Study design issues and sample size

Study design issues and sample size

- Observational studies
  - Types of observational studies
  - Strength and weaknesses
- Experimental studies
  - Randomisation
  - Blinding
  - Placebo-effect
  - Repeatability
- Sample size
  - Effect size
  - Significance level
  - Power

Concept of study design

- **Natural variation** forms the basis for statistics and why we need properly designed research studies
- We wish to assess whether the results of a study are consistent with this natural variation or whether there really is an effect
- The studies that allows us to make firm conclusions about a factor are the ones where everything is the same apart from this factor of interest
  - If we then see differences in the outcome measure that are inconsistent with natural variation, this should be due to the factor of interest

Observational Studies

- No intervention by the investigator
- Some observational studies describe the frequency, patterns and trends of an outcome of interest
- These can be used to help generate research questions
  - A **case report** is a profile of a single patient reported in detail
  - A **case series** is an extension of a case report to include a number of patients with a given condition
  - A **cross-sectional study** involves obtaining information about prevalence and association of a given condition in a population of interest by including 'controls' (i.e. those without the condition)
- Other observational studies can be more analytical than these (e.g. cohort studies, case-control study)

Cohort Studies

"Study outcomes by exposure"

Process:

- Identify a suitable group of subjects **at risk**
- Follow them over time
- Compare health outcome of interest in:
  - Subjects exposed to/have risk factor
  - Subjects not exposed to/do not have risk factor
- No direct intervention by investigator
- These can be carried out prospectively or retrospectively

Case-Control Studies

"Study exposures by outcome"

Process:

- Identify a suitable group of subjects with outcome of interest ('**cases**')
- Select '**controls**' who do not have the outcome of interest from the population who were at risk
- Compare past exposures to risk factor(s) in both cases and controls
- No direct intervention by investigator
- These are carried out retrospectively

Observational Studies: Strengths & Limitations

Strengths:

- Experimental design may not be ethical
- Can be relatively cheap/quick to carry out
- Methods and results are simple to interpret
- Collect detailed information on the risk exposures and health outcomes of interest and target research (e.g. rare exposures/outcomes)
- Information can also be collected for controls

Limitations:

- Results may not be generalizable
- Causation or association?
- Extraneous factors cannot be manipulated by the investigators (i.e. prone to confounding and bias)

Confounding

Confounding is when an observed association between a factor and an outcome is actually due to **the effects of another factor**

This can result in:

- An observed difference when no real difference exists
- No observed difference when a true association does exist
- An underestimate of an effect
- An overestimate of an effect

Confounding reflects the natural relationships between lifestyle, habits and other characteristics

It cannot be removed but can be allowed for in the design and analysis

Confounding - example

Example: the association between smoking and death from liver cirrhosis

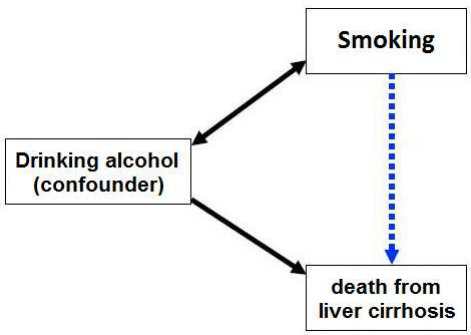
	Non-smokers			Smokers			
	No. deaths/no. of men	Death rate per 1000 (A)		No. deaths/no. of men	Death rate per 1000 (B)		Relative risk (B÷A)
All	9/1000	9		15/1000	15		1.70

It seems that smoking is associated with a higher death rate

?

Confounding - example

- But we also know that:
  - Drinking alcohol is a cause of liver cirrhosis
  - Alcohol drinkers tend to smoke
- So drinking could be a confounder



Confounding - example

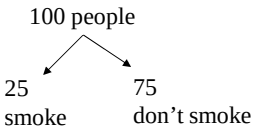
	Non-smokers			Smokers			
	No. deaths/no. of men	Death rate per 1000 (A)		No. deaths/no. of men	Death rate per 1000 (B)		Relative risk (B÷A)
All	9/1000	9		15/1000	15		1.70
Non-drinkers	**2/660**	**3**		**1/340**	**3**		**1**
Drinkers	**7/340**	**21**		**14/660**	**21**		**1**

- To allow for drinking alcohol, we simply divide the data into 2 groups, and then we look again at the association between smoking and death rate
- Conclusion: no association

