Statistical Computing and Empirical Methods

An Introduction

Unit EMATM0061, Data Science MSc

Rihuan Ke

rihuan.ke@bristol.ac.uk

Teaching Block 1, 2024

University of BRISTOL

About me

Unit director of *Statistical Computing and Empirical Methods*

Dr. Rihuan Ke

Lecturer in Statistical Science

Research interests: Machine learning, Image analysis

contact: rihuan.ke@bristol.ac.uk

What is Data Science?

The science of
- extracting information, insight and understanding from data
- using the extracted knowledge to help decision making

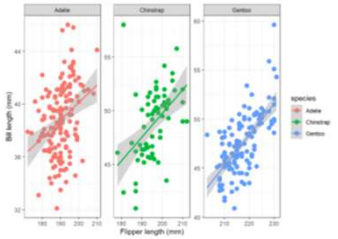
Data Science

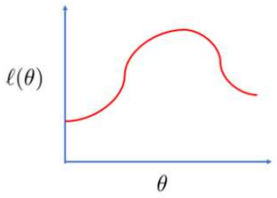
Computer science	Statistics
Algorithmic thinking	Experimental design
Software engineering	Hypothesis testing
Data engineering	Statistical inference
Data mining	Generative modelling
Data visualisation	Sampling theory
…	…

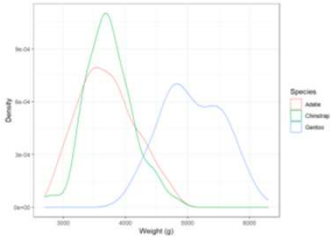
Statistical computing and empirical methods

Main Objective:
To gain a broad understanding of the fundamental statistical principles and methods necessary for a successful career in data science

Topics: data visualisation, data wrangling and data exploration, statistical significance testing, parameter estimation, experimental design, classification, regression analysis, …



