Fundamentals of Experimental Design

Scientific validity and experimental design

Statistical Computing and Empirical Methods
Unit EMATM0061, Data Science MSc

Rihuan Ke
rihuan.ke@bristol.ac.uk

Teaching Block 1, 2024

What we will cover today

We will consider three significant obstacles to valid scientific inference:

1. Measurement distortions
2. Selection bias
3. Confounding variables

We will then consider the role of experimental design in overcoming these challenges.

Validity of scientific research

Validity: Scientific research should use principled approaches to ensure that the results of the investigation actually support the conclusions drawn.

Validity

Our crop yield example: suppose a farmer wants to know if the yields of his wheat fields will change following soil treatment.

We model the difference as i.i.d . draws from a Gaussian $D_1, \cdots, D_n \sim \mathcal{N}(\mu, \sigma^2)$.

For our yield experiment, the computed p-value 0.0183 is below the significance level. Hence, we are justified in rejecting the null $\mu = 0$ and concluding that $\mu > 0$ **provided our assumption that $D_1, \cdots, D_n \sim \mathcal{N}(\mu, \sigma^2)$ (independent) holds.**

Validity? Is the statement that the soil treatment makes a difference valid?

However, the independence assumption is often violated in practice.

- e.g., imagine a crop disease across the fields.

Even if we have shown that $\mu > 0$ does it follow that the change was caused by the treatment?

- If all fields are treated in one year and untreated in another there could be another causal factor e.g. weather.

Random variation vs. threats to validity

Scientific research can reach false conclusions due to random variation.

- e.g. By pure chance your sample average is much greater than the population average.

Random variation can be handled in different ways

1. The role of error due to random variation can be quantified and understood through statistical techniques such as confidence intervals and hypothesis testing.
2. As the size of the data grows the random error typically goes to zero.

Scientific research can reach false conclusions due to problems with the validity of a methodology

Problems with the validity of a methodology are more pernicious and typically cannot be resolved by big data!

Obstacles on the path to valid inference

1. Measurement distortions

2. Selection bias

3. Confounding variables

Measurement distortions:

your data has errors

Valid measurements

A valid measurement is one that accurately reflects the aspect of reality you intend to measure.

Example 1. A scientist wants to understand the effect of coffee on “concentration levels”.

What is a valid measure of someone’s “concentration level”?

- Perhaps we can measure the amount of time taken to perform some arithmetic tasks?
- Or perhaps we can ask people how distracted they feel when reading?

Valid measurements

A valid measurement is one that accurately reflects the aspect of reality you intend to measure.

Example 2. An employer wants to measure people's "computer science ability".

What is a valid measure of someone's ability as a computer scientist?

- Perhaps we can measure how long it takes for them to solve a set of algorithmic problems?
- Or perhaps ask someone to score the presentation of their code?

Valid measurements: proxy measurement

A valid measurement is one that accurately reflects the aspect of reality you intend to measure.

Example 1. A scientist wants to understand the effect of coffee on “concentration levels”.

Example 2. An employer wants to measure people’s “computer science ability”.

“concentration levels” and “computer science ability” are imprecise concepts and not directly measurable

Often we must make do with a proxy measurement: a variable that can be accurately measured and is believed to correlate well with the true variable of interest

- e.g. “Speed to complete tasks A, B, C” rather than “attention level”

It is vital that this choice of proxy is well documented.

Conclusions drawn from the research should also reflect the reliance upon a proxy measurement.

Measurement error

Often we must make do with a proxy measurement: a variable that can be accurately measured and is believed to correlate well with the true variable of interest

…. but you may not be always able to measure the variable accurately

Measurement error is the difference between the measured value of a quantity and its true value.

Measurement error is very common in practice

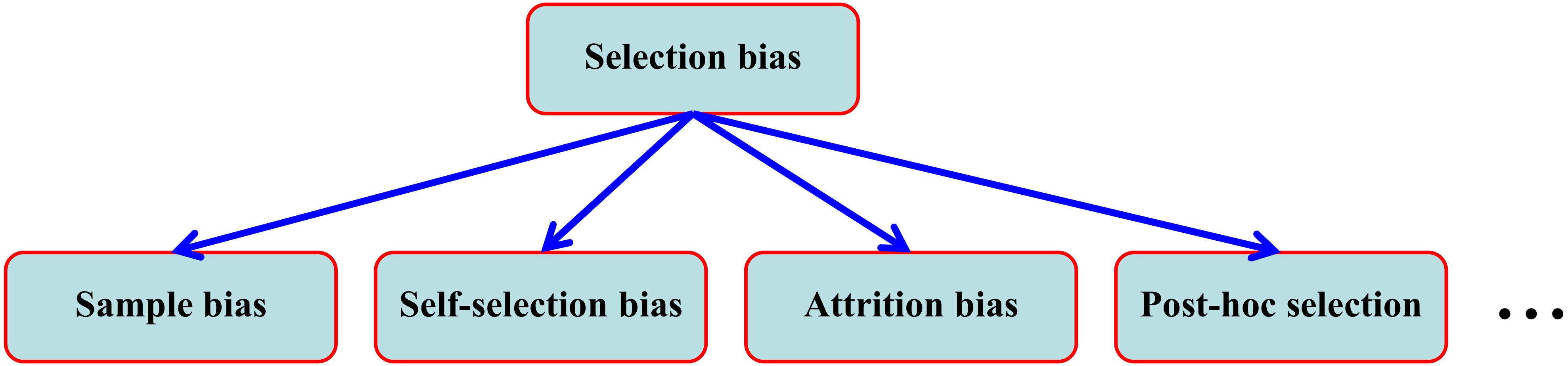
- Miscalibration of measurement instruments e.g. a clock running slightly too fast or slow.
- Rounding errors due to computational constraints e.g. 0.904 converted to 0.9.
- Inaccurate responses to questionnaires e.g. under reporting alcohol consumption.
- Human error e.g. data entry mistakes.

