STAT 210
Applied Statistics and Data Analysis
First Exam

October 23, 2021

This exam is open notes and open book but not open internet. You are not allowed to surf the internet or look for answers to the questions

You are reminded to adhere to the academic integrity code established at KAUST.

Show complete solutions to get full credit. Label your graphs appropriately

Please identify the files you submit with your surname

Question 1 (40 points)

The file `theater.csv` has information on a survey conducted on visitors of a local Swiss theater in terms of age (`Age`), sex (`Sex`), annual income (`Income`), general expenditure on cultural activities (`Culture`), expenditure on theater visits (`Theatre`), and the estimated expenditure on theater visits in the year before the survey was done (`Theatre_ly`).

Note: Sex is coded 1 for females, and theater is spelled `Theatre`.

(a) Load the data into a data frame called `q1.df`. Divide the plotting window into two columns and draw boxplots for `Culture` and `Theatre` as a function of `Sex`. Comment on the results.

(b) The year before the survey was carried out, the average expenditure in culture was 216 Swiss francs. We want to test whether there is a change in the average behavior in this variable.

- What are the hypotheses you wish to contrast?
- What parametric test would be adequate for this?
- What are the assumptions on which this test is based, and why do you think they are satisfied?
- What is the test statistic? What is the corresponding sampling distribution?
- Carry out this test and discuss the results.

(c) Test the hypothesis that women spend more on theater visits than men. What is your conclusion?

(d) What parametric test would be adequate to compare the average expenditure in theater visits in the survey year and the preceding year for the whole population? Carry this test out and discuss your findings.

(e) What non-parametric test or tests would be adequate for (b) and (c)? Perform these tests and compare your results with what you obtained before.

Solution

(a) Loading the data

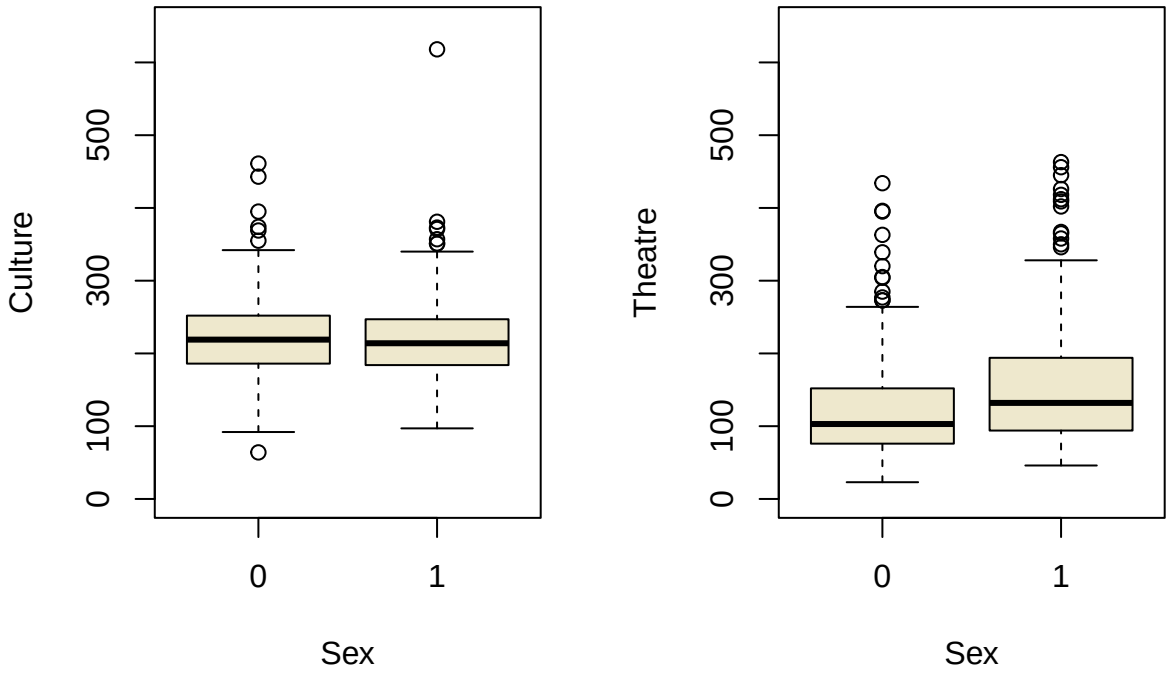
```
q1.df <- read.csv('theatre.csv')
str(q1.df)
```

```
## 'data.frame':    699 obs. of  7 variables:
##  $ X         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age       : int  31 54 56 36 24 25 61 50 53 52 ...
##  $ Sex       : int  1 0 1 1 1 0 1 0 1 0 ...
##  $ Income    : num  90.5 73 74.3 73.6 109 93.1 63.9 46.1 75 68.7 ...
##  $ Culture   : int  181 234 289 185 191 273 184 155 253 291 ...
##  $ Theatre   : int  104 116 276 75 172 168 119 97 152 166 ...
##  $ Theatre_ly: int  150 140 125 130 140 130 195 110 155 150 ...
```

```
attach(q1.df)
```

