Week 3 - Simple and multiple linear regression

PSC 103B

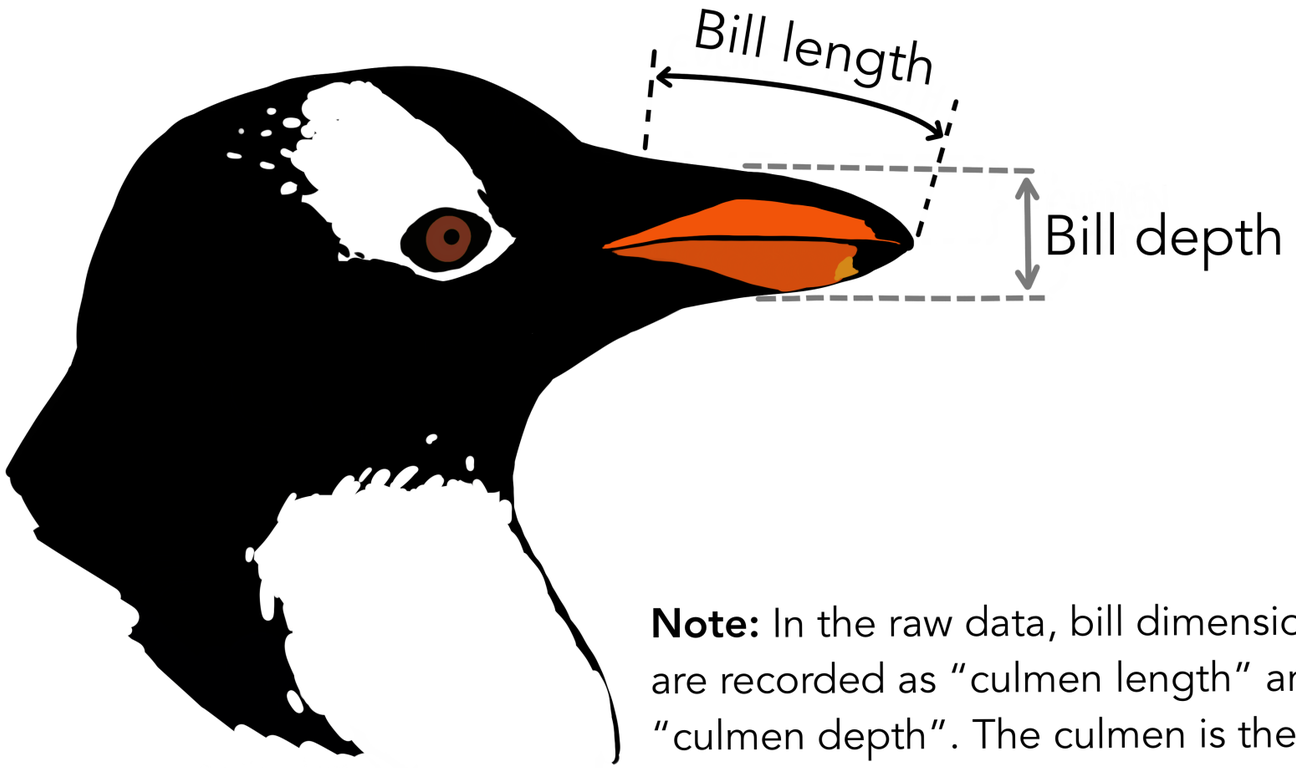
Marwin Carmo

We can access this dataset by installing the `palmerspenguins` package.

```r
install.packages("palmerpenguins")
library(palmerpenguins)
```

```r
dplyr::glimpse(penguins)
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel…
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse…
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, …
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, …
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186…
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, …
$ sex               <fct> male, female, female, NA, female, male, female, male…
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007…
```



Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

- Outcome variable: `bill_length_mm`
- Not all penguins gave data on bill length and there are some missing values.
- The `complete.cases()` function gives the row numbers where there is non-missing values on the variable you give it.

```
1 penguins_subset <- penguins[complete.cases(penguins$bill_length_mm),]
```

- Suppose we were interested in whether male penguins or female penguins had different bill lengths.
- We suspected that male penguins have longer bill lengths than female penguins.
- Let's look at both means

```r
# Average bill lenght for males
mean(penguins_subset$bill_length_mm[penguins_subset$sex == "male"],
     na.rm = TRUE)
```

```
[1] 45.85476
```

```r
# Average bill lenght for females
mean(penguins_subset$bill_length_mm[penguins_subset$sex == "female"],
     na.rm = TRUE)
```

```
[1] 42.09697
```

