Fundamentals of Experimental Design

Scientific validity and experimental design

Henry W J Reeve

henry.reeve@bristol.ac.uk

Include EMATM0061 in the subject of your email.

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.

What will we 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.

Three goals for scientific research

Reliable: Scientific research should be potentially replicable by other scientists.

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

Importance: Scientific research should address important questions.

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.

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.

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

Obstacles on the path to valid inference

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

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 puzzles or 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.

<u>Example 2</u>

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 ask someone to score the presentation of their code?

Valid measurements

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

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

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

Examples

1. Miscalibration of measurement instruments e.g. a clock running slightly too fast or slow.

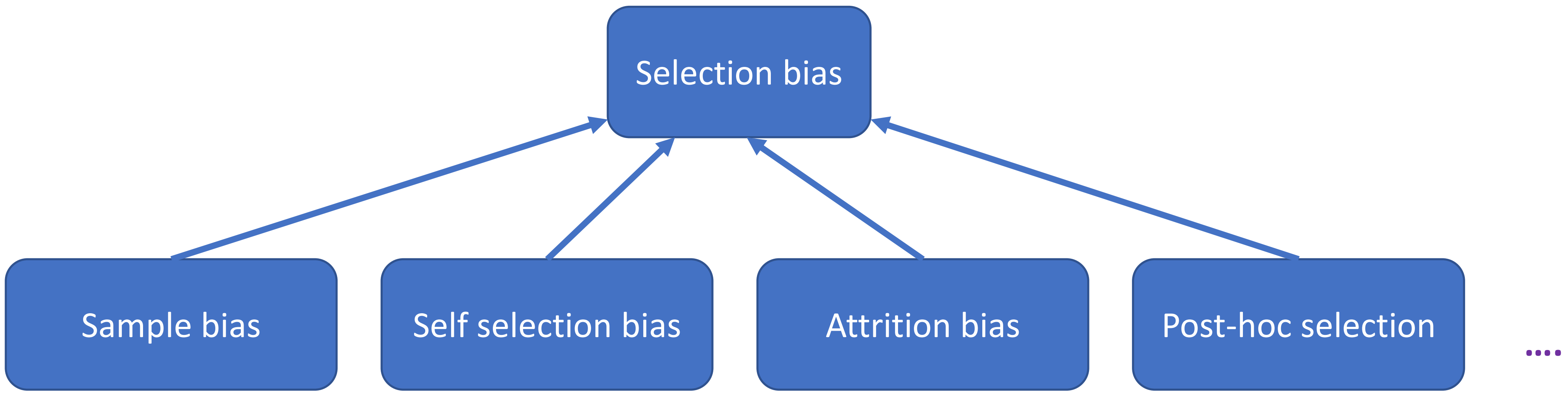
2. Rounding errors due to computational constraints e.g. 0.904 converted to 0.9.

3. Inaccurate responses to questionnaires e.g. under-reporting alcohol consumption.

4. Human error e.g. data entry mistakes.

Selection bias

Selection bias when the data included in the analysis is mis-represents the underlying population of interest.



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

Sample bias

Sample bias 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 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 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.

Examples

- Online reviews of restaurants might disproportionately represent subsets of the population with strong opinions or certain age groups.

- The results from medical trials can be distorted by their over reliance upon student participants.

- 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

Suppose 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 upon the sample itself.

Example

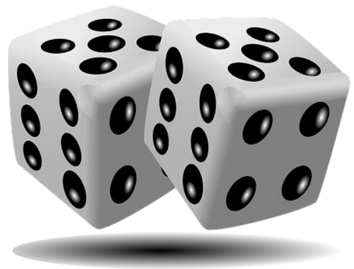
Suppose a scientist is investigating the efficacy of a medical treatment.

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



The ideal solution to selection bias problems is **randomization:**

Data is randomly sampled from the population of interest with uniform weight.

<u>Example</u>:

Our population of interest 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.

Now take a break!

Confounding variables

Correlation does not imply causation!

Sales of

sun glasses

Sales of

ice cream

Confounding variables

Correlation does not imply causation!

