

# Topic 2 – Exploratory Data Analysis (EDA)

ENVX1002 Introduction to Statistical Methods

Dr. Floris van Ogtrop  
The University of Sydney

Feb 2024



THE UNIVERSITY OF  
**SYDNEY**

# Topic 2 - Exploratory Data Analysis

- Summary statistics:
  - measures of centre;
  - measures of spread (dispersion).
- Graphical summaries:
  - bar chart;
  - histogram;
  - boxplot.

# Learning Outcomes

- At the end of this topic students should able to:
  - Calculate “by hand” summary statistics for simple datasets;
  - Manually draw graphical summaries (boxplots and histograms) for simple datasets;
  - Demonstrate proficiency in the use of R and Excel for calculating summary statistics and generating graphical summaries;
  - Describe key features of their data using summary statistics and graphical summaries.

# Types of data

- Numerical:
  - Continuous: yield, weight
  - Discrete: weeds per  $m^2$
- Categorical:
  - Binary: 2 mutually exclusive categories
  - Ordinal: categories ranked in order
  - Nominal: qualitative data

# Presentation of data

**Tables:** Experimental data

**Table 1** Yields of cabbage (mean fresh weight per head in kg) for 24 plots (Source: Mead, Curnow & Hasted (2003)).

Irrigation	Spacing	Field A	Field B	Field C
Frequent	1 (21 in)	1.11	1.03	0.94
Frequent	2 (18 in)	1.00	0.82	1.00
Frequent	3 (15 in)	0.89	0.80	0.95
Frequent	4 (12 in)	0.87	0.65	0.85
Rare	1 (21 in)	0.97	0.86	0.92
Rare	2 (18 in)	0.80	0.91	0.68
Rare	3 (15 in)	0.57	0.72	0.77
Rare	4 (12 in)	0.60	0.69	0.51

# Presentation of data

Tables: Observational data

ID	Eastings	Northings	Yield
1	508562	235514	4.00
2	508533	235469	2.33
3	508562	235591	4.16
.....	.....	.....	.....
99981	508722	235521	3.88
99982	508706	235590	3.76

# Population versus Sample

Before we go calculate averages, we need to think about the difference between population and sample

- We take a sample from a larger population
- What information does the sample give about the population and how reliable is that information?



