Government of Western Australia
Child and Adolescent Health Service

Introductory Biostatistics

Michael Dymock

Biostatistician, Telethon Kids Institute

16 February 2024

Compassion | Excellence | Collaboration | Accountability | Equity | Respect

TELETHON KIDS INSTITUTE Discover. Prevent. Cure.

Perth Children's Hospital Foundation

Acknowledgement of Country

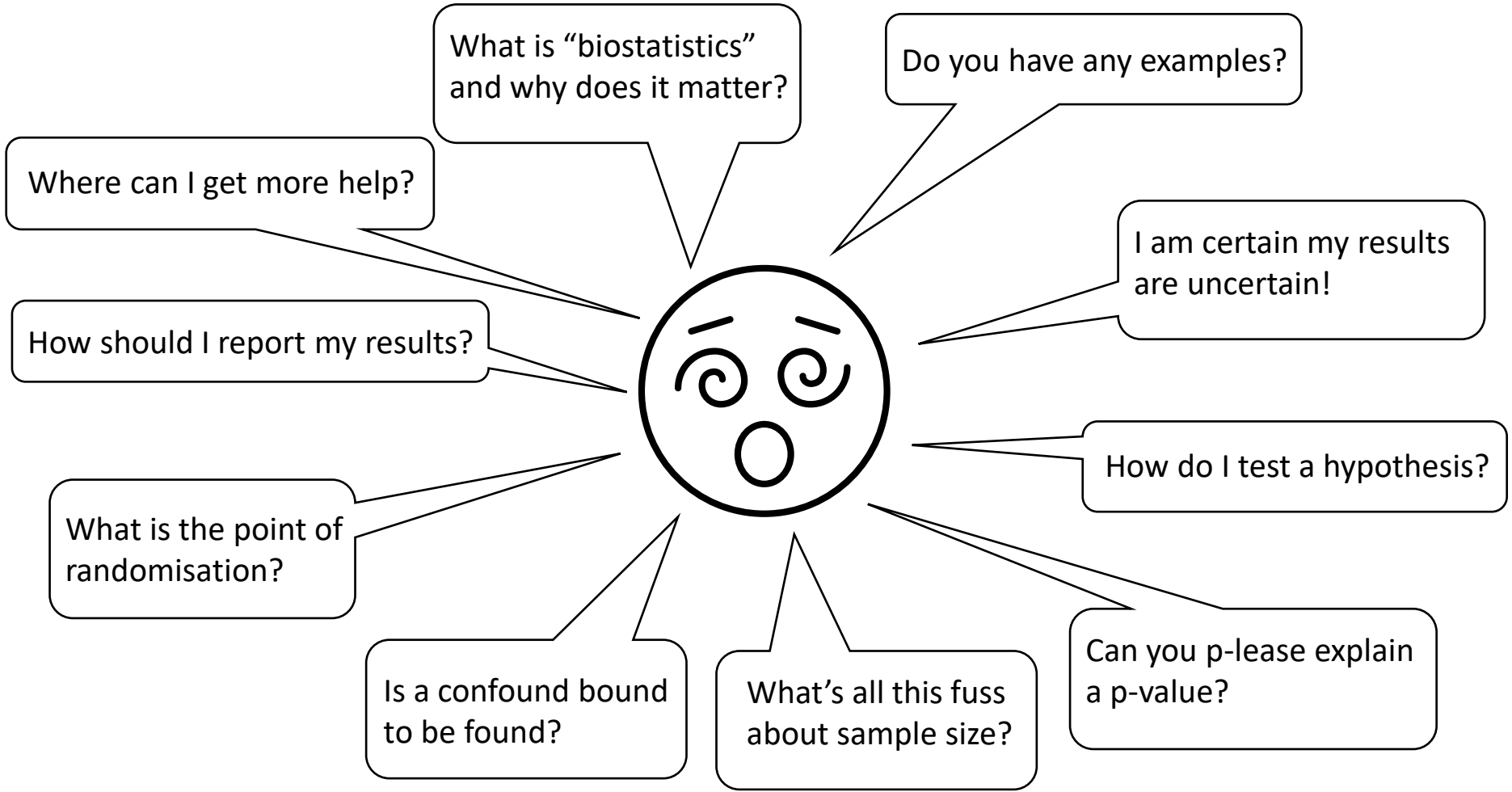
The Child and Adolescent Health Service acknowledge Aboriginal people of the many traditional lands and language groups of Western Australia.
We acknowledge the wisdom of Aboriginal Elders both past and present and pay respect to Aboriginal communities of today.

CAHS Research Education Program

Research Skills Seminar Series

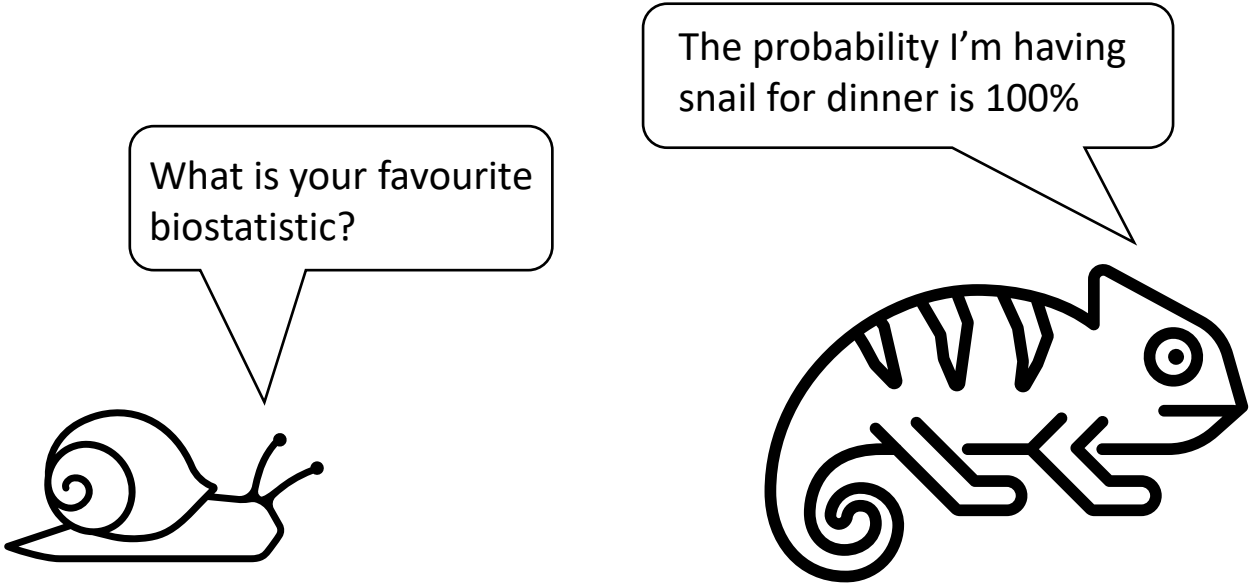
- Over 20 topics across the research process
  - 1h overview
  - Handouts are provided
- Recorded and uploaded
- Feedback
  - Back of handout
  - Emailed link
- Please hold questions to the end
  - Use provided microphone

