Lab 05 - Post-hoc Tests & Factorial ANOVA

PSC-103B

Marwin Carmo

2025-02-13

Since post-hoc tests are the follow-up to a one-way ANOVA we will be using the same Palmer Penguins dataset as last week.

However, I made it available as a download from Canvas, because I adjusted it slightly. In addition to only including penguins who provided information on their bill length, I also randomly chose 68 penguins from each species, because the post-hoc test we are doing requires equal group sizes[1].

Read in the data

```
# you can load in the data set directly through this link
penguins_data <- read.csv("https://shorturl.at/gfDYk")
```

Post-Hoc Tests

Recall the one-way ANOVA we conducted last week we were interested in whether the different species of penguins had, on average, equal bill lengths, or whether there was some difference in the average bill length among the 3 species.

We conducted our ANOVA using the following code:

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

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## species       2   4436  2217.9   241.8 <2e-16 ***
## Residuals   201   1844     9.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

and our p-value was less than .05, indicating that we can reject H_0. However, since the F-test / ANOVA is an omnibus test, a significant p-value only tells us that there is some difference in the means – but it does not tell us which means are different or how.

So, once a researcher has a significant ANOVA result, they would be interested in learning *where* those differences are. This would involve the use of post-hoc tests, which are significance tests a researcher can do *after* a significant ANOVA that can help shed light on which group means are significantly different.

Unlike conducting multiple t-tests at the .05 level, these post-hoc tests are meant to correct for the many comparisons you are doing, so that your chances of a Type 1 error (saying 2 groups have different means when they actually have the same means) stays at .05.

[1]Note that this only applies when conducting the test by hand – the function in R that runs the post-hoc test applies a correction for unequal group sizes as long as the differences aren't too extreme!

There are multiple post-hoc tests available, and they differ in how conservative they are (e.g., some are more likely to find a significant difference than others). The one we are focusing on today, and the one you learned in class, is Tukey's Honest Significant Difference (HSD) Test.

The way that this test works is it computes a value called a "Critical Difference", and so any group differences that are larger than the Critical Difference (CD) in absolute value are significant.

We will go over doing this by hand first, before showing `R`'s function for it.

The formula for the CD is:

$$CD = q \times \sqrt{\frac{MS_w}{n}}$$

where n is the group size within each group, and MS_w is the Mean Squared Within you can get from your ANOVA table.

For n, we know the group size is 68 because I told you that at the beginning. But if you didn't know, you could always check using which function?

```
table(penguins_data$species)
```

```
##
##    Adelie Chinstrap    Gentoo
##        68        68        68
```

```
n <- 68
```

Then, MS_w can be found from our ANOVA table. How do I access the results of the ANOVA again?

```
summary(my_anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## species       2   4436  2217.9   241.8 <2e-16 ***
## Residuals   201   1844     9.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MSW <- 9.2
```

Finally, the last value we need is q, which is a quantile value from a particular distribution called the studentized range distribution. You can look this up in a table, but with `R`, you can also use the built in function called `qtukey()`.

`qtukey()` takes 3 main arguments:

- `p`, which is the probability to the *left* of the critical value,
- `nmeans`, which is the number of groups we're comparing and,
- `df`, which is just df within

Recall, if our test is conducted at the .05 level, then that means the probability *above* the value is .05, so the probability *below* the critical value is .95. Therefore, `p` = .95. You could also set `p` = .05, but that means you also need to set `lower.tail = FALSE` so that `R` knows you're looking at upper-tail probabilities.

Let's use this function to get our quantile value:

```
qtukey(p = .95,
       nmeans = 3, # because we have 3 species
       df = 339) # where could we find this value?
```

```
## [1] 3.329136
```

So our critical value is 3.33. Let's save this for future use:

```
qvalue <- qtukey(p = .95,  nmeans = 3, df = 339)
```

Now, we can go ahead and calculate our CD.

```
cd <- qvalue * sqrt(MSW / n)
```

Now, all we need to do is compare the absolute mean difference between each pair of groups to this critical difference!

