**Ali Sadeghinia**

**ANOVA Project**

**Question 3**

rm(list = ls())

battery=import('question3.xls')

attach(battery)

library(rio)

library(car)

names(battery)

colnames(battery)=tolower(make.names(colnames(battery)))

names(battery)

"sonoran" "minnesotan"

stacked.battery=stack(battery)

**Using the stack command allows us to put all the data into 2 columns with our dependent and independent variables. In R the columns are named as values and ind. Values is our dependent variable column and ind is our independent variable column.**

leveneTest(stacked.battery$values~stacked.battery$ind, data=stacked.battery)

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 1 0.0618 0.8041

112

**With the Levene test we can analyze the homogeneity of the data. Our null hypothesis is that there is a normality in the data set. Since our p-value is bigger than 0.05, we cannot reject the null hypothesis. Thus, the software is indicating that there is a normality or homogeneity in the data.**

battery.out=aov(stacked.battery$values~stacked.battery$ind, data = stacked.battery)

summary(battery.out)

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

stacked.battery$ind 1 6796 6796 245.6 <2e-16 \*\*\*

Residuals 112 3099 28

---

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

10 observations deleted due to missingness

**In this ANOVA test we can see that the p-value is well below 0.05 which rejects our null hypothesis. The null hypothesis in this analysis is that the means of the two batteries cannot be differentiated from each other and the grand mean.**

**Therefore, our alternate hypothesis stands and tells us that there are significant differences between the means (with 95% confidence level).**

battery.out$coefficients

(Intercept) stacked.battery$indminnesotan

49.30769 -15.50124

**The coefficients of the battery.out shows us the mean of the Sonoran batteries as intercept and the difference between the mean of the Minnesotan batteries – Sonoran batteries under indminnesotan.**

minnesotan.mean=mean(minnesotan)

minnesotan.mean

33.80645

sonoran.mean=mean(sonoran, na.rm = TRUE)

sonoran.mean

49.30769

minnesotan.mean-sonoran.mean

-15.50124

**This is just to show the difference in the means of the two battery conditions. we can see that the 2 numbers we got with the coefficients data of the ANOVA test and the difference in the means is the same.**

tukey.battery=TukeyHSD(battery.out)

tukey.battery

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = stacked.battery$values ~ stacked.battery$ind, data = stacked.battery)

$`stacked.battery$ind`

diff lwr upr p adj

minnesotan-sonoran -15.50124 -17.46101 -13.54147 0

**Our p-value or p adjusted value is zero. Rejects our null hypothesis: that there is no difference between the two batteries’ lifespan means. Thus, our alternate hypothesis stands; there is certainly differences between the lifespan of the batteries in the different conditions of weather (with 95% confidence level).**

par(mar=c(5.1,10,4.1,2.1))

plot(tukey.battery, las=2)

![Chart, box and whisker chart

Description automatically generated]()

**Tukey test was not needed for this data and analyzing the data because there are only Minnesotan and Sonoran batteries in our data set.**

**Conclusion**

**There is a normality in the data due to the similarities between each battery lifespan group (Minnesotan and Sonoran). The analysis of variance shows us that the means are significantly different.**

**The Sonoran batteries last longer.**

**This is One-Way ANOVA Test. Two-Way, we can add hot and cold in another column and analyze whether the factors of weather with two levels are also of impact.**

**Question 5**

rm(list = ls())

Photographer=import('question5.xls')

attach(Photographer)

library(rio)

library(car)

names(Photographer)

colnames(Photographer)=tolower(make.names(colnames(Photographer)))

names(Photographer)

"photographer" "time"

leveneTest(Photographer$time~Photographer$photographer, data = Photographer)

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 2 4.1483 0.01612 \*

842

---

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

**Our null hypothesis in the Levene test is that the data is homogeneous. With the p-value below 0.05 we can reject that the data is homogeneous or there is a normality.**

**The data is not homogeneous, or software cannot find normality in the data.**

photo.out=aov(Photographer$time~Photographer$photographer, data = Photographer)

summary(photo.out)

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

Photographer$photographer 2 18671 9335 2162 <2e-16 \*\*\*

Residuals 842 3635 4

---

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

**The null hypothesis is that there is no difference between the means of our data. However, with the p-value that we got from the ANOVA test, we can say that the alternate hypothesis is true.**

**There are differences in the means of the times for each photographer.**

photo.out$coefficients

(Intercept) Photographer$photographerEddie Marrs

5.7753968 11.0326600

Photographer$photographerVivian Rutledge

0.2931896

**We can see that there are differences between the means of the times of the photographers.**

**Intercept is the time mean of Carmen Sternwood, alphabetically sorted. The other numbers are the differences of the means of Carmen’s times and the respective photographer.**

tukey.photo=TukeyHSD(photo.out)

tukey.photo

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Photographer$time ~ Photographer$photographer, data = Photographer)

