University of BRISTOL

Statistical Computing and Empirical Methods

MSc Data Science

Teaching block 1, 2021

Henry W J Reeve

henry.reeve@bristol.ac.uk

Unit EMATM0061

Statistical Computing and Empirical Methods

Dr. Henry WJ Reeve

Lecturer in Statistical Science in the School of Mathematics

Research interests in Machine Learning & High-Dimensional Statistics

Unit director on Statistical Computing and Empirical Methods

Email: henry.reeve@bristol.ac.uk

Subject including: EMATM0061

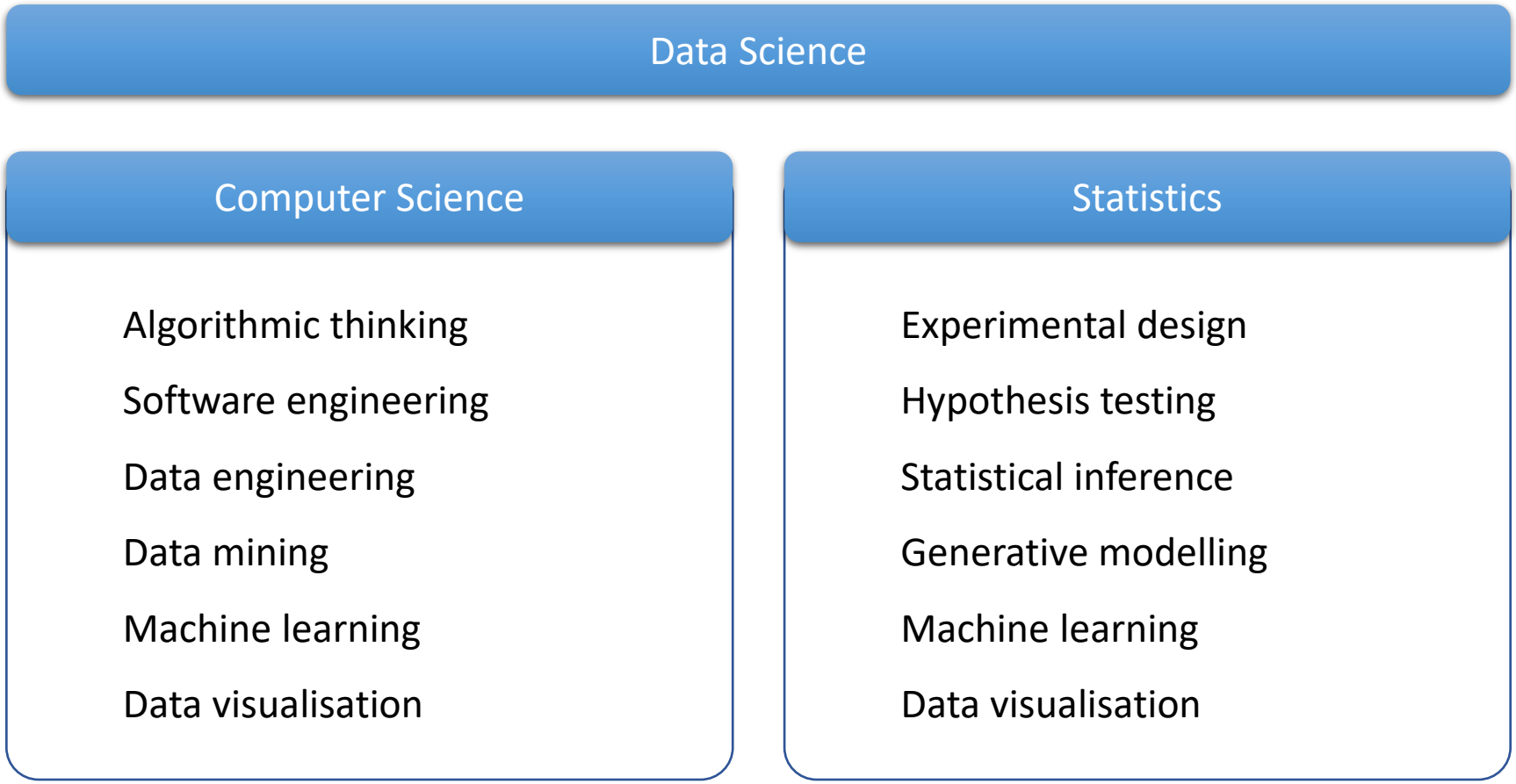
Teaching assistants

Dominic Owens works on change point methods and time series models for multivariate and high-dimensional data.