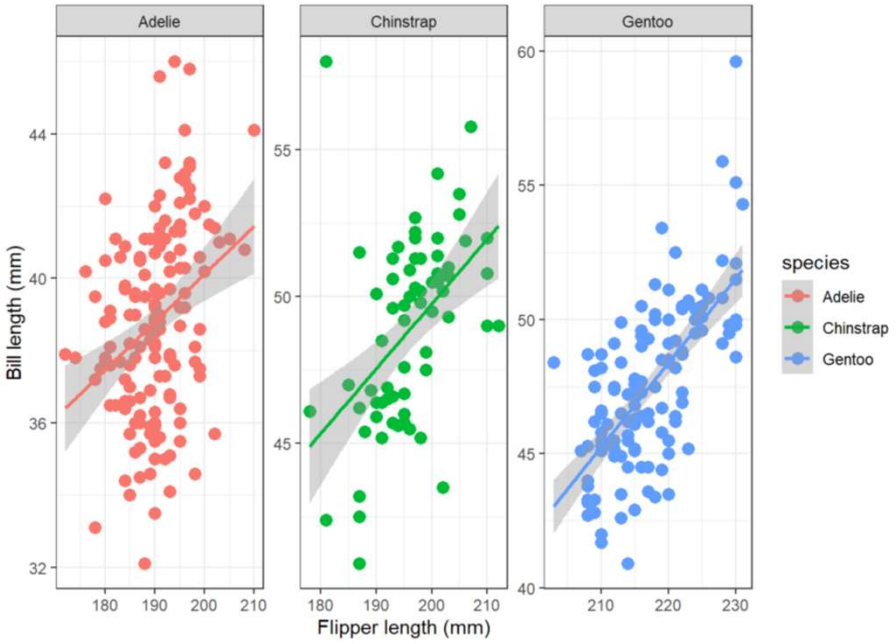


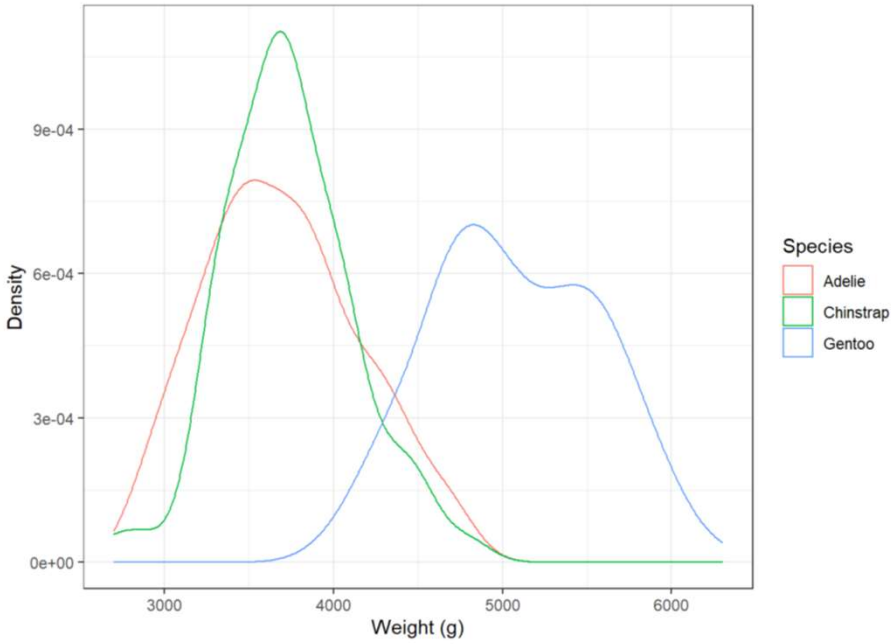


Data visualisation

Data visualisation is crucially important:

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

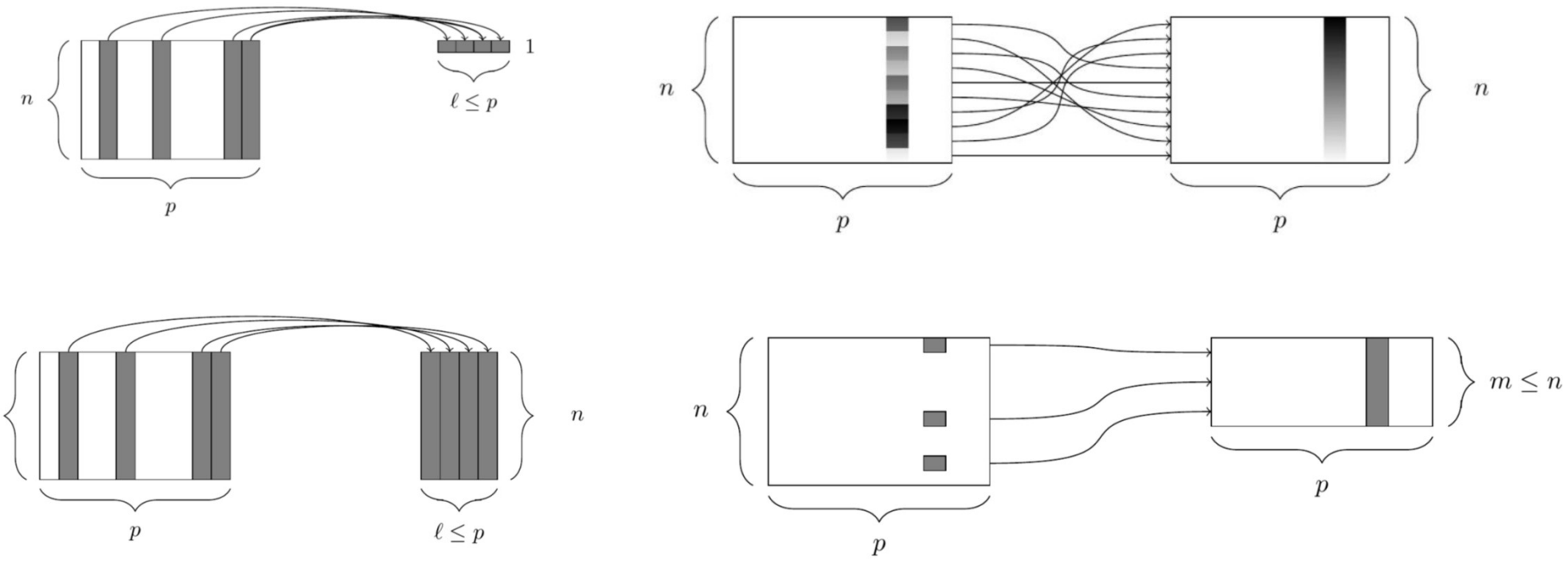




Data wrangling

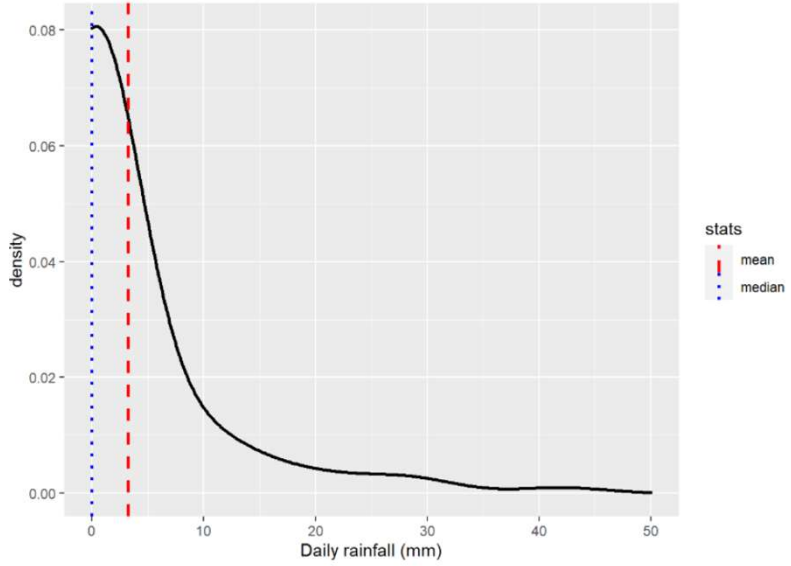
Data wrangling is a crucial skill which involves transforming data from one form to another in preparation for another downstream task:

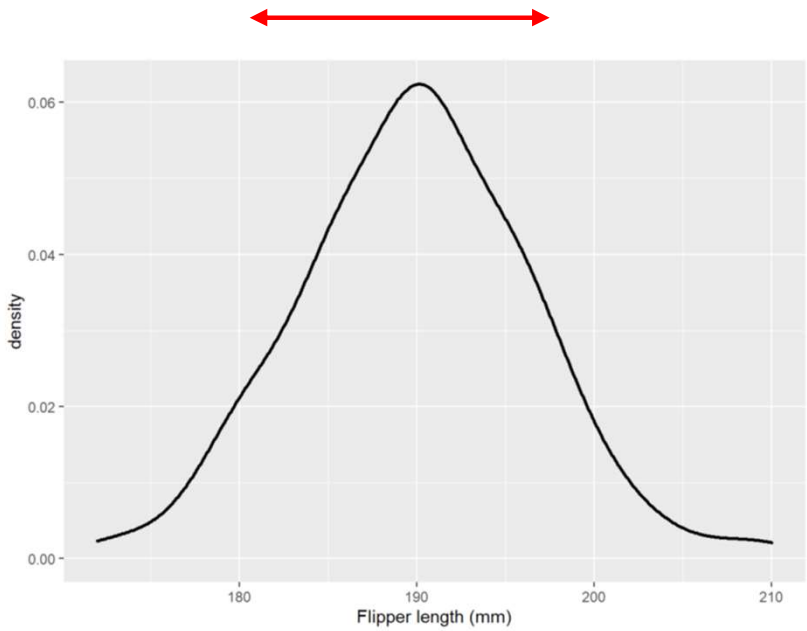
- Reshaping, rearranging, merging, selecting, filtering and aggregating data…



