
Emma Rand

# Laboratory & Professional skills for Bioscientists
# Term 2: Data Analysis in R

## Week 2: Introduction to module 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 every get away from it throughout your degree

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

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.

Stay and understand every workshop, they may seen really hard at the time but it will but 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 give it a go!

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

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!

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_qhIvkLVl3u-1ROawLsWOWkt7U2CGD3fpS2818/edit?usp=sharing

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

# Organisation: Interlinked delivery

Weekly Workshop

1. Preparatory independent study
2. Lecture and slides – introduction to concepts
3. Workshop – apply concepts
4. Follow up Independent study - consolidate understanding

Warnings!

These do not standalone – weekly L.O.

All are needed

Progressive

# Learning Outcomes

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

# 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 | Chi-squared tests |
| 5 | The normal distribution, summary statistics and confidence intervals; user-defined functions, RStudio |
| 6 and 7 | One- and two-sample t-tests and their non-parametric equivalents (2 lectures) |
| 8 | One-way ANOVA and Kruskal-Wallis |
| 9 | Two-way ANOVA incl understanding the interaction |
| 10 | Correlation and regression |

# Other sources of information

- Your previous work!!!
- The help pages
- Online – see the VLE!
- Google
- Practice!!

RStudio


R Graph Catalog

stackoverflow

CrossValidated

Cookbook for R

R-bloggers
R news and tutorials contributed by (573) R bloggers

*"Young man, in mathematics you don't understand things. You just get used to them."*

# Summary of this week

- We explain why we do statistical tests

- Using DataCamp and 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:

- to explain why we need statistical tests and the logic of hypothesis testing (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 → Something we measure

Independent variables
Explanatory variables
The ‘x’ s

| | 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 |

Dependent variables
Response variables
The ‘y’ s

inferences made

# Why do we need statistics?

- *Responses* vary
- We see patterns
- ‘coincidence’ can be common

- The birthday problem

If you are in a group of 25 people, what is the probability of sharing a birthday?

# Why do we need statistics?

- *Responses* vary; 'coincidence' can be common
- Logic of statistical testing: Birthweight
  - National average is 3300 grams with an s.d. = 900
  - Do mothers who live in poverty have babies of average birthweight?
  - Take a sample: 25 mothers who live in poverty has a mean of 3000 grams
  - What should you conclude about the effect of poverty?

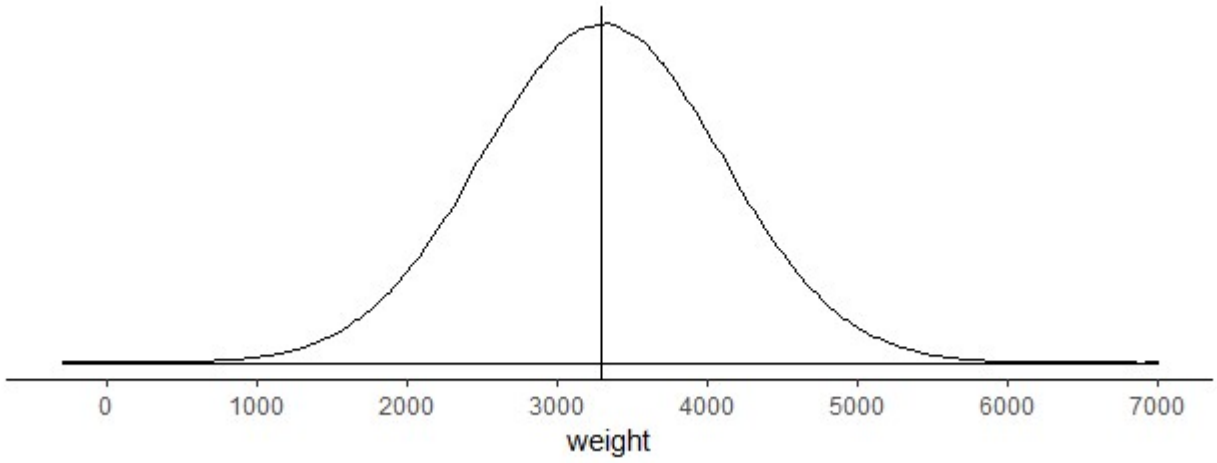
# The logic of ‘hypothesis’ testing

- Do our results (3000 g) seem likely if mean weight is 3300g?
- 3300 g (the National average) is the ‘no effect hypothesis’
- “The null hypothesis”

