**Statistics 147**

**Assignment #6**

**Fall 2017**

**Sarah Ruckman**

**7194**

**Section 002**

1. **Question 1. Using SAS:**

Let μi = true mean time finishing time of dog team i, i = 1

(Trusty Dusty), 2 (White Shadow), 3 (Lakota Dakota)

Let  = true variance for the finishing time of dog team i, i = 1

(Trusty Dusty), 2 (White Shadow), 3 (Lakota Dakota)

1. Test for normality for each of the geysers. (Use Shapiro-Wilk Test.)

**SAS Code:**

options ls = **70** ps = **55** nocenter formdlim = '\*';

/\* ls = linesize, ps = pagesize, nocenter = justifies output, formdlim = overrides the internal page breaks

and replaces them with the designated symbol \*/

ods graphics off;

/\*Turns off extra graphics\*/

/\* Create titles \*/

title1 'Statstics 147 Assignment #6';

title2 'Fall 2017';

title3 'Sarah Ruckman';

title4 'Section 002';

title5 'SAS Question 1';

/\*Read infile using do loops do rows and then columns and there are 2 lines of headers\*/

/\*Create temporary SAS dataset called geysers\*/

**data** geysers;

infile "C:\Users\sarah\Downloads\geysers\_f17.dat" firstobs = **3**;

do rows = **1** to **10**;

do col = **1** to **3**;

/\*Create if then else statements to label the columns and end with an else statement\*/

if col = **1** then name = 'Old Faithful';

else if col = **2** then name = 'Grand Geysir';

else name = 'Strokkur ';

/\*input the values and add @@ to denote that the values are on the same line\*/

input values @@;

/\*Output the results\*/

output;

/\*Close both loops\*/

end;

end;

/\*Sort the data by column using proc sort\*/

**proc** **sort**;

by name;

/\*Print as check without observations\*/

**proc** **print** noobs;

/\*Use proc univariate with the normal options to test normality\*/

/\*Use ods select TestsForNormality to supress printing of everything except the test for normailty\*/

/\*Use by class statement and then var input variable\*/

**proc** **univariate** normal;

ods select TestsForNormality;

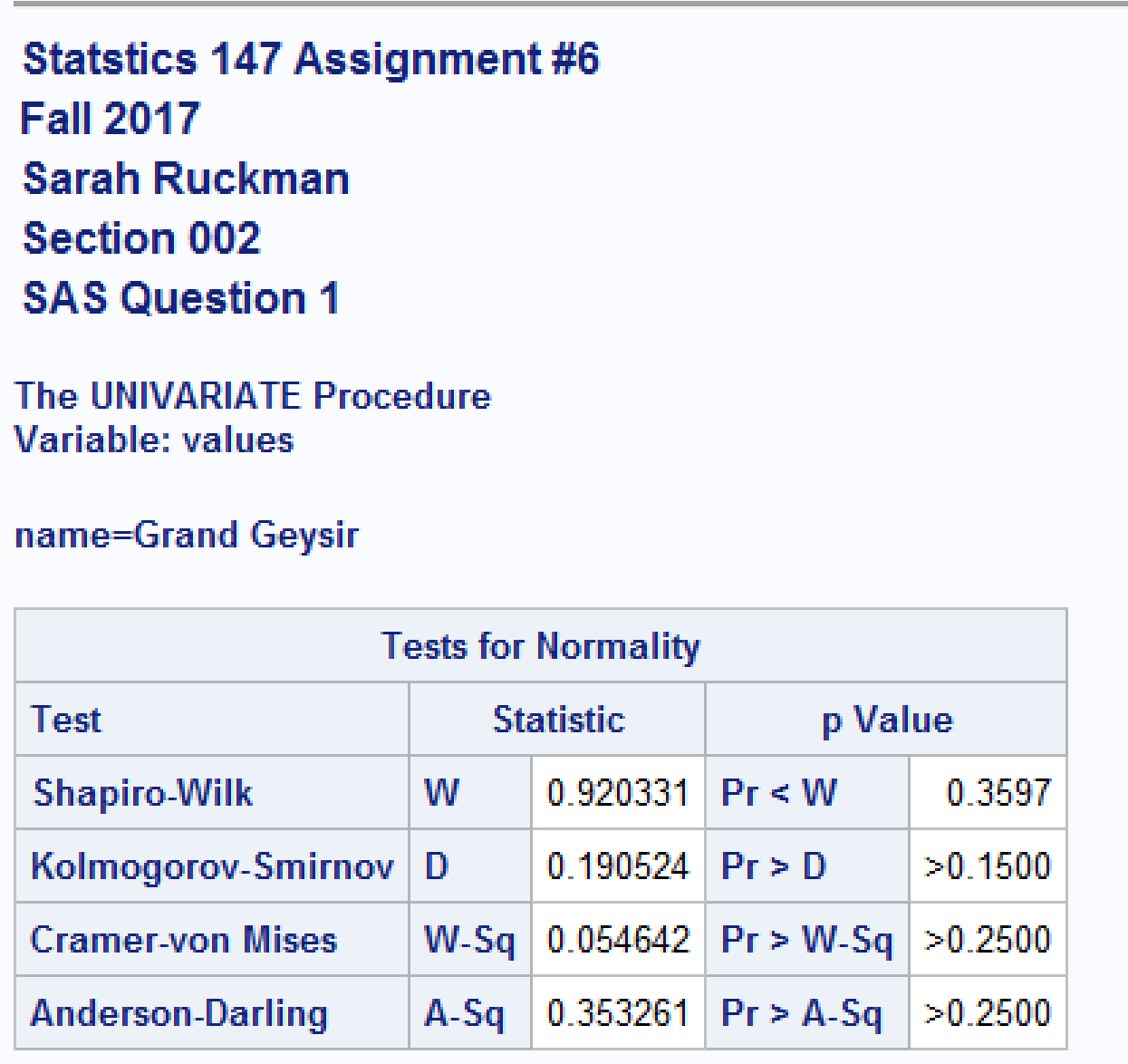
by name;

