

Common mistakes in statistics and how to avoid them

Data and Decision Science Network, Health Innovations

Professor Marijka Batterham, Dr Bradley Wakefield

@ UOW Statistical Consulting

National Institute of Applied Statistics Research Australia



UNIVERSITY
OF WOLLONGONG
AUSTRALIA



WE ACKNOWLEDGE THAT COUNTRY FOR
ABORIGINAL PEOPLES IS AN
INTERCONNECTED SET OF ANCIENT AND
SOPHISTICATED RELATIONSHIPS.

THE UNIVERSITY OF WOLLONGONG
SPREADS ACROSS MANY INTERRELATED
ABORIGINAL COUNTRIES THAT ARE
BOUND BY THIS SACRED LANDSCAPE,
AND INTIMATE RELATIONSHIP WITH
THAT LANDSCAPE SINCE CREATION.

FROM SYDNEY TO THE SOUTHERN
HIGHLANDS, TO THE SOUTH COAST.

FROM FRESH WATER TO BITTER WATER
TO SALT. FROM CITY TO URBAN TO
RURAL.

THE UNIVERSITY OF WOLLONGONG
ACKNOWLEDGES THE CUSTODIANSHIP
OF THE ABORIGINAL PEOPLES OF THIS
PLACE AND SPACE THAT HAS KEPT ALIVE
THE RELATIONSHIPS BETWEEN ALL
LIVING THINGS.

THE UNIVERSITY ACKNOWLEDGES THE
DEVASTATING IMPACT
OF COLONIZATION ON OUR CAMPUSES
FOOTPRINT AND COMMIT OURSELVES TO
TRUTH-TELLING, HEALING, AND
EDUCATION.



Introductions

- Professor Marijka Batterham
 - Director, Stats Consulting
 - Passionate about data literacy
 - Statistics anxiety/ barriers to developing data literacy skills
 - Interest in lifestyle interventions for chronic diseases
 - Use RStudio/SPSS/ChatGPT most often
 - Like learning new methods & exploring new packages
- Dr Brad Wakefield
 - Statistical Consultant in the Stats Consulting Centre.
 - Rstudio is my go-to but commonly use other packages in teaching and consulting.
 - Interests in data privacy, probability theory, statistical inference, and data analytics.
 - Passion for ethical applications of data science methods in research and industry.
 - Enjoys learning and collaborating with other disciplines and solving real-world problems.

9 common mistakes

1. Presenting/analysing data types incorrectly
2. Misunderstanding hypothesis testing, p -values and significance
3. Not checking assumptions
4. Not adjusting for false positive bias
5. p -Hacking
6. Correlation versus causation
7. Under and over fitting models
8. Overstating results - SPIN
9. Failing to consider sample bias

1. Presenting/analysing data types incorrectly

Types of Data – Don't forget the basics!!



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Two types of data (variables/features/attributes)

Variable Types:

- **Quantitative (Numeric/continuous)** - Variables which have a numeric meaning.
- **Categorical**- Variables which describe a categorisation of an attribute.

- Both types have subgroups

Quantitative Variables

- Variables which have a numeric meaning.
 - Discrete - Numeric variable in countable units/ integers
 - eg number of hospitalisations, number of children
 - Continuous - Numeric variable given on any interval
 - eg weight, height, blood pressure, cholesterol levels

Additional classification:

- Interval Data - Differences are meaningful; zero is not. eg temperature °C
- Ratio Data - Differences and ratios are meaningful. Zero represents absence of property eg Age, distance

Categorical Variables

- Variables which describe a categorisation of an attribute.
 - Nominal - Categorical with no ordering structure.
 - Type of pet (eg Dog, Cat, ...)
 - Type of transport (eg Car, Bus, Train,....)
 - Gender, eye colour, blood type, political party, religious affiliation
 - Ordinal - Categorical with an ordering structure.
 - Primary, secondary, tertiary education level

Categorical Data

- We often have categorical data within our data sets.
- Code it numerically to avoid errors

	Age	Marital Status	Location	Education
231655	18	1. Never Married	1. Urban	1. Primary
86582	24	1. Never Married	1. Urban	4. Undergrad
161300	45	2. Married	2. Regional	3. Certificate
155159	43	2. Married	3. Rural	4. Undergrad
11443	50	4. Divorced	1. Urban	2. High School

Analysing categorical Data

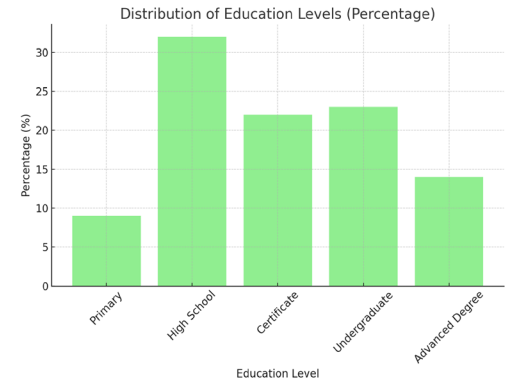
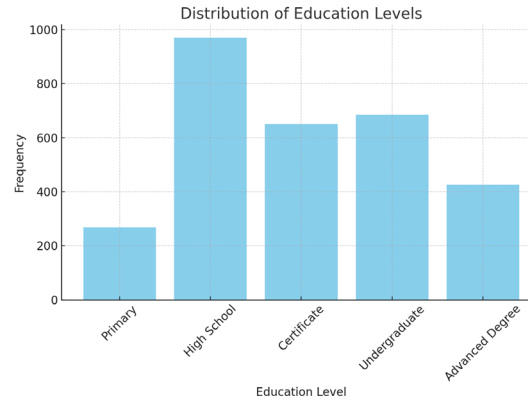
Sometimes we need to reformat the data to analyse it

Name	Age	Gender	Symptoms	Pain in Arm?	Has Fever?
John	52	1. Male	Patient experienced shortness of breath, pain in arm	1. Yes	0. No
Joe	47	1. Male	Pain was experienced by the patient in arm	1. Yes	0. No
Jane	32	2.. Female	Patient exhibited a fever on arrival	0. No	1. Yes
Jack	19	1. Male	Coughing, sore throat, runny nose, fever	0. No	1. Yes

Analysing Categorical Data

To analyse this data we need to use **appropriate** statistics.

Label	Education Level	Frequency	Percentage
1	Primary	268	9
2	High School	971	32
3	Certificate	650	22
4	Undergraduate	685	23
5	Advanced Degree	426	14



Analysing Categorical Data (DOs)

DO

- Use frequency and contingency tables.
- Calculate proportions.
- compare differences in other variables across different groups.
- Use logistic (ordinal/multinomial) regression (when a dependent variable).
- Use bar charts to visualize, or boxplots when comparing groups across a quantitative variable.
- Treat it as a categorical variable in analysis.

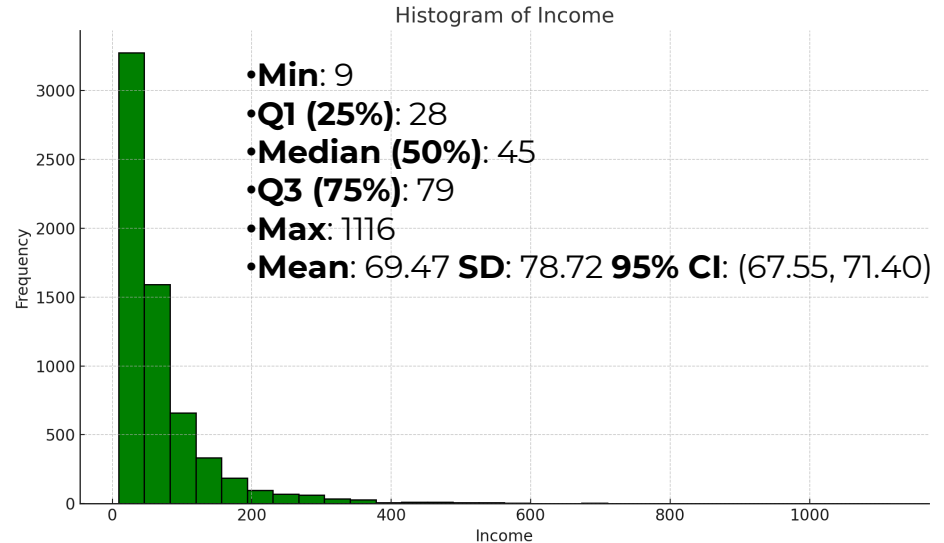
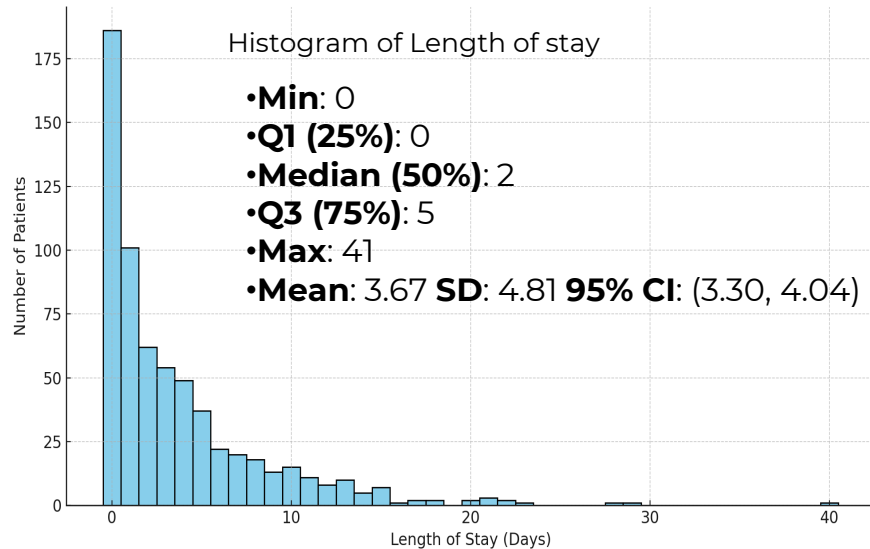
Analysing Categorical Data (DON'TS)

DO NOT

- Compute mean, median, standard deviation.
 - Education data mean 3.01, SD 1.23, median 3 is not meaningful
- Calculate correlations (unless ordinal).
- Fit a linear regression (when a dependent variable).
- Visualise with a line graph or scatterplot (unless ordinal).
- Treat it as a quantitative variable in analysis.

Quantitative Data: Mean or Median?

- If your variable is skewed the mean(SD or CI) do not correctly describe the data, use median, IQR(25th & 75th percentile)
- Particularly if bounded eg by 0 or if there is a high proportion of 0



Often present both

Table 1. *Daily Mean Tally Counts and Estimated Self-Reports of Need-Based Cognitions for Women and Men*

Variable	Men					Women				
	Range	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>n</i>	Range	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>n</i>
Food estimate	2–50	7.6	6.9	5.0	53	2–50	8.0	7.3	6.0	88
Food count	3–111	25.1	23.7	17.7	27	3–53	15.3	10.4	14.9	32
Sleep estimate	0–50	5.8	5.2	4.0	53	0–50	6.2	6.7	4.0	87
Sleep count	3–253	29.0	55.8	10.7	21	2–57	13.4	12.8	8.6	40
Sex estimate	0–50	7.9	8.4	5.0	53	1–50	6.1	7.9	3.0	87
Sex count	1–388	34.2	57.5	18.6	71	1–140	18.6	24.0	9.9	91

Sex on the Brain?: An Examination of Frequency of Sexual Cognitions as a Function of Gender, Erotophilia, and Social Desirability. *Journal of Sex Research*
[Volume 49, Issue 1, 2012](#):pages 69-77

Visualising Data

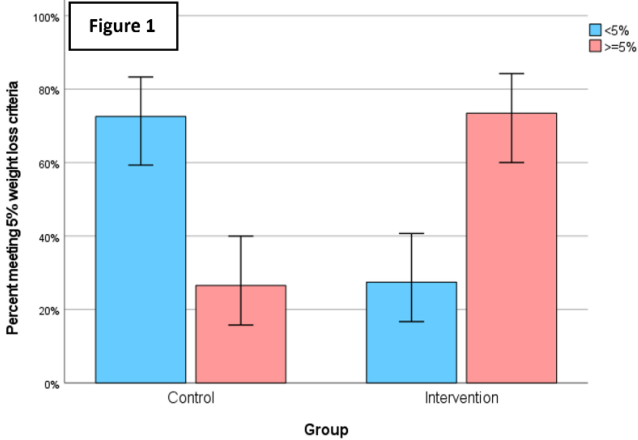


Figure 1 correct use of bar chart to display percentage of participants (with 95%CI) in the control and treatment groups meeting and not meeting the 5% weight loss criteria.