There are more sophisticated methods to do this, that can also allow for several confounders at the same time (multivariable methods)

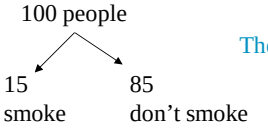
Confounding cannot be removed from a research study, but can be allowed for in the study design and statistical analysis (though may complicate the results)

Bias



If no-one has lied, then the true smoking prevalence is 25%

But if 10 smokers lie *and* no non-smokers lie then the observed smoking prevalence is 15%:



The study result is biased

But if 10 smokers lie *and* 10 non-smokers lie then the observed smoking prevalence is 25%

The study result is not biased

Bias

- **Bias** is usually a design feature of a research study that affects how people (or things) are selected, treated, managed or assessed
- **Bias** is any systematic difference that results in an incorrect estimate of the effect
- **Common types of bias:**
  - Selection bias – the experimenter deliberately chooses the fitter people for one of the study groups
  - Reporting bias – subjects in one of the study groups systematically over/under-report a certain issue
  - Observer bias – the observer tends to systematically over/under-report a particular factor in one of the study groups
  - Measurement bias – a systematic error is introduced by an inaccurate measurement tool (e.g. set of poorly calibrated scales) in one of the study groups
- Bias is a problem because it is:
  - can be difficult to prevent
  - will be difficult to allow for in the analysis because it often cannot be measured

Experimental Research

- Experimental studies involve the investigator intervening in some way to affect the outcome
- These can be laboratory experiments, animal studies or clinical trials
- Experimental research provides data from which firmer conclusions can be made compared to observational studies
- Study design is very important:
  - Must consider all possible confounders and remove any potential biases
  - Suitability? (cost, size, time to complete)
- Key concepts
  - Randomisation
  - Blinding
  - Placebo-effect
  - Repeatability

Experimental Research - Randomisation

- Randomisation is the process by which the allocation of subjects to the study groupings is by chance (i.e. cannot be predicted)
- Importance of randomisation
  - Ensures a fair comparison of study groups
  - Minimises the effect of bias and confounding
  - Helps the study groups to be similar at baseline in terms of known and unknown factors except for the intervention itself
  - If the study groups have similar characteristics, any observed difference in outcome can then be attributed to the factor being tested
- We can ensure that important factors are equally balanced between the study groups at this stage

Experimental Research - Blinding

- Blinding is the process of concealing which group a subject has been (randomly) allocated to
- Knowledge of which group a subject is in can bias the results of a study, so blinding further minimises the potential for bias
- The group allocation could be blinded to subjects, investigators and other researchers
- In research:
  - Single-blind means that only the subjects are blind to the allocated group
  - Double-blind means that neither the subject nor the investigators know the allocated group

Blinding and the placebo-effect

- The placebo-effect is a bias related to the perceptions and expectations of the subject or researcher
- If the study group (and intervention received) is known, then the *placebo-effect* can create an association when there really is none
- Blinding is a way to protect against this
- Blinding can also be done in other studies (e.g. blind review of imaging scans or tissue samples, without knowing the status of the patient/animal/cells)
- Attempting to have some form of blinding (only if possible and useful), can help strengthen study conclusions
- Using objective outcome measures rather than self-reported measures (e.g. quality of life score, pain score) further protects against the placebo-effect

Example of the placebo-effect

Treating angina pectoris

- After 1939 it was believed that ligating the internal mammary artery would increase blood flow to the cardiac muscle
- The symptoms of angina would then diminish
- In practice, about 75% of patients reported improvement after the operation

Example of the placebo-effect

Trial in 1959 of 17 angina patients:
8 patients randomised to receive artery ligation
9 patients randomised to receive skin incision on chest

	Average subjective improvement
Ligation arm:	32%
Not ligated arm:	43%

2 patients demonstrated significant improvement in endurance (i.e. could walk for 10 mins without angina)
Both were in the non-ligated arm

This, and another similar trial, stopped this practice (and saved much morbidity and mortality associated with the operation)

Example of the placebo-effect

- About half of patients with osteoarthritis of the knee report pain relief after arthroscopic surgery, an expensive procedure
- Randomised blinded trial of 180 patients in 2002
  - 60 placebo
  - 61 lavage
  - 59 debridement
- Mean knee pain score after 1 and 2 years (0 to 100=no pain to most pain):
  - Placebo 49 52
  - Lavage 55 54
  - Debridement 52 51
- Conclusion: no evidence of any effect of arthroscopic surgery on knee pain