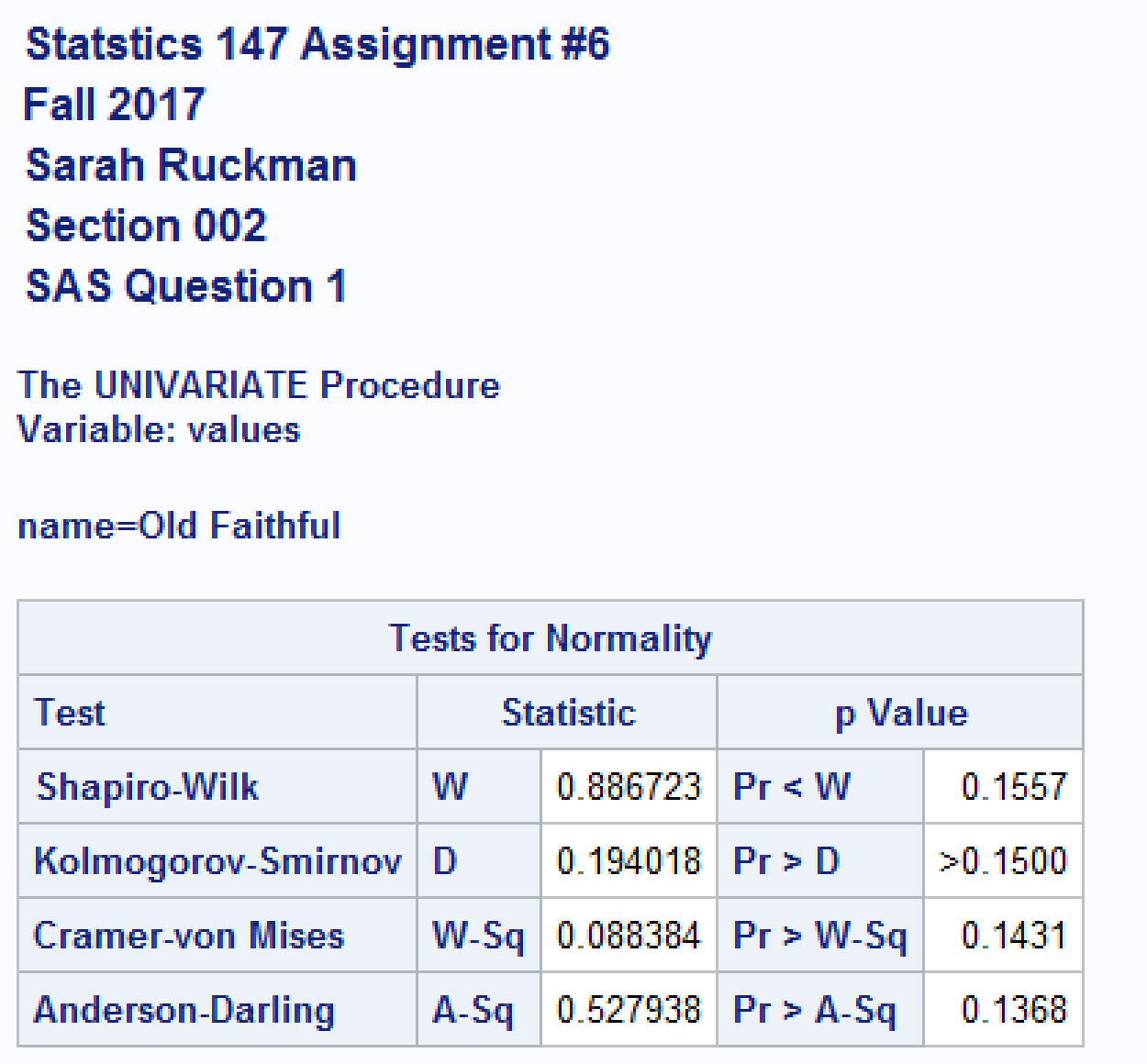
var values;

