Forum 1
Data Skills for Scientists Data: Navigating Uncertainty

2025-09-24

Welcome!

Who am I?

! Important

Replace this text with a little about yourself, lecturer!

Who to ask for help.

❗ Important

Replace this text with a few slides on where to go to ask questions. Include a brief description of what kinds of questions to bring to who. Include such things as:

- Course coordinator contact details and consulting times
- Online discussion forum (if available),
- Maths Learning Centre (if applicable),
- List of teaching team with photos and method of contacting them,
- Canvas site pages with more information, etc.

In this Forum

We will motivate why study statistics in the first place, and begin on your journey to working with data!

But first, we'll go over some important preliminaries:

- Assessment;
- Weekly Activities; and
- Course structure.

Assessment

Assessment

There are three assessment components in this course:

- Quizzes (10%);
- Report (40%); and
- Exam (50%).

Each contributing a different percentage (shown above) to your final grade.

Quizzes

There are 9 quizzes total.

Only your best 8 out of 9 quiz results contribute to your final grade, each contributing 1.25%. Your worst quiz result does not contribute to your final grade.

Report

The report will be submitted in three parts contributing 10%, 15% and 15% to your final grade respectively. The three parts correspond to the modules of the course. Each part builds on the previous parts to extend the analysis of the same data set.

! Important

Enter information about the due dates of the parts of the report here.

Exam

The exam will have two components

- multiple choice questions similar to the quiz questions, and
- short-answer questions similar to the questions in the practicals and report.

You will not be expected to do any computing during the exam.

The exam contributes 50% of your final grade.

Weekly Activities

Routines

Knowing what you need to do each week, and establishing routines to do the work is often the key to success.

💡 Tip

Review this section of the forum slides and plan out when you'll do each weekly activity. Put it in your calendar on your phone so you get notifications to remind you!

Weekly Structure

Almost every week will have

- a video,
- a forum,
- a practical,
- a quiz, and
- additional resources.

Course Website

Lets take a quick tour of the course website now.

Videos

Provide a brief introduction and summary of the topic for that week.

> **Tip**
>
> Watch the video before attending the forum each week!

If you didn't this time, just watch it afterwards. But remember for next week!

Forums

Discuss the topic in more detail with examples, demonstrations, and interactives!

> 💡 Tip
>
> Take notes during the forums, and spend some time after each forum summarising what you've learnt. This will be useful later!

Quizzes

Each topic has a quiz due the following week on

❗ Important

Enter information about the due dates of the quizzes here.

Each quiz can be attempted up to 3 times, and your best result will be recorded.

There is no quiz in the first week.

More on Quizzes

To prepare for the quiz, review the content from the previous topic:

- watch the video,
- review and summarise your notes and the forum slides.

💡 Tip

Use the quizzes as a prompt to review the previous topic and test yourself — go back and review concepts you got wrong on the quiz afterwards

Practicals

In the forums, we will demonstrate analysing data using the programming language R. The practicals will step you through how to do it yourself!

> **Tip**
>
> Attempt practical activities before class!

Programming can be intimidating and at times frustrating, particularly at first.

> **Tip**
>
> Remember, getting stuck and then seeking help is an important part of the learning process!

The First Two Practicals

Are different to the rest. They include extra steps and explanations to help with your first steps in computing. This is why they are long, but try not to be too intimidated!

> **Tip**
>
> Come to your practicals prepared with questions to ask based on where you got confused to make the best use of class time!

The first topic doesn't have a corresponding practical. The first practical is an introduction to computing in R. The second topic is split accross the second and third practicals. After that, each topic has a corresponding practical in the following week.

Outside of Class

Each week, you are expected to spend around 3 hours studying for this course outside of class, not including working on assessments. This includes:

- Watching the video,
- Summarising what you learnt in the forum,
- Working through the practical, and
- Reviewing additional resources.

> **Tip**
>
> In the first two weeks you may need to spend more time working through the practicals in particular!

Course Structure

Data Skills for Scientists

This course is structured into 3 modules, each of which contains 3 topics. The three modules are:

- Data;
- Inference; and
- Regression.

One topic is introduced each teaching week, and has associated activities in the following teaching week.

So the the total of 9 topics are introduced over the first 9 teaching weeks, with the activities for the 9th and final topic in the 10th and final teaching week.