Graphs

```
par(mfrow=c(1,2))
boxplot(Culture ~ Sex, ylim = c(0,650), col = 'cornsilk2')
boxplot(Theatre ~ Sex, ylim = c(0,650), col = 'cornsilk2')
```




```
par(mfrow=c(1,1))
```

I set the same y scales in both graphs, but this was not required. It just makes the comparisons easier. Since the expenditure in theater is part of the expenditure in culture, one would expect the values for theater to be lower than those for culture, which is indeed the case. The first boxplot (for culture) shows very similar distributions for both sexes but the second shows some differences, mainly in the expenditure level. Apparently, women spend more in theater visits than men do. We will test this later on in this question.

(b) We want to compare whether the average culture expenditure $\hat{\mu}_n$ is equal to 216, so the hypotheses are

$$H_0 : \hat{\mu}_n = 216 \qquad \text{vs} \qquad H_1 : \hat{\mu}_n \neq 216$$

The adequate test for comparing the mean annual expenditure between the two years is the t-test, since we have to estimate the variance. In this case, there are 699 subjects in the sample. Since sample size is large, the Central Limit Theorem says that the normal approximation required for the t-test is reasonable. The test statistic for this test is the standardized sample mean

$$\frac{\hat{\mu}_n - 216}{s_n/\sqrt{699}}$$

where s_n is the sample standard deviation. We calculate below the value for this statistic (this was not required)

```
(tn <- (mean(Culture)-216)/(sd(Culture)/sqrt(699)))
```

```
## [1] 1.9678
```

The sampling distribution is the t distribution with 698 degrees of freedom. The following command carries out this test in R

```
t.test(Culture, mu=216)
```

```
##
##  One Sample t-test
##
## data:  Culture
## t = 1.97, df = 698, p-value = 0.049
## alternative hypothesis: true mean is not equal to 216
## 95 percent confidence interval:
##  216.01 223.70
## sample estimates:
## mean of x
##    219.86
```

The p value is just below 0.05 and the decision depends on our choice for α. If we choose 0.05, the null hypothesis is rejected, while if we choose 0.02 or 0.01, we will not reject the null hypothesis.

(c) We now want to compare the average values for two populations (males and females) and the adequate test for this is the t-test. In this case it is justified because the sample sizes for each population are large enough:

```
sum(Sex==0); sum(Sex==1)
```

```
## [1] 309
```

```
## [1] 390
```

In the sample, 309 are males and 390 are females. Since sample sizes are large, the Central Limit Theorem says that the t-test is a reasonable choice. We want to test whether women spend more so we need a one-sided alternative hypothesis. The following command carries out this test in R

```
t.test(Theatre[Sex == 0], Theatre[Sex == 1], alternative = 'less')
```

```
##
##  Welch Two Sample t-test
##
## data:  Theatre[Sex == 0] and Theatre[Sex == 1]
## t = -5.62, df = 694, p-value = 1.4e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##     -Inf -22.227
## sample estimates:
## mean of x mean of y
##    122.13    153.58
```

The p-value is small and we reject the null hypothesis of equal expenditure. Our conclusion is that women spend more on theater visits than men.

(d) Since we have data for both years for the same subjects, the adequate test here is a paired test.

```r
t.test(Theatre_ly, Theatre, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Theatre_ly and Theatre
## t = -1.09, df = 698, p-value = 0.27
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.7075  2.4815
## sample estimates:
## mean of the differences
##                  -3.113
```

The p-value is large, so we do not reject the null hypothesis of equal expenditure.

