# Statistics 147 LAB #6
## 10 pts; Summer 2020

**NAME:** Wesley Chang **ID: (last 4 #s only)** 0996

This lab is designed to give the student practice performing 1-Way Analysis of Variance and Multiple Comparison Techniques using R and SAS.

**REMINDER:** You reject $H_0$ if p-value $<$ specified of $\alpha$. We'll use $\alpha = 0.05$.

**Data Files:** You will need the data file, **plant.dat** (again).

Four chemical plants, producing the same product and owned by the same company, discharge effluent into streams in the vicinity of their locations. To check the extent of the pollution created by the effluents and to determine whether the amount of polluting effluents varies from plant to plant, the company collected random samples of liquid waste from each of the four plants. The data, in pounds per gallon of waste, is given in the table below.

| PlantA | PlantB | PlantC | PlantD |
|---|---|---|---|
| 1.65 | 1.70 | 1.40 | 1.58 |
| 1.72 | 1.85 | 1.75 | 1.77 |
| 1.50 | 1.36 | 1.58 | 1.48 |
| 1.37 | 2.05 | 1.65 | 1.69 |
| 1.60 | 1.80 | 1.55 | 1.65 |
| 1.40 | 2.10 | 1.45 | 1.65 |
| 1.75 | 1.95 | 1.66 | 1.79 |
| 1.38 | 1.65 | 1.70 | 1.58 |
| 1.65 | 1.80 | 1.85 | 1.77 |
| 1.55 | 2.00 | 1.24 | 1.60 |

Let Plant A be sample 1, Plant B be sample 2, Plant C be sample 3, and Plant D be sample 4. (Name your columns: Plant A. Plant B, Plant C, and Plant D.)

Let

| | | |
|---|---|---|
| $\mu_A$ | = | true mean discharge effluent for Plant A |
| $\mu_B$ | = | true mean discharge effluent for Plant B |
| $\mu_C$ | = | true mean discharge effluent for Plant C |
| $\mu_D$ | = | true mean discharge effluent for Plant D |

1. Complete the following using SAS.

   (i) Perform the appropriate test to determine whether there is a significant difference in mean amount of effluent between the four plants.

   (a) First test for normality of each the plants using **proc univariate** and the **Shapiro-Wilks** test.

   Recall, in Labs #3, #4 and #5, you read in the data. So open your SAS program file (lab5su20.sas), save it as **lab6su20.sas**, and add the following lines of code, **right before** the **run** statement.

   **(Modify title1 to show it is Lab 6.)**

   **DO NOT RETYPE THE ENTIRE PROGRAM!**

   **NOTE:** Scroll down to find the new code for Lab #6.

```
/* Set up format of the output */
options nocenter ps = 55 nocenter ls = 78 nodate nonumber formdlim='*';
 /* ls = linesize,  ps = pagesize
           nocenter      justifies the output so it is not centered on the page
           nodate        suppresses printing of today's date on each page of output
           nonumber      suppresses printing of page number on each page of output
           formdlim      overrides the internal page breaks and replaces them
                            with the designated symbol*/


/* Use DM to clear all windows except the editor window */
DM log "odsresults; clear; out; clear; log; clear;";
ods graphics off;

/* Create temporary SAS dataset named lab3su20 */


/* Create titles */
   title1 'Statistics 147 Lab #6, Summer 2020';
   title2 'Your name goes here';
   title3 'Question 1';

/* Open data file plant.dat. Be sure to specify the path indicating where
   you have saved the data file.  The actual data starts on Line 2.*/

data lab3su20;
    infile 'c:\linda\summer2020\su20147\datafiles\plant.dat' firstobs = 2;

    /* Create nested do loops to read in the data
       NOTE: There are 10 rows and four columns of data
       First, Do loop for the rows*/
    do row = 1 to 10;
        /* Do Loop for the columns */
        do plant = 1 to 4;
            /* Use If-Then-Else Structure to name the plants */
            if      plant = 1 then name = 'Plant A';
            else if plant = 2 then name = 'Plant B';
            else if plant = 3 then name = 'Plant C';
            else                   name = 'Plant D';

            /* Input response (data values) */
            input dischrg @@;

            /* Output the data */
            output;

        /* Close the plant loop */
        end;

    /* Close the row loop */
    end;
```

```
run;
/* Print the results */
proc print noobs;
   title4 'Part (i) Read in and Print data';
run;

/* First sort the data according to the variable plant */
proc sort;
/* Modify title4 */
   title4 'Part (ii): Descriptive Statistics';
   by plant;
run;

/* Use proc means to generate the mean and variance for each plant
    n              number of observations
    mean           sample mean
    var            sample variance
    by plant       group the data according to the plant from which the observation
                   was selected
    var discharge  generate mean and variance of the variable dischrg (for each plant) */

proc means n mean var;
   by plant;
   var dischrg;
run;

/* ********************************************************/
/*    NEW CODE BEGINS HERE FOR LAB 6                      */

/* Use proc univariate with the normal options to test normality
   Use ods select TestsForNormality to suppress printing of everything except
   the tests for normality
   Use "by" statement to generate test for each plant */
proc univariate normal;
   ods select TestsForNormality;
   by plant;
   var dischrg;
run;
quit;
```

