Statistics 411/511
Lab 5

Lab Instructions: If you want to work along with the TA in lab, please sit near the front. If you prefer to go to lab but work at your own pace, please sit near the back, and wait for the appropriate time to ask any questions.

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.
- 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. Load the Sleuth3 and ggplot2 R packages. If you are working in Bexell, you'll have to install the packages again as described in item 5(a) of Lab 1.

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

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

   (a) Preview the first few lines of data.

   ```
   > head(case0501)
   ```

   As with most of the other case studies, this data frame contains two columns. The first contains the response variable (lifetime in months) and the second contains a grouping variable.

   (b) Check to see how many groups there are, what they're called, and how R orders them.

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

   (c) Create a boxplot.

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

   As with the two-sample t-test, one-way ANOVA assumes the population standard deviations are equal. Do the boxplots suggest this assumption is 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. The calculations are done by the `aov()` command. However, the output from `aov()` is limited, so we save the "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*. The F-statistic and p-value are shown on the ANOVA table in columns labeled `F value` and `Pr(>F)`. What is the conclusion of the test?

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 as in item 1.

```
> 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
```

This explains 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.

(d) On page 2 of October 5th's lecture notes, we calculated the pooled standard deviation $s_p = 0.97304$ (see also Display 2.8 on page 41 and Section 5.2.2 on page 120 of the textbook). The square of s_p is called the *residual mean square* or *mean squared error* (MSE) and estimates σ^2:

```
> 0.97304^2
```

Find this quantity on the ANOVA table.

(e) In addition to the residual mean square, the ANOVA table has a mean square for "Year." 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 clearly 19.889/1. Check the residual mean square:

```
> 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*.

4. We will discuss degrees of freedom and sums of squares in more detail in lecture. The material below aims to give some intuitive background.

   (a) Residual degrees of freedom quantify the complexity of the 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).

   Refer to the ANOVA table from item 2(f). Find the degrees of freedom for Diet. Check that it follows the same pattern.

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

   (b) Degrees of freedom for Year or Diet represent something different than residual degrees of freedom. The one degree of freedom for Year indicates that a model allowing different means for each year is one parameter more complex than the model that assumes the two years share a common mean. Does this interpretation work for the ANOVA table in the longevity study?

   (c) In the ANOVA table from item 3(b), the sum of squares for Year quantifies the variability in the data attributable to systematic differences between years, whereas the residual sum of squares quantifies the variability in the data **not** attributable to year. Similarly, the sum of squares for Diet quantifies the variability in the data attributable to systematic differences in Diet. That is, the sum of squares for Diet and Year represent the ability of the *model* to explain how the data varies. In these two cases, the model specifies how many different populations there are, each with its own mean.

   The idea behind the ANOVA F-test is to compare the variability explained by the model with the variability not explained by the 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 variability. That's what the model mean squares quantify:

$$\begin{aligned}\text{Mean Square} &= \frac{\text{sum of squares}}{\text{degrees of freedom}} \\ &= \frac{\text{variability in data explained by the model}}{\text{model complexity}} \\ &= \text{explanatory power of the model per unit of complexity}\end{aligned}$$

   The F-statistic is the ratio of the model mean square to the residual mean square. Verify this for the ANOVA table of item 2(f).

```
> 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 allows only one mean.