Psyc 2111, Jones

Summary of Lab Week 11

ANOVA

A simple ANOVA allows us to test whether any number of groups have reliably different means. Start by loading the example data in "lab11-1.txt" on the course website. Run a summary of the data to see what the variables look like. You'll see that we have one variable showing what group each subject is in, and a second showing each subject's score on the dependent variable.

The first thing to look at is the sample group means. You can find the mean of each group separately:

```
> mean(d$X[d$group==1])   #etc.
```

The means differ, but do these differences reflect real differences among the populations? To answer this question, we do an ANOVA. This involves calculating the sums of squares, mean squares, and F statistic. The first step is to calculate the grand mean, which is simply the mean of all the scores (`d$X`), and the means of the separate groups, which you just did (but you may have to do again if you didn't store them as variables). The other information you'll need are the sizes of all the groups, for example

```
> n1 = sum(group==1)
```

Now you can compute the sums of squares. To get the treatment sum of squares, square the difference between each group mean and the grand mean, multiply by the group size, and add up the results for all the groups. To get the residual sum of squares, calculate the sum of squares for each group, and then add them up. For example, you calculate the sum of squares for Group 1 by subtracting all the scores in Group 1 from the mean of that group, then squaring and summing.

```
> group1.deviations = d$X[d$group==1] - M1
> group1.sq.deviations = group1.deviations^2
> SS.group1 = sum(group1.sq.deviations)
```

You can also calculate the total sum of squares, to check your answers for $SS_{\text{treatment}}$ and SS_{residual} (they should add up to SS_{total}). To get the total sum of squares, just subtract each score from the grand mean, square, and sum.

Once you have the sums of squares, the next step is to get the mean squares, by dividing by degrees of freedom. The degrees of freedom for the treatment is one less than the number of groups (k – 1). The degrees of freedom for the residual is the total number of scores, which you can get from `length(d$X)` or `dim(d)`, minus the number of groups.

After you have the mean squares, divide $MS_{\text{treatment}}$ by MS_{residual} to get F. According to the null hypothesis, that there are no differences among the populations, $MS_{\text{treatment}}$ should be about σ^2 (the population variance of the raw scores). MS_{residual} is also about σ^2, meaning we can use MS_{residual} as an estimate of σ^2. If $MS_{\text{treatment}}$ is much larger than MS_{residual}, then the differences among the sample means are greater than would be expected by chance. We would see this as a large value of F. If $MS_{\text{treatment}}$ is just as big as the null hypothesis predicts, then F should be about 1.

In this case, you should find that $MS_{\text{treatment}}$ is bigger than MS_{residual}, and therefore F is bigger than 1. The question is whether F is big enough that we can confidently reject the null hypothesis. To answer this, find the p-value, which is the probability that F would come out as large as it did if the null hypothesis were true (i.e., just by chance). We get this probability from the `pf()` function, remembering to subtract from 1 because `pf()` gives us the probability of a result less than F.

```
> 1 - pf(F,df.treatment,df.residual)
```

If `p < .05`, then we conclude the differences among the group means are reliable. If not, then we conclude the differences can be explained by chance.

Hypothesis testing with the `anova()` function

Now that we've done the ANOVA by hand, we can check it against R's built-in function. The `anova()` function works in the same way as the `summary()` function for testing regression. The input you give it is the "model" produced by the `lm()` (linear model) function. The model is a complicated kind of variable, but you can think of it as R's description to itself of how the dependent variable (`X`) is explained by the treatment (`group`).

To create the model for the ANOVA, we tell `lm()` to try to explain `X` using `group`. Since `group` is a nominal-scale variable, we need to tell `lm()` to treat it as a factor (rather than as a continuous, interval-scale variable).

```
> model = lm(X ~ factor(group))
```

Now we can run the ANOVA on the model. The steps getting to this point—using `lm()` to create a model of `X` explained by `group`—might be a little confusing, but it's just an idiosyncrasy of how R works. The important part to understand is what the output of `anova()` tells us.

```
> anova(model)
```

The output of this command is what's called an ANOVA table. It shows the degrees of freedom, sum of squares, and mean square for the treatment and the residual. Then it shows the F statistic and p-value. All 8 of the entries in this table are numbers you calculated earlier. Check the table against your own results. Seeing how these two sets of values line up should help you understand what `anova()` does. The function produces a lot of numbers, but they're exactly the numbers you calculate yourself (*SS*, *MS*, *F*, *p*) when you do an ANOVA.