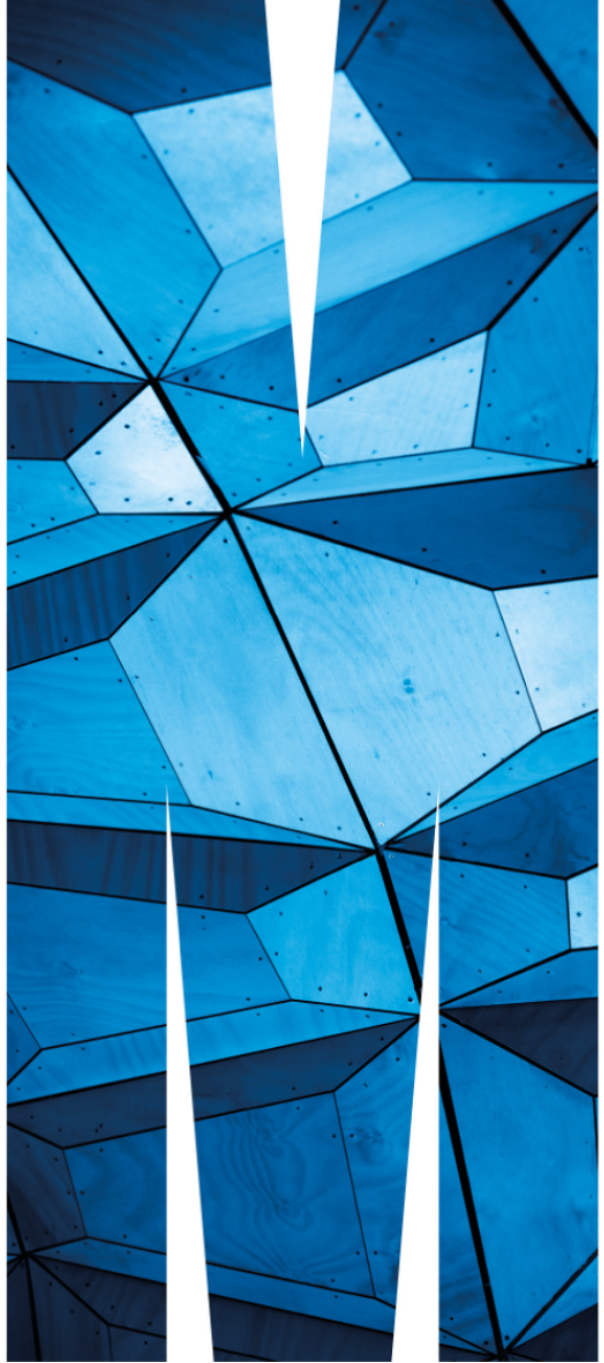
MONASH University

# ETC5512: Wild Caught Data

## Introduction to data collection methods




Lecturer: *Emi Tanaka*

Department of Econometrics and Business Statistics

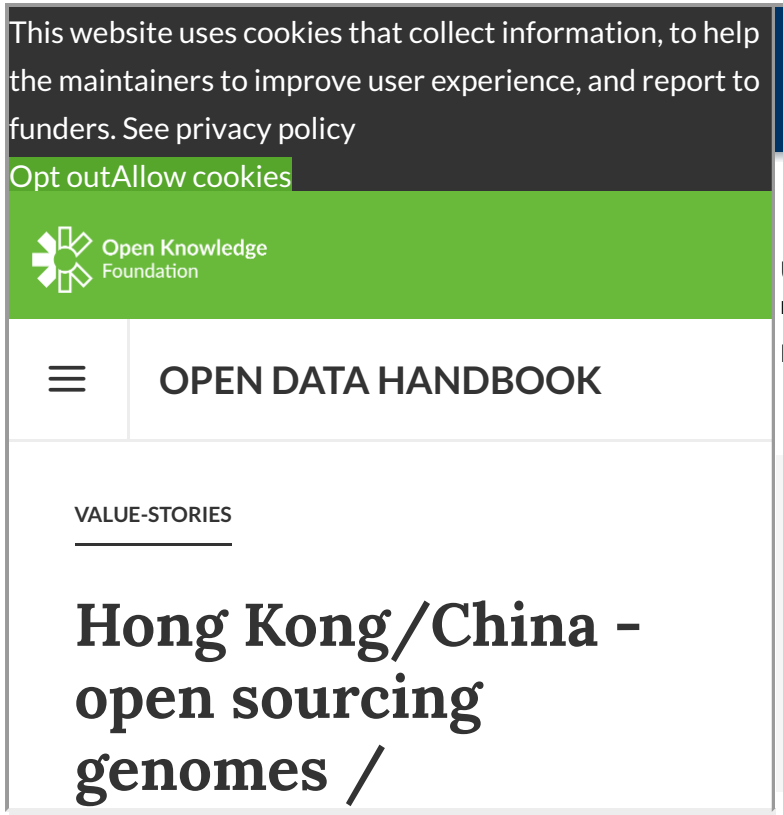
