STAT 210
Applied Statistics and Data Analysis
First Exam

October 22, 2022

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. Writing code is not enough to answer a question. Your comments are more important than the code. Label your graphs appropriately

Please identify the files you submit with your surname

Question 1 (40 points)

The file `motorskills` has information on an experiment to study the motor skills of children and youngsters. In the experiment, the speed with which subjects placed a series of cylinders into a set of holes was measured. The data set has five variables: `Age` in months, `Gender`, `DH`, which denotes the Dominant Hand, i.e., whether the dominant hand is the right or the left hand, `sp1`, which corresponds to the speed with the dominant hand and `sp2` which corresponds to the speed with the non-dominant hand. The speed is measured in cylinders per second. Use $\alpha = 0.02$ for all tests in this question.

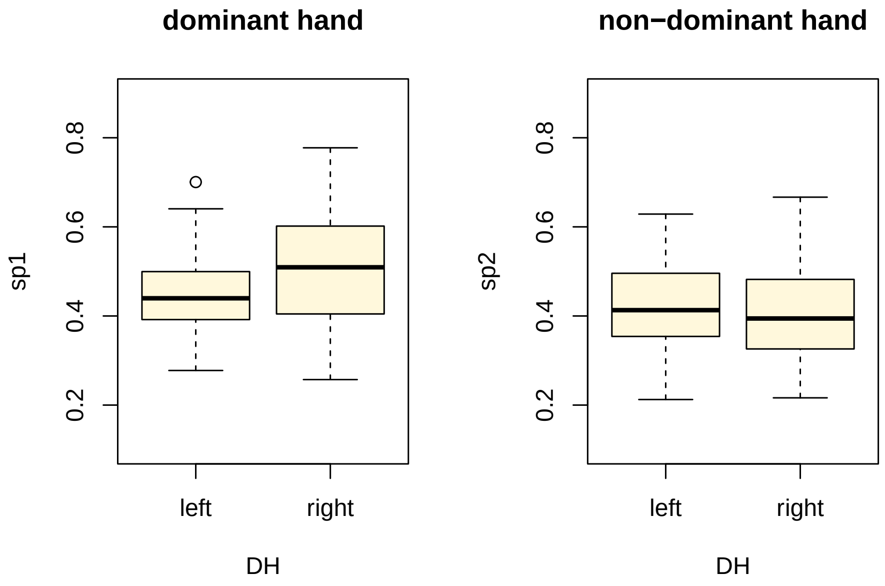
Load the data and store it in a data frame named `df1`.

```
df1 <- read.csv('motorskills.csv')
str(df1)
```

```
## 'data.frame':    67 obs. of  5 variables:
##  $ Age   : int  117 101 135 119 124 127 101 131 119 105 ...
##  $ Gender: chr  "male" "male" "male" "male" ...
##  $ DH    : chr  "right" "right" "right" "right" ...
##  $ sp1   : num  0.353 0.257 0.537 0.444 0.483 ...
##  $ sp2   : num  0.216 0.343 0.497 0.496 0.388 ...
```

(a) Draw boxplots for the speed with the dominant hand as a function of `DH` (dominant hand) and also for the speed with the non-dominant hand as a function of `DH`. Use a common scale. Comment on what you observe.

```
par(mfrow=c(1,2))
boxplot(sp1 ~ DH, data = df1, main = 'dominant hand', col = 'cornsilk',
        ylim = c(0.1,0.9))
boxplot(sp2 ~ DH, data = df1, main = 'non-dominant hand', col = 'cornsilk',
        ylim = c(0.1,0.9))
```



