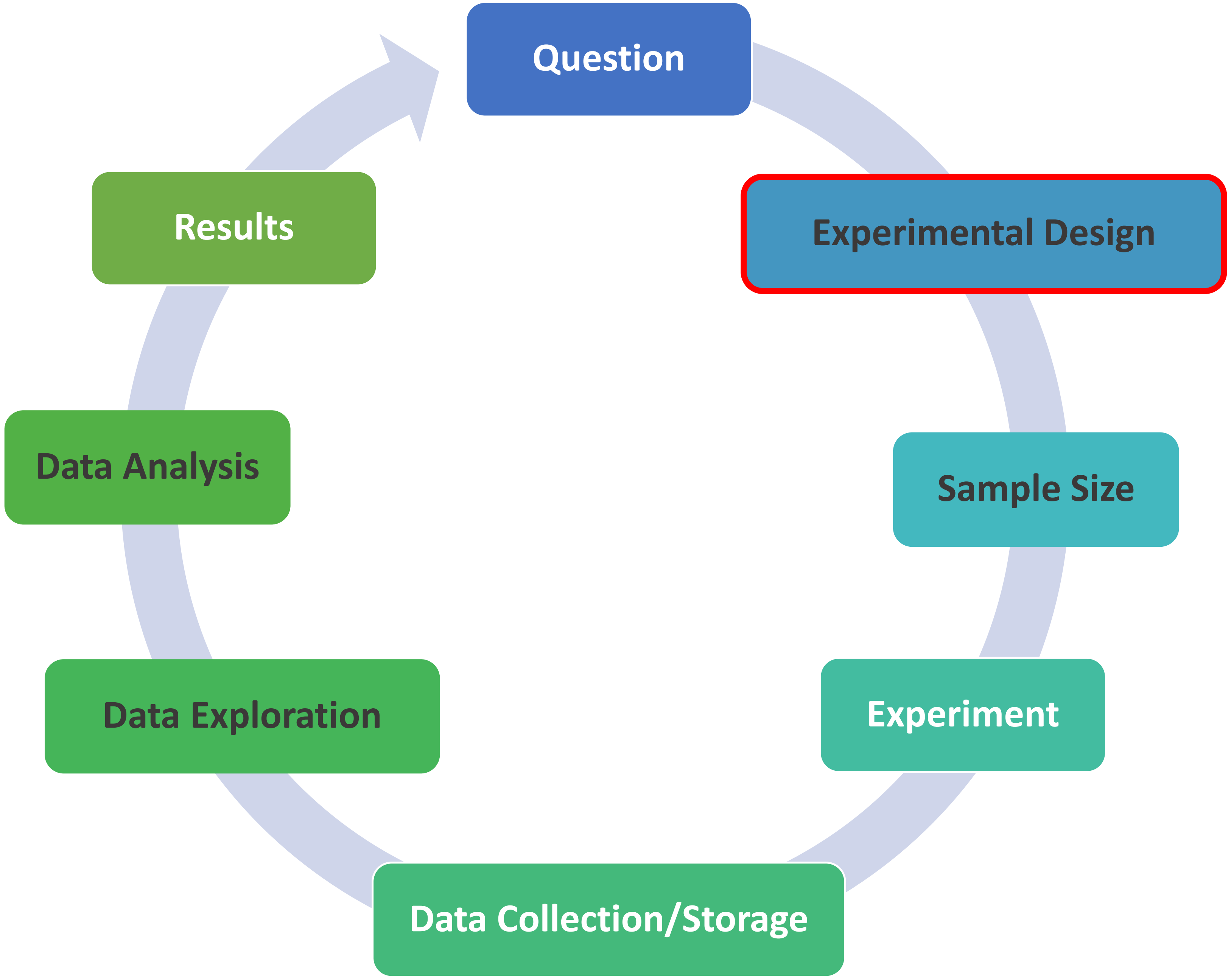
Babraham Institute

Day 1
Experimental design

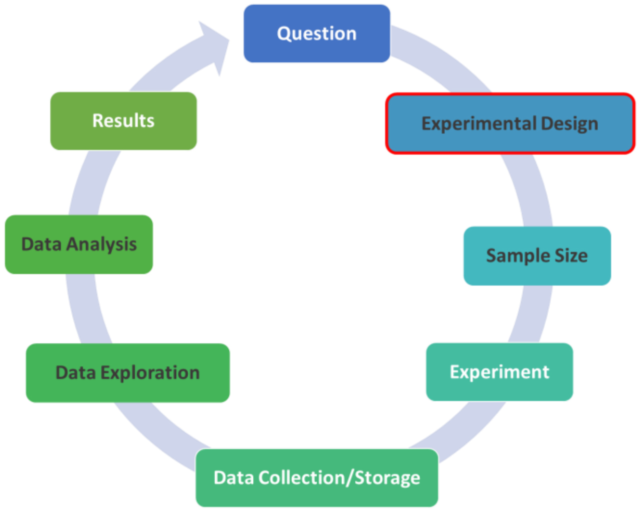
Anne Segonds-Pichon

v2019-06

Babraham Bioinformatics



Question
Experimental Design
Sample Size
Experiment
Data Collection/Storage
Data Exploration
Data Analysis
Results

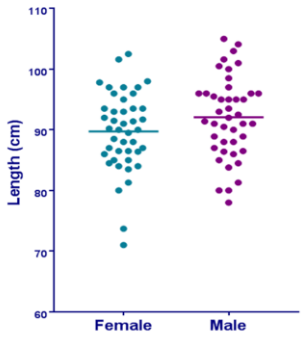


- <u>Universal principles</u>
  - The same-ish questions should always be asked
    - **What is the question?**
    - **What measurements will be made?**
    - **What factors could influence these measurements?**
  - But the answers/solutions will differ between areas

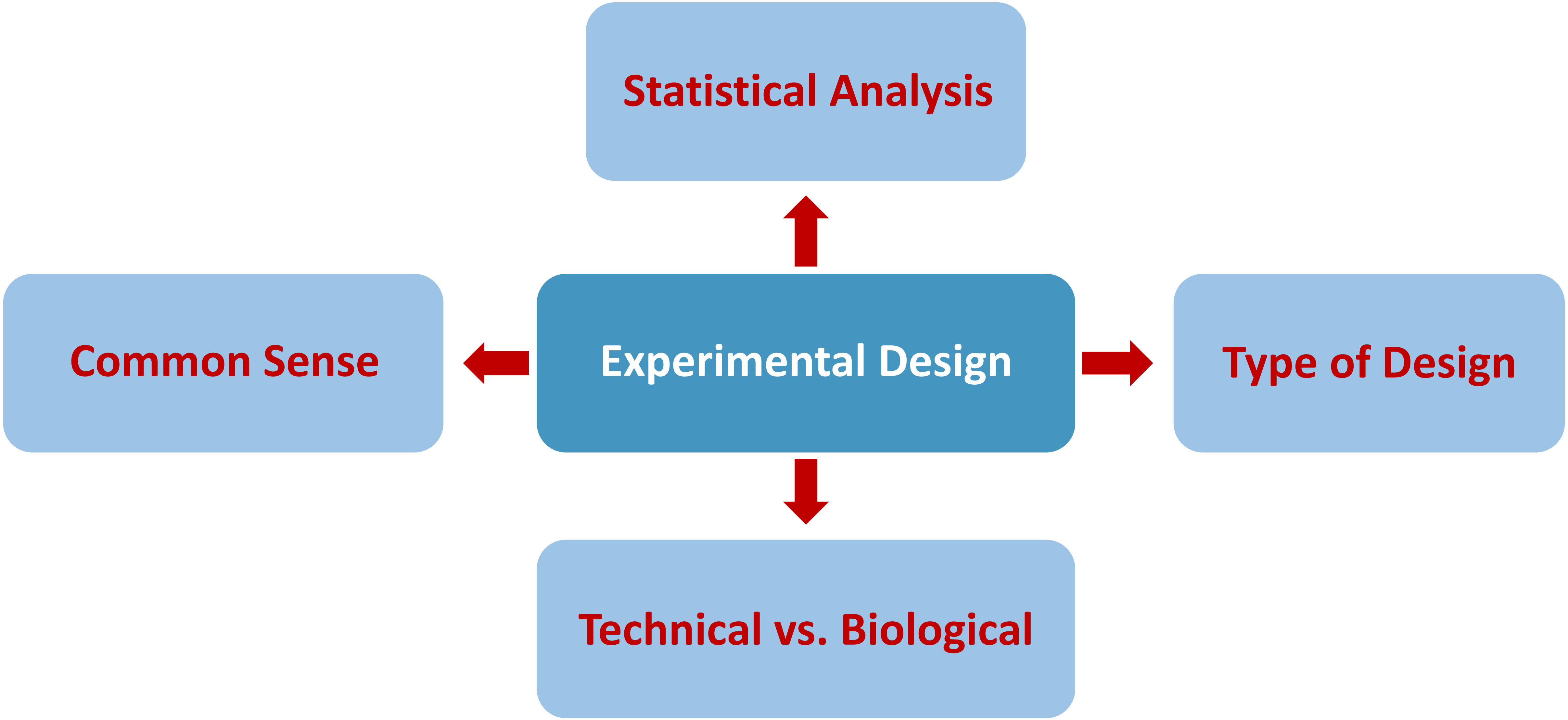
- <u>Examples</u>:
  - **Experimental design** will be affected by the question
    - but also by practical feasibility, factors that may affect causal interpretation …
    - e.g. number of treatments, litter size, number plants per bench …
  - **Sample size** will be affected by ethics, money, model …
    - e.g. mouse/plant vs. cell, clinical trials vs. lab experiment …
  - **Data exploration** will be affected by sample size, access to raw data …
    - e.g. >20.000 genes vs. weight of a small sample of mice

Vocabulary, tradition and software

- People use different words to describe the same data/graphs …
- There are different traditions in different labs, areas of science …
- Different software mean different approaches: R, SPSS, GraphPad, Stata, Minitab …
- <u>Examples:</u>
  - Variable names: qualitative data = attribute
  - Scatterplots in GraphPad Prism = stripchart in R
  - 2 treatment groups in an experiment = 2 arms of a clinical trial
  - Replicate = repeat = sample
  - QQ plots in SPSS versus D'Agostino-Pearson test …
  - Sample sizes



- Very different biological questions, very different designs, sophisticated scientific approach or very simple
  - Similar statistical approach
  - <u>Example</u>:
    - **Data**: Gene expression values from The Cancer Genome Atlas for samples from tumour and normal tissue, **question**: which genes are showing a significant difference? ***t*-test**
    - **Data**: weight from WT and KO mice, **question**: difference between genotypes? ***t*-test**


Statistical Analysis
Common Sense
Experimental Design
Type of Design
Technical vs. Biological

Experimental Design → **Statistical Analysis**

- **Translate the hypothesis into statistical questions**
  - Think about the statistical analyses before you collect any data
- What data will I collect?
- How will it be recorded/produced?
- Will I have access to the raw data?
- I have been told to do this test/use that template, is that right?
- Do I know enough stats to analyse my data?
  - If not: ask for help!

Experimental Design → **Statistical Analysis**

- <u>Example</u>:

  - **Hypothesis**: exercise has an effect on neuronal density in the hippocampus.

  - **Experiment**: 2 groups of mice on 2 different levels of activity:
    - No running or running for 30 minutes per day
    - After 3 weeks: mice are euthanized and histological brain sections are prepared
      - Neuronal density by counting the number of neurons per slide

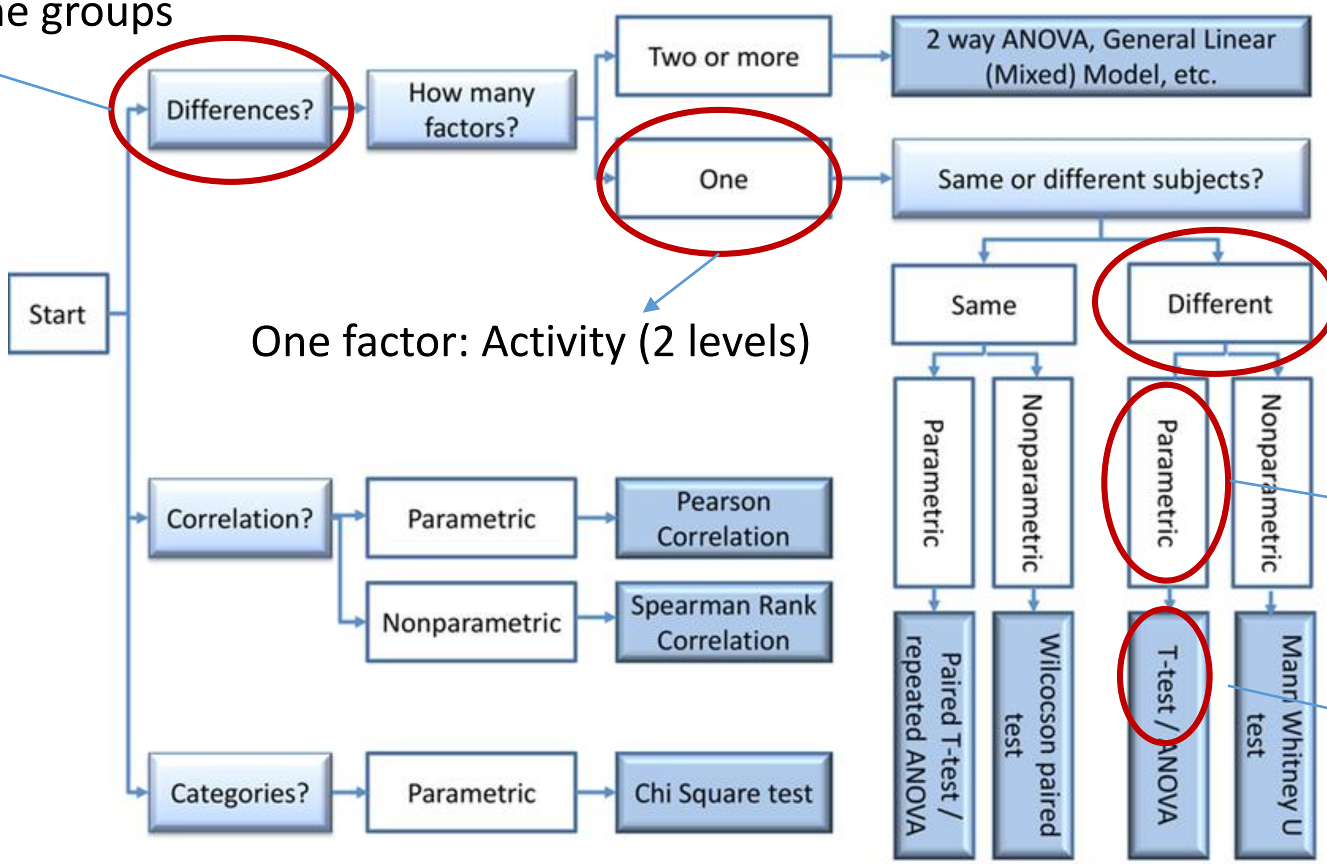
  - **Stats**: <u>one factor</u>: activity and <u>one outcome</u>: number of neurons

Bate and Clark, 2014

Experimental Design → **Statistical Analysis**

- **Experiment:** exercise has an effect on neuronal density in the hippocampus

Difference between the groups

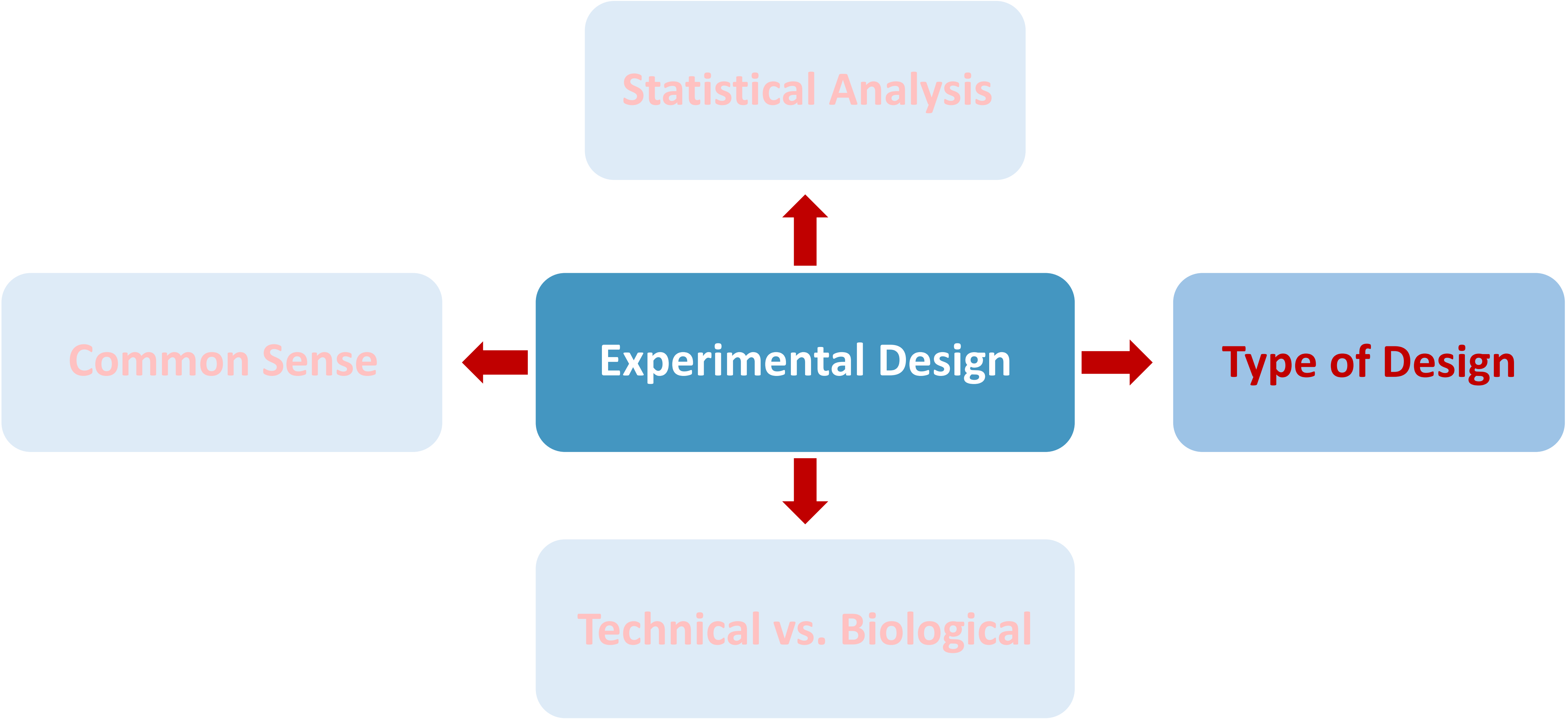


One factor: Activity (2 levels)

2 different groups of mice

Neurons counts:
Normality
Homogeneity of variance

That's the one!


Statistical Analysis
Common Sense
Experimental Design
Type of Design
Technical vs. Biological

Experimental Design → **Type of design**

- **Experimental unit**: cell, tissue sample, leaf, mouse, plant, litter …
  - Neuronal density experiment: experimental unit: **mouse**

- **Factor**:
  - Fixed factor: factor of interest, predictor, grouping factor, arm in controlled trial, independent variable …
    - e.g. : treatment, gender, genotype …
    - Neuronal density experiment: fixed factor: **running**

  - Random factor: factor we need to account for, blocking factor, nuisance factor …
    - e.g. : experiment, batch, plate, lanes …
    - Neuronal density experiment: **uh oh**

- **Key concepts**:
  - Blinding: not always possible, single and double-blinding
  - Randomisation

Experimental Design → Type of design

Completely random
CRD

Simplest:
experimental units randomly allocated to groups
e.g. : treatment …

Complete Randomised block
CRBD

Accounting for **random factors**, **nuisance variables**
e.g. : batch effect, experimental effect, day-to-day variation …

Split-plot

Also **nested design**, **repeated measures**
e.g. : several measures per animal, several treatments per plot, pups in a litter…

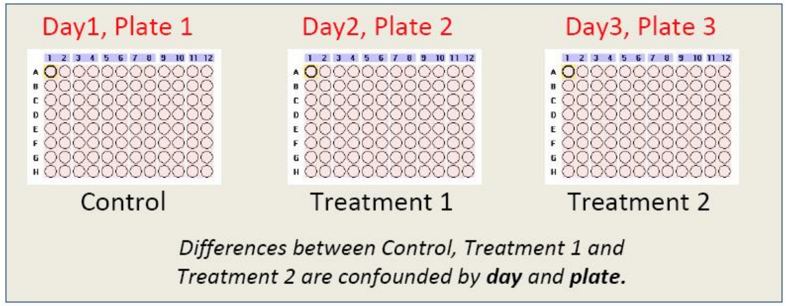
Experimental Design → Type of design

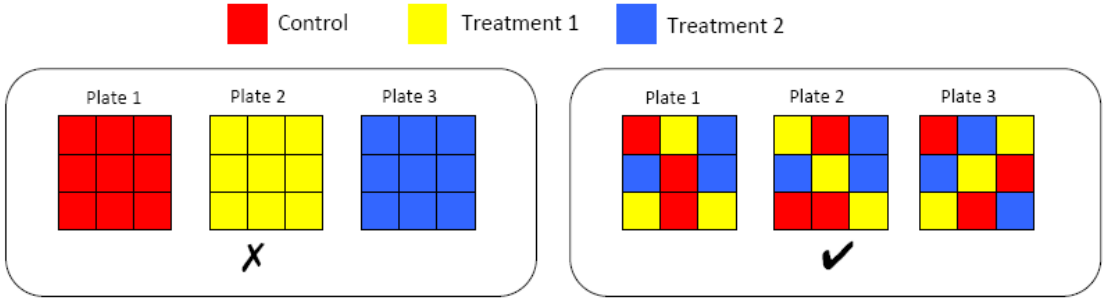
Completely random
CRD

Control	Treatment
Mouse 1	Mouse 6
Mouse 2	Mouse 7
Mouse 3	Mouse 8
Mouse 4	Mouse 9
Mouse 5	Mouse 10

Complete Randomised block
CRBD

Bad design

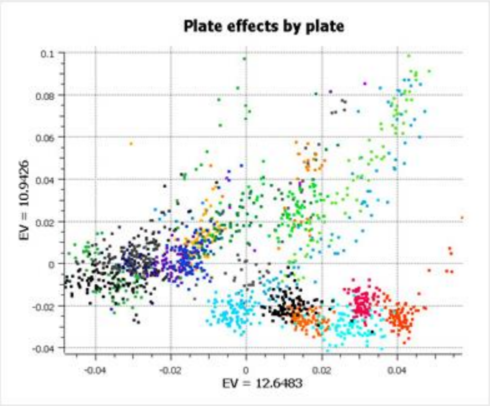




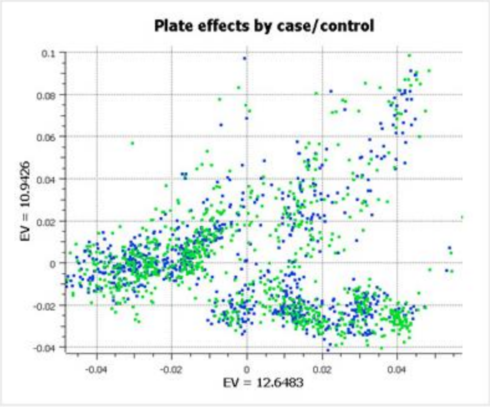
Good design:

GenADA multi-site collaborative study 2010

Alzheimer's study on 875 patients



Controls and Cases

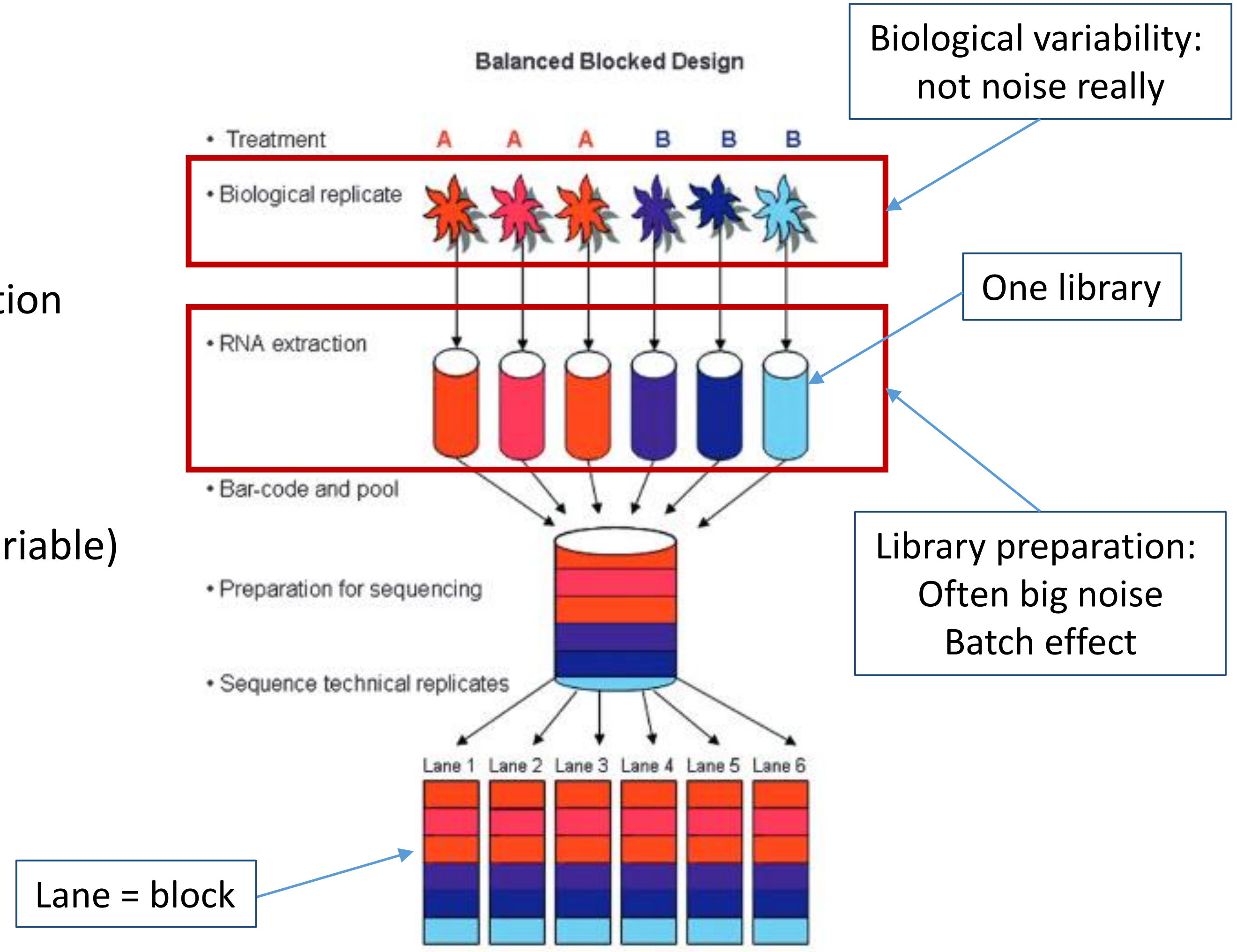


http://blog.goldenhelix.com/?p=322

Experimental Design → Type of design

Complete Randomised block

- **RNA-Seq experiments**: multiplexing allows for randomization
  - Multiplexing: barcodes attached to fragments
  - Barcodes: distinct between libraries (samples)
  - **Important**: identify the sources of noise (nuisance variable)
    - Library preparation: big day-to-day variability
      - **Batch effect**
    - Big variability between runs
    - **Lane effect**

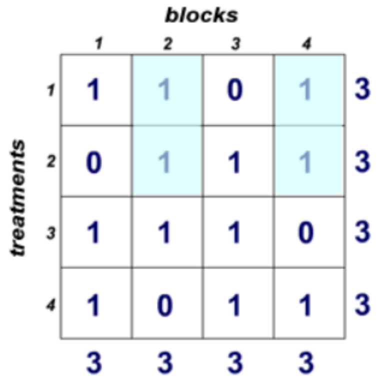


Auer and Doerge, 2010

Experimental Design → Type of design

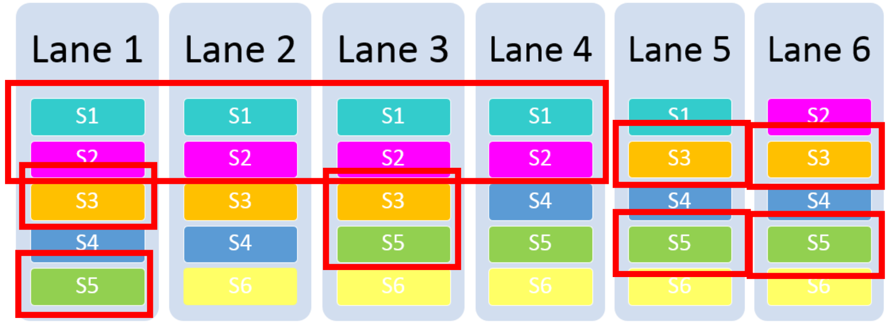
Incomplete Randomised block

- **RNA-Seq experiments**:
  - **Incomplete block design:**
    - All treatments/samples are not present in each block
  - **Balanced Incomplete Block Design** (BIBD):
    - where all pairs of treatments/samples occur together within a **block** an equal number of times



Six samples

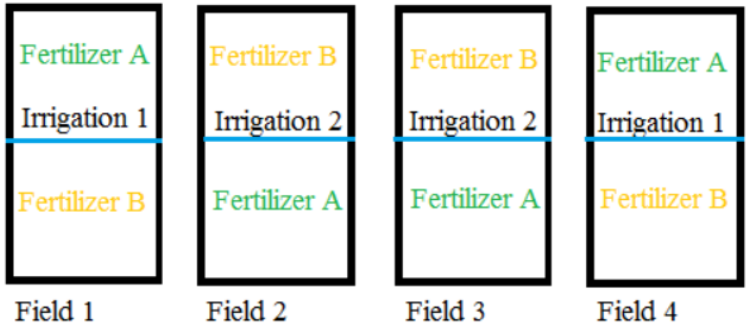
Five samples per lanes



- Statistical analysis:
  - account for missing values
  - e.g.: a model fits blocks then samples

Experimental Design → Type of design

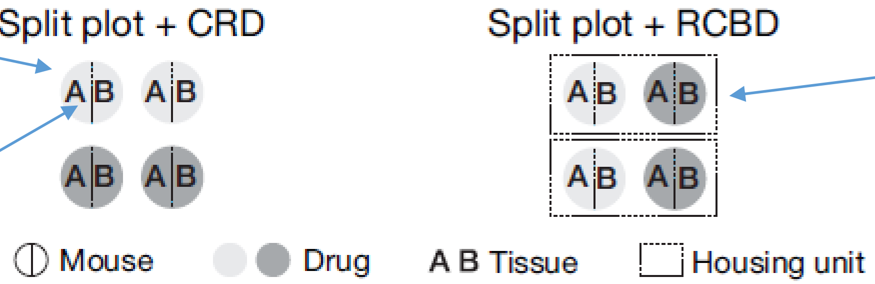
Split-plot : from agriculture: fields are **split** into **plots** and subplots.



- <u>Example</u>: *in vivo* effect of a drug on gene expression on 2 tissues.

Mouse/Plant = plot = random factor

Tissue = subplot (split-plots) = nested factor



Housing unit/Bench = blocking factor = nuisance factor

Krzywinski and Altman, 2015

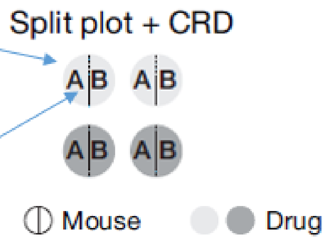
Experimental Design → Type of design

Split-plot

One-factor design: drug

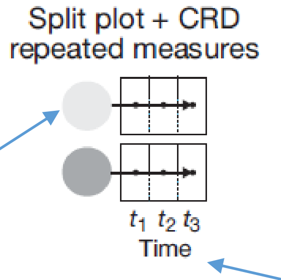
Mouse/Plant = plot
= random factor

Tissue = subplot (split-plots)
= nested factor



- More complex design:
  - **Split-plot + Completely Random Design**: commonly used for repeated measures designs

Two-factor design: drug+time

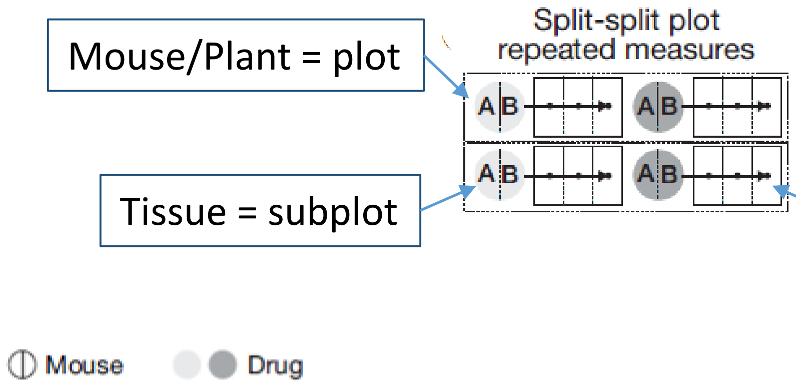


Mouse/Plant = plot

Time = nested within mouse/plant
= repeated measures = subplot

Three-factor design: drug+time+tissue

Mouse/Plant = plot



Tissue = subplot

Time = sub-subplot

Krzywinski and Altman, 2015

Experimental Design → **Type of design**

- Other designs: crossover, sequential ….

Factorial Design : more an arrangement of factors than a design

  - When considering more than one factor

- Back to our neuronal density experiment: exercise has an effect on neuronal density in the hippocampus

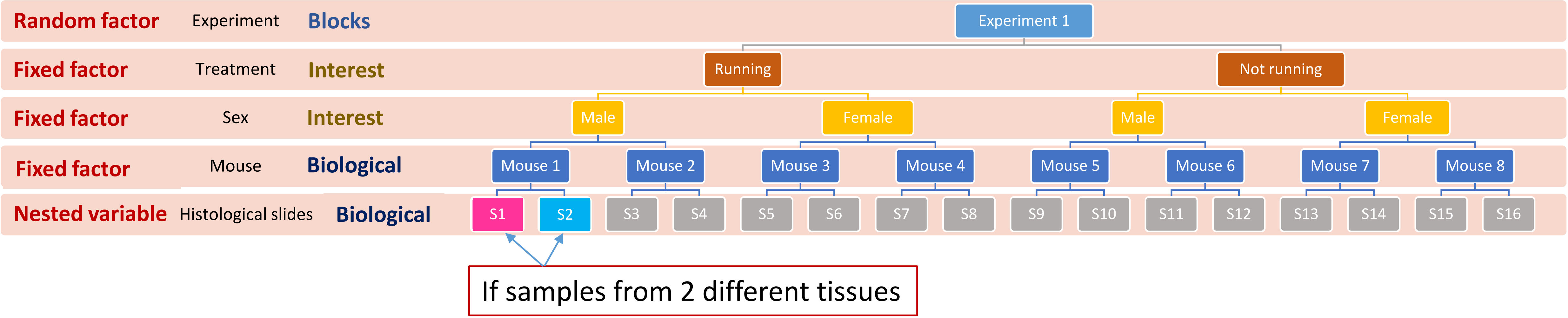
Running	Not running
n mice	n mice

→ **Completely random**

- Not enough: we want to account for:
  - Sex: factor of interest: **factorial design** (2 factors: running and sex)
  - Experimental variability: random factor: **blocking factor (one experiment = one block)**
  - Several histological slides: **nested variable**

Experimental Design → Type of design

- Neuronal density experiment: Complete Randomised block design + **Split-plot**



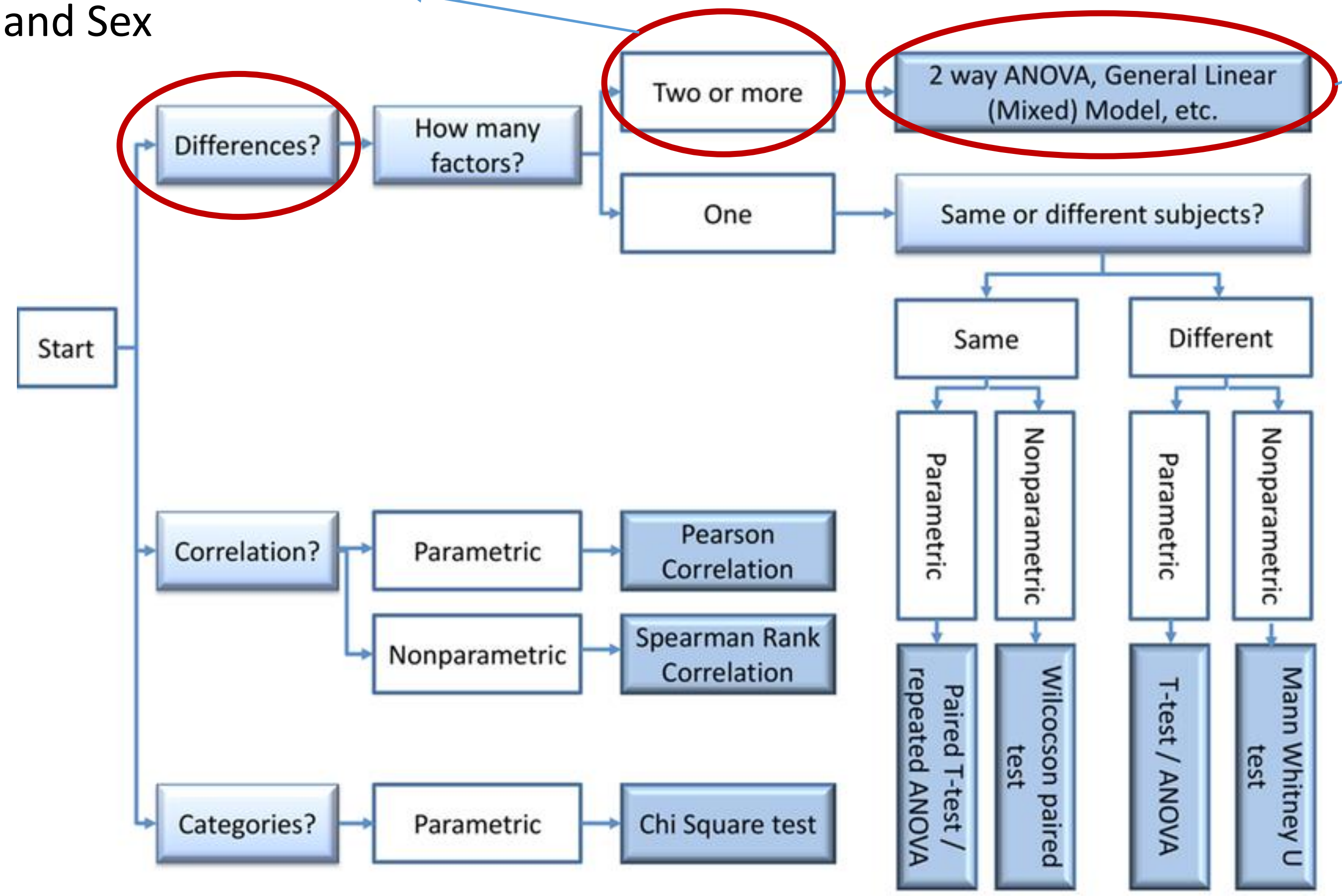
- Rule of thumb: Block what you can, randomize what you cannot
  - **Blocking** is used to remove the effects of a few of the most important nuisance variables (known/controllable)
  - **Randomisation** is then used to reduce the contaminating effects of the remaining nuisance variables (unknown/uncontrollable, lurking).
- Drawing the experimental design can help!

Experimental Design → Statistical Analysis

- **Experiment:** exercise has an effect on neuronal density in the hippocampus

Two factors of interest per experiment:
Activity and Sex

That's the one!



Statistical Analysis

Sample → Statistical inference → Population

- Statistical tests are tools used to quantify our level of confidence in what we see.

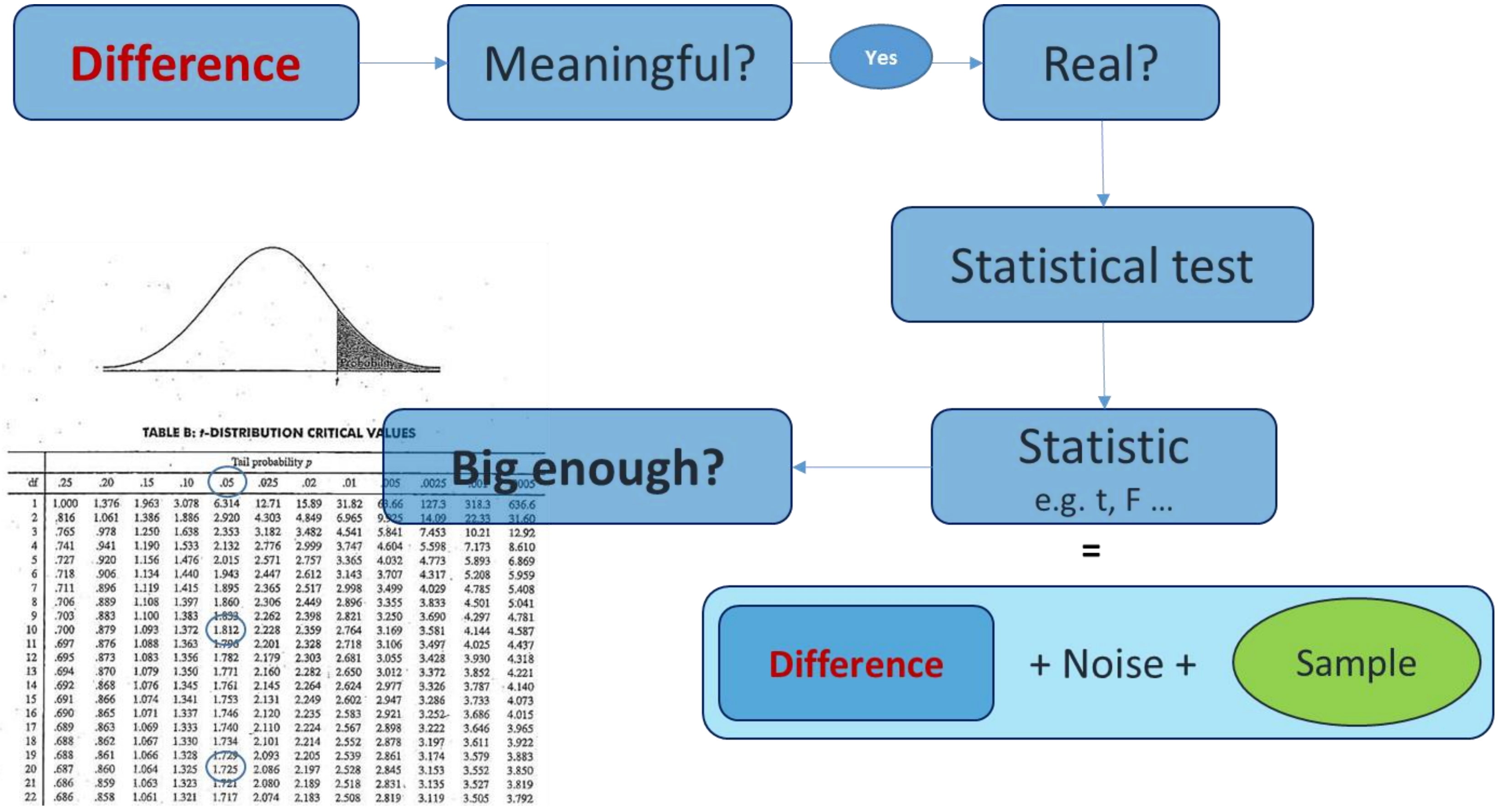
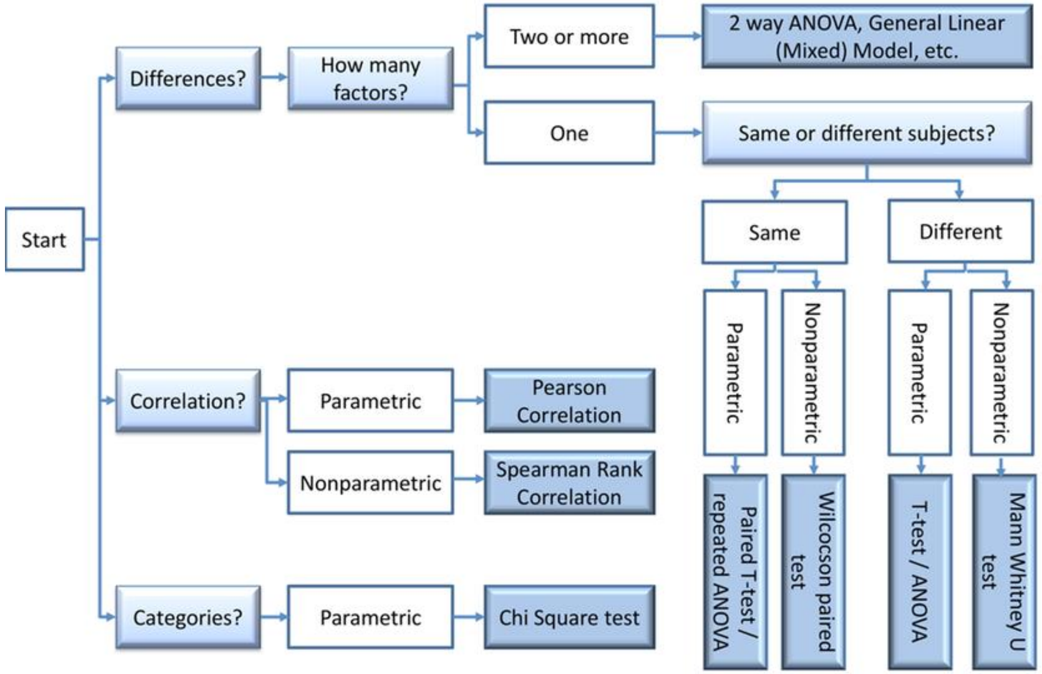


TABLE B: t-DISTRIBUTION CRITICAL VALUES

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	[illegible]	[illegible]
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	[illegible]	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	[illegible]	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792

Statistical Analysis

- **Statistical tests are tools**
  - How do we choose the right tool?

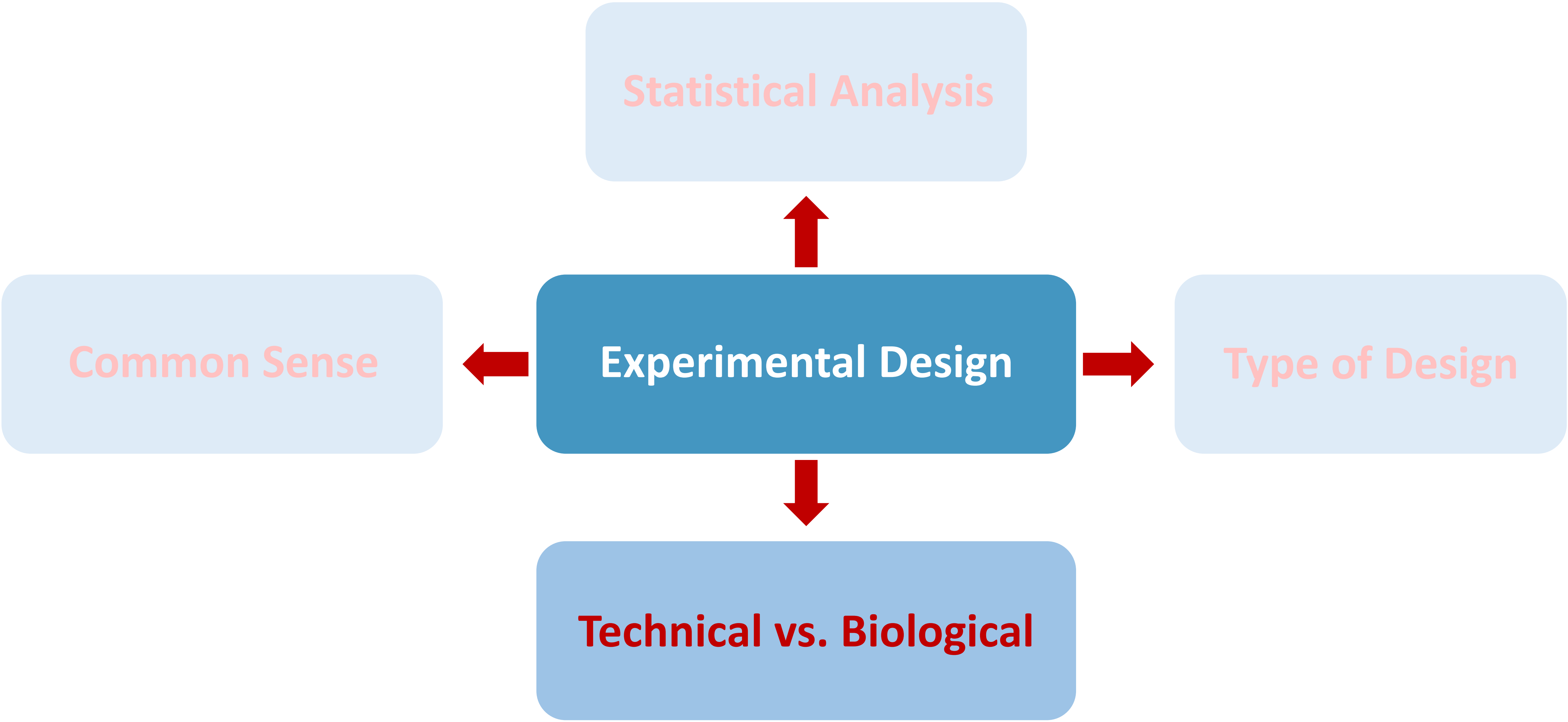


- **The 'job' = the question(s)**
  - The main one: cause → effect
  - What (can) affects that relationship?
    - Both technical and biological

- **Data**
- Nature and behaviour of the data:
  - All statistical tests are associated with assumptions
    - e.g. normality and homogeneity of variance
  - If assumptions not met: bad p-values

- Running a statistical test is easy
  - but making sure it's the right test is not.

- Getting to know the data:
  - Data exploration
  - But also if not one's data:
    - raw or not raw?
    - If normalised/standardised, how?
    - e.g raw counts (qualitative data) vs. normalised (quantitative)


Statistical Analysis
Common Sense
Experimental Design
Type of Design
Technical vs. Biological

Experimental Design → **Technical vs. Biological**

- Definition of **technical** and **biological** depends on the model and the question
  - e.g. mouse, cells …

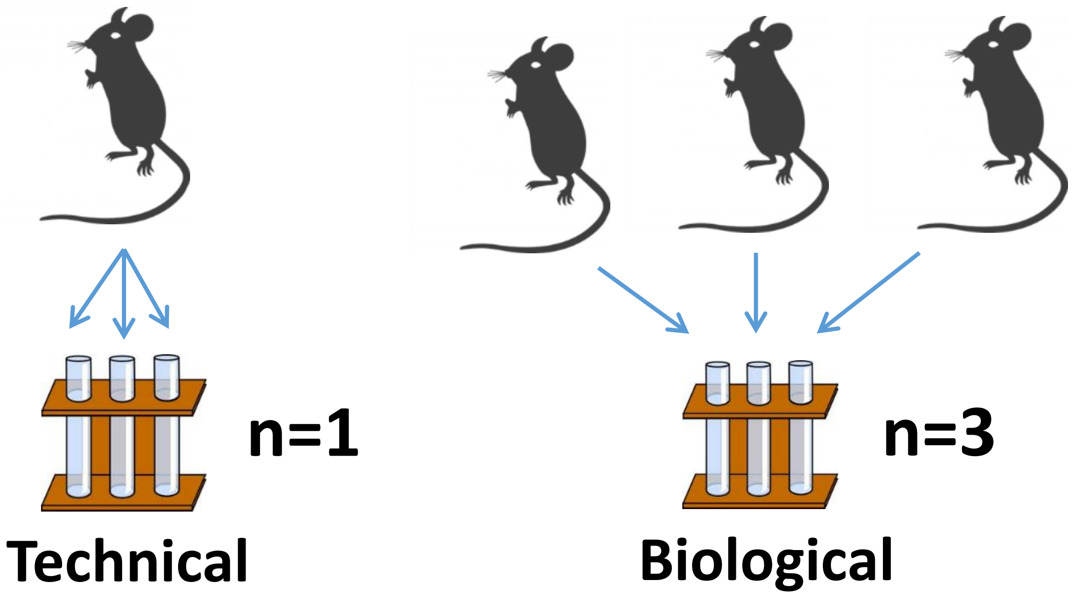
- Question: Why **replicates** at all?
  - To make **proper inference** from sample to general population we need biological samples.

  - Example: difference on weight between grey mice and white mice:
    - cannot conclude anything from one grey mouse and one white mouse randomly selected
      - only 2 biological samples
    - need to repeat the measurements:
      - measure 5 times each mouse: **technical replicates**
      - measure 5 white and 5 grey mice: **biological replicates**

- Answer: Biological replicates are needed to infer to the general population

Always easy to tell the difference?

- Definition of **technical** and **biological** depends on the model and the question.
- The model: mouse, plant … complex organisms in general.
  - Easy: one value per individual organism
    - e.g. weight, neutrophils counts …

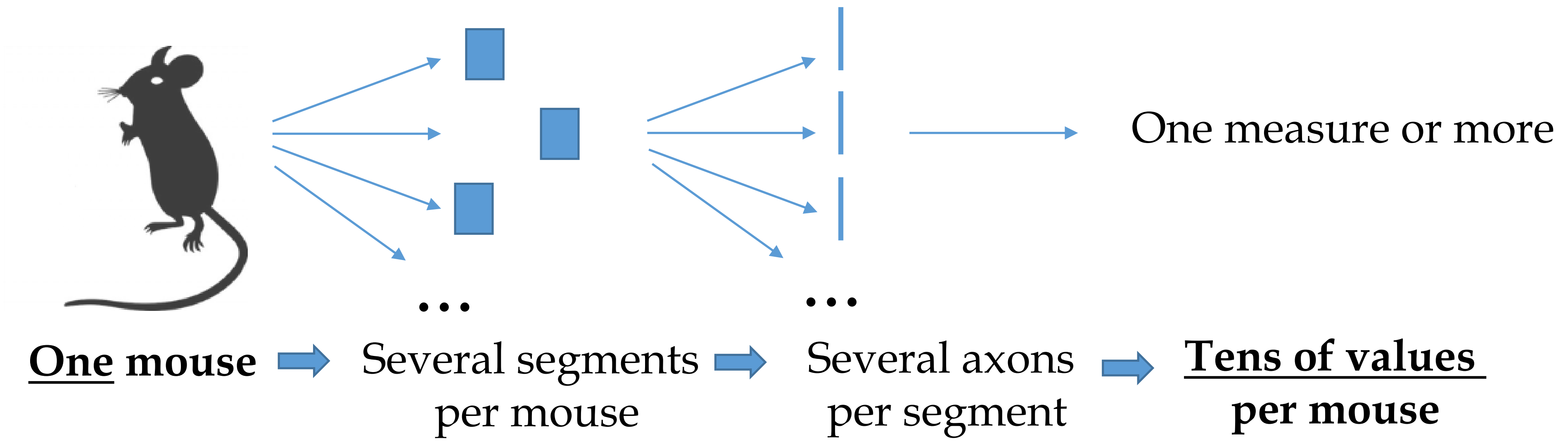


- **What to do?** Mean of technical replicates = 1 biological replicate

Technical vs. Biological

Always easy to tell the difference?

- The model is still: mouse, plant … complex organisms in general.
  - Less easy: more than one value per individual
    - e.g. axon degeneration



- **What to do**? Not one good answer.
  - In this case: mouse = experiment unit (block, split-plot)
    - axons = technical replicates, nerve segments = biological replicates

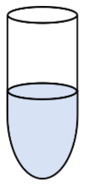
Technical vs. Biological

Always easy to tell the difference?

- The model is : worms, cells …
  - Less and less easy: many ‘individuals’
    - What is ‘n’ in cell culture experiments?

- Cell lines: no biological replication, only technical replication

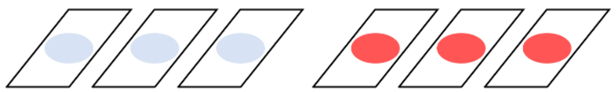
- To make valid inference: valid design



Vial of frozen cells

Control Treatment

Dishes, flasks, wells …
Cells in culture
Point of Treatment

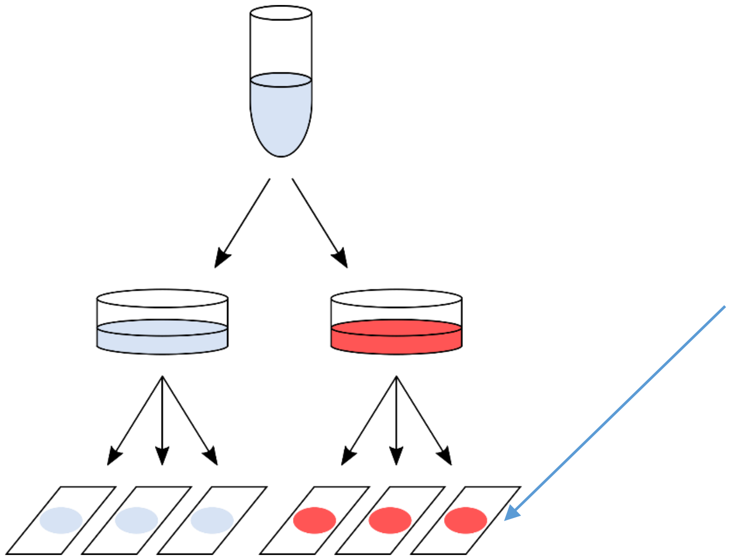


Glass slides
microarrays
lanes in gel
wells in plate
…
Point of Measurements

Technical vs. Biological

Always easy to tell the difference?

