STAT 210
Applied Statistics and Data Analysis
First Exam

Juyi Lin

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. Do not write comments in chunks. Label your graphs appropriately

Please identify the files you submit with your surname

Question 1 (40 points)

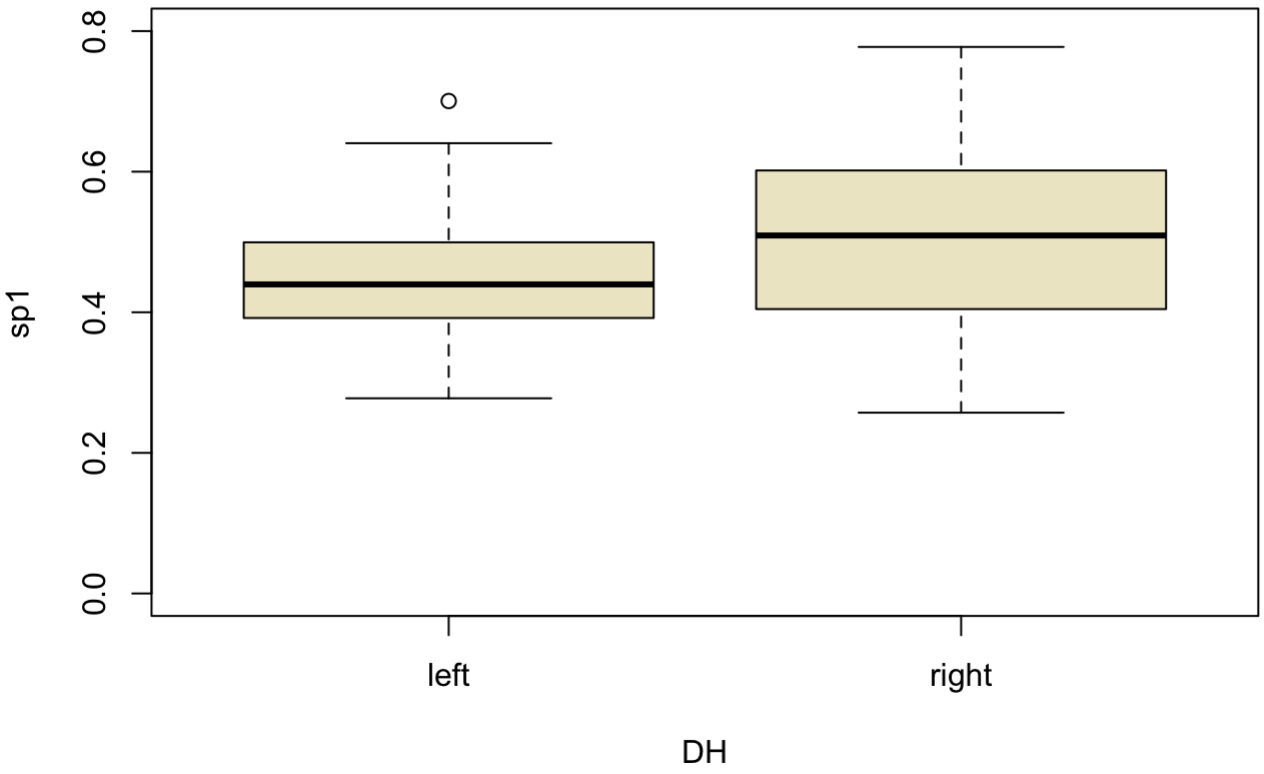
The file `motorskills` has data on an experiment to study the motor skills of children and youngsters. In the experiment, the speed with which subjects placed a series of cylinders into a set of holes was measured. The data set has five variables: `Age` in months, `Gender` , `DH` , which denotes the 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` .