Sales of

sun glasses

Sales of

ice cream

Sunny

weather

Confounding variables

Correlation does not imply causation!

Sales of

sun glasses

Sales of

ice cream

Confounding

variable

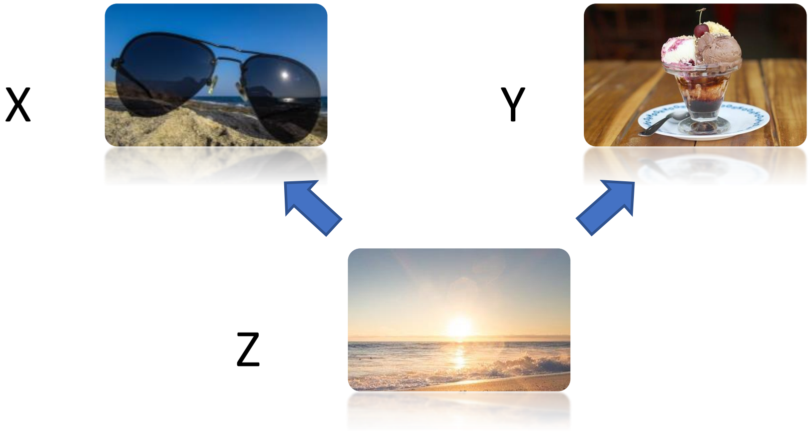
Sunny

weather

Confounding variables

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

A confounder Z is a third variable that has a causal effect upon both X and Y.



Confounding variables

Confounding variables can obscure causal relationships.

Organic vegetables

Longevity

Confounding variables

Confounding variables can obscure causal relationships.

Organic

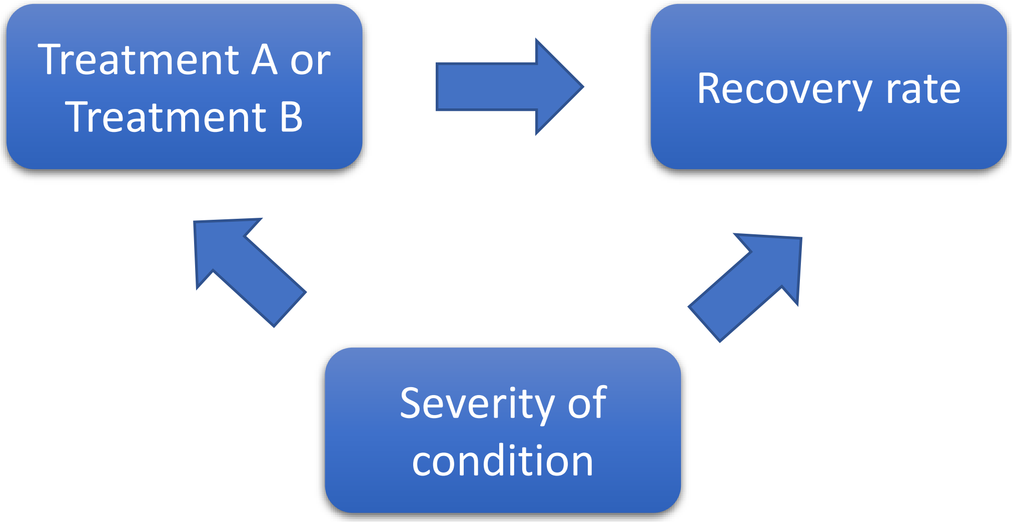
vegetables

Longevity

Wealth

Simpson’s paradox

Confounding variables can obscure causal relationships.



	Treatment A	Treatment B
Less	19/20 (95%)	72/80 (90%)
More	60/80 (75%)	12/20 (60%)
Total	79/100 (79%)	84/100 (84%)

Experimental design

Causation typically cannot be inferred from statistical correlations alone…

Yet, more often causal relationships are what we are most interested in.

Example We want to understand the causal effect of treatment upon a patient’s recovery.

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.**

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

This requires a carefully designed experiment rather than an observational study.

