Stat 4201 Homework 7

Mengqi Zong < $mz2326@columbia.edu$ >

March 18, 2012

Question 1

a)

The F-test shows that there is a significant difference among the group means. Here is the output from R:

```
            Df  Sum Sq  Mean Sq F value   Pr(>F)
Disease      2 0.43282 0.216411  7.9493 0.002522 **
Residuals   22 0.59892 0.027224
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Now we try to calculate the simultaneous confidence interval of Bonferroni Method. Here is the output from MATLAB:

```
    'Controls'      [1]     [2]     [-0.3858]      [-0.1781]      [ 0.0297]
    'Gallstone'     [1]     [3]     [-0.5258]      [-0.3181]      [-0.1103]
    'Ulcer'         [2]     [3]     [-0.3538]      [-0.1400]      [ 0.0738]
```

As we can see, the only confidence interval that does not contain zero is Conrtols-Ulcer: $[-0.5258, -0.1103]$. This indicates that the means of Controls and Ulcer are different.

b)

For the normality assumption, we use q-q plot of residuals to do the analysis. The plot is shown in Fig-1. As we can see, the normality assumption holds.
For the equal variance assumption, I use Bartlett's test. Here is the output from R:

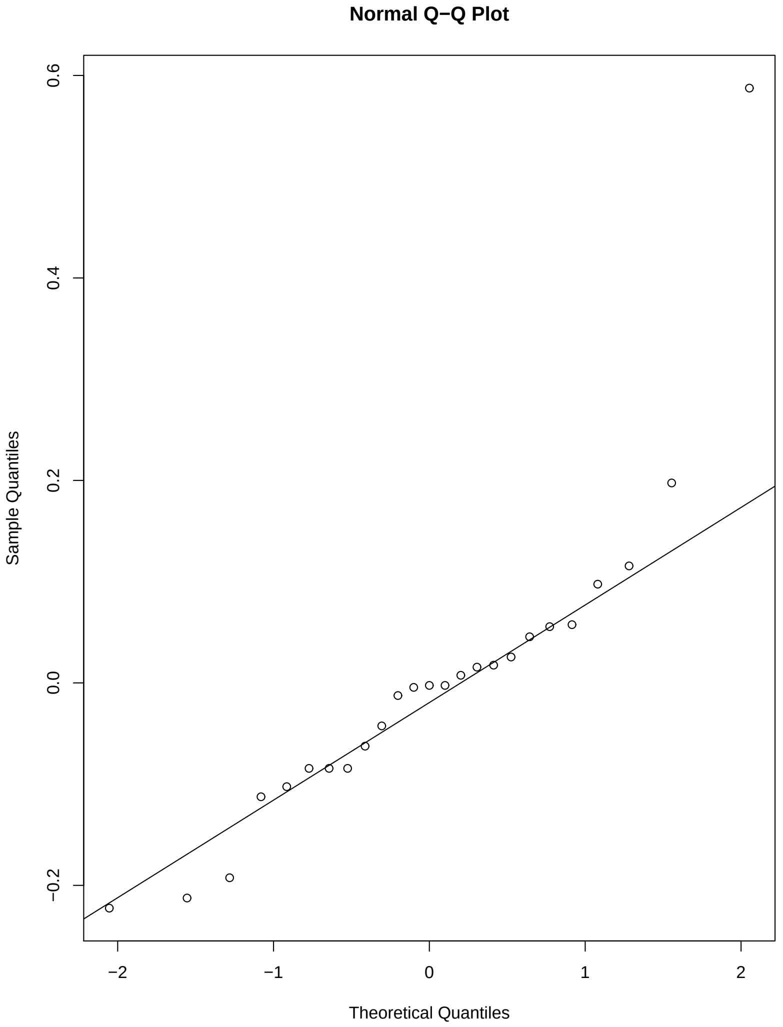

Figure 1: P1: q-q plot for residuals

```
Bartlett test of homogeneity of variances

data:  data.p1$CCK and data.p1$Disease
Bartlett's K-squared = 11.2755, df = 2, p-value = 0.003561
```

As we can see, variance is not equal. However, I found the following opinion on the Internet:

“Hypothesis testing is the wrong tool to use to asses the validity of model assumptions. If the sample size is small, you have no power to detect any variance differences, even if the variance differences are large. If you have a large sample size you have power to detect even the most trivial deviations from equal variance, so you will almost always reject the null. Simulation studies have shown that preliminary testing of model assumption leads to unreliable type I errors.
Looking at the residuals across all cells is a good indicator, or if your data are normal, you can use the AIC or BIC with/without equal variances as a selection procedure.”
As a result, since the sample size is small, the unequal variance test is not reliable. Therefore, we don’t need to test the unequal variance.

c)

The result from Kruskal-Wallis rank sum test is

```
Kruskal-Wallis rank sum test

data:  data.p1$CCK and data.p1$Disease
Kruskal-Wallis chi-squared = 13.8673, df = 2, p-value = 0.0009744
```

As we can see, both parametric and non-parametric methods reject the null hypothesis.

Question 2

a)

I use the two-way ANOVA to o do the analysis. Here is the output of the F-test from R:

```
                  Df Sum Sq Mean Sq F value    Pr(>F)
ADOPTIVE           1 1477.6 1477.63  8.4561 0.0063663 **
BIOLOGIC           1 2291.5 2291.47 13.1135 0.0009445 ***
ADOPTIVE:BIOLOGIC  1    1.9    1.91  0.0109 0.9174370
Residuals         34 5941.2  174.74
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

And the coefficient is:

```
            (Intercept)             ADOPTIVELow             BIOLOGICLow
                  119.6                   -12.1                   -16.0
ADOPTIVELow:BIOLOGICLow
                    0.9
```

As we can see, there is no interaction term. And the SES of biological parents effects larger than the SES of adoptive parents (16.0 : 12.1).

b)

For the normality assumption, we use q-q plot of residuals to do the analysis. The plot is shown in Fig-2. As we can see, the normality assumption holds.

