Lab 04 - T-tests and ANOVA

PSC-103B

Marwin Carmo

2025-02-06

The data

The dataset we will be using today is the Palmers Penguins dataset. We can access this dataset by installing the `palmerspenguins` package.

First, install the `palmerpenguins` package:

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

Now, load all the required packages:

```
library(palmerpenguins)
library(ggplot2)
library(viridis)
```

```
## Loading required package: viridisLite
```

The dataset is now available as the object `penguins`

```
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 2 more variables: sex <fct>, year <int>
```

```
names(penguins)
```

```
## [1] "species"           "island"            "bill_length_mm"
## [4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
## [7] "sex"               "year"
```

The penguins dataset contains information on 344 penguins, collected by the Palmer Station in Antartica as part of the Long Term Ecological Research Network (LTER). There are 3 different species of penguins in this dataset - Adelie, Chinstrap, and Gentoo - and they were studied at 3 different islands in the Palmer Archipelago in Antartica. Researchers collected information on their species, as well as body characteristics (such as weight, bill length, and flipper length).

Review of Two-Sample t-test

For all our examples today, our outcome variable will be bill length. However, not all penguins gave data on bill length and there are some missing values. To make our lives easier, let's subset our data to just the rows where penguins gave us information. The `complete.cases()` function gives the row numbers where there is non-missing values on the variable you give it.

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

Suppose we were interested in whether male penguins or female penguins had different bill lengths. In particular, we suspected that male penguins have longer bill lengths than female penguins, regardless of species. First, let's look at the descriptive statistics for this variable - what is the average bill length for each group?

How could we do that? We need to use the mean() function, but simply typing `mean(penguins_subset$bill_length_mm, na.rm = TRUE)` would not work because that gives you the average bill length collapsing across sex!

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

```
## [1] 43.92193
```

Instead, we need to subset our data into the rows pertaining to male penguins, and the rows pertaining to female penguins. Remember how to subset to particular rows? Use a logical statement, like `penguins$sex == "male"`.

```
mean(penguins_subset$bill_length_mm[penguins_subset$sex == "male"],
     na.rm = TRUE)
```

```
## [1] 45.85476
```

This is saying give me the average bill length, but only using the rows where the sex of the penguin is male. And we can switch it out to get the mean bill length for female penguins.

```
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 which takes the arguments: `tapply(X = target variable, INDEX = grouping variable, FUN = function, extra arguments for the function)` so that you can apply a function to a particular variable separately for each group. Here, the variable we want to apply a function (mean) to is bill length, and the group is the sex.

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

```
##   female     male
## 42.09697 45.85476
```

So there is a numerical difference that suggests male penguins have, on average, longer bills. But is the difference of ~4 mm actually significant? That's why we conduct our t-test! For this test, what is our null and alternative hypotheses?

$H_0 : \mu_{female} = \mu_{male}$, 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.

I am not going to go in-depth on how to conduct the t-test, since this is review from last quarter. But I just want to remind you what the logic of the t-test is: 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.

Remember, our hypothesis was that females have shorter bill lengths than males; since `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.

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.