Jake Spiteri works on nonparametric methods for latent variable models using reproducing kernel Hilbert spaces..

What is Data Science?

Data Science is the science of extracting information, insight and understanding from data!



Why learn Data Science?

Data Science is a fascinating field which can lead the way to an intellectually rewarding career.

- Pharmaceuticals

- Finance

Why learn Data Science?

Data Science is a fascinating field which can lead the way to an intellectually rewarding career.

- Retail

- Marketing

Why learn Data Science?

Data Science is a fascinating field which can lead the way to an intellectually rewarding career.

- Sport

- Academia

Why learn Data Science?

Data Science is a fascinating field which can lead the way to an intellectually rewarding career.

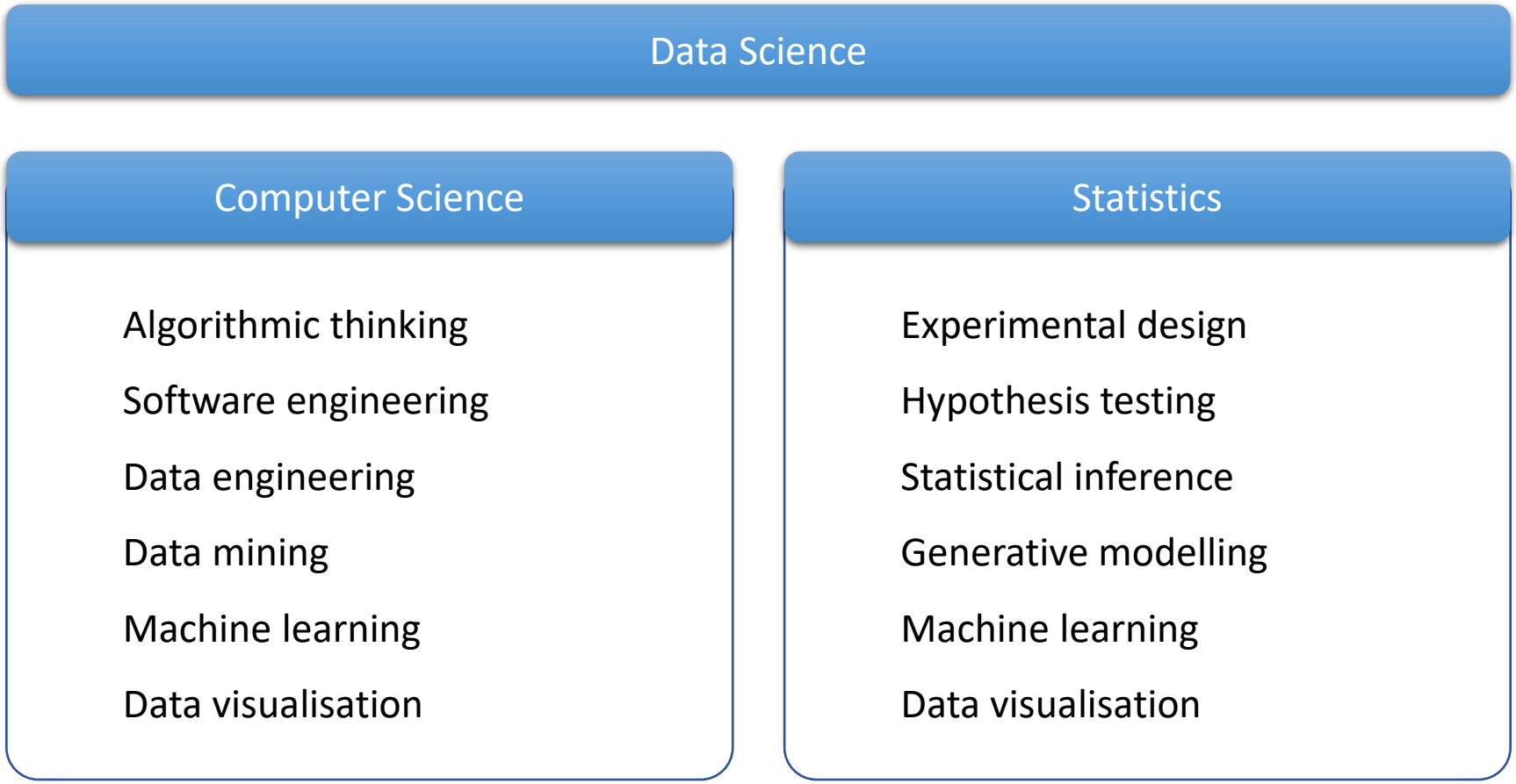





What motivates you?

What is Data Science?

