# [WEEK 1 ] Exploratory Data Analysis

## CS5701 - Quantitative Data Analysis

Dr Isabel Sassoon

Department of Computer Science
Brunel University London
isabel.sassoon AT brunel.ac.uk

21st September 2022

Brunel University London

# Week 1 - Learning Outcomes

This week after the lecture, the lab and the independent practice you should be:

- Aware that statistical analysis is part of a process such as PPDAC
- Able to determine the different data types
- Suggest and use appropriate methods for numerical and graphical exploration depending on the data type
- Able to use R Studio to read in data and run some numerical and graphical exploration

# What is Statistics?

> *"Statistics is concerned with* **collecting, analysing** *and* **interpreting** *data in the best possible way, where the meaning of 'best' depends on the particular circumstances of the practical situation"*

Quote from C. Chatfield, "Problem Solving A Statistician's Guide"

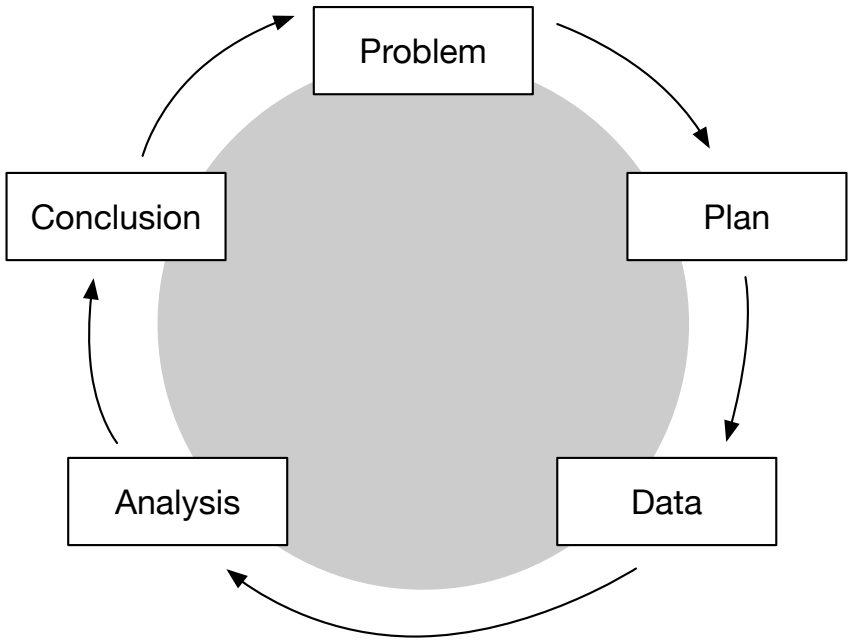
Brunel University London

# Why is Statistics Important?

- There is great potential for using data to better understand the world , make better judgements and Statistics gives us a way of achieving this
- For example Statistics can help us determine if a change in a value or measure is real or just an artefact of randomness - does a treatment work?
- Statistical methods are important for data science

# Statistical Analysis as a Process

One way of articulating this process is: **PPDAC** Problem, Plan, Data, Analysis & Conclusion




In this course we will be mainly focusing on **analysis** and **conclusion** but it is important to be aware of the overall process.

(Ref Wild, Chris J., and Maxine Pfannkuch. "Statistical thinking in empirical enquiry." International statistical review 67.3 (1999): 223-248.)

Brunel University London

# Statistical Analysis as a Process - PPDAC

- **Problem** understand and define the problem, how to answer the question asked
- **Plan** What to measure and how?
- **Data** Collect it, manage it and clean it (if needed)
- **Analysis** Explore the data, test the hypothesis and/or build the model
- **Conclusion** Interpret the results, craft conclusions, communicate the results and (sometimes plan the next phase)

# Examples of Statistical Analysis as a Process - PPDAC

An example we will revisit later is:

- **Problem** we need an estimate for the height of a 12 year old child (maybe as we are ordering sports equipment)
- **Plan** in order to solve the problem and answer the question we should find data about the heights of 12 year old children, explore it and make sure it is representative and extract our estimate from it (perhaps with some confidence interval)
- **Data** find or collect the data - prepare it for analysis
- **Analysis** Estimate the parameter we need to answer our question
- **Conclusion** In this case it is trivial but the output of the analysis is not enough, the questions needs to be answered