Data

The 3 topics in the first module, ‘Data’ are:

- Navigating Uncertainty;
- Summarising and Visualising; and
- Probabilities and Distributions.

Navigating Uncertainty

The remainder of this forum will cover the first topic, in particular:

- How (and why) data is used in science;
- Some examples of common biases and paradoxes;
- Types of data and variables; and
- Experimental and observational study design.

This is a good point to take a quick break!

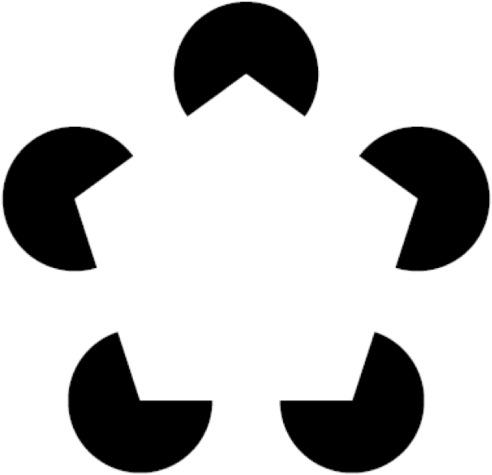
Data versus Information

Why Study Statistics?

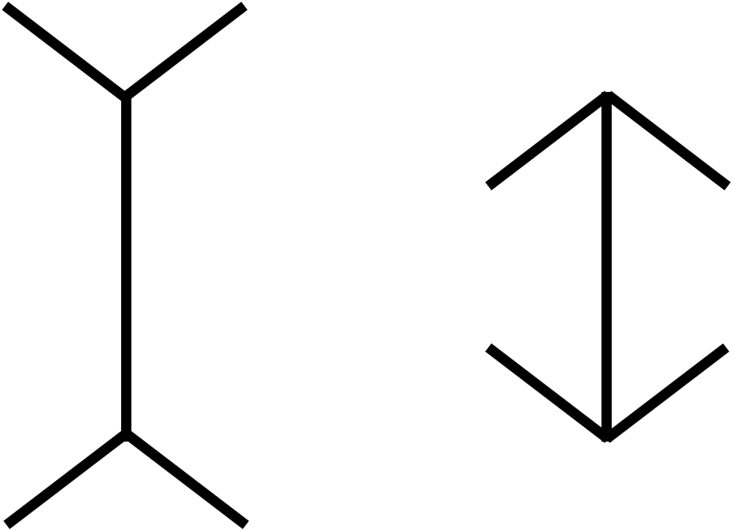
- Our judgement is more subjective than we think.
- We believe we can be objective, but our intuition is often flawed.
- Our intuition is particularly bad at accurately assessing uncertainty.
- Statistics provides objective methods for estimating uncertainty

So we we can use statistics combined with critical thinking to carefully navigate uncertainty and attempt to avoid being misled by our own flawed intuition.

Perception Bias - Illusions

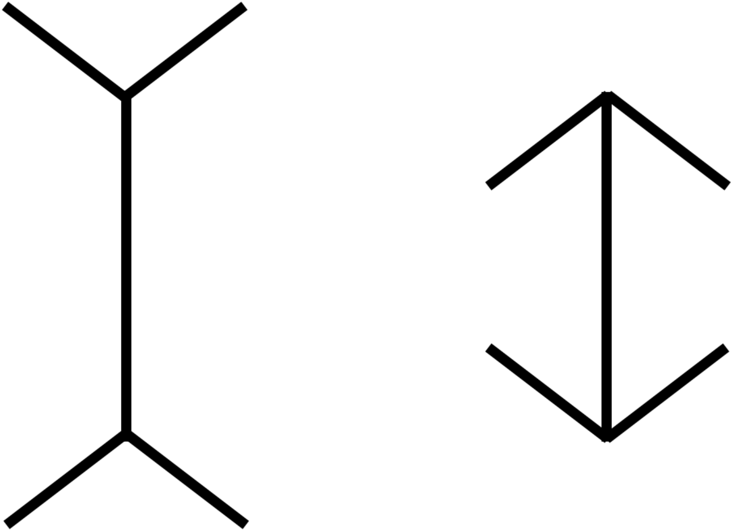


Perception Bias - Illusions



Perception Bias - Illusions

Perception Bias - Illusions



Subjective vs Objective

In order to avoid falling into perception and other biases, it is valuable to learn how to seperate between the subjective and the objective in what we observe.