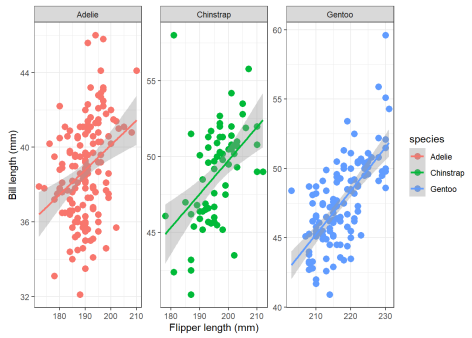
Data Science is the science of extracting information, insight and understanding from data!

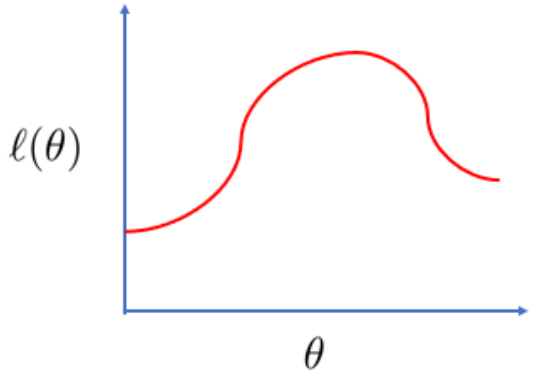


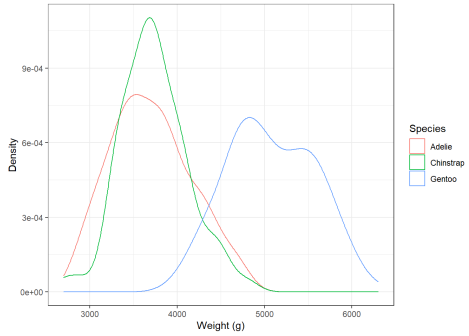
Statistical computing and empirical methods

Objective:

Gain a broad understanding of the fundamental statistical principles and methods necessary for a successful career in Data Science.







Statistical computing and empirical methods

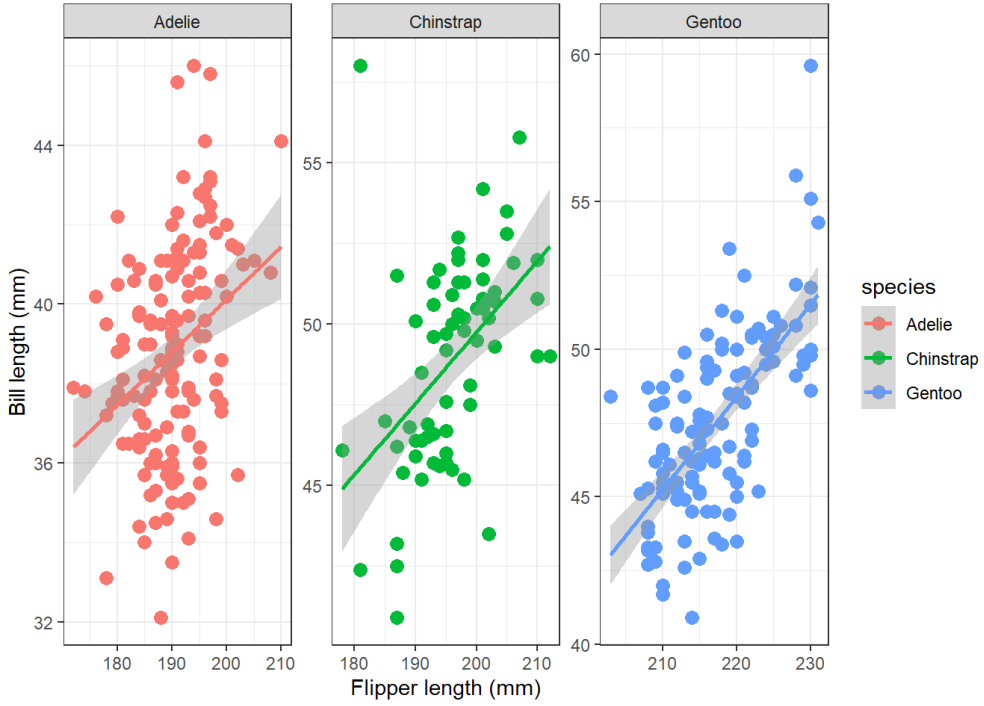
We shall use the **R programming language** and environment:

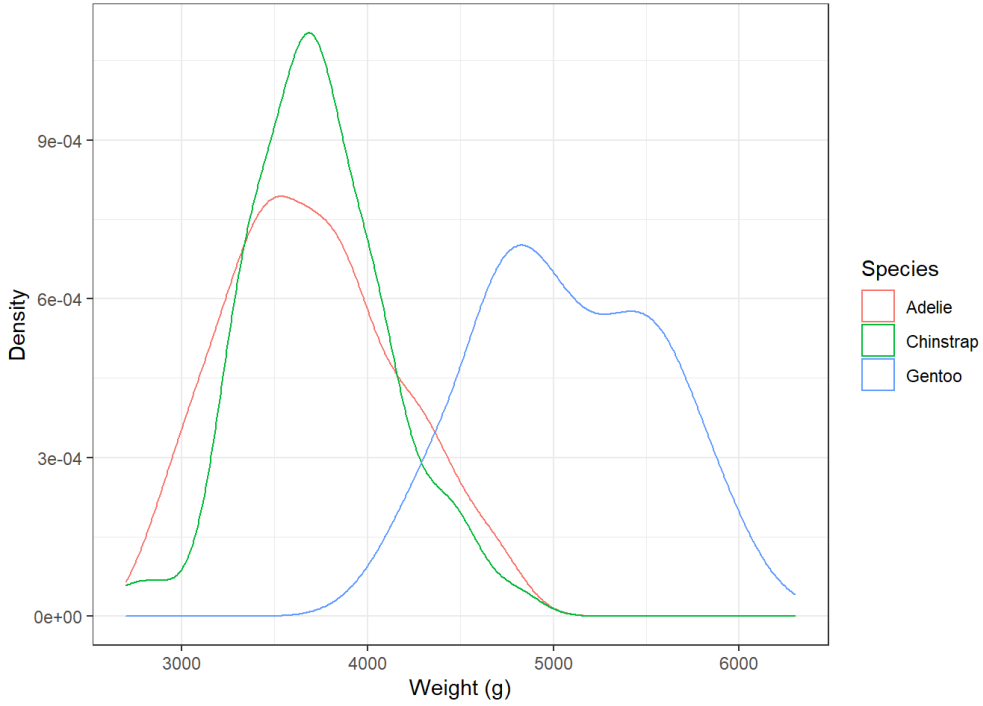
- A vast eco-system of open source tools for statistical computing.
- A rich and diverse community of R enthusiasts spanning industry and academia.
- Straightforward interfaces with other languages;
- Other approaches are available! e.g. Python and Julia;
- The primary focus of this course will be on transferable concepts.

Data visualisation

Data visualisation is crucially important:

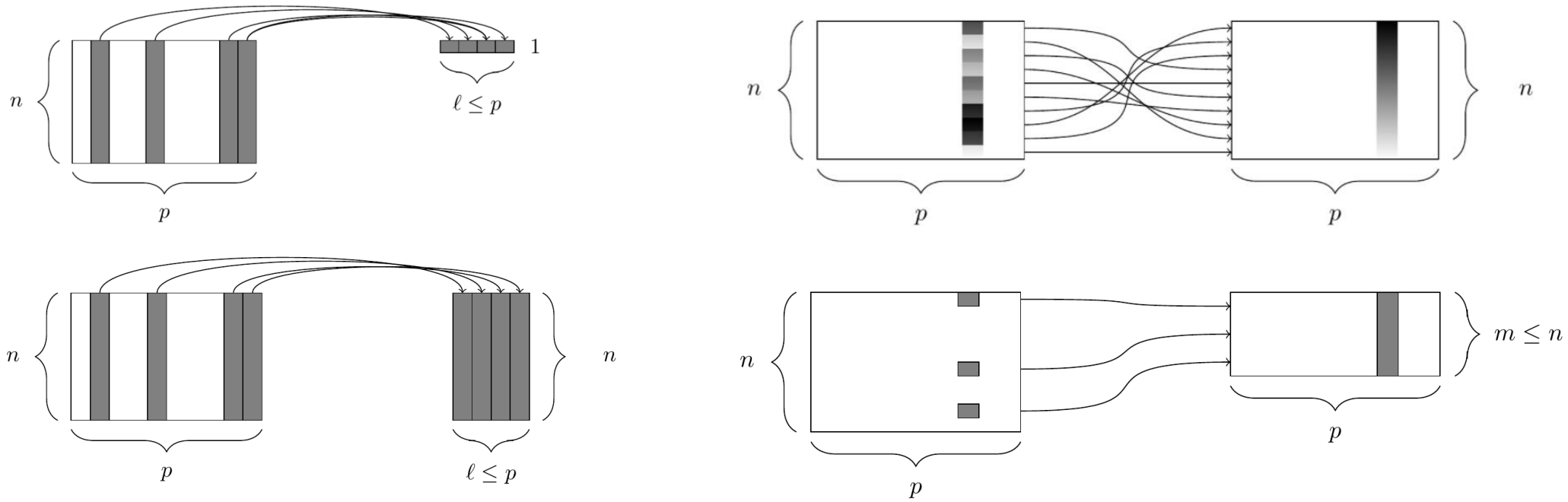
- Exploring your data and gaining preliminary insights.
- Communicating your analysis to your colleagues and clients.





