ST 411/511 Lab 5
Introduction to One-way ANOVA

Objectives for this Lab

- Perform an ANOVA F-test to determine if the mean lifetimes differ for the different diets in case study 5.1.1.
- Obtain an ANOVA table for the finch beak depth data of case study 2.1.1, and compare this analysis to the two-sample t-test we did in Chapter 2 and Lab 2.
- Consider the intuition behind the ANOVA table and the ANOVA F-test.

1. As usual, start up RStudio and open Lab5.R. Load the Sleuth3 and ggplot2 R packages.

```
> library(Sleuth3)
> library(ggplot2)
```

2. We'll start with the diet restriction and longevity case study of Chapter 5.

   (a) View the data.

   ```
   > View(case0501)
   ```

   As with most of the other case study data, this data frame contains two columns. The first contains the response variable (`Lifetime`, measured in months) and the second contains a grouping variable (`Diet`). If you scroll down, you'll see there are more than two groups.

   (b) Check to see how many groups there are, what they're called, how R orders them, and the sample size in each group.

   ```
   > summary(case0501$Diet)
   ```

   (c) Create side-by-side boxplots.

   ```
   > qplot(Diet, Lifetime, data=case0501, geom="boxplot")
   ```

   As with the two-sample t-test, one-way ANOVA assumes the populations are normal with equal standard deviations. Do the boxplots suggest these assumptions are reasonable?

   (d) The two-sample t-test can be generalized to the situation when there are more than two groups, as is the situation here. This analysis tests the null hypothesis that all six population means are equal vs. the alternative hypothesis that at least two means are not equal (**not** that all the means are different). The calculations are done by the `aov()` command. However, the output from `aov()` is limited, so we save the aov "object" in a variable called `case0501_aov`. Presently we will use the `anova()` function to produce the desired output.

   ```
   > case0501_aov <- aov(Lifetime~Diet, data=case0501)
   ```

   Saving the object `case0501_aov` tells R we don't want any output at all.

   (e) You can see what the output from `aov()` looks like by typing the object name.

   ```
   > case0501_aov
   ```

(f) To get an analysis of variance table comparable to Display 5.10, use the `anova()` command on `case0501_aov`.

```
> anova(case0501_aov)
```

The test statistic in an ANOVA is called an *F-statistic*. Under the null hypothesis that all the population means are equal, the F-statistic has an F distribution. F distributions have two degrees-of-freedom parameters, whereas t distributions have only one. More on this later.

The F-statistic and p-value are shown on the ANOVA table in columns labeled `F value` and `Pr(>F)`. What is do you conclude from this p-value?

3. Since one-way ANOVA generalizes the two-sample t-test, we can apply `aov()` and `anova()` to the finch beak depth data of case study 2.1.1 where we first saw the two-sample t-test. This will allow us to recognize some familiar numbers in the ANOVA table.

(a) First, do the two-sample t-test. The ANOVA F-test is inherently a two-sided test, and it assumes equal standard deviations, so perform a comparable t-test:

```
> t.test(Depth~Year, data=case0201, var.equal=TRUE)
```

(b) Now analyze the finch data using `aov()` as in item 2. and obtain the ANOVA table from `anova()`.

```
> case0201_aov <- aov(Depth~Year, data=case0201)
> anova(case0201_aov)
```

Note that the first two arguments to `aov()` are the same as to `t.test()`.

Compare the output from `anova()` and `t.test()`. Find the t-test's p-value and degrees of freedom in the ANOVA table.

(c) The equality of p-values between the two-sample t-test and the one-way ANOVA F-test suggests that they are the same test. In fact, you can check that the square of the t-statistic is the F-statistic:

```
> (-4.5833)^2
```

a relationship that holds whenever there are only two groups. This relationship illustrates why the ANOVA F-test is a two-sided test. The one-sided t-test's p-value depends on the sign of the t-statistic, whereas the F-statistic is always positive.

The ANOVA F-test is the same as a two-sided two-sample t-test when there are two groups. When there are more than two groups, the ANOVA F-test can be thought of as a generalization of the two-sample t-test.

(d) On page 28 of Outline 2, we calculated the pooled standard deviation $s_p = 0.9730406$ (see also Display 2.8 on page 41 and Section 5.2.2 on page 120 of the textbook). s_p estimates the population standard deviation σ, which we assumed was the same for each population. The square of s_p is called the *residual mean square* or *mean squared error* (MSE) and estimates the population variance σ^2:

```
> 0.9730406^2
```

Find this quantity on the ANOVA table (it's been rounded there).

(e) In addition to the residual mean square, the ANOVA table has a mean square for `Year`. Item 5(f) below explains its interpretation.

The mean squares in the ANOVA table are always the corresponding sum of squares (`Sum Sq` in the ANOVA table) divided by the corresponding degrees of freedom (`Df`). Check that this is true for the ANOVA table at hand. Since the degrees of freedom for `Year` are 1, the first mean square is 19.889/1. Check that the residual mean square is the residual sum of squares divided by the residual degrees of freedom:

```
> 166.638/176
```

Note that the residual degrees of freedom are the same as the degrees of freedom for s_p given on page 40 of the *Sleuth*. The residual degrees of freedom are always those associated with the estimate of σ^2.

4. We will want to write confidence intervals to estimate individual population means and differences between two population means. The formulas will be the same as on page 32 of Outline 2, except the pooled standard deviation s_p and its associated degrees of freedom, which we use to get the t-quantile, will come from the ANOVA table.

   (a) In particular, we'll need the sample means.