- Another way to do this is to use the `tapply()` function.
- `tapply(variable, group, function, extra arguments for the function)`

```
tapply(penguins_subset$bill_length_mm,
       penguins_subset$sex, mean, na.rm = TRUE)
```

```
  female     male
42.09697 45.85476
```

- Is the numerical difference of ~4 mm actually significant?
- $H_0 : \mu_{female} = \mu_{male}$, or the average bill length of females is the **same** as the average bill length of males.
- $H_1 : \mu_{female} < \mu_{male}$, or the average bill length of females is **less** than that of males.
- The t-test is trying to see whether the difference you observed between the groups is large given the expected variability of that difference across samples.

- Our hypothesis was that females have shorter bill lengths than males.
- R views the females as Group 1 and males as Group 2 (because female is alphabetically before male). We need to decide our alternative with Group 1 compared to Group 2.

```r
levels(penguins_subset$sex)
```

```
[1] "female" "male"
```

- Using the following syntax, replace the placeholders with the names of the variables we're interested in:

```
t.test(dependent_variable ~ group_variable, data = dataset,
       alternative = "???")
```

Tip

The argument `alternative` specifies the alternative hypothesis and can take any of these three values: `"two.sided"`, `"less"`, or `"greater"`. Think about our hypothesis to choose one of the alternatives.

```r
t.test(bill_length_mm ~ sex, data = penguins_subset, alternative = "less")
```

```
    Welch Two Sample t-test

data:  bill_length_mm by sex
t = -6.6725, df = 329.29, p-value = 5.332e-11
alternative hypothesis: true difference in means between group female and group male is less than 0
95 percent confidence interval:
     -Inf -2.82883
sample estimates:
mean in group female   mean in group male
            42.09697             45.85476
```

> The Welch Two Sample t-test found that female penguins (M = 42.1, SD = 4.90) have, on average, shorter bill lenghts than male penguins (M = 45.9, SD = 5.37), t(329.29) = -6.67, $p < .001$.

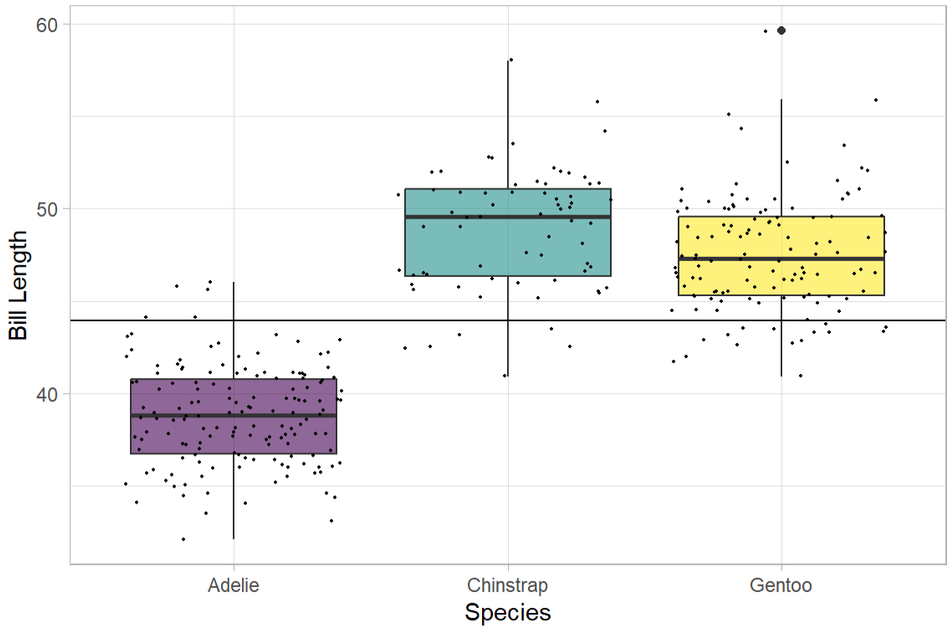
- Notice that R gives us the Welch's t-test by default.
- It is used when the number of samples in each group is different, and the variance of the two data sets is also different. Usually, that is a safe assumption.
- To assume equal variances, set the argument `var.equal = TRUE`.

