STAT 251
Lab 8: Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA)

In the previous labs we learned about hypothesis testing, normally along the lines on whether or not some parameter, like μ, is equal to some constant value that the experimenter claims. What happens, though, if we want to test to figure out whether or not groups are the same? For example, what if we wanted to test on whether or not two shipments of steel rods have the same breaking points. To test the equality of groups, for $k \geq 2$, we build an ANOVA table, which stands for Analysis of Variance.

Note: Technically speaking, ANOVA can be used to compare 2 or more population means. However, when there are only 2 groups to compare, we can still use ANOVA, but which is much more computationally intensive: therefore we just use the 2-sample z or t tests in such cases.

As you've seen in class, the calculations for ANOVA can be quite tedious, and so we'll show you how to do you analysis via R to make things much simpler. To start, let's go to the course website and download: **players.txt** and load it into R using the **read.table()** command.

The data file contains the stats on the members of three hockey teams, the

Vancouver Canucks, Calgary Flames, and the Edmonton Oilers. Now suppose you wanted to test the hypothesis that the hockey players in the three teams have the same mean height. Therefore, our hypotheses are

$$\text{H}_0\text{: } \mu_{canucks} = \mu_{oilers} = \mu_{flames}, \text{ H}_a\text{: at least one } \mu_i \neq \mu_j$$

To do this test, we'll need to use the **lm()** command. As we saw previously, the *lm()* command was used to build linear models and again it comes into play here because we use to build a relationship between the three teams and the height of their players.

```
> hockey = read.table("players.txt", head=T)
> attach(hockey)
> names(hockey)
> model = lm(height ~ team)
```

Note: The response variable is **height** and we are comparing that with respect to explanatory variable **team**, which are either the *Canucks*, the Oilers, or the *Flames*. If, by chance, we use numerical values like 1, 2, 3 to represent the different teams, R would think of the variable as continuous instead of being categorical. To alleviate this, we would have to use the **as.factor()** command, which would tell R that the data is categorical. Therefore, to set up our model:

```
> model = lm(height ~ as.factor(team))
```

Now that we have a model built, we can use perform an ANOVA on our data. Interestingly enough, although we are performing ANOVA on our data, the command we will use is actually the **aov()** command. There is, in fact, an **anova()** command, but interpreting the results is a bit more difficult, and so we'll opt to perform our ANOVA with the former:

```
> fit = aov(model)
> summary(fit)
```

and as we can see in the display output, our familar ANOVA table.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
team	2	5.2706	2.6353	569.38	< 2.2e-16	***
Residuals	72	0.3332	0.0046			

Similar to how we did our analysis in linear regression for significance of covariates, we focus on the p-value. The p-value is the farthest right column, this time signified by Pr(>F), recalling the p-value is the probability of getting a higher statistic (in our case, F value and previously for linear regression a T value) than the current one. Recall, if the p-value was greater than 0.05 then we could assume that the hypothesis was "true" and so we'd fail to reject the null and be done, but where's the fun in that? So, comparing the p-value in the output to $\alpha = 0.05$ (or any suitable level), there is overwhelming evidence to reject the null hypothesis in favor of the alternative hypothesis (that is, the p-value is way less than 0.05 which leads to the rejection of the null).

Alternatively, as you learned in class, you can compare the F-value directly since you know the degrees of freedom from the ANOVA. Therefore, if we looked up the critical F-value for $F_{2,15}$ at a significance level of $\alpha = 0.05$ in a F-table, we find $F_{2,15} = 6.3589$, and comparing it against the F-value calculated by ANOVA, which is 569.38, again we see overwhelming evidence to reject the null. Note: The 3 $*$ symbols indicate how strong the significance, the more $*$'s, the more evidence against the null.