Experimental Research - Repeatability

- Repeatability is the degree to which a measurement provides a similar result each time it is performed on a given subject or specimen
- When measuring a group of subjects, the variability of observed values is a combination of the variability in their true values and measurement error
- Consider a study to measure height in the community:
  - if we measure height twice on a given person and get two different values, then one of the two values must be wrong (invalid)
  - if study measures everyone only once, errors, despite being random, may not balance out
  - final inferences are likely to be wrong (invalid)

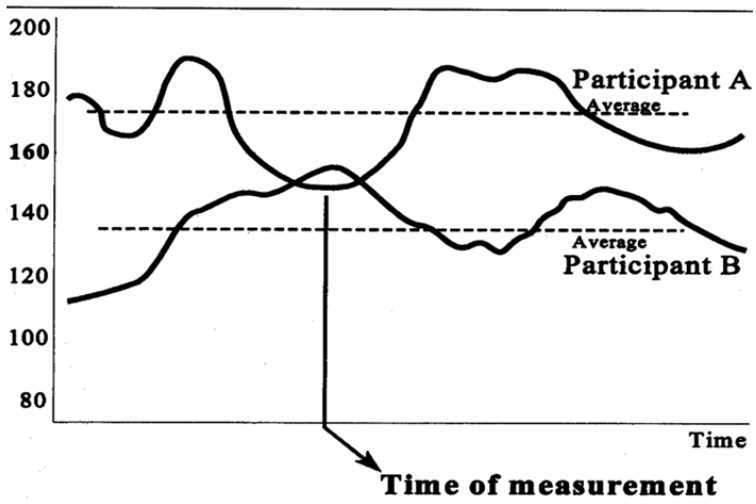
Repeatability - Sources of Measurement Variability

- Observer
  - within-observer (intra-rater)
  - between-observer (inter-rater)
- Instrument
  - within-instrument
  - between-instrument
- Subject
  - within-subject

Repeatability - Sources of Measurement Variability

- e.g. plasma HIV viral load
  - *observer:* measurement to measurement differences in tube filling, time before processing
  - *instrument:* run to run differences in reagent concentration, PCR cycle times, enzymatic efficiency
  - *subject:* biologic variation in viral load

Repeatability – within-subject measurement variability



Sample size

- The size of a study (whether it be on people or a laboratory experiment) is crucial to study design
- A good study design should be able to answer the research question with the minimum number of subjects possible
- If the study is too small:
  - you may miss important differences (because of chance variation)
  - 95% confidence intervals could be wide
  - difficult to make robust conclusions
  - you may see spurious associations
- If too large, you waste resources because you could have found a clear answer sooner

Consider study comparing exposure A with B, and the endpoint is the 1-year death rate

	Number of deaths observed (%)			
No. of patients per arm	**Exposure A**	**Exposure B**	**Difference**	**Comment**
100	15 (15%)	20 (20%)	5 fewer deaths	Difficult to distinguish a real effect from chance
100	5 (5%)	40 (40%)	35 fewer deaths	A difference this big is unlikely to all be due to chance
1000	150 (15%)	200 (20%)	50 fewer deaths	A difference this big is unlikely to all be due to chance

Method of sample size estimation

The method used will depend on the following

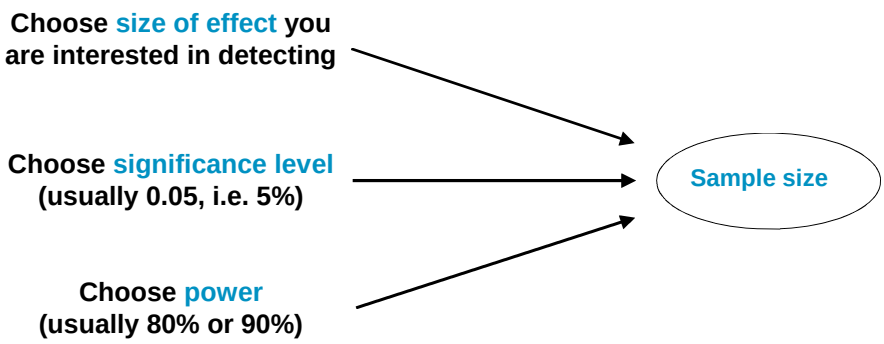
- The type of outcome measure used (i.e. taking measurements on people, counting people, time-to-event data)
- You choose what you think is an important endpoint

- The objectives:
  - Factor A has different effect to Factor B
  - Factor A has similar effect to Factor B
  - Association between two measurements (e.g. in regression)
  - Examining risk factors (e.g. in regression)

- Whether you are comparing two different groups (unpaired data), or comparing two measurements on the same group (paired data)

- Confidence interval for a single percentage or single mean value

Steps in choosing Sample Size



What is an effect?

