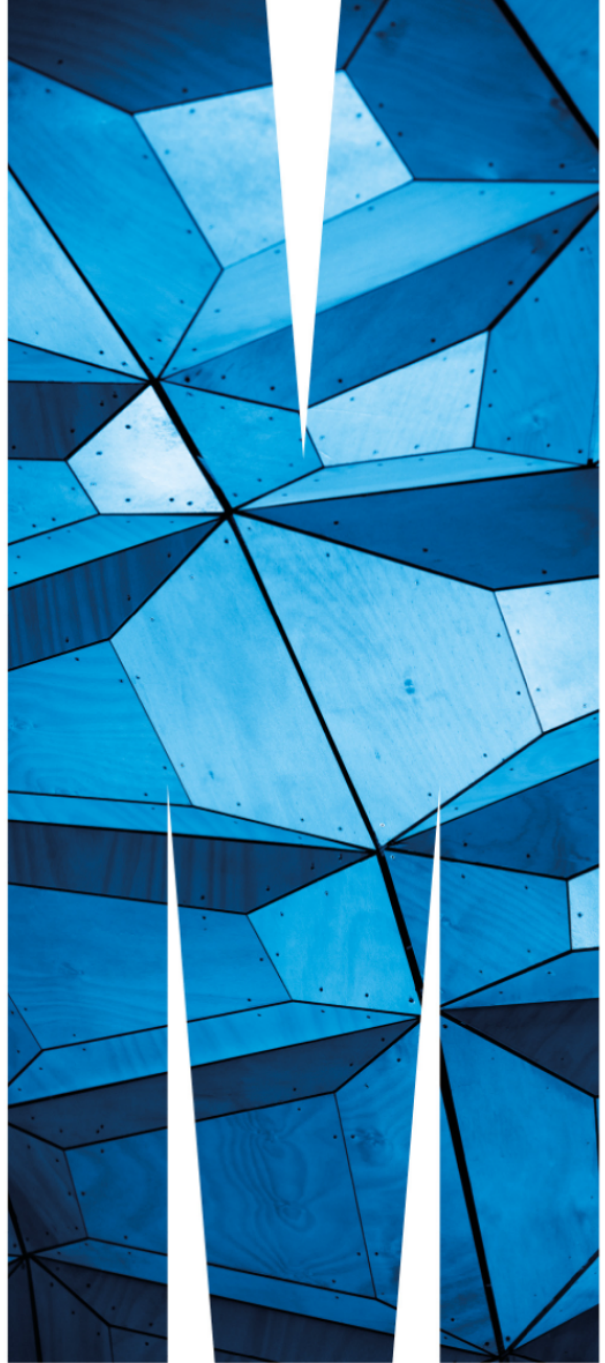
MONASH University

ETC5512: Wild Caught Data

Introduction to data collection methods



Lecturer: *Kate Saunders*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

Week 2

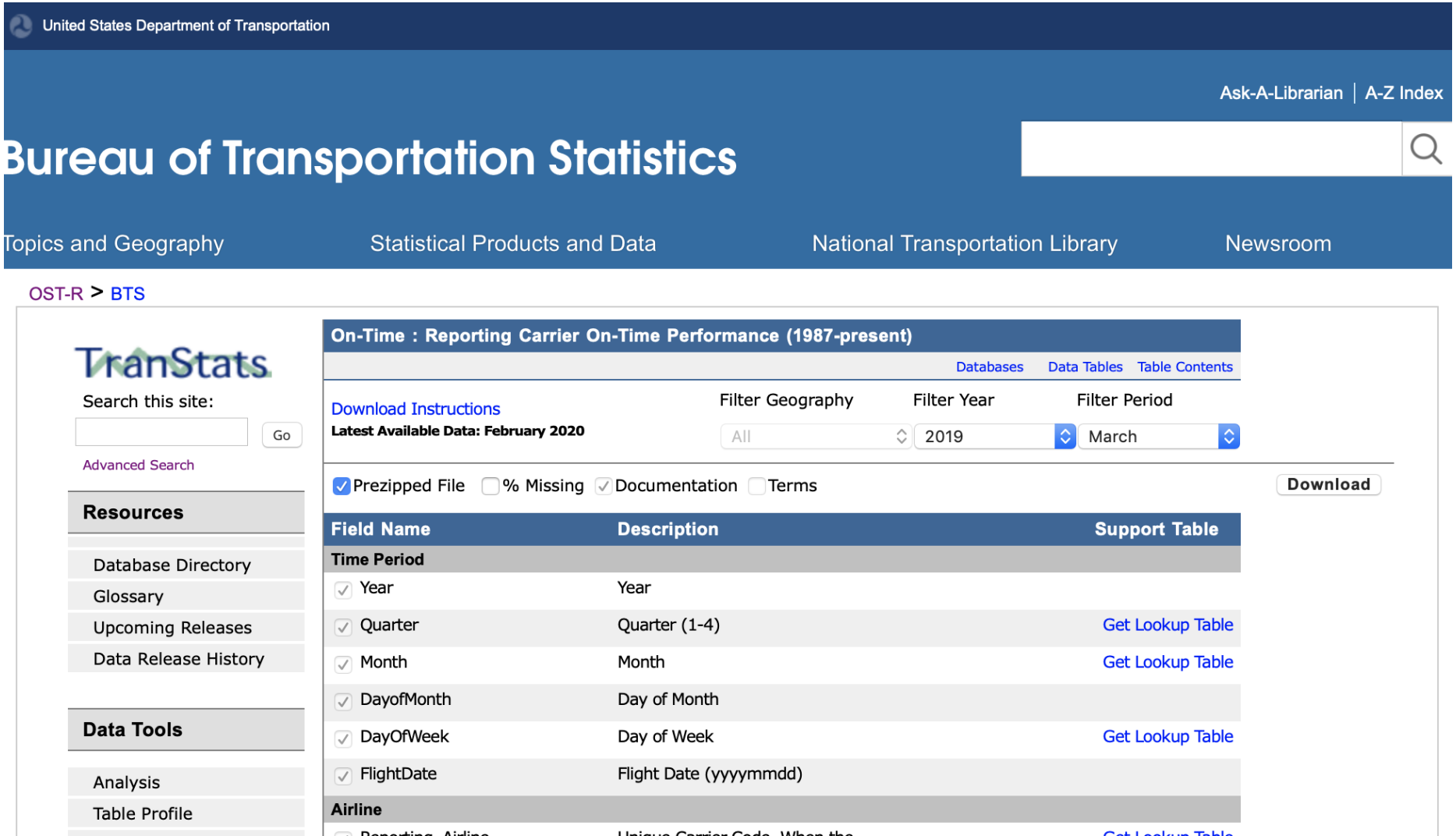
Starts with curiosity

which is

tempered with skepticism

What would you be curious about ... ?