ETC5512.Clayton-x@monash.edu

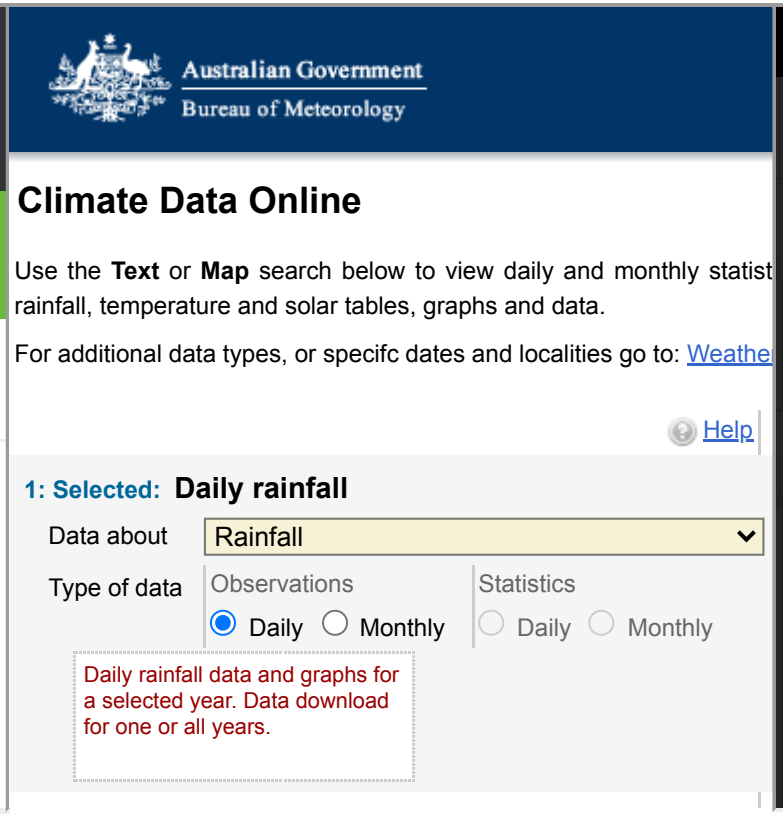

Week 1

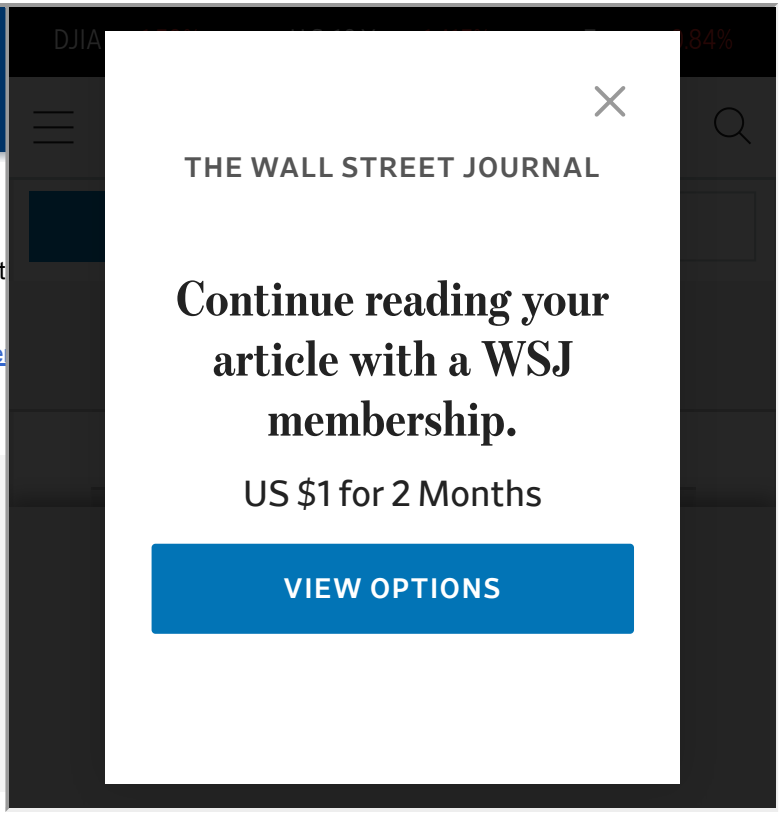
# Starts with a question

# What questions do you have ... ?

- ... about a virus?
- ... forecasting the weather?
- ... about the stock prices?