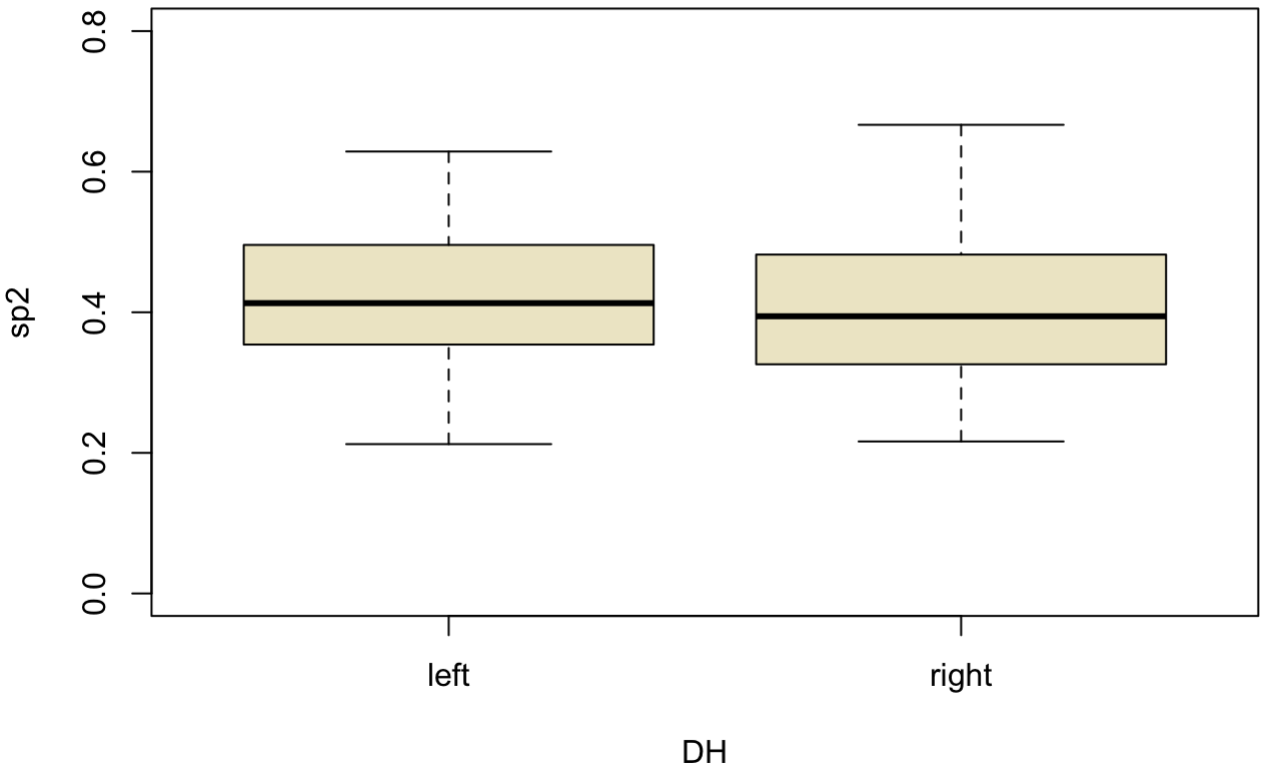
```
df1 = read.csv('motorskills.csv')
```

a. (5 points) 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.

```
boxplot(sp1 ~ DH,ylim = c(0,0.8) , data = df1,  col = 'cornsilk2')
```




```
boxplot(sp2 ~ DH,  ylim = c(0,0.8) , data = df1,  col = 'cornsilk2')
```



I set the same y scales in both graphs. The first boxplot shows the speed with the right dominant hand has higher mean, and wider range. The second boxplot shows the speed with the non-dominant hand, if the right hand is dominant hand, has lower mean. Compare two boxplots, we can find that the speed with the non-dominant hand has lower mean than the speed with the dominant hand.

b. (10 points) 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.

We want to compare the average values for two populations (with dominant and with non-dominant) In this case we can use a two-sample t-test, the test assumes that the sampling distribution for the means is approximately Gaussian.

```
str(df1$sp1); str(df1$sp2)
```

```
##  num [1:67] 0.353 0.257 0.537 0.444 0.483 ...
```

```
##  num [1:67] 0.216 0.343 0.497 0.496 0.388 ...
```

In the sample, both data has 67 samples. Since the sample sizes are large, 67points, the Central Limit Theorem says that this a reasonable assumption.