bakhtiarzein - <https://stock.adobe.com/>

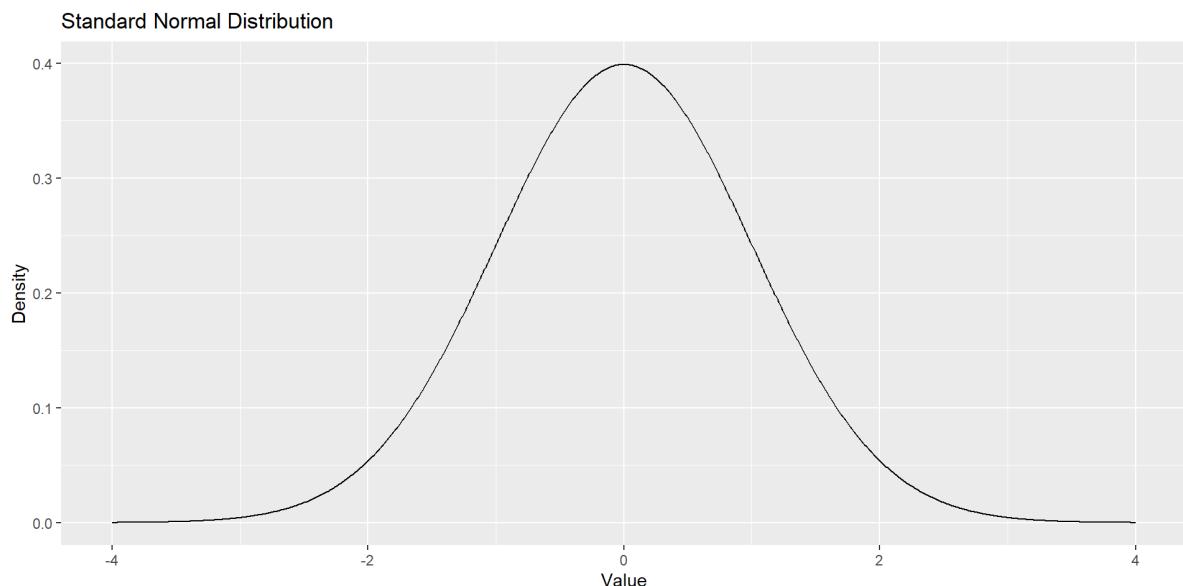
# Descriptive statistics

- Measures of central tendency
  - Mean
  - Median
  - Mode
- Measures of spread or dispersion
  - Range
  - Interquartile range
  - Standard deviation / Variance

```

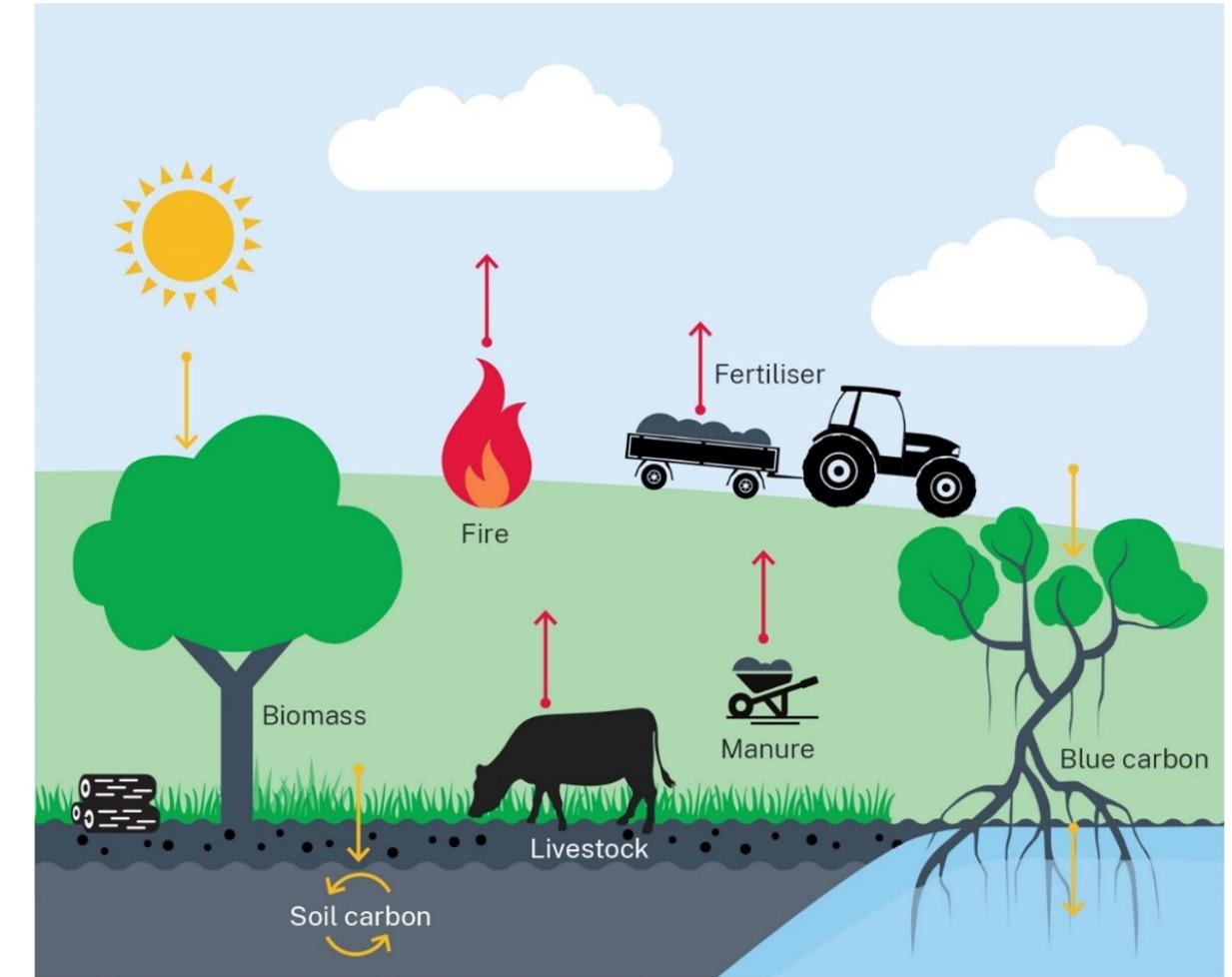
1 library(ggplot2)
2 library(tidyverse)
3
4 # Generate a normal distribution
5 normal_dist <- tibble(x = seq(-4, 4, by = 0.01)) %>%
6   mutate(y = dnorm(x))
7
8 # Plot the distribution
9 ggplot(normal_dist, aes(x = x, y = y)) +
10  geom_line() +
11  ggtitle("Standard Normal Distribution") +
12  xlab("Value") +
13  ylab("Density")

```



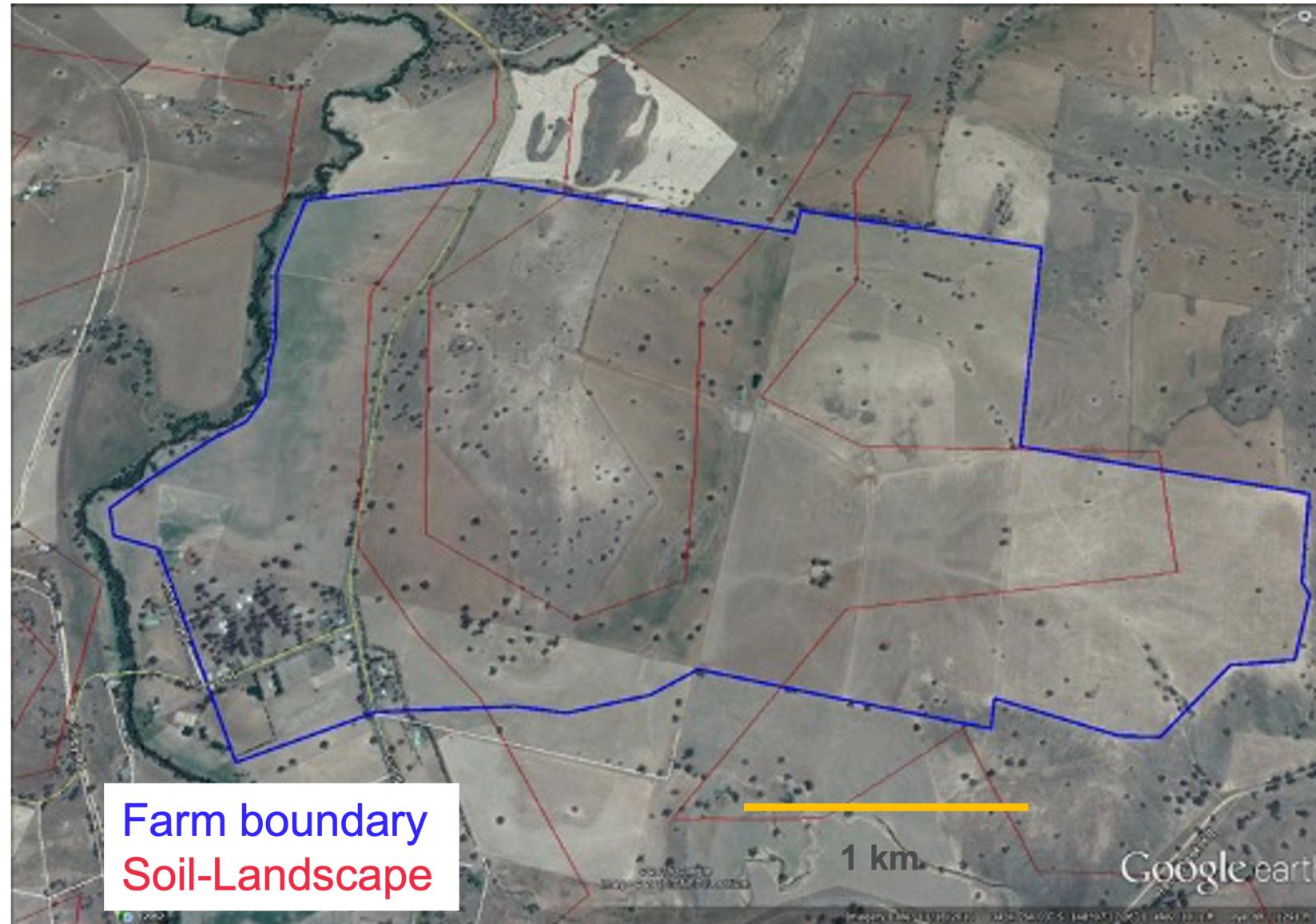
# Motivating example

- Sequestered soil carbon is worth \$35/tonne if measured  
(1 Tonne of Carbon = 1 Australian Carbon Credit Unit = \$AU35 See [Clean Energy Regulator](#))
- It costs \$100 to collect and analyse one soil sample for soil carbon
- The farmer needs an estimate of carbon stored on the property.
- How many samples are needed to give a good estimate of carbon on the property? Is it worth measuring soil carbon for a land holder?



