One-way ANOVA and Post-Hoc Analysis

Scientists are interested in whether organic methods can be used to control harmful insects and limit their effect on sweet corn.[1] In a study of this question, researchers compared the weights of ears of corn under five conditions in an experiment in which sweet corn was grown using organic methods. A total of 60 plots were used in the study. In 12 plots of corn a beneficial soil nematode was introduced. In another 12 plots a parasitic wasp was used. Another 12 plots were treated with both the nematode and the wasp. In a fourth set of 12 plots a bacterium was used. Finally, a fifth set of 12 plots of corn acted as a control in which no special treatment was applied. The plots were all randomly assigned to the treatment conditions. Twenty-five ears of corn from each plot were randomly sampled and each was weighed (in ounces). Those weights were then used to determine an average weight for an ear of corn for each plot.

The data set can be loaded using the code

```
corn <- read.csv("https://aloy.rbind.io/data/cornplot.csv")
```

Notice that the treatment groups were recorded as numeric labels where

treatment label	condition
1	soil nematode
2	parasitic wasp
3	nematode and parasitic wasp
4	bacterium
5	control

Your task is to use a one-way ANOVA model to analyze this experiment and report the results. Recall that the one-way ANOVA model is given by

$$y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}, \quad \text{for } i = 1, 2, 3, 4, 5, \quad \text{and } j = 1, \dots, 12$$

EDA

To begin, explore the data set by completing the following tasks:

Task 1. Create an individual values plot of the data. Comment on the distribution within each treatment group. Are there outliers? Are the distributions skewed? Are there unequal variances?

Task 2. Reflecting on your answers to task 1, are you worried about any of the conditions necessary for one-way ANOVA?

[1] The problem and data set were obtained from *Intermediate Statistical Investigations* by Tintle et al.

ANOVA-based analysis

Now, let's use the one-way ANOVA model to conduct a first analysis.

Task 3. Use R to fit the one-way ANOVA model for this experiment. Remember that we use the `aov()` function to fit the model and the `summary()` function to print the ANOVA table.

> **Note**
>
> Since `Treatment` is coded numerically, you need to convert it to a categorical variable so that R knows what to do. Thus, you should use `factor(Treatment)` as the explanatory variable for your ANOVA model.

```
# Fill in the blanks to fit the ANOVA model
corn_anova <- aov(___ ~ factor(___), data = ___)
summary(___)
```

Task 4. The ANOVA model predicts that every point in a treatment group will be at the group mean. Use the below code to create a plot of the fitted ANOVA model. Comment on the fit of this model to your data.

```
group_means <- corn |>
  group_by(Treatment) |>
  summarize(avg = mean(Weight))

gf_jitter(Weight ~ Treatment, data = corn, width= 0.1) |>
  gf_summary(fun.y = "mean", color = "darkorange", size = 3, geom = "point")
```

Task 5. Construct a plot of the residuals vs. the explanatory variable (Treatment) and a histogram of the residuals for the ANOVA model. Comment on the fit of the ANOVA model.

Post-hoc analysis

In the Daily Prep you saw that the ANOVA F-test indicated that there is a statistically discernible difference in the treatment means. This is great information to have, but the scientists ultimately want to know *which* treatment groups are different and *how* they are different. To answer this question we need to assess how the means of the treatment groups relate to one another. This is sometimes called **post-hoc analysis**. For five treatment groups, this might involve looking at all ten pairwise comparisons.

You may recall from your introductory statistics course that running a large number of hypothesis tests or constructing a large number of confidence intervals quickly increases the chance of making at least one error (i.e., getting a small p-value or a CI that doesn't contain 0 when there's no effect). This is the problem of **multiple comparisons**. To protect ourselves from this issue, we can construct confidence intervals for pairwise comparisons using **Tukey's Honest Significant Difference (HSD)**.

> **Note**
>
> To compute multiple confidence intervals using Tukey's HSD
>
> 1. Choose the desired confidence level across all intervals, $1-\alpha$. Sometimes α is referred to as the

familywise error rate.

2. Find the critical value q from the standardized range distribution based on the number of intervals, $1-\alpha$, and the MSE degrees of freedom.
3. Compute the intervals as $\overline{y}_i - \overline{y}_j \pm \frac{q}{\sqrt{2}}\sqrt{MSE\left(\frac{1}{n_i} - \frac{1}{n_j}\right)}$.

Task 6. Calculate 95% confidence intervals for the 10 pairwise comparisons of each of the treatments using the `TukeyHSD()` command. If you saved your model as `corn_anova`, then this can be done as shown below. The result has rows labeled by what difference is being considered (e.g., `2-1` denotes treatment 2 minus treatment 1), showing the estimated difference in means (`diff`), the lower and upper end points of the confidence interval (`lwr` and `upr`), and an adjusted p-value for a two-sample test (`p adj`).

```
TukeyHSD(corn_anova, conf.level = 0.95)
```

Task 7. Summarize, broadly, what these intervals tell you about how the treatments compare to each other.