Laboratory & professional skills for Bioscientists term 3: Data Analysis in R

Revision: Overview and Developing understanding of term 2 statistics

Overview

- Module LO reminder
- What is the biological question
- Organising your data and analysis
- Overview for Choosing tests
- Figures
- Introduction to the Workshop

Module Learning Outcomes

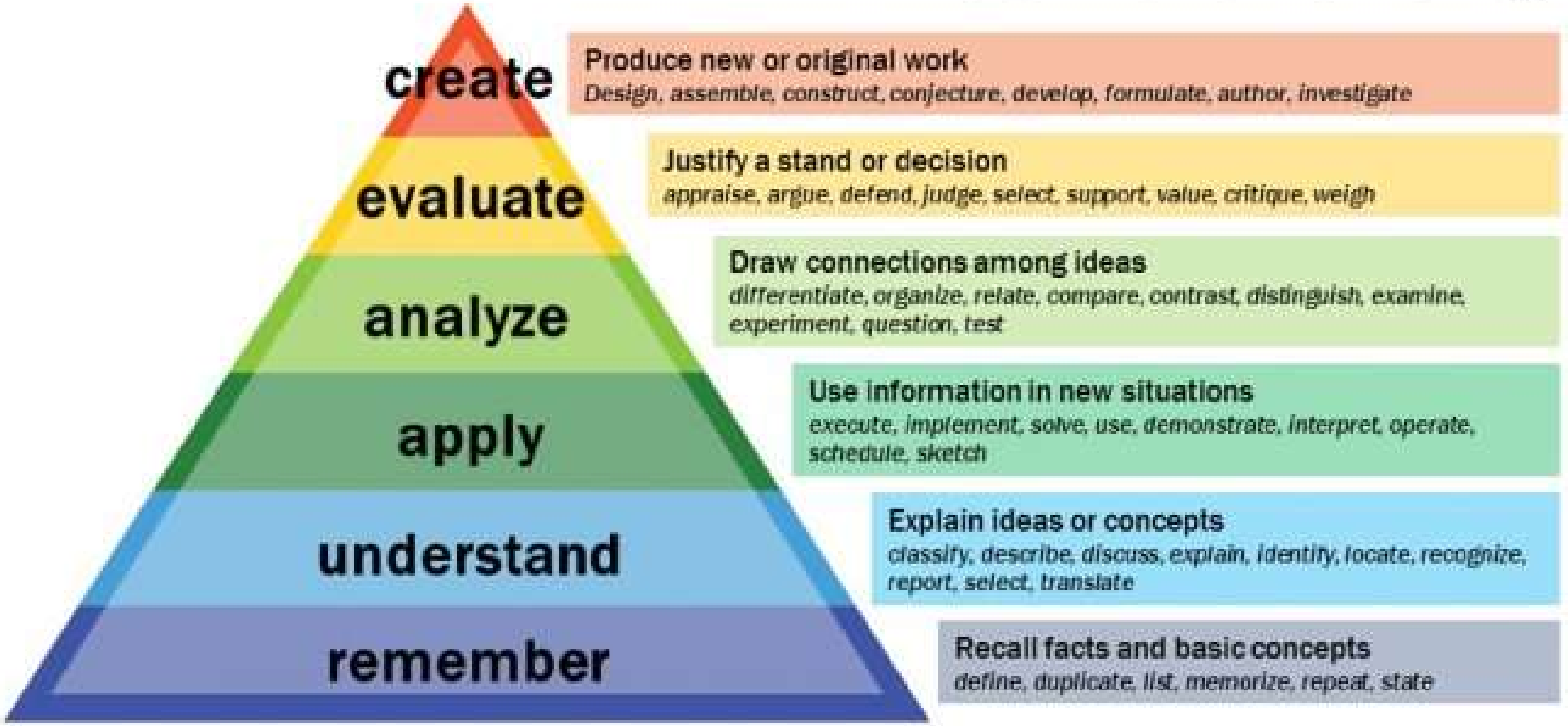
The successful student will be able to:

1. Explain the purpose of data analysis
2. Name, identify and choose classical univariate statistical tests (and some non-parametric equivalents) appropriate to a given scenario and recognise when these are not suitable
3. Use R to perform these analyses on data in a variety of formats
4. Interpret, report and graphically present the results of covered tests

Meeting the learning outcomes will enable you to:

- Write-up your laboratory report
- Design and analyse experiments including those for projects in stages 2, 3, and 4 and year-away
- Evaluate and interpret the data analysis in papers
- Perform well in assessments
- Improve your employability!

Bloom's Taxonomy



What is the biological question?

- Very many ‘statistics’ problems are biology problems
  - Think about your questions before you design experiments
  - What is your response variable?
  - What are your explanatory variables?
  - Generate dummy data

What is the biological question?

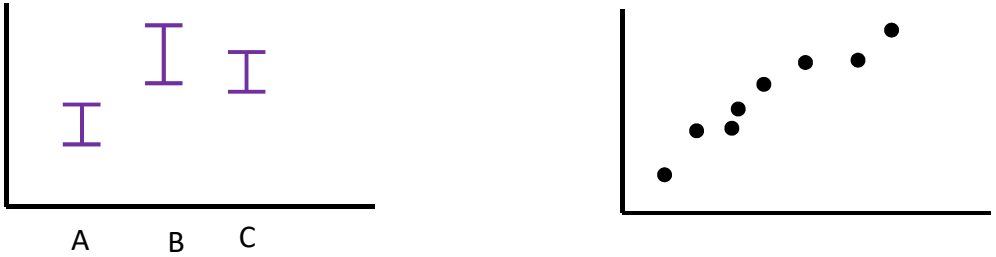
Response: blood pressure

Explanatory variables you want to understand: drug

Explanatory variable for control: sex, age, disease??

What is the biological question?

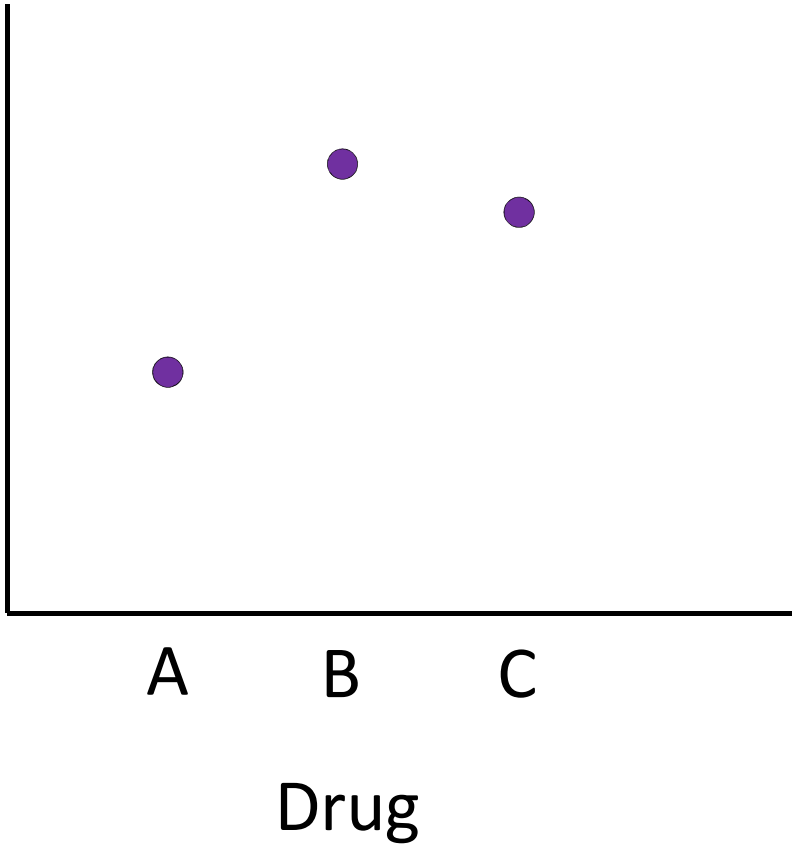
- Tests apply to a particular question



- Statistics are just evidence: as references are to your introduction, statistics are to your results.
- At this level you should be able to outline results are before the statistical analysis