# Phases of the Analysis task in PPDAC

- **Look at the data** Summarise, explore and assess the quality of the data. Modify if necessary.
- **Formulate a sensible model** Use the findings from the data exploration to formulate an approach
- **Fit a model to the data** Estimate the model parameters, test hypotheses and check for model fit and diagnostics. Be prepared to modify the model if necessary.
- **Utilise the model and present the conclusions** Models can be used for descriptive, predictive or comparative purposes. A summary of the data used, the model fitted and findings need to be communicated.

Taken from Chapter 5 - Chatfield.

# Types of Analysis

There are two main types of analysis:

- **Descriptive**: describing the data by using numerical summaries or graphical outputs
- **Inferential**: goes beyond describing the data by using the information from a sample of data to make a conclusion about a larger population.

# Inferential Statistics - Population vs. sample

In **Inferential statistics** the aim is to use the information from a sample to *infer* about the larger population. For example to determine the mean height of a 12 year old child, it is not necessary to measure all 12 year old children in the UK.

- **Population** of interest: this is the entire group of items or individuals about which we want to estimate its parameters. In the example this is all the 12 year old children in the UK this year.
- **Sample** a smaller set of data from the population of interest, that is available to be analysed and upon which the parameters of the populations can be inferred. In the example these can be a set of 10 classes of 12 year olds in London.

This lecture wont focus on inference, we will cover this in later lectures.

Brunel University London

# Descriptive Statistics

- Types of variables
- How to summarise each type numerically
- How to visualise each type using plots

# Descriptive Statistics - Types of Variables

- **Categorical** - records which one of a list of possible categories or attributes is observed for a particular sampling unit. When there are two possible categories it is known as **binary**
- **Numerical** - takes a numerical value, which can be on a discrete, ordinal or continuous scale. Sometimes called **interval** or **continuous**

Note: that a variable can also be referred to as an attribute, a column (for example in a spreadsheet) or a vector of data.

# Worms data

This table has the first 10 rows of the Worms Data set from Crawley's book (Ch 2):

| | Field.Name | Area | Slope | Vegetation | Soil.pH | Damp | Worm.density |
|---|---|---|---|---|---|---|---|
| 1 | Nashs.Field | 3.60 | 11 | Grassland | 4.10 | FALSE | 4 |
| 2 | Silwood.Bottom | 5.10 | 2 | Arable | 5.20 | FALSE | 7 |
| 3 | Nursery.Field | 2.80 | 3 | Grassland | 4.30 | FALSE | 2 |
| 4 | Rush.Meadow | 2.40 | 5 | Meadow | 4.90 | TRUE | 5 |
| 5 | Gunness.Thicket | 3.80 | 0 | Scrub | 4.20 | FALSE | 6 |
| 6 | Oak.Mead | 3.10 | 2 | Grassland | 3.90 | FALSE | 2 |
| 7 | Church.Field | 3.50 | 3 | Grassland | 4.20 | FALSE | 3 |
| 8 | Ashurst | 2.10 | 0 | Arable | 4.80 | FALSE | 4 |
| 9 | The.Orchard | 1.90 | 0 | Orchard | 5.70 | FALSE | 9 |
| 10 | Rookery.Slope | 1.50 | 4 | Grassland | 5.00 | TRUE | 7 |

What types of variables are these?

