

# PS2010 Workshop Code Book

Luke Kendrick

2025-10-23

# Contents

<b>Preface</b>	<b>4</b>
0.1 About this Code Book . . . . .	4
0.2 <b>License</b> . . . . .	4
0.3 <b>Citation</b> . . . . .	4
<b>1 Workshop 1: Data Handling Skills</b>	<b>5</b>
1.1 Exercise 1: Import the Data . . . . .	5
1.2 Exercise 2: Inspect and Check Your Data . . . . .	6
1.3 Exercise 3: Change a Variable Name . . . . .	6
1.4 Exercise 4: Remove a Variable . . . . .	6
1.5 Exercise 5: Filter Cases . . . . .	7
1.6 Exercise 6: Guess Who? . . . . .	7
1.7 Exercise 7: Create a New Variable . . . . .	8
1.8 Exercise 8: Counting and Removing Missing Data . . . . .	9
1.9 Exercise 9: Summary Statistics . . . . .	9
1.10 Exercise 10: Fixing Luke's Broken Code . . . . .	10
<b>2 Workshop 2: Summarising and Describing Data</b>	<b>12</b>
2.1 Exercise 1: Import the Data . . . . .	12
2.2 Exercise 2: Inspect and Check Your Data . . . . .	13
2.3 Exercise 3: Calculate the Stroop Inteference Score . . . . .	13
2.4 Exercise 4: Calculate Descriptive Statistics . . . . .	14
2.5 Exercise 5: Explore Data with Plots . . . . .	15

2.6	Exercise 6: What Does <code>facet_wrap()</code> do? . . . . .	16
2.7	Exercise 7: Save Your Amended Data File . . . . .	16
<b>3</b>	<b>Workshop 3: <i>t</i>-tests</b>	<b>17</b>
3.1	Exercise 1: Import the Data . . . . .	17
3.2	Exercise 2: Inspect and Check Your Data . . . . .	18
3.3	Exercise 3: Check Assumptions . . . . .	18
3.4	Exercise 4: Run the Two-Sample <i>t</i> -Test and ask for Cohen's <i>d</i> . .	19
3.5	Part 2 - Exercise 5: Import the Data . . . . .	20
3.6	Exercise 6: Inspect and Check Your Data . . . . .	21
3.7	Exercise 7: Check Assumptions . . . . .	21
3.8	Exercise 8: Run the Paired <i>t</i> -Test and ask for Cohen's <i>d</i> . . . . .	21
3.9	Exercise 9: Calculate Descriptive Statistics . . . . .	22
<b>4</b>	<b>Workshop 4: One-Way ANOVA (Model, Designs, Assumptions)</b>	<b>24</b>
4.1	Exercise 1: Import the Data . . . . .	25
4.2	Exercise 2: Inspect and Check Your Data . . . . .	25
4.3	Exercise 3: Check Descriptives . . . . .	26
4.4	Exercise 4: Check Assumptions . . . . .	27
4.5	Exercise 5: Interpret the ANOVA Model . . . . .	28
4.6	Part 2 - Exercise 6: Import the Data . . . . .	28
4.7	Exercise 7: Check Descriptives . . . . .	29
4.8	Exercise 8: Check Assumptions . . . . .	29
4.9	Exercise 9: Interpret the ANOVA Model . . . . .	30
<b>5</b>	<b>Workshop 5: One-Way ANOVA (Post-hocs, Contrasts, Write-up)</b>	<b>32</b>
5.1	Exercise 1: Import Data and Run Model . . . . .	32
5.2	Exercise 2: Estimated Marginal Means (EMMs) . . . . .	33
5.3	Part 2 - Exercise 3: Import the Data . . . . .	34
5.4	Exercise 3: Run the ANOVA Model . . . . .	34
5.5	Exercise 4: Post-hoc Pairwise Comparisons . . . . .	35
5.6	Exercise 5: Introducing the Violin Plot . . . . .	36

# Preface

This is the PS2010 Psychological Research Methods and Analysis Workshop Codebook.

## 0.1 About this Code Book

This code book contains information, exercises, and code for the PS2010 workshop sessions.

This resource is a work in progress, and we're continually updating and improving it.

If you spot an error or something that doesn't look quite right, please get in touch:

luke.kendrick@rhul.ac.uk.

## 0.2 License

This work is licensed under Creative Commons Attribution-ShareAlike 4.0 International License (CC-BY-SA 4.0). You are free to share and adapt the material for non-commercial purposes, with appropriate credit and under the same license. If you adapt the material, you must distribute your contributions under the same license.