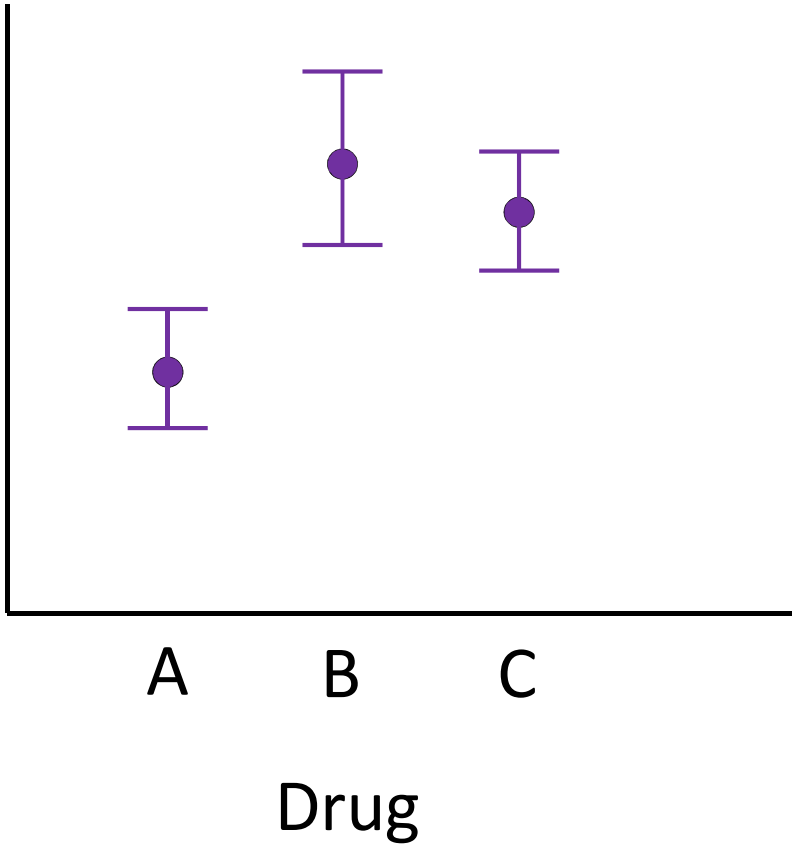
Outline the results

Drug A looks better than drugs B and C

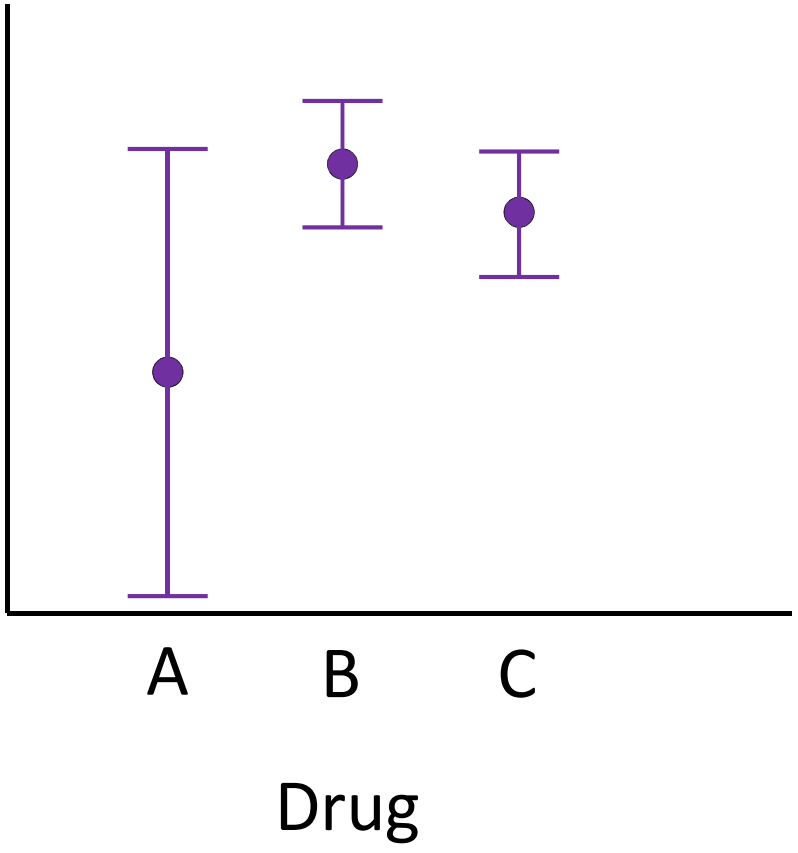


Statistical evidence for results being real

Drug A **is** better than drugs B and C



Or not…



Collecting data

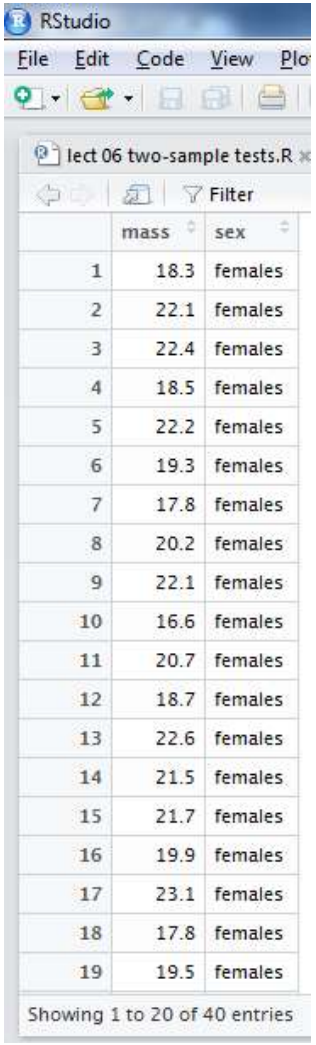
- Design and execute your experiment
- Experimental design and data analysis go hand-in-hand
  - Type of experiment determines statistics
  - But understanding of statistics informs design
- Organise your data in ’tidy’ format
- Workshop

Collecting data

- Organise your data in ’tidy’ format
- Fine to use a spread sheet but save in a plain text format (txt, csv, dat, etc)
- Write data straight into a blank tidy format

Wickham, H. (2014), “Tidy Data,” Journal of Statistical Software, 59, available at http://www.jstatsoft.org/article/view/v059i10

Response all in one column
Additional column for each explanatory
i.e., case (individual) per row


RStudio

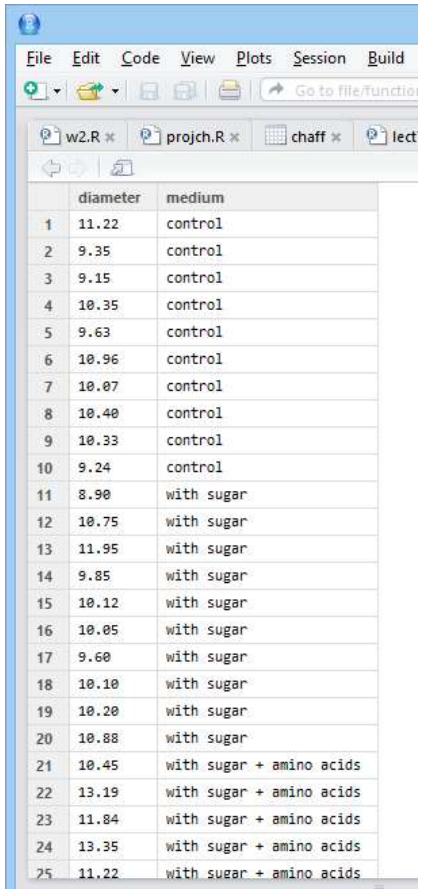
File Edit Code View Plo

lect 06 two-sample tests.R

Filter

	mass	sex
1	18.3	females
2	22.1	females
3	22.4	females
4	18.5	females
5	22.2	females
6	19.3	females
7	17.8	females
8	20.2	females
9	22.1	females
10	16.6	females
11	20.7	females
12	18.7	females
13	22.6	females
14	21.5	females
15	21.7	females
16	19.9	females
17	23.1	females
18	17.8	females
19	19.5	females

