**Biostatistics**

**Lab 10**

*Tasks*

ANOVA in R

*Introduction*

Today will be our second lab using R to do statistical tests.

*Set-Up and Data*

If you have not done so already, download from Blackboard all of the data files necessary for the lab and save them in the folder for today. The file that will be most important for today is the Dryad\_Eldridge\_Ecology.csv data.

Start R Studio and open a new R Notebook file. Save this file in today’s folder where the data files are. Delete the text that comes in the notebook template, and use the first chunk to load libraries. It is good practice to also present your session info, make a last chunk to do this. Make a chunk 1 to read in the data. For the sake of time, simply copy and paste this dplyr pipe to import and clean the data:

Fungus <- read\_csv("Dryad\_Eldridge\_Ecology.csv",

col\_names = TRUE) %>%

mutate(LogNAG = log(NAG),

Community = as\_factor(Community),

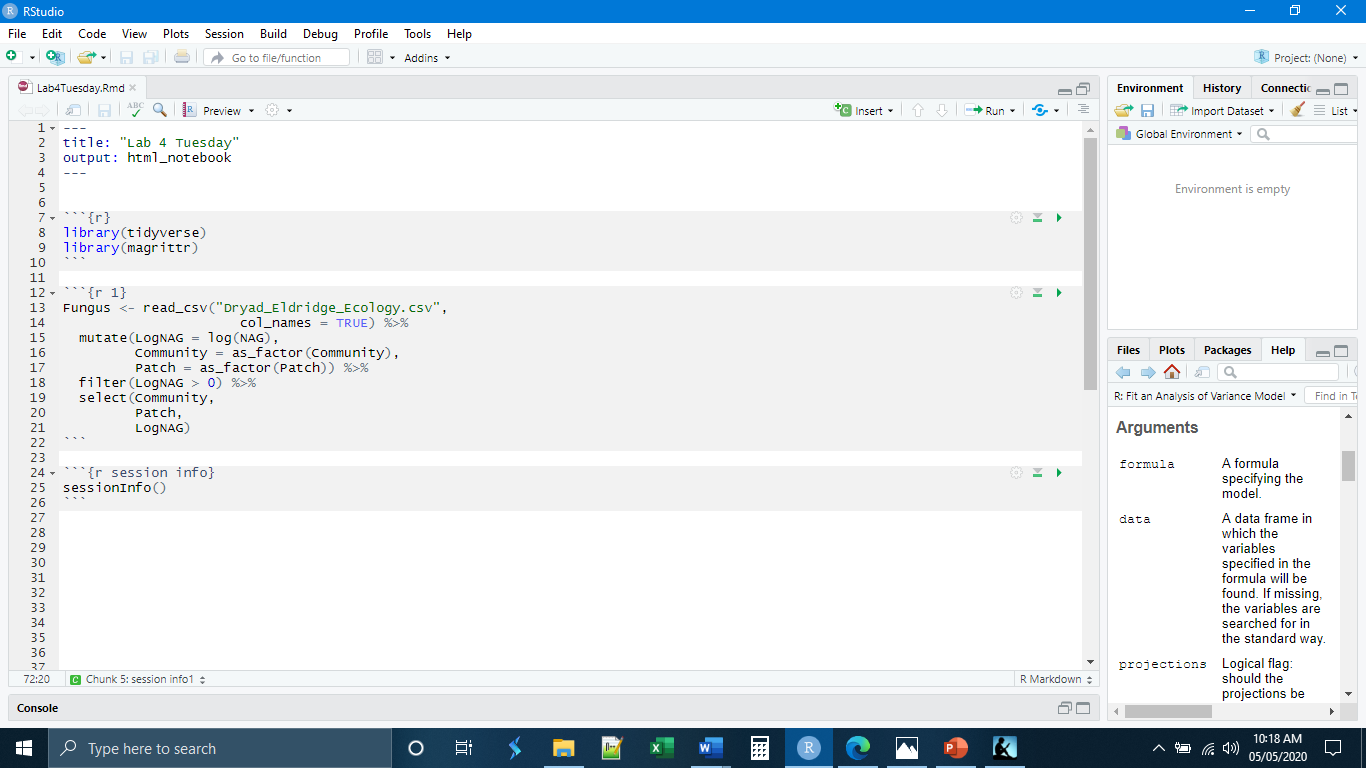
Patch = as\_factor(Patch)) %>%

filter(LogNAG > 0) %>%

select(Community,

Patch,

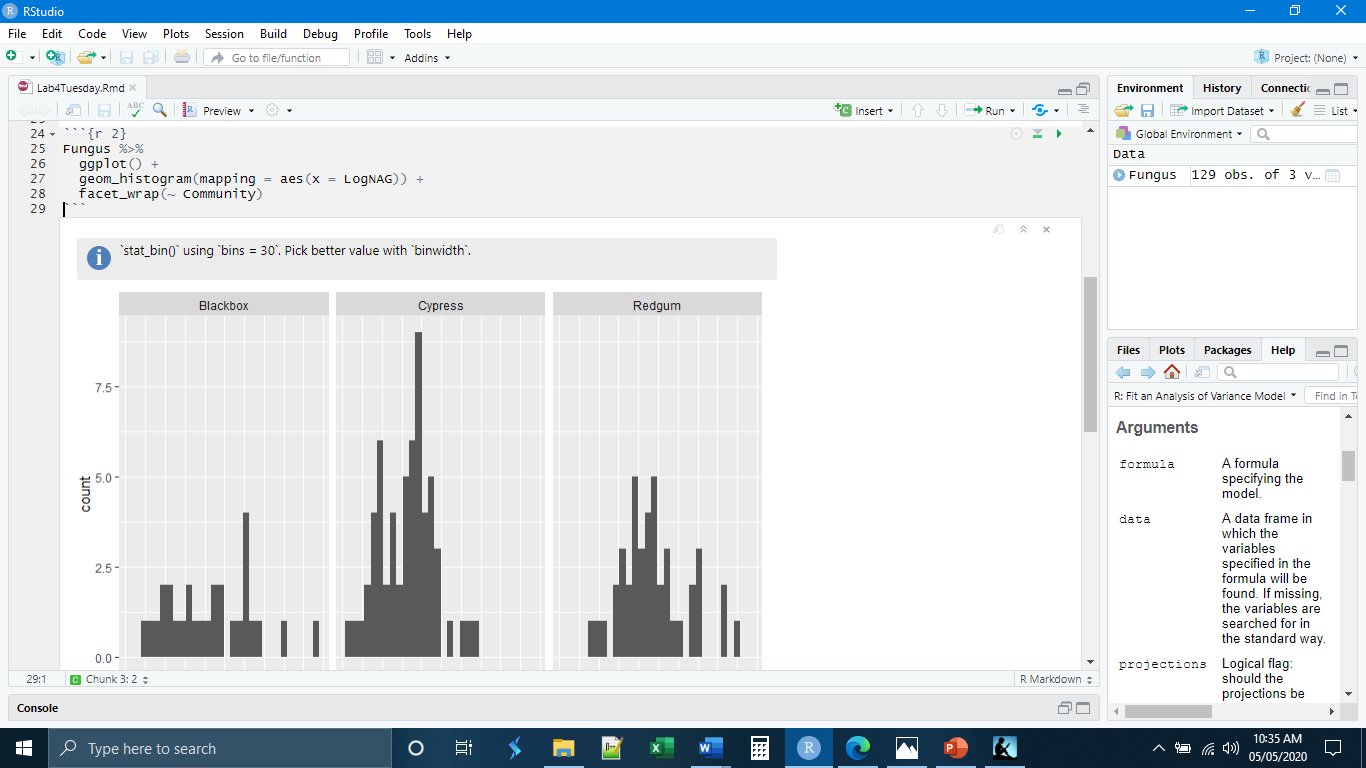
LogNAG)