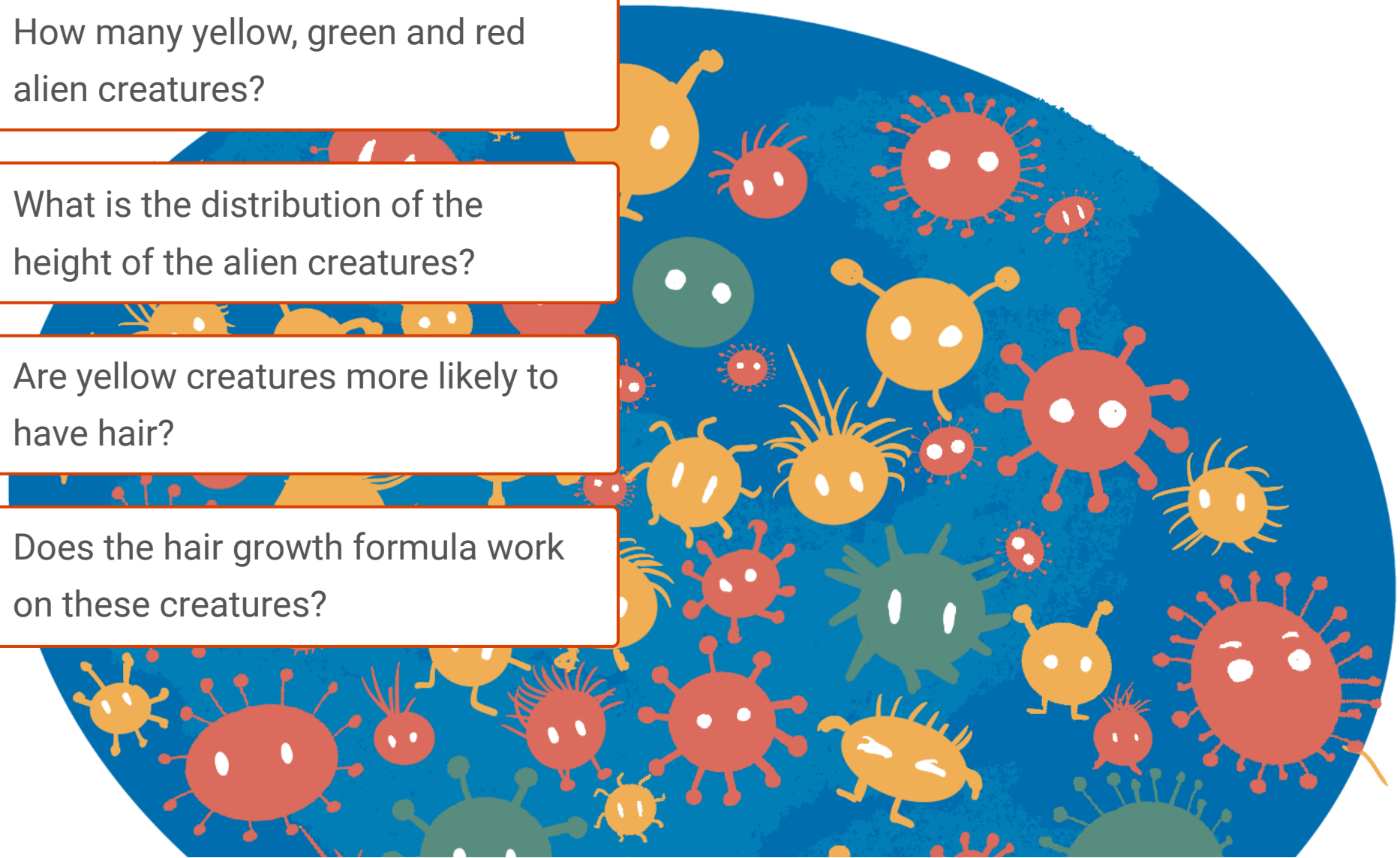


# Planet Monash

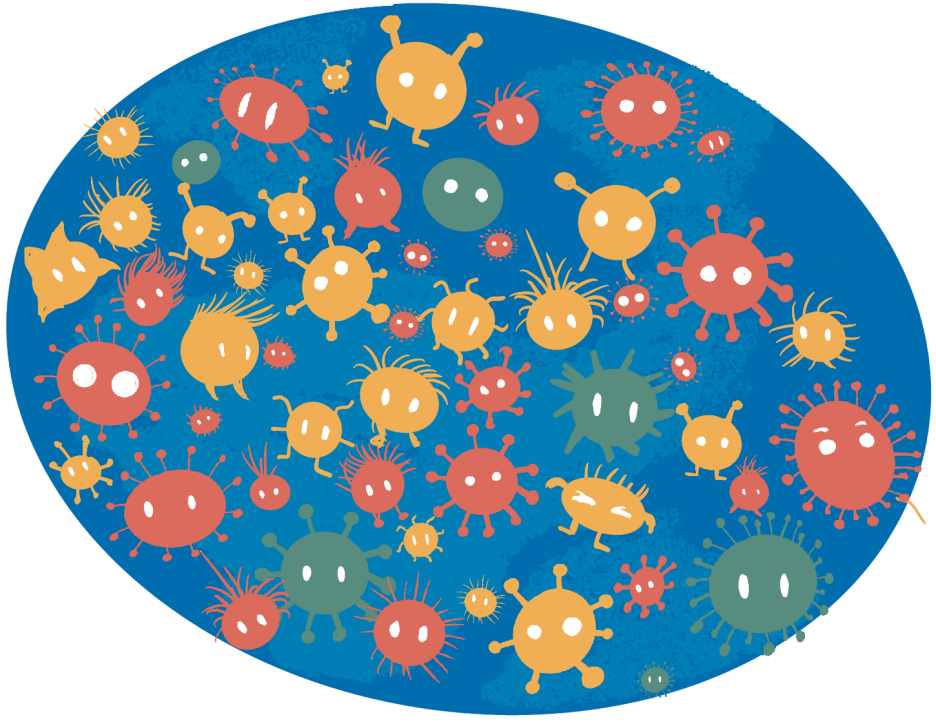
- How many yellow, green and red alien creatures?
- What is the distribution of the height of the alien creatures?
- Are yellow creatures more likely to have hair?
- Does the hair growth formula work on these creatures?



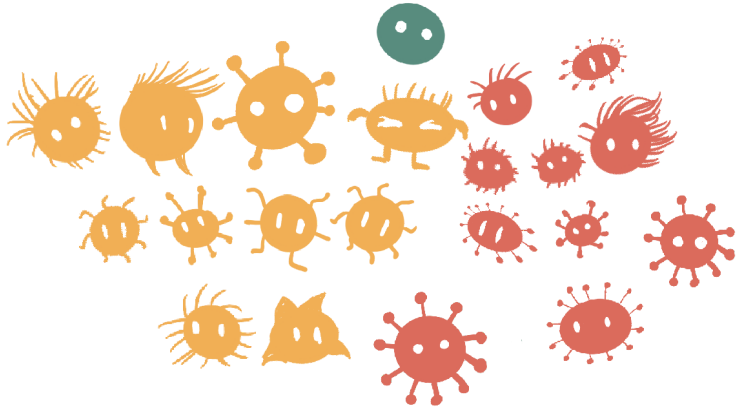
# Now that we have a question ...

# What's the population of interest?

## Population



## Sample



Population parameters:

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

Estimates of population parameters:

Green: 1/21=0.048, Yellow: 10/21=0.476,
Red: 10/21=0.476

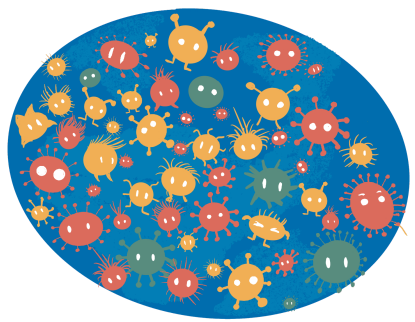
# Sampling the population

Collecting data on the entire population is normally too expensive or infeasible!

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

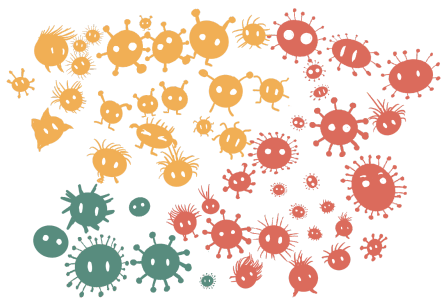
**Simple random sampling**

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



**Stratified random sampling**

Units are drawn from non-overlapping sub-populations.



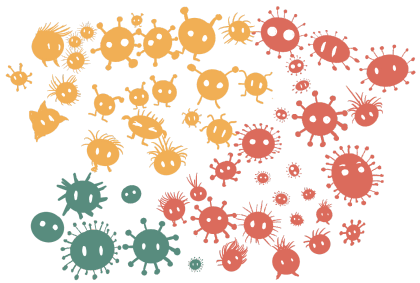
# Goal of sampling schemes

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

- 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** or **covariates** 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 have 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 ideally are sampled *randomly*, but more than often selection are non-random.

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 white pages and stop when you have 20 observations.

You want to get the hair count distribution of the Plant Monash population.

You sample creatures from the Society of Bald Extraterrestrials.

# Reality of data collection ...

- Making an appropriate sampling design is ***hard***.
  - There may be unknown or hidden structures in the population.
  - You may introduce intentional data structures, e.g.
    - Cross-sectional data,
    - Repeated cross-sectional data (e.g. case-control),
    - Panel or longitudinal data (e.g. cohort studies), and so on.
  - You may have unintended or unknown structures in the data, e.g. confounded variables.
  - It's further complicated by:
    - Non-response,
    - Missing data,
    - Mis-measured data,
    - Sample attrition, and so on. 😱

# Observational studies

- Studies mentioned so far has been **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 population 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 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 unit

**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 method 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 the experimental units?** It's the classes.