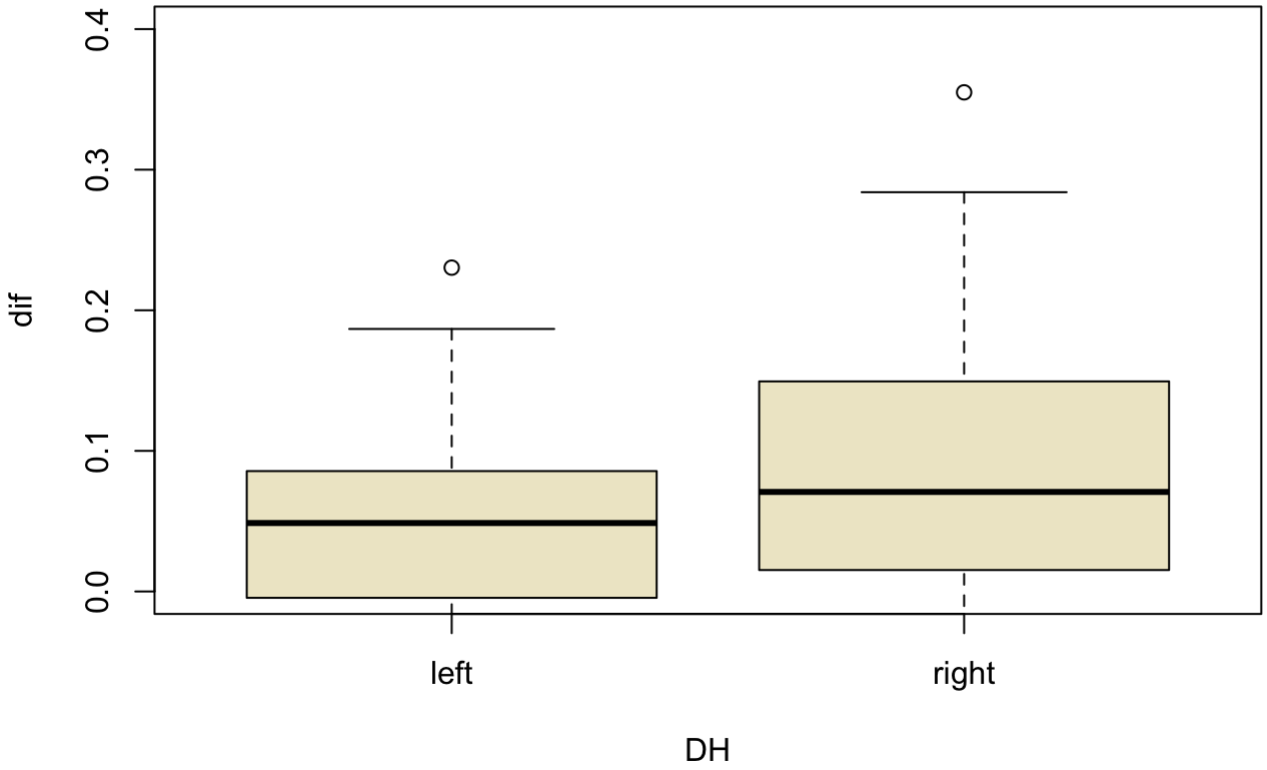
```
t.test(df1$sp1,df1$sp2)
```

```
##
##  Welch Two Sample t-test
##
## data:  df1$sp1 and df1$sp2
## t = 2.99, df = 132, p-value = 0.0033
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.020328 0.099666
## sample estimates:
## mean of x mean of y
##   0.47814   0.41814
```

The small p-value implies that we reject the null hypothesis of equal means for speed. The speed with the dominant hand is not same as with the non-dominant hand.

c. (5 points) 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 <- within(df1, dif <- sp1 - sp2 )
boxplot(dif ~ DH,ylim = c(0,0.4) , data = df1,  col = 'cornsilk2')
```



We can find that the dominant hand is left has less mean and less range in difference between the speed with the dominant hand minus the speed with the non-dominant hand. Right dominant hand people has higher mean, and wider range in difference between the speed with the dominant hand minus the speed with the non-dominant hand.

d. (10 points) 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.

μd be the average for dif[DH==right].

H0:μd=0 vs.H1 :μd != 0

Since we have measurements left-hand and right-hand for the same subjects we should do a t test. the test assumes that the sampling distribution for the means is approximately Gaussian. Since the sample sizes are large, 38 and 29 points, the Central Limit Theorem says that this a reasonable assumption.

