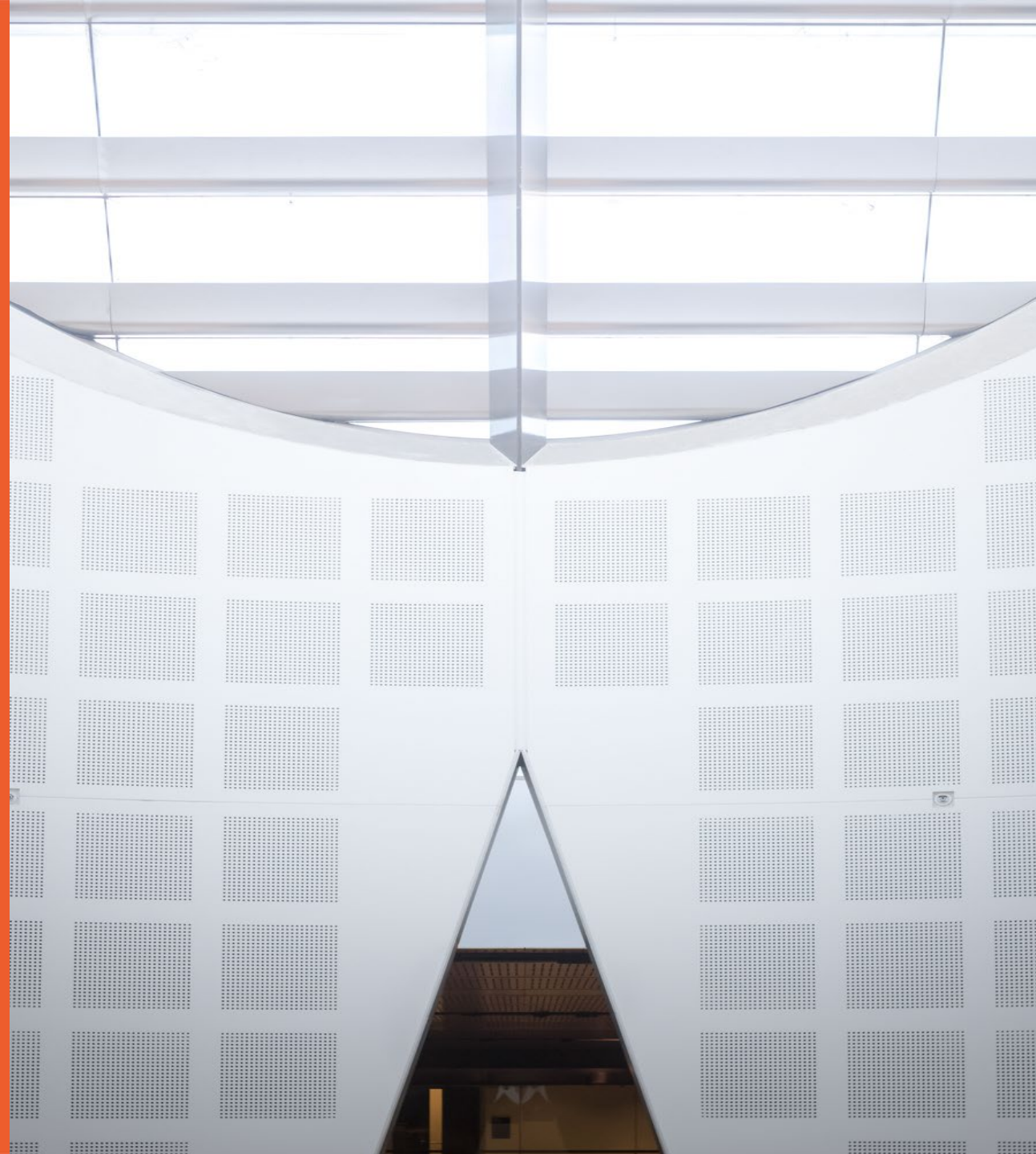
Experimental Design

Presented by
Alexandra Green
Statistical Consultant
Sydney Informatics Hub
Core Research Facilities
The University of Sydney

THE UNIVERSITY OF SYDNEY



Acknowledging SIH

All University of Sydney resources are available to Sydney researchers free of charge. The use of the SIH services including the Artemis HPC and associated support and training warrants acknowledgement in any publications, conference proceedings or posters describing work facilitated by these services.

The continued acknowledgment of the use of SIH facilities ensures the sustainability of our services.

Suggested wording for use of workshops and workflows:

"The authors acknowledge the Statistical workshops and workflows provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

What is a workflow?

- Every statistical analysis is different, but all follow similar paths. It can be useful to know what these paths are

- We have developed practical, step-by-step instructions that we call '*workflows*', that can you can follow and apply to your research

- We have a general research workflow that you can follow from hypothesis generation to publication

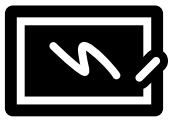
- And statistical workflows that focus on each major step along the way (e.g. experimental design, power calculation, model building, analysis using linear models/survival/multivariate/survey methods)

Statistical Workflows

- Our statistical workflows can be found within our workshop slides
- Statistical workflows are software agnostic, in that they can be applied using any statistical software
- There may also be accompanying software workflows that show you how to perform the statistical workflow using particular software packages (e.g. R or SPSS). We won't be going through these in detail during the workshop. If you are having trouble using them, we suggest you attend our monthly Hacky Hour where SIH staff can help you.

During the workshop

Ask short questions or clarifications during the workshop (either by Zoom chat or verbally). There will be breaks during the workshop for longer questions.

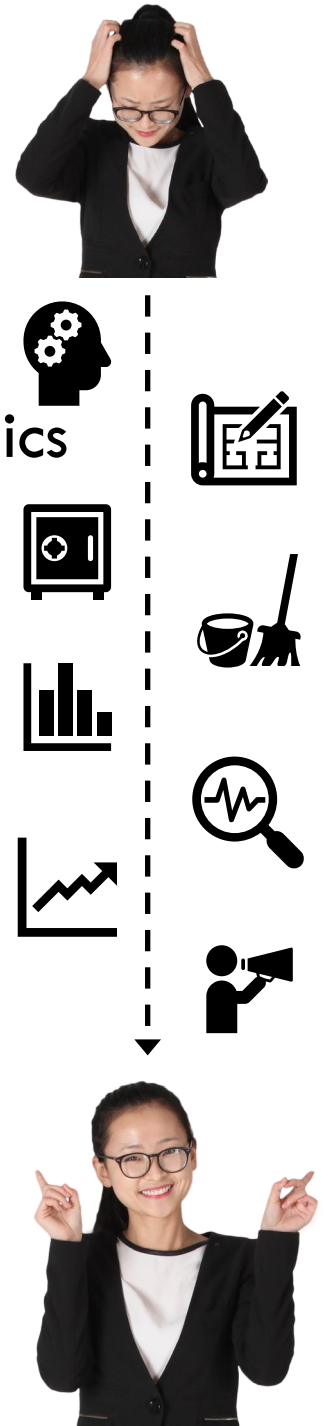


Slides with this blackboard icon are mainly for your reference, and the material will not be discussed during the workshop.

Challenge questions will be encountered throughout the workshop.

General Research Workflow

1. **Hypothesis Generation** (Research/Desktop Review)
2. **Experimental and Analytical Design** (sampling, power, ethics approval)
3. **Collect/Store Data**
4. **Data cleaning**
5. **Exploratory Data Analysis (EDA)**
6. **Data Analysis aka inferential analysis**
7. **Predictive modelling**
8. **Publication**



This workshop

I. The big picture

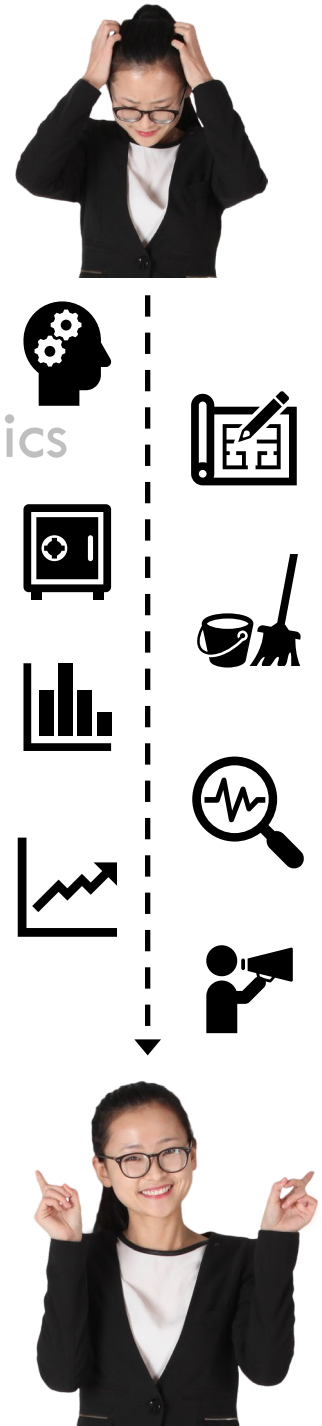
II. A workflow for experimental design:
   I. Step 1: Defining your sample
   II. Step 2: Considering analysis
   III. Step 3: Designing your study

The big picture

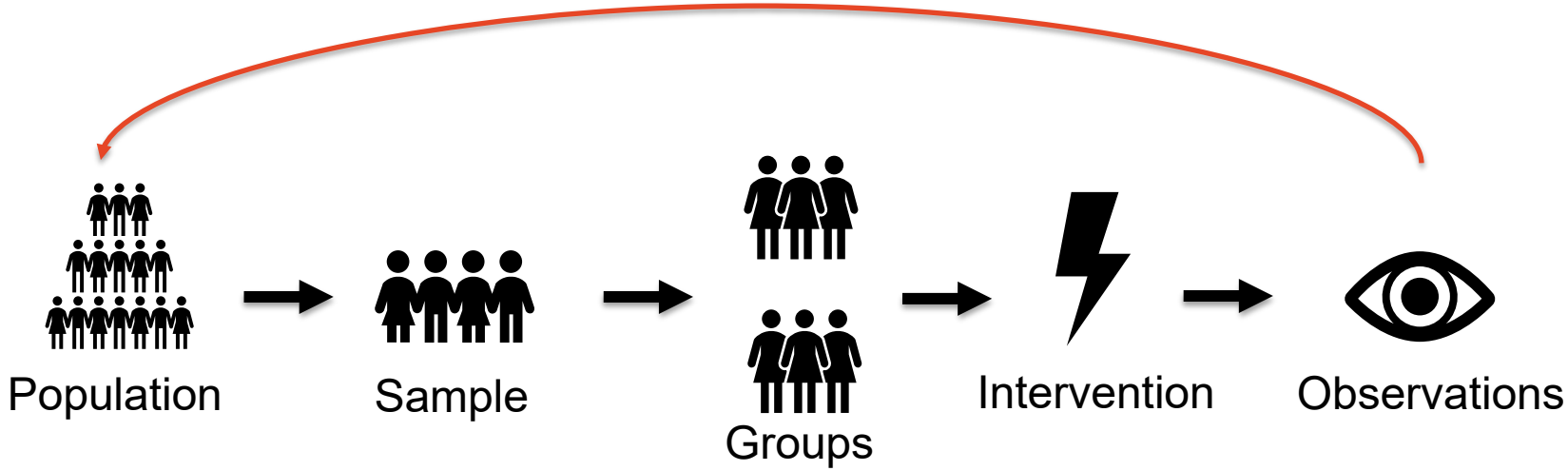
Why do we care about experimental design?

General Research Workflow

1. **Hypothesis Generation** (Research/Desktop Review)
2. **Experimental and Analytical Design** (sampling, power, ethics approval)
3. **Collect/Store Data**
4. **Data cleaning**
5. **Exploratory Data Analysis (EDA)**
6. **Data Analysis aka inferential analysis**
7. **Predictive modelling**
8. **Publication**



Running a study



We want to find out (infer) something about the members of our population so we run a study.

Despite the labour involved, or perhaps because of it, many do not spend enough time thinking about designing a study so that it has the best chance of finding something out (valid statistical inference – the red arrow above)

Experimental design is about being aware of the challenges to statistical inference, and using the best tools available to overcome them

What is the challenge of statistical inference?

- Ronald Reagan loved jellybeans

- If we wanted to estimate the proportion of red jellybeans in this jar, we could count them all, or take a **sample**

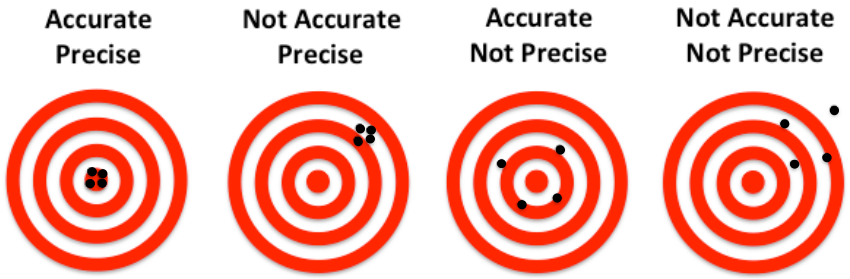
- Taking a sample, and using it to make **estimates** and draw **inferences** about a **population** (in this case all of the jelly beans in the jar) is the basis of statistics.

Estimates are inaccurate and imprecise… but by how much?

- When we take a sample, and run a study there are two 'forces' that make our estimates differ from the true value in the population: **bias** (accuracy) and **error** (precision)

- Think of making an estimate like throwing a dart. There is some true value in the population (bullseye) that you are trying to hit. You usually only make one estimate per study (throw one dart), but imagine for a moment you make more.

- Experimental design is all about controlling **bias** and minimising **error** (variability) in order to estimate the true value as closely as possible



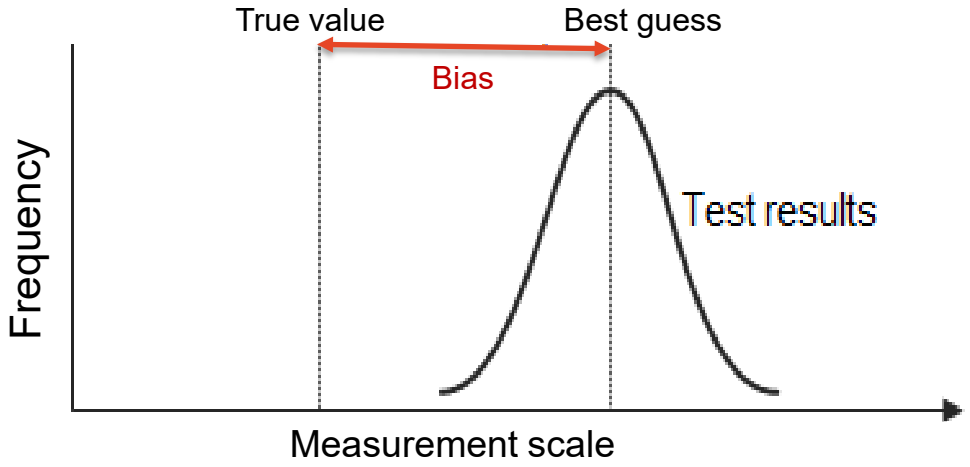
Just like throwing a dart, using our experimental sample to estimate some population statistic is subject to randomness. Unlike throwing a dart – we don't know where the bullseye is!

Why are our estimates biased?