# Observational unit

**Observational units** are units that you measure the response on.

Carrying on from 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)!

# Wheat Yield Trial

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CORRI | SUNLA | RAC80 | VF300 | TINCU | RAC80 | VF655 | VF519 | RAC79 | VG701 | OSPRE | PELSA | VF300 | MEERI | EXCAL |
| CADOU | SUNFI | RAC80 | VF299 | TINCU | RAC75 | SWIFT | TRIDE | MOLIN | HOUTM | VF299 | BT_SC | M5075 | KATUN | (WWH* |
| BLADE | SUNBR | RAC80 | VF508 | CONDO | VF508 | VF299 | AMERY | WI221 | RAC78 | VF508 | BD231 | RAC81 | SUNFI | BATAV |
| BEULA | SHRIK | RAC80 | VF655 | RAC65 | TINCU | SUNBR | RAC79 | RAC81 | (WWH* | RAC82 | VG503 | CADOU | VF519 | WW183 |
| BATAV | ROSEL | RAC80 | (WWH* | M5097 | DGR/M | PEROU | BD231 | VF655 | RAC77 | CONDO | VG878 | RAC80 | RAC82 | RAC65 |
| AROON | PEROU | RAC80 | (WqKP | K2011 | VG878 | JANZ | KITE | WI232 | RAC80 | GOROK | HALBE | VG701 | RAC79 | RAC75 |
| AMERY | PELSA | RAC79 | WI216 | WW183 | VG714 | MACHE | BT_SC | RAC75 | RAC71 | VF302 | RAC80 | STILE | SUNLA | OUYEN |
| YARRA | OXLEY | RAC79 | WW147 | WW147 | KULIN | VF664 | TINCU | VF300 | RAC81 | CUNNI | DGR/M | WW140 | SUNBR | RAC71 |
| WYUNA | OUYEN | RAC79 | WW140 | M4997 | SUNLA | LOWAN | WARBL | AROON | M5097 | K2011 | WILGO | JANZ | M5097 | HOUTM |
| TRIDE | OSPRE | RAC78 | VF519 | WI232 | TASMA | WW147 | STILE | BEULA | VG506 | DOLLA | RAC80 | VG714 | RAC81 | RAC81 |
| TATIA | LOWAN | RAC77 | RAC82 | WI231 | EXCAL | (WqKP | M4997 | RAC81 | RAC77 | ROSEL | RAC79 | KIATA | WW147 | MOLIN |
| STILE | LARK | RAC77 | RAC82 | WI221 | RAC79 | WW183 | BATAV | RAC81 | RAC81 | RAC79 | RAC75 | SPEAR | WYUNA | CORRI |
| SPEAR | KULIN | RAC77 | RAC81 | BD231 | RAC65 | RAC77 | M5075 | CORRI | RAC80 | BEULA | SHRIK | RAC78 | VF664 | RAC80 |
| SCHOM | KITE | RAC75 | RAC81 | VG878 | VF302 | MEERI | SHRIK | SCHOM | KIATA | RAC81 | TASMA | RAC81 | RAC81 | RAC65 |
| MOLIN | KIATA | RAC75 | RAC81 | VG714 | RAC80 | WILGO | CADOU | SUNFI | BLADE | WI216 | SCHOM | RAC81 | TRIDE | MACHE |
| MEERI | KATUN | RAC71 | RAC81 | VG701 | OXLEY | KATUN | YARRA | CONDO | PELSA | RAC77 | KULIN | WARBL | WI232 | PEROU |
| MACHE | HOUTM | RAC69 | RAC81 | VG506 | RAC65 | RAC81 | CUNNI | WI216 | RAC81 | AMERY | KITE | WI221 | RAC81 | YARRA |
| JANZ | HALBE | RAC65 | RAC81 | VG503 | RAC80 | RAC80 | OSPRE | RAC69 | RAC82 | RAC81 | RAC80 | OXLEY | TINCU | VF655 |
| EXCAL | GOROK | WILGO | RAC81 | VG127 | K2011 | OUYEN | WW140 | LARK | RAC81 | RAC77 | VG127 | RAC80 | TINCU | VG506 |
| DGR/M | DOLLA | WARBL | RAC81 | VF302 | RAC82 | VG127 | HALBE | WI231 | WYUNA | WW147 | TATIA | LOWAN | RAC81 | WI231 |
| BT_SC | M5075 | TASMA | RAC81 | VF664 | GOROK | VG503 | RAC81 | ROSEL | TATIA | ANGAS | RAC69 | VF655 | BLADE | M4997 |
| ANGAS | CUNNI | SWIFT | RAC81 | VF655 | WW147 | RAC81 | ANGAS | SPEAR | DOLLA | AROON | (WqKP | LARK | SWIFT | RAC77 |

- A selective breeding experiment with 107 wheat varieties (or *genotypes*) were conducted in South Australia in a field with plots laid out in a rectangular array with 22 rows and 15 columns.
- The breeders want to find a variety with *high yield*.
- The **treatments** are the *107 wheat varieties*.
- The **experimental units** are the *330 plots*.
- The **observational units** are also the *330 plots*.