Appendices

The R code is listed below:

```
# Problem 1
data.p1 <- read.table("ex0525", header=TRUE)

aov.p1 <- aov(CCK~Disease, data = data.p1)
summary(aov.p1)

postscript(file="~/Documents/LaTeX/stat4201-hmwk7/qq.eps",
           onefile=FALSE, horizontal=FALSE)
qqnorm(resid(aov.p1)); qqline(resid(aov.p1), color = 2)
dev.off()

bartlett.test(data.p1$CCK, data.p1$Disease)

kruskal.test <- kruskal.test(data.p1$CCK, data.p1$Disease)

# Problem 2
data.p2 <- read.csv("ex1319.csv", header=TRUE)

aov.p2 <- aov(IQ~ADOPTIVE*BIOLOGIC, data = data.p2)
summary(aov.p2)
```

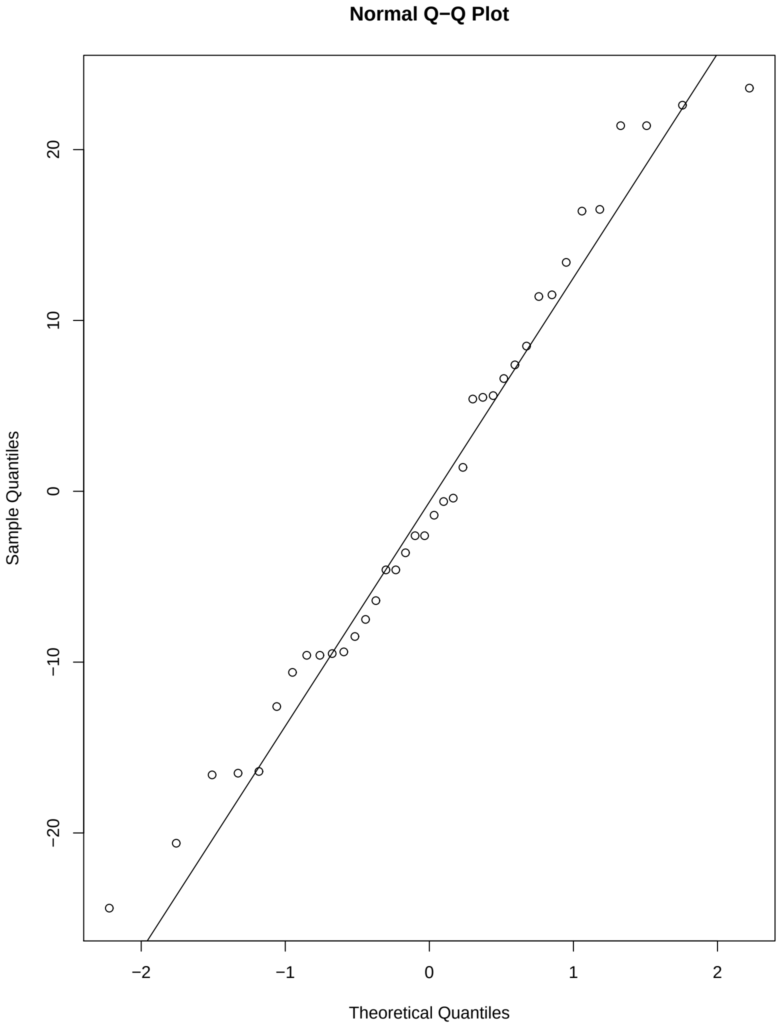

Figure 2: P2: q-q plot for residuals

```
postscript(file="~/Documents/LaTeX/stat4201-hmwk7/qq2.eps",
           onefile=FALSE, horizontal=FALSE)
qqnorm(resid(aov.p2)); qqline(resid(aov.p2), color = 2)
dev.off()
```

```
bartlett.test(data.p1$CCK, data.p1$Disease)
```

The MATLAB code is listed below:

```
CCK = [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]';
Disease = char('Controls', 'Controls', 'Controls', 'Controls',...
               'Controls', 'Controls', 'Controls', 'Controls',...
               'Controls', 'Gallstone', 'Gallstone', 'Gallstone',...
               'Gallstone', 'Gallstone', 'Gallstone', 'Gallstone',...
               'Gallstone', 'Ulcer', 'Ulcer', 'Ulcer', 'Ulcer', 'Ulcer',...
               'Ulcer', 'Ulcer', 'Ulcer');

[p,t,st] = anova1(CCK, Disease,'off');
[c,m,h,nms] = multcompare(st, 'display', 'off', 'ctype', 'bonferroni');
[nms num2cell(c)]
```