



17C

Laboratory & Professional Skills:
Data Analysis

Emma Rand

Data Analysis in R

Week 2: Introduction to module and
to R and RStudio



Data Analysis in R Aims

To explain what matters in choosing methods of data analysis and give you practice in making those decisions.

To train you in analysing data in R specifically and help you develop an understanding of some core and highly transferable concepts in data analysis.

Module Learning Outcomes (MLO)

The successful student will be able to:

1. Explain the purpose of data analysis
2. 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!

What advice or encouragement would you give to a stage 1 student?

You might not like it, but try to like it because you're not going to ever get away from it throughout your degree

Stay and understand every workshop, they may seem really hard at the time but it will be so helpful in your future years. I left early in workshops as I found them too hard and was scared to ask for help, now in final year I'm having to play catch up.

Just get stuck in because it will really help you down the line! Once you gain confidence then it starts to become really enjoyable too!

Practise practise practise!! Just mess around in R as much as possible, understanding the content of the lectures is not enough you have to get to know R's little quirks and what writing a code is like.

GO. TO. THE. WORKSHOPS

Just give it a go!

R is widely used in top institutions, that is a good resource that will allow people to stand out, and that it is way more powerful than it appears to be.

Approach with an open mind. It will be hard if you've never used code before. Use datacamp free tutorials to help.

Rstudio can be daunting, however Emma is an expert in Rstudio and very few other biology degrees can offer the same level of training as you have available through her teaching. Taking the time to learn Rstudio will be really useful if you intend to do a year in industry or finding grad jobs requiring any kind of data analysis skills. Don't skip the workshops during your first year!

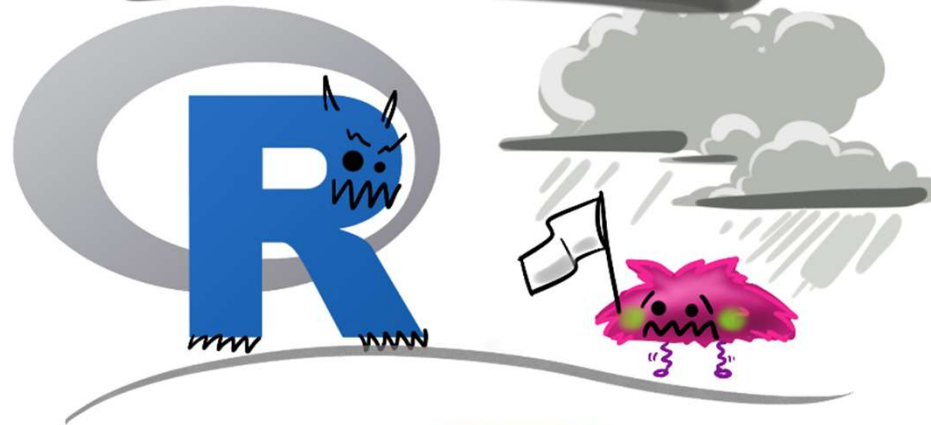
At the beginning it will seem impossible, but don't give up because the more you practice the better you will be able to solve any problems that arise. I would recommend attending all the workshops, as R isn't something you can just read about and understand. It is a lot easier to learn if you can watch someone do it and practice doing it yourself. If you think you are only one that doesn't get it, then you are wrong, because most people will feel the same. I hadn't done any coding before I came to uni and I really didn't like R to begin with, but now I can see how much easier and faster it is to analyse data with R.

All the advice and encouragement

https://docs.google.com/spreadsheets/d/1kN26o_qhIvkLVI3u-1ROawLsWOWkt7U2CGD3fpS2818/edit?usp=sharing

Don't be hungover for your first session because it'll make everything a lot more difficult than it is

at first I was like...



...but now it's like...



Organisation: Interlinked delivery

Weekly Workshop

1. Preparatory independent study
 - a) Introduction to statistical concepts
 - b) Demos: How to do in R
2. Workshop – guided practice applying
3. Follow up Independent study – problems to consolidate understanding

Warnings!

These do not stand alone – weekly L.O.