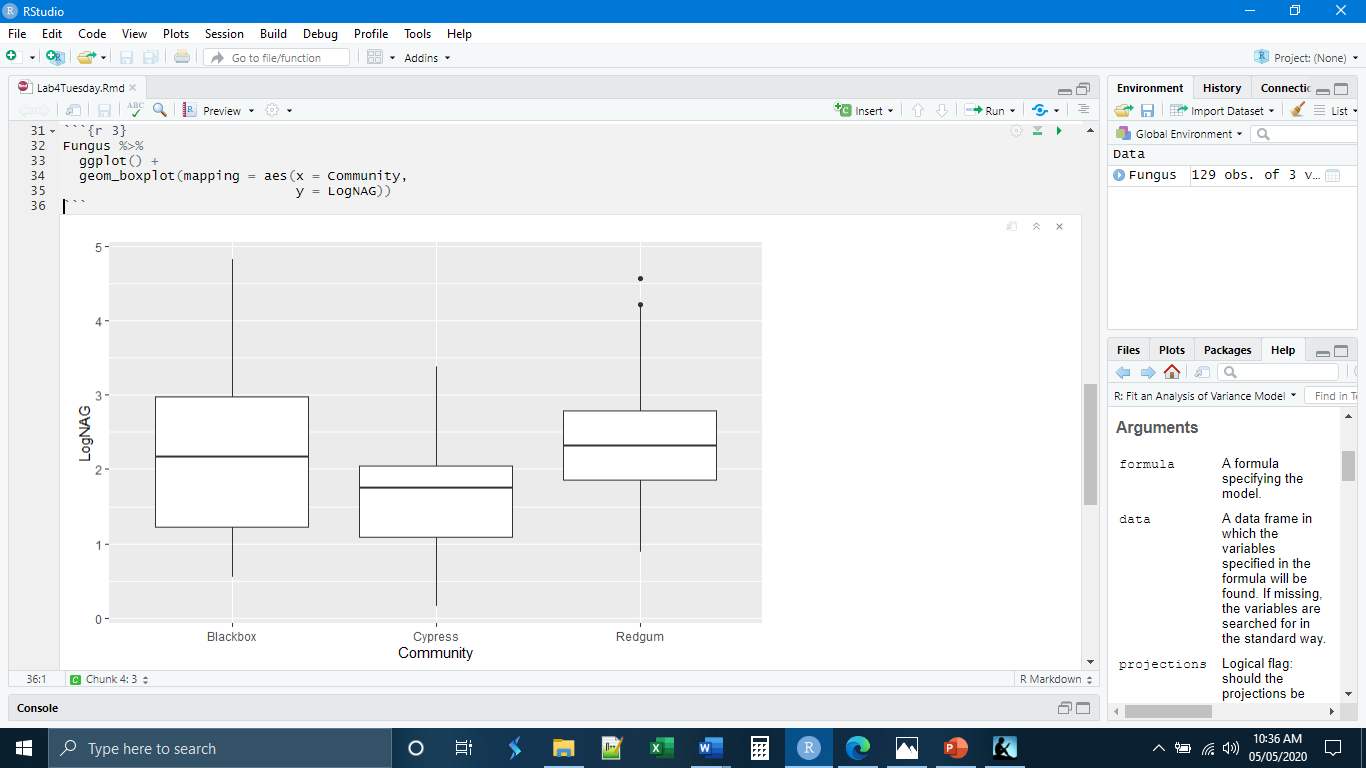
*Checking Assumptions*

As we have discussed in lecture for both the t-Test and ANOVA, there are assumptions that need to be met in order for certain statistical tests to be valid. For ANOVA, we have the standard assumption of the independence of observations. We will assume this is true as we have no way to assess this. We will trust the researchers. In addition to this, ANOVA requires that the distributions of the groups be normal and that the variances of the groups be equal. Having worked with these data last week in our lab on transformations, we know that the NAG variable is not normally distributed and that we need to correct for it (which we did in our dplyr pipe above).

To remind ourselves that these data are broadly normally distributed we can make histograms. Again, it is not perfect. Blackbox, for example, looks pretty flat. Our sample size is large, so we can get away with less-than-perfect normality.



There are multiple ways that we can assess homoscedasticity (evenness of variance), including specific tests. It is also possible to assess evenness of variance visually. We can look to see that the width of our histograms is even. We can also make boxplots and look to see if the widths of the boxes/whiskers are even among the groups. Our data appear to be in good shape.



*One-Way ANOVA*

Our assumptions are met close enough to do our ANOVA. R has a base function that will do the ANOVA. This function is not inside the tidyverse and will involve some subsetting. The first thing that we need to do, however, is to state our hypotheses.