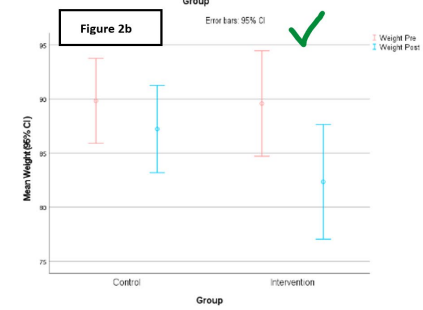
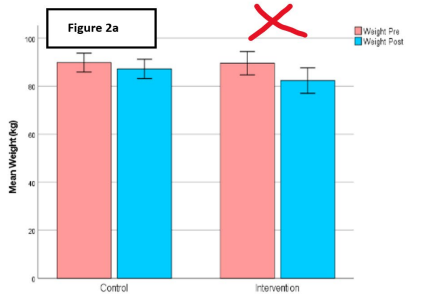
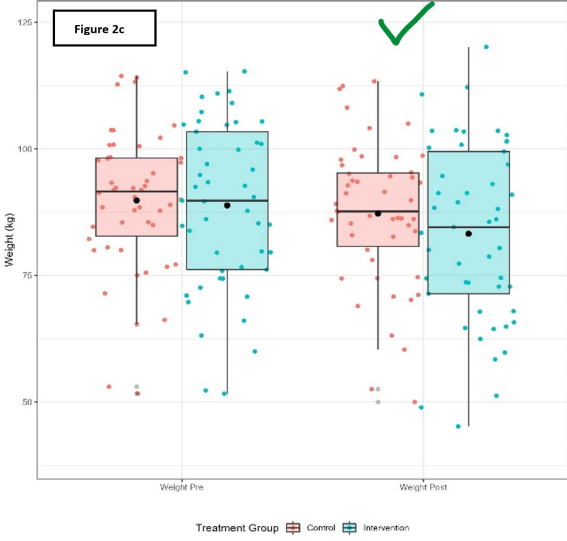


Figure 2a Incorrect use of a bar chart to display mean weights Figure 2b Correct use of a CI plot and Figure 2c boxplot to show mean weights and the distribution of the weight data.



U

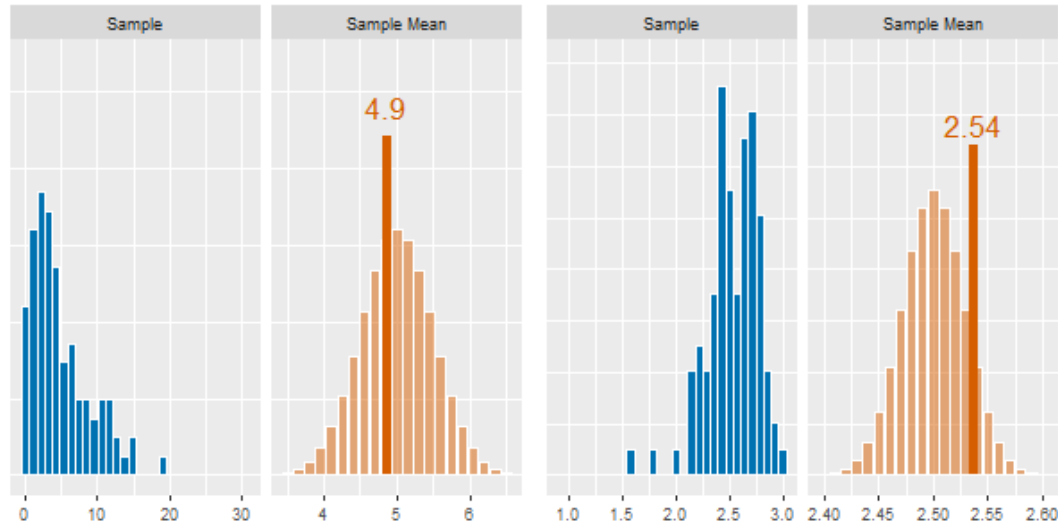
2. Misunderstanding hypothesis testing, p-values and significance

O



How to Estimate the Population?

Given a **random sample**, we can apply **probability theory** ...



Inferential statistics have known probability distributions that relate to the population parameters.

Population vs Sample

The term **mean** can apply to two different quantities

	Population Mean	Sample Mean
	<ul style="list-style-type: none">• Based on the entire population or theoretical distribution.• Notated usually as μ• Is the expected value of the random variable.	<ul style="list-style-type: none">• A calculable statistic based on sample data.• Notated usually as \bar{x} or $\hat{\mu}$.• Is the average value in the sample.

We use **sample mean** as an estimate of **population mean**.

What is Hypothesis Testing?

Hypothesis testing is a statistical analysis used to determine whether there is **sufficient statistical evidence** to reject a specific statement about a population parameter.

i.e. based on my sample data, can I be **reasonably** sure that the mean for the entire population is not equal to 10.

We do not prove things.

At the conclusion of a hypothesis test we either,

- Conclude that we have sufficient evidence to reject a hypothesis.
- Conclude that we do not have sufficient evidence to reject a hypothesis.

The Logic of a Hypothesis Test

We begin every hypothesis test with two hypotheses.

e.g. a null $H_0: \mu = 10$ vs an alternate $H_1: \mu \neq 10$

We **assume the null hypothesis is true.**

Assume the population mean is really 10

We compute a **test statistic** e.g. $T=2.65$.

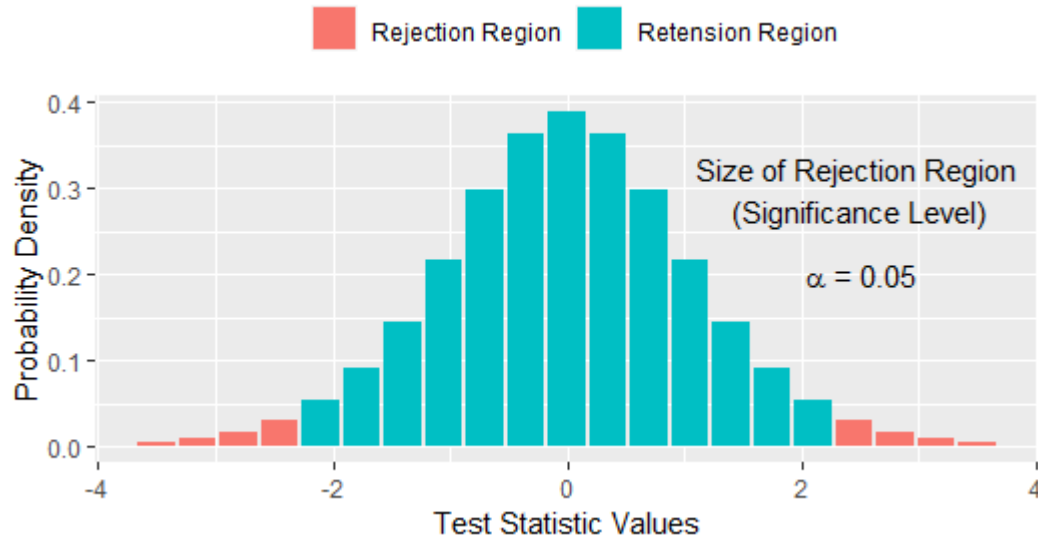
T is given by a formula

Given the null hypothesis and the assumptions of the test, we know the **probability distribution** of the test statistic.

Assuming that the true mean is really 10, we know what are the chances of getting values of T (we know the distribution).

The Logic of a Hypothesis Test

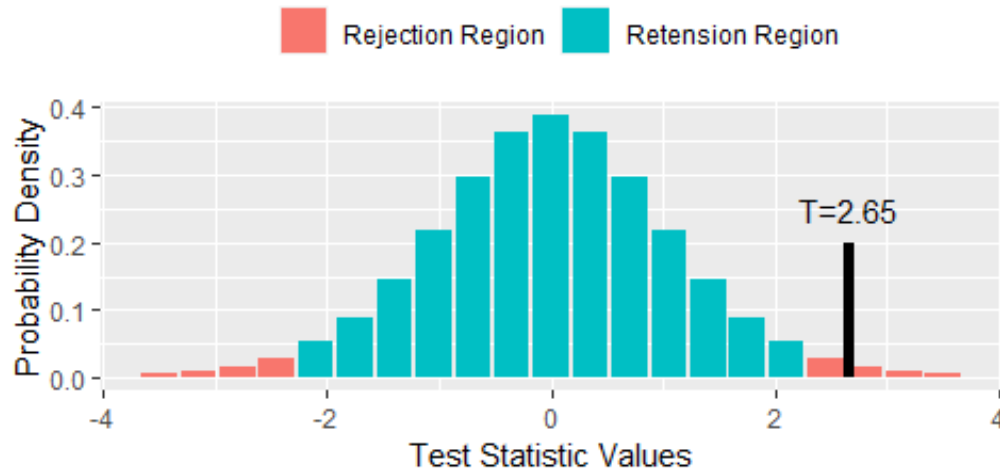
Given a significance level (α), we determine the “extremes” of our test statistic distribution.



The most extreme values with a combined area that add up to α belong to the rejection region.

The Logic of a Hypothesis Test

If our observed test statistic value falls within the rejection region...

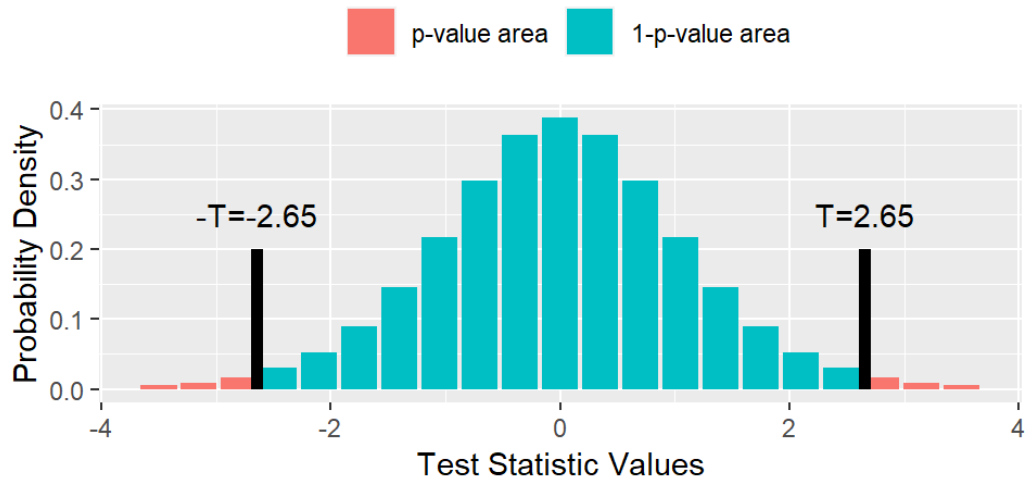


- Our observed value is **extreme** and not consistent with what is expected (were our population mean really 10).
- Evidence suggests our null hypothesis is not true.

