Lab #9: ANOVA and TUKEY tests

Objectives:

1. Column manipulation in SAS
2. Analysis of variance
3. Tukey test
4. Least Significant Difference test
5. Analysis of variance with PROC GLM
6. Levene test for Homoscedasticity

Column Manipulation

Analysis of variance is used when 3 or more different groups of observations need to be compared with each other. The best scenario is for example comparing 3 or more different treatments on a population, or 2 or more treatments and a control, in an experiment, where the number of samples is the same in each group and independence within and between is controlled.

One such example is the dataset Diets.csv which you can find on our MOODLE page and contains the results of weight loss for 75 individuals that followed 3 different programs. The question for the researcher is first, are the outputs of the programs different? If yes which one seems to be better?

First let's import our dataset by using the by now well-known proc import command as follows:

```
PROC IMPORT OUT= Diets
            DATAFILE= "put the complete path here…\Diets.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
```

Using proc print we can view the dataset and notice that it contains the observation number, the gender (0 female, 1 male) the age, the height and the diet number (1,2,3). It also contains 2 columns the pre_weight and post_weight measured in kilos.

```
PROC PRINT DATA = Diets;
RUN;
```

What we would like to have is the difference of weight before and after the diet, so we can compare those averages amongst the three diets. Thus we will create a new dataset, called DietClean that has the difference in weight, and just the type of diet the person followed, disregrading all other information. To do that we will use the same subseting tools we learned in the past, but this time we will create a new column by doing arithmetics on variables, namely:

```
DATA DietsClean;
SET Diets;
WeightDif = pre_weight-post_weight;
KEEP Diet WeightDif;
RUN;
```

Notice that we created a new set by invoking the command **DATA** followed by the name of the new dataset, in this case DietsClean. The command **SET Diets** instructs SAS to use the dataset Diets to create the new columns. The line **WeightDif = pre_weight-post_weight** creates a new variable (column) called WeightDif by subtracting the variable post_weight from the variable pre_weight. It is neat that SAS understands column-wise operations! Finally with the command **KEEP Diet WeightDif** we instruct SAS to keep only those two variables (columns) in the new dataset we created and discard the rest.

Go ahead and use proc print to see how the new dataset looks like.

```
PROC PRINT DATA = DietsClean;
RUN;
```

Before we do ANOVA, let's explore our dataset, and do summary statistics for the different diets, including the boxplots. **PROC UNIVARIATE** will do the trick once again. Always in these types of analyses start with some exploratory statistics. It helps justify the need for ANOVA, regression or other "heavy machinery"!

```
ods graphics on;
TITLE "Statistics By Diet";
PROC UNIVARIATE Data=DietsClean plots;
var WeightDif;
by Diet;
RUN;
```

As you recall, the command **by Diet** will present the summary statistics, of the variable of choice (**var** WeightDif) for the three different diets. Adding the command **plots** creates all the histograms and boxplots, including the side by side boxplot at the end. Already from that plot we can be sure that there is a difference between the diets and probably the last one is the best. But we will analyze it further to be absolutely sure. Just to be on the safe side I added the global command **ods graphics on**, in case you turned it off earlier on since we do need to see the output plots.

ANALYSIS OF VARIANCE

Once again, remember that our null and alternative hypotheses are

$$H_0: \mu_1 = \mu_2 = \mu_3$$
$$H_A: one\ of\ these\ equalities\ is\ not\ true$$

Notice that all three groups have the same amount of observations (25). This is called a "balanced" experiment, and it is preferable. ANOVA WILL work on "unbalanced groups" just fine but it is easier to showcase the validity of it if the groups have the same amount of observations. As a matter of fact, the ANOVA procedure that we will use works better if you have the same amount of observations, and the SAS programmers suggest using a different procedure, **PROC GLM**, when dealing with unbalanced experiments. We'll get to that soon. For now the following code will do the trick:

```
PROC ANOVA Data=DietsClean;
class Diet;
model WeightDif=Diet;
RUN;
```

As you can see a very easy code, and it is also easy to explain. First of all we are using the procedure ANOVA on the dataset DATA=DietsClean. Then we need to use the command **class** which provides the variable (column) that contains the different groups. In this case we have the 3 diets (1, 2, 3) found in the column Diet, hence **class Diet** is used. Then SAS wants to know what variable to analyze based on what class, basically how to create the analysis model. The syntax for that is **model** "the variable of interest" = "the variable of different classes".

A couple of tables pop up after the code is implemented. We have the general description:

Statistics By Diet

The ANOVA Procedure

Class Level Information

Class	Levels	Values
Diet	3	1 2 3

Number of Observations Read	75
Number of Observations Used	75

SAS informs us that it used all the observations (75 out of 75) and that it sees 3 levels (groups) with values 1,2,3 respectively in the column that contains the groups (in this case Diet).