- ... about air travel in the USA?



What would you be curious about ... ?

- ... about people in Australia?

Skip to main content
Archived content. See ABS Website for latest information and statistics

Australian Bureau of Statistics
Search for: Submit search query:
MENU

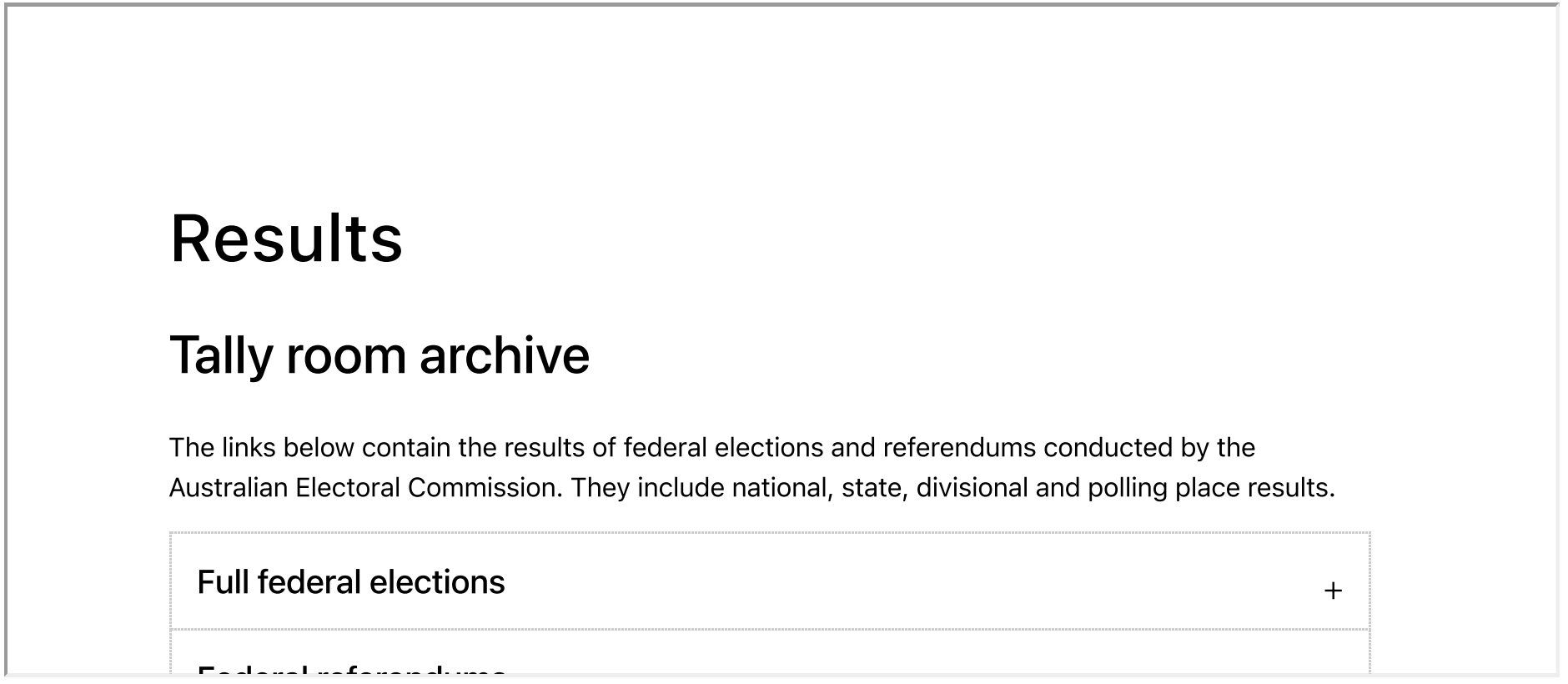
- Statistics
- Census
- Participating in a survey
- About