Selection bias:

your sample is not selected properly

Selection bias

Selection bias occurs when the data included in the analysis misrepresents the underlying population of interest.



Note: The different forms of bias are not mutually exclusive.

Sample bias

Sample bias occurs when some members of your intended population are more likely to be sampled than others.

Example: You want to know what the most popular genre (jazz, folk, rock …) is amongst attendees of a music festival.

You conduct a survey at a two-day music festival and all of your participants are selected on the morning of the first day.

Is your sample representative of the population of festival goers?

- What if the Saturday focuses on folk and the Sunday focuses on rock?

Self selection bias

Self-selection bias occurs whenever participants self select whether or not they are assigned to a group.

Self-selection bias often results in sampling bias.

Example: online reviews of restaurants might disproportionately represent subsets of the population with strong opinions or certain age groups.

Example: the results from medical trials can be distorted by their over reliance upon student participants.

Example: Medical trials can also be distorted by allowing patients to self select which medication they choose.

Attrition bias

Attrition bias occurs when a sample is distorted by participants leaving a study.

Example: A scientist is investigating the efficacy of a new exercise program.

Participants may leave the study if they are not having success.

The sample of remaining participants may not be representative.

Post hoc selection

Post hoc selection occurs whenever a subset of the data is chosen based on the sample itself.

Example: A scientist is investigating the efficacy of medical treatment.

The average performance on the sample is disappointing.

However, there exists a subgroup for which the treatment performs better.

We must not treat the sub-sample as if were the original sample!

Randomized samples

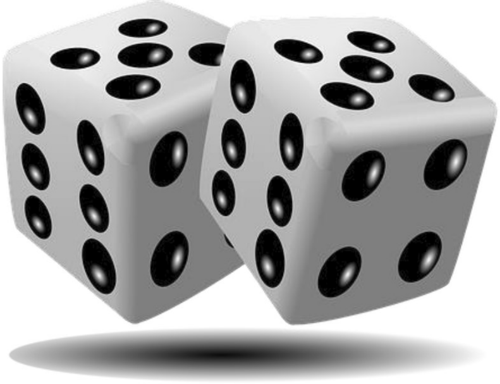
Solution:

The ideal solution to selection bias problems is randomization: Data is randomly sampled from the population of interest with uniform weight.

Example: sampling from the population of interest which is the set of all attendees of a music festival.

We obtain a list of ticket numbers, pick a number uniformly at random and ask the ticket holder to participate.

In reality, problems like self selection and attrition bias are difficult to overcome.



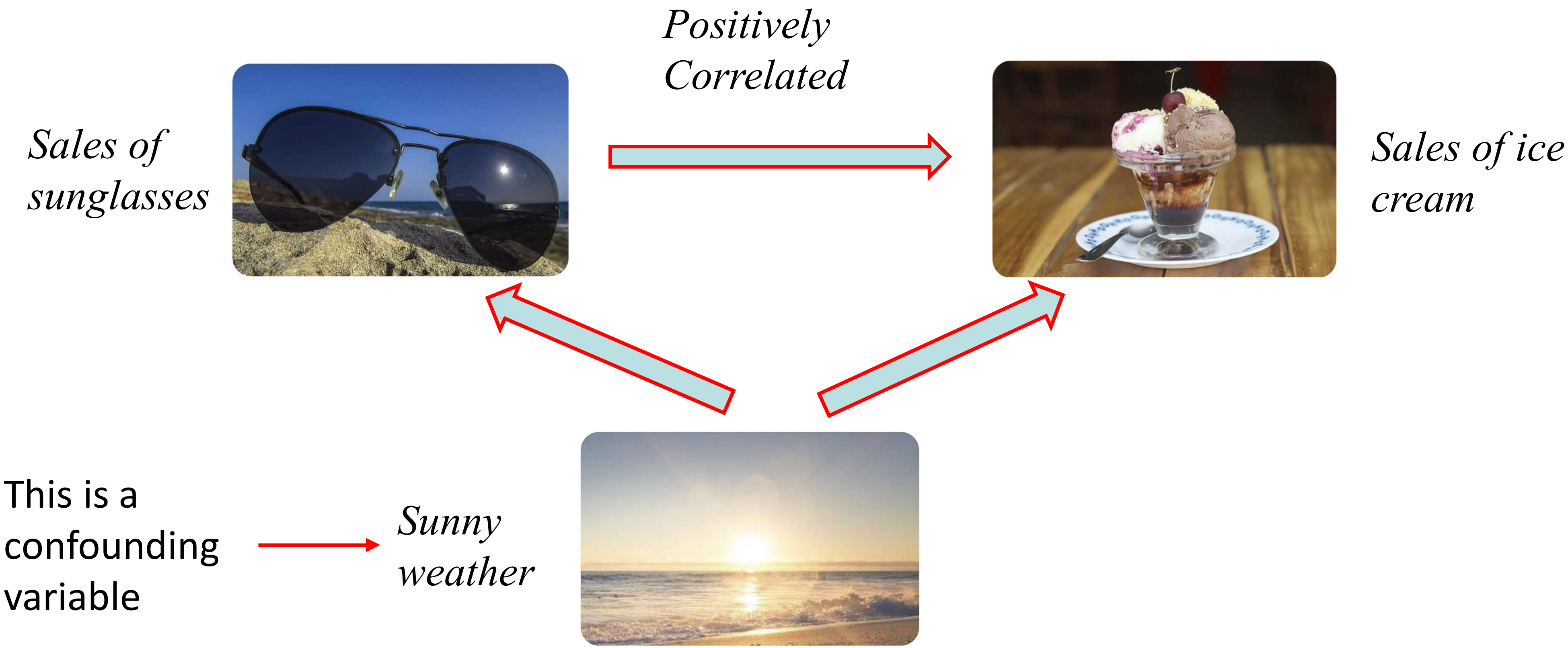
Confounding variables:

