Statistics 147 Assignment #5

Summer 2020

Wesley Chang

0996

R Questions:

1. **Using R, complete the following:**

   (i) Test for underlying normality for each of the dogs. (Use Anderson-Darling Test.) (4 pts)

From the results of ad.test(), we can see that the results of the tests for Cody, Dusty, and Shadow return p-values that are larger than 0.05, which is the alpha at the 95% level. In all three cases, we do not reject the null hypothesis that the data is normally distributed. Therefore, we can assume that all three samples are normal.

```
> ## Part i
> # test for underlying normality for each of the dogs (using Anderson-Darling Test)
>
> # use ad.test, which is in the library nortest
>
> # call nortest library
> library(nortest)
>
> ## Anderson-Darling Test
> # For Cody
> ad.test(Cody)

	Anderson-Darling normality test

data:  Cody
A = 0.28829, p-value = 0.5222

> # For Dusty
> ad.test(Dusty)

	Anderson-Darling normality test

data:  Dusty
A = 0.222, p-value = 0.7426

> # For Shadow
> ad.test(Shadow)

	Anderson-Darling normality test

data:  Shadow
A = 0.16514, p-value = 0.9052

>
```

(ii) Test for equality (homogeneity) of variances. (Use Bartlett's test.) Use $\alpha = 0.05$. (2 pts)

The results of Bartlett's test give us a p-value of 0.3536. If we use alpha = 0.05 level of signficance, we see that the p-value is greater than the alpha, so we do not reject the null hypothesis that the variances are equal. Therefore, we can assume homogeneity of variances.

```
> ## Part ii
> # test for equality (homogeneity) of variances (using Bartlett's test), use a = 0.05
>
> # bartlett.test(x,g), where x is the set of numeric vectors, and g is the set of factor objects
> bartlett.test(values,ind)

	Bartlett test of homogeneity of variances

data:  values and ind
Bartlett's K-squared = 2.0791, df = 2, p-value =
0.3536
```

(iii) Perform the appropriate test(s) of hypothesis to determine whether one can conclude that at least one of the dogs has a significantly different mean finishing time. Use α = 0.05. (4 pts)

Since the above results show that there is normality and a homogeneity of variances, we can test the difference of means. The function aov() gives us a p-value of 0.0125, so we reject the null hypothesis that there is an equality in means at the 0.05 level of significance.

```
> ## Part iii
> # perform the appropriate test(s) of hypothesis to determine wheter one can conclude that
> # at least of the dogs has a significantly different mean finishing time, using a = 0.05
>
> # appropriate test is determined by whether normality and homogeneity assumptions are satisfied
>
> # normality satisfied as per Anderson-Darling Test
> # homogeneity satisfied as per Bartlett's test
>
> # testing difference in means, we can use aov()
> # we set values ~ ind to indicate corresponding variables and group
> # we set data=agility_long to indicate R to read from this data frame
> results_anova <- aov(values ~ ind, data=agility_long)
>
> # print data and a summary of the data
> results_anova
Call:
   aov(formula = values ~ ind, data = agility_long)

Terms:
                      ind Residuals
Sum of Squares   866.0833 1673.7500
Deg. of Freedom         2        21

Residual standard error: 8.927619
Estimated effects may be unbalanced
> summary(results_anova)
            Df Sum Sq Mean Sq F value Pr(>F)
ind          2  866.1   433.0   5.433 0.0125 *
Residuals   21 1673.7    79.7
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(iv) If appropriate, use **Tukey's test** *and* the **p-value method** to determine which mean(s) is(are) significantly different. (Be sure to justify your answer!) (3 pts)

We can use Tukey's test here since the data is independent, there is assumed normality, and assumed homogeneity in variances.

Confidence Interval approach:

Dusty-Cody: 0 does not fall into the confidence interval, so we can conclude that there is a difference in means between Dusty and Cody.

Shadow-Cody: 0 falls into the confidence interval, so we cannot conclude that is a difference in means between Shadow and Cody.

Shadow-Dusty: 0 does not fall into the confidence interval, so we can conclude that there is a difference in means between Shadow and Dusty

P-value method:

Dusty-Cody: The p-value is lower than the 0.05 significance level, so we reject the null hypothesis that there is no difference, so we can conclude that there is a difference in means between Dusty and Cody.

Shadow-Cody: The p-value is higher than the 0.05 significance level, so we do not reject the null hypothesis that there is no difference, and cannot conclude that there is a difference in means between Shadow and Cody

Shadow-Dusty: The p-value is lower than the 0.05 significance level, so we reject the null hypothesis that there is no difference, so we can conclude that there is a difference in means between Shadow and Dusty.

