Laboratory Topic 3

- General format of ANOVA in SAS
- Testing the assumption of homogeneity of variances
  - by "/hovtest"
  - by ANOVA of squared residuals
  - by Analyst
- One-way ANOVA of nested design
  - by Proc Nested
  - by Proc GLM
  - by Analyst
- Obtaining and interpreting Components of Variance
  - by Proc VarComp

ANOVA in SAS

The primary SAS procedures for analysis of variance are Proc ANOVA and Proc GLM (General Linear Model). Proc ANOVA assumes that the dataset is "balanced" (i.e. that it has the same number of replications r for each treatment), though it can be used for single factor ANOVA even if the design is unbalanced. Proc GLM is more general and can be used for data not meeting this restriction; this is the procedure we will be using.

Proc GLM has two required statements: Class and Model.

CLASS: With the Class statement, you declare all the classification variables used in the linear model. All classification variable appearing in the model must first be declared here. The syntax is:

```
Class variables;
```

MODEL: With the Model statement, you use the declared classification variables to build an explanatory linear model for the response variable (explained below). The syntax is:

```
Model dependent variable = independent effects;
```

The Model statement is a minimalist representation of a general linear effects model, where a response (i.e. dependent variable) is explained by a host of known additive deviations from a base mean:

\[ Y_i = \mu + \kappa_1 + \kappa_2 + \ldots + \kappa_n + \epsilon_i \]

For a completely randomized design (CRD) with one treatment variable, the independent effect is simply the one classification variable and the dependent variable is simply the response. But there can be more complex models:

```
Model y = a b ;  -> main effects
Model y = a b a*b ;  -> interaction
```
Example 3.1

In this experiment, the nitrogen fixation capacities of six different rhizobia on clover are compared. The experiment is arranged as a CRD with 6 treatments (i.e. 6 different rhizobia strains) and five independent replications per treatment.

Data Clover;
  Input Culture $ Nlevel;
Cards;
  3DOk1 24.1  3DOk1 19.1  3DOk5 17.9  3DOk7 20.7  3DOk13 14.3  Comp 17.3
  3DOk1 32.6  3DOk5 24.8  3DOk4 16.5  3DOk7 23.4  3DOk13 14.4  Comp 19.4
  3DOk1 27    3DOk5 26.3  3DOk4 10.9  3DOk7 20.5  3DOk13 11.8  Comp 19.1
  3DOk1 28.9  3DOk5 25.2  3DOk4 11.9  3DOk7 18.1  3DOk13 11.6  Comp 16.9
  3DOk1 31.4  3DOk5 24.3  3DOk4 15.8  3DOk7 16.7  3DOk13 14.2  Comp 20.8
;
Proc GLM;
  Class Culture;
  Model Nlevel = Culture;  * Response variable = Class variable;
  Means Culture;  * Gives us the means and stdevs of the treatments;
Proc Plot;
  Plot Nlevel*culture;  * Generates plot of NLevel (y) vs. Culture (x);
Run;
Quit;

Results of the GLM procedure:

Dependent Variable: Nlevel

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>778.5249138</td>
<td>155.7049828</td>
<td>22.87</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>23</td>
<td>156.6075000</td>
<td>6.8090217</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>28</td>
<td>935.1324138</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square Coeff Var Root MSE Nlevel Mean

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culture</td>
<td>5</td>
<td>778.5249138</td>
<td>155.7049828</td>
<td>22.87</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

**Interpretation**  Recall that the null hypothesis of an ANOVA is that all means are equal (H₀: μ₁ = μ₂ = ... = μₙ) while the alternate hypothesis is that at least one mean is not equal (H₁: μᵢ ≠ μⱼ). With a p-value of less than 0.0001, we soundly reject H₀. A significant difference exists among the treatment means. Notice that the R-Square value is simply a measure of the amount of variation explained by the model:

\[ R\text{-Square} = \frac{778.5249138}{935.1324138} = 0.832529 \]
Sums of Squares (SS) When you run an ANOVA, SAS will produce two different tables by default, one featuring "Type I SS," the other featuring "Type III SS." These Types have nothing to do with Type I and Type II errors. We'll discuss these SS Types later in the term; for now, just obey the following:

1. Always report Type III SS when performing ANOVAs.
2. Always report Type I SS when performing regressions.

Testing the Assumption of Homogeneity of Variances

When you perform an ANOVA, you assert that the data under consideration meet the assumptions for which an ANOVA is valid. We briefly discussed normality last week and will revisit it more formally next week. Now we will cover a second assumption, that the variances of all compared treatments are homogeneous.