This is followed by the table we analyzed in class, which is the main table you should report when doing an analysis of variance, namely:

Statistics By Diet

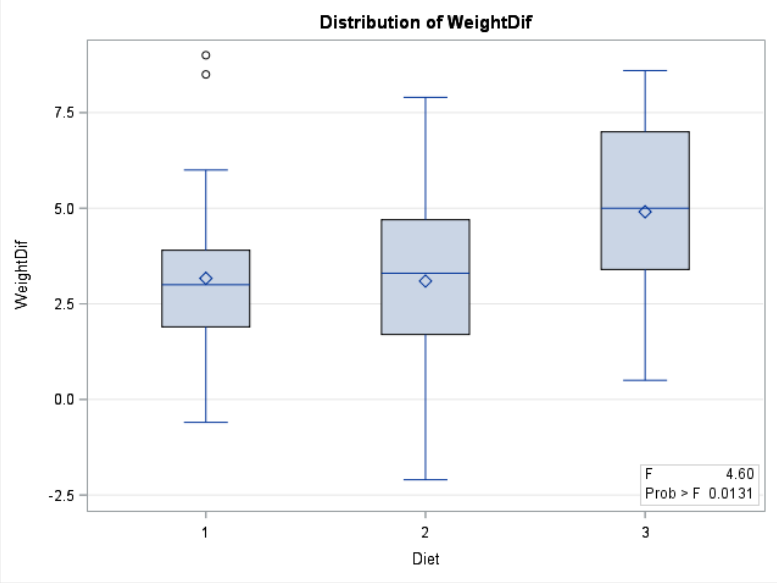
The ANOVA Procedure

Dependent Variable: WeightDif

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	52.6344000	26.3172000	4.60	0.0131
Error	72	411.6624000	5.7175333		
Corrected Total	74	464.2968000			

If we focus on the top right cell, we see that our p-value is 0.0131 which is smaller than the default alpha 0.05. So we **reject** the null hypothesis and claim that at least one of the diets has a different average weight loss than the rest.

We also get the boxplot of the distributions with the F value and the p value as a nice plot to put in our posters!



Tukey Test

Let's proceed now with the post-ANOVA analysis and start with pairwise "adjusted" comparisons, namely the Tukey test. This can be achieved by extending the ANOVA code and adding the means computation, namely:

```
PROC ANOVA Data=DietsClean;
class Diet;
model WeightDif=Diet;
means Diet/ tukey alpha=0.05;
RUN;
```

The line **means Diet/ tukey** instructs SAS to compare the means of the groups found in the variable Diet using the tukey pairwise comparison test we learned in lecture. The confidence level is indicated by the parameter **alpha=0.05.**
What we obtain is a "grouping" of means that look the same, vs groupings of means that look different. In the form of the following table:

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Diet
A	4.9080	25	3
B	3.1680	25	1
B			
B	3.0960	25	2

This implies that the average weight-loss for diets 1 and 2 is similar, therefore we group them together under the arbitrary "B" group symbol. On the other hand, Diet 3 seems different from the other two so it creates its own grouping, again arbitrarily given the letter "A".

Least Significant Difference test

Another interesting pairwise comparison we can do is that of the Least Significant Difference (LSD), or the Fisher procedure as it is known. Remember that this method again performs a T-test comparison, between the differences of all possible pairs of means using the following formula for the standard error:

$$SE = \sqrt{\frac{2 * MSW}{m}}$$

Where *MSW* is the mean square error within groups and *m* is the number of observations in each group. Then instead of outputting the p-values it outputs a symbol to indicate significant differences between the groups and a confidence interval around the pairwise difference between the means. Once again we need to augment the ANOVA code a bit to get:

```
PROC ANOVA Data=DietsClean;
class Diet;
model WeightDif=Diet;
means Diet/ lsd cldiff alpha=0.05;
RUN;
```

Once again, the line **means Diet /lsd cldiff alpha=0.05** instructs SAS to pairwise compare the means for each group found in the variable Diet by using the LSD test. **Cldiff** will output the confidence interval differences and alpha=0.05 is the user defined alpha level.

The output table looks like this:

Comparisons significant at the 0.05 level are indicated by *.**

Diet Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
3 - 1	1.7400	0.1215	3.3585	***
3 - 2	1.8120	0.1935	3.4305	***
1 - 3	-1.7400	-3.3585	-0.1215	***
1 - 2	0.0720	-1.5465	1.6905	
2 - 3	-1.8120	-3.4305	-0.1935	***
2 - 1	-0.0720	-1.6905	1.5465	

As you see all comparisons happen twice, but it is good to have both, depending on which Diet we want to put first or second in our pairwise comparisons. Again we can see that the comparison between 1 and 2 does not yield a significant difference, so once again diets 1 and 2 are almost the same. Diet 3 on the other hand creates significant differences from both 1 and 2, as we can see in the highlighted part of the table.

