EXST SAS Lab
Lab 11: Blocking and Factorials

Objectives

1. Run some contrasts on last week's assignment (Lab10)
2. Do an Analysis of Variance (ANOVA) in **PROC MIXED** including
   - Output of residuals **PROC MIXED**
   - **LSMEANS** with a **TUKEY** adjustment
   - **ODS** output for a macro called **PDMix800.sas**
   - Run a contrast testing for a linear trend and curvature
3. Run **PDMix800.sas** macro
4. Use **PROC UNIVARIATE** to test the residuals for normality

Simple Contrasts

The first part of this week's assignment is to add some contrasts to an ANOVA you did last week. Starting with last week's **PROC MIXED**. You may remove everything except the minimum needed to run the procedure; the **PROC**, **CLASS**, **MODEL** and a **RUN** statement. I would also include appropriate **TITLE#** statements.

```
PROC MIXED data=FungusMedium cl covtest;
   TITLE2 'Analysis of Variance using PROC MIXED';
   TITLE3 'Using Contrasts';
   class Medium;
   model FungusDiameter = Medium / outp=outputstuff;
RUN;
```

We will do some contrasts using this CRD from last week. These are sooo easy that I don't think I need to do a separate example for you. Your assignment from last week had a single treatment with levels equal to "cma, na, pca, pda, rda, twa, wa" assuming that you let SAS order the variables in the default alphabetical order. If you requested "order=data" the order is "wa, rda, pda, cma, twa, pca, na".

If you are in doubt about the order, check the "Class Level Information" segment of the **PROC MIXED** output. Let's assume that you are using the SAS default alphabetical order. The treatment levels are abbreviations of the following treatment levels: Water Agar (WA), Cornmeal Agar (CMA), Potato Carrot Agar (PCA), Potato Dextrose Agar (PDA), Tap Water Agar (TWA), Rice Dextrose Agar (RDA), and Nutrient Agar (NA).

The purpose of contrasts is to test one or more means against one or more other means. For example, we might decide we want to test "water" treatments (WA and TWA) against dextrose treatments (PDA and RDA). The test of hypothesis is $H_0: \frac{\mu_{WA}+\mu_{TWA}}{2} = \frac{\mu_{PDA}+\mu_{RDA}}{2}$, where the mean of the WA and TWA is tested for equality to the mean of PDA and RDA. The need for denominators can be eliminated by multiplying through by 2 and the hypothesis can be expressed as a linear combination; $H_0: (\mu_{WA}+\mu_{TWA})-(\mu_{PDA}+\mu_{RDA})=0$ or $H_0: \mu_{WA}+\mu_{TWA}-\mu_{PDA}-\mu_{RDA}=0$. Note that the coefficients of these means are +1, +1, -1 and -1.

A contrast statement in SAS has the following structure: "**CONTRAST '*SOME DESCRIPTION HERE*' TREATMENTNAME**" and then a series of contrast multipliers. The multipliers have a few restrictions: there must be one for each treatment level mean and they must sum to zero. For last week's ANOVA of the variable MEDIUM the contrast would look something like the following.

```
CONTRAST   'FIRST VERSUS ALL OTHERS'   MEDIUM   -6   1   1   1   1   1   1;
```

Obviously, listing a string of multipliers requires, each corresponding to the coefficient of one of the treatment level means, requires that you know the order of the treatment levels. To facilitate the making of the contrasts in the proper order, I suggest you make a SAS comment line to keep track of the order: cma, na, pca, pda, rda, twa, wa, That SAS comment statement is as follows.

```
*ORDER OF TREATMENT LEVELS              CMA   NA   PCA   PDA   RDA   TWA   WA;
```

Finally, we place our guide comment line above the contrast line and adjust the spacing of the comment line and write the desired contrasts.

```
*Order of Treatment levels                  cma  na pca pda rda twa  wa;
CONTRAST 'First versus all others' MEDIUM   -6   1   1   1   1   1   1;
```