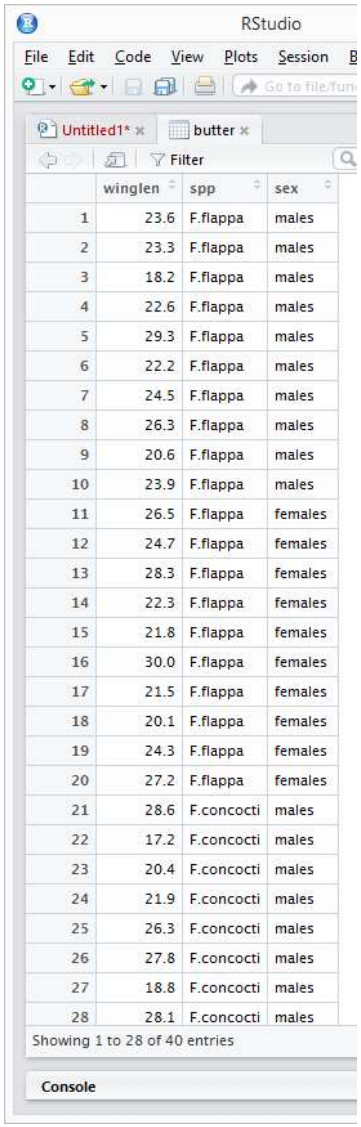
Showing 1 to 20 of 40 entries


File Edit Code View Plots Session Build

w2.R projch.R chaff lect

	diameter	medium
1	11.22	control
2	9.35	control
3	9.15	control
4	10.35	control
5	9.63	control
6	10.96	control
7	10.07	control
8	10.40	control
9	10.33	control
10	9.24	control
11	8.90	with sugar
12	10.75	with sugar
13	11.95	with sugar
14	9.85	with sugar
15	10.12	with sugar
16	10.05	with sugar
17	9.60	with sugar
18	10.10	with sugar
19	10.20	with sugar
20	10.88	with sugar
21	10.45	with sugar + amino acids
22	13.19	with sugar + amino acids
23	11.84	with sugar + amino acids
24	13.35	with sugar + amino acids
25	11.22	with sugar + amino acids

	x	y
1	12.43	24.94
2	14.55	22.98
3	9.41	25.74
4	10.31	25.98
5	10.64	23.16
6	14.48	26.20
7	6.91	27.89
8	9.92	22.99
9	8.38	24.67
10	8.07	24.53


RStudio

File Edit Code View Plots Session B

Untitled1* butter

Filter

	winglen	spp	sex
1	23.6	F.flappa	males
2	23.3	F.flappa	males
3	18.2	F.flappa	males
4	22.6	F.flappa	males
5	29.3	F.flappa	males
6	22.2	F.flappa	males
7	24.5	F.flappa	males
8	26.3	F.flappa	males
9	20.6	F.flappa	males
10	23.9	F.flappa	males
11	26.5	F.flappa	females
12	24.7	F.flappa	females
13	28.3	F.flappa	females
14	22.3	F.flappa	females
15	21.8	F.flappa	females
16	30.0	F.flappa	females
17	21.5	F.flappa	females
18	20.1	F.flappa	females
19	24.3	F.flappa	females
20	27.2	F.flappa	females
21	28.6	F.concocti	males
22	17.2	F.concocti	males
23	20.4	F.concocti	males
24	21.9	F.concocti	males
25	26.3	F.concocti	males
26	27.8	F.concocti	males
27	18.8	F.concocti	males
28	28.1	F.concocti	males

Showing 1 to 28 of 40 entries

Console

Before the experiment: Have some idea what test you'll be doing

- Always better to do BEFORE you design experiments and collect data
- Consider assumptions
- Take appropriate action – transform or chose another test

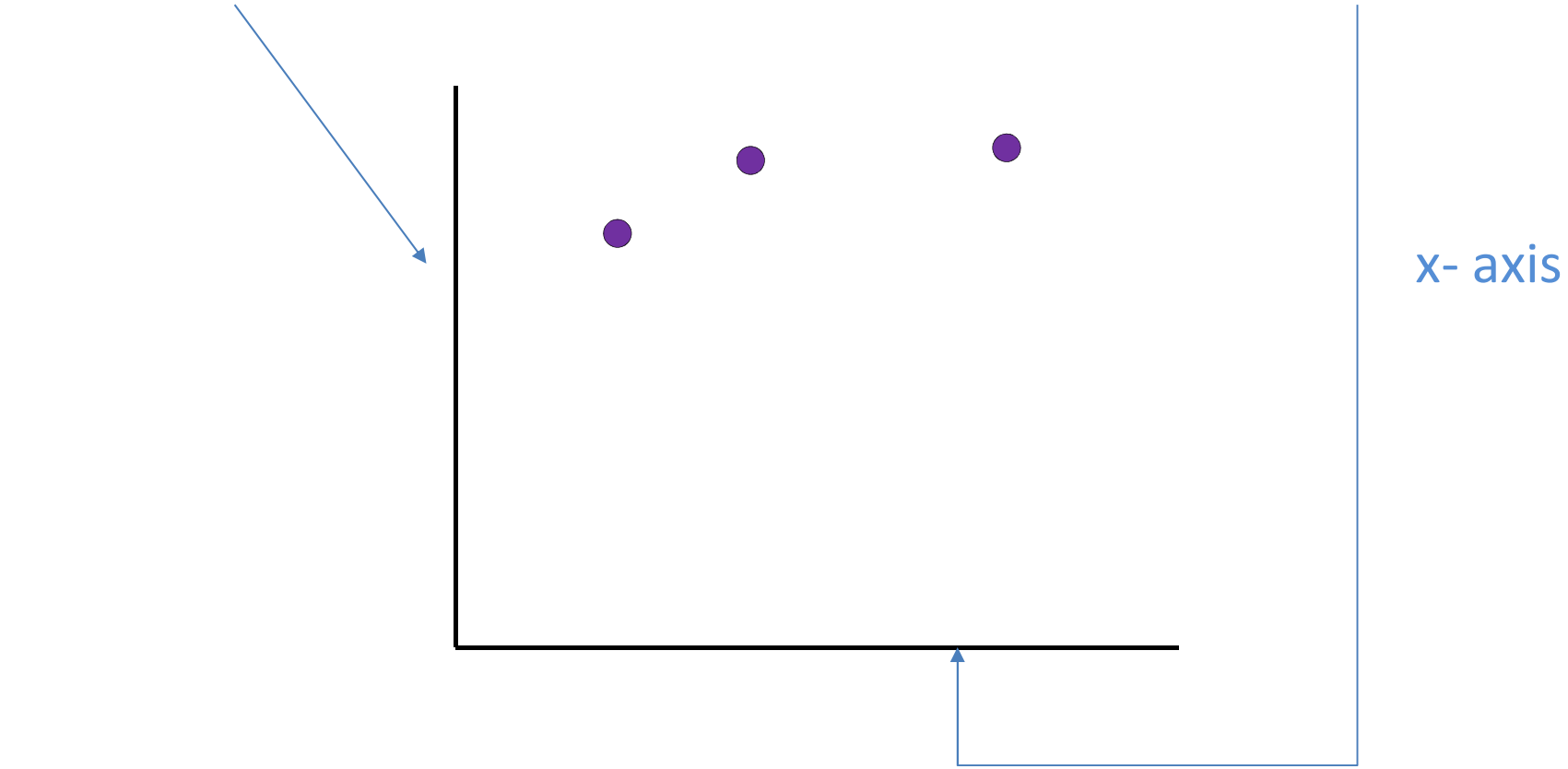
After the experiment: Explore your data

- Frequency histograms
- Scatter plots – geom_point()
- Boxplots
- Descriptive statistics – often absent from reports.
  - Sample sizes/number of cases
  - Variables
  - Means/medians, s.e
- You don’t put all your exploring in your report