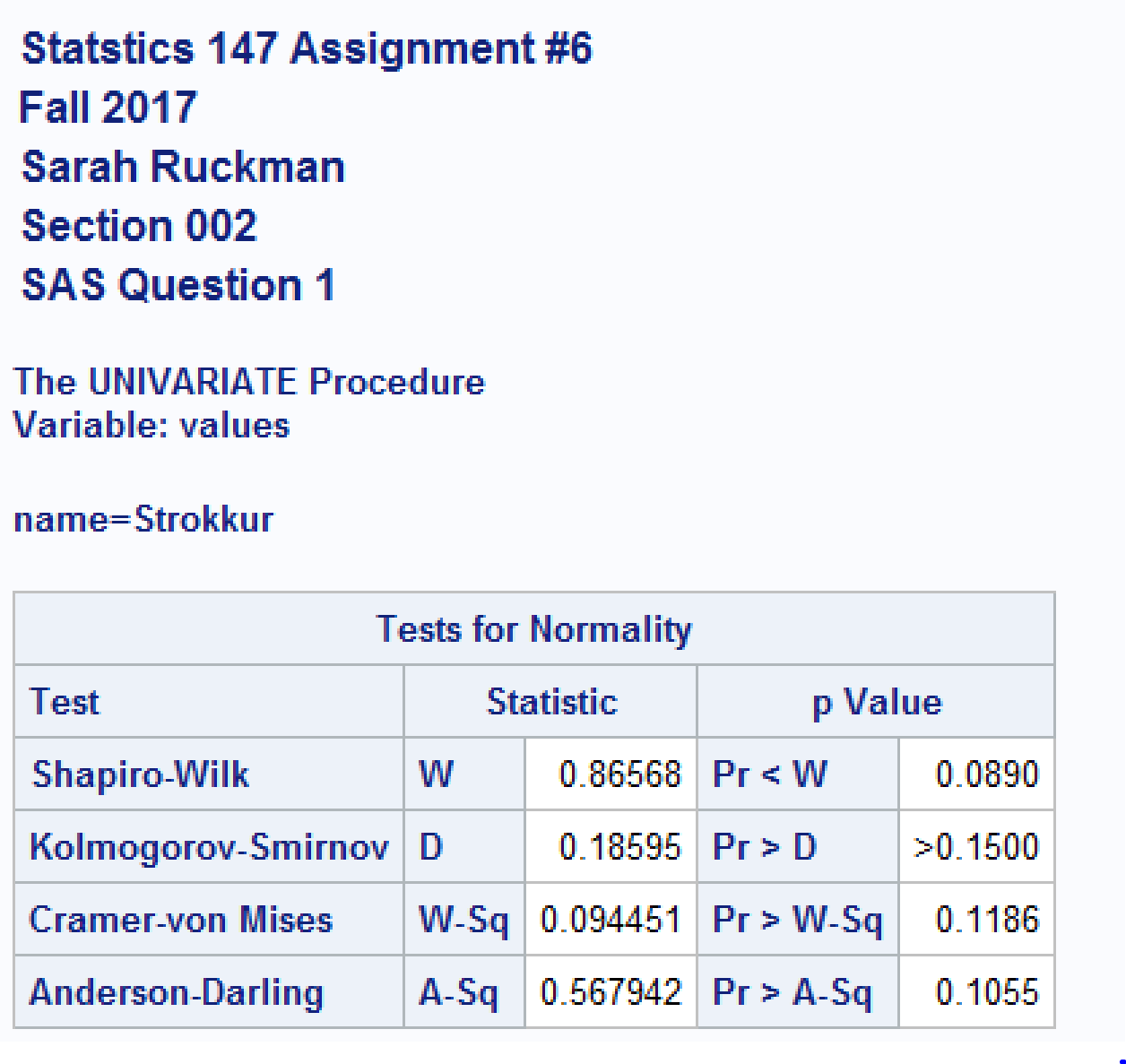
**run**;

**quit**;

**Output:**







First, Old Faithful geyser:

* H0: Old Faithful eruption height is normally distributed
* Ha: Old Faithful eruption height is not normally distributed
* p-value = **0.1557**
* Rejection Region: Reject H0 if p-value < α = 0.05
* Conclusion: **Since the p-value = 0.1557 is greater than α = 0.05, We do not reject H0, it is reasonable to assume that Old Faithful’s eruption height is normally distributed**

Next, Grand Geysir geyser:

* H0: Grand Geysir eruption height is normally distributed
* Ha: Grand Geysir eruption height is not normally distributed
* p-value = **0.3597**
* Rejection Region: Reject H0 if p-value < α = 0.05
* Conclusion: **Since the p-value = 0.3597 is greater than α = 0.05, We do not reject H0, it is reasonable to assume that the Grand Geysir’s eruption height is normally distributed**

Finally, Strokkur geyser:

* H0: Strokkur eruption height is normally distributed
* Ha: Strokkur eruption height is not normally distributed
* p-value = **0.0890**
* Rejection Region: Reject H0 if p-value < α = 0.05
* Conclusion: **Since the p-value = 0.0890 is greater than α = 0.05, We do not reject H0, it is reasonable to assume that Strokkur’s eruption height is normally distributed**

1. Test for equality/homogeneity of variances. (Use Bartlett's test.)

Use α = 0.05.