An “effect size” is used when we are making quantitative comparisons

Comparison	Effect size
Comparing two or more groups:	
Taking measurements on people/objects	Difference between 2 means or medians
Counting people/objects	Relative risk, risk difference
Time-to-event	Hazard ratio, difference in median survival
Comparing two measurements on the same person/object (e.g. regression)	Regression coefficient, correlation coefficient

What is significance level?

- At the end of the study we will perform a significance test and obtain a p-value
- This will tell us how likely we would be to observe an effect as large as the one we have found simply by chance, if there really were no effect
- The p-value can also be thought of as the probability of making the wrong conclusion that an effect exists when in fact there is no real effect (we want this to be small, say 5% or 1%) – i.e. **the error rate**

What is power?

Power is the probability that a given effect is detected if there is truly an effect

Example: Suppose a standard treatment has a cure rate of 75% and a new treatment is expected to have a cure rate of 90%:

At the end of the study we want to be able to say:
"A difference of 90% vs 75% (i.e. 15 percentage points) is statistically significant at the 5% level"

↓

Power: We want an 80% probability of being able to make this statement (or, if there really is a difference of ≥15% the probability of detecting such a difference will be 80%)

Sample size: A trial size of 200 patients is expected to allow this

EXAMPLE OF SAMPLE SIZE

Rashighi M, Agarwal P, Richmond JM, Harris TH, Dresser K, et al. (2014) CXCL10 Is Critical for the Progression and Maintenance of Depigmentation in a Mouse Model of Vitiligo. Sci Transl Med 6: 223ra223

Study design

The overall study design was based on controlled laboratory experimentation using *ex vivo* human tissue samples and a mouse model for *in vivo* mechanistic studies. The research objectives at the outset of the study were to test the hypothesis that IFNγ-inducible chemokines were responsible for the recruitment of autoreactive T cells to the skin. This hypothesis was formed on the basis of previously reported observations in our mouse model (26). Sample size was determined using the approach described by Dell, *et al.* (60). Briefly, each experiment was powered to detect a difference between group means of twice the observed standard deviation, with a power of 0.8 and a significance level of 0.05. Replicate

N=10

Shakur H, Roberts I, Bautista R, et al. CRASH-2 trial collaborators. Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (CRASH-2): a randomised, placebo-controlled trial. Lancet. 2010;3;376:23–32.

Statistical analyses

The statistical analysis plan was sent to all ethics committees and regulatory agencies before unblinding. Because the risk of death might be around 20%, and even a 2% survival difference (corresponding to an RR of death with tranexamic acid of 0·9) would be important, a trial of 20000 patients was planned, which would then have an 85% chance of achieving a two-sided p value of less than 0·01 and a 95% chance of a two-sided p value of less than 0·05.

N=20 000

Sample size determination

Sample size goes up

as effect size (difference) goes down	→	harder to detect small differences than big ones
as power goes up	→	increase the chance of picking up a difference
as significance level goes down	→	decrease chance of saying there is a difference when there really is not

"difference" here means any effect size of measure of association

I.E. DRAWING CONCLUSIONS WITH MORE CERTAINTY REQUIRES MORE SUBJECTS

Sample size
Outcome: counting people (or objects)

% patients alive			Power		
Control	New	Difference	80%	85%	90%
50	60	10	776	886	1038
50	70	20	186	214	248
50	80	30	78	88	104
50	90	40	40	44	52

NB: sample size also depends on the rate in controls as well as the size of the expected difference

- A study of 78 patients is enough to detect a difference of 30% or more (at 80% power)
- But if the true difference is actually smaller, e.g. 10%, it is possible that a study with 78 patients would not produce a statistically significant result

Sample size
Outcome: taking measurements on people/objects

- When the outcome measure of the study involves taking measurements on people (or objects) we calculate:

$$\textbf{Standardized difference } (\Delta) = \frac{\text{Mean value Treatment B} - \text{Mean value Treatment A}}{\text{Standard deviation}}$$