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

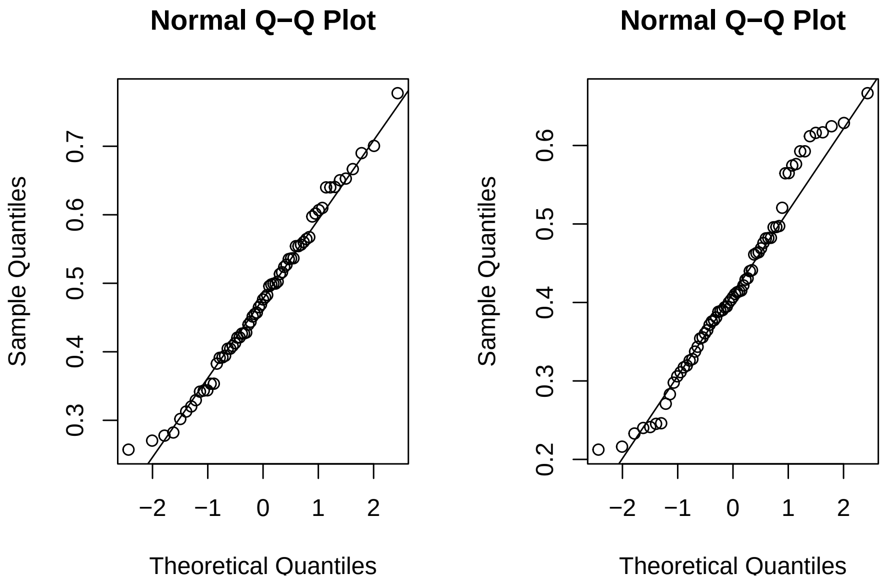
The first boxplot (dominant hand) shows that right-handed subjects achieved a higher median speed and have a wider range of values than left-handed subjects. Particularly, about 50% of right-handed subjects achieved speeds of 0.5 or more while less than 25% of left-handed subjects achieved this range of values.

The second boxplot shows similar distributions for both groups, with left-handed subjects having a slightly higher median speed. Also, values for the non-dominant hand (second plot) are lower than for the dominant hand (first plot).

(b) We want to explore if the speed with the dominant hand is the same as with the non-dominant hand. What parametric test or tests are appropriate here? What are the assumptions, and why do you think they are satisfied? Carry out this test or tests and comment on the results.

Since the speed measurements `sp1` and `sp2` were taken on the same subjects, we have paired data. Therefore, the adequate test here is a t-test for paired data. This test assumes that the data come from a normal distribution or that the sample is large enough for the mean to have an approximate normal distribution. We look at the quantile plots for `sp1` and `sp2`:

```
par(mfrow = c(1,2))
qqnorm(df1$sp1); qqline(df1$sp1)
qqnorm(df1$sp2); qqline(df1$sp2)
```



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

Both quantile plots are good and support the assumption of normality. We confirm this with a Shapiro-Wilk test:

```
shapiro.test(df1$sp1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df1$sp1
## W = 0.987, p-value = 0.71
```

```
shapiro.test(df1$sp2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df1$sp2
## W = 0.971, p-value = 0.12
```

In both cases the p-value is large and we do not reject the null hypothesis of normality. Therefore, we use the t-test.

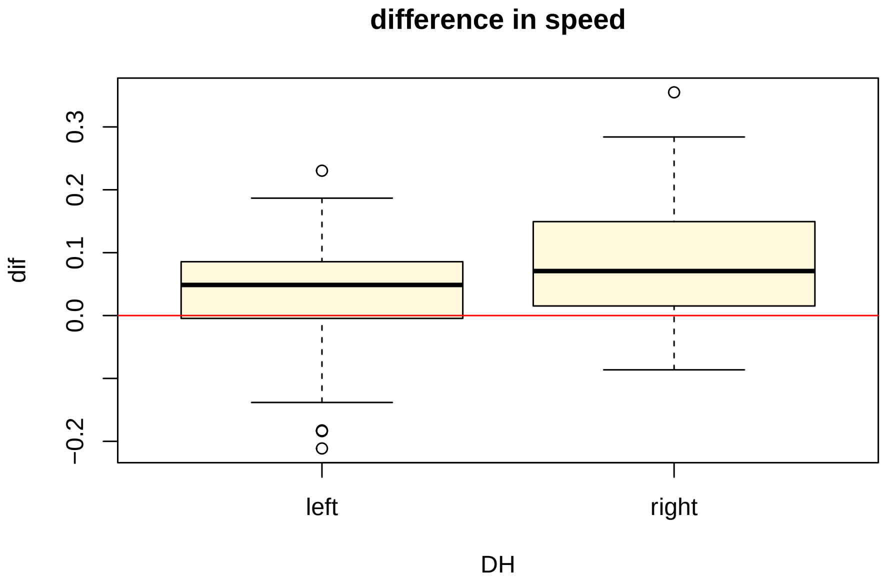
```
t.test(df1$sp1, df1$sp2, paired = T, conf.level = 0.98)
```

```
##
##  Paired t-test
##
## data:  df1$sp1 and df1$sp2
## t = 4.51, df = 66, p-value = 2.7e-05
## alternative hypothesis: true difference in means is not equal to 0
## 98 percent confidence interval:
##  0.028279 0.091715
## sample estimates:
## mean of the differences
##                0.059997
```

We observe that the p-value is very small, so we reject the null hypothesis of equal speed with both hands.