## 0.3 Citation

Kendrick, L. T. (2025). PS2010 Workshop Code Book (Version 1.0). [https://luke-kendrick.github.io/r\\_codebook](https://luke-kendrick.github.io/r_codebook)

# Chapter 1

## Workshop 1: Data Handling Skills

### Aims:

- Practice importing a .csv data file into RStudio using `read_csv()`
- Practice inspecting your data in RStudio.
- Use different data wrangling functions to develop your data handling skills.
- Check basic summary statistics.

### 1.1 Exercise 1: Import the Data

#### Import the Data File `guess_who.csv`

Before you begin, you will need the tidyverse package loaded.

```
install.packages("tidyverse") #install tidyverse if you do not have it.  
library(tidyverse) #loads tidyverse.
```

Next import the data file and store it as an object called `dataset`.

```
dataset <- read_csv("guess_who.csv")
```

If you see an error saying cannot find function `read_csv()` this usually means you have not loaded (or installed) the tidyverse package.

## 1.2 Exercise 2: Inspect and Check Your Data

Take a look at your newly imported data file

Check the top right panel (the environment) and also use the code below to inspect your data set.

```
view(dataset) # this will open the data in a new tab.  
names(dataset) # this will show the variable names.
```

It is really important to look at the variable names as you'll be using them in code later on.

Answer Question 2.1 - 2.2 on your worksheet.

## 1.3 Exercise 3: Change a Variable Name

One of the variable names is quite long. This can be annoying if we have to keep typing it out

Change the variable name `do_you_own_a_pet` to `pet`. The `rename()` function will let you rename a variable.

```
dataset <- dataset %>%  
  rename(pet = do_you_own_a_pet)
```

Check it has worked:

```
names(dataset) # ask for the variable names again
```

## 1.4 Exercise 4: Remove a Variable

We do not really care about the `age` variable for the next few exercises.

Let's remove it.

The code below will create a new object (once we start removing things, it is best to keep the original data file called `dataset` in the environment)

```
mydata <- dataset %>%  
  select(-age)
```

This code will:

- Create a new object called `mydata`.
- Take our original data called `dataset`
- “And then” `%>%`
- Use the `select()` function to remove `age` by placing a minus symbol `-` in front of it.

From now on, we will use the object called `mydata` and not the original data set.

## 1.5 Exercise 5: Filter Cases

**We can select particular cases in our data set**

For example, I could ask: how many people were from the city of Birmingham using the code below:

```
mydata %>%
  filter(city == "birmingham") %>%
  count()
```

Check the console (bottom left panel) for the answer.

This code will:

- Take `mydata` and then...
- Filter it by the `city` variable.
- We use a double equals symbol `==` to specify an exact match.
- I’ve added “birmingham” in speech marks. Note it is lowercase as to match the data set and then...
- `count()` the number of data points.

Adapt the code above to answer question 5.1 on the worksheet.

## 1.6 Exercise 6: Guess Who?

**We can filter based on multiple criteria.**

The code below will show us someone who is from Brighton, has a dog, and does not drink coffee.

```
mydata %>%  
  filter(city == "brighton", pet == "dog", coffee == 0)
```

We can also use less than/more than symbols to filter data, For example, this will show all people who have a maths enjoyment score of less than 20:

```
mydata %>%  
  filter(maths < 20)
```

Use what you have learned above and adapt your code to play GUESS WHO? and complete questions 6.1-6.2 on the worksheet.

## 1.7 Exercise 7: Create a New Variable

**Sometimes we might want to compute new scores or variables**

Add up the three enjoyment scores for `maths`, `science`, and `art` to create an overall score called `total_score`.

```
mydata <- mydata %>%  
  mutate(total_score = maths + science + art)
```

This code will:

- Take `mydata` to overwrite it (ready to add the new variable) and then...
- Use the `mutate()` function to create a new variable named `total_score` which should equal `= maths + science + art`.

View the data set and look for the new column to see it has worked.

```
view(mydata)
```

Now, we can look at who had the highest and lowest total enjoyment score.

`slice_min` will find the row which has the lowest score:

```
mydata %>%  
  slice_min(total_score)
```

`slice_max` will find the row which has the highest score:



```
mydata %>%  
  slice_max(total_score)
```

Answer questions 7.1-7.2 on the worksheet.

## 1.8 Exercise 8: Counting and Removing Missing Data

### Real data sets often are missing data points

Different people have differing views on how to treat missing data points. For today, we will just identify and remove any. If you view the data, you might notice there are some blanks for **degree** as not everyone is studying for one.

```
sum(is.na(mydata$degree))
```

This code will:

- Calculate the total number using `sum()` of...
- Any missing data points (R calls these `is.na`)
- We can then direct to a particular column using `mydata$degree`. This essentially means “look in `mydata` and then the column called `degree`. We use the dollar sign `$` to specify the column.