Overview

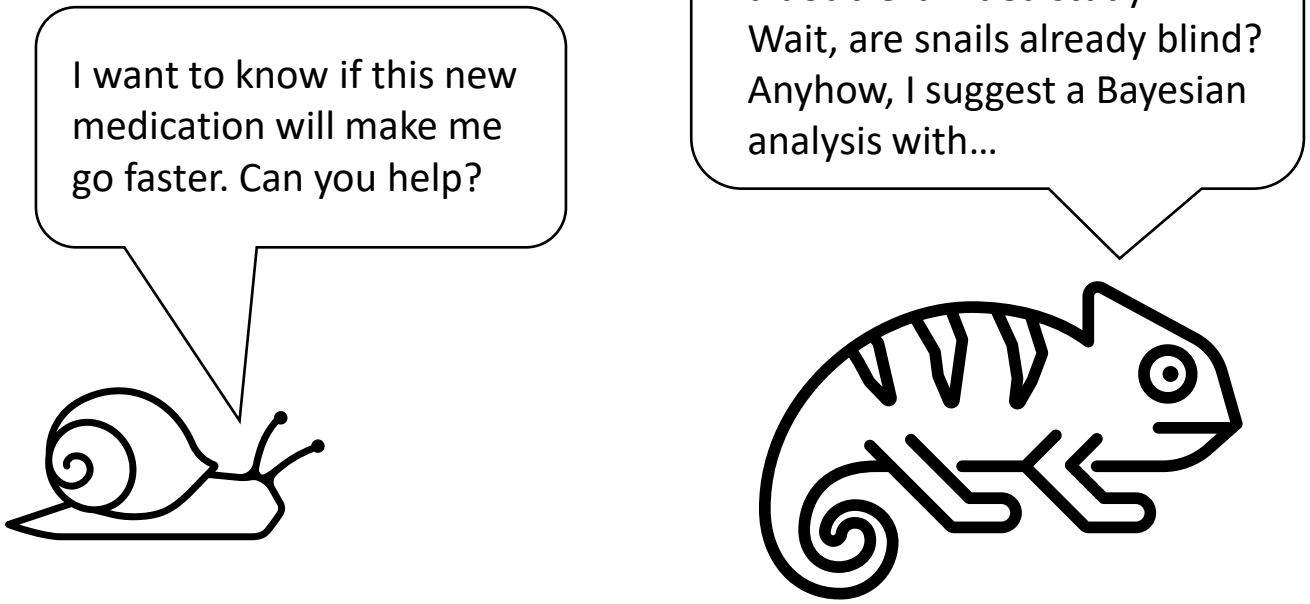

What is “biostatistics” and why does it matter?
Do you have any examples?
Where can I get more help?
I am certain my results are uncertain!
How should I report my results?
How do I test a hypothesis?
What is the point of randomisation?
Is a confound bound to be found?
What’s all this fuss about sample size?
Can you p-lease explain a p-value?

What is “biostatistics” and why does it matter?

