Homework 7 Wenxin Liang wl2455

**Question 1.**

Consider the Duodenal Ulcer data in Problem 25, Chapter 5

Duodenal Ulcers. To clarify the importance of a certain kind of antibody activity (CCK) in gastrointestinal diseases, researchers assessed the CCK activity in the duodenal mucosa of 27 guinea pigs. Of these, 8 had gallstones, 8 had gastric ulcers, and 9 were healthy controls. The following CCK activity was determined by bioassay and measured in Ivy units per milligram of dry weight. (Data from S. Kataoka et al., “Bioassay of Cholecystokinin-Pancreozymin in Duodenal Mucosa”, Lancet (1978): 1043.)

**a. Using an appropriate ANOVA model, determine whether there is a significant different among the group means. Use both an F test and simultaneous confidence interval procedures.**

Based on the data we have in Question 25 chapter 5, we can use the one-way ANOVA as the appropriate ANOVA model.

Therefore, the F test we got,

> data <- c(0.11,0.11,0.11,0.19,0.21,0.22,0.24,0.25,0.31,0.18,0.27,

+ 0.36,0.37,0.39,0.47,0.37,0.57,0.29,0.30,0.40,0.45,0.47,

+ 0.52,0.57,1.10)

> team <- rep(c("Controls","Gallstone","Ulcer"),c(9,8,8))

> team <- factor(team)

> # For F test

> anova1 <- aov(data ~ team)

> summary(anova1)

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

team 2 0.4328 0.21641 7.949 0.00252 \*\*

Residuals 22 0.5989 0.02722

---

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

The hypothesis for F test are:

Since the P value = 0.00252 < 0.05 then we can reject the null hypothesis which is . We conclude that not all the mean are equaled which means there is a significant difference among the group means.

For the simultaneous confidence interval procedures we use several tests to get the conclusion,

> # For simultaneous confidence interval procedures

> # Bonferroni Procedure

> pairwise.t.test(data\_new$data,data\_new$team,p.adj="bonferroni")

Pairwise comparisons using t tests with pooled SD

data: data\_new$data and data\_new$team

Controls Gallstone

Gallstone 0.111 -

Ulcer 0.002 0.311

P value adjustment method: bonferroni

Based on the Bonferroni adjustment, we observed only the Ulcer-Controls comparison is statistically significant. This suggests that the mean of group Gastric Ulcer is statistically different from the Healthy Controls group and the Gallstone group but there is insufficient statistical support to distinguish between Ulcer Gallstone or Gallstone and Controls.

> # Scheffe Procedure

> library(agricolae)

> (scheffe.test(anova1,"team",alpha=0.05,group=TRUE,main = NULL,console =FALSE))

$statistics

Mean CV MSerror CriticalDifference

0.3532 46.71466 0.02722374 0.212449

$parameters

Df ntr F Scheffe

22 3 3.443357 2.624255

$means

data std r Min Max

Controls 0.1944444 0.07143373 9 0.11 0.31

Gallstone 0.3725000 0.11744300 8 0.18 0.57

Ulcer 0.5125000 0.25677951 8 0.29 1.10

$comparison

NULL

$groups

trt means M

1 Ulcer 0.5125000 a

2 Gallstone 0.3725000 ab

3 Controls 0.1944444 b

Based on the Scheffe procedure, we observed only the Ulcer is exactly marked in-group a. This suggests that the mean of group Gastric Ulcer is statistically different from the Healthy Controls group and the Gallstone group. Since Gallstone marked as ab then there there is insufficient statistical support to distinguish between Ulcer Gallstone or Gallstone and Controls. The same conclusion we obtain from the Bonferroni procedure.