…by / HovTest

You can add the option "/ HovTest = Levene" to the Means statement following the Proc GLM. Levene’s test is an ANOVA of the squares of the residuals (i.e. deviations of the observations from their expected values based on the linear model; in this case, this means deviations of the observation from their respective treatment means \((Y_{ij} - \bar{Y}_i)^2\)). To see how this works, modify the Means statement in Example 3.1 to read as follows:

```
Proc GLM;
  Class Culture;
  Model Nlevel = Culture;
  Means Culture / HovTest = Levene;
```

Result of Levene's Test:

```
Levene's Test for Homogeneity of Nlevel Variance
ANOVA of Squared Deviations from Group Means

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culture</td>
<td>5</td>
<td>227.4</td>
<td>45.4814</td>
<td>1.10</td>
<td>0.3868</td>
</tr>
<tr>
<td>Error</td>
<td>23</td>
<td>950.0</td>
<td>41.3047</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Interpretation H₀ for Levene's Test is that the variances of the treatments are homogeneous, and Hₐ is that they are not homogeneous. With a p-value of 0.3868, we fail to reject H₀; thus we have no evidence, at 95% confidence level, that the variances are not homogeneous.
...by ANOVA of Squared Residuals

As stated above, the "/ HovTest = Levene" option performs an ANOVA of the squared residuals. To see that this is the case, we can program SAS the following way:

Example 3.2

```
Proc GLM;
   Class Culture;
   Model Nlevel = Culture;
   Means Culture;
   Output Out = Residual R = Res1 P = Pred1;
   * Creates data set called Residual that includes all the variables as before plus two new ones: Res1 and Pred1.
   R tells SAS to calculate the Residuals and we name that new variable Res1.
   P tells SAS to calculate the Predicted and we name that new variable Pred1;
Proc Print Data = Residual; * Take a look at this data set Residual;
Proc Plot Data = Residual; * It is now necessary to specify the data set;
   Plot Nlevel*Culture;
   Plot Res1*Pred1;
Data Levene; * Creates new data set called Levene (or any name you like);
   Set Residual; * Imports the data from the data set Residual;
   SqRes = Res1*Res1; * Adds to it a new variable called SqRes that is the square of the residuals;
Proc Print Data = Levene; * Take a look at this data set Levene;
Proc GLM Data = Levene; * Perform an ANOVA of the variable SqRes;
   Class Culture;
   Model SqRes = Culture;
Run;
Quit;
```

Results of the GLM Procedure:

Dependent Variable: SqRes

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culture</td>
<td>5</td>
<td>227.4071147</td>
<td>45.4814229</td>
<td>1.10</td>
<td>0.3868</td>
</tr>
</tbody>
</table>

It is no coincidence that we obtain the same p-value (0.3868) as in the previous case where we used the /HovTest command!!
...by Analyst

To use Analyst to test homogeneity of variances and obtain power and sample size:

1. Start the Analyst Application
2. Import the data set "Clover" from the previous SAS session
3. Statistics → ANOVA → One-Way ANOVA → click CULTURE and then INDEPENDENT → click Nelevel and then DEPENDENT
4. Select the desired tests: TESTS → 
   b. Under "Power Analysis," select Perform power analysis;  
      under "Sample size," input from 4 to 30 by 2. 
5. PLOTS → Select Means Plot (specify Standard Error of Mean) and both Residual Plots
6. Run; save and analyze output.

One-Way ANOVA of Nested Design

...by Proc Nested

Though you will need to learn how to analyze nested designs using Proc GLM, we'll begin here by using the simpler Proc Nested, a special procedure in SAS that can be used only for purely nested CRDs (again, for more complicated designs, you will need to use Proc GLM). The following example will introduce this procedure as well as remind you of another way to input data (see Appendix of Lab 1):

Example 3.3  

In this experiment, the effects of different treatments on the growth of mint plants are being measured. It is a purely nested CRD with six levels of treatment, three replications (i.e. pots) per level, and four subsamples per replication (i.e. plants).