- Let's start by thinking about bias in a statistical way: it is *systematically* over- or under-estimating our population property

- Bias may be introduced in the way we **sample, conduct our experiment, measure,** or even **analyse** our data.

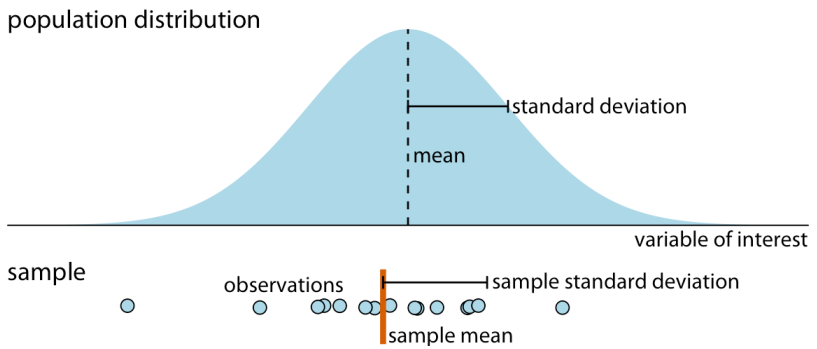
- Some level of bias is inevitable, but too much bias can be fatal: we would reach the wrong conclusion *most of the time* in a poorly designed experiment!

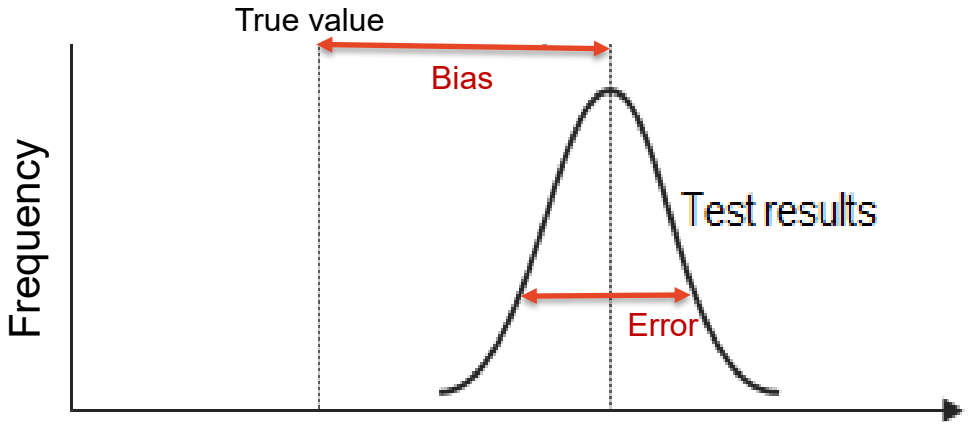


If we ran our experiment many times, and plotted the frequency of our estimates, our best guess of the true value would be the mean of all of our estimates. The distance from this best guess to the true value is our bias. No matter how many times we ran the experiment, or how large our sample was, we would end up with this level of bias.

Why are our estimates imprecise?

- 'Error' in statistical terms is about how close together our estimates are – it's 'error' in the sense of measurement, not making a mistake
- Error exists because the individuals in our population vary. Each sample is different, and so our estimates will vary from study to study*
- Again, error is influenced by the way we sample, measure, conduct or analyse our experiment
- We can increase the precision of our estimates by having larger samples. Choosing an appropriate sample size is a vital and non-trivial part of controlling error in experimental design: see our Power and Sample Size Calculation workshop for more





It's important to realise that although variability in the population affects the variability of your estimates –there's another important ingredient: your sample size. We use variation in our sample population to quantify the uncertainty in our estimates.

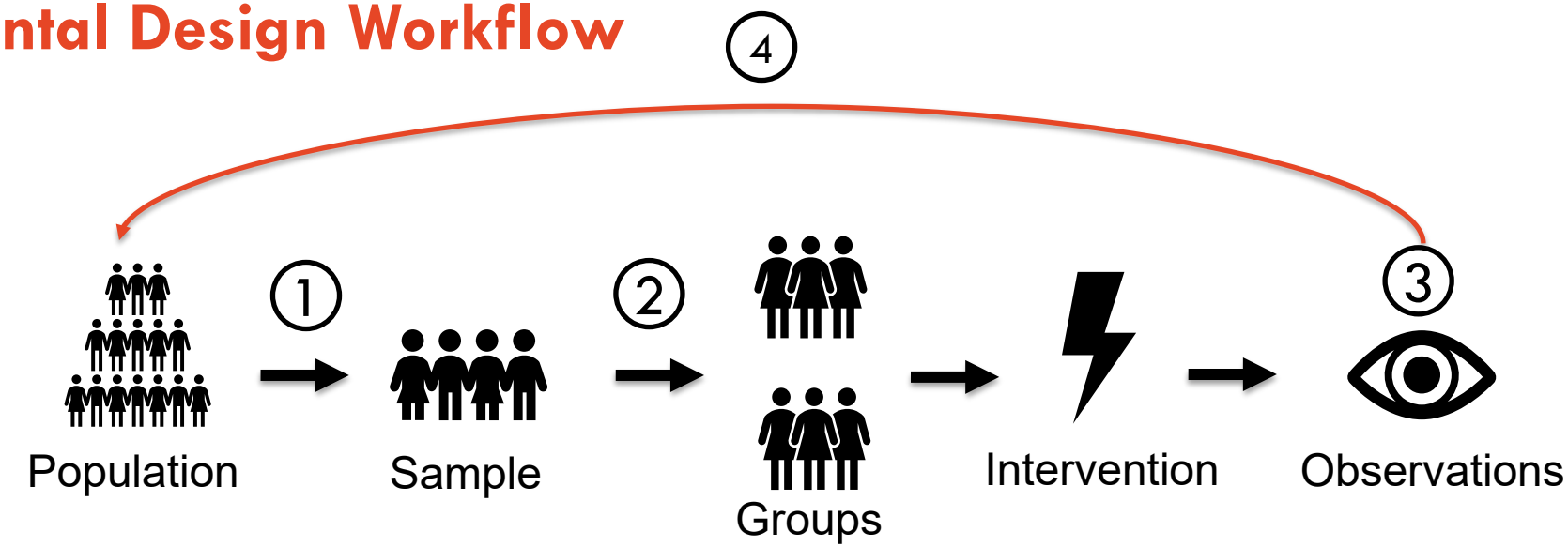
* This is called 'sampling error', but there is also 'non-sampling' error which covers everything else that causes variation between estimates, e.g. mistakes when collecting data

A quick challenge question

- Would you be *most* concerned about bias, error, or both equally in the following studies?
  - A diet study in which respondents were asked to recall their food intake for the previous week every Sunday evening
  - An urban rat behaviour study that used wild rats caught in traps distributed throughout the city
  - A survey on the number of motor vehicle crashes experienced over your life time where you, the respondent was at fault
  - As above, where either party were at fault

A workflow for experimental design

Experimental Design Workflow



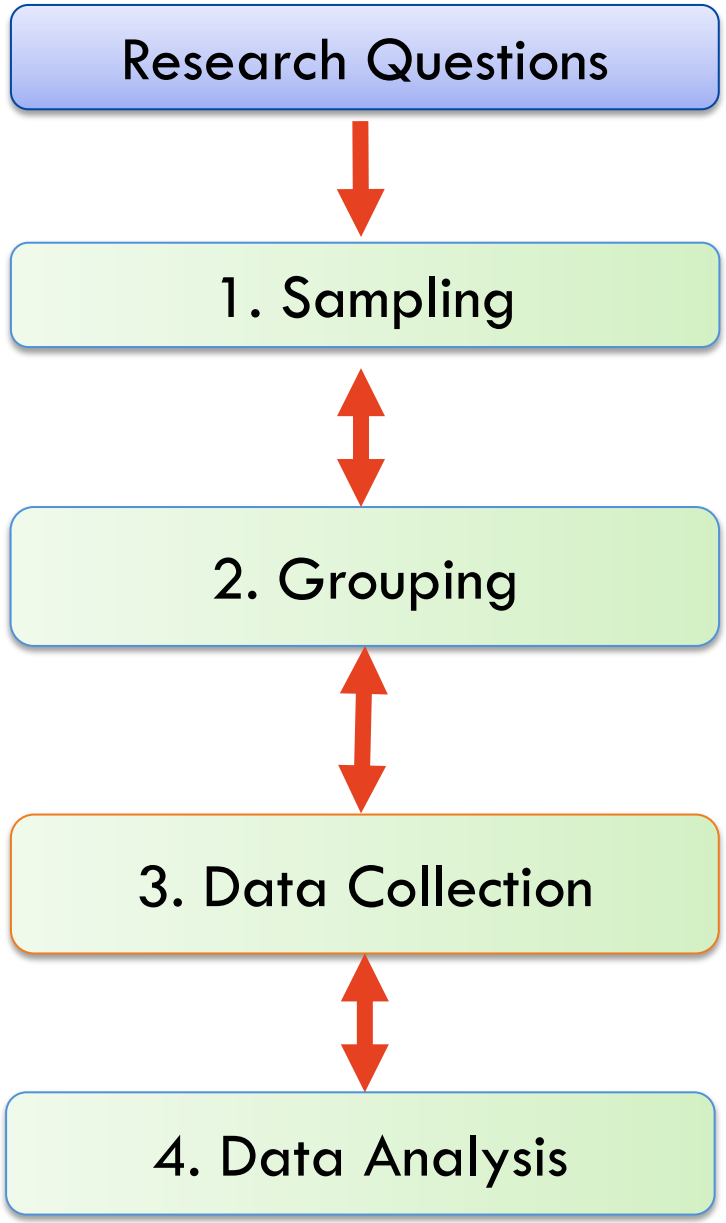
- We need to consider the steps of experimental design in the context of how a study is performed
- The process of experimental design itself should be iterative, so after you have considered the following, cycle back through these steps until you are convinced the design is the best it can be

1. Sampling
2. Grouping
3. Data collection
4. Analysis

Experimental Design Workflow

0) Finalise your research questions and hypotheses then choose your study type

1. Sampling
    - Define your sampling method
    - Consider representativeness
    - Identify your units
    - Consider replication
2. Grouping
    - Choose your blocks or strata
    - Perform randomisation for treatment allocation
3. Data Collection
    - Consider blinding
4. Data Analysis/Inferential Analysis
    - Identify your method



Experimental Design Worksheet

- After attending this workshop (and potentially other stats consulting workshops) you should be able to fill in all these boxes as they apply to your research
- You can bring this information to a statistical consult and/or share it with your research team

Biological Units	
Experimental Units	
Observational Units	

Outcome Variables	Explanatory Variables	Design features	Blocking Factors	Randomisation & blinding	Analysis Method

Step 0) Finalise your research question and hypotheses

What are your research questions? Having a very clear idea of what your research questions is vital for achieving good experimental design.

Avoid the urge to rush into collecting data and just "worry about the analysis later."

You can only get meaningful results from a well-controlled and adequately-powered experiment. The type of analyses possible are also dictated by your experimental design.

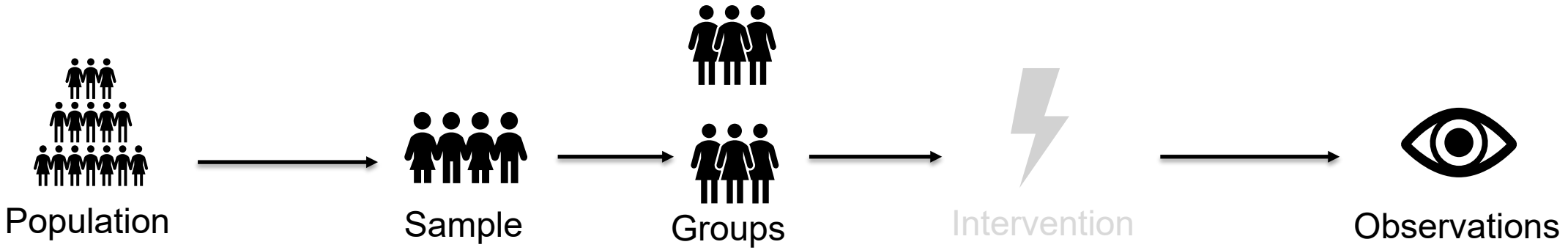
Think of this paradox: *often the sooner you start your experiments, the longer your project will take*. Invest the time in good design.

Step 0) Choose your study type

Study type

- What is the best approach to answer your research question?
- What approaches are possible for you?
- Many approaches are available. Primary research can be divided into observational studies and designed experiments
- We will discuss some widely-used study types

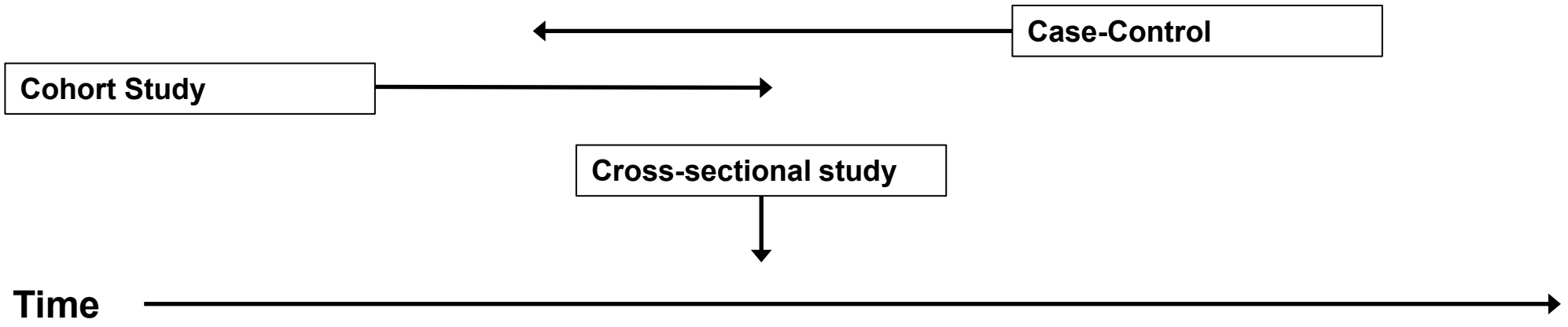
Observational Studies



Observational studies do not involve any experimental intervention

Adapted from: https://www.psychdb.com/teaching/2-intro-critical-appraisal#study-design

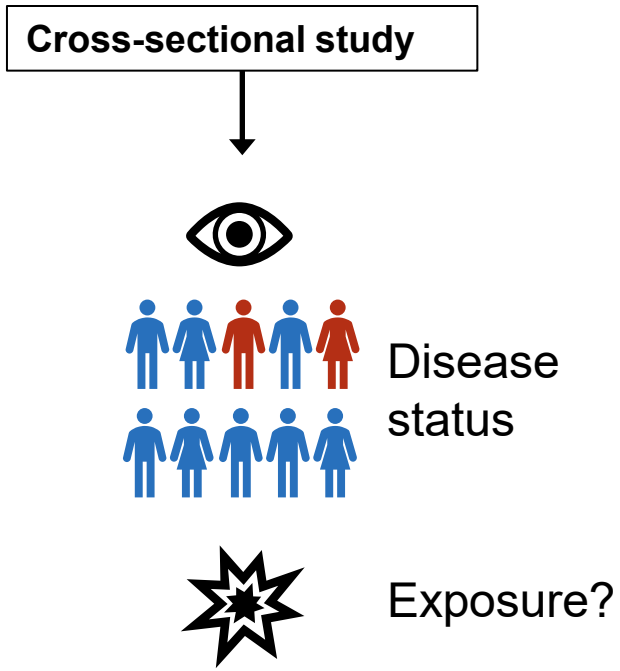
Observational Studies



The type of observational study you will perform depends on when you are collecting your observations, and how you will be sampling from the population.