> # Tukey HSD Procedure

> TukeyHSD(anova1)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = data ~ team)

$team

diff lwr upr p adj

Gallstone-Controls 0.1780556 -0.02334592 0.3794570 0.0897039

Ulcer-Controls 0.3180556 0.11665408 0.5194570 0.0018171

Ulcer-Gallstone 0.1400000 -0.06724040 0.3472404 0.2286986

Only the ulcer-control p-value=0.0018171<0.05, which the comparison is statistically significant. This suggests that the Gastric Ulcer is superior to the Healthy Controls and there is a statistically difference between the Ulcer group mean and the Healthy Control group mean but there is insufficient statistical support to distinguish between Ulcer Gallstone or Gallstone and Controls since there is no significant difference from the other two groups Gallstone-Control, and Ulcer-Gallstone. The same result we had in the Bonferroni procedure and Scheffe procedure.

Therefore, based on both F test and simultaneous confidence interval procedures, we conclude that there is a significant difference among group means.

**b. Assess the assumptions of the ANOVA model**

Based on our one-way ANOVA model we have three assumptions. First we assume the data from each group are independent. Second the residuals are normal distributed. Third there is equal variance among each group.

For the first assumption since all the data from each group are under different condition, since of the 27 guinea pigs, 8 had gallstones, 8 had gastric ulcers, and 9 were healthy controls. Therefore, we can conclude that the first assumption of independence is satisfied.

For the second assumption, we will use Shapiro test to check.

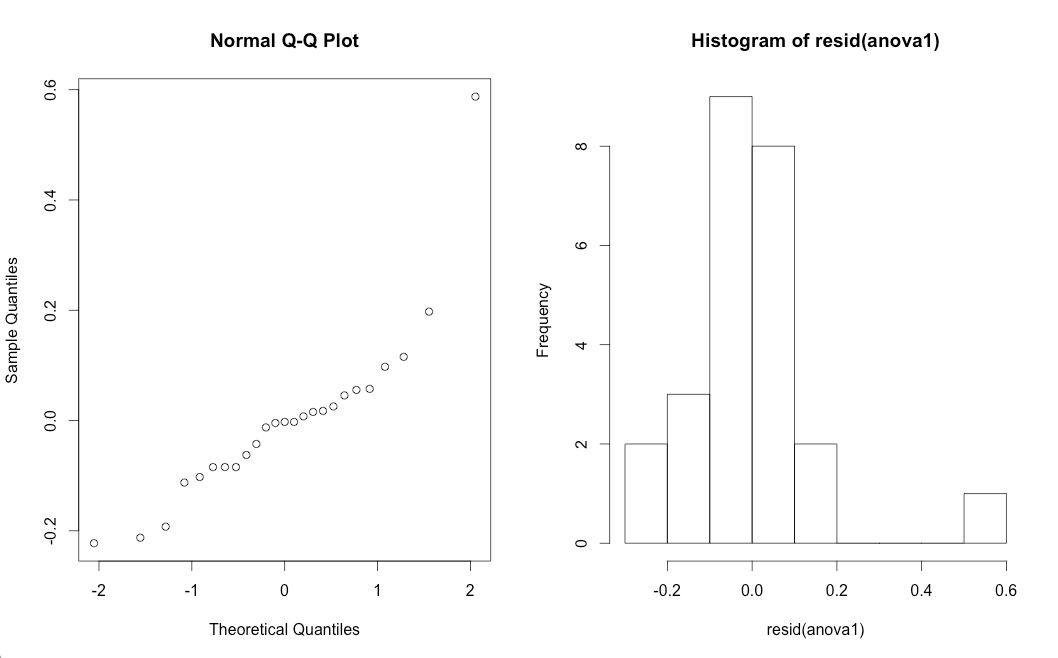
> shapiro.test(resid(anova1))

Shapiro-Wilk normality test

data: resid(anova1)

W = 0.8155, p-value = 0.0004176

Since p-value=0.0004176<0.05. We reject the null, and then we conclude that the residuals are not normally distributed. We can also see that from the qqplot and histogram.



