Computer Practical 4: Data Analysis in R

Dr Alice Risely [a.risely@salford.ac.uk](mailto:a.risely@salford.ac.uk)

Table of Contents

[Learning outcomes for today 2](#_Toc182165440)

[Practical 3 learning outcomes 2](#_Toc182165441)

[Preliminary tasks 2](#_Toc182165442)

[Recap: The five data analysis steps 2](#_Toc182165443)

[How to conduct statistical tests in R 3](#_Toc182165444)

[Uploading a new dataset on earthworms 4](#_Toc182165445)

[Exploring your new dataset 5](#_Toc182165446)

[Subsetting data 6](#_Toc182165447)

[Q1 What is the relationship between habitat type and worm abundance? 6](#_Toc182165448)

[Q2: Is there a relationship between rainfall and worm abundance? 8](#_Toc182165449)

[OPTIONAL: Fit a General Linear Model 11](#_Toc182165450)

## Learning outcomes for today

What you will learn in this practical:

* How to filter data
* How to conduct and interpret a one-way ANOVA and Tukeys posthoc test, using the the five data analysis steps
* How to conduct and interpret a Pearson’s correlation test.

## Practical 3 learning outcomes

In last week’s practical you learned:

* How to upload a dataset into R
* How to explore a new dataset
* How to conduct the five data analysis steps to answer a scientific question
* How to write hypotheses using IF/THEN statements
* How to conduct and interpret **t-tests and Mann-Whitney-Wilcoxon U tests**
* How to write conclusions based on evidence
* Optional: How to apply a General Linear Model (advanced)

## Preliminary tasks

1. **If you have not completed practicals 1 to 3, you need to do these first.**

2. If you are ready to start Practical 4, **DOWNLOAD the fourth practical worksheet** from Blackboard, do not just view it. To download it, click the “…” next to the file and click ‘Download original file’. **Save this file in an appropriately named folder on your OneDrive (e.g., “MODULE NAME/WEEK X”) so you can access it at home.**

3. **Navigate on your web browser to** [**<https://login.rstudio.cloud/>**](https://login.rstudio.cloud/)**.** Assuming you have already registered for practical 1 to 3, then sign in. If prompted, click “Posit Cloud”.

4. **You should already have an existing R studio project from the first practical called ‘R workshop’,** click on this to enter your saved workspace.

5. **Start a new R script for this practical.** Once your project has loaded (it takes a few seconds), click ‘File’ then ‘New file’, then ‘R Script’.

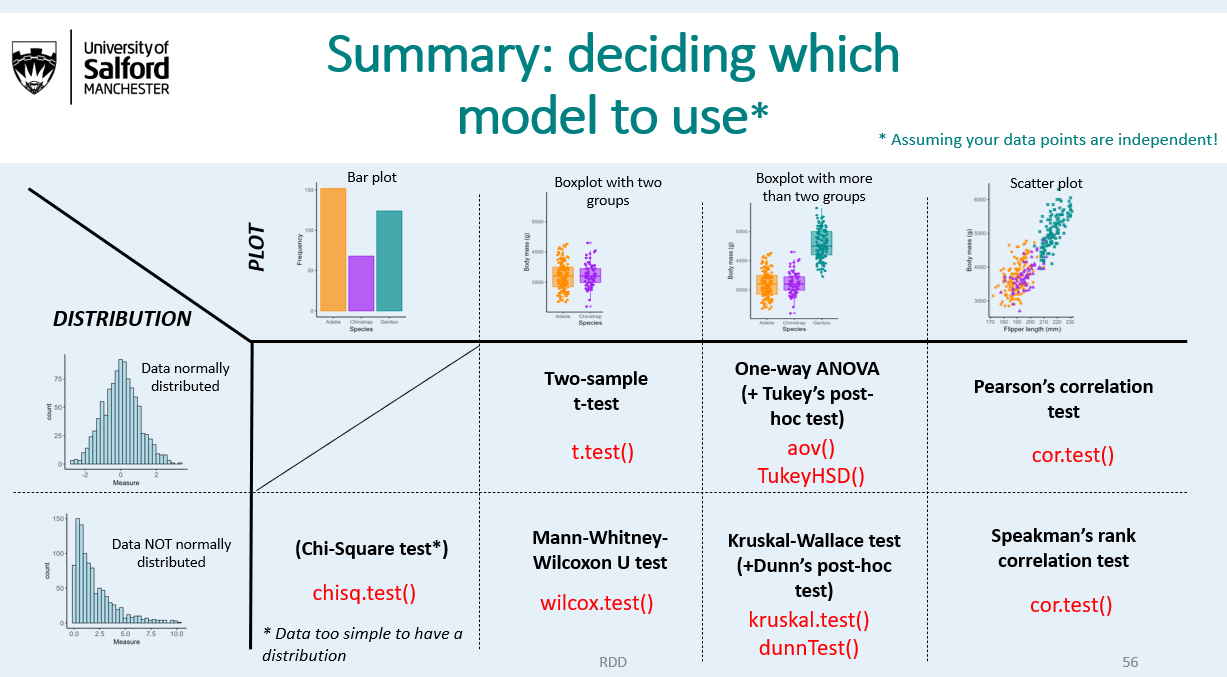
6. **Press the Save button and save your script as “Practical\_4.R”.** This will save all your code you write today.

