BMS PhD Bitesize Training
Introduction to Statistics

30/05/23

Andrew Mason

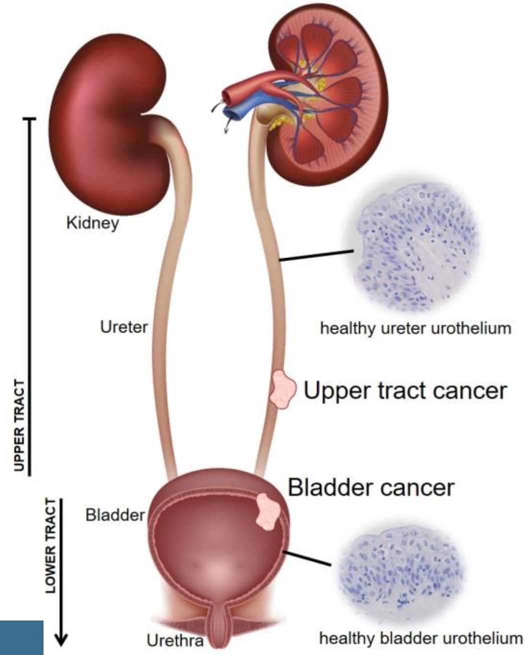
andrew.mason@york.ac.uk
asmasonomics.github.io
@asmasonomics

Introductions

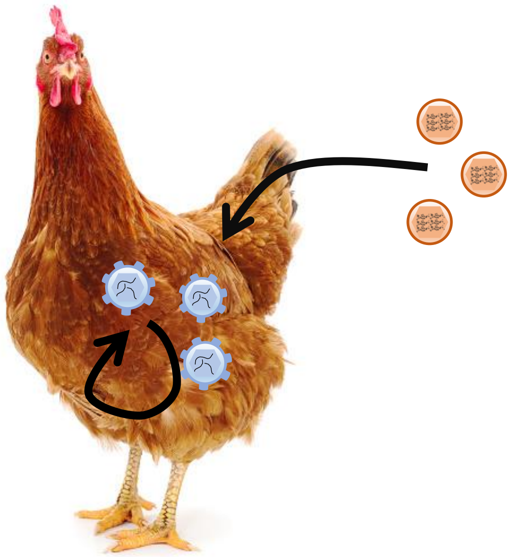
Andrew Mason

Lecturer in Cancer Informatics
Elixir Data Stewardship Fellow
Bladder cancer bioinformatic lead for the 100,000 Genomes Project
Endogenous retrovirus lead for chicken reference genome, pangenome and diversity consortia

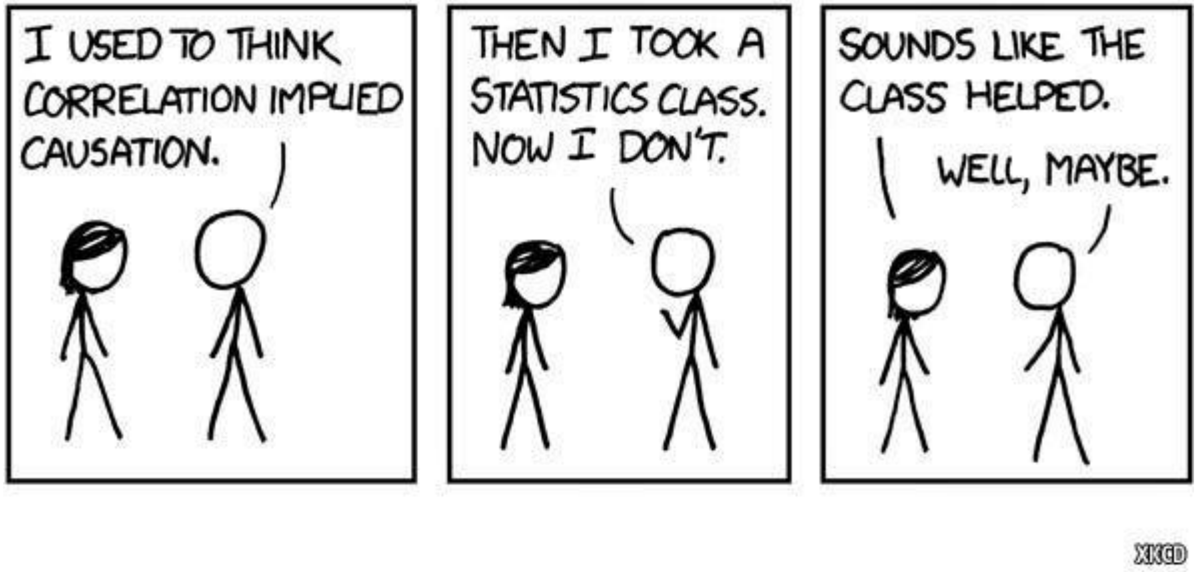
Urothelial carcinoma in humans



Oncogenic viruses in chickens



Data science training in the life sciences: UG, PG, Academics


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.
XKCD

Introduction to statistics:

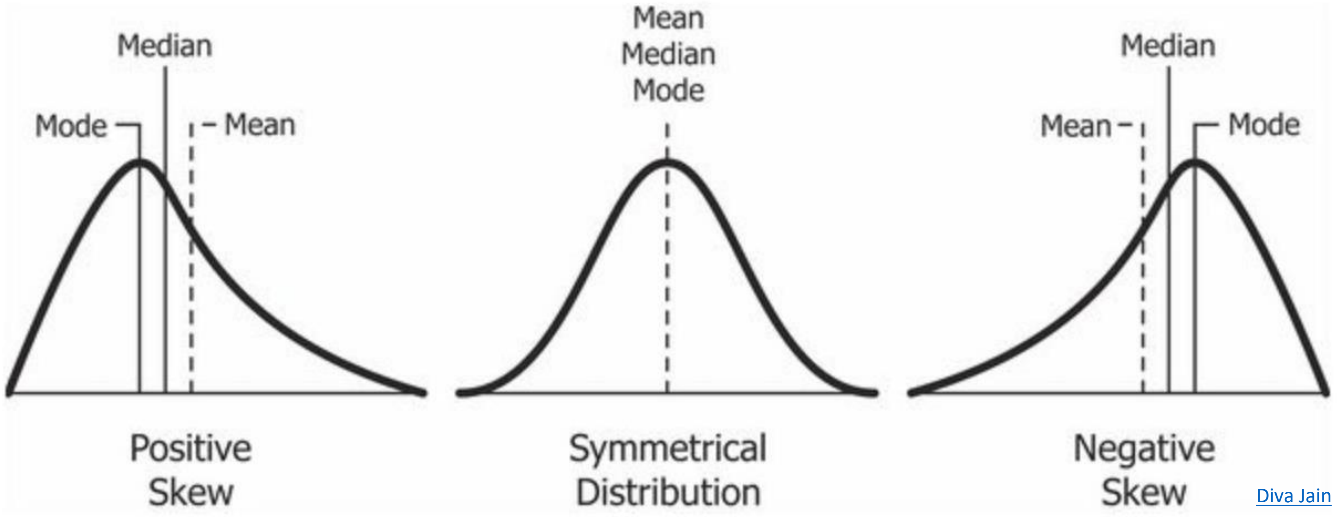
- Better idea of the statistical test to use
- Appreciation of some common pitfalls
- Make the most of your data!
- Nail those stats questions in your viva
- (Maybe) know more about statistics than your supervisor…

Descriptive Statistics (i)

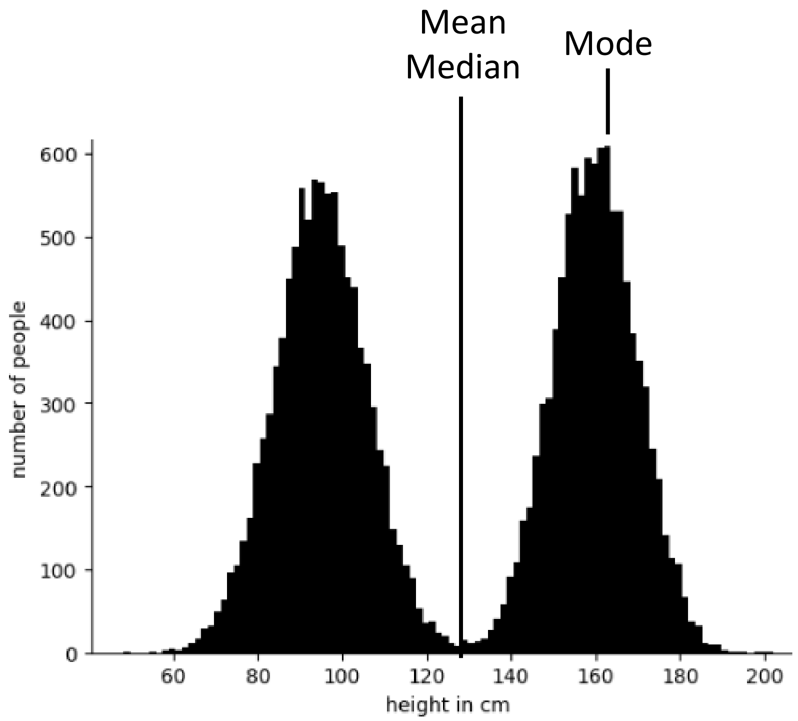
Mean – the “traditional” average (sum/n)

Median – the value in the middle of your dataset

Mode – the most common value in your dataset



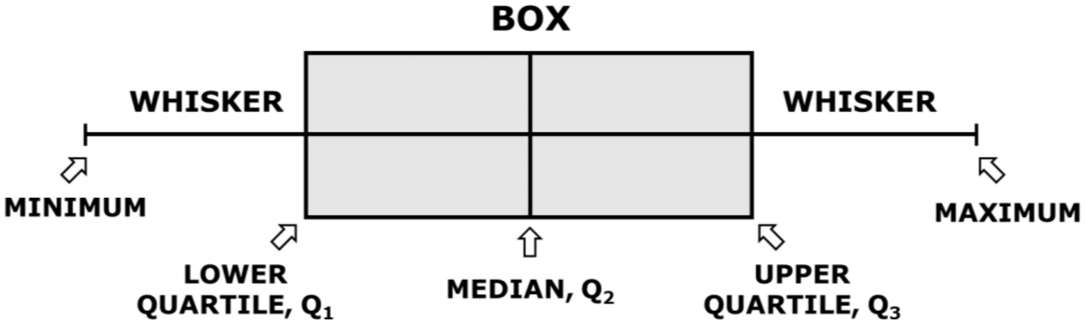
Diva Jain



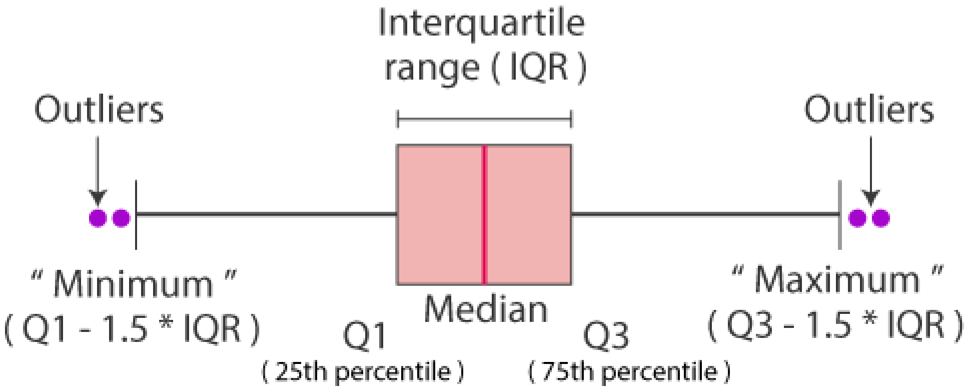
Always graph your data!