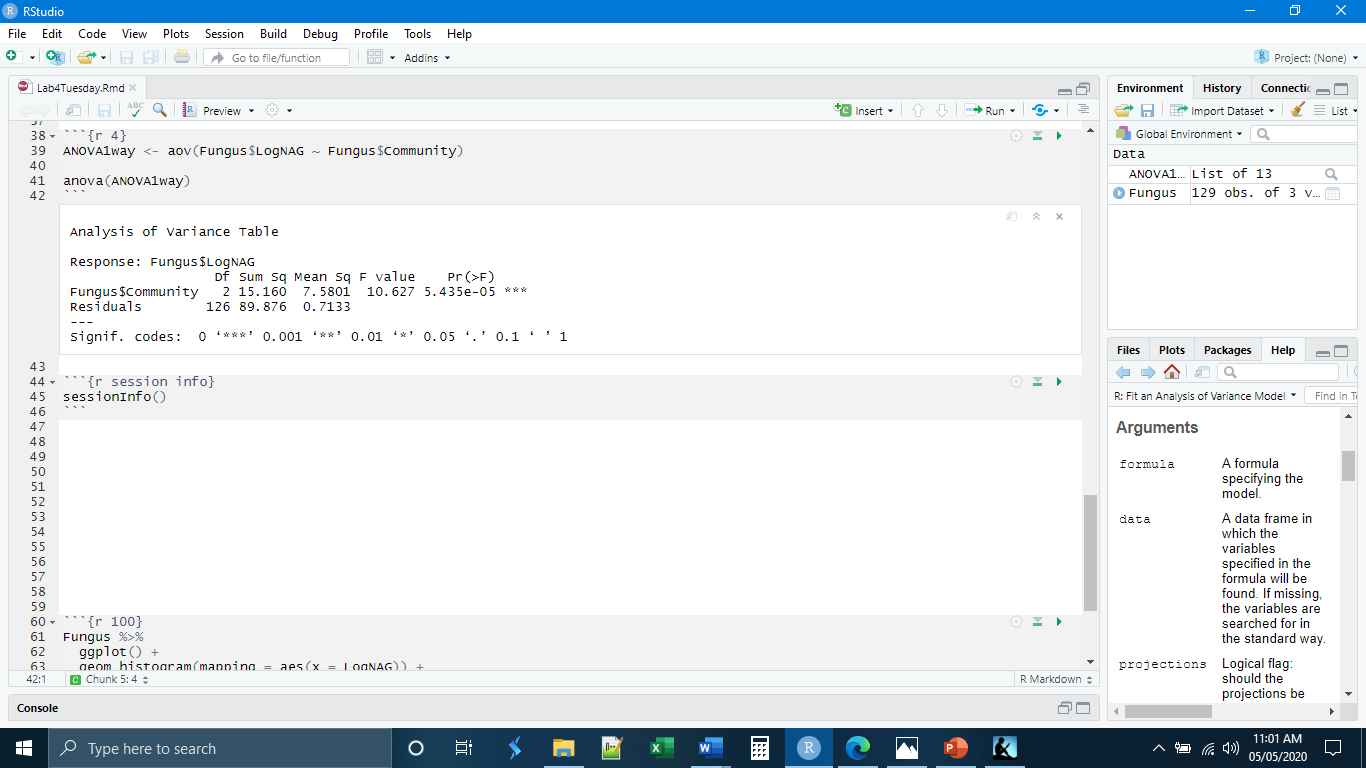
H0: There are no differences in the means among the groups. More specifically, there are no differences in the mean concentration of NAG among the different ecological communities.

HA: There are differences in the means among the groups. More specifically, the mean NAG concentrations differ among the different ecological communities.

The basic syntax requires two lines of code and an intermediary data object :

SavedAnovaFile <- aov(DataObject$ResponseVariable ~ DataObject$GroupVariable)

anova(SavedAnovaFile)

