Penguin Assignment Q4

2022-12-06

Question 04

Load all necessary packages

```
library(palmerpenguins)
library(ggplot2)
library(tidyverse)
library(janitor)
library(dplyr)
library(svglite)
library(ragg)
library(car)
```

Visualising our data frame

Can see that the column headers are messy and not uniform, they need to be cleaned up.

```
head(penguins_raw)

## # A tibble: 6 × 17
##   study…¹ Sampl…² Species Region Island Stage Indiv…³ Clutc…⁴ `Date
Egg` Culme…⁵
##   <chr>     <dbl> <chr>   <chr>  <chr>  <chr> <chr>   <chr>
<date>       <dbl>
## 1 PAL0708       1 Adelie… Anvers Torge… Adul… N1A1    Yes
2007-11-11    39.1
## 2 PAL0708       2 Adelie… Anvers Torge… Adul… N1A2    Yes
2007-11-11    39.5
## 3 PAL0708       3 Adelie… Anvers Torge… Adul… N2A1    Yes
2007-11-16    40.3
## 4 PAL0708       4 Adelie… Anvers Torge… Adul… N2A2    Yes
2007-11-16    NA
## 5 PAL0708       5 Adelie… Anvers Torge… Adul… N3A1    Yes
2007-11-16    36.7
## 6 PAL0708       6 Adelie… Anvers Torge… Adul… N3A2    Yes
2007-11-16    39.3
## # … with 7 more variables: `Culmen Depth (mm)` <dbl>,
## #   `Flipper Length (mm)` <dbl>, `Body Mass (g)` <dbl>, Sex <chr>,
## #   `Delta 15 N (o/oo)` <dbl>, `Delta 13 C (o/oo)` <dbl>, Comments
<chr>, and
## #   abbreviated variable names ¹studyName, ²`Sample Number`, ³
`Individual ID`,
## #   ⁴`Clutch Completion`, ⁵`Culmen Length (mm)`
```

To save a version of our original dataframe

```
write.csv(penguins_raw, "data_raw/penguins_raw.csv")
```

Creating a cleaning function

```
cleaning <- function(data_raw){
  data_raw %>%
    select(-starts_with("Delta")) %>%
    select(-Comments) %>%
    clean_names() %>%
    remove_empty(c("rows","cols"))
}
```

Applying our cleaning function to the raw penguin data, creating a new dataframe called penguins_clean.

```
penguins_clean <- cleaning(penguins_raw)

write.csv(penguins_clean, "data_clean/penguins_clean")

penguins_clean

## # A tibble: 344 × 14
##    study_name sample_nu…¹ species region island stage indiv…²
clutc…³ date_egg
##    <chr>            <dbl> <chr>   <chr>  <chr>  <chr> <chr>   <chr>
<date>
##  1 PAL0708              1 Adelie… Anvers Torge… Adul… N1A1    Yes
2007-11-11
##  2 PAL0708              2 Adelie… Anvers Torge… Adul… N1A2    Yes
2007-11-11
##  3 PAL0708              3 Adelie… Anvers Torge… Adul… N2A1    Yes
2007-11-16
##  4 PAL0708              4 Adelie… Anvers Torge… Adul… N2A2    Yes
2007-11-16
```

```
##   5 PAL0708                5 Adelie… Anvers Torge… Adul… N3A1     Yes
2007-11-16
##   6 PAL0708                6 Adelie… Anvers Torge… Adul… N3A2     Yes
2007-11-16
##   7 PAL0708                7 Adelie… Anvers Torge… Adul… N4A1     No
2007-11-15
##   8 PAL0708                8 Adelie… Anvers Torge… Adul… N4A2     No
2007-11-15
##   9 PAL0708                9 Adelie… Anvers Torge… Adul… N5A1     Yes
2007-11-09
##  10 PAL0708               10 Adelie… Anvers Torge… Adul… N5A2     Yes
2007-11-09
## # … with 334 more rows, 5 more variables: culmen_length_mm <dbl>,
## #   culmen_depth_mm <dbl>, flipper_length_mm <dbl>, body_mass_g
<dbl>,
## #   sex <chr>, and abbreviated variable names ¹sample_number, ²
individual_id,
## #   ³clutch_completion
```

Running a statistical test

First I chose what question I wanted to ask. Is there a difference between male and female body mass in Adelie penguins?