## Recap: The five data analysis steps

For every analysis you do, you need to go through five steps:

1. Write out your null and alternative hypotheses using if/then statements
2. Plot the relationship between your variables of interest (i.e., make a barplot, boxplot, or scatterplot)
3. Check if your numerical data has a normal (bell shaped) distribution
4. Identify and conduct the appropriate statistical test
5. Write out your results using evidence generated from the statistical model.

**To choose the appropriate statistical model, you will need this guide, which contains the name of the test, and the R function to perform the test (in red):**



## How to conduct statistical tests in R

Below is a table of the code you will need to perform the seven statistical tests you need to know for this module. On the right in an example of how to perform the test using the tiger data. With a new dataset (e.g., for your assignment), simply replace the name of the dataset and the x and y variables that you want to test.

|  |  |
| --- | --- |
| **Test name** | **Code** |
| Two sample T-test | t.test(y~x, data = your\_data) |
| Mann-Whitney-Wilcoxon test | wilcox.test(y~x, data = your\_data) |
| One way ANOVA | summary(aov(y~x, data = your\_data))  *OR*  m<-aov(y~x, data = your\_data)  summary(m) |
| Tukeys posthoc test | TukeyHSD(aov(y~x, data = your\_data))  OR  m<-aov(y~x, data = your\_data)  TukeyHSD(m) |
| Kruskal-Wallace test | kruskal.test(y~x, data = your\_data) |
| Pearson’s correlation test | cor.test(your\_data$x , your\_data$y) |
| Spearman’s correlation test | cor.test(your\_data$x , your\_data$y, method = “spearman”) |
| Chi-square test | chisq.test(your\_data$x) |

## Uploading a new dataset on earthworms

We will now analyse a new dataset on earthworms. Earthworms are crucial ecosystem engineers that are important for soil health and you want to know what factors predict earthworm diversity and abundance across different habitats (farmland, forest, and gardens). This dataset is called ‘worm\_data.csv’.



The dataset consists of data from 92 x 1m squared plots across different habitat types.

**At each of these 92 plots, the researchers collected data how how many species of earthworms they found, as well as the total abundance. They recorded the amount of rainfall that had fallen in the last month, and recorded the month (5 = May, 6 = June, and 7 = July).**

#### Exercise 1

Load this new data into R and call it ‘*worm\_data*’. You did this for tiger\_data in the last practical. Download the dataset from Blackboard into an appropriate OneDrive folder.

If you are using R Studio (Posit) Cloud then you have to:

1. Press ‘Files’ in the bottom right window of your R interface
2. Press the ‘Upload’ button (beige with arrow pointing upwards) and press ‘Browse’ in the File to Upload section, and navigate to worm\_data.csv. From there click ‘open’.
3. Once uploaded to your cloud directory, in the top right window press ‘Import dataset’ and follow the instructions.

If you are using R Studio software, you can skip steps 1 and 2.