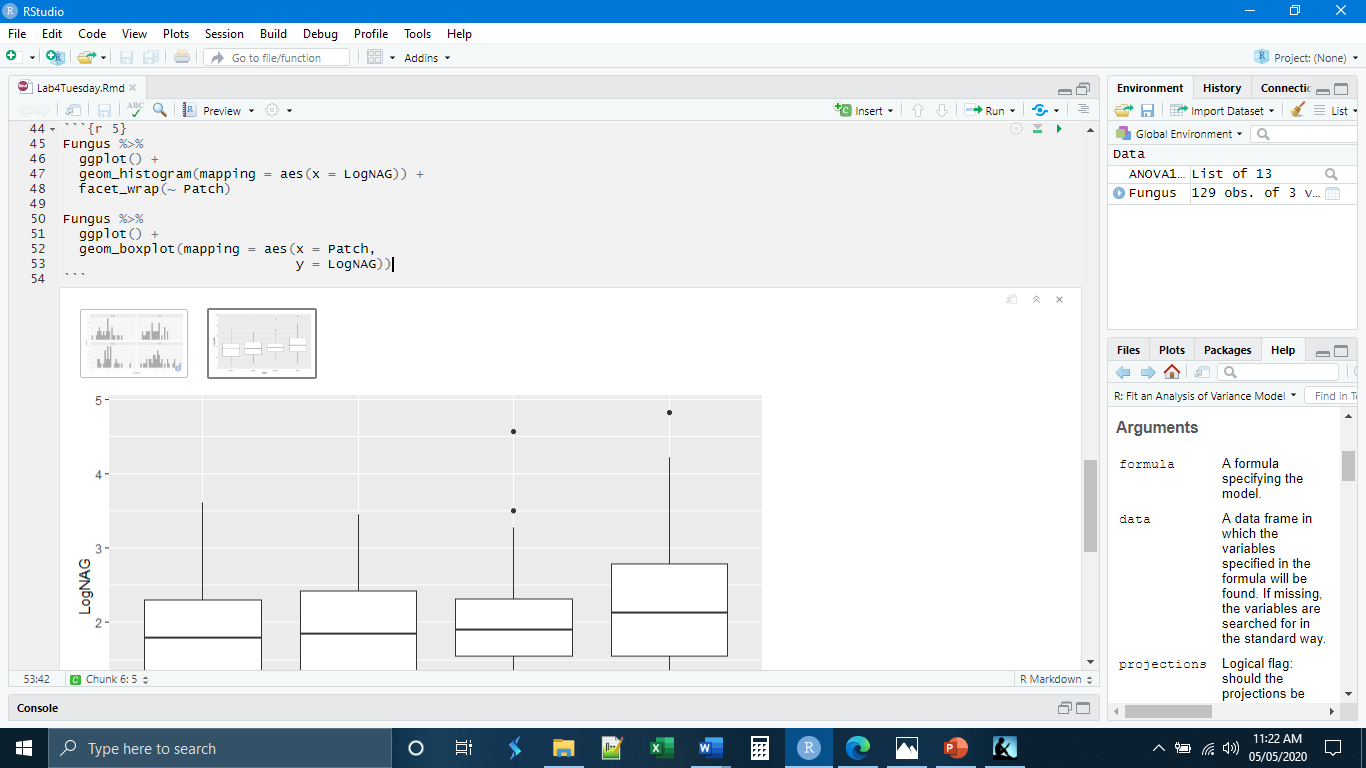


Note that the output that R gives is not exactly what we are used to. The final row of totals (total degrees of freedom and total sum of squares) is missing. These could be calculated if desired. Otherwise, the table looks familiar.

Remember that the interpretation of the ANOVA result is not specific to any pairs of groups. In this case we would accept our alternative hypothesis that there is a difference in means among the groups. We cannot say any more, for example, which groups have different means. We also have to take into consideration here that we did a data transformation, so the interpretation is a little more obscure, as we discussed last week.

*Two-Way ANOVA*

In terms of R code, running a two-way ANOVA is similar, the only difference is that we have to define a second grouping variable. Before we can do that, however, we need to make a quick check to see that our new grouping variable, in this case the patches variable, also has normal distributions of data and evenness of variances.



Again, our distributions look a little flat, but our larger sample size will help to obscure that. Our variances look to be about even. Before actually doing the test, we need to write out our hypotheses. This is a two-way ANOVA. We are testing differences in the mean over two grouping variables. As a result, we need two sets of hypotheses.