Overview of experiments and analysis

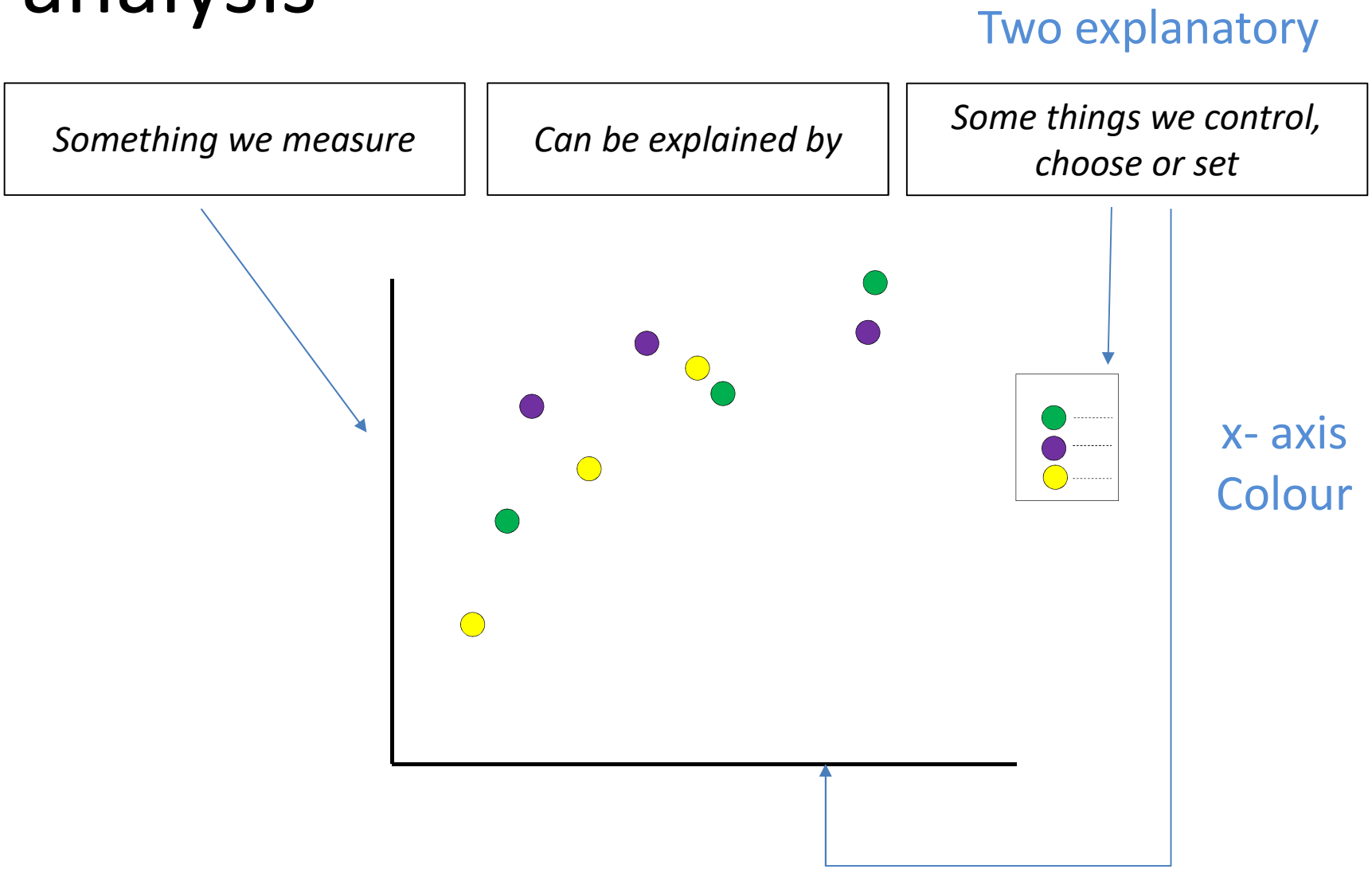
One explanatory

Something we measure	*Can be explained by*	*Some things we control, choose or set*

