# ST 512 - Lab 1 - The basics of SAS

## What is SAS?

SAS is a programming language based in C. Most data manipulation and is done in the DATA step while data analysis is done in steps called PROCs (short for procedures). For instance, to do a correlation analysis there is PROC CORR. Today we will start with the basics: The SAS interface, reading in data, and running a few procedures.

## The SAS Interface

There are five main windows in the SAS environment:

1. *The Program Editor Window*: This is where you spend most of your time working in SAS, writing your program in the editing window. Note: SAS commands and variable names are not case sensitive.
2. *The Log Window*: Once you execute your program, SAS will report back to you in this window. ERROR messages, WARNING messages, and NOTES about your program will appear here.

   Don't underestimate the importance of this window – remember to look here each and every time you execute a SAS program!
3. *The Output/Results Viewer Window*: The output requested in your SAS program will appear in one of these windows, depending on your personal settings. (The default location depends on the version of SAS you are using.)

   Remember that output may be generated even when errors are present in your program! This is why it is important to always check the log window!
4. *The Results Window*: The output is listed by section here. Click on an item and you are taken to that place in the output.
5. *The Explorer Window*: The explorer window allows you to navigate your SAS libraries and their contents (e.g. data sets).

## A Few Notes

- All SAS programs in ST 512 must contain a header that shows: your name, your lab section, the date the file was created, the purpose of the file.
- Typically each statement in SAS begins with a *keyword*. E.g. DATA, or PROC, or OPTIONS.
- Each statement in SAS must end with a semicolon. E.g. PROC PRINT;
- It is good practice to always explicitly state the data set you want to use. E.g. PROC PRINT DATA = Example_01;

# Reading in Data

There two main methods for reading data into SAS:

1. Use SAS code that allows you to type the data directly into SAS. (This is called *instream data*.)
2. Use SAS code (or the IMPORT Wizard) to import the data from an external file. The code used depends on the type of file and its properties. E.g a text file with no headers requires different code than an excel spreadsheet with multiple header rows. (Side note: never use multiple header rows. And don't merge cells!)

Let's get started!

1. Instream Data:
   (a) Go to the Moodle page. Here you will find a file called Cloverdata.txt. You'll need to copy and paste this data as shown below.
   (b) Open SAS and locate the Editor window.
   (c) To create data in SAS we use the DATA step as shown in the code below:

```
data clover;
    input strain $ nitrogen;
    cards;
    <copy and paste your data from cloverdata.txt here>
    ;
run;
```

   (d) This first line creates a data set named CLOVER.
   (e) The INPUT statement names the two variables: STRAIN and NITROGEN
   (f) The dollar sign is used to let SAS know the variable STRAIN is categorical and not numeric
   (g) The CARDS statement tells SAS that instream data is coming next
   (h) Finally, the RUN statement tells SAS that the DATA step is complete.
   (i) Press F8 (or use the running man) to submit your code.
   (j) **As always**, check your log first! Next use the explorer window to confirm the data set looks as expected.
   (k) Alternatively, you can use the following code to place the data set in the output window:

```
proc print data = clover;
run;
```

2. Import the data (using the Wizard):

   (a) Go to the Moodle page. Download the Firm.xls file (make sure you know where it is saved!)

   (b) Go to File $\rightarrow$ Import Data.

   (c) The import wizard will pop up. You can choose a standard source from the drop down menu. Choose the Microsoft Excel Workbook option.

   (d) Hit next and browse to the location of the file.

   (e) Hit ok, now type in the name of the data set you want to create (call this data set FIRM).

   (f) Hit next, SAS will ask if you want to save the commands for importing the file. Hit browse and find the folder where you would like to save the file, type in the file name and hit save. (This gives you a SAS program that shows you how the Import Wizard works. This is very helpful for learning SAS and moving away from using the Import Wizard.)

   (g) Finally, hit finish.

   (h) Check your log to see if there are any errors. Print the data out to check that it was read in correctly.

3. Importing the data (without the Wizard):

   (a) Find the file that contained the commands for importing a file and open it.

   (b) In the future you can use these commands to read in the data rather than using the import wizard.

   (c) Try running this code to see that the results are the same!

# Basic One-Way ANOVA analysis (this should all be review)

## Description of the data set:

The following example studies the effect of bacteria on the nitrogen content of red clover plants. The treatment factor is bacteria strain, and it has six levels. Red clover plants are inoculated with the treatments, and nitrogen content is later measured in milligrams. The data are derived from an experiment by Erdman (1946) and are analyzed in Chapters 7 and 8 of Steel and Torrie (1980).

Some questions we may want to answer from this type of data set:

1. Can we get descriptive statistics for each strain?
2. Are the means are equal at the 0.05 level?
3. Does the model adhere to the assumptions of the One-Way ANOVA model?

Since we've already read in the data, we can analyze it.

1. To answer the first question, we want to compute statistics separately for each strain. This can be done via PROC MEANS; however, we must first sort the data by strain.

```
*Sort the data so we can use the BY statement in PROC MEANS;
proc sort data = clover;
    by Strain;
run;
```

Now we can run the means procedure.

```
*Find summary statistics;
proc means data = clover;
    *Summary statistics for each STRAIN separately;
    by Strain;
    *Response variable is NITROGEN;;
    var nitrogen;
run;
```

Just viewing the summary statistics, do you think the constant variance assumption should be investigated further?

2. To answer the second question we can use PROC GLM.

```
proc glm data = clover;
    *Use STRAIN as a classification (i.e. categorical) variable;
    class strain;
    *Model the NITROGEN level based on the STRAIN;
    model nitrogen = strain;
run;
quit;
```

   (a) What is your conclusion about the equality of means at the 0.05 significance level? Explain.

   (b) What does the quantity MSE estimate?

   (c) What is the reference distribution for the F statistic?

3. Of course the p-value and conclusions are only valid if the assumptions for the model are met. We can check those assumptions by adding PLOTS = ALL in the PROC GLM statement.

```
proc glm data = clover plots=all; *will give residual diagnostic plots;
      class strain;
      model nitrogen = strain;
run;
quit;
```

   (a) To investigate the constant variance assumption, we can look at side-by-side box plots or residual vs predicted plots. A residual is the observed value minus the predicted value. For the $ij^{th}$ observation the residual is

$$r_{ij} = observed \ - \ predicted = y_{ij} - \bar{y}_{i\cdot}$$

      i. What are we looking for in the residual vs predicted value plots? Why?

   If the constant variance assumption is violated, one thing we might do is try a transformation of the data. This will be looked at in a later lab session.

   (b) To check the normality, we look at the QQ-plot (or quantile vs residual plot) and also we might inspect the histogram in the bottom left panel. (Note: the transformation idea above may also solve some non-normality issues.)

      i. What do we look for in the QQ-plot?

      ii. What do we look for in the histogram?

To get some practice, answer the same questions for the Firm data set.

1. Can we get descriptive statistics for each gender?

2. Are the means equal at the 0.05 level?

3. Does the model adhere to the assumptions of the One-Way ANOVA model?