balance

Control (for validity)



Source: `Optimus.com`

- In a designed study, the group of patients (subjects) that are given no treatment or a standard treatment is called the control group.
- Comparison is key to identifying the effects on the response variable. If we have only one treatment group, then there is no way to identify what is and what is not the effect. The control group forms a baseline for comparison, to detect the effect of any other treatments.

Control (for validity)

MY HOBBY:
MY RASH SEEMS TO HAVE SHRUNK BY ABOUT 20% TODAY.
OH GOD SPIDERS
? ? ?
SNEAKING INTO EXPERIMENTS AND GIVING LSD TO THE CONTROL GROUP

(`xkcd.com/790/`, Creative Commons license)

This is a bad idea...

Randomisation (for validity)

Treatments are allocated to the experimental units randomly.

The effect of randomisation is to use randomness to even out the effect of uncontrolled (or unknown) confounding variables.

Randomisation (for validity)

Exercise

Plants.

Stratification / blocking (for precision and validity)

When we know certain factors (e.g., age, gender) have an effect on the response variable (e.g., survival), we ensure these factors even out in the different treatment groups, instead of trusting it to randomisation. This is done by blocking.

We form blocks (strata), which are groups of similar units (e.g., same gender, same age groups).

A blocked experiment uses randomisation of individuals separately within each block. This reduces the natural variation by making comparison on similar units.

Blocking removes a source of confounding (for validity) and also achieves higher precision.

Stratification / blocking (for precision and validity)

Both blocking and randomisation deal with nuisance factors (potential confounders), factors that are not of interest but might influence the outcome of the experiment.

Blocking is used when the nuisance factor is under our control. If the nuisance factor is not under our control, use randomisation.

The general rule is: ‘block what you can, randomise what you can not’.

Stratification / blocking (for precision and validity)

Exercise

Puppies

Replication (for precision)

We need replications. This means enough individuals in each treatment group so that chance variation can be measured and systematic effect can be seen.

The more replications, the more reliable (precise) the comparison of treatments.

Replication (for precision)

- Define your sample size
- Define your replicates
  - Biological
  - Internal
  - Technical

Internal/technical replicates are not a substitute for biological replication.

Example

In a proteomics experiment, we can use

- samples from different patients → biological variation
- same sample across different runs → technical variation (protocol)
- common internal reference spiked sample across runs → sensitivity / reliability of the equipment

Replication (for precision)

Exercise

Dairy cattle.

Blinding (for validity)

A blind experiment is an experiment where the subject (and sometimes the experimenter) does not know which treatment is used, to avoid conscious or unconscious bias on their part, which would invalidate the results.

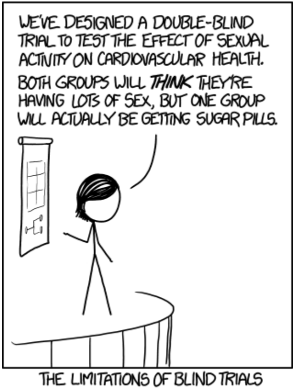
Example

When comparing the effectiveness of two different brands of a medical drug, both the patients and the doctors who administer the drug may be kept in the dark about the brand identities. Otherwise they may tend to prefer the brand they are familiar with.

Single-blind study Subjects do not know which treatment they got.

Double-blind study Neither the subjects nor those providing the treatment know which treatment was given.

Blinding (for validity)


WE'VE DESIGNED A DOUBLE-BLIND TRIAL TO TEST THE EFFECT OF SEXUAL ACTIVITY ON CARDIOVASCULAR HEALTH.
BOTH GROUPS WILL *THINK* THEY'RE HAVING LOTS OF SEX, BUT ONE GROUP WILL ACTUALLY BE GETTING SUGAR PILLS.
THE LIMITATIONS OF BLIND TRIALS

(`xkcd.com/1462/`, Creative Commons license)

But blinding isn't always possible. . .

Balance (for precision)

Balance means each treatment is applied to the same number of individuals (study units).