Let's go through these and examine how they attempt to find an association between 'exposure' to a risk factor, and the development of disease. These classic epidemiology study designs arose from the human health context, but are applicable to any field.

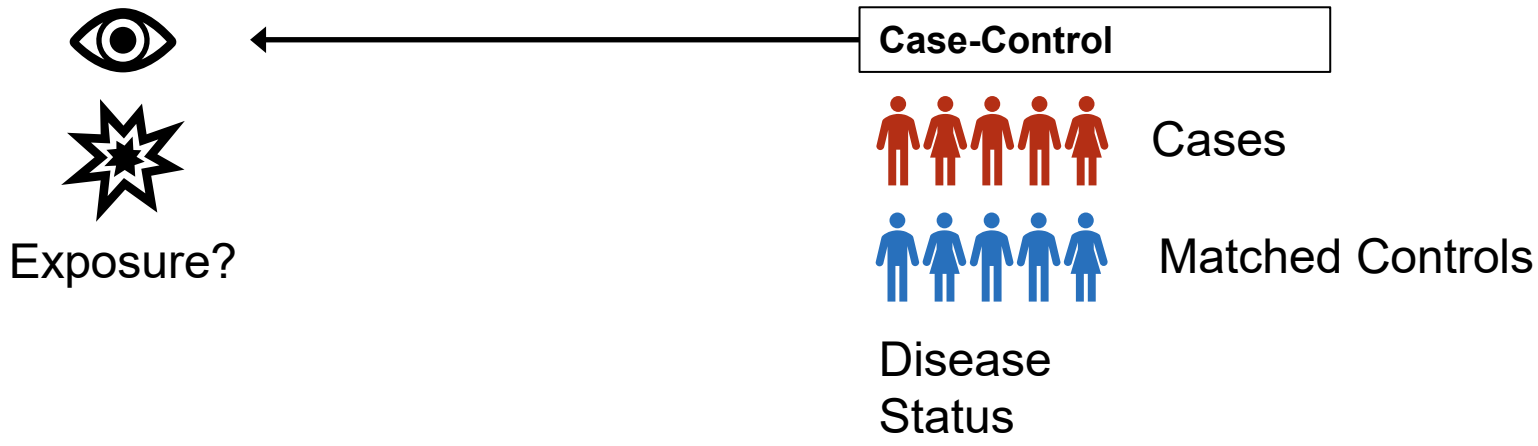
Observational Studies: Cross-sectional



- Look at a single time-point
- Relatively quick and easy to run
- Cannot establish any temporal order of exposure and disease status, only the prevalence of disease in the population at a particular point in time

Adapted from: https://www.psychdb.com/teaching/2-intro-critical-appraisal#study-design

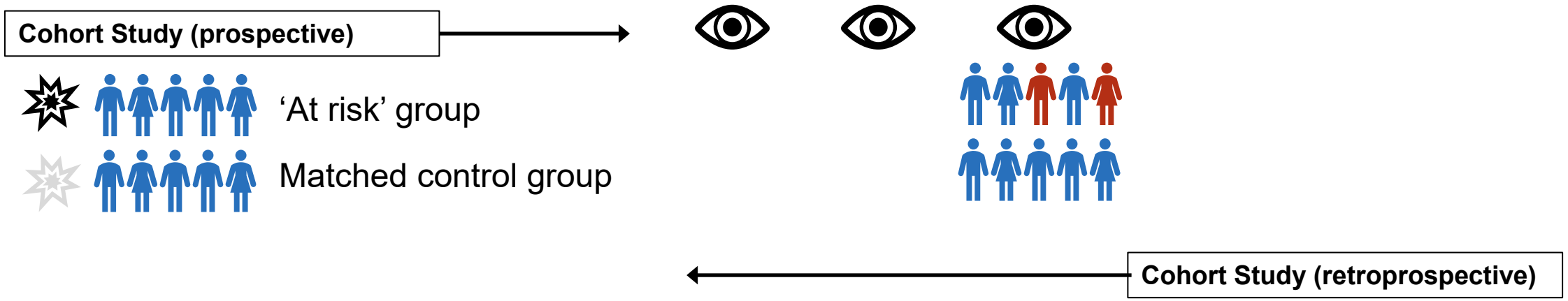
Observational Studies: Case-Control



- Assemble cases and controls matched on several factors, similar except for disease status
- Look at exposure to some potential risk factor to examine association with disease
- Good for studying rare disease with relatively common exposures
- Sampling: Enrich for cases

Adapted from: https://www.psychdb.com/teaching/2-intro-critical-appraisal#study-design

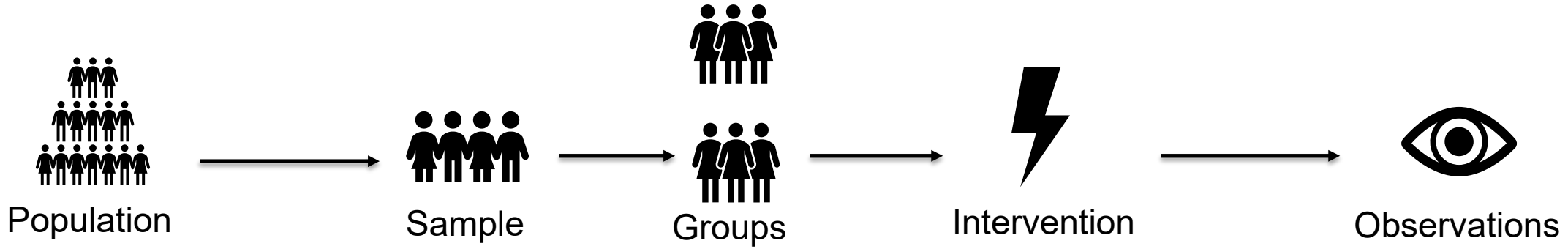
Observational Studies: Cohort Study



- Assemble groups often based on exposure to some risk factor for developing disease and follow up over time to measure development of disease in the cohort and often other outcomes
- Can establish temporal relationship between exposure and development of disease
- Often more expensive and logistically challenging (loss to follow up, longitudinal observation)
- Good for studying rare exposures, and relatively common diseases
- Can be done prospectively (enrolment prior to any disease development) or retrospectively (using existing cohorts that were followed through time)
- Sampling: Enrich for 'at risk' subjects

Adapted from: https://www.psychdb.com/teaching/2-intro-critical-appraisal#study-design

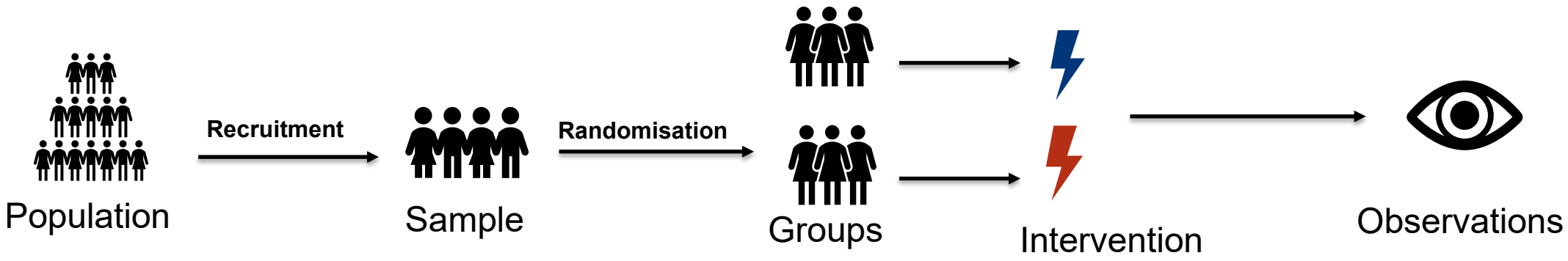
Designed experiments



Designed experiments do involve an experimental intervention. Other factors in the experiment are more likely to be designed and controlled (e.g. taking place in controlled conditions inside a lab or a clinic).

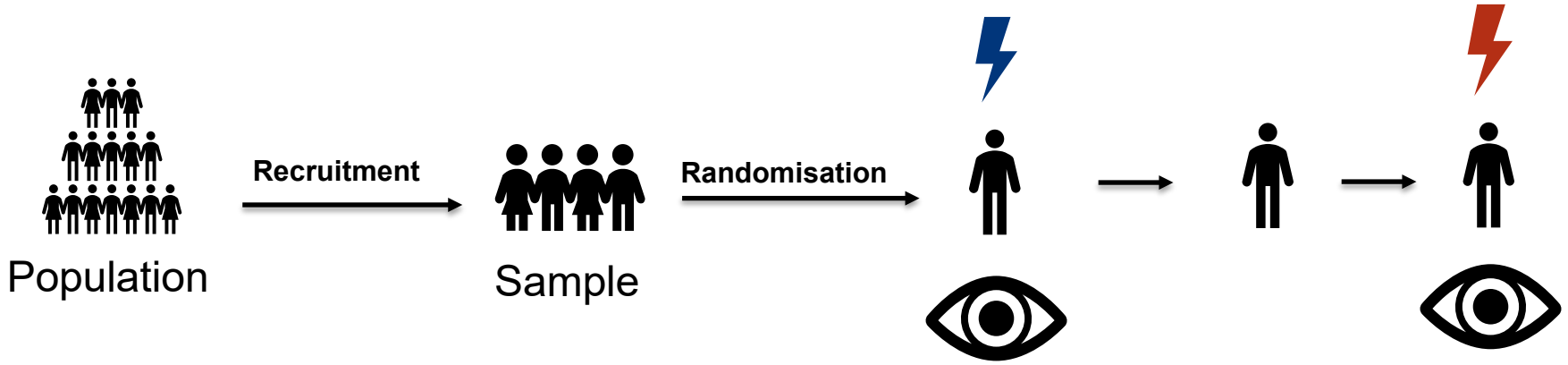
Adapted from: https://www.psychdb.com/teaching/2-intro-critical-appraisal#study-design

Designed experiments: Randomised Control Trial



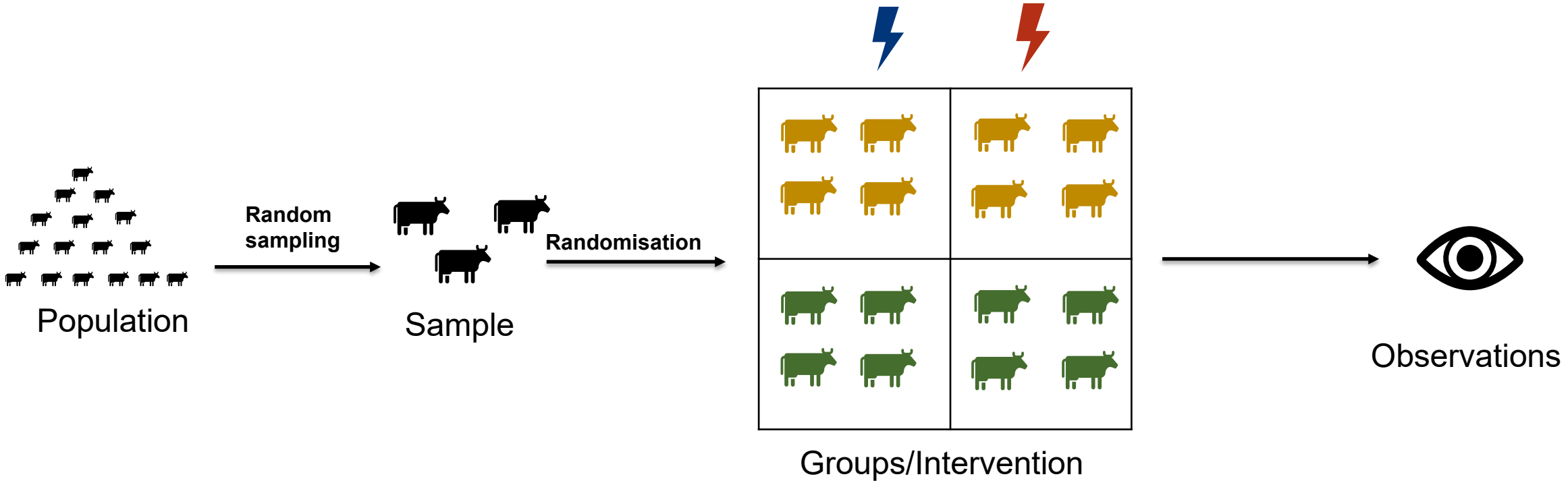
- Considered by many to be the gold standard for a real-world intervention (drug trials, educational interventions, etc)
- Not always possible or ethical to randomise to different treatments and hence perform an RCT
- Sometimes interim analysis performed to determine whether trial should be stopped because of clear demonstration of benefit (and placebo group should receive treatment too)

Designed experiments: Crossover trial



- Crossover trials are a 'within-subject' version of a randomised control trial
- Can be used in contexts where the treatment effect can be 'washed out' in the subjects between different treatments
- More efficient than an RCT (more power for the same number of subjects)
- Not always possible e.g. if the treatment effect is not transient

Designed experiments: Factorial study



- Common in agricultural/engineering disciplines where subject factors can be easily controlled (e.g. easy to source a male calf of a particular breed)
- Also used extensively in psychology in 'within subject' designs
- Most efficient way to estimate the effects of multiple factors, alone or in combination

Observational vs. Designed Experiments

- In studying some phenomena in the real world we are usually interested in **causality.** E.g. In health research we may aim to find modifiable factors that can improve health outcomes.

- **A causal** relationship between refers to changes in an explanatory variable causing a change in an outcome variable (e.g. high protein diet causing weight gain)

Observational vs. Designed Experiments

- In **designed experiments,** one or more explanatory variables are manipulated by the experimenter. This allows you to evaluate a causal hypothesis (e.g. "diet affects weight gain").

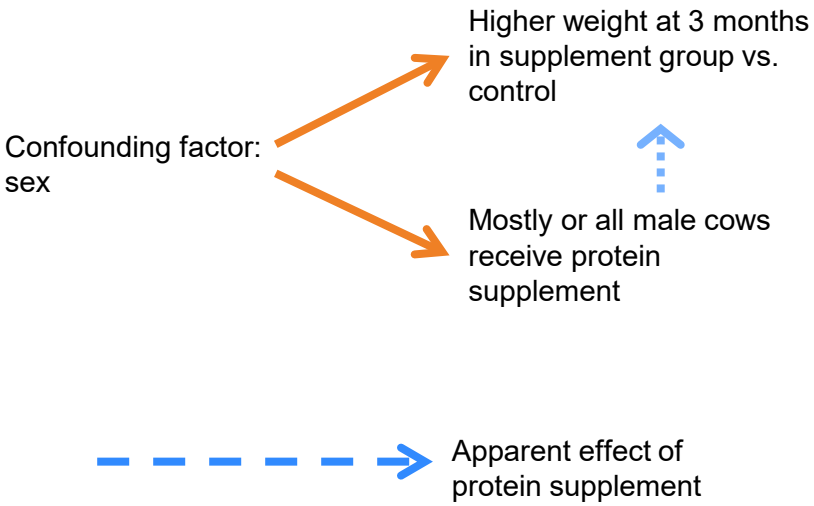
- In an **observational study,** you cannot demonstrate causality.