In the illusion examples:

- The subjective observation is that the line in one image appears longer than in the other.
- The objective observation would involve measuring the line, for example by counting the pixels or using a ruler.

The subjective observation is not wrong *per se* in that it is a valid subjective perception. It does include in it cognitive perception biases however, and so distinguishing it from objective observations is useful when making conclusions.

The terms *information* and *data* are used to make a similar and useful distinction as well.

Information

We see information every day — almost anything is information. It often includes either having been calculated from data and/or some interpretation and context.

For example:

- More respondents answered ‘yes’ to the survey question than expected,
- The patient shows signs of mild hypertension,
- Heart rate trents suggest recovery post-surgery.
- The average temperature this week was 34.5°C

Data

Data, in contrast to information, are measurements or observations that have been directly made, and does not include any interpretation or context.

For example:

- Survey responses: `"yes", "yes", "no", "yes", ...`;
- Systolic blood pressure (mmHg): `120, 130`;
- Heartrate (bpm): `82, 85, 90`.
- Temperature (34.5°C): `34, 35, 32.5, ...`;

Information versus Data

Both data and information are important.

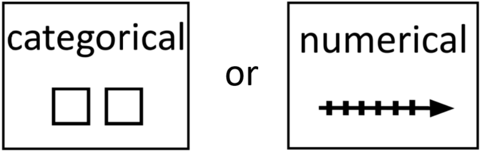
- Data needs associated information in order to be useful to inform decisions.
- Information needs data in order to have substance and to be justified.

Good reporting includes both and clearly distinguishes between the two, highlighting where interpretation and context is being incorporated into information, and including appropriate disclaimers and qualifications alongside conclusions.

Variables[1]

Data is recorded by making measurements or observations, called the values, of **variables**.

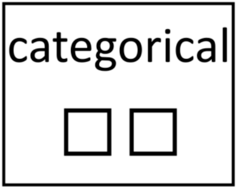
Variables are either



- **Categorical** or qualitative variables are often recorded as words, and how far apart the values are has no meaning.
- **Numerical** or quantitative, interval, or scale variables are recorded as numbers and how far apart the values are does have meaning.

[1] Images are from 'Making sense of statistics' by Dr. David Butler, you may find the accompanying Prezi helpful in reviewing these ideas.

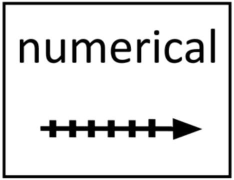
Categorical Variables



Can be further sub-divided into

- **Ordinal** variables whose categories have a natural ordering, e.g. Olympic medals, and
- **Nominal** variables which do not, e.g. colours of cars

Numeric Variables



Can also be further subdivided into

- **Discrete** variables which take one of a discrete set of numeric values, often a count e.g. number of cases of Ebola in a year, the number of heads from tossing a coin 10 times, and
- **Continuous** variables which can take any value in a continuous interval of values, e.g. temperature, time, weight, length.

Examples of Variables

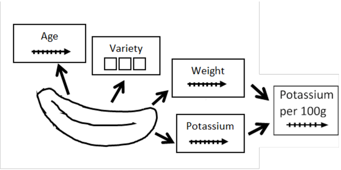
What type of variables are these?

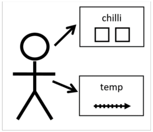
Categorical/Ordinal, Categorical/Nominal, Quantitative/Discrete, Quantitative/Continuous?

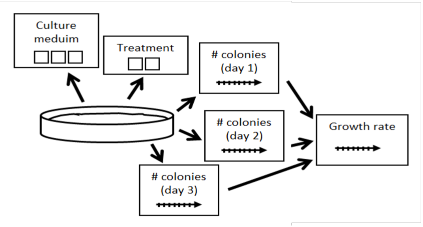
- Concentration of a drug in a blood sample.
- Number of red cars that pass an intersection in an hour.
- Flavours of ice-cream.
- Grades for a subject: F, P, C, D, HD.

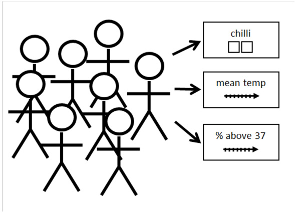
Subjects

Are the source of variables, they are the objects on which you make measurements or observations when you collect data.