If we want to remove them, we can use `filter()` again!

```
mydata <- mydata %>%  
  filter(!is.na(degree))
```

Note: this will overwrite `mydata` and remove the cases.

Answer question 8.1 on the worksheet.

## 1.9 Exercise 9: Summary Statistics

**We might want to know What was the average enjoyment score?**

We can use this code to look across the data set as a whole:

```
summary(mydata)
```

Look through the output in the console (bottom left panel) and answer questions 9.1-9.3 on the worksheet.

## 1.10 Exercise 10: Fixing Luke's Broken Code

Help!! My code below is not working. I need your help to fix it...

Fix the code below to work out how many coffees were drunk by the person from canterbury and is studying medicine. Try running the code first and then work out why it doesn't work!

```
mydata %>%  
  filter(city = "canterburY", degree == medicine)
```

Fix the code below to work out the name of the person who is from London, has a hamster, studies psychology, and did not drink coffee. Try running the code first and then work out why it doesn't work!

```
mydata =  
  filter(city == "London", pet == "hamsta", degree == "psychology", coffee >1)
```

Answer questions 10.1-10.2 on the worksheet.

If you get stuck, use the hints below.

---

Click for a hint

- Check for spelling errors: there are two of them.
- Make sure to use double equals when specifying a label ==.
- Use quote marks when necessary. Some are missing.
- Code is case sensitive. There is a capital letter where there shouldn't be one.
- Use symbols correctly. We want to use the pipe %>% before we filter.
- Use symbols correctly. More than > is not the same as <.

---

---

**Well Done. You have reached the end of the workshop.**

---

## Chapter 2

# Workshop 2: Summarising and Describing Data

### Aims:

- Practice importing a .csv data file into RStudio using `read_csv()`
- Practice inspecting your data in RStudio.
- Calculate mean and standard deviation using the `group_by()` and `summarise()` functions.
- Visually inspect data using plots and describe data distributions.

## 2.1 Exercise 1: Import the Data

### Import the Data File: `stroop.csv`

Before you begin, you will need the tidyverse package loaded.

```
install.packages("tidyverse") #install tidyverse if you do not have it.  
library(tidyverse) #loads tidyverse.
```

Next import the data file and store it as an object called `dataset`.

```
dataset <- read_csv("stroop.csv")
```

If you see an error saying cannot find function `read_csv()` this usually means you have not loaded (or installed) the tidyverse package.

## 2.2 Exercise 2: Inspect and Check Your Data

### Take a look at your newly imported data file

Check the top right panel (the environment) and also use the code below to inspect your data set.

```
view(dataset) # this will open the data in a new tab.  
names(dataset) # this will show the variable names.
```

It is really important to look at the variable names as you'll be using them in code later on.

## 2.3 Exercise 3: Calculate the Stroop Inteference Score

### Sometimes we might want to compute new scores or variables

Calculate the Stroop interference measure. This should be the difference between the incongruent and congruent conditions. Take the `incongruent` reaction times and then subtract the `congruent` reaction times using the code below:

```
mydata <- dataset %>%  
  mutate(int = incongruent - congruent)
```

This code will:

- Create an object called `mydata` before using the original `dataset` and then...
- Use the `mutate()` function to create a new variable named `int` which should equal `incongruent - (minus) congruent`.

View the data set and look for the new column to see it has worked. Note: check the final column to see if `int` has appeared.

```
view(mydata)
```

What exactly is this Stroop Interference thingy? If you want to learn more, see below.

[Click for more information](#)

The interference measure in milliseconds (msecs) is the amount of extra time it took a participant to answer the incongruent (trickier trials because the colours do not match the word) compared to the congruent trials (easier trial because the colours do match the word).

In a sense, it is how many milliseconds slower you are because you need to focus your attention and engage executive functions to fight the urge to read the word rather than name the colour.

A smaller number is perhaps indicative of having better attention/executive functioning processes!

---

Answer questions 3.1-3.2 on the worksheet.

## 2.4 Exercise 4: Calculate Descriptive Statistics

Just as shown in the lecture, use the code below to calculate the mean and standard deviation for Stroop interference or `int`. Remember, we want to use `int` that was calculated in exercise 3.

```
desc <- mydata %>%  
  group_by(NULL1) %>%  
  summarise(mean_int = mean(NULL2),  
            sd_int = sd(NULL2))
```

You will need to change `NULL` to match your data set. Try and give this a go on your own first, but if you aren't sure look below for help.