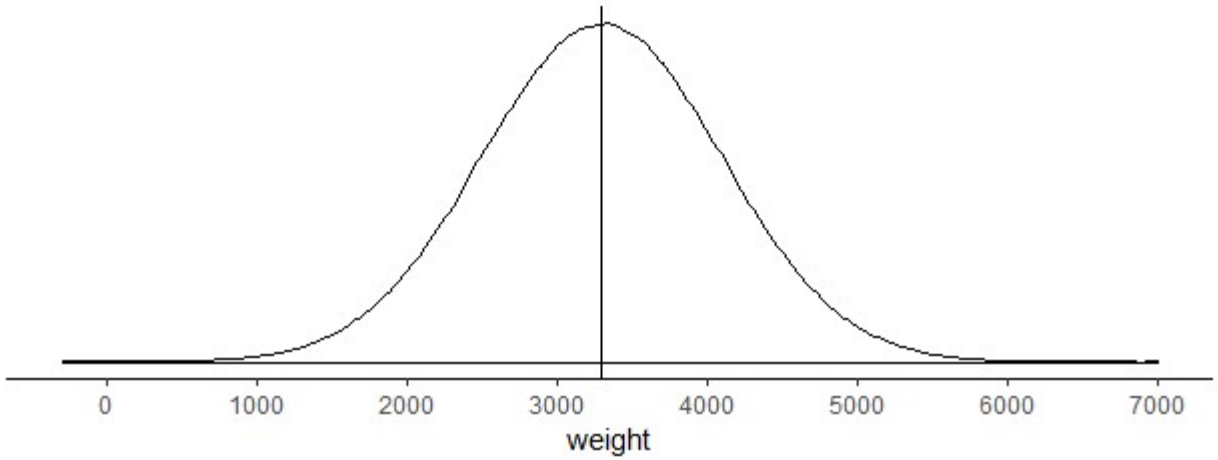
# The null hypothesis. $H_0$

What you expect to happen if nothing interesting biologically is occurring.

We would expect a mean of 3300 if poverty has no effect.



# The null hypothesis. $H_0$




This does not mean every baby is 3300 g or every sample of babies has a mean of 3300 g.

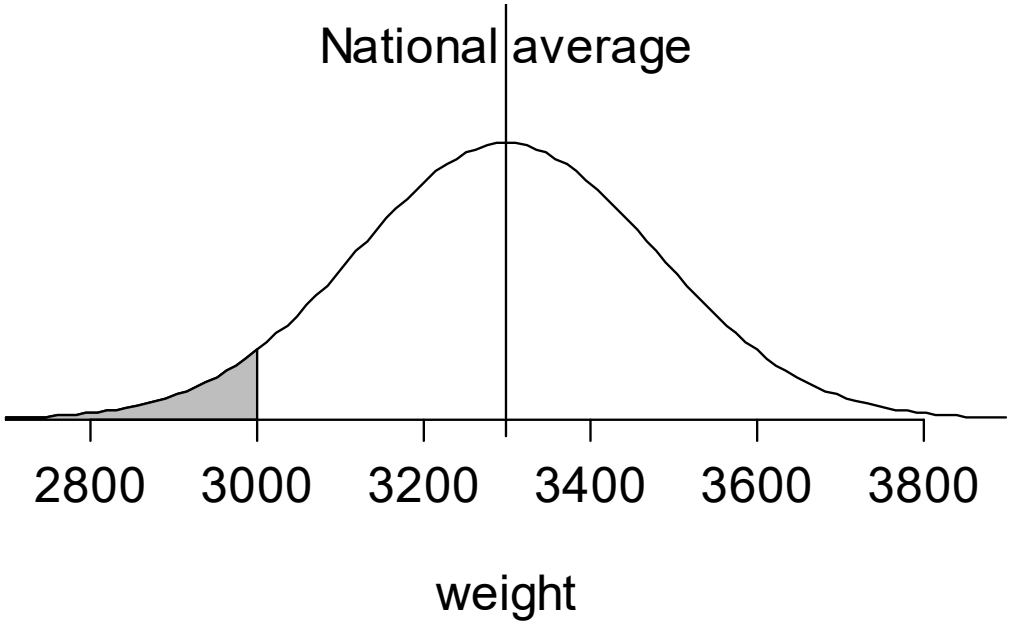
It means we wouldn't expect a sample mean to be "too far" from 3300 g