- Various modelling techniques can be used with observational study data to examine causal hypotheses (see our model building workshop). They are used to examine whether observational study data are consistent with hypothesised causal pathways. This is not the same thing as demonstrating causality.

- **Designed experiments** have much stronger causal power than observational studies. There may be strong ethical, or practical reasons stopping you from performing a designed experiment to answer your research question.

Observational vs. designed experiments: Confounding variables

- **Internal validity** is the validity of the experiment for the sample chosen, including proper control of confounding variables
- A *confounding factor* is some variable that may partly or fully explain the apparent (or "uncontrolled effect") of an explanatory variable of interest
- We control for confounders in **designed experiments** by using randomisation and potentially by including other explanatory variables as covariates in our analysis
- In an **observational study**, we don't have experimental intervention, so we instead design our study to sample participants with similar characteristics (e.g. matched pairs), and/or attempt to adjust for differences in characteristics analysis
- When confounding is so strong that the effects cannot be disentangled (e.g. *all* male cows receive protein supplement) the study is referred to as **confounded**



Other lines of evidence to examine causality

Bradford Hill Criteria

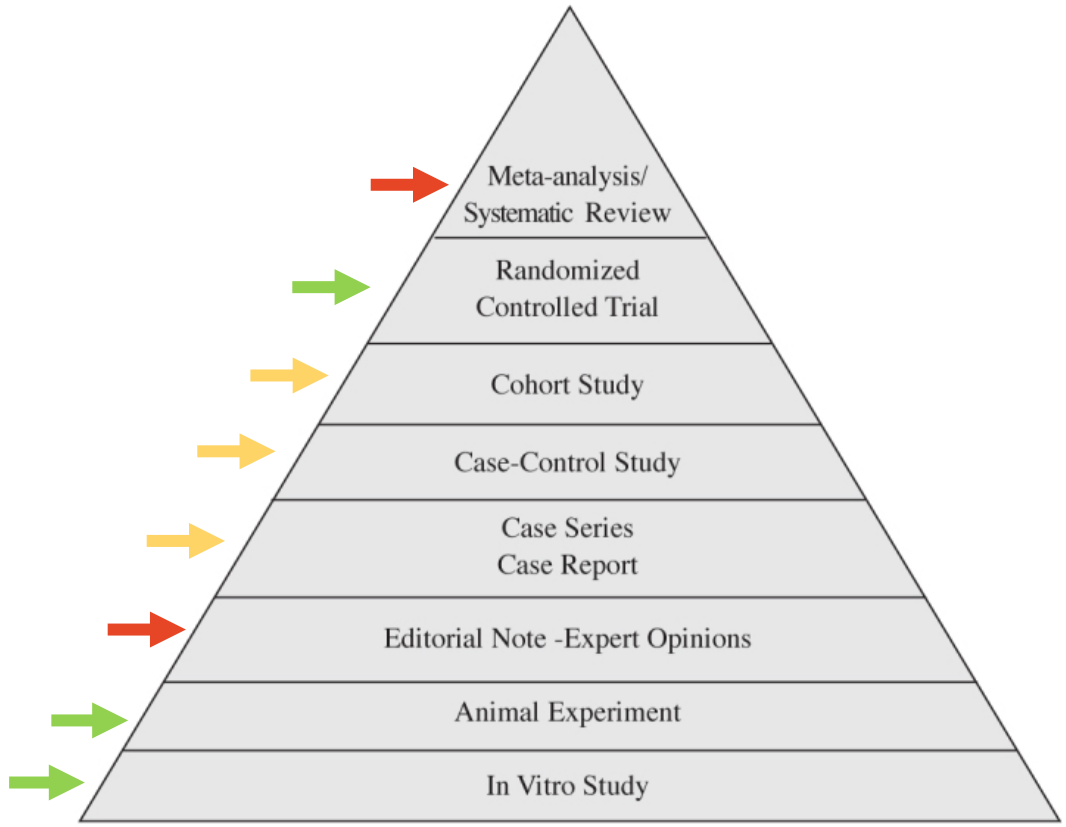
1. **Strength** (effect size): A small association does not mean that there is not a causal effect, though the larger the association, the more likely that it is causal.
2. **Consistency** (reproducibility): Consistent findings observed by different persons in different places with different samples strengthens the likelihood of an effect.
3. **Specificity**: Causation is likely if there is a very specific population at a specific site and disease with no other likely explanation. The more specific an association between a factor and an effect is, the bigger the probability of a causal relationship.
4. **Temporality**: The effect has to occur after the cause (and if there is an expected delay between the cause and expected effect, then the effect must occur after that delay).
5. **Biological gradient**: Greater exposure should generally lead to greater incidence of the effect. However, in some cases, the mere presence of the factor can trigger the effect. In other cases, an inverse proportion is observed: greater exposure leads to lower incidence.
6. **Plausibility**: A plausible mechanism between cause and effect is helpful (but Hill noted that knowledge of the mechanism is limited by current knowledge).
7. **Coherence**: Coherence between epidemiological and laboratory findings increases the likelihood of an effect. However, Hill noted that "... lack of such [laboratory] evidence cannot nullify the epidemiological effect on associations".
8. **Experiment**: "Occasionally it is possible to appeal to experimental evidence".
9. **Analogy**: The effect of similar factors may be considered.

Choose your study type

Is there a best study type?

- Consider this evidence pyramid
- Some will involve be designed experiments and others are observational, while others involve synthesis of primary literature
- The optimal study type will depend on what is known and the feasibility of each study type in your chosen field

Evidence pyramid for medical studies

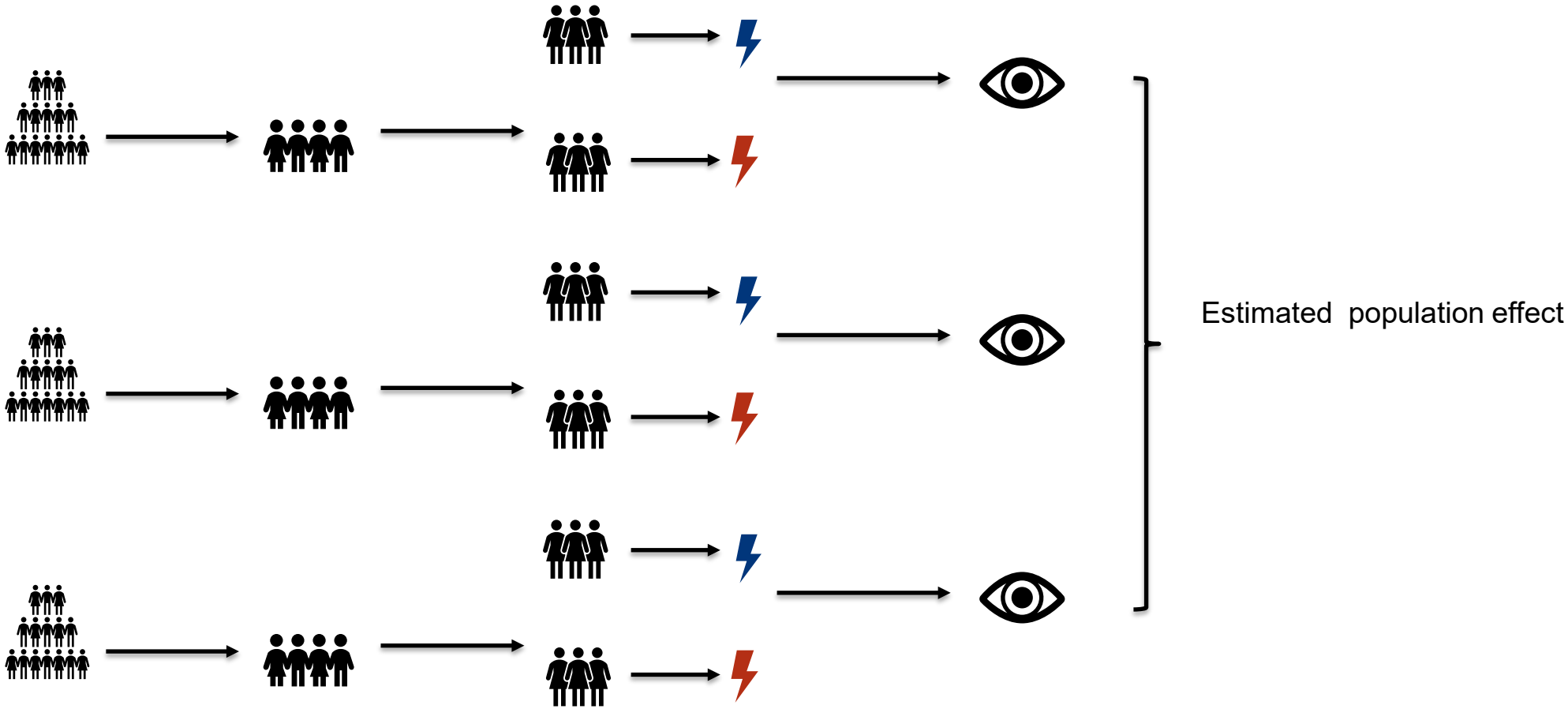


Study designs in medicine
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4318396/

The evidence pyramid and introduction to randomized controlled trials. *Pandis N Am J Orthod Dentofacial Orthop. 2011 Sep; 140(3):446-7.*

Synthesis of primary literature: systematic review or meta analysis

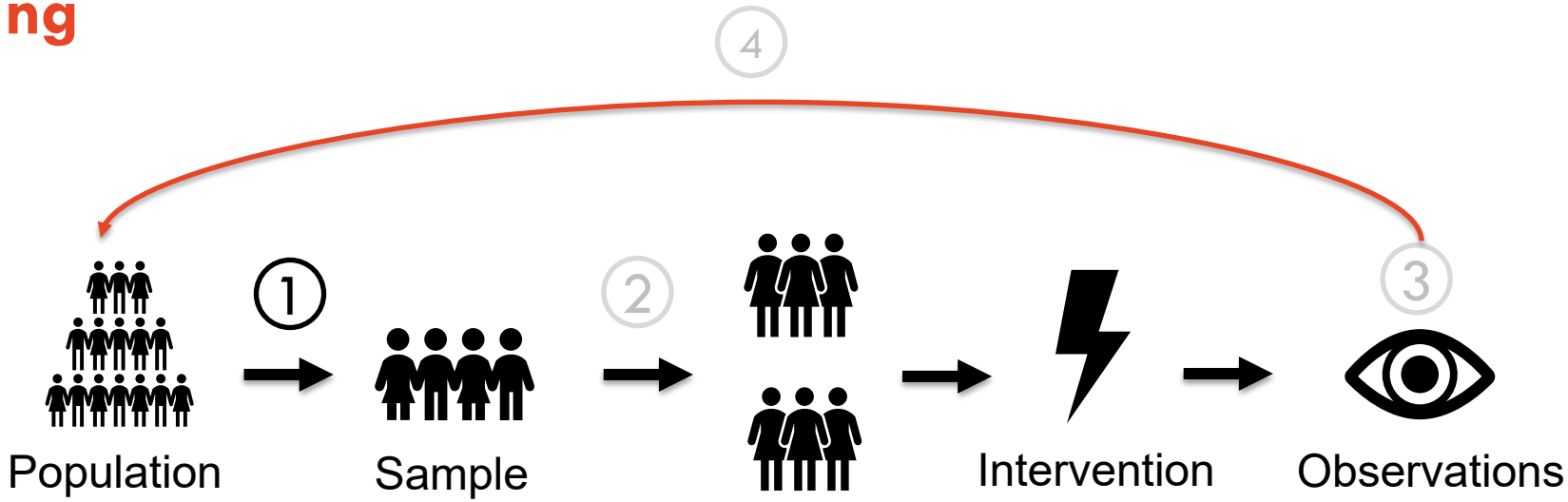
- Meta-analysis combines the estimates from multiple primary studies*



* See our meta-analysis workshop for more detail

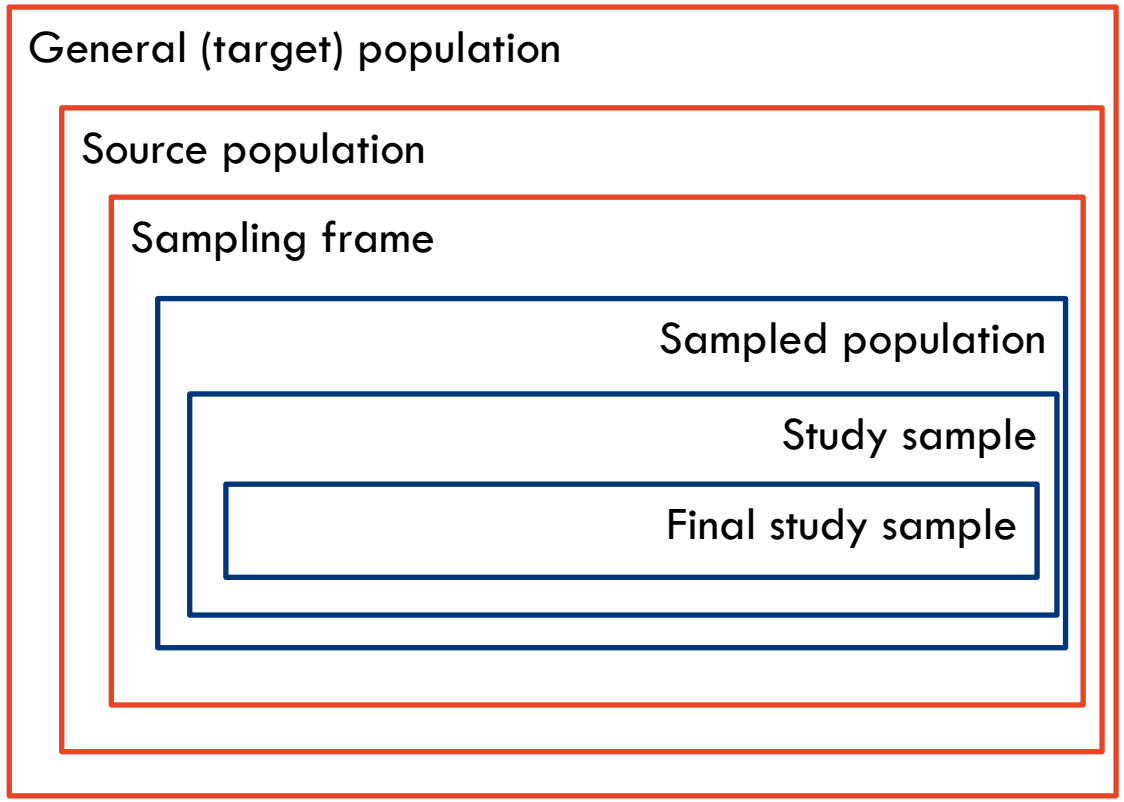
Step 1) Sampling

1. Sampling



- In estimating some population property, we need to consider 'external validity'
- External validity is how generalisable the study is to a wider population and depends on the size and representativeness of the sample used
- It is the [final] sample on which we take observations that we are using to infer something about our population

Generalisability to your target population



- Your **target population** is the population about which you would like to make inferences
- From a subset of the target population called the **source population,** you define your sampling frame
- The **sampling frame** is a list of individuals from which you draw your sample, you may exclude some individuals if you find they aren't part of your target population
- You conduct your study and after excluding some participants to obtain your **study sample,** you run your study. You end up with complete data for all individuals in a **final study sample**

Would the estimates obtained from each of these sets of individuals be the same (or similar enough) as the estimates for the larger box? Is the smaller box representative of the target population?

