

# Lab 4: t-tests and One-Way ANOVA

PSC 103B

Marwin Carmo

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

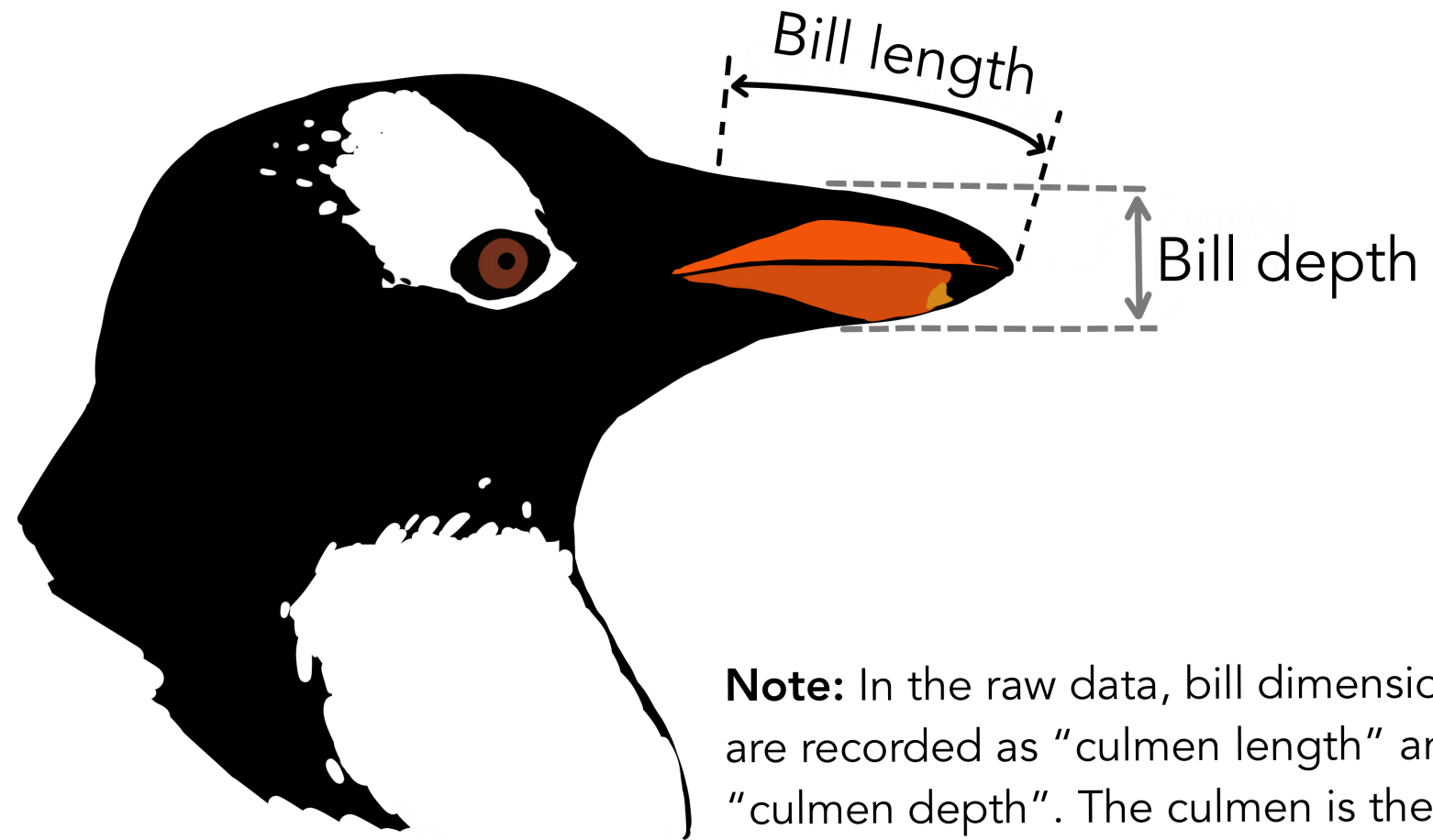
```
1 install.packages("palmerpenguins")
2 library(palmerpenguins)
```

```
1 dplyr::glimpse(penguins)
```

Rows: 344

Columns: 8

```
$ species      <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel...
$ island       <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen...
$ 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, 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

```
1 # Average bill length for males
2 mean(penguins_subset$bill_length_mm[penguins_subset$sex == "male"],
3       na.rm = TRUE)
```

```
[1] 45.85476
```

```
1 # Average bill length for females
2 mean(penguins_subset$bill_length_mm[penguins_subset$sex == "female"],
3       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)`

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

```
female    male  
42.09697 45.85476
```

- Is the numerical difference of ~4 mm actually significant?
- $(H_0: \mu_{\text{female}} = \mu_{\text{male}})$ , or the average bill length of females is the **same** as the average bill length of males.
- $(H_1: \mu_{\text{female}} < \mu_{\text{male}})$ , 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.

```
1 levels(penguins_subset$sex)
```

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



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

```
1 t.test(dependent_variable ~ group_variable, data = dataset,  
2         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.

```
1 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 lengths 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**.