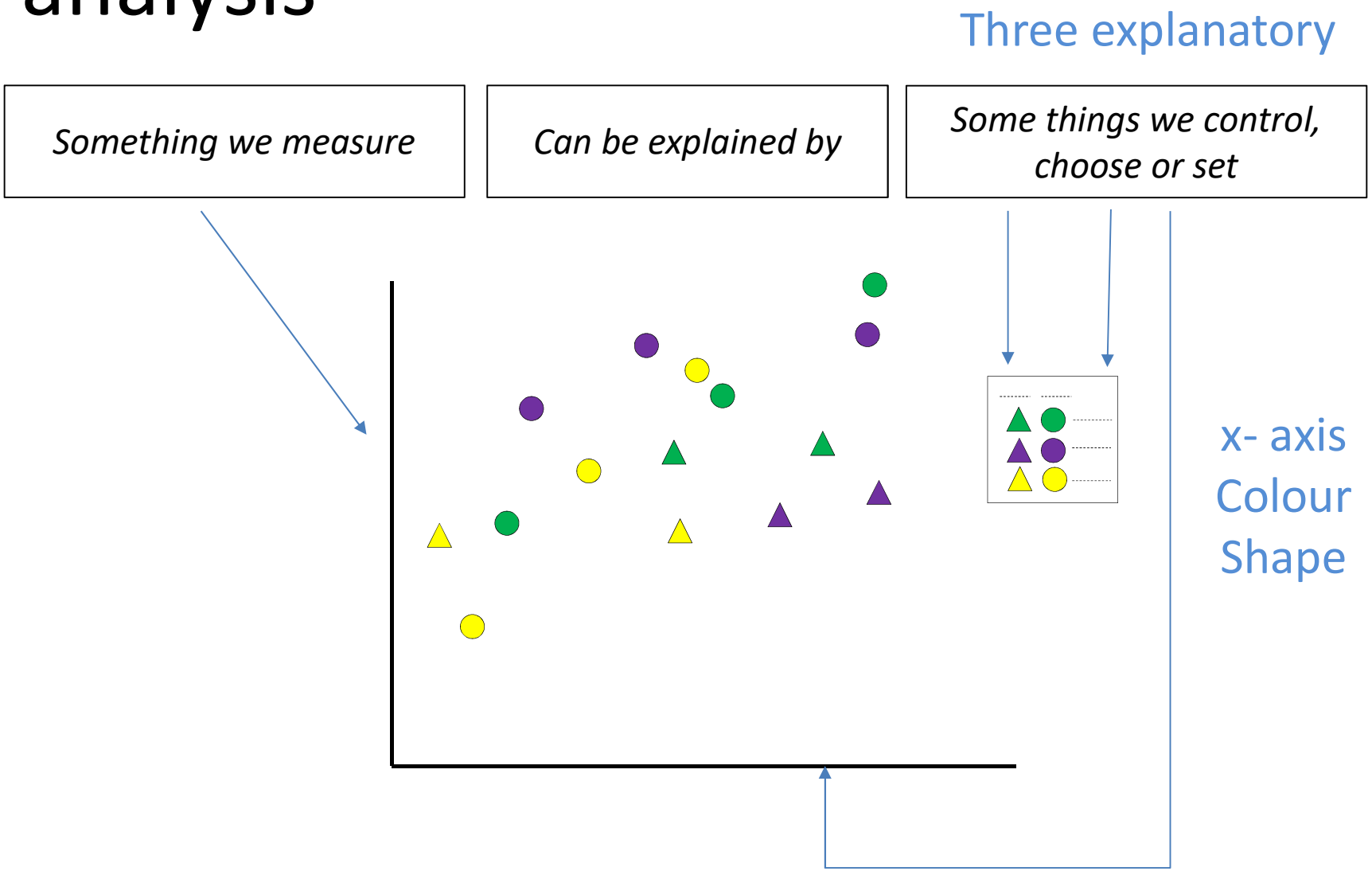


x- axis

Overview of experiments and analysis



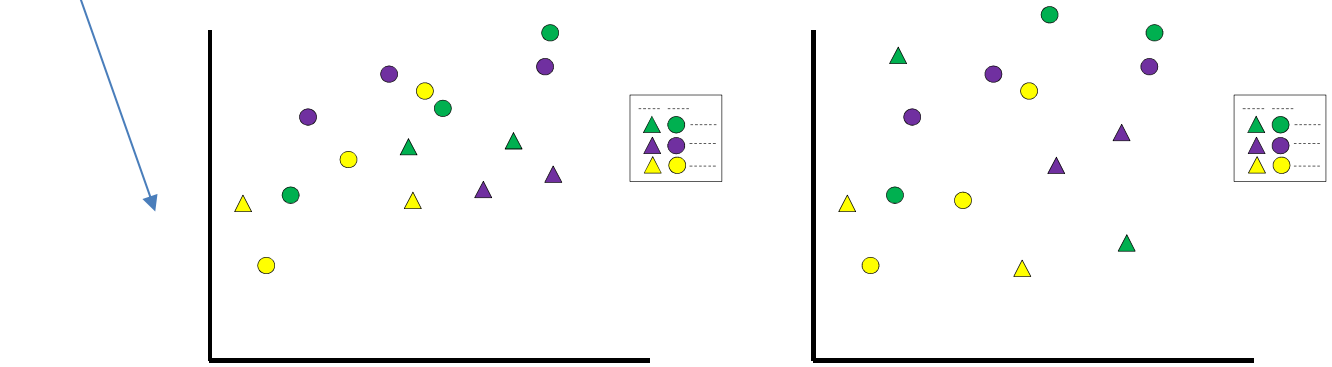
Overview of experiments and analysis



Overview of experiments and analysis

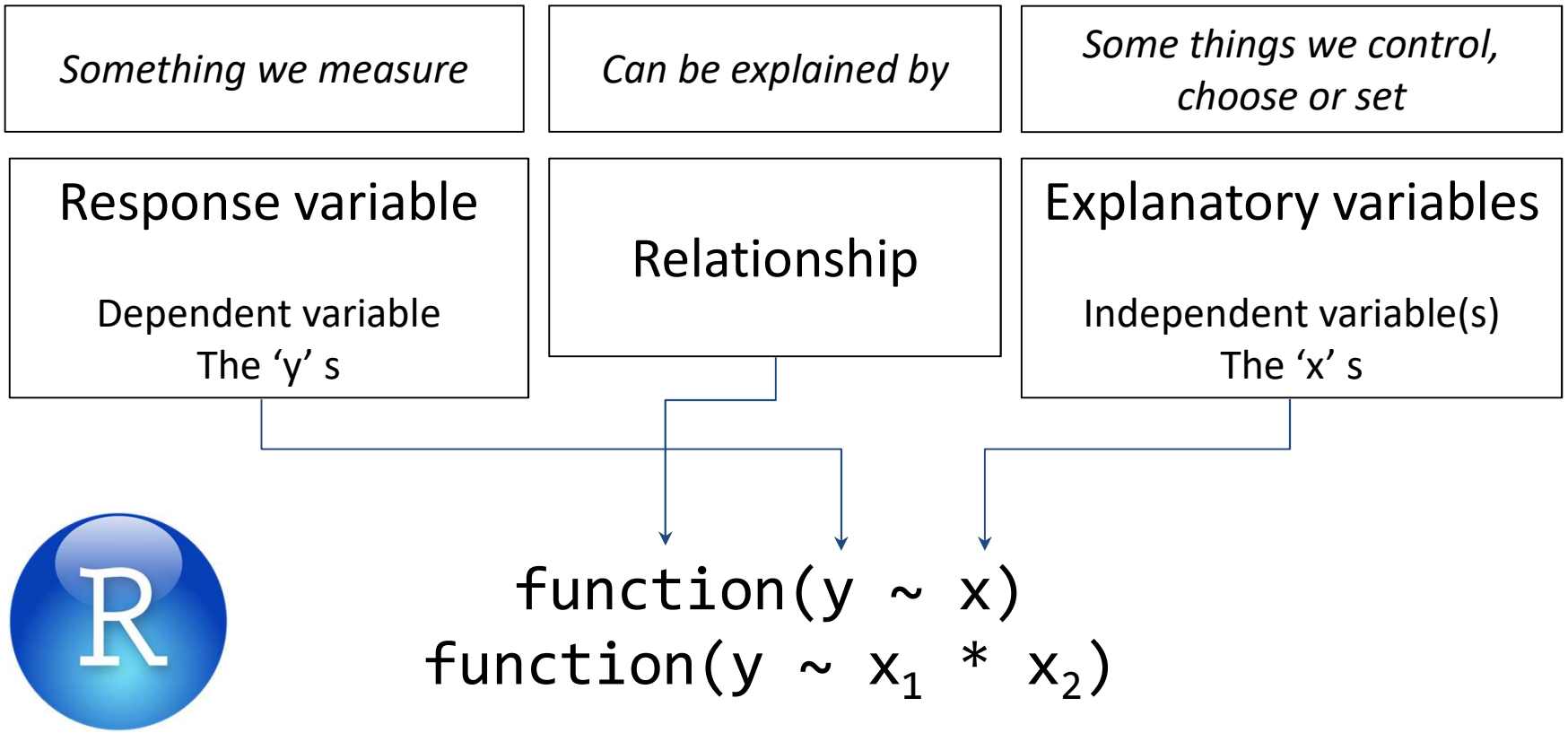
Four explanatory

Something we measure	*Can be explained by*	*Some things we control, choose or set*



x- axis
Colour
Shape
Facet

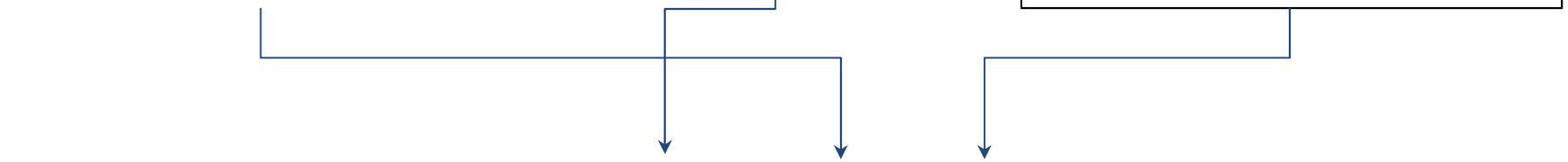
Overview of experiments and analysis



Choosing the analysis

The type of values the variables can take and the number of variables determines the test