	Power		
Δ	80%	85%	90%
0.1	3142	3594	4206
0.2	788	900	1054
0.3	352	402	470
0.4	200	228	266
0.5	128	146	172
1.0	34	38	46

Sample size
Outcome: taking measurements on people/objects

Outcome measure	Units of measure	Mean in Group A	Mean in Group B	Standard deviation	Standardized difference
Blood pressure	mmHg	90	85	6	**0.8**
Cholesterol	mmol/L	6	4.7	1.6	**0.8**

We have two different measurements (blood pressure and cholesterol) but the standardized difference between Group A and Group B is the same

The sample size would be the same

Sample size
Outcome: Time-to-event data

There are several different methods, depending on how you want to describe the effect size (power and significance level the same as before):

- Can specify the survival (event) rate in each group
- Can specify the median survival in each group, with length of recruitment time and length of follow up time
- Can specify one event rate (or median survival) and the hazard ratio

Sample size determination and effect size

- Sample size estimation is influenced by the effect size
- Sample size always involve making guesses about the effect size we are interested in (though people often over-estimate it)
- The effect size used to estimate sample size should be
  - realistic, e.g. from prior evidence (look at the literature)
  - clinically useful (talk to colleagues)
- Choose effect size first, then look at sample size - do not choose a sample size first, then say that the corresponding effect size is the one you are interested in
- The main statistical analysis should be based on the same endpoint and effect size specified in the sample size calculation (e.g. if the sample size is based on comparing the mean difference, the main analysis should be an unpaired t-test, not a comparison of the percentage above or below a given cut-off)

Sample size and other considerations

- Each bit of information used in estimating sample size can vary, but what you are interested in is whether the study needs to have 500 subjects instead of 100, rather than 500 subjects instead of 490
- In studies examining several factors at the same time**,** the sample size needs to be large enough to detect all the main effects of interest
- Consider time & cost to recruit subjects and obtain your outcome measures
- To examine associations (e.g. in regression), the more subjects the better, especially if you want to look at several factors simultaneously (multivariable analyses)
  - For counting people or time-to-event data, an approximate guide is to have ≥10 events for each factor you want to look at
- May have to inflate your sample size if participant dropout or failed lab analysis is anticipated (as you will have missing data for these subjects)

Sample size and other considerations

- If sample size calculation gives something too big, then:
  - Reconsider the effect size or other parameters (but don't choose unrealistic effect sizes)
  - Acknowledge small study and be aware of the problems that may arise
  - See how results fit in with other similar research
  - Consider calling your study a pilot or feasibility study (these are generally small, <50 people) and they are not powered to make direct comparisons, but rather to have a preliminary look

Result is not statistically significant

"*Result is not statistically significant*"

Possible reasons for this:

- There really is no difference, or the difference is smaller than expected
- There is a difference, but by chance we picked a sample that did not show this
- There is a difference but the study had **insufficient power** to detect it - the study was too small

Study design and sample size further information

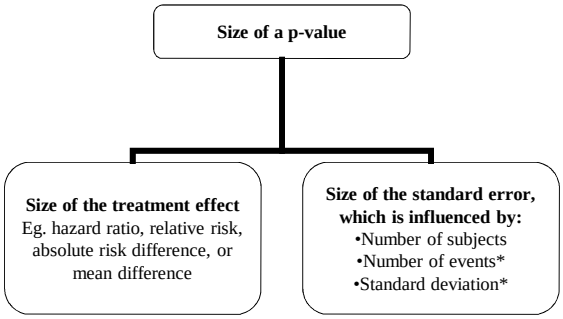
- *Case-control and cohort studies:*
- http://www.cdc.gov/EpiInfo/

- http://www.sph.emory.edu/~cdckms/sample%20size%202%20grps%20case%20control.html

- *All studies*
- Dupont WD and Plummer WD: PS power and sample size program available for free on the Internet. Controlled Clin Trials,1997;18:274. http://ps-power-and-sample-size-calculation.software.informer.com/

- Sample size tables for clinical studies. Machin et al. Wiley Blackwell 2009 (includes software on CD)

Sample size determination and effect size



The size of a p-value depends separately on the 2 items above

*The number of subjects is relevant to all studies, but number of events is more important when the trial endpoint is based on counting people/time-to-event, and standard deviation when endpoint is taking measurements on people

- **P-values get smaller** with large treatment effects, or small standard errors (seen with large studies, many events or small standard deviations).
- **P-values get larger** with small treatment effects, or large standard errors (seen with small studies, few events or large standard deviations).