Lab 4 - Assumptions & Multiple Comparisons

ENVX2001 Applied Statistical Methods

Semester 1, 2026

> 💡 **Learning outcomes**
>
> At the end of this Lab students should be able to:
>
> - test the assumptions of ANOVA using residual diagnostics;
>
> - use plotting and Tukey's tests to determine which pairs of groups are significantly different;
>
> - use R to perform the analyses.
>
> All of the data for this practical is in the **Data4.xlsx** file.

Exercise 1 - Diatoms in streams (Walk-through)

Here we will test the assumptions using residual diagnostics and finding significant differences using plots and Tukey's test. The data is found in the **Diatoms** worksheet.

> **Question 1.1**
>
> *(i)* Importing and processing data, then fitting an ANOVA model
>
> ```
> CODE
> # write your code here
> ```

> **Question 1.2**
>
> *(ii)* Statistical test of the assumption of constant variance
>
> Statistics is made up of different tribes and some tribes use hypothesis testing to see if a dataset meets the assumptions of normality and constant variance. One option is the Bartlett's test for constant variances. The mechanics are not important but the function and syntax are shown below. The hypotheses are:
>
> - $H_0 : \sigma^2_{BACK} = \sigma^2_{LOW} = \sigma^2_{MED} = \sigma^2_{HIGH}$
>
> - $H_1 : not\ all\ \sigma^2_i\ are\ equal\ (i = BACK, LOW, MED, HIGH)$
>
> We prefer to use numerical and graphical diagnostics, e.g. residuals plots, but this is more to show you other possibilities. You can use this as a different line of evidence for testing assumptions if you wish.

> **⚠ Warning**
>
> It won't work if the data is non-normal and only use it if the data has one treatment factor with a completely randomised design!

```
CODE
#
```

Question 1.3

(iii) Identify significant differences

In Topic 3 we used the `emmeans` package to extract means for each group and their associated 95% CI. The `emmeans` is useful to produce a plot showing the mean and 95% CI which is a nice way to present the results.

```
CODE
#
```

Question 1.4

(iv) The another approach is to use a **Tukey's test** which we can extract using the `emmeans(model-goes-here, pairwise ~ your-treatment)` function from the `emmeans` package. Note there are many other post-hoc tests that have come and gone, but we will just focus on one of them.

Question 1.5

(v) Another way to present the results is to use the `plot()` function to show the confidence intervals and the comparisons among them.

Question 1.6

An alternative base R function for Tukey tests is `TukeyHSD()`. It creates a set of confidence intervals on the differences between the means of the levels of a factor. Also plot the results from this function.

> **💡 Tip**
>
> If the confidence interval does not cross over 0, then that pair significantly differs from each other.

```
CODE
#
```

Exercise 2 - Mean comparisons, residual diagnostics and back-transformations

In this exercise will add a layer of complexity by considering a transformation. If our data does not meet the assumptions we need to transform the data, possible transformations are the square root (weak) and log (high). When we transform the data we need to be careful about how we interpret the results.

Concentration of prolactin (units g/L) in the pituitary glands of nine-spined stickleback fish was assessed. The fish were kept in either saltwater or freshwater prior to assay and were different batches were examined on three successive occasions. Cysts tend to develop in fish when kept in saltwater and sometimes develop in freshwater populations. The four different groups of fish were used in a preliminary experiment to examine the effects of cysts, whether induced by saltwater or normally present, on the prolactin production of the pituitary gland.

The four groups of fish were codes as follows, with 10 fish per group:

- A = saltwater cysts, day 1;
- B = freshwater, no cysts, day 2;
- C = freshwater, no cysts, day 2;
- D = freshwater, cysts, day 3.

The data is found in the **Prolactin** worksheet.