(e) For (b) we have the Wilcoxon one sample test

```r
wilcox.test(Culture, mu=216)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Culture
## V = 124991, p-value = 0.53
## alternative hypothesis: true location is not equal to 216
```

We would reach the same conclusion with this test. For (c) we have Wilcoxon's test for two samples

```r
wilcox.test(Theatre[Sex == 0],Theatre[Sex == 1], alternative = 'less')
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Theatre[Sex == 0] and Theatre[Sex == 1]
## W = 44669, p-value = 2.1e-09
## alternative hypothesis: true location shift is less than 0
```

Again, we reach the same conclusion.

Question 2 (25 points)

The data set `Auto` in the `ISLR` package has information on nine variables for 392 vehicles. For this question, we will be only interested in two of them, `mpg` and `origin`. The first variable, `mpg`, corresponds to fuel consumption in miles per gallon for each vehicle, while `origin` is coded as 1 (American), 2 (European), and 3 (Japanese).

(a) Create a data frame named `q2.df` with the two columns corresponding to these variables. Do a boxplot of `mpg` as a function of `origin`. Comment on what you observe.

(b) Using the information in `mpg`, add a factor `fmpg` to `q2.df` created according to the following rule: if `mpg` is below 20, the value for the factor is `low`; if `mpg` is between 20 and 35, the value is `med`, and if `mpg` is above 35, the value is `high`. One way to do this is using the function `cut`. Also, change the labels in the `origin` factor to `Am`, `Eu`, and `Jap`.

(c) Produce a table of `origin` and `fmpg` and do a mosaic plot. The table should have `origin` as rows and `fmpg` as columns. Comment on what you observe. Produce a second table with proportions calculated relative to the different levels of `origin`. Again, comment on what you observe.

(d) We want to determine whether the fuel consumption categories that we created are homogeneously distributed for the different origins of the vehicles.

- Which test or tests do you know that can be used for this?
- What are the underlying assumptions?
- Are they satisfied in this case?
- Carry out all the tests you mentioned and discuss the results.
- What are your conclusions?

Solution

```
library(ISLR)
str(Auto)
```

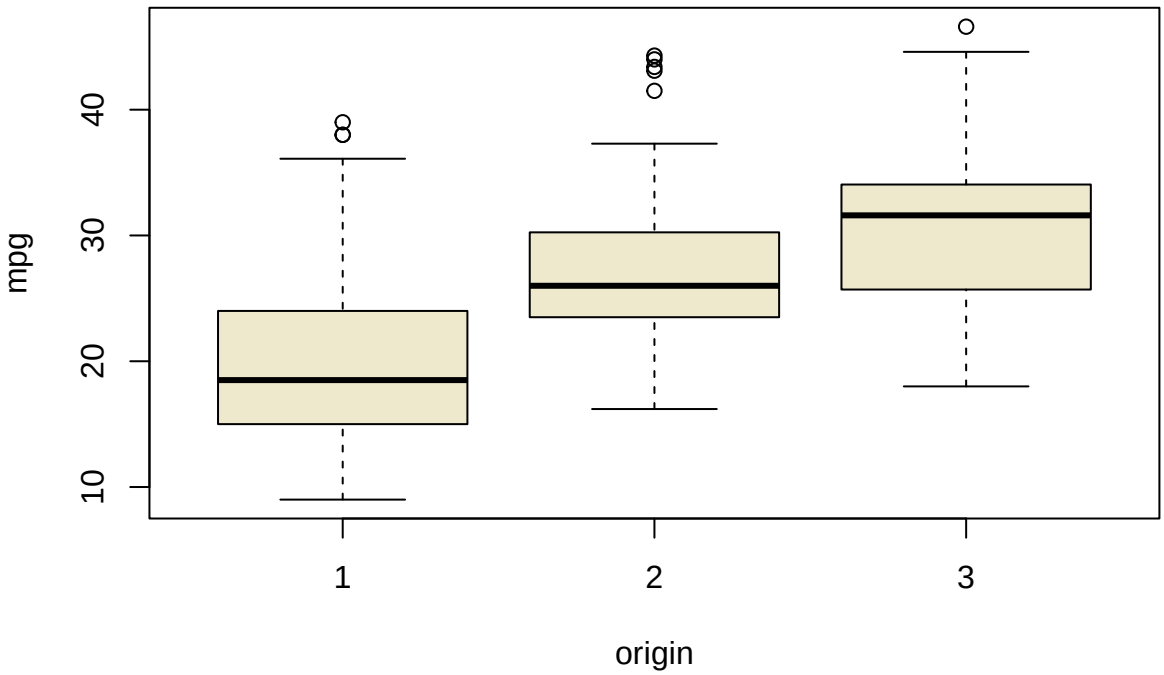
```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 1
```