We can calculate these mean differences by hand, by getting the mean for each group. How can we get all the groups' means?

```
penguin_means <- tapply(penguins_data$bill_length_mm,
                        penguins_data$species,
                        mean, na.rm = TRUE)
penguin_means
```

```
##    Adelie Chinstrap    Gentoo
##  38.38529  48.83382  47.60588
```

This `tapply()` function gives a vector of the means for each group. So we can subset this vector to get the mean of each group.

```
adelie_mean <- penguin_means[1]
chinstrap_mean <- penguin_means[2]
gentoo_mean <- penguin_means[3]
```

Now we can get the differences between the means - there are 3 groups, so there are 3 possible differences (Adelie - Chinstrap, Adelie - Gentoo, and Chinstrap - Gentoo).

```
adelie_mean - chinstrap_mean
```

```
##    Adelie
## -10.44853
```

```
adelie_mean - gentoo_mean
```

```
##    Adelie
## -9.220588
```

```
chinstrap_mean - gentoo_mean
```

```
## Chinstrap
##  1.227941
```

Recall our CD was 1.21 – given that, which of these differences are significant? All of them are, because all differences are greater than 1.21! Therefore, we can say that Adelie has a significantly lower average bill length than both Chinstrap and Gentoo (the difference in means was negative), and Chinstrap has a significantly longer bill length, on average, than Gentoo (the difference was positive).

This was all pretty manageable, because we only had 3 groups. But imagine if we had more groups! It could get pretty annoying to do this by hand. Luckily, `R` has a built-in function for this, called `TukeyHSD()`, and the only argument you need to give it is the ANOVA object you saved earlier:

```
TukeyHSD(my_anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = bill_length_mm ~ species, data = penguins_data)
```

```
##
## $species
##                          diff       lwr          upr     p adj
## Chinstrap-Adelie 10.448529  9.222071 11.674988263 0.0000000
## Gentoo-Adelie     9.220588  7.994129 10.447047086 0.0000000
## Gentoo-Chinstrap -1.227941 -2.454400 -0.001482326 0.0496478
```

The output of the `TukeyHSD()` function is the difference between the groups (with Group1 - Group2), the lower and upper limits of a confidence interval, and an adjusted p-value. If the adjusted p-value is less than .05, the difference is significant. Do the results of this function match what we calculated earlier?

Factorial ANOVA

So far, we have been working with what is called One-Way ANOVA where we look at only 1 grouping variable However, ANOVA is super flexible, and can actually handle more than 1 grouping variable. This is called Factorial ANOVA, and let's us examine in one test: is there an effect of Grouping Variable 1 (at least one mean is different in those groups), is there an effect of Grouping Variable 2, and is there an interaction between the 2 variables (so that the effect of Grouping Variable 1 depends on Grouping Variable 2, or vice versa).

Let's assume we were interested in whether body mass differed not only among the 3 penguin species, but also whether it differed among males and females.

If we were to do this as 2 separate ANOVAs, we would have 2 sets of hypotheses.

One for the species of penguin:

- $H_0 : \mu_{Adelie} = \mu_{Chinstrap} = \mu_{Gentoo}$ (there is no difference in average body mass among the species).
- H_A: At least one mean is significantly different from the rest.

One for the islands:

- $H_0 : \mu_{female} = \mu_{male}$ (there is no difference in average body mass among male and female penguins).
- $H_A : \mu_{female} \neq \mu_{male}$ (note that because there are only 2 groups, we can be more specific with our alternative hypothesis).

But since we're conducting this as a factorial ANOVA, we also have a hypothesis for the interaction:

- H_0: There is no interaction between the species of penguin and their sex.
- H_A: There is an interaction between the species of penguin and their sex .

And now, we can test these hypotheses in `R`!