H0: The difference between male and female Adelie Penguin body mass mean is 0 . µ1 = µ2

H1: The difference between male and female Adelie Penguin body mass mean is not 0. µ1 ≠ µ2

As I am comparing the means of two independent groups to each other I will run a two-sample t-test. To do this I need to remove data from my dataframe that I am not interested in; i.e. non-Adelie penguins.

```
adelie0 <- penguins_clean[-c(153:344), ]
adelie <- na.omit(adelie0)
adelie

## # A tibble: 146 × 14
##    study_name sample_nu…¹ species region island stage indiv…²
clutc…³ date_egg
##    <chr>            <dbl> <chr>   <chr>  <chr>  <chr> <chr>   <chr>
<date>
##   1 PAL0708                1 Adelie… Anvers Torge… Adul… N1A1     Yes
```

```
2007-11-11
##  2 PAL0708              2 Adelie… Anvers Torge… Adul… N1A2    Yes
2007-11-11
##  3 PAL0708              3 Adelie… Anvers Torge… Adul… N2A1    Yes
2007-11-16
##  4 PAL0708              5 Adelie… Anvers Torge… Adul… N3A1    Yes
2007-11-16
##  5 PAL0708              6 Adelie… Anvers Torge… Adul… N3A2    Yes
2007-11-16
##  6 PAL0708              7 Adelie… Anvers Torge… Adul… N4A1    No
2007-11-15
##  7 PAL0708              8 Adelie… Anvers Torge… Adul… N4A2    No
2007-11-15
##  8 PAL0708             13 Adelie… Anvers Torge… Adul… N7A1    Yes
2007-11-15
##  9 PAL0708             14 Adelie… Anvers Torge… Adul… N7A2    Yes
2007-11-15
## 10 PAL0708             15 Adelie… Anvers Torge… Adul… N8A1    Yes
2007-11-16
## # … with 136 more rows, 5 more variables: culmen_length_mm <dbl>,
## #   culmen_depth_mm <dbl>, flipper_length_mm <dbl>, body_mass_g
<dbl>,
## #   sex <chr>, and abbreviated variable names ¹sample_number, ²
individual_id,
## #   ³clutch_completion
```

Before I can run the t-test I need to test its assumptions. First I test that both populations’ (male and female) body mass are normally distributed and then test whether they have similar variance.

Creating a dataframe with only female Adelie penguins.

```
adelieF <- adelie %>% filter(sex == 'FEMALE')
adelieF

## # A tibble: 73 × 14
##    study_name sample_nu…¹ species region island stage indiv…²
clutc…³ date_egg
##    <chr>            <dbl> <chr>   <chr>  <chr>  <chr> <chr>   <chr>
<date>
##  1 PAL0708              2 Adelie… Anvers Torge… Adul… N1A2    Yes
2007-11-11
##  2 PAL0708              3 Adelie… Anvers Torge… Adul… N2A1    Yes
```

```
2007-11-16
##  3 PAL0708              5 Adelie… Anvers Torge… Adul… N3A1    Yes
2007-11-16
##  4 PAL0708              7 Adelie… Anvers Torge… Adul… N4A1    No
2007-11-15
##  5 PAL0708             13 Adelie… Anvers Torge… Adul… N7A1    Yes
2007-11-15
##  6 PAL0708             16 Adelie… Anvers Torge… Adul… N8A2    Yes
2007-11-16
##  7 PAL0708             17 Adelie… Anvers Torge… Adul… N9A1    Yes
2007-11-12
##  8 PAL0708             19 Adelie… Anvers Torge… Adul… N10A1   Yes
2007-11-16
##  9 PAL0708             21 Adelie… Anvers Biscoe Adul… N11A1   Yes
2007-11-12
## 10 PAL0708             23 Adelie… Anvers Biscoe Adul… N12A1   Yes
2007-11-12
## # … with 63 more rows, 5 more variables: culmen_length_mm <dbl>,
## #   culmen_depth_mm <dbl>, flipper_length_mm <dbl>, body_mass_g
<dbl>,
## #   sex <chr>, and abbreviated variable names ¹sample_number, ²
individual_id,
## #   ³clutch_completion
```

Creating a dataframe with only male Adelie penguins.