- Design 1: As bad as it can get

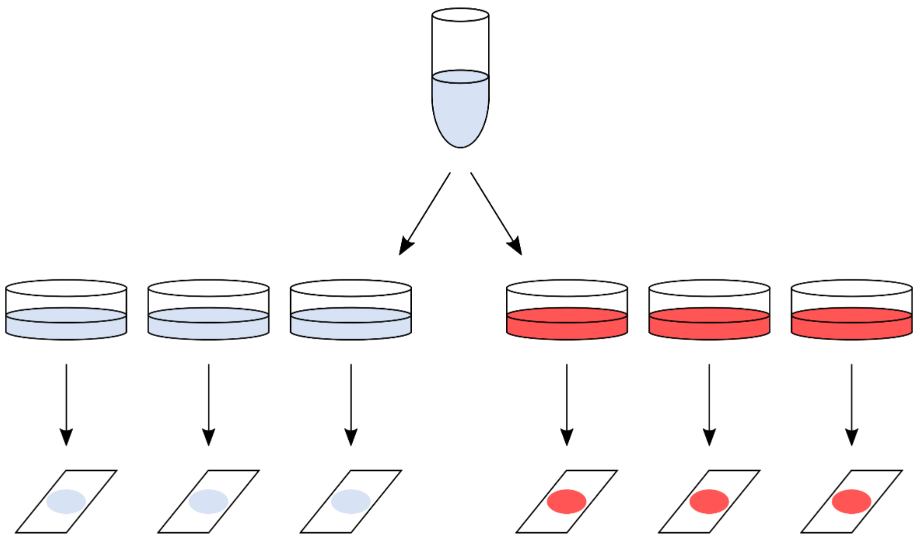


One value per glass slide
e.g. cell count

- After quantification: 6 values
  - But what is the sample size?
    - **n = 1**
      - no independence between the slides
      - variability = pipetting error

Always easy to tell the difference?

- <u>Design 2</u>: Marginally better, but still not good enough

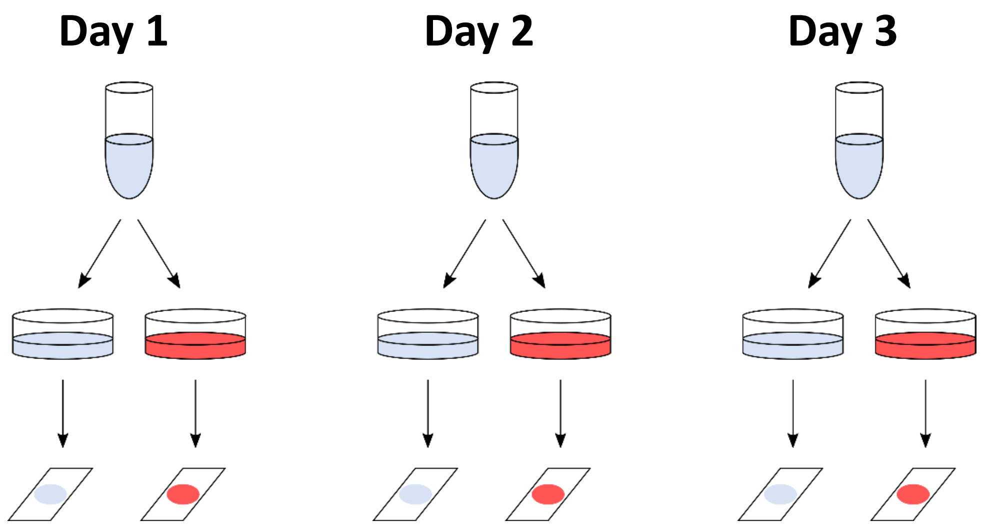


Everything processed on the same day

- After quantification: 6 values
  - But what is the sample size?
    - **n = 1**
      - no independence between the plates
      - variability = a bit better as sample split higher up in the hierarchy

Always easy to tell the difference?

- <u>Design 3</u>: Often, as good as it can get

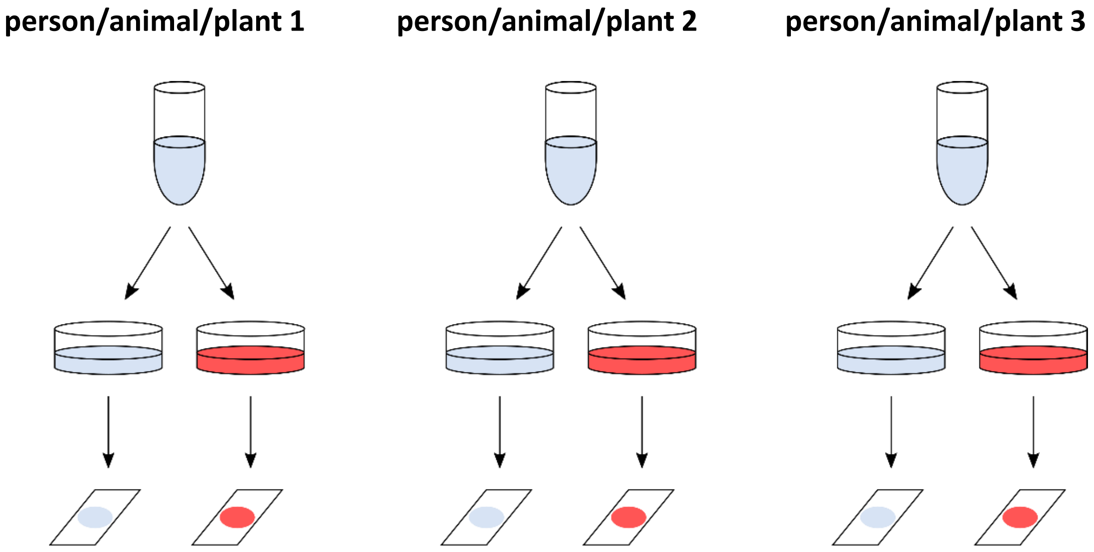


- After quantification: 6 values
  - But what is the sample size?
    - **n = 3**
      - Key difference: the whole procedure is repeated 3 separate times
      - Still technical variability but done at the highest hierarchical level
      - Results from 3 days are (mostly) independent
      - Values from 2 glass slides: paired observations

Technical vs. Biological

Always easy to tell the difference?

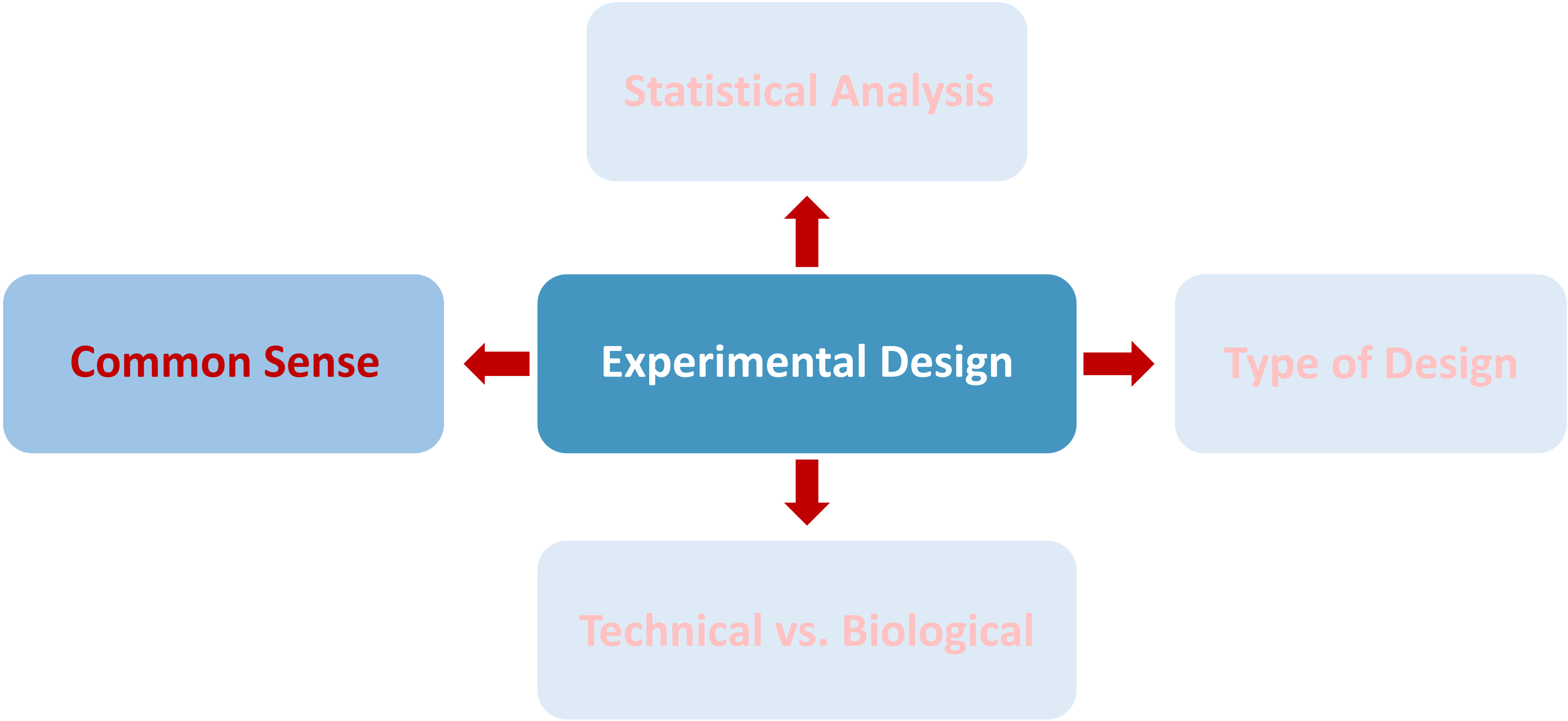
- Design 4: The ideal design



- After quantification: 6 values
    - But what is the sample size?
        - **n = 3**
            - Real biological replicates

Technical and biological replicates
What to remember

- Take the time to identify technical and biological replicates
- Try to make the replications as independent as possible
- Never ever mix technical and biological replicates
- The hierarchical structure of the experiment needs to be respected in the statistical analysis (nested, blocks …).

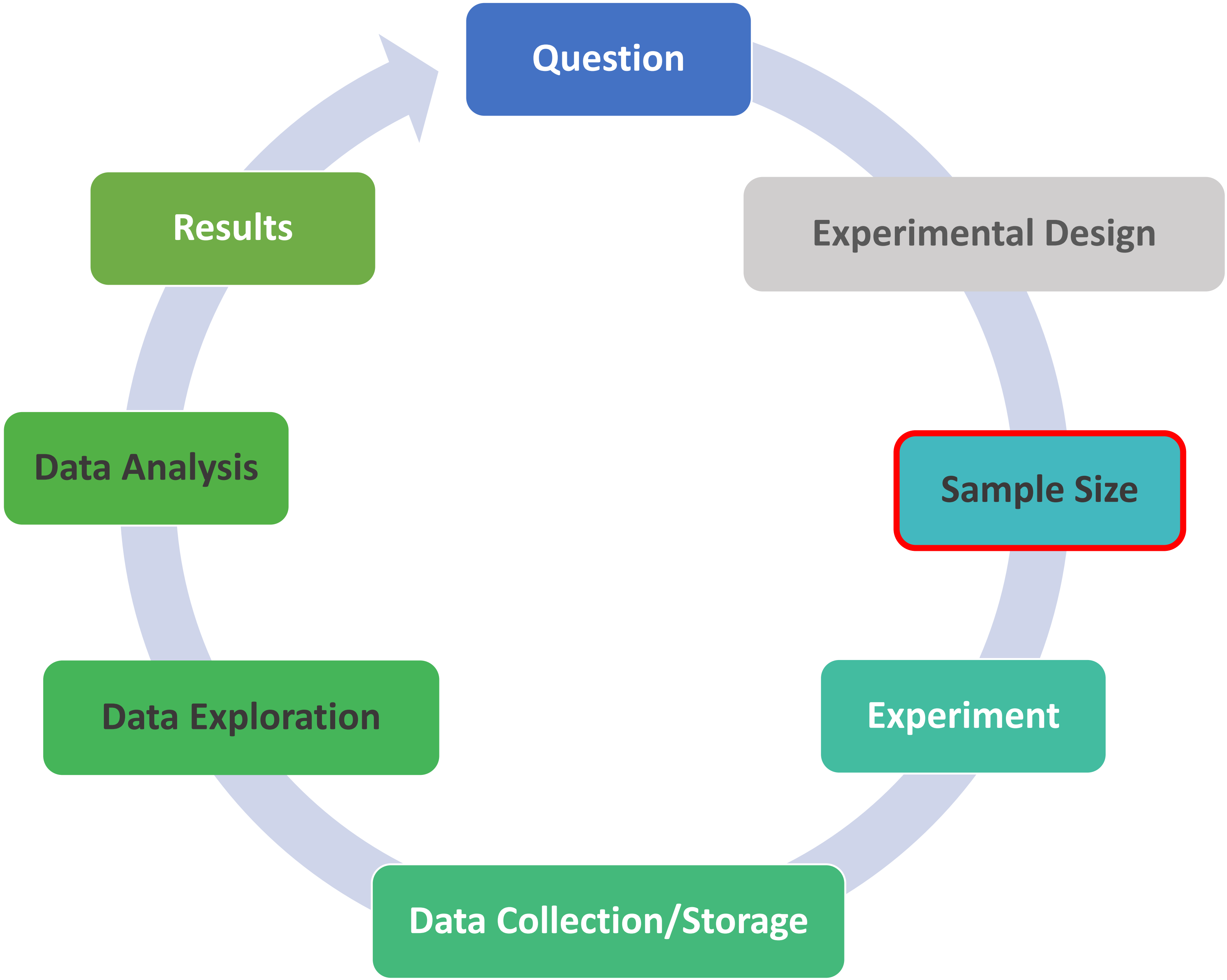

Statistical Analysis
Common Sense
Experimental Design
Type of Design
Technical vs. Biological

Experimental Design → Common Sense

- Design your experiment to be analysable
- The gathering of results or carrying out of a procedure is not the end goal
  - Think about the analysis of the data and design the experiment accordingly
- Imagine how your results will look
- Ask yourself whether these results will address your hypothesis
- Don’t get fixated on being able to perform a cool technique or experimental protocol.
- Don’t be overwhelmed (or try not to be).
- **Draw your experiment and imagine all that can go wrong at each step**

Day 1
Power Analysis

Anne Segonds-Pichon
v2019-06

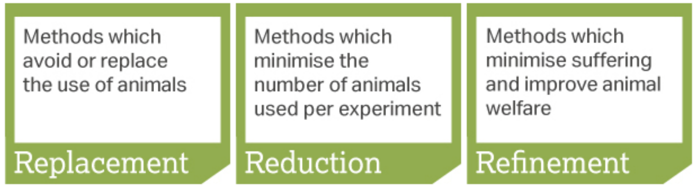

Question
Experimental Design
Sample Size
Experiment
Data Collection/Storage
Data Exploration
Data Analysis
Results

Sample Size: Power Analysis

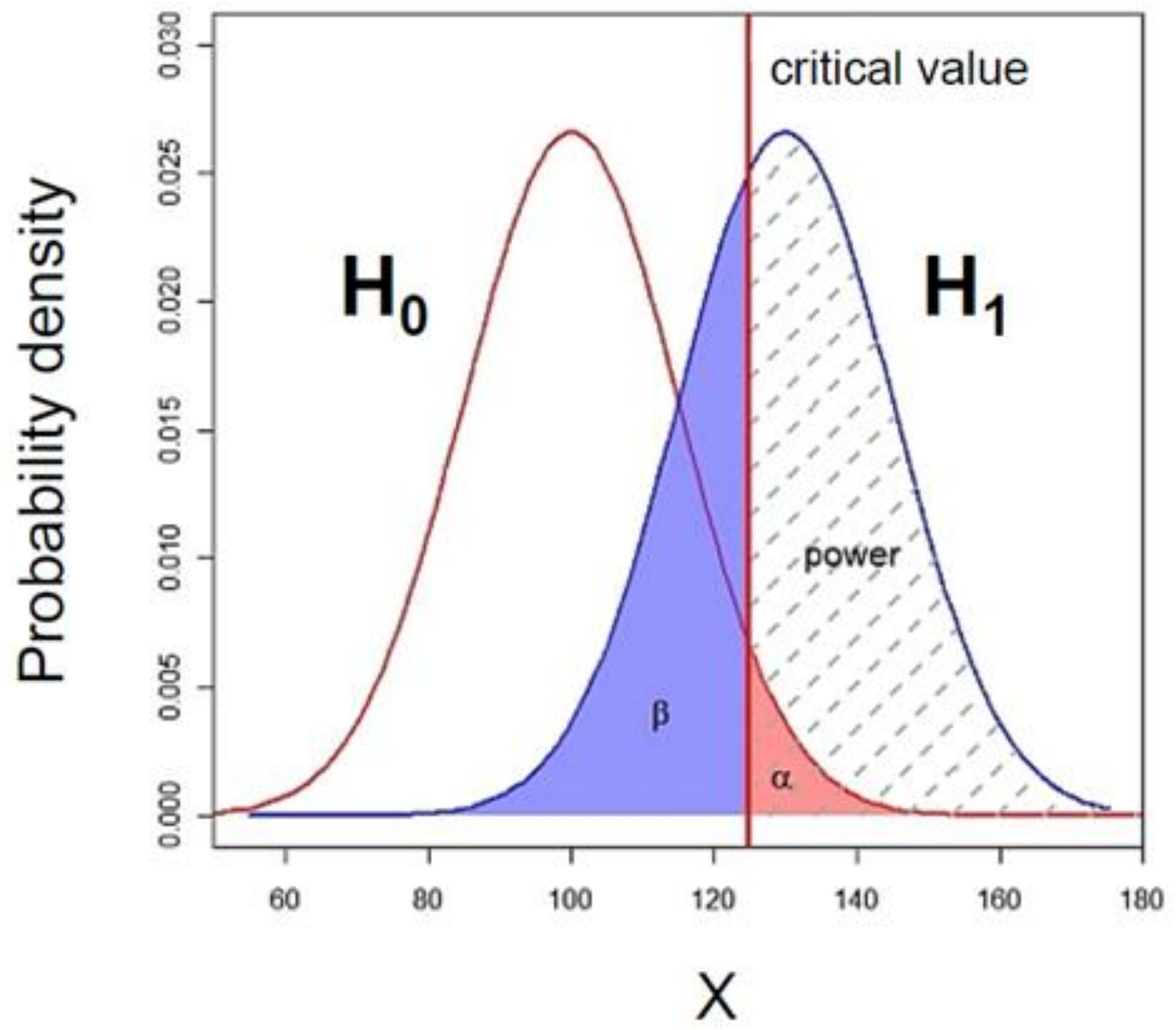
- **Definition of power**: probability that a statistical test will reject a false null hypothesis (H_0).
  - **Translation**: the probability of detecting an effect, given that the effect is really there.

- **In a nutshell**: the bigger the experiment (big sample size), the bigger the power (more likely to pick up a difference).

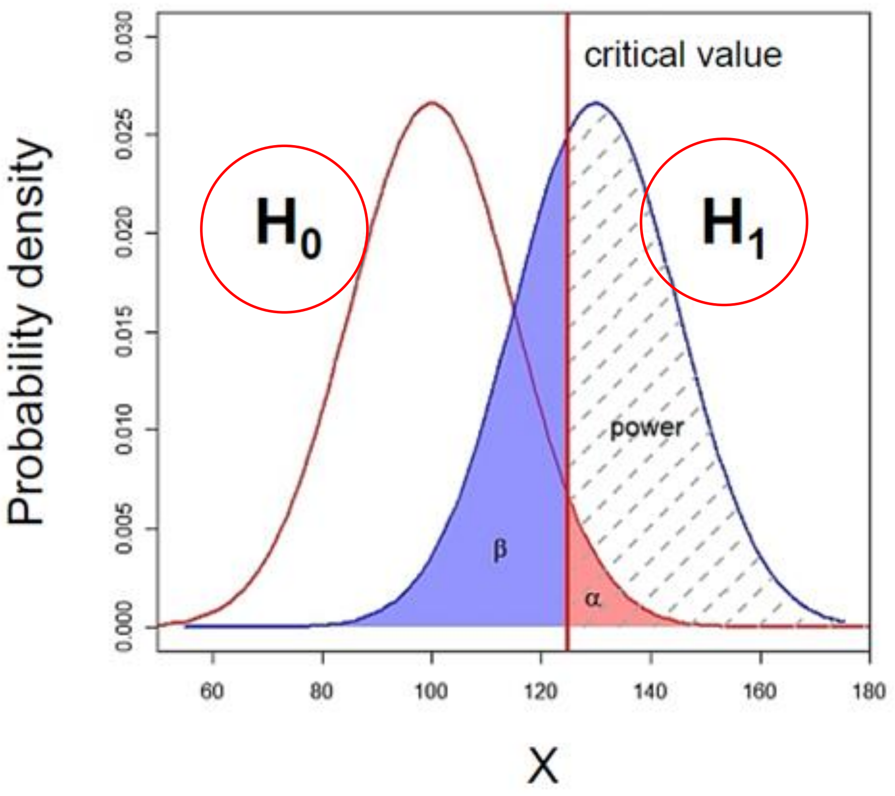
- Main output of a **power analysis**:
  - Estimation of an appropriate **sample size**
    - **Too big**: waste of resources,
    - **Too small**: may miss the effect (p>0.05)+ waste of resources,
    - **Grants**: justification of sample size,
    - **Publications:** reviewers ask for power calculation evidence,
    - **Home office**: the 3 Rs: Replacement, **Reduction** and Refinement.



What does Power look like?

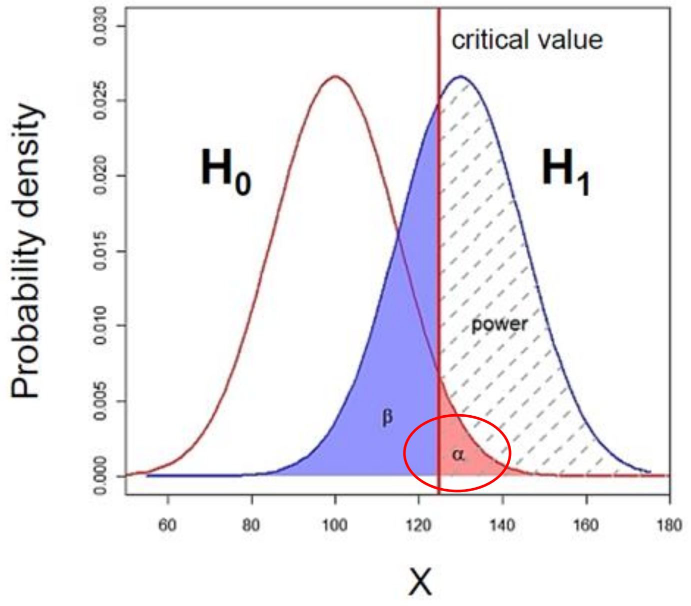


What does Power look like? Null and alternative hypotheses



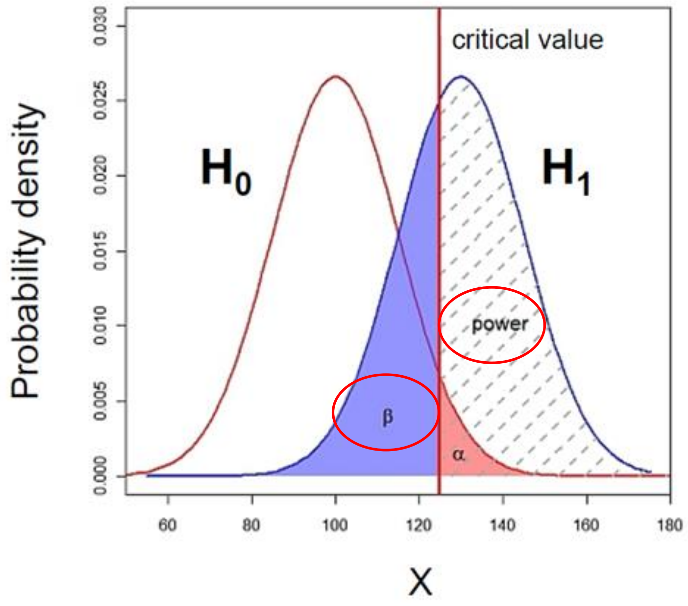
- Probability that the observed result occurs if $\mathbf{H_0}$ is true
  - H_0 : **Null hypothesis** = absence of effect
  - H_1: **Alternative hypothesis** = presence of an effect

What does Power look like? Type I error α



- **α :** the threshold value that we measure p-values against.
  - For results with 95% level of confidence: **α = 0.05**
  - = probability of **type I error**
- **p-value**: probability that the observed statistic occurred by chance alone
- **Statistical significance**: comparison between **α** and the **p-value**
  - p-value < 0.05: reject H_0 and p-value > 0.05: fail to reject H_0

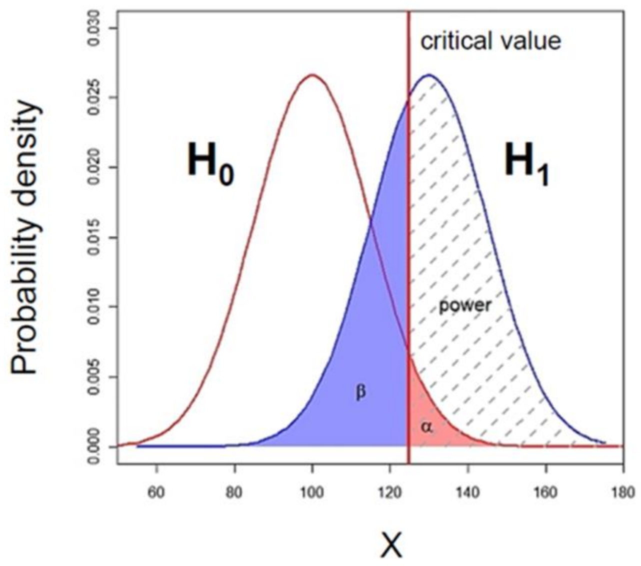
What does Power look like? Power and Type II error β



- **Type II error (β)** is the failure to reject a <u>false</u> H_0
  - Probability of missing an effect which is really there.
  - **Power**: probability of detecting an effect which is really there

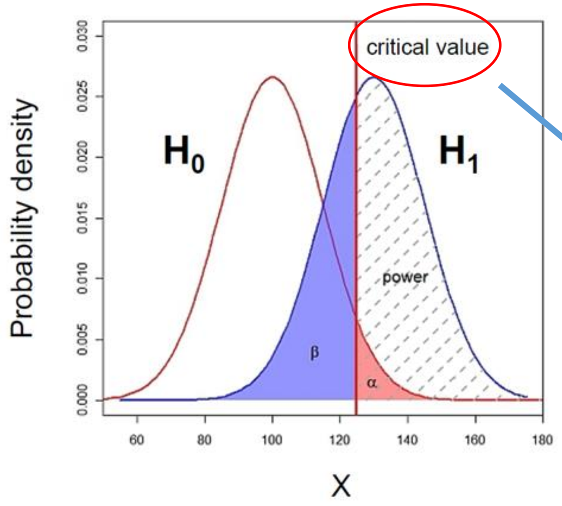
  - Direct relationship between **Power** and **type II error**:
    - **Power** = 1 – **β**

What does Power look like? Power = 80%



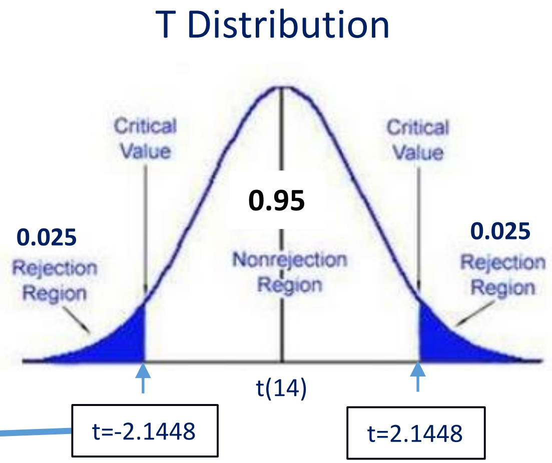
- **Type II error (β)** is the failure to reject a <u>false</u> H_0
  - Probability of missing an effect which is really there.
  - **Power**: probability of detecting an effect which is really there
    - Direct relationship between **Power** and type II error:
    - if **Power** = 0.8 then β = 1- **Power** = 0.2 (20%)
  - Hence a true difference will be missed 20% of the time
  - **General convention: 80%** but could be more
  - Cohen (1988):
    - For most researchers: Type I errors are four times more serious than Type II errors so **0.05 * 4 = 0.2**
    - Compromise: 2 groups comparisons:
      - 90% = +30% sample size
      - 95% = +60%s sample size

What does Power look like? Critical value



Example: 2-tailed t-test with n=15 (df=14)

df	0.20	0.10	0.05	0.02	0.01	0.001
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728



- In **hypothesis testing**, a **critical value** is a point on the test distribution that is compared to the **test statistic** to determine whether to reject the null **hypothesis**
  - Example of test statistic: t-value
- Absolute value of **test statistic** > **critical value** = statistical significance
  - Example: t-value > critical t-value -> p<0.05

To recapitulate:

- The null hypothesis (H_0): H_0 = no effect

- The aim of a statistical test is to reject or not $H_{0.}$

Statistical decision	True state of H_0	
	H_0 True (no effect)	**H_0 False (effect)**
Reject H_0	**Type I error α** **False Positive**	Correct **True Positive**
Do not reject H_0	Correct **True Negative**	**Type II error β** **False Negative**

- Traditionally, a test or a difference are said to be "**significant**" if the probability of type I error is: **α =< 0.05**
- **High specificity** = low **False Positives** = low **Type I error**
- **High sensitivity** = low **False Negatives** = low **Type II error**

Sample Size: Power Analysis

The power analysis depends on the relationship between 6 variables:

- the **difference** of biological interest
- the **variability** in the data (**standard deviation**)

} **Effect size**

- the significance level (5%)
- the desired power of the experiment (80%)
- the **sample size**
- the alternative hypothesis (ie one or two-sided test)

The effect size: what is it?

- The **effect size**: minimum meaningful effect of biological relevance.
  - Absolute difference + variability

- How to determine it?
  - Substantive knowledge
  - Previous research
  - Conventions

- **Jacob Cohen**
  - Author of several books and articles on power
  - Defined small, medium and large effects for different tests

Test	Relevant effect size	Effect Size Threshold		
		Small	Medium	Large
t-test for means	d	0.2	0.5	0.8
F-test for ANOVA	f	0.1	0.25	0.4
t-test for correlation	r	0.1	0.3	0.5
Chi-square	w	0.1	0.3	0.5
2 proportions	h	0.2	0.5	0.8

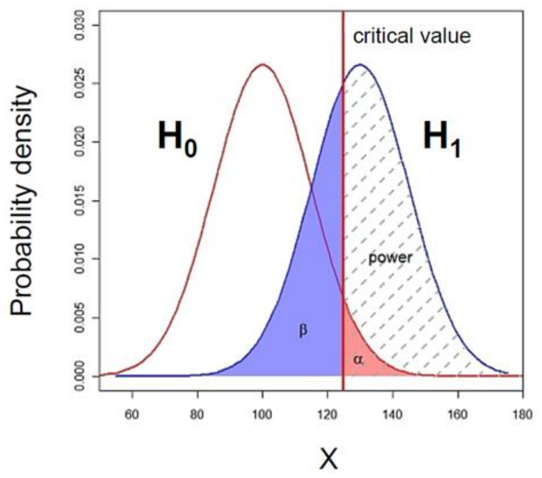
The effect size: how is it calculated?
The absolute difference

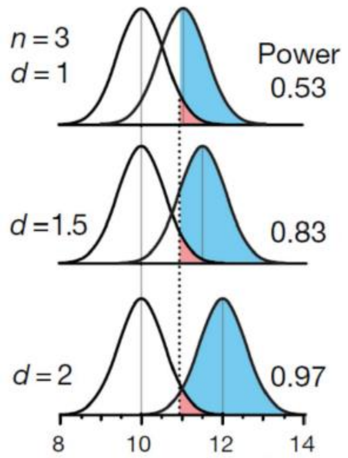
- It depends on the type of difference and the data
  - Easy example: comparison between 2 means

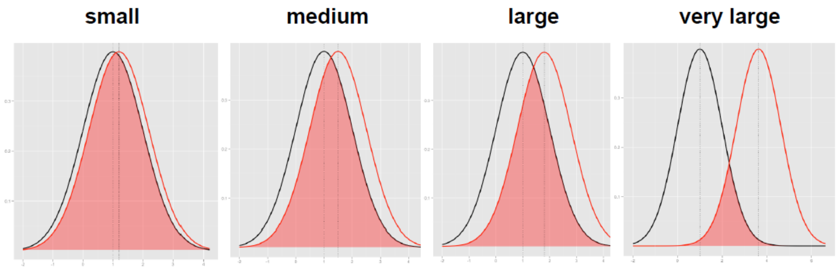
$$\text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{Standard Deviation}}$$

Absolute difference

- The bigger the effect (the absolute difference), the bigger the power
= the bigger the probability of picking up the difference







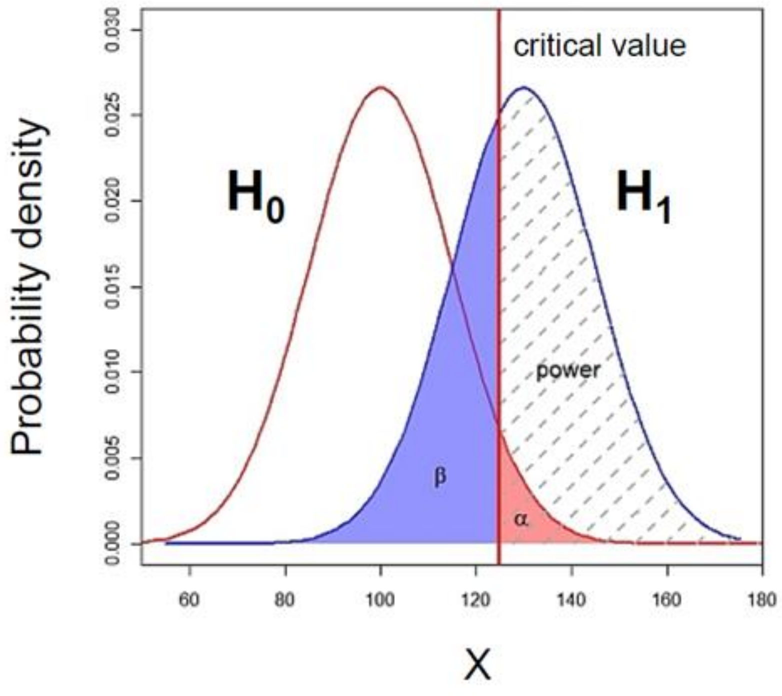
http://rpsychologist.com/d3/cohend/

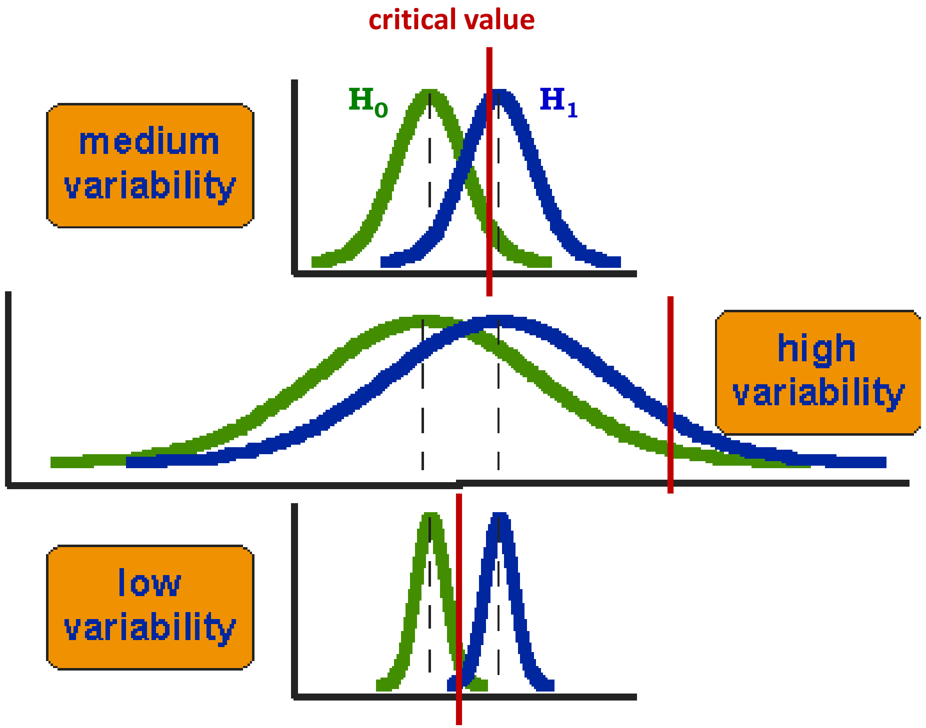
The effect size: how is it calculated?

The standard deviation

- The bigger the variability of the data, the smaller the power

$$\text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{Standard Deviation}}$$





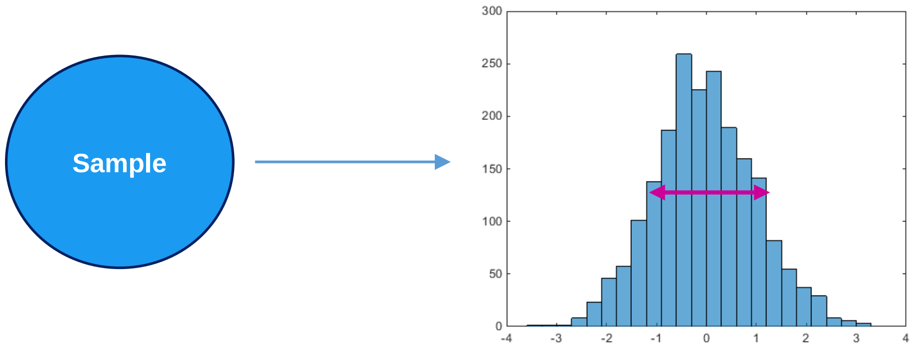
Power Analysis

The power analysis depends on the relationship between 6 variables:

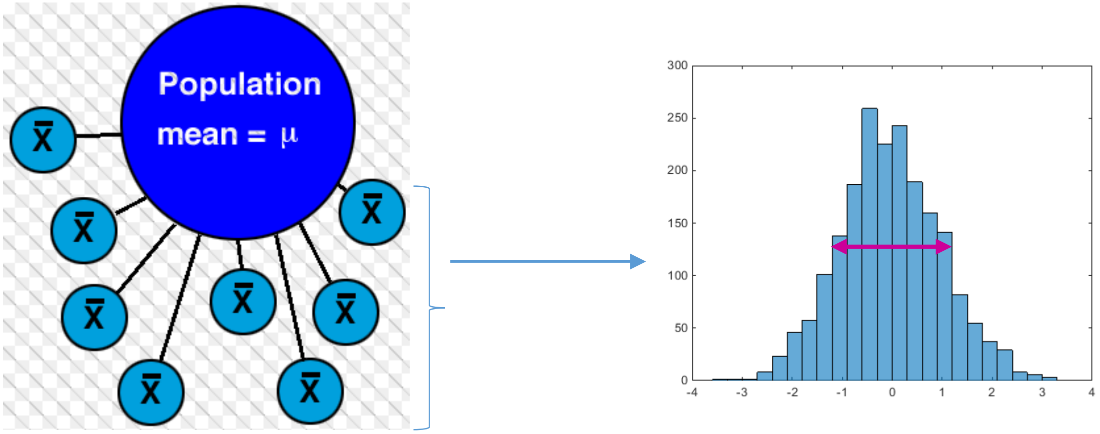
- the **difference** of biological interest
- the **standard deviation**
- **the significance level (5%) ($p < 0.05$) α**
- **the desired power of the experiment (80%) β**
- the **sample size**
- the alternative hypothesis (ie one or two-sided test)

The sample size

- Most of the time, the output of a power calculation.
- **The bigger the sample, the bigger the power**
  - but how does it work actually?
- In reality it is difficult to reduce the variability in data, or the contrast between means,
  - most effective way of improving power:
    - increase the sample size.
- The standard deviation of the sample distribution= Standard Error of the Mean: **SEM** =SD/√N
    - SEM decreases as sample size increases

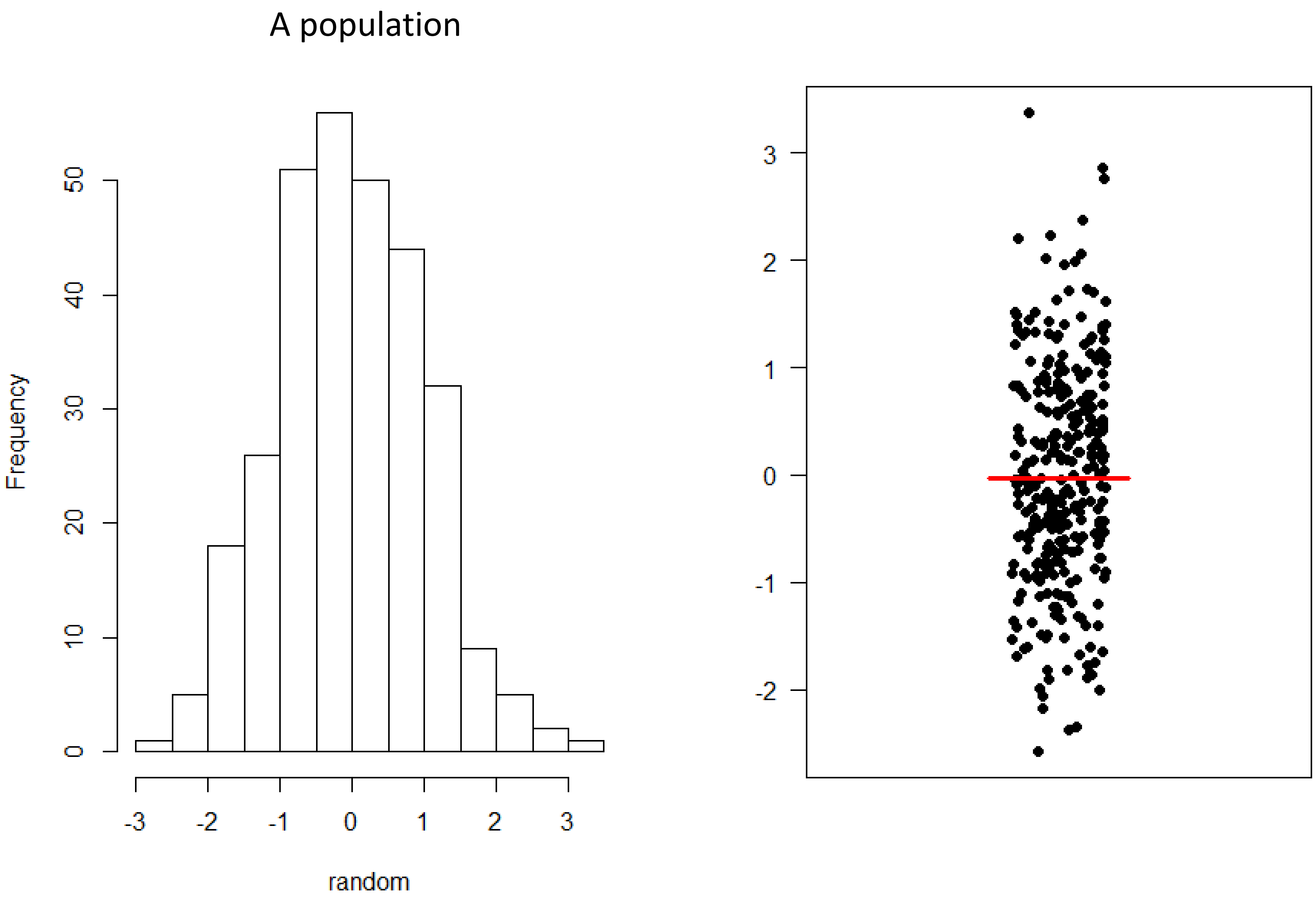


Standard deviation

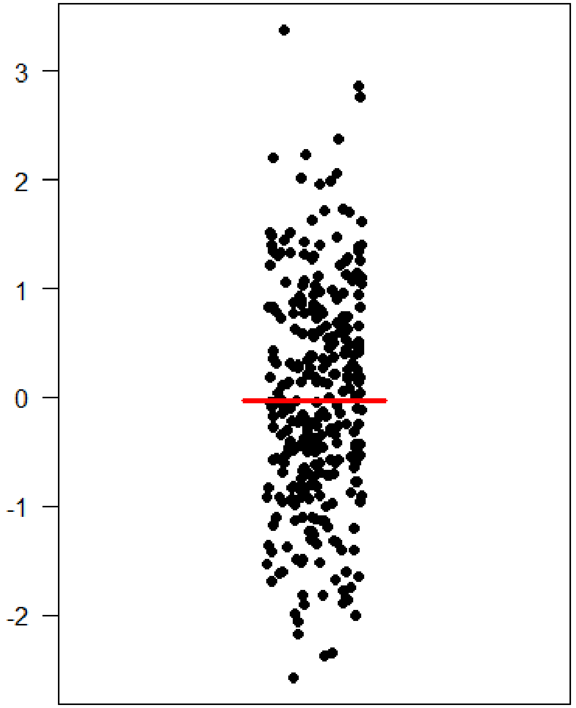


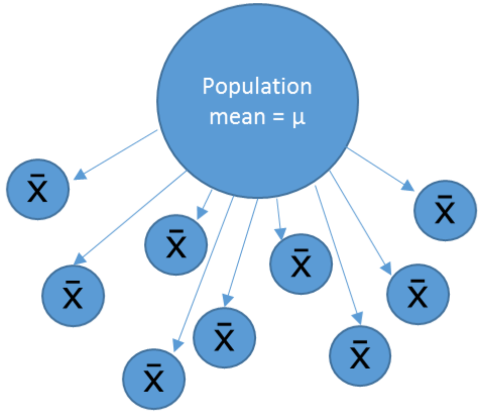
SEM: standard deviation of the sample distribution

The sample size



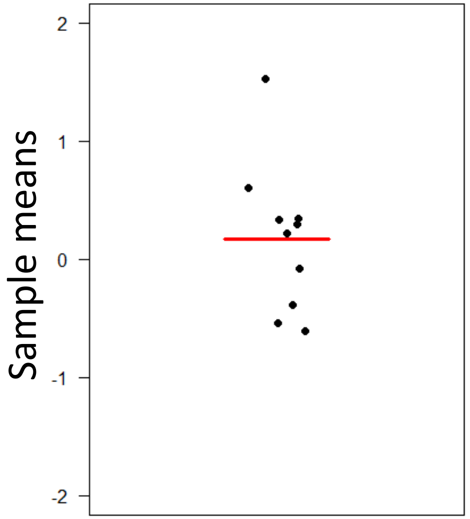
The sample size



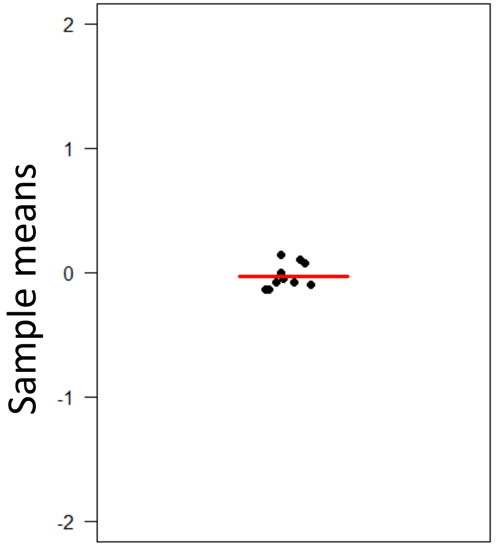


'Infinite' number of samples
Samples means = X̄

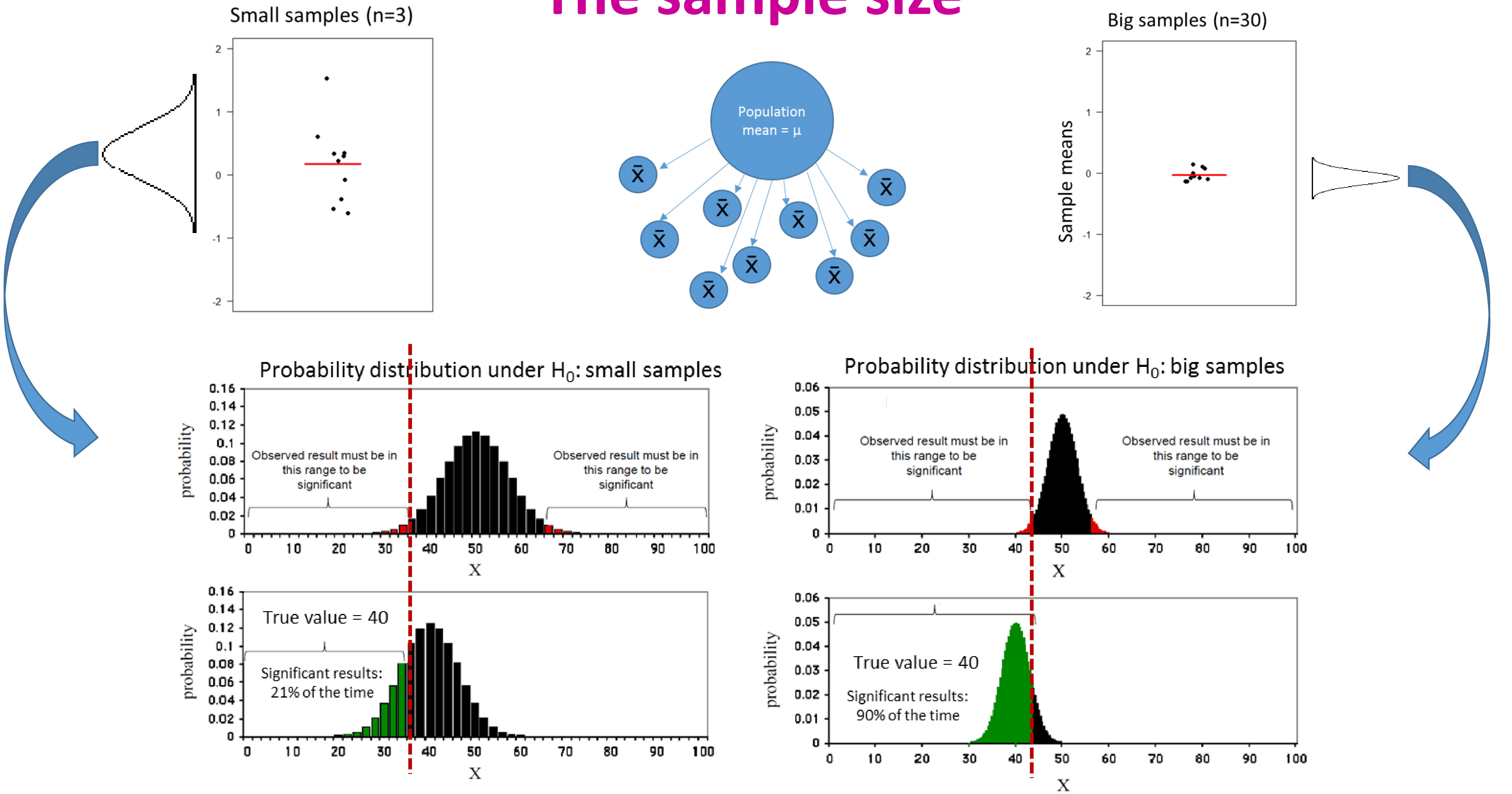
Small samples (n=3)



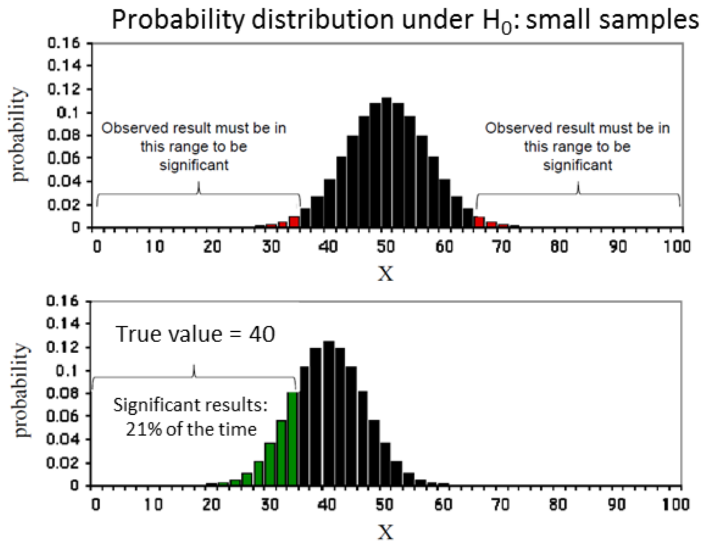
Big samples (n=30)

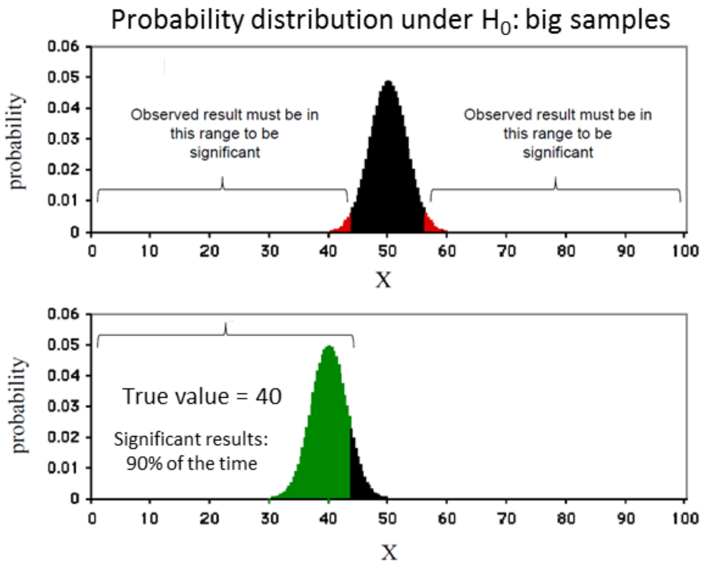


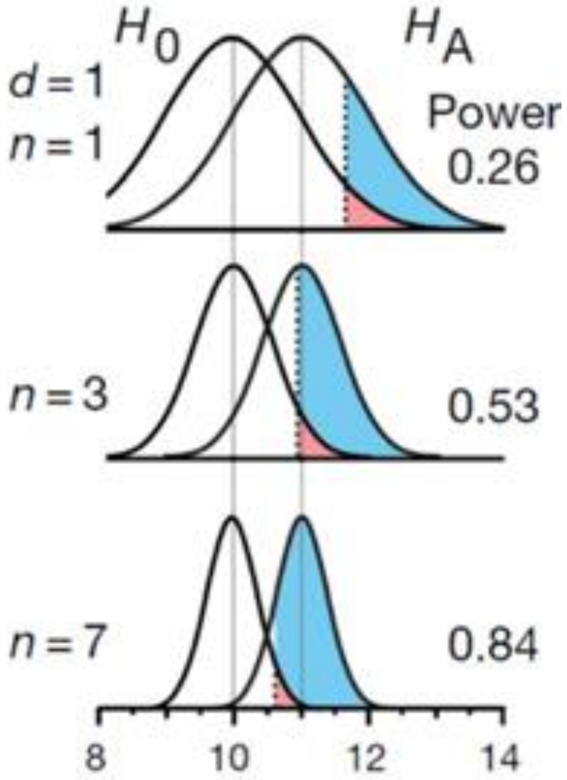
The sample size



The sample size





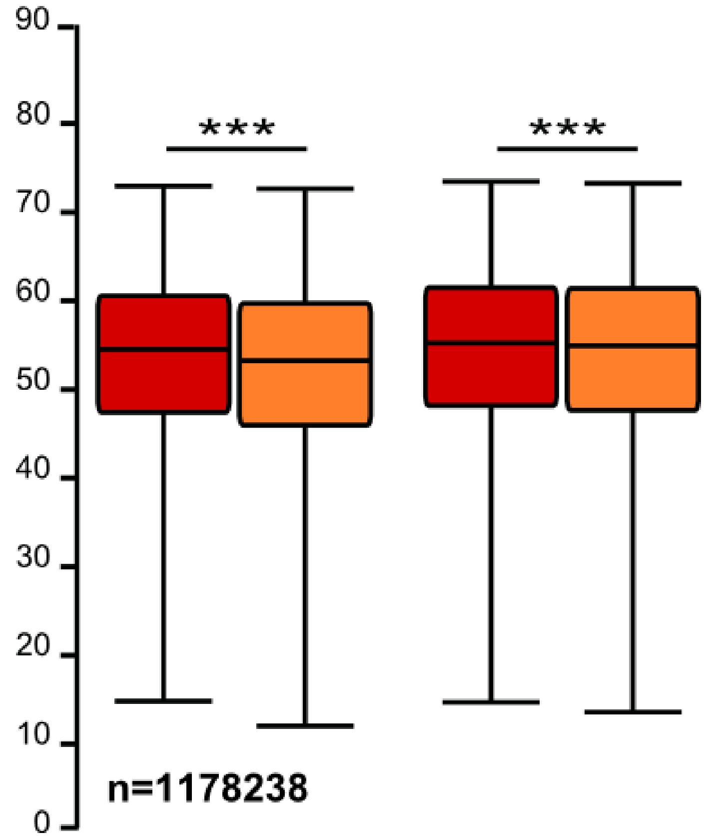


The sample size: the bigger the better?

- It takes huge samples to detect tiny differences but tiny samples to detect huge differences.

- What if the tiny difference is meaningless?
  - Beware of **overpower**
  - Nothing wrong with the stats: it is all about interpretation of the results of the test.