Since there are only have 6 degrees of freedom, no more than 6 contrasts should be done. The treatment levels and abbreviations were: **Water** Agar (WA), **Cornmeal** Agar (CMA), **Potato** Carrot Agar (PCA), **Potato Dextrose** Agar (PDA), Tap **Water** Agar (TWA), **Rice Dextrose** Agar (RDA), and **Nutrient** Agar (NA). The contrasts needed are as follows:

1) Waters versus all others (WA and TWA versus all the rest) – 2 means versus 5 means
2) Cornmeal versus Potato (CMA versus PCA and PDA) – 1 mean versus 2 means
3) Dextrose versus others (PDA and RDA versus all the rest) – 2 means versus 5 means
4) Dextrose versus waters (PDA and RDA versus WA and TWA) – 2 means versus 2 means
5) waters versus vegetables (WA and TWA versus CMA, PCA, PDA and RDA) – 2 means versus 4
6) nutrient versus potato (NA versus PCA and PDA) – 1 mean versus 2 means

The first contrast tests the mean of water treatments (2 treatment levels) against the rest (5 levels). Assigning "5" to the 2 and assigning "2" to the 5 gives a sum of zero if you remember to make one of the two sides negative. The order is determined by the comment line we wrote as a guide.

```
*Order of Treatment levels                  cma  na pca pda rda twa  wa;
CONTRAST '2 Waters versus others' MEDIUM    2   2   2   2   2   -5  -5;
```

The second contrast tests the mean of the cornmeal treatment (1 treatment level) against the two potato treatments (2 levels). Assigning "1" to the 2 levels and assigning "2" to the 1 level gives a sum of zero. All the levels that we don't want to include get a zero. Again, the order is determined by the comment line guide.

```
*Order of Treatment levels                  cma  na pca pda rda twa  wa;
CONTRAST '2 Waters versus others' MEDIUM    2   2   2   2   2   -5  -5;
CONTRAST 'Cormeal versus Potato ' MEDIUM   -2   0   1   1   0   0   0;
```

Now, you write the remaining 4 contrasts. Good luck.

3) Dextrose versus others (PDA and RDA versus all the rest) – 2 means versus 5 means
4) Dextrose versus waters (PDA and RDA versus WA and TWA) – 2 means versus 2 means
5) waters versus vegetables (WA and TWA versus CMA, PCA, PDA and RDA) – 2 means versus 4
6) nutrient versus potato (NA versus PCA and PDA) – 1 mean versus 2 means

Randomized Block Design and Factorial Treatment Arrangement

The second part of this week's assignment includes much of the same tasks as last week (reinforcement!), like the **MACRO** and test of normality. There is also a **CONTRAST** statement. **PROC MIXED** will be done as a two-way Analysis of Variance and as a Randomized Block Design (RBD or RCBD) instead of the Completely Randomized Design (CRD) we did last week. The random component will necessitate a **RANDOM** statement.

The example

The dataset for this example was not taken from the textbook, The example is a CSV file from *The Data and Story Library* , hosted by the Stat. Dept. at Carnegie Mellon University http://lib.stat.cmu.edu/DASL/Datafiles/Stepping.html. The description of the data set states as follows.