Descriptive statistics (ii)

Box plots



One way to determine outliers



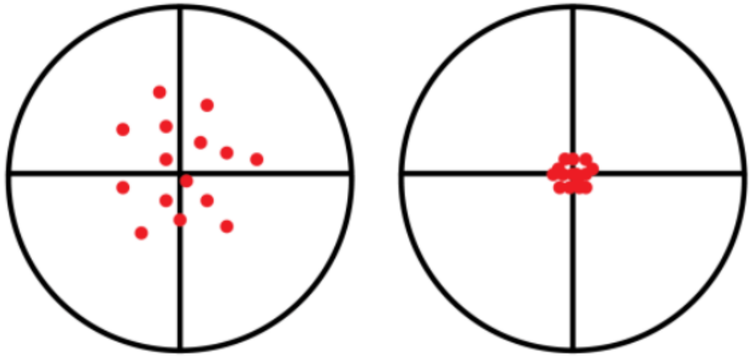
Descriptors

- Range (w/wo outliers)
- IQR
- 95% confidence intervals

Descriptive statistics (iii)

Mean – the “traditional” average (sum/n)

How accurate have we been in measuring the mean of the population?



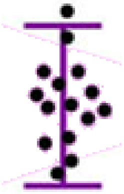
Standard deviation – how much do the individual measurements differ from the mean value?

Low SD gives us more confidence in our assessment of the mean value

Descriptive statistics (iv)

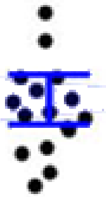
Standard deviation vs Standard error **of the mean**

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \mu|^2}{N}}$$

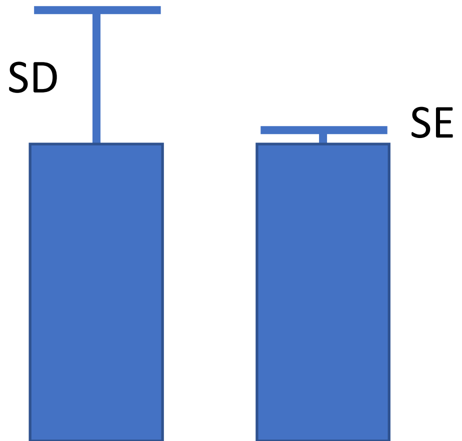


SD – how well do my **individual** measurements support **my observed mean** value?

$$SE = \frac{\mathrm{SD}}{\sqrt{n}}$$

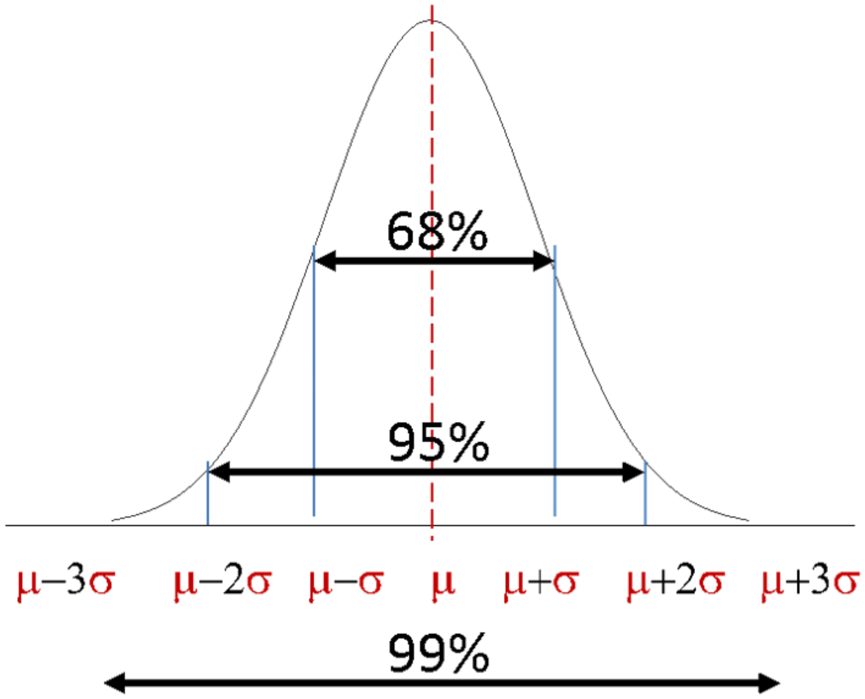


SE – how well do my **repeat** measurements support the **actual population mean** value?



