STAT115: Introduction to Biostatistics

University of Otago
Ōtākou Whakaihu Waka

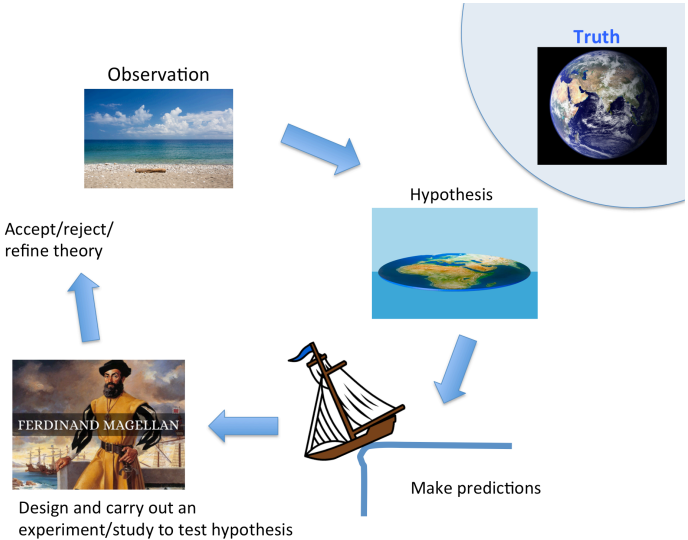
Lecture 2: Data and Discovery

- Saw previously that statistics is about learning from data
- Today we will explore how this fits into the research process
- Will consider how data is collected
- Will examine different types of data

Statistics and Research

- The *scientific method* is a process we use to gain knowledge and understanding through testing of hypotheses.
- This knowledge is gained through collection and analysis of data.
- Statistics provides much of the framework and mechanisms for implementing the scientific method.

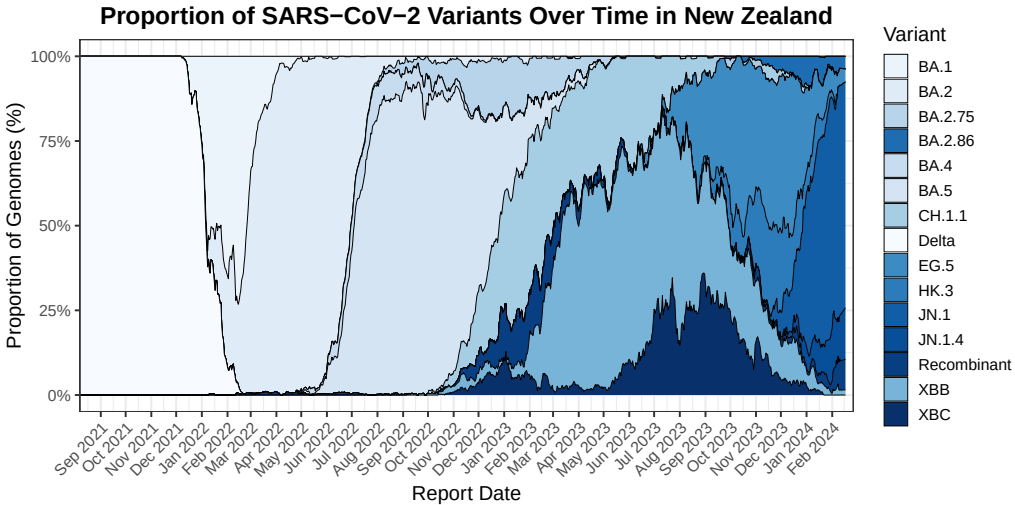
The Scientific Method



Data Collection

- Need to ensure data are suitable to address research question.
- How were data collected?
  - Experiment?
  - Observational study/survey?
- Are data representative?
  - Does sample of data differ systematically from target population?
  - Do we have enough data?
- Will look at these questions again in depth later in the course.

Remember these data?



The Perils of Poor Data Collection

No amount of statistical intervention can circumvent flawed subject matter models or salvage valid conclusions from poorly designed studies...[a]

[a] Marie Davidian and Thomas Louis (2012). Why Statistics? Editorial. *Science* **336 (6077)**, 12.

Types of Data

- Data can be stuctured by *case* and *variable*.
- The *case* identifies basic unit on which data is recorded.
  - E.g. patient ID, petri dish in lab experiment
- A *variable* is a measured characterisic of each unit.
  - E.g. systolic blood pressure, bacterial count
- Variables can be *numerical* or *categorical*.
  - Finer distinctions can be made.