```
str(df1$dif[df1$DH=='right']); str(df1$dif[df1$DH=='left'])
```

```
##  num [1:38] 0.1371 -0.0863 0.0392 -0.0524 0.0947 ...
```

```
##  num [1:29] 0.10218 0.0742 -0.00904 0.12367 0.08331 ...
```

```
t.test(df1$dif[df1$DH == 'right'],
       df1$dif[df1$DH == 'left'])
```

```
##
##  Welch Two Sample t-test
##
## data:  df1$dif[df1$DH == "right"] and df1$dif[df1$DH == "left"]
## t = 2.46, df = 60, p-value = 0.017
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.011798 0.115512
## sample estimates:
## mean of x mean of y
##  0.087549  0.023894
```

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

e. (10 points) 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.

For (b) we have the Wilcoxon two sample test. This non-parametric test is useful when the assumption of normality is not justified.

```
wilcox.test( df1$sp1,df1$sp2)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df1$sp1 and df1$sp2
## W = 2889, p-value = 0.0042
## alternative hypothesis: true location shift is not equal to 0
```

We would reach the same conclusion with this test.

for (d) we have Wilcoxon's test for two samples.

```
wilcox.test(df1$dif[df1$DH == 'right'],df1$dif[df1$DH == 'left'], alternative ='two.sided')
```

```
## Warning in wilcox.test.default(df1$dif[df1$DH == "right"],
## df1$dif[df1$DH == : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df1$dif[df1$DH == "right"] and df1$dif[df1$DH == "left"]
## W = 706, p-value = 0.051
## alternative hypothesis: true location shift is not equal to 0
```

We will not reject the null hypothesis. So that the average value for this difference in speed (`dif`) is the same for right-handed and left-handed subjects.

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` .

a. (3 points) 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.