What most people think biostatistics is…



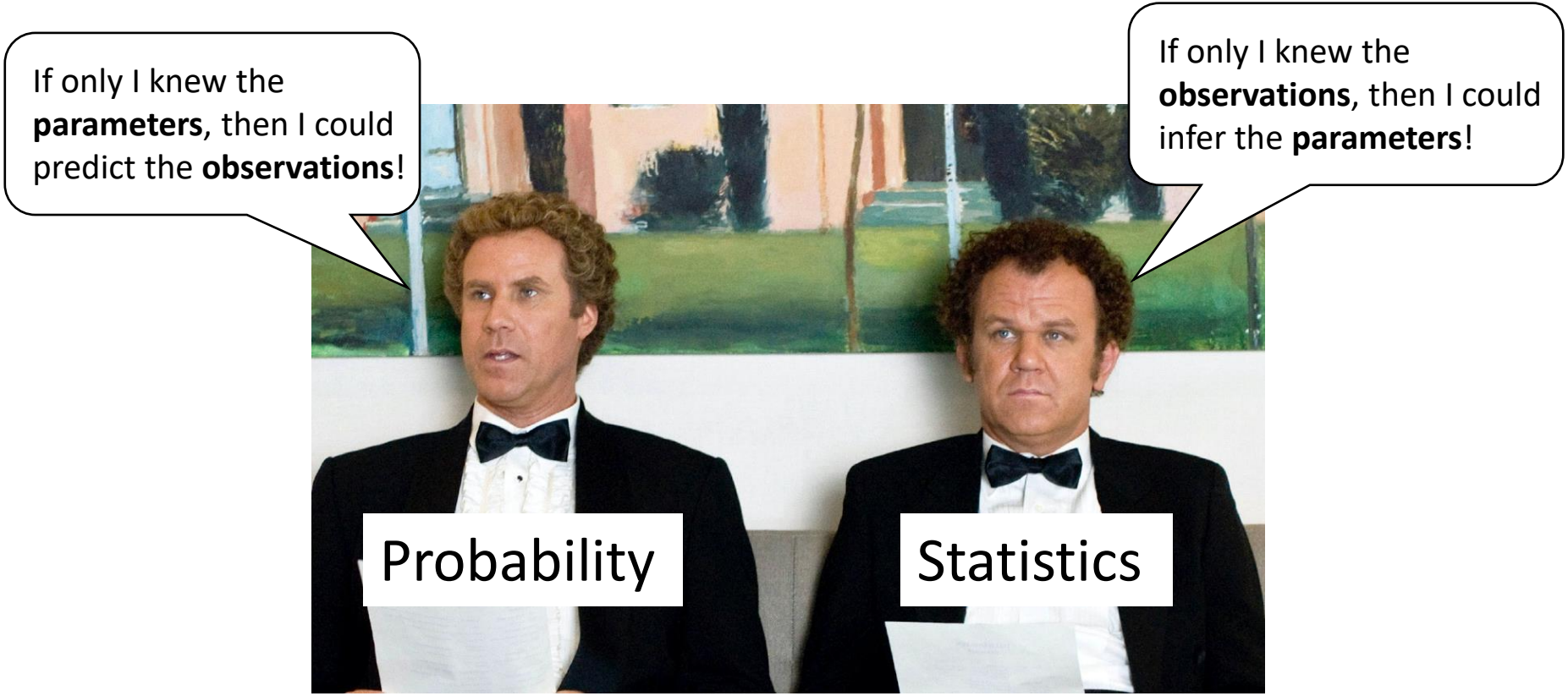
What biostatistics actually is…



More formally…

- Biostatistics can be conceptualised as the **methodology** for the **design**, **analysis** and **interpretation** of studies using health, medical or biological data
  - **Study design** is important to efficiently answer your question of interest, save resources and conduct ethical research
  - **Rigorous analysis and modelling** is important to compute the “correct” results
  - **Appropriate interpretation** of the results is important to draw justifiable conclusions

Probability vs Statistics



Probability vs Statistics

More formally…

- We can use **probability distributions** to understand the behaviour of the world around us
  - E.g., a clinician can make an **informed** decision on prescribing a treatment if they understand its behaviour
    (e.g., **mean** and **variance**)
- We can use **statistical methods** to infer the probability distributions of interest
  - E.g., by collecting and analysing data, we can estimate **parameters** (e.g., **mean** and **variance**)

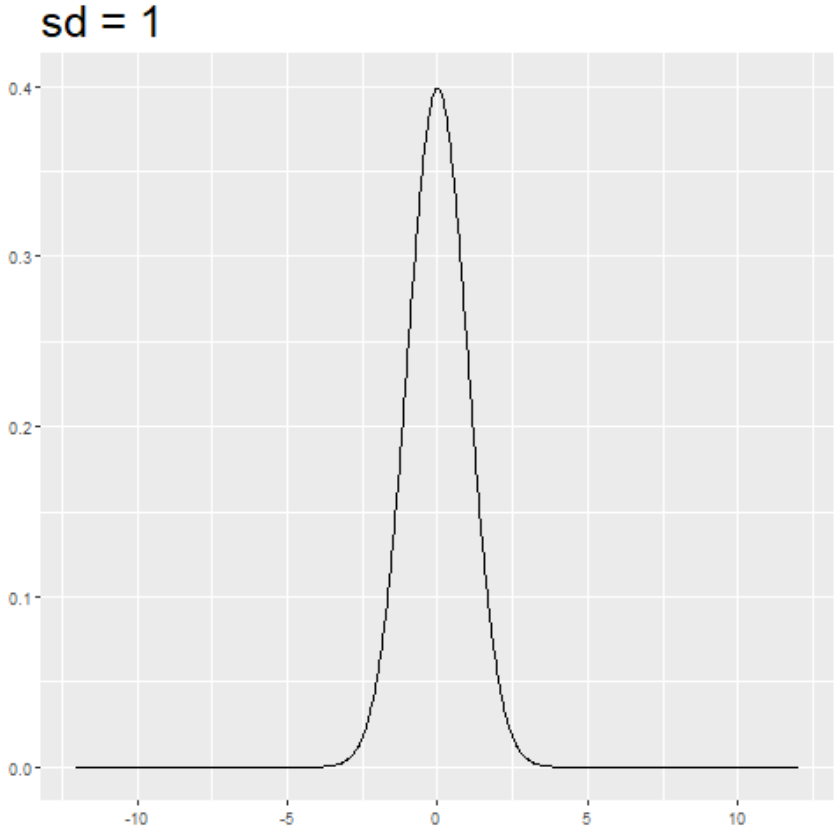
Do you have any examples?

A generic problem

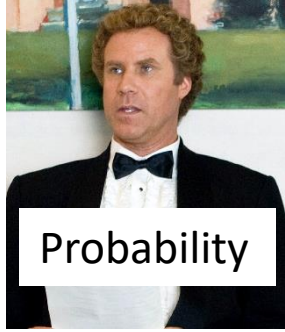
- Suppose that we are interested in testing if a new treatment improves some generic biological measure (e.g., blood pressure, survival time, infection status, hospital admission time, lung capacity)
  - Assume the measure is **continuous** and higher values are preferable
  - Denote the **mean** value **without treatment** as α
  - Denote the **mean** value **with treatment** as $\alpha + \beta$
  - Denote the **variance** of the response as σ^2
- We want find out:
  - What is my best guess for the value of β?
  - How certain/uncertain am I in this guess?
  - Once I have data, can I make statements with evidence about the value of β?