Source: Gilmour et al. (1997) Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agric Biol Env Statistics*, 2, 269-293.

# Replications

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CORRI | SUNLA | RAC80 | VF300 | TINCU | RAC80 | VF655 | VF519 | RAC79 | VG701 | OSPRE | PELSA | VF300 | MEERI | EXCAL |
| CADOU | SUNFI | RAC80 | VF299 | TINCU | RAC75 | SWIFT | TRIDE | MOLIN | HOUTM | VF299 | BT_SC | M5075 | KATUN | (WWH* |
| BLADE | SUNBR | RAC80 | VF508 | CONDO | VF508 | VF299 | AMERY | WI221 | RAC78 | VF508 | BD231 | RAC81 | SUNFI | BATAV |
| BEULA | SHRIK | RAC80 | VF655 | RAC65 | TINCU | SUNBR | RAC79 | RAC81 | (WWH* | RAC82 | VG503 | CADOU | VF519 | WW183 |
| BATAV | ROSEL | RAC80 | (WWH* | M5097 | DGR/M | PEROU | BD231 | VF655 | RAC77 | CONDO | VG878 | RAC80 | RAC82 | RAC65 |
| AROON | PEROU | RAC80 | (WqKP | K2011 | VG878 | JANZ | KITE | WI232 | RAC80 | GOROK | HALBE | VG701 | RAC79 | RAC75 |
| AMERY | PELSA | RAC79 | WI216 | WW183 | VG714 | MACHE | BT_SC | RAC75 | RAC71 | VF302 | RAC80 | STILE | SUNLA | OUYEN |
| YARRA | OXLEY | RAC79 | WW147 | WW147 | KULIN | VF664 | TINCU | VF300 | RAC81 | CUNNI | DGR/M | WW140 | SUNBR | RAC71 |
| WYUNA | OUYEN | RAC79 | WW140 | M4997 | SUNLA | LOWAN | WARBL | AROON | M5097 | K2011 | WILGO | JANZ | M5097 | HOUTM |
| TRIDE | OSPRE | RAC78 | VF519 | WI232 | TASMA | WW147 | STILE | BEULA | VG506 | DOLLA | RAC80 | VG714 | RAC81 | RAC81 |
| TATIA | LOWAN | RAC77 | RAC82 | WI231 | EXCAL | (WqKP | M4997 | RAC81 | RAC77 | ROSEL | RAC79 | KIATA | WW147 | MOLIN |
| STILE | LARK | RAC77 | RAC82 | WI221 | RAC79 | WW183 | BATAV | RAC81 | RAC81 | RAC79 | RAC75 | SPEAR | WYUNA | CORRI |
| SPEAR | KULIN | RAC77 | RAC81 | BD231 | RAC65 | RAC77 | M5075 | CORRI | RAC80 | BEULA | SHRIK | RAC78 | VF664 | RAC80 |
| SCHOM | KITE | RAC75 | RAC81 | VG878 | VF302 | MEERI | SHRIK | SCHOM | KIATA | RAC81 | TASMA | RAC81 | RAC81 | RAC65 |
| MOLIN | KIATA | RAC75 | RAC81 | VG714 | RAC80 | WILGO | CADOU | SUNFI | BLADE | WI216 | SCHOM | RAC81 | TRIDE | MACHE |
| MEERI | KATUN | RAC71 | RAC81 | VG701 | OXLEY | KATUN | YARRA | CONDO | PELSA | RAC77 | KULIN | WARBL | WI232 | PEROU |
| MACHE | HOUTM | RAC69 | RAC81 | VG506 | RAC65 | RAC81 | CUNNI | WI216 | RAC81 | AMERY | KITE | WI221 | RAC81 | YARRA |
| JANZ | HALBE | RAC65 | RAC81 | VG503 | RAC80 | RAC80 | OSPRE | RAC69 | RAC82 | RAC81 | RAC80 | OXLEY | TINCU | VF655 |
| EXCAL | GOROK | WILGO | RAC81 | VG127 | K2011 | OUYEN | WW140 | LARK | RAC81 | RAC77 | VG127 | RAC80 | TINCU | VG506 |
| DGR/M | DOLLA | WARBL | RAC81 | VF302 | RAC82 | VG127 | HALBE | WI231 | WYUNA | WW147 | TATIA | LOWAN | RAC81 | WI231 |
| BT_SC | M5075 | TASMA | RAC81 | VF664 | GOROK | VG503 | RAC81 | ROSEL | TATIA | ANGAS | RAC69 | VF655 | BLADE | M4997 |
| ANGAS | CUNNI | SWIFT | RAC81 | VF655 | WW147 | RAC81 | ANGAS | SPEAR | DOLLA | AROON | (WqKP | LARK | SWIFT | RAC77 |