Strata

- We may be able to group subjects who are similar to each other into **strata**
- An example of strata is the age group of your subjects
- Defining strata allows us to adjust our sampling to ensure representativeness of our sample to the target population (e.g. we could sample to match the distribution of the population into age groups)
- In a designed experiment, the allocation of treatments should be balanced in the different strata

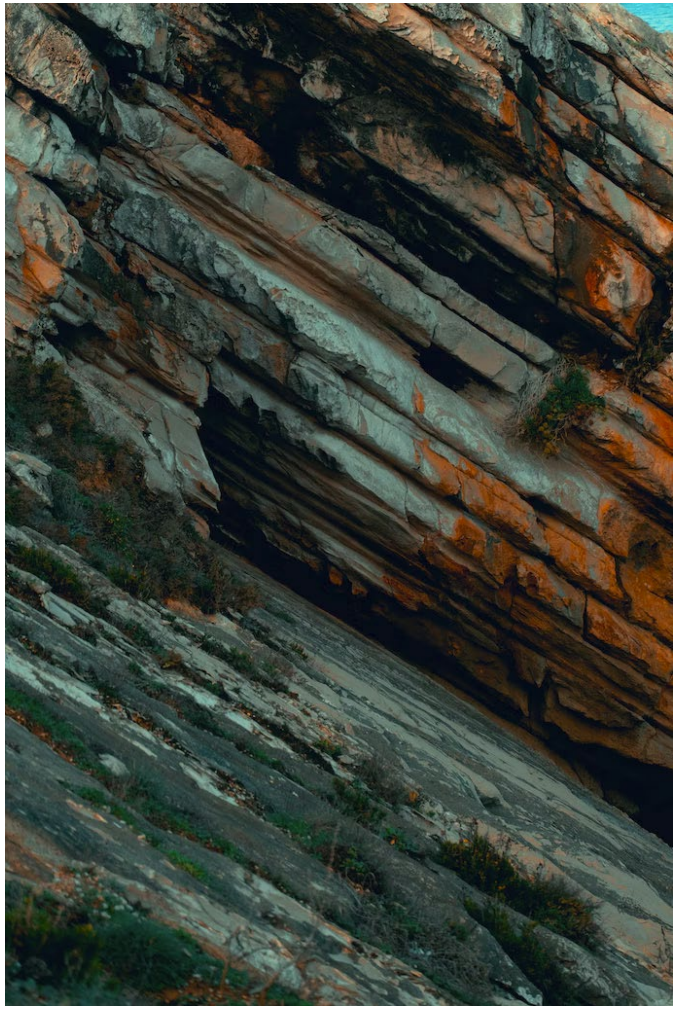


Photo by Bea Vallejo on Unsplash

Sampling methods: HILDA example

- The sampling method used for the Household, Income and Labour Dynamics in Australia (HILDA) survey. This is a panel survey, meaning that data is collected on the same panel of individuals, in this case in annual 'waves'

Target population: All people living in households in Australia

Sampling frame: "the **reference population for Wave 1 of the HILDA Survey was all members of private dwellings in Australia,** with the main exception being the exclusion of people living in remote and sparsely populated areas."

Sampling method: "Households were selected using a multi-staged approach designed to ensure representativeness of the reference population. First, **a stratified random sample of 488 1996 Census Collection Districts (CDs)**, each of which contains approximately 200 to 250 households, **was selected from across Australia.** Within each of these areas, depending on the expected response and occupancy rates of the area, **a random sample of 22 to 34 dwellings was selected**…. Nonetheless, despite the region-based stratification, Wave 1 of the HILDA Survey was an equal-probability sample; in particular, the smaller states and territories were not over-sampled. This reflects the focus of the HILDA Survey on producing nationwide population estimates."

Source: HILDA Statistical report

Sampling methods: HILDA example

- Sample (initial):

"Of the 11,693 households selected for inclusion in the sample in 2001, 7,682 households agreed to participate, resulting in a household response rate of 66%. The 19,914 residents of those households form the basis of the 'main sample' that is interviewed in each subsequent year (or survey wave), but with interviews only conducted with people aged 15 years or older"

- Sample (subsequent):

"Table A1 presents the number of households, respondents and children under 15 years of age in each wave. In Wave 19, interviews were obtained with a total of 17,462 people, of which 13,748 were from the original sample and 3,714 were from the top-up sample. Of the original 13,969 respondents in 2001, 7,142, or 60.7% of those still in scope (that is, alive and in Australia), were still participating at Wave 19."

Source: HILDA Statistical report

Sampling methods: some other examples

A study to measure blood glucose concentrations in people with diabetes

- Your target population: people with diabetes in Australia
- Your source population: people presenting to a hospital in NSW with diabetes
- Your sampling frame: the list of 523 people who presented to the hospital in a 3 month period
- Your sample: The first 200 people from your sampling frame who consented to participating in your study
- Your final sample: 184 people with complete data

A study to evaluate the effectiveness

- Your target population: Gamers in Australia who play more than 10 hours a week of video games
- Your source population: 50 000 gamers in Australia who were presented with a Facebook recruitment ad
- Your sampling frame: the first 5000 people who volunteered for the study and met the inclusion criteria
- Your sample: 200 randomly selected from the sampling frame who were selected
- Your final sample: 40 people who showed up at your clinic and completed a gaming addiction treatment program

External Experimental validity: Example

Study to evaluate the effect of a feed supplement on the growth of calves

Study Design:

- 2 groups: 1: Standard feed and 2: Standard feed with supplement
- All calves are the same breed - Charolais
- All based on one cattle farm (sampling frame: a list of all cattle IDs)

What larger source population does this sample represent?

What general (target) population might we wish to make inferences about?

External Experimental validity: Example

Study to evaluate the effect of a feed supplement on the growth of calves

Conclusions will be valid for Charolais calves raised under local conditions.

Findings of the study may not be valid for all calves:

- Other breeds
- Grass fed?
- Other

External Experimental validity: Example

Study to evaluate the effect of a feed supplement on the growth of calves

An expanded study could now include:

- Sex: male, female
- Feed type: grass and grain
- Breed: Charolais, Hereford
- Climate: temperate, arid

This will expand the external validity of the study to cover a much wider population, but make the study potentially much more difficult to carry out.

Compromise is often necessary. The tension in experimental design is often between what you would *like* to find out and what you feasibly *can* find out in a single experiment.

Generalisability vs. bias

- You may argue that Charolais calves are representative of all calves in a wider population of interest for your outcome of interest (serves as an appropriate model for all cattle breeds)

- The danger is when you wish to make inferences within a wider population, without adequately sampling that population or being reasonably sure that the measured individuals in your population are representative

- These issues are where your domain expertise can help you optimise your experimental design

Generalisability and controlled conditions

- You may plan to perform an experiment under laboratory conditions, perhaps using a model organism or in cells grown *in vitro*

- The question of generalisability always arises when experiments are not performed under 'real world' conditions. Is bias introduced by performing the experiments in a lab compared to the real world?
  - The various forms of validity to assess model organisms for human disease

- On the other hand, controlling conditions often results in far less variability than in the real world (e.g. inbred mouse strains vs. outbred mice). This may be necessary to study subtle effects. We effectively trade off less error for potentially more bias.

- In practice, a combination of lab-based and real-world experiments are often necessary to fully characterise some biological phenomenon

A quick challenge question

- Can you make conclusions regarding causality in the following studies?
  - A randomised control trial where subjects received either a novel drug, or a placebo and their symptoms were examined after 3 weeks
  - An observational (cross-sectional) study examining at the rate of disease in subjects of age 60 who routinely took aspirin (for any reason), vs. those that didn't
  - An observational cohort study examining developmental delay in children. Recruited children of the same age and following them over a 10 year period. Comparing outcomes in those who attended pre-school and those that didn't

1. The units in your experiment

So what are the units in your experiment? Understanding the types of units will help you recognise design and analysis considerations

Types of Units (adapted from Lazic, 2016)

- Sampling Unit/biological unit – is the entity (animal/plant/thing) about which inferences are made

- Experimental Unit – is the entity that is randomly and independently assigned to experimental conditions or treatments

- Observational Unit – is the entity on which measurements are taken

Lazic, Stanley E. *Experimental Design for Laboratory Biologists : Maximising Information and Improving Reproducibility* . Cambridge, United Kingdom: Cambridge University Press, 2016. Print.

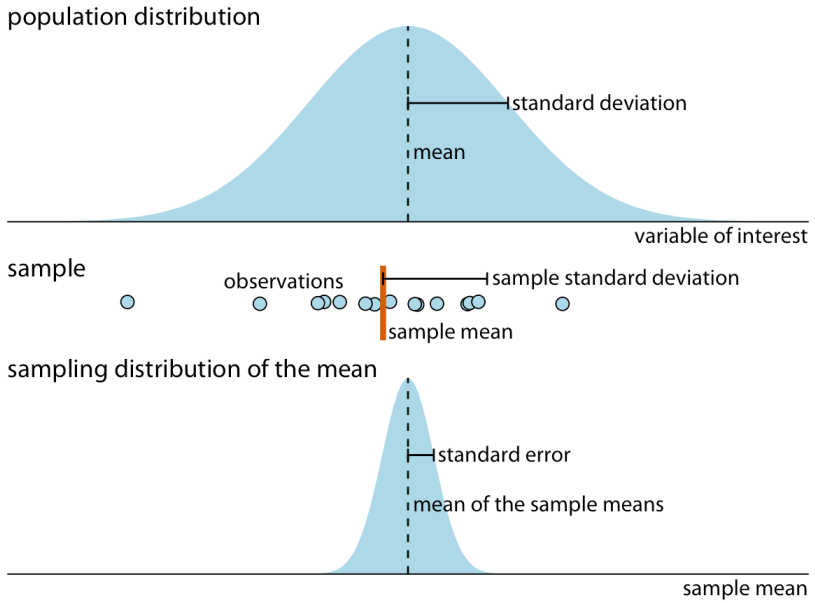
The units in your experiment

Why do these different units matter so much?

One of the most important assumptions in many statistical methods is statistical independence: simply stated, making one measurement gives you no information about another measurement

If this assumption is violated by taking repeated measurements on the same individual, or measurements on individuals that are clustered together (e.g. in a family, or a classroom, or a hospital), we need to choose an appropriate method

These methods partition the *variability* between measurements to different sources: e.g. within-subject and between subject, so that accurate estimates and valid inference can be performed



Recall that the variability in the population affects the variability of our estimates – so it matters that different units have different variability. We need methods that take this into account so that we can accurately infer the variability of our estimates (accurate confidence intervals and p-values)

Lazic, Stanley E. *Experimental Design for Laboratory Biologists : Maximising Information and Improving Reproducibility* . Cambridge, United Kingdom: Cambridge University Press, 2016. Print.

Experimental Unit leads to N

In a simple experiment, sampling unit = experimental unit = observational unit. And your sample size is the number of subjects that you have

Things can get complicated when your experimental unit is not the same as your sampling/biological unit. When considering replication, you may need to consider two levels: the number of individuals per cluster, and the number of clusters in your study

Animal experiments can be used to illustrate different forms of clustering. In a simple experiment where an experimental drug is administered to mice, the drug may be administered to each mouse. But what about an animal experiment when the treatment is applied to the mother (mare, sow, ewe, etc) and the measurements are carried out on babies in the litter?

Replication

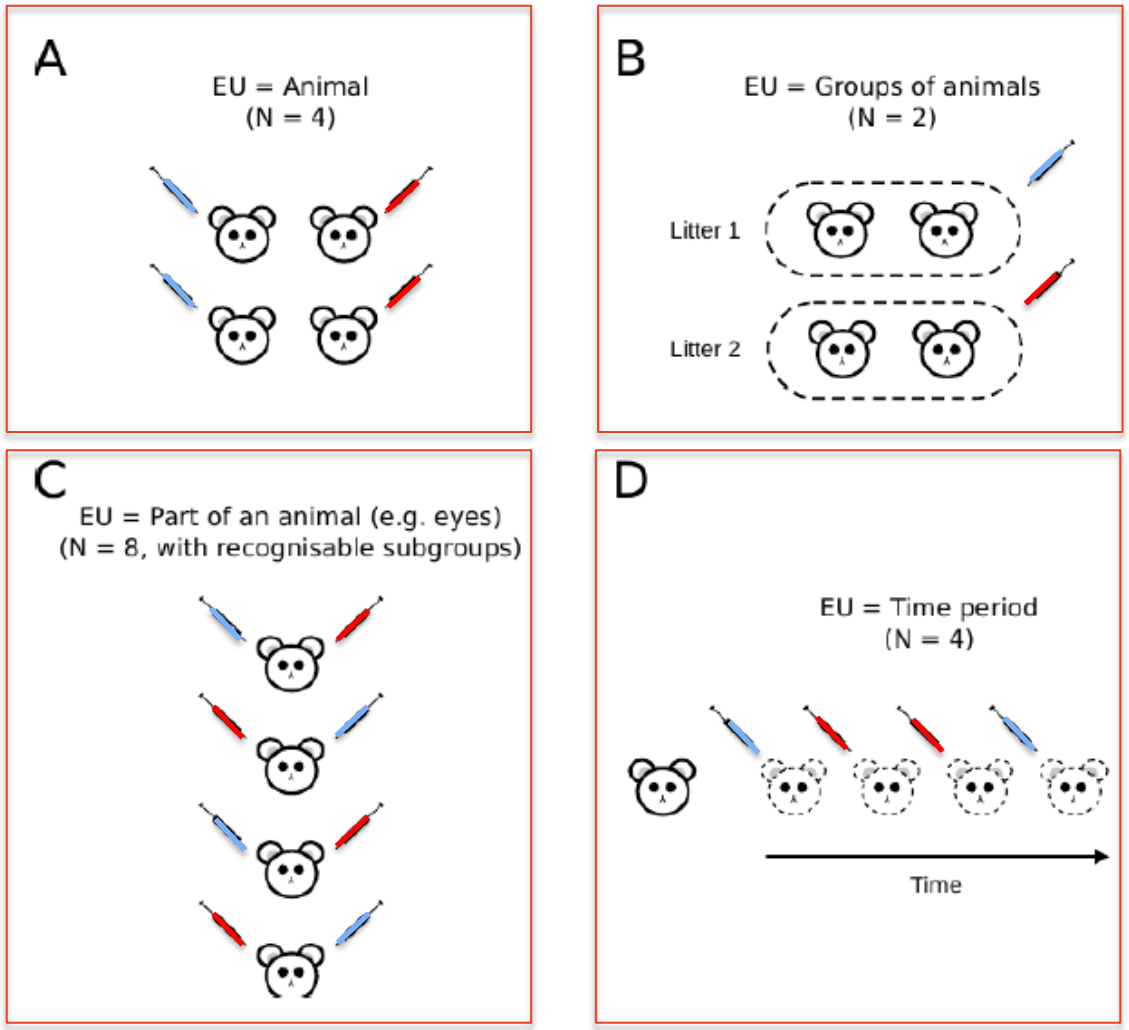
- Replication is a requirement for statistical inference. You must be able to quantify how your underlying measurements vary to make probabilistic, inferential statements. Generally, 3 replicates per group is the minimum required for a simple group comparison.