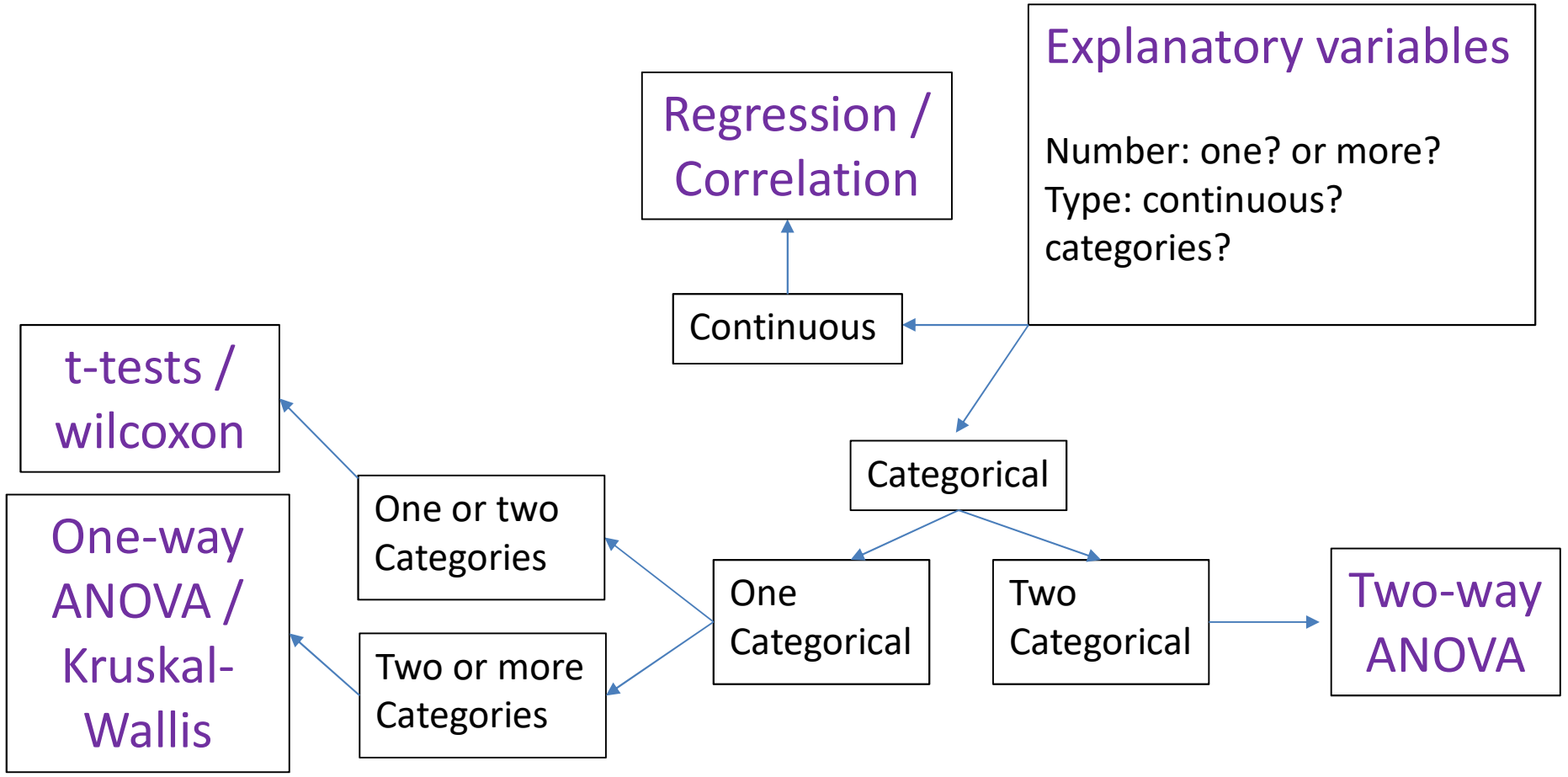
Type of Response variable	*can be explained by*	Explanatory variables
Continuous or discrete? If continuous, normal?	What relationship links the predictors to the response? Linear?	Number: one? or more? Type: continuous? categories?



```
function(y ~ x)
function(y ~ x1 * x2)
```

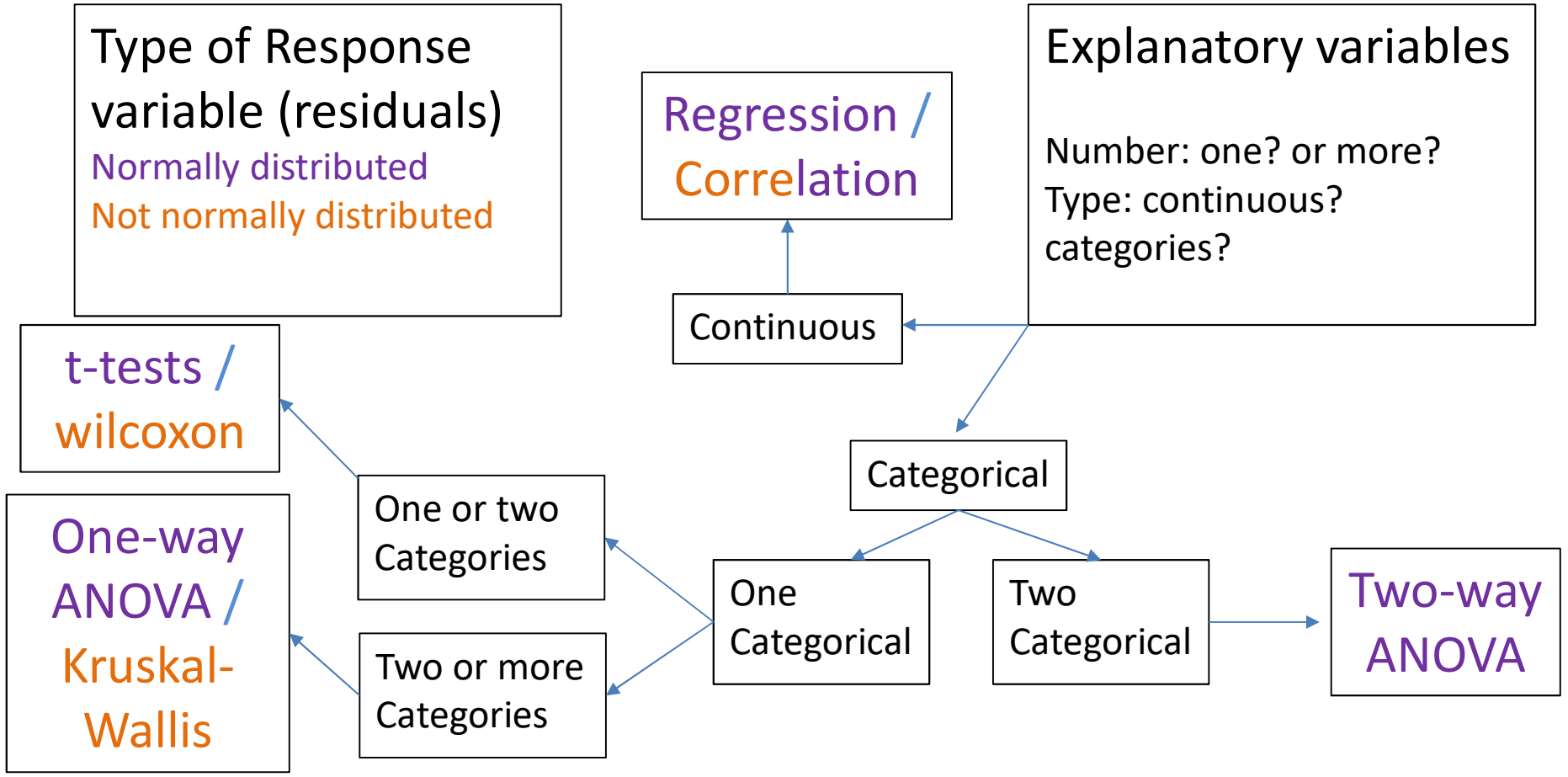
Choosing the analysis: simplification

The type of values the variables can take and the number of variables determines the test