- The varieties `VF655`, `TINCURRIN` and `WW1477` have a **replication** of 6, the remaining 104 varieties each have a replication of 3.
- Treatment **replications are essential** in an experiment; without any replication, no treatment variation can be measured nor distinguished from unit variation.
- More replications are desirable for accuracy, however, there is always a tension to balance between accuracy and the cost of the experiment.

# Pseudo-replication

Carrying on from the teaching example...

- Suppose there were 30 students in each class.
- The treatments were the two teaching method confounded with each professor.
- There were two professors and 10 classes.
- Each professor was randomly assigned to 5 classes, so each professor manages 150 students.

**What are the replications of each treatment?** It's 5.

The treament of repetition as replication in the analysis is referred to as **pseudo-replication**.

# Systematic Design of Experiments

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BEULA | DGR/M | K2011 | M5075 | OXLEY | RAC75 | RAC79 | RAC81 | RAC81 | SPEAR | TATIA | VF508 | VG701 | WI232 | YARRA |
| BD231 | DGR/M | K2011 | M4997 | OXLEY | RAC75 | RAC79 | RAC81 | RAC81 | SHRIK | TATIA | VF508 | VG506 | WI232 | YARRA |
| BD231 | CUNNI | K2011 | M4997 | OUYEN | RAC75 | RAC79 | RAC80 | RAC81 | SHRIK | TASMA | VF508 | VG506 | WI231 | YARRA |
| BD231 | CUNNI | JANZ | M4997 | OUYEN | RAC71 | RAC79 | RAC80 | RAC81 | SHRIK | TASMA | VF302 | VG506 | WI231 | WYUNA |
| BATAV | CUNNI | JANZ | LOWAN | OUYEN | RAC71 | RAC79 | RAC80 | RAC81 | SCHOM | TASMA | VF302 | VG503 | WI231 | WYUNA |
| BATAV | CORRI | JANZ | LOWAN | OSPRE | RAC71 | RAC79 | RAC80 | RAC81 | SCHOM | SWIFT | VF302 | VG503 | WI221 | WYUNA |
| BATAV | CORRI | HOUTM | LOWAN | OSPRE | RAC69 | RAC79 | RAC80 | RAC81 | SCHOM | SWIFT | VF300 | VG503 | WI221 | WW183 |
| AROON | CORRI | HOUTM | LARK | OSPRE | RAC69 | RAC78 | RAC80 | RAC81 | ROSEL | SWIFT | VF300 | VG127 | WI221 | WW183 |
| AROON | CONDO | HOUTM | LARK | MOLIN | RAC69 | RAC78 | RAC80 | RAC81 | ROSEL | SUNLA | VF300 | VG127 | WI216 | WW183 |
| AROON | CONDO | HALBE | LARK | MOLIN | RAC65 | RAC78 | RAC80 | RAC81 | ROSEL | SUNLA | VF299 | VG127 | WI216 | WW147 |
| ANGAS | CONDO | HALBE | KULIN | MOLIN | RAC65 | RAC77 | RAC80 | RAC81 | RAC82 | SUNLA | VF299 | VF664 | WI216 | WW147 |
| ANGAS | CADOU | HALBE | KULIN | MEERI | RAC65 | RAC77 | RAC80 | RAC81 | RAC82 | SUNFI | VF299 | VF664 | WARBL | WW147 |
| ANGAS | CADOU | GOROK | KULIN | MEERI | RAC65 | RAC77 | RAC80 | RAC81 | RAC82 | SUNFI | TRIDE | VF664 | WARBL | WW147 |
| AMERY | CADOU | GOROK | KITE | MEERI | RAC65 | RAC77 | RAC80 | RAC81 | RAC82 | SUNFI | TRIDE | VF655 | WARBL | WW147 |
| AMERY | BT_SC | GOROK | KITE | MACHE | RAC65 | RAC77 | RAC80 | RAC81 | RAC82 | SUNBR | TRIDE | VF655 | VG878 | WW147 |
| AMERY | BT_SC | EXCAL | KITE | MACHE | PEROU | RAC77 | RAC80 | RAC81 | RAC82 | SUNBR | TINCU | VF655 | VG878 | WW140 |
| (WqKP | BT_SC | EXCAL | KIATA | MACHE | PEROU | RAC77 | RAC80 | RAC81 | RAC81 | SUNBR | TINCU | VF655 | VG878 | WW140 |
| (WqKP | BLADE | EXCAL | KIATA | M5097 | PEROU | RAC77 | RAC80 | RAC81 | RAC81 | STILE | TINCU | VF655 | VG714 | WW140 |
| (WqKP | BLADE | DOLLA | KIATA | M5097 | PELSA | RAC77 | RAC80 | RAC81 | RAC81 | STILE | TINCU | VF655 | VG714 | WILGO |
| (WWH* | BLADE | DOLLA | KATUN | M5097 | PELSA | RAC75 | RAC80 | RAC81 | RAC81 | STILE | TINCU | VF519 | VG714 | WILGO |
| (WWH* | BEULA | DOLLA | KATUN | M5075 | PELSA | RAC75 | RAC79 | RAC81 | RAC81 | SPEAR | TINCU | VF519 | VG701 | WILGO |
| (WWH* | BEULA | DGR/M | KATUN | M5075 | OXLEY | RAC75 | RAC79 | RAC81 | RAC81 | SPEAR | TATIA | VF519 | VG701 | WI232 |