Subjects versus Variables

In describing data:

- **Subjects**, also called individuals, cases, observations, etc. are the objects being measured and observed, while
- **Variables** are the characteristics of the subjects being measured and observed.

Sometimes, additional variables that are not part of the data can be calculated from the data. These are still variables, but they are not data, they are information!

Rectangular Data

There are many different forms of data, but in this course we will only deal with rectangular data.

In rectangular data each variable is recorded for each subject.

This means the data can be arranged into a rectangular array, with

- each row corresponding to a subject, and
- each column corresponding to a variable.

Example - palmerpenguins

Below is a subset of the palmerpenguins data[2] as an example

species	island	bill_length_mm	bill_depth_mm
Adelie	Torgersen	39.1	18.7
Adelie	Dream	43.2	18.5
Gentoo	Biscoe	50.5	15.9
Chinstrap	Dream	50.6	19.4

[2] Read more about the palmerpenguins data set

Example - `palmerpenguins`

In example shown,

- subjects are the penguins, and
- variables are the `species` of the penguins, which `island` they were observed on (both categorical, ordinal), the `length` and `depth` of their bill, measured in mm (both numerical, continuous).

Two-way Table

The two categorical variables `species` and `island` can be summarised in a two-way table also known as a contingency table:

	Adelie	Chinstrap	Gentoo
Biscoe	44	0	124
Dream	56	68	0
Torgersen	52	0	0

Aggregate data

This is an example of aggregate data, as it aggregates the `pamerpenguins` data by counting how many penguins are in each combination of `species` and `island`.

It is considered information, as it summarises and is calculated from the `pamerpenguins` data.

It can also be considered data, but to do so we change what is considered subjects and variables:

- subjects would be the 3 islands and
- variables would be the number of penguins of a particular species observed (all numerical, discrete).

So, Why Study Statistics?

So far, we've discussed how

> *It helps structure our reasoning to seperate subjective from objective, and be mindful of bias when interpreting information and data.*

But even if we only consider data that consists of objective measurements, there is still always some uncertainty associated with it, and

> *It also provides a structure for dealing with uncertainty.*

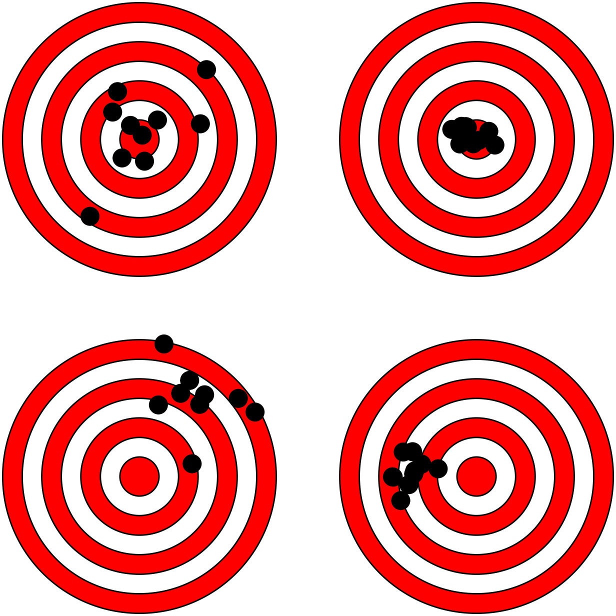
It can be useful to separate the uncertainty associated with data into two categories: accuracy and precision.

Accuracy and Precision

Accuracy is closely associated to validity, and can be interpreted as the opposite of bias. It describes how well data represents the thing is it trying to measure on average.

Precision is closely associated to reliability, and can be interpreted as the opposite of variability. It describes how consistently the data reproduces similar values in similar situations.

Accuracy and Precision



Reproducibility

Reproducibility is not only an important element of scientific inquiry in terms of the precision of measurements, but it is also important in terms of the reporting of results!

> **! Important**
>
> This would be a good place to add a live-coding demo on using quarto and explaining its value in science for reproducibile reporting.

Variable versus Variable

The word “variable” has a very specific meaning in statistics, as described earlier.

However in computing, the same word “variable” has a completely different meaning — it refers to a value stored in the memory of a computer that can be referred too by its name.