Alternatively you can use the code:

**worm\_data<- read.csv(“pathway\_to\_your\_data/worm\_data.csv”)**, making sure to input the correct pathway to where you saved the data.

You should now have worm\_data in your working environment.

## Exploring your new dataset

When you import a new dataset, you should first explore it and check everything is ok.

str(worm\_data) # check data types  
  
head(worm\_data) # check first 6 rows  
  
dim(worm\_data) # check how many rows and columns are in the dataset

#### Exercise 2

What are the names of your variables in your dataset? You can either look at your dataset or use the ***names()*** function.

\*\* WRITE YOUR ANSWERS HERE \*\*  
  
column 1 =   
column 2 =   
column 3 =   
column 4 =   
column 5 =   
column 6 =  
  
---

#### Exercise 3

1. What is the average and median earthworm species diversity?
2. What is the average and maximum earthworm abundance found?
3. How many plots were surveyed in each habitat type?
4. What is your overall sample size (number of plots surveyed)

\*\* WRITE YOUR ANSWERS HERE \*\*

a)

b)

c)

d)  
  
---

## Subsetting data

In exercise 2d, you should have got the following output:



There are two plots that seem to be from unknown habitat. You want to remove these from your dataset before you proceed with analysis.

There are two ways you can subset (or ‘filter’) data. The first is using the function subset().

Lets remove the two rows that are from unknown habitat, and call our new filtered dataset ‘new\_worm\_data’. The code you need is basically saying ‘Keep anything that is NOT ’Unknown’ in the column habitat\_type, and call our new dataset *new\_worm\_data*.

new\_worm\_data <- subset(worm\_data, habitat\_type != "Unknown") #   
  
dim(new\_worm\_data) # 90 rows  
dim(worm\_data) # 92 rows

The ‘!=’ operator means NOT. Some other operators you may need to know are:

|  |  |
| --- | --- |
| Operator | Meaning |
| == | Exactly equal to |
| != | NOT equal to |
| < | Less than |
| > | More than |

Lets practice by making two filtered datasets, one just with plots from farmland, another only from plots sampled in June, and another with only rows with high species diversity.

# keep only plots from farmland  
worm\_data\_farmland <- subset(worm\_data, habitat\_type == "Farmland")  
View(worm\_data\_farmland)  
  
# keep only plots sampled in June  
worm\_data\_june <- subset(worm\_data, month == 6)  
View(worm\_data\_june)  
  
# Keep only plots with over 6 species of earthworm (ie 7 or more)  
worm\_data\_high\_diversity<- subset(worm\_data, worm\_diversity > 6)  
View(worm\_data\_high\_diversity)

## Q1 What is the relationship between habitat type and worm abundance?

Lets move on to analysing our earthworm data. Earthworms are crucial for soil health, yet are sensitive to disturbance. You think that pesticide use might negatively impact earthworm abundance on farmland by directly killing earthworms. You decide to test this.

Step 1: Lets write out the null and alternative hypotheses:

***NULL: If farmland pesticides don’t kill or impact earthworms, then I expect there to be equal numbers of earthworms across farmland, forest and garden habitats***

***ALTERNATIVE: If farmland pesticides kill or negatively impact earthworms, then I expect farmland to have fewer earthworms than garden or forest habitats.***

#### Exercise 4

Step 2: Plot the relationship between habitat type and worm abundance. Make sure you use the new\_worm\_data where you have removed the ‘Unknown’ habitats. What does this plot tell you?

**Remember: you need to load the ggplot2 package before plotting with *library(ggplot2)*.**

\*\* Copy your plot and answers here \*\*  
  
  
  
---

#### Exercise 5

Step 3: Check if your numerical variable is normally distributed of not by looking at a histogram and doing a Shapiro-Wilk test. Is it normally distributed?

**Remember: If the p value for your Shapiro-Wilk test is over 0.05, that is evidence your data IS normally distributed.**

\*\* Copy your plot and Shapiro Wlk test results here \*\*  
  
  
  
  
---

#### Exercise 6

Step 4: Use the ‘how to choose a statistical test’ framework to decide which statistical test you need based on steps 1 and 2 above. Write the name of the test below.

