Analysis of Variance

In this lab, we will be learning how to perform Analysis of Variance (ANOVA) in SAS and Excel. We will be using the dataset `pets.xls` which may be downloaded from Blackboard.

The Pets dataset corresponds to the example data discussed in lecture notes from Chapter 13. Briefly, the study consists of three groups of people whose heart rates were measured after being exposed to psychological stressors. Subjects in Group 1 had their dog present, subjects in Group 2 had a human friend present, and subjects in Group 3 were alone.

Research Question: Among these three populations, is there any evidence of a difference in heart rate?

The SAS code in this exercise can be found in the file Lab9.sas, which is posted on Blackboard.

1 SAS

1.1 Read In The Data

We will read the dataset into SAS using the SAS Import Wizard. Make sure the Excel file `pets.xls` is closed before trying to import into SAS.

1. From the File menu in SAS, select 'Import Data ...'.
2. Check 'Standard Data Source' and select 'Microsoft Excel 97/2000/2002/2003 Workbook' from the pulldown menu.
3. When the window pops up, click Browse and locate your file ('H:\bios600\Data\pets.xls'). Click Open and then OK.
4. Select the table 'Sheet1$' (Sheet title) from the pulldown menu and hit Next.
5. Select 'WORK' to put the file in the Work library. Name your SAS dataset in the 'Member:' box by typing 'Pets'.
6. Click Finish.

After completing these steps, click on the Log to view it. You should see the following statement in blue.

```
NOTE: WORK.PETS data set was successfully created.
```

1.2 Descriptive Statistics

Our first step is to explore the data using descriptive statistics on the variable `HrtRate`. We are interested in describing `HrtRate` separately for each of the three groups, as well as combined across group. We use the `SORT` and `MEANS` procedures for this.

```
PROC SORT DATA=pets;
  BY group ;
RUN ;

*Heart Rate stats separately by group;
PROC MEANS DATA=pets N NMISS MEAN STD MIN MAX MAXDEC=4;
  CLASS group ;
  VAR hrtrate;
RUN ;

*Heart Rate stats for all groups combined;
PROC MEANS DATA=pets N NMISS MEAN STD MIN MAX MAXDEC=4;
  VAR hrtrate;
RUN ;
```

1.3 ANOVA

Suppose we would like to perform a test on the variable `hrtrate`, compared between the three levels of the `group` variable. That is, we would like to test whether the average Heart Rate is the same between the three Groups at significance level $\alpha = 0.05$. This corresponds to Analysis of Variance with the following hypotheses.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_A : \text{at least one mean } \mu_i \text{ differs from the others.}$$

To carry out the F-test for ANOVA, we use the `ANOVA` procedure. We implement the above hypothesis test in SAS by submitting the following four lines of code.

```
PROC ANOVA DATA=Pets ;
  CLASS Group ;
  MODEL HrtRate = Group ;
RUN ; QUIT ;
```

Assuming everything runs correctly, SAS will perform the desired hypothesis test and generate the results. The output is shown on the next page.

```
                                The SAS System
                              The ANOVA Procedure

                           Class Level Information

                          Class         Levels    Values
                          Group              3    1 2 3

                       Number of Observations Read          45
                       Number of Observations Used          45

********************************************************************************

                                The SAS System
                              The ANOVA Procedure

Dependent Variable: HrtRate   HrtRate
                                        Sum of
   Source                   DF         Squares     Mean Square    F Value    Pr > F
   Model                     2     2387.685004     1193.842502      14.08    <.0001
   Error                    42     3561.309490       84.793083
   Corrected Total          44     5948.994494


                     R-Square     Coeff Var      Root MSE    HrtRate Mean
                     0.401359      11.16916      9.208316        82.44410


   Source                   DF        Anova SS     Mean Square    F Value    Pr > F
   Group                     2     2387.685004     1193.842502      14.08    <.0001
```

SAS output often includes lots of things you don't need. The most important thing that you'll need from this output is the second table, which displays the ANOVA table corresponding to the F-test for this data. Remember that the ANOVA table contains all the elements we need for our hypothesis test.

Note that the `Source` column in the ANOVA table does not look exactly like the ANOVA tables from lecture. SAS applies different names to the rows, but everything else is the same. Just remember that 'Model'='Between', 'Error'='Within', and 'Corrected Total'='Total'.

From the ANOVA table, we have $F = 14.08$ and $p < 0.0001$, leading us to reject H_0. We conclude that the mean heart rate is not the same for all three groups.