(c) Define a new variable named dif in the data frame. The value for this variable is the difference between the speed with the dominant hand minus the speed with the non-dominant hand. Do a boxplot of this new variable as a function of DH. Comment on the graph.

```
df1$dif <- df1$sp1 - df1$sp2
boxplot(dif ~ DH, data = df1, main = 'difference in speed', col = 'cornsilk')
abline(h=0, col = 'red')
```



Values for right-handed subjects seem to be higher. In both cases, we see that about 75% of the subjects were faster with the dominant hand and around 25% were not. Both distributions look symmetric. The variance for right-handed subjects is bigger than for left-handed.

(d) We now want to test whether the average value for this difference in speed (dif) is the same for right-handed and left-handed subjects. What parametric test or tests are adequate here? What are the assumptions, and why do you think they are satisfied? Carry out this test or tests and comment on the results.

In this case we are comparing two groups, right-handed versus left-handed, and we want to test whether the mean speeds are the same. The adequate test here is a two-sample t-test because we need to estimate the standard deviation. The test assumes that the populations follow a normal distribution or that the samples are large enough for the normal approximation to the distribution of the mean to be valid. The sample sizes are

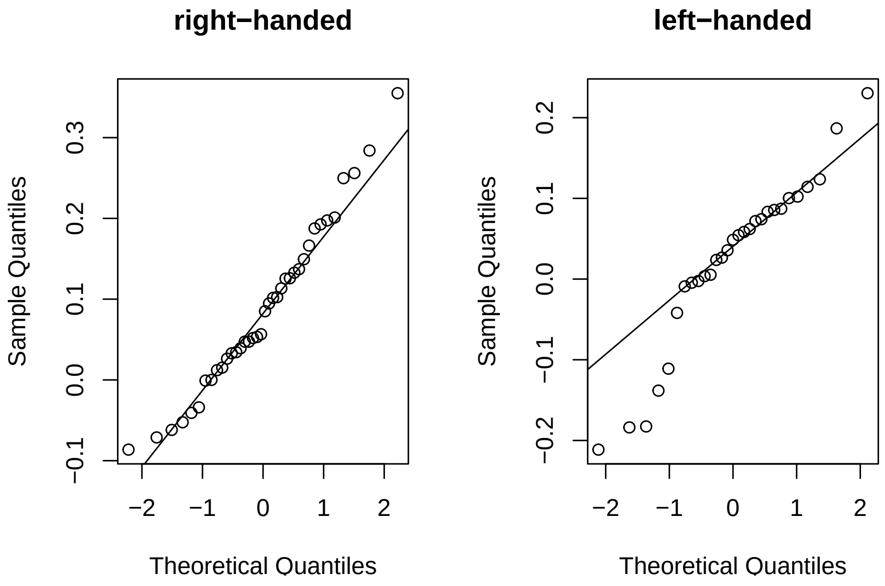
```
sum(df1$DH=='right'); sum(df1$DH=='left')
```

```
## [1] 38
```

```
## [1] 29
```

which are moderate. To check whether the groups have a normal distribution we do quantile plots

```
par(mfrow = c(1,2))
qqnorm(df1$dif[df1$DH=='right'], main = 'right-handed'); qqline(df1$dif[df1$DH=='right'])
qqnorm(df1$dif[df1$DH=='left'], main = 'left-handed'); qqline(df1$dif[df1$DH=='left'])
```



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

The plot for right-handed subjects looks fine but for left-handed subjects the points in the lower tail are far from the reference line. To check whether normality is a reasonable assumption, we do a Shapiro-Wilk test.

```
shapiro.test(df1$dif[df1$DH=='right'])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df1$dif[df1$DH == "right"]
## W = 0.974, p-value = 0.52
```

```
shapiro.test(df1$dif[df1$DH=='left'])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df1$dif[df1$DH == "left"]
## W = 0.93, p-value = 0.053
```