The normal distribution

Normal distribution = Gaussian distribution



Mean determines the peak of the curve
SD determines the shape of the curve

95% of the data points fall within +/- 2SD from the mean

100% - 95% = 5% = 0.05

This is where the significance threshold comes from…

Parametric testing

Parametric testing requires data to be normally distributed (ish)

- Equal variance between groups

- Groups are independent measures

- No distribution altering outliers

t test

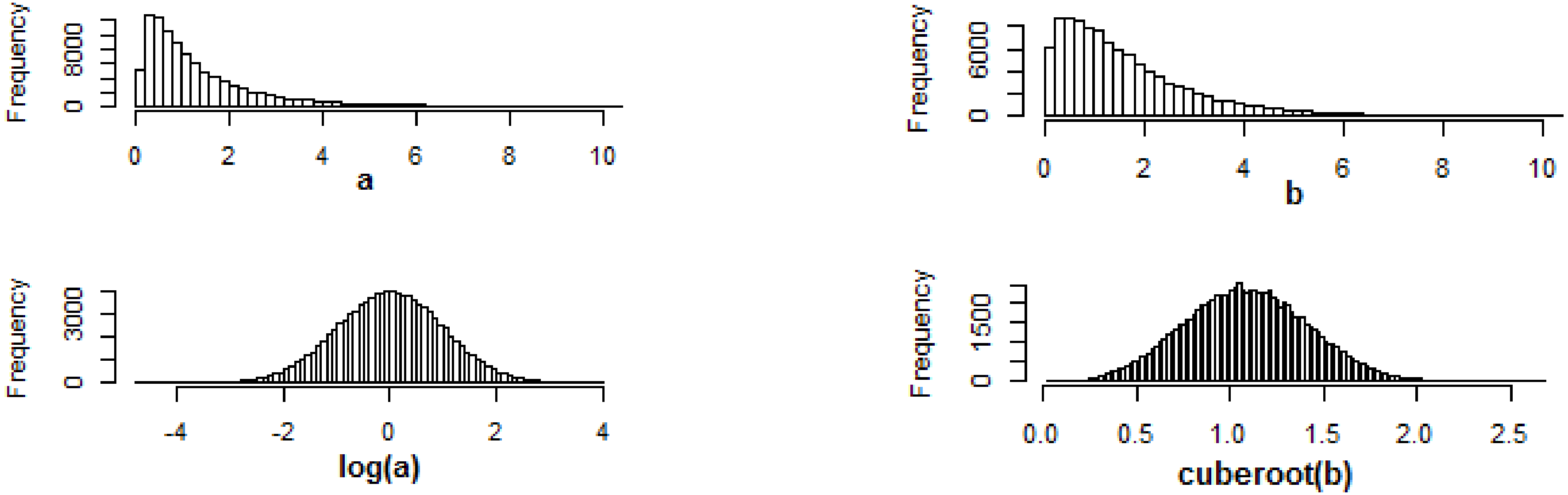
ANCOVA

Pearson correlation

ANOVA

Can you make your data normal?

Data transformations are legit!

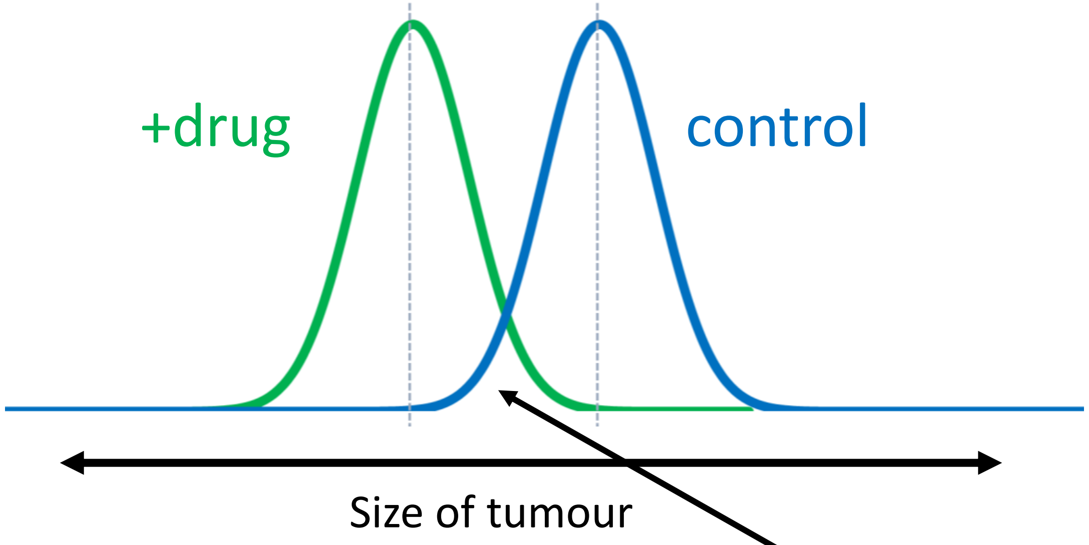


Just be careful when:

- reporting results (is it understandable?)
- considering outliers
- drawing error bars

NON PARAMETRIC TESTING DOES EXIST!!

t test logic



Has the tumour significantly decreased in size with addition of the drug?

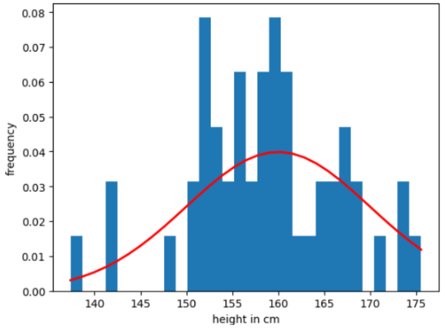
What is the overlap here, what proportion of the data could be found in both distributions?

Significance – do you trust it?