Sample statistics

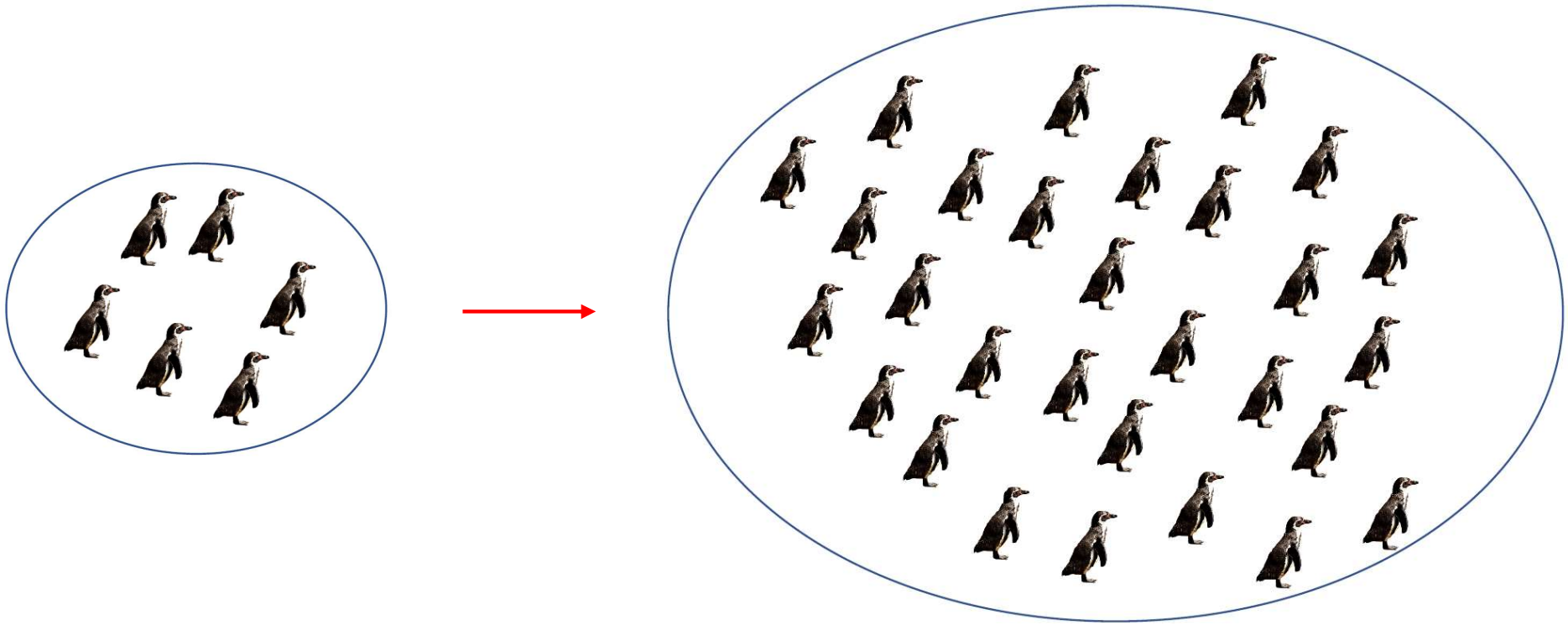
We will learn about the basic data types, see how sample statistics give us useful summary information about our data and discuss the concept of outliers.