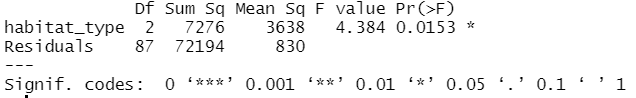
\*\* Write your answer here \*\*  
  
  
  
---

#### Exercise 7

Conduct the statistical test, using the code on your cheatsheet or the code above in the ‘How to conduct Statistical Tests in R’ section. Is the effect of habitat type on earthworm abundance significant?

\*\* Copy your test output here \*\*  
  
  
  
---

You should of got the following output:



From this output you can see that there is a significant effect of habitat type on earthworm abundance (**One-way ANOVA, F = 4.4, p = 0.015**).

**Remember: if your p-value is less than 0.05, that means there is a significant association between your two variables. You are 95% certain that your alternative hypothesis is correct.**

#### Tukeys Posthoc Test

The one-way ANOVA tells you that habitat type is significantly associated with worm abundance. **However, you want to know which habitats are significanlly different from each other in their worm abundance (Farmland-Forest, Forest-Garden, or Farmland-Garden?).**

A graph with arrows and black dots

Description automatically generated with medium confidence

To do this, you need to do a Tukeys posthoc test that tests for significant differences between each pair of habitats.

#### Exercise 8

Conduct a pairwise Tukey’s posthoc test on your ANOVA model. The output will tell you how big the difference is between each pair (“diff”) and the p value for that comparison.

A screenshot of a computer code

Description automatically generated

**Which habitats are significantly different from each other in their worm abundance, and which aren’t? Which habitat pair has the biggest difference in their earthworm abundance?**

\*\* Write your answers here \*\*  
  
  
  
---

#### Exercise 9

Step 5: Below I have written the conclusions, but have left gaps for you to fill in the evidence. You need to add the test name, the effect size, and the p-value for your main test (first gap), and the effect size and p value for your posthoc test.

*There is evidence from the 90 sampled plots that habitat type significantly affects earthworm abundance (\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_), with farmland having the lowest abundance of earthworms. A Tukey’s posthoc test showed that the biggest difference in earthworm abundance was between farmland and forest (\_\_\_\_\_\_\_\_), with farmland having the lowest abundance and forest the highest. Gardens and forest did not have significantly different worm abundance (\_\_\_\_\_\_\_\_). Our results suggest that pesticide use in farmland negatively affects earthworm abundance.*

\*\* Copy the filled in section below \*\*  
  
  
  
---

## Q2: Is there a relationship between rainfall and worm abundance?

Earthworms often come to the surface during wet conditions, making them more detectable. You therefore hypothesize that levels of recent rainfall might positively affect worm abundance. You decide to test this.

#### Exercise 10

Step 1: Write out your null and alternative hypotheses, using IF/THEN statements.

Look at the hypotheses above and in your handout for guidance. I have provided the mechanisms above (i.e. why would you expect more worms when it is raining? What are earthworms doing?).

\*\* Write your hypotheses here\*\*  
  
  
---

#### Exercise 11

Step 2: Plot the relationship between worm abundance and rainfall. Make sure you add neat and capitalised axis labels. Add a linear trend line to your plot. Make sure the variable you are interested in understanding (worm abundance) is on the y axis.

Extra: If you like, colour the points by habitat type.

\*\* Copy your plot here\*\*  
  
  
---

#### Exercise 12

Step 3: Check whether your numberical data (in this case, BOTH worm abundance AND rainfall) are normally ditributed. You need to make two histograms and two shapiro tests for this step.