The way you enter the formula for a Factorial ANOVA is very similar to adding an interaction in a regression in `R` where we "multiply" our two predictors to include an interaction between them.

```
factorial_anova <- aov(body_mass_g ~ species * sex,
                       data = penguins_data)
summary(factorial_anova)
```

```
##              Df   Sum Sq  Mean Sq F value   Pr(>F)
## species       2 93491533 46745767 466.899  < 2e-16 ***
## sex           1 19697662 19697662 196.741  < 2e-16 ***
## species:sex   2  1480819   740409   7.395 0.000801 ***
## Residuals   196 19623451   100120
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

Note that this ANOVA table looks slightly different from the one presented in lecture - mainly we do not have an overall row for the "Model", but just the 2 factors and the interaction; however, if you wanted to, you could calculate the Sums of Squares and df for the model yourself by adding the Sums of Squares and df for the individual factors and the interaction (and use those to calculate the Mean Squares).

Which results are significant? What does this tell us?

The main effect of species: the p-value is less than .05, indicating a significant effect of species on bill length. At least one species has an average body mass that is significantly different from another species. How could we go about testing which species differed in their average body mass? A post-hoc test, like Tukey's HSD.

```
TukeyHSD(factorial_anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = body_mass_g ~ species * sex, data = penguins_data)
##
## $species
##                       diff        lwr       upr     p adj
## Chinstrap-Adelie   81.5957  -47.03703  210.2284 0.2940693
## Gentoo-Adelie    1484.3284 1355.22009 1613.4366 0.0000000
## Gentoo-Chinstrap 1402.7327 1274.09993 1531.3654 0.0000000
##
## $sex
##                 diff      lwr      upr p adj
## male-female 620.0387 532.1883 707.8891     0
##
## $`species:sex`
##                                       diff        lwr       upr     p adj
## Chinstrap:female-Adelie:female   139.0809  -73.34912  351.5109 0.4148774
## Gentoo:female-Adelie:female     1289.3750 1069.42303 1509.3270 0.0000000
## Adelie:male-Adelie:female        653.5417  426.71455  880.3688 0.0000000
## Chinstrap:male-Adelie:female     550.8456  338.41558  763.2756 0.0000000
## Gentoo:male-Adelie:female       2119.3074 1911.58473 2327.0301 0.0000000
## Gentoo:female-Chinstrap:female  1150.2941  922.17622 1378.4120 0.0000000
## Adelie:male-Chinstrap:female     514.4608  279.70677  749.2148 0.0000000
## Chinstrap:male-Chinstrap:female  411.7647  190.89051  632.6389 0.0000034
## Gentoo:male-Chinstrap:female    1980.2266 1763.87585 2196.5772 0.0000000
## Adelie:male-Gentoo:female       -635.8333 -877.41522 -394.2514 0.0000000
## Chinstrap:male-Gentoo:female    -738.5294 -966.64730 -510.4115 0.0000000
## Gentoo:male-Gentoo:female        829.9324  606.19154 1053.6733 0.0000000
## Chinstrap:male-Adelie:male      -102.6961 -337.45009  132.0579 0.8067463
## Gentoo:male-Adelie:male         1465.7658 1235.26271 1696.2688 0.0000000
## Gentoo:male-Chinstrap:male      1568.4618 1352.11114 1784.8125 0.0000000
```

This gives us a LOT of info, but basically, we care about the effect of species (not sex or the interaction). And we can see that the main effect of species is that Gentoo has a much higher body mass than Adelie or Chinstrap, while the difference between Adelie and Chinstrap is not significant.

The main effect of sex: the p-value is also less than .05 so we can reject H0 and (since there are only 2 groups) say that there is a significant difference between the body mass of male and female penguins. We could look at the means for male and female penguins to see what this difference is.