What are p -values?

p -values are often used with statistical software to perform hypothesis testing.

p -values are the probability, under the null hypothesis, of achieving test statistic values at least as extreme as the result actually observed.



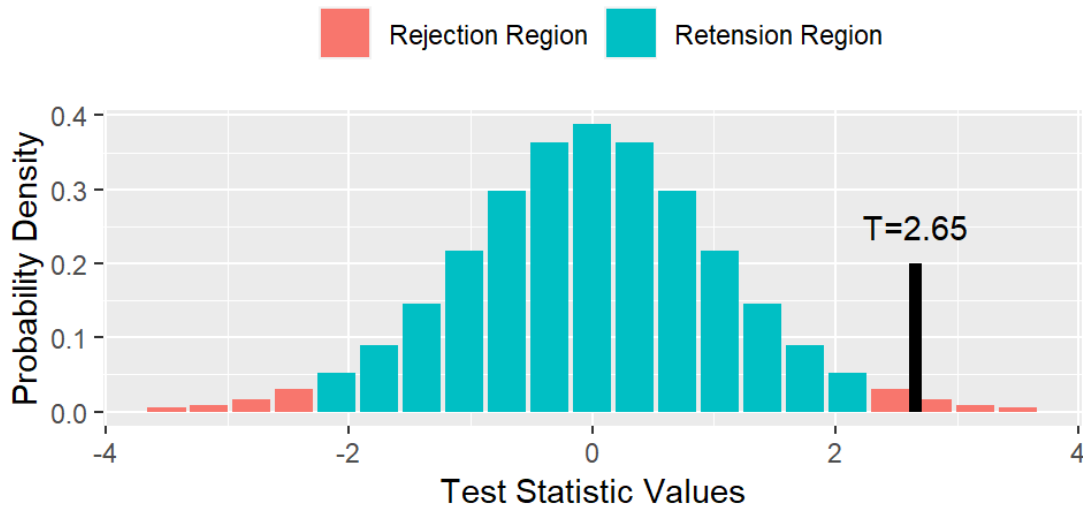
The total area in orange is the p -value.

What are p -values?

To determine whether to reject a null hypothesis,

- If $p\text{-value} \leq \alpha$ **reject the null hypothesis.**
- If $p\text{-value} > \alpha$ **do not reject the null hypothesis.**

Consider the p -value for the value below.



“Significance” is Significant

In statistics, **significant** means “significant with respect to some hypothesis test and significance level”.

Do not use the term **significant** unless you have performed a hypothesis test.

Statistically significant results account for the size of the difference relative to the variation in the data.

Results with large differences in mean can be described as “considerable” or “notable” but not significant.

“Significance” is Significant

Example 1

Consider the difference between 100 observations of the two variables $X \sim N(200, 1000^2)$ and $Y \sim N(300, 1000^2)$.

With sample means $\bar{x} = 63$ and $\bar{y} = 205$.

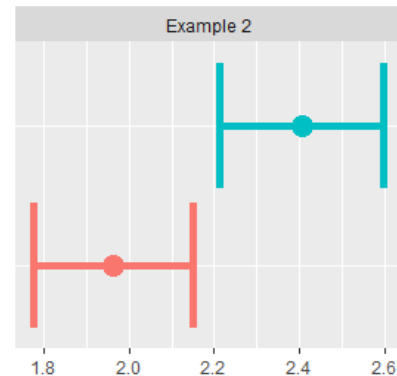
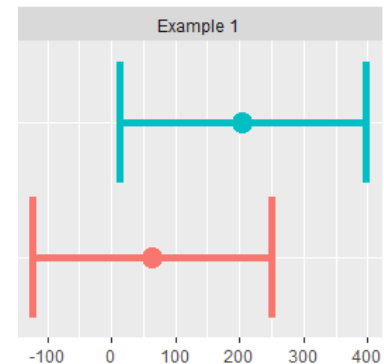
Performing a Student's t-test we conclude there **is not a significant difference** in means.

Example 2

Consider the difference between 100 observations of the two variables $X \sim N(2.1, 1)$ and $Y \sim N(2.5, 1)$.

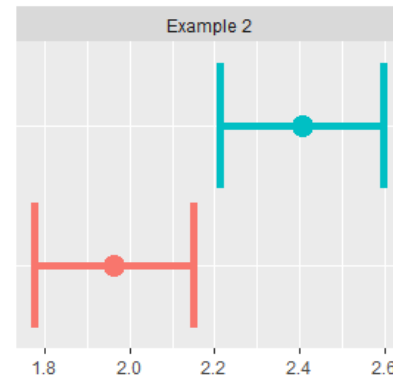
With sample means $\bar{x} = 1.96$ and $\bar{y} = 2.40$.

Performing a Student's t-test we conclude there **is a significant difference** in means.



What is a Confidence Interval?

A 95% confidence interval for a parameter is an interval, based on sample statistics, that if we were to repeatedly compute for different random samples of the population indefinitely would contain the true parameter value 95% of the time.



- The true value of a parameter is **not random** in frequentist statistics.
- Computing probabilities of the true value being any value(s) **does not make sense**.
- **Do not say that there is 95% probability your true value is in the confidence interval.**
- **You can say that you are 95% confident that your interval contains your true value.**
- Confidence intervals give a sense of the uncertainty in the estimate within the scale of the estimate.
- Some statisticians and journals prefer CIs over p -values.

Degrees of Freedom

The degrees of freedom of a statistical analysis are the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample.

Degrees of freedom can be thought of as the number of independent pieces of information used to obtain an estimate.

Consider the sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

This estimate has $n-1$ degrees of freedom.

3. Not checking assumptions



Assumption Checking

It wouldn't be statistics without assumptions

EVERY TEST HAS ASSUMPTIONS

ASSUMPTIONS ALLOW US TO INFER BASED ON PROBABILITY

A **parametric** assumption refers to an assumption about how our distribution is “parameterised.” Parametric tests require parametric assumptions.

Non-parametric tests still have assumptions usually around independence between observations and consistency in distribution (i.i.d.)

Assumption Checking

ALWAYS CHECK THE SPECIFIC ASSUMPTIONS OF YOUR ANALYSIS

Results may be invalid when assumptions are not met

Results may not make sense when assumptions are not met

There are often multiple assumptions to models to consider including:

- Constant variances, homogeneity of variance, homoscedasticity,...
- Independence between observations ...
- Sufficient sample sizes (asymptotic methods) ...
- Linearity and no multicollinearity ...
- Normality of the data or residuals

All relevant assumptions should be checked.

Assumption Checking

BE CAREFUL WITH ASSUMING NORMALITY

Normality is a common parametric assumption, but it is not always about the raw values.

Some analyses require normal distributed data.

- e.g. t-tests, ANOVA, pearson correlations, etc

Some analyses require aspects (like residuals) to be normally distributed, but not the actual data.

- e.g. **linear regression**, ARIMA modelling etc.

Some analyses do not have a parametric assumption (non-parametric) ... but they do have others.

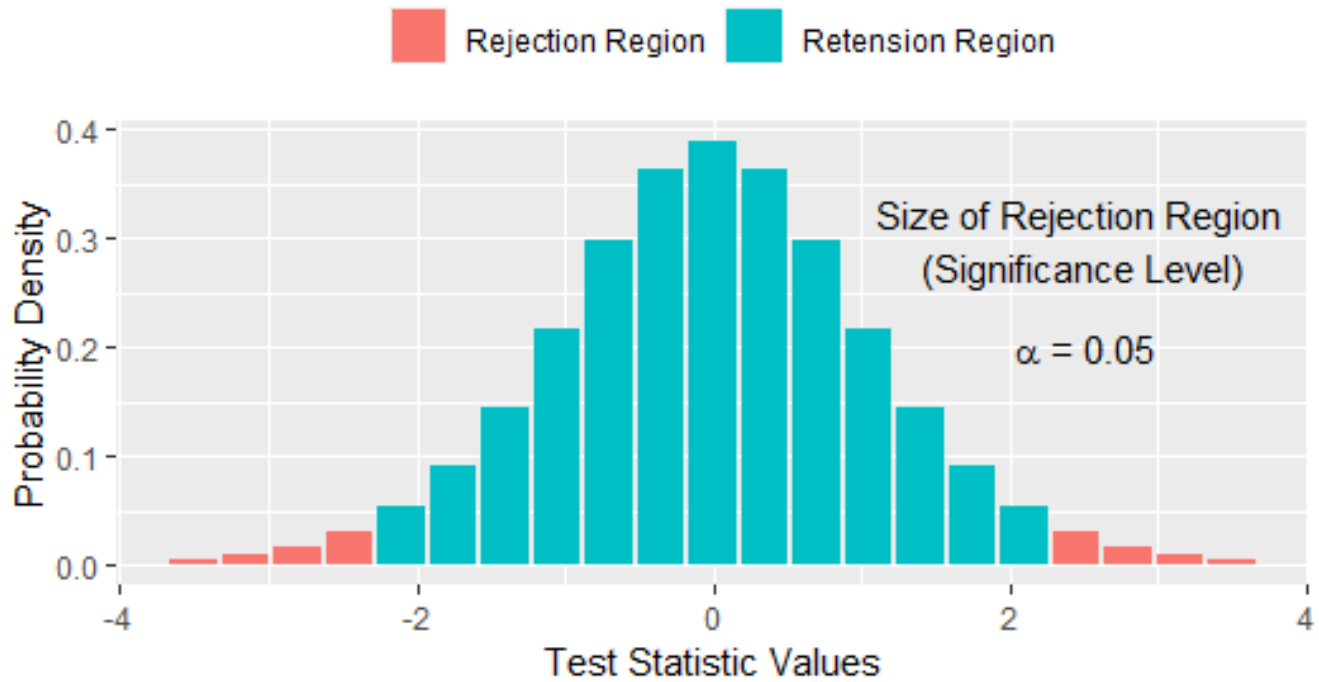
- e.g. Wilcoxon test, Mann-Whitney test, Kruskal-Wallis test, etc.

Remember to still check other assumptions!

4. Not adjusting for false positive bias

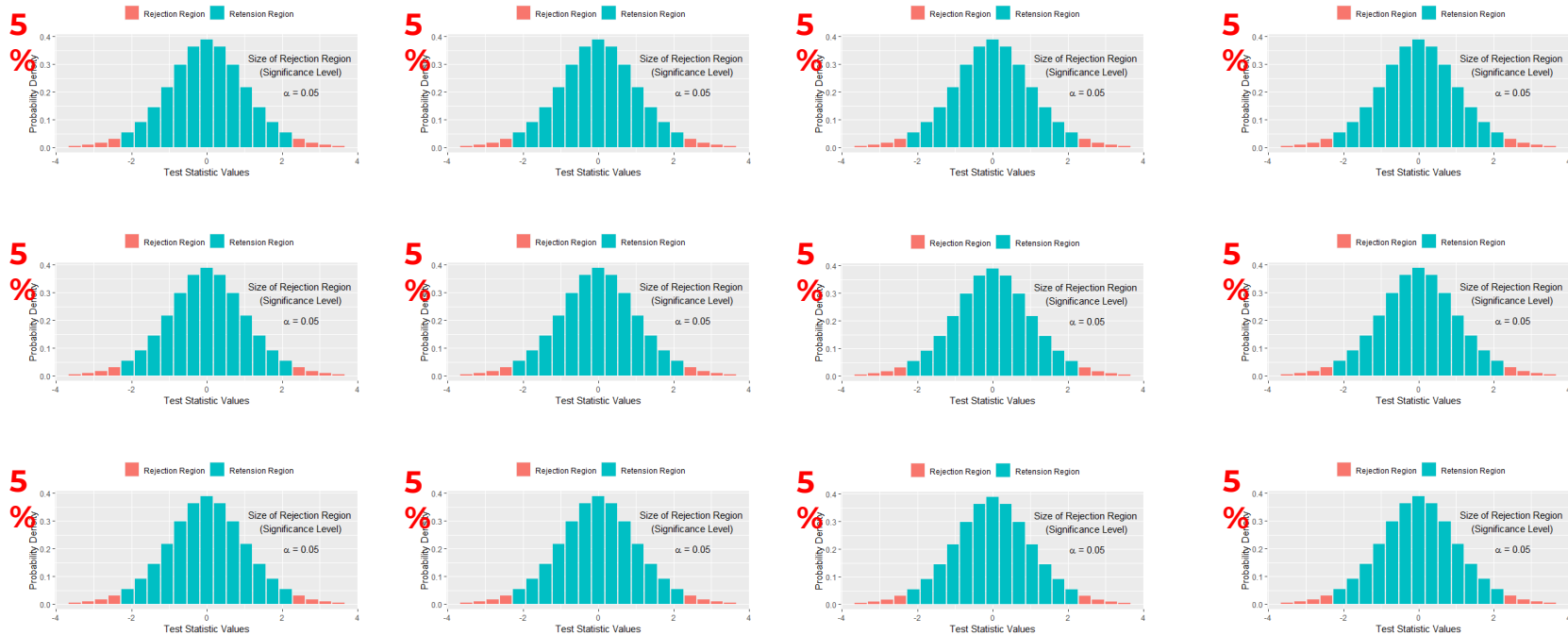


REMEMBER THIS PLOT...