**SAS Code:**

options ls = **70** ps = **55** nocenter formdlim = '\*';

/\* ls = linesize, ps = pagesize, nocenter = justifies output, formdlim = overrides the internal page breaks

and replaces them with the designated symbol \*/

ods graphics off;

/\*Turns off extra graphics\*/

/\* Create titles \*/

title1 'Statstics 147 Assignment #6';

title2 'Fall 2017';

title3 'Sarah Ruckman';

title4 'Section 002';

title5 'SAS Question 1';

/\*Read infile using do loops do rows and then columns and there are 2 lines of headers\*/

/\*Create temporary SAS dataset called geysers\*/

**data** geysers;

infile "C:\Users\sarah\Downloads\geysers\_f17.dat" firstobs = **3**;

do rows = **1** to **10**;

do col = **1** to **3**;

/\*Create if then else statements to label the columns and end with an else statement\*/

if col = **1** then name = 'Old Faithful';

else if col = **2** then name = 'Grand Geysir';

else name = 'Strokkur ';

/\*input the values and add @@ to denote that the values are on the same line\*/

input values @@;

/\*Output the results\*/

output;

/\*Close both loops\*/

end;

end;

/\*Sort the data by column using proc sort\*/

**proc** **sort**;

by name;

/\*Print as check without observations\*/

**proc** **print** noobs;

/\*Use proc univariate with the normal options to test normality\*/

/\*Use ods select TestsForNormality to supress printing of everything except the test for normailty\*/

/\*Use by class statement and then var input variable\*/

**proc** **univariate** normal;

ods select TestsForNormality;

by name;

var values;

/\*Use proc glm to generate appropriate output for homogenity of variances and ANOVA\*/

/\*Use format:

class name of classification variable

model dependent = class

means class/HOVTEST = bartlett\*/

**proc** **glm**;

class name;

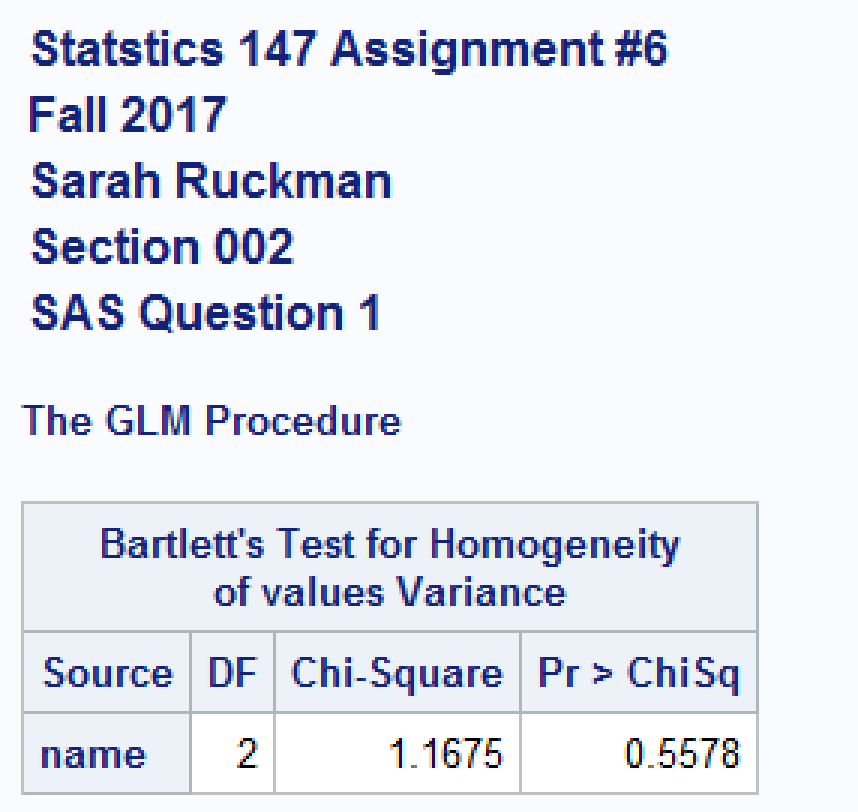
model values = name;

means name /HOVTEST=bartlett;