```sas
Title 'Nested Design';
Data Mint;
  do Trtmnt = 1 to 6;
    do Pot = 1 to 3;
      do Plant = 1 to 4;
        Input Growth @@; * The @@ says the input line has values;
        Output; * for more than 1 observation;
      end;
    end;
  end;
Cards;
3.5 4.0 3.0 4.5
2.5 4.5 5.5 5.0
3.0 3.0 2.5 3.0
5.0 5.5 4.0 3.5
3.5 3.5 3.0 4.0
```

Reads Trtmnt 1, Pot 1, Plant 1, Growth = 3.5
Then Trtmnt 1, Pot 1, Plant 2, Growth = 4.0
And so on...

Next line starting at 2.5 represents
Trtmnt 1, Pot 2, Plant 1, Growth = 2.5... Etc.
Proc Print; * Always use this when you input data as above to;  
   ID Trtmt; *be sure SAS read it correctly;
   Var Plant Pot Growth;

Proc Nested; * Use Proc Nested only with one-way ANOVA's;  
   Class Trtmt Pot; * List classes in hierarchical order;
   Var Growth; * Then use 'Var' command to identify the response variable;
Run;
Quit;

With Proc Nested, the only things different from Proc GLM are:

1) The Class statement must have the treatments listed in hierarchical order (high → low)
2) Instead of a Model statement, there is a Var statement listing the response variable

Results of the Nested Procedure:

Coefficients of Expected Mean Squares

<table>
<thead>
<tr>
<th>Source</th>
<th>Trtmt</th>
<th>Pot</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trtmt</td>
<td>12</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Pot</td>
<td>0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Error</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Nested Random Effects Analysis of Variance for Variable Growth

<table>
<thead>
<tr>
<th>Variance Source</th>
<th>Sum of Squares</th>
<th>F Value</th>
<th>Pr &gt; F</th>
<th>Term</th>
<th>Mean Square</th>
<th>Variance Component of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>255.913194</td>
<td></td>
<td></td>
<td></td>
<td>3.604411</td>
<td>4.053356</td>
</tr>
<tr>
<td>Trtmt</td>
<td>179.642361</td>
<td>16.69</td>
<td>&lt;.0001</td>
<td>Pot</td>
<td>35.928472</td>
<td>69.4398</td>
</tr>
<tr>
<td>Pot</td>
<td>25.833333</td>
<td>2.30</td>
<td>0.0186</td>
<td>Error</td>
<td>2.152778</td>
<td>7.5169</td>
</tr>
<tr>
<td>Error</td>
<td>50.437500</td>
<td></td>
<td></td>
<td></td>
<td>0.934028</td>
<td>23.0433</td>
</tr>
</tbody>
</table>

Growth Mean 5.78472222
Standard Error of Growth Mean 0.70640396
Things to notice:

- The variance components add up to the total, giving you insight into the % variation due to each.
- The correct F value for treatment is 16.69 (MS-Trtmt / MS-pot), not 38.47 (MS-Trtmt / MSE).

... by Proc GLM

The example below illustrates the steps necessary for achieving the same result using Proc GLM. Again, while it is more complicated, it is important to understand because Proc Nested is only for Completely Randomized Designs (CRDs). For more complicated designs, you will need to use Proc GLM.

Example 3.4

```
Proc GLM;
Class Trtmt Pot;  * Plant not included because it is the error term;
Model Growth = Trtmt Pot(Trtmt);  * As an ID variable, 'Pot' only has meaning once you specify the "Treatment";
Random Pot(Trtmt);  * Must declare "Pot" as a random variable;
Test H = Trtmt E = Pot(Trtmt);  * Here we request a customized F test with the correct error term
Hypothesis H = Trtmt, Error E = Pot(Trtmt);
Proc VarComp Method = Type1;  * To obtain VARiance COMPonents analysis;
Class Trtmt Pot;
Model Growth = Trtmt Pot(Trtmt);
Run;
Quit;
```

Results of the GLM Procedure:

```
Dependent Variable: Growth

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>17</td>
<td>205.4756944</td>
<td>12.0868056</td>
<td>12.94</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>54</td>
<td>50.4375000</td>
<td>0.9340278</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>71</td>
<td>255.9131944</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trtmt</td>
<td>5</td>
<td>179.6423611</td>
<td>35.9284722</td>
<td>38.47</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pot(Trtmt)</td>
<td>12</td>
<td>25.8333333</td>
<td>2.1527778</td>
<td>2.30</td>
<td>0.0186</td>
</tr>
</tbody>
</table>
```

This is the *incorrect F test for Trtmt* because it is using the wrong error term. By default, SAS computes F values using the residual MS as the error term (0.934 in this case). The TEST statement, however, allows you to request additional F tests using other error terms. "H" specifies which effects in the preceding model are to be used as hypothesis (numerator). "E" specifies one and only one effect to be used as the error term (denominator).
Results of the TEST statement:

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trtmt</td>
<td>\text{Var(\text{Error}) + 4 Var(\text{Pot(Trtmt)}) + Q(Trtmt)}</td>
</tr>
<tr>
<td>Pot(Trtmt)</td>
<td>\text{Var(\text{Error}) + 4 Var(\text{Pot(Trtmt)})}</td>
</tr>
</tbody>
</table>

Dependent Variable: Growth

Tests of Hypotheses Using the Type III MS for Pot(Trtmt) as an Error Term

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trtmt</td>
<td>5</td>
<td>179.6423611</td>
<td>35.9284722</td>
<td>16.69</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

This is the correct F test for treatment since it uses MS-pot as the error term!

Also note the analysis of components of variance similar to the one done by Proc Nested:

<table>
<thead>
<tr>
<th>Variance Component</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var(Trtmt)</td>
<td>2.81464</td>
</tr>
<tr>
<td>Var(Pot(Trtmt))</td>
<td>0.30469</td>
</tr>
<tr>
<td>Var(Error)</td>
<td>0.93403</td>
</tr>
</tbody>
</table>

It is worth noting that the same F value will result if you simply consider the averages of the subsamples, though the MS's will be different. Again, if you are not concerned with the components of the variance, there is no need to carry out a nested analysis. Just average the subsamples and treat it as an un-nested experiment.

… by Analyst

To carry out an ANOVA of a nested design using Analyst, follow these steps:

1. Start the Analyst Application
2. Import the data set "Mint" from the previous SAS session
3. Statistics \rightarrow ANOVA \rightarrow Linear Models \rightarrow Select \textit{Growth} as Dependent variable, \textit{Trtmt} and \textit{Pot} as Class variables
4. Model \rightarrow Select \textit{Trtmt} and \textit{Pot} from the Independent Variables and click Add to add them to the model as main effects
5. To create the nested effect of Pot(Trtmt), select \textit{Trtmt} from the Independent variables and select \textit{Pot} from Effects in model \rightarrow Click Nest \rightarrow OK.
6. Tests \rightarrow Select \textit{Pot(Trtmt)} from the effects \rightarrow Click \textit{Error} \rightarrow Select \textit{Trtmt} \rightarrow Click Add (the other box will then show: H=TRTMT E=POT(DESIGN)) \rightarrow OK \rightarrow OK.
Results from Analyst:

Dependent Variable: Growth

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>17</td>
<td>205.4756944</td>
<td>12.0868056</td>
<td>12.94</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>54</td>
<td>50.4375000</td>
<td>0.9340278</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>71</td>
<td>255.9131944</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square 0.802912  Coeff Var 16.70696  Root MSE 0.966451  Growth Mean 5.784722

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trtmt</td>
<td>5</td>
<td>179.6423611</td>
<td>35.9284722</td>
<td>38.47</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pot(Trtmt)</td>
<td>12</td>
<td>25.8333333</td>
<td>2.1527778</td>
<td>2.30</td>
<td>0.0186</td>
</tr>
</tbody>
</table>

Tests of Hypotheses Using the Type III MS for Pot(Trtmt) as an Error Term

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trtmt</td>
<td>5</td>
<td>179.6423611</td>
<td>35.9284722</td>
<td>16.69</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

And once again, we've obtained the correct F value for Trtmt.

As you see in this example, it is possible to use the Model dialogue box in Analyst to define the effects to be included in the model, as well as their relationships to one another. You can select standard models with a simple click or build custom models with the Add, Cross, Nest, Factorial, and Polynomial buttons. We'll see much more of this as the quarter progresses.
Appendix

For those of you interested, here's another way you could have coded the data input routine for the first example (useful depending on how you're presented the dataset):

```sas
Data Clover;
    Do Culture = 1 to 6;
        Do Rep = 1 to 5;
            Input Nlevel @@;
            Output;
        End;
    End;
Cards;
24.1 32.6 27.0 32.1 33.0
17.7 24.8 27.9 25.2 24.3
17.0 19.4 9.1 11.9 15.8
20.7 21.0 20.5 18.8 18.6
14.3 14.4 11.8 11.6 14.2
17.3 19.4 19.1 16.9 20.8
;
```

Or if you wished to keep the names of the treatments in:

```sas
Data Clover;
    Do Culture = '3Dok01', '3Dok05', '3Dok04', '3Dok07', '3Dok13', 'Comp00';
        Do Rep = 1 to 5;
            Input Nlevel @@;
            Output;
        End;
    End;
Cards;
24.1 32.6 27.0 32.1 33.0
17.7 24.8 27.9 25.2 24.3
17.0 19.4 9.1 11.9 15.8
20.7 21.0 20.5 18.8 18.6
14.3 14.4 11.8 11.6 14.2
17.3 19.4 19.1 16.9 20.8
;
```