$`Photographer$photographer`

diff lwr upr p adj

Eddie Marrs-Carmen Sternwood 11.0326600 10.5774350 11.4878851 0.0000000

Vivian Rutledge-Carmen Sternwood 0.2931896 -0.1027144 0.6890935 0.1914205

Vivian Rutledge-Eddie Marrs -10.7394705 -11.1579089 -10.3210321 0.0000000

**With Tukey test we can see which photographers the software has been able to find differences between means of their times (with 95% confidence level).**

**Vivian Rutledge-Carmen Sternwood**

**The software cannot identify the differences between these two photographers. The p-value indicates that we cannot reject the null hypothesis since it is bigger than 0.05. The null hypothesis is that there is no difference between the means of the two photographers.**

**We fail to reject the null hypothesis in this one case.**

par(mar=c(5.1,15,4.1,2.1))

plot(tukey.photo, las = 2)

![A picture containing chart

Description automatically generated]()

**The confidence interval crossing the 0 vertical line is the one that the software is telling us that it could not find significant differences between the two photographers time means (with 95% confidence level).**

**Conclusion**

**The Levene test’s p-value is below 0.05 and rejects the null; the data is not homogeneous.**

**Analysis of Variance also has a below 0.05 p-value which means that there are differences between at least one factor-level is different.**

**Found more in detail which ones are different and which ones are not so different with the Tukey test.**

**Question 8**

rm(list = ls())

cities=import('question8.xls')

attach(cities)

library(rio)

library(car)

names(cities)

[1] "chicago" "los.angeles" "miami" "minneapolis"

colnames(cities)=tolower(make.names(colnames(cities)))

stacked.cities=stack(cities)

stacked.cities

**Stack command allows us to stack the data into 2 columns: dependent and independent variables.**

leveneTest(stacked.cities$values~stacked.cities$ind, data = stacked.cities)

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 3 2.9495 0.03224 \*

572

---

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

**P-value is below 0.05 which means that our null hypothesis is rejected. Our null hypothesis with the Levene test is that our variables are homogeneous. Thus, we can confirm that our data is not normal/homogeneous.**

cities.out=aov(stacked.cities$values~stacked.cities$ind, data = stacked.cities)

summary(cities.out)

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

stacked.cities$ind 3 1447922 482641 102.7 <2e-16 \*\*\*

Residuals 572 2687280 4698

---

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

**The ANOVA test also shows us a very significant p-value. This also rejects the null hypothesis.**

**The null hypothesis for the ANOVA test is that there is no difference between the means of TV times and the grand mean of all.**

**Therefore, we can reject the null and support the alternate hypothesis that there are differences in the means(At least 1 variance has a different mean than others).**

tukey.cities=TukeyHSD(cities.out)

tukey.cities

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = stacked.cities$values ~ stacked.cities$ind, data = stacked.cities)

$`stacked.cities$ind`

diff lwr upr p adj

los.angeles-chicago -88.625000 -109.43821 -67.811791 0.0000000

miami-chicago -103.472222 -124.28543 -82.659013 0.0000000

minneapolis-chicago 7.097222 -13.71599 27.910431 0.8159645

miami-los.angeles -14.847222 -35.66043 5.965987 0.2565646

minneapolis-los.angeles 95.722222 74.90901 116.535431 0.0000000

minneapolis-miami 110.569444 89.75624 131.382653 0.0000000

**With the information from the Tukey test, we can see that the software finds differences with 95% confidence level between:**

**los.angeles-chicago (difference was recognized between the 2)**

**miami-chicago (difference was recognized between the 2)**

**minneapolis-los.angeles (difference was recognized between the 2)**

**minneapolis-miami (difference was recognized between the 2)**

**In other words, in those cases, the null hypothesis is rejected.**

**The null hypothesis: there is no difference between the means of each of them.**

**The software can identify differences between the TV time means of all of them.**

**On the other hand, we can see that the software was not able to identify the differences with 95% confidence interval between:**

**minneapolis-chicago**

**miami-los.angeles**

**Thus, we cannot reject the null hypothesis. The one with lower p-value is the one with bigger difference between the means.**

par(mar=c(5.1,12,4.1,2.1))

plot(tukey.cities, las = 2)

![Chart, box and whisker chart

Description automatically generated]()

**As it can be seen in the Tukey plot as well; the confidence intervals crossing the 0-vertical line are the ones that the software was not able to recognize a significant difference between the cities’ TV means.**

**Conclusion**

**Levene test equality of variances test has a p-value of below 0.05; There is at least one factor-level that is not homogeneous in the data (more than one for this case).**

**ANOVA test also has a p-value of below 0.05; There is at least one factor-level that there are differences between the means of the tv times (more than one for this case as well).**

**Tukey test shows the differences more in detail.**

**The ones in the plot above crossing the 0 vertical line are the ones that the software could not find significant differences between statistically.**