5% of the time a true null hypothesis will be rejected by random chance!
This is a false positive!

WHAT IF WE DID MORE THAN ONE TEST?



\approx **46%** chance at least one true null hypothesis will be rejected by random chance – *assuming independence*.

Example: Simulated BMI Data

- We randomly sample BMI, Gender, and Age Group to be independent and for there to be **no differences** between groups.
- **BUT LOOK AT THE POST-HOC TESTS**

Group 1		Group 2		p-value
<25 years	Female	25-39 years	Female	0.020
<25 years	Female	40-59 years	Female	0.549
<25 years	Female	60+ years	Female	0.171
25-39 years	Female	40-59 years	Female	0.411
25-39 years	Female	60+ years	Female	0.195
40-59 years	Female	60+ years	Female	0.204
<25 years	Male	<25 years	Female	0.059
<25 years	Male	25-39 years	Female	0.230
<25 years	Male	40-59 years	Female	0.801
<25 years	Male	60+ years	Female	0.407
<25 years	Male	25-39 years	Male	0.101
<25 years	Male	40-59 years	Male	0.008
<25 years	Male	60+ years	Male	0.049
25-39 years	Male	<25 years	Female	0.874
25-39 years	Male	25-39 years	Female	0.110

Group 1		Group 2		p-value
25-39 years	Male	40-59 years	Female	0.283
25-39 years	Male	60+ years	Female	0.766
25-39 years	Male	40-59 years	Male	0.136
25-39 years	Male	60+ years	Male	0.395
40-59 years	Male	<25 years	Female	0.086
40-59 years	Male	25-39 years	Female	0.456
40-59 years	Male	40-59 years	Female	0.443
40-59 years	Male	60+ years	Female	0.833
40-59 years	Male	60+ years	Male	0.655
60+ years	Male	<25 years	Female	0.260
60+ years	Male	25-39 years	Female	0.378
60+ years	Male	40-59 years	Female	0.711
60+ years	Male	60+ years	Female	0.960
25-39 years	Male	40-59 years	Female	0.283
25-39 years	Male	60+ years	Female	0.766

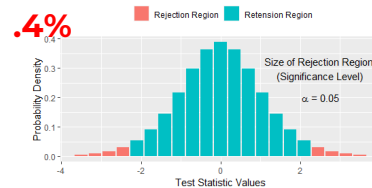
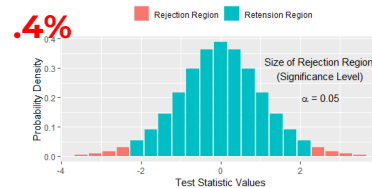
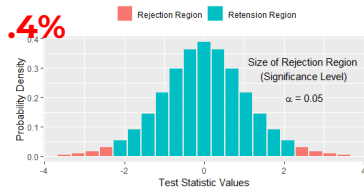
THIS OVERALL RATE IS CALLED THE FALSE POSITIVE DISCOVERY RATE

So how do we address it?

- Limit hypothesis testing to **planned analyses**.
- Perform a p-value **correction** when performing multiple comparisons.
- Ensure interpretations acknowledge false positive bias.
- Significance isn't everything – use p-values, effect sizes, and confidence intervals.

WHAT IF WE DID MORE THAN ONE TEST?

BUT USING $\alpha_{adj} = \frac{5\%}{12} = 0.4\%$



$\approx 4.6\%$ chance at least one true null hypothesis will be rejected by random chance – *assuming independence*.

Example: Simulated BMI Data (Revisited)

- We randomly sample BMI, Gender, and Age Group to be independent and for there to be **no differences** between groups.
- **BUT LOOK AT THE POST-HOC TESTS AFTER ADJUSTING**

Group 1		Group 2		Adjusted p-value
<25 years	Female	25-39 years	Female	0.548
<25 years	Female	40-59 years	Female	1
<25 years	Female	60+ years	Female	1
25-39 years	Female	40-59 years	Female	1
25-39 years	Female	60+ years	Female	1
40-59 years	Female	60+ years	Female	1
<25 years	Male	<25 years	Female	1
<25 years	Male	25-39 years	Female	1
<25 years	Male	40-59 years	Female	1
<25 years	Male	60+ years	Female	1
<25 years	Male	25-39 years	Male	1
<25 years	Male	40-59 years	Male	0.227
<25 years	Male	60+ years	Male	1
25-39 years	Male	<25 years	Female	1
25-39 years	Male	25-39 years	Female	1

Group 1		Group 2		Adjusted p-value
25-39 years	Male	40-59 years	Female	1
25-39 years	Male	60+ years	Female	1
25-39 years	Male	40-59 years	Male	1
25-39 years	Male	60+ years	Male	1
40-59 years	Male	<25 years	Female	1
40-59 years	Male	25-39 years	Female	1
40-59 years	Male	40-59 years	Female	1
40-59 years	Male	60+ years	Female	1
40-59 years	Male	60+ years	Male	1
60+ years	Male	<25 years	Female	1
60+ years	Male	25-39 years	Female	1
60+ years	Male	40-59 years	Female	1
60+ years	Male	60+ years	Female	1
25-39 years	Male	40-59 years	Female	1
25-39 years	Male	60+ years	Female	1

5. *p*-Hacking



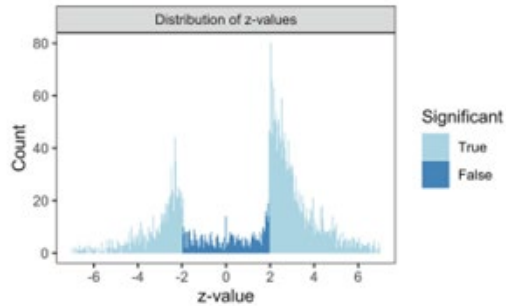
p -Hacking: what is it?



- Refers to researchers “searching” for significant results, by collecting or selecting data or statistical analyses until nonsignificant results become significant.
- Term developed out of the reproducibility crisis
- ‘publish or perish’ culture encourages searching for significance

p -Hacking: How do we know it exists?

- Publication bias for significant results (journal/researcher)



JSMS 2023:26:164-168

p -Hacking: How is it done?

- Results from “Researcher degrees of freedom”, the process of decision making along the analysis pathway
- Stop collecting data once $P < 0.05$
- Analyse many variables but only report those with $P < 0.05$
 - selective choice of dependent variables
 - Selective choice of independent variables
- Collect and analyse data on many groups but only report those with $P < 0.05$ (subgroup analysis)
- Use covariates which result in $P < 0.05$
- Exclude observations/outliers so $P < 0.05$
- Transform the data so $P < 0.05$
- Redefine scales (composite scores, item deletions in scales SPSS)
- Discretising variables (eg long/short exposure, left/right political orientation) arbitrary cut-off
- Alternate hypothesis tests (use a different method)
- Favourable imputation methods
- Incorrect rounding of P values
- Often multiple combinations of these

A “real” example

- “We asked 20 University of Pennsylvania undergraduates to listen to either “When I’m Sixty-Four” by The Beatles or “Kalimba” by Mr Scruff.
- Then, in an ostensibly unrelated task, they indicated their birth date (mm/dd/yyyy) and their father’s age.
- We used father’s age to control for variation in baseline age across participants.
- An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to “When I’m Sixty-Four” (adjusted M = 20.1 years) rather than to “Kalimba” (adjusted M = 21.5 years), $F(1, 17) = 4.92, p = .040$.”

What really happened.

Simmons et al Psychological Science 2011;22(11):1359-66

Table 3. Study 2: Original Report (in Bolded Text) and the Requirement-Compliant Report (With Addition of Gray Text)

Using the same method as in Study 1, we asked 20³⁴ University of Pennsylvania undergraduates to listen only to either “**When I’m Sixty-Four**” by The Beatles or “**Kalimba**” or “Hot Potato” by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. **Then, in an ostensibly unrelated task, they indicated only their birth date (mm/dd/yyyy) and** how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with “computers are complicated machines,” **their father’s age**, their mother’s age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as “the good old days,” and their gender. **We used father’s age to control for variation in baseline age across participants.**

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to “When I’m Sixty-Four” (adjusted M = 20.1 years) rather than to “Kalimba” (adjusted M = 21.5 years), $F(1, 17) = 4.92, p = .040$. Without controlling for father’s age, the age difference was smaller and did not reach significance ($M_s = 20.3$ and 21.2 , respectively), $F(1, 18) = 1.01, p = .33$.

A multitude of sins

- Researcher's use their “degrees of freedom” to employ multiple strategies

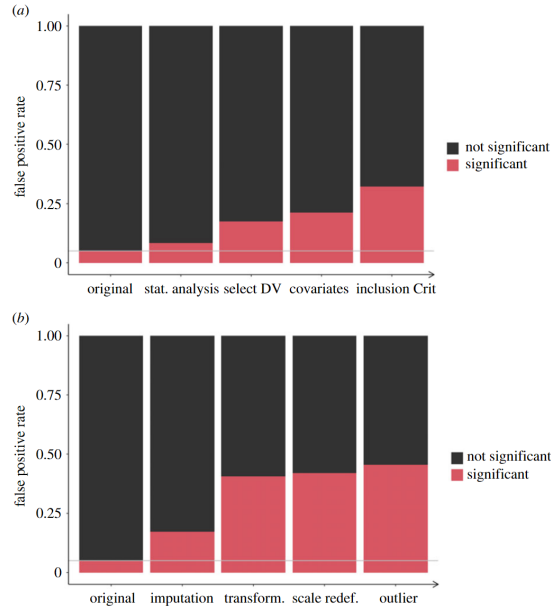


Figure 13. Two scenarios for *p*-hacking ‘workflows’ and their effect on false-positive rates. (a) *p*-hacking workflow in a *t*-test. (b) *p*-hacking workflow in a regression analysis. *p*-hacking strategies are applied sequentially; at each stage, incorrect rounding with a rounding level of $p < 0.051$ is applied in addition. The solid grey line shows the nominal α -level of 5%.

t-test: 1. 3 different analyses, 2. 5 correlated DV, 3. 3 correlated covariates, 4. restrict sample based on 3 binary grouping variables.

regression: 1. 5 different imputation methods, 2. different transformations to DV and IV, 3. random scale redefinition (5 item scale) 4. different outlier removal strategies.

p-Hacking : How to stop it

- Just DON'T do it, use your “researcher degrees of freedom” wisely
- Preregistration with clearly defined statistical analysis plans
- Publishing protocol papers BEFORE you collect the data
- Provide code for all data management and manipulation as well as the analysis
- Bayesian analysis maybe less prone to p hacking

6. Correlation vs Causation



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Correlation vs Causation

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between variables (-1, 0, 1).