For right-handed subjects, the p-value is large but for left-handed it is close to 5%. Since we are testing at a level of 2%, we do not reject the null hypothesis of normality, but the p-value we get with this test may be inacurate. The t-test is

```
t.test(dif ~ DH, data = df1, conf.level = 0.98)
```

```
##
##  Welch Two Sample t-test
##
## data:  dif by DH
## t = -2.46, df = 60, p-value = 0.017
## alternative hypothesis: true difference in means between group left and group right is not equal to
## 98 percent confidence interval:
##  -0.125618 -0.001692
## sample estimates:
##  mean in group left mean in group right
##            0.023894            0.087549
```

Since the *p*-value is below 0.02, we reject the null hypothesis of equal means.

(e) For the problems in (b) and (d), what non-parametric tests can be applied? What are the assumptions for these tests? Why do you think they are satisfied? Carry out these tests and comment on the results.

Wilcoxon's test can be applied in both cases. It assumes that the sample or samples come from a continuous symmetric distribution. For both problems we saw that the assumption of normality is supported by the tests and normal data are continuous and symmetric. For the problem in (b) we have

```
wilcox.test(df1$sp1, df1$sp2, paired = T, conf.level = 0.98)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  df1$sp1 and df1$sp2
## V = 1769, p-value = 2.3e-05
## alternative hypothesis: true location shift is not equal to 0
```

The *p*-value is similar to what we obtained before and we reach the same conclusion, the average speeds are not the same.

For the problem in (d)

```
wilcox.test(dif ~ DH, data = df1, conf.level = 0.98)
```

```
## Warning in wilcox.test.default(x = c(0.10218458, 0.07420087,
## -0.00904005000000002, : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dif by DH
## W = 396, p-value = 0.051
## alternative hypothesis: true location shift is not equal to 0
```

Now the *p*-value is 0.051, above 0.02, and we would not reject the null hypothesis of equal means with this result. Since the assumption of normality for left-handed subjects was doubtful, perhaps this test is more reliable in this case.

Question 2 (25 points)

The data set `bloodpress` has information on blood pressure for 321 males over 20 years old. The set has two variables, `Age`, the age of the subject in years, and `BP`, the blood pressure classified into three levels, `Low`, `Normal`, and `High`. Read the data and store it in a file named `df2`.

```
df2 <- read.csv('bloodpress.csv')
```

(a) Check whether `BP` has been stored as a factor. If not, transform it into a factor. The levels should be in the order `Low`, `Normal`, and `High`. If `BP` has been stored as a factor, verify if the levels are in the correct order as stated above. If they are not, modify the variable so that they are.

```
str(df2)
```

```
## 'data.frame':    321 obs. of  2 variables:
##  $ BP : chr  "Low" "Low" "Low" "Low" ...
##  $ Age: int  20 27 25 28 22 23 22 25 21 22 ...
```

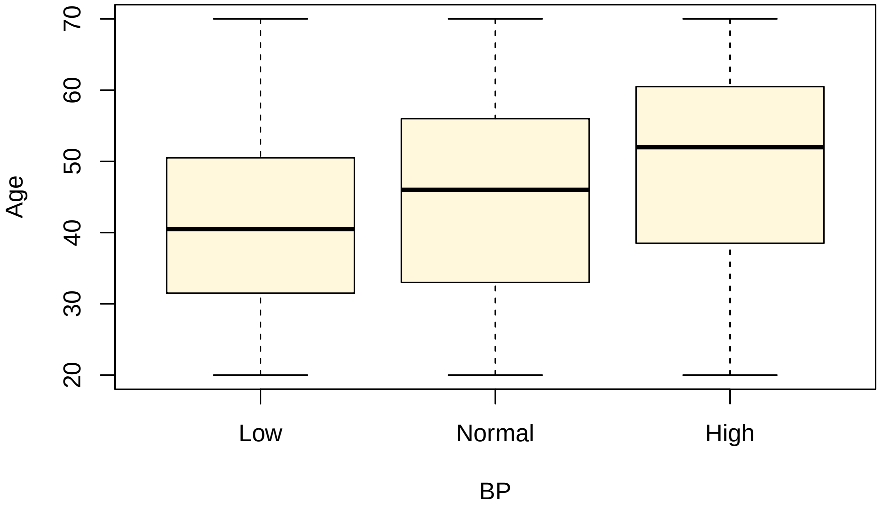
We see that `BP` is stored as a character. We change it to a factor specifying the levels in order

```
df2$BP <- factor(df2$BP, levels = c('Low','Normal','High'))
str(df2)
```

```
## 'data.frame':    321 obs. of  2 variables:
##  $ BP : Factor w/ 3 levels "Low","Normal",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age: int  20 27 25 28 22 23 22 25 21 22 ...
```