- What should we do if we have more than two groups we are interested in comparing?
- Our question is the same as a t-test - are there differences in the average score across the groups? We can't use a t-test because a t-test is **limited to 2 groups**.
- Running multiple t-tests increases our Type 1 error rate - the probability of finding a significant difference when there is none.
- One-way ANOVA lets us examine whether **multiple groups** differ in average scores.

- Let us apply this to the example of whether bill length differs across the different species of penguins.
- $H_0 : \mu_{Adelie} = \mu_{Chinstrap} = \mu_{Gentoo}$ or in other words, the average bill length is the same for all 3 species of penguins. The alternative hypothesis is:
- H_A: At least one of the means is different, or H_0 is not true.

- The alternative hypothesis is a bit more complicated. It can be that:
  - $\mu_{Adelie} \neq \mu_{Chinstrap} = \mu_{Gentoo}$ or,
  - $\mu_{Adelie} = \mu_{Chinstrap} \neq \mu_{Gentoo}$ or,
  - $\mu_{Adelie} \neq \mu_{Chinstrap} \neq \mu_{Gentoo}$,
  - etc.
- To capture all those possibilities, we need an alternative hypothesis that is a bit more vague (H_0 is not true, or at least one mean is different).

- What can you say by looking at this plot?



- On face value, the means of the groups are different, but there is also a lot of variability within each group around that group mean.
- ANOVA quantifies how much variation we see between groups is due to actual, significant group differences and how much is just due to sampling variation.

- If H_0 were true, we would expect the amount of variance due to individual differences to be larger than the amount of variance that is due to group differences
- If H_0 **were not true**, and there were actual group differences, then we expect the variation between groups to be larger than the residual variance (which is the variance due to sampling error/non-group differences).

We can compare between and within groups variability with the F-ratio:

$$
\begin{aligned}
F &= \frac{\text{Between groups variability}}{\text{Within groups variability}} \\
&= \frac{\text{Group effects} + \text{Ind diffs} + \text{Error}}{\text{Ind diffs} + \text{Error}}
\end{aligned}
$$

If the group effect is zero, F-ratio will be close to one.

$$F = \frac{MS_{Between}}{MS_{Within}} = \frac{\frac{SS_{Between}}{df_{Between}}}{\frac{SS_{Within}}{df_{Within}}}$$

Let’s walk through an example.

- If H_0 were true, then our best guess for the score of a new penguin would be the **grand mean** (or the mean of the entire sample), since group membership wouldn't tell us anything useful.
- We can compare each group's mean to this grand mean.
- If the group means are all similar, then the variance will be small.
- If the group means are different, then the variance will be large.

- First, let us calculate the mean of each group.
- How would you do it in R?

```r
tapply(penguins_subset$bill_length_mm, penguins_subset$species, mean,
       na.rm = TRUE)
```

```
   Adelie Chinstrap    Gentoo
 38.79139  48.83382  47.50488
```

We can make a dataframe that contains the group means and the grand means, to make it easier to calculate the $SS_{Between}$.

```r
penguin_means <- data.frame(
  GroupMean = tapply(penguins_subset$bill_length_mm,
                     penguins_subset$species, mean,
                     na.rm = TRUE),
  GrandMean = mean(penguins_subset$bill_length_mm,
                                 na.rm = TRUE))

penguin_means
```

```
           GroupMean GrandMean
Adelie      38.79139  43.92193
Chinstrap   48.83382  43.92193
Gentoo      47.50488  43.92193
```

Exercise

Create a new column in `penguin_means` named `mean_deviations` containing the difference between each group mean and the grand mean.

```
1 penguin_means$mean_deviations <- penguin_means$GroupMean - penguin_means$GrandMean
2
3 penguin_means
```

```
           GroupMean GrandMean mean_deviations
Adelie      38.79139  43.92193       -5.130539
Chinstrap   48.83382  43.92193        4.911894
Gentoo      47.50488  43.92193        3.582948
```

- The mean deviations do not tell us much yet. We are first trying to estimate the $SS_{Between}$
- If you recall from class, this tell us the variability of group means around the grand mean scaled by group sample size:

$$
\begin{aligned}
SS_{Between} &= \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2 \\
&= n_1(\bar{X}_1 - \bar{X})^2 + n_2(\bar{X}_2 - \bar{X})^2 + n_3(\bar{X}_3 - \bar{X})^2
\end{aligned}
$$