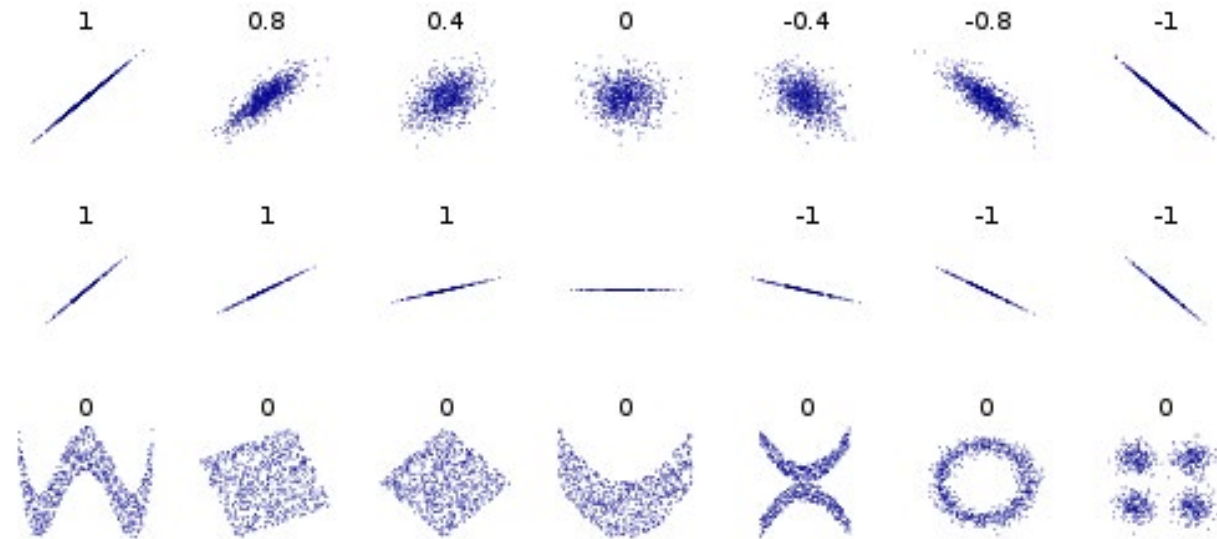
A **linear** correlation between two variables suggests as one increases, the other increases (if positively correlated) or decreases (if negatively correlated).

Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events.

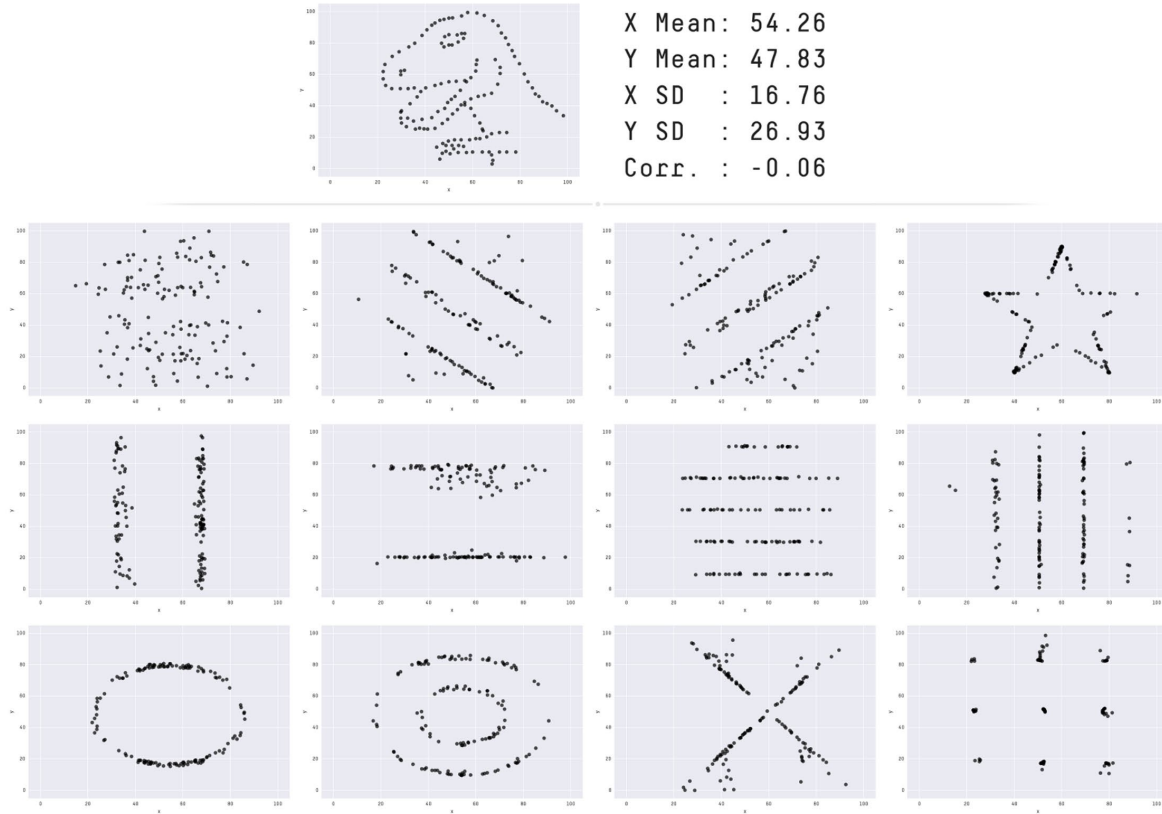
A correlation between variable X and Y does not mean X causes Y (or vice versa).

ALWAYS PLOT AND
VISUALISE YOUR DATA

If there is no correlation it does not mean there is no relationship



Source: wikipedia



ABOVE: Fig. 2. The Datasaurus Dozen. While different in appearance, each dataset has the same summary statistics (mean, standard deviation, and Pearson's correlation) to two decimal places.

Lurking variables

- Often there is a 3rd variable that maybe causing the relationship

A \longrightarrow B

A \longrightarrow C

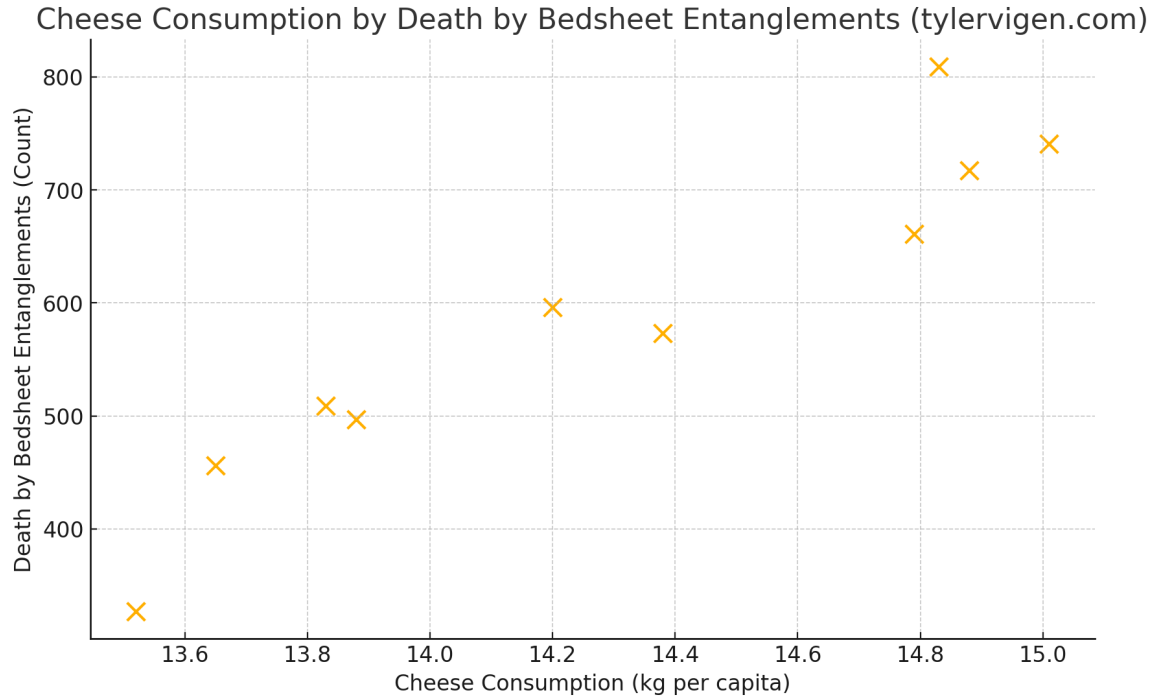
If A increases, then B & C increase together, appears B \longrightarrow C

A \longrightarrow B \longrightarrow C

If you only measure A & C, it appears A \longrightarrow C

- Time is a common third variable

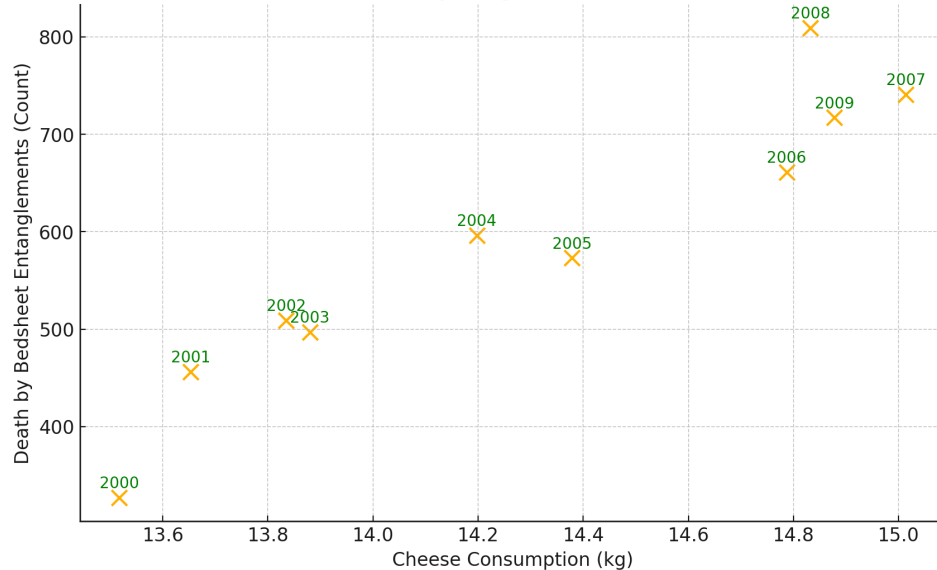
Spurious Correlations



Source: <https://www.tylervigen.com/spurious-correlations>

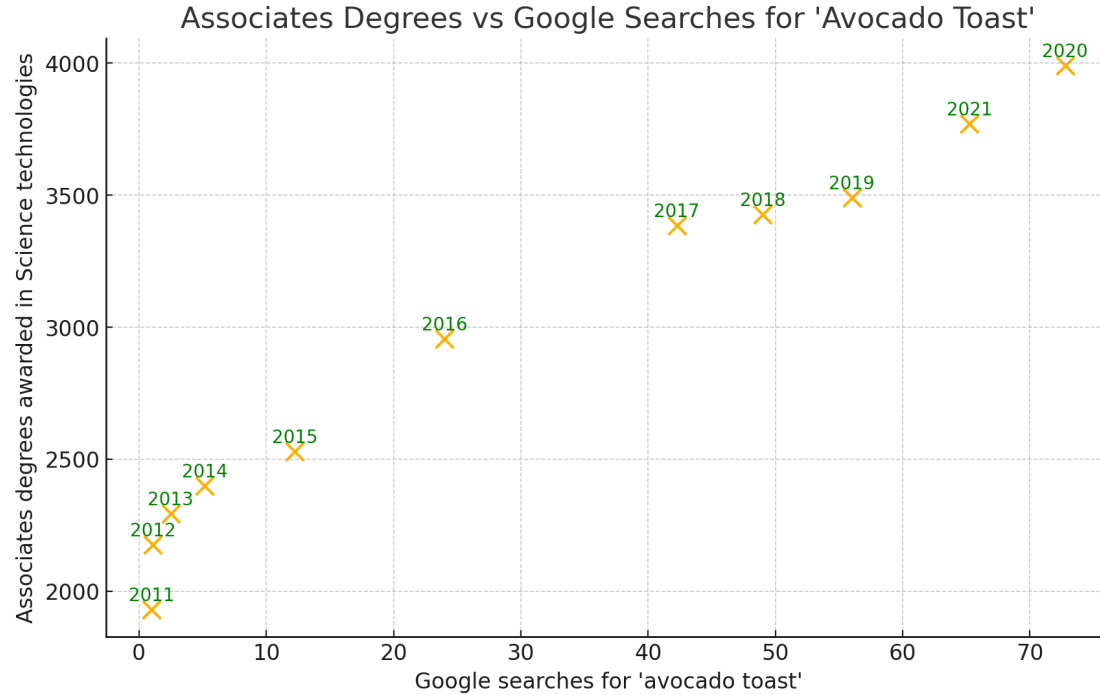
Spurious Correlations

Cheese Consumption vs Death by Bedsheet Entanglements
(tylervigen.com)