**NOTE:** You may remove the remainder of the code! Then save and execute your code.

Complete the following. (Recall if $p-value < \alpha = 0.05 \Rightarrow$ reject $H_0$ : data is normal. Thus if $p-value > \alpha = 0.05 \Rightarrow$ ok to assume normality.)

**For Plant A:**

Tests for Normality

| Test | --Statistic--- | | -----p Value------ | |
|---|---|---|---|---|
| Shapiro-Wilk | W | 0.920447 | Pr < W | 0.3607 |

♣ $H_0$: Plant A is normally distributed

- ♣ $H_a$: Plant A is not normally distributed
- ♣ p-value = 0.3607
- ♣ RR: Reject $H_0$ if p-value $< \alpha = 0.05$
- ♣ **Conclusion:** Since the p-value = 0.3607 is (less than / greater than) [circle your choice] $\alpha = 0.05$, (reject / do not reject) [circle your choice] $H_0 \rightarrow$ it (is / is not) [circle your choice] reasonable to assume the Plant A is normally distributed.

**For Plant B:**

| Test | Tests for Normality --Statistic--- | | -----p Value------ | |
|---|---|---|---|---|
| Shapiro-Wilk | W | 0.940178 | Pr < W | 0.5550 |

- $H_0$: Plant B is normally distributed
- $H_a$: Plant B is not normally distributed
- p-value = 0.5550
- RR: Reject $H_0$ if p-value $< \alpha = 0.05$
- **Conclusion:** Since the p-value = 0.5550 is (less than, greater than) [circle your choice] $\alpha = 0.05$, (reject, do not reject) [circle your choice] $H_0 \rightarrow$ it (is, is not) [circle your choice] reasonable to assume the Plant B is normally distributed.

**For Plant C:**

| Test | Tests for Normality --Statistic--- | | -----p Value------ | |
|---|---|---|---|---|
| Shapiro-Wilk | W | 0.976499 | Pr < W | 0.9416 |

- $H_0$: Plant C is normally distributed
- $H_a$: Plant C is not normally distributed
- p-value = 0.9416
- RR: Reject $H_0$ if p-value $< \alpha = 0.05$
- **Conclusion:** Since the p-value = 0.9416 is (less than, greater than) [circle your choice] $\alpha = 0.05$, (reject, do not reject) [circle your choice] $H_0 \rightarrow$ it (is, is not) [circle your choice] reasonable to assume the Plant C is normally distributed.

**For Plant D:**

| Test | Tests for Normality --Statistic--- | | -----p Value------ | |
|---|---|---|---|---|
| Shapiro-Wilk | W | 0.939475 | Pr < W | 0.5472 |

- $H_0$: Plant D is normally distributed
- $H_a$: Plant D is not normally distributed
- p-value = 0.5472
- RR: Reject $H_0$ if p-value $< \alpha = 0.05$
- **Conclusion:** Since the p-value = 0.5472 is (less than, greater than) [circle your choice] $\alpha = 0.05$, (reject, do not reject) [circle your choice] $H_0 \rightarrow$ it (is, is not) [circle your choice] reasonable to assume the Plant D is normally distributed.