> **Question 2.1**
>
> *(i)* Import the data into R, perform some exploratory data analysis to make **tentative** suggestions about differences between means and the likelihood of the data meeting the assumptions.
>
> ```
> CODE
> ```

```
#
```

Question 2.2

(ii) Fit an ANOVA model and test the assumption of normality using a QQ plot and a histogram - both based on standardised residuals.

```
CODE
#
```

Question 2.3

(iii) Assess the assumption of constant variance by:

- examine the plot of the standardised residuals against fitted values;

```
CODE
#
```

- From the performance package, use `check_model()` to assess the assumptions of the model. This function will check the assumptions of normality and constant variance;

```
CODE
#
```

- using the Bartlett's test;

```
CODE
#
```

- using `check_homogeneity()` from the `performance` package;

> 💡 Tip
>
> Note there are many tests: `method = c("bartlett", "fligner", "levene", "auto")`

```
CODE
#
```

- calculating the ratio of the larges SD:smallest SD to see if it is below 2:1;

```
CODE
#
```

Question 2.4

(iv) The data does not meet the assumptions so log transform ('log' function) the response and repeat *(ii)* and *(iii)* to test the assumptions;

```
CODE
#
```

In R you can transform data in the model formula see below or you could create a new column in your data frame, for example `fish$logProlactin←log(fish$Prolactin)`.

Question 2.5

(v) If the assumptions are met and there is significant F-test perform Tukey tests and identify which pairs are significantly different.

```
CODE
#
```

Question 2.6

(vi) One issue is that we have performed our hypothesis testing on the log scale. This means there are some steps to be made if we wish to interpret the data on the original scale; e.g. provide a 95% CI on the original scale. We will step through these.

Suppose the biologist was primarily interested in comparing the prolactin concentrations for A (saltwater cysts, day 1) vs B (freshwater, no cysts, day 1).

- Find the means and CI from the output of the `TukeyHSD()` or `emmeans()` function.
- The CI and mean are on the log scale, so back-transform the difference in the means (`exp()` function), the lower and upper end-point 95% CI. Note that the upper and lower tail are not of equal length on the original scale.

```
CODE
#
```

Question 2.7

(vii) Now have an estimate of the difference in the means on the original scale. It actually corresponds to a ratio on the original scale. The reason is based on log laws, we can write the difference between 2 logged numbers (A and B) as a log of their ratio (A/B);

$$\log(A) - \log(B) = \log\left(\frac{A}{B}\right).$$

If we back-transform the log of their ratio we get the ratio on the original scale;

$$e^{\log\left(\frac{A}{B}\right)} = \frac{A}{B}.$$

So the back-transformed difference between the pairs of the means is a ratio.

Note: if we were to back-transform the group means on the log scale we would get the geometric mean on the original scale.

Provide a biological interpretation for this estimate and confidence interval. Use the CI to decided if there is a significant difference between Treatment A and Treatment B.

Exercise 3 - Broiler Chickens

This exercise is an analysis of a set of growth data. It is an open question for you to gain more practice.

The effect of weight gain in dressed broiler chickens was determined after five generations of selection. Group A was bred by using only the heaviest 10% in each generation; groups B and C were bred using respectively the heaviest 30% and 50%; group D was obtained by crossing groups A and C of the previous generation. The dressed weights (kg) of 25 birds from each group have been recorded.

The data is found in the **Broilers** worksheet.

> **Question 3.1**
>
> *(i)* Write down the null and alternate hypothesis. What is the treatment factor, and how many levels does it have? What are the sample sizes for each group (n_i)?

> **Question 3.2**
>
> *(ii)* Import the data into R, and then obtain some numerical and graphical summaries of the data, by each group. How would you interpret these data? From these summaries, is the assumption of homogeneity of variances met? What about normality? Try a formal Bartlett's test using the `bartlett.test` or `check_homogeneity()` function. Use residual diagnostics to assess the assumptions.
>
> ```
> CODE
> #
> ```

> **Question 3.3**
>
> *(iii)* Note that the results of the analysis can only be used when the assumptions of the analysis have been met. If you believe that the assumptions are met, then what would your conclusions of the analysis of variance be? You should use the `summary()` function applied to your `aov()` object to obtain the ANOVA table.
>
> ```
> CODE
> #
> ```

> **Question 3.4**
>
> *(iv)* Without any formal analysis, consider the result of the group means in relation to the group treatment - i.e. type of selection. Would this pattern be expected? If appropriate perform a Tukey test.
>
> ```
> CODE
> #
> ```