Data wrangling

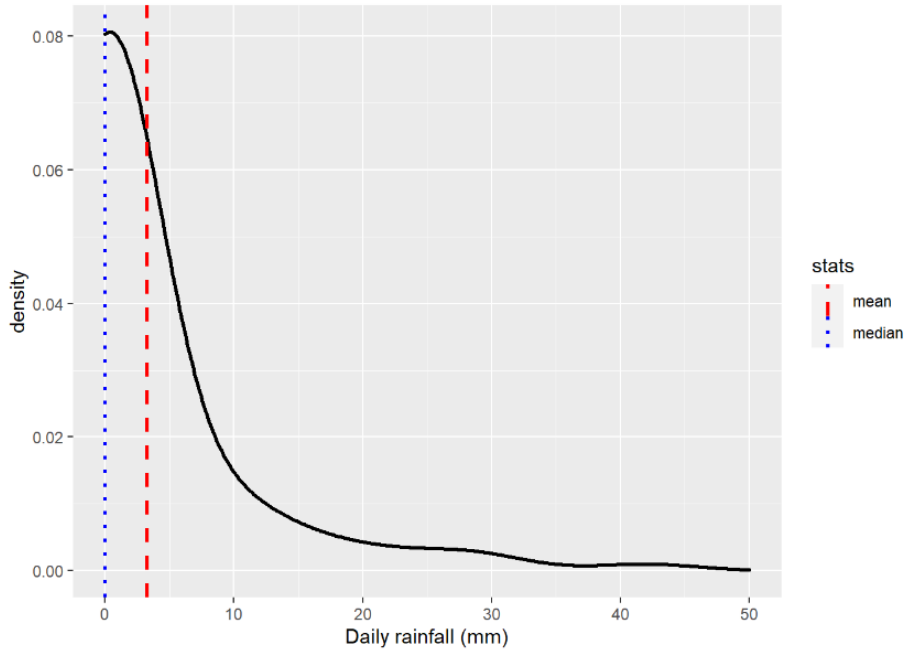
Data wrangling is a crucial skill which involves transforming data from one form to another in preparation for another down-stream task: Reshaping, rearranging, merging, selecting, filtering and aggregating data.

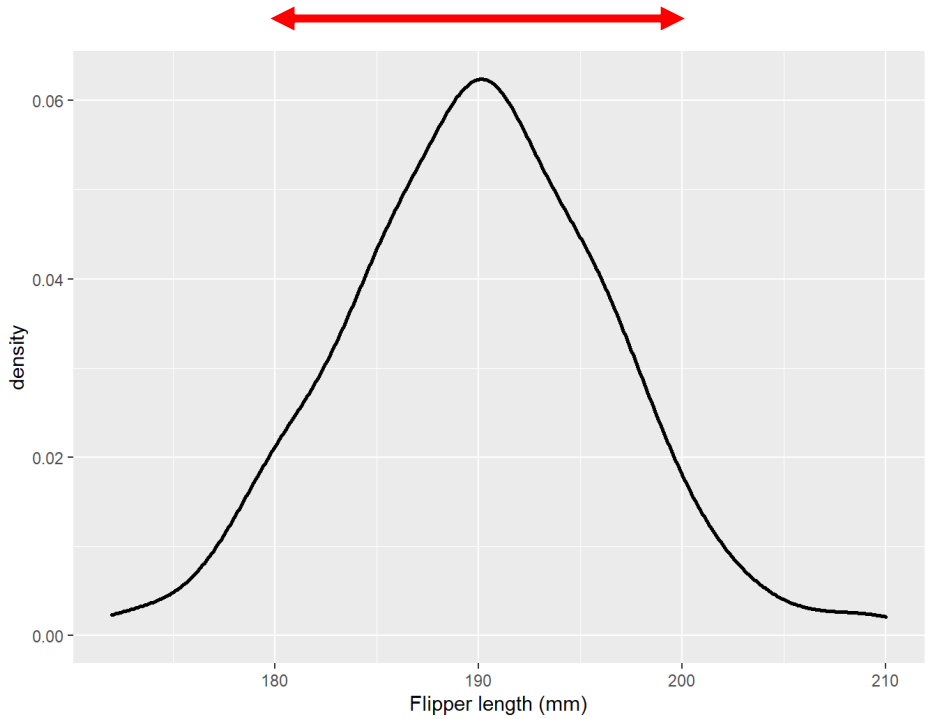


Diagrams from Baumer et al. Modern Data Science with R, 2017.