Source: <https://www.energy.nsw.gov.au/business-and-industry/programs-grants-and-schemes/primary-industries-carbon-farming>

# Motivating example



Google earth image with farm and soil-landscape boundaries

# Motivating example

- Soil carbon content was measured at 6 points across a farm
  - The amount at each location was 48, 56, 90, 78, 86, 271 (t/ha)
- We will now get into some formulas and calculations ;o)

# Sigma notation

- $\Sigma$ , is the greek capital letter called sigma, refers to the sum
- It is a convenient way to represent long sums

$$\sum_{i=1}^n x_i = x_1 + x_2 + x_3 + \dots + x_n$$



- `sum(c(48, 56, 78, 86, 90, 271))`

```
1 total_c <- sum(c(48, 56, 78, 86, 90, 271))
2 print(total_c)
```

[1] 629



- `=SUM(A1:A6)`

# Centre: Arithmetic mean

- Population mean ( $\mu$ ): sum of all values of a variable divided by the number of objects in the population;

$$\mu = \frac{\sum_{i=1}^N y_i}{N}$$

- Sample mean ( $\bar{y}$ ) is based on a subset of  $n$  objects from a population of size  $N$

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$



- `mean(c(48, 56, 78, 86, 90, 271))`

```
1 mean_c <- mean(c(48, 56, 78, 86, 90, 271))
2 print(mean_c)
```

[1] 104.8333



- `=AVERAGE(A1:A6)`

# Centre: Median

- Median is the middle number of a set of ordered observations
- Population median  $M$ : sum of all values of a variable divided by the number of objects in the population;

Population median:  $M = \left(\frac{N+1}{2}\right) th$  sorted value

Sample median:  $\tilde{y} = \left(\frac{n+1}{2}\right) th$  sorted value



- `median(c(48, 56, 78, 86, 90, 271))`

```
1 median_c <- median(c(48, 56, 78, 86, 90, 271))
2 print(median_c)
```

[1] 82



- `=MEDIAN(A1:A6)`

# Centre: Mode

Mode is the most commonly occurring number in a set of observations



```
1 mode_function <- function(x) {  
2   uniq_x <- unique(x)  
3   uniq_x[which.max(tabulate(match(x, uniq_x)))]  
4 }  
5 data_vector <- c(1, 2, 4, 4, 3, 5, 4)  
6 mode_value <- mode_function(data_vector)  
7 print(mode_value)
```

[1] 4



=MODE.SNGL(A1:A7)

# Spread: Range

- Difference between largest and smallest observations in a group of data
- Note that we also refer to spread as measures of dispersion



- `max(c(48, 56, 78, 86, 90, 271)) - min(c(48, 56, 78, 86, 90, 271))`

```
1 range_c <- max(c(48, 56, 78, 86, 90, 271)) - min(c(48, 56, 78, 86, 90, 271))
2 print(range_c)
```

[1] 223



- `=MAX(A1:A6) - MIN(A1:A6)`

# Spread: Inter-quartile range (IQR)

- Median divides dataset into 2, quartile divides it into 4:
  - 25% observations  $\leq$  1st quartile (Q1)
  - 50% observations  $\leq$  Median (Q2)
  - 75% observations  $\leq$  3rd quartile (Q3)

Let's take an easy example

1 2 3 4 5 6 7 8 9

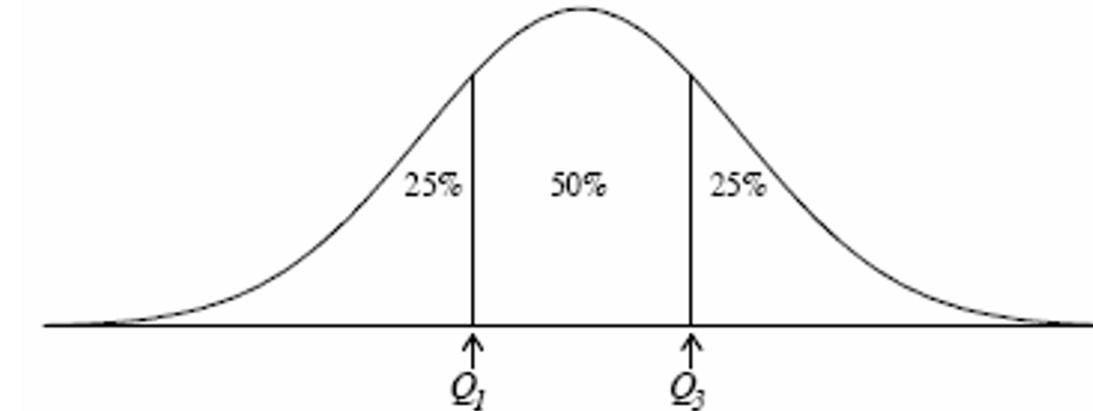
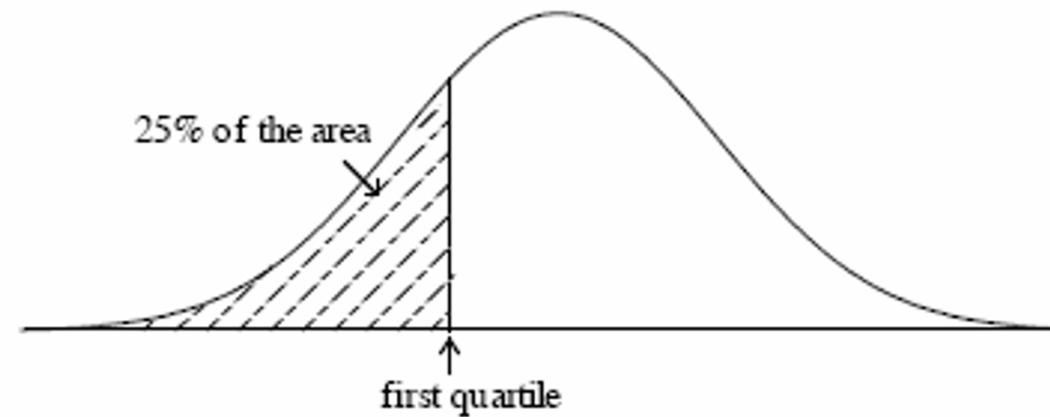
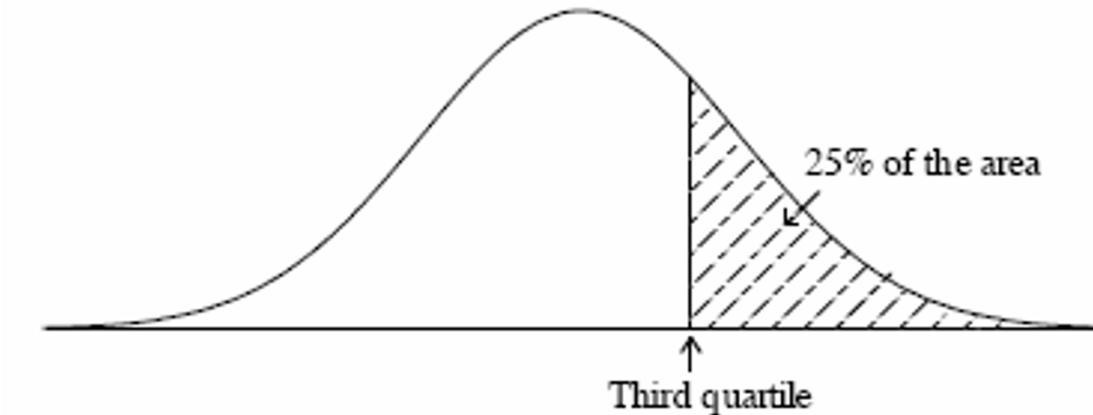
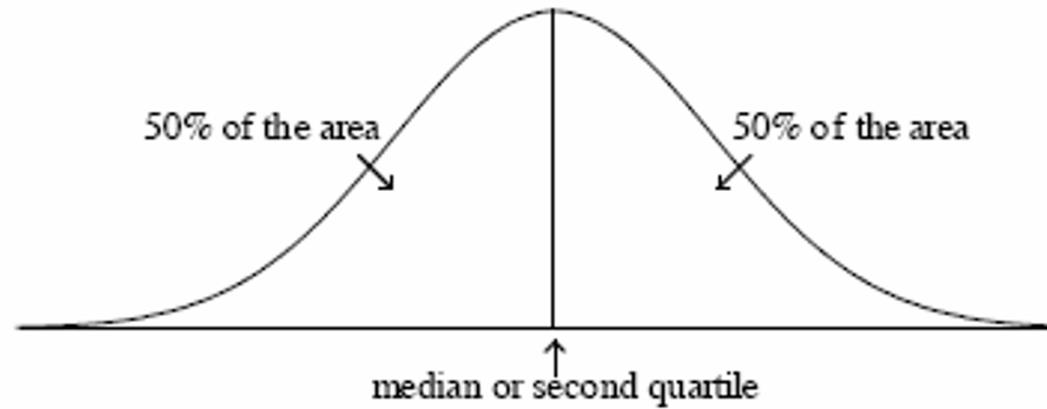
What is Q1, Median, Q3?

```
1 quantile(c(1,2,3,4,5,6,7,8,9))
```

```
0% 25% 50% 75% 100%
 1    3    5    7    9
```