(a) We use `subset` to create this data frame and then we graph the boxplot

```
q2.df <- subset(Auto, select = c(mpg, origin))
str(q2.df)
```

```
## 'data.frame':    392 obs. of  2 variables:
##  $ mpg   : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ origin: num  1 1 1 1 1 1 1 1 1 1 ...
```

```
boxplot(mpg ~ origin, data = q2.df, col = 'cornsilk2')
```



We can see that American cars have a much smaller average `mpg` than Europeans or Japanese cars. European and Japanese cars are closer, but with higher values for the Japanese cars.

(b) Using the function `cut`, we can create the new factor and then we change the labels for `origin`.

```
q2.df$fmpg <- cut(q2.df$mpg,c(0,20,35,50),labels = c('low','med','high'))
q2.df$origin <- factor(q2.df$origin, labels = c('Am','Eu','Jap'))
str(q2.df)
```

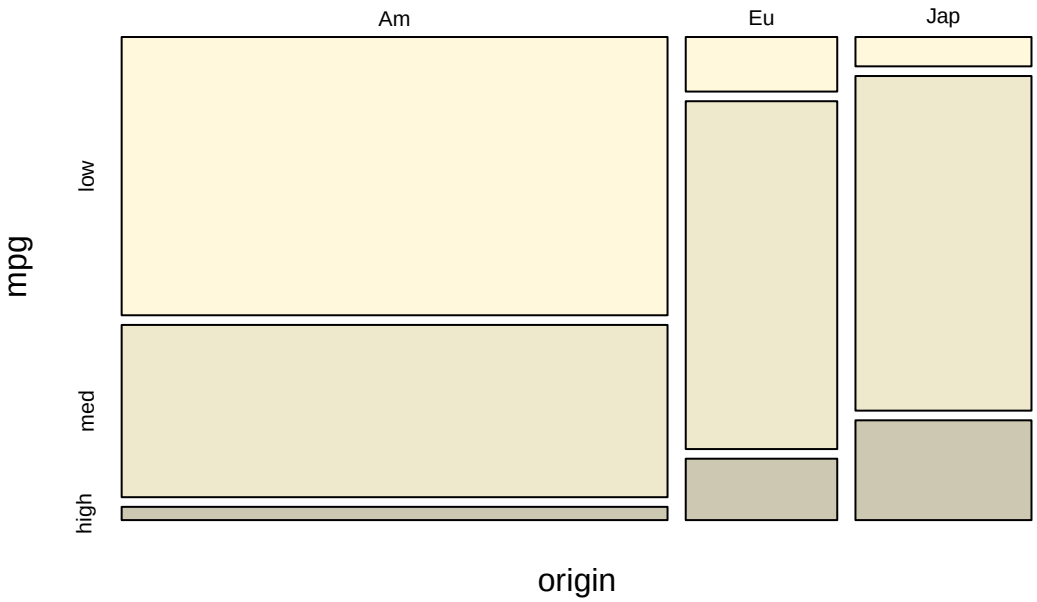
```
## 'data.frame':    392 obs. of  3 variables:
##  $ mpg   : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ origin: Factor w/ 3 levels "Am","Eu","Jap": 1 1 1 1 1 1 1 1 1 1 ...
##  $ fmpg  : Factor w/ 3 levels "low","med","high": 1 1 1 1 1 1 1 1 1 1 ...
```

(c) Table and mosaic plot:

```
(q2.tbl <- with(q2.df, table(origin, fmpg)))
```

```
##       fmpg
## origin low med high
##    Am  147  91    7
##    Eu    8  51    9
##    Jap   5  57   17
```

```
mosaicplot(q2.tbl, ylab = 'mpg', color = c('cornsilk1','cornsilk2','cornsilk3'),
           main = 'mpg according to origin of car')
```



We see in the plot that the proportions of the different mpg categories for European and Japanese cars are similar, although Japanese cars are more frequent on the `high` category and less frequent on the `low`. The proportions for the American cars contrast sharply with the other two. Most cars (more the half) fall in the `low` category and there are very few in the `high` category. This points to American cars being different from the rest, and European and Japanese having similar distributions. The plot also shows that more than half of the cars in the sample come from America.