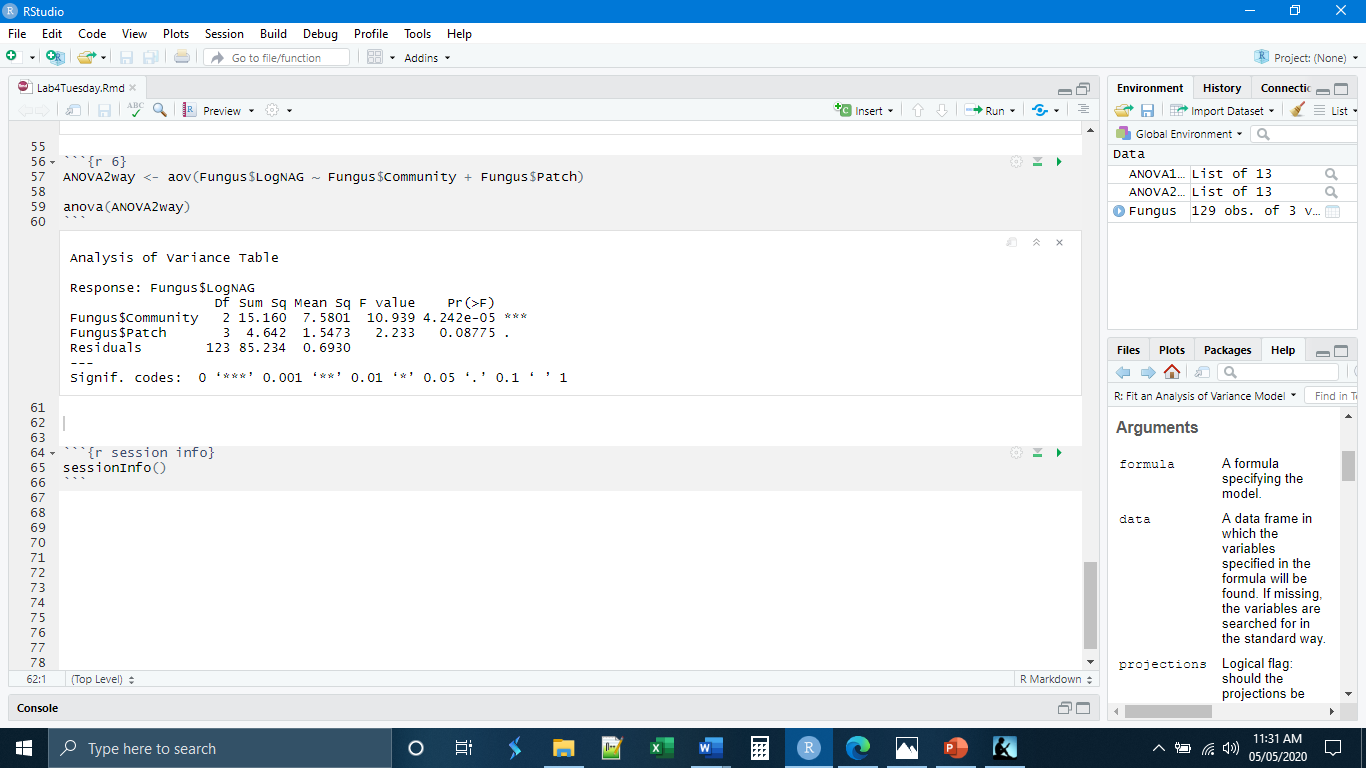
H0: The mean NAG concentration is the same among all the communities.

HA: The mean NAG concentration is different among the different communities.

H0: The mean NAG concentration is the same among all the patch types.