```
df2  =  read.csv('bloodpress.csv')
str(df2$BP)
```

```
##  chr [1:321] "Low" "Low" "Low" "Low" "Low" "Low" "Low" "Low" ...
```

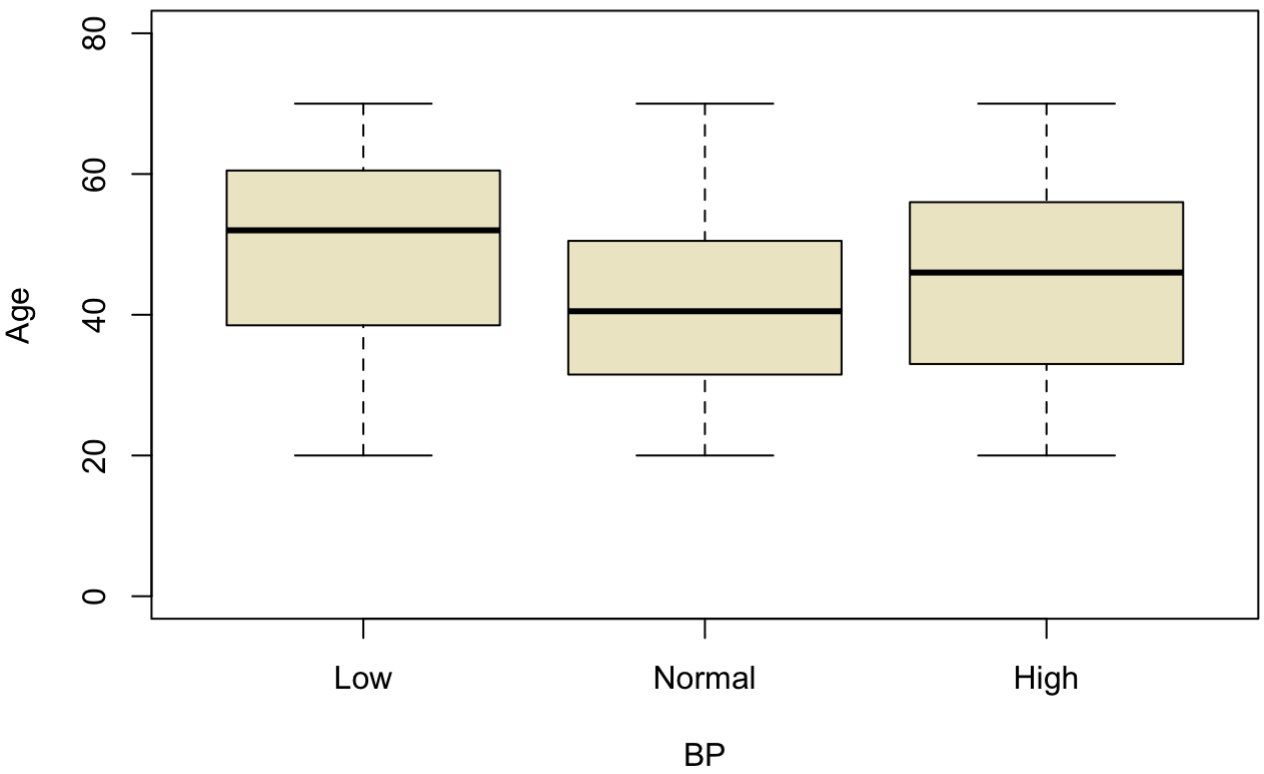
It is not a factor

```
df2$BP <- factor(df2$BP, labels = c('Low','Normal','High'))
df2 = df2[order(df2$BP),]
str(df2$BP)
```

```
##  Factor w/ 3 levels "Low","Normal",..: 1 1 1 1 1 1 1 1 1 1 ...
```

b. (4 points) Boxplot `Age` as a function of `BP` . Comment on what you observe.

```
boxplot(Age ~ BP,ylim = c(0,80) , data = df2,  col = 'cornsilk2')
```



The Low blood pressure have highest mean age, the normal pressure have lowest mean age.
The height of all three boxes are similar. Normal box is symmetric.

c. (5 points) 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` .

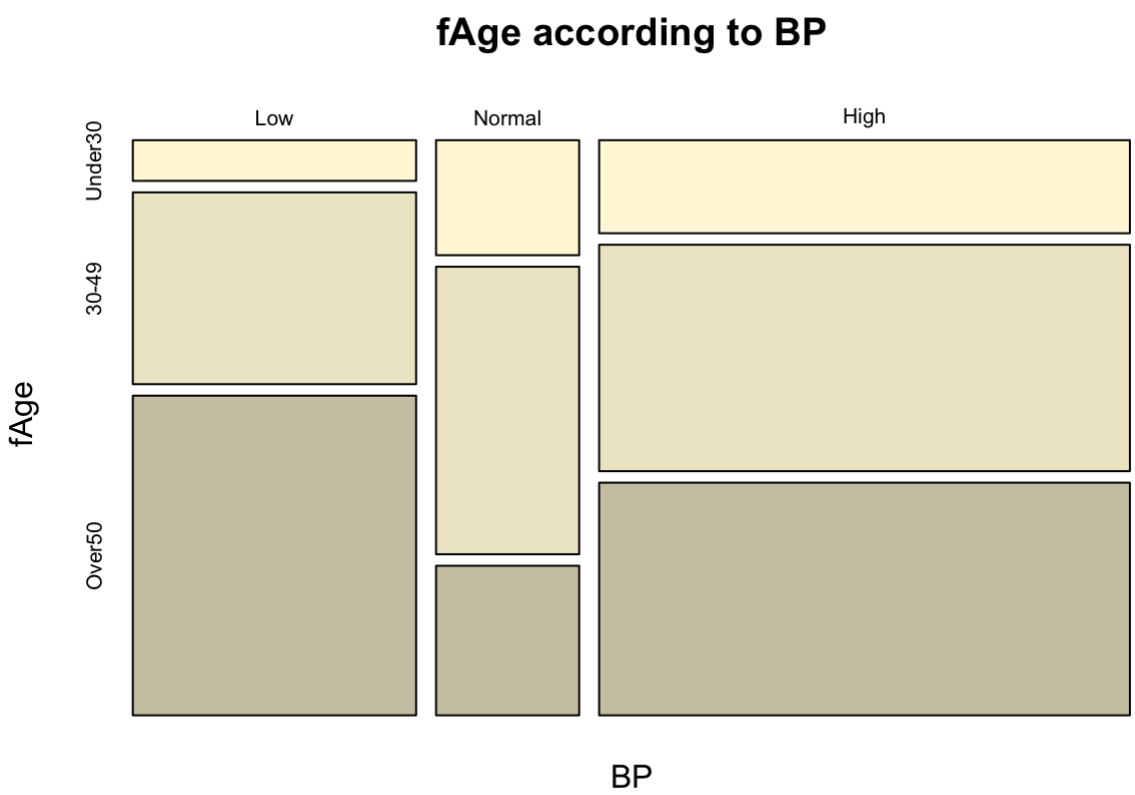
```
df2 <- within(df2, fAge <- cut(df2$Age,breaks = c(-Inf,29,49,Inf),labels = c("Under30","30-49",
"Over50") ) )
```

d. (5 points) 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.