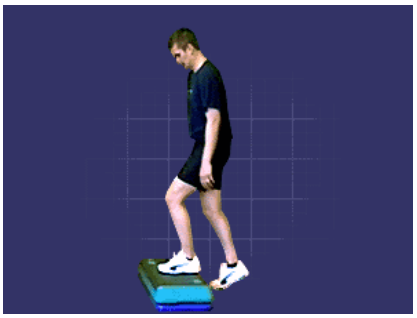
> "An experiment was conducted by students at The Ohio State University, Fall 1993, to study the relationship between a person's heart rate and the frequency at which that person stepped up and down on steps of various heights. The response variable "heart rate" was measured in beats per minute. There were two different step heights: 5.75 inches (coded as 0), and 11.5 inches (coded as 1) and three stepping frequencies: 14 steps/min. (coded as 0), 21 steps/min. (coded as 1), and 28 steps/min. (coded as 2). Each subject performed the activity for three minutes, kept on pace by the beat of an electric metronome. The subject's pulse was taken for 20 seconds before and after each trial and the subject rested between trials until the heart rate returned to near the beginning rate. Another experimenter kept track of the time. Each subject was always measured and timed by the same pair of experimenters to reduce variability in the experiment, and each pair of experimenters was treated as a block."

```
Order,Block,Height,Frequency,RestHR,HR
 1,   5,   0,   0,   87,    93
 2,   1,   1,   1,   87,   111
 3,   6,   1,   2,   81,   120
 4,   5,   0,   2,   75,   123
 5,   1,   0,   1,   81,    96
 6,   6,   1,   0,   84,    99
 7,   1,   1,   0,   84,    99
 8,   5,   1,   1,   90,   129
 9,   6,   0,   1,   75,    90
10,   1,   0,   0,   78,    87
11,   6,   0,   0,   84,    84
12,   5,   0,   1,   90,   108
13,   1,   0,   2,   78,    96
14,   6,   1,   1,   84,    90
15,   5,   1,   2,   90,   147
16,   2,   0,   0,   60,    75
18,   2,   0,   1,   63,    84
19,   2,   1,   2,   69,   135
21,   2,   1,   0,   69,   108
25,   2,   0,   2,   69,    93
17,   4,   1,   1,   96,   141
20,   4,   1,   0,   87,   120
22,   4,   0,   0,   90,    99
23,   4,   1,   2,   93,   153
27,   4,   0,   2,   87,   129
24,   3,   1,   1,   72,    99
26,   3,   0,   1,   69,    93
28,   3,   1,   0,   78,    93
29,   3,   0,   2,   72,    99
30,   3,   1,   2,   78,   129
```

After the usual heading information, the input statements were as follows.

```
title1 "Assignment 10 - Stepping - Heart Rate example";

data SteppingData;
  INFILE 'SteppingData.csv' dlm=',' dsd missover firstobs=2;
  input Order Block Height Frequency RestHR HR;
  HRIncrease = HR - RestHR;
     if height eq 0 then Ht = 5.75;
     if height eq 1 then Ht = 11.5;
     if frequency = 0 then freq = 14;
     if frequency = 1 then freq = 21;
     if frequency = 2 then freq = 28;
datalines; run;
;
run;
```



Source:http://news.bbc.co.uk/sport2/hi/health_and_fitness/4271662.stm

You will notice that the original data was coded as simple integers. This is not necessary, and I prefer the real values of the treatments in order to simplify interpretation later. So, I included 5 "**IF**" statements, renaming the variables and reassigning the values. Since the variables are primarily used in the **CLASS** statement of the **PROC MIXED** procedure, they can be either quantitative (numeric) or qualitative (alpha-numeric). If, however, we wish to use them in a graphic later it is best to have the original quantitative values. Your textbook generally uses the actual quantitative values, so you will not need to reassign any values using "**IF**" statements like the example. Lucky you.

First I ran a **PROC FREQ** to evaluate the data. I wanted to make sure that each combination of the two treatments was present and I to see if the experiment is balanced (i.e. has equal numbers of observations in each treatment combination).

```
proc freq data=SteppingData;
   TITLE2 'Number of observations in each category';
   table Ht*Freq / norow nocol nopercent;
run;
```

The table below shows that all six combinations of the two treatments exist, and that the frequency of replicates in each combination is the same.

```
Table of Ht by freq
Ht        freq
Frequency|      14|      21|      28|  Total
---------+--------+--------+--------+
    5.75 |      5 |      5 |      5 |     15
---------+--------+--------+--------+
    11.5 |      5 |      5 |      5 |     15
---------+--------+--------+--------+
Total          10       10       10       30
```

So the analysis has no missing cells and it is balanced. If one of the cells was empty of replicates this would be termed a missing cell. This can be a serious issue for two-way, or factorial, analyses.