- Remember the important first step of power analysis
  - **What is the effect size of biological interest?**



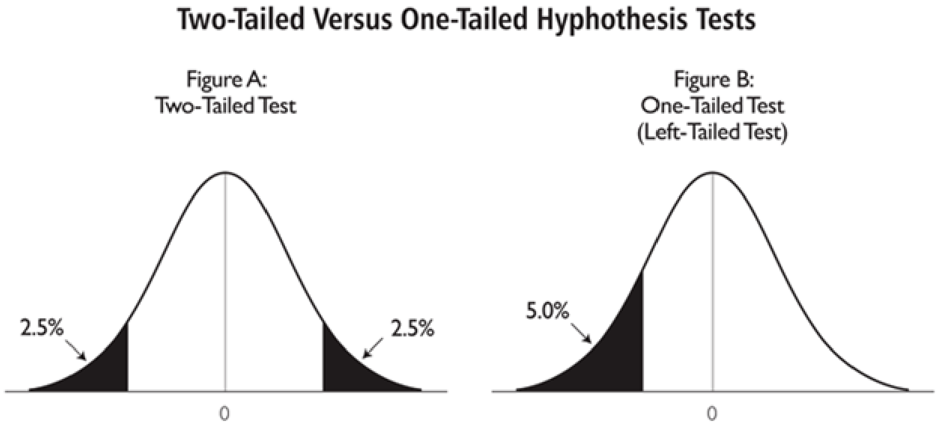
Power Analysis

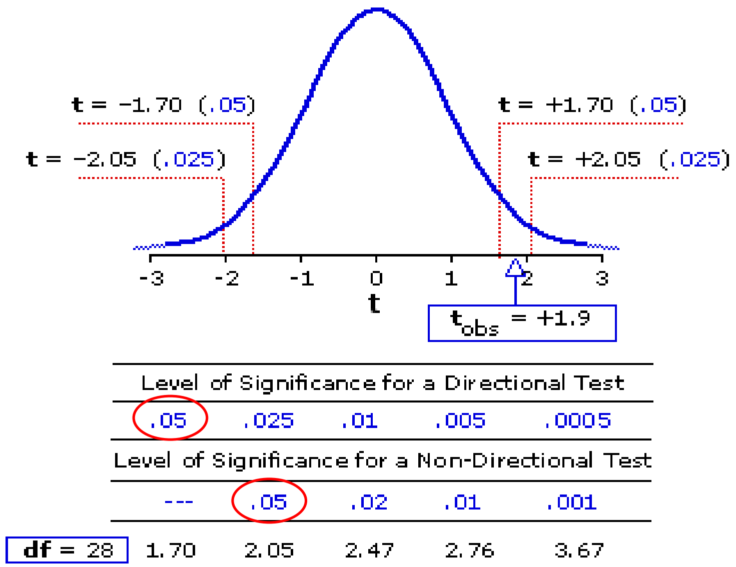
The power analysis depends on the relationship between 6 variables:

- the **effect size** of biological interest
- the **standard deviation**
- **the significance level (5%)**
- **the desired power of the experiment (80%)**
- the **sample size**
- the alternative hypothesis (ie one or two-sided test)

The alternative hypothesis: what is it?

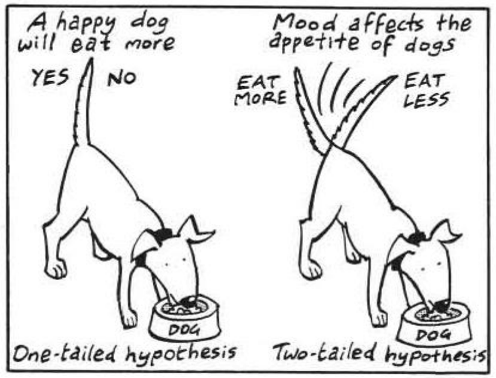
- One-tailed or 2-tailed test? One-sided or 2-sided tests?





- Is the question:
  - Is the there a difference?
  - Is it bigger than or smaller than?

- Can rarely justify the use of a one-tailed test
- Two times easier to reach significance with a one-tailed than a two-tailed
  - Suspicious reviewer!



Hypothesis

↓

Experimental design
Choice of a Statistical test

↓

Power analysis

↓

Sample size

↓

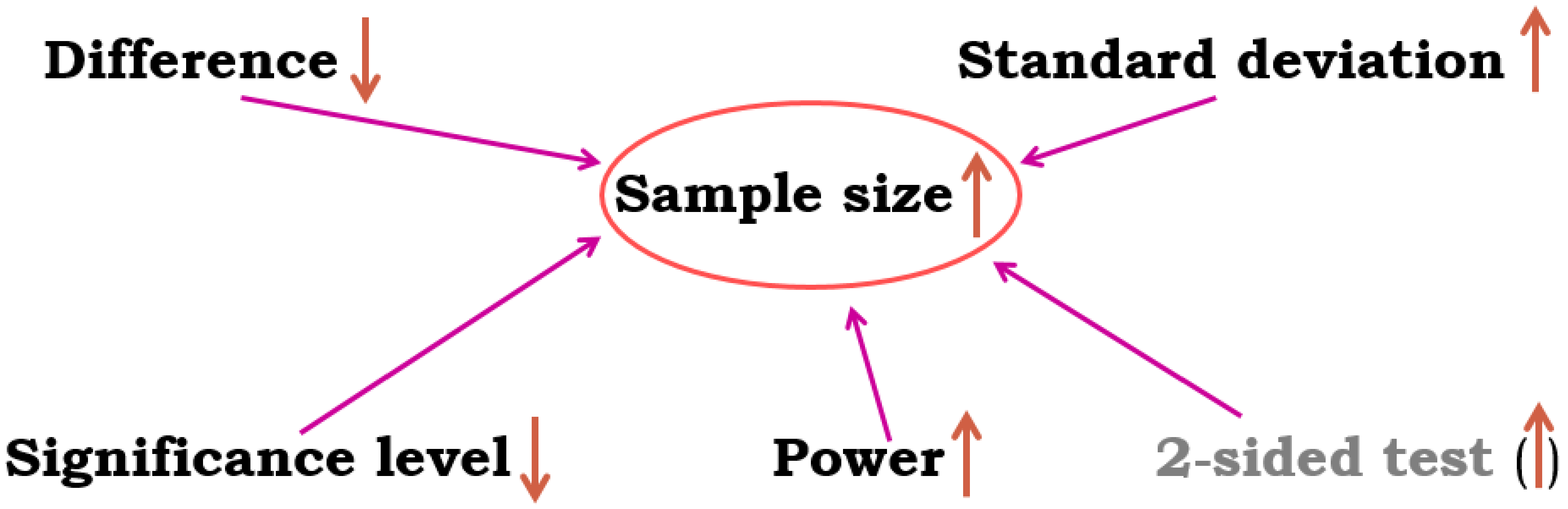
Experiment(s)

↓

(Stat) analysis of the results

- **Fix any five of the variables and a mathematical relationship can be used to estimate the sixth.**

e.g. What sample size do I need to have a 80% probability (**power**) to detect this particular effect (**difference** and **standard deviation**) at a 5% **significance level** using a **2-sided test**?



- **Good news**:

there are packages that can do the power analysis for you ... providing you have some prior knowledge of the key parameters!

difference + standard deviation = effect size

- **Free packages**:
  - R
  - **G*Power** and InVivoStat
  - Russ Lenth's power and sample-size page:
    - http://www.divms.uiowa.edu/~rlenth/Power/

- Cheap package: StatMate (~ $95)

- Not so cheap package: MedCalc (~ $495)

Power Analysis
Let’s do it

- **Examples of power calculations**:
  - Comparing 2 proportions: **Exercise 1**
  - Comparing 2 means: **Exercise 2**

Sample Size: Power Analysis

Exercise 1:

- Scientists have come up with a solution that will reduce the number of lions being shot by farmers in Africa: painting eyes on cows' bottoms.
- Early trials suggest that lions are less likely to attack livestock when they think they're being watched
  - Fewer livestock attacks could help farmers and lions co-exist more peacefully.
- Pilot study over 6 weeks:
  - 3 out of 39 unpainted cows were killed by lions, none of the 23 painted cows from the same herd were killed.

http://www.sciencealert.com/scientists-are-painting-eyes-on-cows-butts-to-stop-lions-getting-shot

Sample Size: Power Analysis

Exercise 1:

- **Questions**:
  - Do you think the observed effect is meaningful to the extent that such a ‘treatment’ should be applied?
    Consider ethics, economics, conservation …
  - Run a power calculation to find out how many cows should be included in the study.

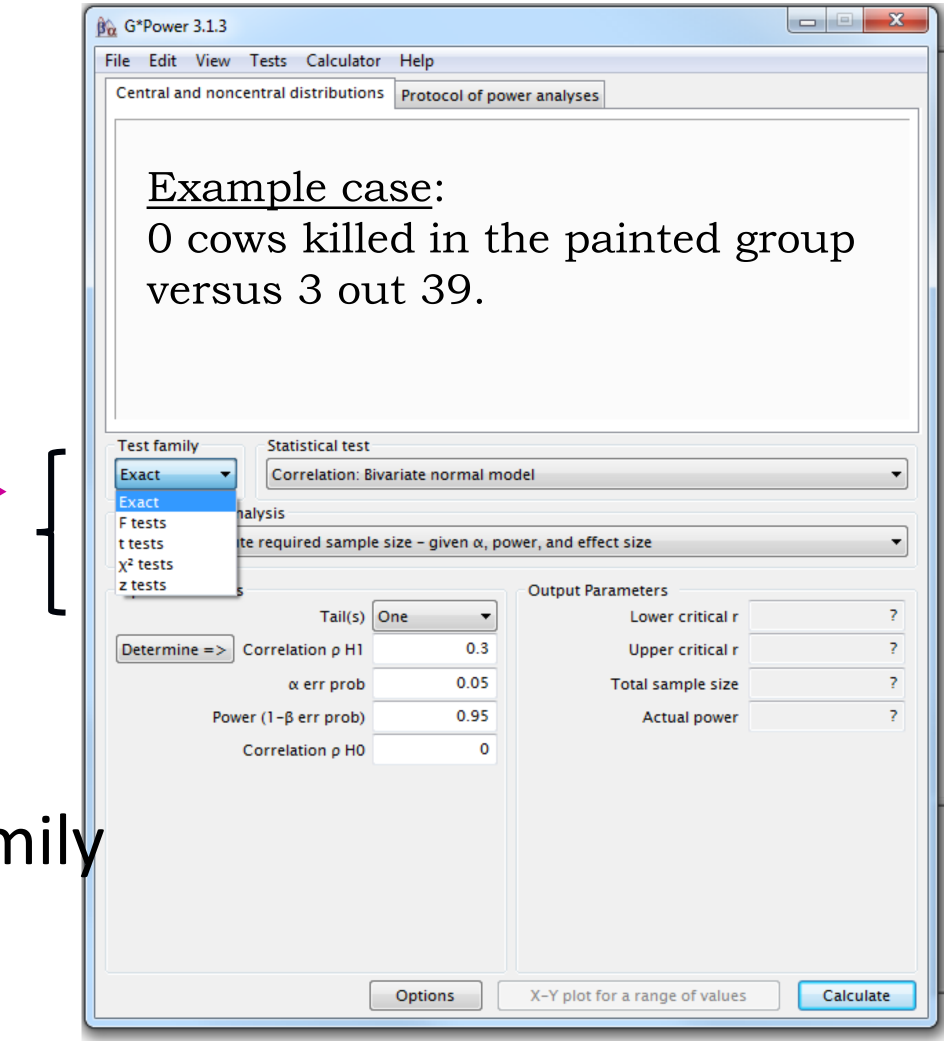
- **Effect size**: measure of distance between 2 proportions or probabilities

- Comparison between 2 proportions: **Fisher’s exact test**

http://www.sciencealert.com/scientists-are-painting-eyes-on-cows-butts-to-stop-lions-getting-shot

Power Analysis
Comparing 2 proportions

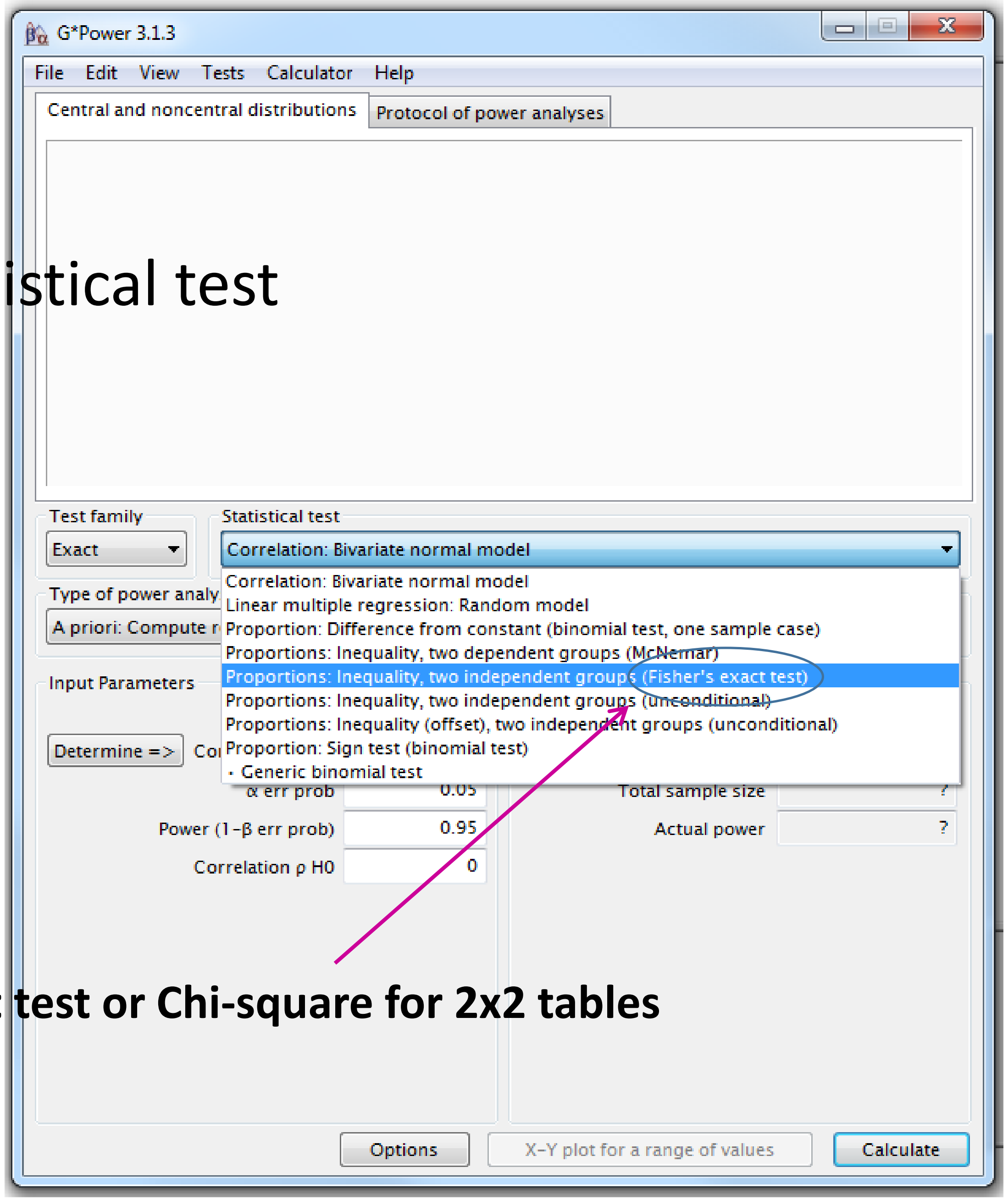
Four steps to Power



Step1: choice of Test family

G*Power

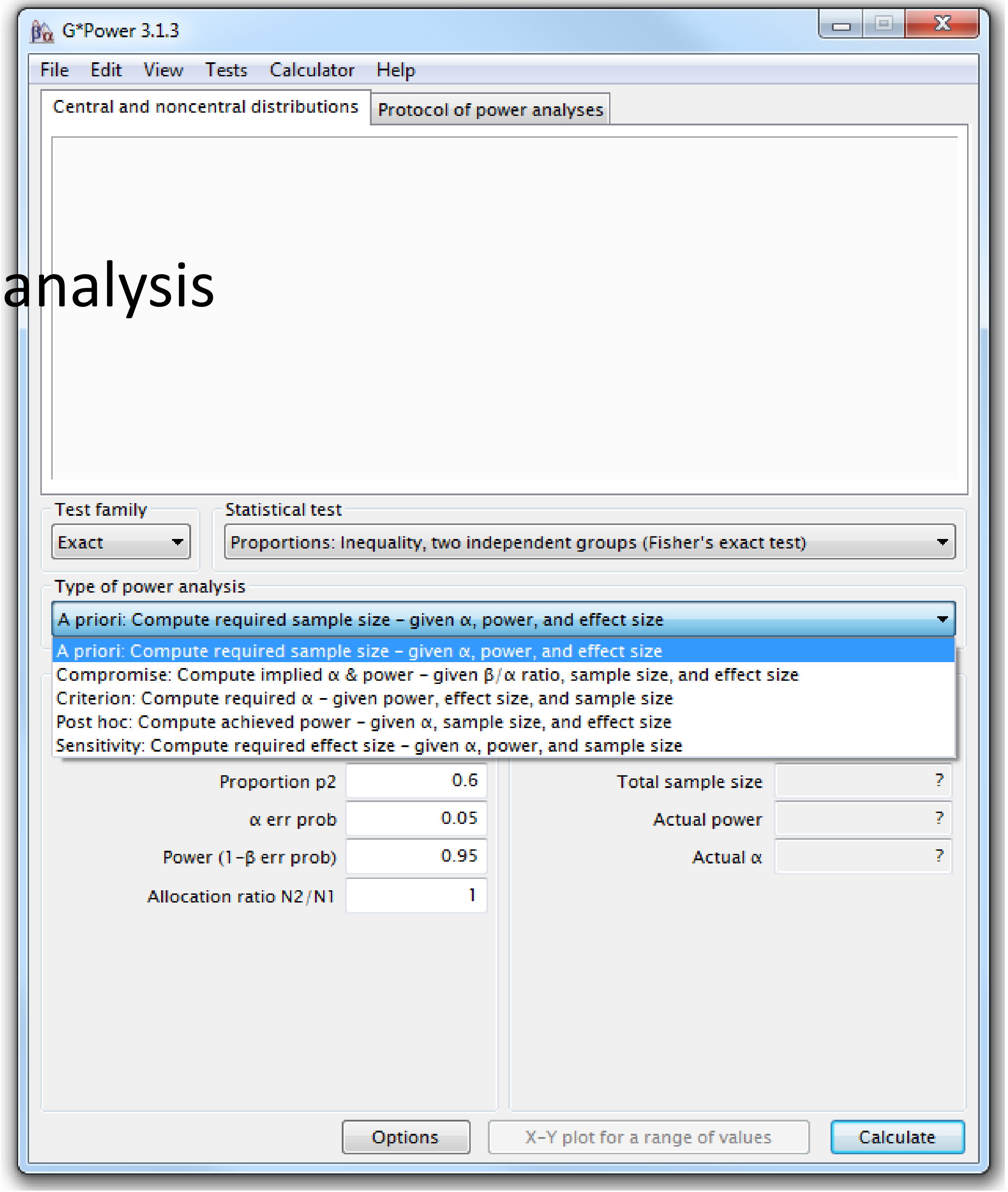
Step 2 : choice of Statistical test



Fisher's exact test or Chi-square for 2x2 tables

G*Power

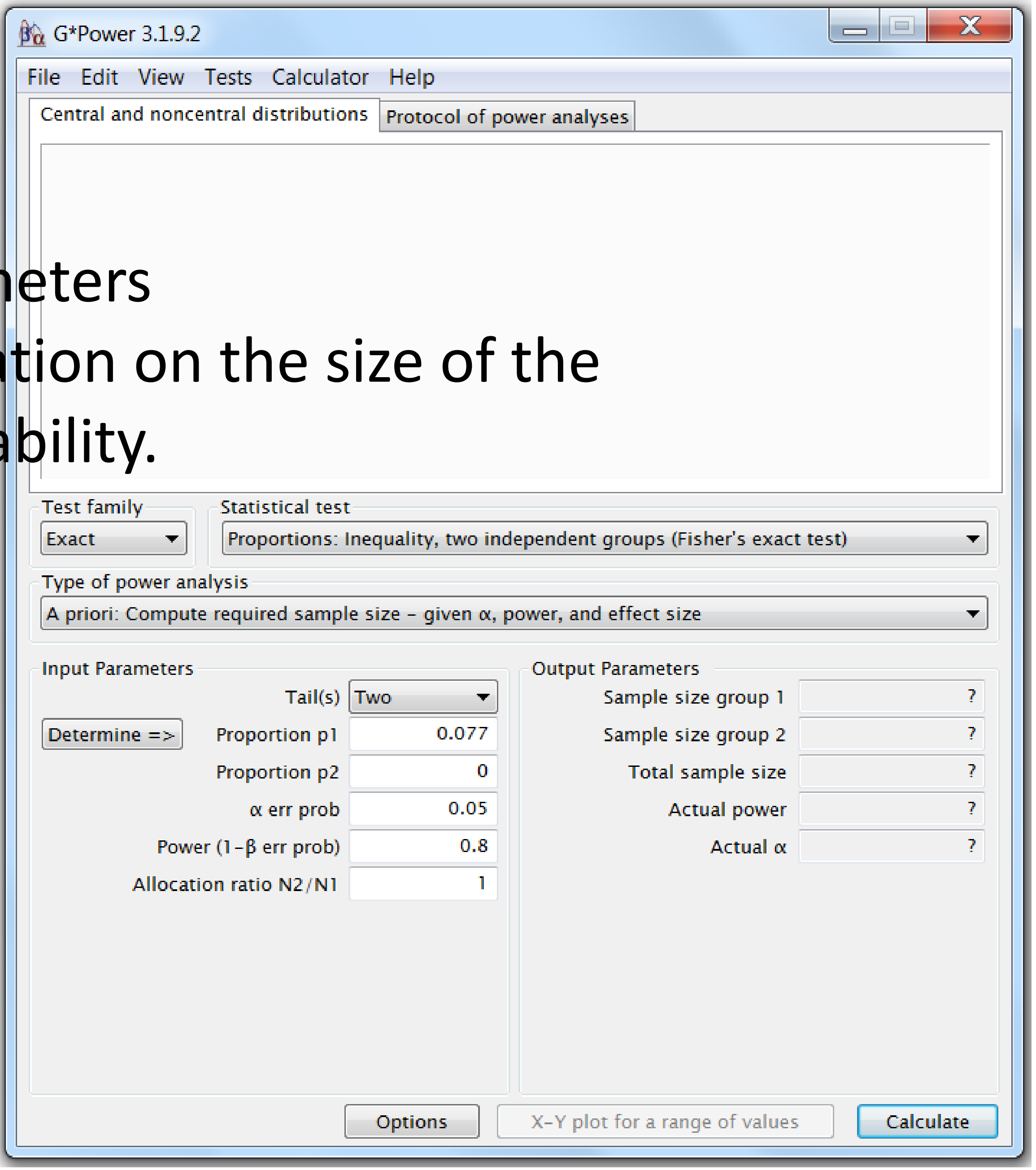
Step 3: Type of power analysis



G*Power

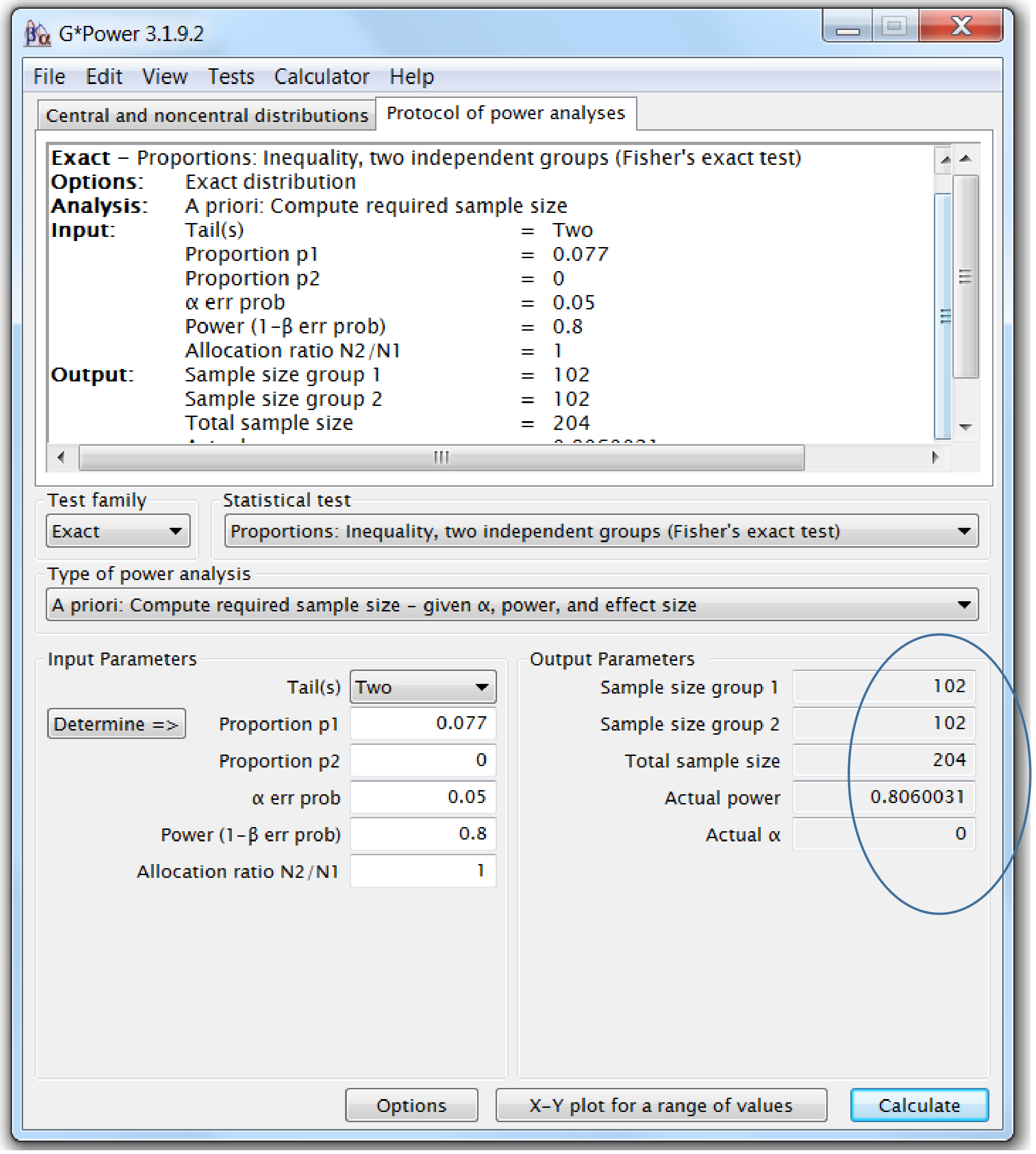
Step 4: Choice of Parameters

Tricky bit: need information on the size of the difference and the variability.



G*Power

- To be able to pick up such a difference, we will need 2 samples of about **102 cows** to reach significance (p<0.05) with 80% power.



Sample Size: Power Analysis

Exercise 2:

- Pilot study: 10 arachnophobes were asked to perform 2 tasks:

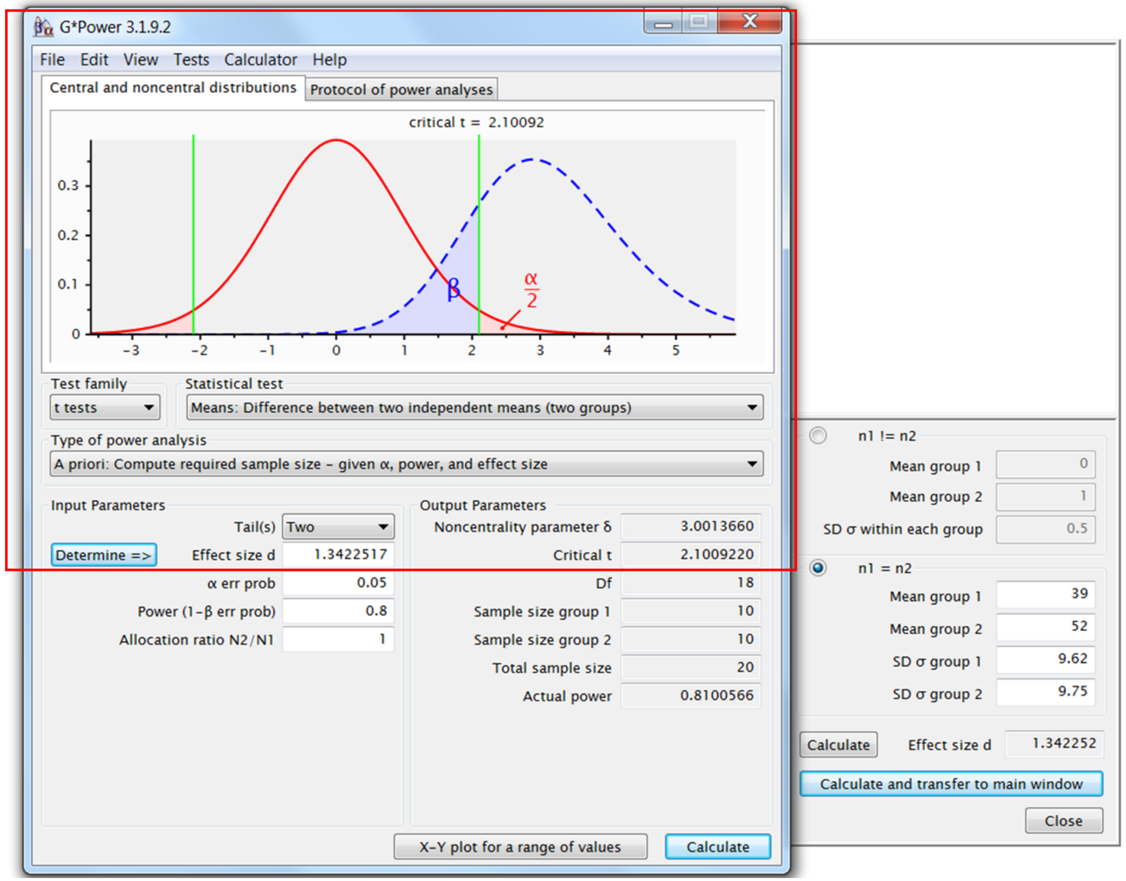
Task 1: Group1 (n=5): to play with a big hairy tarantula spider with big fangs and an evil look in its eight eyes.
Task 2: Group 2 (n=5): to look at pictures of the same hairy tarantula.
Anxiety scores were measured for each group (0 to 100).

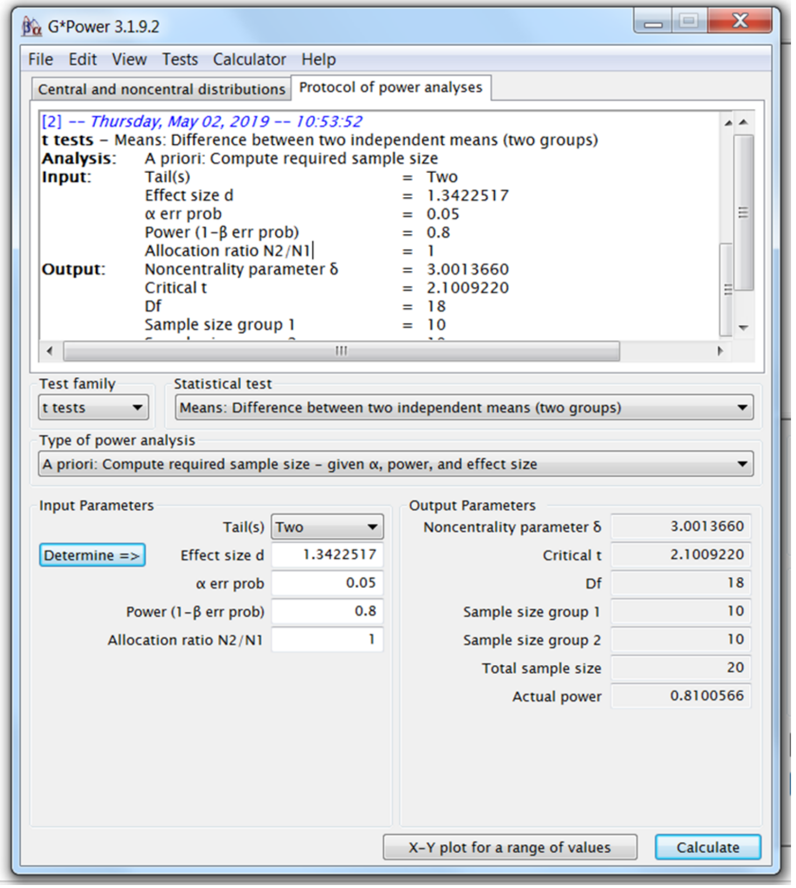
Picture	Real Spider
25	45
35	40
45	55
40	55
50	65

- Use the data to calculate the values for a power calculation

- Run a power calculation (assume balanced design and parametric test)

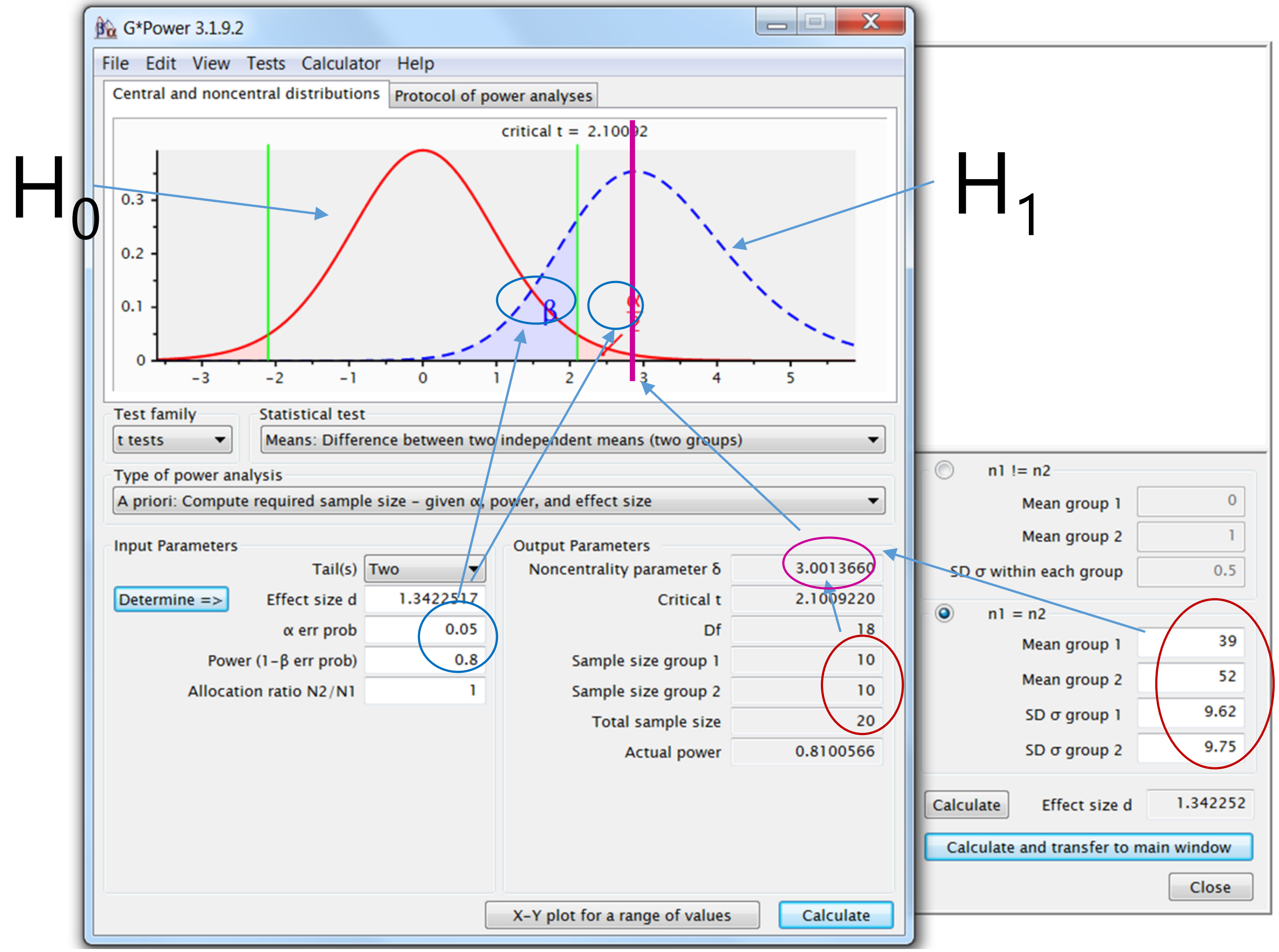
Power Analysis





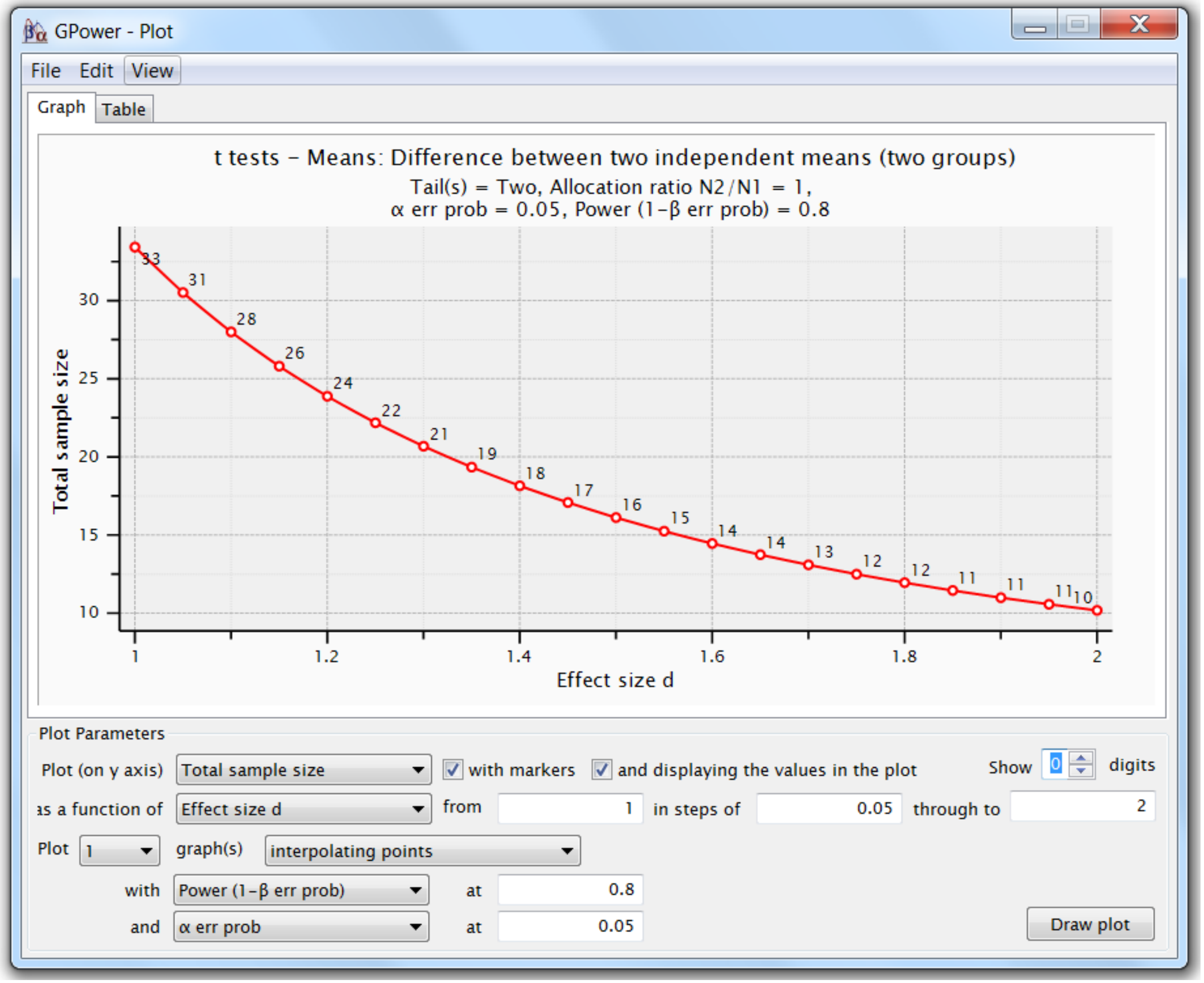
- To reach significance with a t-test, providing the preliminary results are to be trusted, and be confident about the difference between the 2 groups, we need about **20 arachnophobes** (2*10).

Power Analysis



Power Analysis

- For a range of sample sizes:



Unequal sample sizes

- Scientists often deal with unequal sample sizes
  - No simple trade-off:
    - if one needs 2 groups of 30, going for 20 and 40 will be associated with decreased power.
  - **Unbalanced design = bigger total sample**
  - Solution:

    Step 1: power calculation for equal sample size

    Step 2: adjustment

$$N = \frac{2n(1+k)^2}{4k}$$

$$n_1 = \frac{N}{(1+k)}$$

$$n_2 = \frac{kN}{(1+k)}$$

- Cow example: balanced design: **n = 102**

but this time: unpainted group: 2 times bigger than painted one (k=2):

- Using the formula, we get a total:

$N=2*102*(1+2)^2/4*2 = 230$

Painted butts **(n_1)=77** Unpainted butts **(n_2)=153**

- Balanced design: **n = 2*102 = 204**
- Unbalanced design: **n= 77+153 = 230**

Sample Size: Power Analysis

Non-parametric tests

- Non-parametric tests: do not assume data come from a Gaussian distribution.
  - Non-parametric tests are based on ranking values from low to high
  - Non-parametric tests not always less powerful

- Proper power calculation for non-parametric tests:
  - Need to specify which kind of distribution we are dealing with
    - Not always easy

- Non-parametric tests never require more than 15% additional subjects providing that the distribution is not too unusual.

- **Very crude rule of thumb for non-parametric tests**:
  - Compute the sample size required for a parametric test and add 15%.

Sample Size: Power Analysis

- What happens if we ignore the power of a test?
  - Misinterpretation of the results

- p-values: never ever interpreted without context:
  - **Significant p-value (<0.05)**: exciting! Wait: what is the difference?
    - >= smallest meaningful difference: exciting
    - < smallest meaningful difference: not exciting
      - very big sample, too much power

  - **Not significant p-value (>0.05)**: no effect! Wait: how big was the sample?
    - Big enough = enough power: no effect means no effect
    - Not big enough = not enough power
      - Possible meaningful difference but we miss it

Day 2

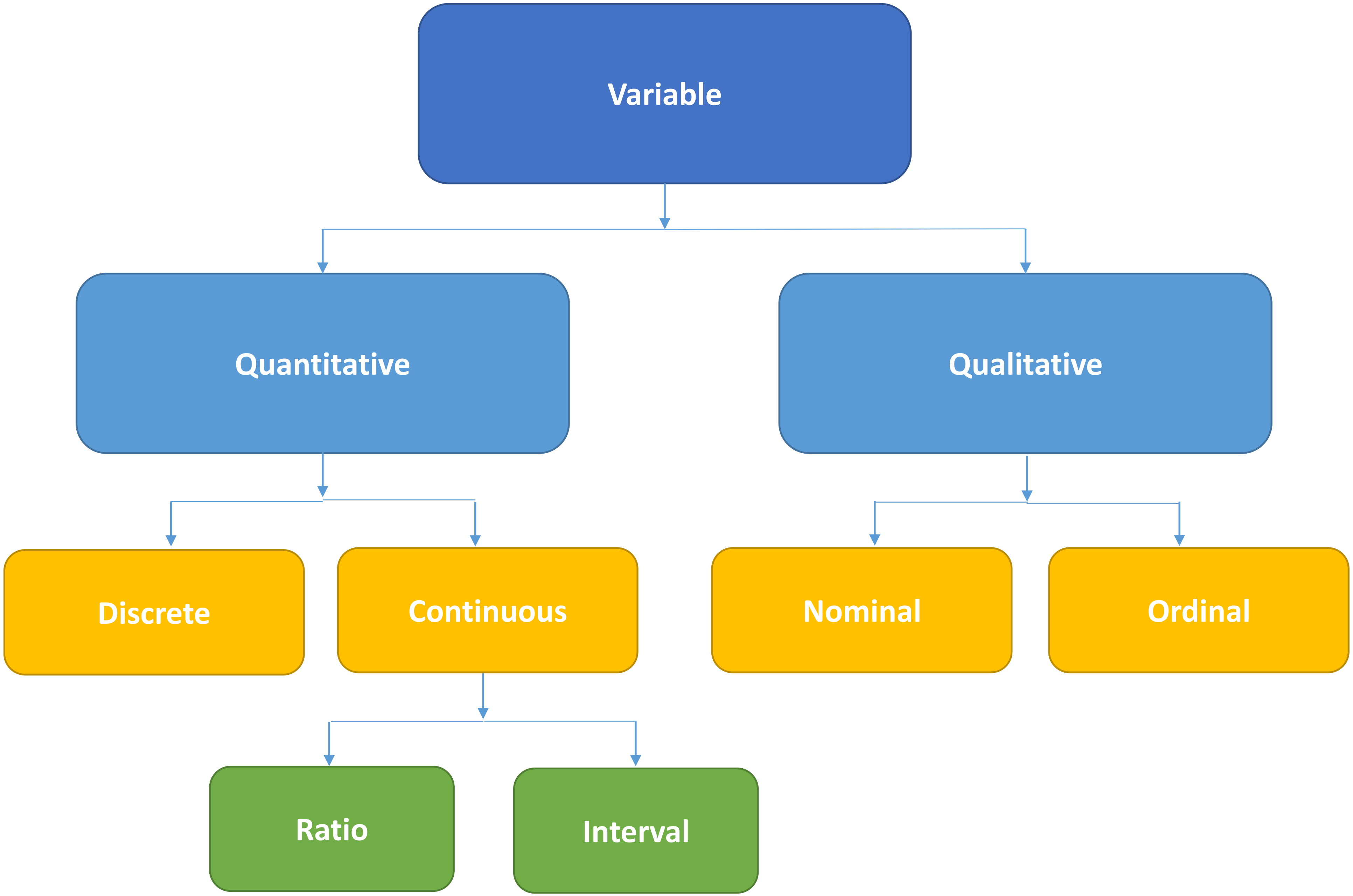
escriptive statistics and data exploration

Anne Segonds-Pichon

v2019-06

Babraham Institute

Babraham Bioinformatics


Variable
Quantitative
Qualitative
Discrete
Continuous
Nominal
Ordinal
Ratio
Interval

Quantitative data

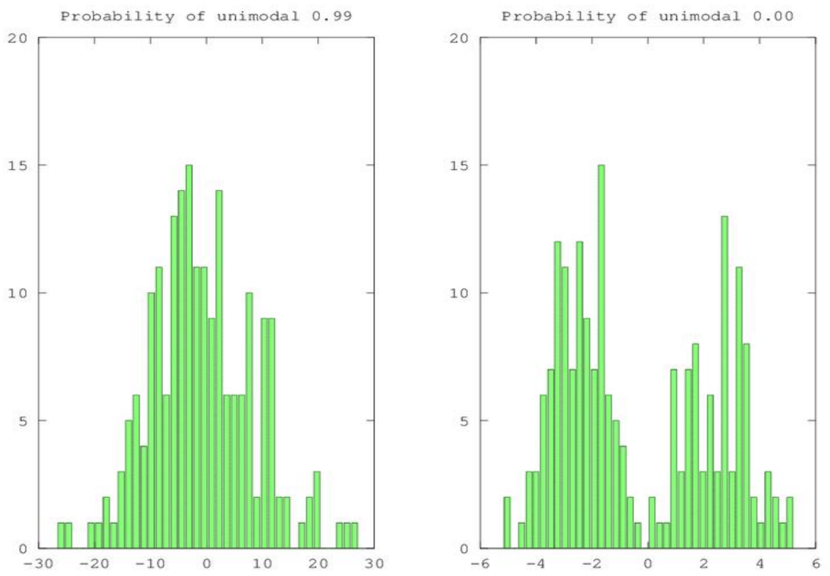
- They take numerical values (units of measurement)

- Discrete: obtained by counting
  - Example: number of students in a class
  - values vary by finite specific steps
- or continuous: obtained by measuring
  - Example: height of students in a class
  - any values

- They can be described by a series of parameters:
  - Mean, variance, standard deviation, standard error and confidence interval

Measures of central tendency
Mode and Median

- **Mode:** most commonly occurring value in a distribution



- **Median**: value exactly in the middle of an ordered set of numbers

Example 1: 18 27 34 52 54 59 61 68 78 82 85 87 91 93 100, Median = 68
Example 2: 18 27 27 34 52 52 59 61 68 68 85 85 85 90, Median = 60

Measures of central tendency
Mean

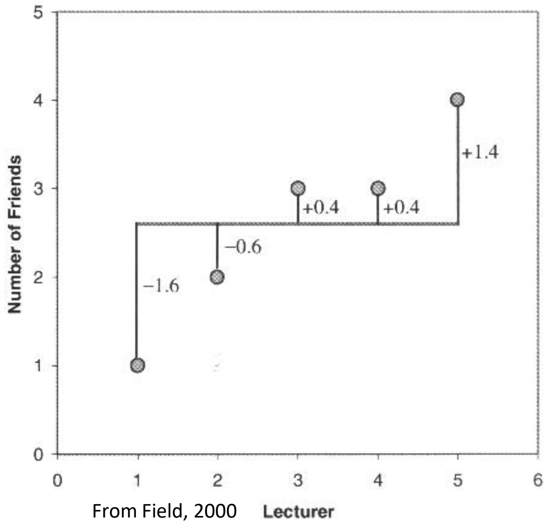
- Definition: average of all values in a column

- It can be considered as a model because it summaries the data
  - Example: a group of 5 lecturers: number of friends of each members of the group: 1, 2, 3, 3 and 4
    - Mean: (1+2+3+3+4)/5 = 2.6 friends per person
      - Clearly an hypothetical value

- How can we know that it is an accurate model?
  - Difference between the real data and the model created

Measures of dispersion

- Calculate the magnitude of the differences between each data and the mean:



From Field, 2000

- Total error = sum of differences

$$= 0 = \Sigma(x_i - \overline{x}) = (-1.6)+(-0.6)+(0.4)+(1.4) = 0$$

No errors !

- Positive and negative: they cancel each other out.

Sum of Squared errors (SS)

- To avoid the problem of the direction of the errors: we square them
  - Instead of sum of errors: sum of squared errors (SS):

$$(SS) = \Sigma(x_i - \overline{x})(x_i - \overline{x})$$

$$= (1.6)^2 + (-0.6)^2 + (0.4)^2 + (0.4)^2 + (1.4)^2$$

$$= 2.56 + 0.36 + 0.16 + 0.16 + 1.96$$

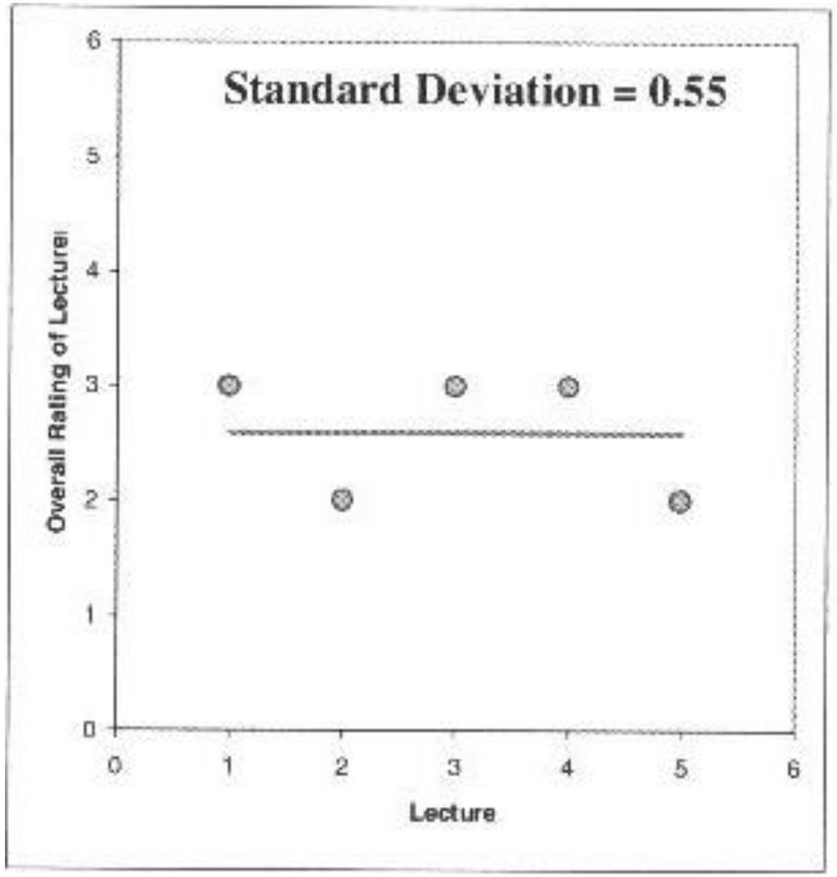
$$= 5.20$$

- SS gives a good measure of the accuracy of the model
  - But: dependent upon the amount of data: the more data, the higher the SS.
  - <u>Solution</u>: to divide the SS by the number of observations (N)
    - As we are interested in measuring the error in the sample to estimate the one in the population we divide the SS by N-1 instead of N and we get the **variance** (S^2) = SS/N-1

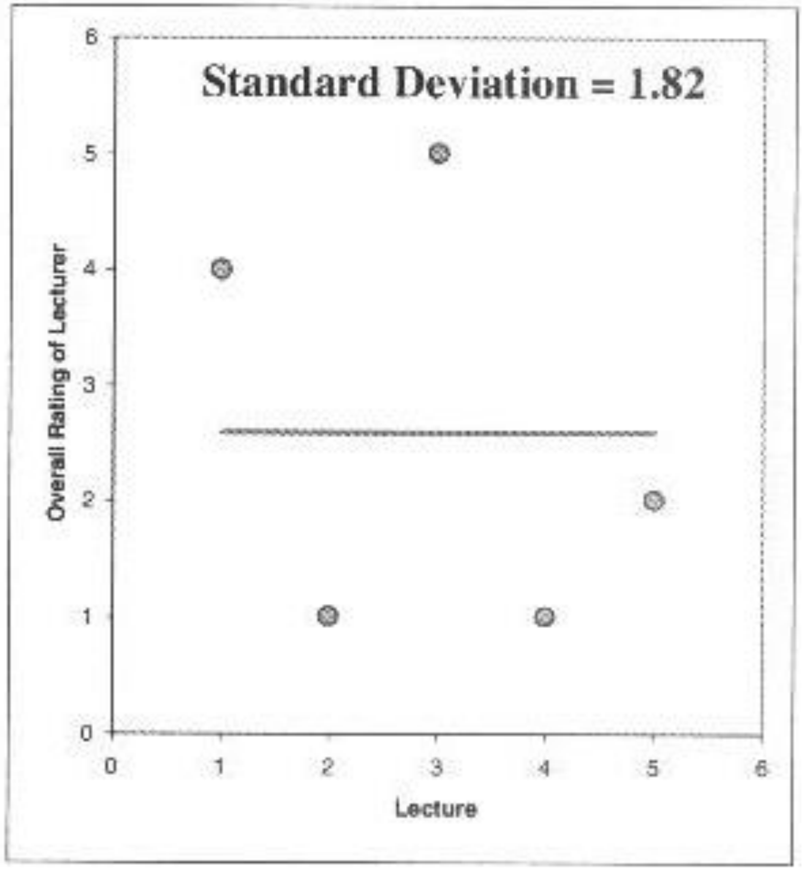
Variance and standard deviation

- $variance\ (s^2) = \frac{SS}{N-1} = \frac{\Sigma\,(x_i - \overline{x})^2}{N-1} = \frac{5.20}{4} = 1.3$
- Problem with variance: measure in squared units
  - For more convenience, the square root of the variance is taken to obtain a measure in the same unit as the original measure:
    - the **standard deviation**
      - S.D. = √(SS/N-1) = √(s^2) = s = $\sqrt{1.3} = 1.14$
    - The standard deviation is a measure of how well the mean represents the data.

Standard deviation



Small S.D.:
data close to the mean:
mean is a good fit of the data

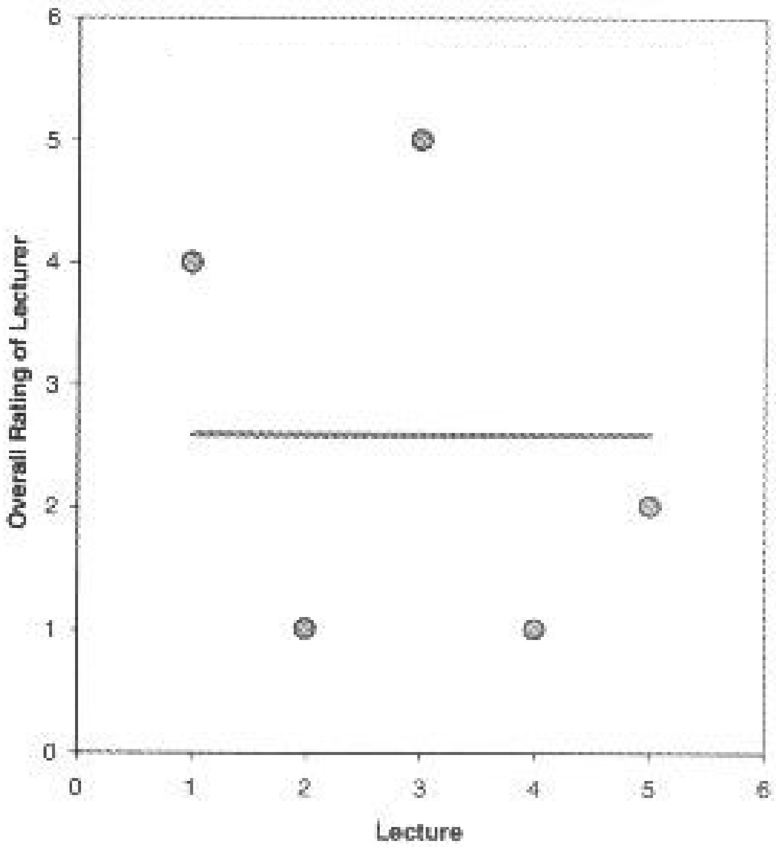


Large S.D.:
data distant from the mean:
mean is not an accurate representation

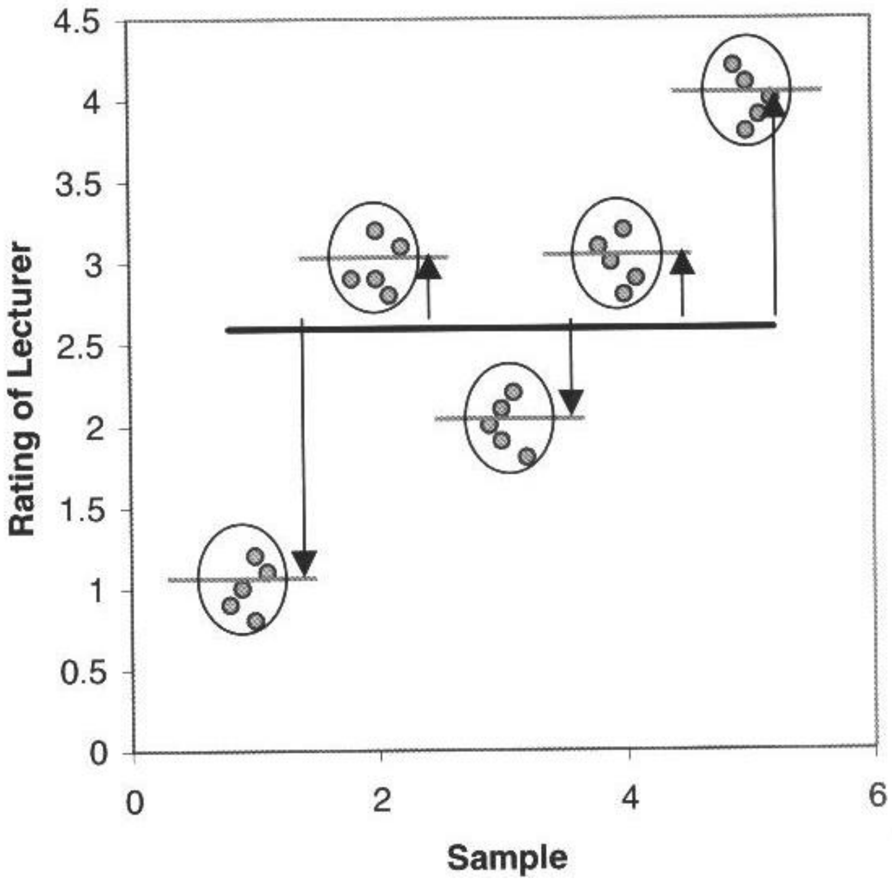
SD and SEM (SEM = SD/√N)

- What are they about?
  - The **SD** quantifies **how much the values vary** from one another: **scatter or spread**
    - The SD does not change predictably as you acquire more data.
  - The **SEM** quantifies **how accurately** you know the **true mean** of the population.
    - Why? Because it takes into account: **SD + sample size**
    - The SEM gets smaller as your sample gets larger
      - Why? Because the mean of a large sample is likely to be closer to the true mean than is the mean of a small sample.