A quick note on the normal distribution

- A **probability distribution** (something that allocates probabilities over a set of possible outcomes)
- Has both a **mean** parameter and a **variance** parameter and is **symmetric**



Build a model



- We call α, β and σ^2 the **model parameters**
- Denote subjects $i = 1,2, \ldots, N$, with treatment indicator $x_i \in \{0,1\}$ and response $y_i \in \mathbb{R}$
  - E.g., the 7th subject may have: $i = 7, x_7 = 1, y_7 = 23$
- If we assume a **normal distribution**, we could say:

$$y_i \sim \mathrm{N}(\alpha + \beta \times x_i, \sigma^2)$$

mean

variance

Estimate the parameters

- Do we just guess the values of α, β and σ^2?
- Our estimates should be **informed** by data
- We could use **linear regression** (or another method)
- Formulas exist but you would rarely (if ever) do this by hand
- Instead, use a computer program like R, SAS or Stata (even Excel)

What could the output look like?

Parameter	Estimate	Standard Error	P-Value
α	30.2	4.3	??
β	5.7	1.1	??
σ	2.1	1.7	??

- For the control arm: $y_i \sim \mathrm{N}(30.2, 2.1^2)$
- For the treatment arm: $y_i \sim \mathrm{N}(35.9, 2.1^2)$

But how should I interpret the uncertainty in these estimates??

I am certain my results are uncertain!

What do we mean by uncertainty?

- In **frequentist** statistics, parameters have **unknown** but **fixed** values
- Because they are **unknown**, we cannot be sure how close our guess/estimate is to the true **fixed** value
- But we can estimate our **uncertainty** in the estimation itself
- We usually do this using **confidence intervals**

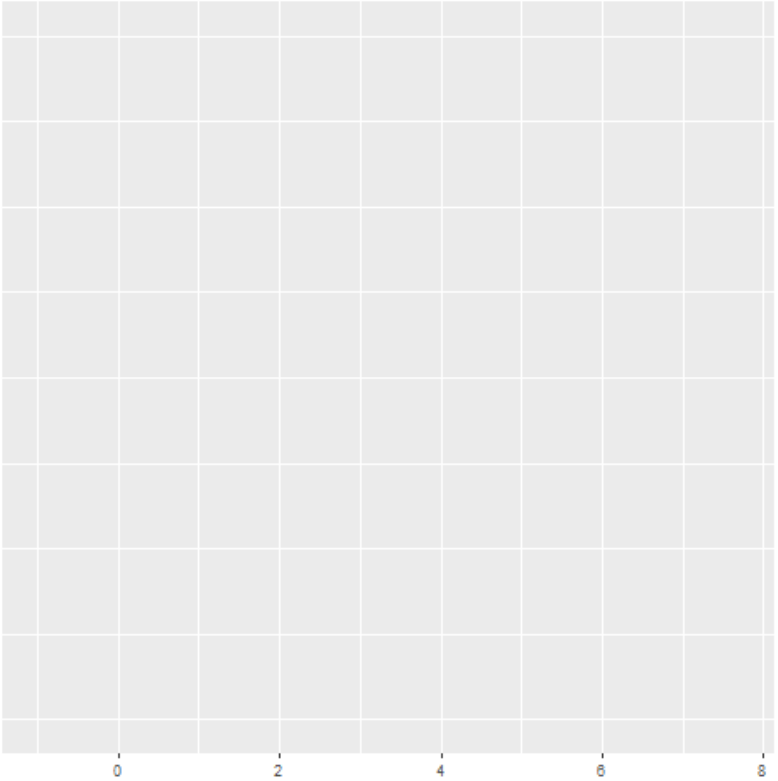
Unconfident computing confidence intervals?

- Just like parameter estimation, formulas exist but you would rarely (if ever) do this by hand
- Usually, confidence intervals are provided in the computer output
- A generic formula looks like: $\overline{x} \pm z_{\alpha/2} \times \frac{s}{\sqrt{n}}$

Parameter	Estimate	Confidence Interval
α	30.2	(21.8, 38.6)
β	5.7	(3.5, 7.9)
$\alpha + \beta$	35.9	(28.4, 43.4)

Interpreting a confidence interval

- Assume $\beta = 5$
- If I repeatedly compute a 95% confidence interval on similar datasets (sampled via the same methods etc.) then 95% of those intervals will contain the true (fixed) value of the parameter



How do I test my hypothesis?

What is a hypothesis test?

- We **assess** the claim of a **hypothesis** against the evidence
- Specifically, we assess the evidence that a **model parameter** takes on a certain value or lies within a certain range
- E.g., one may **hypothesise** that $\beta > 0$ (i.e., that the treatment has some positive benefit)
- We can test this claim using the two **hypotheses:**
  - **Null hypothesis**: $H_0: \beta = 0$
  - **Alternative hypothesis**: $H_1: \beta > 0$

The philosophical argument…

- Proof by **contradiction**
  - Suggest Theory X
  - Find a contradiction (or counter example) to Theory X
  - Therefore, Theory X is false
- Scientific arguments or theories (rarely) can ever be proven
- Instead, we gather evidence to **support** or **counter** a theory
- With a hypothesis test, we aim to assess evidence that **counters** the claim of the null hypothesis, thus **supporting** the alternative hypothesis
- **BUT** the failure to find counter evidence **does not** prove the null hypothesis
- We do this with p-values!

Can you p-lease explain a p-value?

P-values

- A p-value is the *a priori* probability of observing the data (or more extreme) under the **assumption** that the null hypothesis is **true**
- A small p-value is therefore evidence that the data were unlikely to be observed **if** the null hypothesis was **true**
- This is the **counter evidence** against the null hypothesis, and so we **reject** it
- We may therefore claim that the alternative hypothesis is **supported** by the evidence

Back to the example…

- Recall:
  - Null hypothesis: $H_0: \beta = 0$
  - Alternative hypothesis: $H_1: \beta > 0$
- The p-value will be computed via the **linear regression**
- E.g., if the p-value = 0.02 this means that the probability we would have observed data **at least as extreme** as the data we did in fact observe, whilst **assuming** $\beta = 0$, is 0.02

How convinced of the evidence do I need to be?