# Spread: Inter-quartile range (IQR)

- $IQR = Q_3 - Q_1$



Source: Nicholas (1999)

# Spread: Inter-quartile range (IQR)

## Quartiles



- `quantile(c(48, 56, 78, 86, 90, 271))`

```
1 quant_c <- quantile(c(48, 56, 78, 86, 90, 271))
2 print(quant_c)
```

```
0%   25%   50%   75%   100%
48.0  61.5  82.0  89.0 271.0
```



- `=QUARTILE.INC(A1:A6, 1)` - first quartile

# Spread: Inter-quartile range (IQR)

IQR



- `IQR(c(48, 56, 78, 86, 90, 271))`

```
1 iqr_c <- IQR(c(48, 56, 78, 86, 90, 271))
2 print(iqr_c)
```

[1] 27.5



- `=QUARTILE.INC(A1:A6, 3)-QUARTILE.INC(A1:A6, 1)` - third quartile - first quartile

# Spread: Variance

- Describes variability around the arithmetic mean

Population variance:  $\sigma^2 = \frac{\sum_{i=1}^N (y_i - \mu)^2}{N}$

Sample variance:  $s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$



```
- var(c(48, 56, 78, 86, 90, 271))
```

```
1 var_c <- var(c(48, 56, 78, 86, 90, 271))
2 print(var_c)
```

[1] 6904.167



```
- =VAR.S(A1:A6)
```

# Spread: Standard deviation

- Describes variability around the **arithmetic mean**
  - Variance is in *units<sup>2</sup>* as it is based on squared deviations from the mean
  - Standard deviation describes variability around the mean in original units
  - Standard deviation  $\sqrt{\text{variance}}$  or the variance

$$\text{Population standard deviation: } \sigma = \sqrt{\frac{\sum_{i=1}^N (y_i - \mu)^2}{N}}$$

$$\text{Sample variance: } s^2 = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}$$

# Spread: Standard deviation

- R's sd function always calculates the sample standard deviation
- The denominator of sample standard is  $n-1$  (Bessel's correction) this is an important concept in statistics. A key property is that it gives a more accurate estimate of the population variance and standard deviation when working with a sample.



- `sd(c(48, 56, 78, 86, 90, 271))`

```
1 sd_c <- sd(c(48, 56, 78, 86, 90, 271))
2 print(sd_c)
```

[1] 83.09132



- `=STDEV.S(A1:A6)`

# Spread: Coefficient of variation

- Let's take an example where we measured both nitrogen and carbon in our soil such that:

Soil nitrogen (%): 2 16 22 45 65 93

- How could we find out which measurements have a greater spread given they have very different units (%) versus t/ha)?
- It turns out we can use the CV

$$CV = \left( \frac{s}{\bar{y}} \right) \times 100$$

# Spread: Coefficient of variation

- Looking at the calculations below, which is more variable, Carbon or Nitrogen?



- `sd(c(48, 56, 78, 86, 90, 271))`

```
1 cv_c <- sd(c(48, 56, 78, 86, 90, 271))/mean(c(48, 56, 78, 86, 90, 271))*100
2 print(cv_c)
```

[1] 79.2604

```
1 cv_n <- sd(c(2, 16, 22, 45, 65, 93))/mean(c(2, 16, 22, 45, 65, 93))*100
2 print(cv_n)
```

[1] 84.10661



- `=(STDEV.S(A1:A6)/AVERAGE(A1:A6))*100`

# Robustness (to outliers)

- Which summary statistics should I use to describe centre?
  - Example: 48, 56, 8, 86, 90, 27
  - Example: 48, 56, 8, 86, 90, 271

mean -  $\bar{y}$ :

```
1 mean(c(48, 56, 8, 86, 90, 27))
```

```
[1] 52.5
```

```
1 mean(c(48, 56, 8, 86, 90, 271))
```

```
[1] 93.16667
```

median -  $\tilde{y}$ :

```
1 median(c(48, 56, 8, 86, 90, 27))
```

```
[1] 52
```

```
1 median(c(48, 56, 8, 86, 90, 271))
```

```
[1] 71
```

# Robustness (to outliers)

- Which summary statistics should I use to describe spread?
  - Example: 48, 56, 8, 86, 90, 27
  - Example: 48, 56, 8, 86, 90, 271

variance -  $s^2$ :

```
1 var(c(48, 56, 8, 86, 90, 27))
```

```
[1] 1038.3
```

```
1 var(c(48, 56, 8, 86, 90, 271))
```

```
[1] 8472.167
```