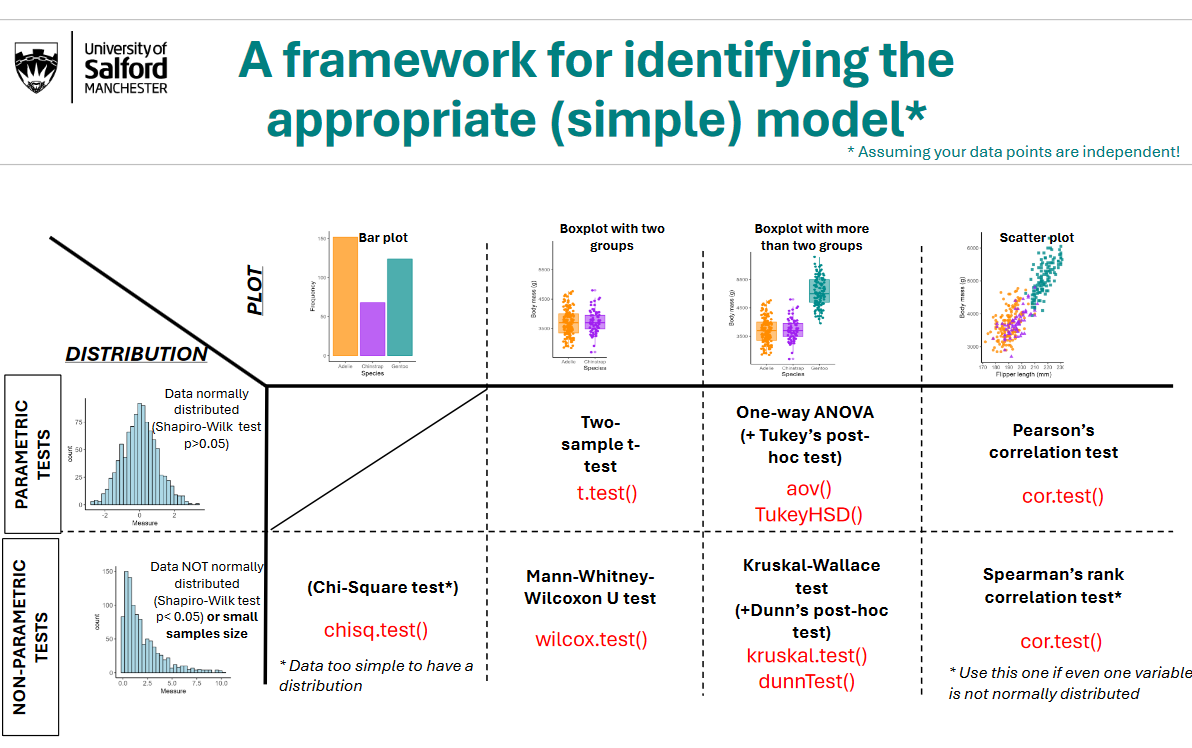
Is worm abundance normally distributed? Is rainfall normally distributed?

\*\* Write your answers here, include your histograms\*\*  
  
Results for worm abundance  
  
  
  
Results for rainfall  
  
  
  
---

#### Exercise 13

Step 4: Based on your plot and your normality tests, identify the appropriate statistical test, using the statistical test framework.

**IMPORTANT: BOTH variables need to be normally distributed to choose a parametric test. If one is not normally distributed, you need a non-parametric test.**



Name the correct test below:

\*\* Write your answer below\*\*  
  
  
  
---

#### Exercise 14

Conduct the correct statistical test. You will need the function cor.test(). Use the cheatsheet to identify the correct code, and copy the code and output below:

\*\*Copy the code and test output below\*\*  
  
  
  
---

#### Exercise 15

Step 5: Write your formal results results section based on the results of your model. Remeber, you need to have your conclusions statement, followed by information on the test you used, the effect size, the p value, and the sample size. Use the “How to write conclusions” handout to guide you.

**Remember: if your p-value is < 0.05, the association between your two variables is statistically significant.**

\*\*Write your results here\*\*  
  
  
  
---

## OPTIONAL: Fit a General Linear Model

Last week in the ‘Optional’ section you learned how to fit a General Lienar Model (GLM) to your data, using the code:

model <- lm(y ~ x + y, data = your\_data)  
  
summary(model)

Where ‘y’ is the variable you want to predict, and x and y are variables you think explain variation in y.

#### Optional Exercise 16

Fit a general linear model to understand the effect of habitat type AND rainfall on worm abundance. Which variable has the biggest effect on worm abundance? Try to write a results section on this question based on the results of the GLM.