(b) Boxplot `Age` as a function of `BP`. Comment on what you observe.

```
boxplot(Age ~ BP, data = df2,col = 'cornsilk')
```



We see that higher blood pressure is asociated to older age. However, there is a large overlap in the three distributions, and the whiskers for the three plots cover the whole range of values.

(c) Using the information in `Age`, add a factor `fAge` to `df2` created according to the following rule: if the subject has less than 30 years, the value for the factor is `Under30`. If the subject is between 30 and 49 years old, the value is `30-49`, and if the subject is 50 or more, the value is `Over50`. One way to do this is using the function `cut`.

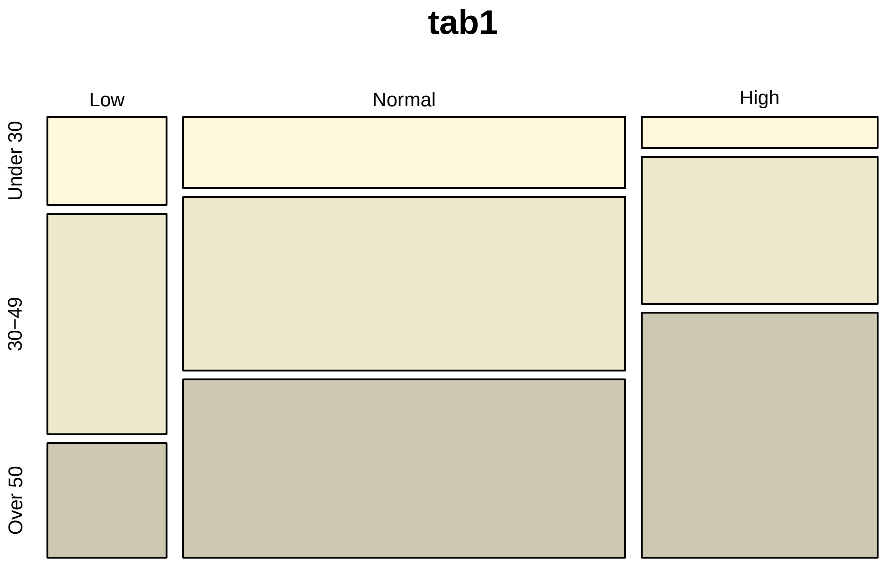
```
df2$fAge <- cut(df2$Age, breaks = c(0,29,49,70),
                labels = c('Under 30', '30-49', 'Over 50'))
```

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

```
(tab1 <- table(df2$BP,df2$fAge))
```

```
##         
##          Under 30 30-49 Over 50
##   Low          10    25      13
##   Normal       30    73      75
##   High          7    33      55
```

```
mosaicplot(tab1,color = c('cornsilk1','cornsilk2','cornsilk3'))
```



From the graph we see that about half of the subjects have `Normal` blood pressure. The proportion of subjects in the `Over 50` group increases when we move from `Low` to `Normal` to `High`. For the group `30-49` the proportions of `Normal` and `High` are similar, while the proportion of `Low` is bigger than the other two. For the age group `Under 30`, the proportions diminish as we move from `Low` to `Normal` to `High`.

We now produce the table of relative proportions

```
round(prop.table(tab1,2),3)
```

```
##
##          Under 30 30-49 Over 50
##   Low       0.213 0.191   0.091
##   Normal    0.638 0.557   0.524
##   High      0.149 0.252   0.385
```

From the table we see that for all age groups, more than half of the subjects belong to the `Normal` blood pressure class. The proportion of subjects in the `High` blood pressure class increases as we move from young to old, while for the `Low` blood pressure group this is the other way round.