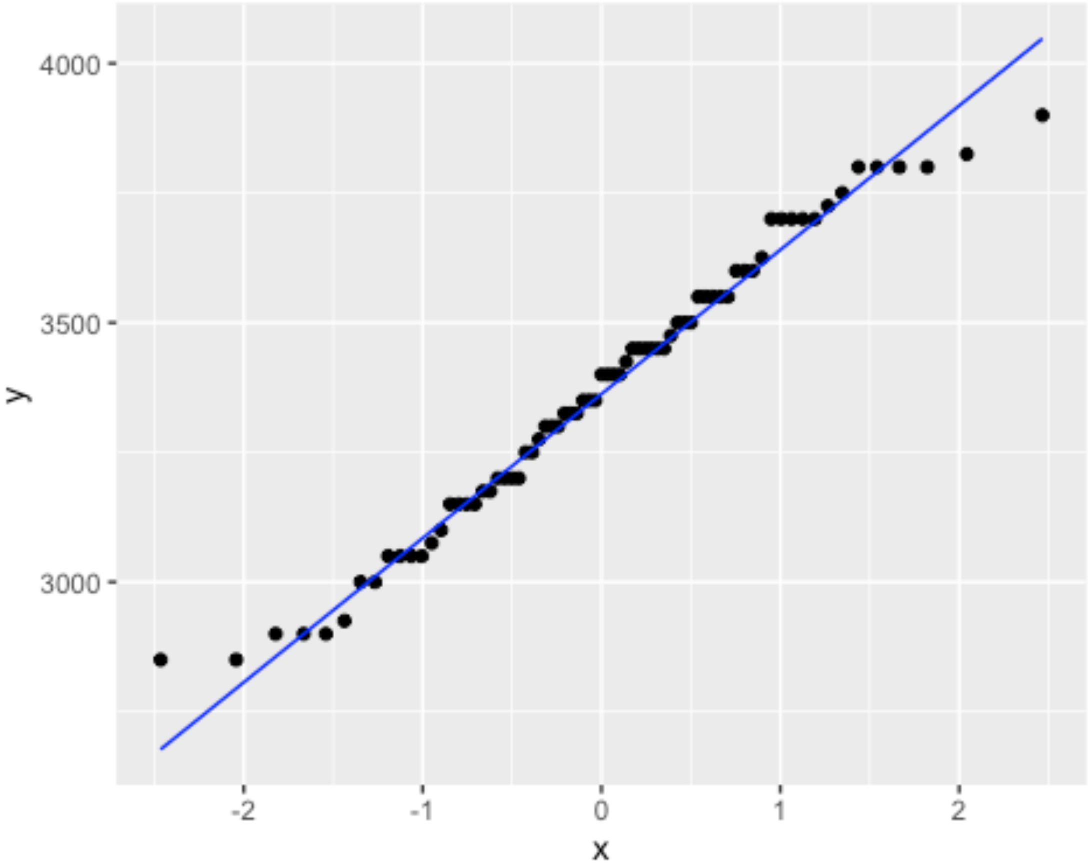
```
adelieM <- adelie %>% filter(sex == 'MALE')
adelieM
```

```
## # A tibble: 73 × 14
##    study_name sample_nu…¹ species region island stage indiv…²
clutc…³ date_egg
##    <chr>            <dbl> <chr>   <chr>  <chr>  <chr> <chr>   <chr>
<date>
##  1 PAL0708              1 Adelie… Anvers Torge… Adul… N1A1    Yes
2007-11-11
##  2 PAL0708              6 Adelie… Anvers Torge… Adul… N3A2    Yes
2007-11-16
##  3 PAL0708              8 Adelie… Anvers Torge… Adul… N4A2    No
2007-11-15
##  4 PAL0708             14 Adelie… Anvers Torge… Adul… N7A2    Yes
```

```
2007-11-15
##  5 PAL0708                 15 Adelie… Anvers Torge… Adul… N8A1     Yes
2007-11-16
##  6 PAL0708                 18 Adelie… Anvers Torge… Adul… N9A2     Yes
2007-11-12
##  7 PAL0708                 20 Adelie… Anvers Torge… Adul… N10A2    Yes
2007-11-16
##  8 PAL0708                 22 Adelie… Anvers Biscoe Adul… N11A2    Yes
2007-11-12
##  9 PAL0708                 24 Adelie… Anvers Biscoe Adul… N12A2    Yes
2007-11-12
## 10 PAL0708                 25 Adelie… Anvers Biscoe Adul… N13A1    Yes
2007-11-10
## # … with 63 more rows, 5 more variables: culmen_length_mm <dbl>,
## #   culmen_depth_mm <dbl>, flipper_length_mm <dbl>, body_mass_g
<dbl>,
## #   sex <chr>, and abbreviated variable names ¹sample_number, ²
individual_id,
## #   ³clutch_completion
```