The two-way Analysis of Variance is done in **PROC MIXED**. I have labeled each line in the **MIXED** procedure below to discuss each line individually.

```
a  PROC MIXED data=SteppingData cl covtest;
b     TITLE2 'Analysis of Variance using PROC MIXED';
c     class HT FREQ BLOCK;
d     model HRIncrease = HT | FREQ / outp=outputstuff;
e     random block;
f     contrast 'linear' FREQ   -1   0   1;
g     contrast 'curve ' FREQ   -1   2  -1;
h     lsmeans HT | FREQ / adjust=tukey cl;
i    ods output diffs=ppp lsmeans=mmm;
j    *ods exclude diffs lsmeans;
k  run;
l  %include 'C:\pdmix800.sas';
m  %pdmix800(ppp,mmm,alpha=0.05,sort=yes);
n  RUN;
```

The first four lines are similar to those you used last week. The vertical bar between the two treatments, **HT** and **FREQ**, will cause an interaction between those two variables to be included in the model (**HT** × **FREQ**). Note that this bar also occurs in the **LSMEANS** statement.

The model statement on line "d" represents the linear model for the two-way ANOVA, $Y_{ijk} = \mu + \beta_i + \tau_{1j} + \tau_{2k} + (\tau_1\tau_2)_{jk} + \varepsilon_{ijk}$, , where Y_{ijk} is the variable to be analyzed, μ represents the overall mean (correction factor), τ_{1j} represents the effect of the first treatment (**HT**), τ_{2k} represents the effect of the second treatment (**FREQ**), $(\tau_1\tau_2)_{jk}$ represents the effect of the interaction between the two treatments (**HT*FREQ**), and ε_{ijk} represents the within group deviations, or residuals. In SAS a vertical bar between treatments requests tests of the treatment effects themselves (called main effects) and the interaction of all treatments separated by bars.

In this particular case the error term comes from the pooling of all the block interactions together. The variance estimates are pooled and the degrees of freedom are $((b-1)(t_1-1) + (b-1)(t_2-1) + (b-1)(t_1-1)(t_2-1))$. This is typical of Randomized Block Designs and provides a proper error term when the assumptions are met. There is one oddity about this particular experiment. Usually, each of the six treatment combinations would occur in each block yielding $2 \times 3 \times 6 = 36$. However, this block design is incomplete, so not every subject (block) performed all 6 combinations of the treatment. As a result, there are a total of only 30 observations instead of 36 (an incomplete block design). The sources of variation and their degrees of freedom would factor out as follows.

Source	**degrees of freedom**	**d.f.**
Block	$b-1$	5
Treatment 1 (height)	t_1-1	1
Treatment 2 (frequency)	t_2-1	2
Interaction (ht × freq)	$(t_1-1)(t_2-1)$	2
Experimental Error (pooled block interactions)	$(b-1)(t_1-1) + (b-1)(t_2-1) + (b-1)(t_1-1)(t_2-1)$	19 (25 if blocks were complete)
Total	bt_1t_2-1	29 (35 if blocks were complete)

It would be possible to factor out some individual block interactions (e.g. **BLOCK*HT**, **BLOCK*FREQ** and **BLOCK*HT*FREQ**), but random block effects represent random error and we commonly make an *a priori* decision to pool these into a single error term. The random components are then just the block and the random error from the block interactions. The latter are not specified in the **RANDOM** statement and become the residuals.

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr > Z	Alpha	Lower	Upper
Block	55.1605	42.6183	1.29	0.0978	0.05	18.4948	616.98
Residual	57.3395	18.6034	3.08	0.0010	0.05	33.1620	122.32

As a result there are two components in the “Covariance Parameter Estimates”, the block, which are the subjects participating in the experiment, and the residual error.