```
tapply(penguins_data$body_mass_g, penguins_data$sex,
       mean, na.rm = TRUE) # males weigh more than females
```

```
##   female     male
## 3805.529 4559.439
```

The interaction: the interaction **is** significant, so the difference in body mass between male and female penguins depends on the species (or the difference of body mass between species depends on the sex of the penguin).

Now that we have a significant interaction, we can try to uncover what that interaction is by graphing the means. In this graph, we would have one grouping variable on the x-axis and the other grouping variable as different lines on the graph.

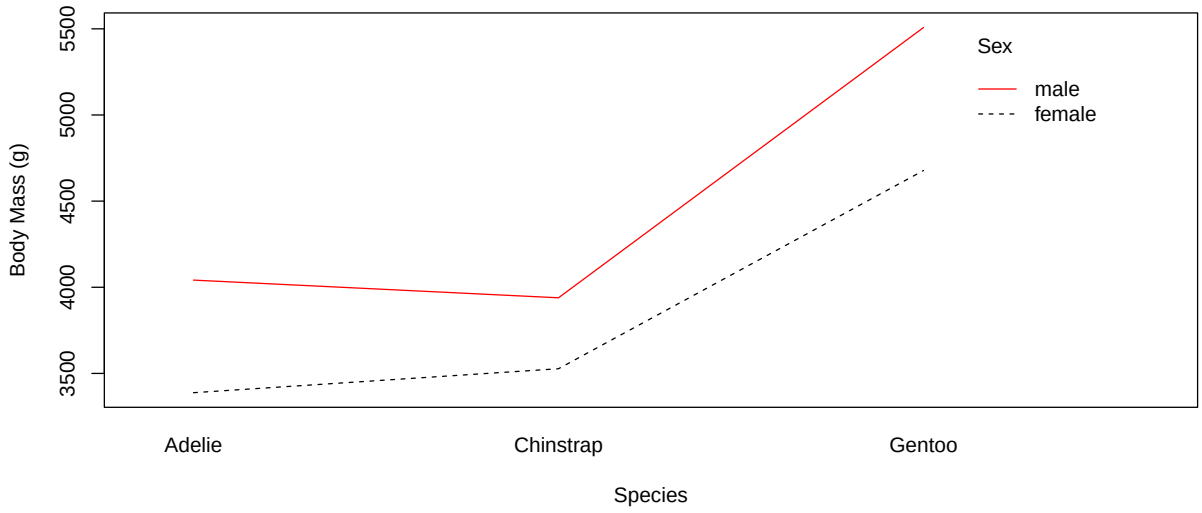
To do this, we will use the function `interaction.plot()`. This function takes the following arguments:

- `x.factor`: which grouping variable do you want on the x-axis?
- `trace.factor`: which grouping variable do you want as different lines?
- `response`: what is the outcome variable?
- `fun`: what summary statistic are you graphing? Default is mean, which is what we want!
- `type`: Set equal to "l" to graph lines (not points)

You can find the other function arguments that help make this graph pretty using `?interaction.plot`.

Let's say that I want species on the x-axis, and sex as the different lines:

```
interaction.plot(x.factor = penguins_data$species,
                 trace.factor = penguins_data$sex,
                 response = penguins_data$body_mass_g,
                 fun = mean, # note that I could omit this argument
                 type = "l",
                 col = c("black", "red"),
                 xlab = "Species", ylab = "Body Mass (g)",
                 trace.label = "Sex")
```



Note that from the interaction plot, we can sort of see the significant main effects we discussed earlier.

If we found the average of the male penguins (red line) and the female penguins (black line) across all species we can see that males, on average, weigh more than females.

Then, if we average the body mass for each species, we can see how Gentoo penguins weigh more than Adelie and Chinstrap, but Adelie and Chinstrap are pretty close to each other.

What does this plot tell us about the interaction? We can see that for all 3 species, males weigh more than females, so the interaction is not that the effect is reversed. Instead, the interaction appears to be that the magnitude of the difference is different depending on the species. In particular, males and females have a much smaller difference in the Chinstrap species than the other 2 species.