```
(df2tab <- with(df2, table(BP,fAge)))
```

```
##         fAge
## BP       Under30 30-49 Over50
##   Low          7    33     55
##   Normal      10    25     13
##   High        30    73     75
```

```
mosaicplot(df2tab, color = c('cornsilk1','cornsilk2','cornsilk3'),
           main = 'fAge according to BP')
```



We see in the plot that the proportions of low pressure are have most partition in Over50. early Half Low BP is Over50. This points to Normal BP being different from the rest, its have much more 30-49 fAge. For 3 BP, there are very few in the Under30 category.

```
prop.table(df2tab,2)
```

```
##         fAge
## BP        Under30    30-49   Over50
##   Low    0.148936 0.251908 0.384615
##   Normal 0.212766 0.190840 0.090909
##   High   0.638298 0.557252 0.524476
```

We see that 52% of Over50 sample are in the High BP category, in contrast with only around 9% in the Normal category. For 30-49 and Under30 the proportions are similar, with over 55% in the High category and more BP in High than in Other.

e. (8 points) 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 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 under30 categories.

```
colSums(prop.table(df2tab))%*%t(rowSums(prop.table(df2tab))) *321
```

```
##         Low Normal   High
## [1,] 13.910  7.028 26.062
## [2,] 38.769 19.589 72.642
## [3,] 42.321 21.383 79.296
```

All values are above five, so the conditions for the test are satisfied.

```
chisq.test(df2tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df2tab
## X-squared = 15, df = 4, p-value = 0.0048
```

The test gives a p-value smaller than 0.01 , so there is strong evidence to reject the null hypothesis of homogeneous distributions.

We now do Fisher's test:

```
fisher.test(df2tab)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  df2tab
## p-value = 0.0037
## alternative hypothesis: two.sided
```

Again, the p-value is smaller than 0.01 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.

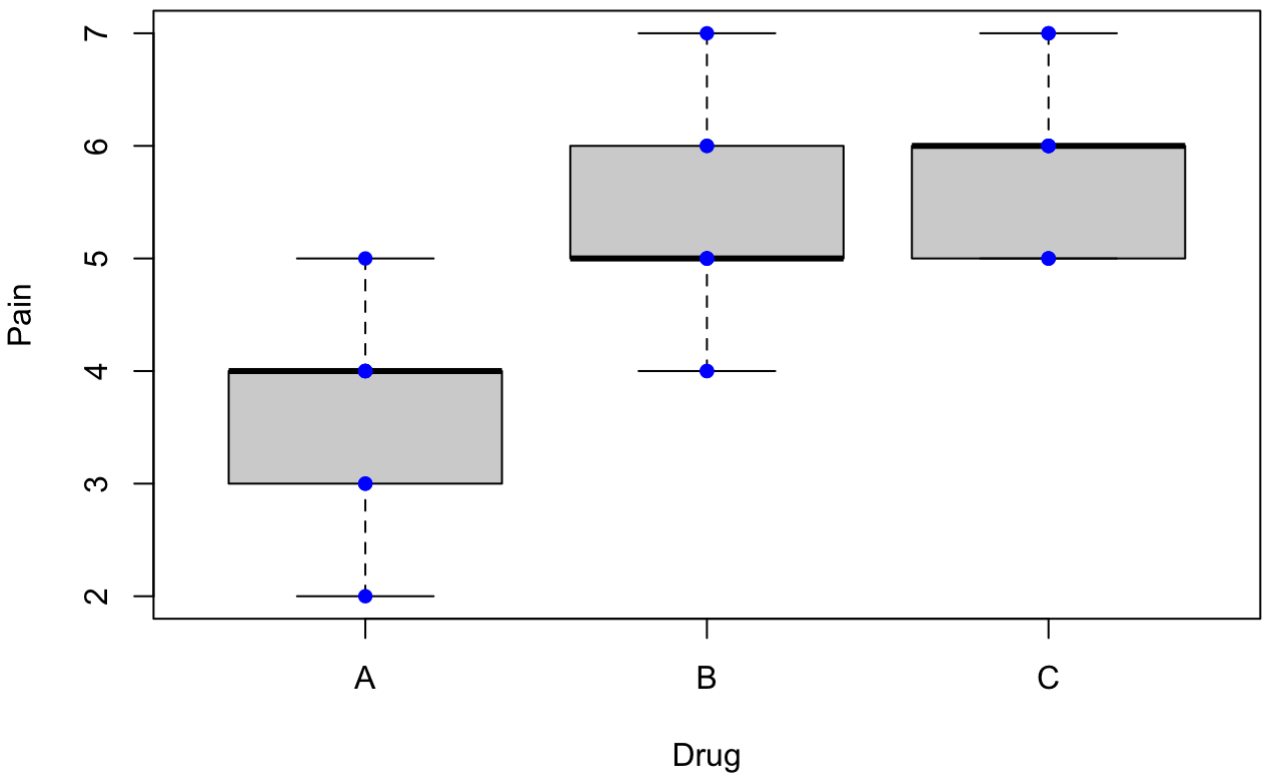
```
df3  =  read.csv('migraine.csv')
str(df3)
```

```
## '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$Drug <- as.factor(df3$Drug)
```

a. (4 points) Do boxplots for `Pain` as a function of `Drug` (all the boxplots should appear on the same panel). Add the points to this graph. Comment on what you observe.

```
boxplot(Pain~Drug,data = df3)
points(Pain ~ Drug, data = df3, pch = 16, col = 'blue')
```



We observe that B drug and C drug have more pain than A drug. And A , B Drug has more Pain vairance than C Drug.

b. (8 points) 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?

```
fit1 <- lm(Pain~Drug, data=df3)
anova(fit1)
```

```
## 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 smaller than 0.02 , so we conclude that there is a difference in the drugs.

c. (10 points) 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 level from this table.

```
summary(fit1)
```

```
##
## 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
```

```
print(fit1)
```

```
##
## Call:
## lm(formula = Pain ~ Drug, data = df3)
##
## Coefficients:
## (Intercept)        DrugB        DrugC
##        3.67         1.56         2.22
```

the meaning of the numbers in the `Estimate` column is the the fitted linear model estimated coefficient.

DrugA : 3.6667 DrugB: 1.5556 DrugC :2.2222

d. (5 points) What are the estimated values for the variance and standard deviation of the errors in this experiment?