1.4 Post-Hoc Comparisons

In the analysis above, we reject the null hypothesis H_0 and we conclude that there are differences between the three mean heart rates (i.e., the average heart rate is not the same for all three groups). However, we don't know where these differences are.

1.4.1 LSD

To identify pairwise differences, we use the post-hoc LSD method to compute the adjusted p-values. That is, we will perform the modified 2-sided 2-sample t-test for all three pairwise comparisons at $\alpha = 0.05$, which we set for ANOVA. In general, we are testing $H_0 : \mu_i = \mu_j$ against $H_A : \mu_i \neq \mu_j$ with test statistic

$$t_{i\ vs\ j} = \frac{\bar{x}_i - \bar{x}_j}{adjusted\ SE_{\bar{x}_i - \bar{x}_j}} = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MS_W\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

which has $N - 3$ degrees of freedom. We can do all three pairwise comparisons at once in SAS with one slight modification to the previous code.

```
PROC ANOVA DATA=Pets ;
  CLASS Group ;
  MODEL HrtRate = Group ;
  MEANS Group / LSD CLDIFF;
RUN ;
```

The same output as before is printed, with the addition of new output corresponding to the newly added `MEANS` statement. The new output is printed on the next page. The information we want is contained in the second table, which provides all pairwise mean differences ($\bar{x}_i - \bar{x}_j$), as well as the LSD-adjusted confidence interval.

SAS does not print the LSD-adjusted test statistics and p-values for each test, but it does give us the information to make a decision about each test. For each entry in the second table that demonstrates a significant difference via LSD method (adjusted $p < \alpha$), a series of three asterisks (***) is printed to the right.

To illustrate this, consider the first pairwise comparison t-test with $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 \neq \mu_2$. The output tells us that

$$\widehat{\bar{x}_1 - \bar{x}_2} = -17.842 \quad 95\%\text{ CI for } \mu_1 - \mu_2 : (-24.628, -11.056)(***)$$

The (***) tells us that we reject H_0 at $\alpha = 0.05$ and conclude that the average heart rate in Group 1 is significantly different than in Group 2.

For this example, since (***) is printed to the right of every entry, we can reject all three tests and conclude that all three pairs of means are significantly different from all the others.

```
                              The SAS System
                            The ANOVA Procedure

                        t Tests (LSD) for HrtRate
NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise
                                  error rate.

                        Alpha                          0.05
                        Error Degrees of Freedom         42
                        Error Mean Square           84.79308
                        Critical Value of t          2.01808
                        Least Significant Difference  6.7856


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

                          Difference
                Group        Between      95% Confidence
              Comparison       Means          Limits

                2 - 3           8.801      2.015   15.587  ***
                2 - 1          17.842     11.056   24.628  ***
                3 - 2          -8.801    -15.587   -2.015  ***
                3 - 1           9.041      2.255   15.827  ***
                1 - 2         -17.842    -24.628  -11.056  ***
                1 - 3          -9.041    -15.827   -2.255  ***
```

1.4.2 Bonferroni

If we want to use the Bonferroni adjustment instead of the LSD correction method, then we simply change the word LSD to BON in the code from §1.4.1. The output format is the same, but the values of the confidence interval will be different.

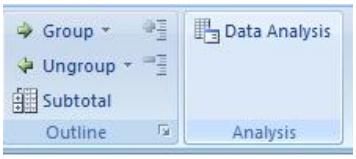
```
PROC ANOVA DATA=Pets ;
  CLASS Group ;
  MODEL HrtRate = Group ;
  MEANS Group / BON CLDIFF;
RUN ;
```

2 Excel

2.1 Load Analysis ToolPak

First, open up the dataset `pets.xls` in Excel. Click on the word 'Data' at the top of the screen to open up the Data ribbon ('Data' is fifth from left, starting with the word 'Home').

With the Data ribbon open, look at the right-most section of the ribbon menu. If the last two sections are 'Outline' and 'Analysis' and look like the picture below, then the Analysis ToolPak is ready to use.



If the last section of your Data ribbon only has the 'Outline' section but NOT the 'Analysis' section, then use the following steps to load the Analysis ToolPak in Excel 2007. (Directions for other versions of Excel can be found in Week1a_GetSoftware.pdf posted on Blackboard under Course Documents → Lab Exercises → Week 1.)

1. Click the Microsoft Office Button (top left corner)  , and then click Excel Options (bottom right).
2. Click Add-Ins from menu at left, and then in the Manage box (bottom), select Excel Add-ins.
3. Click Go.
4. In the Add-Ins available box, select the Analysis ToolPak check box, and click OK.

   * If Analysis ToolPak is not listed in the Add-Ins available box, click Browse to locate it.