Analysis of Variance with PROC GLM

So far we have been working with a balanced experiment, where all the subgroups have the same number of samples. Suppose though we want to use unbalanced groups. For example, let's explore the dataset Tomato.csv found on MOODLE. In this dataset the researchers have catalogued 3 versions of the same tomato breed M82. Besides the regular one, they have created two strands A, B by selective breeding with other varieties and they have catalogued various characteristics of the roots and plants after a few weeks of cultivation.

Our question is to check if there is a difference in the average length of the roots among the three varieties which we will answer by doing an unbalanced one way ANOVA on the last column titled "length".

Let's import that dataset

```
PROC IMPORT OUT= Tomato
            DATAFILE= "put the complete path here…\Tomato.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
```

Using proc print we can see that the dataset contains 42 (!?!) observations. The main variables of interest are Genotype and Length. We could create a smaller dataset that contains only these 2 but we will proceed with all of them for now.

```
PROC PRINT DATA = Tomato;
RUN;
```

Before we analyze our data, we should sort it by genotype using the code:

```
PROC SORT DATA = Tomato;
By = Genotype;
RUN;
```

Let us try some summary statistics first by the different genotypes using the code:

```
ods graphics on;
TITLE "Statistics By Genotype";
PROC UNIVARIATE Data=Tomato plots;
var Length;
by Genotype;
RUN;
```

One of the things we get is that the three groups M82, A and B have different number of observations (13, 14 and 15 respectively). This could be due to errors in inputting, failures in crops, failures in saving the data and others. In this case the experimenter could either

a) Make all groups equal by removing the appropriate amount of observations at random from the groups with extra observations.
b) Utilize an unbalanced one way ANOVA through PROC GLM.

Here we will present how the second procedure is implemented including LSD test. As a matter of fact, the code is VERY similar to PROC ANOVA, but instead the mean procedure is PROC GLM as follows.

```
Title "Unbalanced ANOVA for Tomato varieties";
PROC GLM Data=Tomato;
class Genotype;
model Length=Genotype;
means Genotype/ lds cldiff alpha=0.05;
RUN;
```

We basically use PROC GLM (general linear model) instead of PROC ANOVA. The procedure GLM is very useful and can be thought of as model exploration which includes ANOVA as one of its procedures. More on that later.

Let's explore the tables now

Class Level Information

Class	Levels	Values
Genotype	3	Line A Line B M82

Number of Observations Read	42
Number of Observations Used	42

This table gives us general information about our model. We have three different groups, line a, line b and M82 and we used all 42 observations in the analysis. The familiar table:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	25918.388	12959.194	0.45	0.6425
Error	39	1129472.088	28960.823		
Corrected Total	41	1155390.476			

follows which contains the p value. Since that is large, we fail to reject the null hypothesis and conclude that there does not seem to be a difference between the average lengths of roots in the three varieties.

The pairwise comparisons using the Fishers LSD test also give no significant pairs (no *** in any comparison) as we can see here:

Comparisons significant at the 0.05 level are indicated by ***.				
Genotype Comparison	**Difference Between Means**	**95% Confidence Limits**		
Line A - M82	25.54	-104.90	155.97	
Line A - Line B	59.71	-68.20	187.63	
M82 - Line A	-25.54	-155.97	104.90	
M82 - Line B	34.18	-98.41	166.76	
Line B - Line A	-59.71	-187.63	68.20	
Line B - M82	-34.18	-166.76	98.41	

Levene test for Homoscedasticity:

To test if the variances for the different groups are the same of statistically different we can use again proc glm to produce the Levene test. The code is similar as before namely:

```
Title "Anova with Levene test for Tomato varieties";
PROC GLM Data=Tomato;
class Genotype;
model Length=Genotype;
means Genotype/ HOVTEST alpha=0.05;
RUN;
```

What was added was the command HOVTEST after the means, which instructs SAS to run a Levene test after finishing with the ANOVA computations.
After getting again all the information about the ANOVA tables as before we also get the table:

Levene's Test for Homogeneity of Length Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Genotype	2	1.198E10	5.9899E9	3.02	0.0605
Error	39	7.745E10	1.9859E9		

As we can see the p value is slightly larger than 0.05, so we fail to reject the null hypothesis and conclude (*marginally*) that the variances amongst the three groups are probably the same.