# Worms data

| | Field.Name | Area | Slope | Vegetation | Soil.pH | Damp | Worm.density |
|---|---|---|---|---|---|---|---|
| 1 | Nashs.Field | 3.60 | 11 | Grassland | 4.10 | FALSE | 4 |
| 2 | Silwood.Bottom | 5.10 | 2 | Arable | 5.20 | FALSE | 7 |
| 3 | Nursery.Field | 2.80 | 3 | Grassland | 4.30 | FALSE | 2 |
| 4 | Rush.Meadow | 2.40 | 5 | Meadow | 4.90 | TRUE | 5 |
| 5 | Gunness.Thicket | 3.80 | 0 | Scrub | 4.20 | FALSE | 6 |
| 6 | Oak.Mead | 3.10 | 2 | Grassland | 3.90 | FALSE | 2 |
| 7 | Church.Field | 3.50 | 3 | Grassland | 4.20 | FALSE | 3 |
| 8 | Ashurst | 2.10 | 0 | Arable | 4.80 | FALSE | 4 |
| 9 | The.Orchard | 1.90 | 0 | Orchard | 5.70 | FALSE | 9 |
| 10 | Rookery.Slope | 1.50 | 4 | Grassland | 5.00 | TRUE | 7 |

- `Field.Name, Vegetation, Damp` are **categorical**.
- `Area, Slope, Worms` are **numerical**

When you use data make sure to check the metadata or data description - sometimes a number actually can represent a category....

# Another example

What type of variables are `extra` and `group`?

| | extra | group | ID |
|---|---|---|---|
| 1 | 0.70 | 1 | 1 |
| 2 | -1.60 | 1 | 2 |
| 3 | -0.20 | 1 | 3 |
| 4 | -1.20 | 1 | 4 |
| 5 | -0.10 | 1 | 5 |
| 6 | 3.40 | 1 | 6 |
| 7 | 3.70 | 1 | 7 |
| 8 | 0.80 | 1 | 8 |
| 9 | 0.00 | 1 | 9 |
| 10 | 2.00 | 1 | 10 |
| 11 | 1.90 | 2 | 1 |
| 12 | 0.80 | 2 | 2 |
| 13 | 1.10 | 2 | 3 |
| 14 | 0.10 | 2 | 4 |
| 15 | -0.10 | 2 | 5 |
| 16 | 4.40 | 2 | 6 |
| 17 | 5.50 | 2 | 7 |
| 18 | 1.60 | 2 | 8 |
| 19 | 4.60 | 2 | 9 |
| 20 | 3.40 | 2 | 10 |

# Another example

| | extra | group | ID |
|---|---|---|---|
| 1 | 0.70 | 1 | 1 |
| 2 | -1.60 | 1 | 2 |
| 3 | -0.20 | 1 | 3 |
| 4 | -1.20 | 1 | 4 |
| 5 | -0.10 | 1 | 5 |
| 6 | 3.40 | 1 | 6 |
| 7 | 3.70 | 1 | 7 |
| 8 | 0.80 | 1 | 8 |
| 9 | 0.00 | 1 | 9 |
| 10 | 2.00 | 1 | 10 |
| 11 | 1.90 | 2 | 1 |
| 12 | 0.80 | 2 | 2 |
| 13 | 1.10 | 2 | 3 |
| 14 | 0.10 | 2 | 4 |
| 15 | -0.10 | 2 | 5 |
| 16 | 4.40 | 2 | 6 |
| 17 | 5.50 | 2 | 7 |
| 18 | 1.60 | 2 | 8 |
| 19 | 4.60 | 2 | 9 |
| 20 | 3.40 | 2 | 10 |

- `extra` seems to be numerical or continuous
- `group` numerical, categorical or binary are all possible options (without context)

Brunel University London

# Descriptive Statistics

- It is important to investigate what each variable contains - performing **Exploratory Data Analysis** or **EDA**
- There are different ways of describing or exploring each variable and the approach will depend on their type: categorical vs numerical
- In each case it is possible to display information about the variable as a numerical summary or a graph
- It is also possible to look at how variables relate to each other, but we will cover this later in the module

The phrase *exploratory data analysis* was coined by John W. Tukey.

Brunel University London

# Categorical Data - Descriptive Statistics

There are different ways to numerically summarise a **categorical** variable:

- Frequencies or counts
- Relative Frequencies
- Relative Cumulative Frequencies

# Categorical Data - Descriptive Statistics

Frequency table for `Vegetation`

| | Frequency |
|---|---|
| Arable | 3 |
| Grassland | 9 |
| Meadow | 3 |
| Orchard | 1 |
| Scrub | 4 |

Brunel University London

# Categorical Data - Descriptive Statistics

Frequency table for `Vegetation`

| | Frequency | Cumulative Frequency |
|---|---|---|
| Arable | 3 | 3 |
| Grassland | 9 | ? |
| Meadow | 3 | ? |
| Orchard | 1 | ? |
| Scrub | 4 | ? |

What are the values to replace ?