**run**;

**quit**;

**Output**:

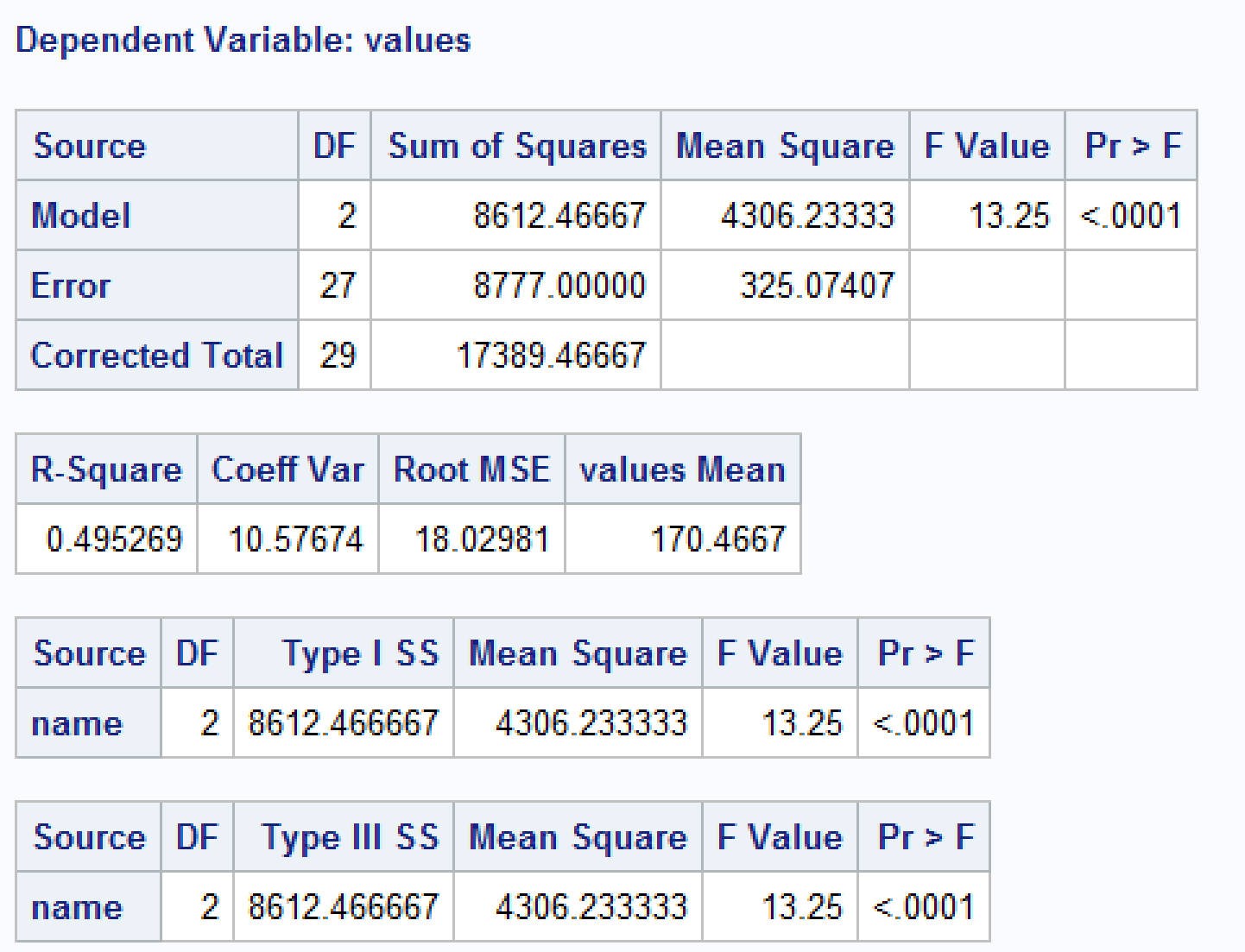


* 
* 
* p-value = **0.5578**
* Rejection Region: Reject H0 if p-value < α = 0.05
* Conclusion: **Since the p-value = 0.5578 is greater than α = 0.05, We do not reject H0, it is reasonable to assume that the variances are equal**

1. Perform the appropriate test(s) of hypothesis to determine whether one can conclude that at least one of the geysers has a significantly different mean eruption height. Use α = 0.05.

**See SAS Code Above.**

**Output:**



* +  (cannot conclude that at least one of the geysers yields a significantly different mean eruption height)
  +  (can conclude that at least one of the geysers yields a significantly different mean eruption height)
  + p-value = **< 0.0001**
  + Rejection Region: Reject H0 if p-value < α = 0.05
  + Conclusion: **Since the p-value <0.0001 is less than α = 0.05, we reject H0 and it is reasonable to assume that at least one of the mean eruption heights is significantly different**

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

**SAS Code:**

options ls = 70 ps = 55 nocenter formdlim = '\*';

/\* ls = linesize, ps = pagesize, nocenter = justifies output, formdlim = overrides the internal page breaks

and replaces them with the designated symbol \*/

ods graphics off;