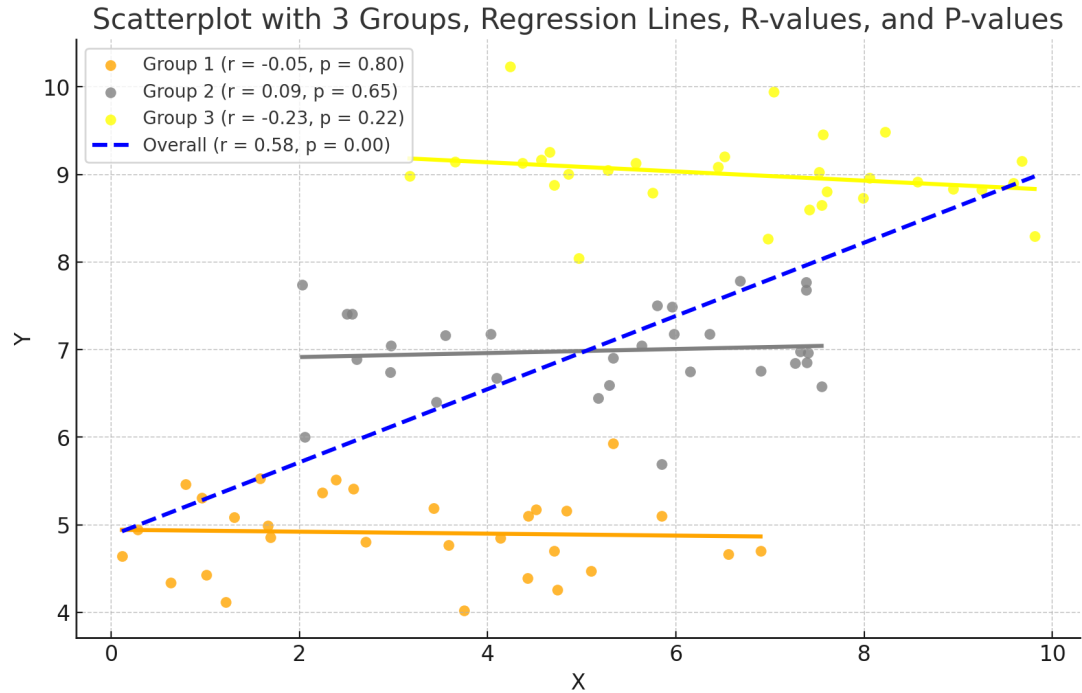
Variable	Correlation	Years
Master's degrees awarded in Health professions and related programs	r=0.99	10yrs
Gasoline pumped in Guatemala	r=0.99	32yrs
Hydropower energy generated in Peru	r=0.98	32yrs
Associates degrees awarded in Philosophy and religious studies	r=0.98	11yrs
The number of school teachers in Florida	r=0.98	12yrs
Number of Las Vegas Hotel Room Check-ins	r=0.98	24yrs
Wind power generated in Morocco	r=0.98	22yrs
Wind power generated in South Africa	r=0.98	19yrs
The number of truck drivers in Virginia	r=0.98	12yrs
The number of nurse practitioners in Hawaii	r=0.98	10yrs
Bachelor's degrees awarded in Mathematics and statistics	r=0.98	10yrs
Popularity of the first name Elena	r=0.98	32yrs
Johnson & Johnson's stock price (JNJ)	r=0.98	20yrs
Total geothermal power generated globally	r=0.98	32yrs
Average number of comments on Mark Rober YouTube videos	r=0.98	11yrs
The number of physician assistants in Texas	r=0.97	19yrs
Google searches for 'avocado toast'	r=0.97	14yrs
Berkshire Hathaway's stock price (BRK.B)	r=0.97	20yrs
Wind power generated in Turkiye	r=0.97	24yrs
Annual Revenue of Walt Disney Company	r=0.97	31yrs
Cigna's stock price (CI)	r=0.97	20yrs
Wind power generated in Chile	r=0.97	21yrs
Votes for the Democratic Presidential candidate in Massachusetts	r=0.97	8yrs
Electricity generation in Nepal	r=0.97	32yrs
Google's Advertising Revenue	r=0.97	21yrs
Google's annual advertising revenue	r=0.97	21yrs
Google searches for 'reddit'	r=0.97	14yrs
Electricity generation in Guatemala	r=0.97	32yrs
Associates degrees awarded in Engineering	r=0.97	11yrs
The number of dentists in Florida	r=0.97	18yrs
Geothermal power generated in Germany	r=0.97	18yrs
Solar power generated in Suriname	r=0.97	12yrs
Votes for the Democratic Presidential candidate in Arizona	r=0.97	8yrs
Constellation Brands' stock price (STZ)	r=0.97	20yrs
Google searches for 'how to learn python'	r=0.96	15yrs
Lockheed Martin's stock price (LMT)	r=0.96	20yrs
Google's Annual Global Revenue	r=0.96	20yrs
Total likes of The Game Theorists YouTube videos	r=0.96	13yrs
US Annual Tax Revenue	r=0.96	32yrs
Popularity of the first name Sage	r=0.96	32yrs
Gender pay gap in the U.S.	r=0.96	32yrs
PepsiCo's stock price (PEP)	r=0.96	20yrs
The number of solar photovoltaic installers in California	r=0.96	10yrs
The number of nurse practitioners in California	r=0.96	10yrs
Popularity of the first name Cyrus	r=0.96	32yrs
The Home Depot's stock price (HD)	r=0.96	20yrs
Intel Corporation's annual revenue	r=0.96	32yrs
The Charles Schwab Corporation's stock price (SCHW)	r=0.96	20yrs
McDonald's stock price (MCD)	r=0.96	20yrs

Source: <https://www.tylervigen.com/spurious-correlations>

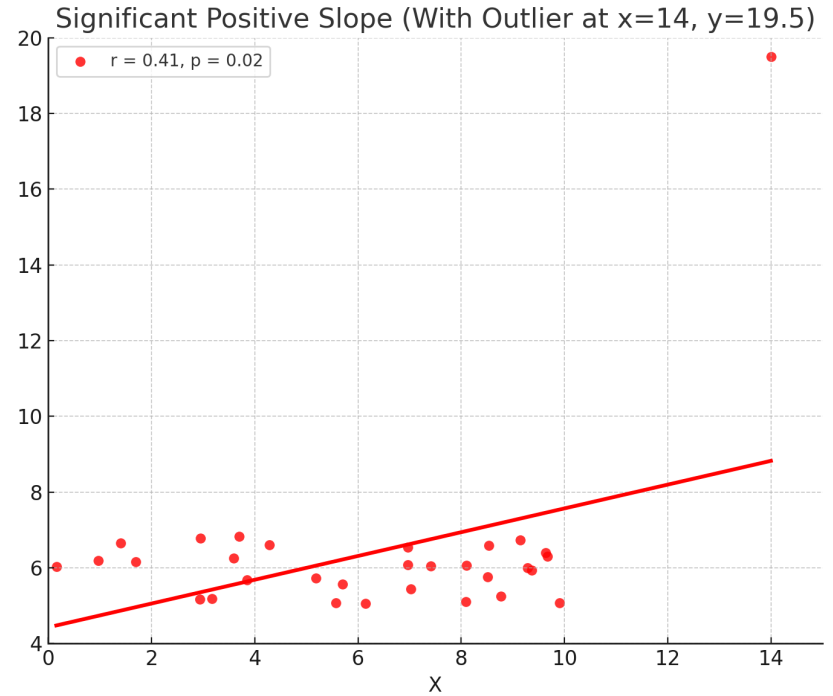
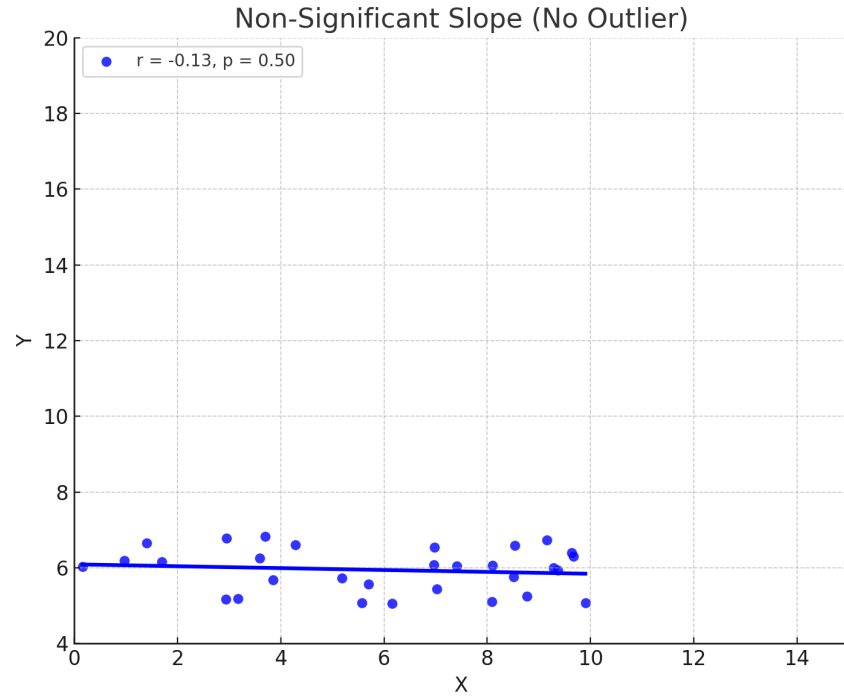


Source: <https://www.tylervigen.com/spurious-correlations>

Clusters



Outliers



Correlation vs Causation

In order to make **causal inference** we usually need to conduct an experiment or a controlled study.

A controlled study involves controlling external sources of variation and performing a specific intervention.

Inference around **correlation** cannot be made on uncontrolled studies.

Implementing a good **study or experimental design** ensures you can make quality inference.