Inter quartile range -  $IQR$ :

```
1 IQR(c(48, 56, 8, 86, 90, 27))
```

```
[1] 46.25
```

```
1 IQR(c(48, 56, 8, 86, 90, 271))
```

```
[1] 39
```

# Categorical data

- Different types with examples:
  - Binary: We spray insects and see how many die
  - Nominal: We count how animals, and their species, are in a forest
  - Ordinal: Different disease levels for a plant, no disease, moderate, severe
- We can count the number observations belonging to each class, called frequency, f.
  - Can present as a frequency table

Plant disease severity	Frequency
None	3
Moderate	5
Severe	2

# Graphical summaries

- Visualisation of data is useful for identifying
  - outliers
  - shape and distribution
  - communicating results
  - suggest modelling strategies
- Bar chart
- Strip chart
- Boxplot
- Histogram

# Bar chart

- We first tabulate the data

```
1 # Load necessary library
2 library(ggplot2)
3
4 # Your disease data
5 disease <- c("None", "Moderate", "None", "Severe", "Moderate", "Moderate", "Severe", "Moderate", "None", "Moderate")
6
7 # Order factors from no disease to severe disease
8 disease = factor(disease, levels = c("None", "Moderate", "Severe"))
9
10 # Create a frequency table
11 disease_tbl <- table(disease)
12 print(disease_tbl)
```

disease	None	Moderate	Severe
	3	5	2

# Bar chart

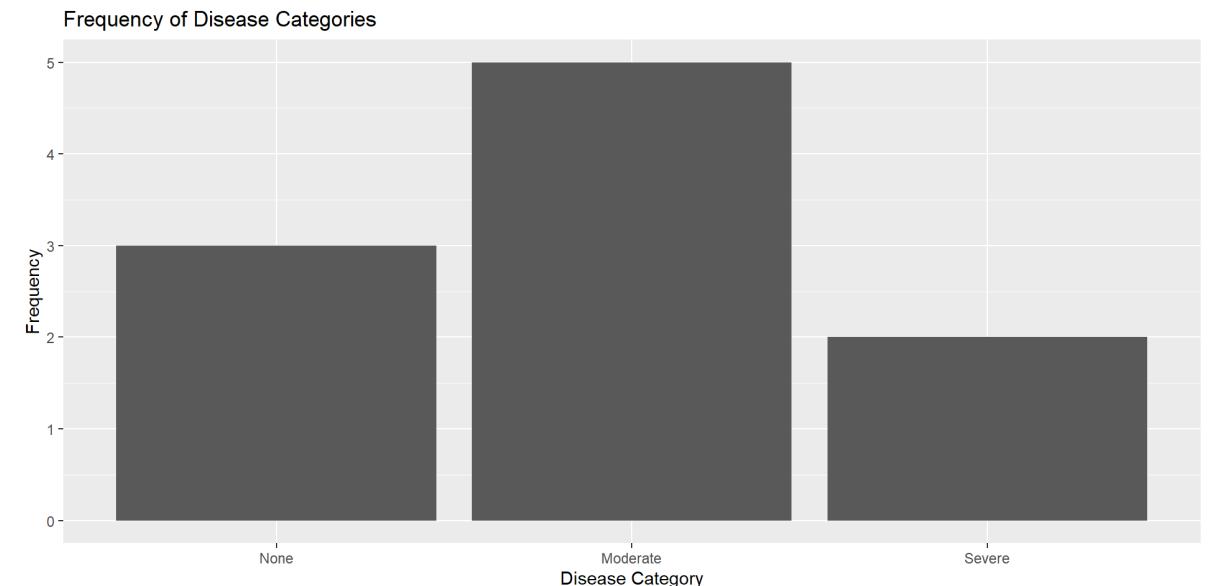
- We then plot the table in ggplot

```

1 # Convert the table to a data frame for ggplot2
2 disease_df <- as.data.frame(disease_tbl)
3
4 # Rename the columns appropriately
5 names(disease_df) <- c("Disease", "Frequency")
6
7 # Create the bar plot
8 p <- ggplot(disease_df, aes(x = Disease, y = Frequency)) +
9   geom_bar(stat = "identity") +
10  ggtitle("Frequency of Disease Categories") +
11  xlab("Disease Category") +
12  ylab("Frequency")