Probability theory

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

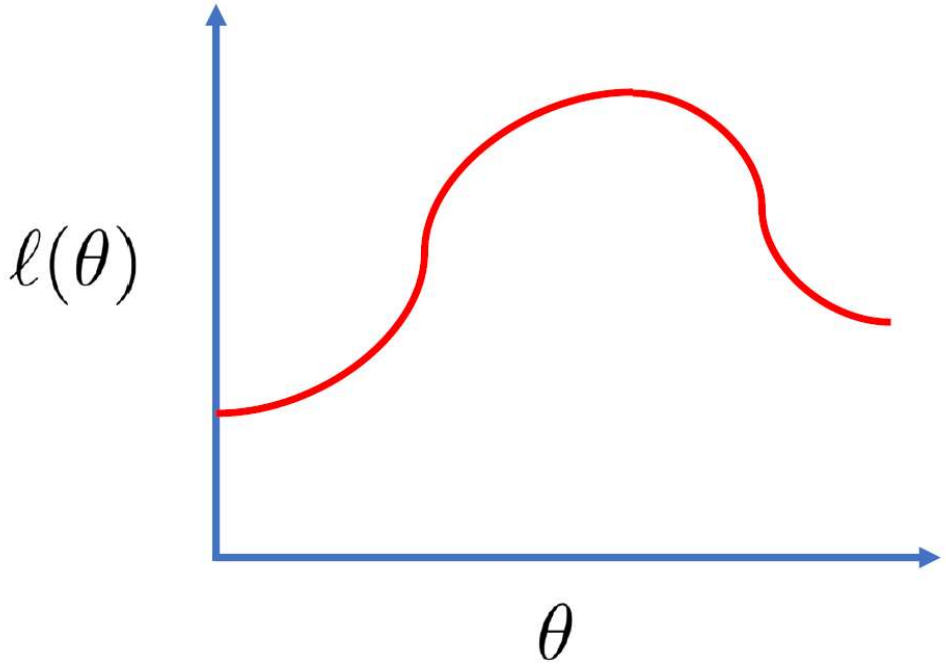


sample

population

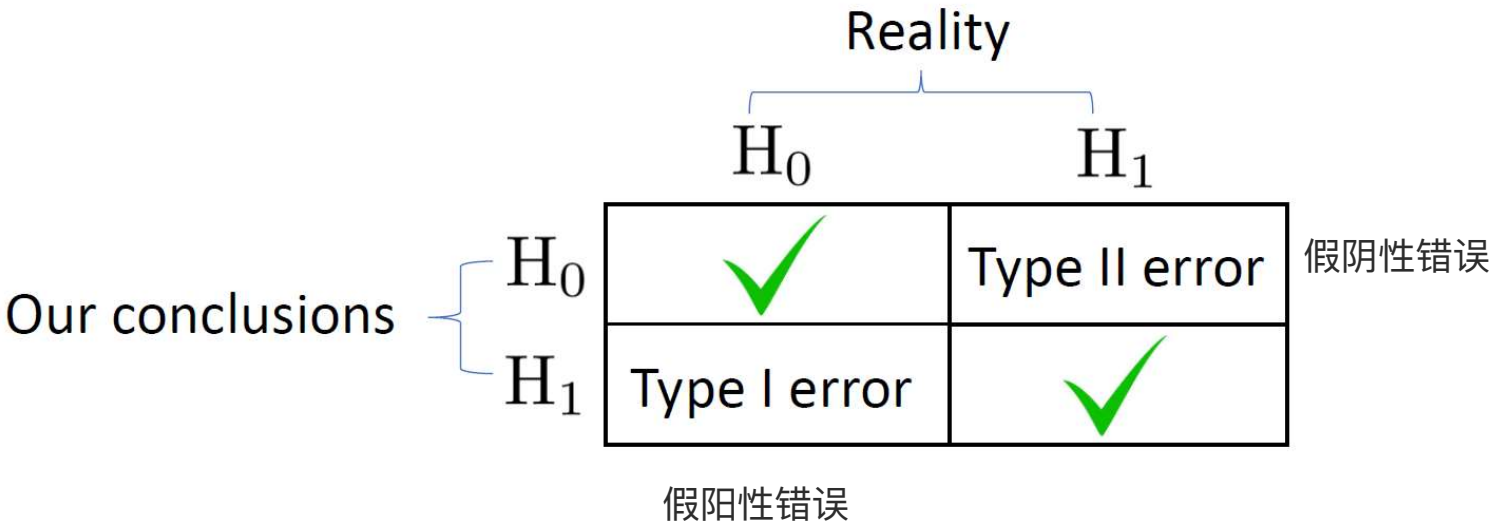
Statistical estimation

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



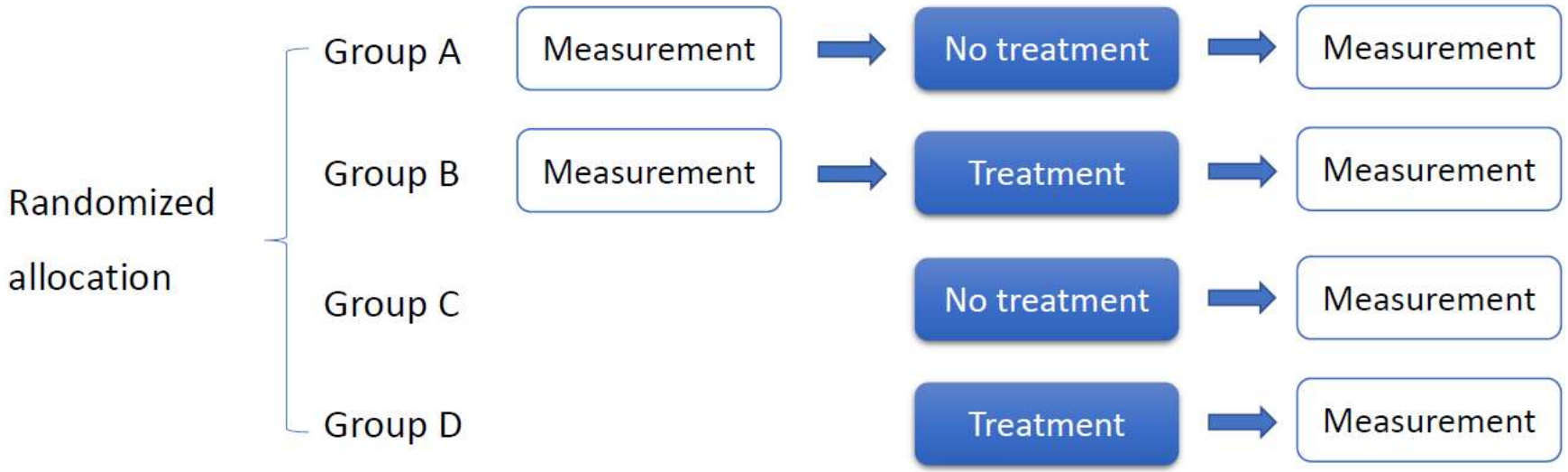
Statistical hypothesis testing

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

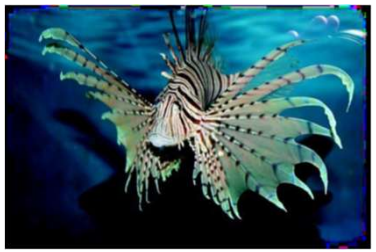





Regression

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.







A fish image → weight (a real number)

R programming

We shall use the R programming language and environment:

- A vast ecosystem of *open-source tools and packages* 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.

Statistical computing and empirical methods

Main Objective:
To gain a broad understanding of the fundamental statistical principles and methods necessary for a successful career in data science

- *Data visualisation*
- *Data wrangling,*
- *Sample statistics*
- *Probability theory*
- *Statistical estimation*
- *Statistical hypothesis testing*
- *Experimental design*
- *Classification*
- *Regression*
- *R programming*

Unit teaching & assessment plan

Asynchronous video lectures

In-person lectures

Assignments

Computer labs

} Every week (except the reading week)

Summative assessment

Lectures and computer labs

Each week there will be several video lectures

They can be viewed at a time of your choosing, but it is recommended to do so before the in-person lectures and computer labs

In the labs, you will work through your weekly assignments.

During the computer labs the teaching assistants and I will provide assistance. You can ask questions.

Attendance for the computer labs is optional.