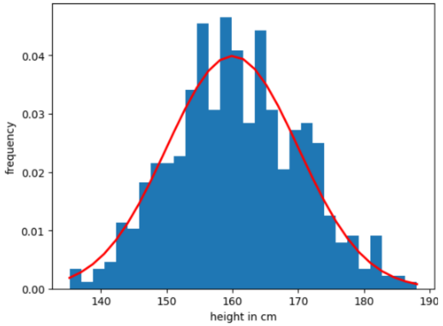
1) Small n vs big n

e.g. height dataset - mean 160cm, SD 10cm

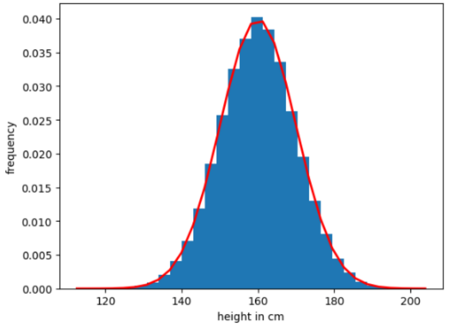
n=50



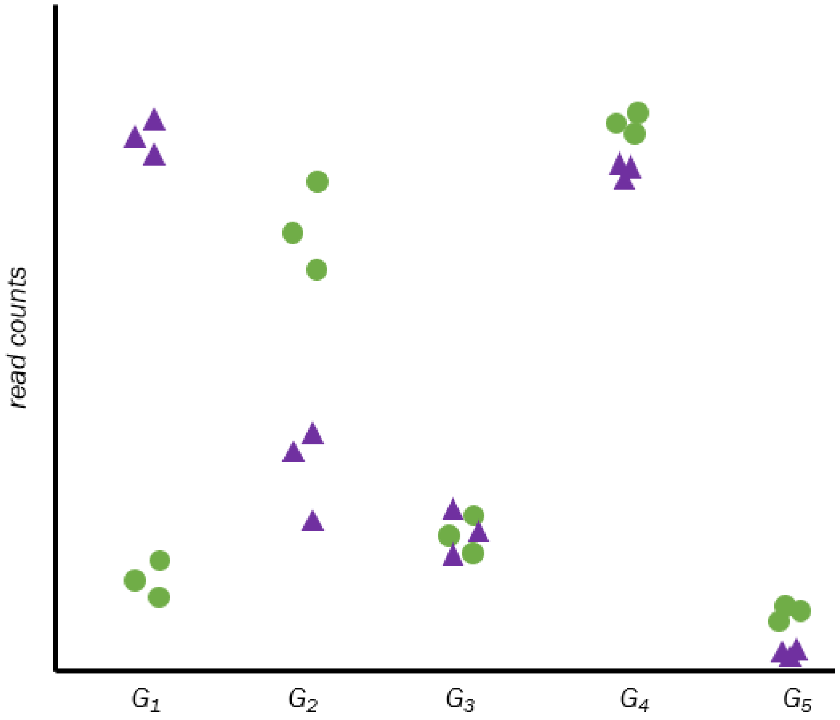
n=500



n=50,000

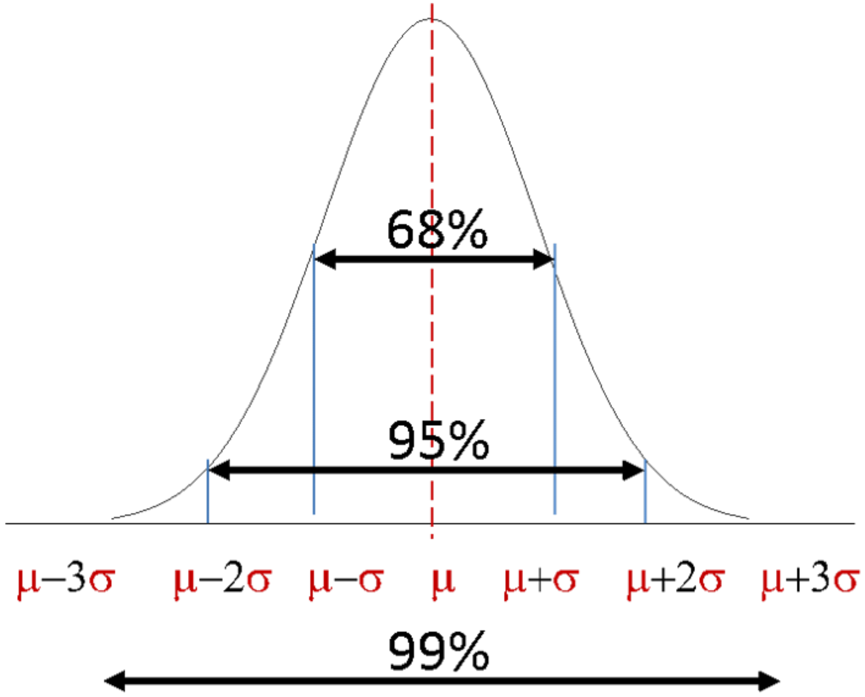


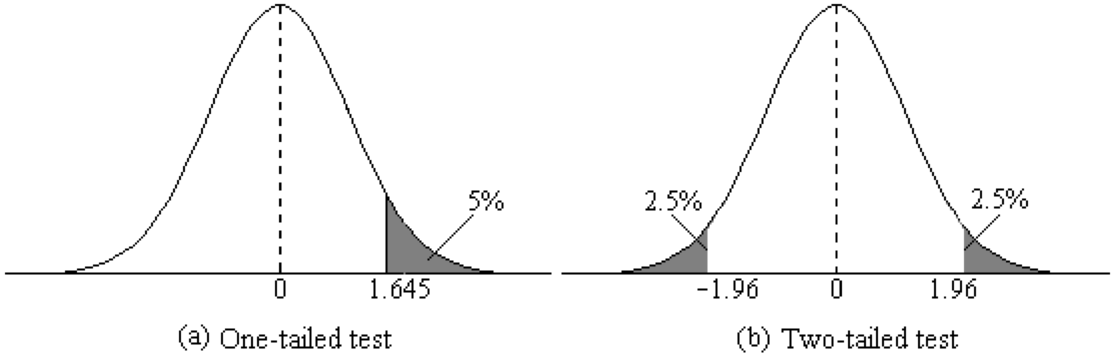
2) Biological vs Statistical significance



One tailed vs two tailed

Normal distribution = Gaussian distribution





Null hypothesis is typically that “there is no difference”