Sample statistics

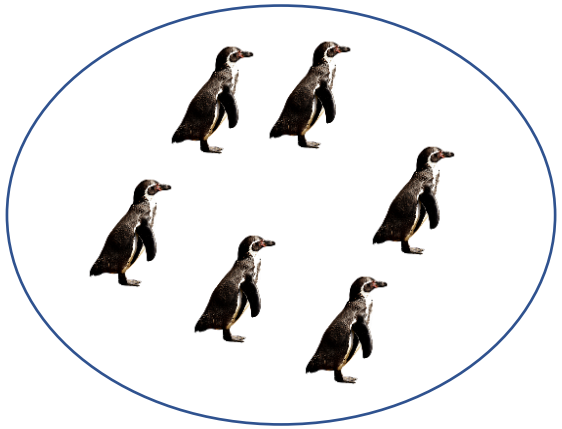
We will learn about the basic data types, see how sample statistics give us useful summary information about our data and discuss the concept of outliers.

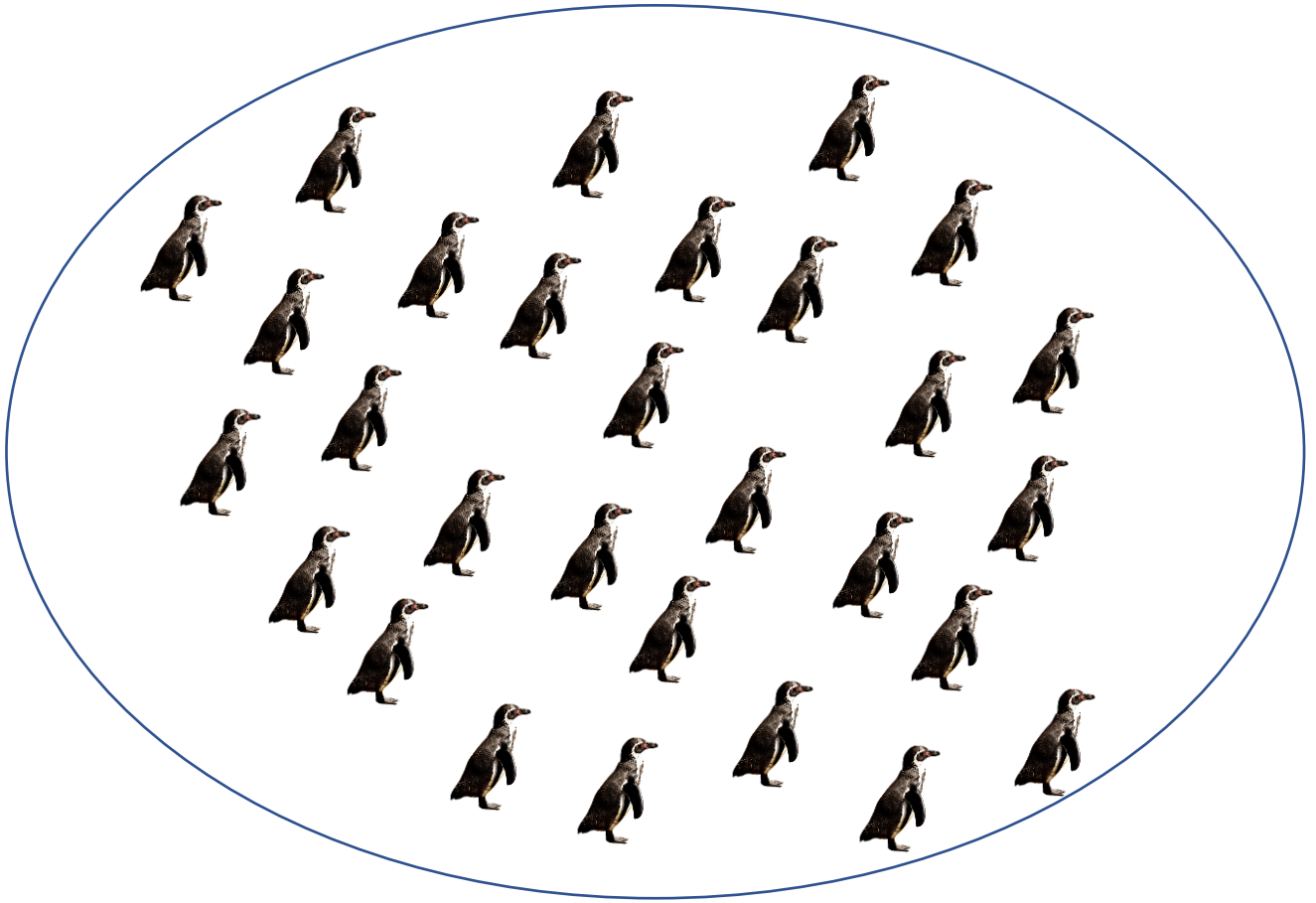




Probability theory

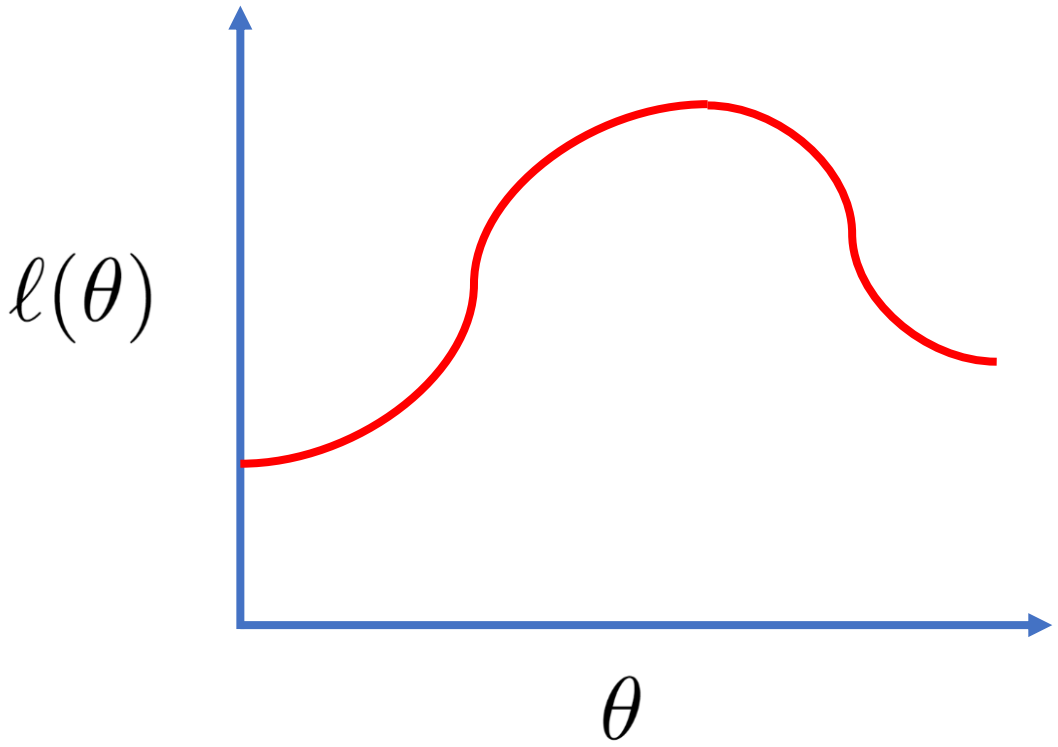
We will use probability theory to understand the relationship between sample statistics, random samples and the underlying populations they represent.





Statistical estimation with maximum likelihood

We will introduce fundamental concepts from the theory of statistical estimation such as bias, variance and the maximum likelihood paradigm.



Statistical hypothesis testing

Hypothesis testing provides a rigorous statistical methodology for deciding whether or not we have sufficient information to reject a hypothesis in favor of a suitable alternative.

Experimental design

Through careful experimental design we can increase the likelihood that the statistical properties of our data sample provide meaningful information concerning the research question of interest.

Classification

Classification models are functions which assign a feature vector to a corresponding category. Machine learning provides an array of algorithms which learn a classification model based on a labelled data set.






Regression

Regression models are functions which map a feature vector to a real number. Machine learning provides an array of algorithms which learn a regression model based on a labelled data set.

Statistical computing and empirical methods

Objective:

Gain a broad understanding of the fundamental statistical principles and methods necessary for a successful career in Data Science.