> By Catalogue Number

2916.0 - Census of Population and Housing - QuickStats, Community Profiles and

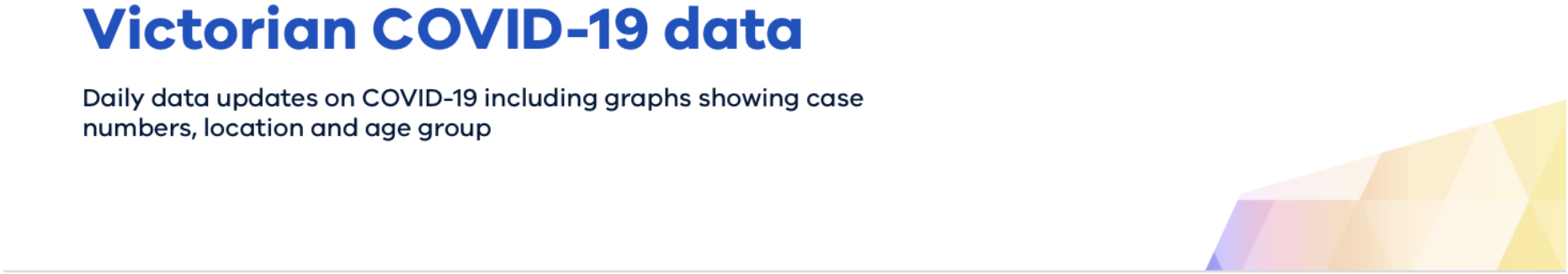
What would you be curious about ... ?

- ... about how people vote in Australia?



What would you be curious about ... ?

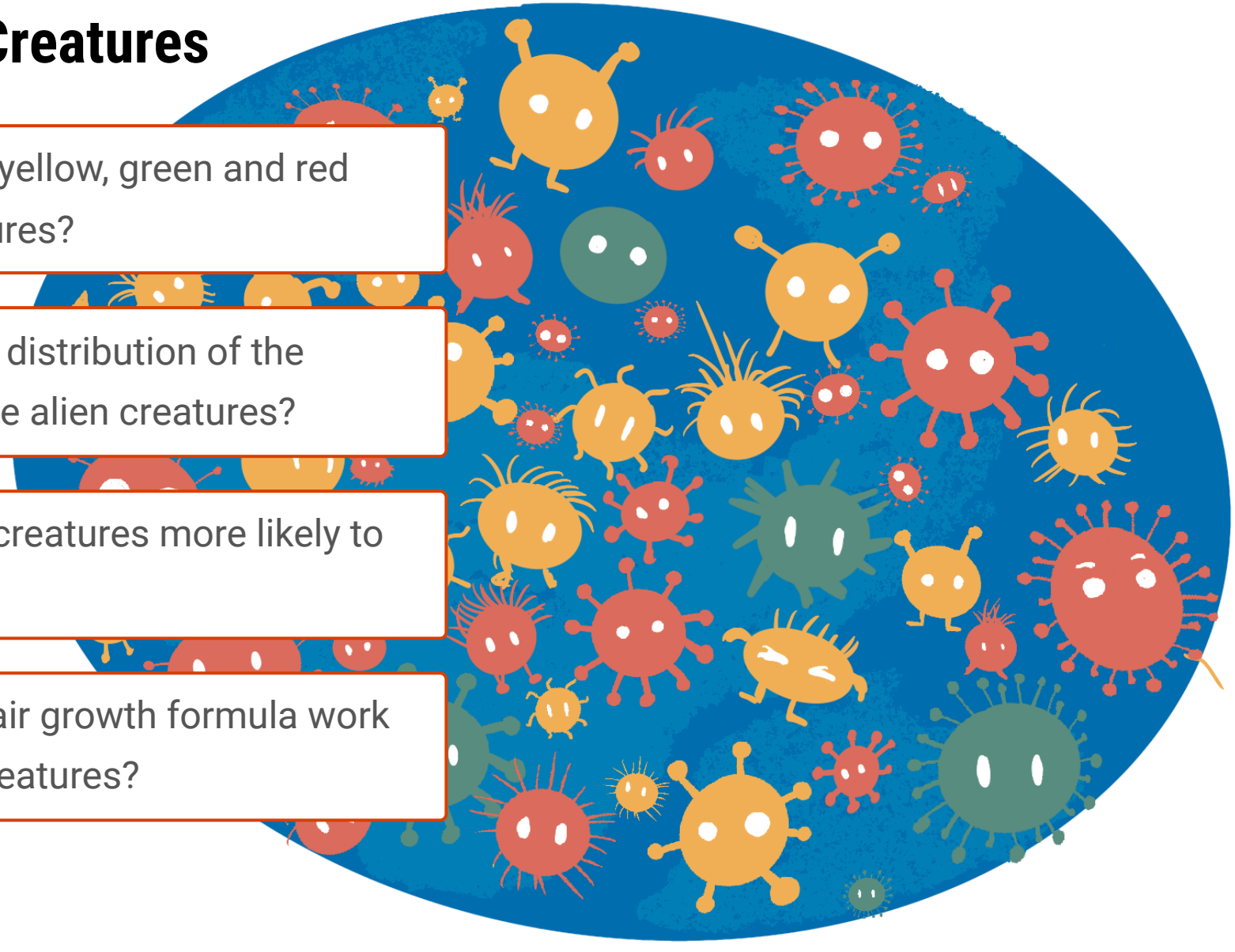
- ... About COVID in Victoria



Curiosity usually corresponds to asking questions

Planet Cute Creatures

- How many yellow, green and red alien creatures?
- What is the distribution of the height of the alien creatures?
- Are yellow creatures more likely to have hair?
- Does the hair growth formula work on these creatures?