```
> ## Part iv
> # if appropriate, use Tukey's test and the p-value method to determine which means(s) is(are)
> # significantly different (justify answer)
>
> # We concluded in Part iii that there is a significant difference such that not all means are e
>
> # we now can use Tukey's test
> # TukeyHSD(results_anova,conf.level)
> # we pass on the results_anova data frame to TukeyHSD and generate results at the 95% level
> TukeyHSD(results_anova,conf.level=0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = values ~ ind, data = agility_long)

$ind
                diff        lwr        upr     p adj
Dusty-Cody   -12.000 -23.251345 -0.7486546 0.0352459
Shadow-Cody    1.375  -9.876345 12.6263454 0.9491664
Shadow-Dusty  13.375   2.123655 24.6263454 0.0180987
```

SAS Questions:

2. **Using SAS, complete the following.** Modify your existing **SAS** program file (from Assignments #3 and #4) to complete the following.

(i) Test for underlying normality for each of the dogs. (Use Shapiro-Wilk Test.) (4 pts)

According to the results of the Shapiro-Wilk test, each dog's p-value is larger than the level of significance 0.05. This means that we do not reject the null hypothesis that each dog's results are normal, and can assume normality in all three samples.

```
title5 'Part i';
title6 'test for normality';
/* test for normality for each of the dogs, using Shapiro-Wilk test */

/* 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 name;
    var score;
run;
```

Statistics 147 Assignment #5
Summer 2020
Wesley Chang
SAS Question 1
Part i
test for normality

The UNIVARIATE Procedure
Variable: score

name=Cody

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.912591	Pr < W	0.3727
Kolmogorov-Smirnov	D	0.13869	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.040069	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.28829	Pr > A-Sq	>0.2500

Statistics 147 Assignment #5
Summer 2020
Wesley Chang
SAS Question 1
Part i
test for normality

The UNIVARIATE Procedure
Variable: score

name=Dust

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.942452	Pr < W	0.6354
Kolmogorov-Smirnov	D	0.146037	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.030387	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.222003	Pr > A-Sq	>0.2500

Statistics 147 Assignment #5
Summer 2020
Wesley Chang
SAS Question 1
Part i
test for normality

The UNIVARIATE Procedure
Variable: score

name=Shad

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.980723	Pr < W	0.9664
Kolmogorov-Smirnov	D	0.139337	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.024628	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.165141	Pr > A-Sq	>0.2500

(ii) Test for equality (homogeneity) of variances. (Use Bartlett's test.) Use $\alpha = 0.05$. (2 pts)

According to the results of Bartlett's test, we get a p-value of 0.3536, which is larger than the level of significance 0.05. We do not reject the null hypothesis that there is equality of variances and can therefore assume homogeneity of variances.

```
title5 'Part ii';
title6 'Test for equality (homogeneity of variances), using Bartletts test';

/* Barlett's test can be conducted using proc glm with the HOVTEST option */

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

proc glm;
    class name;
    model score = name;
    means name /HOVTEST = bartlett;
run;
```

Statistics 147 Assignment #5
Summer 2020
Wesley Chang
SAS Question 1
Part ii
Test for equality (homogeneity of variances), using Bartletts test

The GLM Procedure

Class Level Information		
Class	Levels	Values
name	3	Cody Dust Shad

Number of Observations Read	24
Number of Observations Used	24

Statistics 147 Assignment #5
Summer 2020
Wesley Chang
SAS Question 1
Part ii
Test for equality (homogeneity of variances), using Bartletts test

The GLM Procedure

Bartlett's Test for Homogeneity of score Variance			
Source	DF	Chi-Square	Pr > ChiSq
name	2	2.0791	0.3536

Statistics 147 Assignment #5
Summer 2020
Wesley Chang
SAS Question 1
Part ii
Test for equality (homogeneity of variances), using Bartletts test

The GLM Procedure

Level of name	N	score Mean	score Std Dev
Cody	8	77.1250000	5.9865922
Dust	8	65.1250000	10.3432172
Shad	8	78.5000000	9.8125284

(iii) Perform the appropriate test(s) of hypothesis to determine whether one can conclude that at least one of the dogs has a significantly different mean finishing time. Use α = 0.05. (4 pts)

From parts i and ii, we concluded that there is normality and homogeneity in variances. From this, we can continue on with testing the difference in means. The output was already generated in the proc glm function, so we refer to that section for the p-value. Since the p-value of that result is 0.0125, we reject the null hypothesis that the means are equal at the 0.05 level of significance. Therefore, we can conclude that there is a significant difference in means.