SD and SEM



The SD quantifies the scatter of the data.



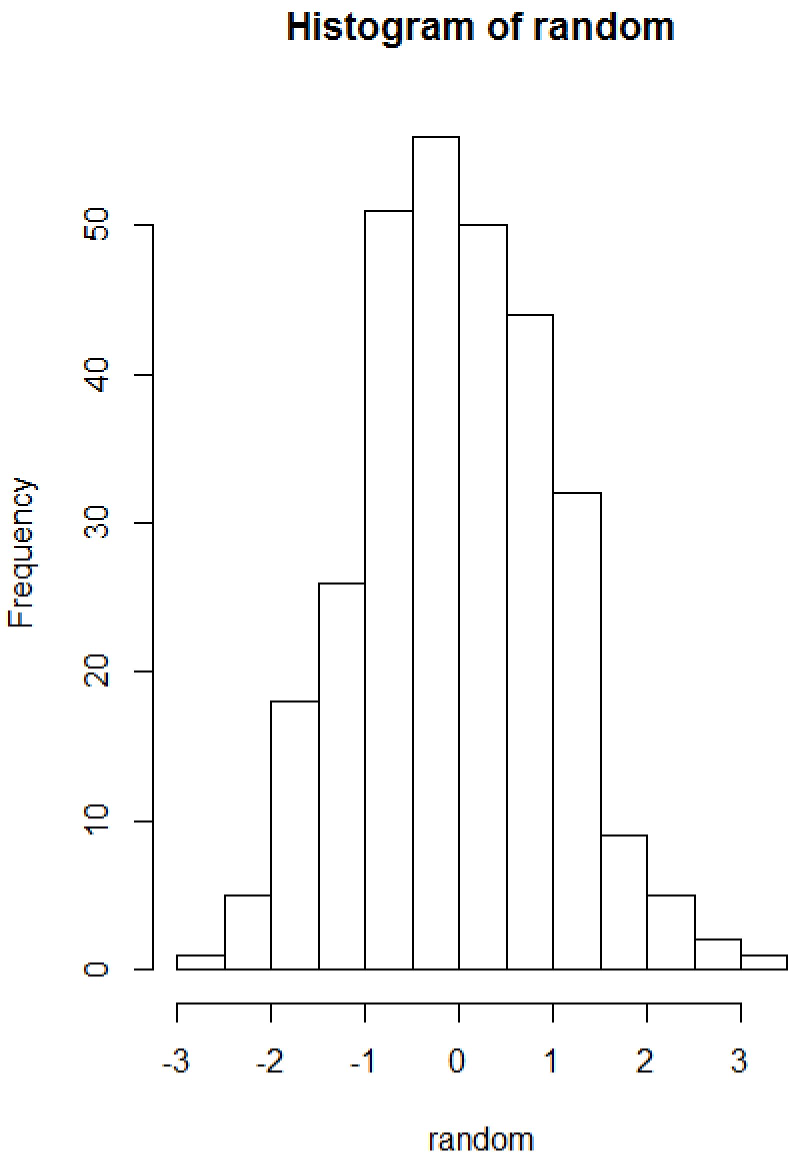
The SEM quantifies the distribution of the sample means.

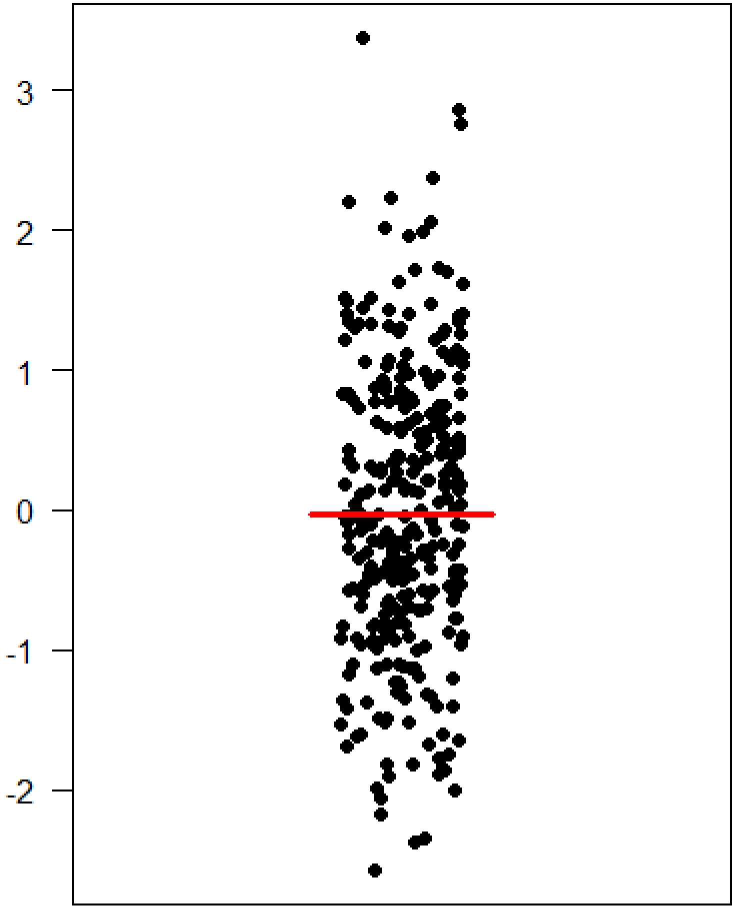
SD or SEM ?

- If the scatter is caused by biological variability, it is important to show the variation.
  - Report the SD rather than the SEM.
    - Better even: show a graph of all data points.

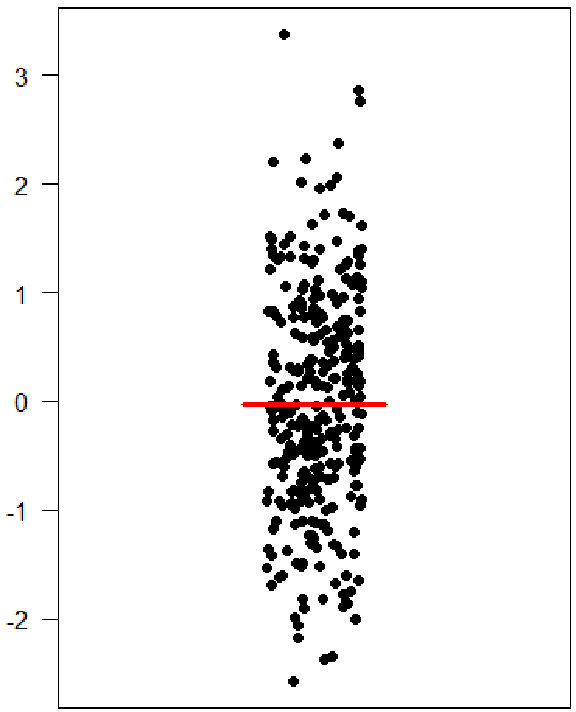
- If you are using an in vitro system with no biological variability, the scatter is about experimental imprecision (no biological meaning).
  - Report the SEM to show how well you have determined the mean.

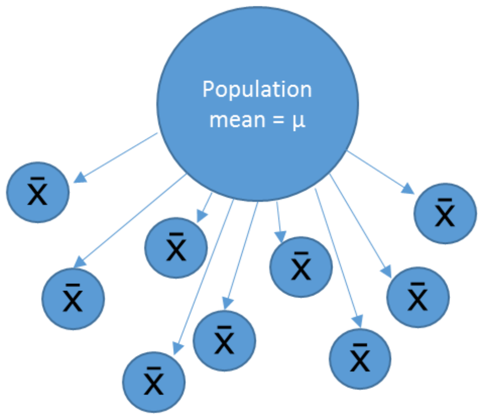
The SEM and the sample size





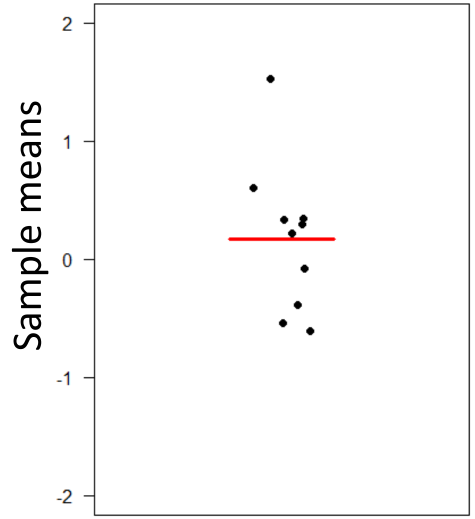
The SEM and the sample size



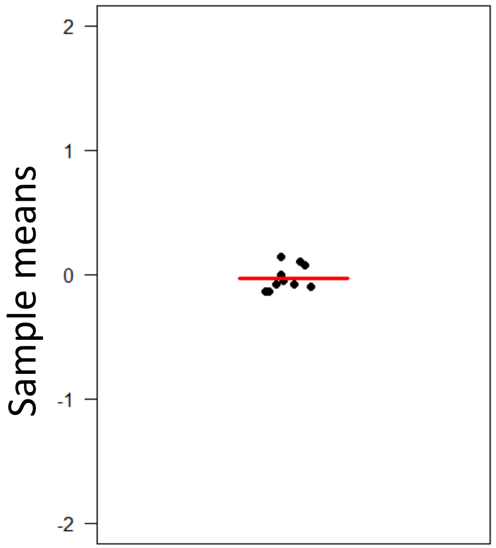


'Infinite' number of samples
Samples means = $\bar{x}$

Small samples (n=3)

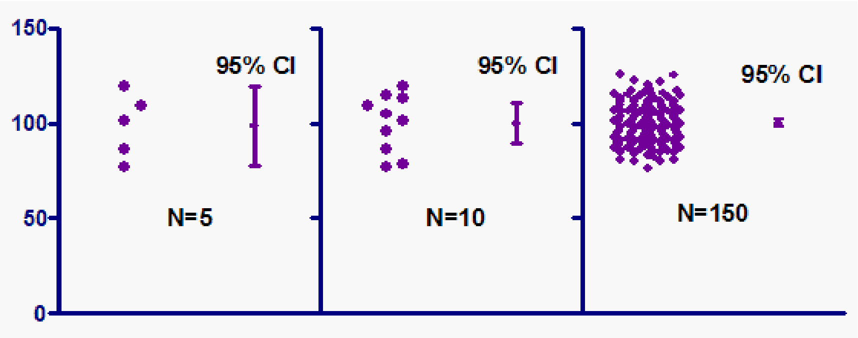


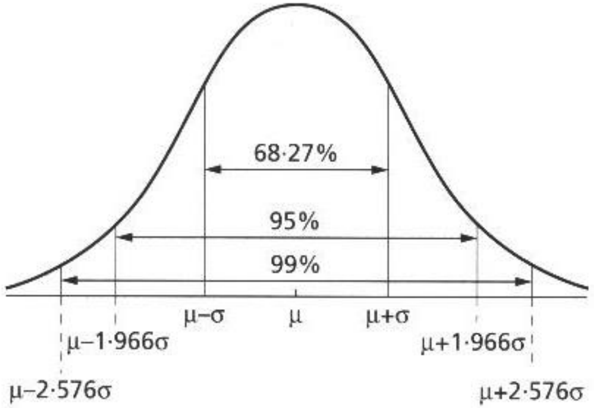
Big samples (n=30)



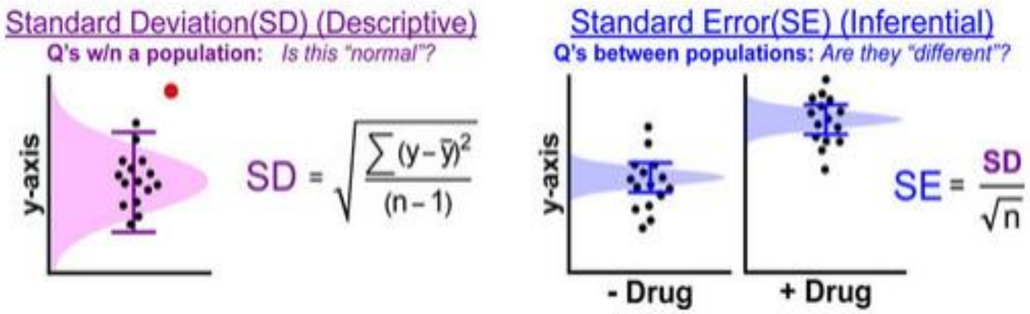
Confidence interval

- Range of values that we can be 95% confident contains the true mean of the population.
  - So limits of 95% CI: **[Mean - 1.96 SEM; Mean + 1.96 SEM]** (SEM = SD/√N)





Error bars	Type	Description
Standard deviation	Descriptive	Typical or average difference between the data points and their mean.
Standard error	Inferential	A measure of how variable the mean will be, if you repeat the whole study many times.
Confidence interval usually 95% CI	Inferential	A range of values you can be 95% confident contains the true mean.



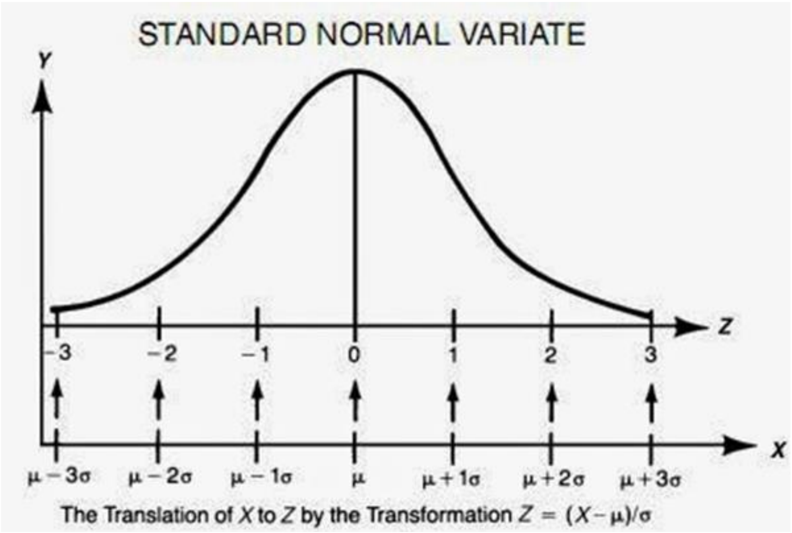
Z-score

- Standardisation of normal data with mean μ and standard deviation σ

$$Z = \frac{x - \mu}{\sigma}$$

- Example: μ=50 and σ=1.
  - A variable with value x=60 has a z-score=1

$$z = \frac{X - \mu}{\sigma} = \frac{60 - 50}{10} = 1$$



Z-score

$$Z = \frac{x - \mu}{\sigma}$$

- Probability that a given value is found in a normally distributed sample with known μ and σ.
- Beyond a **threshold**, values 'do not belong' or are very unlikely to be found in such a sample.
  - Threshold = 1.96
  - Normal distribution: 95% of observations lie within μ ± 1.96σ (Z=1.96)
  - Probability to find values beyond ± 1.96σ is =<5% ($p<0.05$)

The value 0.975 = a *z*-value of 1.96

4. Statistical Tables

Areas (probabilities) under a normal distribution

P(-1.96 < z > +1.96) is (2 × 0.025) = 0.05

The **left column** gives the **first decimal place** and the top row gives the second decimal place. So the area (probability) corresponding to z_1 = 0.23, for example, is in the row labelled 0.2 and the column headed .03, value = 0.5910).

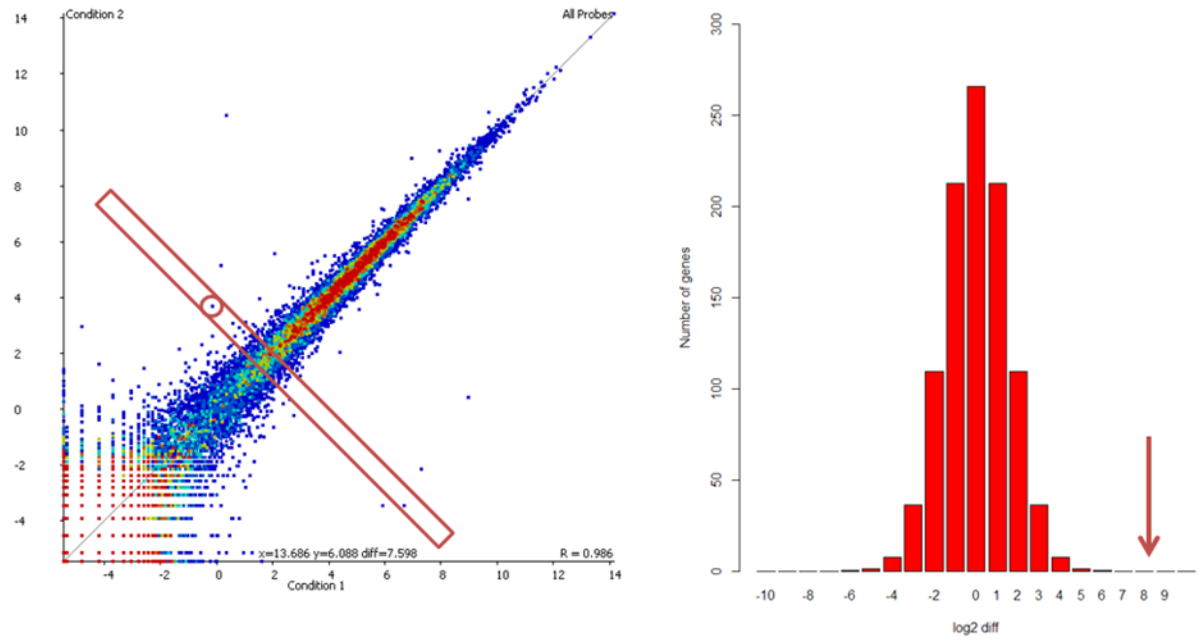
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	**.9750**	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

Z-score application

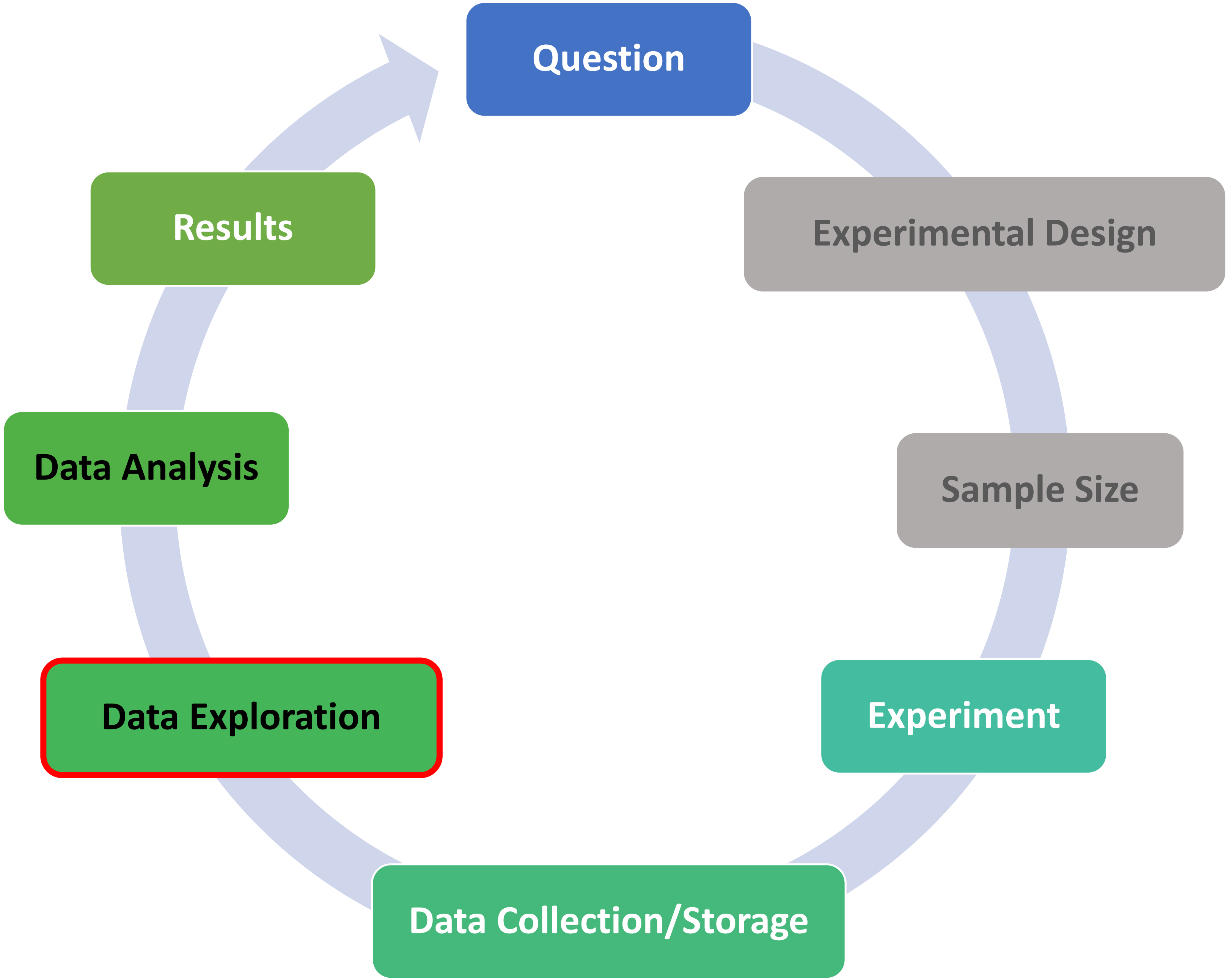
RNA-seq analysis

- Differential gene expression: Noise
  - Length of gene and level of expression

- Lowly expressed genes = highest fold changes
  - Often biologically meaningless

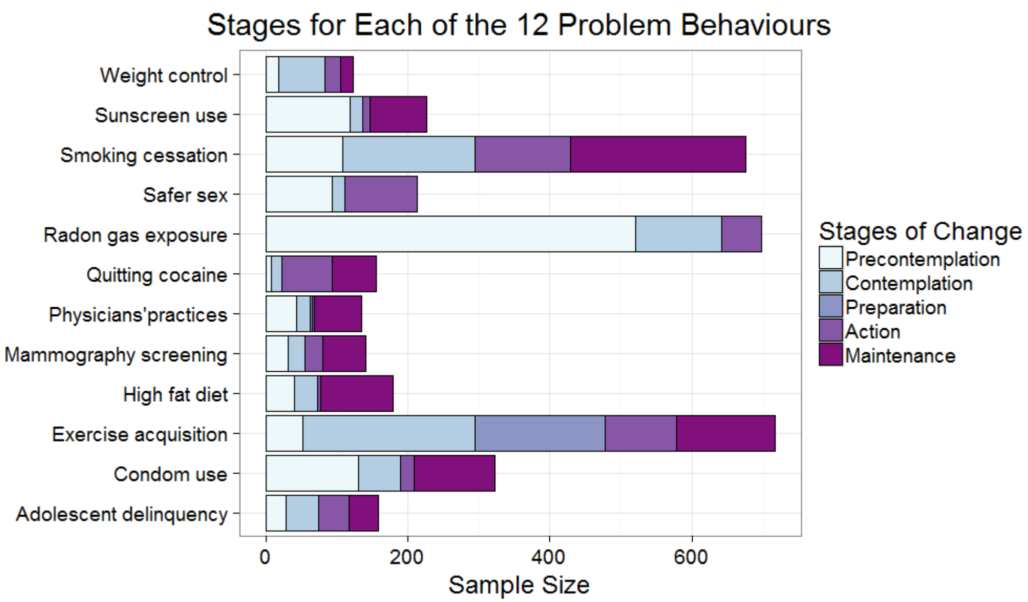


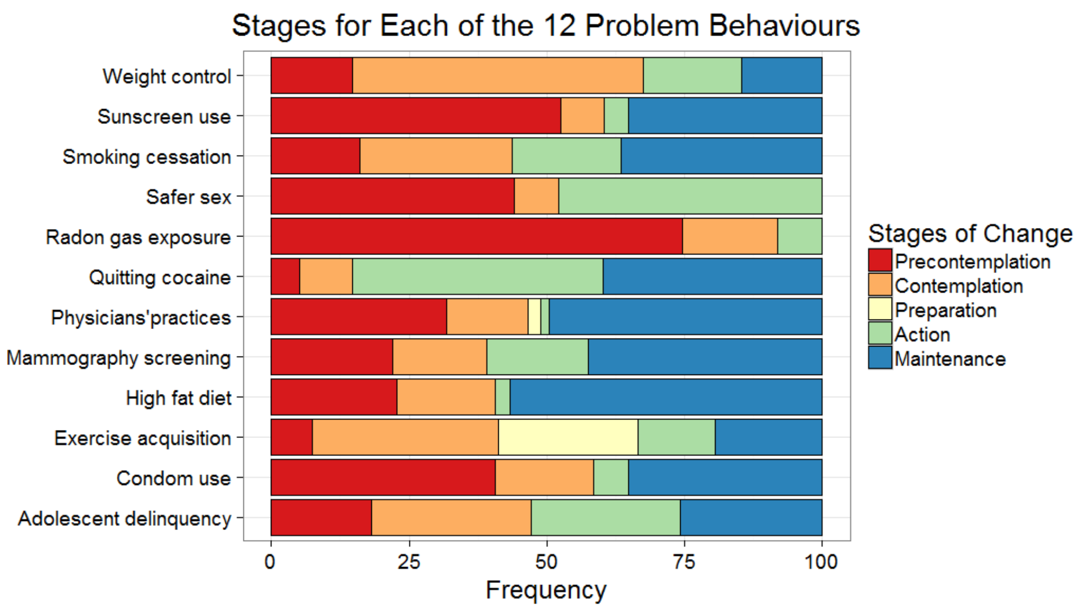
Graphical exploration of data

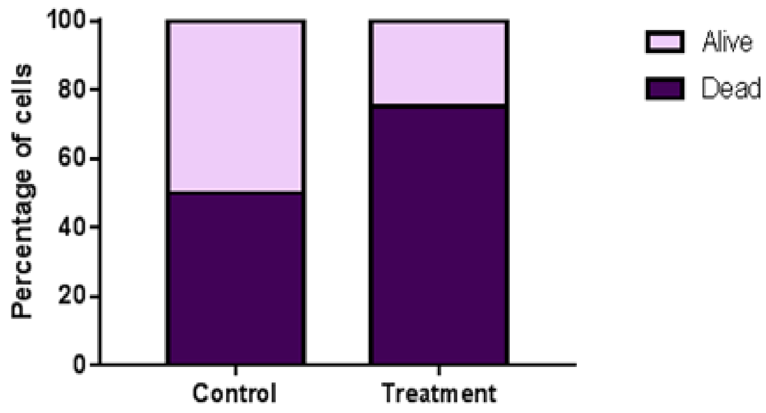

Question
Experimental Design
Sample Size
Experiment
Data Collection/Storage
Data Exploration
Data Analysis
Results

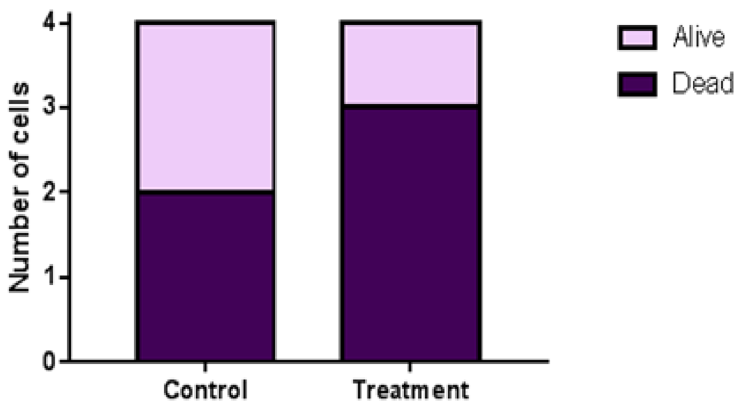
Data Exploration

Categorical data



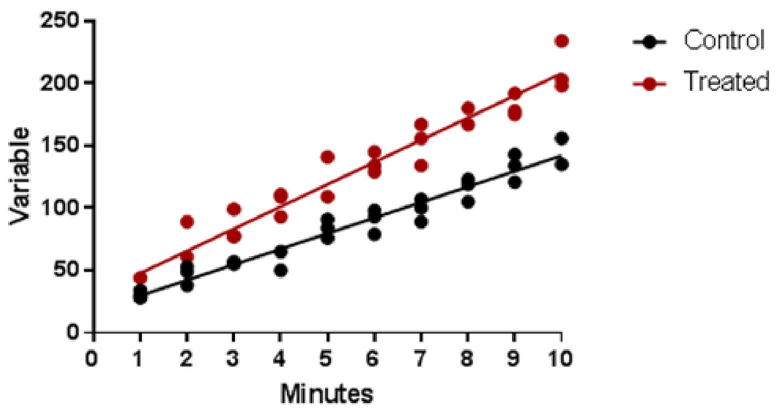


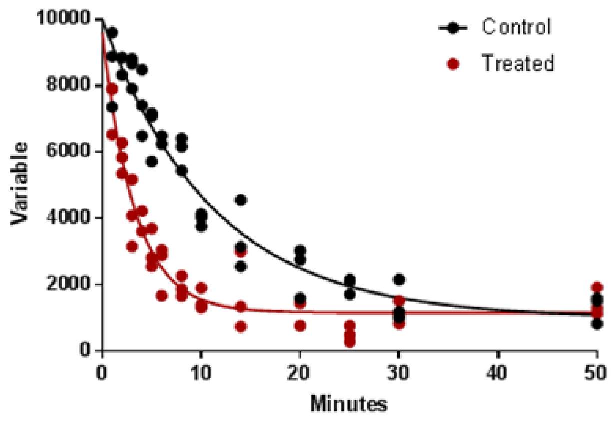


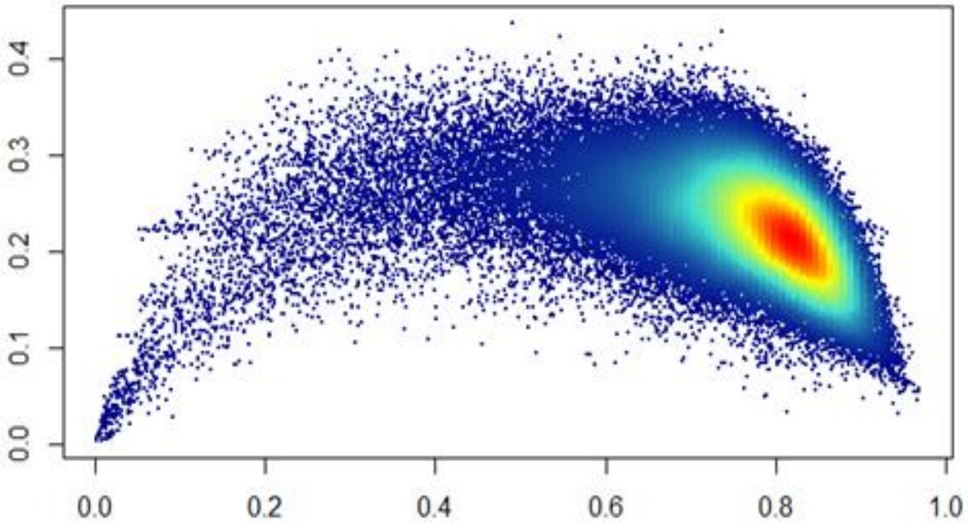


Data Exploration

Quantitative data: Scatterplot



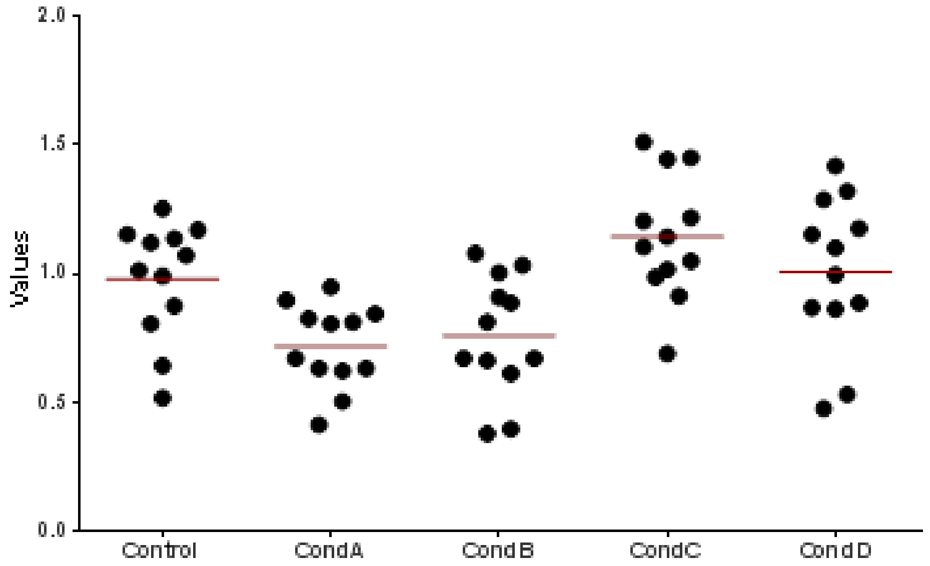




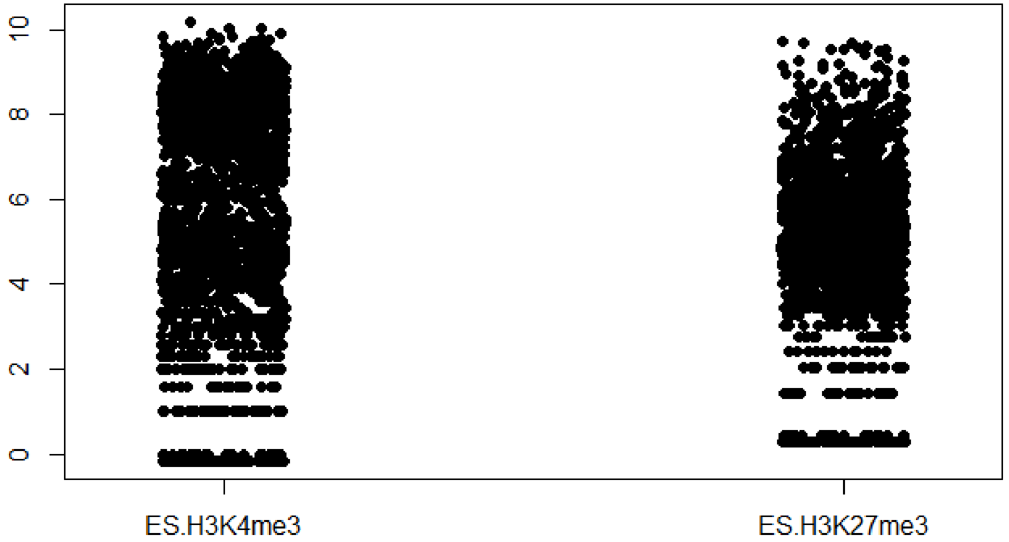
Data Exploration

Quantitative data:

Scatterplot/stripchart



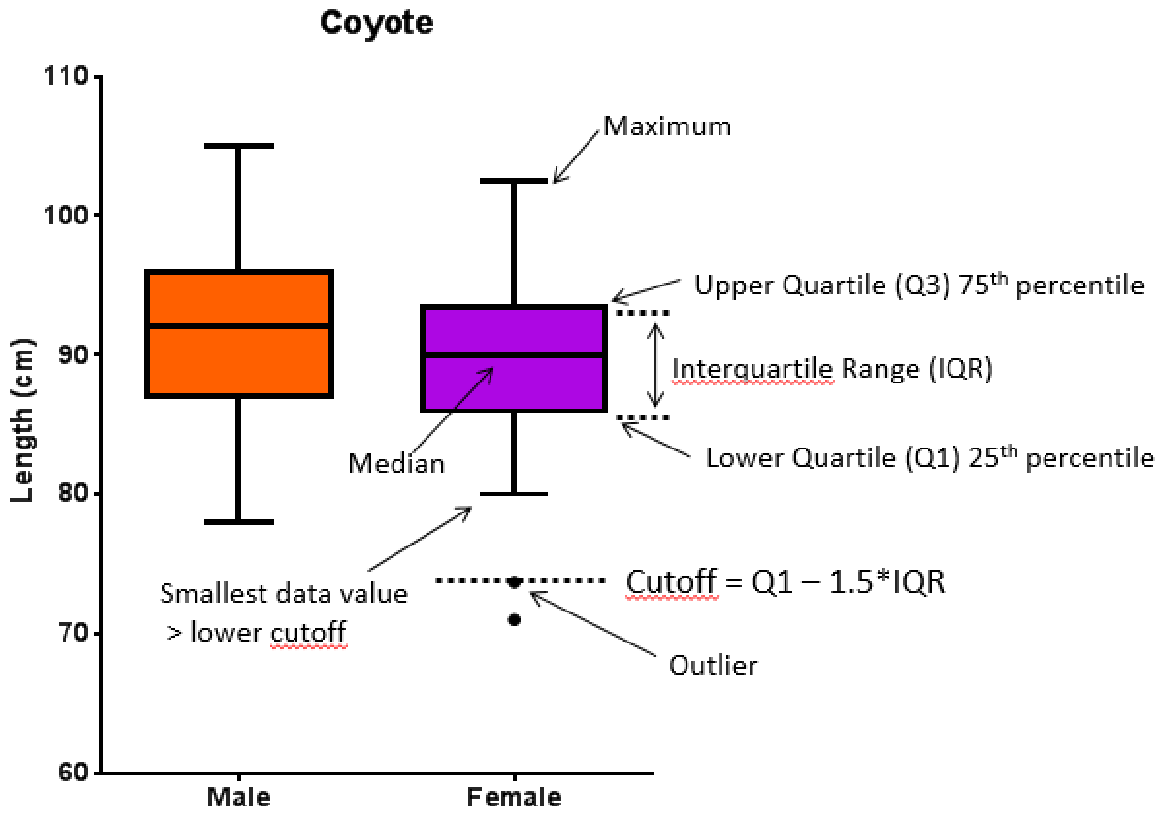
Small sample

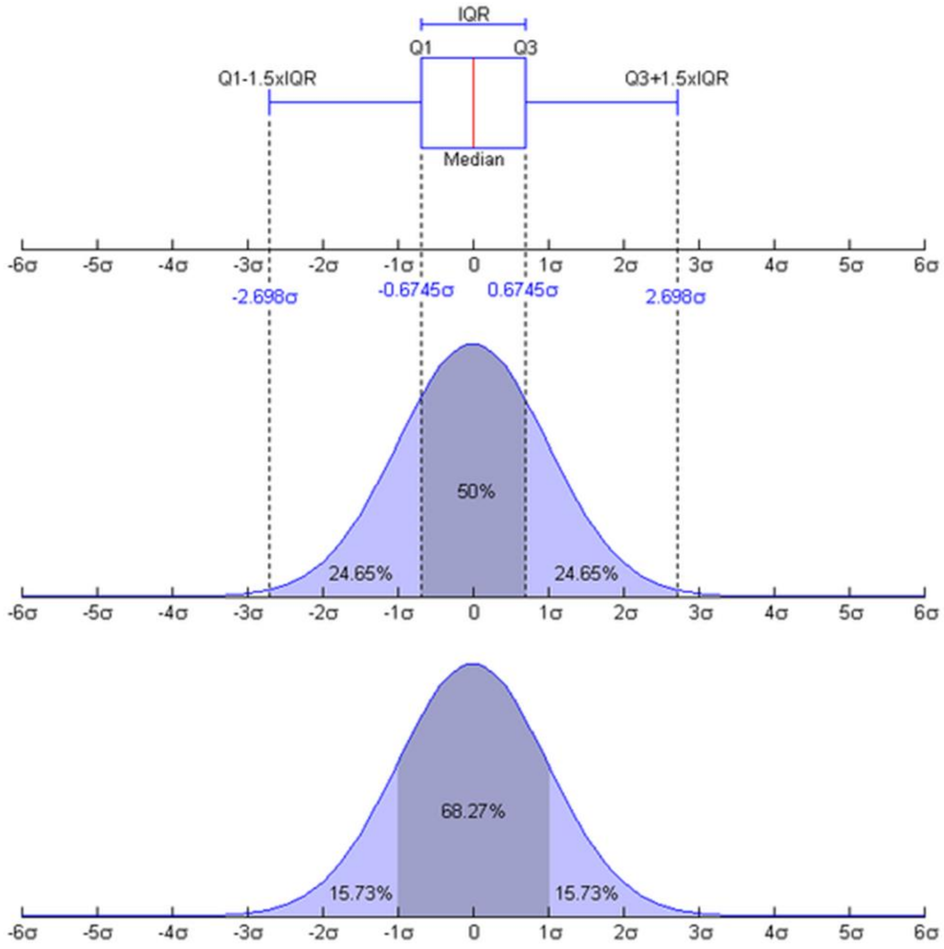


Big sample

Data Exploration

Quantitative data: Boxplot

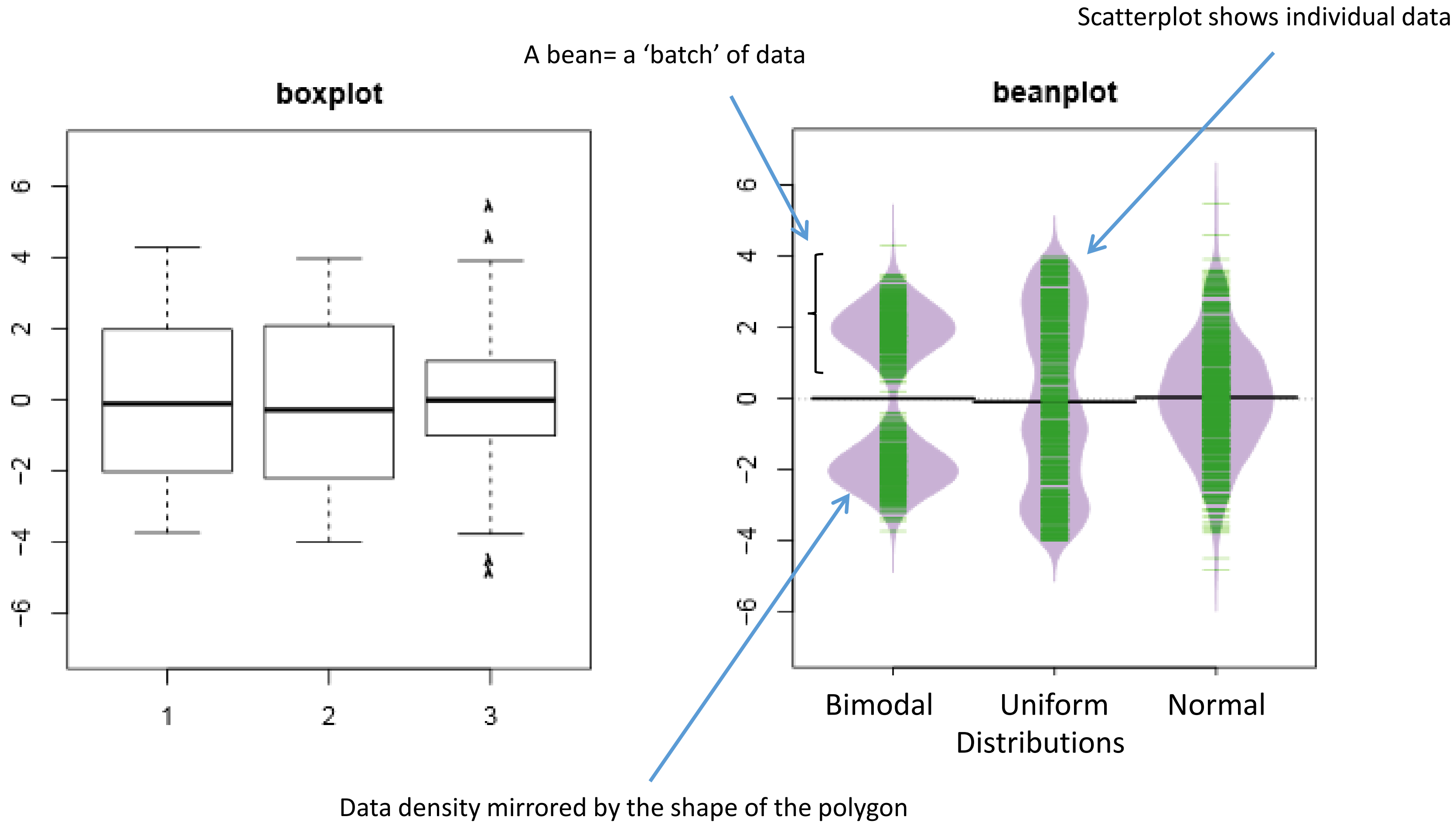




Data Exploration

Quantitative data:

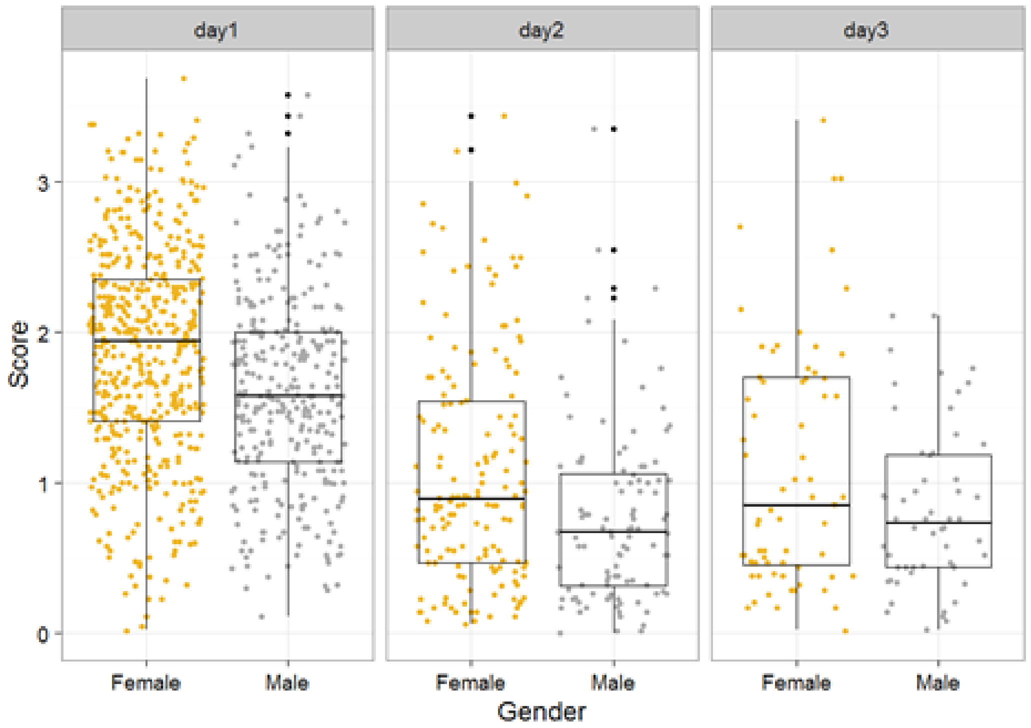
Boxplot or Beanplot

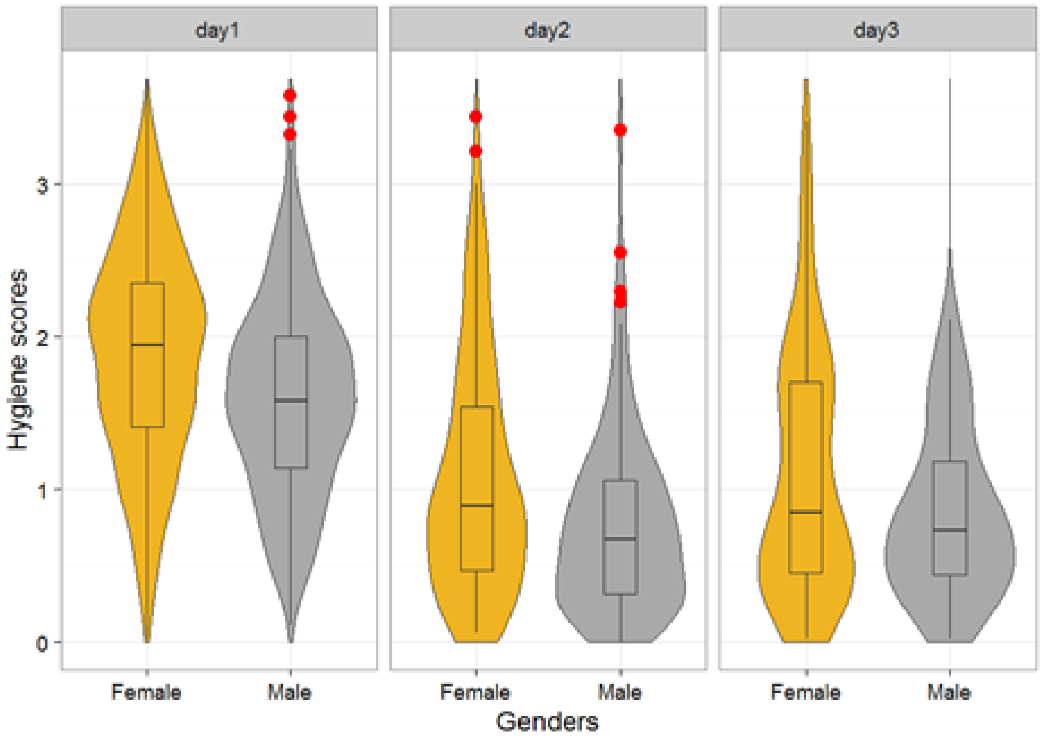


Data Exploration

Quantitative data:

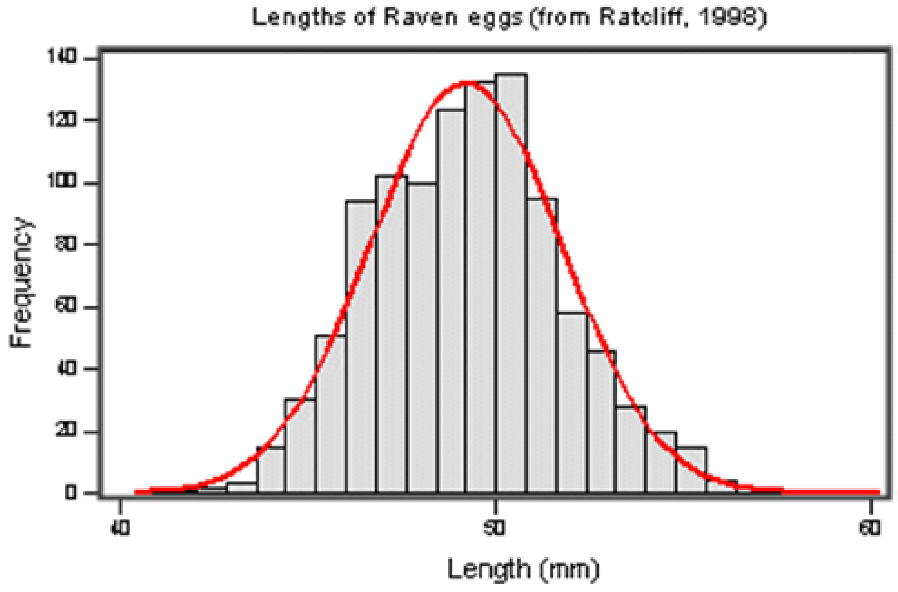
Boxplot and Beanplot and Scatterplot



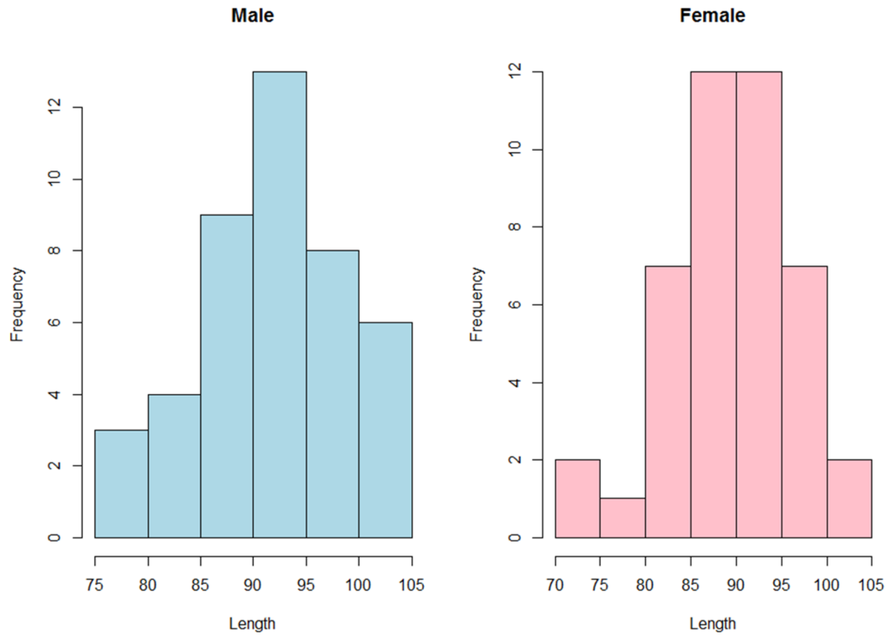


Data Exploration

Quantitative data: Histogram



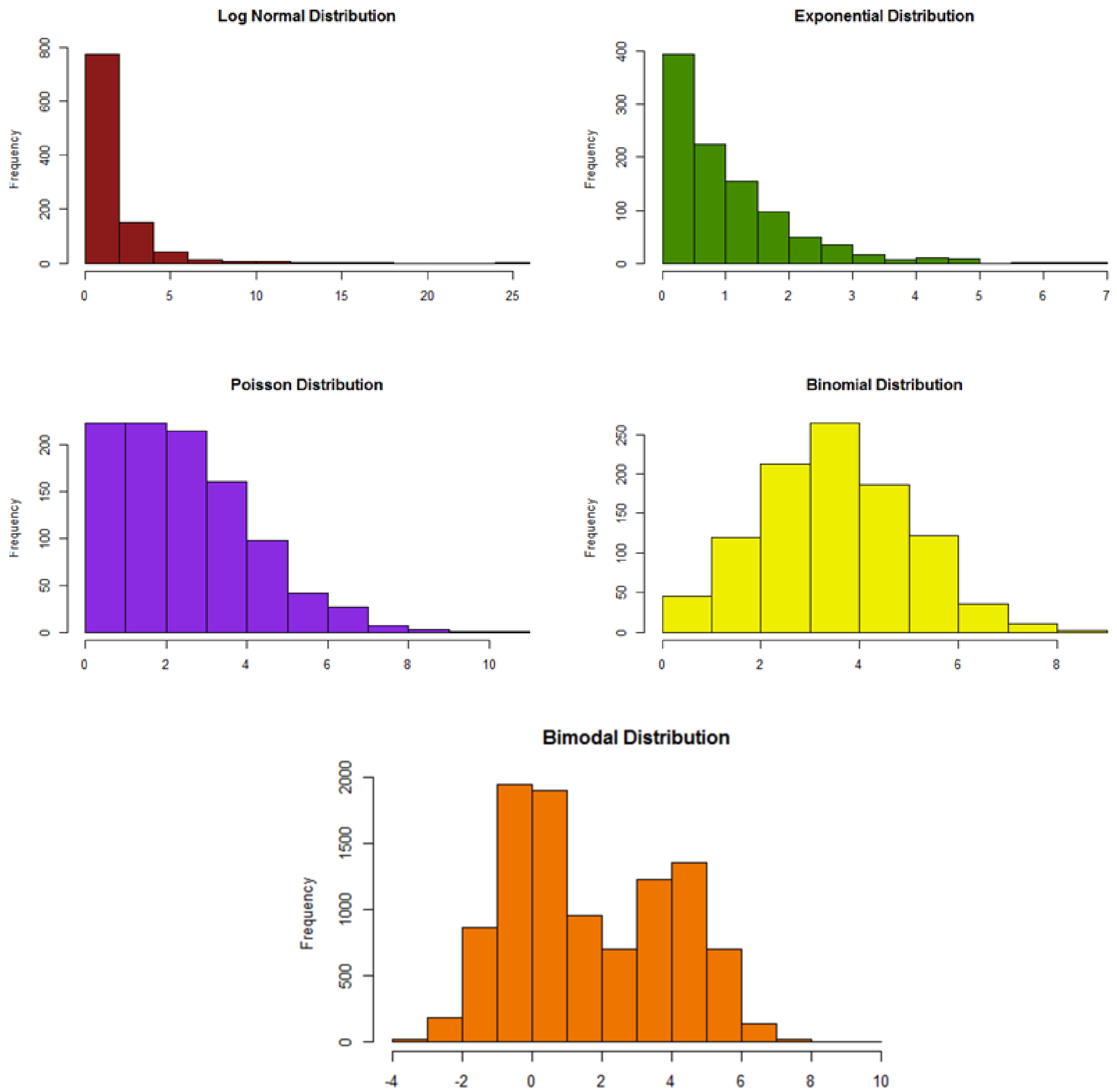
Big sample



Small sample

Data Exploration

Quantitative data: Histogram (distribution)

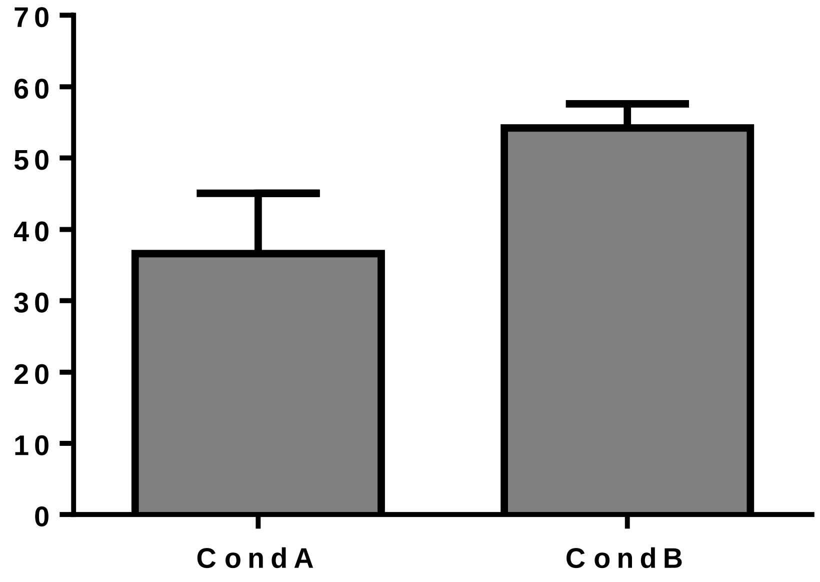


Data Exploration

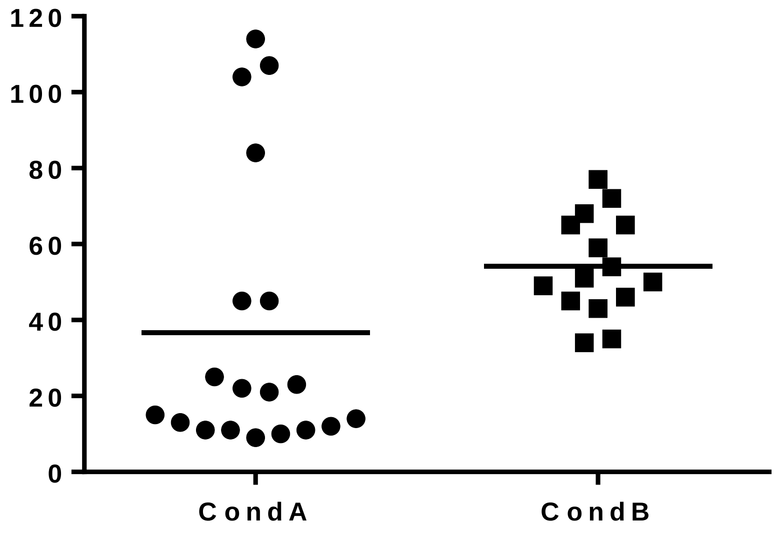
Plotting is not the same thing as exploring

- <u>One experiment</u>: change in the variable of interest between CondA to CondB.
  - ❖ Data plotted as a **bar chart**.