/\*Turns off extra graphics\*/

/\* Create titles \*/

title1 'Statstics 147 Assignment #6';

title2 'Fall 2017';

title3 'Sarah Ruckman';

title4 'Section 002';

title5 'SAS Question 1';

/\*Read infile using do loops do rows and then columns and there are 2 lines of headers\*/

/\*Create temporary SAS dataset called geysers\*/

data geysers;

infile "C:\Users\sarah\Downloads\geysers\_f17.dat" firstobs = 3;

do rows = 1 to 10;

do col = 1 to 3;

/\*Create if then else statements to label the columns and end with an else statement\*/

if col = 1 then name = 'Old Faithful';

else if col = 2 then name = 'Grand Geysir';

else name = 'Strokkur ';

/\*input the values and add @@ to denote that the values are on the same line\*/

input values @@;

/\*Output the results\*/

output;

/\*Close both loops\*/

end;

end;

/\*Sort the data by column using proc sort\*/

proc sort;

by name;

/\*Print as check without observations\*/

proc print noobs;

/\*Use proc univariate with the normal options to test normality\*/

/\*Use ods select TestsForNormality to supress printing of everything except the test for normailty\*/

/\*Use by class statement and then var input variable\*/

proc univariate normal;

ods select TestsForNormality;

by name;

var values;

/\*Use proc glm to generate appropriate output for homogenity of variances and ANOVA\*/

/\*Use format:

class name of classification variable

model dependent = class

means class/HOVTEST = bartlett\*/

proc glm;

class name;

model values = name;

means name /HOVTEST=bartlett;

/\*Add means class/Tukey LSD and clm Tukey LSD CLDIFF to test for significant differences\*/

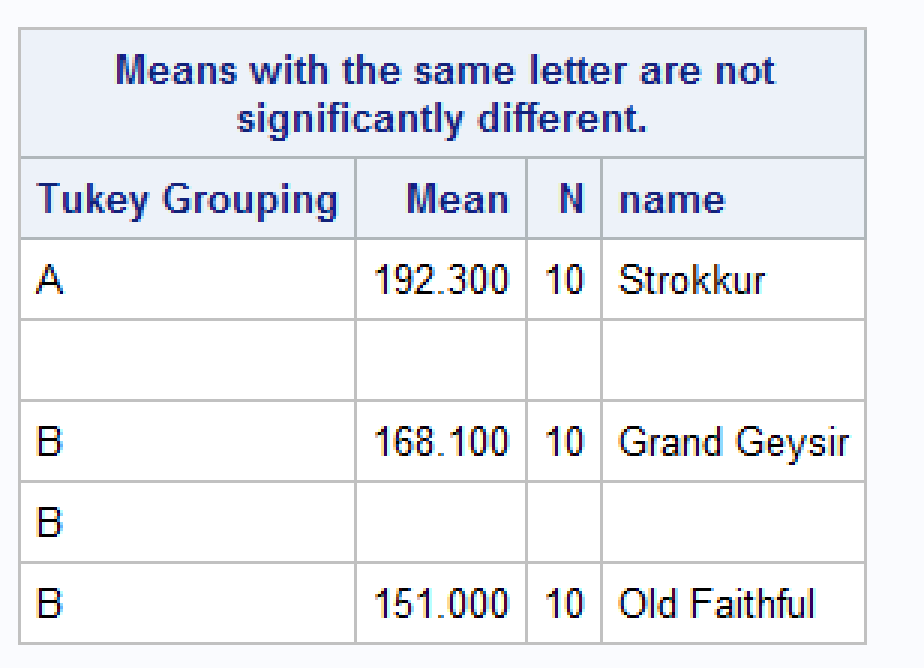
means name /Tukey LSD;