(e) We want to determine whether the blood pressure levels are homogeneously distributed in the age groups we created. 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?

We can use the Chi-square test and Fisher's exact test. The first requires that in the matrix of expected values, all entries be at least 5. We check this after the test.

```
(tab1.tst <- chisq.test(tab1))
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 15, df = 4, p-value = 0.0048
```

The p value is small and we reject the null hypothesis of homogeneous distributions across the age groups. The matrix of expected values is

```
tab1.tst$expected
```

```
##
```

```
##          Under 30  30-49 Over 50
##   Low       7.028 19.589  21.383
##   Normal   26.062 72.642  79.296
##   High     13.910 38.769  42.321
```

All values are above 5 so the chi-square approximation is valid.

The other test is Fisher's exact test:

```
fisher.test(tab1)
```

```
## 
##  Fisher's Exact Test for Count Data
## 
## data:  tab1
## p-value = 0.0037
## alternative hypothesis: two.sided
```

The *p*-value is of the same order of magnitude as for the chi-square test, and we reach the same conclusion.

Question 3 (35 points)

A pharmaceutical company did an experiment to compare three different pain relievers for treating migraines. The data is stored in the file `migraine`. In the experiment, 27 volunteers participated, and nine were randomly selected for each pain reliever. The subjects were instructed to take the drug during their next migraine headache episode and to report their pain on a scale of 1 = no pain to 10 = extreme pain 30 minutes after taking the drug.

Read the data file into a data frame named `df3`. Make sure the data are read correctly. If `Drug` has character mode, transform it into a factor.

```
df3 <- read.csv('migraine.csv', header = T)
str(df3)
```

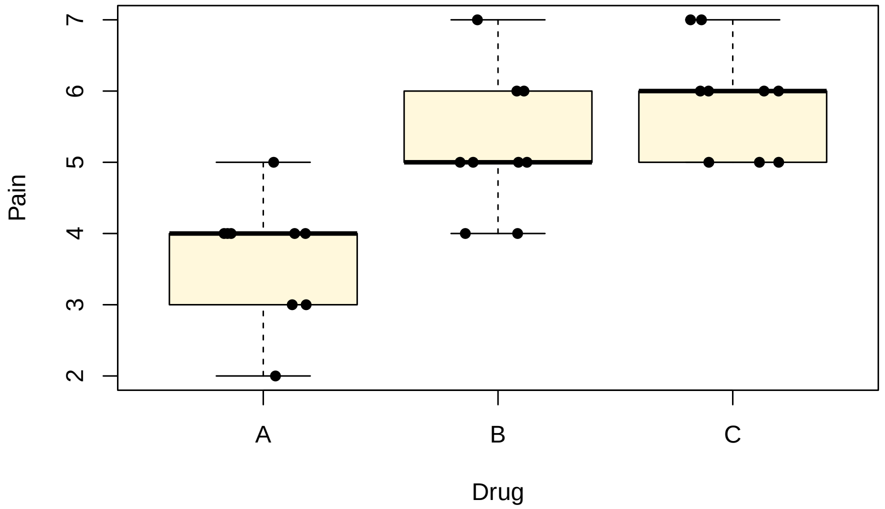
```
## 'data.frame':    27 obs. of  2 variables:
##  $ Pain: int  4 5 4 3 2 4 3 4 4 6 ...
##  $ Drug: chr  "A" "A" "A" "A" ...
```

```
df3$Drug <- factor(df3$Drug)
str(df3)
```

```
## 'data.frame':    27 obs. of  2 variables:
##  $ Pain: int  4 5 4 3 2 4 3 4 4 6 ...
##  $ Drug: Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 2 ...
```

(a) Do boxplots for `Pain` as a function of `Drug`. Add the points to this graph. Comment on what you observe.

```
boxplot(Pain ~ Drug, data = df3, col = 'cornsilk')
points(Pain ~ jitter(as.numeric(Drug)), data = df3, pch = 16)
```



We observe from the graph that drugs B and C have distributions with similar range of values, while drug A has lower values and is likely to have a different effect on `Pain` than the other two. The assumption of equal variances seems valid.