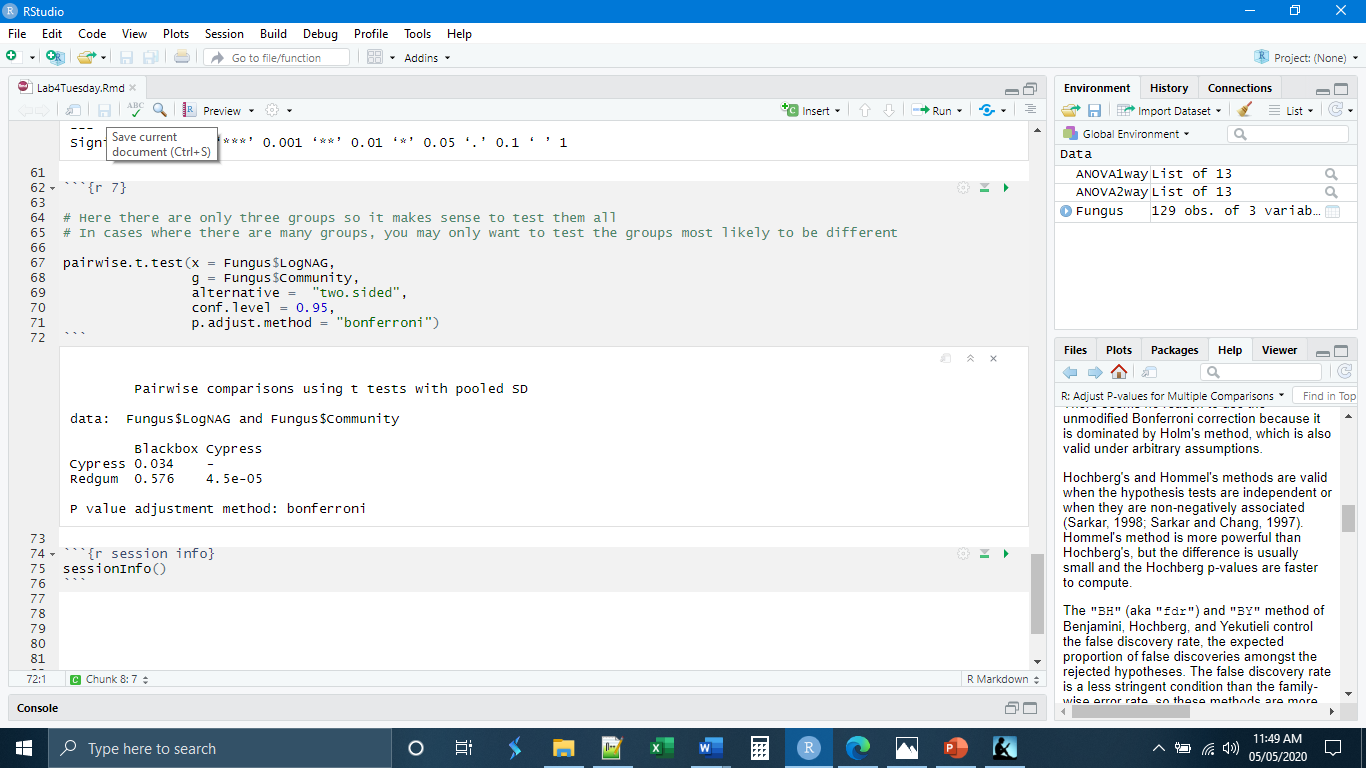
HA: The mean NAG concentration is different among the different patch types.

The two-way ANOVA results indicate that there is a difference in the means among the communities (which matches what we observed in the one-way ANOVA result), but that there is not a difference in the means among the different patches.



*Multiple Tests*

If the results of our ANOVA indicate that there are differences among the groups, our next question will naturally be, which groups are different? The way to determine which groups are different would be through multiple t-Tests. While it is tempting to look at the ANOVA and then decide which means you want to test against each other, methodologically, it is better to decide beforehand which groups you will compare using multiple comparisons. In our specific case here, there are only three groups in the community variable, and we can simply make all of the comparisons. If our data set were different and there were many groups, it would probably be advisable to only make a few comparisons. In that case, you would want to identify which comparisons you wanted to make prior to doing the ANOVA.