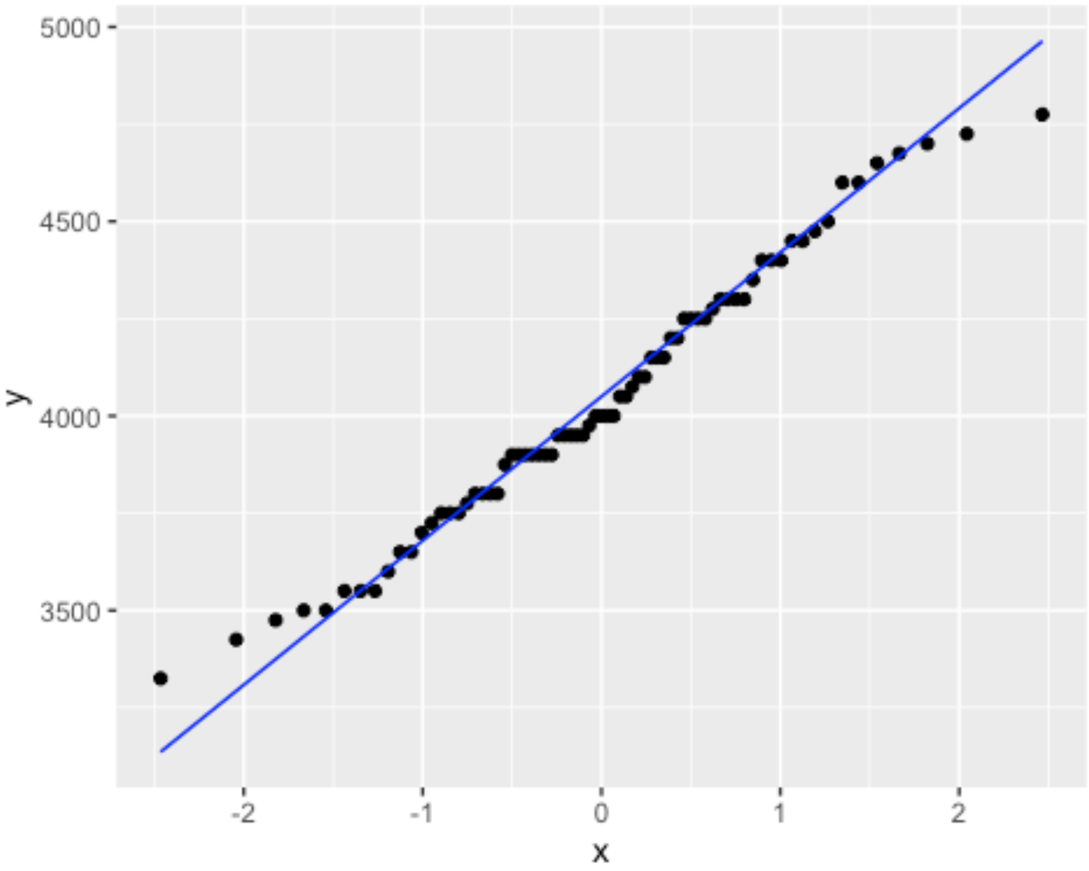
Testing normality in females using a qqplot.

```
ggplot(adelieF, aes(sample = body_mass_g)) +
  geom_qq() +
  geom_qq_line(colour = "blue")
```



Testing normality in males using a qqplot

```
ggplot(adelieM, aes(sample = body_mass_g)) +
  geom_qq() +
  geom_qq_line(colour = "blue")
```



I would conclude that both data are normally distrubuted; at the tail ends there are some slight deviations however t-tests are robust to some deviation from normality.

To check that the variance between the populations are equal I use a Levene test To do this the ‘car’ package is required. H0: The two varainces are equal. H1: The two variances are not equal.

```
leveneTest(data = adelie, body_mass_g ~ sex, centre = mean)

## Levene's Test for Homogeneity of Variance (center = median: mean)
##         Df F value  Pr(>F)
## group    1  3.8664 0.05118 .
##        144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the p-value is greater than 0.05 the two variances are not significantly different from each other and so we do not reject the null.

Now that my assumptions have been tested and met I can carry out my two-sample t-test.