Think about:

- For `NULL1`: What is the name of the variable you will split the data file by (e.g., what is the grouping variable/independent variable called in `mydata`)
- For `NULL2`: What is the name of the score that you want to find the mean and standard deviation for (e.g., what is the dependent variable called in `mydata`)

---

[Click for a hint](#)

```
desc <- mydata %>%  
  group_by(drink) %>%  
  summarise(mean_int = mean(int),  
            sd_int = sd(int))
```

---

If you look in the environment (top right panel) you will see a new object called **desc**. This is where your descriptive statistics are stored. I called it **desc** but you can call it anything you like. It is best to keep object names short and informative. We can now view that object using the **view()** function.

```
view(desc)
```

Answer questions 4.1-4.2 on the worksheet.

## 2.5 Exercise 5: Explore Data with Plots

Generate a box plot:

```
ggplot(mydata, aes(x = drink, y = int)) +  
  geom_boxplot(width = .4)
```

Generate histograms:

```
ggplot(mydata, aes(x = int, fill = drink)) +  
  geom_histogram(colour = "black") +  
  facet_wrap(~ drink)
```

Generate density plots:

```
ggplot(mydata, aes(x = int, fill = drink)) +  
  geom_density(alpha = .5) +  
  facet_wrap(~ drink)
```

Answer questions 5.1-5.2 in the worksheet.

## 2.6 Exercise 6: What Does `facet_wrap()` do?

Re-run the density plot code, except this time delete the final line. This will show what `facet_wrap()` does. What do you notice about the plot now?

Use this code without `facet_wrap()`.

```
ggplot(mydata, aes(x = int, fill = drink)) +  
  geom_density(alpha = .5)
```

Answer question 6.1 on the worksheet.

## 2.7 Exercise 7: Save Your Amended Data File

Your current data file has the `int` column in, calculated in exercise 3. You need to save it so you can use it for next week's workshop. You can overwrite your original `.csv` file using this code, which will save today's data set.

```
write.csv(mydata, "stroop.csv")
```

This means the `stroop.csv` file on your computer will be updated and ready to use next week! Make sure you know where it has saved on your computer before you leave. You will need this file next week!

---

**Well Done. You have reached the end of the workshop.**

---



## Chapter 3

# Workshop 3: *t*-tests

### Aims:

- Practice running and interpreting a two-sample *t*-test in RStudio.
- Practice running and interpreting a paired *t*-test in RStudio.

### Part One

Part one of today's workshop will involved running a two-sample *t*-test, which is appropriate for an independent measures design with **two** groups.

You should use the same data file from last week, as you will be aiming to investigate whether Stroop interference scores (msecs) significantly differ between the Red Cow group and the control group.

You must have completed the week 2 workshop before starting this one.

## 3.1 Exercise 1: Import the Data

### Import the Data File: stroop.csv

Before you begin, you will need the following packages:

```
install.packages("tidyverse") #install if needed.  
install.packages("rstatix")  #install if needed.  
install.packages("car")      #install if needed.  
library(tidyverse)           #load package.  
library(rstatix)              #load package.  
library(car)                  #load package.
```

Then import the data file. Make sure it has the `int` score that you calculated last week. If not, you will need to go back and complete workshop 2.

```
mydata <- read_csv("stroop.csv")
```

## 3.2 Exercise 2: Inspect and Check Your Data

Take a look at your newly imported data file

Check the top right panel (the environment) and also use the code below to inspect your data set.

```
view(mydata) # this will open the data in a new tab.  
names(mydata) # this will show the variable names.
```

It is really important to look at the variable names as you'll be using them in code later on.

## 3.3 Exercise 3: Check Assumptions

### 3.3.1 Check Heteroscedasticity with Levene's Test

We do not need to do this because we will just use a Welch's *t*-test instead.

If you need to justify this decision (e.g., for a 3rd Year Project), the following papers might help:

- Delacre et al. (2017)
- Ruxton (2006)
- For a gentler explanation, see also this blog post from Daniel Lakens.

### 3.3.2 Check Normality with Shapiro-Wilk Test and Histograms

We need to check normality for both groups separately. We can filter the groups:

```
#first create a data set that contains the control group only
control <- mydata %>%
  filter(drink == "control")
#then run the test
shapiro.test(control$int)

#then create a data set that contains Red Cow drinkers only
redcow <- mydata %>%
  filter(drink == "redcow")
#then run the test
shapiro.test(redcow$int)
```

We use a dollar sign \$ to point R to a particular column. For example, when using `shapiro.test(redcow$int)` you are saying to run the Shapiro Test on the `redcow` only data set and the specific column called `int`

Again, we want the  $p$ -values to be not significant. A non-significant  $p$ -value means the data are roughly normally distributed. If the  $p$ -value is significant, this could be an issue as it indicates the data are not normally distributed.