   ** If you get prompted that the Analysis ToolPak is not currently installed on your computer, click Yes to install it.

The Analysis ToolPak should now be loaded. Click on the Data Ribbon again, and confirm that the last two sections at right are 'Outline' and 'Analysis', as in the picture above.

2.2 ANOVA: Single Factor

2.2.1 Convert Data to Short Format

Before we can perform ANOVA in Excel, we must first organize the data into **short format**, since the ToolPak does not work with data in long format. To do this, copy-paste the HrtRate data from subjects with Group=1 into column D, which is titled Group 1.

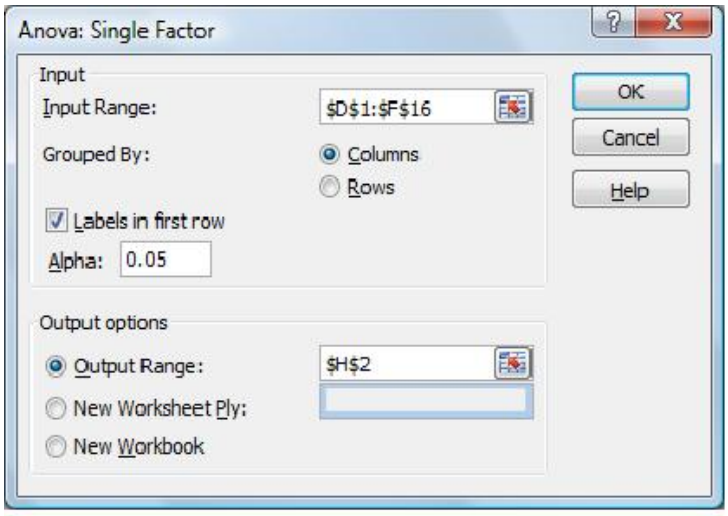
1. Left-click and hold down on cell B2. Then drag the cursor down to cell B16, so that all the cells in between are highlighted. This corresponds to the HrtRate data for subjects in Group 1.
2. In this highlighted area, right-click and select Copy from the menu.
3. Right-click on cell D2 (underneath 'Group 1' heading) and select Paste from the menu.

After completing these steps, the cells D2:D16 should now exactly match cells B2:B16. Do the same for Group 2 and Group 3 data, so that all of your data is now contained in cells D2:D16, E2:E16 and F2:F16.

2.2.2 ANOVA

Now we can perform ANOVA on this data using the 'ANOVA: Single Factor' tool. Follow the steps below to apply ANOVA to the heart rate data. A picture of the resulting input values is shown on the next page.

1. Activate the Data ribbon by clicking 'Data' at the top of the screen (fifth from left).
2. In the rightmost section called 'Analysis', click Data Analysis button to open the Analysis ToolPak dialog box.
3. Select 'ANOVA: Single Factor' and press OK.
4. In the Input Range box, press the button with the diagonal red arrow  , then select the entire set of data, including the Group labels.
5. Press the button with the downward red arrow  .
6. Check the box 'Labels in first row'.
7. Set Alpha to 0.05 (or whatever value you prefer).
8. Select Output Range, press the diagonal red arrow button and click cell H2. Press the downward red arrow button.
9. Check to make sure your menu is filled out exactly like the picture on the next page. Then press OK.



You should obtain the following output. If your results match those shown below, then you have done everything correctly!

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Group 1	15	1102.246	73.4830	99.397559
Group 2	15	1369.877	91.3251	69.574810
Group 3	15	1237.862	82.5241	85.406879

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2387.685	2	1193.843	14.0794798	2.09E-05	3.219942
Within Groups	3561.309	42	84.79308			
Total	5948.994	44				

2.3 Post-Test Comparisons

In the analysis above, we reject the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$. Therefore, we conclude that there are differences between the three mean heart rates (i.e., the average heart rate is not the same for all three groups). However, we don't know where these differences are.

To identify pairwise differences, we use the post-hoc LSD method to compute the adjusted p-values. That is, we will perform the modified 2-sample t-test for all three pairwise comparisons. There is no tool for this in the Analysis ToolPak, so we will rely on built-in Excel functions.

2.3.1 Compare Groups 1 and 2 (LSD)

We would like to perform the modified 2-sample t-test comparing Groups 1 and 2. In this setup, our hypotheses are $H_0 : \mu_1 = \mu_2$ vs $H_A : \mu_1 \neq \mu_2$. The LSD-adjusted test statistic is

$$t_{1v2} = \frac{\bar{x}_1 - \bar{x}_2}{adjusted\ SE_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{MS_W \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

and it have $df_W = 42$ degrees of freedom. We use the same $\alpha = 0.05$ from the original ANOVA F-test and compute the LSD-adjusted 2-sided p-value by completing the following steps.