(b) Fit an analysis of variance model to this data using the function `lm` and print the anova table. Use $\alpha = 0.02$ for your test. What do you conclude from this analysis?

```
model1 <- lm(Pain ~ Drug, data = df3)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: Pain
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Drug       2   23.4   11.70    15.2 5.4e-05 ***
## Residuals 24   18.4    0.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is small and we reject the null hypothesis that all treatment levels have the same effect on `Pain`. At least two of the drugs have different effects.

(c) Using the function `summary`, obtain a summary table for the model fitted in (b). What is the meaning of the numbers in the `Estimate` column? Obtain the estimate for the mean response for each treatment from this table.

```
summary(model1)
```

```
##
## Call:
## lm(formula = Pain ~ Drug, data = df3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.667 -0.667  0.111  0.333  1.778
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.667      0.292   12.55  4.9e-12 ***
## DrugB          1.556      0.413    3.76  0.00095 ***
## DrugC          2.222      0.413    5.38  1.6e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.877 on 24 degrees of freedom
## Multiple R-squared:  0.559,  Adjusted R-squared:  0.523
## F-statistic: 15.2 on 2 and 24 DF,  p-value: 5.37e-05
```

Let us use the notation $\hat{\mu}_A, \hat{\mu}_B$, and $\hat{\mu}_C$ for the mean observed responses for drugs A, B, and C. The first value in the `Estimate` column –named `(Intercept)`– corresponds to the average value for the observations corresponding to drug A, i.e., this is $\hat{\mu}_A$. The second and third rows in the `Estimate` column correspond to differences between the averages of observed values for drugs B and C and the average for drug A, i.e., they correspond to $\hat{\mu}_B - \hat{\mu}_A$ and $\hat{\mu}_C - \hat{\mu}_A$, respectively. Therefore, to obtain the mean response for drug B we need to add the value for drug A (`Intercept`) and the value for drug B (`DrugB`), and similarly for drug C:

$$\hat{\mu}_B = 3.667 + 1.556 = 5.223, \quad \text{and} \quad \hat{\mu}_C = 3.667 + 2.222 = 5.889$$

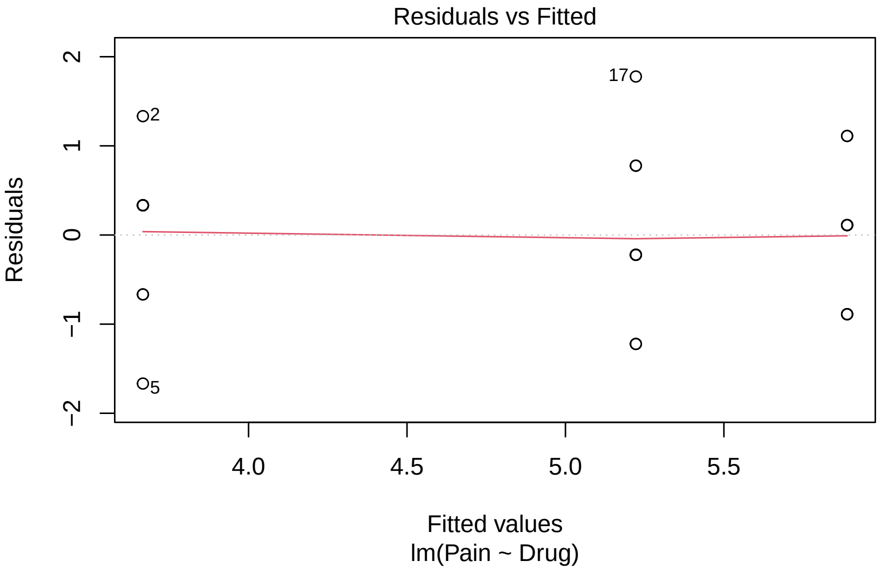
(d) What are the estimated values for the variance and standard deviation of the errors in this experiment?

The estimated value for the variance of the errors is the mean square error, which can be read from the anova table: 0.77. The standard deviation is the square root of this number: 0.8775.

(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.

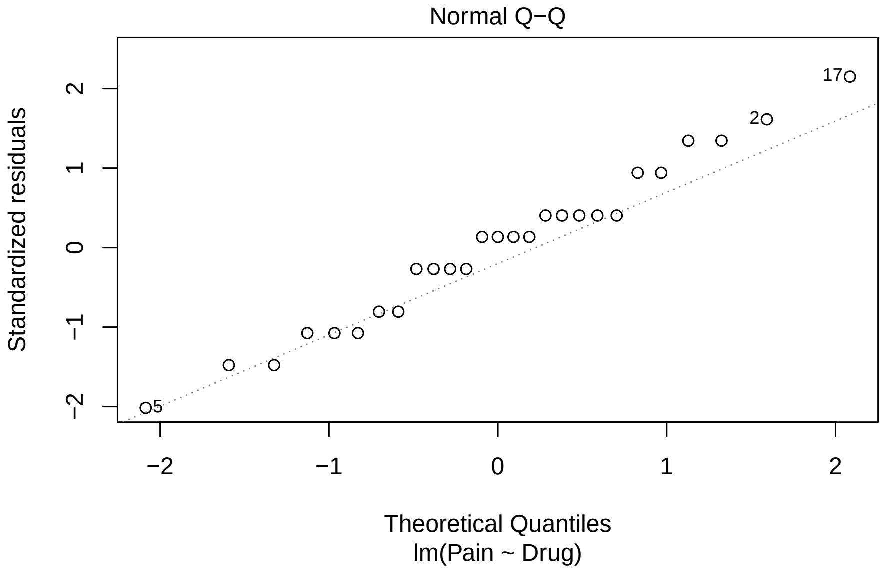
The model assumes that the errors are independent and have a common normal distribution with mean zero and variance σ^2.