Choosing the analysis: simplification

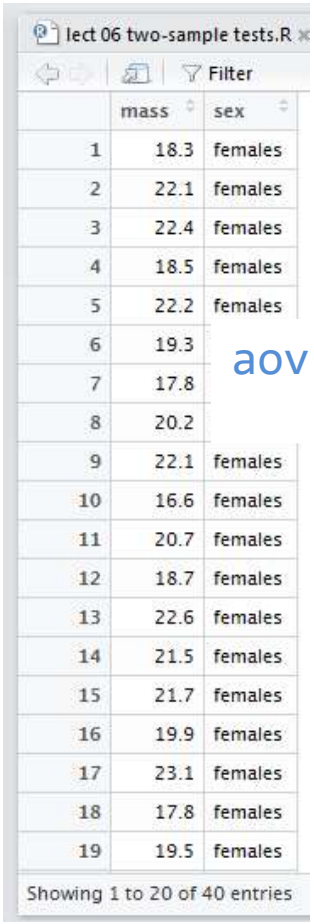
The type of values the variables can take and the number of variables determines the test



The ‘formula’ in `t.test()`, `lm()`, `aov()`, `wilcox.test()` etc, etc are COLUMNS

`aov(winglen ~ spp *sex, …)`

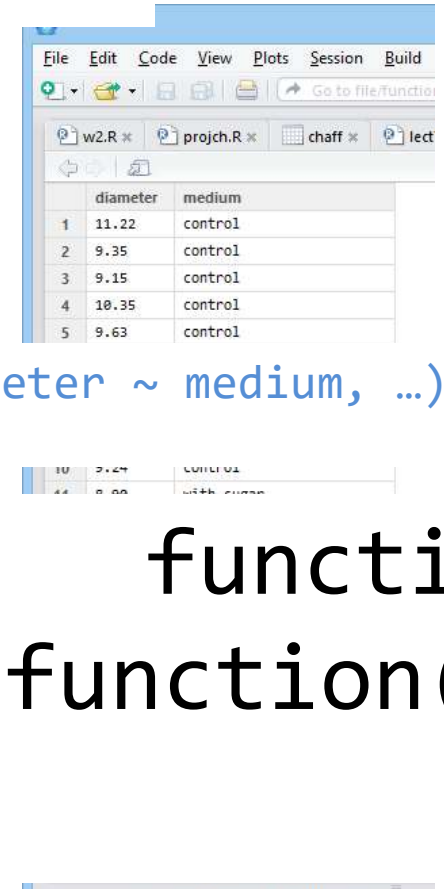
`t.test(mass ~ sex, …)`


lect 06 two-sample tests.R

Filter

	mass	sex
1	18.3	females
2	22.1	females
3	22.4	females
4	18.5	females
5	22.2	females
6	19.3	
7	17.8	
8	20.2	
9	22.1	females
10	16.6	females
11	20.7	females
12	18.7	females
13	22.6	females
14	21.5	females
15	21.7	females
16	19.9	females
17	23.1	females
18	17.8	females
19	19.5	females

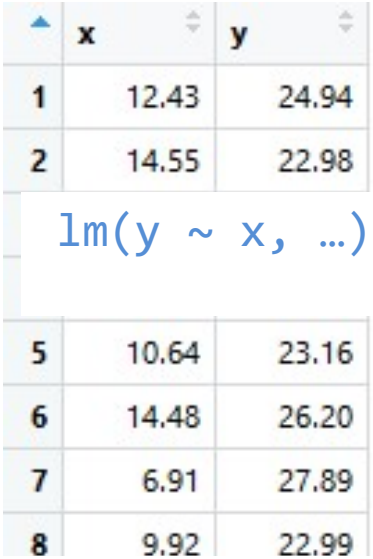
Showing 1 to 20 of 40 entries


File Edit Code View Plots Session Build

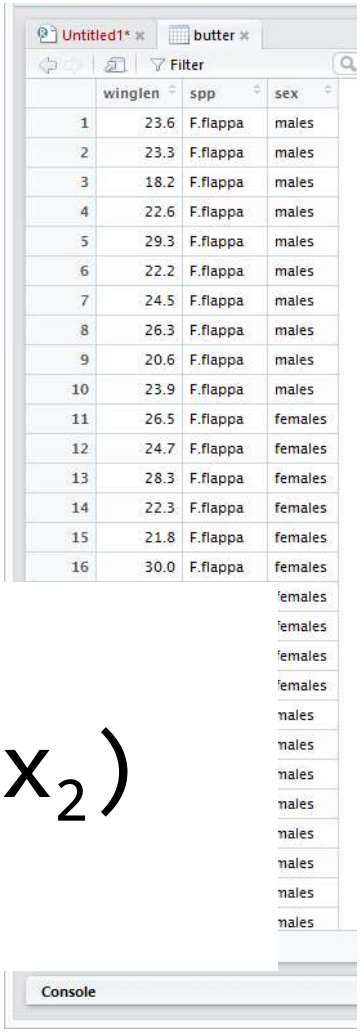
w2.R projch.R chaff lect

	diameter	medium
1	11.22	control
2	9.35	control
3	9.15	control
4	10.35	control
5	9.63	control

`aov(diameter ~ medium, …)`


	x	y
1	12.43	24.94
2	14.55	22.98
5	10.64	23.16
6	14.48	26.20
7	6.91	27.89
8	9.92	22.99

`lm(y ~ x, …)`


Untitled1* butter

Filter

	winglen	spp	sex
1	23.6	F.flappa	males
2	23.3	F.flappa	males
3	18.2	F.flappa	males
4	22.6	F.flappa	males
5	29.3	F.flappa	males
6	22.2	F.flappa	males
7	24.5	F.flappa	males
8	26.3	F.flappa	males
9	20.6	F.flappa	males
10	23.9	F.flappa	males
11	26.5	F.flappa	females
12	24.7	F.flappa	females
13	28.3	F.flappa	females
14	22.3	F.flappa	females
15	21.8	F.flappa	females
16	30.0	F.flappa	females

Console

`function(y ~ x)`

`function(y ~ x1 * x2)`

Non-parametric alternatives

- Non-parametric tests make fewer assumptions
- Based on the ranks rather than the actual data
- Null hypotheses are about the *mean rank* (not the mean)
- More conservative (less likely to be significant)
- P values maybe estimates (may generate a warning). You can't usually 'fix' that

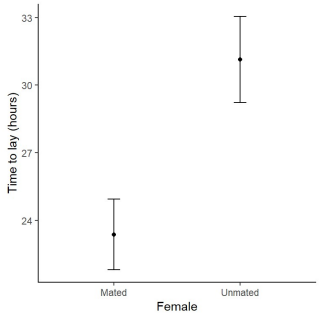
Execute the test

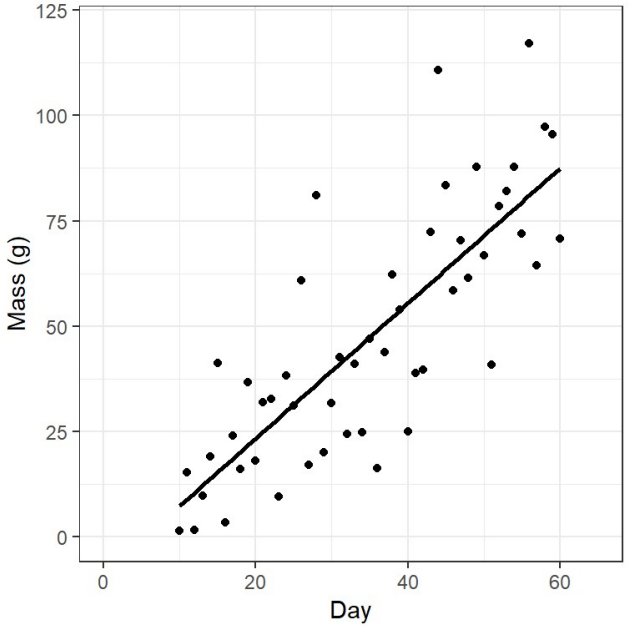
- Do results make sense based on your data exploration?
- Chose APPROPRIATE figures and/or tables
- Report results scientifically
  - Appropriate descriptive statistics
  - Results of test – significance, magnitude, direction
  - Explaining results – leave for the Discussion