(b) Next test for equality (homogeneity) of variances using **proc glm** and the **HOVTEST** option with Bartlett's test. To accomplish this, add the following lines of code **right after the test for normality code**.

```
/* Use prog glm to generate appropriate output */
/*    class     name of classification variable
      model  dependent = class
      means  class/ HOVTEST = bartlett */

proc glm;
    class plant;
    model dischrg = plant;
    means plant /HOVTEST = bartlett;

run;
```

Complete the following:

```
The GLM Procedure

Bartlett's Test for Homogeneity of dischrg Variance

Source        DF     Chi-Square     Pr > ChiSq

plant          3     5.4307         0.1428      (p-value)
```

- ♠ $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
- ♠ $H_a$ : at least one of the $\sigma_i^2$ is different
- ♠ p-value = 0.1428
- ♠ Rejection Region: Reject $H_0$ if p-value < $\alpha = 0.05$
- ♠ **Conclusion:** Since the p-value = 0.1428 is (less than, greater than) [circle your choice] $\alpha = 0.05$, (reject, do not reject) [circle your choice] $H_0 \rightarrow$ it (is, is not) [circle your choice] reasonable to assume equality (homogeneity) of variances.

(c) Perform the appropriate test to determine whether there is a significant difference in mean amount of effluent between the four plants. This output already appears in your output file. So just toggle back to your output and complete the following:

```
 Dependent Variable: dischrg
                                   Sum of
Source                      DF       Squares    Mean Square   F Value   Pr > F

Model                        3    0.44029000     0.14676333   5.29      0.0040
Error                       36    0.99810000     0.02772500
Corrected Total             39    1.43839000
```

- ♣ $H_0$ : $\mu_1 = \mu_1 = \mu_1 = \mu_4$ (all $\alpha_i = 0$) (cannot conclude that at least one of the plants yields a significantly different mean discharge)

- ♣ $H_a$ : At least one $\mu_i$ is different (At least one $\alpha_i \neq 0$) (can conclude that at least one of the plants yields a significantly different mean discharge)
- ♣ p-value = 0.0040
- ♣ **Conclusion:** Since the p-value = 0.0040 is (less than, greater than) [circle your choice] $\alpha = 0.05$, (reject, do not reject) [circle your choice] $H_0 \rightarrow$ it (is, is not) [circle your choice] reasonable to assume that at least one of the plants yields a significantly different mean amount of effluent.

(d) Finally use Tukey's test to determine which treatment means are different. (Note: If you did not conclude there was a significant difference in treatment means, just state that!)

To get the Tukey's test information, add the following lines of code

```
\begin{verbatim}
/* New code */
/* means class / tukey */
    means plant / tukey;
```

right after the following lines of code (right before the **run** statement).

```
proc glm;
    class plant;
    model dischrg = plant;
    means plant /HOVTEST = bartlett;
```

Please complete the partial listing of the output and complete the table below.

## Using Grouping Method:

```
t Tests (LSD) for dischrg
Means with the same letter are not significantly different.
t Grouping  Mean      N   plant

A         1.82600     10    2

B A       1.6560      10    4

B         1.58300     10    3

B         1.55700     10    1
```

| Comparison | Common letter? Yes or No | Can conclude Sig Diff? Yes or No |
|---|---|---|
| Plant A vs Plant B | no | yes |
| Plant A vs Plant C | yes | no |
| Plant A vs Plant D | yes | no |
| Plant B vs Plant C | no | yes |
| Plant B vs Plant D | yes | no |
| Plant C vs Plant D | yes | no |

## Using the Confidence Interval Method:

One can use the **cldiff** command in the means statement to generate the simultaneous confidence intervals for the difference between each pair of means.