```

```
1 print(p)
```



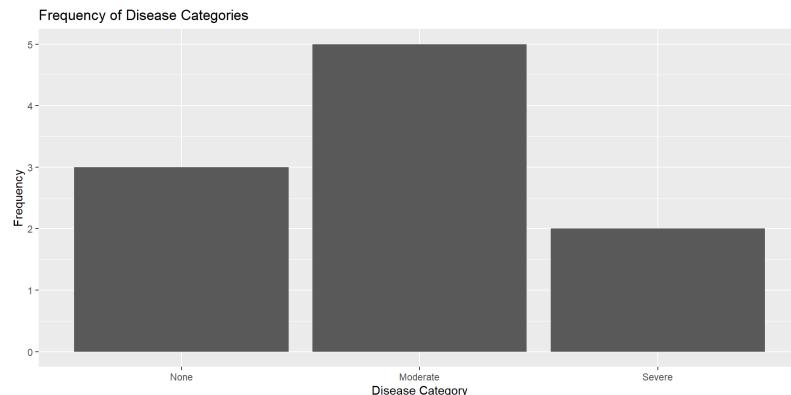
# Bar chart

- Using `tidyverse`

```

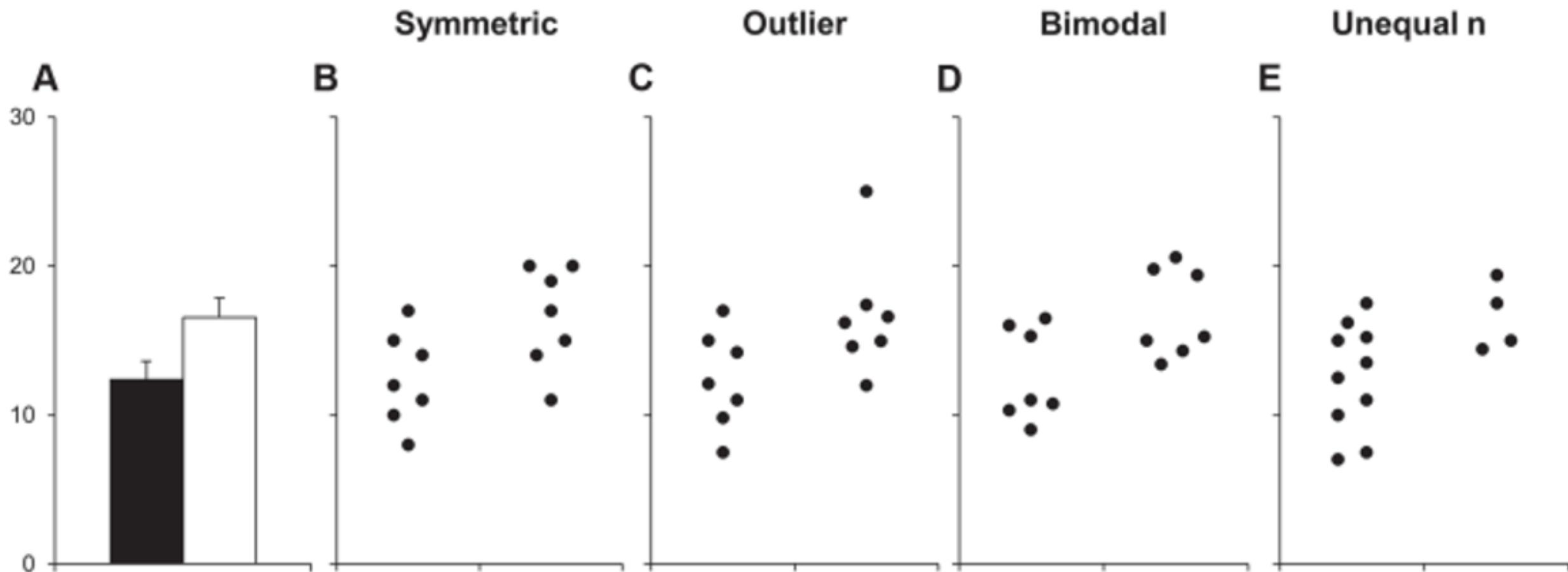
1 # Load necessary libraries
2 library(tidyverse)
3
4 # Your disease data
5 disease <- c("None", "Moderate", "None", "Severe", "Moderate", "Moderate", "Severe", "Moderate", "None", "Moderate")
6
7 # Convert to tibble and count occurrences
8 disease_data <- tibble(disease) %>%
9   mutate(disease = factor(disease, levels = c("None", "Moderate", "Severe"))) %>%
10  count(disease, name = "Frequency")
11
12 # Create the bar plot
13 ggplot(disease_data, aes(x = disease, y = Frequency)) +
14   geom_bar(stat = "identity") +
15   ggtitle("Frequency of Disease Categories") +
16   xlab("Disease Category") +
17   ylab("Frequency")

```



# Bar chart

- NOTE: Bar charts should generally not used for continuous numerical data



Source: Weissgerber et al. (2015)

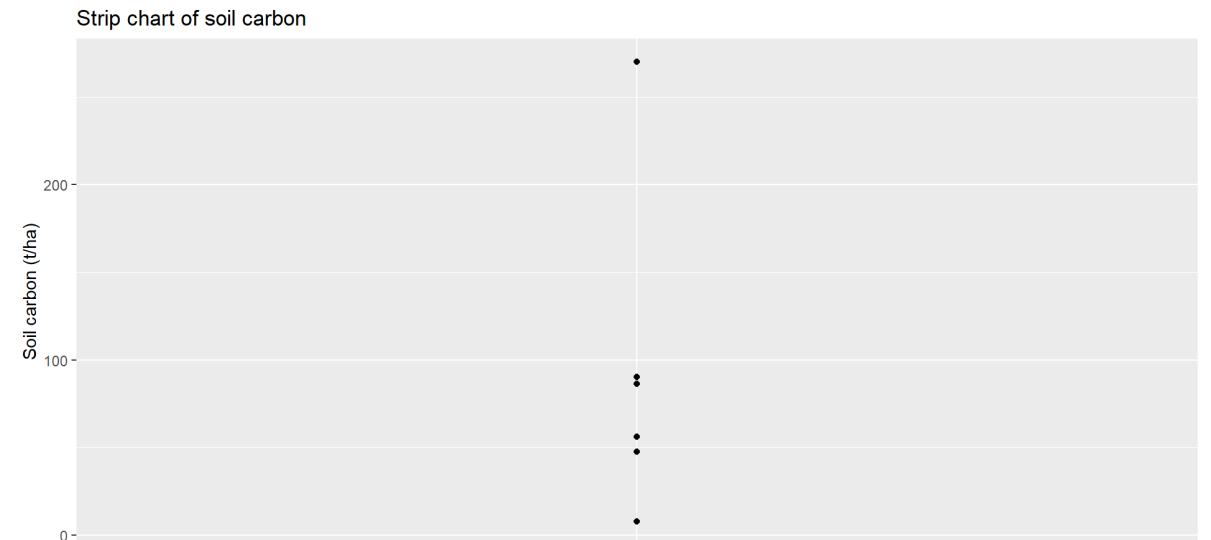
# Strip chart

- Often if we have a small data set (1-5 data points), we can use a stripchart to visualise our data. We will demonstrate using our soil carbon data set.
- What do we notice from the plot?

```

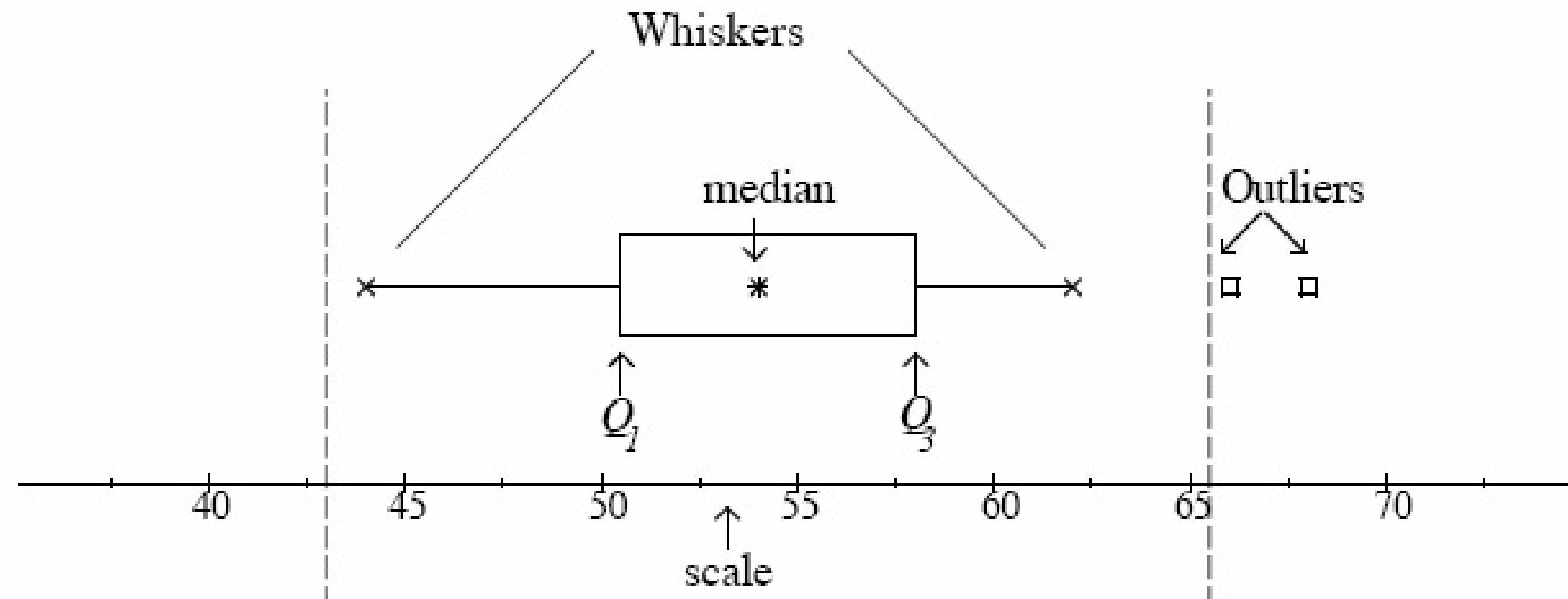
1 # Load necessary library
2 library(ggplot2)
3
4 # Your data
5 soil_c <- c(48, 56, 8, 86, 90, 271)
6
7 # Convert to a data frame
8 soil_c_df <- data.frame(Value = soil_c)
9
10 # Create the strip chart
11 p <- ggplot(soil_c_df, aes(x = "", y = Value)) +
12   geom_jitter(width = 0) +
13   ggtitle("Strip chart of soil carbon") +
14   xlab("") +
15   ylab("Soil carbon (t/ha)")
```

```
1 print(p)
```