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

Assignments, summative assessment

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

The summative assessment, in the form of a coursework, will count 100% towards your final grade

You will write a report demonstrating your understanding of the main concepts covered by the unit

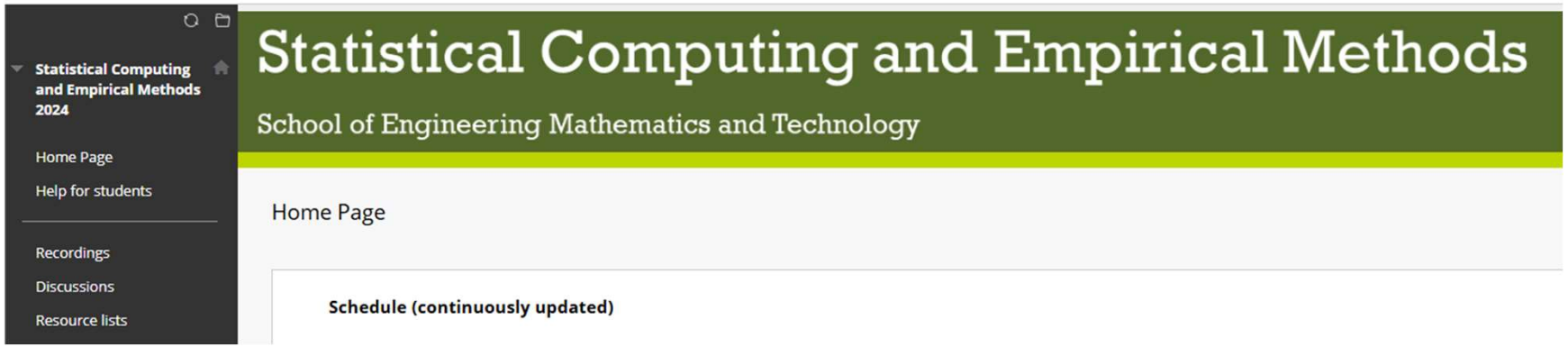
Regulations on plagiarism, extenuating circumstances and late submission policies, from Blackboard

Course webpage at Blackboard

Visit our blackboard page to access

- video lectures
- assignments
- computer lab sessions info (where & when)
- schedules
- announcements