correlation ≠ causation

Confounding variables

Correlation does not imply causation!

<u>Example</u>: the increased sales of sun glass may not be the reason for the increased sales of ice cream, but the sunny weather is!



Positively Correlated

Sales of sunglasses

Sales of ice cream

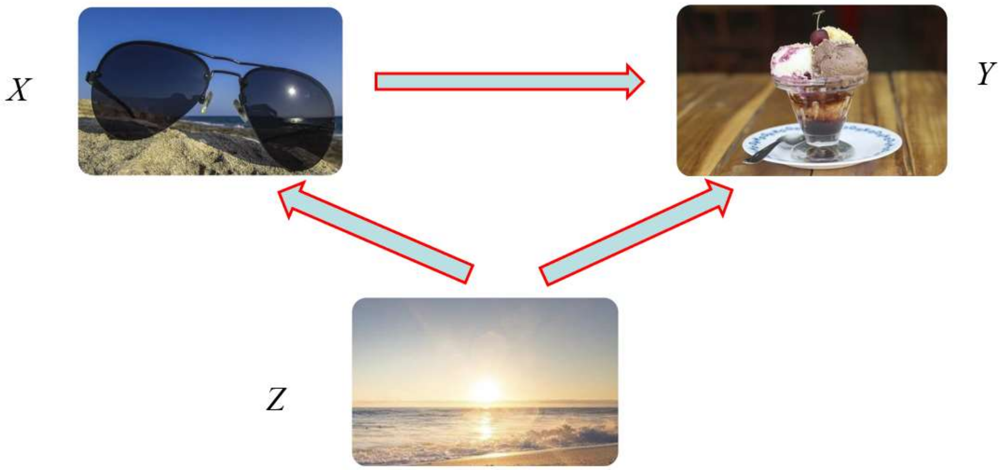
This is a confounding variable

Sunny weather

Confounding variables

Suppose we want to understand the causal relationship between two variables X and Y.

A confounding variable Z is a third variable that has a causal effect on both X and Y.



Note: the sunny weather (Z) has a causal effect on both the sale of sunglasses (X) and the sales of ice cream (Y)

Confounding variables & correlation

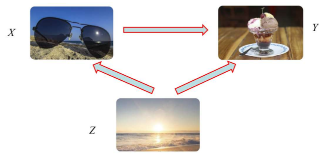
Correlation might be a consequence of causal relationships, but …

… the correlation of two variables might also be a consequence of the confounding variable:

Sunny weather (confounding variable) leads to better sales of sunglasses

Sunny weather (confounding variable) leads to better sales of ice-cream

So the sales of sunglasses and the sales of ice-cream are correlated



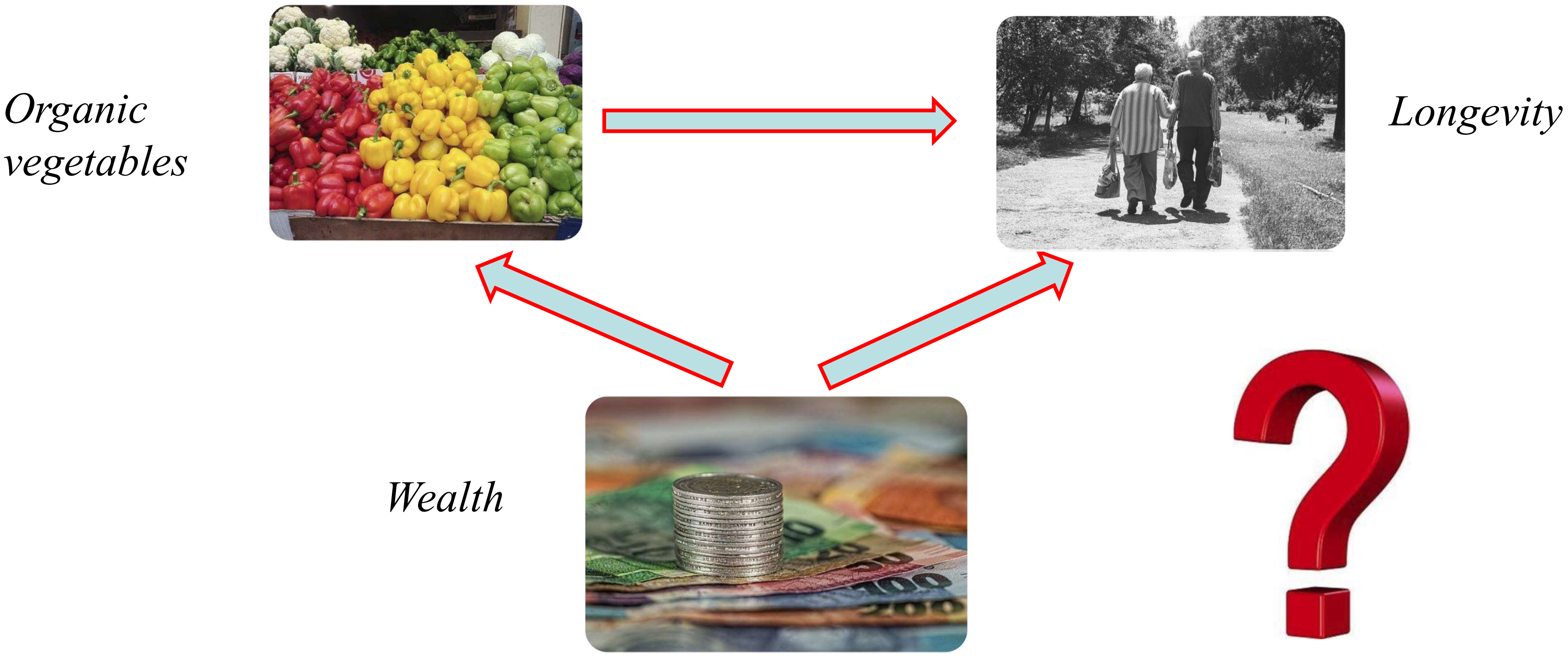
So confounding variables are the troublemakers (as we can not conclude causal relationships by correlation!)

Maybe we want to remove the correlation caused by confounding variables, through experimental design? We can use the idea of randomized intervention (which will be introduced later)

Confounding variables

Confounding variables can obscure causal relationships.

Example: Does having organic vegetables imply an increase in life expectancy?



Organic vegetables

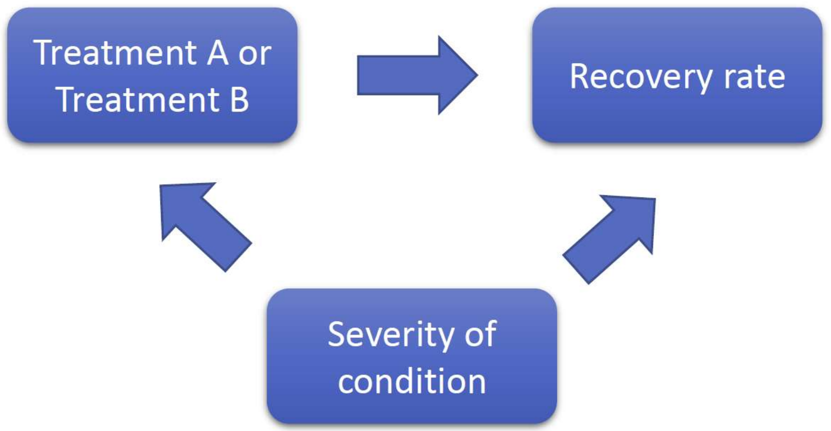
Longevity

Wealth

e.g. a wealthy person may have access to more organic vegetables as well as better medical treatment

Confounding variables: Simpson's paradox

Confounding variables can obscure causal relationships.



In each of the two patient groups (poor conditions vs good conditions), the success rate for treatment A is better than that of B.

Surprisingly, the overall success rate for B is better than A (when computed for all patients)

This surprising phenomenon is due to the confounding variable (severity of condition)

Success rates for two different treatments

	Treatment A	**Treatment B**
Patients in good conditions	19/20 (95%)	72/80 (90%)
Patients in poor conditions	60/80 (75%)	12/20 (60%)
All patients	79/100 (79%)	84/100 (84%)

Experimental design

While correlation is what we observed, more often causal relationships are what we are most interested in.

Unfortunately, causation typically cannot be inferred from statistical correlations alone…

- The observed positive correlation between the sales of sun grass and the sales of ice cream does not imply a causal relationship between them
- The observed positive correlation between having more organic vegetables and increased life expectancy does not imply a causal relationship between them
- The better treatment for a medical condition does not imply higher success rate

These are because of the confounder variables

Fortunately, the impact of confounder variables can be reduced or removed by careful experimental design.