The fiction



The truth

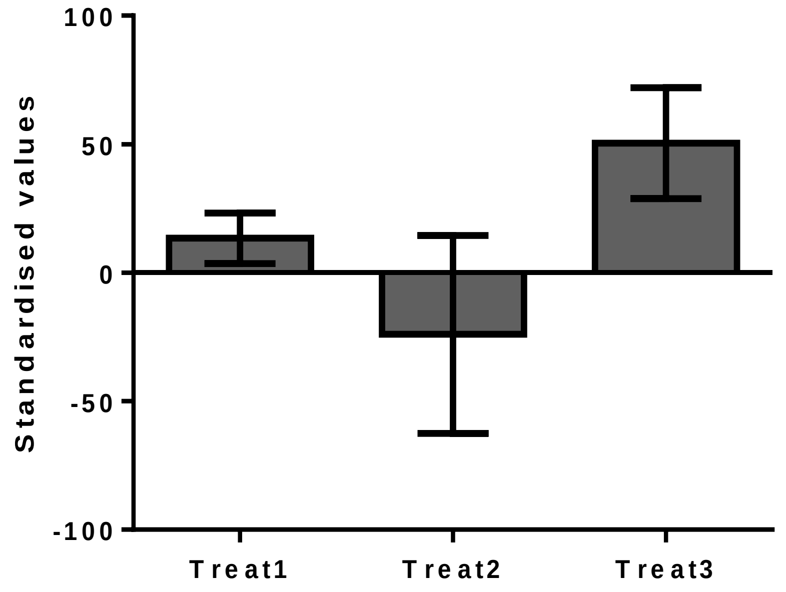


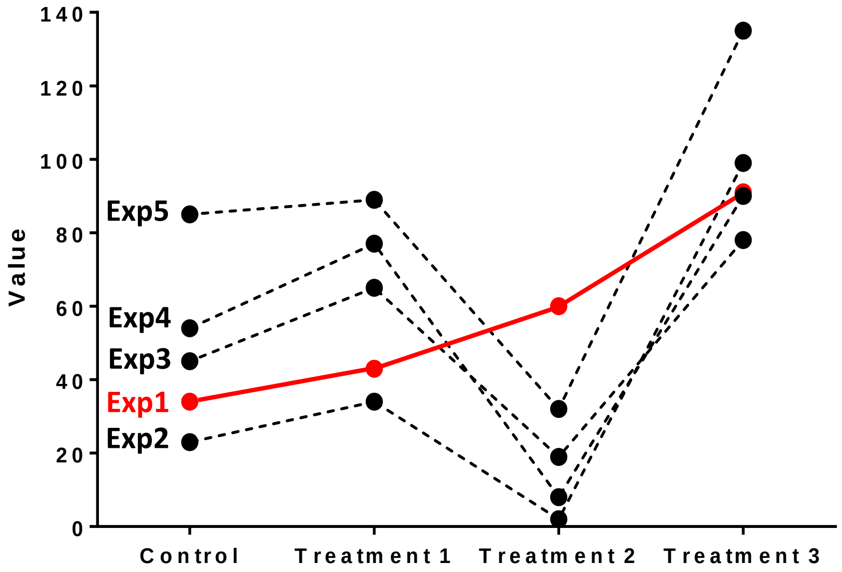
Data Exploration

Plotting (and summarising) is (so) not the same thing as exploring

- Five experiments: change in the variable of interest between 3 treatments and a control.
  - ❖ Data plotted as a **bar chart**.

The truth (if you are into bar charts)



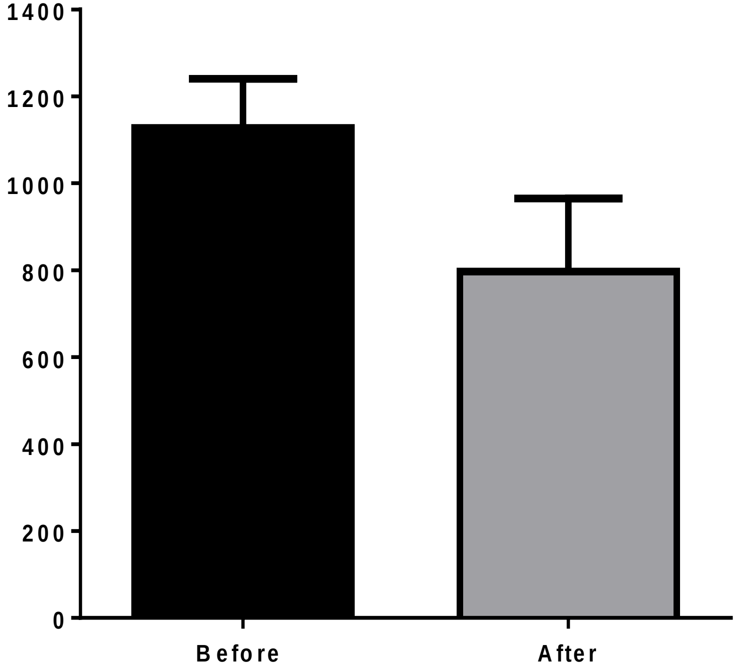


Data Exploration

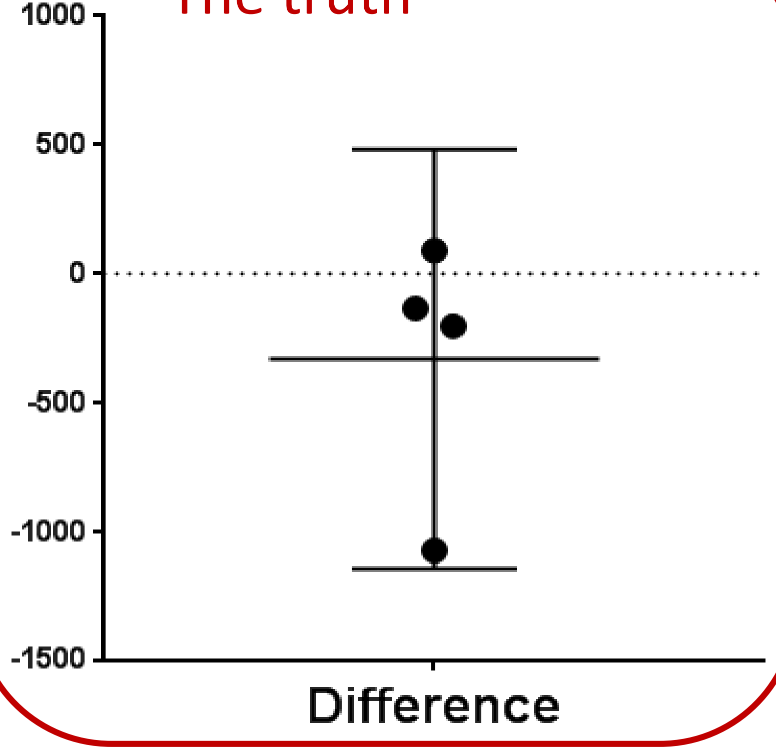
Plotting (and summarising and choosing the wrong graph) is (definitely) not the same thing as exploring

- Four experiments: Before-After treatment effect on a variable of interest.
- Hypothesis: Applying a treatment will decrease the levels of the variable of interest.
  - Data plotted as a **bar chart**.

The fiction



The truth



Days 2 and 3
Analysis of Quantitative data

Anne Segonds-Pichon

v2019-06

Babraham Institute

Babraham Bioinformatics

Outline of this section

- Assumptions for parametric data
- Comparing two means: **Student's *t*-test**
- Comparing more than 2 means
  - One factor: **One-way ANOVA**
  - Two factors: **Two-way ANOVA**
- Relationship between 2 continuous variables:
  - Linear: **Correlation**
  - Non-linear: **Curve fitting**
- **Non-parametric tests**

Introduction

- **Key concepts to always keep in mind**
  - Null hypothesis and error types
  - Statistics inference
  - Signal-to-noise ratio

The null hypothesis and the error types

- The null hypothesis (H_0): H_0 = no effect
  - e.g. no difference between 2 genotypes
- The aim of a statistical test is to reject or not $H_{0.}$

Statistical decision	True state of H_0	
	H_0 True (no effect)	**H_0 False (effect)**
Reject H_0	**Type I error α** **False Positive**	Correct **True Positive**
Do not reject H_0	Correct **True Negative**	**Type II error β** **False Negative**

- Traditionally, a test or a difference is said to be "**significant**" if the probability of type I error is: **α =< 0.05**
- **High specificity** = low **False Positives** = low **Type I error**
- **High sensitivity** = low **False Negatives** = low **Type II error**

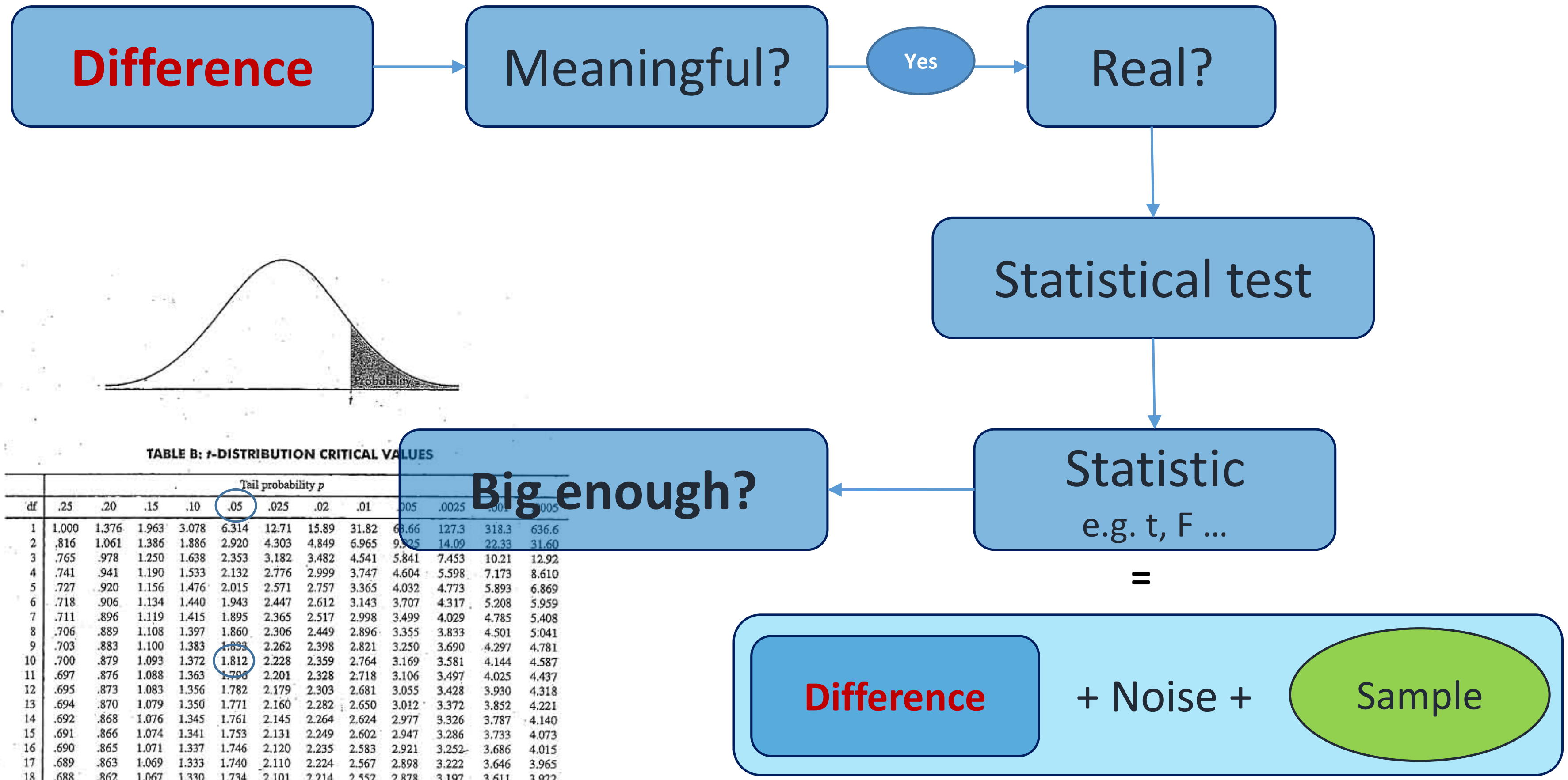
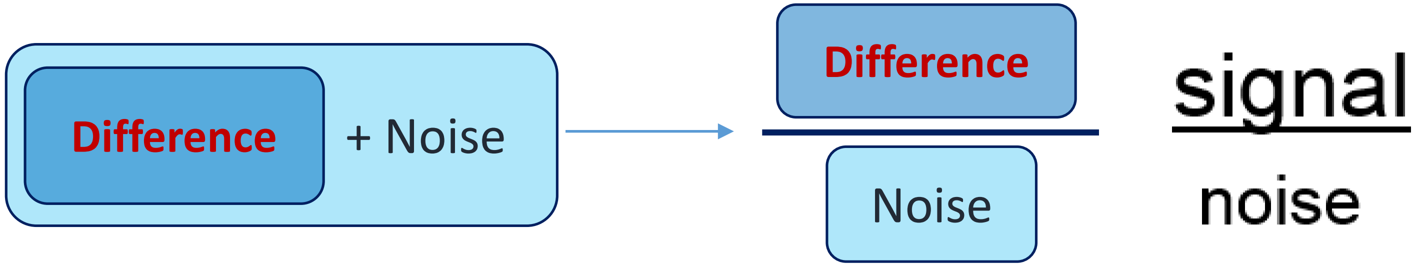


TABLE B: t-DISTRIBUTION CRITICAL VALUES

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792

Signal-to-noise ratio

- Stats are all about understanding and controlling variation.



signal / noise — If the **noise is low** then the **signal is detectable** …
= statistical significance

signal / noise — … but if the **noise** (i.e. interindividual variation) **is large** then the **same signal will not be detected**
= no statistical significance

- In a statistical test, the ratio of signal to noise determines the significance.

Analysis of Quantitative Data

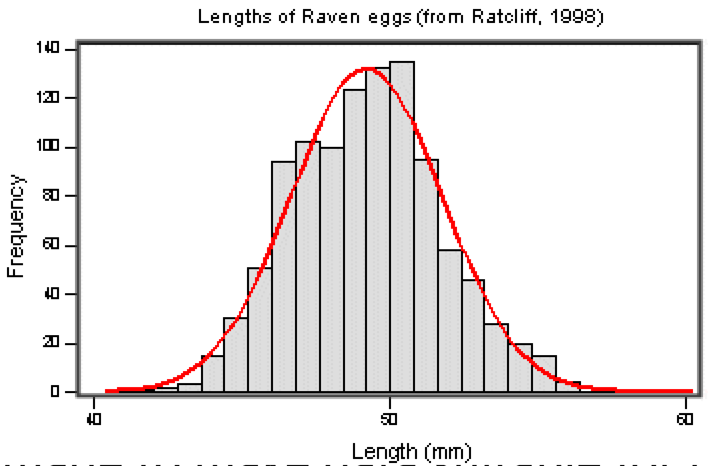
- Choose the correct statistical test to answer your question:
  - They are 2 types of statistical tests:
    - **Parametric tests** with 4 assumptions to be met by the data,
    - **Non-parametric tests** with no or few assumptions (e.g. Mann-Whitney test) and/or for qualitative data (e.g. Fisher's exact and χ^2 tests).

Assumptions of Parametric Data

- All parametric tests have 4 basic assumptions that must be met for the test to be accurate.

1) Normally distributed data

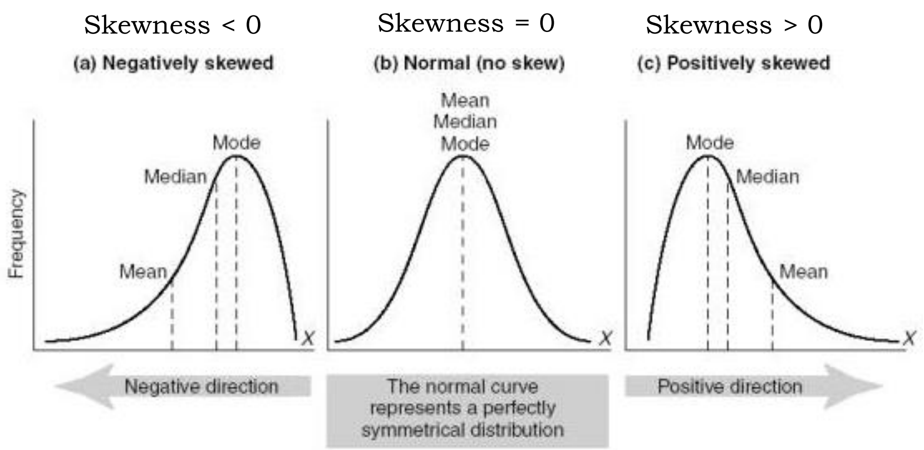
- Normal shape, bell shape, Gaussian shape



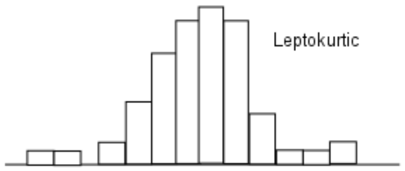
- Transformations can be made to make data suitable for parametric analysis.

Assumptions of Parametric Data

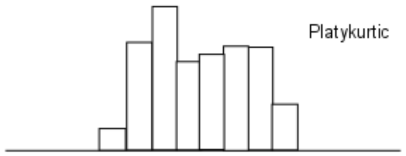
- Frequent departures from normality:
  - Skewness: lack of symmetry of a distribution



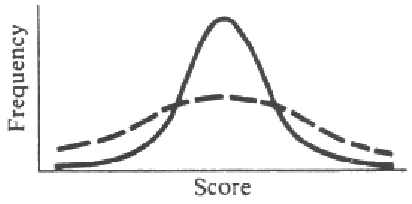
  - Kurtosis: measure of the degree of 'peakedness' in the distribution
    - The two distributions below have the same variance approximately the same skew, but differ markedly in kurtosis.



More peaked distribution: kurtosis > 0



Flatter distribution: kurtosis < 0



(e) Platykurtic and leptokurtic

Assumptions of Parametric Data

2) ***Homogeneity in variance***

- The variance should not change systematically throughout the data

3) ***Interval data (linearity)***

- The distance between points of the scale should be equal at all parts along the scale.

4) ***Independence***

- Data from different subjects are independent
  - Values corresponding to one subject do not influence the values corresponding to another subject.
  - Important in repeated measures experiments

Analysis of Quantitative Data

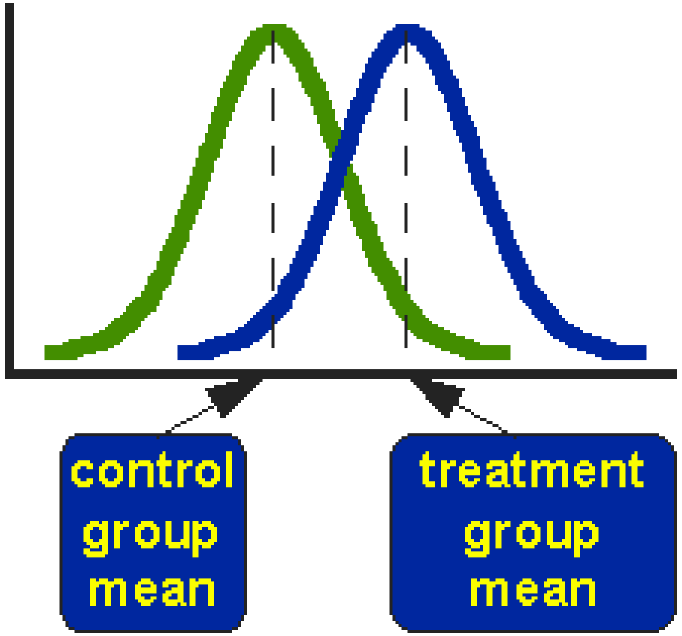
- **Is there a difference between my groups regarding the variable I am measuring?**
  - e.g. are the mice in the group A heavier than those in group B?
    - Tests with 2 groups:
      - Parametric: **Student's *t*-test**
      - Non parametric: **Mann-Whitney/Wilcoxon rank sum test**
    - Tests with more than 2 groups:
      - Parametric: **Analysis of variance (one-way and two-way ANOVA)**
      - Non parametric: **Kruskal Wallis**
- **Is there a relationship between my 2 (continuous) variables?**
  - e.g. is there a relationship between the daily intake in calories and an increase in body weight?
    - Test: **Correlation** (parametric or non-parametric) and **Curve fitting**

Comparison between 2 groups
Parametric data

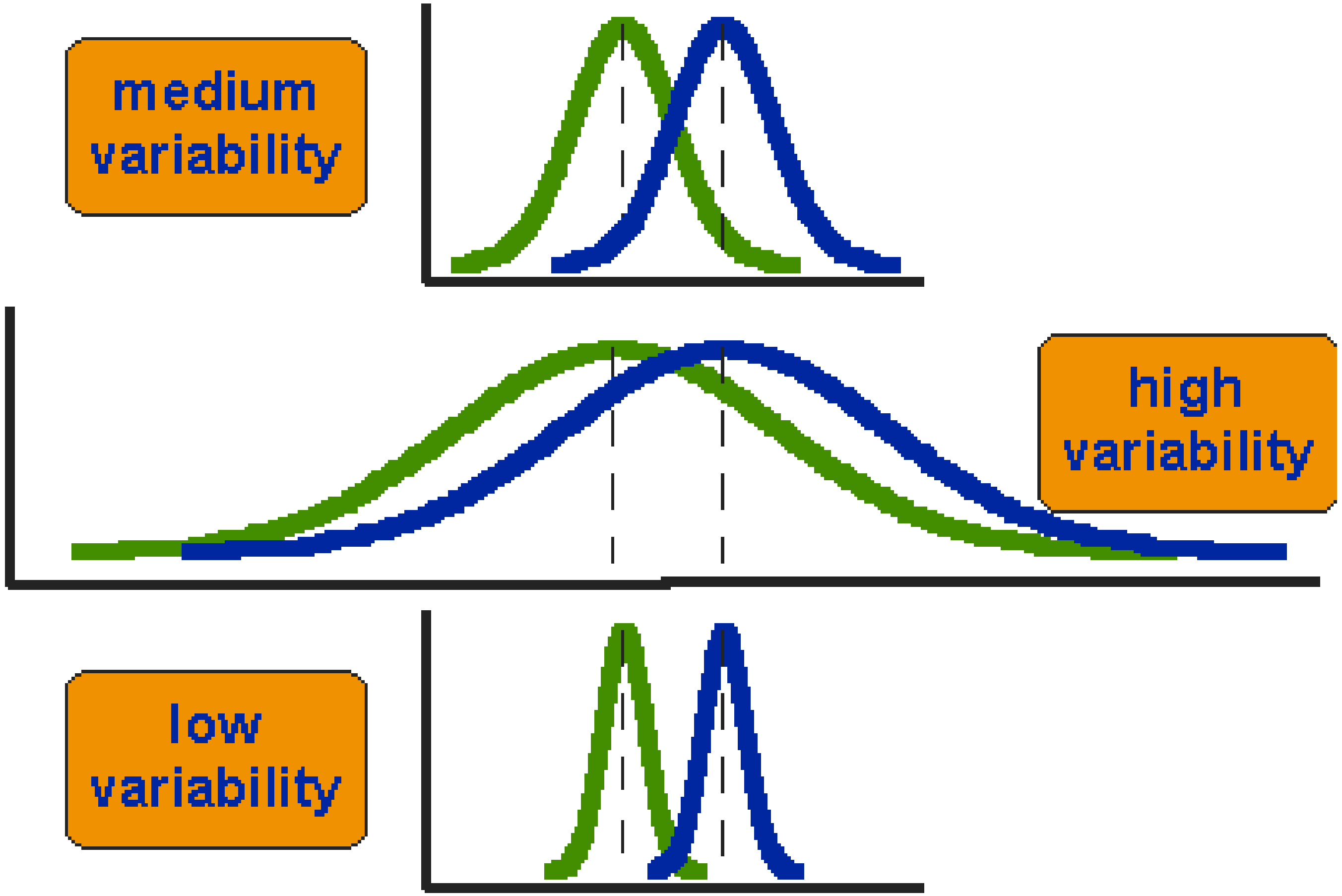
Babraham Bioinformatics

Comparison between 2 groups:
Student's *t*-test

- **Basic idea**:
  - When we are looking at the differences between scores for 2 groups, we have to judge the difference between their means relative to the spread or variability of their scores.
    - Eg: comparison of 2 groups: control and treatment



Student's *t*-test

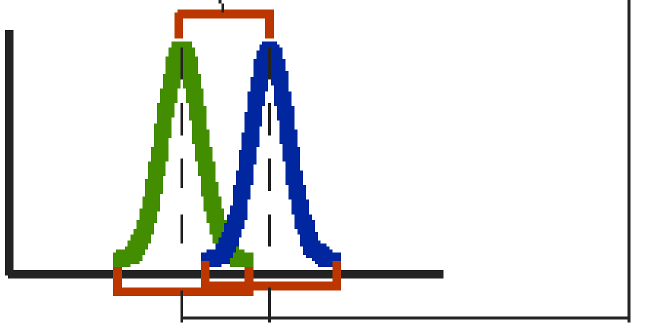


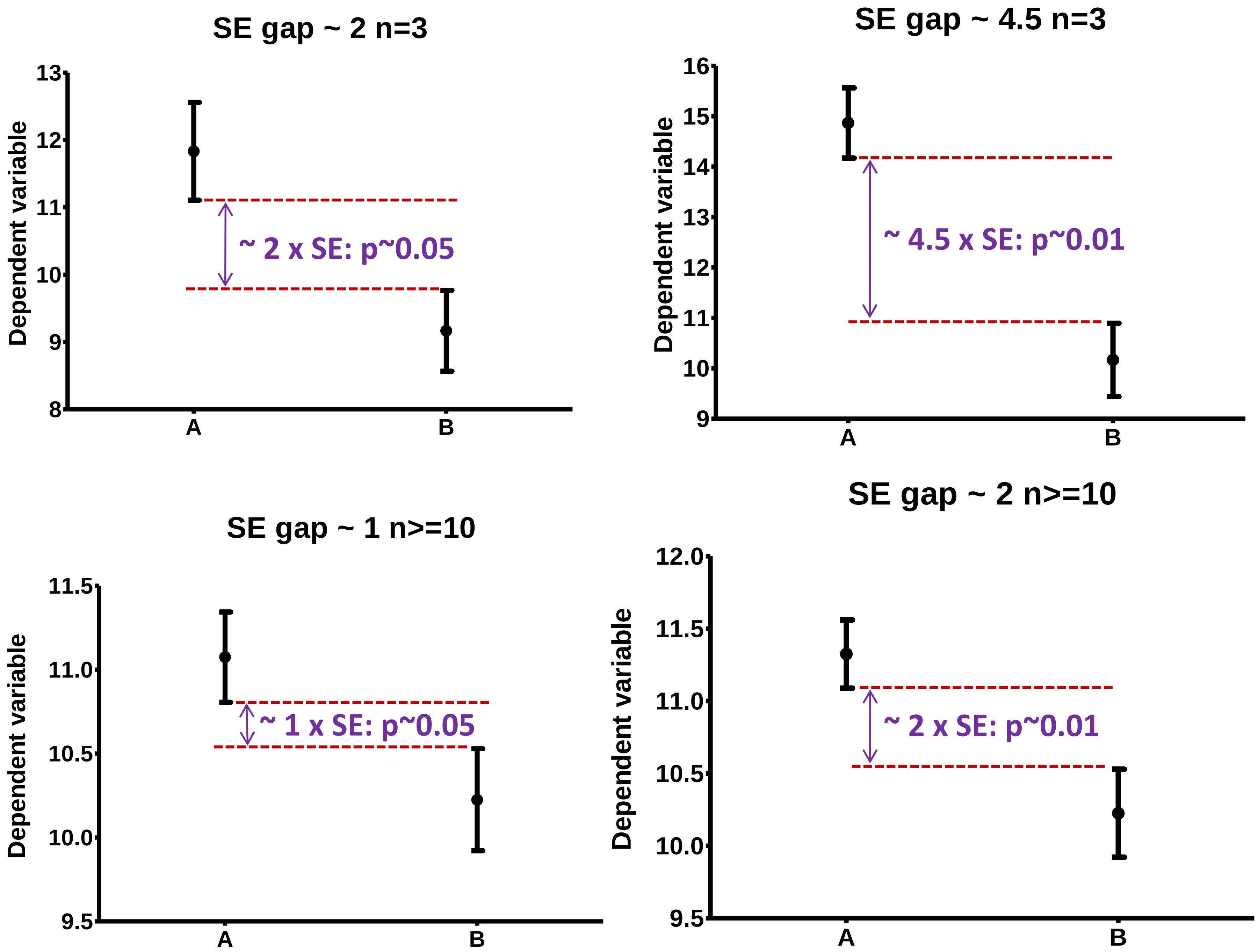
Student's *t*-test

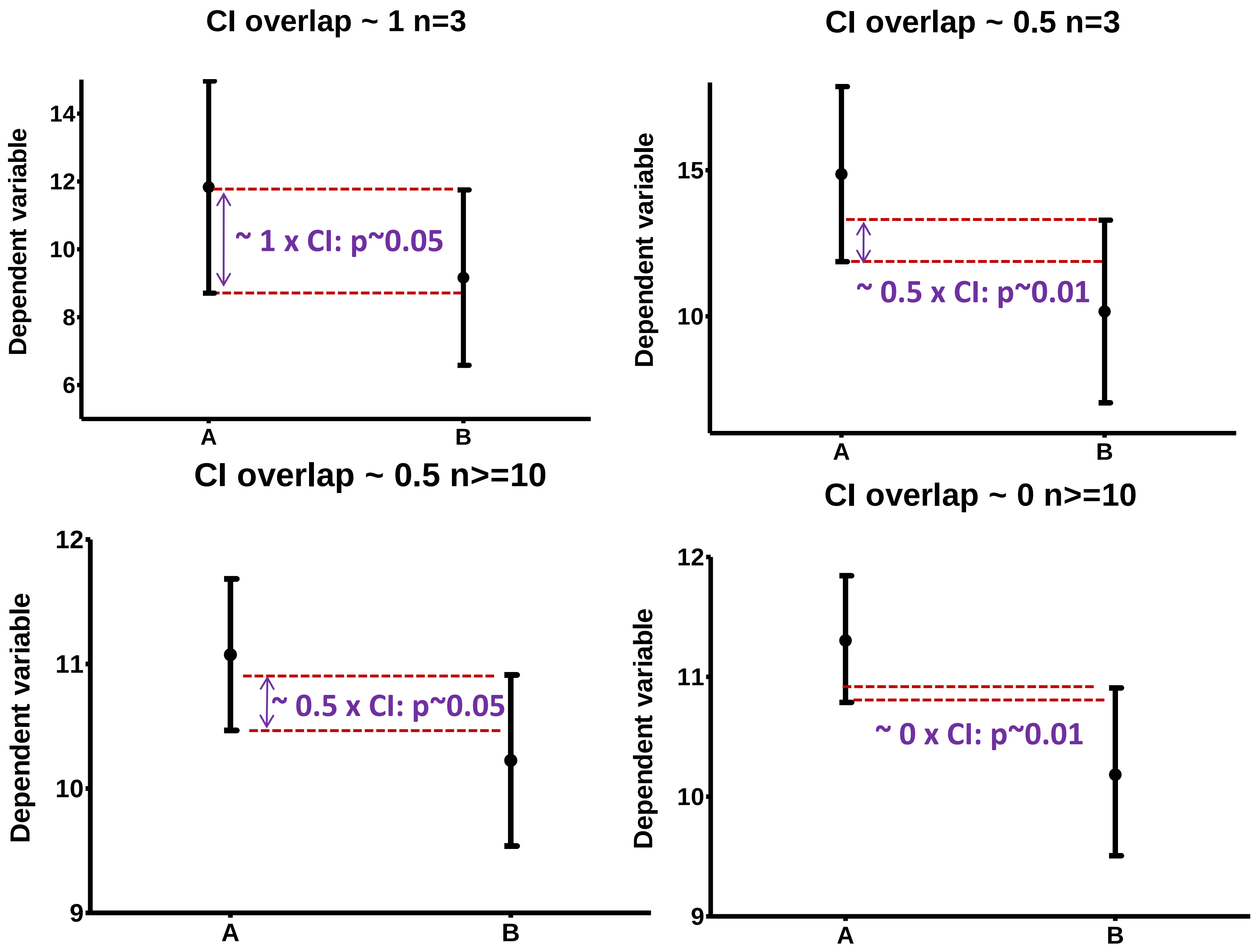
$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

$$= \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}}$$

$$= \text{t-value}$$




SE gap ~ 2 n=3
Dependent variable
~ 2 x SE: p~0.05
A
B
SE gap ~ 4.5 n=3
Dependent variable
~ 4.5 x SE: p~0.01
A
B
SE gap ~ 1 n>=10
Dependent variable
~ 1 x SE: p~0.05
A
B
SE gap ~ 2 n>=10
Dependent variable
~ 2 x SE: p~0.01
A
B



CI overlap ~ 1 n=3
~ 1 x CI: p~0.05
CI overlap ~ 0.5 n=3
~ 0.5 x CI: p~0.01
CI overlap ~ 0.5 n>=10
~ 0.5 x CI: p~0.05
CI overlap ~ 0 n>=10
~ 0 x CI: p~0.01
Dependent variable
A
B

Student's *t*-test

- 3 types:
  - **Independent t-test**
    - compares means for two independent groups of cases.
  - **Paired t-test**
    - looks at the difference between two variables for a single group:
      - the second 'sample' of values comes from the same subjects (mouse, petri dish …).
  - One-Sample t-test
    - tests whether the mean of a single variable differs from a specified constant (often 0)

Example: coyotes.xlsx

- Question: do male and female coyotes differ in size?

- **Sample size**
- **Data exploration**
- **Check the assumptions for parametric test**
- **Statistical analysis: Independent t-test**

Exercise 3: Power analysis

- Example case:

No data from a pilot study but we have found some information in the literature.

In a study run in similar conditions as in the one we intend to run, **male coyotes** were found to measure: **92cm+/- 7cm (SD)**.

We expect a **5% difference** between genders.

- **smallest biologically meaningful difference**

G*Power

Independent t-test

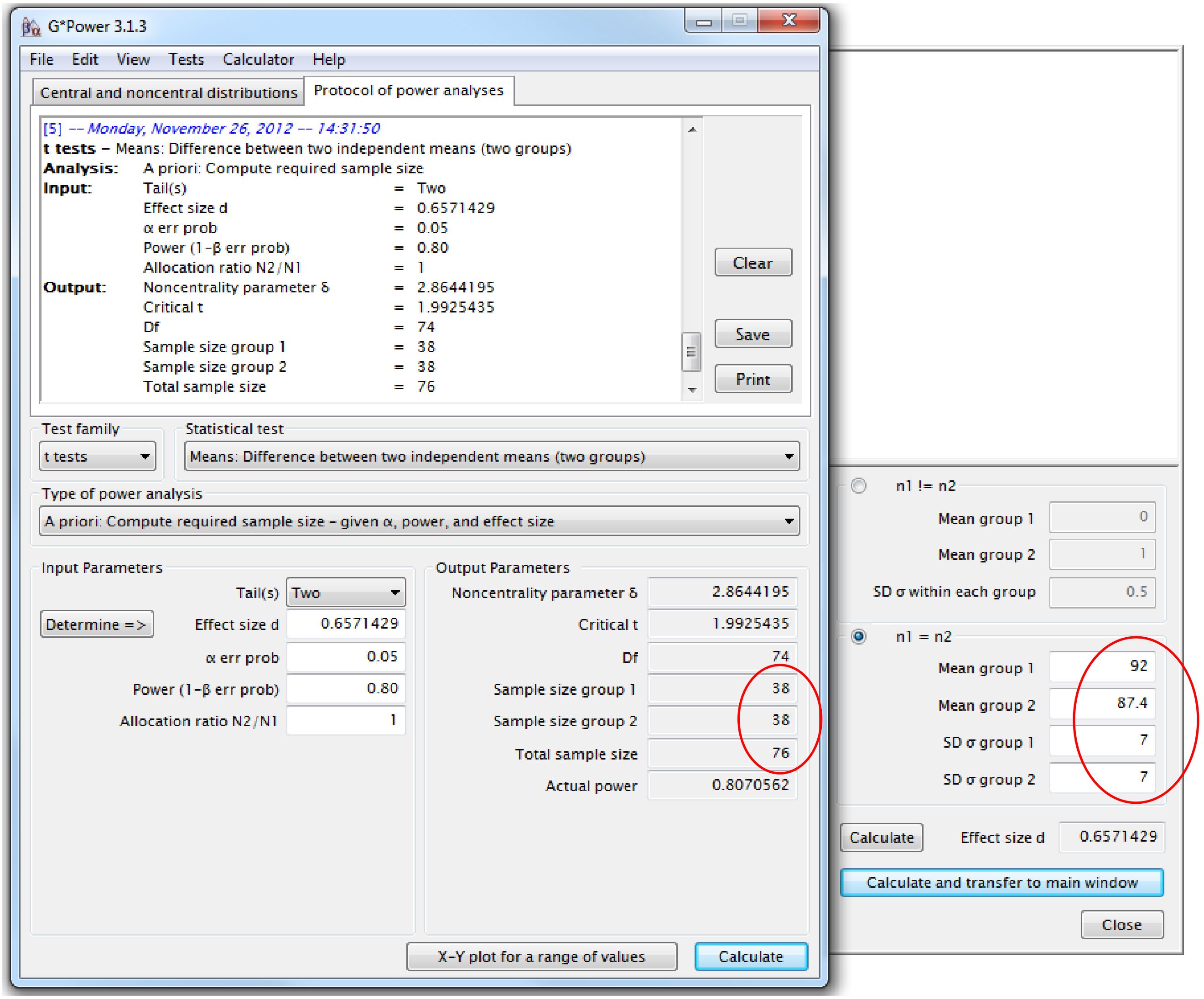
***A priori* Power analysis**

Example case:

You don't have data from a pilot study but you have found some information in the literature.

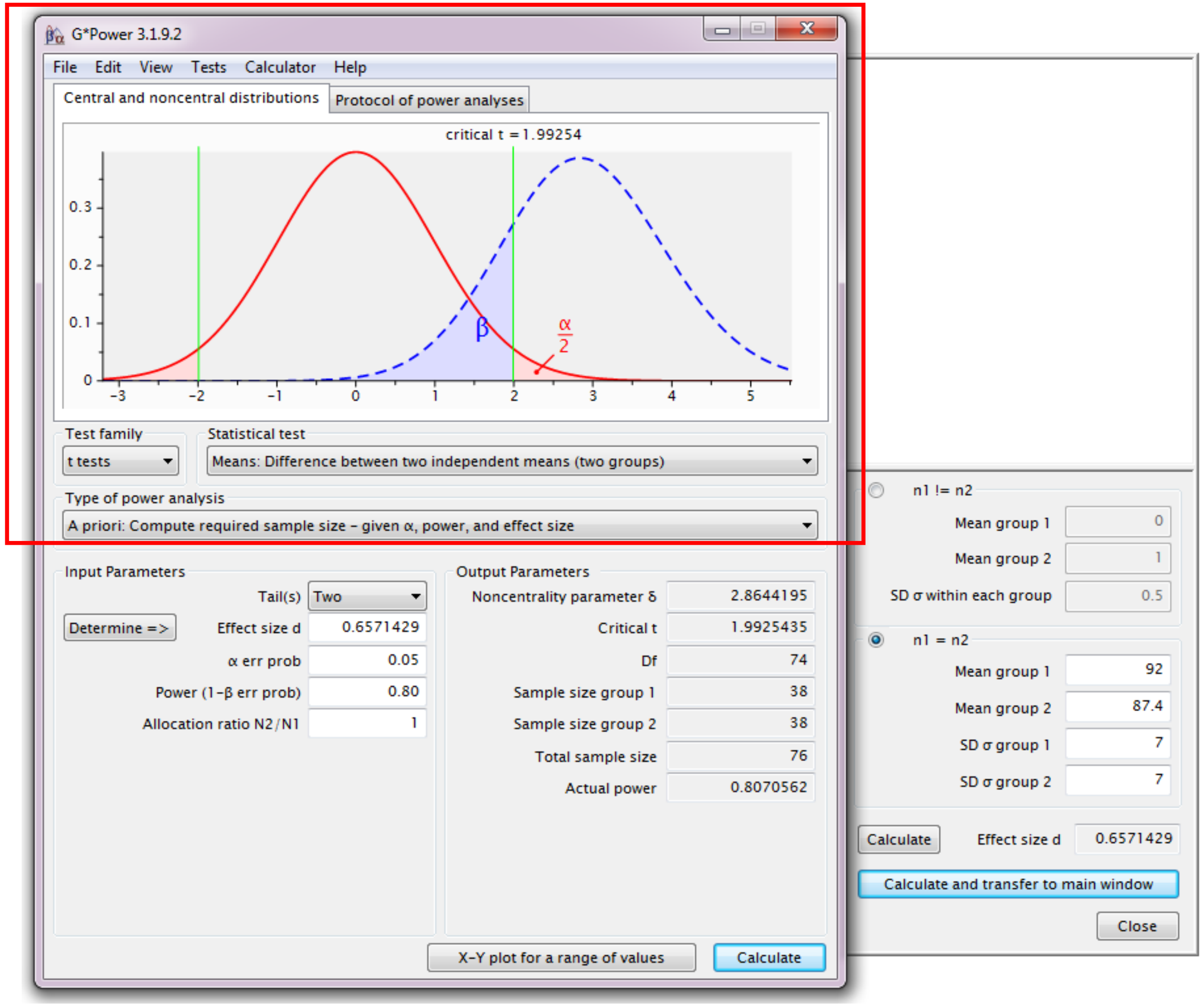
In a study run in similar conditions to the one you intend to run, male coyotes were found to measure:
92cm+/- 7cm (SD)

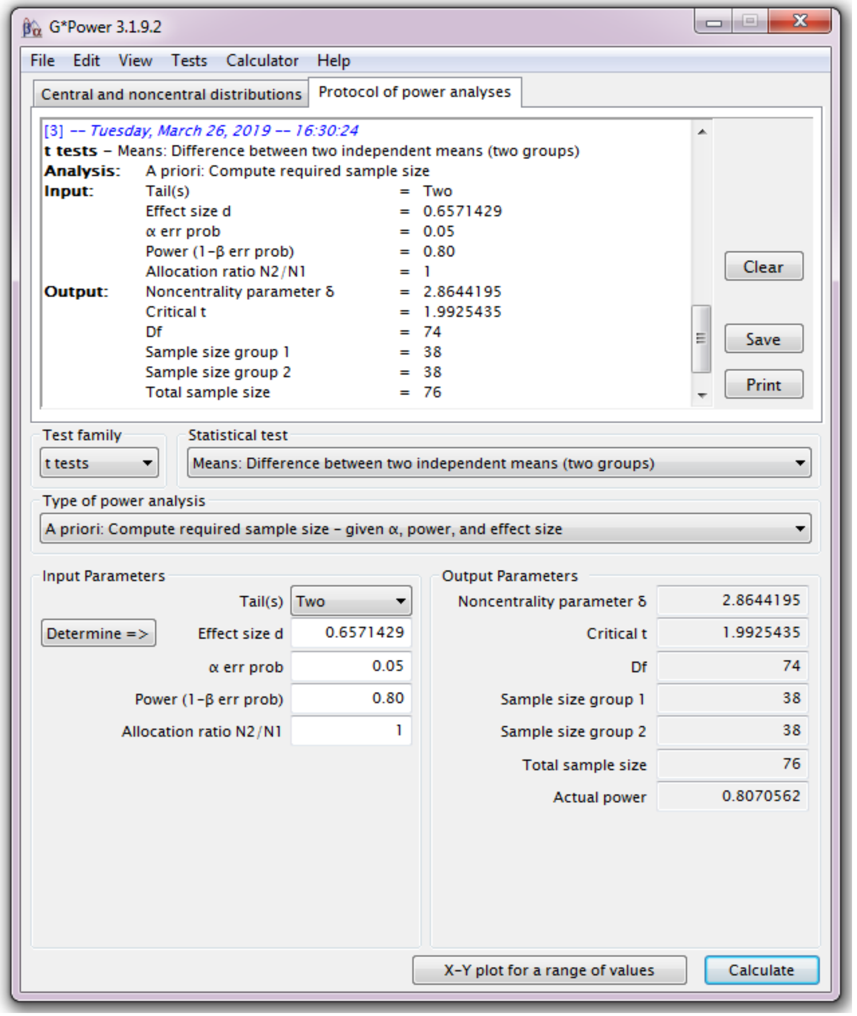
You expect a 5% difference between genders with a similar variability in the female sample.



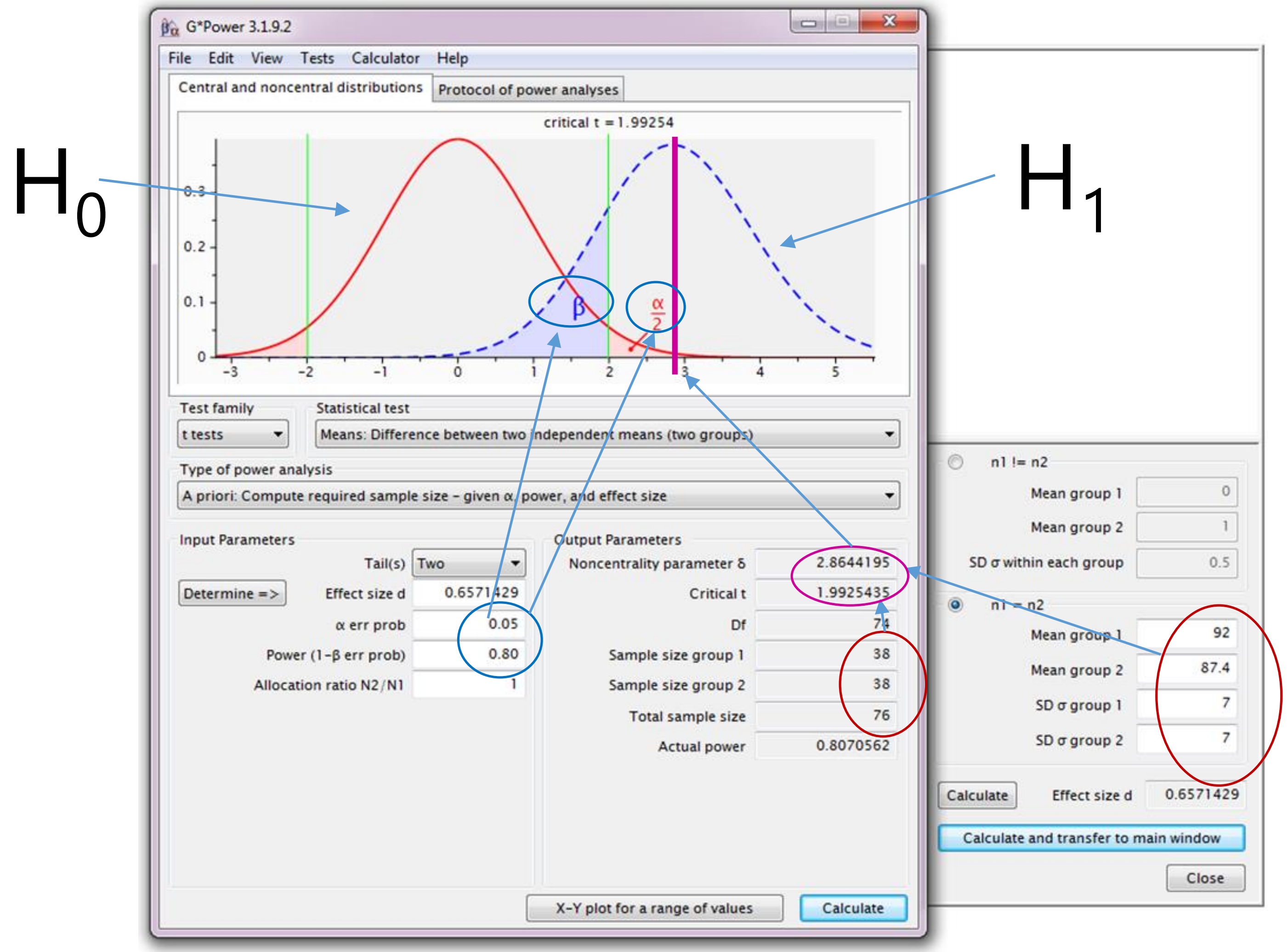
You need a sample size of n=76 (2*38)

Power Analysis



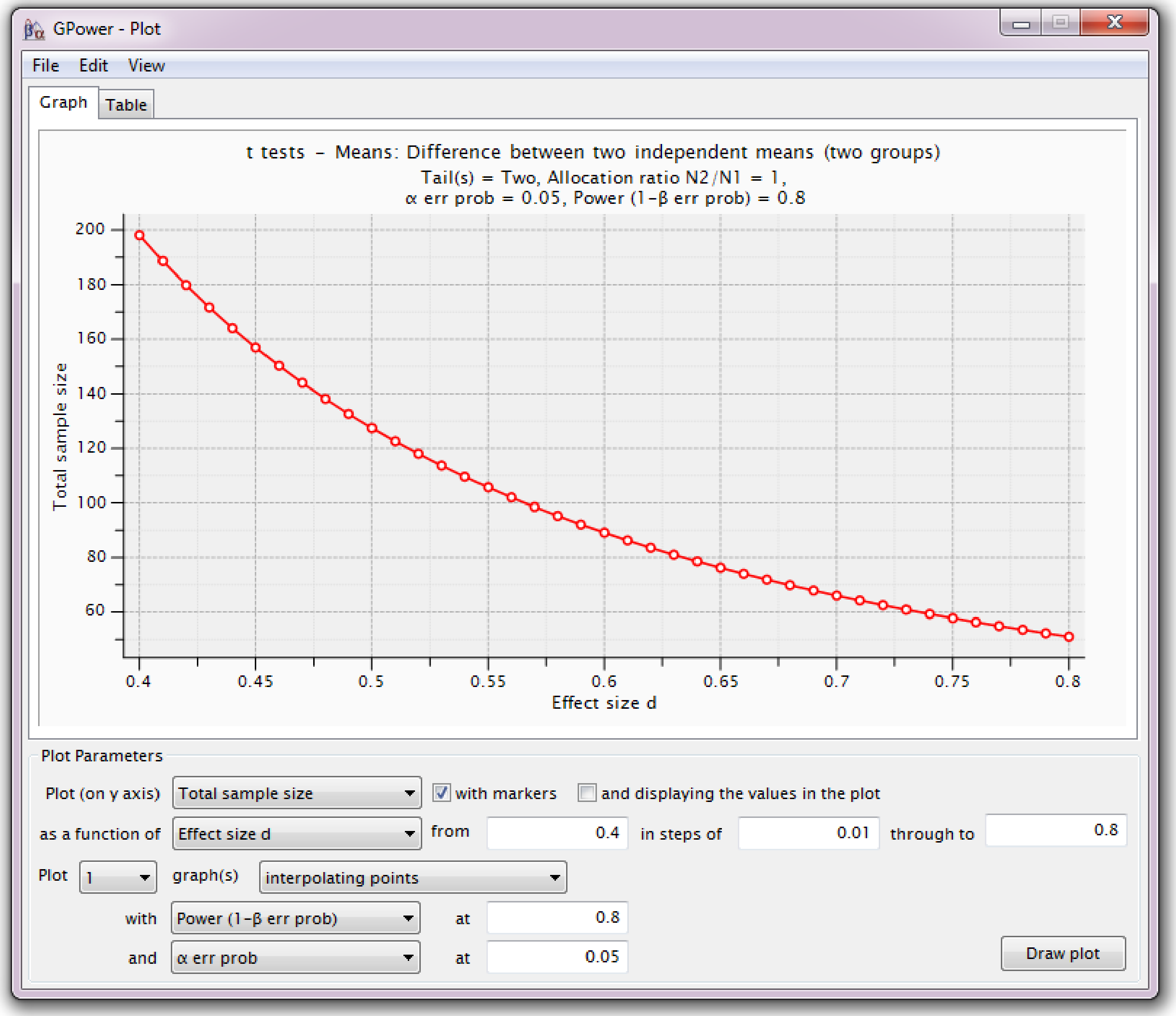


Power Analysis



Power Analysis

For a range of sample sizes:



Data exploration $\neq$ plotting data

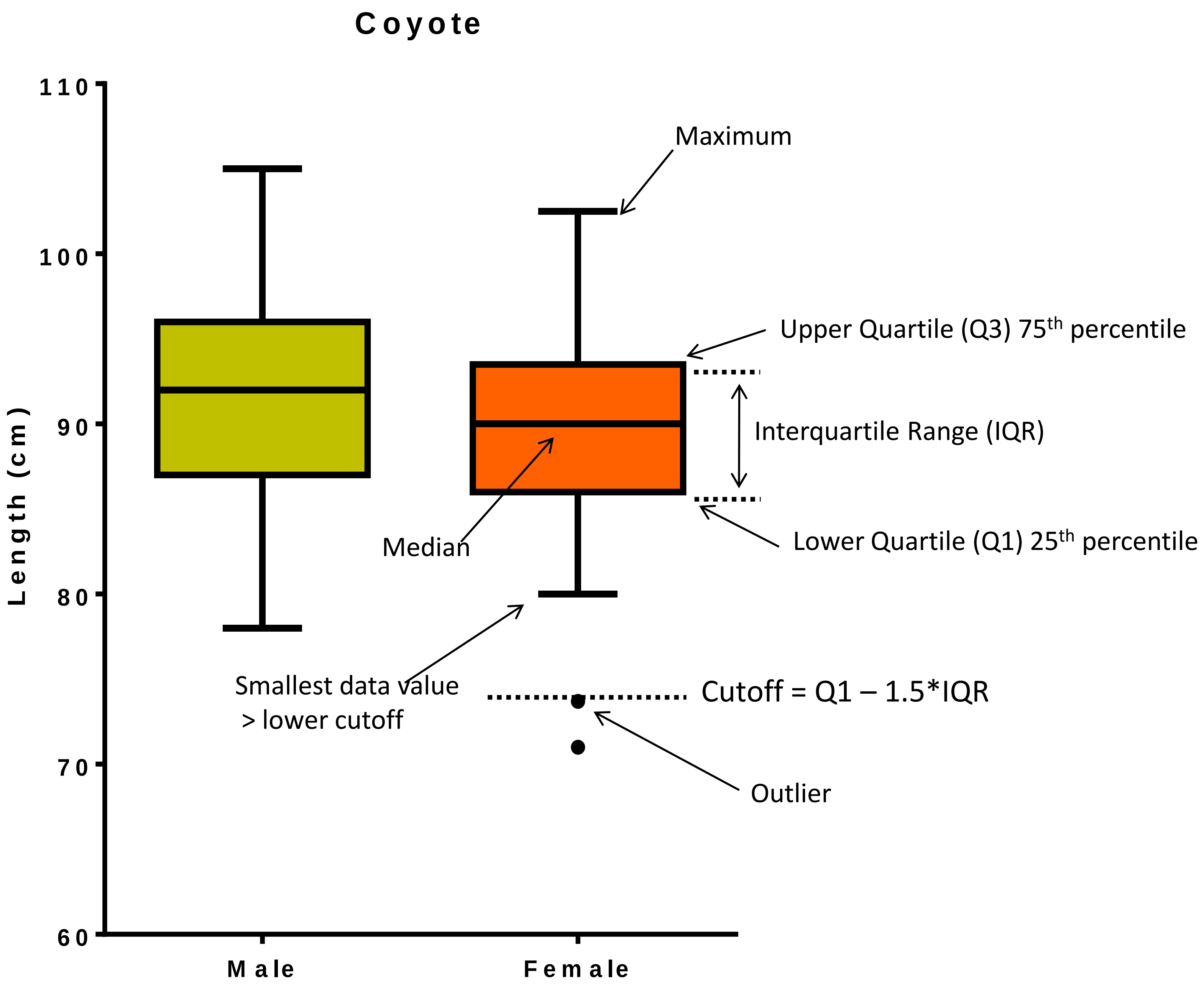
Babraham Bioinformatics

Exercise 4: Data exploration

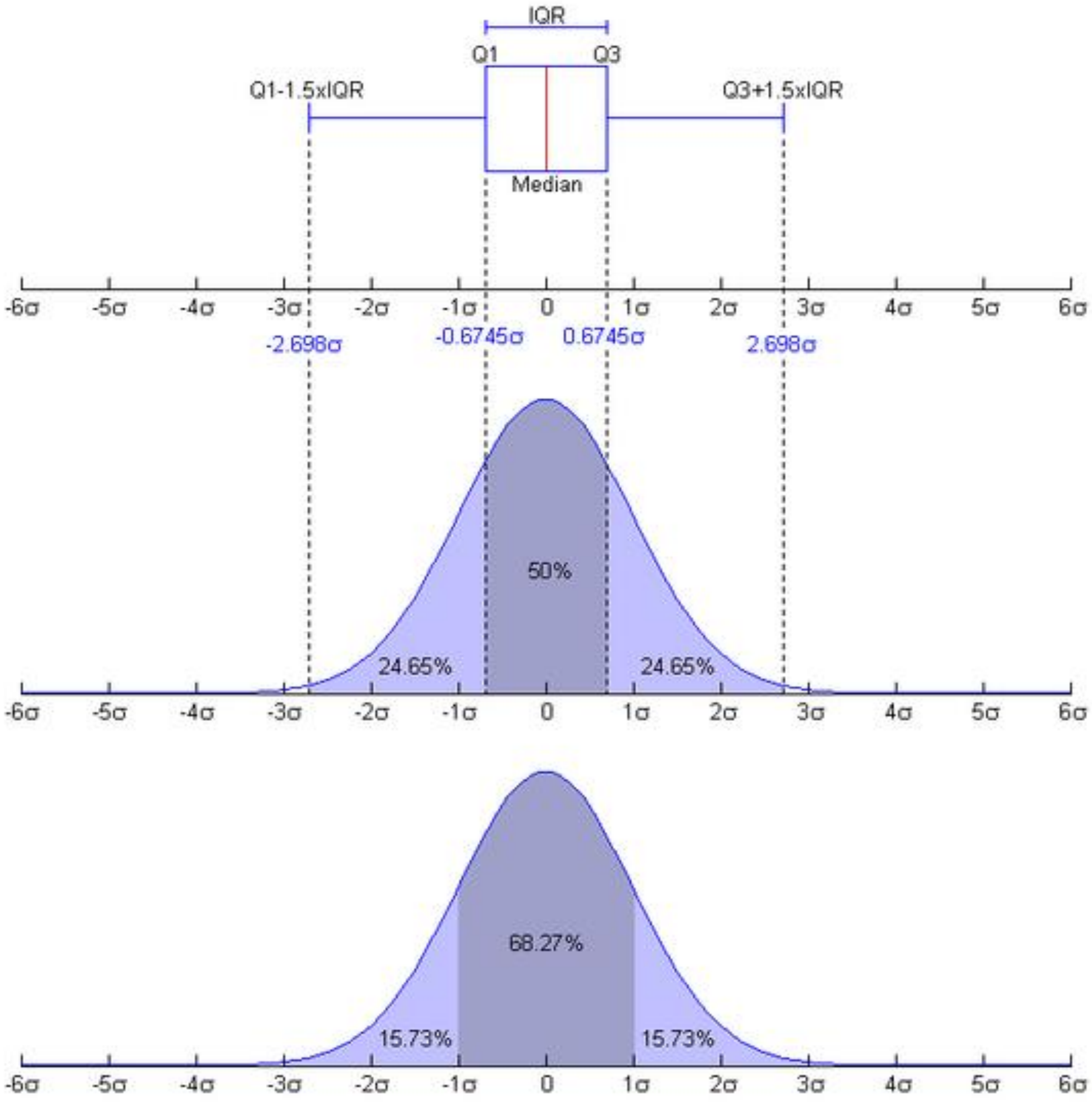
- The file contains individual body length of male and female coyotes.