All are needed

Progressive

Overview of topics

Week	Topic
2	Introduction to module, data analysis and RStudio including first figure
3	Hypothesis testing, data types, reading data in to R and saving figures in reports
4	The normal distribution, summary statistics and confidence intervals; user-defined functions, RStudio
5	One- sample t-tests and their non-parametric equivalents
6	Two-sample t-tests and their non-parametric equivalents
7	One-way ANOVA and Kruskal-Wallis
8	Two-way ANOVA incl understanding the interaction
9	Correlation and regression
10	Chi-squared tests

Summary of this week

- We explain why we do statistical tests
- Using RStudio we learn how to use the command line, basic functions and arguments; navigate the panes; and what the workspace, scripts and history are

Learning objectives for the week

By actively following the lecture and practical and carrying out the independent study the successful student will be able to:

- explain why we need data analysis (MLO 1)
- use the R command line as a calculator and to assign variables (MLO 3)
- Create and use the basic data types in R (MLO 3)
- find their way around the RStudio windows (MLO 3)
- create, use and save a script file to run R commands (MLO 3)
- search and understand manual pages (MLO 3)

Foundations of statistical testing: Science overview

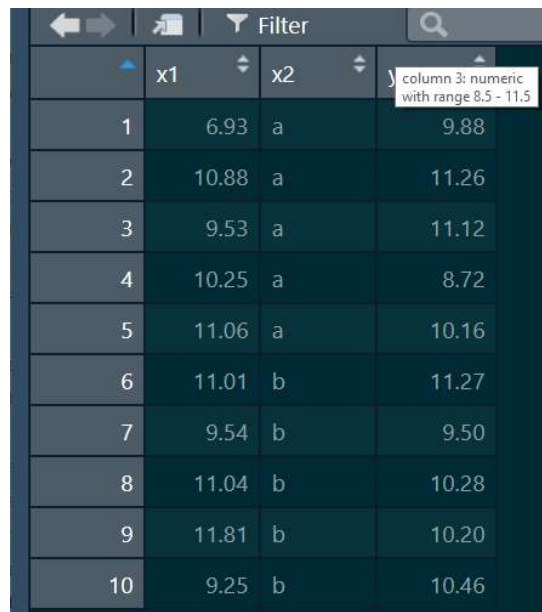
- 'Experiments'

Some things we control,
choose or set

Independent variables
Explanatory variables
The 'x' s

Something
we measure

Dependent variables
Response variables
The 'y' s



	x1	x2	
1	6.93	a	9.88
2	10.88	a	11.26
3	9.53	a	11.12
4	10.25	a	8.72
5	11.06	a	10.16
6	11.01	b	11.27
7	9.54	b	9.50
8	11.04	b	10.28
9	11.81	b	10.20
10	9.25	b	10.46

inferences made

Why do we need statistics?

If a drug reduces everybody's blood pressure by exactly the same amount we don't need statistics! The drug is effective!

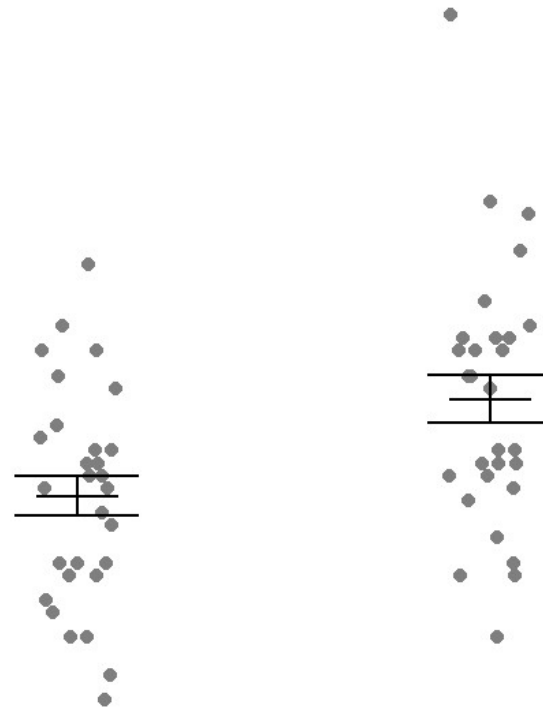
If every beetle in population A is exactly 400mg and every beetle in population B is exactly 398mg we don't need statistics! Beetles in population A are heavier!

Why do we need statistics?

- But *Responses* vary!

Is the difference
between two means
real?

Or just random
variation?



Why do we need statistics?

- *Responses* vary
- humans see patterns
- 'coincidence' can be common

The logic of 'hypothesis' testing

- Have a 'null' hypothesis': no difference
- Calculate probability of getting your data if that null hypothesis is true
- If the probability is less than 0.05 reject the null hypothesis
- Frequentist/classical statistics
- N.b. 0.05 is an agreed but arbitrary level

See you next week!