The estimated variance comes from the anova table and is 0.7685. the standard deviation is

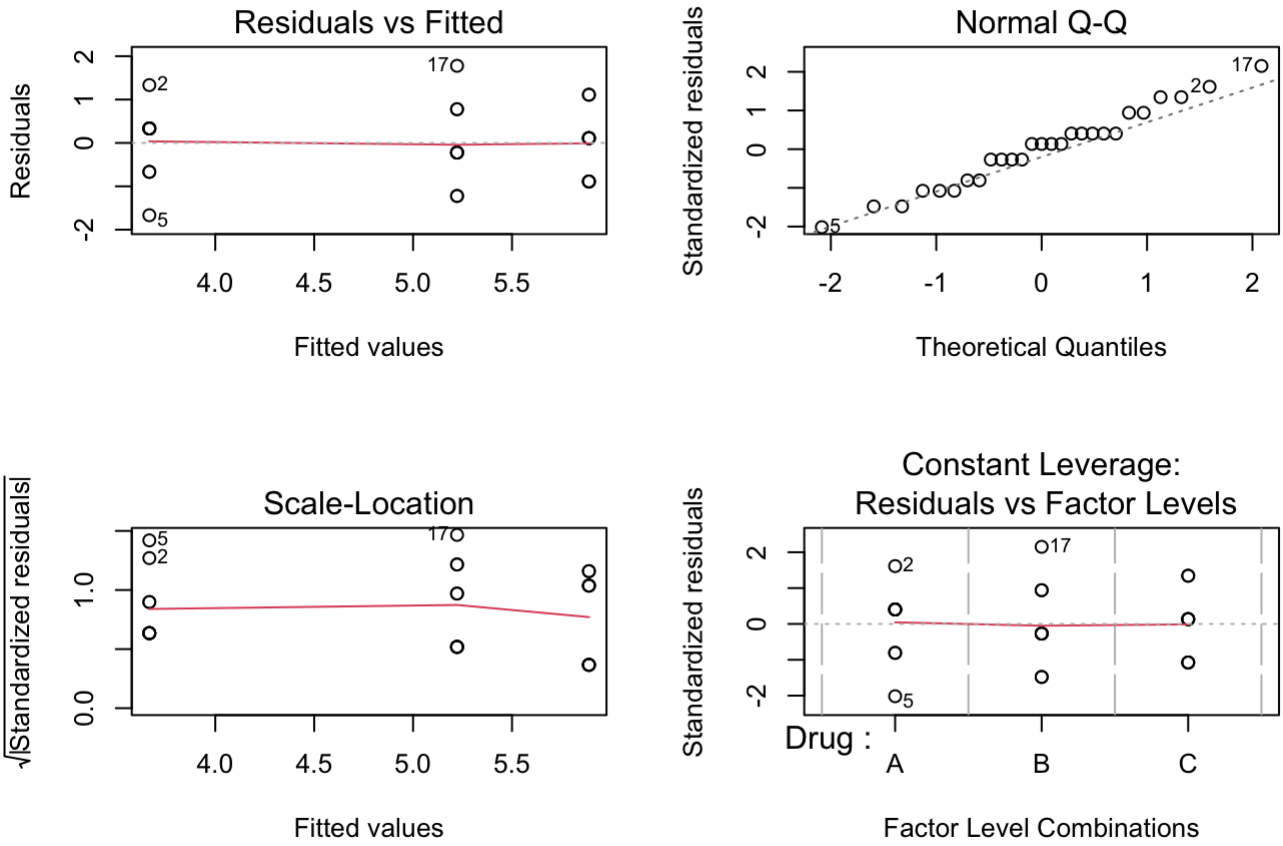
```
Va = 0.7685
sqrt(Va)
```

```
## [1] 0.87664
```

e. (8 points) What are the assumptions on which the analysis of variance model is based? Draw diagnostic plots for checking these assumptions and discuss the results.

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(fit1)
```



The normal quantile plot is very good, and shows that the residuals follow a normal distribution. Three points, 2, 17, and 5, are singled out in the graphs as having largest residuals. The scale-location plot shows an decrease in the average value towards the largest fitted values, but the decrease 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.