```
title5 'Part ii and iii';
title6 'Test for equality (homogeneity of variances), using Bartletts test; and difference in means';

/* Barlett's test can be conducted using proc glm with the HOVTEST option */
/* we can also test difference in means */

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

proc glm;
    class name;
    model score = name;
    means name /HOVTEST = bartlett;
run;
```

Statistics 147 Assignment #5
Summer 2020
Wesley Chang
SAS Question 1
Part ii
Test for equality (homogeneity of variances), using Bartletts test

The GLM Procedure

Dependent Variable: score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	866.083333	433.041667	5.43	0.0125
Error	21	1673.750000	79.702381		
Corrected Total	23	2539.833333			

R-Square	Coeff Var	Root MSE	score Mean
0.341000	12.13266	8.927619	73.58333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
name	2	866.0833333	433.0416667	5.43	0.0125

Source	DF	Type III SS	Mean Square	F Value	Pr > F
name	2	866.0833333	433.0416667	5.43	0.0125

(iv) If appropriate, use Tukey's test and both the grouping and confidence interval methods to determine which mean(s) is(are) significantly different. (Be sure to justify your answer!) (3 pts)

We can use Tukey's test here since the data is independent, there is assumed normality, and assumed homogeneity in variances.

Confidence Interval approach:

Dusty-Cody: 0 does not fall into the confidence interval, so we can conclude that there is a difference in means between Dusty and Cody.

Shadow-Cody: 0 falls into the confidence interval, so we cannot conclude that is a difference in means between Shadow and Cody.

Shadow-Dusty: 0 does not fall into the confidence interval, so we can conclude that there is a difference in means between Shadow and Dusty

Grouping method:

Group A: Includes Shadow and Cody

Group B: Dusty

There is a significant difference in means between Group A, which contains Shadow and Cody and Group B, which only contains Dusty.

```
title5 'Part iv';
title6 'using Tukeys test to determine which means are significantly different';

/* append the code from parts ii and iii to include
        means name / tukey
    to perform Tukey's test
    */

proc glm;
    class name;
    model score = name;
    means name /HOVTEST = bartlett;
    /* for grouping method */
    means name / tukey;
    /* for confidence interval method */
    means name /tukey cldiff;
run;
```

Statistics 147 Assignment #5
Summer 2020
Wesley Chang
SAS Question 1
Part iv
using Tukeys test to determine which means are significantly different

The GLM Procedure

Tukey's Studentized Range (HSD) Test for score

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	21
Error Mean Square	79.70238
Critical Value of Studentized Range	3.56462
Minimum Significant Difference	11.251

Means with the same letter are not significantly different.			
Tukey Grouping	Mean	N	name
A	78.500	8	Shad
A			
A	77.125	8	Cody
B	65.125	8	Dust

Statistics 147 Assignment #5
Summer 2020
Wesley Chang
SAS Question 1
Part iv
using Tukeys test to determine which means are significantly different

The GLM Procedure

Tukey's Studentized Range (HSD) Test for score

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	21
Error Mean Square	79.70238
Critical Value of Studentized Range	3.56462
Minimum Significant Difference	11.251

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

name Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Shad - Cody	1.375	-9.876	12.626	
Shad - Dust	13.375	2.124	24.626	***
Cody - Shad	-1.375	-12.626	9.876	
Cody - Dust	12.000	0.749	23.251	***
Dust - Shad	-13.375	-24.626	-2.124	***
Dust - Cody	-12.000	-23.251	-0.749	***

R Code:

```
# Statistics 147 Assignment #5
# Summer 2020
# Wesley Chang

# open file agility.dat, read into data set agility
setwd("C:/Users/wesle/iCloudDrive/Summer 2020 (UCR)/STAT 147 (Session A)/Assignments/5")
agility <- read.table(file = "agility.dat", header=TRUE, skip=1)
agility

# attach column names, and verify
attach(agility)
names(agility)
Cody
Dusty
Shadow

# convert data frame from wide to long
agility_long <- stack(agility)
agility_long

# attach column names, and verify
attach(agility_long)
names(agility_long)
values
ind

### R Question 1

## Part i
# test for underlying normality for each of the dogs (using Anderson-Darling Test)

# use ad.test, which is in the library nortest

# call nortest library
library(nortest)

## Anderson-Darling Test
# For Cody
ad.test(Cody)
# For Dusty
ad.test(Dusty)
# For Shadow
ad.test(Shadow)
```

```
## Part ii
# test for equality (homogeneity) of variances (using Bartlett's test), use a =
0.05

# bartlett.test(x,g), where x is the set of numeric vectors, and g is the set of
factor objects
bartlett.test(values,ind)


## Part iii
# perform the appropriate test(s) of hypothesis to determine wheter one can conclude
that
# at least of the dogs has a significantly different mean finishing time, using a
= 0.05

# appropriate test is determined by whether normality and homogeneity assumptions
are satisfied

# normality satisfied as per Anderson-Darling Test
# homogeneity satisfied as per Bartlett's test
# we can proceed with testing for equality of means

# testing difference in means, we can use aov()
# we set values ~ ind to indicate corresponding variables and group
# we set data=agility_long to indicate R to read from this data frame
results_anova <- aov(values ~ ind, data=agility_long)