```
t.test(data = adelie, body_mass_g ~ sex, var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  body_mass_g by sex
## t = -13.126, df = 144, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group
FEMALE and group MALE is not equal to 0
## 95 percent confidence interval:
##  -776.2484 -573.0666
## sample estimates:
## mean in group FEMALE   mean in group MALE
##             3368.836             4043.493
```

The p-value produced is significantly lower than 0.05 meaning we can be confident that there is a difference between the two means. Therefore we reject the null hypothesis and conclude there is likely a difference between mean male and mean female body weight in Adelie penguins.

Creating the figure

```
body_mass_hist <- ggplot(data = adelie, aes(x = body_mass_g, fill =
sex)) +
  geom_histogram(position = "Identity", bins = 30, alpha = 0.6) +
#position = "Identity" is what produces the overlapping histograms
  scale_fill_manual(values = c("#06238b", "#f4c1dc")) + #choosing the
colours for sex
  labs(x = "Body mass (g)",
       y = "Number of penguins", fill = "Sex",
       title = "Histogram comparing the body mass of male and female
Adelie penguins",
       caption = "Fig.1") +
  theme_light()

body_mass_hist
```

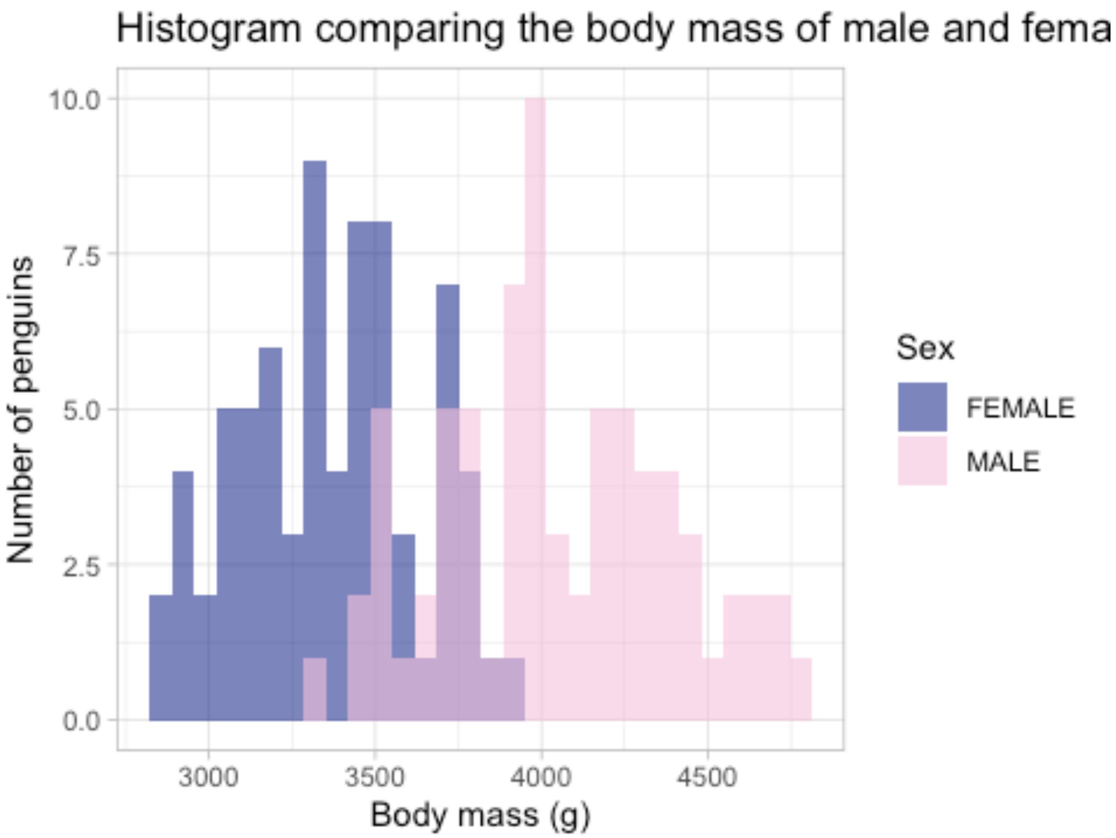


Fig.1

I decided to produce two overlapping histograms so that you can clearly see where the two peaks are as well as seeing overlap between the two populations. From the graph can see that female body mass peaks around 3300g while male peaks around 4000g. These two visually distinct peaks give us greater confidence that our t-test is correct. Furthermore, I have chosen two colours from the opposite ends of the batlow scale in order for the two different populations to be visible to people across the colourblind spectrumn.

Saving the figure

As a png.

```
body_mass_hist

agg_png("Penguin Project/figures/highres.png",
       width = 800, height = 600, units = "px",
        res=300,
       scaling = 0.4
        )
body_mass_hist
```

To save the image to your directory you need to change the bit between “” to your chosen filepath.

As a vector.

```
body_mass_hist

svglite("Penguin Project/figures/8*6.svg",
        width = 8, height = 6)
body_mass_hist
```