Question: do male and female coyotes differ in size?

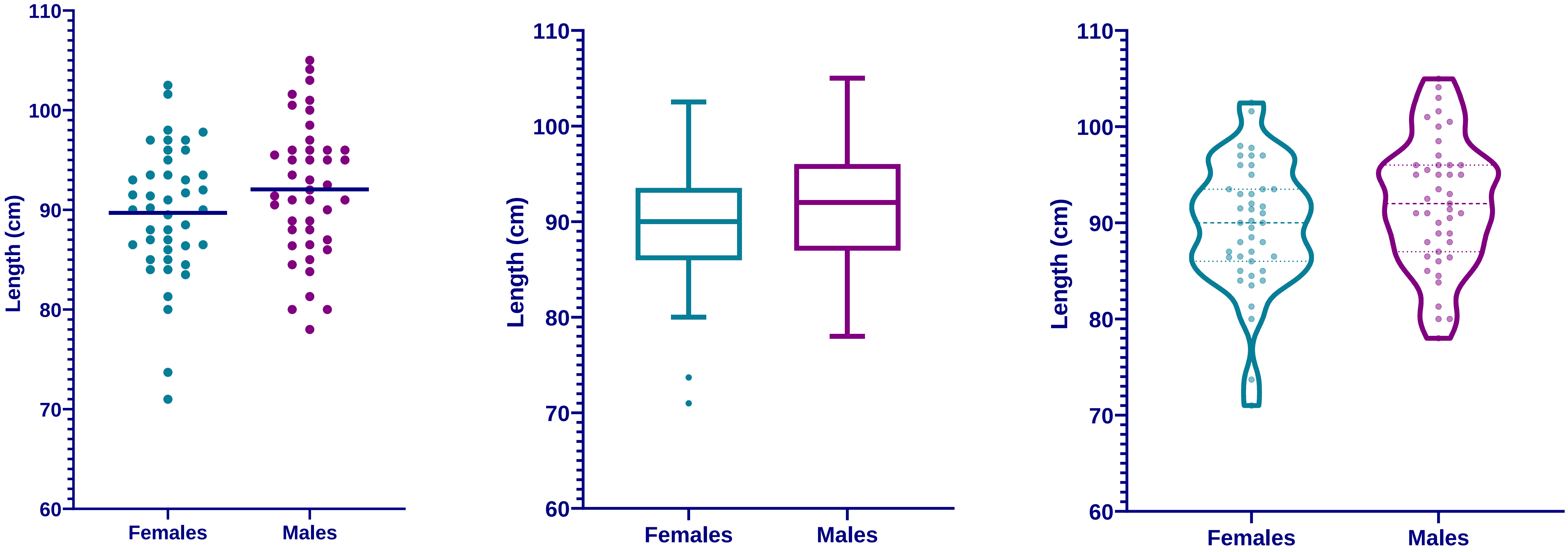
  - Plot the data as stripchart, boxplot and violinplot



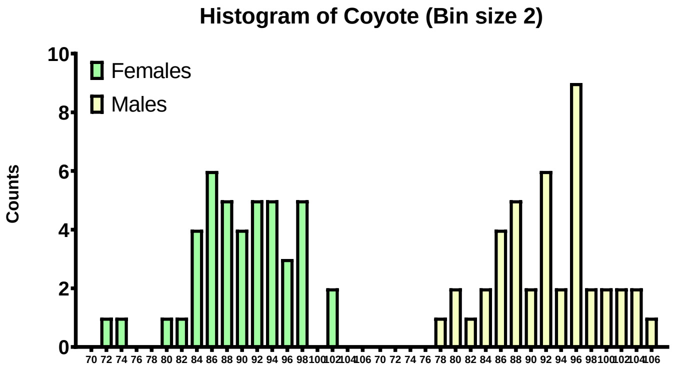
Coyote
110
100
90
80
70
60
Length (cm)
Male
Female
Maximum
Upper Quartile (Q3) 75th percentile
Interquartile Range (IQR)
Median
Lower Quartile (Q1) 25th percentile
Smallest data value
> lower cutoff
Cutoff = Q1 – 1.5*IQR
Outlier

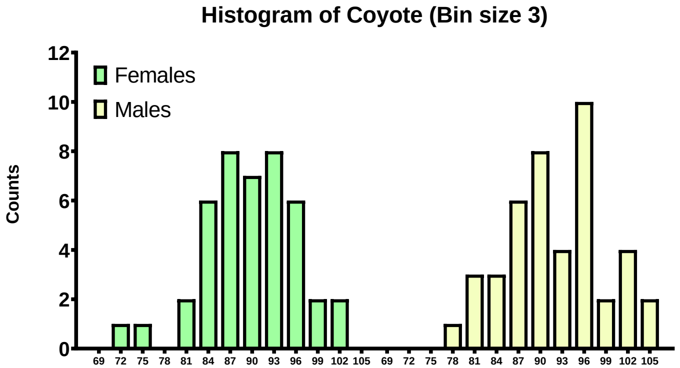

IQR
Q1
Q3
Q1-1.5xIQR
Q3+1.5xIQR
Median
-6σ -5σ -4σ -3σ -2σ -1σ 0 1σ 2σ 3σ 4σ 5σ 6σ
-2.698σ
-0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
-6σ -5σ -4σ -3σ -2σ -1σ 0 1σ 2σ 3σ 4σ 5σ 6σ
15.73%
68.27%
15.73%
-6σ -5σ -4σ -3σ -2σ -1σ 0 1σ 2σ 3σ 4σ 5σ 6σ

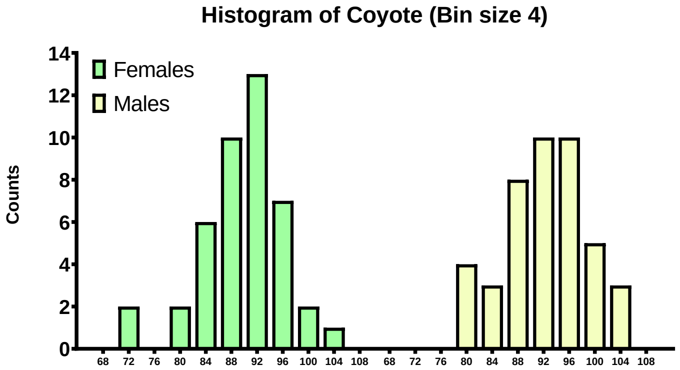
Exercise 4: Exploring data - *Answers*



Assumptions for parametric tests







Normality ☑

	Col. stats	A Females	B Males
1	Number of values	43	43
2			
3	Minimum	71.00	78.00
4	25% Percentile	86.00	87.00
5	Median	90.00	92.00
6	75% Percentile	93.50	96.00
7	Maximum	102.5	105.0
8			
9	Mean	89.71	92.06
10	Std. Deviation	6.550	6.696
11	Std. Error of Mean	0.9988	1.021
12			
13	Lower 95% CI of mean	87.70	90.00
14	Upper 95% CI of mean	91.73	94.12
15			
16	Sum	3858	3958
17			
18	**D'Agostino & Pearson normality test**		
19	K2	4.203	0.5080
20	P value	0.1223	0.7757
21	Passed normality test (alpha=0.05)?	Yes	Yes
22	P value summary	ns	ns
23			
24	**Shapiro-Wilk normality test**		
25	W	0.9700	0.9845
26	P value	0.3164	0.8190
27	Passed normality test (alpha=0.05)?	Yes	Yes
28	P value summary	ns	ns

Independent *t*-test: results

	Unpaired t test	
1	Table Analyzed	Coyote
2		
3	Column A	Females
4	vs.	vs.
5	Column B	Males
6		
7	**Unpaired t test**	
8	P value	0.1045
9	P value summary	ns
10	Significantly different (P < 0.05)?	No
11	One- or two-tailed P value?	Two-tailed
12	t, df	t=1.641, df=84
13		
14	**How big is the difference?**	
15	Mean of column A	89.71
16	Mean of column B	92.06
17	Difference between means (A - B) ± SEM	-2.344 ± 1.428
18	95% confidence interval	-5.185 to 0.4964
19	R squared (eta squared)	0.03107
20		
21	**F test to compare variances**	
22	F, DFn, Dfd	1.045, 42, 42
23	P value	0.8870
24	P value summary	ns
25	Significantly different (P < 0.05)?	No
26		
27	**Data analyzed**	
28	Sample size, column A	43
29	Sample size, column B	43
30		

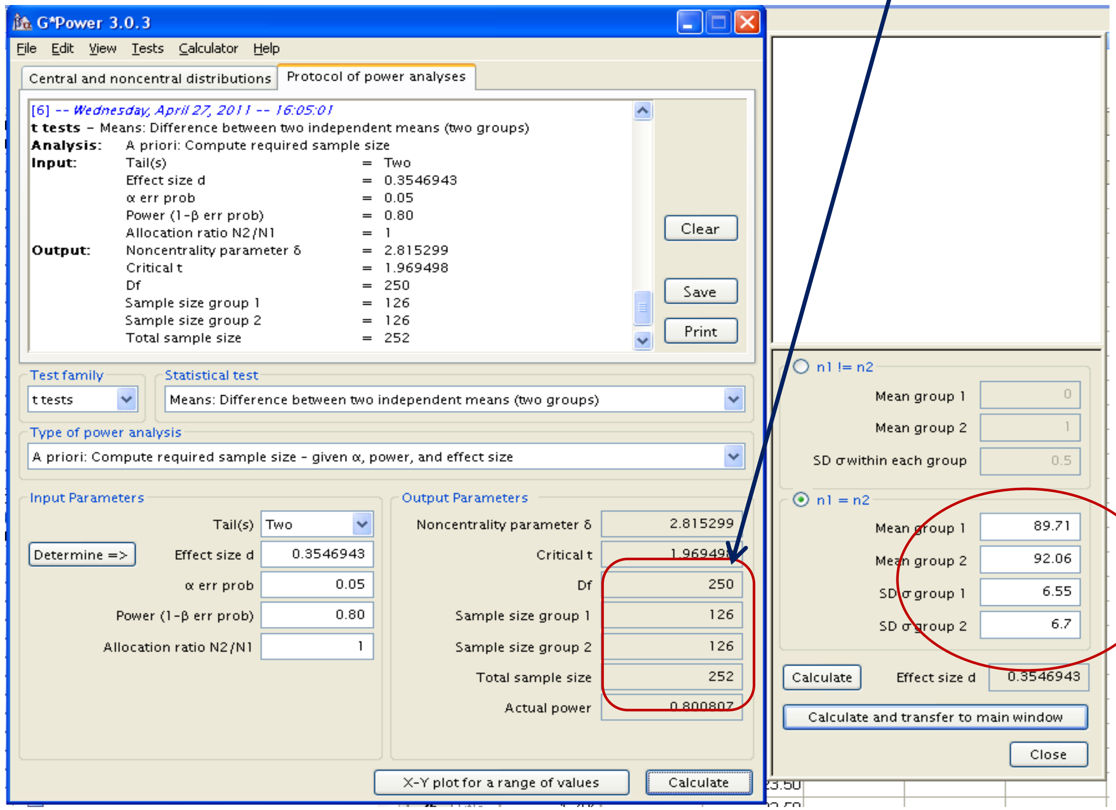
Males tend to be longer than females but not significantly so (p=0.1045)

Homogeneity in variance ☑

What about the power of the analysis?

Power analysis

You would need a sample 3 times bigger to reach the accepted power of 80%.



Col. stats	A Females	B Males
1 Number of values	43	43
2		
3 Minimum	71.00	78.00
4 25% Percentile	86.00	87.00
5 Median	90.00	92.00
6 75% Percentile	93.50	96.00
7 Maximum	102.5	105.0
8		
9 Mean	89.71	92.06
10 Std. Deviation	6.550	6.696
11 Std. Error of Mean	0.9988	1.021
12		
13 Lower 95% CI of mean	87.70	90.00
14 Upper 95% CI of mean	91.73	94.12
15		
16 Sum	3858	3958
17		
18 D'Agostino & Pearson normality test		
19 K2	4.203	0.5080
20 P value	0.1223	0.7757
21 Passed normality test (alpha=0.05)?	Yes	Yes
22 P value summary	ns	ns
23		
24 Shapiro-Wilk normality test		
25 W	0.9700	0.9845
26 P value	0.3164	0.8190
27 Passed normality test (alpha=0.05)?	Yes	Yes
28 P value summary	ns	ns

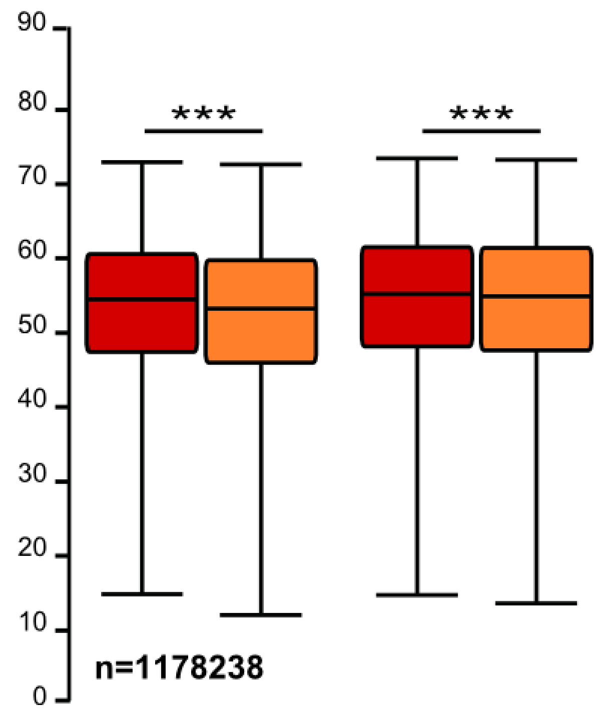
But is a 2.3 cm difference between genders biologically relevant (<3%) ?

Sample size: the bigger the better?

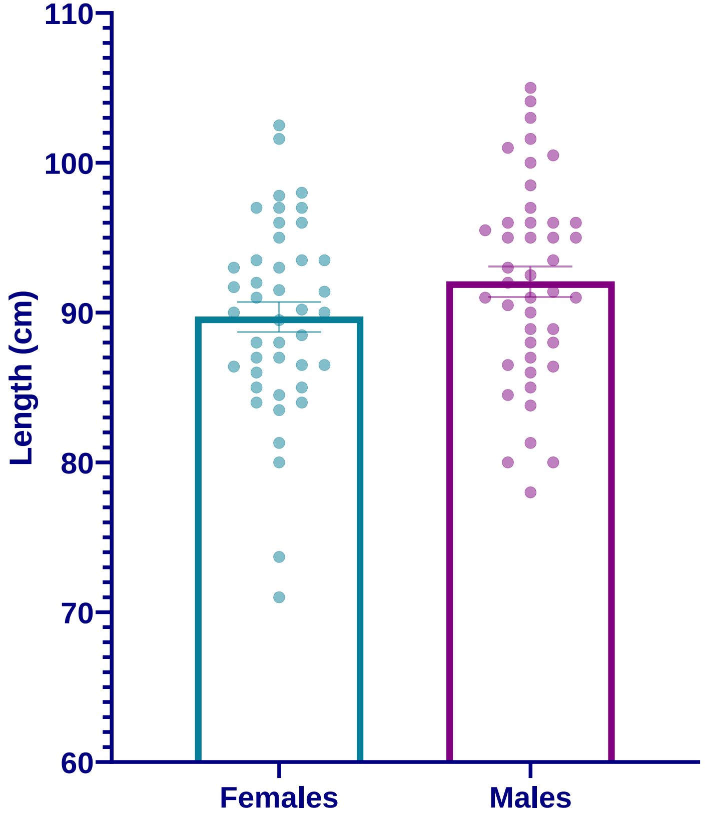
- It takes huge samples to detect tiny differences but tiny samples to detect huge differences.

- What if the tiny difference is meaningless?
  - Beware of **overpower**
  - Nothing wrong with the stats: it is all about interpretation of the results of the test.

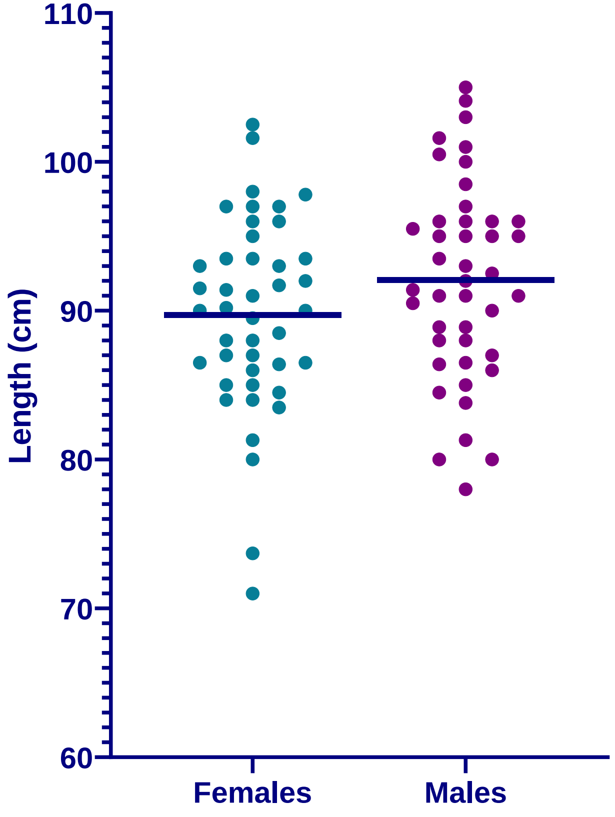
- Remember the important first step of power analysis
  - **What is the effect size of biological interest?**



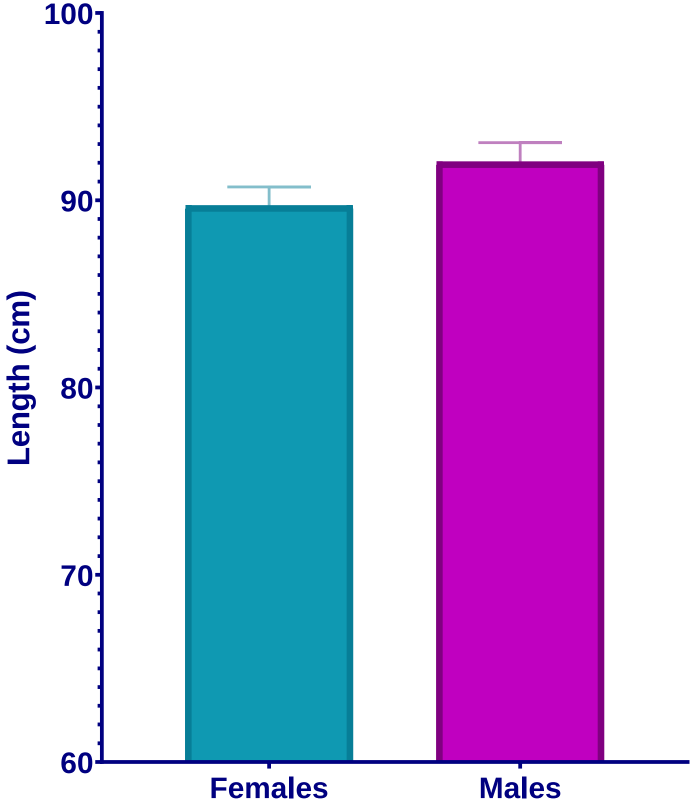
Coyotes



110
100
90
80
70
60
Length (cm)
Females
Males



110
100
90
80
70
60
Length (cm)
Females
Males



100
90
80
70
60
Length (cm)
Females
Males

Exercise 5: Dependent or Paired *t*-test

working memory.xlsx

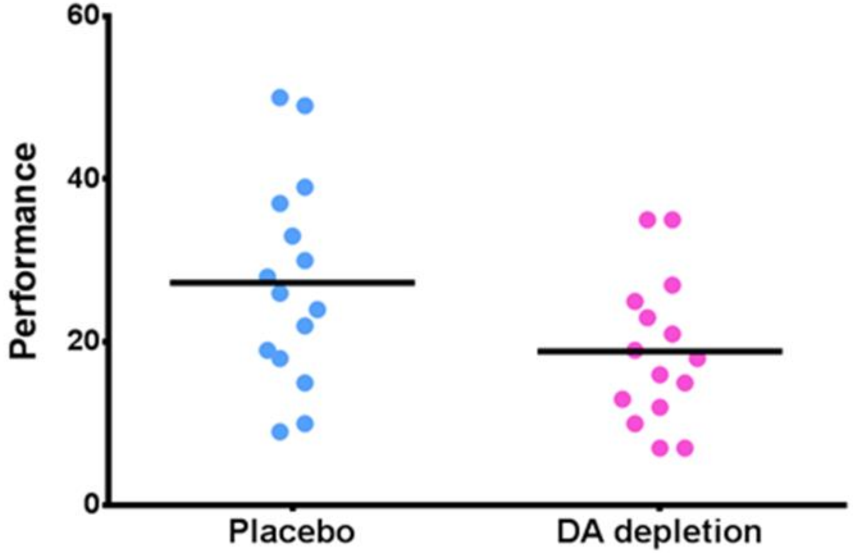
A group of rhesus monkeys (n=15) performs a task involving memory after having received a placebo. Their performance is graded on a scale from 0 to 100. They are then asked to perform the same task after having received a dopamine depleting agent.

Is there an effect of treatment on the monkeys' performance?

Another example of *t*-test:

working memory.xlsx

	Col. stats	A Placebo Y	B DA depletion Y
1	Number of values	15	15
2			
3	Minimum	9.000	7.000
4	25% Percentile	18.00	12.00
5	Median	26.00	18.00
6	75% Percentile	37.00	25.00
7	Maximum	50.00	35.00
8			
9	Mean	27.27	18.87
0	Std. Deviation	12.65	8.911
1	Std. Error of Mean	3.265	2.301
2			
3	Lower 95% CI of mean	20.26	13.93
4	Upper 95% CI of mean	34.27	23.80
5			
6	D'Agostino & Pearson omnibus normality test		
7	K2	0.6754	0.9815
8	P value	0.7134	0.6122
9	Passed normality test (alpha=0.05)?	Yes	Yes
0	P value summary	ns	ns
1			
2	Sum	409.0	283.0

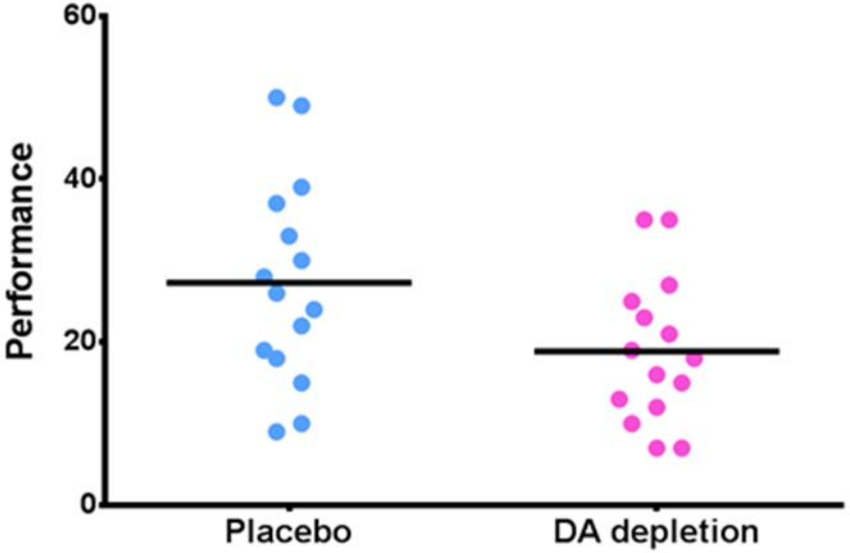


Normality ☑

Another example of *t*-test:

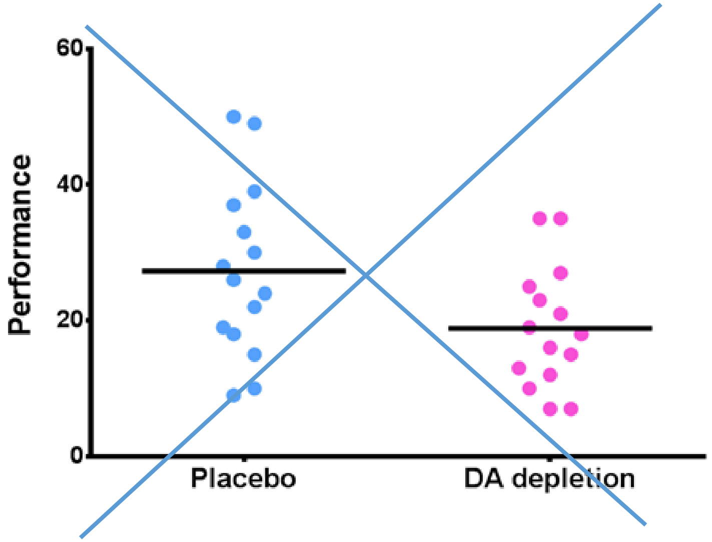
working memory.xlsx

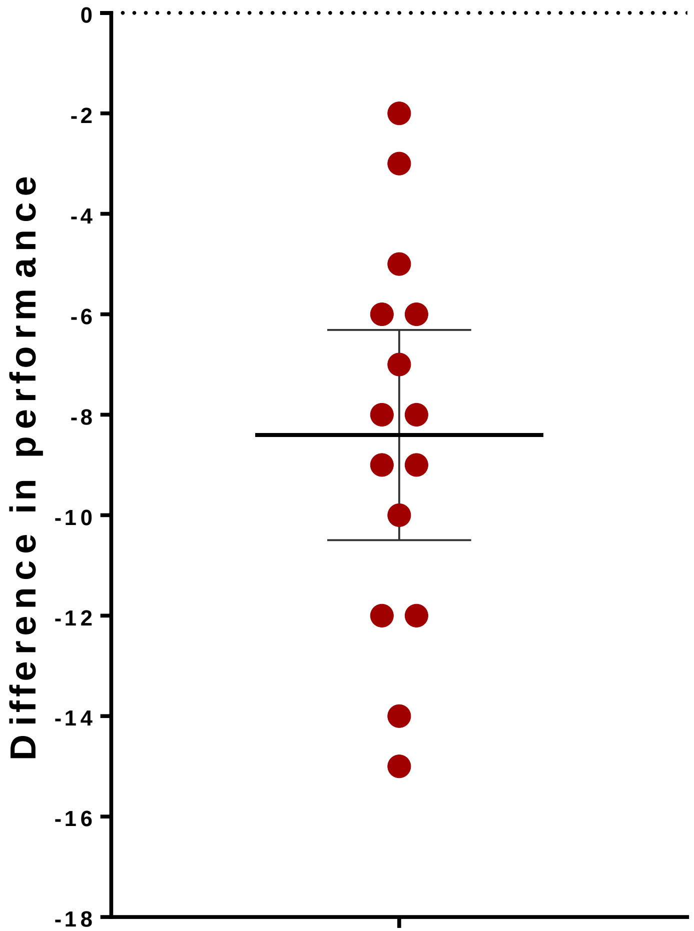
	Paired t test	
1	Table Analyzed	Working memory
2		
3	Column A	Placebo
4	vs.	vs.
5	Column B	DA depletion
6		
7	**Paired t test**	
8	P value	<0.0001
9	P value summary	****
10	Significantly different (P < 0.05)?	Yes
11	One- or two-tailed P value?	Two-tailed
12	t, df	t=8.616, df=14
13	Number of pairs	15
14		
15	**How big is the difference?**	
16	Mean of differences	8.400
17	SD of differences	3.776
18	SEM of differences	0.9749
19	95% confidence interval	6.309 to 10.49
20	R squared (partial eta squared)	0.8413
21		
22	**How effective was the pairing?**	
23	Correlation coefficient (r)	0.9986
24	P value (one tailed)	<0.0001
25	P value summary	****
26	Was the pairing significantly effective?	Yes
27		



Paired *t*-test: Results

working memory.xlsx





Comparison between 2 groups
Non-Parametric data

Babraham Bioinformatics

Non-parametric test:

Mann-Whitney = Wilcoxon rank test

- Non-parametric equivalent of the t-test.
- **What if the data do not meet the assumptions for parametric tests?**
  - The outcome is a rank or a score with limited amount of possible values: non-parametric approach.

- **How does the Mann-Whitney test work?**

Group 1	Group 2
5	8
7	9
3	6

→

Real values	Ranks
3	1
5	2
6	3
7	4
8	5
9	6
Mean	3.5

→

	Group 1	Group 2
	2	5
	4	6
	1	3
Sum	7	14

- Statistic of the Mann-Whitney test: **W (U)**
  - W = sum of ranks – mean rank: W_1=3.5 and W_2=10.5
  - Smallest of the 2 Ws: W_1 + sample size = **p-value**

Exercise 6: smelly teeshirt.xlsx

- Hypothesis: Group body odour is less disgusting when associated with an in-group member versus an out-group member.
- Study: Two groups of Cambridge University students are presented with one of two smelly, worn t-shirts with university logos.

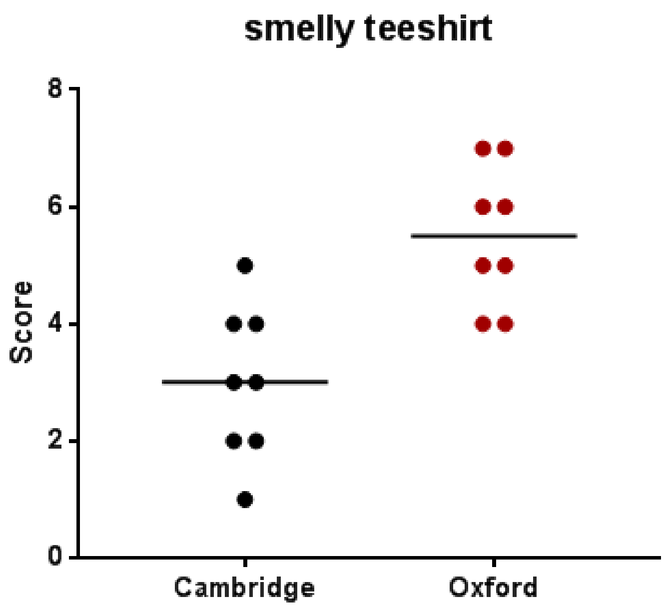
- **Question**: are Cambridge students more disgusted by worn smelly T-shirts from Oxford or Cambridge? Disgust score: 1 to 7, with 7 the most disgusting

  - Explore the data with an appropriate combination of 2 graphs
  - Answer the question with a non-parametric approach
  - What do you think about the design?

Exercise 6: smelly teeshirt.xlsx

- **Question**: are Cambridge students more disgusted by worn smelly T-shirts from Oxford or Cambridge?

Disgust score: 1 to 7, with 7 the most disgusting



	Mann-Whitney test	
1	Table Analyzed	smelly teeshirt
2		
3	Column B	Oxford
4	vs.	vs.
5	Column A	Cambridge
6		
7	Mann Whitney test	
8	P value	0.0037
9	Exact or approximate P value?	Exact
10	P value summary	**
11	Significantly different (P < 0.05)?	Yes
12	One- or two-tailed P value?	Two-tailed
13	Sum of ranks in column A,B	41 , 95
14	Mann-Whitney U	5
15		

- A paired design would have been better.

Non-parametric test:
Wilcoxon's signed-rank

- Non-parametric equivalent of the paired t-test
- **How does the test work?**

Before	After	Differences
9	3	-6
7	4	-3
10	4	-6
8	5	-3
5	6	1
8	2	-6
7	7	0
9	4	-5
10	5	-5

Ranking	Ranks
0	
1	1
3	2.5
3	2.5
5	4.5
5	4.5
6	7
6	7
6	7

	Negative rank	Positive rank
	-1	
	-2.5	
	-2.5	
		4.5
	-4.5	
	-7	
	-7	
	-7	
Sum	-31.5	4.5

- Statistic of the Wilcoxon's signed-rank test: **T (W)**
  - Here: Wilcoxon's T = 4.5 (smallest of the 2 (absolute value))
  - N = 9 (we ignore the 0 difference): T + N → **p-value**

Exercise 7: botulinum.xlsx

	Before	After
1	9	3
2	7	4
3	10	4
4	8	5
5	9	6
6	8	2
7	7	4
8	9	4
9	10	5

A group of 9 disabled children with muscle spasticity (or extreme muscle tightness limiting movement) in their right upper limb underwent a course of injections with botulinum toxin to reduce spasticity levels.
A second group of 9 children received the injections alongside a course of physiotherapy.
A neurologist (blind to group membership) assessed levels of spasticity pre- and post-treatment for all 18 children using a 10-point ordinal scale.

Higher ratings indicated higher levels of spasticity.

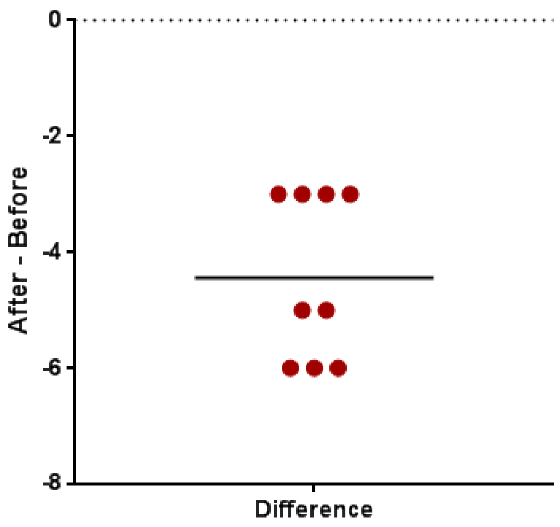
- **Question**: do botulinum toxin injections reduce muscle spasticity levels?
  - Score: 1 to 10, with 10 the highest spasticity

Exercise 7: botulinum.xlsx

	Before	After
1	9	3
2	7	4
3	10	4
4	8	5
5	9	6
6	8	2
7	7	4
8	9	4
9	10	5

- **Question**: do botulinum toxin injections reduce muscle spasticity levels?

	Wilcoxon test	
1	Table Analyzed	botulinum
2		
3	Column B	after
4	vs.	vs.
5	Column A	before
6		
7	Wilcoxon matched-pairs signed rank test	
8	P value	0.0039
9	Exact or approximate P value?	Exact
10	P value summary	**
11	Significantly different (P < 0.05)?	Yes
12	One- or two-tailed P value?	Two-tailed
13	Sum of positive, negative ranks	0 , -45
14	Sum of signed ranks (W)	-45
15	Number of pairs	9



Answer: There was a significant difference pre- and post- treatment in ratings of muscle spasticity. (T=-45, p=0.004).
Note: T=W

Comparison between more than 2 groups

One factor

Comparison of more than 2 means

- Running multiple tests on the same data increases the **familywise error rate**.

- What is the familywise error rate?
  - The error rate across tests conducted on the same experimental data.

- One of the basic rules ('laws') of probability:
  - The Multiplicative Rule: The probability of the joint occurrence of 2 or more independent events is the product of the individual probabilities.

$$P(A,B) = P(A) \times P(B)$$

For example:

$$P(2\ Heads) = P(head) \times P(head) = 0.5 \times 0.5 = 0.25$$

Familywise error rate

- **Example**: All pairwise comparisons between 3 groups A, B and C:
  - A-B, A-C and B-C

- Probability of making the Type I Error: **5%**
  - The probability of not making the Type I Error is 95% (=1 – 0.05)

- Multiplicative Rule:
  - Overall probability of no Type I errors is: 0.95 * 0.95 * 0.95 = 0.857

- So the probability of making at least one Type I Error is 1-0.857 = 0.143 or **14.3%**
    - The probability has increased from 5% to 14.3%

- Comparisons between 5 groups instead of 3, the familywise error rate is **40%** (=1-$(0.95)^n$)

Familywise error rate

- <u>Solution</u> to the increase of familywise error rate: correction for multiple comparisons
  - **Post-hoc tests**

- Many different ways to correct for multiple comparisons:
  - Different statisticians have designed corrections addressing different issues
    - e.g. unbalanced design, heterogeneity of variance, liberal vs conservative

- However, they all have **one thing in common**:
  - the more tests, the higher the familywise error rate: the more stringent the correction

- Tukey, Bonferroni, Sidak, Benjamini-Hochberg …
  - Two ways to address the multiple testing problem
    - **Familywise Error Rate** (FWER) vs. **False Discovery Rate** (FDR)

Multiple testing problem

- **FWER**: **Bonferroni**: α_{adjust} = 0.05/n comparisons e.g. 3 comparisons: 0.05/3=0.016
  - Problem: very conservative leading to loss of power (lots of false negative)
  - 10 comparisons: threshold for significance: 0.05/10: 0.005
  - Pairwise comparisons across 20.000 genes ☹

- **FDR**: **Benjamini-Hochberg**: the procedure controls the expected proportion of "discoveries" (significant tests) that are false (false positive).
  - Less stringent control of Type I Error than FWER procedures which control the probability of at least one Type I Error
  - More power at the cost of increased numbers of Type I Errors.

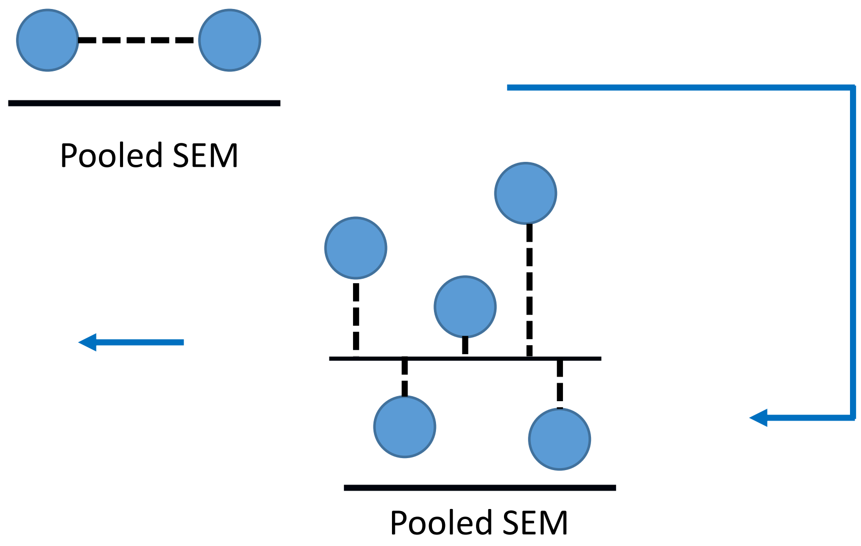
- **Difference between FWER and FDR**:
  - a p-value of 0.05 implies that 5% of all tests will result in false positives.
  - a FDR adjusted p-value (or **q-value**) of 0.05 implies that 5% of significant tests will result in false positives.

Analysis of variance

- Extension of the 2 groups comparison of a *t*-test but with a slightly different logic:

- $t\text{-test} = \dfrac{\text{mean1} - \text{mean2}}{\text{Pooled SEM}}$

- $\text{ANOVA} = \dfrac{\text{variance between means}}{\text{Pooled SEM}}$



- ANOVA compares variances:
  - If variance between the several means > variance within the groups (random error) then the means must be more spread out than it would have been by chance.

Analysis of variance

- The statistic for ANOVA is the **F ratio**.

- F = $\dfrac{\text{Variance between the groups}}{\text{Variance within the groups (individual variability)}}$

- F = $\dfrac{\text{Variation explained by the model (= systematic)}}{\text{Variation explained by unsystematic factors (= random variation)}}$

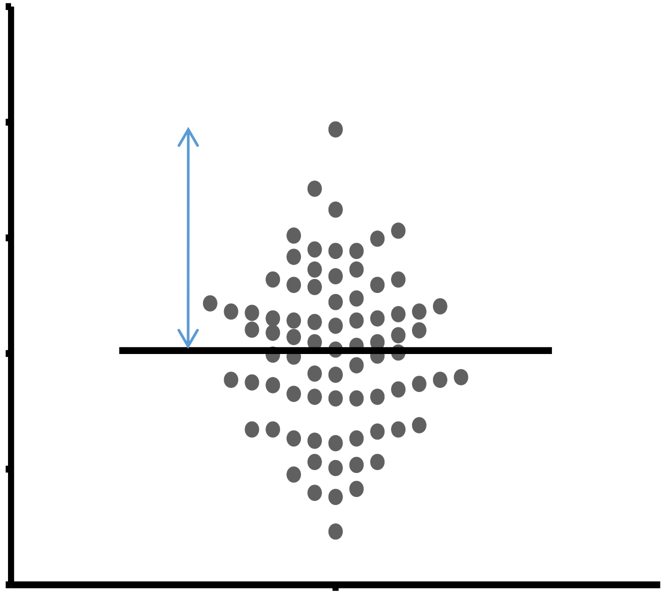
- If the variance amongst sample means is greater than the error/random variance, then F>1
  - In an ANOVA, we test whether F is significantly higher than 1 or not.

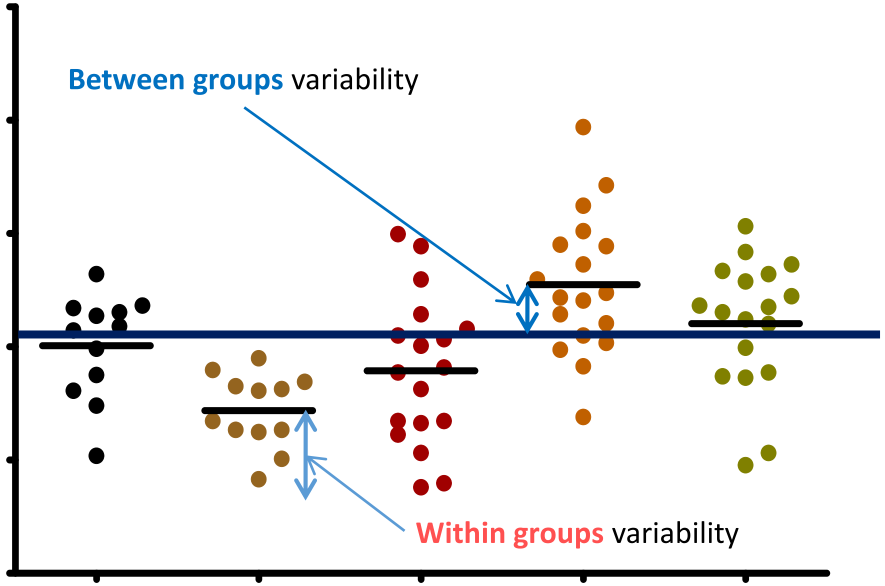
Analysis of variance

Source of variation	Sum of Squares	df	Mean Square	F	p-value
Between Groups	2.665	4	0.6663	8.423	<0.0001
Within Groups	5.775	73	0.0791		
Total	8.44	77			

In Power Analysis:
Pooled SD=√MS(Residual)

- Variance (= SS / N-1) is the mean square
  - df: degree of freedom with df = N-1


Total sum of squares



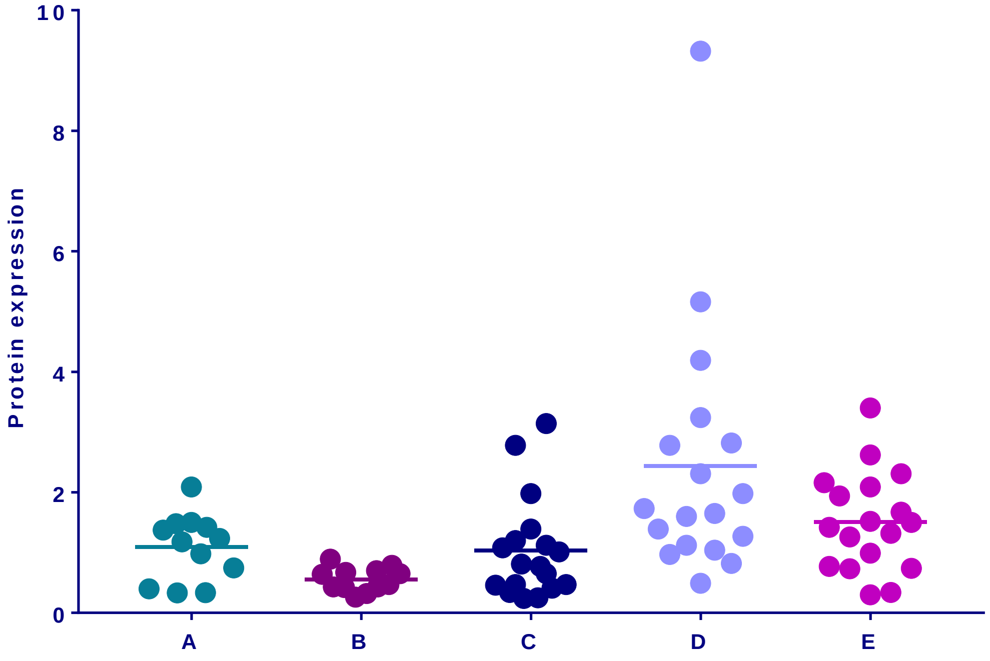
Exercise 8: One-way ANOVA

protein expression.xlsx

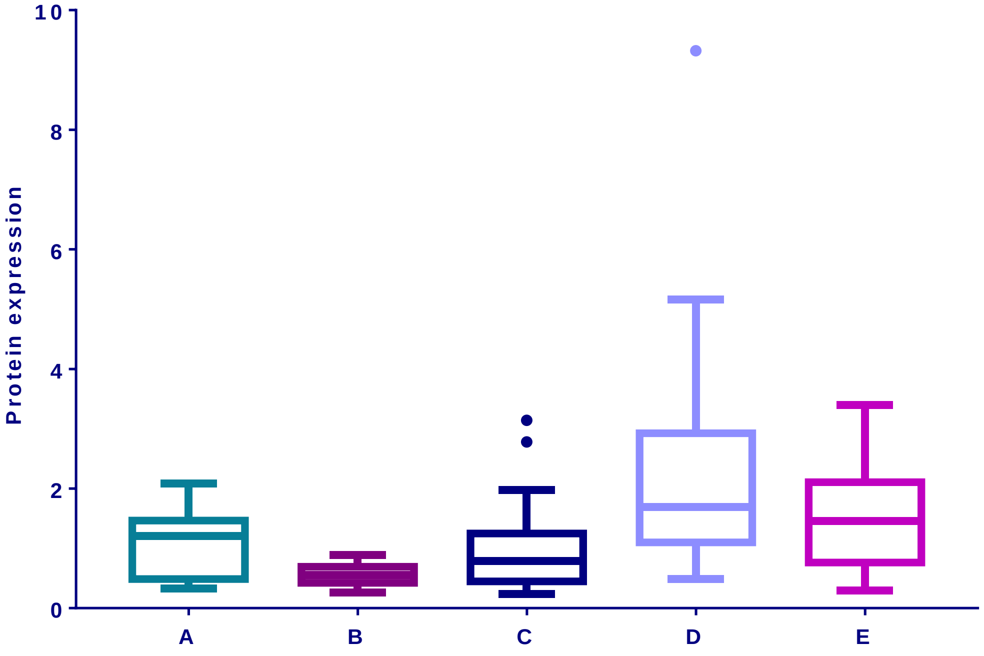
- **Question**: is there a difference in protein expression between the 5 cell lines?

- **1 Plot the data**

- **2 Check the assumptions for parametric test**



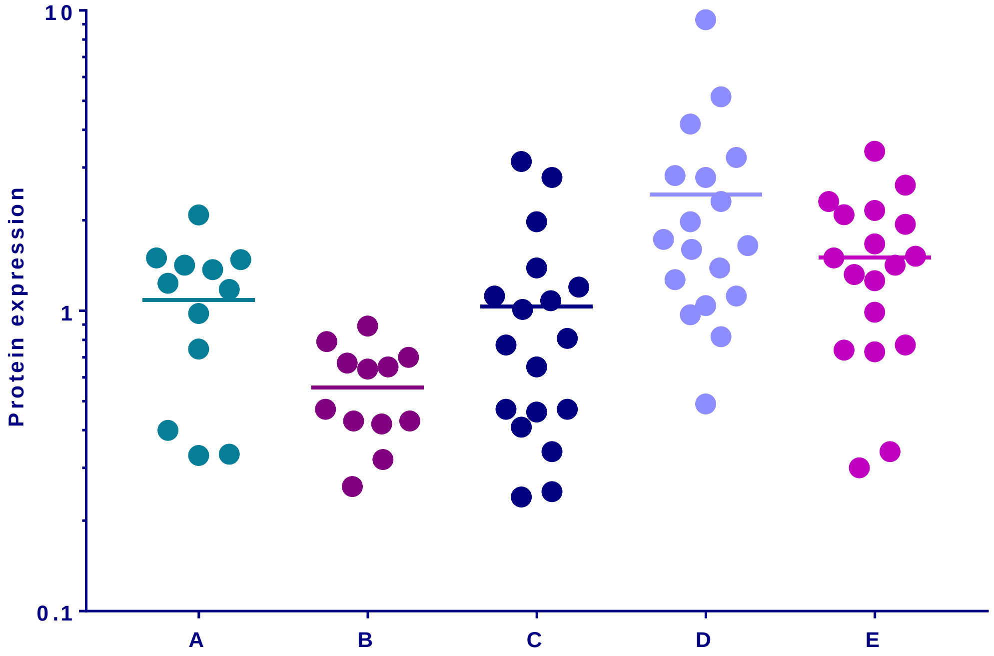
Protein expression
10
8
6
4
2
0
A
B
C
D
E

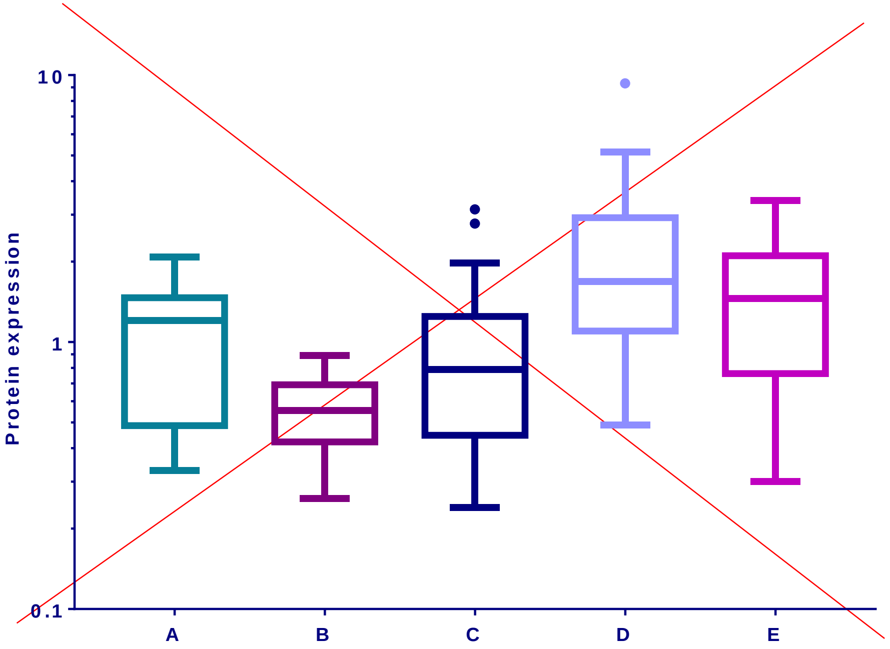


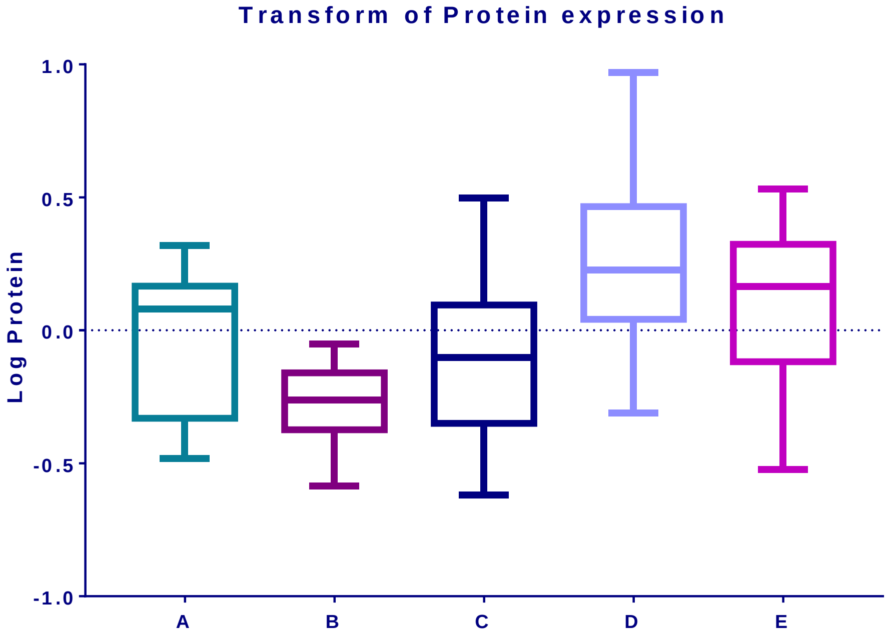
Protein expression
10
8
6
4
2
0
A
B
C
D
E

Parametric tests assumptions

	Col. stats	A	B	C	D	E
		A	B	C	D	E
1	Number of values	12	12	18	18	18
2						
3	Minimum	0.3300	0.2600	0.2400	0.4900	0.3000
4	25% Percentile	0.4864	0.4225	0.4475	1.100	0.7625
5	Median	1.206	0.5550	0.7900	1.690	1.460
6	75% Percentile	1.465	0.6925	1.248	2.925	2.108
7	Maximum	2.088	0.8900	3.140	9.320	3.400
8						
9	Mean	1.088	0.5558	1.032	2.438	1.504
10	Std. Deviation	0.5469	0.1947	0.8364	2.108	0.8179
11	Std. Error of Mean	0.1579	0.05620	0.1971	0.4968	0.1928
12						
13	Lower 95% CI of mean	0.7409	0.4321	0.6157	1.390	1.098
14	Upper 95% CI of mean	1.436	0.6795	1.448	3.486	1.911
15						
16	Sum	13.06	6.670	18.57	43.88	27.08
17						
18	**D'Agostino & Pearson normality test**					
19	K2	0.1236	0.7508	9.375	22.59	1.280
20	P value	0.9401	0.6870	0.0092	<0.0001	0.5274
21	Passed normality test (alpha=0.05)?	Yes	Yes	No	No	Yes
22	P value summary	ns	ns	**	****	ns
23						


10
1
0.1
Protein expression
A
B
C
D
E


10
1
0.1
Protein expression
A
B
C
D
E


Transform of Protein expression
1.0
0.5
0.0
-0.5
-1.0
Log Protein
A
B
C
D
E

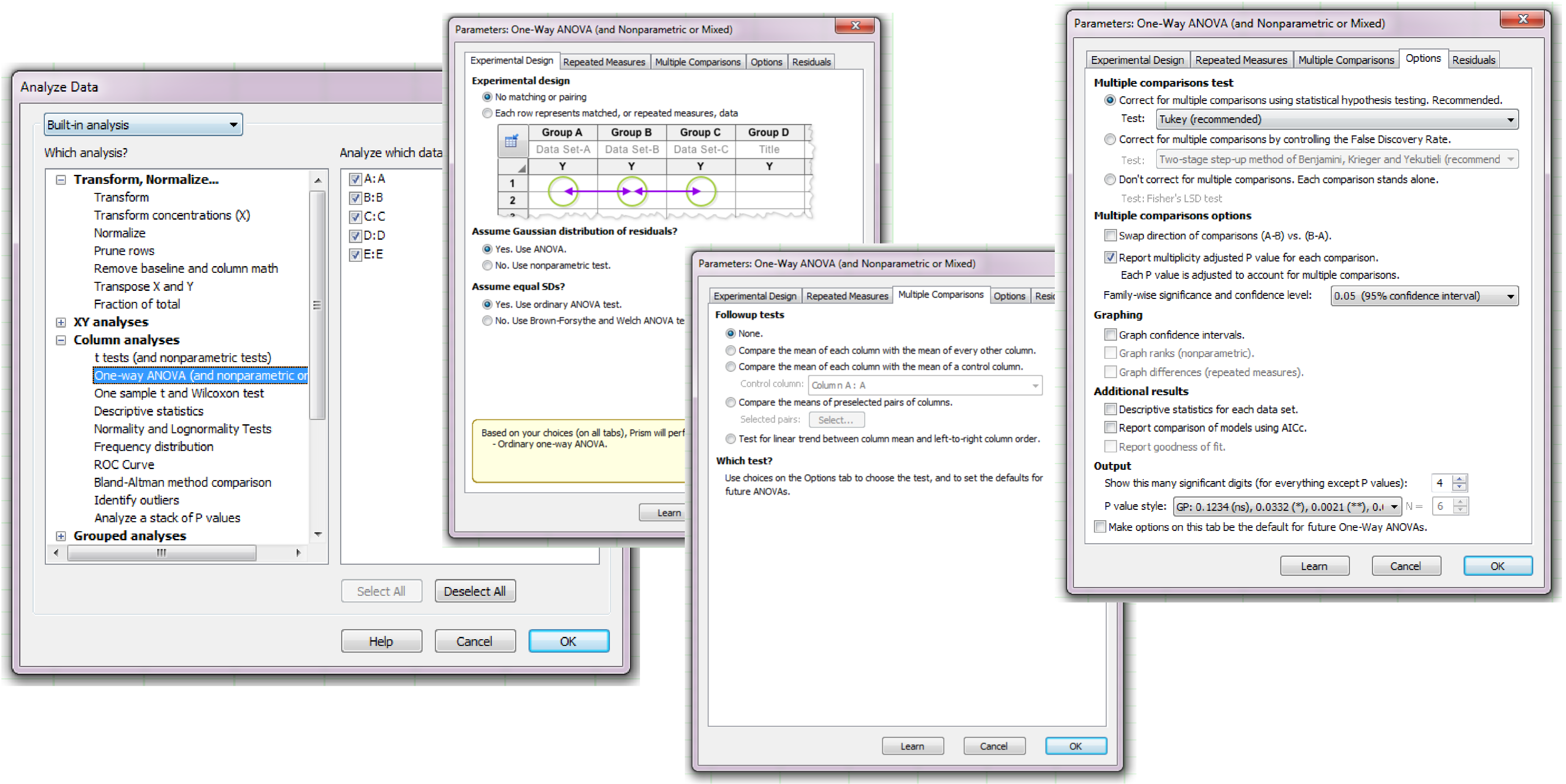
Parametric tests assumptions

Col. stats	A	B	C	D	E
	A	B	C	D	E
1 Number of values	12	12	18	18	18
2					
3 Minimum	-0.4815	-0.5850	-0.6198	-0.3098	-0.5229
4 25% Percentile	-0.3303	-0.3742	-0.3497	0.04117	-0.1178
5 Median	0.08140	-0.2609	-0.1025	0.2278	0.1642
6 75% Percentile	0.1659	-0.1597	0.09514	0.4653	0.3237
7 Maximum	0.3196	-0.05061	0.4969	0.9694	0.5315
8					
9 Mean	-0.03123	-0.2817	-0.1064	0.2740	0.1018
10 Std. Deviation	0.2764	0.1632	0.3307	0.3112	0.2873
11 Std. Error of Mean	0.07978	0.04711	0.07796	0.07336	0.06772
12					
13 Lower 95% CI of mean	-0.2068	-0.3854	-0.2709	0.1193	-0.04104
14 Upper 95% CI of mean	0.1444	-0.1780	0.05803	0.4288	0.2447
15					
16 Sum	-0.3747	-3.380	-1.916	4.933	1.833
17					
18 **D'Agostino & Pearson normality test**					
19 K2	2.037	0.6827	0.5884	0.8869	2.902
20 P value	0.3611	0.7108	0.7451	0.6418	0.2344
21 Passed normality test (alpha=0.05)?	Yes	Yes	Yes	Yes	Yes
22 P value summary	ns	ns	ns	ns	ns

Analysis of variance: Post hoc tests

- The ANOVA is an "omnibus" test: it tells you that there is (or not) a difference between your means but not exactly which means are significantly different from which other ones.
  - To find out, you need to apply **post hoc** tests.
  - These post hoc tests should only be used when the ANOVA finds a significant effect.