Only use a one-tailed test when there is no possibility of there being the alternate direction of response

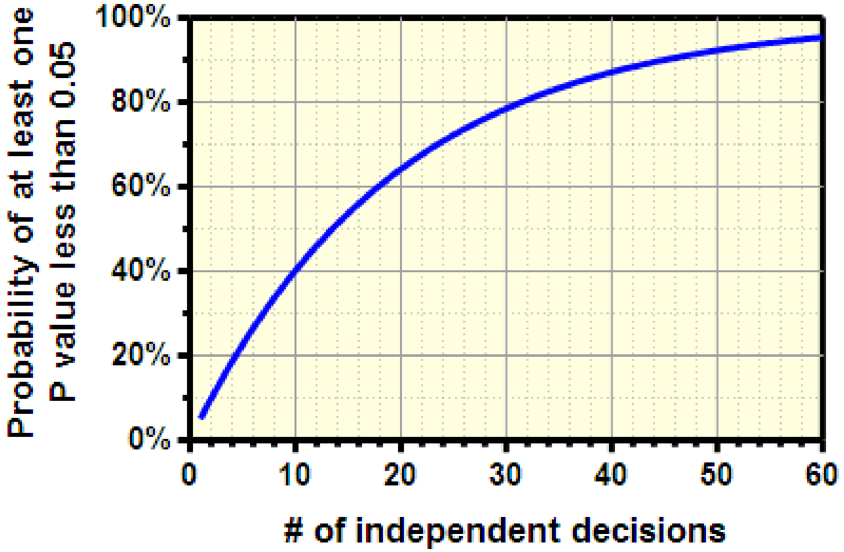
p hacking

WHY IS EVERYONE OPPOSED TO P-HACKING? IT'S BRILLIANT!
LITERALLY *ANYTHING* CAN BE TRUE! WE CAN CONTROL THE ECONOMY! WE CAN BE TELEPATHIC! WE CAN CURE CANCER!
P-HACKING MAKES YOU A WIZARD! IF NOT FOR ALL THE NAYSAYERS, WE COULD WIELD TOTAL POWER OVER THE ENTIRE UNIVERSE!
ARE YOU SURE THAT'S TRUE?
NOT YET, BUT IF I CAN JUST PICK THE RIGHT DATA...