When reporting the Shapiro-Wilk test, you just need to report the test statistics ( $w = XX$ ) and the  $p$ -value. Here is an example what it could look like:

$$W = 0.98, p = .875$$

You can also visually check the data with a quick histogram:

```
hist(control$int)
hist(redcow$int)
```

Check the plots panel, and use the blue arrow to switch between the two plots.

### 3.4 Exercise 4: Run the Two-Sample $t$ -Test and ask for Cohen's $d$

Run the  $t$ -test.

```
t.test(int ~ drink, data = mydata, var.equal = FALSE, alternative = "two.sided")
```

Ask for Cohen's  $d$ :

```
cohens_d(data = mydata, int ~ drink, var.equal = FALSE)
```

Interpret your *t*-test.

- Is the test significant?
- What is the effect size?
- If significant, how do the groups differ (e.g., use descriptive statistics to interpret the difference)?
- Hint: re-use the code from last week to find the mean and standard deviation for the two groups.

Answer questions 4.1-4.3 on the worksheet.

## 3.5 Part 2 - Exercise 5: Import the Data

### Part Two Study

Part two of today's workshop will involve running a paired *t*-test, which is appropriate for a repeated measures design with **two** conditions.

Let's switch it up a bit and use a different scenario and experiment with a new data set.

Here we will look at how someone's resting heart rate might change as a result of drinking a can of Red Cow.

In this study the resting heart rate in beats per minute (BPM) was measured in a group of participants both **before** and **after** consuming a can of Red Cow.

The independent variable is time point: before, after. The dependent variable is BPM.

### Import the Data

We will call the object **dat** (short name for data)

```
dat <- read_csv("bpm.csv")
```

### 3.6 Exercise 6: Inspect and Check Your Data

Take a look at your newly imported data file

Check the top right panel (the environment) and also use the code below to inspect your data set.

```
view(dat) # this will open the data in a new tab.  
names(dat) # this will show the variable names.
```

### 3.7 Exercise 7: Check Assumptions

We do need to check normality, as heteroscedasticity does not apply to repeated measures designs. However, we need to ensure the **difference** score is normally distributed.

```
diff <- dat$after - dat$before #this will calculate the difference score.  
shapiro.test(diff) # run the Shapiro test  
hist(diff) # also visually inspect data
```

You should interpret and report this in the same way as earlier (exercise 3).

Answer questions 7.1-7.2 on the worksheet.

### 3.8 Exercise 8: Run the Paired $t$ -Test and ask for Cohen's $d$

A paired  $t$ -test is a little different compared to the two-sample  $t$ -test.

```
t.test(NULL1, NULL2, paired = TRUE)
```

- Change NULL1 to the column with the first condition.
- Change NULL2 to the column with the second condition.

(Hint: you will need to use the dollar sign  $\$$  to specify which data set and column, e.g., `dat$NULL1` and `dat$NULL2`).

Try yourself first, but if you need, check the solution below.

Click for a hint

```
t.test(dat$before, dat$after, paired = TRUE)
```

---

Annoyingly, we need to use a different package for Cohen's d for a paired t-test.

```
install.packages("effectsize") #install if needed.
library(effectsize)
effectsize::cohens_d(dat$before, dat$after, paired = TRUE)
```

Answer questions 8.1-8.2 on the worksheet. Hint: you will also need the descriptive statistics (see exercise 9 below).

### 3.9 Exercise 9: Calculate Descriptive Statistics

The final thing to do is to convert the data from wide format to long format. Often with repeated measures when asking for descriptive statistics or plots, we need the data in long format.

```
longd <- dat %>%
  pivot_longer(
    before:after,
    names_to = "time",
    values_to = "bpm"
  )
```

Code explanation:

- Create a new object called `longd` where we will store the long data.
- Base it on the original data set called `dat` and then `%>%`
- `pivot_longer()`
- The first argument should include the columns which contain your dependent variable. In this case we will use `before:after` which will then take all and any columns from `before` through to `after` (these are the only columns so it will just take the two of them).
- Next we use `names_to =` to tell R what we want to call our independent variable. I have used “time” as it was time point for this study.

- Then we use `values_to =` to tell R what we want to call our dependent variable. We measure beats per minute, so I have just called this “bpm”.
- Pay careful attention to the lay out of this code. For example, notice where brackets open and close, and the placement of commas to move on the the next line.

Once you have created the new long data set called `longd` we can use it to calculate descriptive statistics. But first check it to see the difference.

```
view(longd)
names(longd)
```

Now adapt the code below to ask for the mean and standard deviation.

```
desc <- longd %>%
  group_by(NULL1) %>%
  summarise(mean_bpm = mean(NULL2)
            sd_bpm = sd(NULL2))
```

Change `NULL1` to the name of the independent variable in the `long` data file, and `NULL2` to the dependent variable. If this is tricky, use `names(longd)` if you need a reminder of the variable names and revisit your notes from last week’s workshop.