For the third assumption, since bartlett’s test is highly dependent on normality assumption so we cannot use the bartlett’s test here.

We can use the Levene-type test to check,

> levene.test(cck,group)

modified robust Brown-Forsythe Levene-type test based on the absolute deviations

from the median

data: cck

Test Statistic = 1.3046, p-value = 0.2915

We use levene test to verify, since the p-value=0.2915 > 0.05, so we fail to reject the null hypothesis, which is the variance is equal. Therefore, the results show that there does not exist unequal variance.

**c. Compare the results to those obtained using a non-parametric procedure**

Since using a non-parametric procedure we use the Kruskal- Wallis test to check.

kruskal.test(data,team)

Kruskal-Wallis rank sum test

data: data and team

Kruskal-Wallis chi-squared = 13.8673, df = 2, p-value = 0.0009744

Since the p-value=0.0009744<0.05, we will reject null hypothesis. We conclude that there is a significant difference among group means, which is the same result in part a.

Question 2

Consider the IQ scores data of Display 13.24, problem 19, Chapter 13.

A 1989 study investigated the effect of heredity and environment on intelligence. From adoption registers in France, researchers selected samples of adopted children whose biological parents and adoptive parents came from either the very highest or the very lowest socio- economic status (SES) categories (based on years of education and occupation). They attempted to obtain samples of size 10 from each combination: (1) high adoptive SES and high biological SES, (2) high adoptive SES and low biological SES, (3) low adoptive SES and high biological SES, and (4) low SES for both parents. It turned out, however, only eight children belonged to combination three. The 38 selected children were given intelligence quotient (IQ) tests. The scores are reported in Display 13.24. (Data from C. Capron and M. Duyme, “Children’s IQs and SES of Biological and Adoptive Parents in a Balanced Cross-fostering Study,” *European Bulletin of Cognitive Psychology* 11(3) (1991): 323–48.) Does the difference in mean scores for those with high and low SES biological parents depend on whether the adoptive parents were high or low SES? If not, how much is the mean IQ score affected by the SES of adoptive parents, and how much is it affected by the SES of the biological parents? Is one of these effects larger than the other? Analyze the data and write a report of the findings.

**a. Do problem 19**

**1. Does the difference in mean scores for those with high and low SES biological parents depend on whether the adoptive parents were high or low SES?**

Based on the data structure we have, we perform two-way ANOVA.

> summary(aov(ex1319$IQ~ex1319$Adoptive\*ex1319$Biological))

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

ex1319$Adoptive 1 1478 1477.6 8.456 0.006366 \*\*

ex1319$Biological 1 2291 2291.5 13.114 0.000944 \*\*\*

ex1319$Adoptive:ex1319$Biological 1 2 1.9 0.011 0.917437

Residuals 34 5941 174.7

---

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

Since the p-value of interaction term=0.917437>0.05 is not significant and the other two terms are significant which 0.006<0.05 and 0.0009<0.05 so that the difference in mean scores for those with high and low bio does not depend on whether the adoptive parents were high or low. Here we will eliminate the interaction term, and then we will fit the Ordinary Least Square model.

**2. If not, how much is the mean IQ score affected by the SES of adoptive parents, and how much is it affected by the SES of the biological parents? Is one of these effects larger than the other?**

We use the Tukey HSD procedure to check the influence,

> anova3=aov(ex1319$IQ~ex1319$Adoptive\*ex1319$Biological)

> TukeyHSD(anova3)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = ex1319$IQ ~ ex1319$Adoptive \* ex1319$Biological)