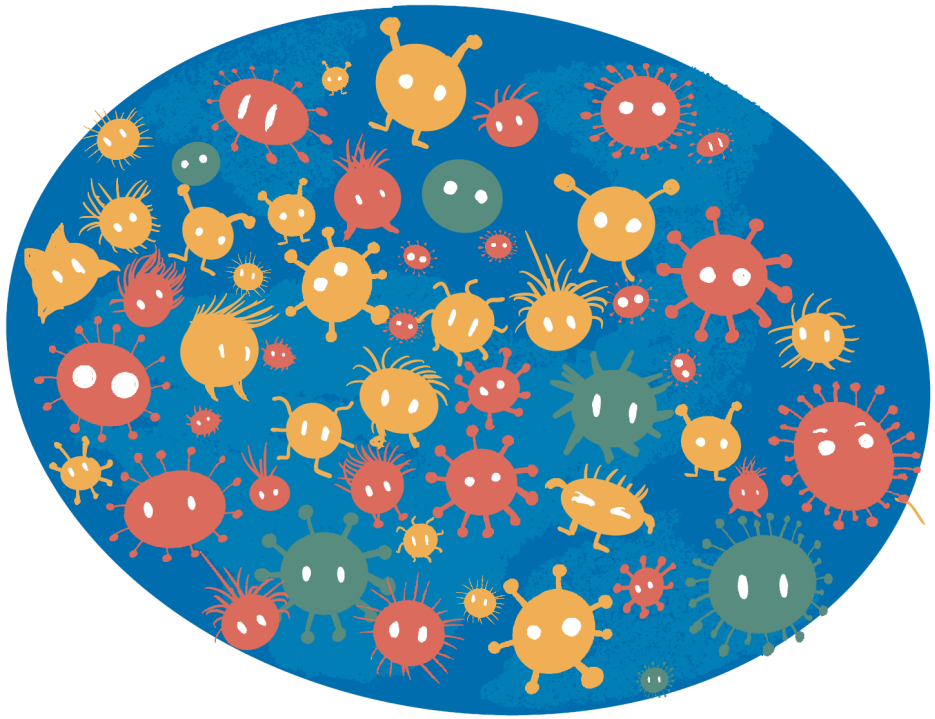


Now that we have a question ...

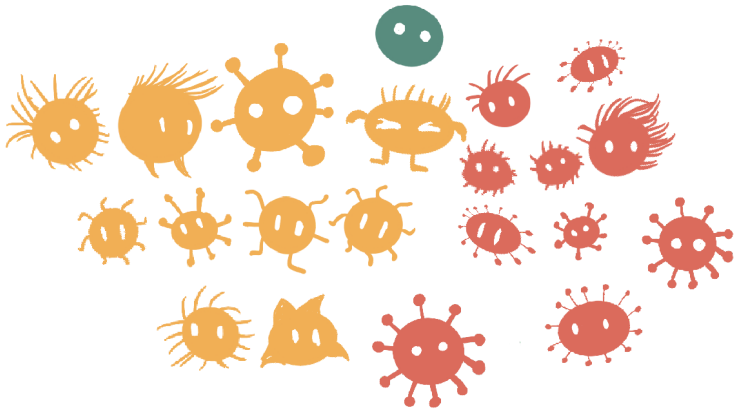
…we can find/use/collect data to obtain answers.

How do we get the data?

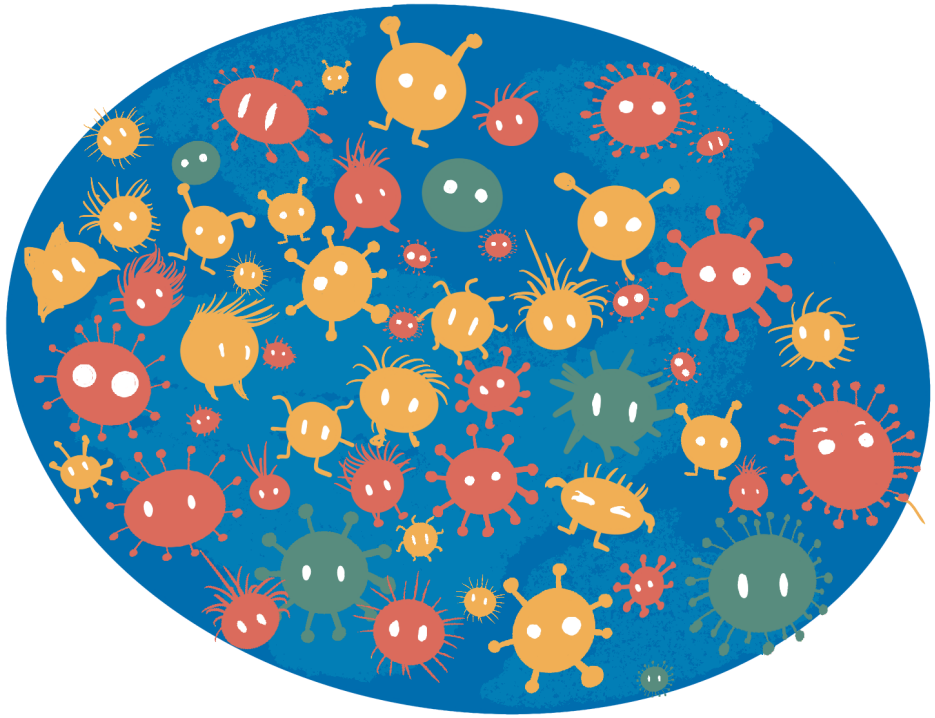
What is the population?