- For a hypothesis test we need to *a priori* set a **threshold** or **significance level**
- i.e., how small does the p-value need to be to convince me that the null hypothesis is false
- Historically, and preferably in the eyes of grant review panels, this is set at the magical value of 5%
- P-values under 5% are “good” otherwise we just try again or file it away and pretend it never happened

Is this an issue?

- The significance level is also called **type one error**
- This is the probability that I will **incorrectly** reject the null hypothesis
- Things happen by chance!
- If I were to test a faulty claim over and over again with a significance level of 5% then 5% of the time I would **incorrectly** claim it to be true!

What’s all this fuss about sample size?

Why do we care about sample size?

Decreasing the sample size

Increasing the sample size

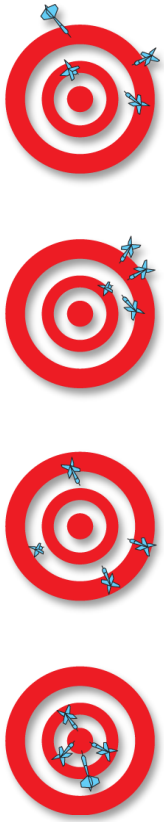
- Save resources!
- Ethics??

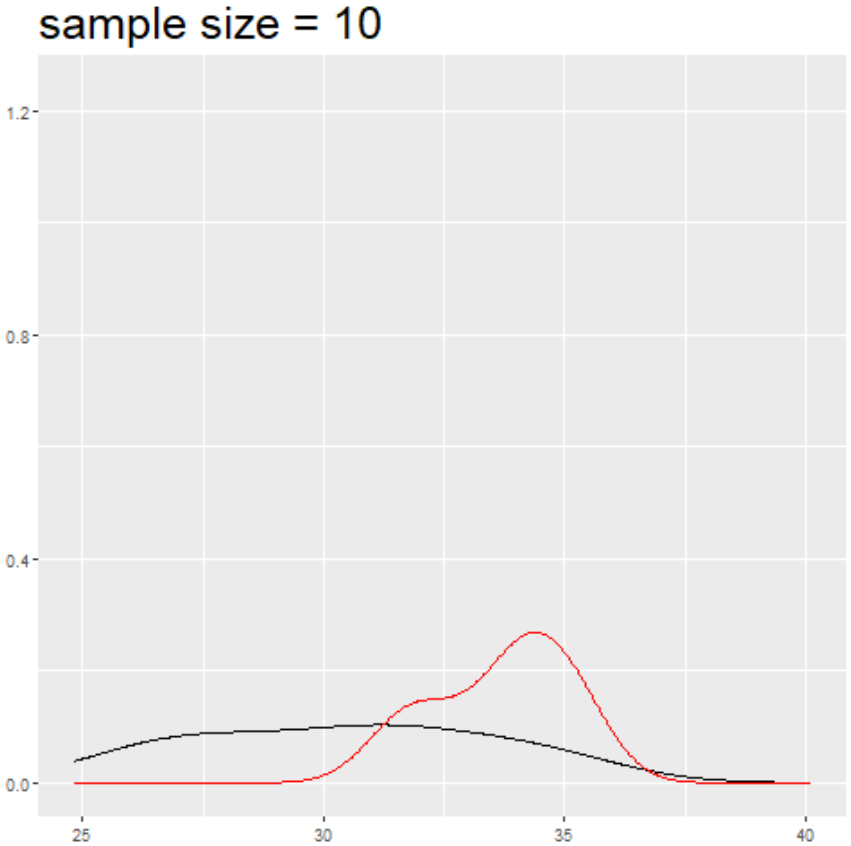
- Increase precision!
- Ethics??

From a (purely) statistical point of view…

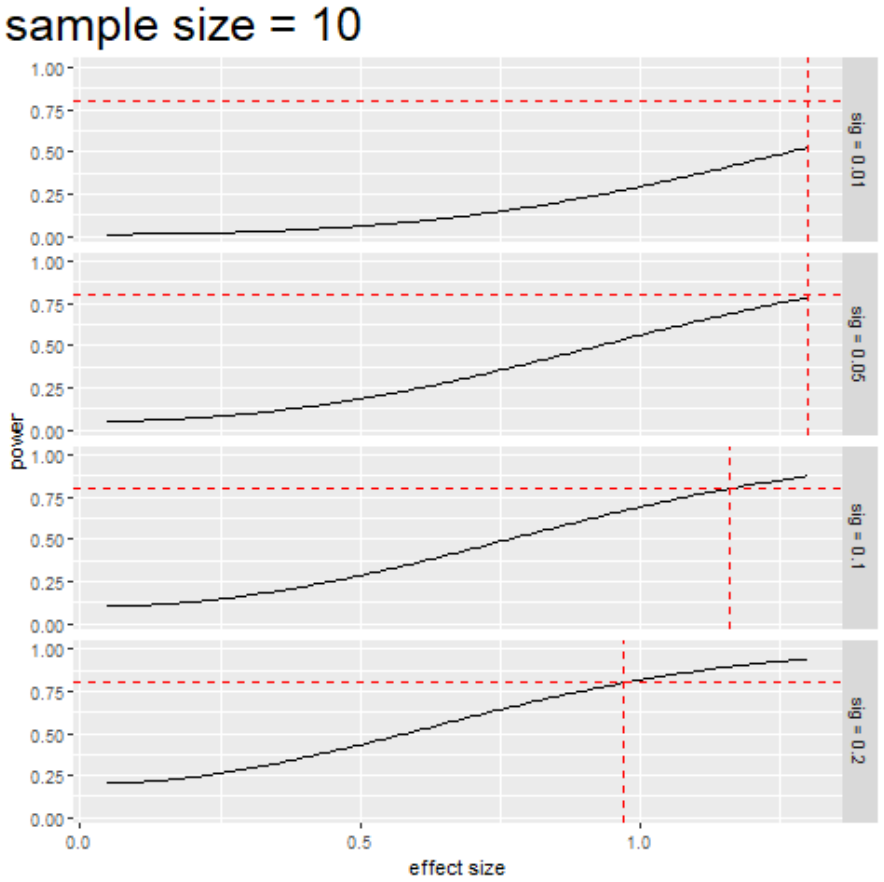
- Large sample sizes are always preferable **with caution**
- At **study design**, we compute the required sample size to achieve the desirable **type one error** and **power**
    - Although, this is usually done backwards!
- **Before** seeing data, the sample size can help us understand the possible behaviour of the trial and guide our interpretation of the results
- **After** seeing data, the sample size no longer matters!