$`ex1319$Adoptive`

diff lwr upr p adj

Low-High -12.48889 -21.21688 -3.760901 0.0063663

$`ex1319$Biological`

diff lwr upr p adj

Low-High -15.5284 -24.25638 -6.800407 0.0009593

$`ex1319$Adoptive:ex1319$Biological`

diff lwr upr p adj

Low:High-High:High -12.1 -29.03486 4.83486428 0.2349558

High:Low-High:High -16.0 -31.96634 -0.03365684 0.0493567

Low:Low-High:High -27.2 -43.16634 -11.23365684 0.0003149

High:Low-Low:High -3.9 -20.83486 13.03486428 0.9243234

Low:Low-Low:High -15.1 -32.03486 1.83486428 0.0946322

Low:Low-High:Low -11.2 -27.16634 4.76634316 0.2493435

From the outcome we can see that when we fixed the adoptive parents SES is “High”, which is the term named “High: Low –High: High”, and the p-value of it is 0.0493567 < 0.05, shows that the biological parents SES differ significantly when the adoptive parents SES is “High”.

The same as before when we fixed the adoptive parents SES is “Low”, which means the term “Low: Low –Low: High”, whose p-value is 0.0946322 < 0.05, indicating that the biological parents SES differ not significantly when the adoptive parents SES is “Low”.

To sum up, the biological parents SES differ significantly when adoptive parents SES “High”, but differ not significantly when adoptive parents SES “Low”.

> anova4=aov(ex1319$IQ~ex1319$Biological\*ex1319$Adoptive)

> TukeyHSD(anova4)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = ex1319$IQ ~ ex1319$Biological \* ex1319$Adoptive)

$`ex1319$Biological`

diff lwr upr p adj

Low-High -16.22222 -24.95021 -7.494234 0.0006102

$`ex1319$Adoptive`

diff lwr upr p adj

Low-High -11.58765 -20.31564 -2.859666 0.0107774

$`ex1319$Biological:ex1319$Adoptive`

diff lwr upr p adj

Low:High-High:High -16.0 -31.96634 -0.03365684 0.0493567

High:Low-High:High -12.1 -29.03486 4.83486428 0.2349558

Low:Low-High:High -27.2 -43.16634 -11.23365684 0.0003149

High:Low-Low:High 3.9 -13.03486 20.83486428 0.9243234

Low:Low-Low:High -11.2 -27.16634 4.76634316 0.2493435

Low:Low-High:Low -15.1 -32.03486 1.83486428 0.0946322

From the outcome we can see that when we fixed the biological parents SES is “High”, which is the term named “High: Low –High: High”, and the p-value of it is 0.2349558 > 0.05, shows that the biological adoptive SES is not differ significantly when the biological parents SES is “High”.

The same as before when the biological parents SES is “Low” is fixed, which is the term named “Low: Low –Low: High”, and the p-value of it is 0.2493435 > 0.05, shows that the biological adoptive SES is not differ significantly when the biological parents SES is “High”.

As a result, there is no doubt that the difference in mean scores for those with high and low SES biological parents depend on whether the adoptive parents were high or low SES.

To sum up, the adoptive parents SES are not differ significantly when adoptive parents SES “High”, also are not differ significantly when adoptive parents SES “Low”.

In conclusion, the biological parents SES depend on the adoptive parents, but the adoptive parents SES don’t depend on the biological parents

> anova3=aov(ex1319$IQ~ex1319$Adoptive\*ex1319$Biological)

> ano=as.data.frame(summary(anova3)[[1]])

> inf=rbind(ano[1,2]/sum(ano[,2]),ano[2,2]/sum(ano[,2]),ano[3,2]/sum(ano[,2]),ano[4,2]/sum(ano[,2]))

> colnames(inf)=c("Accountable Percentage")

> rownames(inf)=c("adoptive","biological","adoptive&biological","Residuals")

> inf

Accountable Percentage

adoptive 0.1521417544

biological 0.2359372142

adoptive&biological 0.0001962357

Residuals 0.6117247957

By definition, Sum Sq is the sum of squared of deviations. The Sum Sq of a predictor means how much variance (deviation) of the response variable this predictor could account for. We can calculate the percentages of each predictor’s Sum Sq of the total Sum Sq and integrate them into the form above.