This is desirable when possible, as it simplifies the analysis, and gives the most precise comparison.

It is sometimes defeated by nature (e.g., some patients withdraw from the study).

Two main types of study

- Experimental study: The investigator assigns the treatment to the individuals in the study with the objective of observing the outcome and comparing the results.
  - Only way to show causation
  - aka controlled experiments (e.g., clinical trials)
- Observational study: The investigator selects individuals with different treatments. The outcome is then observed and the results compared.
  - Cohort studies
  - Case-control studies
  - Cross-sectional studies

Observational studies

A cohort study identifies and follows a cohort over a period of time. (Prospective study)

Exposure (e.g., smoking status, a drug treatment) is not assigned, the investigator is just an observer.

Commonly we identify the exposure status at the start of the study and then compare outcomes in the exposed cohort and unexposed cohort.

Because exposure is not assigned, there is no randomisation. Thus the study will observe association, but not causality.

Observational studies

In many cohort studies, only a tiny minority who are at risk actually develop the disease. Thus we need a very large sample, which is expensive.

In a case-control study, we sample cases and controls separately to avoid this problem.

Disease status is identified at the start of the study, then we look back to record their exposure status. (Retrospective study)

Observational studies

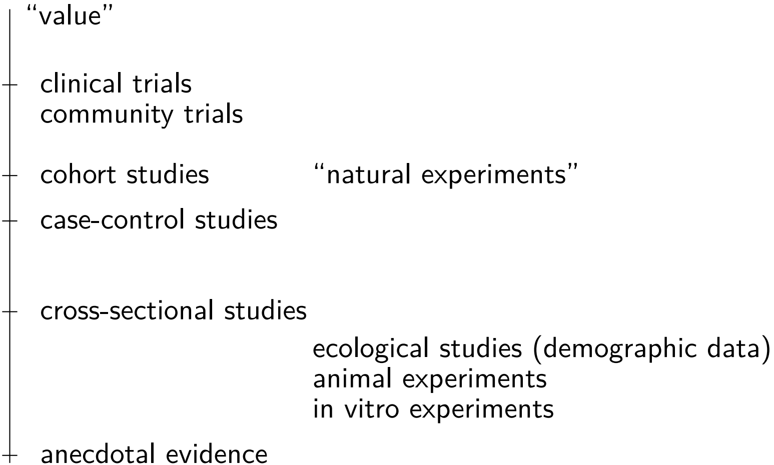
A cross-sectional study is an observational study in which exposure and outcome are determined at the same point in time.

In contrast, prospective and retrospective (longitudinal) studies contain information from more than one point in time.

Cross-sectional studies cannot determine temporal relationships, but are usually cheaper.

They are good for estimating the prevalence of disease (the proportion of the population which have it).

Review of study types: the hierarchy of evidence



“value”

clinical trials
community trials

cohort studies “natural experiments”

case-control studies

cross-sectional studies

ecological studies (demographic data)
animal experiments
in vitro experiments

anecdotal evidence

Hypothesis testing

Sample size calculation is conducted via power analysis.

To understand power analysis, we need to understand statistical tests and hypothesis testing.

Definition

A statistical test is based on the concept of **proof by contradiction** with the following steps:

1. State the null hypothesis H_0,
2. State the research hypothesis (alternative hypothesis) H_1,
3. Conduct statistical test,
4. Decide if we reject H_0 and draw conclusions

Hypothesis testing

We need to simplify the research question into two competing hypotheses:

- the null hypothesis (H_0), believed to be true (or has not been proved yet), against
- the alternative (research) hypothesis (H_1) that states the statistical hypothesis test to establish

→ we carry out an experiment to reject the null hypothesis.

Example

H_0: *on average, there is no difference between the two drugs*, against
H_1: *on average, the two drugs have different effects*

Type I and type II errors

- The probability of wrongly rejecting H_0 when H_0 is true is called the *significance level*, generally denoted α (usually arbitrarily set to $\alpha = 5\%$).
- *Type I error*: we reject the null hypothesis when it is true (probability of type I error is α)
- *Type II error*: we do not reject the null hypothesis when it is false and the research hypothesis is true (probability of type II error is β)
- The **power** of a statistical test is the probability that it correctly rejects the null hypothesis when the null hypothesis is false, i.e. the ability of a test to detect an effect, if the effect actually exists. Power $= 1 - \beta$.