Randomized intervention

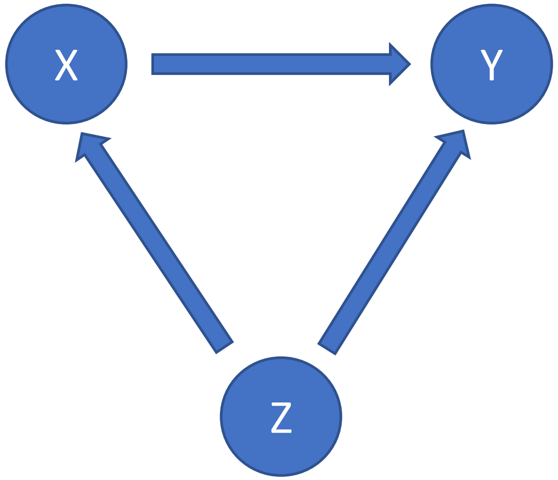
We want to study the causal effect of an independent variable X on a dependent variable Y.

$$X \longrightarrow Y$$

Example The independent variable X is the treatment and the dependent variable Y is recovery or not.

Randomized intervention

We want to study the causal effect of an independent variable X on a dependent variable Y.



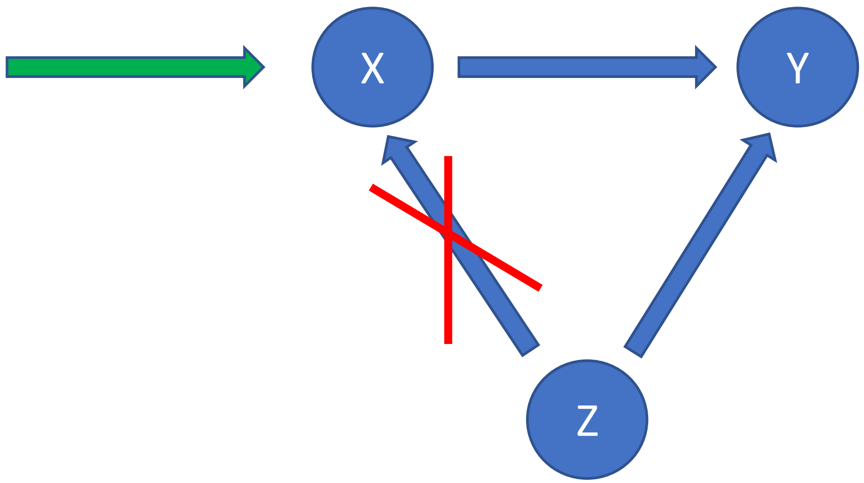
Example The independent variable X is the treatment and the dependent variable Y is recovery or not.

Randomized intervention

We want to study the causal effect of an independent variable X on a dependent variable Y.

Randomized

intervention



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

Simple between-groups experiment

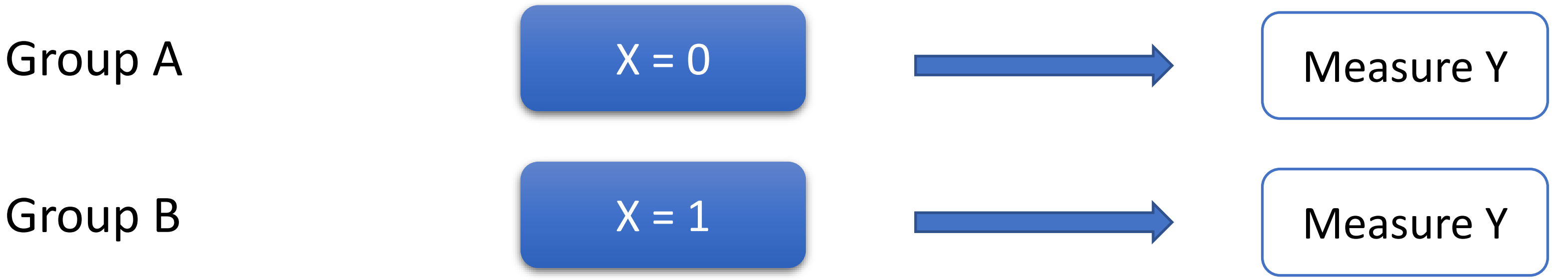
We want to study the causal effect of a binary independent variable $X \in \{0,1\}$ on dependent variable Y.