Brunel University London

# Categorical Data - Descriptive Statistics

Frequency table for `Vegetation`

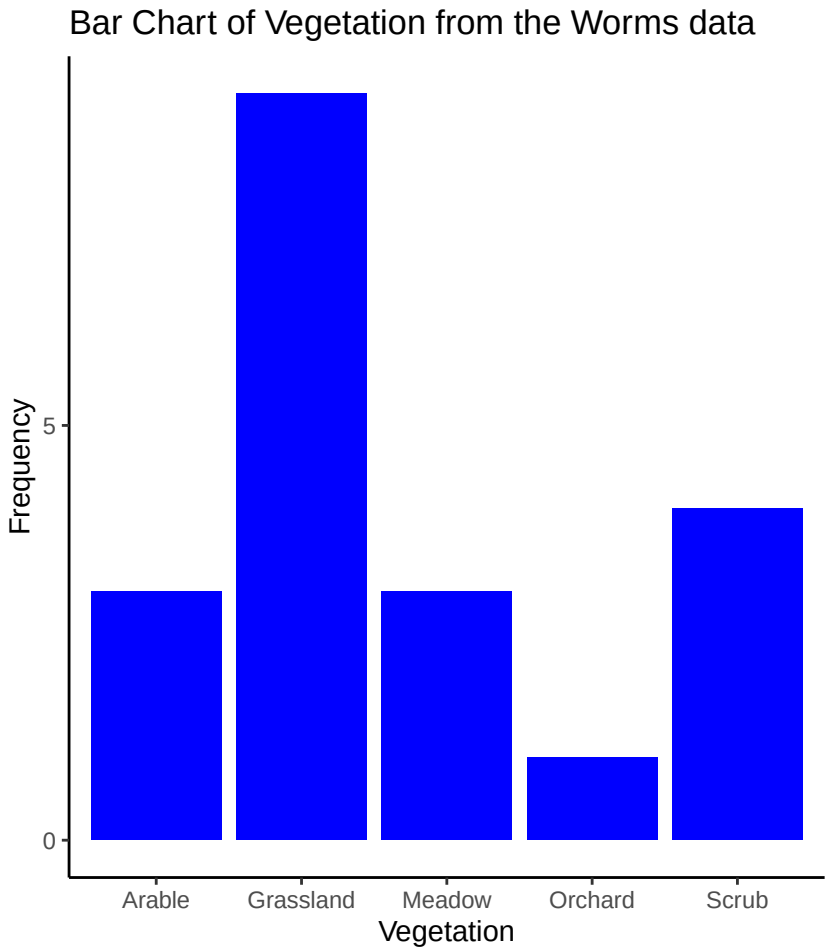
| | Frequency | Cumulative Frequency |
|---|---|---|
| Arable | 3 | 3 |
| Grassland | 9 | 12 |
| Meadow | 3 | 15 |
| Orchard | 1 | 16 |
| Scrub | 4 | 20 |