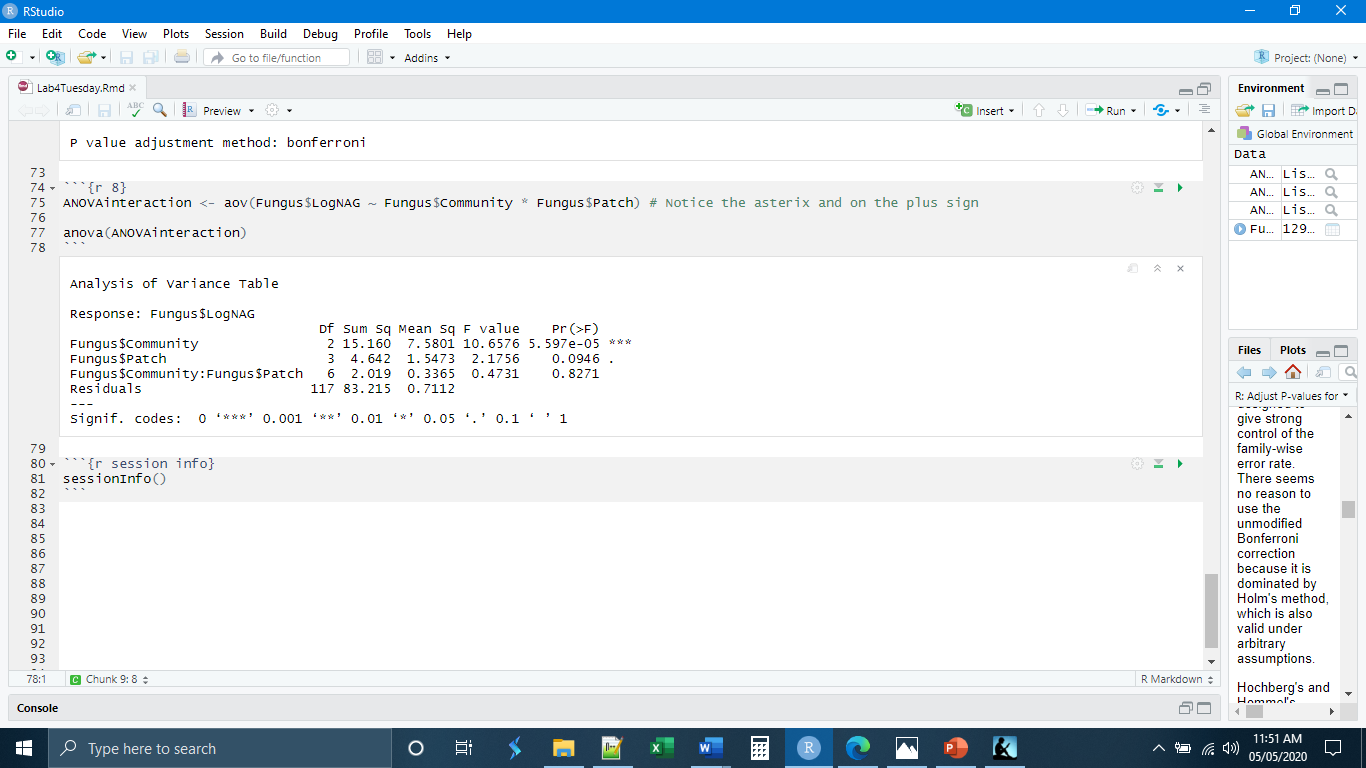
*Testing the Interaction Term*

In biology, different aspects of a system may interact to produce unexpected results. For example, what if we were trying to determine the factors that were contributing to the survival of a particular species in a particular ecosystem. We might find that increased rainfall increased the survival of this organism by 20%. We might also find that increased temperature promotes the survival of this organism by 15%. Each factor increases the survival of the organism. We would expect that if the organism experienced both increased rainfall and increased temperature that its survival probability would increase by 35% (20 +15). What if, instead we found that increased rain and temperature actually increased survival by 60%? This would suggest that the two factors are interacting in some way to produce an increased effect on the organism. Or, another example. Say drug A decreases blood pressure by 10% and drug B decreases blood pressure by 5%, but together they decrease blood pressure by 50%. Here too it appears that the two factors (drugs) are interacting to produce an effect beyond what would be predicted. One of the advantages of using an ANOVA over a series of pairwise t-Tests is that ANOVA is actually capable of testing the interaction of two factors. This can only be done if you have multiple observations in each of the group combinations because we need to be able to account for variation within the groups.

Doing this in R only requires changing the + to a \* in the code for a two-way ANOVA. We now have to form an additional hypothesis, however. In the case of our test, that would be:

H0: There is no interaction between the community and the patch type.

HA: There is an interaction between the community and the patch type.



Notice that our result has changed slightly. This is because our residual sum of squares has changed. Part of that residual sum of squares has been used to account for the interaction. If we interpret this result, we see that 1) mean NAG concentrations do differ among communities, 2) mean NAG concentrations do not differ among patch types, and 3) there is no interaction between the community and the patch type.

*Independent Exercises*

**1.** Make a new chunk called “chunk 3” and write code to

- Read in the Triplaris\_Lab4 data (from Blackboard) and save it as a new data object

- Conduct a one-way ANOVA with Country as your grouping variable and LeafWidth as your response variable.

- At the bottom of the chunk, in a comment or series of comments, write out the two hypotheses (null and alternative) for this ANOVA. Also write in a comment the conclusion (reject null or fail to reject) based on the ANOVA.

**2.** Make a new chunk called “chunk 4” and write code to

- Check the evenness of the variances among the four country groups in the Triplaris\_Lab4 data

(Hint: Do this graphically with boxplots as we did in class)

- Are the variances even? Record your answer as a comment in the chunk.