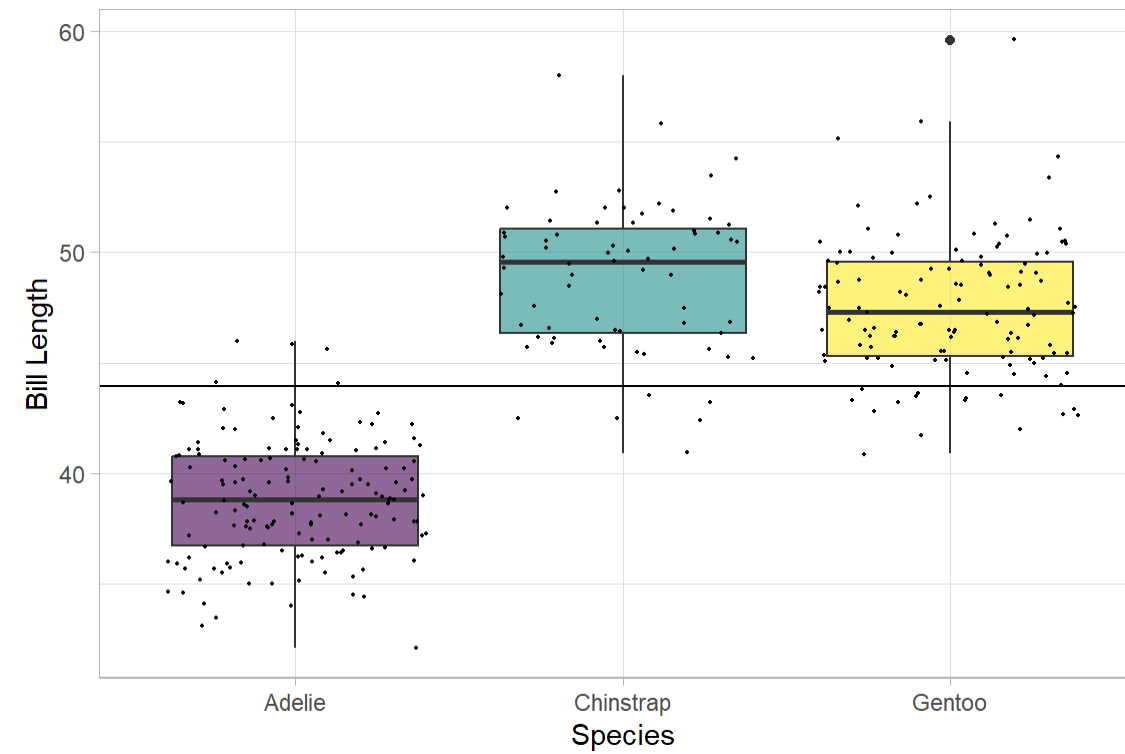
- What should we do if we have more than two groups we are interested in comparing?
- Our question is the same as a t-test - are there differences in the average score across the groups? We can't use a t-test because a t-test is **limited to 2 groups**.
- Running multiple t-tests increases our Type 1 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.
- $(H_0: \mu_{\text{Adelie}} = \mu_{\text{Chinstrap}} = \mu_{\text{Gentoo}})$  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:
  - $(\mu_{\text{Adelie}} \neq \mu_{\text{Chinstrap}} = \mu_{\text{Gentoo}})$  or,
  - $(\mu_{\text{Adelie}} = \mu_{\text{Chinstrap}} \neq \mu_{\text{Gentoo}})$  or,
  - $(\mu_{\text{Adelie}} \neq \mu_{\text{Chinstrap}} \neq \mu_{\text{Gentoo}})$ ,
  - 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:

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

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

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

```
1 tapply(penguins_subset$bill_length_mm, penguins_subset$species, mean,  
2        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  $(SS_{\text{Between}})$ .

```
1 penguin_means <- data.frame(  
2   GroupMean = tapply(penguins_subset$bill_length_mm,  
3                       penguins_subset$species, mean,  
4                       na.rm = TRUE),  
5   GrandMean = mean(penguins_subset$bill_length_mm,  
6                     na.rm = TRUE) )  
7  
8 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  $SS_{\text{Between}}$
- If you recall from class, this tell us the variability of group means around the grand mean scaled by group sample size:

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

- 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.

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

Adelie	Chinstrap	Gentoo
151	68	123

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

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

```
[1] 7194.317
```

- Now we need to calculate  $(SS_{\text{Within}})$ , 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.

```
1 penguins_adelie <- penguins_subset$bill_length_mm[penguins_subset$species == "Adelie"]  
2  
3 penguins_chinstrap <- penguins_subset$bill_length_mm[penguins_subset$species == "Chinstrap"]  
4  
5 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:  $(SS = \sum(X_i - \bar{X})^2)$

```
1 penguins_adelie_dev <- sum((penguins_adelie -
2                             mean(penguins_adelie, na.rm = TRUE))^2, na.rm = TRUE)
3
4 penguins_chinstrap_dev <- sum((penguins_chinstrap -
5                                 mean(penguins_chinstrap, na.rm = TRUE))^2, na.rm = TRUE)
6
7 penguins_gentoo_dev <- sum((penguins_gentoo -
8                             mean(penguins_gentoo, na.rm = TRUE))^2, na.rm = TRUE)
```

And now we add up all these deviations to get **SSW**

```
1 SSW <- penguins_adelie_dev + penguins_chinstrap_dev + penguins_gentoo_dev
2
3 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:

$$(df_{\text{between}} = k - 1)$$

$$(df_{\text{within}} = N - k - 1)$$

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

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

```
1 nrow(penguins_subset)
```

```
[1] 342
```

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

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

```
1 MSB <- SSB / dfB  
2 MSW <- SSW / dfW
```



- $MS_{\text{Between}}$  describes the amount of variance that can be attributed to the differences between groups.
- $MS_{\text{Within}}$  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.

```
1 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:

```
1 my_anova <- aov(bill_length_mm ~ species, data = penguins_subset)
2 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?

```
1 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
```