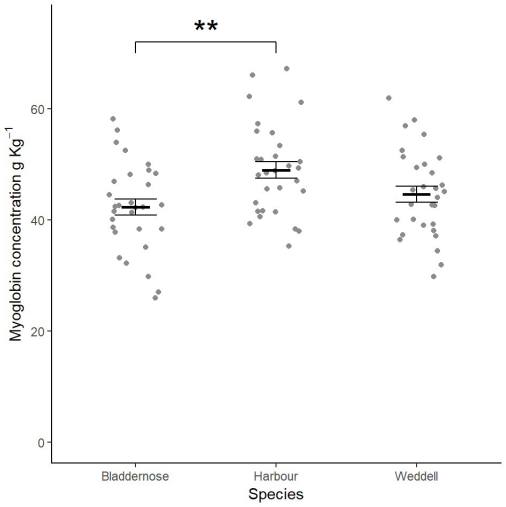
Figures

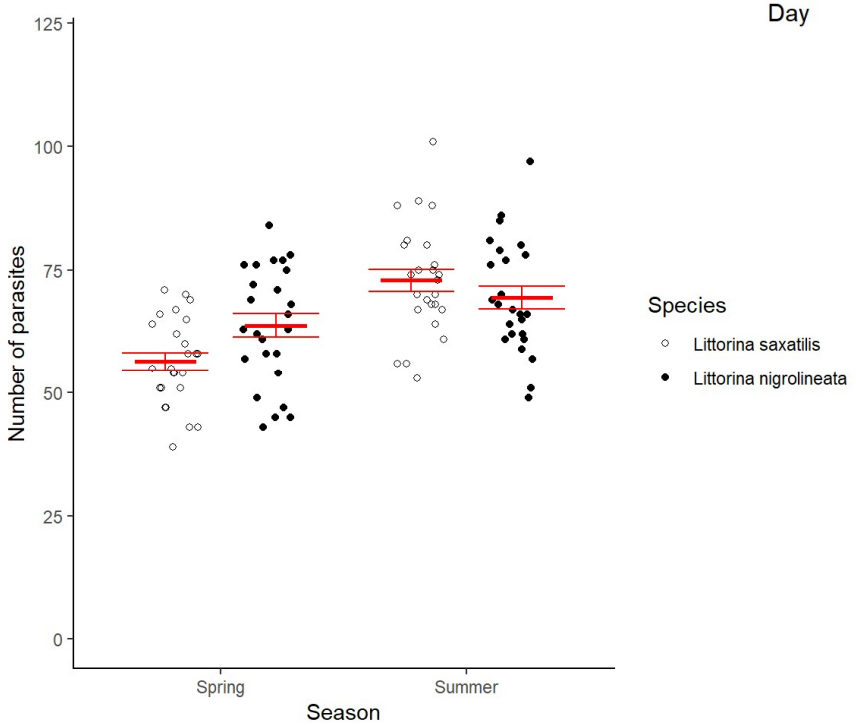
- Should match the test
  - t-tests, ANOVA tests on means thus figures show means
  - Wilcoxon, Mann-Whitney tests on ranks thus figures use medians/mean ranks
  - Correlation should NOT have line of best fit
  - Regression should have line of best fit
- Likewise descriptive statistics

Parametric

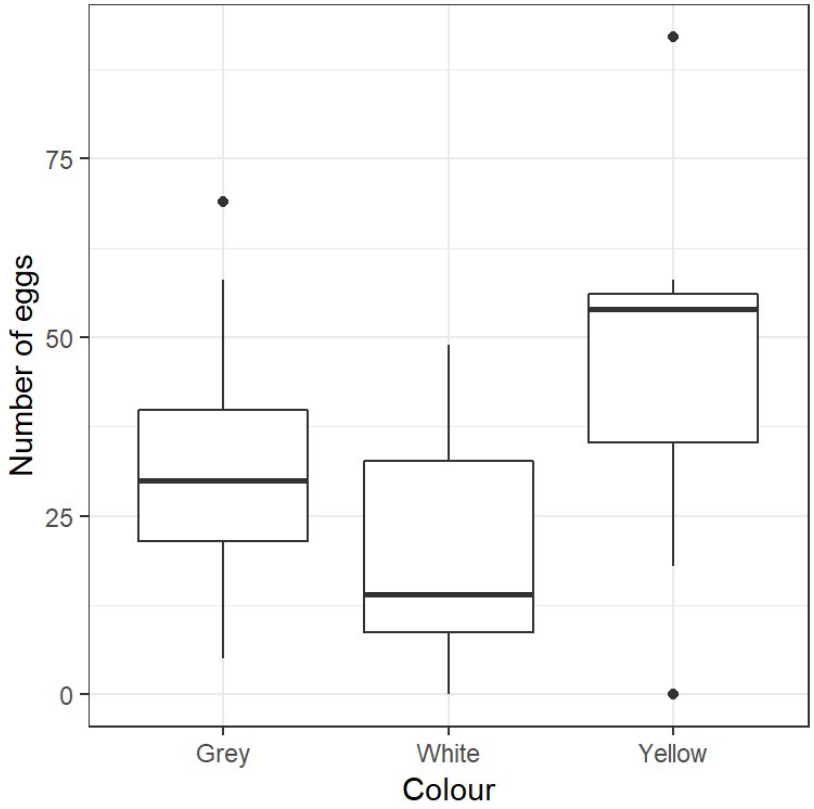

Time to lay (hours)
33
30
27
24
Mated
Unmated
Female

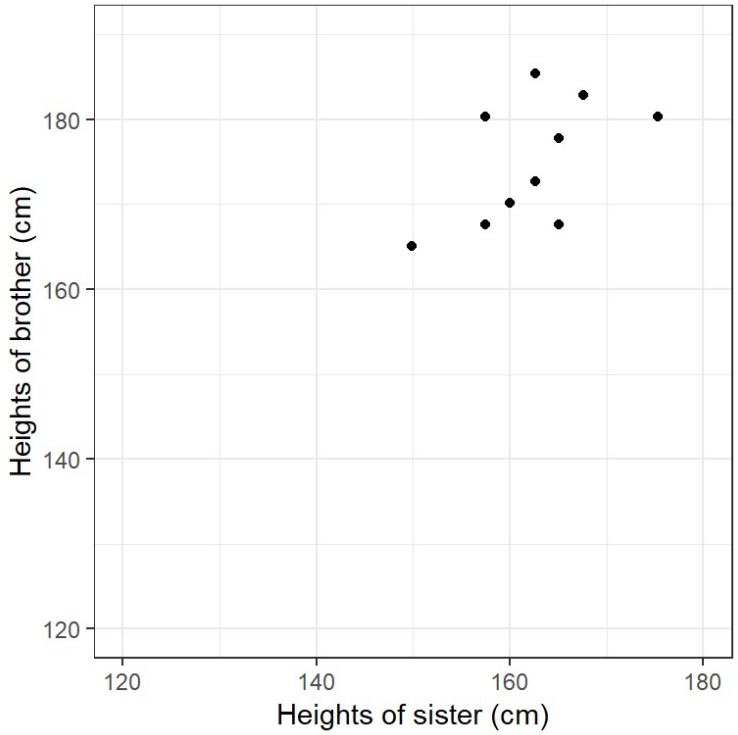

Mass (g)
125
100
75
50
25
0
0
20
40
60
Day


**
Myoglobin concentration g Kg⁻¹
60
40
20
0
Bladdernose
Harbour
Weddell
Species


Number of parasites
125
100
75
50
25
0
Spring
Summer
Season
Species
Littorina saxatilis
Littorina nigrolineata

Non-parametric





Introduction to the practical

Designed to help you:

1. think about response and explanatory variables
2. organisation of data before experiment
3. understand how tests works
4. see the links between linear models
5. practice with figures

Scenarios for biological phenomena

1. identify an appropriate design and statistical test for the general research question
2. generate the data using the random number functions that would give the effects specified.
3. create figures to accompany the results