Types of power analyses

Prior analysis: done before a study takes place.

- Controls both type I error α and type II error β and therefore the power of the statistical test $1-\beta$
- Used to determine the necessary sample size of a test

You need to provide:

1. Statistical test used
2. Desired α
3. Desired power $1-\beta$
4. Effect size expected to be detected* (measure of the difference between H_0 and H_1)

*This is the tricky bit

Types of power analyses

Post-hoc analysis: after a study is conducted.

- Given sample size, α and specified effect size, we obtain the power.
- Provides critical evaluation of β associated with a false decision in favor of H_0

The effect size depends on the statistical test

Example

If we plan to compare a particular variable between two groups of samples using a $t-$test, then the effect size is the difference between the two means divided by the common within-group standard deviation.

This means that *prior to the experiment*, we need to decide on:

- The null and research hypothesis

  $\updownarrow$

- The statistical test

  $\updownarrow$

- The experimental design

Effect size is important in sample size calculation

If the effect size used in the calculation is larger than the true effect, then the sample size calculated for the experiment will not be enough to achieve the target power.

Unless planned at the beginning of the experiment, sample size usually cannot be increased and there is a large risk the experiment will fail.

Specifying the effect size

- Select a 'relevant' difference (i.e. from a clinician's or patient's perspective)
- Select a realistic difference based on prior evidence and information (e.g. pilot data, existing data)

  (In high-throughput experiments, if we use univariate tests, this becomes complicated as we estimate the effect size of pilot data for *each* gene / protein / bacteria).

Example in R for a single t-test

Input (see ?power.t.test)

- delta: true difference in means
- sd: standard deviation (default $= 1$)
- sig.level: type I error α
- type: choose btw 'two.sample', 'one.sample', 'paired' test
- alternative: choose btw 'two.sided', 'one.sided' test

Example:

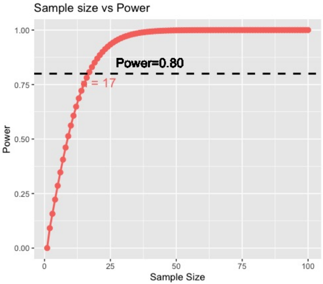
```
power.t.test(power = .80, delta = 1, type = 'two.sample',
             alternative = 'two.sided')
```

Output for a single t-test

```
 Two-sample t test power calculation

              n = 16.71477
          delta = 1
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```



Here we ran this command line for different power values

The R package `ssize.fdr` would provide more robust sample size calculation

Considerations for more than one test (e.g transcriptomics)

Assume you have conducted a univariate statistical test on some pilot transcriptomics data, and ranked each gene to their (decreasing) p-values.

To estimate the effect size for the power analysis you can calculate:

- The mean effect size of the top % DE genes (% defined by you, e.g top 1% of all genes)
- The effect size of each of the top % DE genes, estimate sample size for each and provide a range value
- The effect size of the top gene (might be optimistic)
- The effect size of all genes (might be pessimistic)

Power calculation tools

Besides R packages and R functions, I use the extensive desktop software G*Power.

G*Power requires some advanced knowledge in conducting power analysis (and choosing the appropriate statistical test!).

See small demo

`https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower`

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175-191.

'To consult the statistician after an experiment is finished is often merely to ask them to conduct a post mortem examination.

They can perhaps say what the experiment died of.'

(Sir Ronald Fisher 1890-1962).

- When you design your experiment, consider your research hypothesis, potential confounders, and the statistical method you will use to analyse your data.
- Learn from previous mistakes / previous data.
- Experiments are expensive and you can't go back. Dealing with batch effects after the data are generated would take an extra 80% time in analysing data.
- Sample size calculations need to be considered with caution.