Sample size increases precision





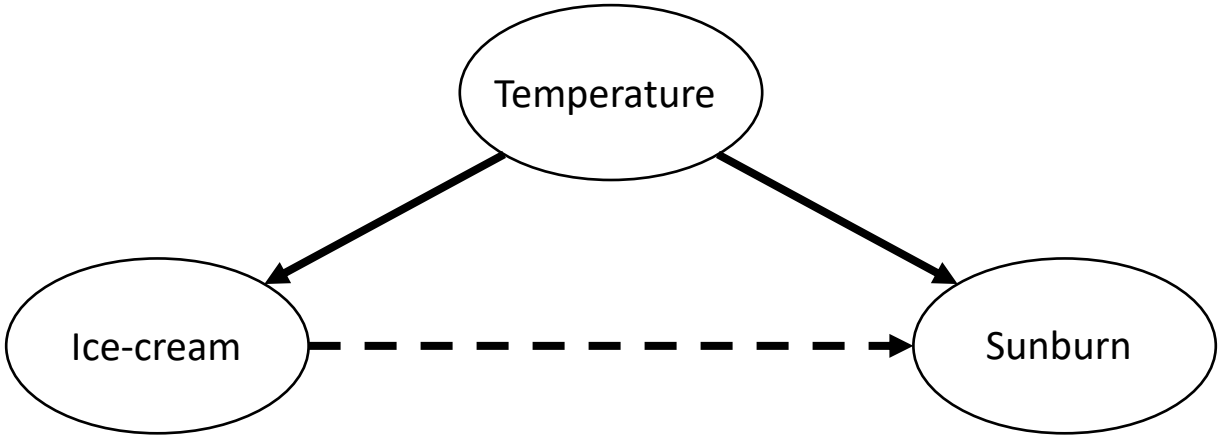
Power


sample size = 10
power
effect size
0.0
0.5
1.0
0.00
0.25
0.50
0.75
1.00
sig = 0.01
sig = 0.05
sig = 0.1
sig = 0.2

Is a confound bound to be found?

Expounding confounding

- A **confound** is an external variable that **influences** both the **dependent** and **independent** variables in the analysis
- Consider the relationship between **eating ice-cream**, **getting sunburnt** and **hot temperatures**



Simulate some data

- Let's simulate some **observational data** and see what happens
- Assume:
  - **Temperature -** $t_i \sim \text{Bernoulli}(0.3)$
  - **Ice-cream -** $x_i \sim \text{Bernoulli}(0.2 + 0.6t_i)$
  - **Sunburn -** $s_i \sim \text{Bernoulli}(0.1 + 0.3t_i)$
- Run analyses **with** and **without** including temperature to assess the relationship between **ice-cream consumption** and **sunburn** for 200 participants

Without temperature

Model: $s_i \sim \text{Bernoulli}^*(\alpha + \beta \times x_i)$

Output:

Parameter**	Estimate	Confidence Interval	P-Value
α	0.12	(0.05, 0.17)	
β	3.35	(0.05, 0.17)	<0.01

With temperature

Model: $s_i \sim \text{Bernoulli}^*(\alpha + \beta \times x_i + \gamma \times t_i)$

Output:

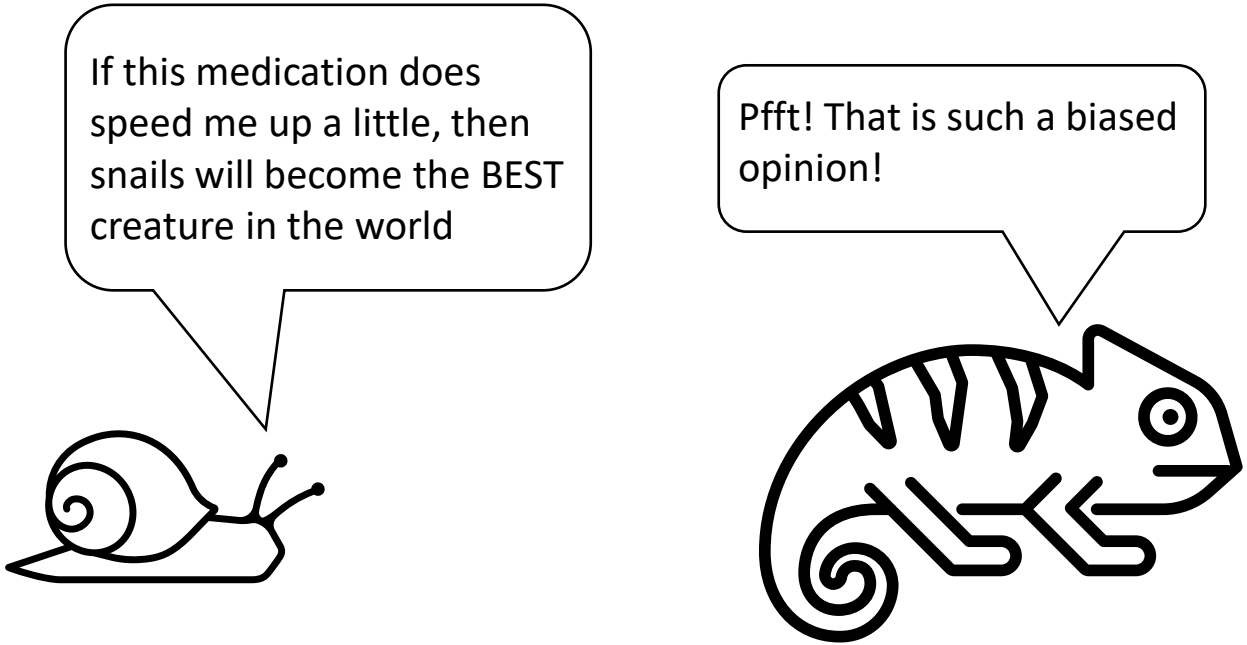
Parameter**	Estimate	Confidence Interval	P-Value
α	0.10	(0.05, 0.17)	
β	1.36	(0.49, 3.66)	0.54
γ	4.74	(1.79, 13.32)	<0.01