- True replication occurs when you have multiple independent measurements at the level of each experimental unit.

- **Pseudo-replication** occurs when repeated measurements on a unit are not independent of each other, and **technical replication** occurs when you take repeated measurements on an observational unit to increase the precision of that measurement (the average of the measurements usually stands in as a single measurement)

- The amount of replication needed to reach a desired level of statistical power is the focus of our Power and Sample Size Calculation workshop

Experimental Unit leads to N

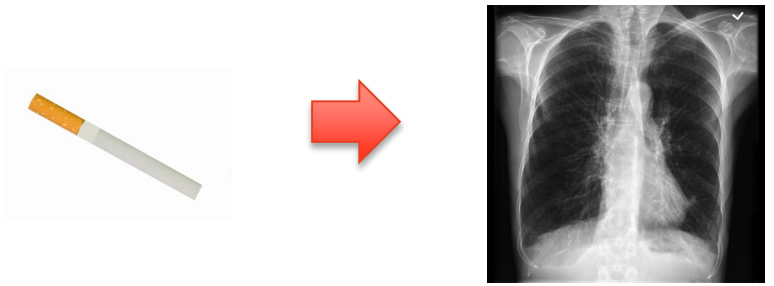


Step 2) Grouping

Terminology of experimental design

- We introduced the different types of variables in Research Essentials
- In the language of experimental design, explanatory variables are often referred to as **factors** (e.g. confounding factors)
- The value of the factor for an individual (e.g. smoker, non-smokers) are referred to as levels of the factor
- When we manipulate a factor to test its effect, it becomes a **treatment** variable
- Other explanatory variables are referred to as design variables (often including blocking variables)

From Our Research Essentials Workshop



Smoking	**Lung disease**
Predictor	Response
Explanatory variable	Outcome
~~Independent variable~~	Dependent variable

Experimental controls

- Treatment **variables**/factors usually have at least two **levels:**
  - A new/experimental treatment
  - Control: i.e. no new treatment

- Having at least one control level/control group is usually essential to compare the treatment level to:
  - Placebo is often used for a drug treatment
  - Sham surgery, a form of placebo for surgical interventions
  - Where a drug is made up in a solution for delivery to animals or cells, 'Vehicle' contains everything other than the drug
  - Wild-type cell line, as opposed to mutant
  - Standard of care may be used in a clinical experiment where a new clinical intervention is being tested

Experimental Design – Randomisation

What is randomisation?

Random allocation of treatments to experimental units

Why randomise? So we can avoid:

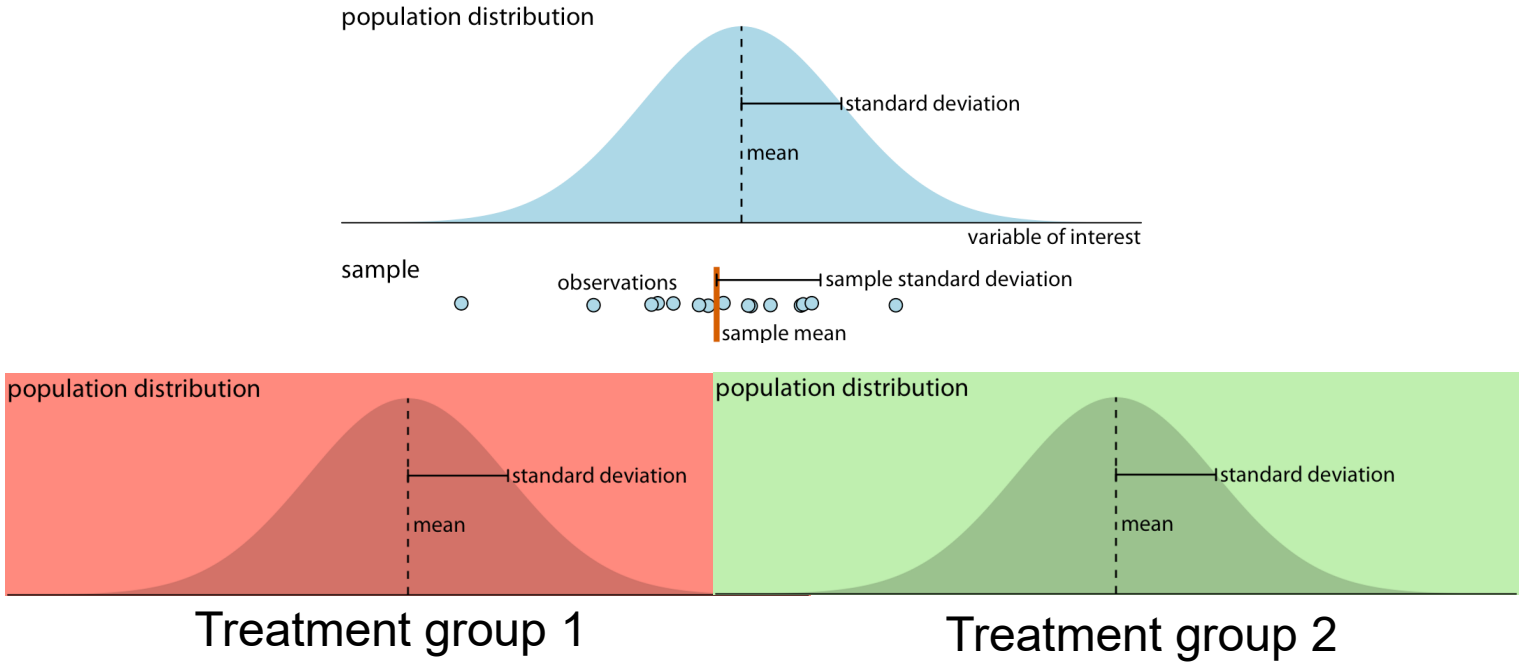
- Systematic bias – e.g. allocating all the drug treatments first, then the placebos
- Selection bias – e.g. subconsciously (or consciously!) choosing healthy patients for the treatment
- Unknown unknowns – potential confounding factors we don't even know exist.

Randomisation greatly strengthens our causal inference by minimising the effect of confounding variables

A philosophical point: randomisation is *not* about perfect balance

Unknown-unknowns cannot be accounted for in an observational study. Some have theorised that all factors including unknowns-unknowns could be perfectly balanced between groups in a randomised experiment. In practice, there are too many unknown-unknowns to achieve perfect balance of all factors across treatment groups with randomisation.

Randomisation does not rely on this perfect balance (which in theory would result in no variability between trials!), instead it allows us to balance the major sources of variability and end up with treatment groups with potentially different means, and individuals that vary to the same extent as individuals in our target population would (after any blocking). Thus we can make as precise estimates as possible in the presence of inevitable randomness.



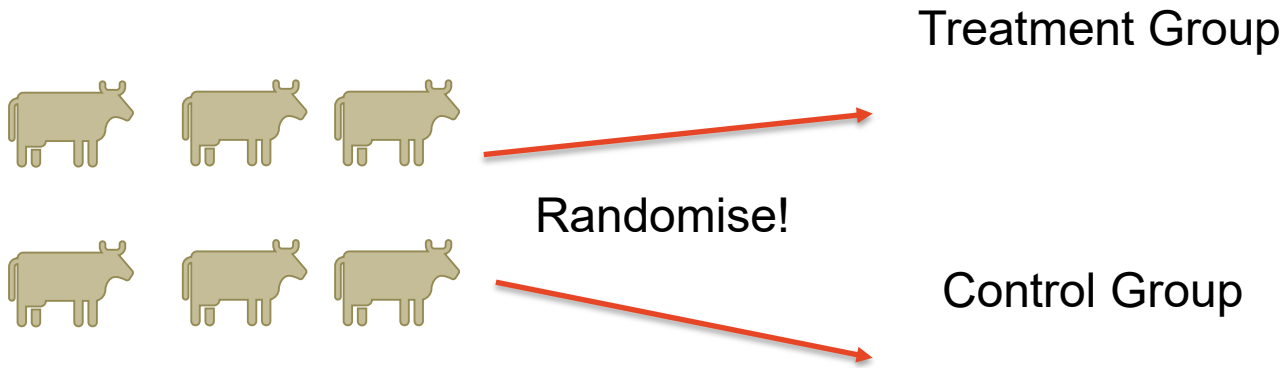
https://errorstatistics.com/2020/04/20/s-senn-randomisation-is-not-about-balance-nor-about-homogeneity-but-about-randomness-guest-post/

Experimental Design – CRD

Completely Randomised Design (unstructured design)

Example: Evaluate the effect of a feed supplement on the growth of calves

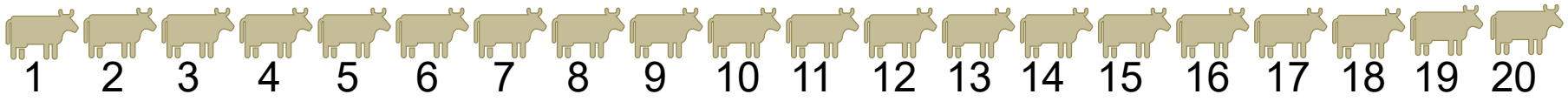
- Suppose that we have no information about the calves (subjects) that we might otherwise use.
- In this case we treat all subjects the same and use randomisation to eliminate allocation biases. (e.g. so the biggest, pushiest cows don't end up in one group)



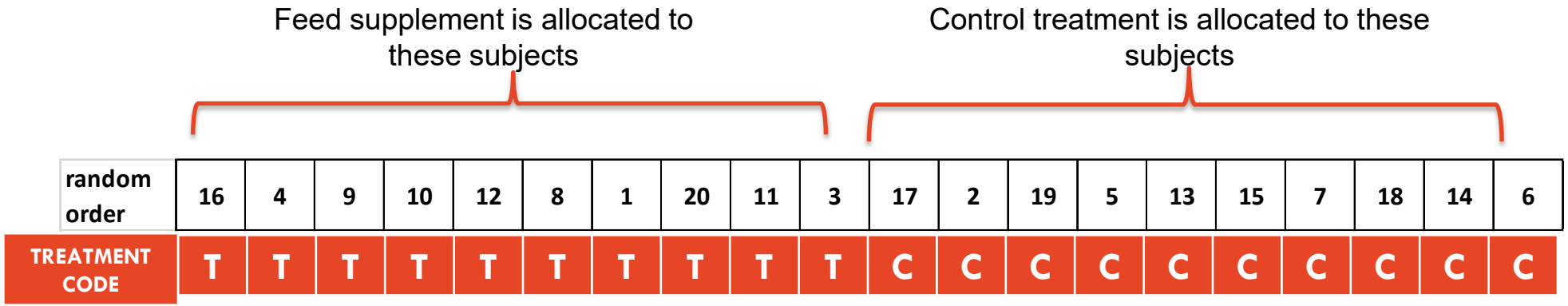
Completely Randomised Design

A1	B1
16	0.031141412
4	0.041331209
9	0.044095909
10	0.132242434
…	…

- Suppose we have 20 subjects and 2 treatments (T and C)
- Assign an ID number to each subject from 1 to 20



- Generate a randomly ordered sequence of numbers 1 to 20 (eg from Excel)
- In Excel use formulae: with IDs in A1; B1=rand(); copy down 20 rows; sort on B1



Completely Randomised Design

How will this experiment be analysed?

We have 2 treatment groups: control and feed supplement

We randomised their treatment ignoring (for the moment) other factors.

Two ways to analyse depending on our design:

1. If we recorded their weight at the beginning of the experiment then we can use a simple linear model to compare the weight gain in each group after adjusting for their baseline weight
2. If we didn't record their weight at the beginning of the experiment we can perform an unpaired t-test

For various reasons, option 1. is probably better

Reducing variability between treatment groups

- Recall the two challenges that sampling poses to estimation of the true effect
  - Bias
  - Error

- Randomisation prevents introduction of allocation biases and strengthens causal inference

- What about the known-knowns: factors other than the treatment that are known to affect our outcome, e.g. the sex of the calves? Can we do anything with them?
  - Randomisation also helps reduce potential bias caused by these factors being imbalanced in different groups
  - However we can also minimise their contribution to error by using blocking

Blocking

- In practice, we may be able to group subjects who are more similar to each other into **blocks.** This is the designed experiment equivalent of **strata** in observational studies

- A blocking variable is a variable that is thought to affect the outcome, but is typically not of interest to the experimenter.

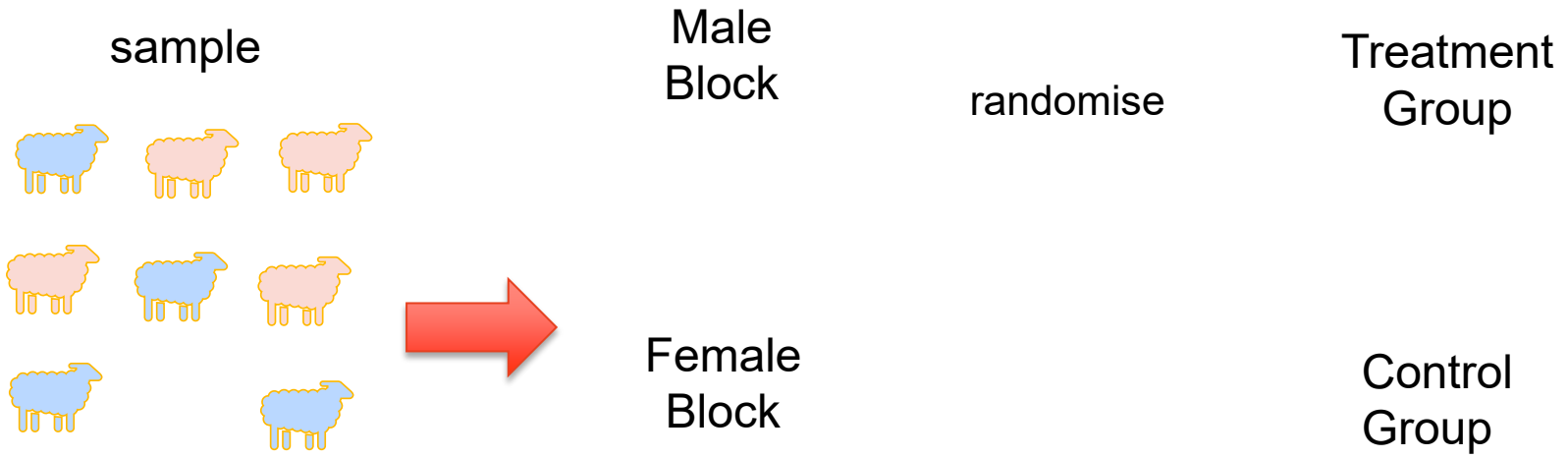
- Balancing your blocking variables within treatment groups will minimise error in your estimation of the treatment effects

- You may have more than one blocking variable, but it is often only feasible to have a few, so choose those which have the greatest potential effect on your outcome

- Rule of thumb: “Block what you can, randomise [for] what you cannot” – George Box