smbc-comics.com

Fishing exercises vs the art of data science

Multiple correction

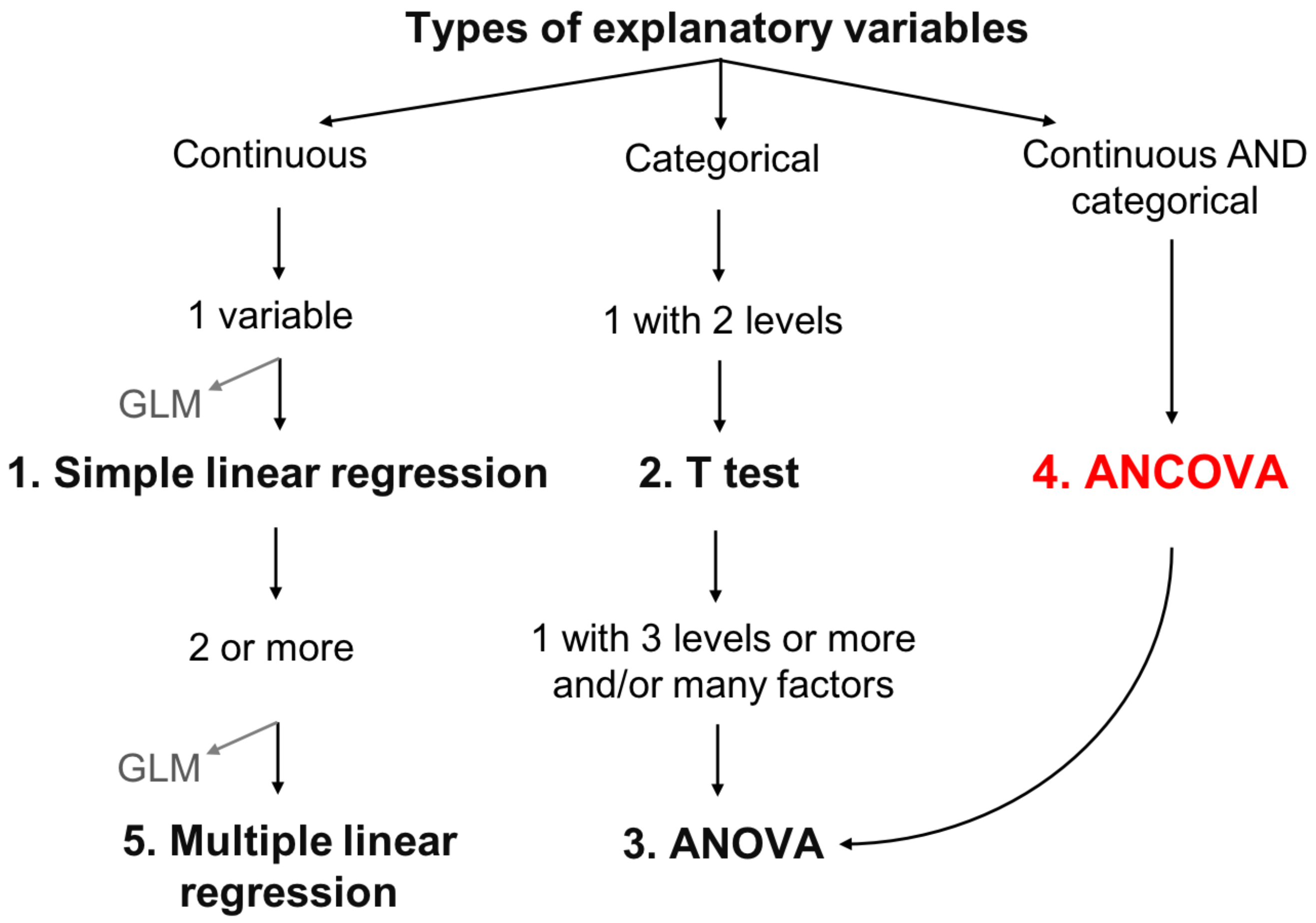


Increasing chance of a false positive when you do lots of tests

Bonferroni correction
Benjamini-Hochberg
Etc etc

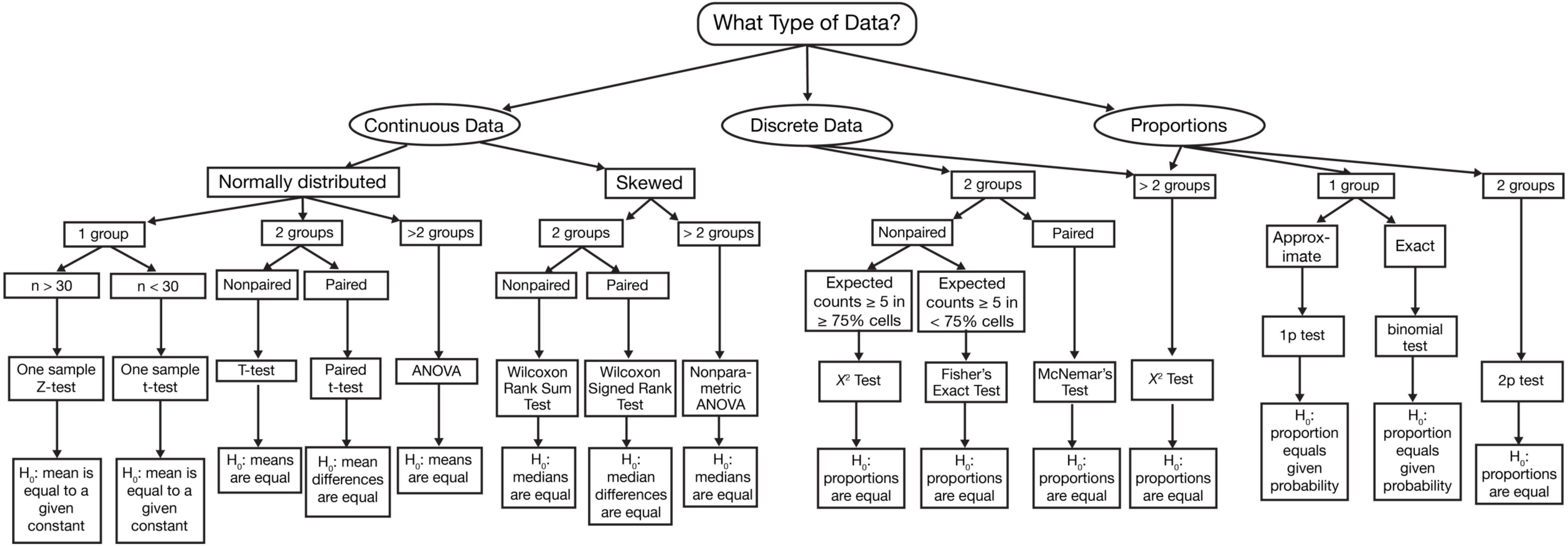
Choice of parametric testing

t test vs ANOVA vs ANCOVA (and what is post hoc testing?!)

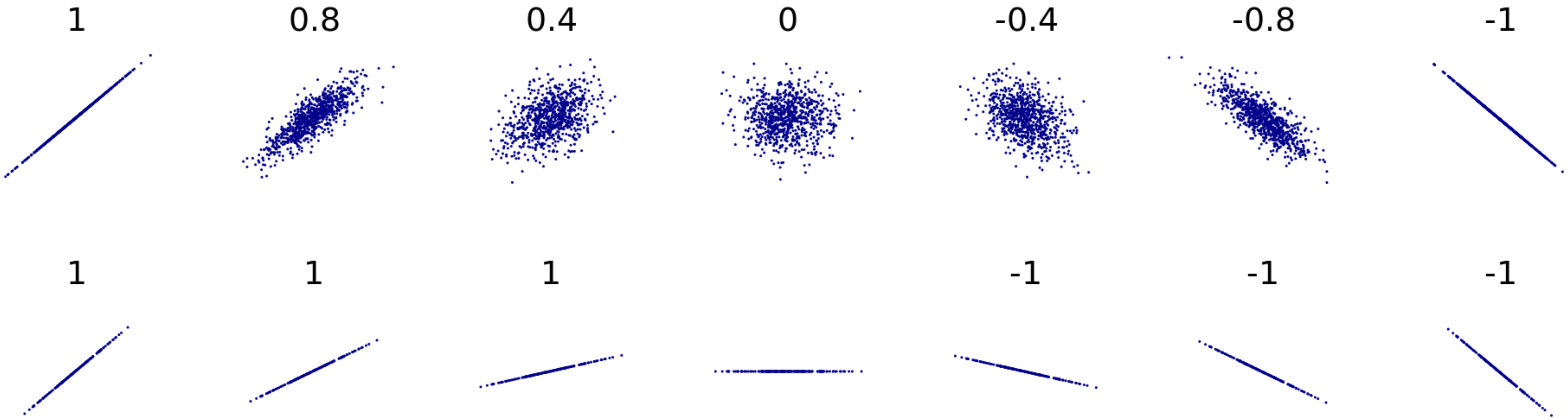

Types of explanatory variables
Continuous
Categorical
Continuous AND categorical
1 variable
1 with 2 levels
GLM
1. Simple linear regression
2. T test
4. ANCOVA
2 or more
1 with 3 levels or more and/or many factors
GLM
5. Multiple linear regression
3. ANOVA

Flow chart

Flow chart: which test statistic should you use?