- The treatments appear to be randomly ordered before.
- Why don't we order the treatments in a **systematic order** like on the left?
- Isn't this easier to manage the experiment?

# Randomisation

- Treatment 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 use 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 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.
- Both the not vaccinated (no consent) and placebo/control group did not receive the treatment but why is the rate of polio cases less in the not vaccinated (no consent) group?

# 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*.

# Pop Quizzes

# Observational or experimental data?

The Academic Performance Index is computed for all California schools based on standardised testing of students. The data sets contain information and characteristics for 100 schools.

OBSERVATIONAL

# Observational or experimental data?

The response is the length of odontoblasts in 60 guinea pigs. Each animal received one of three dose levels of vitamin C by one of two delivery methods by the technician.

EXPERIMENTAL

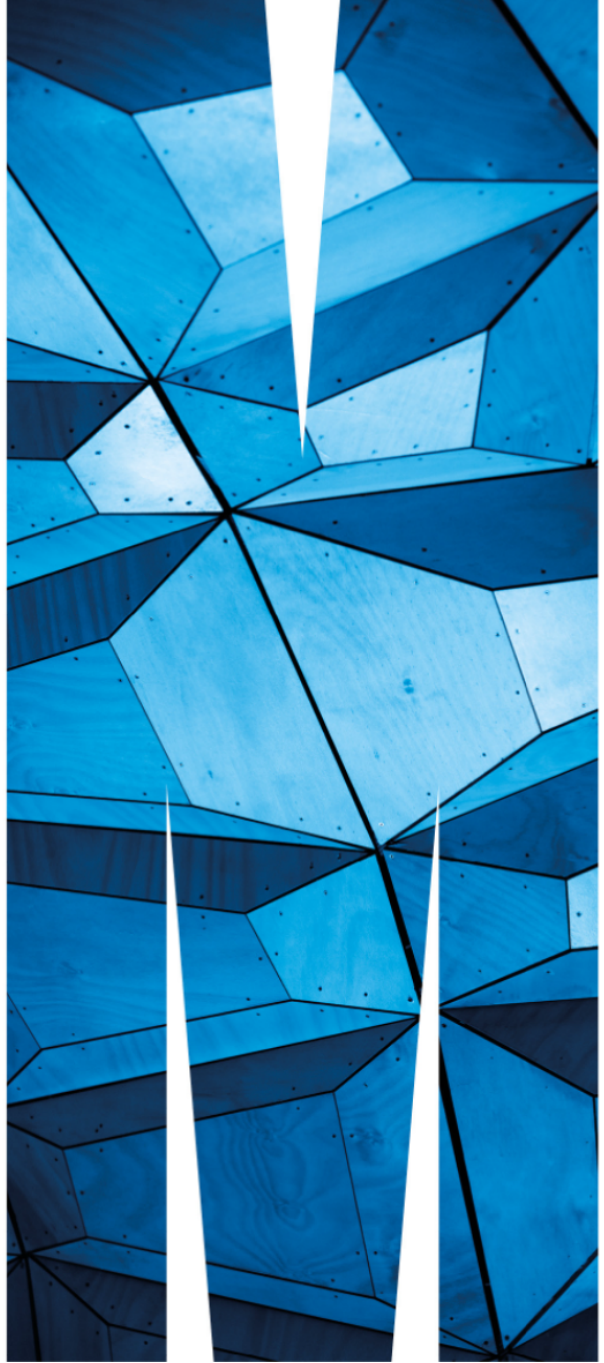
# Observational or experimental data?

Can people really tell the difference between different flavours associated with the color of the skittles? You blind your friends so they can't see the color and collect data on their guess after giving them one skittle at a time.

EXPERIMENTAL

MONASH University





Lecturer: *Emi Tanaka*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

Week 1