The of the test of treatments showed that the main effects (**HT** and **FREQ**) would result in rejection of the null hypothesis ($H_0 : \mu_1 = \mu_2 = ... = \mu_t$). The hypothesis of no significant interaction effect was not rejected.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Ht	1	19	34.92	<.0001
freq	2	19	29.19	<.0001
Ht*freq	2	19	2.27	0.1301

Line “e” of the SAS program is a **RANDOM** statement. This statement has not been used previously. Any source of variation in the model which is a random effect should go in this statement. This would include any random treatment variables (not very common) or more likely block variables and nested error terms. This example has only a single blocking variable and the residual. The residual comes from the pooled block interactions. There are no additional sources of variation, so the pooled block interactions would be omitted from **RANDOM** statement and automatically picked up as the residual error term.

The treatment **HT** has only two levels, so a contrast is not different from the test done by **LSMEANS**. However, the other treatment, **FREQ** is quantitative so there are some special contrasts that may be of interest to test for a straight line trend (line f) or a curved trend (line g). These exist for any series of equally spaced quantitative treatments. The test for a linear trend and curved (quadratic) trend are given below.

```
*Order of Treatment levels          14    21    28;
CONTRAST 'Linear trend'    FREQ     -1     0     1;
CONTRAST 'Curved trend'    freq     -1     2    -1;
```

The results of the contrast indicated that there may be a linear trend, but no curvature.

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
linear	1	19	56.44	<.0001
curve	1	19	1.94	0.1801

After modifying the **LSMEANS** statement (line h) to include both of the treatments and the interaction, the remaining statements, lines h through n are the same. Likewise, the **PROC UNIVARIATE** with a test of normality that follows done on the residuals output from the ANOVA is not new.

The remaining lines (i through n) run the macro to do a range test with the LSMeans output.

Effect=Ht Method=Tukey-Kramer(P<0.05) Set=1

Obs	Ht	freq	Estimate	Standard Error	Alpha	Lower	Upper	Letter Group
1	11.5	_	35.7082	3.6168	0.05	28.1381	43.2784	A
2	5.75	_	19.0918	3.6168	0.05	11.5216	26.6619	B

Effect=freq Method=Tukey-Kramer(P<0.05) Set=2

Obs	Ht	freq	Estimate	Standard Error	Alpha	Lower	Upper	Letter Group
3	_	28	41.7204	3.8805	0.05	33.5984	49.8424	A
4	_	21	24.6330	3.8805	0.05	16.5110	32.7550	B
5	_	14	15.8466	3.8805	0.05	7.7246	23.9686	C

Effect=Ht*freq Method=Tukey-Kramer(P<0.05) Set=3

Obs	Ht	freq	Estimate	Standard Error	Alpha	Lower	Upper	Letter Group
6	11.5	28	53.6136	4.5813	0.05	44.0248	63.2025	A
7	5.75	28	29.8272	4.5813	0.05	20.2384	39.4161	B
8	11.5	21	29.1864	4.5813	0.05	19.5975	38.7752	B
9	11.5	14	24.3247	4.5813	0.05	14.7359	33.9136	B
10	5.75	21	20.0796	4.5813	0.05	10.4907	29.6684	BC
11	5.75	14	7.3685	4.5813	0.05	-2.2204	16.9573	C

Recall the analysis of variance indicated that the means in the main effects were significantly different, but not the interactions. The letter groupings for the interaction appear to show significant differences, but these apparent differences are due to differences among the main effects, not due to an interaction effect. In making the judgment as to whether to interpret difference among means, depend on the ANOVA tests, not the apparent pairwise differences among the means.