# Boxplot

- We can overlay our strip chart with a boxplot. This shows use the min/max, quartiles and median and can also show outliers.
- We generally use boxplots when we have more than 5 data points.
- See lecture notes for creating boxplot by hand.



Source: Nicholas (1999)

# Boxplot

- Here is a slightly larger data set from a trial where creeping bentgrass turf was laid in an experiment to assess root growth. Eighty (80) “plugs” were randomly sampled 4 weeks after laying. Root growth was measured by averaging the length (mm) of the ten longest roots in each plug.

```
1 root_length <- c(108, 102, 100, 135, 113, 109, 92, 97, 73, 65,  
2 68, 74, 93, 97, 118, 121, 103, 99, 90, 90,  
3 99, 102, 106, 90, 92, 97, 100, 92, 80, 99,  
4 103, 103, 115, 85, 96, 86, 85, 86, 91, 90,  
5 94, 93, 93, 99, 109, 115, 110, 94, 107, 88,  
6 101, 89, 117, 91, 112, 101, 91, 81, 80, 67,  
7 69, 80, 86, 81, 65, 90, 99, 93, 90, 102,  
8 72, 70, 90, 90, 87, 89, 90, 96, 108, 86)
```

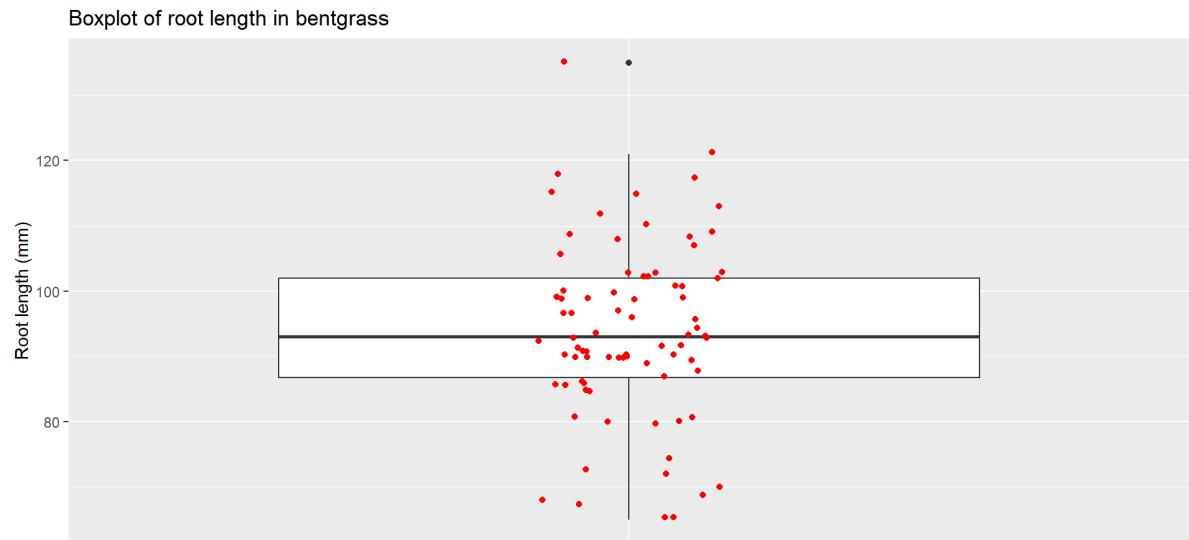
# Boxplot

- The follow produces a boxplot and also includes the jittered data points (red coloured)
- Note that there is on outlier which is the black data point at the top of the plot

```

1 # Load necessary library
2 library(ggplot2)
3
4 # Convert to a data frame
5 root_length_df <- data.frame(Value = root_length)
6
7 # Create the strip chart
8 ggplot(root_length_df, aes(x = "", y = Value)) +
9   geom_boxplot() +
10  geom_jitter(width = 0.1, col = "red") +
11  ggtitle("Boxplot of root length in bentgrass") +
12  xlab("") +
13  ylab("Root length (mm)")