# How far is too far? Distributions

We calculate the probability of 3000 g if we expect 3300 g on average

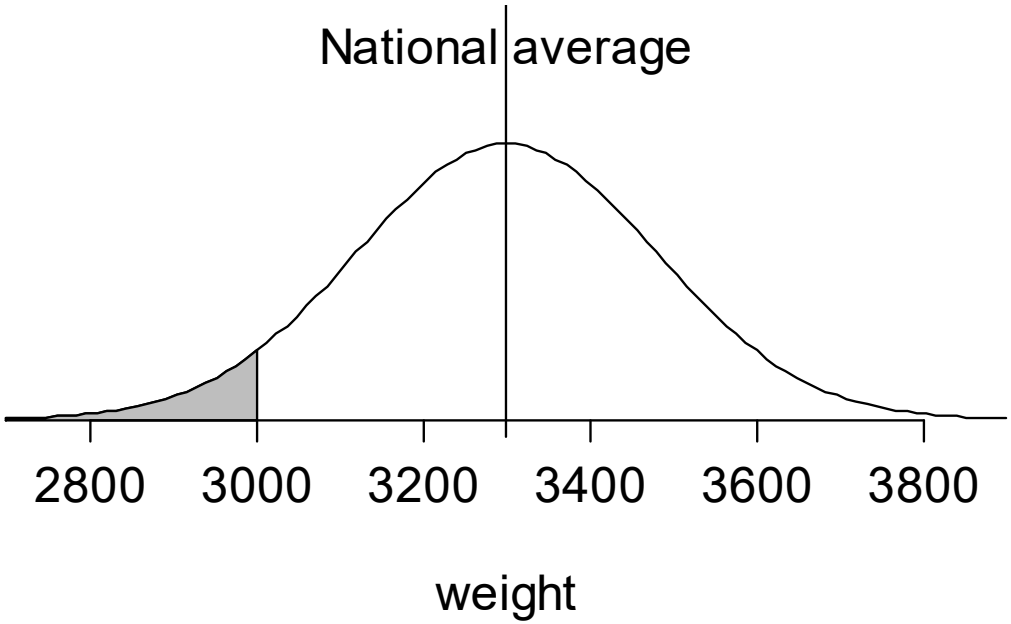
What is P(3000) *or lower* from a distribution with mean 3300