Pairwise comparison of Means

Now you may question, how do we determine if any of the mean heights are the same? To do this, we have to do pairwise comparisons. The number of comparisons to be made is

$$K = \begin{pmatrix} k \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} = 3$$

Therefore, we get the following hypotheses to compare one by one (not simultaneously anymore which we were doing in ANOVA):

1. H_0: $\mu_{canucks} = \mu_{oilers}$ vs. H_a: $\mu_{canucks} \neq \mu_{oilers}$
2. H_0: $\mu_{oilers} = \mu_{flames}$ vs. H_a: $\mu_{oilers} \neq \mu_{flames}$
3. H_0: $\mu_{canucks} = \mu_{flames}$ vs. H_a: $\mu_{canucks} \neq \mu_{flames}$

To do the comparison, we first we need to grab some vital information, which includes the number of players on the team and their respective means. It is easy enough to look through the data to figure which lines correspond to which teams, but let's bypass that and use some R-code:

```
> u1 = mean(height[which(team == "Can")])
> u2 = mean(height[which(team == "Oil")])
> u3 = mean(height[which(team == "Fla")])
> n1 = length(which(team == "Can"))
> n2 = length(which(team == "Oil"))
> n3 = length(which(team == "Fla"))
```

Therefore, if we were to compare the Canucks against the Oilers,

$$\hat{\Delta}_{1,2} = \hat{\mu}_{canucks} - \hat{\mu}_{oilers} = u1 - u2 = -0.6492$$
$$se(\hat{\Delta}_{1,2}) = \sqrt{MSE}\sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}} = \sqrt{0.0046}\sqrt{\frac{25+25}{25 \cdot 25}} = 0.019183$$

where MSE came from our ANOVA table. Now, recalling that for a 95% CI we need to modify our t-value,

$$t - value = t_{(\delta),(n-k)} = t_{(\frac{\alpha}{K}),(n-k)} = t_{(\frac{0.05}{3}),(75-3)} = t_{(0.016667),(72)} = 2.451$$

To calculate the t-value, you can normally look it up on a student t-distribution table, but because we sometimes deal with odd ratios, the following is the code to look up a t-value based on a certain number of degrees of freedom.

> **qt((1-0.0166667/2),72)**

or

> **k <- length(unique(Type))** #number of groups
> **C <- choose(k,2)** #compute kC2 (Bonferroni)
> **qt((1-(0.05/(2*C))),(length(Type)-k))**

or, for three groups only, `k = length(unique(team)) = choose(3,2) = 3`, we could simply write:

> **qt((1-(0.05/length(unique(team)))/2),**
> **(length(team)-length(unique(team))))**

Note: The way it is set up is to keep it similar to the form of $t_{(1-\frac{\alpha}{2}),(n-1)}$
Therefore, our 95% confidence interval is:

$$\left(\hat{\Delta}_{1,2} - t_{(\frac{\alpha}{2*K}),(n-k)} \cdot se(\hat{\Delta}_{1,2}),\ \hat{\Delta}_{1,2} + t_{(\frac{\alpha}{2*K}),(n-k)} \cdot se(\hat{\Delta}_{1,2})\right)$$
$$(-0.6492 - 2.451 \cdot 0.019183,\ -0.6492 + 2.451 \cdot 0.019183)$$
$$(-0.6962, -0.6022)$$

To do this in R, we do as follows:

$>$ **(u1-u2)-qt((1-(0.05/length(unique(team)))/2),**
(length(team)-length(unique(team))))
$\times$ **sqrt((n1+n2)/(n1** $\times$ **n2)**
$\times$**(sum(fit$residuals^2)/fit$df.residual))**

And similar to how we analyze any of our previous confidence intervals, since $\Delta_{1,2} = 0 \notin (-0.6962, -0.6021779)$, we can conclude that the mean heights of the Canucks and the Oilers is different and that, in general, the Oilers are taller than the Canucks. And similarly, we would compare Canucks against Flames and Oilers against Flames; recalling that K = 3.

Skipping the details, we find that a 95% CI for Canucks vs. the Flames is: $(-0.3598, -0.2658)$, which implies that again there is a difference and that in general, the Flames are taller than the Canucks. For a 95% CI for Oilers vs. Flames is: (0.2894, 0.3834, which implies that there is a difference and that in general, the Oilers are taller than the Flames. Therefore, we conclude that the mean heights of the three hockey teams are different and that in general, the Oilers are the tallest, followed by the Flames, followed by the Canucks: $\mu_{Oilers} \geq \mu_{Flames} \geq \mu_{Canucks}$.