Make note of the mean and standard deviation for the **before** and **after** conditions.

Try yourself first, but if you need, check the solution below.

---

Click for a hint

```
desc <- longd %>%
  group_by(time) %>%
  summarise(mean_bpm = mean(bpm),
            sd_bpm = sd(bpm))

view(desc)
```

---

**Well Done. You have reached the end of the workshop.**

---

## Chapter 4

# Workshop 4: One-Way ANOVA (Model, Designs, Assumptions)

### Aims:

- Practice running one-way ANOVAs for both repeated and independent measures designs.
- Practice checking ANOVA assumptions and interpreting the ANOVA model output.

### Part One

Part one of today's workshop will involve running a one-way ANOVA for a repeated measures design.

Mild cognitive impairment (MCI) occurs when someone begins to have difficulty with their memory or other cognitive abilities, however, these are (1) not usually significant enough to interfere with general activities of daily living and (2) do not fully meet the criteria for dementia.

This data set contains data from a new memory assessment used for individuals with MCI across a four year period (each participant was assessed once per year).

Using a one-way ANOVA, can you determine what happens to memory scores over time?



## 4.1 Exercise 1: Import the Data

### Import the Data File

Before you begin you will need a few packages loaded.

```
library(tidyverse) # should already be installed.  
  
install.packages("afex") # install if needed  
install.packages("car") #install if needed  
library(afex)  
library(car)
```

Next import the data file and store it as an object called `dat`.

```
dat <- read_csv("mci.csv")
```

## 4.2 Exercise 2: Inspect and Check Your Data

### Take a look at your newly imported data file

Check the top right panel (the environment) and also use the code below to inspect your dataset.

```
view(dat) # open data in new tab  
names(dat) # print variable names to console
```

Is this data in wide format or long format?

If the data is wide, you will need to convert it to long format using the code below.

```
mydata <- dat %>%  
  pivot_longer(cols = year1:year4,  
               names_to = "time",  
               values_to = "memory")
```

Here is an explanation of this code line by line. This one is a little tricky.

- First, create a new object called `mydata` and base it on the existing `dat` and then...

- Pivot the data (from long to wide) by taking all columns starting with `year1` through to `year4`. By stating `year1:year4` you are telling to take all columns from `year1` through to `year4`, including anything in between. Essentially you are stating the first and final levels of your independent variable.
- `names_to` should include a name you assign to the independent variable. In this study it is time point but we will just use `time` as it is shorter.
- `values_to` should include a name you assign to the dependent variable. In this study it was memory assessment score, but we will just use `memory`.

Run the code and then check `mydata` to see if it worked.

```
view(mydata)
names(mydata)
```

From now on we will only use the `mydata` object, as it is in long format.

### 4.3 Exercise 3: Check Descriptives

Using the code below, generate descriptive statistics to look at the mean memory score across time.

```
desc <- mydata %>%
  group_by(time) %>%
  summarise(m_memory = mean(memory),
            sd_memory = sd(memory))

view(desc) # view the descriptive stats
```

Also, visualise the data using a box plot or density plot.

```
ggplot(mydata, aes(x = time, y = memory)) +
  geom_boxplot(width = .4)
```

```
ggplot(mydata, aes(x = memory, fill = time)) +
  geom_density(alpha = .5)
```

`facet_wrap()` is missing from the density plots, however, if you feel it would be beneficial to aid interpretation, you can add it.

## 4.4 Exercise 4: Check Assumptions

### Run the ANOVA model

Before we start looking at assumptions, we do need to run and store the initial ANOVA model.

```
mod <- aov_ez(id = "id",  
             dv = "memory",  
             within = "time",  
             type = 3,  
             include_aov = TRUE,  
             data = mydata)
```

### Checking Normality

First, look to see if the assumption of normality was met. You should look at three things.

- 1) The QQ-Plot
- 2) The histogram
- 3) The Shapiro-Wilk test

```
#testing normality, looking at normal distribution of residuals from the model.  
residuals <- residuals(mod) #pulls residuals from the model
```

```
# produce QQ-Plot  
qqPlot(residuals) #produces a qq-plot of residuals
```

```
# produce histogram of residuals  
hist(residuals, breaks = 20, main = "Histogram of Residuals", xlab = "Residuals", col = "lightblue")
```

```
# Shapiro test on residuals  
shapiro.test(residuals) #Shapiro test for residuals
```

### Checking Sphericity

To check sphericity, we need to produce the model output.

```
summary(mod) # provides the output of the main ANOVA model with Sphericity output.
```

Look through the output to find: “Mauchly Tests for Sphericity”.