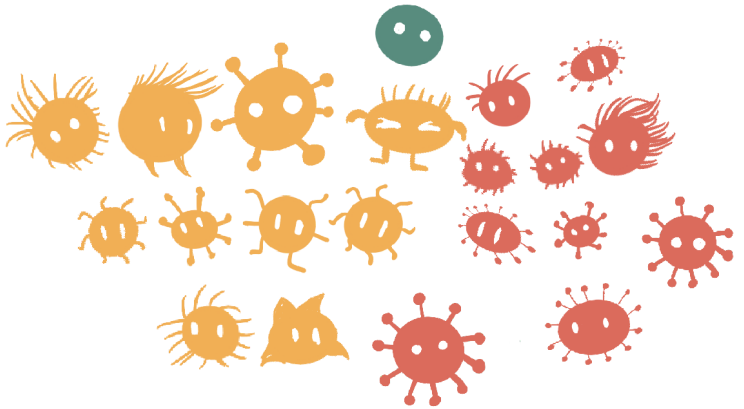
It is rare to have resources to measure ALL of the population, take we a sample



Population:



Sample:



Parameters (typically don't know the values):

Green: 6/53=0.113, Yellow: 22/53=0.416,
Red: 25/53=0.472

The statistics:

Green: 1/21=0.048, Yellow: 10/21=0.476,
Red: 10/21=0.476, are estimates of parameters

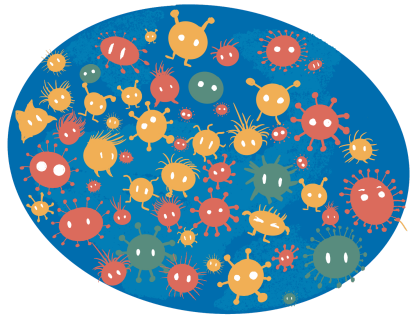
Sampling the population

Collecting data on the entire population is normally too expensive or infeasible! (If we can, call it a census.)

- We therefore collect data only on a subset of the population.
- **How should we sample the population?** There are many sampling schemes.

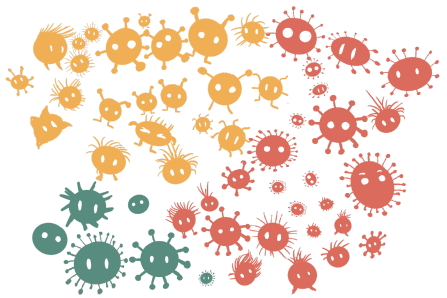
Simple random sampling

Every unit in the population has the same sample probability to be drawn.



Stratified random sampling

Units are drawn from non-overlapping sub-populations.



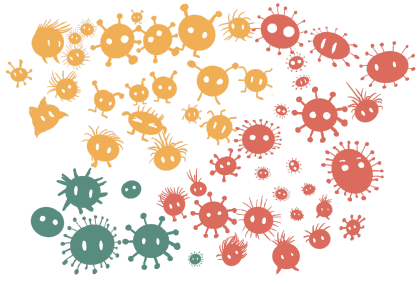
Goal of sampling schemes

The **goal of a sampling scheme** is to get accurate information from the sample in order to answer your question about the population.

- This involves identifying:
  - the **population of interest** (e.g. if studying about male baldness pattern, your population of interest is the biologically male population),
  - what **responses** (dependent variables) or **covariates** (explanatory or independent or predictor variables) to capture and how to measure it (e.g. do you collect their age? Which range of age they are in? Their hair count? The thickness of the hair?),
  - the **sample size** (how many samples do we need?),
  - any **structure** that will be in the data (e.g. population structures, repeated cross-sectional data, panel or longitudinal data), and
  - any **restrictions** (e.g. ethical concerns, limitation on collecting data).

Sampling strategies

- Sampling strategies combine knowledge about the population with statistical methods.
- For example,
  - designing so your sample estimates give (theoretically) unbiased estimates of the population parameters,
  - sample so the data will be representative of the subpopulations (e.g. stratified random sampling), or
  - oversampling or undersampling to compensate for imbalance in classes.



What might go wrong with a simple random sampling of 10 creatures from this population?

Random and non-random selections

- Units (population members) ideally are sampled *randomly*, but often selections are made in a non-random manner.

If I survey every 10th household in a street, is that a random selection?

What do you think can go wrong if we don't sample randomly?

- What's wrong with these examples?

You want to know the attitude of the creatures about working at home.

You call phone numbers listed in the order of telephone directory and stop when you have 20 observations.

You want to get the hair count distribution of the Planet Cute Creatures population.

You sample creatures from the Society of Bald Extraterrestrials.

Reality of data collection ...

- Designing a data collection is ***hard***.
  - There may be unknown or hidden structures in the population.
  - It may add complex structural elements, e.g.
    - Cross-sectional, repeated cross-sectional (e.g. case-control),
    - Panel or longitudinal (e.g. cohort studies), and so on.
    - Clusters or hierarchies (e.g. students in schools in states)
- You may have introduce unintended or unknown structures in the data, e.g. confounded variables.
- It's further complicated by:
  - Non-response,
  - Missing data,
  - Mis-measured data,
  - Dropouts, and censoring,
  - ..... 😱

Observational studies

- Sampling from a population typically yields data considered to be an **observational studies**. Almost all open data are from observational studies.

- An **observational study** aims to draw inferences about a population from a sample where independent variables are *not* intentionally allocated to units within the sample for the purpose of a study.
- Data considered in observational studies are **observational data**.

Examples:

Who will win the 2022 Australian federal election?

Survey households

Where are the best schools?

Government administrative data

Who are buying my products?

Customer database

Experimental studies

- A scientific claim generally need to be validated by an *experimental study*.