*Using a logistic regression model (details omitted)
**Interpret α as the intercept and β and γ as odds ratios

Hounding the confounding

- If we adjust for a confound in the analysis, then its effect disappears
- Without the adjustment it appears that ice-cream is incorrectly causally related to sunburn
- However! It is still true to say that ice-cream and sunburn are **associated** as knowing one informs on the other
- **Caveat:** in complex phenomena it is **not always correct to adjust ("control")** for every variable – it depends on the causal structure!

Bias


If this medication does speed me up a little, then snails will become the BEST creature in the world
Pfft! That is such a biased opinion!

Bias

- Where a **systematic error** in the design, recruitment, data collection or analysis, results in an **incorrect estimation** of the **true effect** of the exposure/intervention on the outcome
- Bias is specifically the tendency to **skew the estimation** in one direction **on average**
- Can be handled with careful design, randomisation and appropriate analysis

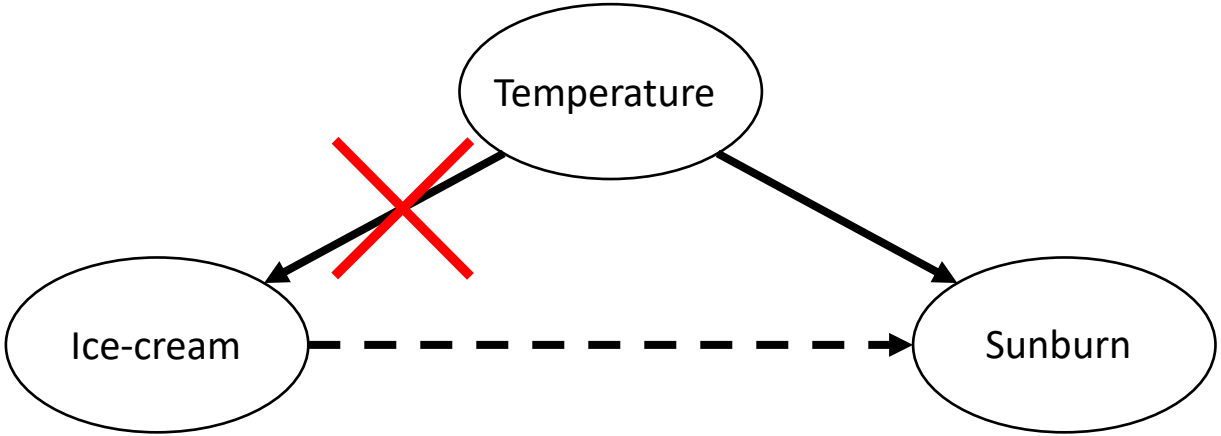
Some examples

- **Selection bias –** some designs only attract “healthy volunteers”
- **Recall bias –** events are easier to remember events than non-events
- **Efficacy bias –** if treatment is known the effect may be distorted (placebo)
- **Survival bias –** outcomes are dependent on participants surviving
- **Analysis bias –** adjusting for a “collider” may introduce bias

What is the point of randomisation?

Why use randomisation?

- **Randomisation** breaks causal relationships
- What if we randomised participants to consume ice-cream (or not)?



Simulate some (more) data

- Assume:
  - **Temperature -** $t_i \sim \text{Bernoulli}(0.3)$
  - **Ice-cream -** $\boldsymbol{x_i \sim \text{Bernoulli}(0.5)}$
  - **Sunburn -** $s_i \sim \text{Bernoulli}(0.1 + 0.3t_i)$

Without temperature

Model: $s_i \sim \text{Bernoulli}^*(\alpha + \beta \times x_i)$

Same statistical model

Output:

Parameter**	Estimate	Confidence Interval	P-Value
α	0.20	(0.12, 0.33)	
β	1.03	(0.49, 2.15)	0.94

P-value no longer significant

Odds ratio accurate

*Using a logistic regression model (details omitted)

**Interpret α as the intercept and β and γ as odds ratios

What just happened?

- Because we **actively randomised** to ice-cream consumption (instead of just **passively observing**), we removed the confounder (temperature)
- It no longer matters whether we include temperature in the model (doing so may increase the precision of estimation but it is not necessary)

Observational studies vs randomised designs

- In an **observational study** the **independent variable** (such as a treatment) is **not under the control** of the researcher
- In a **randomised design** (such as an RCT), the **independent variable** is **randomly** allocated to participants
- This “breaks” the links with any **uncontrolled** variables

Randomised designs

- We may potentially make **causal inference** instead of only **association**
- We may **control** for other variables in order to rule out alternative explanations (e.g., see if temperature affects sunburn in addition to ice-cream consumption)
- These designs may **decrease** the **variability** in the outcome because the participant characteristics and outcome measurement are controlled
- **Caveat:** this will influence **generalisability**

How should I report my results?

What should I report?

- **Estimated effect** – gives direction and magnitude
- **Estimated uncertainty** - confidence intervals or analogous
- **Evidence for conclusion** – p-values or analogous
- **Conclusion** – e.g., treatment is beneficial

But interpret with caution!