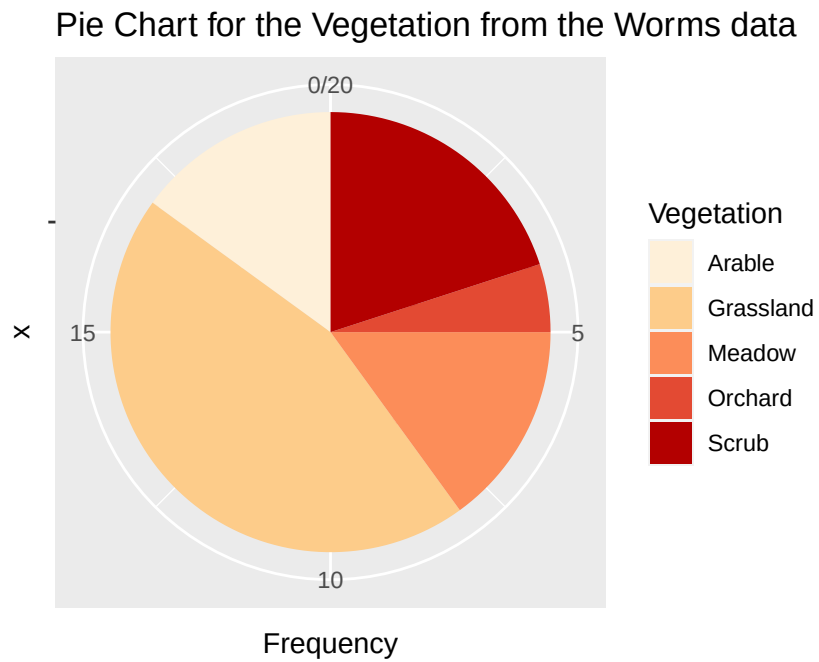
# Categorical Data - Visualising

**Categorical** Variables can be summarised visually using; **Bar Charts** or **Pie charts**

- **Categorical** Variables can be summarised visually using; **Bar Charts** or **Pie charts**
- **Bar Charts** can be used to represent the frequencies of each of the different categories, The y-axis has the frequency and the x-axis the categories.
- **Pie Charts** can be used to represent the frequencies of each of the different categories as different slices of the pie.

# Categorical Data- Visualising



# Numerical Data - Descriptive Statistics

When describing the contents of a numerical variable we can look at different aspects of its distribution, such as

- Measures of location - such as the mean
- Measures of spread and variability
- Extreme values

# Numerical Data - Descriptive Statistics - Measures of Location

- **Mode** the most frequent observation in the data set, there can be more than one
- **Mean** (average) this is calculated as the sum of the observations in the variable divided by the number of the observations.

$$\bar{x} = \frac{1}{n}\Sigma_{i=1}^{n} x_i$$

assuming we have an attribute $X$ which has the following observations $x_1, x_2, \ldots, x_n$

- **Median** when the set of observations is ordered from smallest to largest, the **median** is the value in the middle, or if there are an even number of observations then the mean of the two middle observations is used.

Note: the **mean** is affected by extreme values, the **median** is not.