   ```
   > with(case0501, unlist(lapply(split(Lifetime, Diet), mean)))
   ```

   Read the command from the inside out. The `split()` function produces an R "list" with six elements, each containing the data for one of the groups. `lapply()` ("list" apply) takes each element of the list, applies the `mean()` function to it, and returns a list of the six means. Finally, `unlist()` converts the list to a vector.

   (b) We can use an analogous command to get the sample sizes, which we also need for confidence intervals.

   ```
   > with(case0501, unlist(lapply(split(Lifetime, Diet), length)))
   ```

   The `summary()` command in 2(b) also gives sample sizes, so we could have used that instead.

5. We will discuss degrees of freedom and sums of squares in more detail in lecture. The material below aims to give some intuitive background to the ANOVA table and the ANOVA F-test. These are new ideas. Don't be concerned if they're not immediately completely clear.

   (a) We will take a closer look at the ANOVA table from item 2(f). We got the ANOVA table from R by giving the aov object to the `anova()` command.

   ```
   > anova(case0501_aov)
   ```

   (b) The ANOVA F-test is a comparison between two models. The *null model* is the one given by the null hypothesis. It requires that all the population means are equal. The other model is one given by the alternative hypothesis. This model allows each population to have a different mean. The textbook refers to these models as *full* and *reduced* models. If you take ST 412/512, you will spend a lot of time thinking about full and reduced models. For our ANOVA F-test in Chapter 5,

   $$\text{reduced model} \quad = \quad \text{null model}$$

$$
\begin{aligned}
&= \text{all } \mu_i \text{ are equal} \\
&= \text{“equal means model”} \\
\text{full model} &= \text{alternative model} \\
&= \text{not all } \mu_i \text{ are equal} \\
&= \text{“separate means model”}
\end{aligned}
$$

The null model is very simple. The alternative model is more complex. A more complex model always fits the data better, but we don't want a model that's too complex because it will be harder to interpret, and we run the risk of "overfitting" our data. The F-statistic compares how well the two models fit, taking into account model complexity.

(c) Residual degrees of freedom quantify the complexity of a statistical model compared to the amount of information in the data set. In the one-sample case, the residual degrees of freedom are $n-1$, and in the two-sample case, they are $n-2$. Both of these are sample size minus number of mean parameters (one for each separate population). The sample size quantifies the amount of information in the sample, and the number of mean parameters quantifies the complexity of the model.

Refer to the ANOVA table. Find the degrees of freedom for `Residuals`. Check that it follows the same pattern.

```
> nrow(case0501) # Find total sample size
> length(unique(case0501$Diet)) # How many different groups?
```

(d) Degrees of freedom for grouping variables such as `Year` in `case0201` and `Diet` in `case0501` represent something different than residual degrees of freedom. The one degree of freedom for `Year` in the ANOVA table from item 3(b) indicates that a model allowing different population means for each year is one parameter more complex than the model that assumes both years share a common population mean. Does this kind of interpretation work for the degrees of freedom for `Diet` in the ANOVA table in the longevity study?

This difference in complexity between full and reduced models is called the *extra degrees of freedom*. It's the number of extra parameters in the full model.

(e) The residual sum of squares quantifies the variation in the data not explained by the full model. This variation is sometimes called *noise*.

(f) In the ANOVA table from item 3(b), the sum of squares for `Year` quantifies the variation in the data attributable to systematic differences between years. Similarly, the sum of squares for `Diet` quantifies the variation in the data attributable to systematic differences among the different diets. That is, the sum of squares for `Diet` and `Year` represent the ability of the full model to explain how the data vary between the different groups. The variation explained by the full model is sometimes called the *signal*.

In lecture and in the textbook, the sum of squares for the grouping variable (`Year` or `Diet`) is called the *extra sum of squares*, because it's calculated by subtracting the residual sum of squares for the full model from the residual sum of squares for the reduced model. The extra sum of squares is the extra variation explained by the full model over the reduced model.

(g) Side-by-side boxplots illustrate two sources of variation (variation explained by the full model and variation not explained by the full model) if we also plot sample means. Here's some `ggplot()` code to do this.

```
> ggplot(data=case0501, aes(x=Diet, y=Lifetime)) +
+   geom_boxplot() +
+   stat_summary(fun=mean, geom="point", shape=3, size=3)
```

This is the same plot as in item 2(c), but the sample means appear as plus symbols on the boxes. The sample means estimate the population means, so the variation among population means is illustrated by the fluctuating vertical positions of the pluses.

The height of the boxes illustrates the variation *not* explained by the full model. Here, the pluses are more widely scattered than the average height of the boxes, suggesting the full model explains more variation than not.

(h) The idea behind the ANOVA F-test is to compare the variation explained by the full model with the variation not explained by this model. However, the test also needs to account for the complexity of the model, since a more complex model will be flexible enough to explain more variation. That's what the mean square for the grouping variable quantifies:

$$\begin{aligned} \text{Mean Square for Diet} &= \frac{\text{Diet sum of squares}}{\text{Diet degrees of freedom}} \\ &= \frac{\text{extra variation explained by the model}}{\text{extra model complexity}} \\ &\approx \text{explanatory power of the full model per unit of model complexity} \end{aligned}$$

The F-statistic is the ratio of the "model" mean square (here the mean square for `Diet`) to the residual mean square. Verify this in the ANOVA table.

```
> 2546.8/44.6
```

The numerator of this F-statistic is **much** larger than the denominator, indicating that the model explains much more variability in the data than it fails to explain, even after allowing for model complexity.

The small p-value that results from the large F-statistic indicates that null hypothesis is not credible. The model that allows different means for all six populations is more plausible than the model that requires they all have the same mean.