- In an **experimental study**, a causal variable of interest (referred to as *treatment*) is administered to recipients while holding other covariates at controlled settings to observe responses.
- Data from an experiment are referred to as **experimental data**.

Examples:

Is the vaccine effective against flu?

The data of whether the person who was administered the vaccine or placebo caught the flu afterwards.

Which fertilizer brand is most effective for wheat yield?

Yield data from crop field trial with plots treated with one of the three fertilizer brands.

Experimental units

Experimental units are recipients of the allocated treatment such that no sub-division of it can receive another treatment independently.

- Prof Android delivers their lecture by reciting word-to-word from the text in a monotone.
- Prof Alien delivers their lecture by transmitting the information directly to the students mind.
- You want to see if one of the methods is more effective.
- Students in class 1, 3, 4, 7 and 10 have Prof Android.
- Students in class 2, 5, 6, 8 and 9 have Prof Alien.

What are the experimental units? It's the classes.

Observational units

Observational units are units that you measure the response on.

Carrying on from the previous example...

- Students all sit for the same exam.
- You record the exam mark for each student.

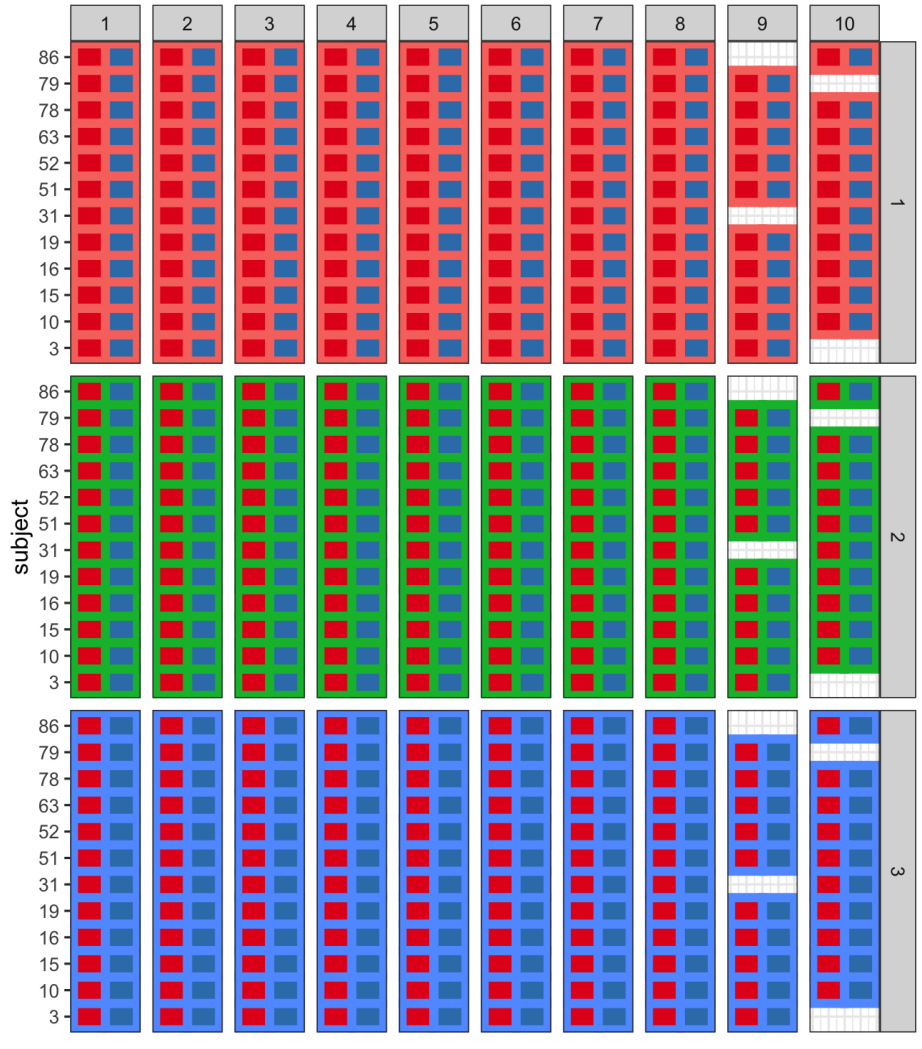
What are the observational units? It's the students.

- Note:
  - *observational unit* is not the *observation* (the response)!
  - Sometimes the experimental units *are the same as* the observational units.

Example: french fries (hot chips)

This is data from a 10 week sensory experiment, 12 individuals assessed taste of french fries on several scales (how potato-y, buttery, grassy, rancid, paint-y do they taste?), fried in one of 3 different oils, replicated twice.