$p = 0.048$



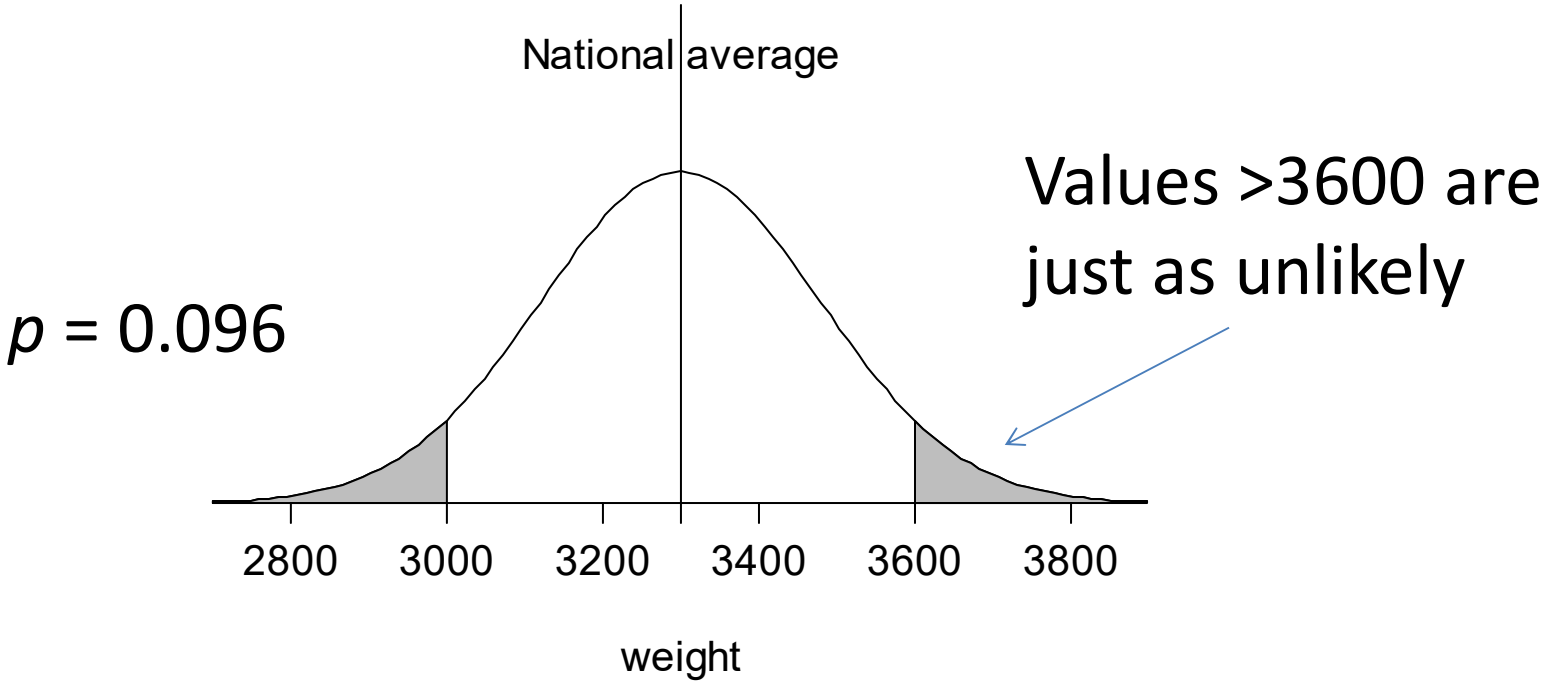

Don’t worry how p was calculated

# Possible, but not likely



# But more appropriately

What is P(3000) or a mean *as unlikely or more unlikely* from a distribution with mean 3300?



- 0.096 is probability of a sample like this if poverty has no effect on birthweight
- How far is too far? i.e., Is that a big or small probability?
- Use 0.05 (1 in 20)

# 0.05 is the (arbitrary) significance level used

- If probability of our result or one as extreme or more extreme is ≤ 0.05 we REJECT our null hypothesis.

$p \leq 0.05$, test is significant

- If probability of our result or one as extreme or more extreme is > 0.05 we DO NOT REJECT our null hypothesis.

$p > 0.05$, test is not significant

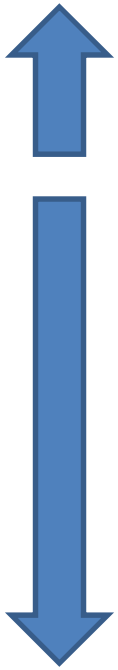
# The logic of ‘hypothesis’ testing

- Have a ‘null’ hypothesis’
- 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

# 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:



- to explain why we need statistical tests and the logic of hypothesis testing (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)