To accomplish this task, add the following lines of code right **before** the run statement:

```
/* Confidene Interval approach using cldiff */
means plant / tukey cldiff;
```

Save and execute your program and complete the following.

| plant Comparison | Difference Between Means | Simultaneous 95% Confidence Limits | |
|---|---|---|---|
| 2 - 4 | 0.17000 | -0.03055 | 0.37055 |
| 2 - 3 | 0.24300 | 0.04245 | 0.44355 |
| 2 - 1 | 0.26900 | 0.06845 | 0.46955 |
| 4 - 2 | -0.17000 | -0.37055 | 0.03055 |
| 4 - 3 | 0.07300 | -0.12755 | 0.27355 |
| 4 - 1 | 0.09900 | -0.10155 | 0.29955 |
| 3 - 2 | -0.24300 | -0.44355 | -0.04245 |
| 3 - 4 | -0.07300 | -0.27355 | 0.12755 |
| 3 - 1 | 0.02600 | -0.17455 | 0.22655 |
| 1 - 2 | -0.26900 | -0.46955 | -0.06845 |
| 1 - 4 | -0.09900 | -0.29955 | 0.10155 |
| 1 - 3 | -0.02600 | -0.22655 | 0.17455 |

**Note:** If the given interval contains the value 0, one cannot conclude there is a significant difference in the means. If the interval does not contain the value 0, then one can conclude there is a significant difference in the two means being compared.

Complete the following.

| Comparison | Interval | 0 in the Interval? Yes or No | Can conclude Sig Diff? Yes or No |
|---|---|---|---|
| Plant A vs Plant B | -0.47,-0.07 | no | yes |
| Plant A vs Plant C | -0.23,0.17 | yes | no |
| Plant A vs Plant D | -0.30,0.10 | yes | no |
| Plant B vs Plant C | 0.04,0.44 | no | yes |
| Plant B vs Plant D | -0.03,0.37 | yes | no |
| Plant C vs Plant D | -0.27,0.13 | yes | no |

2. Complete the following using **R**. (Make sure you make the columns accessible!)

Open and execute your **R script** from Lab #5. Save it as **lab6_su20_XX**, where XX = initials of your name. Then move your cursor to the end of the script, so you can add more code.

(i) Perform the appropriate test to determine whether there is a significant difference in mean amount of effluent between the four plants.

(a) First test for normality using the **Anderson-Darling** test.

Complete the following. (Recall if $p-value < \alpha = 0.05 \Rightarrow$ reject $H_0$ : data is normal. Thus if $p-value > \alpha = 0.05 \Rightarrow$ ok to assume normality.)

To accomplish this task, add the following lines of code to the end of your **R script**: **For Plant A:**

```
# New for Lab 6
# Load nortest package
# Use ad.test( ) for Plant A
ad.test(PlA)
```

Make sure your cursor is in the **R Editor** window. Save your script and select **Edit → Run All**.

Complete the following:

```
        Anderson-Darling normality test
data:  PlA
```

A = 0.31703, p-value = 0.478

- $H_0$: Plant A is normally distributed
- $H_a$: Plant A is not normally distributed
- p-value = 0.478
- RR: Reject $H_0$ if p-value $< \alpha = 0.05$
- **Conclusion:** Since the p-value = 0.478 is (less than, greater than) [circle your choice] $\alpha = 0.05$, (reject, do not reject) [circle your choice] $H_0 \rightarrow$ it (is, is not) [circle your choice] reasonable to assume the Plant A is normally distributed.

**For Plant B:**
To accomplish this task, add the following lines of code to the end of your **R script**:

```
# Use ad.test( ) for Plant B
ad.test(PlB)
```

Make sure your cursor is in the **R Editor** window. Save your script and then

▲ highlight the new text you just typed.

▲ From the main menu, select **Edit → Run line or selection**.

Complete the following:

```
        Anderson-Darling normality test
data:  PlB

A = ____0.26083____, p-value = ____0.6242____
```

♣ $H_0$: Plant B is normally distributed

♣ $H_a$: Plant B is not normally distributed

♣ p-value = ____0.6242____

♣ RR: Reject $H_0$ if p-value $< \alpha = 0.05$

♣ **Conclusion:** Since the p-value = ____0.6242____ is (less than, greater than) [circle your choice] $\alpha = 0.05$, (reject, do not reject) [circle your choice] $H_0 \rightarrow$ it (is, is not) [circle your choice] reasonable to assume the Plant B is normally distributed.