```



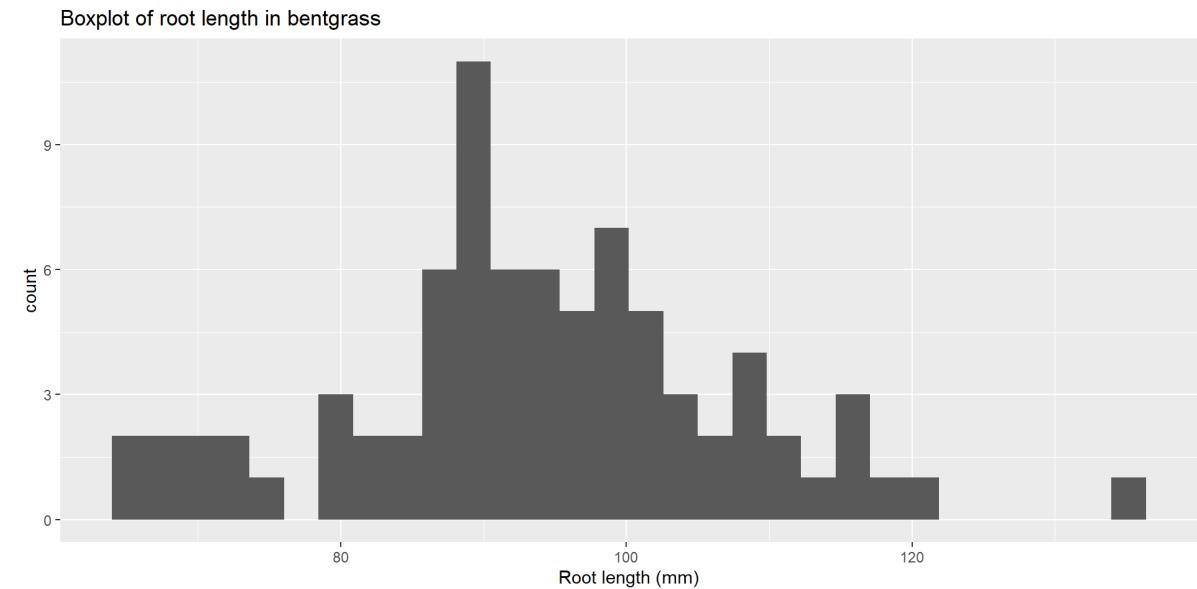
# Histogram

- Based on frequency table
  - Height of each bar proportional to frequency - need to group data
  - We can use histograms to describe the shape of distributions for continuous data sets that are larger than 20 data points.

```

1 # Load necessary library
2 library(ggplot2)
3
4 # Convert to a data frame
5 root_length_df <- data.frame(Value = root_length)
6
7 # Create the strip chart
8 ggplot(root_length_df, aes(Value)) +
9   geom_histogram() +
10  ggttitle("Boxplot of root length in bentgrass") +
11  xlab("Root length (mm)")

```



# Summary

- The following is a rough guide for plotting *continuous* data

observations	graphics	R function
1-5	raw data	stripchart
6-20	boxplot	boxplot
20+	histogram	hist

- Remember for *categorical* data we use **Tables** and **Bar Charts**

# Symmetry

- Throughout this unit you will be assessing the shape of distributions, in particular you will be looking at whether the distribution (histogram) of the data is symmetrical in shape;
- For small data sets, you will generally compare the mean and median;
  - if the mean and the median are similar it indicates that the data is symmetrical.
  - if the mean and the median are not similar it indicates that the data is skewed.

```
1 mean(soil_c)
```

```
[1] 93.16667
```

```
1 median(soil_c)
```

```
[1] 71
```

- What can we conclude from the mean and median of our soil carbon data?

# Symmetry

- For larger data sets, we can look at the mean, median and the histogram to determine if it is symmetrical.

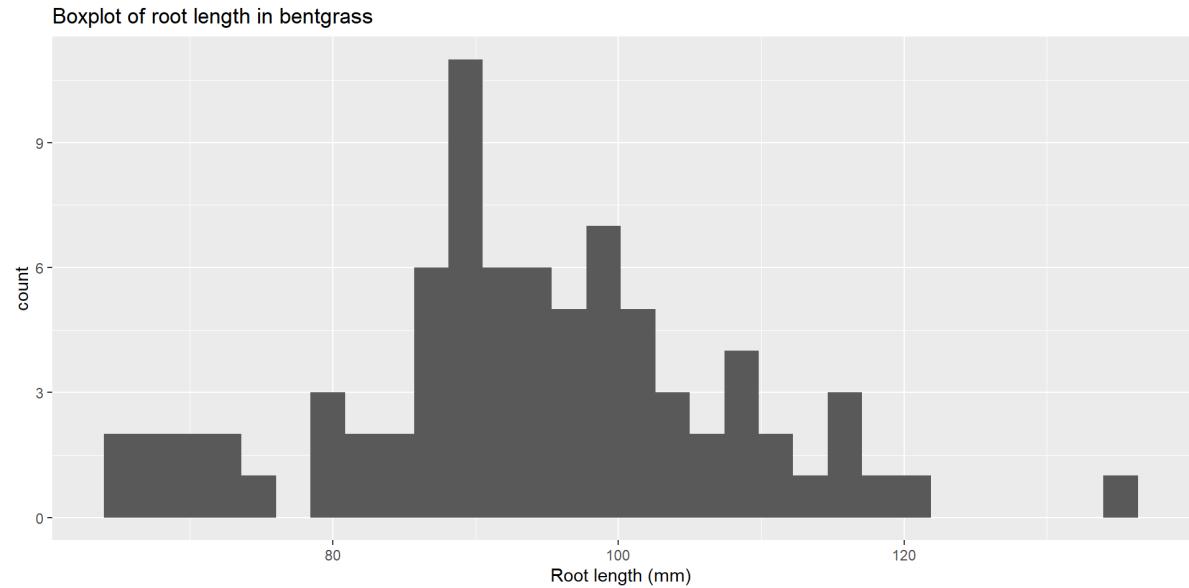
```
1 mean(root_length)
```

[1] 93.8625

```
1 median(root_length)
```

[1] 93

```
1 # Create the strip chart
2 ggplot(root_length_df, aes(Value)) +
3   geom_histogram() +
4   ggtitle("Boxplot of root length in bentgrass") +
5   xlab("Root length (mm)")
```



# Symmetry

- We can also calculate skewness using the following equation

$$g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left( \frac{y_i - \bar{y}}{s} \right)^3$$

- in RStudio, we can use the `skewness` function found in the `e1071` package

```
1 library(e1071)
2 skewness(root_length)
```

```
[1] 0.08316635
```

- if  $|g_1| < 1.0$  then the dataset is approximately symmetrical. If  $g_1 > 1.0$  then the data is positively skewed and if  $g_1 < -1.0$  then the data is negatively skewed

# Reading

- Canvas site
- Notes
- Quinn & Keough (2002)
  - Chapter 2. Sections 2.1-2.2, p. 14-17.
  - Chapter 4. Sections 4.1, p. 58-61 (stop at scatterplot)
- Mead et al. (2002).
  - Chapter 1.
  - Chapter 2. Sections 2.1-2.3, p. 9-19

# References

- J. Nicholas (1999). Introduction to descriptive statistics. Mathematics Learning Centre, University of Sydney.
- T. L. Weissgerber, N. M. Milic, S. J. Winham and V. D. Garovic (2015). Beyond bar and line graphs: time for a new data presentation paradigm. PLOS Biology. 13. e1002128.

# Thanks!

This presentation is based on the [SOLES Quarto reveal.js template](#) and is licensed under a [Creative Commons Attribution 4.0 International License](#).



THE UNIVERSITY OF SYDNEY