We take a random sample from our population of interest.

Our sample is then partitioned randomly into two groups – Group A and Group B.

We intervene so those in Group A receive no treatment X=0 and those in Group B receives treatment X=1.

After some time has elapsed the dependent variable Y is measured for those in both groups.



Simple between-groups experiment

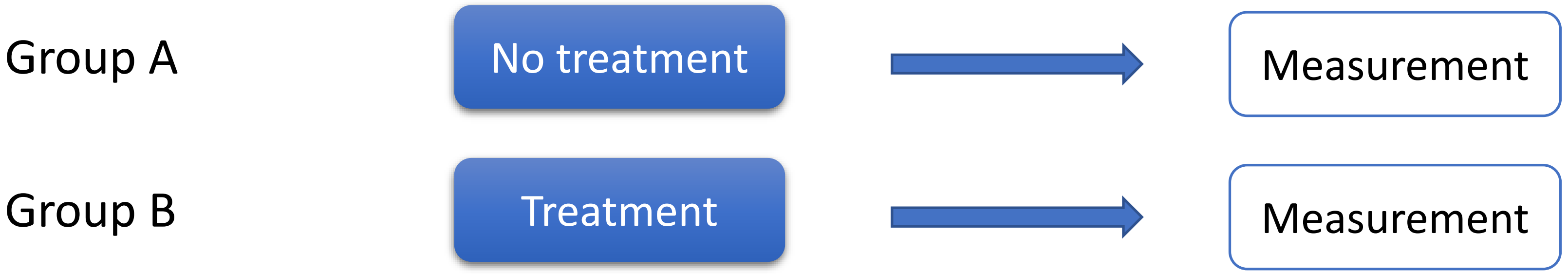
<u>Example</u>

We want to understand the effect of blood pressure medication X on blood pressure level Y.

A random sample is randomly partitioned into a control group and treatment group.

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.

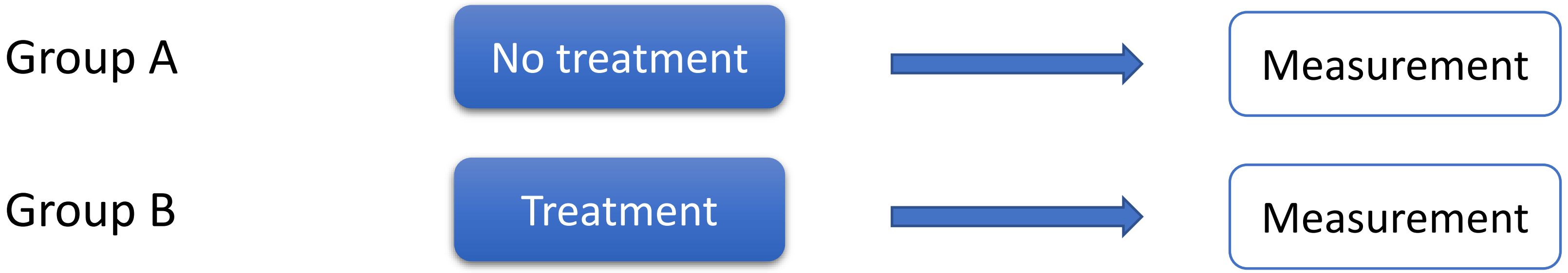


Statistical analysis of simple between-groups experiment

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.

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!



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 on speed by randomly assigning to a new species!

Too expensive: Testing the psychological effects of driving a sports car by randomly giving people sports cars.

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

Pre-test and post test

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

Pre-test and post test

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

Problem What if there were a large pre-test difference between the two groups?