Independent variable & dependent variable

Question: we want to study the causal effect of a variable X upon another variable Y.

The variable X is referred to as the independent variable and Y the dependent variable

Example: We want to understand the causal effect of treatment (X) on a patient's recovery (Y).

We are interested in what happens to Y when we intervene on X, not the correlation between X and Y.

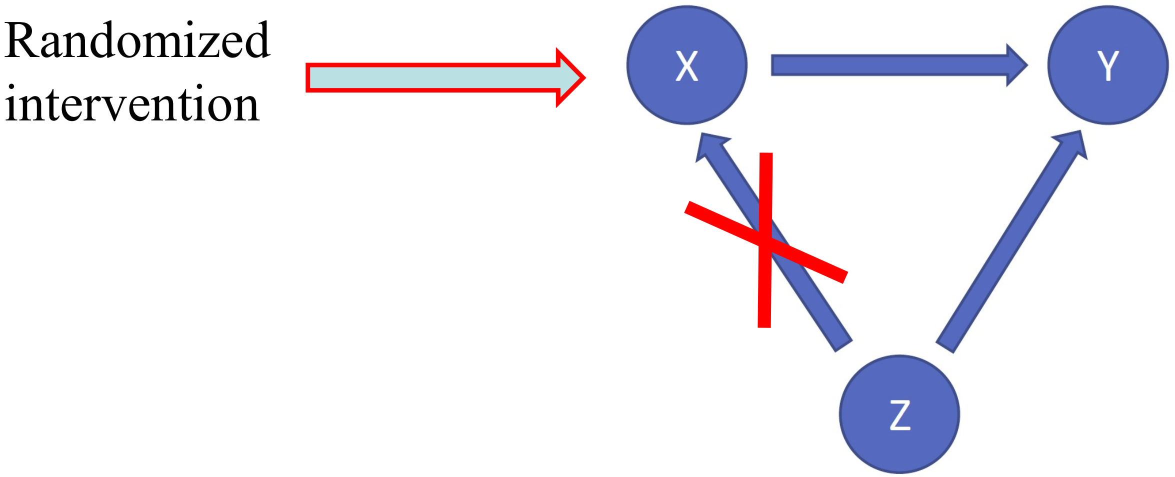
This requires a carefully designed experiment (e.g., randomized intervention) rather than an observational study.

Randomized intervention:

removing the impact of confounding variables

Randomized intervention

We want to study the causal effect of an independent variable X on a dependent variable Y (for example, the independent variable X is the treatment and the dependent variable Y is recovery or not)



By intervening on X we **block** the dependency of X upon any possible confounders Z.

Intuitively, if we block the dependency, we “control” the Z factor and hence are able to see how Y changes with respect to X

Randomized intervention

We want to study the causal effect of an independent variable X on a dependent variable Y (for example, the independent variable X is the treatment and the dependent variable Y is recovery or not)

By intervening on X we **block** the dependency of X upon any possible confounders Z.

There are different Randomized intervention designs (details will be given later), e.g.,:

1. simple between-groups experiment (post test only control group design)

2. pre test/post test control group design

3. Solomon’s four-group design

4. Within subjects designs

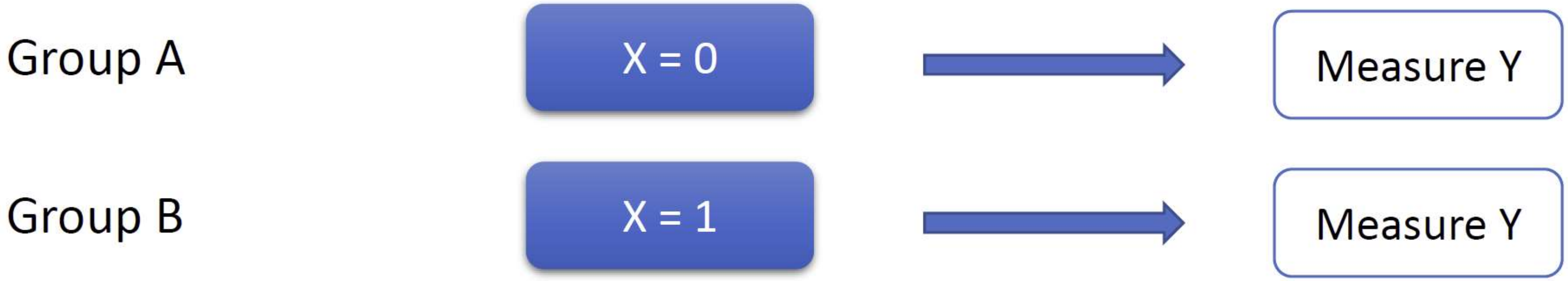
1. Simple between-groups experiment

Question: Suppose that we want to study the causal effect of a binary independent variable X $\in\{0,1\}$on dependent variable Y. We want to remove the impact of Z.