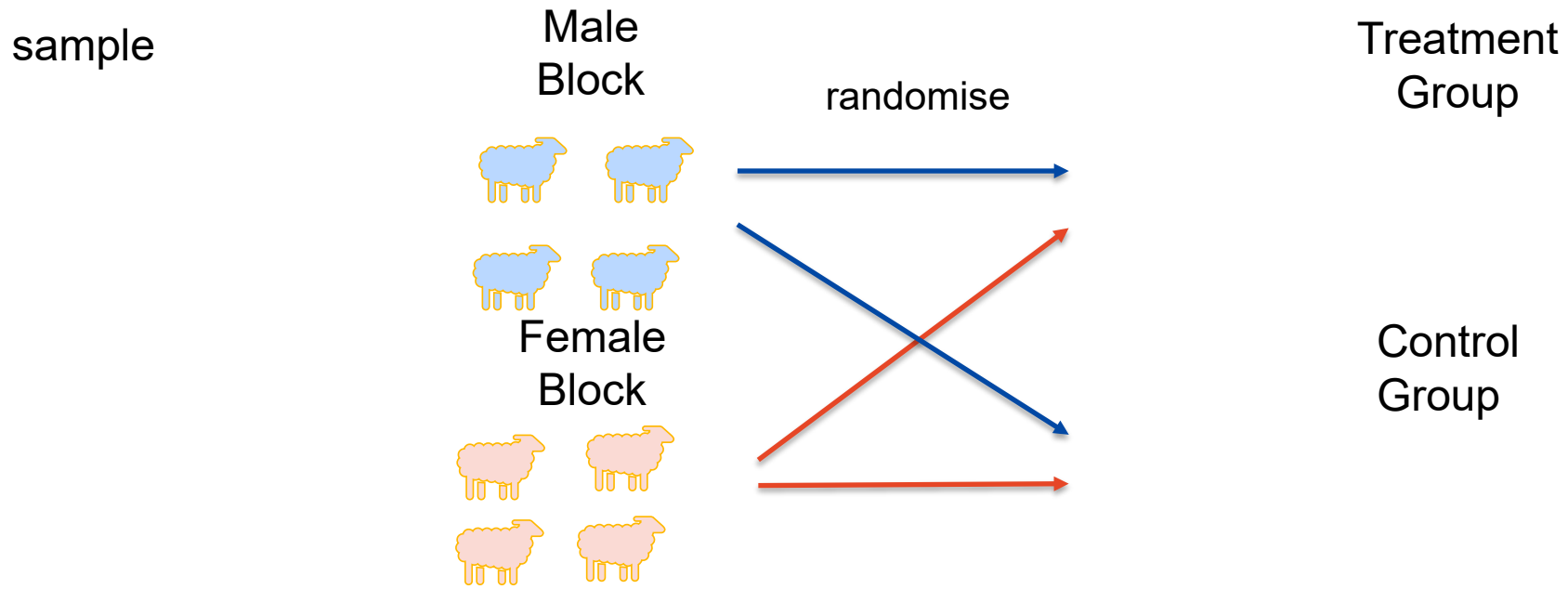
Experimental Design – Randomised Block Design

- **Randomised Block Design (RBD)**
- Example: Evaluate the effect of a feed supplement on the growth of sheep
- Suppose now that we are able to source equal numbers of males and females.
- Use sex as a block variable and randomise within blocks.



Experimental Design – Randomised Block Design

- Randomised Block Design (RBD)
- Example: Evaluate the effect of a feed supplement on the growth of sheep.
- Suppose now that we are able to source equal numbers of males and females.
- Use sex as a block variable and randomise within blocks.



Randomised Block Design

- Sex will be a block variable.
- Male Block

Treatment Code	T	T	T	T	T	C	C	C	C	C
Random order	10	8	2	1	4	9	3	7	6	5

- Female Block

Treatment Code	T	T	T	T	T	C	C	C	C	C
Random order	14	20	19	11	15	18	12	17	16	13

- The allocation is randomised within each block.
  Codes for M 1~10, codes for F 11~20.
- What would be the disadvantage of not blocking for sex in this case?
- How will this experiment be analysed?

Randomised Block Design

- If we do not think males and females have potentially different responses to treatment then we could use an ANOVA model that adjusts for block and tests for the effect of supplement

  $Y = \beta_o + (\text{supplement})\beta_1 + (\text{male})\beta_2 + \varepsilon$

- If we hypothesise difference in male and female response to protein supplement we use an interaction model

  $Y = \beta_o + (\text{supplement})\beta_1 + (\text{male})\beta_2 + (\text{supplement*male})\beta_3 + \varepsilon$

Main effect of interest: β_1
Blocking variable effect: β_2
Interaction is the additional effect of supplement for male (interaction): β_3

- See linear models series of workshops for more details on analysis using ANOVA

Allocation in randomised control trials

- In Randomised Control Trials you may have heard the term 'block randomisation' to refer to allocation of patients to treatments. This is very different to our Randomised Block Design
- In RCTs we are randomly allocating treatments to participants who enter the study at different times. Before the trial we come up with a randomisation sequence.
- We use shorter blocks of sequence to avoid imbalance in the number in each treatment group at any point during trial
- E.g. for a two-arm trial with (T)reatment and (C)ontrol:
  - There are six possible blocks of size 4: TCTC, CTCT, TCCT, CTTC, TTCC, CCTT
  - If we were recruiting a maximum of sixteen patients, we could randomly choose four of these blocks for our random allocation sequence: TCTC CTCT CTCT CCTT
- To preserve blinding (concealment of allocation), you can use the above technique with random block sizes
- If we have defined strata in an RCT, we can come up with a sequence for each strata, which is referred to as a 'stratified randomisation'

Latin Square design

Used to create a balanced design with more than one blocking factor.

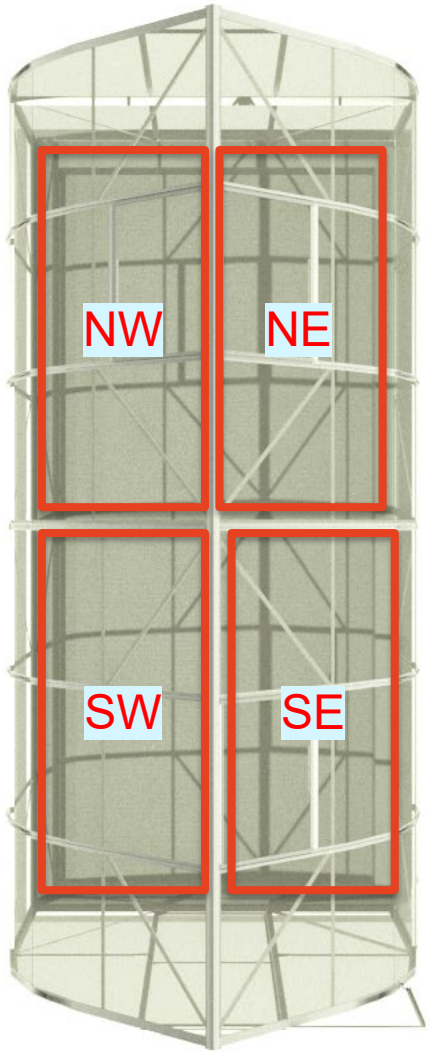
Example: Growing plants in a greenhouse using different fertilisers

Treatments: Fertiliser A, B, C & D

Row block: shelf position 1,2,3,4

Column block: Corner position NE, NW, SE, SW

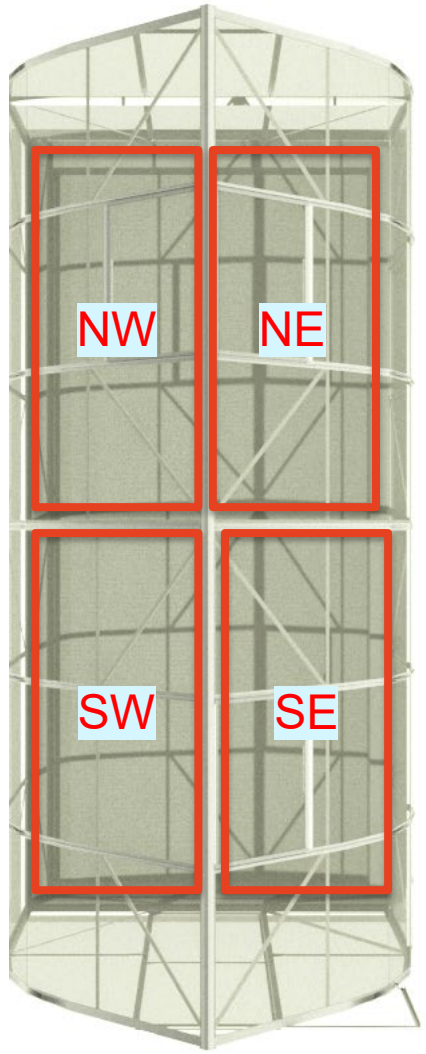
we can create a **4 x 4 Latin Square design**.



Latin Square design

Example: Growing plants in a greenhouse.

Row = Greenhouse corner	Column = Shelf position: 1	2	3	4
NW	A	B	C	D
NE	D	A	B	C
SW	C	D	A	B
SE	B	C	D	A



Each treatment occurs once per shelf position and once per corner position.

Incomplete block design

- The previous examples worked out well because the size of the block matched the number of treatments (four shelves, four corner positions and four fertilizers)
- The advantage of a complete block design is because all blocks contain all treatments, there is no confounding of treatment effect with the effect of block membership, analysis is simpler and more efficient (less replicates required for same accuracy)
- In reality, our experiments may not work out so nicely. We may want to use blocks that do not have as large a block size as the number of treatments
- In this case we can use a form of incomplete block design, we still aim for balance across the experiment, but cannot achieve it within a single block

Balanced incomplete block design

- Every *pair* of treatments occurs together in a block the same number of times.
- Example: 4 treatments A, B, C & D
- Block size is only 3 (e.g. 3 shelves, not 4)

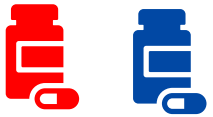
	Column = Shelf position		
	1	**2**	**3**
Block 1	A	B	C
Block 2	A	B	D
Block 3	A	C	D
Block 4	B	C	D

Don't forget, blocks could be batches, days, cycles, fields, etc

Step 3) Data collection

Experimental Design – Blinding

- Blinding is another method to avoid bias



- Blinding can reduce or eliminate confounding bias due to conscious or unconscious preferences or expectations

<u>In RCTs:</u>

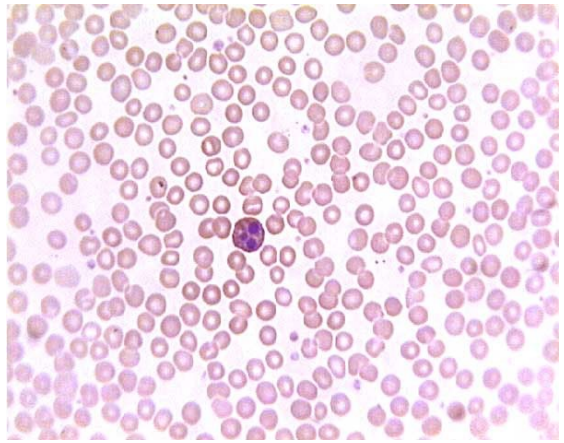
- Blind trials (or single blind) – the subject does not know if they are in the treatment or the placebo group
- Double-Blind trials – Both the subject and the technician are not aware of the assigned treatment
- Open trial – All the treatment information is known to the subject and technician/experimenter

Experimental Design – Blinding

– Laboratory experiments can also benefit from blinding to prevent bias.

Example 1: Histology cell counts

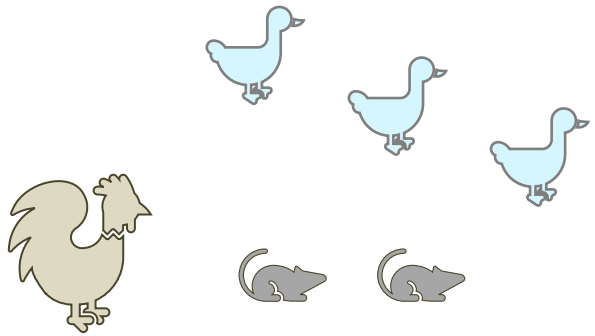
- Counting cells requires judgement (e.g. location sampling, recognising cell types)
- The technician should not know the identity of the specimens
- Use an ID code to anonymise the samples. Randomisation of processing order will also help



Example 2: Animal behaviour

- Many animals respond to the way they are handled
- The technician should (ideally) not know the identity of the animal's treatment group.

Can you use blinding in your research to guard against unconscious bias?



Questions so far?

Step 4) Consider analysis

Why consider analysis of your data, before you have even collected it?

- First and foremost, you should have some idea of the power of your experiment to know whether it is worth doing as designed
- In order to calculate the power of an experiment, you need an analysis plan
- More generally, thinking 'ahead' to the analysis of your data may reveal aspects of your experiment that you hadn't thought about (controls, number of groups, group size), which will in turn affect what data you collect… "an important way to ensure that you collect all the data you need and that you use all the data you collect[1]"
- "Also important for research integrity and quality, guarding against "'data-driven results'[2]"
- Our research essentials workshop discusses possible analysis methods for different variable types, and other workshops will help you develop a detailed analysis plan once you have chosen your statistical methods

- 1. https://www.cdc.gov/globalhealth/healthprotection/fetp/training_modules/9/creating-analysis-plan_pw_final_09242013.pdf
- 2. https://nceph.anu.edu.au/files/Data_Analysis_Plan_Guide_20131125_0.pdf

Analysis of repeated measures

- **Repeated Measures Design (or within-subjects design)**
- Repeated measures are not technical replicates when they represent another aspect of the same subject/sample, typically observations over time
- Repeated measures are pseudoreplicates (not independent observations)
- There are specific statistical procedures to deal with RM's, discussed in our Linear Models workshops
- A paired t-test is a simple repeated measures extension to an 'unpaired' t-test

- It is recommended to calculate the amount of replication required for your experiment using power calculation. The topic of our Power and Sample Size Calculation workshop.

Consider a linear model for your data analysis

To analyse your data appropriately you may need to think about what kind of linear model you need. Even if your outcome is categorical, even if you plan to do a simple group comparison (such as a t-test or ANOVA) you should consider which model is appropriate

A linear model allows you to adjust for confounding factors that may affect your outcome.

One or more fixed Factors

- These are usually the explanatory variables chosen by the experimenter. They have defined levels or categories and for factors of interest, we want to quantify the difference between them.

Your model may also include random Factors

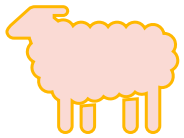
- These are usually incidental to the purpose of the experiment (such as blocking variables).
- The levels of the random factor should be chosen from a larger population of possible values of the variable.
- We don't estimate the difference between levels of a random effect, rather we use the effect to partition variance and thereby reduce within group variance.
- You may use multiple levels of random effects (e.g. individuals within households within districts)

Discussed in further detail in the Model Building and Linear Models workshops 

Conclusions

Challenge Question – Sheep vaccine experiment

Research Question: Does the use of a new vaccine result in a different incidence of parasite infection compared to the standard treatment?



I would like 12 sheep in each group (total n = 24)

I have 12 sheep aged 1yr and 12 sheep aged 2yrs

Q: How should you allocate the treatments to the sheep?

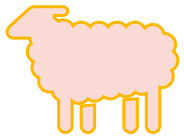
a. vaccinate 6 of the younger sheep and 6 of the older sheep with each treatment.

b. vaccinate12 younger sheep with the new vaccine and 12 older sheep with the standard vaccine treatment.