**For Plant C:**
To accomplish this task, add the following lines of code to the end of your **R script**:

```
# Use ad.test( ) for Plant C
ad.test(PlC)
```

Make sure your cursor is in the **R Editor** window. Save your script and then

▲ highlight the new text you just typed.

▲ From the main menu, select **Edit → Run line or selection**.

Complete the following:

```
        Anderson-Darling normality test
data:  PlC

A = ____0.17604____, p-value = ____0.8942____
```

♣ $H_0$: Plant C is normally distributed

♣ $H_a$: Plant C is not normally distributed

♣ p-value = ____0.8942____

♣ RR: Reject $H_0$ if p-value $< \alpha = 0.05$

- ♣ **Conclusion:** Since the p-value = 0.8942 is (less than, greater than) [circle your choice] $\alpha = 0.05$, (reject, do not reject) [circle your choice] $H_0 \rightarrow$ it (is, is not) [circle your choice] reasonable to assume the Plant C is normally distributed.

**For Plant D:**
To accomplish this task, add the following lines of code to the end of your **R script**:

```
# Use ad.test( ) for Plant D
ad.test(PlD)
```

Make sure your cursor is in the **R Editor** window. Save your script and then

- ▲ highlight the new text you just typed.
- ▲ From the main menu, select **Edit → Run line or selection**.

Complete the following:

```
        Anderson-Darling normality test
data:  PlD

A = 0.29755, p-value = 0.5202
```

- ♣ $H_0$: Plant D is normally distributed
- ♣ $H_a$: Plant D is not normally distributed
- ♣ p-value = 0.5202
- ♣ RR: Reject $H_0$ if p-value $< \alpha = 0.05$
- ♣ **Conclusion:** Since the p-value = 0.5202 is (less than, greater than) [circle your choice] $\alpha = 0.05$, (reject, do not reject) [circle your choice] $H_0 \rightarrow$ it (is, is not) [circle your choice] reasonable to assume the Plant D is normally distributed.

(b) Next test for equality (homogeneity) of variances using the **bartlett.test** function.

**REMINDER:** The data must be stacked for the remaining procedures!

**Recall:** Once the data has been stacked, **values** represents the columns of data values and **ind** indicates the column names (i.e., the plants).

To accomplish this task, add the following lines of code to the end of your **R script**:

```
# Stack the data and make the columns accessible
stack_plants <- stack(plant_data)
attach(stack_plants)
names(stack_plants)
# Use bartlett.test(values,ind) to test for homogeneity of variances
# values = data values and ind = classes
bartlett.test(values,ind)
```

Make sure your cursor is in the **R Editor** window. Save your script and then

- ▲ highlight the new text you just typed.
- ▲ From the main menu, select **Edit → Run line or selection**.

```
        Bartlett test of homogeneity of variances
data:  values and ind

Bartlett's K-squared = 5.4307, df = 3, p-value = 0.1428
```

- ♠ $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
- ♠ $H_a$ : at least one of the $\sigma_i^2$ is different
- ♠ p-value = 0.1428
- ♠ Rejection Region: Reject $H_0$ if p-value $<$ $\alpha = 0.05$
- ♠ **Conclusion:** Since the p-value = 0.1428 is (less than, greater than) [circle your choice] $\alpha = 0.05$, (reject, do not reject) [circle your choice] $H_0 \rightarrow$ it (is, is not) [circle your choice] reasonable to assume equality (homogeneity) of variances.

(c) Perform the appropriate test to determine whether there is a significant difference in mean amount of effluent between the four plants.