Intuition: construct two groups (of a sample) such that the behaviour of X is different within the two groups, but the behaviour of the confounder Z is the same

Approach:

1. Take a random sample from our population of interest.
2. Partition the sample **randomly** into two groups – Group A and Group B
3. We intervene so that those in Group A receive no treatment X=0 and those in Group B receive treatment X=1.
4. After some time has elapsed the dependent variable Y is measured for those in both groups



Group A

X = 0

Measure Y

Group B

X = 1

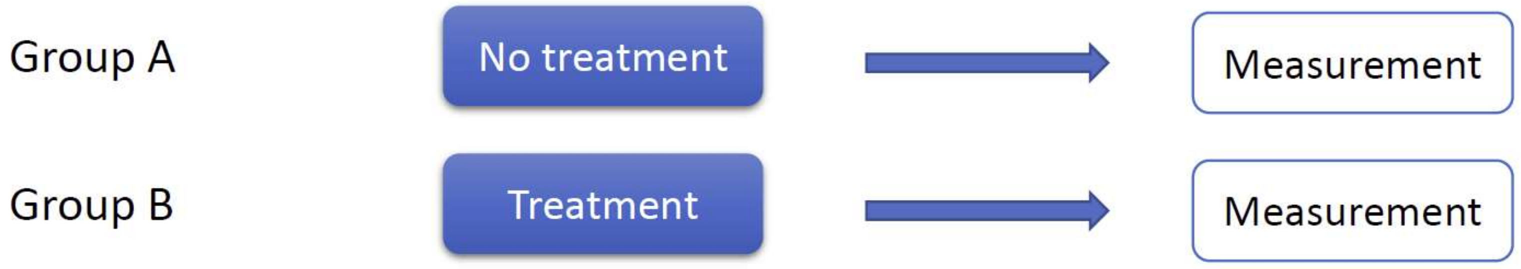
Measure Y

Simple between-groups experiment

Example: we want to understand the effect of blood pressure medication X on blood pressure level Y.

- Obtain a sample randomly
- A random sample is **randomly** partitioned into a control group (Group A) and a treatment group (Group B)
- Those in the control group are untreated (X=0) and those in the treatment group receive the treatment (X=1).
- After some time has elapsed the blood pressure Y is measured for those in both groups.

Note: the randomness of the partition into two groups is the key here. Think about the example of Simpson’s paradox



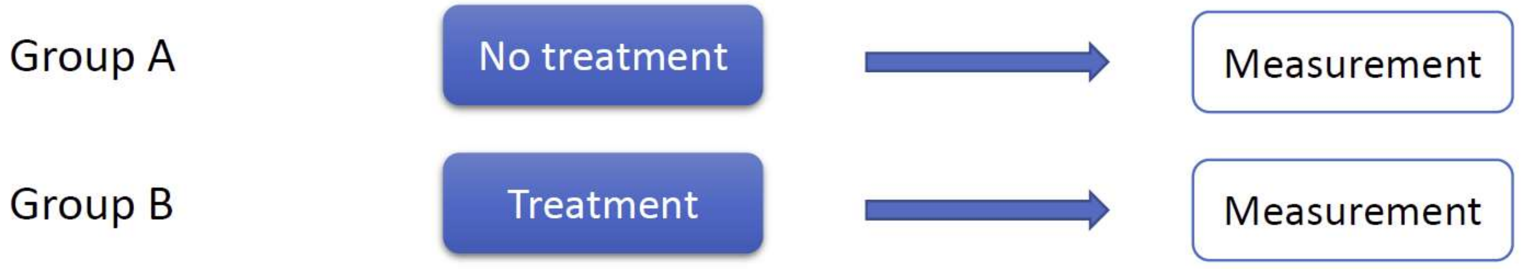
Statistical analysis of simple between-groups experiment

An unpaired t-test could be applied to test for a difference of means between groups.

In general, a difference of population means could be due to the presence of an unobserved confounder.

However, by conducting an experiment with a randomized intervention we conclude the difference of means between the dependent variable Y between the two groups was caused by the difference in X!

- Note that the behaviour of confounder Z is the same between the two groups (randomized)



Experimental designs vs. Observational studies

Advantages:

- Experimental designs enable us to make clear inferences about causal effects!

Disadvantages:

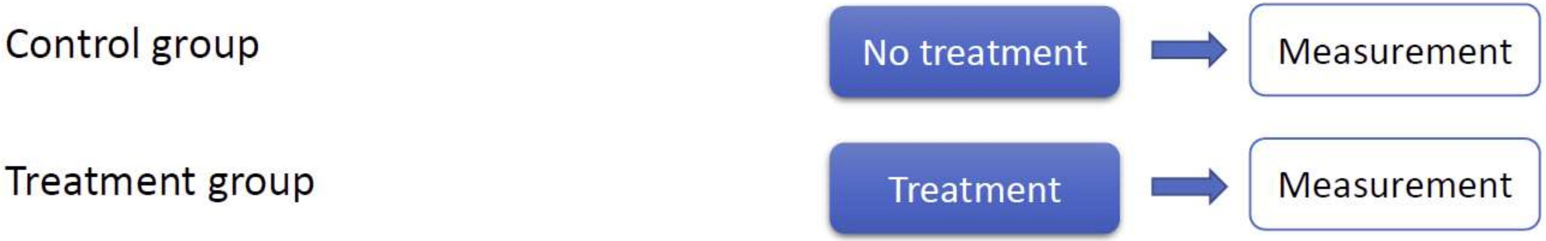
There are many situations where performing a randomized intervention would be

- Unethical: Testing the effect of drug addiction by insisting people take addictive substances.
- Impossible: Testing the effect of species (X) on the speed of running (Y) by randomly assigning an animal to a new species! (does not make sense)
- Too expensive: Testing the psychological effects of driving a sports car by randomly giving people sports cars, which are expensive.