> ⚠ Warning
>
> This is often the case that the same word has different meanings in different contexts, so beware of the confusion!

This is a good point to take a quick break!

Lurking Variables

Lurking Variables

When studying the relationship between variables, there may exist **lurking variables** not included in the study that would dramatically change the nature of the relationship between the variables being studied if it was taken into account.

⚠ Warning

Failing to account for lurking variables can lead to misleading conclusions regarding the relationship between the two variables in question!

Simpson's paradox

When dealing with categorical variables, a lurking variable creates subgroups, and an association or relationship that is consistent for all subgroups can dramatically change and even completely reverse when the data from all the subgroups are combined together!

This reversal is called **Simpson's paradox**.

Example - College acceptance rates

gender	accepted	not-accepted	total
female	88	112	200
male	198	162	360

gender	accepted	not-accepted
female	0.44	0.56
male	0.55	0.45

Example - College acceptance rates

It seems as though men are accepted more often than women!

There is lurking variable here — there are many different departments in the college, for this example we've chosen two:

dept	female	male	total
business	120	120	240
humanities	80	240	320

Example - College acceptance rates

If we just look at the School of Business Studies:

gender	accepted	not-accepted	total
female	24	96	120
male	18	102	120

gender	accepted	not-accepted
female	0.20	0.80
male	0.15	0.85

We see women are accepted at a higher rate than men!

Example - College acceptance rates

If we look at the School of Humanities:

gender	accepted	not-accepted	total
female	64	16	80
male	180	60	240

gender	accepted	not-accepted
female	0.80	0.20
male	0.75	0.25

We see the same thing! Women are accepted at a higher rate than men!

Example - College acceptance rates

So to summarise:

- In the School of Business studies, 20% of women applicants are accepted, but only 15% of men applicants are accepted.
- In the School of Humanities, 80% of women are accepted, but only 75% of men are accepted.
- Overall, 44% of women are accepted and 55% of men are accepted.

How is this possible?!

Example - College acceptance rates

Lurking variable: Applications were split between the School of Business Studies (240) and the School of Humanities (320).

Two factors combine to create Simpson's paradox:

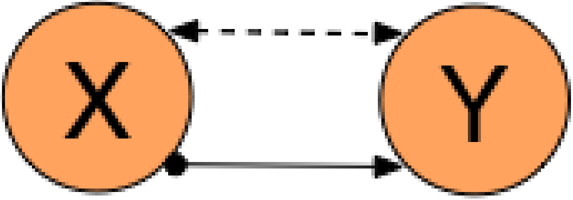
- The proportion men who apply to the School of Humanities is higher than the School of Business Studies, and
- the overall acceptance rate is higher in the School of Humanities than in the School of Business Studies.

Although both schools have lower acceptance rates for men than women, because more men apply to the School of Humanities, which has a higher acceptance rate overall, when the data are combined from both departments it looks as though the college as a whole is biased against women, when this is actually not the case (in this scenario)!

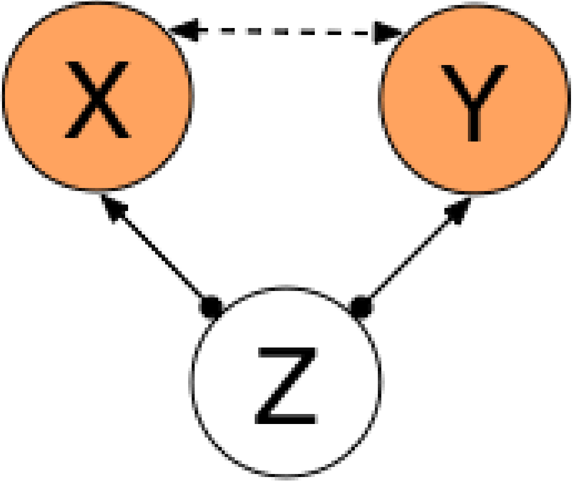
Causation

Association, however strong, does NOT imply causation.