```
plot(model1, which = 1)
```



In this plot the red line is close to 0 and is almost horizontal. This shows that the residuals are centered and have a symmetric distribution. The dispersion of the three groups of points is similar,

```
plot(model1, which = 2)
```



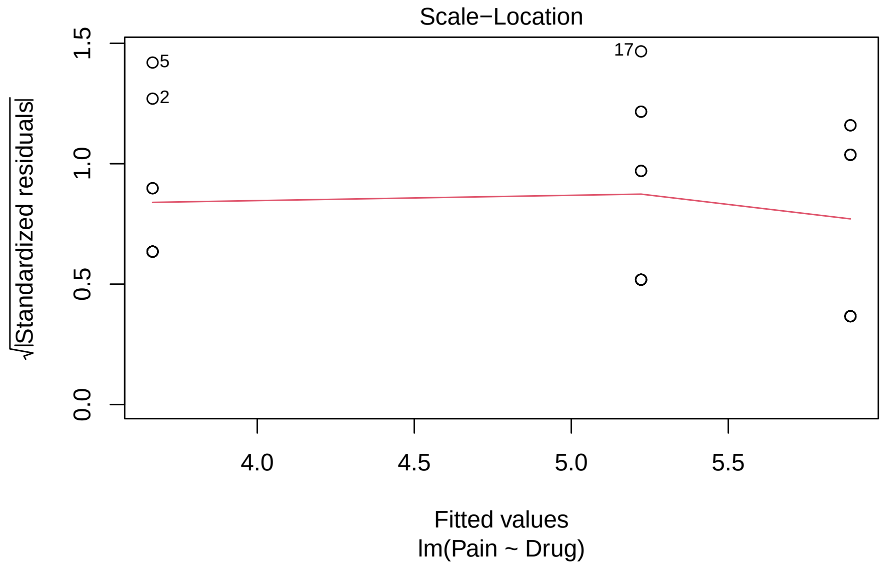
For the quantile plot, the points are close to the reference line and the graph supports the assumption of normality. To have a more solid base for this conclusion, we do a Shapiro-Wilk test on the residuals

```
shapiro.test(residuals(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model1)
## W = 0.979, p-value = 0.83
```

The p-value is large and we do not reject the null hypothesis of normality.

```
plot(model1, which = 3)
```



In the third plot the red line is almost horizontal, indicating that the variability within each group is similar and supporting the assumption of equal variances. There are no points beyond level 1.5, indicating that there are no outliers in the data.

In conclusion, the three plots support that the assumptions are valid for this experiment.