Even when experimental data is available it is typically far more expensive than observational data.

2. pre-test/post-test control group design

The simple between-groups experiment is often referred to as a post-test only control group design

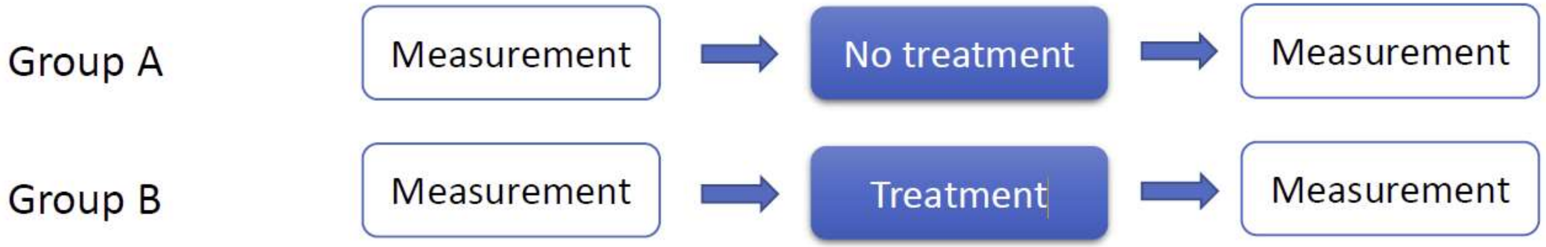


Control group: No treatment → Measurement

Treatment group: Treatment → Measurement

<u>Problem</u>: What if there were a large pre-test difference between the two groups?

An alternative is the pre-test/post-test control group design:

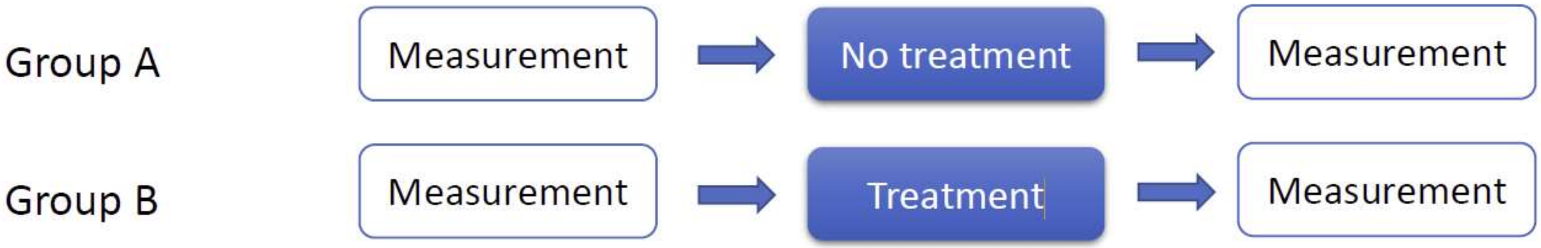


Group A: Measurement → No treatment → Measurement

Group B: Measurement → Treatment → Measurement

pre-test/post-test control group design

An alternative is the pre-test/post-test control group design:



Group A

Measurement → No treatment → Measurement

Group B

Measurement → Treatment → Measurement

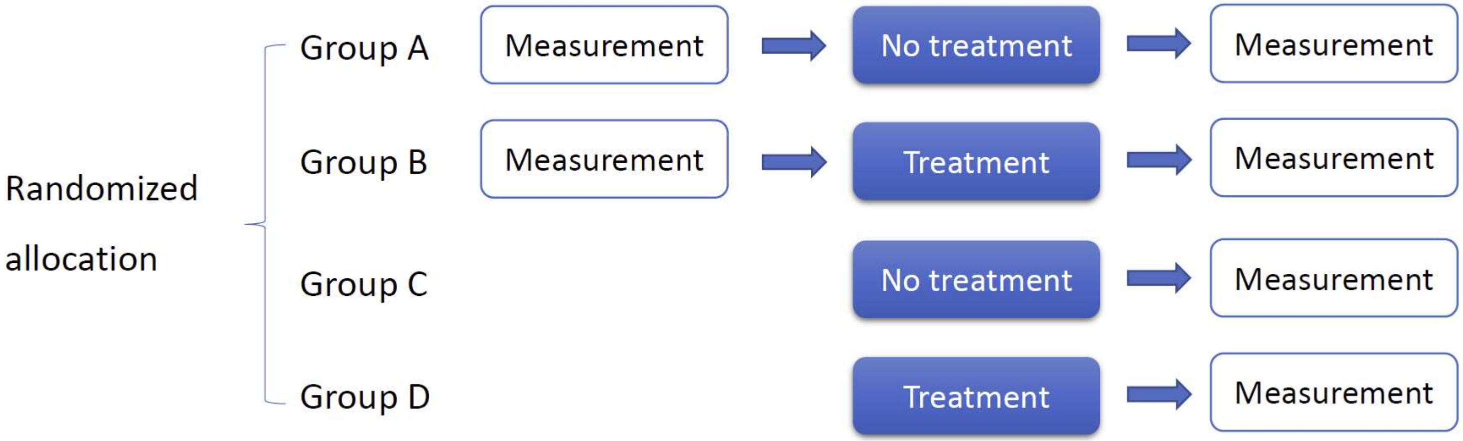
Problem: The act of measurement may affect the outcome of the trials