Option b will result in age and treatment being _____

Challenge Question – Sheep vaccine experiment

Research Question: Does the use of a new vaccine result in a different incidence of parasite infection compared to the standard treatment?



I would like 12 sheep in each group (total n = 24)

I have 12 sheep aged 1yr and 12 sheep aged 2yrs

Q: How should you allocate the treatments to the sheep?

a. vaccinate 6 of the younger sheep and 6 of the older sheep with each treatment.

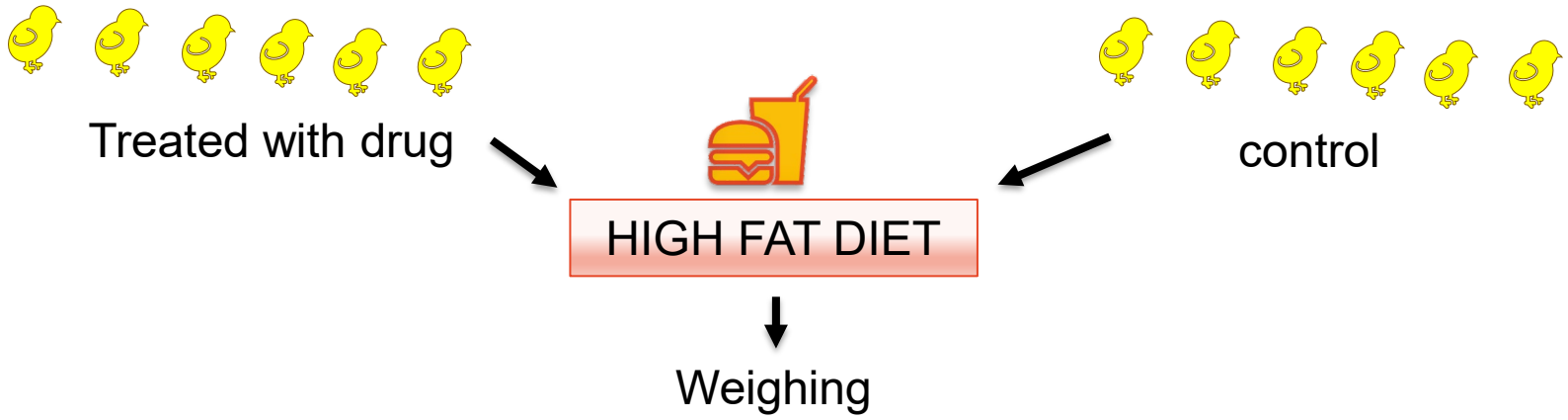
b. vaccinate12 younger sheep with the new vaccine and 12 older sheep with the standard vaccine treatment.

Option b will result in age and treatment being confounded/aliased*

Scenario 2 – Chicken Drug & Diet experiment

Research Question: Groups of treated and untreated chickens are placed on a high fat diet. What is the effect on weight gain?

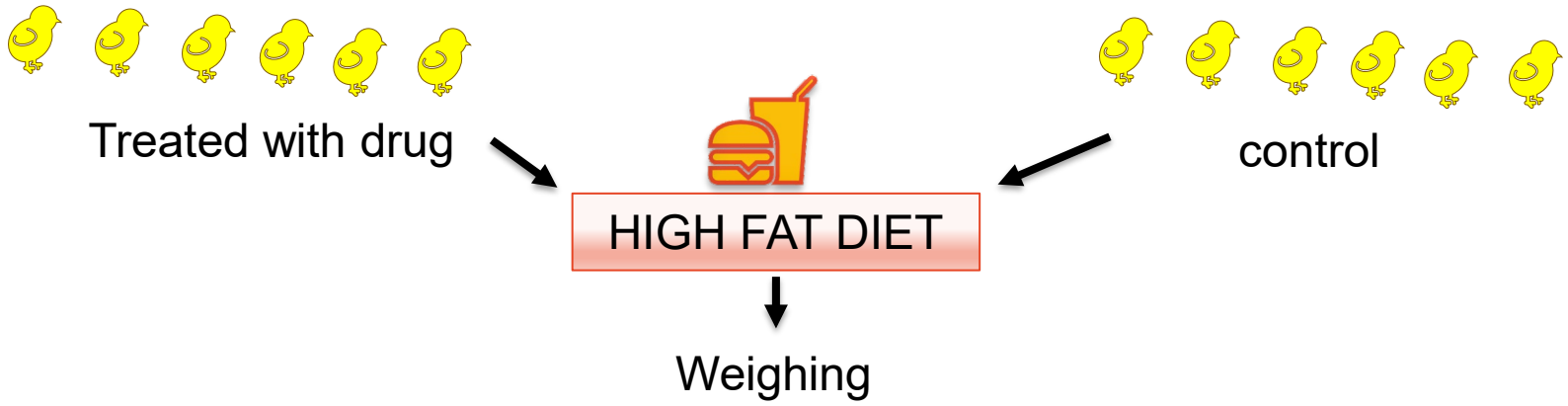
- What are the outcome and explanatory variables?
- What are the treatment variables?
- What are the blocking variables?
- What factors should be fixed in the analysis?



Scenario 2 – Chicken Drug & Diet experiment

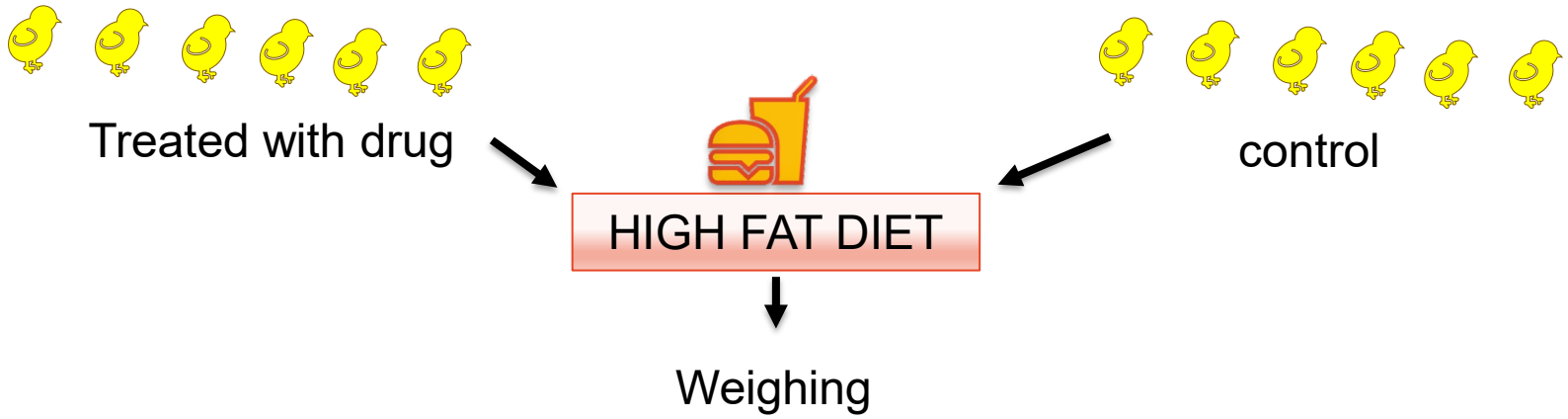
Q. *First what are the:*

Biological Units	
Experimental Units	
Observational Units	



Scenario 2 – Chicken Drug & Diet experiment

Biological Units	chicks
Experimental Units	chicks?
Observational Units	chicks



Worksheet - Chicken Drug & Diet experiment

Fill in details

Outcome Variables	Explanatory Variables	Design features	Blocking Factors	Randomisation & blinding

Worksheet - Chicken Drug & Diet experiment

Outcome variables	Explanatory Variables	Design features	Blocking Factors	Randomisation & blinding
• Weight gain	• Drug treatment (y/n)	• Chick breed • High Fat Diet • Feeding routine • Feeding ad libitum? • Time of day for weighing • Housing – number of chicks per cage?	• Sex? • Chick age, batch, etc	• Drug treatment allocation • Order of handling • Order of weighing

Analytical Design – what statistical test will be used?

Worksheet – Your research

What does your research experimental design look like?

What are the biological, experimental and observational units?

Biological Units	
Experimental Units	
Observational Units	

Worksheet – Your research

Outcome Variables	Explanatory Variables	Design features	Blocking Factors	Randomisation & blinding

Experimental Design – what statistical test will be used?

Experimental design - summary

1. Wherever possible, reduce or eliminate variation due to factors other than the explanatory variable of interest (often treatment variable).
2. Sometimes undesirable variation cannot be avoided due to things beyond your control.
3. Use blocking variables in your design to manage factors that are most likely to cause variation.
4. Use randomisation to prevent bias due to allocation and increase precision in the face of unknown variation outside your control.
5. Use replication to improve precision of your estimated effects.

- Use the general experimental design workflow in this workshop and note the double arrows between stages of the design: there will often be an iterative process of improvement:
  1. Point out the problems
  2. Discuss the implications
  3. Propose a way forward

- Be alert, not alarmed: your experiment doesn't have to be perfect but it should be well thought out through a process of experimental design

Challenge questions

Q. Types of Units

- _____ Unit – is the entity on which measurements are taken
- _____ Unit – is the entity (animal/plant) about which inferences are made
- _____ Unit – is the entity that is randomly and independently assigned to experimental conditions.

The sample size n = number of _____ units

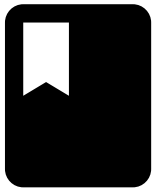
Challenge questions

Q. Types of Units

- Observational Unit – is the entity on which measurements are taken
- Biological/sampling Unit – is the entity (animal/plant/thing) about which inferences are made
- Experimental Unit – is the entity that is randomly and independently assigned to experimental conditions.

The sample size n = number of experimental units

Other resources



Books on Experimental Design

- ”The Design of Experiments” by Fisher, Ronald Aylmer, 1935.
- “Experimental Design for Laboratory Biologists: Maximising Information and Improving Reproducibility” by S.E. Lazic
- “Statistics for Experimenters” by Box, Hunter & Hunter
- Interactive E-book on Experimental Design http://cast.massey.ac.nz

Books on Causality

- “The Book of Why” by Judea Pearl (interesting ideas on causality, confounding, approaches to data)

Books on Bias and Statistical thinking

- “Thinking, Fast and Slow” by Daniel Kahneman

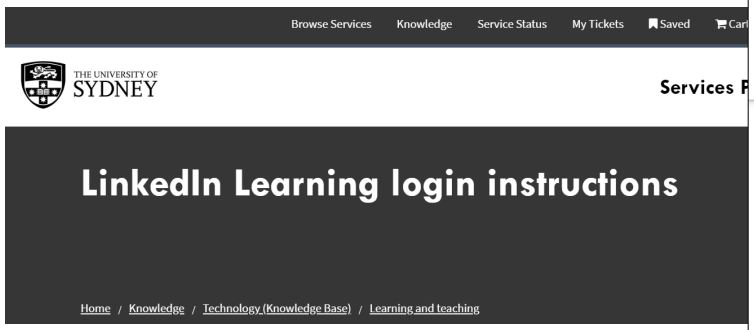
Methods for randomising a sequence

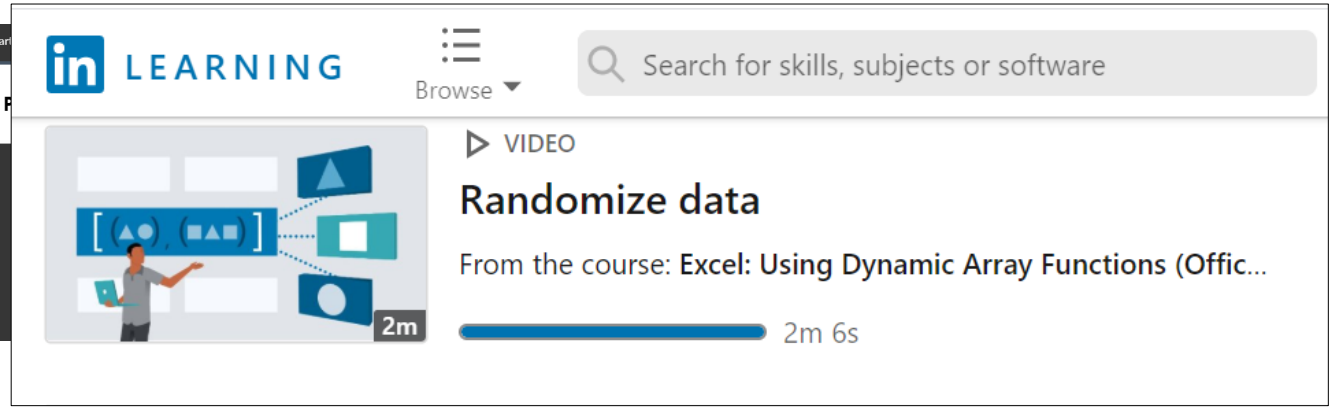
We saw how random sequences are useful for:

- Random sampling of individuals from a sampling frame
- Random allocation of treatments (randomisation)

Other resources for Randomisation:

- Another nifty random number sequence generator is at: www.random.org/sequences
- Have a look at this video on LinkedIn Learning (through your USyd account, check instructions on Services Portal)





Further Assistance at Sydney University

SIH

- Statistical Consulting website: containing our workshop slides and our favourite external resources (including links for learning R and SPSS)
- Hacky Hour an informal monthly meetup for getting help with coding or using statistics software
- 1on1 Consults can be requested on our website (click on the big red 'contact us' link)

SIH Workshops

- Create your own custom programmes tailored to your research needs by attending more of our Statistical Consulting workshops. Look for the statistics workshops on our training page.
- Other SIH workshops
- Sign up to our mailing list to be notified of upcoming training

Other

- Open Learning Environment (OLE) courses
- Linkedin Learning

How to use our workshops

Workshops developed by the Statistical Consulting Team within the Sydney Informatics Hub form an integrated modular framework. Researchers are encouraged to choose modules to *create custom programmes tailored to their specific needs.* This is achieved through:

- Short 90 minute workshops, acknowledging researchers rarely have time for long multi day workshops.
- Providing statistical workflows appliable in any software, that give practical step by step instructions which researchers return to when analysing and interpreting their data or designing their study e.g. workflows for designing studies for strong causal inference, model diagnostics, interpretation and presentation of results.
- Each one focusing on a specific statistical method while also integrating and referencing the others to give a holistic understanding of how data can be transformed into knowledge from a statistical perspective from hypothesis generation to publication.

For other workshops that fit into this integrated framework refer to our training link page under statistics https://www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training.html#stats

A reminder: Acknowledging SIH

All University of Sydney resources are available to Sydney researchers free of charge. The use of the SIH services including the Artemis HPC and associated support and training warrants acknowledgement in any publications, conference proceedings or posters describing work facilitated by these services.

The continued acknowledgment of the use of SIH facilities ensures the sustainability of our services.

Suggested wording for use of workshops and workflows:

"The authors acknowledge the Statistical workshops and workflows provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

We value your feedback

We want to hear about you and whether this workshop has helped you in your research. What worked and what didn't work.

We actively use the feedback to improve our workshops.

Completing this survey really does help us and we would appreciate your help! It only takes a few minutes to complete (*promise!*)

You will receive a link to the anonymous survey by email.