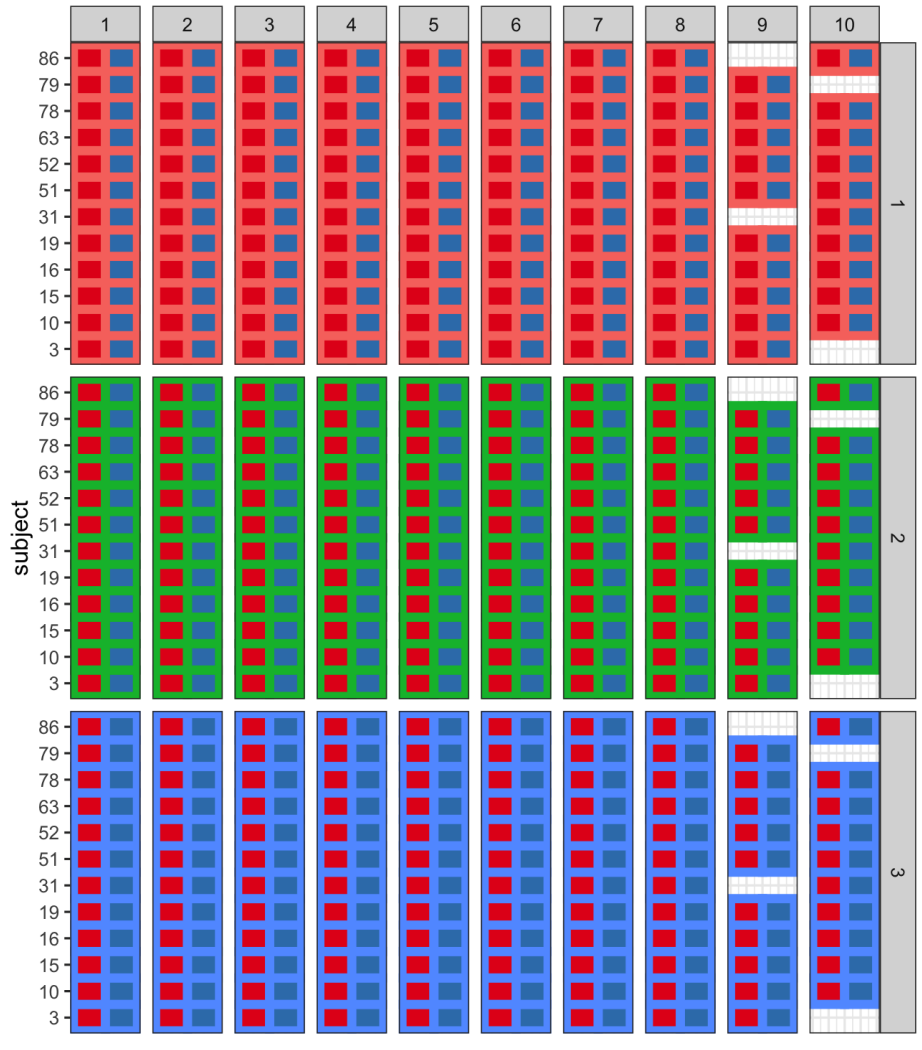
Data is available in the deprecated R package reshape and was one of the examples that inspired the tidyverse tools.



Example: french fries (hot chips)

10 week sensory experiment, 12 individuals assessed taste of french fries on several scales, fried in one of 3 different oils, replicated twice.

- The **treatment** is oil, and there are 3 of them.
- The **experimental units** are batches of chips.
- The **observational units** are the tasters.
- **Replication** is the two batches of each oil for each week.
- Weeks could be considered to be **blocks**, because the taste might change as the oil ages.
- The **outcome** or measured variable is the rating factor. There are five taste factors recorded.
- **Randomisation** applied to order of tasting (probably), but tasters should be **blind** to the type of oil.



Systematic Design of Experiments

Randomisation applied to order of tasting (probably)

- Why don't we order the treatments in a **systematic order**?
- Isn't this easier to manage the experiment?

Systematic designs are prone to **bias** and **confounding**.

Randomisation

- Treatments should be allocated *randomly* to experimental units.
- This avoids:
  - **systematic bias** - e.g. all flu vaccine A tested in January (summer) and all flu vaccine B tested in July (winter).
  - **selection bias** - e.g. giving the treatment that you are testing to the sick patients and placebo to those that are healthy.
  - **other bias** - e.g. the lab technician giving the treatment to the first rat that is taken out of the cage.

Blocking

Blocks are used to group the experimental units into alike units.

- If well done, blocking can lower the variance of treatment contrasts which increase power.
- A non-homogeneous block (i.e. units within block are *not* alike) can decrease the power of the experiment.

You can form blocks from:

- **Natural discrete divisions** between experimental units.
  E.g. in experiments with people, the gender make an obvious block.
- Grouping experimental units with similar **continuous gradients**.
  E.g., if the experiment is spread out in time or space and there exists no obvious natural boundaries, then an arbitrary boundary may be chosen to group experimental units that are contiguous in time or space.

The Salk Vaccine Field Trial

- The first polio epidemic hit the United States in 1916 claiming hundreds of thousands of victims, especially children.
- National Foundation for Infantile Paralysis (NFIP) was ready to test the vaccine developed by Jonas Salk in the real world.
- A controlled experiment was proposed to test the effectiveness of the vaccine on grade 1, 2 and 3 children at selected school districts though the country where the risk of polio was high.
- In total two million children were involved although not all parents consented to their children to be vaccinated.

Design for the NFIP Study

Vaccinate all grade 2 children whose parents would consent, leaving children in grades 1 and 3 as controls.

- Can grade 2 children whose parents did not consent be included as control?
- What are the potential issues with such a design?
- Polio is a contact disease. Would incidences of disease be higher in grade 2?

Randomised controlled trial

An alternate vaccine trial randomly assigned the vaccine and placebo to children.

Vaccine Results

The NFIP Study

Group	Participants	Rate
Vaccinated (Grade 2)	221,998	25
Control (Grade 1 & 3)	725,173	54
Not Vaccination (Grade 2, no consent)	123,605	44
Incomplete Vaccination (Grade 2, incomplete)	9,904	40

Randomised controlled trial

Group	Participants	Rate
Vaccinated	200,745	28
Placebo	201,229	71
Not Vaccination (no consent)	338,778	46
Incomplete Vaccination	8,484	24

- The rate is the number of polio cases per 100,000 in each group.
- RCT and NFIP trial sampled from school districts with similar exposures to the polio virus.