# print data and a summary of the data
results_anova
summary(results_anova)


## Part iv
# if appropriate, use Tukey's test and the p-value method to determine which means(s)
is(are)
# significantly different (justify answer)

# We concluded in Part iii that there is a significant difference such that not all
means are equal

# we now can use Tukey's test
# TukeyHSD(results_anova,conf.level)
# we pass on the results_anova data frame to TukeyHSD and generate results at the
95% level
TukeyHSD(results_anova,conf.level=0.95)
```

```
# Statistics 147 Assignment #5
# Summer 2020
# Wesley Chang

# open file agility.dat, read into data set agility
setwd("C:/Users/wesle/iCloudDrive/Summer    2020    (UCR)/STAT    147    (Session
A)/Assignments/5")
agility <- read.table(file = "agility.dat", header=TRUE, skip=1)
agility

# attach column names, and verify
attach(agility)
names(agility)
Cody
Dusty
Shadow

# convert data frame from wide to long
agility_long <- stack(agility)
agility_long

# attach column names, and verify
attach(agility_long)
names(agility_long)
values
ind

### R Question 1

## Part i
# test for underlying normality for each of the dogs (using Anderson-Darling Test)

# use ad.test, which is in the library nortest

# call nortest library
library(nortest)

## Anderson-Darling Test
# For Cody
ad.test(Cody)
# For Dusty
ad.test(Dusty)
# For Shadow
ad.test(Shadow)
```

```
## Part ii
# test for equality (homogeneity) of variances (using Bartlett's test), use a =
0.05

# bartlett.test(x,g), where x is the set of numeric vectors, and g is the set of
factor objects
bartlett.test(values,ind)


## Part iii
# perform the appropriate test(s) of hypothesis to determine wheter one can conclude
that
# at least of the dogs has a significantly different mean finishing time, using a
= 0.05

# appropriate test is determined by whether normality and homogeneity assumptions
are satisfied

# normality satisfied as per Anderson-Darling Test
# homogeneity satisfied as per Bartlett's test
# we can proceed with testing for equality of means

# testing difference in means, we can use aov()
# we set values ~ ind to indicate corresponding variables and group
# we set data=agility_long to indicate R to read from this data frame
results_anova <- aov(values ~ ind, data=agility_long)

# print data and a summary of the data
results_anova
summary(results_anova)


## Part iv
# if appropriate, use Tukey's test and the p-value method to determine which means(s)
is(are)
# significantly different (justify answer)

# We concluded in Part iii that there is a significant difference such that not all
means are equal

# we now can use Tukey's test
# TukeyHSD(results_anova,conf.level)
# we pass on the results_anova data frame to TukeyHSD and generate results at the
95% level
TukeyHSD(results_anova,conf.level=0.95)
```

SAS Code:

```
title1 'Statistics 147 Assignment #5';
title2 'Summer 2020';
title3 'Wesley Chang';

options nocenter ps = 55 nocenter ls = 78 nodate nonumber formdlim='*';
DM log "odsresults; clear; out; clear; log; clear;";
ods graphics off;


title4 'SAS Question 1';

data agility;
      infile 'C:\Users\wesle\iCloudDrive\Summer 2020 (UCR)\STAT 147 (Session 
A)\Assignments\5\agility.dat' firstobs = 3;

      /* do loop for rows */
      do row = 1 to 8;
            /* do loop for columns */
            do dog = 1 to 3;
                  /* if then structure to name dogs */
                  if                  dog = 1 then name = 'Cody';
                  else if             dog = 2 then name = 'Dusty';
                  else                             name = 'Shadow';

                  /* input statement */
                  input score @@;
                  output;
            end;
      end;
run;

/* print out results */
proc print noobs data = agility;
run;

/* add code to sort the data by the name of the dog */
proc sort data = agility;
      by dog;
run;

proc print noobs data = agility;
      title6 'Print to check sorted';
run;


title5 'Part i';
title6 'test for normality';
/* test for normality for each of the dogs, using Shapiro-Wilk test */

/* 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 name;
      var score;
run;


title5 'Part ii and iii';
title6 'Test for equality (homogeneity of variances), using Bartletts test; and
difference in means';

/* Barlett's test can be conducted using proc glm with the HOVTEST option */
/* we can also test difference in means */

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

proc glm;
      class name;
      model score = name;
      means name /HOVTEST = bartlett;
run;


title5 'Part iv';
title6 'using Tukeys test to determine which means are significantly different';

/* append the code from parts ii and iii to include
            means name / tukey
      to perform Tukey's test
      */

proc glm;
      class name;
      model score = name;
      means name /HOVTEST = bartlett;
      /* for grouping method */
      means name / tukey;
      /* for confidence interval method */
      means name /tukey cldiff;
run;
```