From the outcome we could easily see that the “adoptive” accounts for 15.21% of the total variance and the “biological” accounts for 23.59% of the total variance. We can conclude that numerically, the effect of biological parents SES is larger than the effect of the adoptive parents SES.

**b. Assess the validity of all assumptions.**

Based on our one-way ANOVA model we have three assumptions. First we assume the data from each group are independent. Second the residuals are normal distributed. Third there is equal variance among each group.

For the first assumption since all the data from each group are under different condition, since from adoption registers in France, researchers selected samples of adopted children whose biological parents and adoptive parents came from either the very highest or the very lowest socio- economic status (SES) categories (based on years of education and occupation). Therefore, we can conclude that the first assumption of independence is satisfied.

For the second assumption, we will use Shapiro test to check.

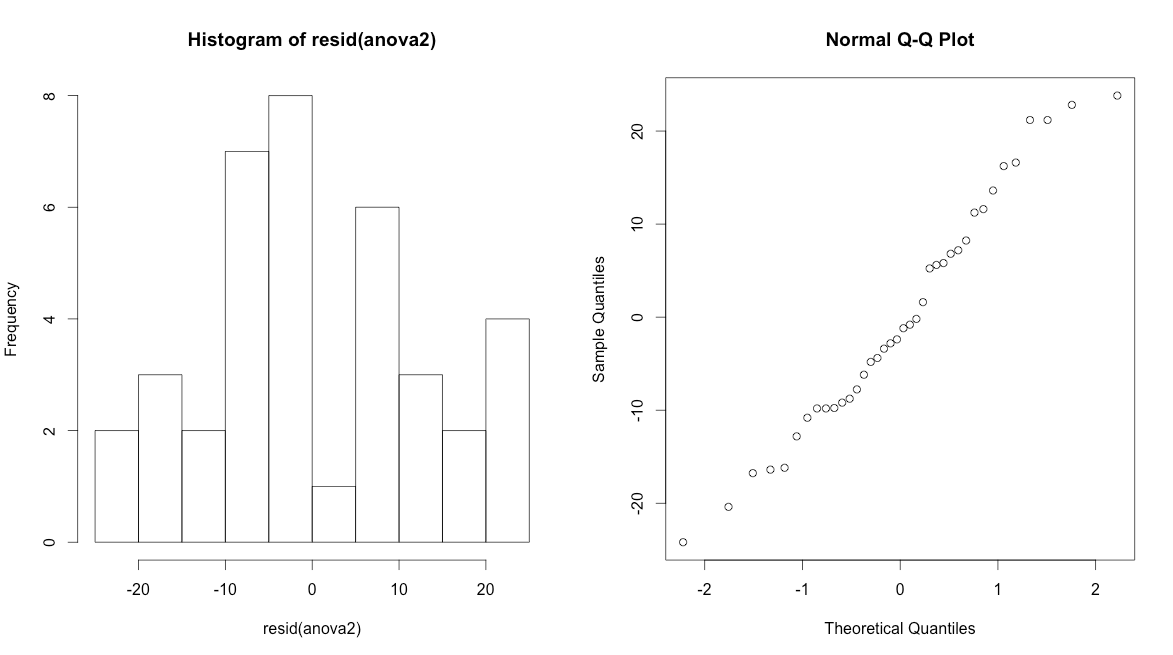
> shapiro.test(resid(anova2))

Shapiro-Wilk normality test

data: resid(anova2)

W = 0.9716, p-value = 0.437

Since 0.437 > 0.05 we fail to reject the null hypothesis, we conclude the residuals are normally distributed. Also we can use see the qqplot and histogram to check.



For the third assumption, we use the Bartlett test to check the homogeneity of variance.