Example: Measuring blood pressure might cause participants to be more health conscious

3. Solomon's four-group design

<u>Problem</u>: The act of measurement may affect the outcome of the trials

Solomon's four-group design



Randomized allocation

Group A: Measurement → No treatment → Measurement

Group B: Measurement → Treatment → Measurement

Group C: No treatment → Measurement

Group D: Treatment → Measurement

We can apply pre-tests and assess the effect of pre-testing by comparing A & B with C & D.

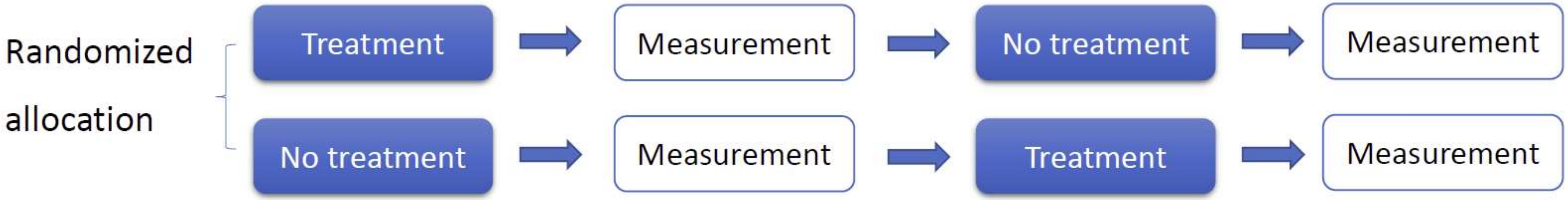
<u>Problem</u>: Requires twice the data!

4. Within-subjects designs

Pre-test only, pre-test/post-test and Solomon’s four-group are all between-group designs.

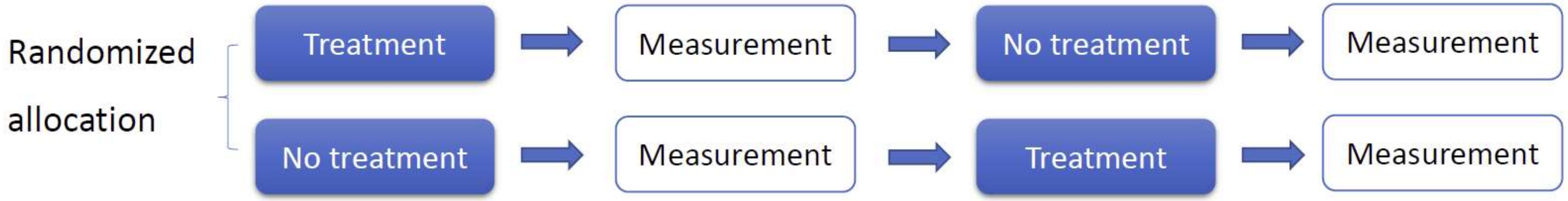
Between-groups (a.k.a. independent measures): Each individual receives at most one treatment

Within-subjects (a.k.a. repeated measures): Each individual receives multiple treatments



Statistical analysis: This produces paired data so we can often apply a paired t-test, for example, a paired t-test

Within-subjects vs. between groups designs



Advantages

- Within-subjects designs are typically more sensitive as they reduce the role of between-subject variation

- Within-subjects designs are often more cost-effective as typically fewer participants are required

Disadvantages

- There are situations where different treatment conditions preclude each other:

  - e.g. Consider an experiment which compares different techniques for learning to drive (a participant can not learn to drive twice)

- There is also a risk of carry-over effects

  -e.g. Consider fatigue or adaptation in an experiment which measures concentration level with logical puzzles.

What have we covered?

We considered three significant obstacles to valid scientific inference:

1. Measurement distortions
2. Selection bias
3. Confounding variables.

We then considered the role of experimental design in overcoming these challenges.

We considered between-groups design: post-test only, pre-test/post-test and Solomon four-group.

We also compared between-groups designs with within-subjects designs.

Thanks for listening!

Dr. Rihuan Ke

rihuan.ke@bristol.ac.uk

Statistical Computing and Empirical Methods
Unit EMATM0061, MSc Data Science