Pre-test and post test

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 and post test

The **pre-test/post test control group design**:

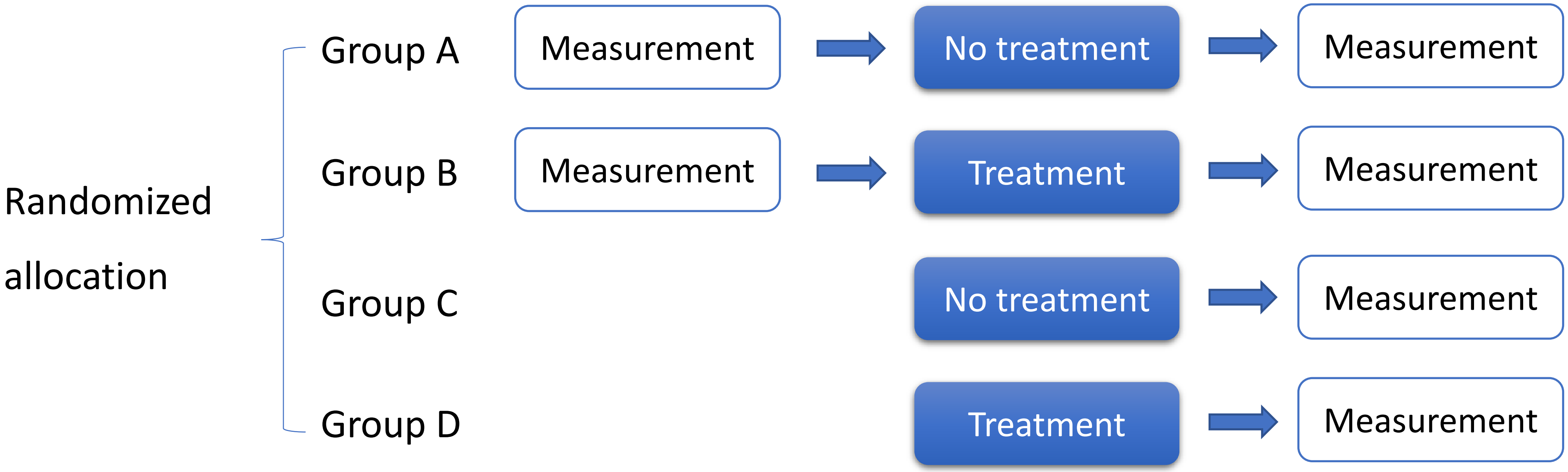
Group A	Measurement →	No treatment →	Measurement
Group B	Measurement →	Treatment →	Measurement

Problem: The act of measurement may effect the outcome of the trial.

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

Solomon’s four group design

<u>Problem</u>: The act of measurement may effect the outcome of the trial.



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

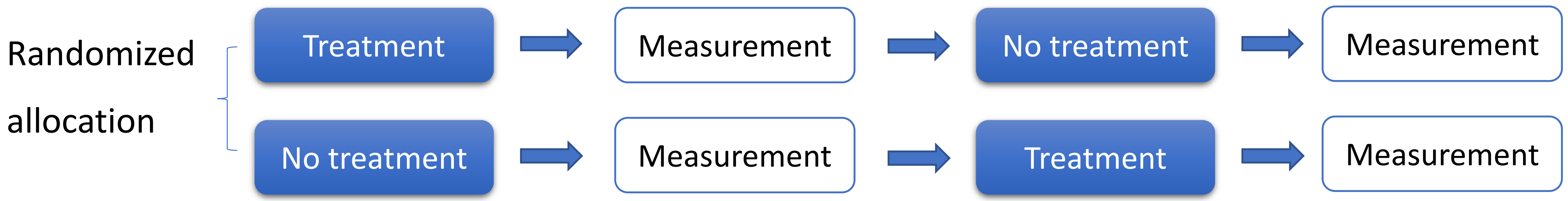
<u>Problem</u> Requires twice the data!

Within-subjects designs

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

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

Within subjects (a.k.a. repeated measures): Each individual receives multiple treatment conditions.



This produces **paired data** so we can often apply a **paired 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 less 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.

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 did we cover today?

- 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!

henry.reeve@bristol.ac.uk

Include EMATM0061 in the subject of your email.