Bradford-Hill Criteria for Causality

- Temporal relationship
- Strength of relationship
- Dose-response
- Consistency
- Plausibility
- Consideration of alternative explanations
- Experiment
- Specificity
- Coherence

Austin Bradford Hill: The Environment and Disease: Association or Causation? Proceedings of the Royal Society of Medicine, 58 (1965): 295-300.

Don’t forget to report the methods!

- Results require **justification** and (ideally) should be **reproducible**
- Statistical methods should be detailed including any **assumptions**, the **software** used and the **computational specifications**
- **Ideally**, code should either accompany the publication or be made available online (or at least on request)

Consult standardised reporting guidelines

- Reporting guidelines provide
  - Checklists, depending on type of study design
  - Ensure that published research includes sufficient details for us to critically assess the quality of research

CONSORT, TREND, STROBE, REMARK, STREGA, PRISMA

These guidelines should also be consulted:

before you Design the Study

before you Plan the Analysis

before you Write the Paper

before you Submit the Paper

Where can I get more help?

Where can I find a Statistician?

Perth Children's Hospital:

Free advice through **Telethon Clinical Research Centre**

Telethon Kids Institute (consultancy service):
Biometrics@telethonkids.org.au

UWA (consultancy service):

consulting-cas@uwa.edu.au

The Centre for Applied Statistics, UWA, offers free advice to UWA postgraduate research students

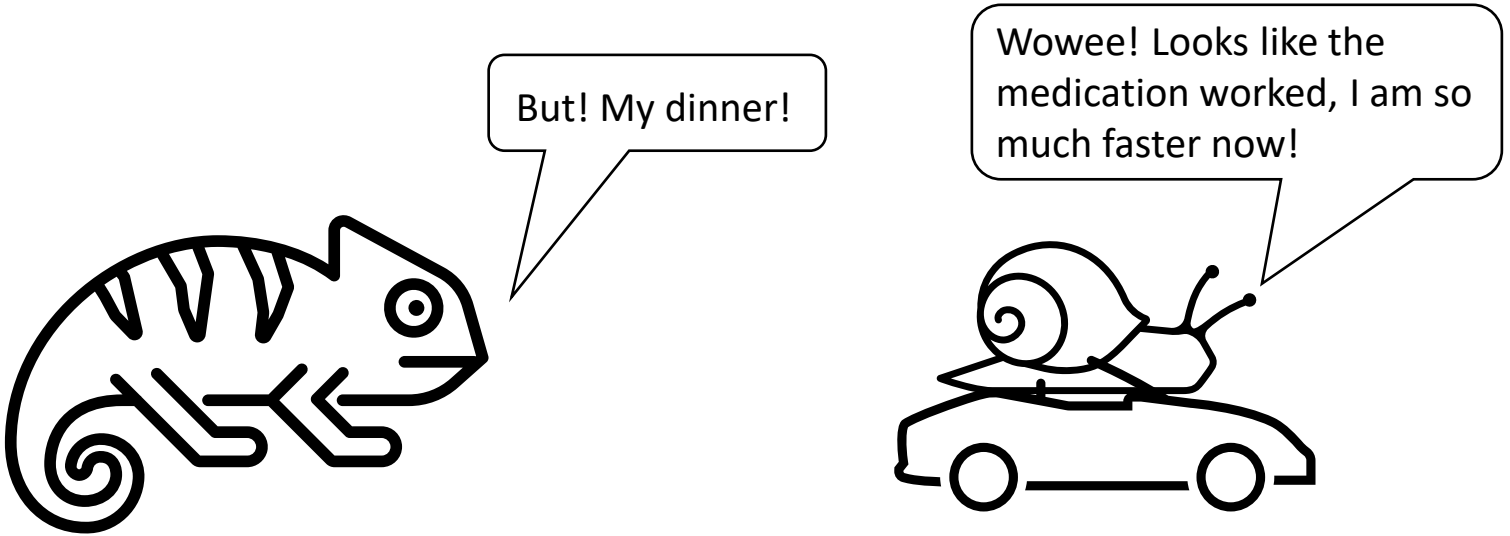
More in handouts

Checklist for talking to a Statistician

- Clear hypothesis
- Proposed study design
- Primary endpoint & estimate of variability
- Clinically relevant effect size
- Estimate of feasible sample size based on budget or potential annual patient recruitment
- Important confounders & source of bias
- Similar publications or systematic reviews
- Bring them cookies

How can I learn more about statistics?

- In the absence of large, randomised, well-controlled clinical trials to address every research question we all need to increase our **statistical literacy**



How can I learn more about statistics?

In person at Perth Children's Hospital:

Attend CAHS REP Research Skills Seminars.

In person at UWA:

The Centre for Applied Statistics provides short courses in statistics which are heavily discounted for students.

Joint Clinical-Statistical Supervision:

If one of your supervisors is a statistician, then you will have "unlimited" access to statistical knowledge/training.

How can I learn more about statistics?

Online: **Data Science Specialization**
Johns Hopkins University

FAQ: You can access the course for free via https://www.coursera.org/specializations/jhu-data-science#courses

This will allow you to explore the course, watch lectures, and participate in discussions for free. To be eligible to earn a certificate, you must either pay for enrolment or qualify for financial aid.

Links in your handouts

Coming up next

8 Mar Using Social Media in Research
Dr Amy Page, UWA

22 Mar Introduction to Good Clinical Practice
Alexandra Robertson, CAHS

Register → **researcheducationprogram.eventbrite.com.au**

We love feedback

A survey is included in the back of your handout, or complete online
https://tinyurl.com/surveyIntroBiostats

✉ ResearchEducationProgram@health.wa.gov.au cahs.health.wa.gov.au/ResearchEducationProgram

✉ ResearchEducationProgram@health.wa.gov.au
cahs.health.wa.gov.au/ResearchEducationProgram