The groups labelled variously as Not Vaccination (no consent), Control and Placebo group did not receive the vaccine. Why is the rate of polio cases different?

Possible explanations

- Higher income parents would more likely consent to treatment than lower-income parents.
- Children of higher income parents are more vulnerable to polio.
- Many forms of polio are hard to diagnose and in borderline cases.

Limitations in (social) experiments

- Cooperation needed from participants
- Ethical objections
- Substitution bias
- Sample attrition
- Hawthorne effect

Basically, designing and running experiments are *hard*.

Taxonomy of types of data studies

- **Experimental data**: the gold standard of data collection, but very difficult
- **Observational data**:
  - **census**: all (or close to all) members of the population are measured
  - **survey sample**: each member of the population has a known probability of being selected into the sample, eg cohort study, cross-sectional, case-control, cluster/hierarchical, multi-stage
  - **non-random sample**: it is not known how the sample relates to the population
  - **censored**: events might happen outside of observation interval, eg observed up to 30,000km but brake failure was at 45,000 km
  - **occurrences**: only when an incident is observed is it recorded, eg wildlife sightings, warranty claims, complaints

Knowing how the sample of data relates to the population is an essential ingredient for making inferential statements and making decisions with data.

Pop Quizzes

Observational (what type) or experimental data?

Airline traffic (on-time performance database) in the USA as available from https://www.bts.gov. Records on every commercial flight operated in the USA since the 1980s, that has carried passengers.

Observational, census

Always ask yourself "What is missing?"

Observational (what type) or experimental data?

National Longitudinal Survey of Youth 1979
https://www.nlsinfo.org/content/cohorts/NLSY79
Measures people born between 1957 and 1964. At the time of first interview, respondents' ages ranged from 14 to 22.

OBSERVATIONAL, SURVEY SAMPLE

Always ask yourself "What is the population?"

Observational (what type) or experimental data?

Atlas of Living Australia at https://www.ala.org.au. The Atlas of Living Australia (ALA) is a collaborative, digital, open infrastructure that pulls together Australian biodiversity data from multiple sources, making it accessible and reusable.

The ALA helps to create a more detailed picture of Australia's biodiversity for scientists, policy makers, environmental planners and land managers, industry and the general public, and enables them to work more efficiently.

OBSERVATIONAL, OCCURRENCE

Always ask yourself "What is missing?" and "What is the population?"

Observational (what type) or experimental data?

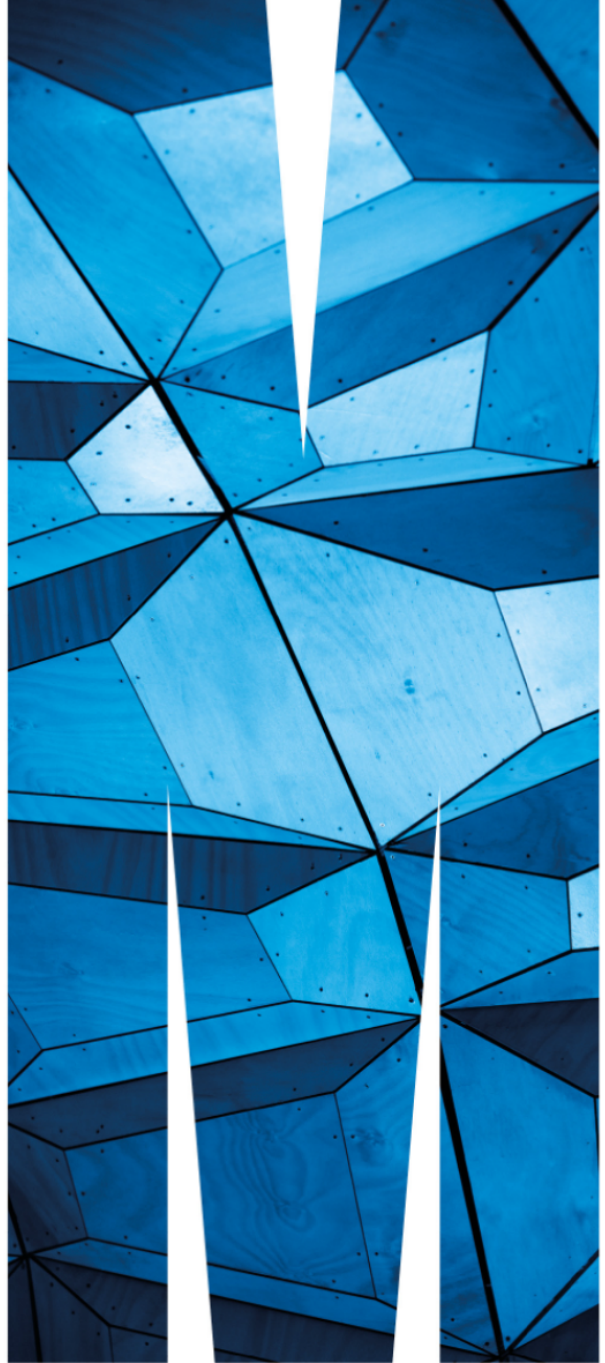
The US National Institute of Health provides a catalog of medical studies including many COVID studies. Here is one that studies the "Safety and Efficacy of C21 in Subjects With COVID-19".

EXPERIMENTAL

What are the treatments? Experimental units? Outcome measure? Randomisation?

Slides originally developed by Professor Di Cook and Dr Emi Tanaka, Slides maintained by Dr. Kate Saunders

MONASH University



Lecturer: *Kate Saunders*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

Week 2