ALWAYS PLOT AND
VISUALISE YOUR DATA

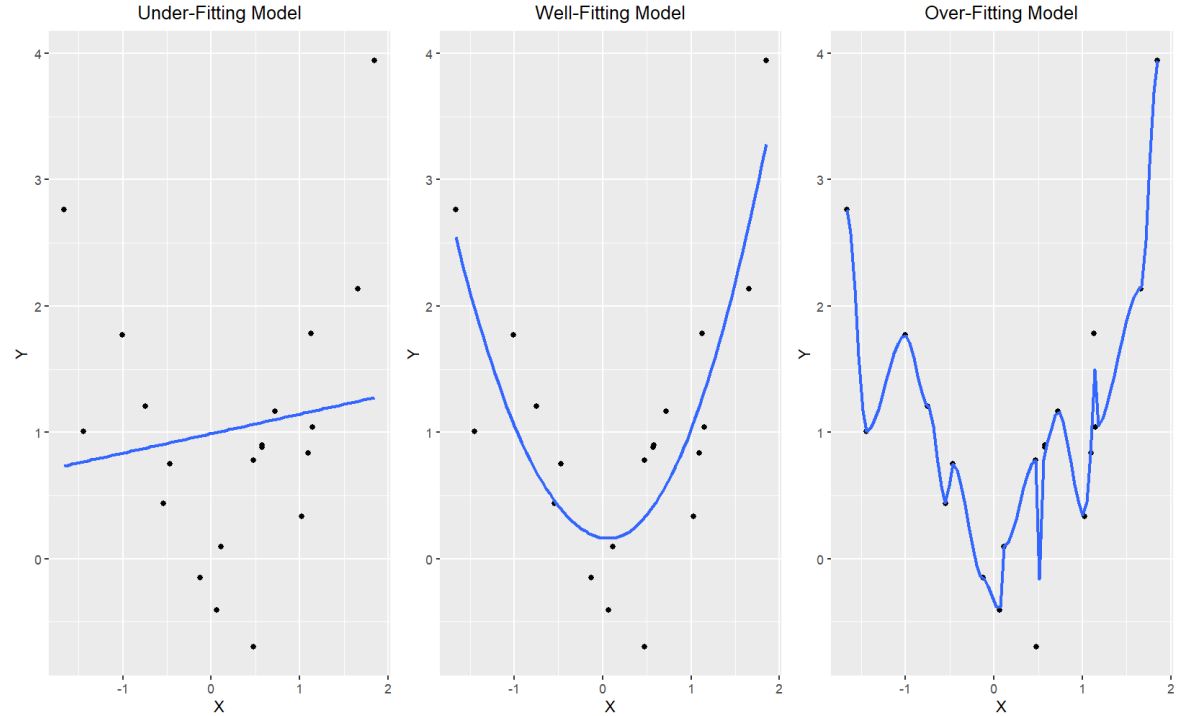
7. Under & over fitting



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

What is Over-Fitting?

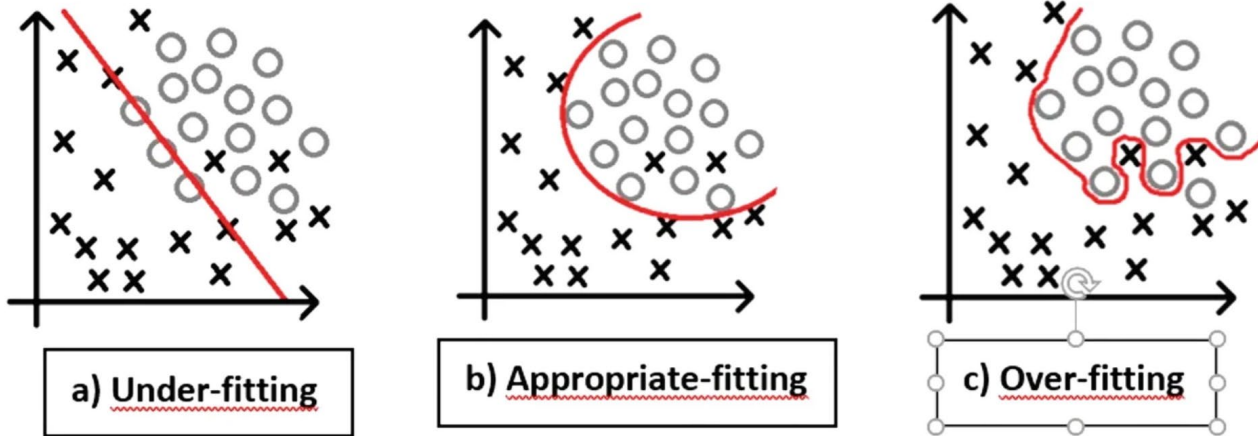
Over-fitting is the process of performing analysis too “closely” to an insufficient number of data points leading to non-generalisable results.



- Over-fit models have too many parameters needing to be estimated from the data available to estimate them.
- Fits noise and inaccurate data
- Common issue in machine learning
- Common in multiple regression – too many variables

Fig. 4.1

From: Overfitting, Model Tuning, and Evaluation of Prediction Performance



Schematic illustration of three models for classification: (a) M1 with underfitting, (b) M2 with appropriate fitting, and (c) M3 with overfitting

Montesinos López, O.A., Montesinos López, A., Crossa, J. (2022). Overfitting, Model Tuning, and Evaluation of Prediction Performance. In: Multivariate Statistical Machine Learning Methods for Genomic Prediction. Springer, Cham. https://doi.org/10.1007/978-3-030-89010-0_4

8. Overstating results: Spin and type II error



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

SPIN

- “specific reporting strategies, whatever their motive, to highlight that the experimental treatment is beneficial, despite a statistically nonsignificant difference for the primary outcome, or to distract the reader from statistically nonsignificant results”
- Boutron et al JAMA, May 26, 2010—Vol 303, No. 20
- Common in many disciplines, many SR,MA surgery, oncology, bariatric surgery
- Spin may consist of; spin in the title, a focus on a statistically significant secondary outcome and linguistic spin (eg “particularly large”, “much greater”)
- Analysis of 567,758 RCTs found the most prevalent phrases were “marginally significant” (7,735 RCTs), “all but significant” (7,015), “a nonsignificant trend” (3,442), “failed to reach statistical significance” (2,578), and “a strong trend” (1,700). PLOS Biology 2022 <https://doi.org/10.1371/journal.pbio.3001562>

Overstating results: Spin

- "Analysis of CCI levels across the adaptation phase showed that the effect of group **approached significance** ($P=0.052$). However, the effect of time was not statistically significant ($p=0.2$)."
- "For flexion, there was a **nonsignificant trend** for reduced matched torques in the self-initiated condition ($p=0.085$). Moreover, there was a 9% increase in H-reflex amplitude following the experimenter-initiated trials, although this increase was not statistically significant ($p=0.14$)"
- Héroux ME, et al. BMJ Open 2022;12:e060976.

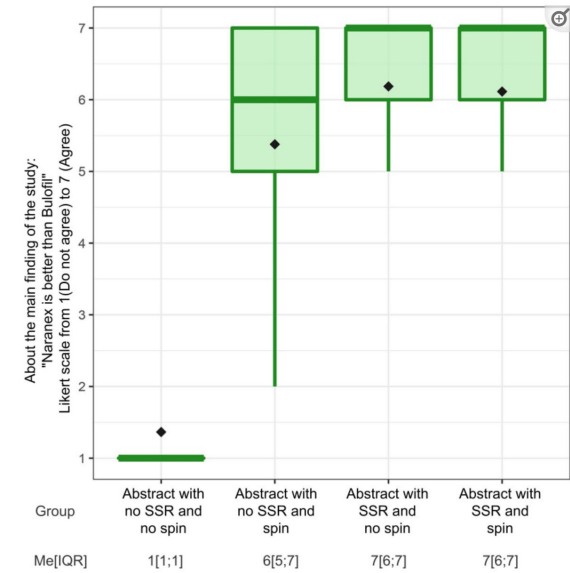
- In all, 110 eligible RCTs with **nonsignificant primary outcomes** were appraised. The title was reported with spin in 8 (7%) articles. Forty-four (40%) included abstracts and 39 (35%) main texts were classified as having spin in at least 1 section. The level of spin was high in 16 (14%) abstract and 19 (19%) main-text “Conclusions” sections. Twenty-five articles **(23%)** recommended the intervention of interest despite a nonsignificant primary outcome.
- *Annals of Surgery* [265\(6\):p 1141-1145, June 2017](#). Reporting of RCTs with non significant outcomes published in high-impact surgical journals

The effect of Spin; an RCT

- 297 health students and professionals
- Rated 4 abstracts based on “Naranex better than Bulofil”

Figure 2

<p>Results In total, 120 patients were randomised to receive either Naranex (n=63) or Bulofil (n=57). The pain disability (RMDQ score) at 6 months was reduced by 4.87 ± 1.83 in the Naranex group and by 3.56 ± 1.52 in the Bulofil group: between-group difference of 1.31 points (-0.08 to 2.70, p=0.06), no statistically significant difference. We found no statistically significant difference in pain intensity: between-group difference of 0.59 points (-0.09 to 1.27, p=0.07). Healthcare resource use, safety and tolerability were not statistically different: patients with non-severe adverse events were 5/63 in the Naranex group and 6/57 in the Bulofil group (p=0.89).</p>	<p>Results In total, 120 patients were randomised to receive either Naranex (n=63) or Bulofil (n=57). The pain disability (RMDQ score) at 6 months was reduced by 4.87 ± 1.83 in the Naranex group and by 3.56 ± 1.52 in the Bulofil group. Naranex improved the RMDQ score at 6 months compared with Bulofil: between-group difference of 1.31 points (-0.08 to 2.70, p=0.06). Moreover, the RMDQ score was much better for compliant people in the Naranex group compared with those in the Bulofil group: between-group difference of 2.4 points (0.26 to 4.54, p=0.02). For women, pain intensity improvement was much better with Naranex compared with Bulofil: between-group difference of 0.91 points (0.10 to 1.72, p=0.04). Non-severe adverse events were less common in the Naranex group (5/63) than in the Bulofil group (6/57), p=0.89.</p>	<p>Results In total, 120 patients were randomised to receive either Naranex (n=63) or Bulofil (n=57). The pain disability (RMDQ score) at 6 months was reduced by 4.87 ± 1.83 in Naranex group and by 3.56 ± 1.52 in Bulofil group: statistically significant between-group difference of 1.31 points (0.08 to 2.54, p=0.04). We found a statistically significant difference in pain intensity in favour of Naranex: between-group difference of 0.59 points (0.09 to 1.09, p=0.04). Healthcare resource use, safety and tolerability were not statistically different: patients with non-severe adverse events were 5/63 in the Naranex group and 6/57 in the Bulofil group (p=0.89).</p>	<p>Results In total, 120 patients were randomised to receive either Naranex (n=63) or Bulofil (n=57). The pain disability (RMDQ score) at 6 months was reduced 4.87 ± 1.83 in Naranex group and by 3.56 ± 1.52 in the Bulofil group. Naranex improved the RMDQ score at 6 months compared with Bulofil: statistically significant between-group difference of 1.31 points (0.08 to 2.54, p=0.04). Moreover, the RMDQ score was much better for compliant people in the Naranex group compared with those in the Bulofil group: between-group difference of 2.4 points (0.26 to 4.54, p=0.02). Pain intensity improvement was better with Naranex: between-group difference of 0.59 points (0.09 to 1.09, p=0.04). That improvement was much better for women: 0.91 points (0.10 to 1.72, p=0.04). Non-severe adverse events were less common in the Naranex group (5/63) than in the Bulofil group (6/57), p=0.89.</p>
---	--	--	---



Readers' assessment of the superiority of 'Naranex' compared with 'Bulofil' readers' assessment of the superiority of Naranex compared with Bulofil after reading their allocated abstract of a randomised controlled trial reported with or without SSRs and with or without spin. Scores are based on a Likert scale, ranging from 0 (do not agree) to 7 (agree). Boxes represent median observations (horizontal rule) with 25th and 75th percentiles of observed data (top and bottom of the box). The diamonds represent the mean. The end of the vertical line represents the minimum values. IQR, considering first and third quartiles. Me, median; SSR statistically significant result.