Explaining association



Commmon response

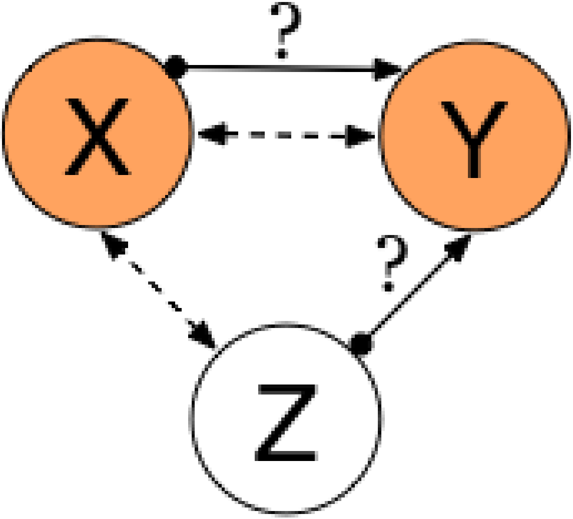


Common response

“Beware the lurking variable’ ’ is good advice when thinking about an association between two variables. The observed relationship between the variables can be explained by a lurking variable. Both x and y may change in response to changes in z.

Example: Ice cream sales and shark attacks. What is the lurking variable?

Confounding



Confounding

Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables.

Example: Smoking and lung capacity. What are some possible lurking variables?

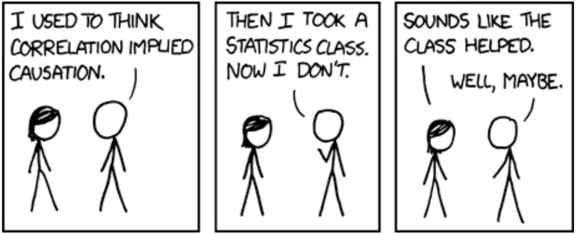
Establishing causation

It appears that lung cancer is associated with smoking.

How do we know that both of these variables are not being affected by an unobserved third (lurking) variable?

For instance, what if there is a genetic predisposition that causes people to both get lung cancer and become addicted to smoking, but the smoking itself doesn't CAUSE lung cancer?

Causation[3]


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

[3] http://xkcd.com/552/

This is a good point to take a quick break!

Study Design

Elements of statistics

Some of the key elements of statistical thinking we will discuss in this course include

- Descriptive statistics,
- Inferential statistics, and
- Study design.

Descriptive Statistics

Summarising and visualising data so as to make the major features of the data more transparent is called **Descriptive Statistics**.

We have touched on this in the discussion of data, variables and Simpson's paradox, but the next forum will be dedicated to exploring this topic.

Inferential Statistics

Drawing and presenting conclusions about a population on the basis of data collected on a sample from that population.

There are many different ways to do inference, depending on what kinds of question you are trying to answer, but they all involve some method of quantifying our uncertainty since any inferential conclusion can never be absolutely certain, there is always some uncertainty and quantifying it is a major role of statistics since our intuition is often very bad at estimating uncertainty.

Inferential Statistics

In the second and third modules of this course we will introduce inferential methods to:

- Estimate an average value of a population;
- Test a hypothesis; and
- Make predictions.

But as a general introduction to the types of questions that can be answered with inferential statistics, you may like to take a look at Dr. David Butlers Prezi slides on the topic.

Study Design

Planning for optimal collection of data given the available resources is critical to being able to adequately address the questions we want too.

Assessing the design used to collect data is also critical in forming conclusions that are appropriate and objectively justified.

Considering possible lurking variables is an important part of study design, and other important factors in study design are what we will discuss for the remainder of this forum.

Observational Study versus Designed Experiments

An **observational study** observes subjects and measures variables of interest but does not actively intervene or assign subjects into different treatment groups. The purpose is to describe some group or situation.

A **designed experiment** deliberately imposes some treatment or conditions on subjects to measure their responses. The purpose is to study whether the treatment **causes** a change in the other measured variables.

Treatments and Response Variable

- **Treatment:** The conditions applied to the subject in a designed experiment.
- **Response variable:** the measurements made on the subjects after they have been given the treatments.

Example - Rats in a maze

To examine learning in rats, 100 rats were timed through a maze, and then rewarded for completing the maze with either bread or cheese. This was repeated several times to allow them to learn, and their final time was recorded.

Is this an observational study or a designed experiment?

What are the subjects?

What are the treatments?

What is the response variable?