1. Compute adjusted variance $MS_W \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$ and put it in cell I24.
   a) Click cell I24.
   b) Click in formula bar to activate (blank bar immediately above spreadsheet).
   c) Type: =K14*(1/I6+1/I7)
   d) Press Enter.
2. Compute adjusted standard deviation from ANOVA and put it in cell J24.
   a) Click cell J24.
   b) Click formula bar and type: =SQRT(I24)
3. Compute adjusted LSD pairwise test statistic t_{1v2} and put it in cell K24.
   a) Click cell K24.
   b) Click formula bar and type: =(K6-K7)/J24
4. Compute adjusted LSD p-value, using TDIST() function, and put it in cell L24. The format of this function is TDIST(positive test stat, degrees of freedom, 2-tails).
   a) Click cell L24.
   b) Click formula bar and type: =TDIST(-K24,42,2)

Since $p = 0.00000392$ is less than $\alpha = 0.05$, we reject the H_0, and conclude that the mean heart rate in Group 1 is significantly different from the mean heart rate in Group 2.

We have now applied the LSD post-hoc comparison between Groups 1 and 2. We can modify these steps to obtain the remaining two pairwise group comparisons (Group 1 vs 3, and Group 2 vs 3). The results from all three post-hoc tests are displayed in Table 1.

	Variance	Std Dev	t	p
G1 v G2	11.30574	3.362402	-5.30634	3.91735E-06
G1 v G3	11.30574	3.362402	-2.68886	0.010237514
G2 v G3	11.30574	3.362402	2.617482	0.012262924

Table 1:

2.3.2 Compare Groups 1 and 3 (LSD)

As in §2.3.1, we compare Groups 1 and 3 with an LSD-adjusted 2-sample t-test having $df = df_W = 42$. We have $H_0 : \mu_1 = \mu_3$ vs. $H_A : \mu_1 \neq \mu_3$ and test statistic

$$t_{1v3} = \frac{\bar{x}_1 - \bar{x}_3}{adjusted\ SE_{\bar{x}_1 - \bar{x}_3}} = \frac{\bar{x}_1 - \bar{x}_3}{\sqrt{MS_W \left(\frac{1}{n_1} + \frac{1}{n_3}\right)}}$$

1. Compute adjusted variance $MS_W \left(\frac{1}{n_1} + \frac{1}{n_3}\right)$ and put it in cell I25 using formula "=K14*(1/I6+1/I8)".
2. Compute adjusted standard deviation from ANOVA and put it in cell J25 using formula "=SQRT(I25)".
3. Compute adjusted LSD pairwise test statistic t_{1v3} and put it in cell K25 using formula "=(K6-K8)/J25".
4. Compute adjusted LSD p-value, using TDIST() function, and put it in cell L25 using formula "=TDIST(-K25,42,2)".

2.3.3 Compare Groups 2 and 3 (LSD)

Finally, we compare Groups 2 and 3 with an LSD-adjusted 2-sample t-test having the same $df = df_W = 42$. We have $H_0 : \mu_2 = \mu_3$ vs. $H_A : \mu_2 \neq \mu_3$ and test statistic

$$t_{2v3} = \frac{\bar{x}_2 - \bar{x}_3}{adjusted\ SE_{\bar{x}_2 - \bar{x}_3}} = \frac{\bar{x}_2 - \bar{x}_3}{\sqrt{MS_W \left(\frac{1}{n_2} + \frac{1}{n_3}\right)}}$$

1. Compute adjusted variance $MS_W \left(\frac{1}{n_1} + \frac{1}{n_3}\right)$ and put in cell I26 using formula "=K14*(1/I7+1/I8)".
2. Compute adjusted standard deviation from ANOVA and put it in cell J26 using formula "=SQRT(I26)".

3. Compute adjusted LSD pairwise test statistic t_{2v3} and put it in cell K26 using formula "=(K7-K8)/J26".

4. Compute adjusted LSD p-value, using TDIST() function, and put it in cell L26 using formula "=TDIST(K26,42,2)".

2.3.4 Bonferroni Adjustment

If we want to use the Bonferroni correction instead of LSD, then we simply perform the LSD method using the steps above and modify the resulting p-values. We use the following formula, where c is the total number of pairwise comparisons being tested.

$$p_{\text{Bon}} = c * p_{\text{LSD}}$$

For the pets data, we have $c = 3$, so we just multiply all the LSD p-values by 3.