One-Way Analysis of variance



Analysis of variance: results

Ordinary one-way ANOVA
ANOVA results

1	Table Analyzed	Transform of Protein expression				
2	Data sets analyzed	A-E				
3						
4	**ANOVA summary**					
5	F	8.127				
6	P value	<0.0001				
7	P value summary	****				
8	Significant diff. among means (P < 0.05)?	Yes				
9	R square	0.3081				
10						
11	**Brown-Forsythe test**					
12	F (DFn, DFd)	0.9831 (4, 73)				
13	P value	0.4222				
14	P value summary	ns				
15	Are SDs significantly different (P < 0.05)?	No				
16						
17	**Bartlett's test**					
18	Bartlett's statistic (corrected)	5.829				
19	P value	0.2123				
20	P value summary	ns				
21	Are SDs significantly different (P < 0.05)?	No				
22						
23	**ANOVA table**	**SS**	**DF**	**MS**	**F (DFn, DFd)**	**P value**
24	Treatment (between columns)	2.691	4	0.6727	F (4, 73) = 8.127	P<0.0001
25	Residual (within columns)	6.043	73	0.08278		
26	Total	8.734	77			
27						
28	**Data summary**					
29	Number of treatments (columns)	5				
30	Number of values (total)	78				

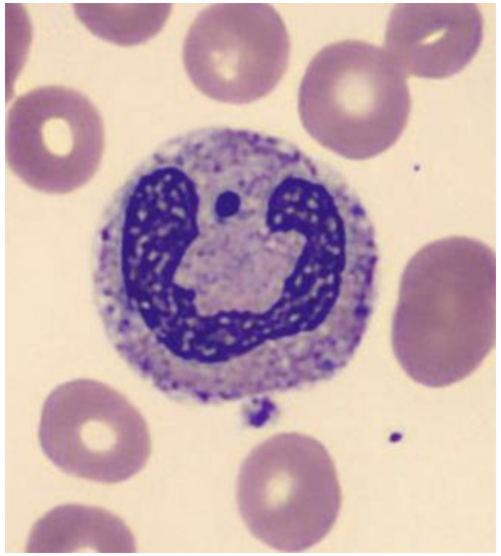
Homogeneity of variance ☑

F=0.6727/0.08278=8.13

Ordinary one-way ANOVA
Multiple comparisons

1	Number of families	1							
2	Number of comparisons per family	10							
3	Alpha	0.05							
4									
5	**Tukey's multiple comparisons test**	**Mean Diff.**	**95.00% CI of diff.**	**Significant?**	**Summary**	**Adjusted P Value**			
6	A vs. B	0.2505	-0.07808 to 0.5790	No	ns	0.2177	A-B		
7	A vs. C	0.07521	-0.2247 to 0.3751	No	ns	0.9555	A-C		
8	A vs. D	-0.3053	-0.6052 to -0.005359	Yes	*	0.0440	A-D		
9	A vs. E	-0.1331	-0.4330 to 0.1669	No	ns	0.7275	A-E		
10	B vs. C	-0.1753	-0.4752 to 0.1247	No	ns	0.4807	B-C		
11	B vs. D	-0.5557	-0.8557 to -0.2558	Yes	****	<0.0001	B-D		
12	B vs. E	-0.3835	-0.6834 to -0.08360	Yes	**	0.0055	B-E		
13	C vs. D	-0.3805	-0.6487 to -0.1122	Yes	**	0.0015	C-D		
14	C vs. E	-0.2083	-0.4765 to 0.05998	No	ns	0.2021	C-E		
15	D vs. E	0.1722	-0.09604 to 0.4405	No	ns	0.3839	D-E		
16									
17	**Test details**	**Mean 1**	**Mean 2**	**Mean Diff.**	**SE of diff.**	**n1**	**n2**	**q**	**DF**
18	A vs. B	-0.03123	-0.2817	0.2505	0.1175	12	12	3.016	73
19	A vs. C	-0.03123	-0.1064	0.07521	0.1072	12	18	0.9920	73
20	A vs. D	-0.03123	0.2740	-0.3053	0.1072	12	18	4.026	73
21	A vs. E	-0.03123	0.1018	-0.1331	0.1072	12	18	1.755	73
22	B vs. C	-0.2817	-0.1064	-0.1753	0.1072	12	18	2.311	73
23	B vs. D	-0.2817	0.2740	-0.5557	0.1072	12	18	7.330	73
24	B vs. E	-0.2817	0.1018	-0.3835	0.1072	12	18	5.058	73
25	C vs. D	-0.1064	0.2740	-0.3805	0.09590	18	18	5.611	73
26	C vs. E	-0.1064	0.1018	-0.2083	0.09590	18	18	3.071	73
27	D vs. E	0.2740	0.1018	0.1722	0.09590	18	18	2.540	73
28									

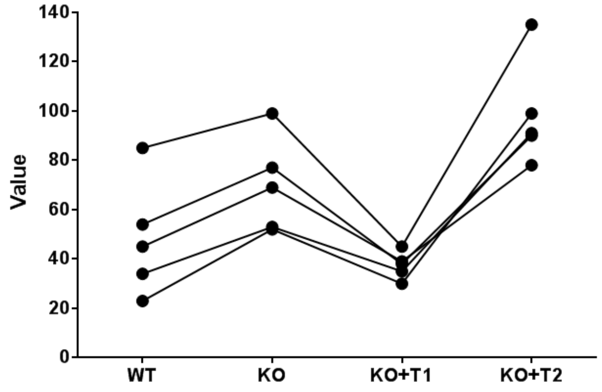
Exercise 9: neutrophils.xlsx

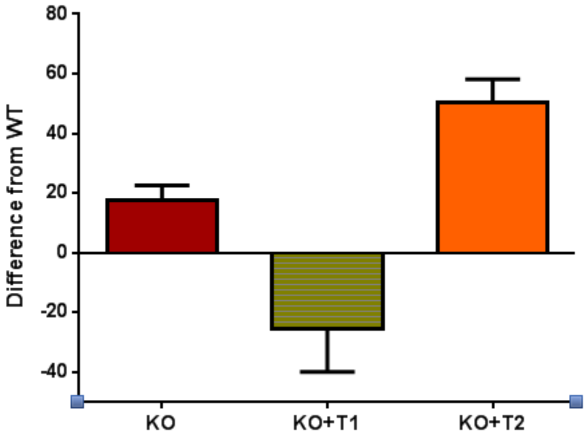


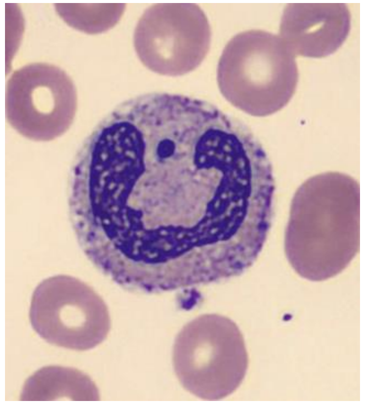
- A researcher is looking at the difference between 4 cell groups. He has run the experiment 5 times. Within each experiment, he has neutrophils from a WT (control), a KO, a KO+Treatment 1 and a KO+Treatment2.

- **Question**: Is there a difference between KO with/without treatment and WT?

Exercise 9: neutrophils.xlsx







	ANOVA					
1	Table Analyzed	Repeated measures one-way ANOVA data2				
2						
3	Repeated measures ANOVA summary					
4	Assume sphericity?	No				
5	F	28.57				
6	P value	0.0002				
7	P value summary	***				
8	Statistically significant (P < 0.05)?	Yes				
9	Geisser-Greenhouse's epsilon	0.6916				
10	R square	0.8772				
11						
12	Was the matching effective?					
13	F	8.239				
14	P value	0.0020				
15	P value summary	**				
16	Is there significant matching (P < 0.05)?	Yes				
17	R square	0.2522				
18						
19	ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
20	Treatment (between columns)	10948	3	3649	F (2.075, 8.299) = 28.57	P = 0.0002
21	Individual (between rows)	4209	4	1052	F (4, 12) = 8.239	P = 0.0020
22	Residual (random)	1533	12	127.7		
23	Total	16689	19			
24						

Dunnett's multiple comparisons test	Mean Diff.	95% CI of diff.	Significant?	Summary	Adjusted P Value	A-?	
WT vs. KO	-21.80	-30.91 to -12.69	Yes	**	0.0022	B	KO
WT vs. KO+T1	10.80	-19.02 to 40.62	No	ns	0.4941	C	KO+T1
WT vs. KO+T2	-50.40	-78.53 to -22.27	Yes	**	0.0067	D	KO+T2

Answer: There is a significant difference from WT for the first and third groups.

Comparison between more than 2 groups

One factor

What about power analysis?

Babraham Bioinformatics

Comparison of more than 2 means

- Different ways to go about power analysis in the context of ANOVA:
  - η^2 : explained proportion variance of the total variance.
    - Can be translated into effect size d.
    - Not very useful: only looking at the omnibus part of the test
  - Minimum power specification: looks at the difference between the smallest and the biggest means.
    - All means other than the 2 extreme one are equal to the grand mean.
  - Smallest meaningful difference
    - Works like a post-hoc test.

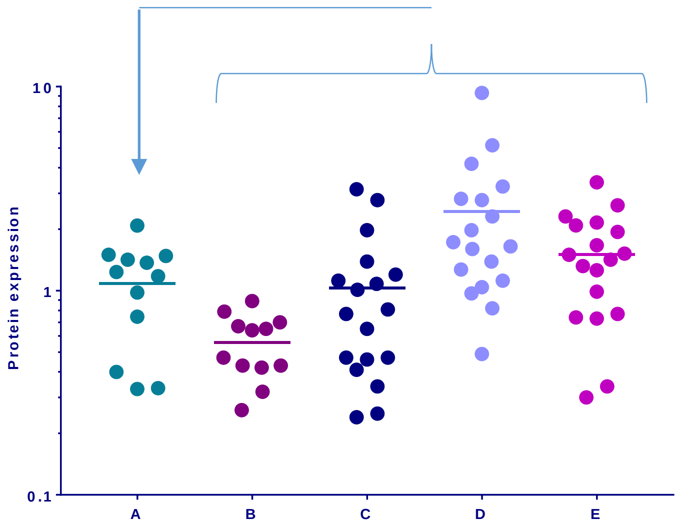
Power Analysis
Comparing more than 2 means

- Research example: Comparison between 4 teaching methods
- Smallest meaningful difference
  - Same assumptions:
    - Equal group sizes and equal variability (SD = 80)
  - 3 comparisons of interest: vs. Group 1
  - Smallest meaningful difference: group 1 vs. Group 2
    - t-test: Mean 1 = 550, SD = 80 and mean 2 = 598, SD = 80
    - Power calculation like for a t-test but with a Bonferroni correction (adjustment for multiple comparisons)

Power Analysis

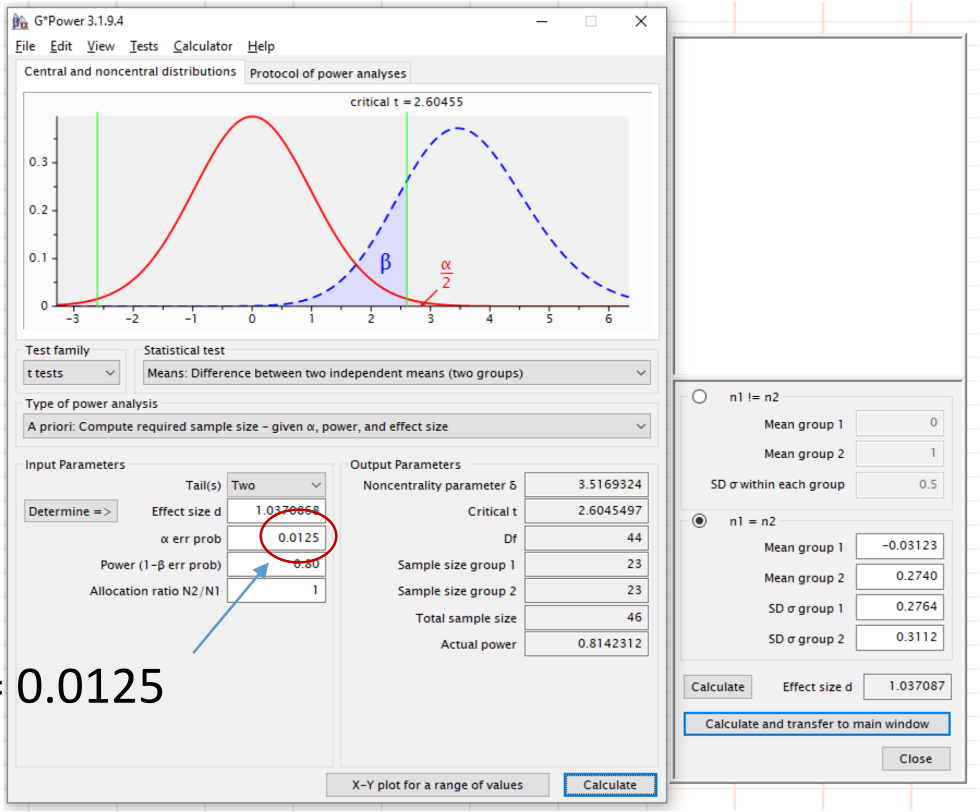
Comparing more than 2 means

- Smallest meaningful difference
  - Power calculation like for a t-test but with a Bonferroni correction.
  - Protein expression example:
    - Comparisons vs. cell line A.
    - Meaningful difference: D vs. A



<u>Bonferroni correction</u>

3 comparisons: 0.05/4 = 0.0125



Comparison between more than 2 groups

One factor

Non-Parametric data

Babraham Bioinformatics

Non Parametric approach:
Kruskal-Wallis

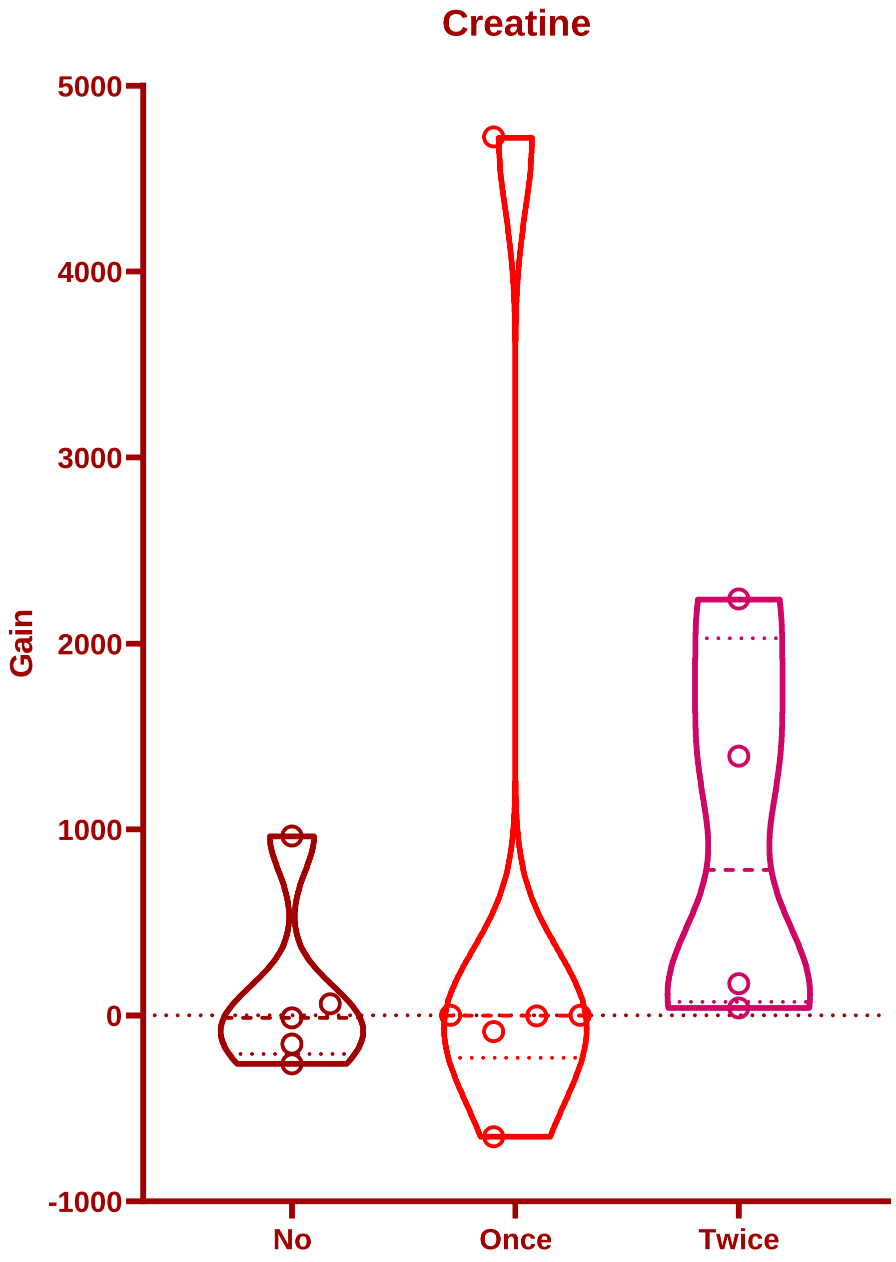
- Non-parametric equivalent of the one-way ANOVA
- It is a test based on ranks
- `kruskal.wallis()` produces omnibus part of the analysis
- Post-hoc test associated with Kruskal-Wallis: **Dunn test**
- `dunn.test()` gives both Kruskall-Wallis and pairwise comparisons results ## dunn.test package ##
- Statistic associated with Kruskal-Wallis is H and it has a Chi^2 distribution
- The Dunn test works pretty much like the Mann-Whitney test.

Exercise 10: creatine.xlsx

- Creatine, a supplement popular among body builders
- Three groups: No creatine; Once a day; and Twice a day.
- **Question**: does the average weight gain depend on the creatine group to which people were assigned?

Exercise 10: creatine.xlsx

	Kruskal-Wallis test ANOVA results	
1	Table Analyzed	Creatine
2		
3	**Kruskal-Wallis test**	
4	P value	0.1458
5	Exact or approximate P value?	Exact
6	P value summary	ns
7	Do the medians vary signif. (P < 0.05)?	No
8	Number of groups	3
9	Kruskal-Wallis statistic	3.868
10		
11	**Data summary**	
12	Number of treatments (columns)	3
13	Number of values (total)	15
14		



Comparison between more than 2 groups

Two factors

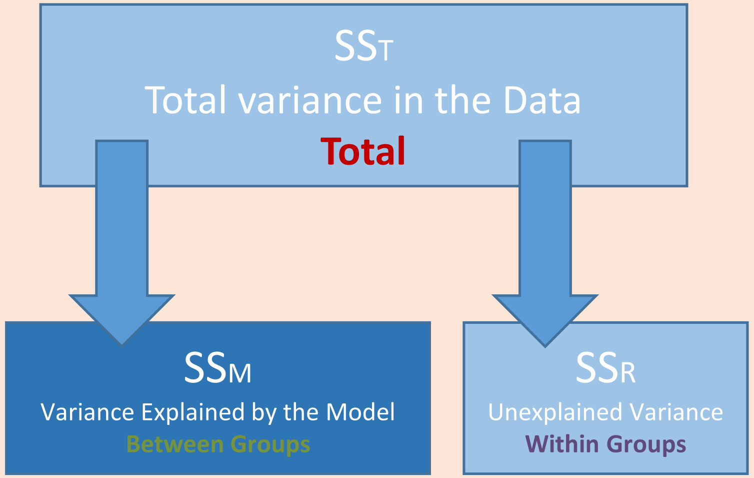
Babraham Bioinformatics

Two-way Analysis of Variance (Factorial ANOVA)

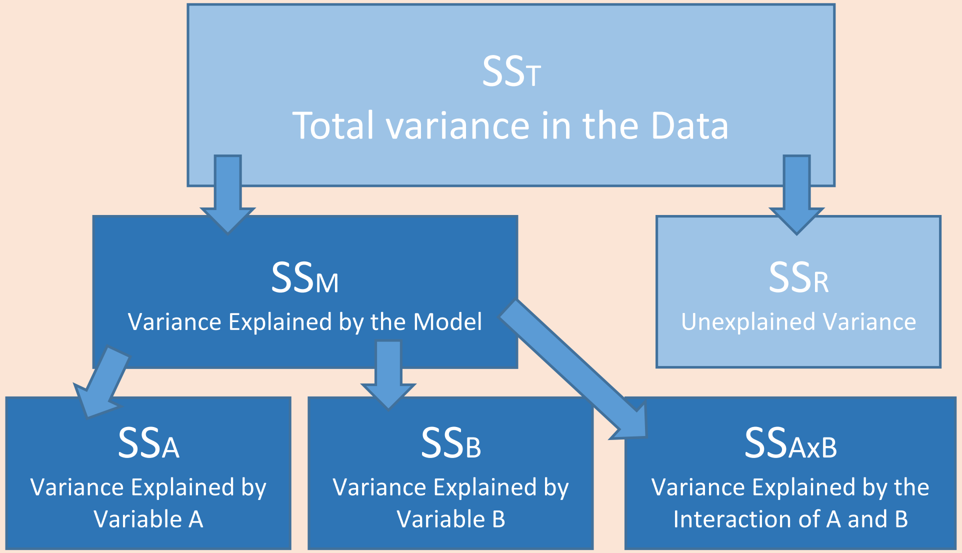
Source of variation	Sum of Squares	Df	Mean Square	F	p-value
Variable A (Between Groups)	2.665	4	0.6663	8.42	<0.0001
Within Groups (Residual)	5.775	73	0.0791		
Total	8.44	77			

Source of variation	Sum of Squares	Df	Mean Square	F	p-value
Variable A * Variable B	1978	2	989.1	F (2, 42) = 11.91	P < 0.0001
Variable B (Between groups)	3332	2	1666	F (2, 42) = 20.07	P < 0.0001
Variable A (Between groups)	168.8	1	168.8	F (1, 42) = 2.032	P = 0.1614
Residuals	3488	42	83.04		

One-way ANOVA= 1 predictor variable



2-way ANOVA= 2 predictor variables: A and B



Two-way Analysis of Variance

Alcohol	None		2 Pints		4 Pints	
Gender	Female	Male	Female	Male	Female	Male
	65	50	70	55	45	30
	70	55	65	65	60	30
	60	80	60	70	85	30
	60	65	70	55	65	55
	60	70	65	55	70	35
	55	75	60	60	70	20
	60	75	60	50	80	45
	55	65	50	50	60	40

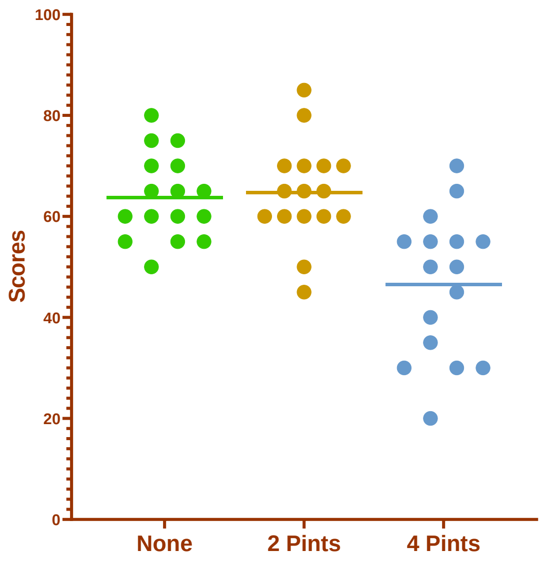
Example: goggles.xlsx

- The 'beer-goggle' effect
  - The term refers to finding people more attractive after you've had a few beers. Drinking beer provides a warm, friendly sensation, lowers your inhibitions, and helps you relax.

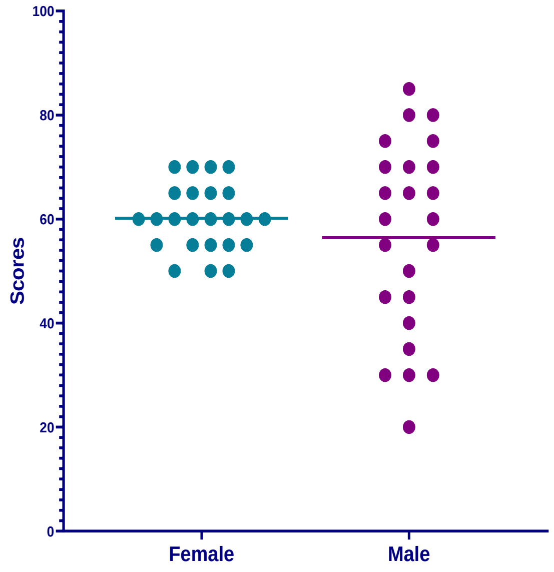
- Study: effects of alcohol on mate selection in night-clubs.
- Pool of independent judges scored the levels of attractiveness of the person that the participant was chatting up at the end of the evening.
- **Question**: is subjective perception of physical attractiveness affected by alcohol consumption?
  - Attractiveness on a scale from 0 to 100

Two-way Analysis of Variance

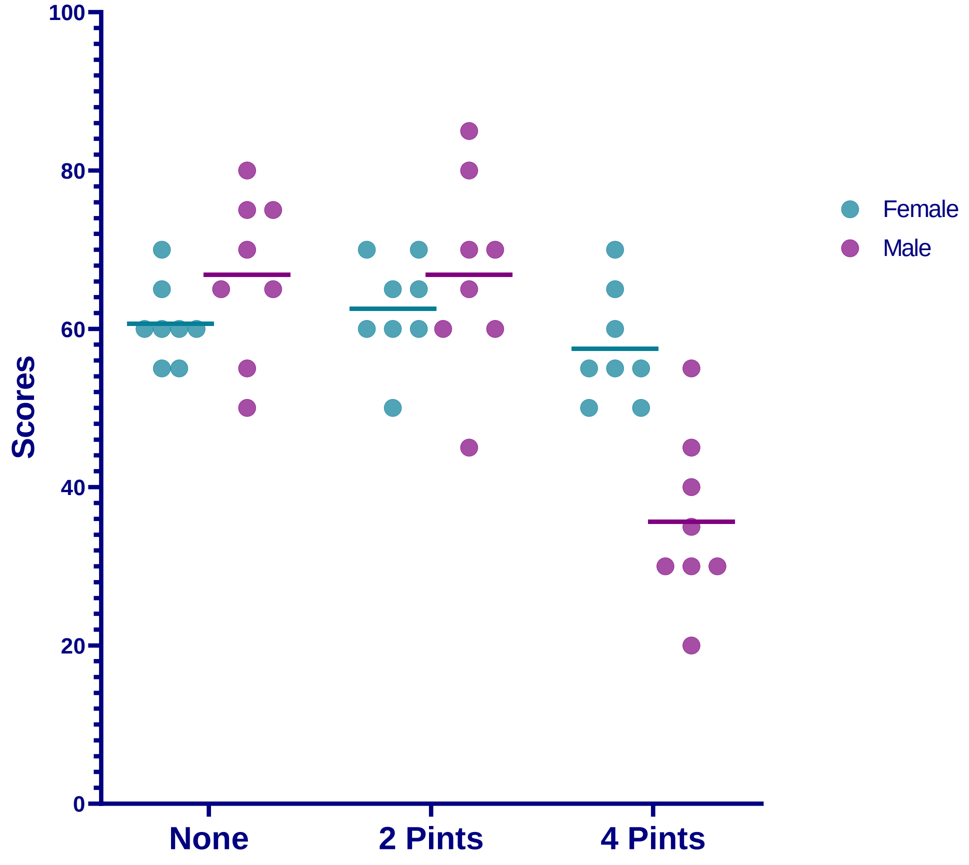
Main effect of Alcohol



Main effect of Gender



Interaction
between Alcohol and Gender



Two-way Analysis of Variance

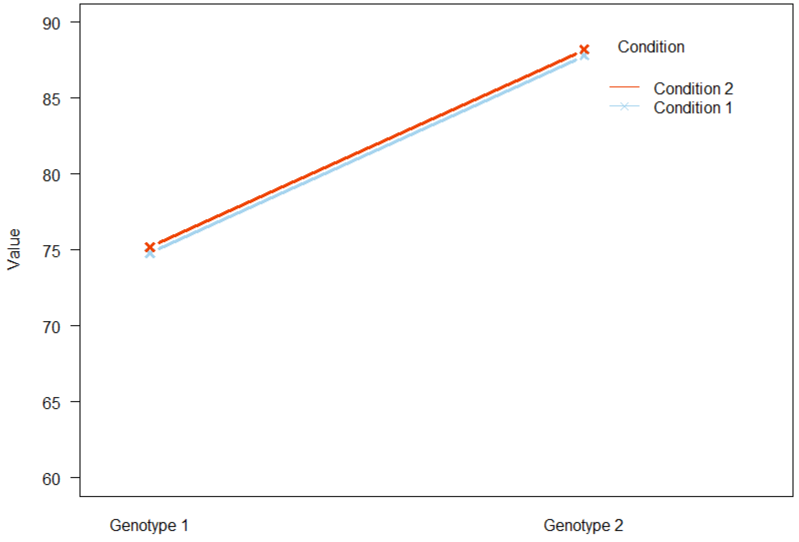
- **Interaction plots: Examples**
  - Fake dataset:
    - 2 factors: **Genotype** (2 levels) and **Condition** (2 levels)

Genotype	Condition	Value
Genotype 1	Condition 1	74.8
Genotype 1	Condition 1	65
Genotype 1	Condition 1	74.8
Genotype 1	Condition 2	75.2
Genotype 1	Condition 2	75
Genotype 1	Condition 2	75.2
Genotype 2	Condition 1	87.8
Genotype 2	Condition 1	65
Genotype 2	Condition 1	74.8
Genotype 2	Condition 2	88.2
Genotype 2	Condition 2	75
Genotype 2	Condition 2	75.2

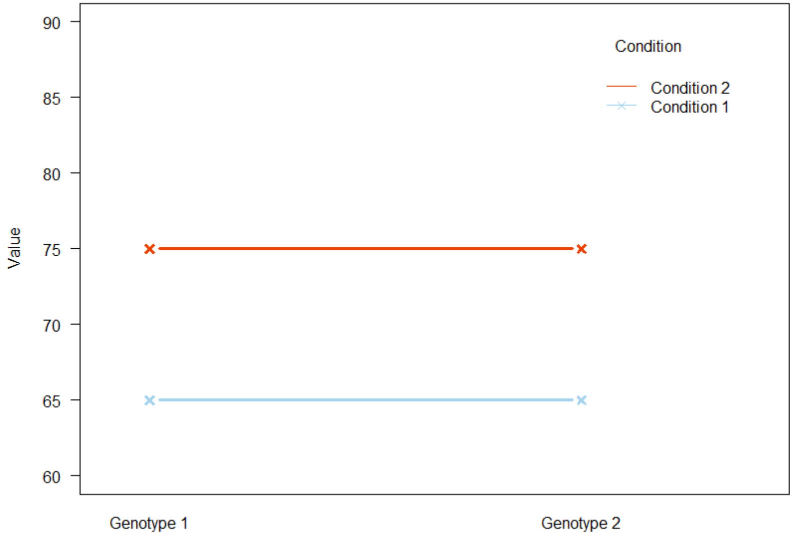
Two-way Analysis of Variance

- <u>Interaction plots</u>: **Examples**
  - <u>2 factors</u>: **Genotype** (2 levels) and **Condition** (2 levels)

<u>Single Effect</u>



Genotype Effect

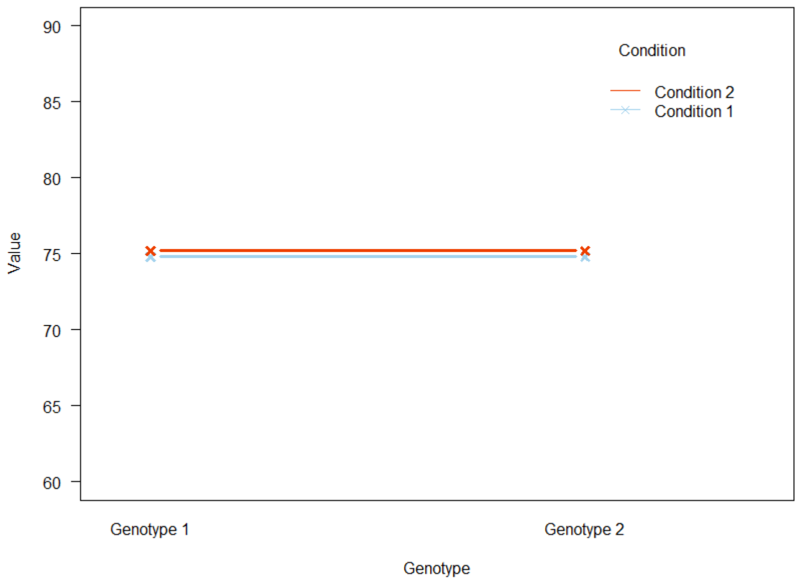


Condition Effect

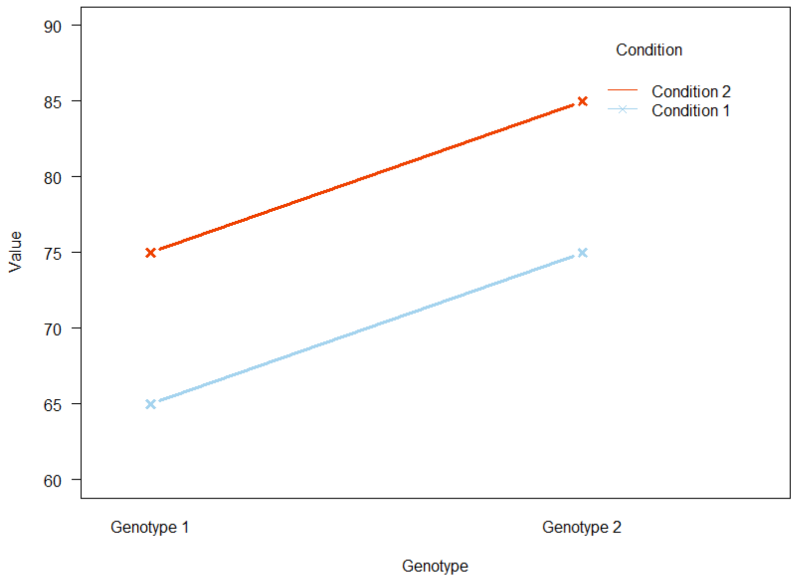
Two-way Analysis of Variance

- **Interaction plots: Examples**
  - 2 factors: **Genotype** (2 levels) and **Condition** (2 levels)

Zero or Both Effect



Zero Effect

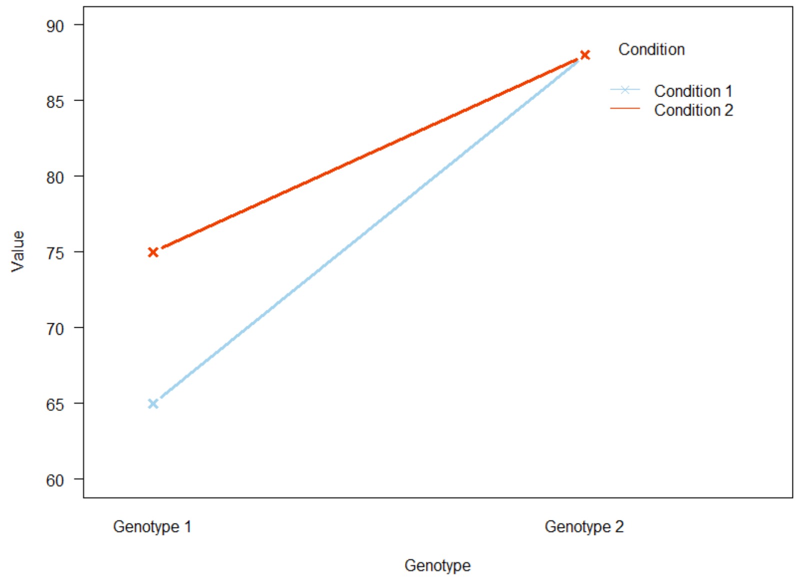


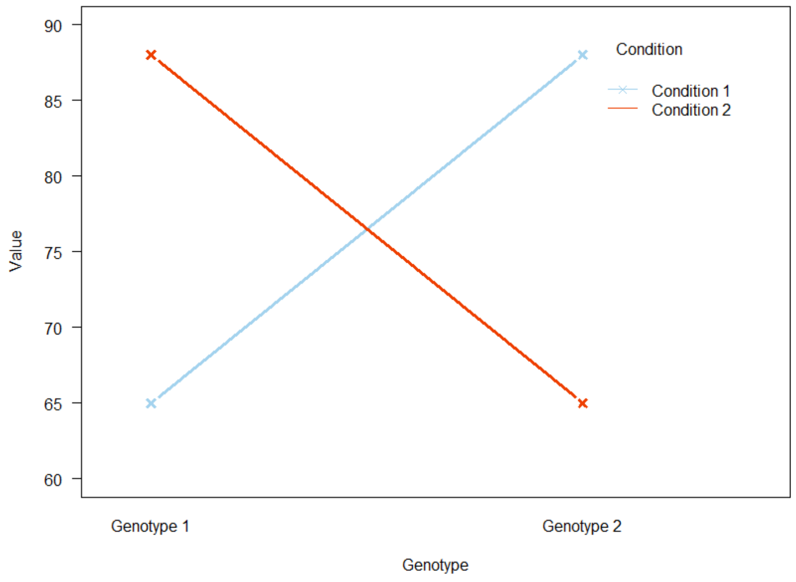
Both Effect

Two-way Analysis of Variance

- **<u>Interaction plots</u>: Examples**
  - <u>2 factors</u>: **Genotype** (2 levels) and **Condition** (2 levels)

Interaction

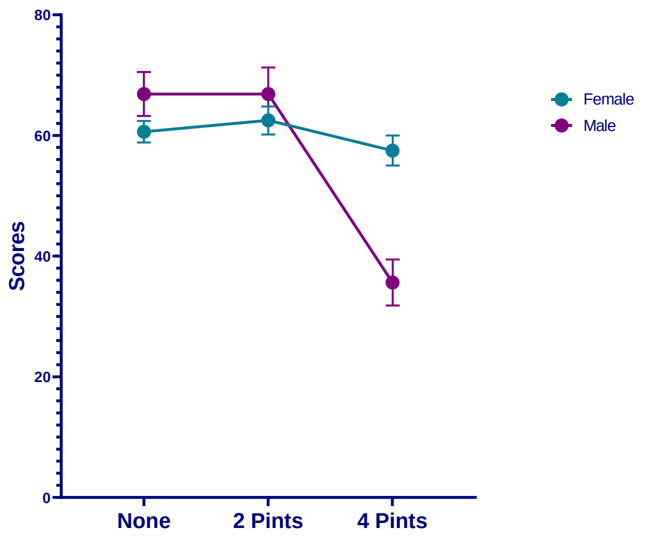




Two-way Analysis of Variance

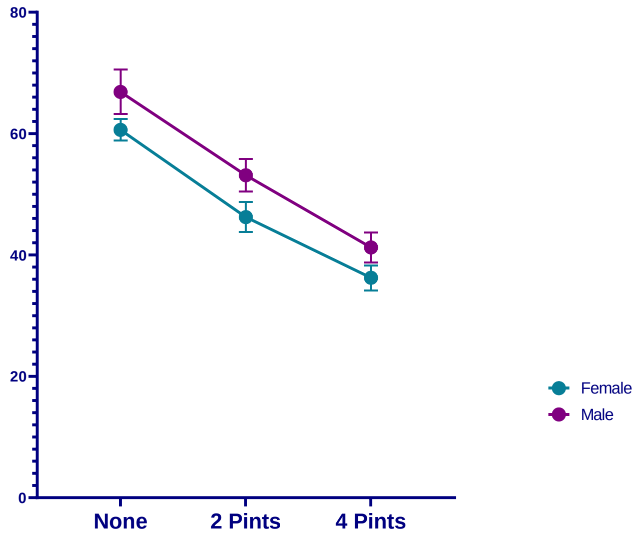
With significant interaction (real data)

ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
Interaction	**1978**	**2**	**989.1**	**F (2, 42) = 11.91**	**< 0.0001**
Alcohol Consumption	3332	2	1666	F (2, 42) = 20.07	< 0.0001
Gender	168.8	1	168.8	F (1, 42) = 2.032	0.1614
Residual	3488	42	83.04		

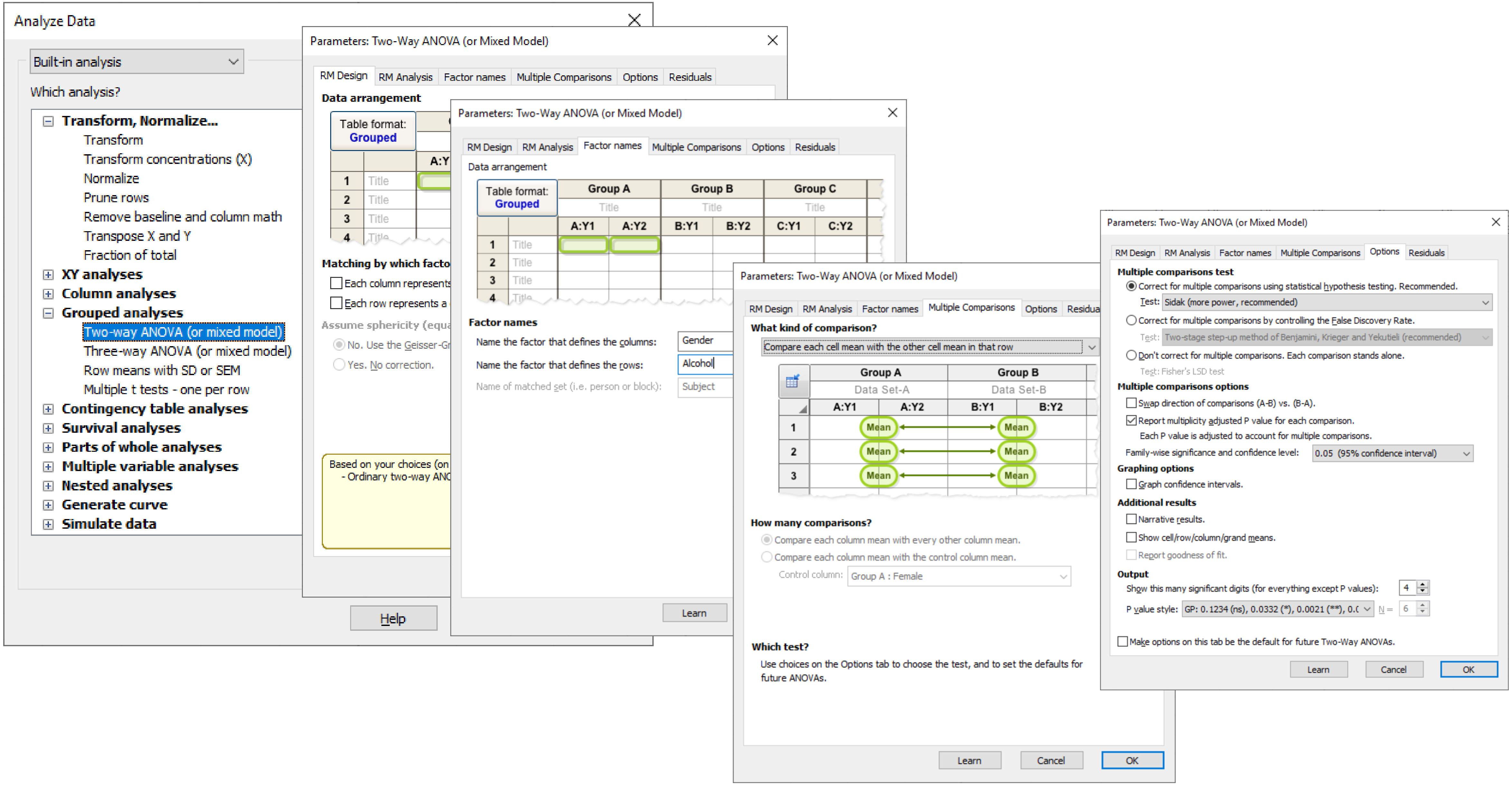


Without significant interaction (fake data)

ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
Interaction	7.292	2	3.646	F (2, 42) = 0.06872	0.9337
Alcohol Consumption	**5026**	**2**	**2513**	**F (2, 42) = 47.37**	**< 0.0001**
Gender	**438.0**	**1**	**438.0**	**F (1, 42) = 8.257**	**0.0063**
Residual	2228	42	53.05		



Two-way Analysis of Variance

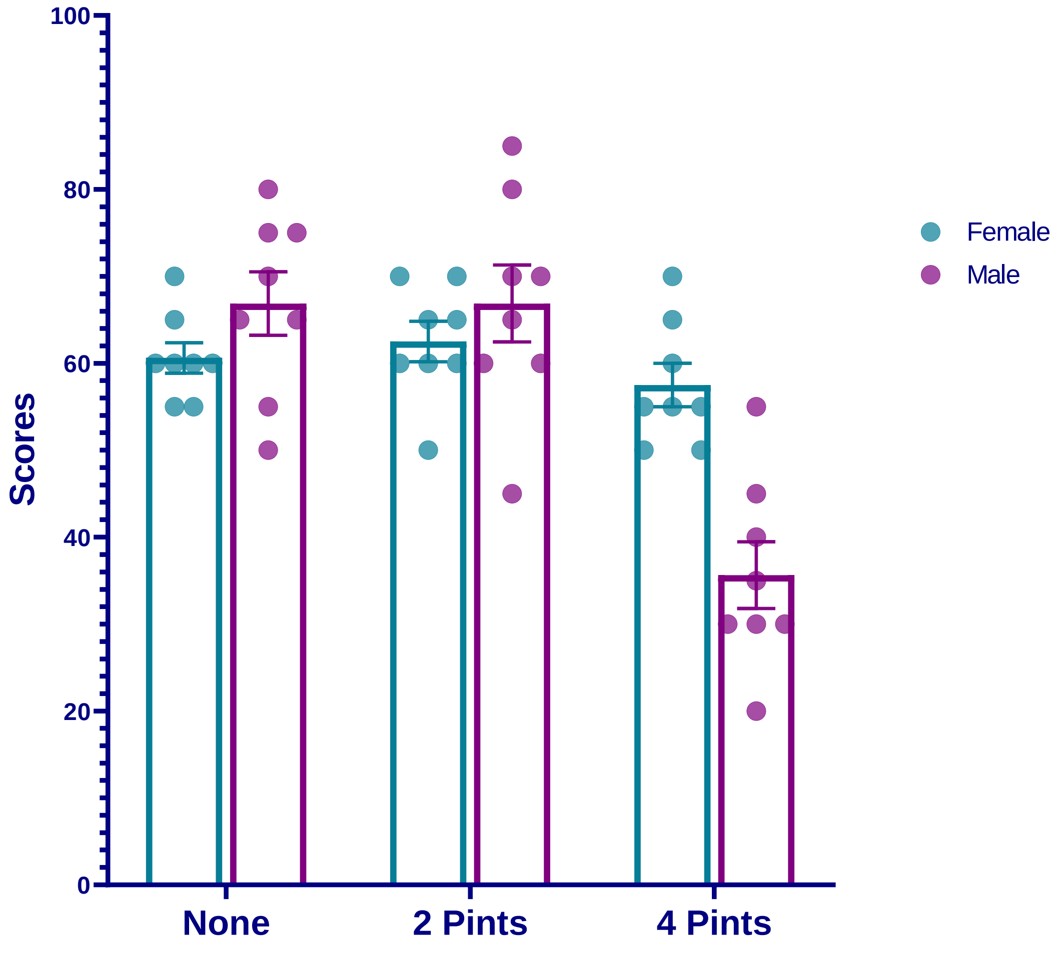


Two-way Analysis of Variance

2way ANOVA
ANOVA results

1	Table Analyzed	data for 2-way				
2						
3	**Two-way ANOVA**	Ordinary				
4	Alpha	0.05				
5						
6	**Source of Variation**	**% of total variation**	**P value**	**P value summary**	**Significant?**	
7	Interaction	22.06	<0.0001	****	Yes	
8	Alcohol Consumption	37.16	<0.0001	****	Yes	
9	Gender	1.882	0.1614	ns	No	
10						
11	**ANOVA table**	**SS**	**DF**	**MS**	**F (DFn, DFd)**	**P value**
12	Interaction	1978	2	989.1	F (2, 42) = 11.91	P<0.0001
13	Alcohol Consumption	3332	2	1666	F (2, 42) = 20.07	P<0.0001
14	Gender	168.8	1	168.8	F (1, 42) = 2.032	P=0.1614
15	Residual	3488	42	83.04		
16						

Tukey's multiple comparisons test	Mean Diff.	95.00% CI of diff.	Significant?	Summary	Adjusted P Value
None:Female vs. None:Male	-6.250	-19.85 to 7.351	No	ns	0.7432
None:Female vs. 2 Pints:Female	-1.875	-15.48 to 11.73	No	ns	0.9984
None:Female vs. 2 Pints:Male	-6.250	-19.85 to 7.351	No	ns	0.7432
None:Female vs. 4 Pints:Female	3.125	-10.48 to 16.73	No	ns	0.9826
None:Female vs. 4 Pints:Male	25.00	11.40 to 38.60	Yes	****	<0.0001
None:Male vs. 2 Pints:Female	4.375	-9.226 to 17.98	No	ns	0.9278
None:Male vs. 2 Pints:Male	0.000	-13.60 to 13.60	No	ns	>0.9999
None:Male vs. 4 Pints:Female	9.375	-4.226 to 22.98	No	ns	0.3287
None:Male vs. 4 Pints:Male	31.25	17.65 to 44.85	Yes	****	<0.0001
2 Pints:Female vs. 2 Pints:Male	-4.375	-17.98 to 9.226	No	ns	0.9278
2 Pints:Female vs. 4 Pints:Female	5.000	-8.601 to 18.60	No	ns	0.8796
2 Pints:Female vs. 4 Pints:Male	26.88	13.27 to 40.48	Yes	****	<0.0001
2 Pints:Male vs. 4 Pints:Female	9.375	-4.226 to 22.98	No	ns	0.3287
2 Pints:Male vs. 4 Pints:Male	31.25	17.65 to 44.85	Yes	****	<0.0001
4 Pints:Female vs. 4 Pints:Male	21.88	8.274 to 35.48	Yes	***	0.0003

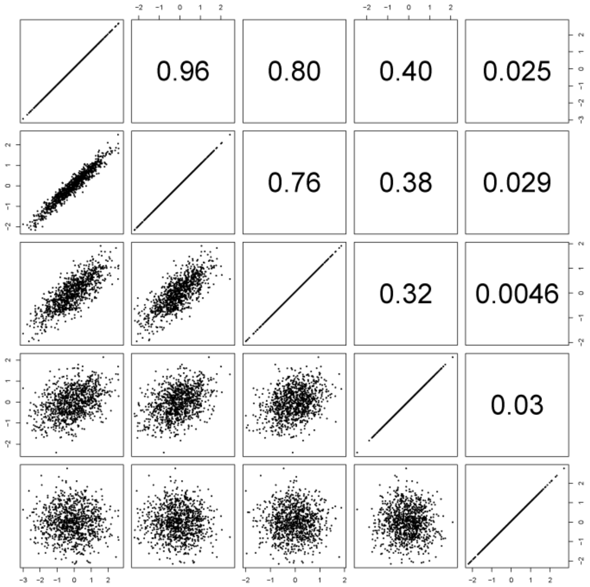


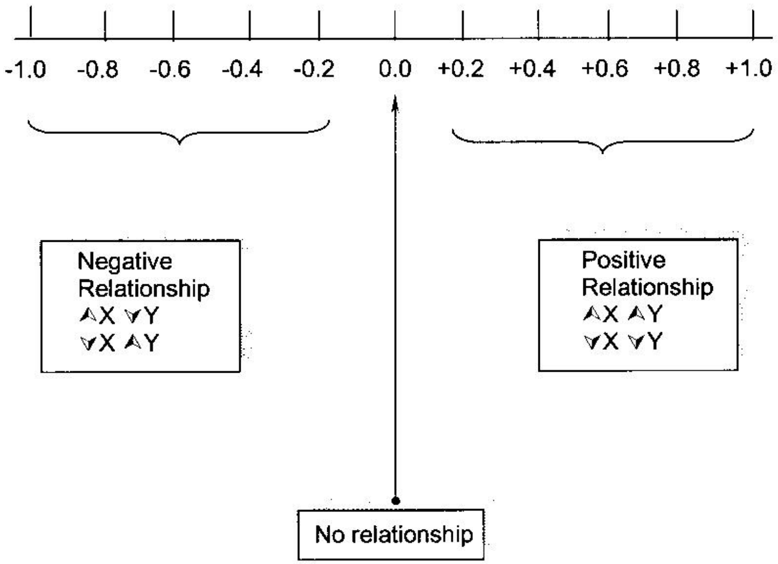
Association between 2 continuous variables
Linear relationship

Babraham Bioinformatics

Correlation

- A correlation coefficient is an index number that measures:
  - The magnitude and the direction of the relation between 2 variables
  - It is designed to range in value between -1 and +1





Correlation

- Assumptions for correlation
  - Regression and linear Model (lm)

- **Linearity**: The relationship between X and the mean of Y is linear.
- **Homoscedasticity**: The variance of residual is the same for any value of X.
- **Independence:** Observations are independent of each other.
- **Normality:** For any fixed value of X, Y is normally distributed.