We now produce the second table using `prop.table`

```
prop.table(q2.tbl,1)
```

```
##       fmpg
## origin      low      med     high
##    Am  0.600000 0.371429 0.028571
##    Eu  0.117647 0.750000 0.132353
##    Jap 0.063291 0.721519 0.215190
```

We see that 60% of American cars are in the `low` category, in contrast with only around 2% in the `high` category. For Japanese and European cars the proportions are similar, with over 72% in the `med` category and more cars in `high` than in 'low.

(d) We have the Chi-square test and Fisher's exact test. Both compare observed and expected values for the contingency table. The first uses a Chi-square approximation for the sampling distribution of the test statistic and requires that the expected value for each cell in the table be at least 5. This is may not be true due to the small number of cars in certain categories.

To check this, we calculate the table of expected values.

```
colSums(prop.table(q2.tbl))%*%t(rowSums(prop.table(q2.tbl))) *392
```

```
##           Am      Eu     Jap
## [1,] 100.000 27.7551 32.2449
## [2,] 124.375 34.5204 40.1046
## [3,]  20.625  5.7245  6.6505
```

All values are above five, so the conditions for the test are satisfied. Another way of obtaining this table is shown below, after doing the chi square test.

The test is executed with the command

```
chisq.test(q2.tbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  q2.tbl
## X-squared = 110, df = 4, p-value <2e-16
```

The test gives a p-value which is practically zero, so there is strong evidence to reject the null hypothesis of homogeneous distributions.

We can verify that the test is valid by extracting the table of expected values from the output of the test:

```
chisq.test(q2.tbl)$expected
```

```
##       fmpg
## origin     low     med    high
##    Am  100.000 124.375 20.6250
##    Eu   27.755  34.520  5.7245
##    Jap  32.245  40.105  6.6505
```

We now do Fisher's test:

```
fisher.test(q2.tbl)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  q2.tbl
## p-value <2e-16
## alternative hypothesis: two.sided
```

Again, the p-value is practically zero and we reach the same conclusion.

Question 3 (35 points)

The data in the file `sulfa.txt` has the results of an experiment to study the effect of sulfamerazine (`Sulfa`) on the amount of hemoglobin (`Hemo`) in trouts. The trouts were placed at random in four different containers, and the fish food added contained, respectively, 0, 5, 10, and 15 grams of sulfamerazine per 100 pounds of fish (coded 1, 2, 3, and 4). The measurements were made on ten randomly selected fish from each container after 35 days.

a) Read the data file into a data frame named `q3.df`. Make sure the data are read correctly. Add a factor `fSulfa` with the information in the variable `Sulfa`.

(b) Do boxplots for `Hemo` as a function of `Sulfa` (all the boxplots should appear on the same panel). Add the points to this graph. Comment on what you observe.

(c) Fit an analysis of variance model to this data. Use $\alpha = 0.02$ for your test. What do you conclude from this analysis?

(d) Find the estimate for the mean response for each treatment. Find also the effects, and include the standard errors in each case. What are the estimated values for the variance and standard deviation in this experiment?

(e) What are the assumption on which the analysis of variance model is based? Draw diagnostic plots for checking these assumptions and discuss the results.

Solution

(a) Read the data

```
q3.df <- read.table('trout.txt', header = T)
str(q3.df)
```

```
## 'data.frame':    40 obs. of  2 variables:
##  $ Sulfa: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Hemo : num  6.7 7.8 5.5 8.4 7 7.8 8.6 7.4 5.8 7 ...
```