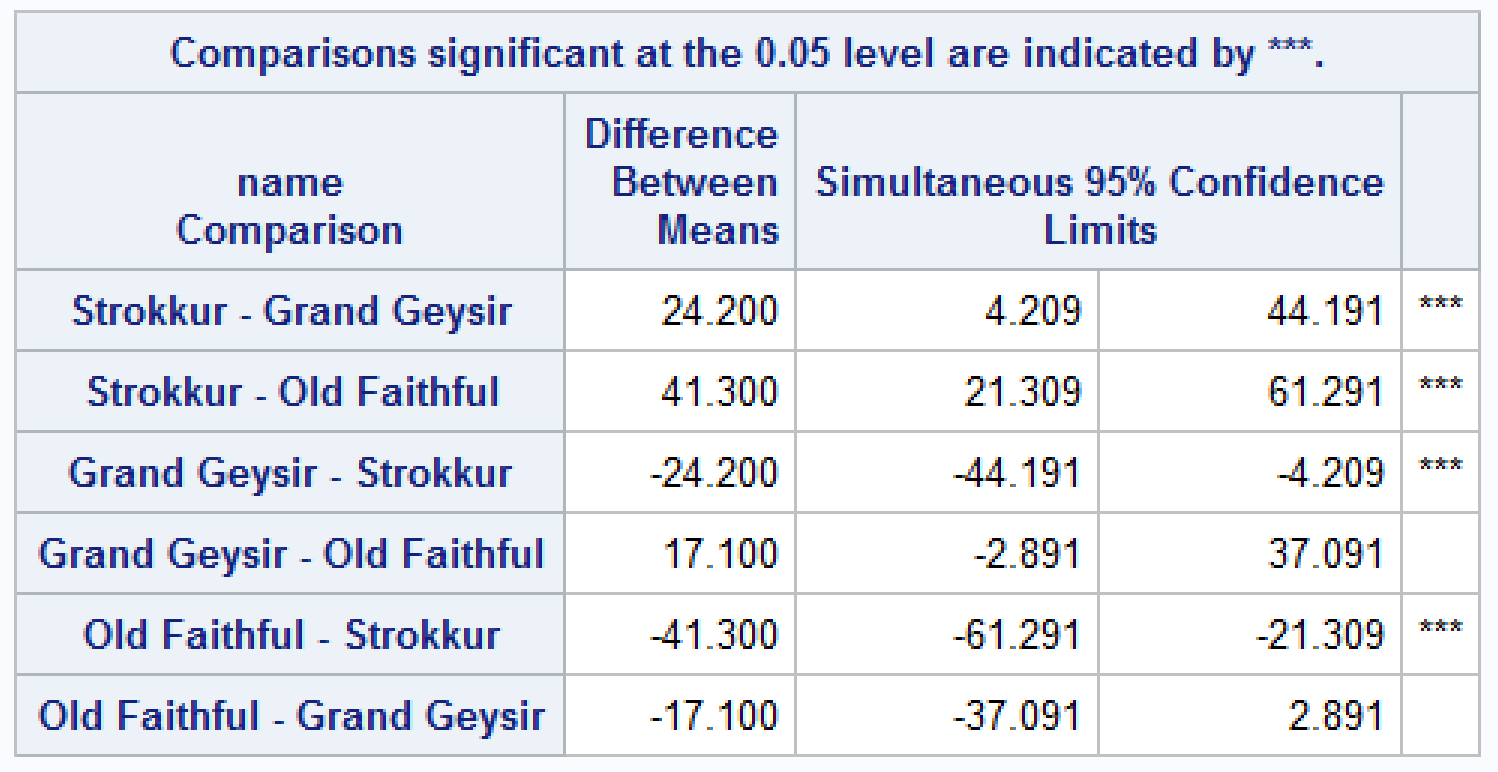
means name / CLM Tukey LSD CLDIFF;

**run**;

**quit**;

**Output:**





**Using the grouping method**

|  |  |  |
| --- | --- | --- |
| **Pair Comparison** | **Same Letter**  **(Yes or No)** | **Can conclude a significant difference? (Yes or No)** |
| Old Faithful vs Grand Geysir | Yes | No |
| Old Faithful vs Strokkur | No | Yes |
| Grand Geysir vs Strokkur | No | Yes |

**Using the confidence interval approach**

|  |  |  |  |
| --- | --- | --- | --- |
| **Pair Comparison** | **Confidence Interval** | **0 in the interval**  **Yes or No** | **Can conclude a**  **sig diff? (Yes or No)** |
| Old Faithful vs Grand Geysir | **(-37.091, 2.891)** | **Yes** | **No** |
| Old Faithful vs Strokkur | **(-61.291, -21.309)** | **No** | **Yes** |
| Grand Geysir vs Strokkur | **(-44.191, -4.209)** | **No** | **Yes** |

1. **Question 2**. Using R
   1. Test for normality for each of the geysers. (Use Anderson-Darling Test.)

> #Statistics 147 Assignment 6

> #Fall 2017

> #Sarah Ruckman

> #Section 002

> #R Question 2

> #Use a read.table function to read in the data

> geysers<-read.table("C:\\Users\\sarah\\Downloads\\geysers\_f17.dat",header=TRUE,skip=1)

> #Print as check

> geysers

OF GG S

1 125 150 195

2 153 146 180

3 140 161 205

4 135 178 210

5 165 150 174

6 184 190 204

7 189 185 200

8 136 176 175

9 146 165 170

10 137 180 210

> #Attach the columns to be used separately

> attach(geysers)

> #Use names function to see names

> names(geysers)

[1] "OF" "GG" "S"

> #Print as check

> OF

[1] 125 153 140 135 165 184 189 136 146 137

> GG

[1] 150 146 161 178 150 190 185 176 165 180

> S

[1] 195 180 205 210 174 204 200 175 170 210

> #Load package nortest

> local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)

+ if(nchar(pkg)) library(pkg, character.only=TRUE)})

> #Use ad.test() for OF, GG, and S to test for normal distributions

> ad.test(OF)

Anderson-Darling normality test

data: OF

A = 0.52794, p-value = 0.132

First, Old Faithful geyser:

* H0: Old Faithful eruption height is normally distributed
* Ha: Old Faithful eruption height is not normally distributed
* p-value = **0.132**
* Rejection Region: Reject H0 if p-value < α = 0.05
* Conclusion: **Since the p-value = 0.132 is greater than α = 0.05, we do not reject H0, it is reasonable to assume that Old Faithful’s eruption height is normally distributed.**

**R Code:**

> ad.test(GG)

Anderson-Darling normality test

data: GG

A = 0.35326, p-value = 0.3872

Next, Grand Geysir geyser:

* H0: Grand Geysir eruption height is normally distributed
* Ha: Grand Geysir eruption height is not normally distributed
* p-value = **0.3872**
* Rejection Region: Reject H0 if p-value < α = 0.05
* Conclusion: **Since the p-value = 0.3872 is greater than α = 0.05, we do not reject H0, it is reasonable to assume that the Grand Geysir’s eruption height is normally distributed.**

**R Code:**

> ad.test(S)

Anderson-Darling normality test

data: S

A = 0.56794, p-value = 0.1046

Finally, Strokkur geyser:

* H0: Strokkur eruption height is normally distributed
* Ha: Strokkur eruption height is not normally distributed
* p-value = **0.1046**
* Rejection Region: Reject H0 if p-value < α = 0.05
* Conclusion: **Since the p-value = 0.1046 is greater than α = 0.05, we do not reject H0, it is reasonable to assume that Strokkur’s eruption height is normally distributed.**
  1. Test for equality/homogeneity of variances. (Use Bartlett's test.)

Use α = 0.05.

**R Code:**

> #Next stack the data and make the colums accessible using the stack() and attach() function

> stack\_geysers<-stack(geysers)

> #Print as check

> stack\_geysers

values ind

1 125 OF

2 153 OF

3 140 OF

4 135 OF

5 165 OF

6 184 OF

7 189 OF

8 136 OF

9 146 OF

10 137 OF

11 150 GG

12 146 GG

13 161 GG

14 178 GG

15 150 GG

16 190 GG

17 185 GG

18 176 GG

19 165 GG

20 180 GG

21 195 S

22 180 S

23 205 S

24 210 S

25 174 S

26 204 S

27 200 S

28 175 S

29 170 S

30 210 S

> attach(stack\_geysers)

> name(stack\_geysers)

Error in name(stack\_geysers) : could not find function "name"

> names(stack\_geysers)

[1] "values" "ind"

> #Use bartlett.test(values,ind) to test for homogeneity of variances

> #values=data values and ind = classes

> bartlett.test(values,ind)

Bartlett test of homogeneity of variances

data: values and ind

Bartlett's K-squared = 1.1675, df = 2, p-value = 0.5578

* 
* 
* p-value = **0.5578**
* Rejection Region: Reject H0 if p-value < α = 0.05
* Conclusion: **Since the p-value = 0.5578 is greater than α = 0.05, we do not reject H0, it is reasonable to assume homogeneity of variances**
  1. Perform the appropriate test(s) of hypothesis to determine whether one can conclude that at least one of the geysers has a significantly different mean eruption height. Use α = 0.05.

**R Code:**

> #Use the aov function to generate an ANOVA

> #Format: aov(values~ind,data=stack\_geysers)

> results<-aov(values~ind,data=stack\_geysers)

> #Summarize results to view

> summary(results)

Df Sum Sq Mean Sq F value Pr(>F)

ind 2 8612 4306 13.25 9.8e-05 \*\*\*

Residuals 27 8777 325

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

* +  (cannot conclude that at least one of the geysers yields a significantly different mean eruption height)
  +  (can conclude that at least one of the geysers yields a significantly different mean eruption height)
  + p-value = **9.8 x 10-5**
  + Rejection Region: Reject H0 if p-value < α = 0.05
  + Conclusion: **Since the p-value = 9.8 x 10-5 is less than α = 0.05, we reject H0, it is reasonable to assume that at least one of the geysers yields a significantly different mean eruption height**

* 1. 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!)

**Using the p-value approach:**

**R Code:**

> #Since we rejected the null we need to use Tukey's test for multiple comparisons

> #Use TukeyHSD(results,conf.level=0.95)

> TukeyHSD(results,conf.level=0.95)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = values ~ ind, data = stack\_geysers)

$ind diff lwr upr p adj

GG-OF 17.1 -2.891982 37.09198 0.1044180

S-OF 41.3 21.308018 61.29198 0.0000635

S-GG 24.2 4.208018 44.19198 0.0153041

|  |  |  |  |
| --- | --- | --- | --- |
| **Pair Comparison** | **p-value** | **p-value < α = 0.05?**  **Yes or No** | **Can conclude a significant difference? (Yes or No)** |
| Old Faithful vs Grand Geysir | **0.1044180** | **No** | **No** |
| Old Faithful vs Strokkur | **0.0000635** | **Yes** | **Yes** |
| Grand Geysir vs Strokkur | **0.0153041** | **Yes** | **Yes** |