Correlation

- Assumptions for correlation
  - Regression and linear Model (lm)

- **Outliers**: the observed value for the point is very different from that predicted by the regression model.

- **Leverage points**: A leverage point is defined as an observation that has a value of x that is far away from the mean of x.

- **Influential observations**: change the slope of the line. Thus, have a large influence on the fit of the model.

❖ **One method to find influential** points is to compare the fit of the **model with** and **without** each observation.

- Bottom line: **influential outliers** are problematic.

Correlation

- Most widely-used correlation coefficient:
  - Pearson product-moment correlation coefficient “r”

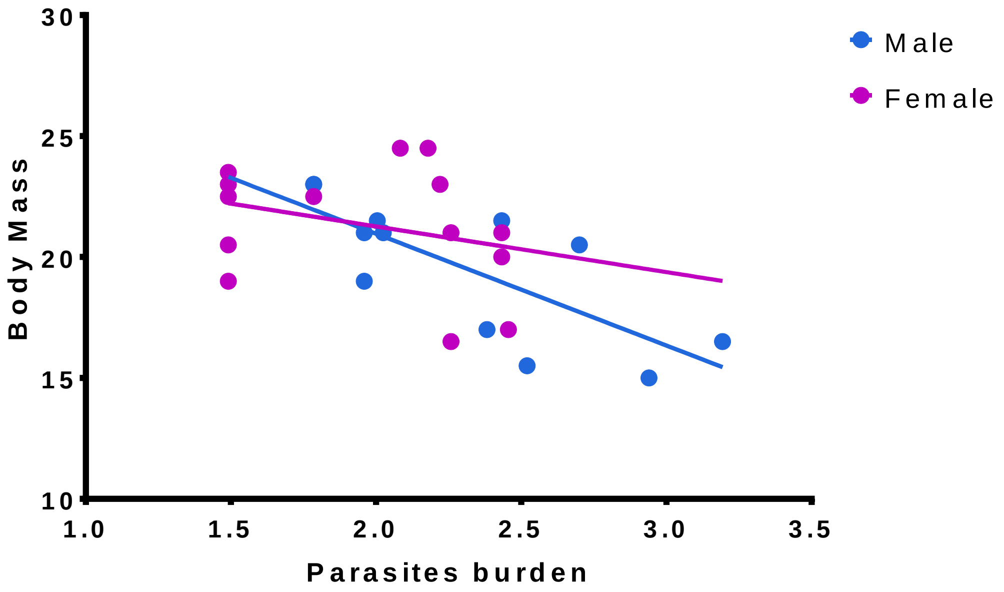
$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

    - The 2 variables do not have to be measured in the same units but they have to be proportional (meaning linearly related)
  - Coefficient of determination:
    - r is the correlation between X and Y
    - r^2 is the coefficient of determination:
      - It gives you the proportion of variance in Y that can be explained by X, in percentage.

Correlation

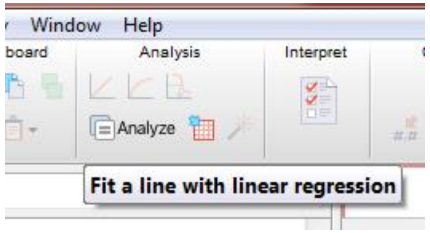
Example: roe deer.xlsx

- Is there a relationship between parasite burden and body mass in roe deer?



Correlation

Example: roe deer.xlsx



Linear reg. Tabular results	A Male	B Female
1 **Best-fit values**		
2 Slope	-4.621	-1.888
3 Y-intercept	30.20	25.04
4 X-intercept	6.536	13.26
5 1/slope	-0.2164	-0.5297
6		
7 **Std. Error**		
8 Slope	1.287	1.721
9 Y-intercept	3.025	3.453
10		
11 **95% Confidence Intervals**		
12 Slope	-7.490 to -1.753	-5.637 to 1.861
13 Y-intercept	23.46 to 36.94	17.51 to 32.56
14 X-intercept	4.902 to 13.47	5.738 to +infinity
15		
16 **Goodness of Fit**		
17 R square	0.5630	0.09119
18 Sy.x	1.980	2.512
19		
20 **Is slope significantly non-zero?**		
21 F	12.89	1.204
22 DFn, DFd	1, 10	1, 12
23 P value	0.0049	0.2940
24 Deviation from zero?	Significant	Not Significant
25		
26 **Equation**	Y = -4.621*X + 30.20	Y = -1.888*X + 25.04
27		
28 **Data**		
29 Number of X values	12	26
30 Maximum number of Y replicates	1	1
31 Total number of values	12	14
32 Number of missing values	0	12

There is a negative correlation between parasite load and fitness but this relationship is only significant for the males(p=0.0049 vs. females: p=0.2940).

Correlation	A PL vs. Male	B PL vs. Female
1 **Pearson r**		
2 r	-0.7504	-0.3020
3 95% confidence interval	-0.9256 to -0.3099	-0.7176 to 0.2722
4 R squared	0.5630	0.09119
5		
6 **P value**		
7 P (two-tailed)	0.0049	0.2940
8 P value summary	**	ns
9 Significant? (alpha = 0.05)	Yes	No
10		
11 **Number of XY Pairs**	12	14

Association between 2 continuous variables
Linear relationship
Diagnostic

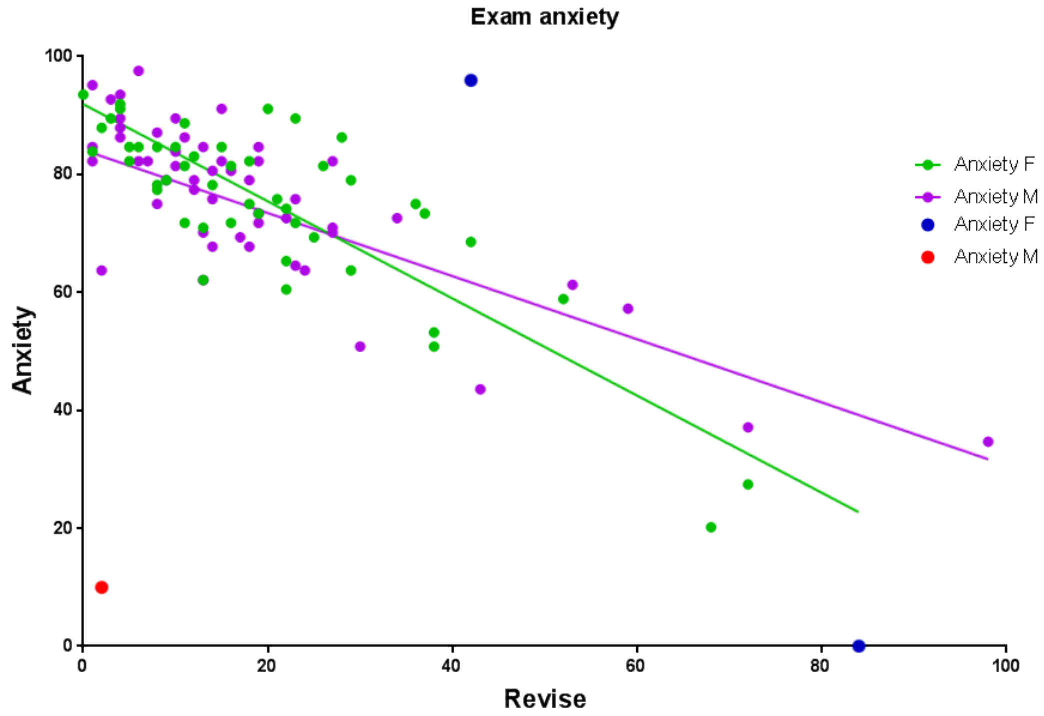
Babraham Bioinformatics

Correlation

exam anxiety.xlsx

• **Question**: Is there a relationship between time spent revising and exam anxiety? And, if yes, are boys and girls different?

- <u>**Focus**</u>: how good is the model?



Correlation

exam anxiety.xlsx

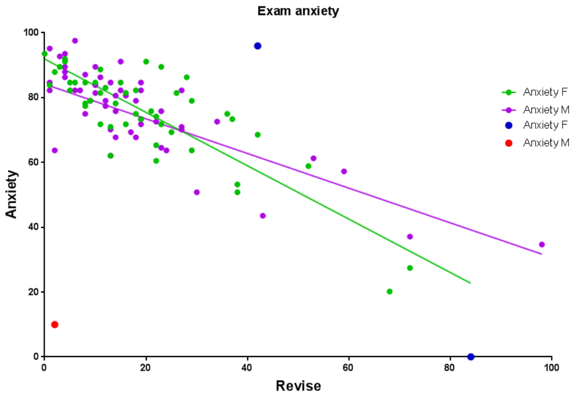
- **Question**: Is there a relationship between time spent revising and exam anxiety? And, if yes, are boys and girls different?

- **Focus**: how good is the model?

	Nonlin fit – Table of results	Anxiety F	Anxiety M	Global (shared)
		Y	Y	Y
1	Comparison of Fits			
2	Null hypothesis			Slope same for all data sets
3	Alternative hypothesis			Slope different for each data set
4	P value			0.0299
5	Conclusion (alpha = 0.05)			Reject null hypothesis
6	Preferred model			Slope different for each data set
7	F (DFn, DFd)			4.852 (1, 99)
8				
9	Slope different for each data set			
10	Best-fit values			
11	YIntercept	91.94	84.19	
12	Slope	-0.8238	-0.5353	
13	Std. Error			
14	YIntercept	2.279	2.621	
15	Slope	0.08173	0.1016	
16	95% CI (profile likelihood)			
17	YIntercept	87.36 to 96.52	78.93 to 89.46	
18	Slope	-0.988 to -0.6596	-0.7394 to -0.3312	
19	Goodness of Fit			
20	Degrees of Freedom	49	50	
21	R square	0.6746	0.3568	
22	Absolute Sum of Squares	5322	8845	
23	Sy.x	10.42	13.3	

Normality of Residuals		
D'Agostino & Pearson omnibus K2	14.43	68.42
P value	0.0007	<0.0001
Passed normality test (alpha=0.05)?	No	No
P value summary	***	****
Number of points		
# of X values	51	103
# Y values analyzed	51	52
Outliers (not excluded, Q=1%)	2	1

Correlation	Revise vs. Anxiety F	Revise vs. Anxiety M
	Y	Y
Pearson r		
r	-0.8214	-0.5974
95% confidence interval	-0.8945 to -0.7055	-0.7483 to -0.3877
R squared	0.6746	0.3568
P value		
P (two-tailed)	<0.0001	<0.0001
P value summary	****	****
Significant? (alpha = 0.05)	Yes	Yes



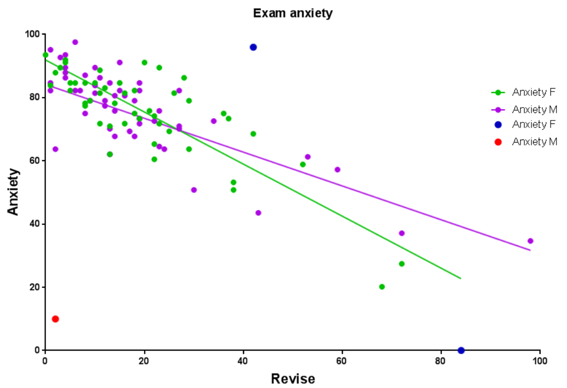
Correlation
exam anxiety.xlsx

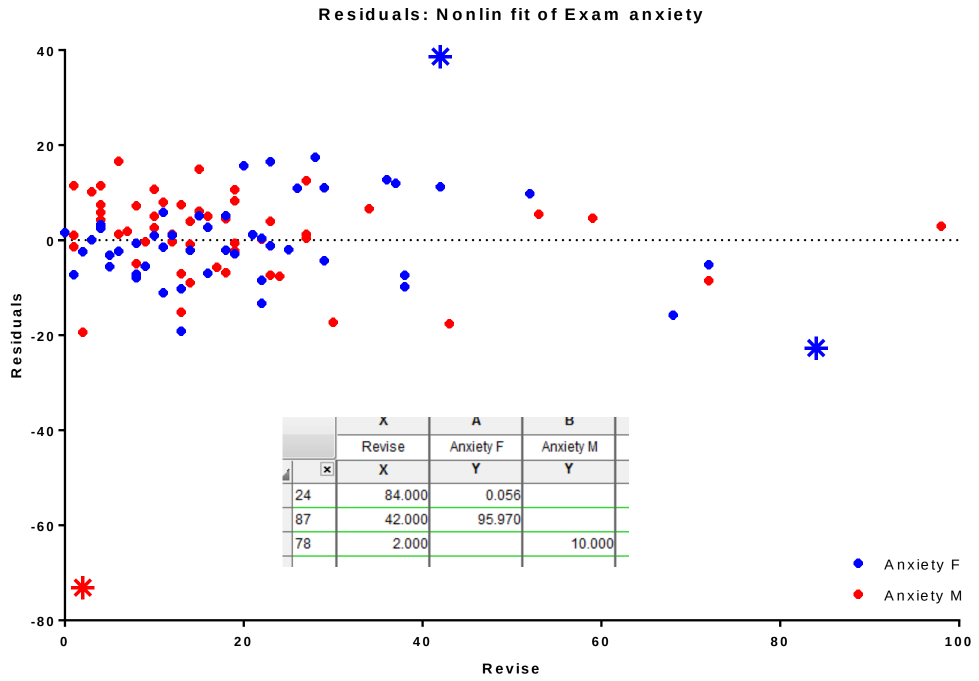
- **Question**: Is there a relationship between time spent revising and exam anxiety? And, if yes, are boys and girls different?

- <u>**Focus**</u>: how good is the model? **Diagnostic**: we don't like students 24, 87 and 78

Normality of Residuals		
D'Agostino & Pearson omnibus K2	14.43	68.42
P value	0.0007	<0.0001
Passed normality test (alpha=0.05)?	No	No
P value summary	***	****

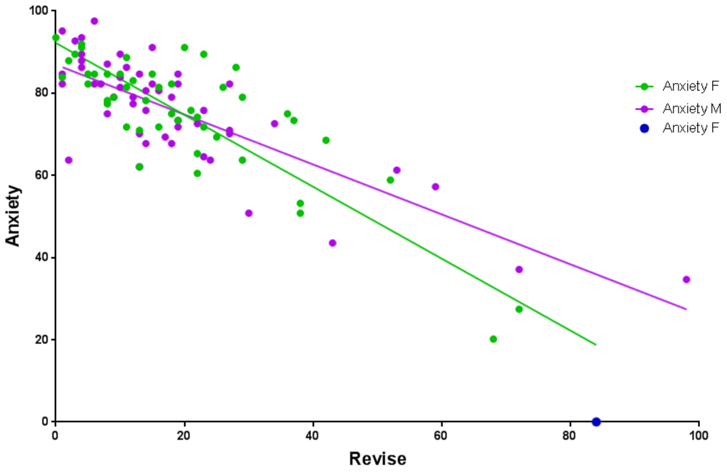
Number of points		
# of X values	51	103
# Y values analyzed	51	52
Outliers (not excluded, Q=1%)	2	1





Correlation

exam anxiety.xlsx



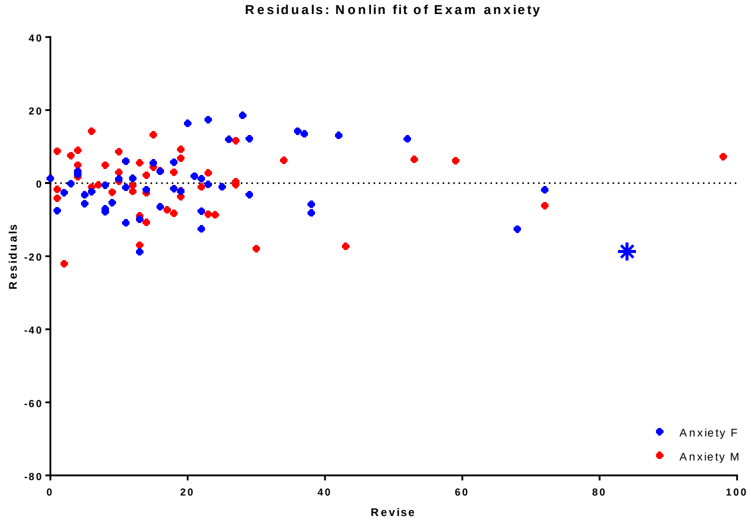


Table of results	Anxiety F	Anxiety M	Global (shared)
	Y	Y	Y
Comparison of Fits			
Null hypothesis			Slope same for all data sets
Alternative hypothesis			Slope different for each data set
P value			0.0056
Conclusion (alpha = 0.05)			Reject null hypothesis
Preferred model			Slope different for each data set
F (DFn, DFd)			8.022 (1, 97)
Slope different for each data set			
Best-fit values			
YIntercept	92.25	86.97	
Slope	-0.875	-0.6075	
Std. Error			
YIntercept	1.936	1.648	
Slope	0.07033	0.06326	
95% CI (profile likelihood)			
YIntercept	88.35 to 96.14	83.66 to 90.29	
Slope	-1.016 to -0.7336	-0.7347 to -0.4804	
Goodness of Fit			
Degrees of Freedom	48	49	
R square	0.7633	0.653	
Absolute Sum of Squares	3759	3306	
Sy.x	8.849	8.213	

Normality of Residuals		
D'Agostino & Pearson omnibus K2	0.5158	5.132
P value	0.7727	0.0768
Passed normality test (alpha=0.05)?	Yes	Yes
P value summary	ns	ns

Correlation	Revise vs. Anxiety F	Revise vs. Anxiety M
	Y	Y
Pearson r		
r	-0.8737	-0.8081
95% confidence interval	-0.9267 to -0.7866	-0.8863 to -0.6851
R squared	0.7633	0.653
P value		
P (two-tailed)	<0.0001	<0.0001
P value summary	****	****
Significant? (alpha = 0.05)	Yes	Yes

Association between 2 continuous variables
Linear relationship
Non-parametric

Babraham Bioinformatics

Non-Parametric:
Spearman Correlation Coefficient

- Only really useful for ranks (either one or both variables)
- **ρ (rho)** is the equivalent of r and calculated in a similar way

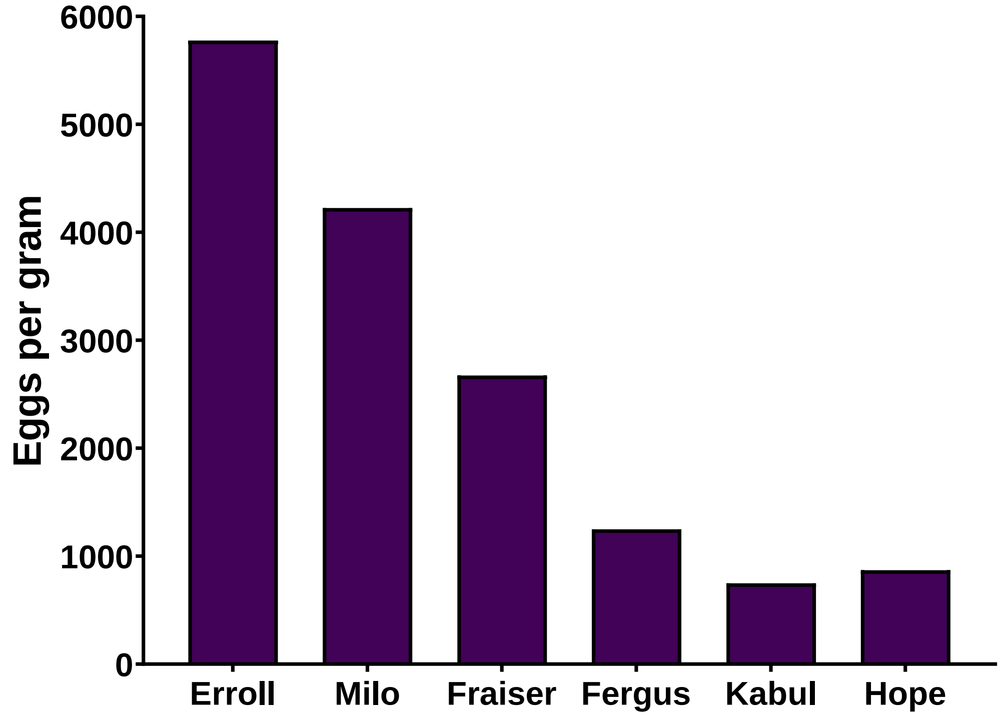
- **Example**: **dominance.xslx**
  - Six male colobus monkeys ranked for dominance
  - Question: is social dominance associated with parasitism?
    - Eggs of *Trichirus* nematode per gram of monkey faeces

Monkey	Dominance	Eggs.per.gram
Erroll	1	5777
Milo	2	4225
Fraiser	3	2674
Fergus	4	1249
Kabul	5	749
Hope	6	870

Non-Parametric:

Spearman Correlation Coefficient

	Correlation	Dominance vs. Eggs per gram
1	**Spearman r**	
2	r	-0.9429
3	95% confidence interval	
4		
5	**P value**	
6	P (two-tailed)	0.0167
7	P value summary	*
8	Exact or approximate P value?	Exact
9	Significant? (alpha = 0.05)	Yes
10		
11	**Number of XY Pairs**	6
12		



- **Answer**: the relationship between dominance and parasitism is significant (ρ =-0.94, p=0.017) with high ranking males harbouring a heavier burden.

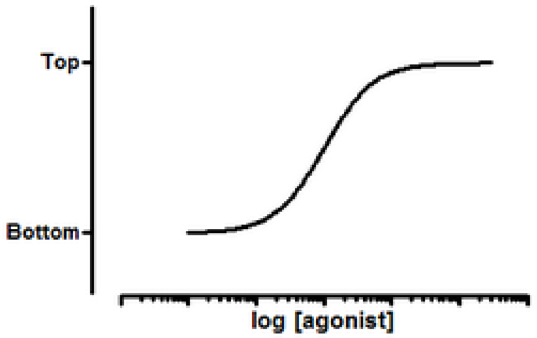
Association between 2 continuous variables

Non-linear relationship

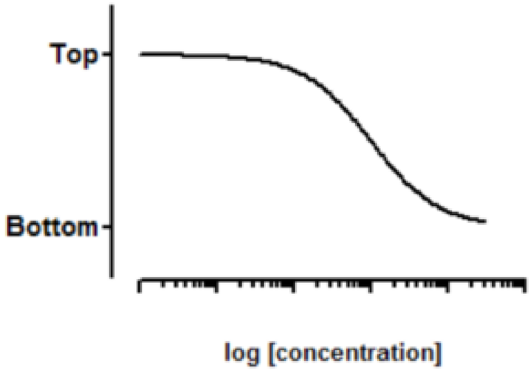
Curve fitting

- **Dose-response curves**
  - Nonlinear regression
  - Dose-response experiments typically use around 5-10 doses of agonist, equally spaced on a logarithmic scale
  - Y values are responses

- The aim is often to determine the **IC50** or the **EC50**
  - **IC50 (I=Inhibition)**: concentration of an agonist that provokes a response half way between the maximal (Top) response and the maximally inhibited (Bottom) response.
  - **EC50 (E=Effective):** concentration that gives half-maximal response



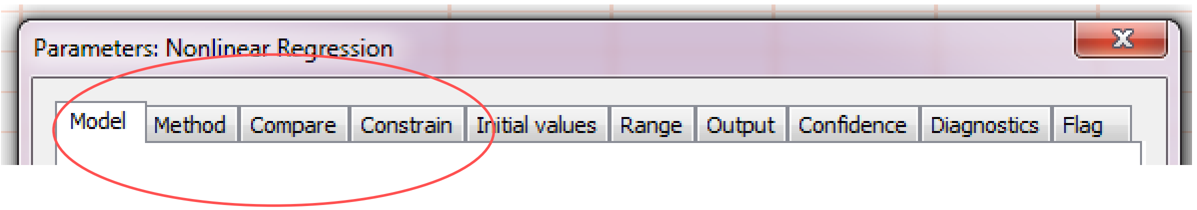
Stimulation:
Y=Bottom + (Top-Bottom)/(1+10^((LogEC50-X)*HillSlope))

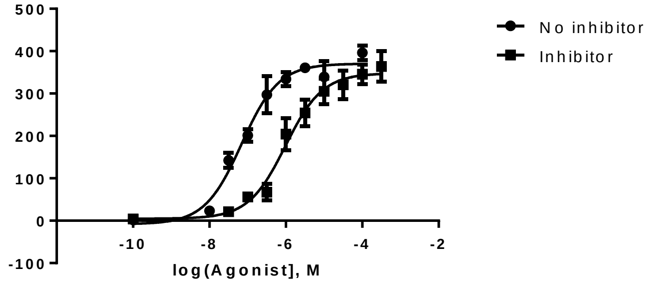


Inhibition:
Y=Bottom + (Top-Bottom)/(1+10^((X-LogIC50)))

Curve fitting

Example: inhibition data.xlsx





Step by step analysis and considerations:

1- Choose a **Model**:

not necessary to normalise

should choose it when values defining 0 and 100 are precise

variable slope better if plenty of data points (variable slope or 4 parameters)

2- Choose a **Method:** outliers, fitting method, weighting method and replicates

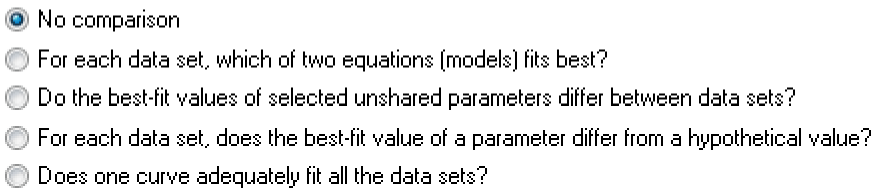
3- **Compare** different conditions:

Diff in parameters

Diff between conditions for one or more parameters

Constraint vs no constraint

Diff between conditions for one or more parameters



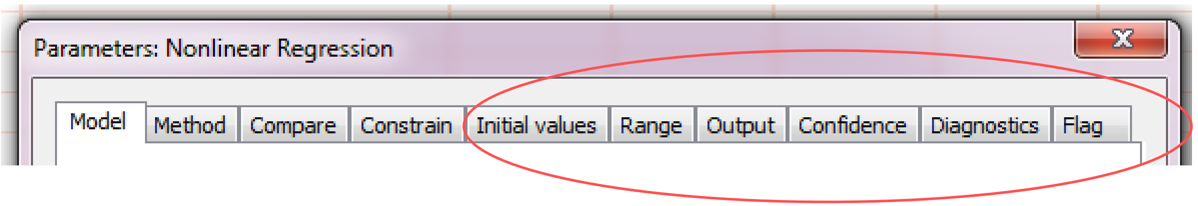
4- **Constrain**:

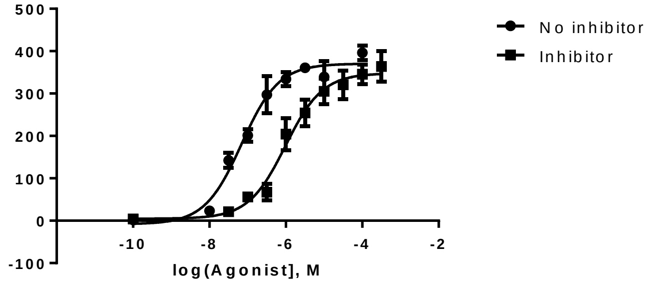
depends on your experiment

depends if your data don't define the top or the bottom of the curve

Curve fitting

Example: inhibition data.xlsx





Step by step analysis and considerations:

5- **Initial values**:

defaults usually OK unless the fit looks funny

6- **Range**:

defaults usually OK unless you are not interested in the x-variable full range (ie time)

7- **Output**:

summary table presents same results in a … summarized way.

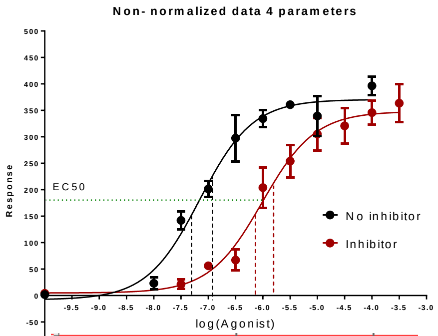
8 – **Confidence**: calculate and plot confidence intervals

9- **Diagnostics**:

check for normality (weights) and outliers (but keep them in the analysis)
check Replicates test
residual plots

Curve fitting

Example: inhibition data.xlsx

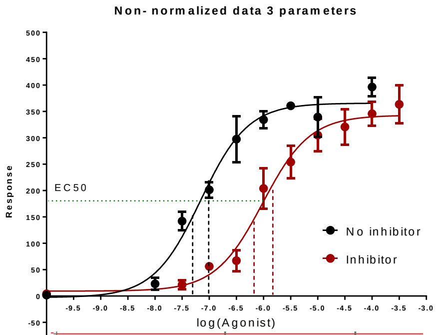


LogEC50	-7.158	-6.011

LogEC50 same for all data sets
LogEC50 different for each data set
< 0.0001
Reject null hypothesis
LogEC50 different for each data set
64.86 (1,48)

95% Confidence Intervals		
Bottom	-41.39 to 24.94	-22.15 to 31.56
Top	348.3 to 392.6	323.1 to 373.0
LogEC50	-7.324 to -6.991	-6.185 to -5.837
HillSlope	0.6347 to 1.159	0.6095 to 1.186
EC50	4.739e-008 to 1.020e-007	6.538e-007 to 1.455e-006

R square	0.9663	0.9653

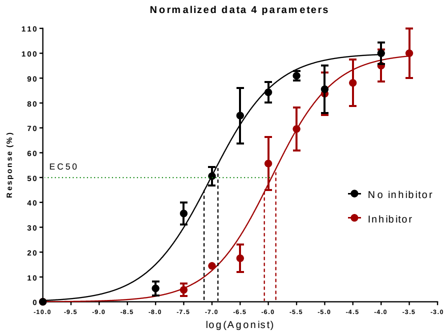


LogEC50	-7.159	-6.017

LogEC50 same for all data sets
LogEC50 different for each data set
< 0.0001
Reject null hypothesis
LogEC50 different for each data set
101.0 (1,50)

95% Confidence Intervals		
Bottom	-30.74 to 24.78	-11.65 to 30.07
Top	348.2 to 383.2	324.3 to 361.4
LogEC50	-7.312 to -7.006	-6.175 to -5.859
EC50	4.875e-008 to 9.858e-008	6.677e-007 to 1.385e-006

R square	0.9655	0.9648

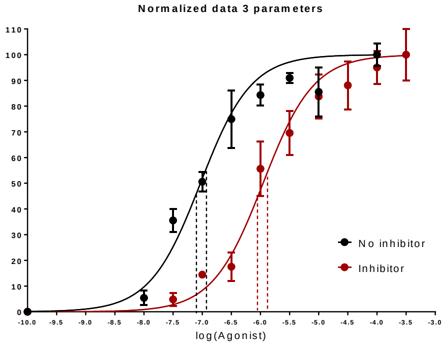


LogEC50	-7.017	-5.943

LogEC50 same for all data sets
LogEC50 different for each data set
< 0.0001
Reject null hypothesis
LogEC50 different for each data set
162.8 (1,52)

95% Confidence Intervals		
LogEC50	-7.137 to -6.897	-6.057 to -5.830
HillSlope	0.6094 to 0.9184	0.6467 to 0.9460
EC50	7.295e-008 to 1.268e-007	8.763e-007 to 1.481e-006

R square	0.9580	0.9635



LogEC50	-7.031	-5.956

One curve for all data sets
Different curve for each data set
< 0.0001
Reject null hypothesis
Different curve for each data set
175.0 (1,54)

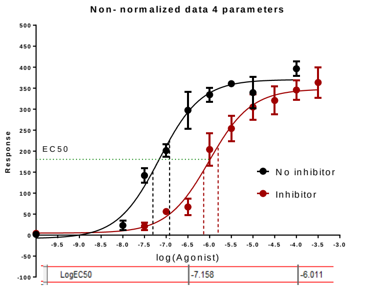
95% Confidence Intervals		
LogEC50	-7.144 to -6.917	-6.064 to -5.848
EC50	7.179e-008 to 1.209e-007	8.633e-007 to 1.420e-006

R square	0.9476	0.9568

Curve fitting

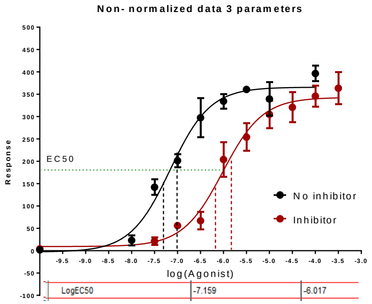
Example: inhibition data.xlsx

	No inhibitor	Inhibitor
Replicates test for lack of fit		
SD replicates	22.71	25.52
SD lack of fit	41.84	32.38
Discrepancy (F)	3.393	1.610
P value	0.0247	0.1989
Evidence of inadequate model?	Yes	No



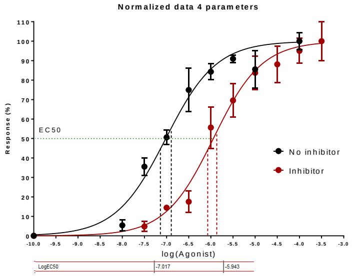
No inhibitor	Inhibitor
-7.158	-6.011

	No inhibitor	Inhibitor
Replicates test for lack of fit		
SD replicates	22.71	25.52
SD lack of fit	39.22	30.61
Discrepancy (F)	2.982	1.438
P value	0.0334	0.2478
Evidence of inadequate model?	Yes	No



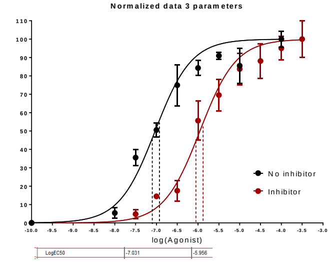
No inhibitor	Inhibitor
-7.159	-6.017

	No inhibitor	Inhibitor
Replicates test for lack of fit		
SD replicates	5.755	7.100
SD lack of fit	11.00	8.379
Discrepancy (F)	3.656	1.393
P value	0.0125	0.2618
Evidence of inadequate model?	Yes	No



No inhibitor	Inhibitor
-7.017	-5.943

	No inhibitor	Inhibitor
Replicates test for lack of fit		
SD replicates	5.755	7.100
SD lack of fit	12.28	9.649
Discrepancy (F)	4.553	1.847
P value	0.0036	0.1246
Evidence of inadequate model?	Yes	No



No inhibitor	Inhibitor
-7.031	-5.956

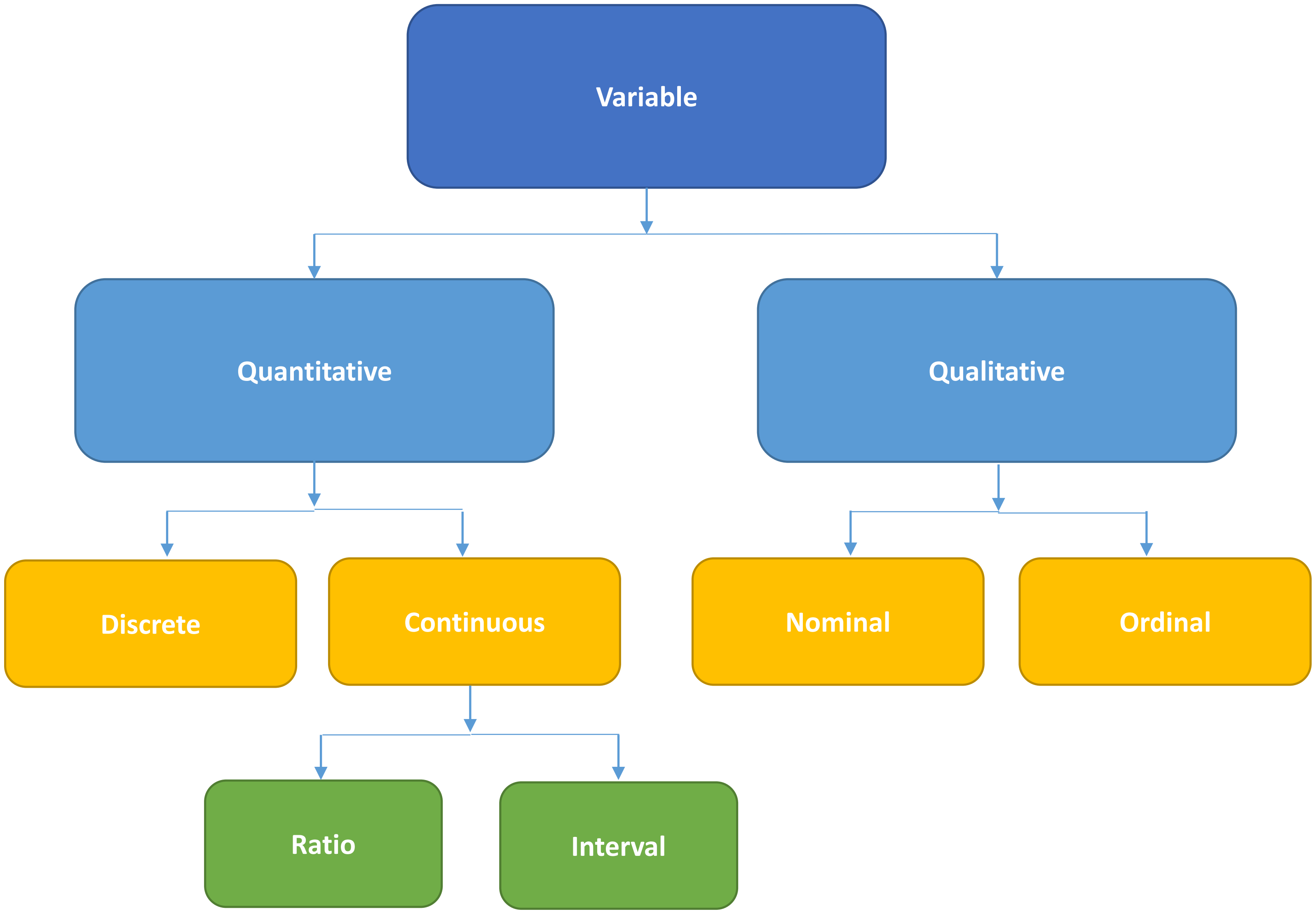
Day 3
Analysis of Qualitative data

Anne Segonds-Pichon

v2019-06

Babraham Institute

Babraham Bioinformatics


Variable
Quantitative
Qualitative
Discrete
Continuous
Nominal
Ordinal
Ratio
Interval

Qualitative data

- = not numerical
- = values taken = usually names (also *nominal*)
  - e.g. causes of death in hospital

- Values can be numbers but not numerical
  - e.g. group number = numerical label but not unit of measurement

- Qualitative variable with intrinsic order in their categories = *ordinal*

- Particular case: qualitative variable with 2 categories: *binary* or *dichotomous*
  - e.g. alive/dead or male/female

Fisher's exact and Chi²

Example: cats and dogs.xlsx

- Cats and dogs trained to line dance
- 2 different rewards: food or affection
- **Question**: Is there a difference between the rewards?

- **Is there a significant relationship between the 2 variables?**
  - does the reward significantly affect the likelihood of dancing?

- To answer this type of question:
  - **Contingency table**
  - **Fisher's exact or Chi² tests**

	Food	Affection
Dance	?	?
No dance	?	?

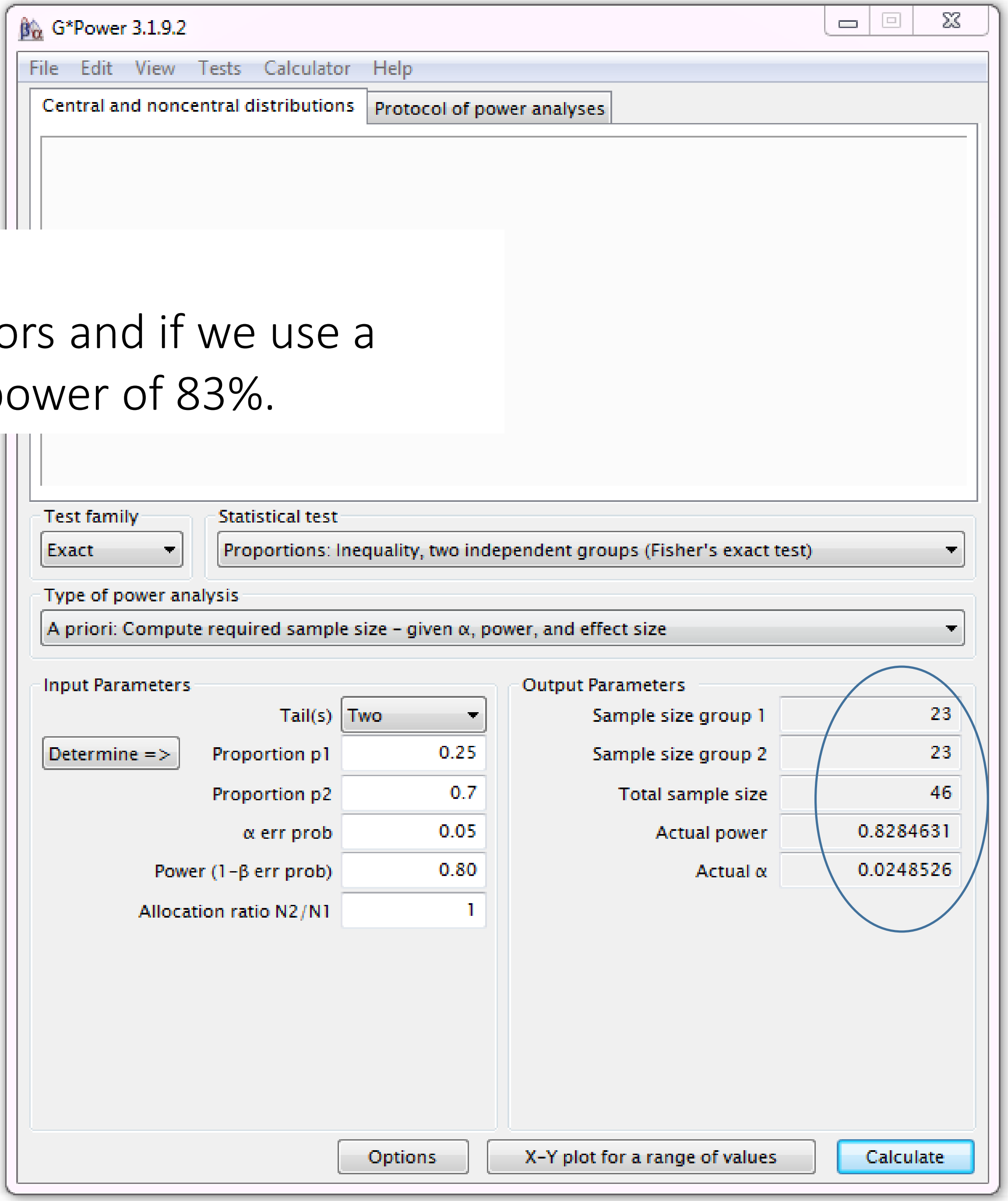
But first: **how many cats** do we need?

Exercise 11: Power calculation

- Preliminary results from a pilot study: **25%** line-danced after having received affection as a reward vs. **70%** after having received food.
  - **How many cats** do we need?

Exercise 11: Power calculation

Output:
If the values from the pilot study are good predictors and if we use a sample of **n=23 for each group**, we will achieve a power of 83%.



Chi-square and Fisher's tests

- Chi^2 test very easy to calculate by hand but Fisher's very hard
- Many software will not perform a Fisher's test on tables > 2x2
- **Fisher's test more accurate** than Chi^2 test on **small samples**
- **Chi^2 test more accurate** than Fisher's test on **large samples**
- Chi^2 test assumptions:
    - 2x2 table: no expected count <5
    - Bigger tables: all expected > 1 and no more than 20% < 5
- Yates's continuity correction
    - All statistical tests work well when their assumptions are met
    - When not: probability Type 1 error increases
    - Solution: corrections that increase p-values
        - Corrections are dangerous: no magic
        - Probably best to avoid them

Chi-square test

- In a chi-square test, the observed frequencies for two or more groups are compared with expected frequencies by chance.

$$\chi^2 = \Sigma \frac{(Observed\ Frequency - Expected\ Frequency)^2}{Expected\ Frequency}$$

  - With observed frequency = collected data

- **Example with ‘cats and dogs’**

Chi-square test

Did they dance? * Type of Training * Animal Crosstabulation

Animal				Type of Training: Food as Reward	Type of Training: Affection as Reward	Total
Cat	Did they dance?	Yes	Count	26	6	32
			% within Did they dance?	81.3%	18.8%	100.0%
		No	Count	6	30	36
			% within Did they dance?	16.7%	83.3%	100.0%
	Total		Count	32	36	68
			% within Did they dance?	47.1%	52.9%	100.0%
Dog	Did they dance?	Yes	Count	23	24	47
			% within Did they dance?	48.9%	51.1%	100.0%
		No	Count	9	10	19
			% within Did they dance?	47.4%	52.6%	100.0%
	Total		Count	32	34	66
			% within Did they dance?	48.5%	51.5%	100.0%

Did they dance? * Type of Training * Animal Crosstabulation

Animal				Type of Training: Food as Reward	Type of Training: Affection as Reward	Total
Cat	Did they dance?	Yes	Count	26	6	32
			Expected Count	15.1	16.9	32.0
		No	Count	6	30	36
			Expected Count	16.9	19.1	36.0
	Total		Count	32	36	68
			Expected Count	32.0	36.0	68.0
Dog	Did they dance?	Yes	Count	23	24	47
			Expected Count	22.8	24.2	47.0
		No	Count	9	10	19
			Expected Count	9.2	9.8	19.0
	Total		Count	32	34	66
			Expected Count	32.0	34.0	66.0

Example: expected frequency of cats line dancing after having received food as a reward:

Direct counts approach:

Expected frequency=(row total)*(column total)/grand total
= 32*32/68 = **15.1**

Probability approach:

Probability of line dancing: 32/68
Probability of receiving food: 32/68

Expected frequency:(32/68)*(32/68)=0.22: **22% of 68 = 15.1**

For the cats:

$Chi^2 = (26-15.1)^2/15.1 + (6-16.9)^2/16.9 + (6-16.9)^2/16.9 + (30-19.1)^2/19.1 = 28.4$

Is 28.4 big enough for the test to be significant?

Is 28.4 big enough for the test to be significant?

Student's *t*-test

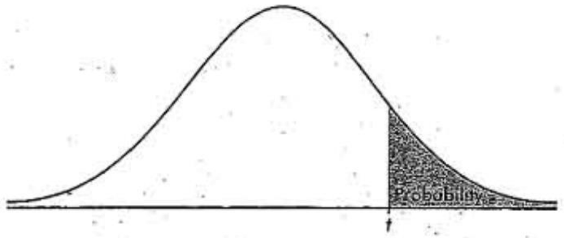


TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

df	Tail probability *p* .25	.20	.15	.10	.05	.025	.02	.01	.005	.0025
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428

χ^2 test

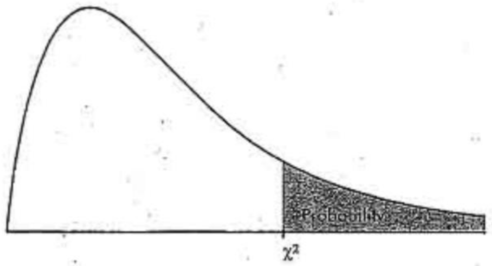


TABLE C: χ^2 CRITICAL VALUES

df	Tail probability *p* .25	.20	.15	.10	.05	.025	.02	.01	.005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19

Results

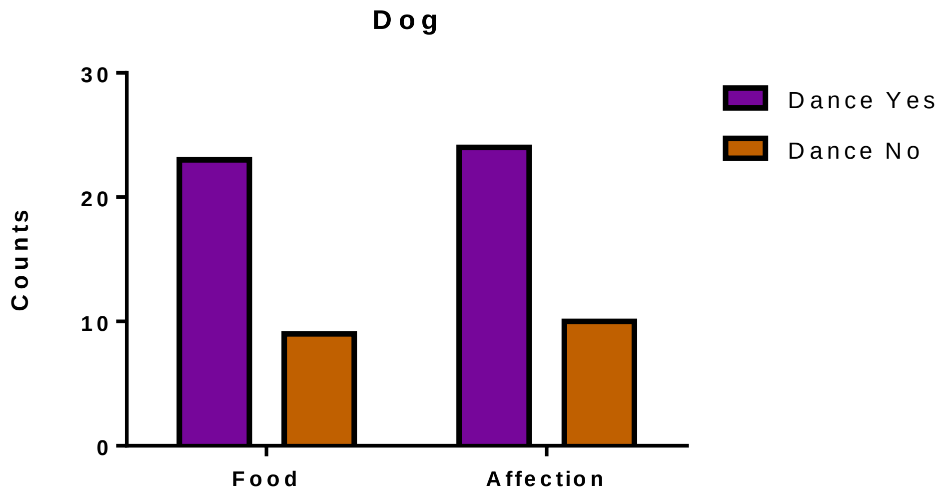
Table Analyzed	Cat
P value and statistical significance	
Test	Chi-square
Chi-square, df	28.36, 1
z	5.32[illegible]
P value	<0.0001
P value summary	****
One- or two-sided	Two-sided
Statistically significant (P < 0.05)?	Yes

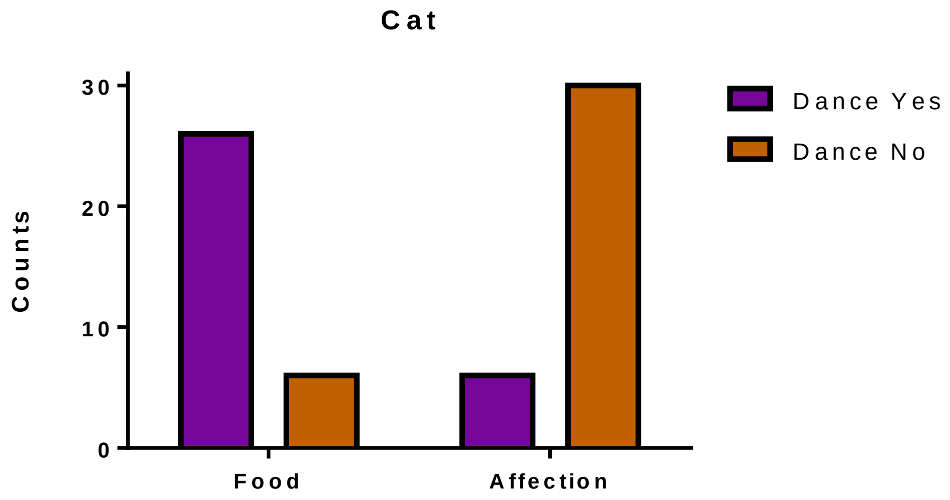
1	Table Analyzed	Cat
2		
3	Fisher's exact test	
4		
5	P value	< 0.0001
6	P value summary	***
7	One- or two-sided	Two-sided
8	Statistically significant? (alpha<0.05)	Yes

Table Analyzed	Dog
P value and statistical significance	
Test	Chi-square
Chi-square, df	0.01331, 1
z	0.1154
P value	0.9081
P value summary	ns
One- or two-sided	Two-sided
Statistically significant (P < 0.05)?	No

Table Analyzed	Dog
P value and statistical significance	
Test	Fisher's exact test
P value	>0.9999
P value summary	ns
One- or two-sided	Two-sided
Statistically significant (P < 0.05)?	No

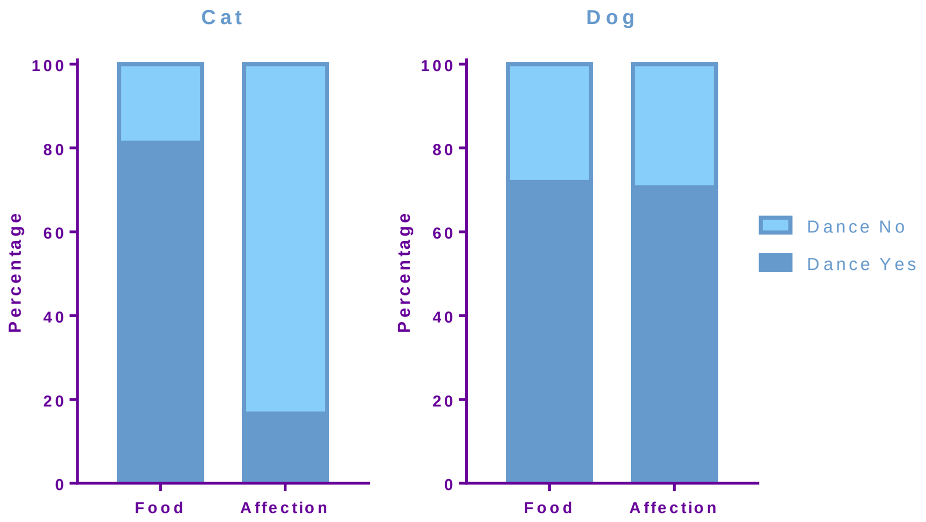
Fisher's exact test: results





- **In our example:**

cats are more likely to line dance if they are given food as reward than affection ($p<0.0001$) whereas dogs don't mind ($p>0.99$).



Exercise 12: Cane toads

	Infected	Uninfected
Rockhampton	12	8
Bowen	4	16
Mackay	15	5

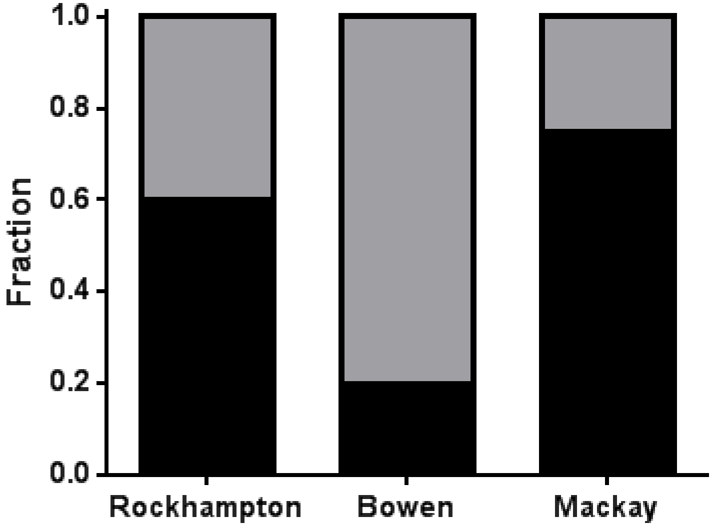
- A researcher decided to check the hypothesis that the proportion of cane toads with intestinal parasites was the same in 3 different areas of Queensland.

From Statistics Explained by Steve McKillup

- **Question**: Is the proportion of cane toads infected by intestinal parasites the same in 3 different areas of Queensland?

Exercise 12: Cane toads

Table Analyzed	Cane toad
Chi-square	
Chi-square, df	12.95, 2
P value	0.0015
P value summary	**
One- or two-tailed	NA
Statistically significant? (alpha<0.05)	Yes
Data analyzed	
Number of rows	3
Number of columns	2



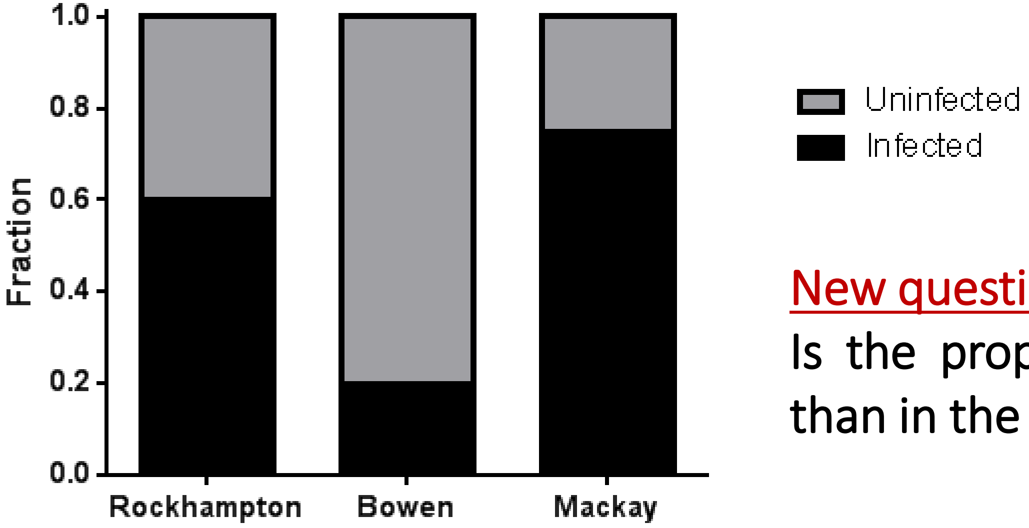
Uninfected
Infected

Answer:

The proportion of cane toads infected by intestinal parasites varies significantly between the 3 different areas of Queensland (p=0.0015), the animals being more likely to be parasitized in Rockhampton and Mackay than in Bowen.

Exercise 12: Cane toads

Table Analyzed	Cane toad
Chi-square	
Chi-square, df	12.95, 2
P value	0.0015
P value summary	**
One- or two-tailed	NA
Statistically significant? (alpha<0.05)	Yes
Data analyzed	
Number of rows	3
Number of columns	2



New question:

Is the proportion of infected cane toads lower in Bowen than in the other 2 areas?

Exercise 12: Cane toads

P value and statistical significance	
Test	Fisher's exact test
P value	0.0225

P value and statistical significance	
Test	Fisher's exact test
P value	0.0012