```
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
```

Is our p-value significant? What does that lead us to conclude?

Yes, our ***p*-value is significant**, since our p-value is less than .05. Therefore, the difference we observed between the sexes is unlikely to be due to chance, and we can reject our null hypothesis. This means we have evidence that the average bill length of females is shorter than that of males!

Here's how you could write-up your results:

> 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.

Note that `R` uses **Welch's t-test** by default. This test is appropriate when **the number of samples in each group is different**, and the **variance of the two data sets is also different**. With psychological data, it is usually safe to assume that it is the case. But if you want to assume equal variances, set the argument `var.equal = TRUE`.

One-Way ANOVA

We often have more than 2 groups than we are interested in comparing - e.g., perhaps you have 2 different treatments and a control group, or year in school. Our question is the same as a t-test - are there differences in the average score across the groups - but we can't use a t-test, because a t-test is limited to 2 groups.

We also can't just conduct a t-test a bunch of times for every pair of groups, because that increases our Type 1 error rate - in other words, our probability of finding a significant difference when there is none increases when we conduct multiple tests.

This is why we do the one-way ANOVA! One-way ANOVA let's us examine whether multiple groups differ in their average scores.

Let us apply this to the example of whether bill length differs across the different species of penguins.

The null and alternative hypotheses of a one-way ANOVA are:

$H_0 : \mu_{Adelie} = \mu_{Chinstrap} = \mu_{Gentoo}$ 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.

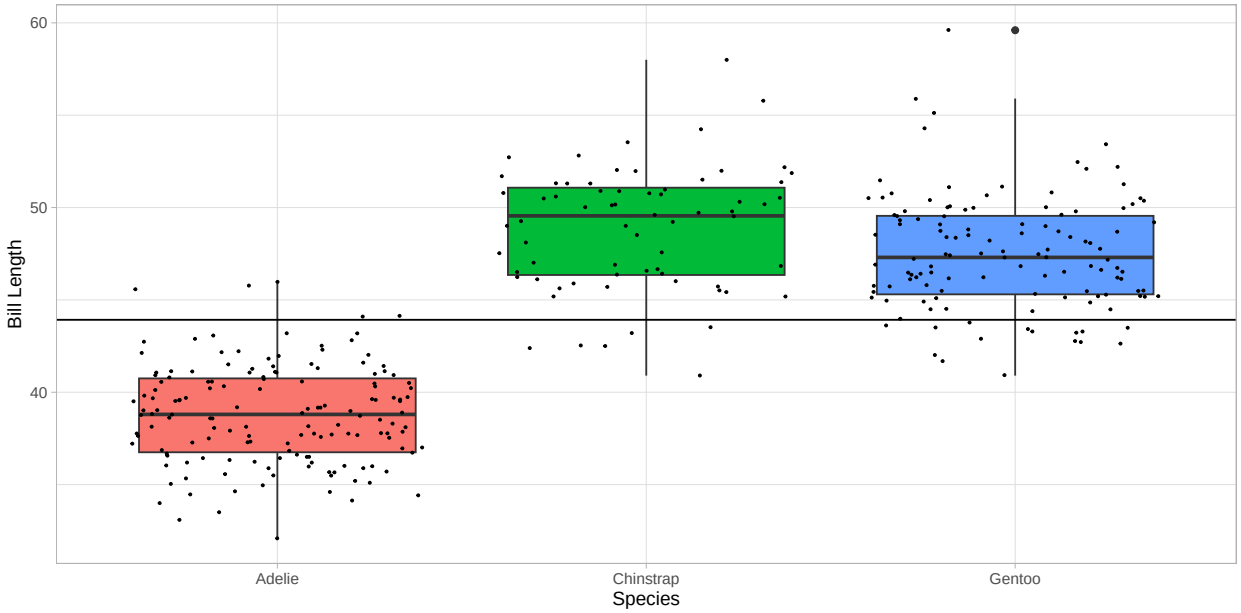
Notice that the null hypothesis is pretty straightforward, and is just a simple extension of the null hypothesis in the t-test: there are no differences between our groups!

However, the alternative hypothesis is a bit more complicated, because there is not just one way that the null hypothesis could be not true. We could have the mean of Adelie penguins is less than Chinstrap and Gentoo, but Chinstrap and Gentoo have equal means or the mean of Adelie is less than Chinstrap which is less than Gentoo, etc. Therefore, we need an alternative hypothesis that is a bit more vague (H0 is not true, or at least one mean is different) in order to capture all those possibilities.

What is the logic of a one-way ANOVA? Well, remember that the logic of a t-test is whether the observed difference between the groups is larger than we would expect based on the variability of that difference across multiple samples. The logic of ANOVA is similar, but we just need to change how we quantify these things because there are more than 2 groups.

First, let's visualize the data.

```
ggplot(data = penguins_subset, aes(x = species, y = bill_length_mm, fill = species))+
  geom_boxplot() +
  geom_jitter(color = "black", size = 0.4) +
  labs(x = "Species", y = "Bill Length") +
  theme_light() +
  theme(legend.position = "none") +
  geom_hline(yintercept = mean(penguins_subset$bill_length_mm, na.rm = TRUE))
```



Looking at this plot, we can see that yes, on face value, the means of the groups are different from each other. But this could be entirely due to sampling error, individual differences, or measurement error, because there is also a lot of variability within each group around that group mean. With an ANOVA, we want to see how much of the variation that we see between groups is due to actual, significant group differences and how much of it is just due to sampling variation.

If the null hypothesis were true, and there were no group differences, then we would expect the amount of total variation in our sample that is due to sampling error/people being different from each other to be more than the same as the amount of variance that is due to group differences. There would be a lot of noise or variation within each group, so it could be possible that in a new sample, perhaps the means weren't so different.

However, if the null hypothesis 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). This is because we expect people to be more clustered around their group mean and less spread out if there were actual group differences, so in a new sample, the group means are still likely to be different.

This is a lot to take in, so let's walk through an example to try and make things clearer!

Between-Group Variance

If the null hypothesis were true and there were no meaningful group differences, 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. Therefore, to calculate the between-group variance and see how much variation in our outcome is due to group differences, we compare each group's mean to this grand mean. If there are actual group differences, we would expect this difference to be large!

First, let us calculate the mean of each group. We could do the same subsetting approach as before, but that's rather cumbersome (especially when there are more groups). Thankfully, we also learned how to calculate these means using `tapply()`.

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

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

To calculate the between-group variance, we can make a dataframe that contains the group means and the grand means, to make it easier to calculate the $SS_{Between}$.