It will provide a test statistic, reported using  $W = XX.XX$

It will provide a  $p$ -value statistic. For this test, a non-significant  $p$ -value means the assumption was met. A non-significant  $p$ -value is one that is a larger value than .05. A significant  $p$ -value is one smaller than .05.

It looks as though this  $p$ -value is smaller than .05 and so the assumption is violated. uh-oh!

## 4.5 Exercise 5: Interpret the ANOVA Model

For an easier and more streamlined output to interpret the model, use the code below:

```
mod$anova_table
```

You should interpret this the same way as was explained in the lecture. Particularly noting the  $F$ , degrees of freedom (two values), and  $p$ -value (noting whether it was significant or not).

This model simply tells us: Should we reject the null hypothesis.

As this is a significant result,  $F(2.15, 105.54) = 127.16$ ,  $p < .001$ , then we can reject our null hypothesis as at least one of the conditions differs. How they differ we do not yet know and you will learn about that next week...

## 4.6 Part 2 - Exercise 6: Import the Data

### Part Two

In this part of the workshop you should conduct a one-way ANOVA for an independent measures design. The general interpretation is the same as above, but as covered in the lecture, the assumptions differ slightly.

In this study, participants were placed into one of three groups based on their circadian chronotype: in other words, are they an early bird (morning person), night owl (evening person), or neither (do not fit into either category). This means we have three independent groups.

You can try this out yourself here: [link](#). Probably best to try this after your workshop.

The research question here is a simple one: does the early bird really catch the worm? or more scientifically put...who is more alert when completing an attention task at 9am.

The independent variable is circadian chronotype, 3 levels. The dependent variable is **score** on an attention task from 0-100 (100 is highly alert/top score).

## Import the data

The file is called `alert.csv`. Import the data file and save it as an object called `mydata1` (as to not overwrite anything from Part 1 in case you want to look back)

## 4.7 Exercise 7: Check Descriptives

Amend the code below to produce descriptive statistics.

```
names(mydata1) # might be useful to check variable names
view(mydata1)  # might be useful to look at the data

desc1 <- mydata1 %>%
  group_by(NULL1) %>%
  summarise(mean_score = mean(NULL2),
            sd_score = sd(NULL2))
```

Change:

- `NULL1` to the variable you want to split the data by. This should be the independent variable. This should be the dependent variable as it is names in `mydata1`.
- `NULL2` to the variable you want to calculate the mean and sd for. This should be the dependent variable as it is names in `mydata1`.

Also use this code to visualise the data to see what is going on.

```
ggplot(mydata1, aes(x = group, y = score)) +
  geom_boxplot(width = .4)

ggplot(mydata1, aes(x = score, fill = group)) +
  geom_density(alpha = .5)
```

## 4.8 Exercise 8: Check Assumptions

Before we start looking at assumptions, we do need to run and store the initial ANOVA model. There are **four** `NULL` values you need to change in the code below to match your new data set.

We will save the model as an object called `mod1`.

```
mod1 <- aov_ez(id = "NULL",
               dv = "NULL",
               between = "NULL",
               type = 3,
               include_aov = TRUE,
               data = NULL)
```

### Checking Normality

First, look to see if the assumption of normality was met. You should look at three things.

- 1) The QQ-Plot
- 2) The histogram
- 3) The Shapiro-Wilk test

```
#testing normality, looking at normal distribution of residuals from the model.
residuals <- residuals(mod1) #pulls residuals from the model, note name is `mod1`

# produce QQ-Plot
qqPlot(residuals) #produces a qq-plot of residuals

# produce histogram of residuals
hist(residuals, breaks = 20, main = "Histogram of Residuals", xlab = "Residuals", col = "red")

# Shapiro test on residuals
shapiro.test(residuals) #Shapiro test for residuals
```

### Checking Homogeneity of Variance

Run and interpret Levene's test using the code below:

```
leveneTest(score ~ group, mydata1)
```

## 4.9 Exercise 9: Interpret the ANOVA Model

Use the code below to interpret the ANOVA model. Decide if the model is significant and whether you can reject the null hypothesis. Also check back at the descriptive statistics to see if you can work out what could be going on.

```
mod1$anova_table
```

---

Well Done. You have reached the end of the workshop

---

## Chapter 5

# Workshop 5: One-Way ANOVA (Post-hocs, Contrasts, Write-up)

### Aims:

- Practice using a post-hoc test to interpret an ANOVA
- Practice using a planned contrast to interpret an ANOVA result.
- Practice writing-up the results from an ANOVA, using APA style.