To accomplish this task, add the following lines of code to the end of your **R script**:

```
# USE THE aov FUNCTION TO GENERATE ANOVA INFORMATION
# GENERAL FORMAT FOR 1-WAY CRD: aov(response~factor, data = dataname)
results2 = aov(values~ind,data=stack_plants)
summary(results2)
```

Make sure your cursor is in the **R Editor** window. Save your script and then

- ▲ highlight the new text you just typed.
- ▲ From the main menu, select **Edit → Run line or selection**.

```
            Df Sum Sq Mean Sq F value  Pr(>F)
ind          3 0.4403 0.14676   5.294 0.00397
Residuals   36 0.9981 0.02772
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- ♣ $H_0$ : $\mu_1 = \mu_1 = \mu_1 = \mu_4$ (all $\alpha_i = 0$) (cannot conclude that at least one of the plants yields a significantly different mean discharge)
- ♣ $H_a$ : At least one $\mu_i$ is different (At least one $\alpha_i \neq 0$) (can conclude that at least one of the plants yields a significantly different mean discharge)

♣ p-value = 0.00397

♣ **Conclusion:** Since the p-value = 0.00397 is (less than, greater than) [circle your choice] $\alpha = 0.05$, (reject, do not reject) [circle your choice] $H_0 \rightarrow$ it (is, is not) [circle your choice] reasonable to assume that at least one of the plants yields a significantly different mean amount of effluent.

(d) Finally use Tukey's test to determine which treatment means are different. (Note: If you did not conclude there was a significant difference in treatment means, just state that!)

To accomplish this task, add the following lines of code to the end of your **R script**:

```
> # Use TukeyHSD test for multiple comparisons
> TukeyHSD(results2,conf.level=0.95)
```

Make sure your cursor is in the **R Editor** window. Save your script and then

▲ highlight the new text you just typed.

▲ From the main menu, select **Edit → Run line or selection**.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = values ~ ind, data = stack_plants)
$ind
```

| | diff | lwr | upr | p adj |
|---|---|---|---|---|
| P1B-P1A | 0.269 | 0.06844949 | 0.46955051 | 0.0048673 |
| P1C-P1A | 0.026 | -0.17455051 | 0.22655051 | 0.9451443 |
| P1D-P1A | 0.099 | -0.10155051 | 0.29955051 | 0.55707007 |
| P1C-P1B | -0.243 | -0.44355051 | -0.04244949 | 0.0123554 |
| P1D-P1B | -0.170 | -0.37055051 | 0.03055051 | 0.1210406 |
| P1D-P1C | 0.073 | -0.12755051 | 0.27355051 | 0.7614324 |

**Using the Confidence Interval Method:**

**Note:** If the given interval contains the value 0, one cannot conclude there is a significant difference in the means. If the interval does not contain the value 0, then one can conclude there is a significant difference in the two means being compared.

| **Comparison** | **0 in the Interval? Yes or No** | **Can conclude Sig Diff? Yes or No** |
|---|---|---|
| Plant A vs Plant B | no | yes |
| Plant A vs Plant C | yes | no |
| Plant A vs Plant D | yes | no |
| Plant B vs Plant C | no | yes |
| Plant B vs Plant D | yes | no |
| Plant C vs Plant D | yes | no |

**Using the p-value Method:**

Recall,

- ♠ If p-value $\not< \alpha$, one **CANNOT** conclude a significant difference in the means.
- ♠ If p-value $< \alpha$, one **CAN** conclude a significant difference in the means.

Complete the following:

| **Pair Comparison** | **p-value** | **p-value $< \alpha = 0.05$ (Yes or No)** | **Can conclude a significant difference? (Yes or No)** |
|---|---|---|---|
| Plant A vs Plant B | 0.005 | yes | yes |
| Plant A vs Plant C | 0.985 | no | no |
| Plant A vs Plant D | 0.551 | no | no |
| Plant B vs Plant C | 0.012 | yes | yes |
| Plant B vs Plant D | 0.121 | no | no |
| Plant C vs Plant D | 0.761 | no | no |

**You have now successfully completed Lab #6. Please turn in your lab worksheet and make sure your work area is neat and clean. Don't forget your flash drive, if you used one!**

**Please thank Ruihan for all her wonderful help this quarter! Have a good day!**

***Luke & Ruihan***