Blackboard page visit blackboard: *https://www.ole.bris.ac.uk*

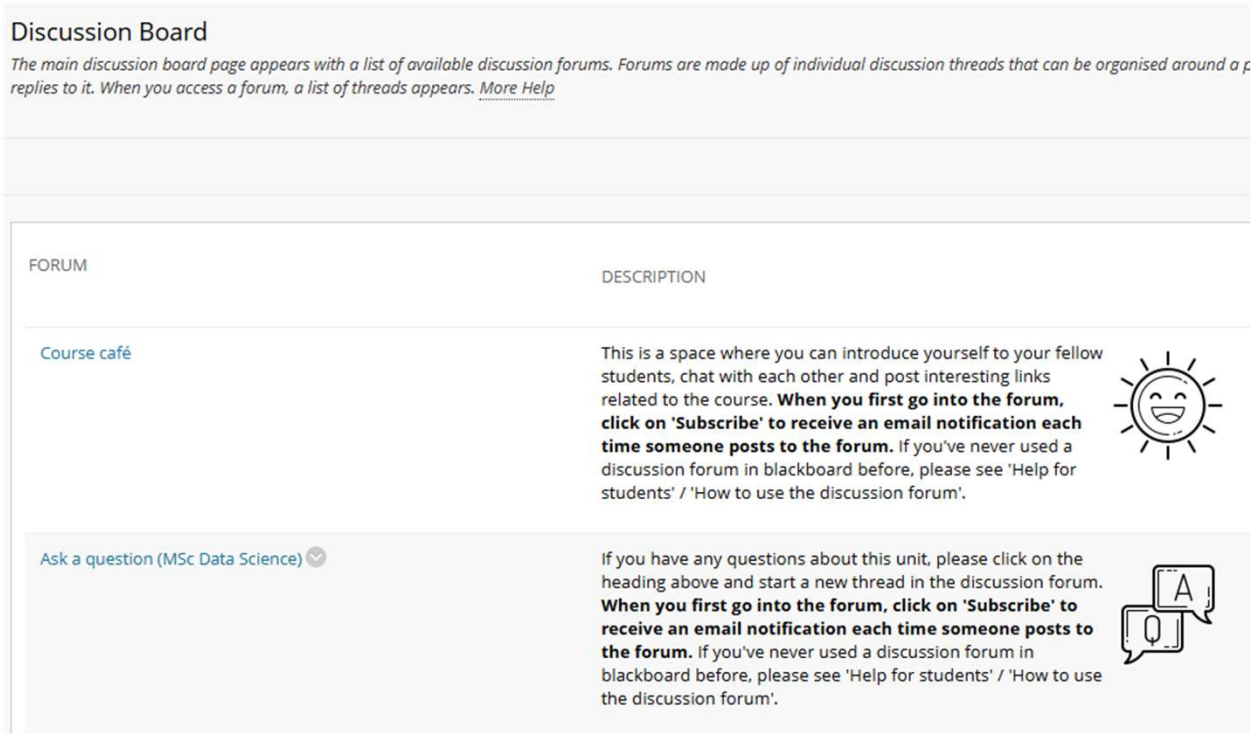


Blackboard "Discussion Boards"

Course café

Available at the "Discussion" tab of the course webpage

- Have a chat with your fellow students
- Post interesting links which you think might be relevant to the course



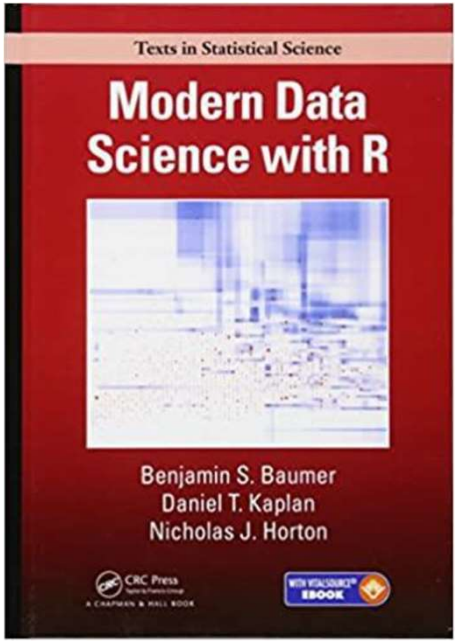
Ask a question

- Ask questions about the course.
- Answer your fellow students' questions whenever you can.

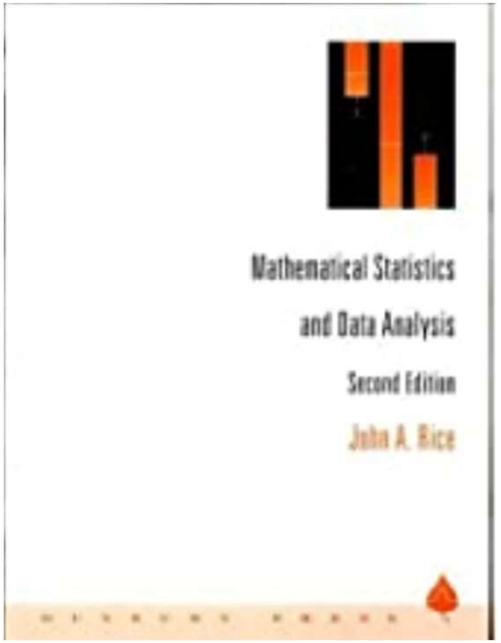
Please be careful to always be polite and considerate on the Blackboard discussions

Recommended reading

Modern Data Science with R
By B. Baumer, D. Kaplan & H. Horton



Mathematical Statistics and Data Analysis
By John Rice



Other books can be found at the "Resource lists" tab on Blackboard

University of BRISTOL

Thanks for listening!

Dr. Rihuan Ke

rihuan.ke@bristol.ac.uk

Statistical Computing and Empirical Methods
Unit EMATM0061, MSc Data Science