### Part One

## 5.1 Exercise 1: Import Data and Run Model

Last week you ran a one-way repeated measures ANOVA looking at memory assessment score over time. This week you will need the same dataset and the code (below) to run the initial model.

Remember - you already checked the assumptions **and** the ANOVA model was significant, in which case you will now practice using a planned contrast.

To save time, use the code below to load necessary packages, import the data, and run the model. Try running **ALL** of this code at once to save time (highlight it all and run it). Make sure your working directory is set up correctly first to find the `mci.csv` file.



```
install.packages("afex") # install if needed
install.packages("car") #install if needed
```

```
library(tidyverse)
library(afex)
library(car)

dat <- read_csv("mci.csv")
mydata <- dat %>%

  pivot_longer(cols = year1:year4,
               names_to = "time",
               values_to = "memory")

mod <- aov_ez(id = "id",
             dv = "memory",
             within = "time",
             type = 3,
             include_aov = TRUE,
             data = mydata)

mod$anova_table
```

## 5.2 Exercise 2: Estimated Marginal Means (EMMs)

The code above should import the data, switch it to long format, and run the model.

As the model was significant, we need to pull out the estimated marginal means.

We need a specific package for this:

```
install.packages(emmeans) # install if needed.
library(emmeans)
```

Next, pull out the EMMs and save them as an object called `emms`.

```
emms <- emmeans(mod, ~ time)
```

Now let's say that the alternate hypothesis for this study was that there would be a linear decline in memory scores over time, we could use a polynomial contrast. This will look for a trend. Polynomial contrasts are only suitable for

an independent variable which could be considered “continuous” in nature, e.g., time point.

Use the code below to run the contrast.

```
poly <- contrast(emms, method = "poly")  
  
print(poly) # print the results to the console
```

Look through the output. Remember, you should look for  $p$ -values that are less than .05 ( $p < .05$ ).

Is there a significant linear, quadratic, or cubic trend?

Additionally, use the code below to visualise this.

```
#let's visualise the data to help our interpretation  
ggplot(mydata_long, aes(x = time, y = score, fill = time)) +  
  stat_summary(fun = mean, geom = "line", color = "black", size = 1, aes(group = 1)) +  
  theme_classic() +  
  labs(title = "Plot of Memory Score Across Time", # add a title  
       x = "Time Point", # X-axis label  
       y = "Memory Score") # Y-axis label
```

## 5.3 Part 2 - Exercise 3: Import the Data

### Part Two

Part two of this workshop will use the same dataset and ANOVA model from part two of workshop 4 (last week).

Ensure you have the correct data file loaded.

The file is called `alert.csv`. Import the data file and save it as an object called `mydata1` (as to not overwrite anything from Part 1 in case you want to look back).

## 5.4 Exercise 3: Run the ANOVA Model

Next, run the ANOVA model again and save it as an object.

We will save the model as an object called `mod1`.

```
mod1 <- aov_ez(id = "id",
               dv = "score",
               between = "group",
               type = 3,
               include_aov = TRUE,
               data = mydata1)

mod1$anova_table
```

Last week you would have seen that the model was significant (p-value was less than .05). This means we reject the null hypothesis and at least **one** of the groups are significantly different.

## 5.5 Exercise 4: Post-hoc Pairwise Comparisons

To run post-hoc tests, you first need to pull out the estimated marginal means using `emmeans()` and apply the necessary p-value adjustment.

For this exercise, use the Holm adjustment.

```
emmeans(mod1,
         pairwise ~ group,
         adjust = "holm")
```

Look at the output and interpret it. Which conditions differ from one another? Remember to look for *p*-values that are less than .05.

Also look for effect sizes, so you can report them alongside the rest of your analysis.

```
cohens_d(mydata1, score ~ group)
```

Use a box plot to also help your interpretation:

```
#how about a box plot too

ggplot(mydata, aes(x = group, y = alert, fill = group)) +
  geom_boxplot() +
  theme_classic()
```

## 5.6 Exercise 5: Introducing the Violin Plot

Previously, you have looked at density plots and box plots. Ever wondered if there was a handy way to look at both the distributions and the box plot at the same time? No, probably not. Well I will show you anyway.

Use this code to generate a violin plot:

```
ggplot(mydata1, aes(x = group, y = score, fill = group)) +  
  geom_violin(trim = FALSE) + # violin plot without trimming tails  
  geom_boxplot(width = 0.1, color = "black", alpha = 0.5) + # overlay a boxplot for a  
  theme_classic() + # Classic theme  
  labs(title = "Violin Plot of Alertness Score by Group", # add a title  
        x = "Group", # x-axis label  
        y = "Alertness Score") # y-axis label
```

---

Well Done. You have reached the end of the workshop

---