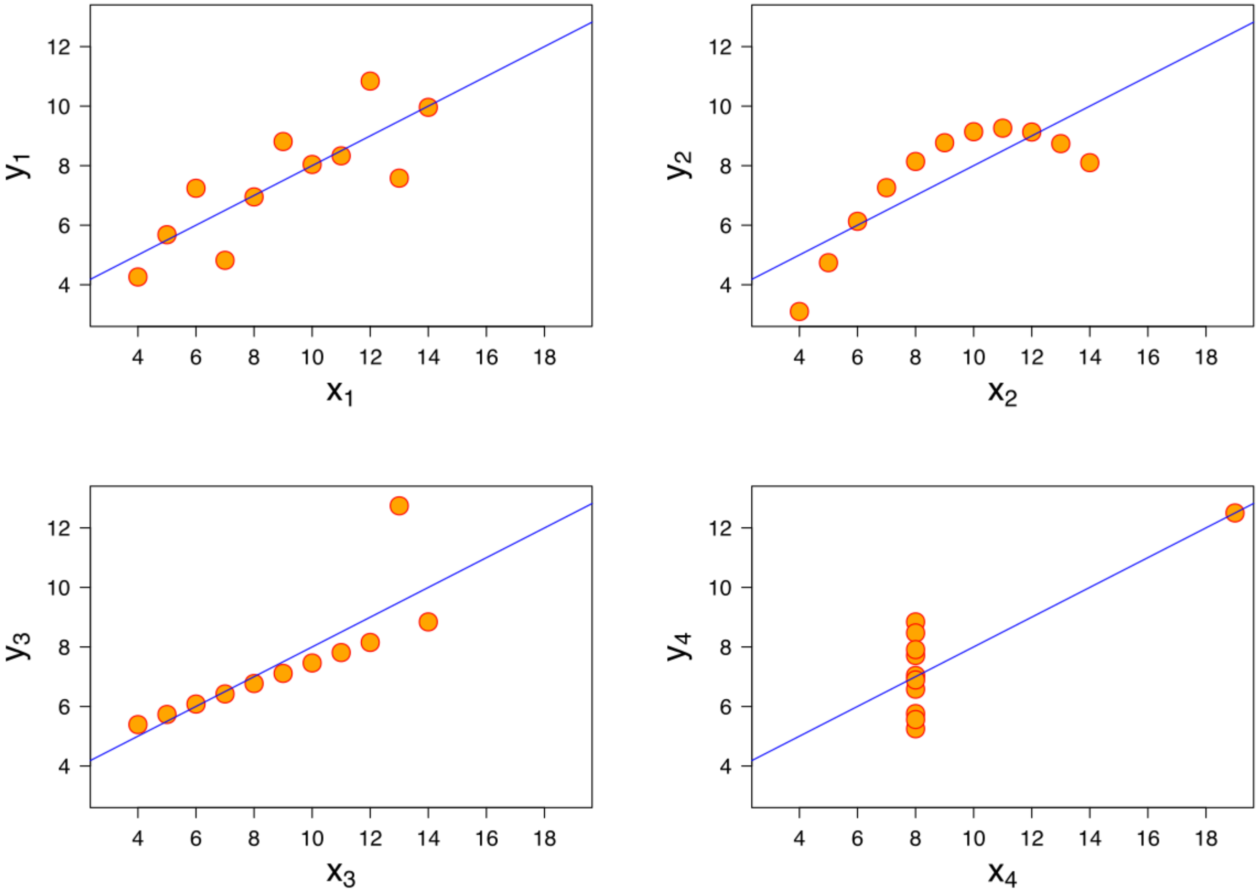


Correlations (i)



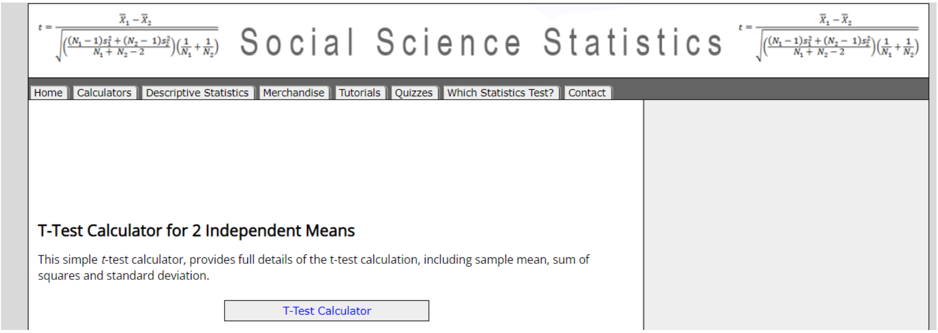
Correlations (ii)

Cautionary tale – Anscombe's quartet



Same considerations for parametric vs non-parametric
 - often Pearson vs Spearman

Where should I do my stats?



$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}$ Social Science Statistics $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}$

T-Test Calculator for 2 Independent Means

This simple *t*-test calculator, provides full details of the t-test calculation, including sample mean, sum of squares and standard deviation.

https://www.socscistatistics.com/

How often will you be doing stats like this?

Summary resources

This presentation, including a recording, will (shortly) be on my website:

https://tinyurl.com/BMS-Stats-May2023

Materials for previous “Introduction to cBioPortal” course:

https://tinyurl.com/Intro-cBioPortal-Jan2023

Elixir Research Data Management “Bites” on sequencing data:

https://tinyurl.com/RDM-Seq-Videos