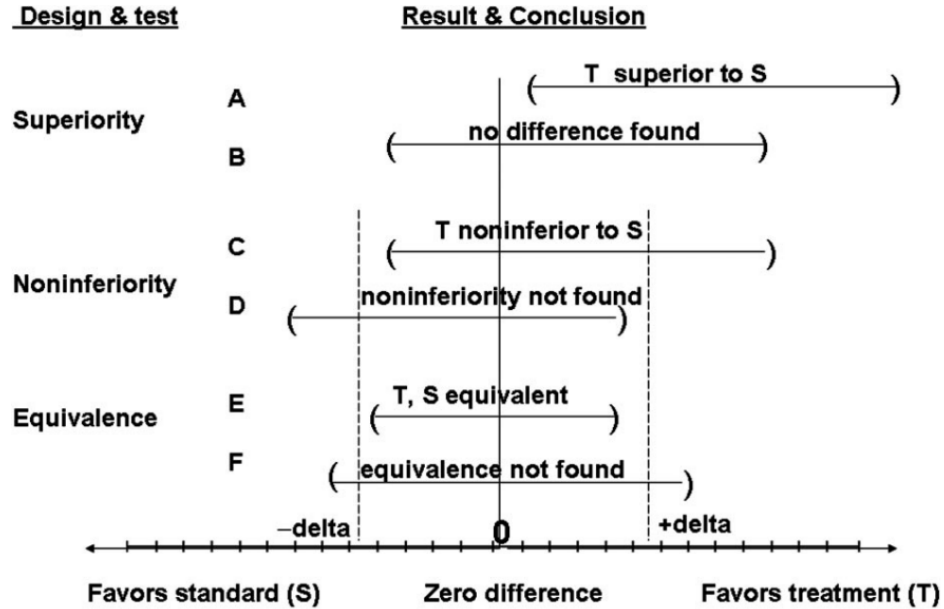
So, what do you say?

- There was no statistically significant difference!!

- There was no statistically significant difference **does not mean** they are not different or that they are the same (Equivalent). $P > 0.05$ usually shows there is “an absence of evidence of a difference”
- Statistics notes: Absence of evidence is not evidence of absence, Bland & Altman 1995 BMJ doi: <https://doi.org/10.1136/bmj.311.7003.485>
- If you want to show two treatments/groups are the “same” you need to swap paradigms to assess **Equivalence**

Batterham et al BJJ doi:10.1017/S0007114516000040

Equivalence testing



9. Failing to consider sample bias

IT IS NOT ALL ABOUT SAMPLE SIZE



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Q: What makes data **representative**?

A: IT'S ABOUT WHO WE ASK AND HOW WE SELECT THEM

or in other words

A: IT'S ABOUT THE SAMPLING DESIGN

NOTE: ASKING MORE STATISTICIANS WON'T GET ME A MORE REPRESENTATIVE PICTURE

How do we obtain a representative sample?

WE SELECT THE SAMPLE

Start with a list of everyone (sampling frame) and use a probability method.

Simple Random Sampling

Systematic Sampling

Stratified Sampling

Cluster Sampling

But we can't usually force people to participate

Problems:

- Expensive
- Time consuming
- Needs a sampling frame
- Non-response bias

THE SAMPLE SELECTS US

Advertise and hope people self-select to participate

Convenience Sampling

Online Polls

But who will see the advertisement?

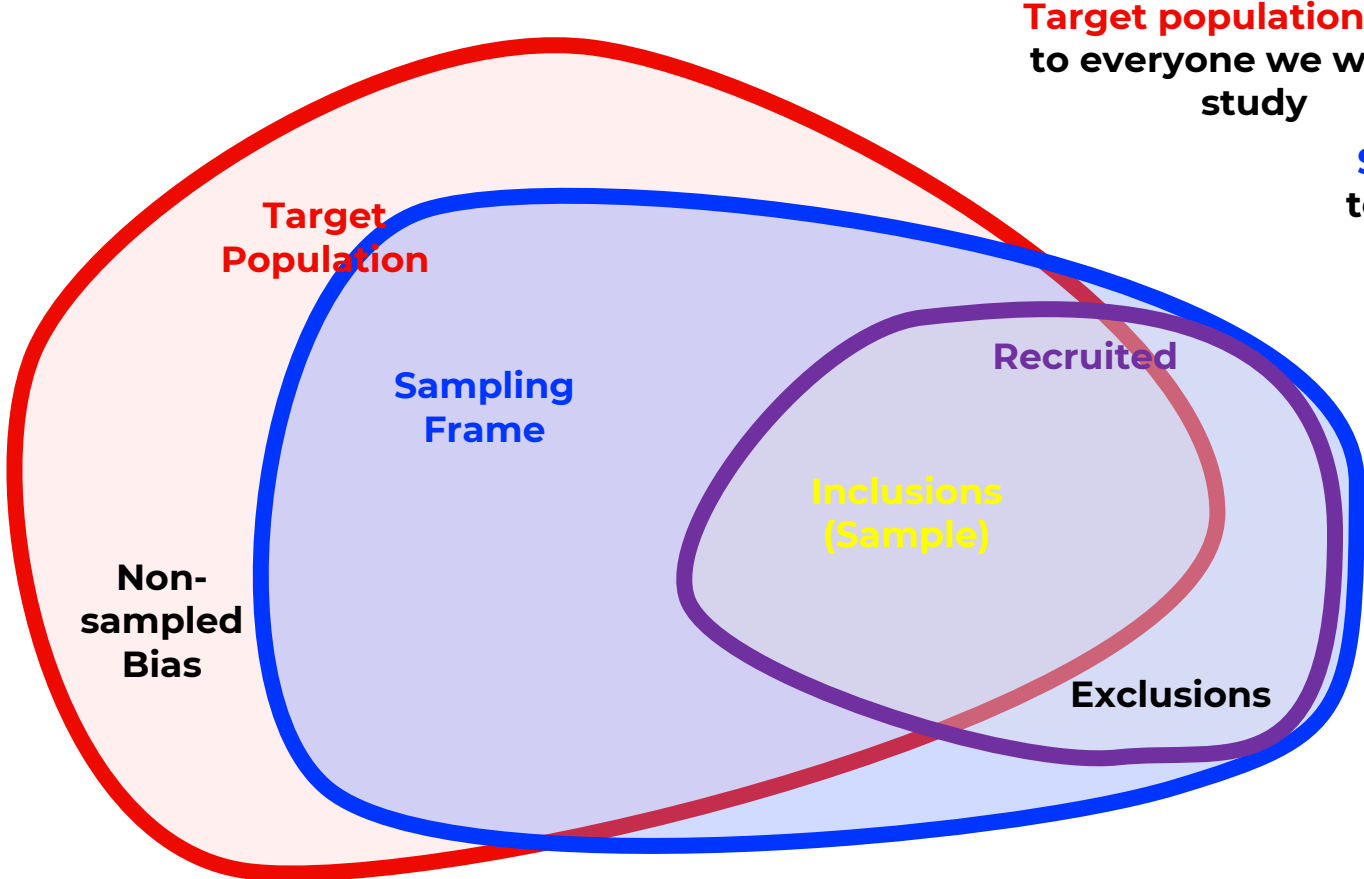
Problems:

- Not necessarily representative
- No way to measure the level of bias
- **Voluntary response bias**

How many **Democrats** do you think watch NEWSMAX?



Population vs Sampling Frame



Target population refers to everyone we want to study

Sampling Frame refers to everyone who can be selected

Sample refers to everyone who will be studied

Selection Bias refers to when some subjects are more / less likely to be selected.

Things to keep in mind...

- Were any people less likely to be recruited?
- Be careful when making generalisations – representativity is a high bar.
- More data does not necessarily mean more representative.
- Weighting does not necessarily “fix” all problems.
- More non-response = More opportunity for bias.
- Emotive issues are more susceptible to volunteer bias.
- Margin of error / statistical testing does **not** account for this kind of bias (non-sampling bias).
- Checking demographics with benchmarks is a good way to assess evidence of sample bias.

A final comment



UNIVERSITY
OF WOLLONGONG
AUSTRALIA

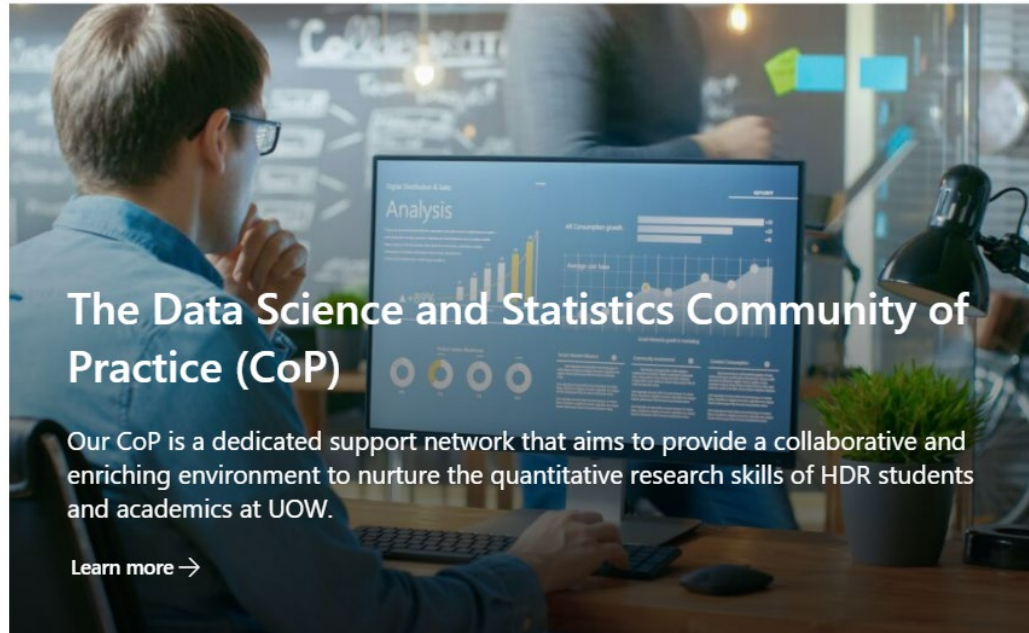
Check your work

- Missing axes
- Incorrect labels
- No labels on figures and columns
- Incorrect data, incorrect p-values
- Watch auto correct, the term “statically significant” has been published > 1000 times in journals instead of “statistically significant”
- eg “This statically significant result proved that this gaming act offers a potential instructional benefit beyond it’s novelty” in JAMA
- Barnett and White “Significance” March 2024:11-13

Keep your code

- Keep a record of all the data manipulations and analyses you have done
- This can be done in SPSS if you are using the workbook mode
- Easy to do if you are using R

The Data Science and Statistics CoP



The Data Science and Statistics Community of Practice (CoP)

Our CoP is a dedicated support network that aims to provide a collaborative and enriching environment to nurture the quantitative research skills of HDR students and academics at UOW.

Learn more →

[Click here to visit](#)

SCC website...

<https://www.uow.edu.au/niasra/our-research/statistical-consulting-centre/>

also have a look at the NIASRA website...

<https://www.uow.edu.au/niasra/>

Data & Decision Science Initiative

four key areas of focus

Research: virtual network and working groups of Data and Decision Science researchers

- Focal point for coordinating the development of Data Science at UOW
- Composed of researchers actively using or interested in Data Science methods
- Themed meetings emphasising translation: Data and Decision Science Network (DDSN)
- Strategically collaborations through the DDSI give a competitive advantage in translation

Education: Training in data science and reproducibility of research.

- Internal and external training and education in data science
- Upskilling research students & staff (particularly ECRs) in data & decision science methods
- Workshops (GRS, Statistical Consulting Centre)

T shaped graduates: Reviewing service subjects to refocus on data science.

- Review of service subjects in statistics and quantitative methods to give data science focus
- Graduates literate in data science and reproducible research

External/Industry engagement: Capitalising on existing links

- Provide enhanced opportunities for external engagement