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.
- Calculate s_p in the two-sample case, and observe that the ANOVA table contains s_p^2.
- Write a confidence interval for the difference between two population means in case study 5.1.1.
- 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 one of the population means is different than the others (**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()`. The `aov()` function always makes the equal variance assumption.

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, you can think of the ANOVA F-test as a generalization of the two-sample t-test.

(d) Page 26 of Outline 2 gives the following formula for the pooled standard deviation s_p in the two-sample case.

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}. \tag{1}$$

The pooled standard deviation s_p estimates the population standard deviation σ, which we assumed was the same for each population. We have noted that the expression on the inside of the square root of (1) is a weighted average of the sample variances, s_1^2 and

s_2^2, where the weights are the sample sizes minus 1. Thus a larger sample will have more weight, which is appropriate because a larger sample contains more information about the parameter σ.

Display 2.8 on page 41 of the textbook calculates s_p "by hand" for the finch study. We will do the same calculation in R. We need the two sample standard deviations and the two sample sizes. Putting R commands inside parentheses tells R to show the results. Compare your results to Display 2.8.

```
> (s1 <- with(case0201, sd(Depth[Year==1976])))
> (s2 <- with(case0201, sd(Depth[Year==1978])))
> (n1 <- with(case0201, length(Depth[Year==1976])))
> (n2 <- with(case0201, length(Depth[Year==1978])))
> (sp <- sqrt(((n1-1) * s1^2 + (n2-1) * s2^2)/(n1 + n2 - 2)))
```

Be careful with parentheses in the last command! We need to enclose numerator and denominator in parentheses, then enclose everything in parentheses after `sqrt`. RStudio can help. Click to the right of any parenthesis, and RStudio will highlight the matching parenthesis.

The square of s_p is called the *residual mean square* or *mean squared error* (MSE) and estimates the population variance σ^2:

```
> sp^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 how to interpret the mean square of a grouping variable.

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. We will need to pay attention to the residual degrees of freedom anytime we use the MSE to estimate σ^2.

4. When we have more than two groups, we will still want to write confidence intervals to estimate differences between two population means. The formula will be almost the same as on page 32 of Outline 2:

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{df}(1-\alpha/2) \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

except the pooled standard deviation s_p and its associated degrees of freedom df, which we use to get the t-quantile, will come from the ANOVA table, and the subscripts on the $\overline{Y}_i$'s and n_i's will reflect the two means of interest.

Let's write a confidence interval for the difference between population mean lifetimes for the N/R50 and lopro diets.

(a) We will need the sample means.

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

Read this 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.

The sample means we will need are 42.29718 and 39.68571, the third and sixth sample means according to R's ordering.

(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.

The sample sizes we will need are 71 and 56.

(c) Find s_p^2 and residual degrees of freedom in the ANOVA table from 2(f). We need the residual degrees of freedom to calculate the t-quantile using R's `qt()` function, as in item 5(e) of Lab 3.

```
> qt(0.975, 343)
```

The first argument to `qt()` is the area to the left of the t-quantile. We are writing a 95% confidence interval, so that area is 0.975. The second argument to `qt()` is the residual degrees of freedom from the ANOVA table in 2(f).

(d) Calculate the confidence interval. The N/R50 and lopro treatments are treatments 3 and 6 in R's ordering, so the formula looks like

$$\overline{Y}_3 - \overline{Y}_6 \pm t_{df}(1-\alpha/2) \cdot \sqrt{\frac{s_p^2}{n_3} + \frac{s_p^2}{n_6}}$$

where $\overline{Y}_3 = 42.29718$, $\overline{Y}_6 = 39.68571$, $n_3 = 71$, and $n_6 = 56$ from items 4(a) and 4(b), and $df = 343$ and $s_p = \sqrt{44.6}$ from the ANOVA table in 2(f).

```
> 42.29718-39.68571 - qt(0.975, 343)*sqrt(44.6)*sqrt(1/71 + 1/56)
> 42.29718-39.68571 + qt(0.975, 343)*sqrt(44.6)*sqrt(1/71 + 1/56)
```

The confidence interval is $(0.2638417, 4.959098)$. We will write a statistical conclusion for this interval in class.

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. This material will not be on the midterm.

(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)
```

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

Both the full and reduced models assume that the populations are normally-distributed around their means with the same standard deviation σ.

Summarizing, for our ANOVA F-test in Chapter 5,

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

(b) 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, i.e. fitting the randomness in the particular sample we have. The F-statistic compares how well the two models fit, taking into account model complexity.

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 $n_1 + n_2 - 2$. Both of these are total 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?
```

(c) 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.

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 compared to the reduced model. Does this interpretation work for the degrees of freedom for `Diet` in the ANOVA table in the longevity study?

(d) The sum of squares for `Diet` quantifies the variation in the data attributable to systematic differences among the different diets' population means. That is, the sum of squares for

`Diet` represents 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*.

The *residual* sum of squares quantifies the variation in the data not explained by the full model. This we sometimes call *noise*.

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.

(e) Side-by-side boxplots illustrate the two sources of variation (variation explained by the full model and variation not explained by the full model) if we also plot sample means. The sample means estimate the population means. Varying population means is how the full model accommodates variation in the data. This code to put sample means on the boxplots:

```
> qplot(Diet, Lifetime, data=case0501, 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 variation in the data explained by the full model is illustrated by the fluctuating vertical positions of the pluses.

The height of the boxes shows the difference between the third and first quartiles of each sample, a measure of the spread of the data within each sample. We can view this as an illustration of 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 it leaves unexplained.

(f) 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:

$$\begin{aligned}\text{Mean Square for Diet} &= \frac{\text{extra variation explained by full model over reduced model}}{\text{extra complexity of full model over reduced model}} \\ &\approx \text{improvement in explanatory power of full model over reduced model per unit of added model complexity}\end{aligned}$$

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.