In the **PROC MIXED** analysis an output dataset was created called “outputstuff”. This dataset will contain a number of variables produced by the analysis including the variables used in the analysis, the predicted value (**PRED**) and the residual (**RESID**). Tests of normality in ANOVA, and later in regression, are done on the residuals. A **PROC UNIVARIATE** can be used for the test of normality of the residuals (**RESID**).

```
The UNIVARIATE Procedure
Variable:  Resid  (Residual)

                            Moments
N                          30    Sum Weights                 30
Mean                        0    Sum Observations             0
Std Deviation      6.27113144    Variance            39.3270896
Skewness           0.00771324    Kurtosis            -0.2857301
Uncorrected SS      1140.4856    Corrected SS         1140.4856
Coeff Variation             .    Std Error Mean      1.14494672

                   Tests for Normality
Test                  --Statistic---    -----p Value------
Shapiro-Wilk          W     0.988361    Pr < W      0.9801
Kolmogorov-Smirnov    D     0.088463    Pr > D     >0.1500
Cramer-von Mises      W-Sq  0.034545    Pr > W-Sq  >0.2500
Anderson-Darling      A-Sq  0.194781    Pr > A-Sq  >0.2500

   Stem Leaf                     #  Boxplot                  Normal Probability Plot
     12 5                        1     |        13+                                                      *++
     10                                |          |                                                  ++++
      8 47                       2     |          |                                               +*+*
      6 165                      3     |          |                                           **+*
      4 882                      3  +-----+       |                                        ***+
      2 57                       2  |     |       |                                      **+
      0 53                       2  |  +  |       |                                  ++**
     -0 2075310                  7  *-----*       |                              ******
     -2                             |     |       |                           +++
     -4 8740                     4  +-----+       |                        *****
     -6 2970                     4     |          |                   * *+*
     -8                                |          |                  +++
    -10 8                        1     |          |              +++*
    -12 3                        1     |       -13+   ++*
        ----+----+----+----+                       +----+----+----+----+----+----+----+----+----+----+
                                                       -2        -1         0        +1        +2
```

The results indicate a very nearly normal distribution of the residuals in this example. Also, there do not appear to be any outliers. At this point we are still assuming that the variances are homogeneous. If this was our research, we would probably also add a box plot and run a second **PROC MIXED** with a **GROUP** statement to check for homogeneity of variance. But since you already know these applications, we don't have to include this in every class exercise. Right?

Assignment 11

Part 1 – Using last week's assignment about a study of growth mediums for fungi, complete the four contrasts below on those treatment means (**1 point each**). Turn in your log and the contrast output.

3) Dextroses versus all others (2 versus 5; PDA and RDA versus all the rest) ..... **(1 point)**

4) Dextroses versus waters (2 versus 2; PDA and RDA versus WA and TWA) ..... **(1 point)**

5) Waters versus vegetables (2 versus 4; WA and TWA versus CMA, PCA, PDA and RDA) ..... **(1 point)**

6) Nutrient versus potato (1 versus 2; NA versus PCA and PDA) ..... **(1 point)**

Part 2

The objective of this experiment was to measure the effect of water stress on nitrogen fixation in four cowpea varieties (**VARIETY**). Plants were grown in a greenhouse and watered with 5, 10, and 15 ml of the appropriate nutrient solution (**SOLUTION**). Fifty-five days after planting, the nitrogen nodules were removed from the plants and counted and weighed. The entire experiment was replicated three times (these are **BLOCK**s). The response variable to be analyzed is the **WEIGHT** of the nitrogen nodule. The dataset is DATATAB_10_24.csv.

Source:http://letsplantsomething.wordpress.com/2010/08/03/cowpea-cultural-management/

Answer all questions about hypothesis tests by stating the outcome (REJECT the null hypothesis or FAIL to reject the null hypothesis) and give a P-value where possible. Turn in your log and the pertinent sections of the list output or results viewer output. You may write answers to questions on the log, or on a separate page.

Task 1: Do an ANOVA using **PROC MIXED** on the two treatments and interaction.

1b) include a **LSMEANS** statement with **ADJUST=TUKEY**. ..... **(1 point)**

1c) include the statements for Saxton's macro. ..... **(1 point)**

Question 1: Are there statistically significant differences for either of the two treatments (variety and solution) .....**(1 point)**

Question 2: Is there a statistically significant interaction? ..... (**1 point**)

Question 3: Are there any significant contrasts for solution (linear or curved)? ..... (**1 point**)

Task 2: Test the residuals from the ANOVA for the assumption of normality.

Question 4: Is the hypothesis of "normality" rejected? .....(**1 point**)