Example: Male Heart Attack Patients in Auckland

Subset only

ID	Eject	Vol	Sten	Time	Outcome	Age	Smoke	Beta	Chol	Surg
399	63	195	0	136	0	36	1	1	61	0
400	66	144	50	65	1	52	0	0	55	0
287	54	145	40	136	0	47	0	0	62	0
81	39	237	87	136	0	39	0	0	56	3
288	59	94	0	135	0	47	1	0	63	0
407	67	117	73	53	1	57	0	0	62	2

- Rows correspond to cases
- Columns correspond to variables

Numerical Variable

Continuous

- *Continuous* numerical variables can (in principle) take any value in a range.
- Examples from heart attack data:
  - `Vol` is measurement of heart size, in ml
  - `Time` is months since heart attack
  - `Age` is age in years
  - `Chol` is total cholesterol in mmoles per litre
- In practice will be subject to some rounding.

Numerical Variable

Discrete

- *Discrete* numerical variables take only certain numerical values, typically whole numbers.
- Value of a discrete variable reflects meaningful magnitude – not just a label.
- Often a count of something.
- Examples:
  - Number of cases of cancer diagnosed during a day.
  - Number of children in a family (0,1,2,3,4,...).

Categorical Variable

Binary

- *Binary* or *dichotomous* variable allocates each case to one of two categories.
- Examples
  - Individual is pet-owner, or not a pet-owner.
  - Coin lands heads or tails.
- Can be represented by text labels (e.g. yes/no, H/T), or numbers (usually 0/1).
- Examples from heart attack data:
  - `Smoke` $= 1$ if smoker, `Smoke` $= 0$ if not.
  - `Beta` $= 1$ if on beta-blockers, `Beta` $= 0$ if not.

Categorical Variable

General

- In general, caterogical variable can have 3 or more categories.
- Each case belongs to just one (i.e. no overlaps).
- Labels can be text or numerical.
- Examples:
  - Blood group A/B/AB/O.
  - Species.
- Examples from heart attack data:
  - `Outcome`: 0 = alive, 1 = sudden cardiac death, 2 = death within 30 days of heart attack, 3 = death from heart failure, 4 = death during surgey, 5 = noncardiac death
  - `Surg`: 0 = no surgery, 1 = surgery as part of trial, 2 = surgery within 1 year, 3 = surgery 1–5 years, 4 = surgery >5 years

Categorical Variable

Nominal and Ordinal

- Categorical variable is *nominal* if there is no natural (or relevant) ordering:
  - Blood group: A/B/AB/O.
  - Species
  - `Surg` and `Outcome` from heart attack data.

  Categorical variable is *ordinal* if there is a natural ordering:
  - Exam result: fail/pass/merit/destinction
  - Degree of pain: minimal/moderate/severe/unbearable.
- Sometimes ordinal variables are analysed as discrete numeric:
  - E.g. responses on *Likert* scale, often 1–5.
  - Likert scale often used on questionnaires: indicate level of agreement with a statement

Ratios and Proportions

- A *ratio* is a fraction given by one quantity over another.
- Example:
  - In a class with 10 boys and 20 girls, the ratio of boys to girls is $10/20 = 1/2 = 0.5$
  - The ratio of girls to boys is $20/10 = 2$
- A *proportion* is fraction of one quantity when compared to the whole.
- Example:
  - In class above, proportion of boys is $\frac{10}{10+20} = \frac{1}{3}$
  - Proportion of girls is $\frac{20}{10+20} = \frac{2}{3}$

Percentages

- Proportions are often expressed in terms of *percentages*.
- To convert proportions to percentages, multiply by 100 and add a % sign.
- To convert percentages to proportions, divide by 100 and remove % sign.
- Examples:
  - $30\% = 0.3$, $56\% = 0.56$
  - `Ejec` in heart attach data is percentage blood ejected in one beat
  - `Sten` is percentage narrowed vessles (stenosis score)
- Specification as a percentage or proportion can imply important loss of information
  - 25% heads in 4 coin tosses: no reason to doubt coin is fair
  - 25% heads in 400 coin tosses: every reason to doubt coin is fair

Rates

- *Rates* are like ratios for quantities with different units.
- Examples:
  - Rates of infection, e.g. number of new diagnoses of HIV in NZ per year.
  - Number of road accidents per 1000km travelled.
- Usual practice is to simplify rates to a 'per unit' measure.
- Typically choose unit for convenience.
  - E.g. Cystic fibrosis has rate of about 30 per 100,000 births.
- In medicine, important to distinguish between *incidence* and *prevalence* rates.
  - Incidence rate is number of *new* cases per unit time and population size
  - Prevalence rate is number of *existing* cases at given time per population size

Summary

- Data is the raw material for research
- A dataset comprises information on variables recorded about individuals or things.
- Variables can be classified in various ways: e.g. numerical versus categorical