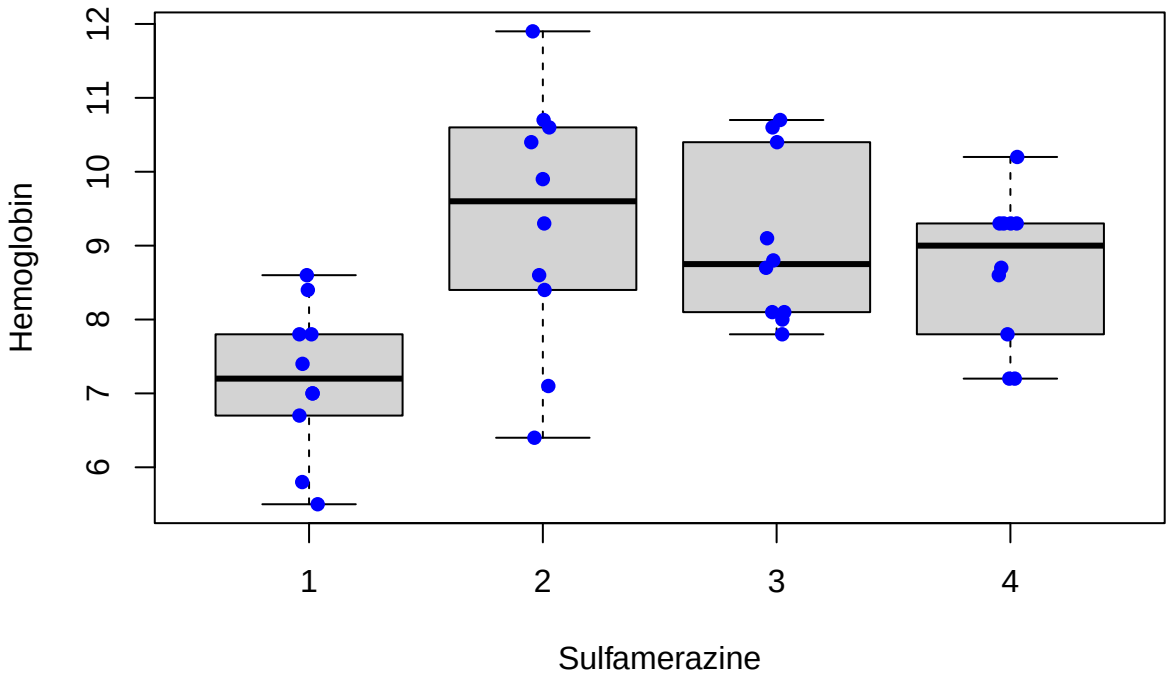
Add a factor

```
q3.df$fSulfa <- factor(q3.df$Sulfa)
str(q3.df)
```

```
## 'data.frame':    40 obs. of  3 variables:
##  $ Sulfa : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Hemo  : num  6.7 7.8 5.5 8.4 7 7.8 8.6 7.4 5.8 7 ...
##  $ fSulfa: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
```

(b) Boxplots:

```
plot(Hemo ~ fSulfa, data = q3.df, xlab='Sulfamerazine', ylab = 'Hemoglobin')
points(Hemo ~ jitter(Sulfa, amount = 0.05), data = q3.df, pch = 16, col = 'blue')
```



We observe that there is an initial increment in the amount of hemogoblin, when we add a small amount of `Sulfa`, but increasing the dosage has the effect of reducing the amount of hemogoblin. Since the amount of `Sulfa` increases uniformly in steps of 5 grams per 100 pounds of fish, we can also interpret the x-axis as a numerical variable and observe that there seems to be a quadratic relation between the variables. Also, the size of the boxes show some variability. This is something we will have to check with the diagnostic plots.

(c) Anova model:

```
model1 <- aov(Hemo ~ fSulfa, data = q3.df)
summary(model1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## fSulfa       3   26.8    8.93     5.7 0.0027 **
## Residuals   36   56.5    1.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is small, so we conclude that there is a difference in the treatments.

(d) Mean responses with standard errors

```
(means <- model.tables(model1, 'means', se = TRUE))
```

```
## Tables of means
## Grand mean
##
## 8.5625
##
##  fSulfa
## fSulfa
##    1    2    3    4
## 7.20 9.33 9.03 8.69
##
## Standard errors for differences of means
##         fSulfa
```

```
##         0.5601
## replic.     10
```

Effects with standard errors:

```
model.tables(model1, se = TRUE)
```

```
## Tables of effects
##
##  fSulfa
## fSulfa
##       1       2       3       4
## -1.3625  0.7675  0.4675  0.1275
##
## Standard errors of effects
##         fSulfa
##         0.3961
## replic.     10
```

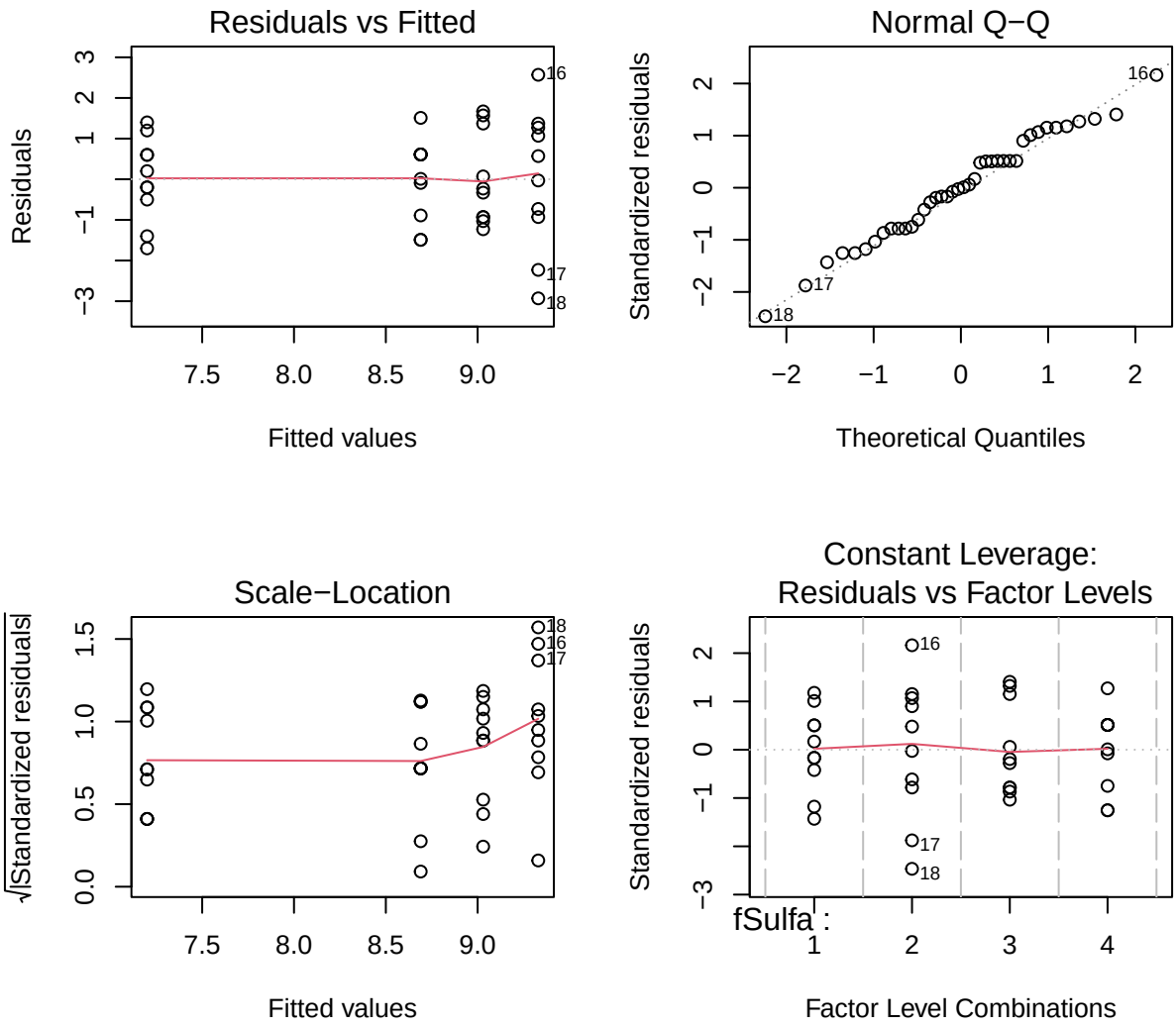
The estimated variance comes from the anova table and is 1.569. the standard deviation is

```
Va = 1.569
sqrt(Va)
```

```
## [1] 1.2526
```

(e) The model is based on the assumption that the experimental errors are independent, normally distributed random variables with mean zero and equal variance. We use the diagnostic plots to check the assumptions.

```
par(mfrow = c(2,2))
plot(model1)
```




```
par(mfrow=c(1,1))
```

The normal quantile plot is very good, and shows that the residuals follow a normal distribution. The residuals vs. fitted values shows that the residuals for one of the treatment levels have a wider spread than the other levels, but the difference does not seem to be important. Three points, 16, 17, and 18, are singled out in the graphs as having largest residuals, and they all belong to the same treatment level. The scale-location plot shows an increase in the average value towards the largest fitted values, but the increase does not seem to be significant. Taking into account all the graphs, it seems reasonable to conclude that the assumptions on which the model is based are satisfied.