```
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
```

Now we can find the difference between each group mean and the grand mean.

```
penguin_means$mean_deviations <- penguin_means$GroupMean - penguin_means$GrandMean

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 deviations from the grand mean are now in a separate column. Now we can square each deviation, but before we can add these squared deviations together, we need to multiply it by the corresponding group sample size. To get the size of each group, we can use the `table()` function.

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

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

Let's add this to our table.

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

Now we can square the deviations, multiply them by the group size, and add them all up. This is the sum of squares between.

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

```
## [1] 7194.317
```

Within-Group Variance

Now that we have our $SS_{Between}$, we also need to calculate our residual variance, which is represented by SS_{Within}. This is quantified by the difference from each individual's score to their group's mean. It tells us how clustered people are around their group mean - remember, we want to compare this to the size of the group differences represented by `SSB`. If this SS_{Within} is bigger than $SS_{Between}$, and people are super spread out around their group mean, it's unlikely that the differences we see are representative of actual group differences.

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"]
```

Now we can calculate the sum of squared deviations from the group mean separately for each group. 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`.

```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
```

ANOVA Calculation

Once we have `SSB` and `SSW`, it's pretty straightforward to calculate the test statistic. First, we need to get the df for each variance. So that we can divide the sum of squares by the df. This is like when we divide the sum of squared deviations by n-1 to get the variance - we need something that is more representative of an average.

The formulas for the df are:

$df_{between} = k - 1$

$df_{within} = N - k$

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

In this case, we have 3 groups and how many penguins?

```
nrow(penguins_subset)
```

```
## [1] 342
```

```
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.

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

$MS_{Between}$ describes the amount of variance that can be attributed to the differences between groups, and 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. If H_0 was true and there were no group differences, what do we expect our F ratio to be? We expect it to be close to 1, because if there were no group differences, then $MS_{Between}$ would be similar to or smaller than MS_{Within} However, if there were significant group differences (H0 was not true), then $MS_{Between}$ would be larger than MS_{Within} and our F ratio would be greater than 1.

So our F-statistic is just the ratio of those 2 variances, and we can get our p-value by looking at the F-distribution with degrees of freedom (dfB, dfW). In `R`, this is done using the pf() function, which is like the `pnorm()` function we used before but for the F-distribution!

When we calculate the p-value, we always look at the upper tail of the distribution only, and then compare that value to our alpha level (which is .05 in this case).

```
Fstat <-  MSB / MSW

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

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

With this p-value do we reject or fail to reject our null? What does this tell us about our groups? 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?

Unfortunately, the F-test is what is referred to as 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. That is what we need post-hoc tests for, which will be covered next week.

Well, after all that work, it might be obvious that there is a much easier way to do this in `R`! To do this in `R`, we use the `aov()` function. The `aov()` function is like the `t.test()` function: `aov(outcome ~ group,`

```
data = dataset)
```

```
aov(bill_length_mm ~ species, data = penguins_subset)
```

```
## Call:
##    aov(formula = bill_length_mm ~ species, data = penguins_subset)
##
## Terms:
##                   species Residuals
## Sum of Squares  7194.317  2969.888
## Deg. of Freedom        2       339
##
## Residual standard error: 2.959853
## Estimated effects may be unbalanced
```

Well this is not helpful! All we get are the Sum of Squares and df - what about the F stat and p-value?

Just like in regression, you need to use the `summary()` function to get more information.

```
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?

Now, 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?

```
mass_anova <- aov(body_mass_g ~ species, data = penguins_subset)
summary(mass_anova)
```

```
##              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
```