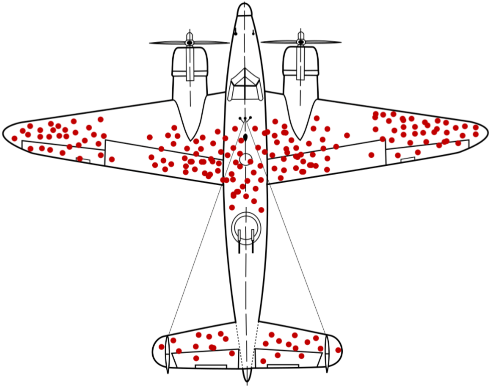
Selection Bias

How subjects are selected to be included in a study is always important, as it informs how results can be interpreted.

Selection bias describes a situation in which some subjects are more likely to be included in a study than others, which can skew results when the subjects that are more likely to be included in the study are not representative of the group being studied as a whole.

WWII Bombers

During world war II, the US military was studying how to reduce aircraft casualties, and mapped bullet holes on aircraft returning from the front. Here is a hypothetical visualization of the damage[4]:



[4] Martin Grandjean (vector), McGeddon (picture), US Air Force (hit plot concept), CC BY-SA 4.0

WWII Bombers

Planes had many bullet holes in the wings and tail, but never the engines.

The US military concluded that the wings and tail are vulnerable to receiving bullets, and armour should be added to these areas.

However the mathematician and statistician Abraham Wald offered a contradictory insight based on his understanding of survivorship bias (a type of selection bias): don't armour the wings and tail, armour the engine instead.[5]

Why?

[5] You can read more about this here.

Surveys

A **survey** is a type of observational study in which participants are asked to report some information.

Selection bias is a critical element of interpreting survey results, and generally can be thought of as having two elements:

- How participants were selected to be sent the the survey, and
- Which and how many of the participants sent the survey choose to respond. The percentage of participants sent the survey who respond is called the **response rate**.

Example – Student Feedback

At the end of this course a survey will be sent to all enrolled students.

- Only students enrolled in the course at the end of the study period will be sent the survey, students who initially enrolled but then withdrew won't.
- The response rate is often low, many students won't respond. Students with strong opinions (either negative or positive) will often be more likely to respond, while those with moderate opinions will be less likely to respond.

and this influences the interpretation of results dramatically!

How to Assess a Study

When assessing a study, consider the following questions as starting points:

- Is it an experimental or observational study? If observational, is it a survey?
- What are the subjects and variables, and how are the variables measured?
- What are some of the most likely sources of bias that could be invovled?

Note on Observational studies

Recall that an observational study is one where the subjects assign themselves to the treatment or control.

The main issue is often: “was the control group really similar to the treatment group - apart from the exposure of interest?’ ’

Summary

Summary

After this forum, you should be able to:

- Know what is expected of you in this course,
- Distinguish between data and information,
- Identify the subjects and variables in a study,
- Identify if a variable is numeric (and discrete or continuous) or categorical (and ordinal or nominal),
- Distinguish between accuracy and precision,
- Identify various types of bias, including perception bias, selection/ observation bias, and recognise Simpson's paradox.
- Identify potential lurking and confounding variables.
- Distinguish between experimental and observational studies and identify surveys.

Additional Detail on Experimental Design

Principles of Experimental Design

Are:

- Control,
- Randomisation,
- Blocking and
- Replication.

Control

The first principle of experimental design, control, refers to controlling unwanted sources of variation in the response variable.

Method of Comparison

Statisticians use the **method of comparison**. They want to know the effect of a treatment on a response. To achieve this they compare the responses of a treatment group with a **control** group.

The control group does not get the treatment of interest. It gets a placebo or the standard treatment.

Method of Comparison - Causality

This element of experimental design — controlling which subjects get the treatment of interest and which get the control treatment — is what distinguishes a designed experiment from an observational study, and it is what allows for conclusions about causality to be inferred.

Control - Rats in a maze

In the rat study final time to complete the maze is compared between rats that are given bread (the control group) and those that are given cheese (the treatment of interest).

Placebo

A placebo is a treatment that looks like, tastes like, smells like the active treatments, but does not contain any active ingredients.

Sometimes, subjects that take a placebo will experience an effect to the response variable compared to a control who don't take any treatment, and this is called **the placebo effect**.

Blinding

To compensate for the placebo effect, and other confounding factors, controlled experiments are often blinded — meaning that subjects are not told which treatment they are being given.