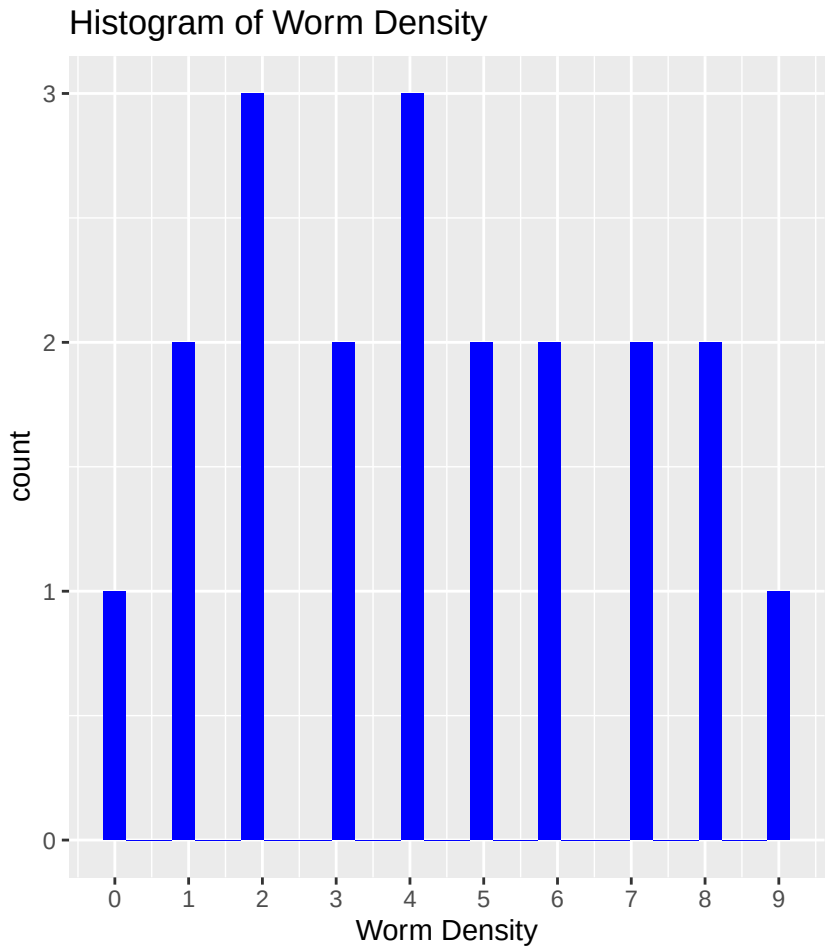
Brunel University London

# Numerical Data - Descriptive Statistics - Measures of Variation

There are different ways of quantifying the spread:

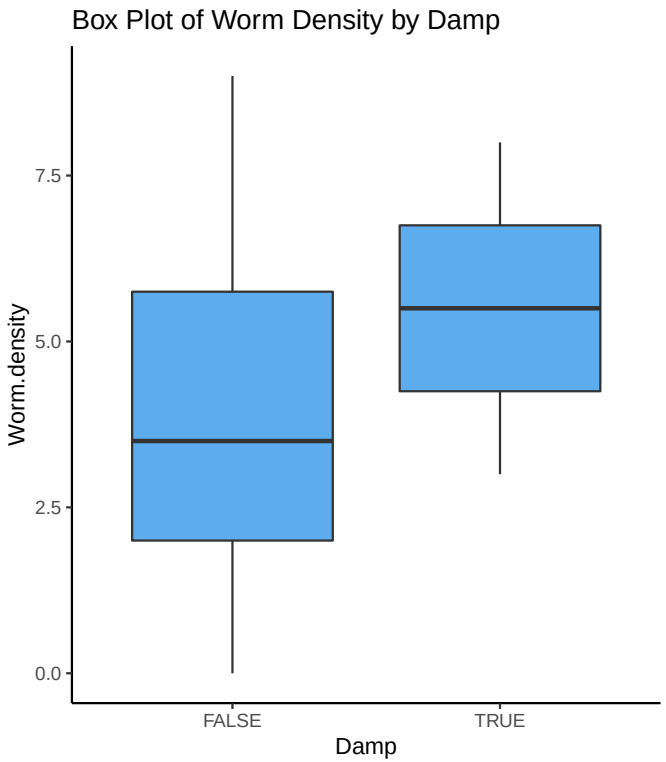
- The **range**: defined as *maximum value in the variable - minimum value*
- The **Inter Quartile Range - IQR** defined as the *75th Percentile - 25th Percentile*, this will measure the spread of the central 50% of the data
- The **Standard Deviation** - measures the spread of the observations from the mean

# Numerical Data - Visualising - Histogram




- A **Histogram** shows the frequency of values in the data.
- For example: the graph on the left, we can read that:
  - In the data there is only one garden with a worm density of 9
  - The most frequent values are: 4 and 2. They each appear 3 times in the data.

# Numerical Data - Visualising - Box Plot




- A **Box Plot** or **Box and Whisker** Plot displays the data using the median, the 25th and 75th percentile and the minimum and maximum.
  - A **Box Plot** can also display extreme observations or outliers
  - The middle of the box is the *median*
  - The edges of the box are the 25th and 75th percentiles
- In the example to the left:
  - the median `worm density` when `Damp` is `TRUE` is higher then when it is `FALSE`
  - the spread of the data is larger when `Damp` is `FALSE`

# Summary

In this lecture:

- Introduced the stages of statistical analysis
- Outlined the different types of variables
- Introduced EDA and methods for descriptive statistics for different types of variables

In the Labs - Using RStudio to read in data, explore the data with numerical summaries and graphs.

# QDA - What's next?

- This afternoon in the labs you will have a chance to do this independently
- We will be around to answer questions
- Look out for the material for next week (what to read before the lecture)