> bartlett.test(split(IQ,list(ado,bio)))

Bartlett test of homogeneity of variances

data: split(IQ, list(ado, bio))

Bartlett's K-squared = 0.699, df = 3, p-value = 0.8734

Since see from the normal test that the residuals are normally distributed, so we can use bartlett’s test to test whether the variance are equal. The p-value for bartlett’s test is 0.8734>0.05, so we fail to reject the null hypothesis, we conclude that the variance are equal.

The following code:

# Homework 7

# Question 1

# Part a

data <- c(0.11,0.11,0.11,0.19,0.21,0.22,0.24,0.25,0.31,0.18,0.27,

0.36,0.37,0.39,0.47,0.37,0.57,0.29,0.30,0.40,0.45,0.47,

0.52,0.57,1.10)

team <- rep(c("Controls","Gallstone","Ulcer"),c(9,8,8))

team <- factor(team)

data\_new <- data.frame(team,data)

# For F test

anova1 <- aov(data ~ team)

summary(anova1)

# For simultaneous confidence interval procedures

# Bonferroni Procedure

pairwise.t.test(data\_new$data,data\_new$team,p.adj="bonferroni")

# Scheffe Procedure

library(agricolae)

(scheffe.test(anova1,"team",alpha=0.05,group=TRUE,main = NULL,console =FALSE))

# Fisher's Least Significant Difference

df <- df.residual(anova1)

MSerror <- deviance(anova1)/df

(LSD.test(data\_new$data,data\_new$team,df,MSerror))

# Tukey HSD Procedure

TukeyHSD(anova1)

# Dunnett's Procedure

#library(multcomp)

#Dunnett\_Test <- glht(anova1)

#summary(Dunnett\_Test)

# Part b

bartlett.test(data ~ team)

shapiro.test(resid(anova1))

Controls=c(0.11,0.11,0.11,0.19,0.21,0.22,0.24,0.25,0.31)

Gallstone=c(0.18,0.27,0.36,0.37,0.39,0.47,0.37,0.57)

Ulcer=c(0.29,0.30,0.40,0.45,0.47,0.52,0.57,1.10)

par(mfrow=c(1,3))

#plot(density(Controls),main="CCK level of Controls Group")

#plot(density(Gallstone),main="CCK level of Gallstone Group")

#plot(density(Ulcer),main="CCK level of Ulcer Group")

#plot(density(team),main="CCK level of Groups")

#shapiro.test(Controls)$p.value

#shapiro.test(Gallstone)$p.value

#shapiro.test(Ulcer)$p.value

par(mfrow=c(1,2))

qqnorm(resid(anova1))

hist(resid(anova1))

# Part c

kruskal.test(data,team)

# Question 2

# Part a

# sub question 1

library("Sleuth3")

summary(aov(ex1319$IQ~ex1319$Adoptive\*ex1319$Biological))

mylm <- lm(ex1319$IQ~ex1319$Adoptive+ex1319$Biological)

summary(mylm)

anova2=aov(ex1319$IQ~ex1319$Adoptive+ex1319$Biological)

anova3=aov(ex1319$IQ~ex1319$Adoptive\*ex1319$Biological)

TukeyHSD(anova3)

anova4=aov(ex1319$IQ~ex1319$Biological\*ex1319$Adoptive)

TukeyHSD(anova4)

ano=as.data.frame(summary(anova3)[[1]])

inf=rbind(ano[1,2]/sum(ano[,2]),ano[2,2]/sum(ano[,2]),ano[3,2]/sum(ano[,2]),ano[4,2]/sum(ano[,2]))

colnames(inf)=c("Accountable Percentage")

rownames(inf)=c("adoptive","biological","adoptive&biological","Residuals")

inf

# Part b

# assumption

par(mfrow=c(1,1))

plot(mylm)

shapiro.test(resid(anova2))

par(mfrow=c(1,2))

hist(resid(anova2))

qqnorm(resid(anova2))

bartlett.test(ex1319$IQ~ex1319$Adoptive+ex1319$Biological)