- The R programming language
- Data visualisation
- Data wrangling
- Exploratory data analysis
- Statistical estimation
- Hypothesis testing
- Confidence intervals
- Experimental design
- Classification
- Regression

Summative assessment

Objective:

Gain a broad understanding of the fundamental statistical principles and methods necessary for a successful career in Data Science.

Summative assessment:

You will write a report demonstrating your understanding of these concepts.

This report will take the form of a R markdown document and will count 100% towards your final grade.

Regulations on plagiarism, extenuating circumstances and late submission policies can be found on the central Blackboard page for the Data Science MSc.

More details to follow!

Asynchronous video lectures

Every week there will be several video lectures.

These can be viewed at a time of your choosing, but it is important to do so before the computer labs.

Video lectures can be found in Blackboard within the “Recordings” section.

Before the first computer lab on Wednesday the 29th of September please aim to watch the following two lectures:

“Lecture 1: Introduction to R and R Studio”;

“Lecture 2: Reproducible Data Science”.

However, if you don’t have time to watch both lectures just watch as much as you can.

Assignments

- Assignments will be provided throughout the course;
- These are not mandatory and will not count towards your final grade;
- However, completing these assessments will develop your skill set and improve your understanding;
- We will work through the assignments during the weekly computer labs;
- Assignments will be made available before the lab via the “Assignment” tab within Blackboard.

Weekly computer labs

Every week there will be a computer lab

These will be held in rooms 1.07 & 1.08 within the Merchant Venturers Building

Lectures will be held every Wednesday (except reading week) starting 29th September from 9:00-12:00.

You will work through your weekly assignments in this lab.

During the computer labs I and the two teaching assistants (Dom & Jake) will provide assistance.

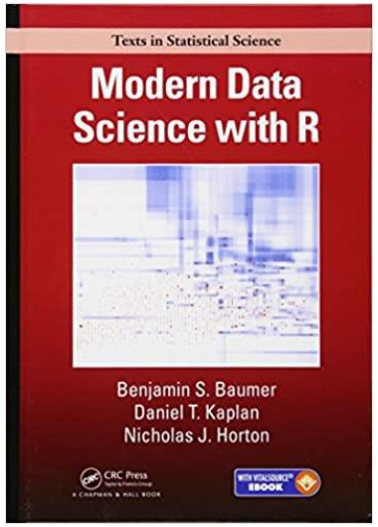
Attendance for the computer labs is optional.

However, it is strongly recommended that you attend at least the first part of the computer lab.

Recommended reading

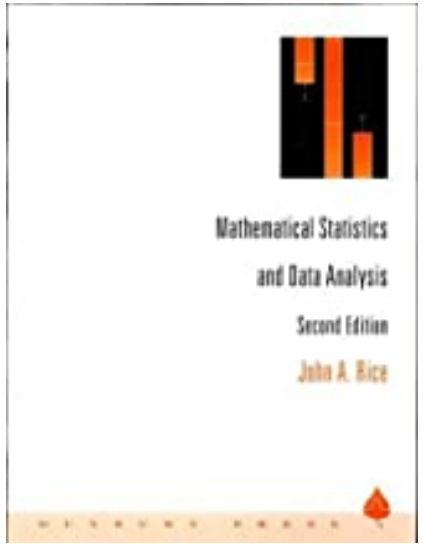
Modern Data Science with R

By Baumer, Kaplan & Horton



Mathematical Statistics and Data Analysis

By John Rice



Other books can be found in the Resource List on Blackboard.

Discussion boards

Course cafe

Introduce yourself to your fellow students

Post interesting links which you think might be relevant to the course.

Ask a question

Ask questions about the course.

Answer your colleagues questions whenever you can.

Please be careful to always be polite and considerate on the Blackboard discussions.

The Data Science MSc Blackboard

You should all have access to the central Black board page for the Data Science MSc:

On Blackboard go to EMAT Data Science MSc.

Time management: Information about how to manage your time effectively on the MSc.

Late submissions Information about the universities policies on late submission of assessed work.

Plagiarism Information about the universities policies on plagiarism.

Extenuating circumstances: Information about university policies on extenuating circumstances.

Wellbeing

If you are experiencing a difficult time and would like to talk to someone within SCEEM it is recommended that you talk to your Academic Personal Tutor in the first instance.

Bristol University Wellbeing Services

Website: www.bristol.ac.uk/students/support/wellbeing/
Email: wellbeing-access@bristol.ac.uk

University of BRISTOL

Thanks for listening!

Henry W J Reeve

Any questions to: henry.reeve@bristol.ac.uk

With subject including: EMATM0061

Statistical Computing & Empirical Methods