This is not always possible (for example in the rat study!) since the treatments must be indistinguishable to the subjects in order to be able to blind the study.

Double and Triple Blinding

An experiment can also be double-blinded, where not only the subjects do not know which treatment they are being given, but the experimenters also don't know which treatment they are giving.

The gold-standard in clinical trials is the triple-blind study, which is not only double-blinded, but in addition the data analysts are also blinded.

Randomisation

Randomisation refers to the method by which subjects are allocated to treatment groups, specifically — randomly.

Randomly allocating subjects between the treatment of interest and a control group produces two groups of subjects that we expect to be similar in all respects before the treatments are applied.

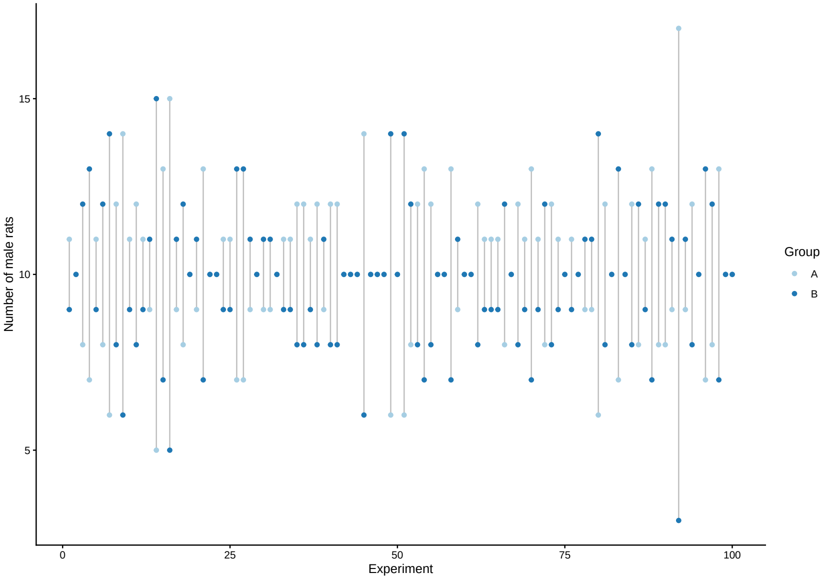
- It is used to prevent **selection bias**.
- It also aims to average out lurking variables.

Example

Consider a population of 60 rats where 40 rats are female and 20 rates are male. So $1/3$ of the rats are male.

We allocate 30 rats to Group A and 30 rats to Group B.

Example



Blocking

A **block** is a group of subjects that are known before the experiment to be similar in some way that is expected to affect the response to the treatments.

In a **block design**, the random assignment of subjects to treatments is carried out separately within each block.

Blocking — Fertiliser

To test different fertilisers it may be desirable to block by soil type: for clay soil, apply the two different treatments (fertilisers), and then also do so for sandy soil, etc.

If only one soil type is used, this would be an example of control, rather than blocking, but this would also limit the interpretation of results to that soil type only.

Blocking — College acceptance rates

In the Simpson's paradox example earlier, separating out the analysis of acceptance rates to consider each department separately is an example of blocking in an observational study.

Blocking — Twin studies

Twin studies, in which one identical twin is allocated to the control group and the other to the treatment group, is an example of blocking, since identical twins share almost exactly the same genes.

Blocking - Matched Pairs

A **matched pairs design** is a randomized blocked experiment in which each block consists of a matching pair of similar experimental units.

A twins study is an example of a matched pairs design, and in this case which twin gets allocated to which treatment would be randomised.

Blocking - Matched Pairs

Sometimes, a "pair' ' in a matched pairs design consists of a single subject that receives both treatments.

In some situations the order of the treatments can influence the response and so in which order the subjects get the treatments can be randomised.

In other situations, the two measurments might correspond to a "before" and "after" measurement, where subjects are given some intervention between the two measurements.

Replication

We aim to have enough subjects in each group to reduce chance variation in the results.

An experiment is called **balanced** if it has the same number of subjects in each treatment group.

General guideline

Overall, the main aims in experimental design are to:

> *To make the treatment groups as similar as possible except for the treatment:*
>
> *Control what you can, block what you can't control, and randomize to create comparable groups.*
>
> *Then, collect as much data as possible so that effects from variables you can't control or block can be averaged out.*