- So, before we can square the deviations and add them, we need to multiply it by the corresponding group sample size.
- To get the size of each group, we can use the `table()` function.

```
table(penguins_subset$species)
```

```
   Adelie Chinstrap    Gentoo 
      151        68       123 
```

```
penguin_means$SampleSize <- table(penguins_subset$species)
```

```
SSB <- sum(penguin_means$mean_deviations^2 * penguin_means$SampleSize)
SSB
```

```
[1] 7194.317
```

- Now we need to calculate SS_{Within}, or the residual variance, the difference from each individual's score to their group's mean.
- To calculate this, we need to get each penguin's observation and each penguin's group mean in the same dataframe.
- One way to do this is create smaller vectors for each species.

```r
penguins_adelie <- penguins_subset$bill_length_mm[penguins_subset$species == "Adelie"]

penguins_chinstrap <- penguins_subset$bill_length_mm[penguins_subset$species == "Chinstrap"]

penguins_gentoo <- penguins_subset$bill_length_mm[penguins_subset$species == "Gentoo"]
```

Exercise

Calculate the sum of squared deviations from the group mean separately for each group, and save them in three different objects: `penguins_adelie_dev`, `penguins_chinstrap_dev`, and `penguins_gentoo_dev`.

Tip

Note that we have to use `na.rm = TRUE` twice: one to calculate the value of the mean, but also to use the `sum()` function since not all penguins have a bill length, so using `mean()` or `sum()` on something with a `NA` value leads to a `NA`.

Here's what you're calculating: $SS = \sum(X_i - \bar{X})$

```r
penguins_adelie_dev <- sum((penguins_adelie -
                            mean(penguins_adelie, na.rm = TRUE))^2, na.rm = TRUE)

penguins_chinstrap_dev <-  sum((penguins_chinstrap -
                                mean(penguins_chinstrap, na.rm = TRUE))^2, na.rm = TRUE)

penguins_gentoo_dev <- sum((penguins_gentoo -
                            mean(penguins_gentoo, na.rm = TRUE))^2, na.rm = TRUE)
```

And now we add up all these deviations to get SSW

```r
SSW <- penguins_adelie_dev + penguins_chinstrap_dev + penguins_gentoo_dev

SSW
```

```
[1] 2969.888
```

- Now that we have SSB and SSW, we need to get the df for each variance.
- The formulas for the df are:

$$df_{between} = k - 1$$
$$df_{within} = N - k - 1$$

- where k is the number of groups and N is the total sample size.

- We know we have 3 groups, but how many penguins?

```r
nrow(penguins_subset)
```

```
[1] 342
```

```r
dfB <- 3 - 1
dfW <- 342 - 3
```

- Now we can calculate Mean Squared Between and Mean Squared Within by dividing each sum of squares by the df.

```r
MSB <- SSB / dfB
MSW <- SSW / dfW
```

- $MS_{Between}$ describes the amount of variance that can be attributed to the differences between groups.
- MS_{Within} describes the amount of variance that can be attributed to chance or sampling error (basically, whatever cannot be described by group differences).
- We compare these 2 to calculate our F-statistic.

```
1 Fstat <-  MSB / MSW
```

- We can get the p-value of F by looking at the F-distribution with degrees of freedom (dfB, dfW).
- In R, this is done using the pf() function.

```
pf(Fstat, df1 = dfB, df2 = dfW, lower.tail = FALSE)
```

```
[1] 2.694614e-91
```

- What can we conclude?
- Since our p-value is less than .05, we would reject H_0, and conclude that at least one of the groups have an average bill length that is not equal to the rest.
- Which species, though, is different?
- The F-test is an omnibus test, so although it can tell us that there are significant group differences in bill length, it does **not** tell us which groups are different.
- We find that with post-hoc tests (next week!)

- R obviously has a way simpler solution to do the anova.
- The `aov()` function: `aov(outcome ~ group, data = dataset)`
- To get meaningful results, we need to wrap the object created by `aov()` around the `summary()` function:

```r
my_anova <- aov(bill_length_mm ~ species, data = penguins_subset)
summary(my_anova)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
species       2   7194    3597   410.6 <2e-16 ***
Residuals   339   2970       9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Does this match what we got before? Is our conclusion the same?

Exercise

Run an One-Way ANOVA using `aov()` to investigate if there is an overall difference in `body_mass_g` between `species` in the `penguins_subset` dataset. What can you conclude?

```r
summary(aov(body_mass_g ~ species, data = penguins_subset))
```

```
             Df    Sum Sq  Mean Sq F value Pr(>F)
species       2 146864214 73432107   343.6 <2e-16 ***
Residuals   339  72443483   213698
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```