Assignment 2: Statistical Inference

Chalkiopoulos Georgios | $p3352124$

November 28, 2021

Exercise 1. Let $Y = (Y_1, \ldots, Y_n)$ denotes n independent random variables with $Y_i \sim P(\theta x_i), \theta \in \Theta = (0, \infty)$ and $x_i > 0$ known constants, $i = 1, \ldots, n$.

1. *Show that the joint probability function $f_Y(y|\theta)$ is a member of the exponential family of distributions.*

We will start the calculation by calculating the joint probability density function:

$$\begin{aligned}
f_y(y|\theta) &= \prod_{i=1}^{n} f_{yi}(y_i|\theta x_i) \\
&= \prod_{i=1}^{n} \frac{(\theta x_i)^{y_i} e^{-\theta x_i}}{y_i!} \\
&= \frac{1}{\prod_{i=1}^{n} y_i!} \prod_{i=1}^{n} exp\{log(\theta x_i) y_i\} e^{-\theta x_i} \\
&= \frac{e^{-\sum_i^n \theta x_i}}{\prod_{i=1}^{n} y_i!} exp\left\{\sum_i^n log(\theta x_i) y_i\right\}
\end{aligned}$$

Based on the above result we conclude the the joint probability function $f_Y(y|\theta)$ is a member of the exponential family of distributions, with where for:

$$f(y|\theta) = h(y)c(\theta)exp\left\{\sum_{i=1}^{k} w_i(\theta)t_i(y)\right\}$$

we have:

$$h(y) = \frac{1}{\prod_{i=1}^{n} Y_i!}$$

$$c(\theta) = exp\left\{-\sum_i^n \theta x_i\right\}$$

$$w_i(\theta) = exp\left\{\sum_{i}^{n} log(\theta x_i)\right\}$$
$$t_i(y) = exp\left\{Y_i\right\}$$

2. *Use the Cramer-Rao inequality to show that* $\delta(Y) = \sum Y_i / \sum x_i$ *is the UMVUE of* θ*.*
Hint: *Notice that the data is not an iid sample, so be careful in the calculation of the denominator of the Cramer-Rao lower bound*

We will start by obtaining a point estimate using the method of maximum likelihood (MLE). Specifically:

$$\begin{aligned}
log(l(\theta)) &= log\left(\frac{e^{-\sum_{i=1}^{n}\theta x_i} exp\{\sum_i^n log(\theta x_i) Y_i\}}{\prod_{i=1}^{n} Y_i!}\right)\\
&= -\sum_{i=1}^{n}\theta x_i + \sum_{i}^{n} log(\theta x_i) Y_i - \prod_{i=1}^{n} Y_i!\\
&= -\sum_{i=1}^{n}\theta x_i + \sum_{i}^{n} log(\theta) Y_i + \sum_{i}^{n} log(x_i) Y_i - \prod_{i=1}^{n} Y_i!
\end{aligned}$$

$$\frac{\partial log(l(\theta))}{\partial\theta} = 0 \Rightarrow -n\sum_{i=1}^{n} x_i + \frac{n}{\theta}\left(\sum_{i}^{n} Y_i\right) = 0$$
$$\Rightarrow \hat{\theta} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i}^{n} x_i}$$

Moreover:
$$\frac{\partial^2 log(l(\theta))}{\partial^2\theta} = -n\frac{\sum_{i=1}^{n} Y_i}{\theta^2} < 0$$

The latter is always smaller than zero, for $\theta = \hat{\theta}$, since $\hat{\theta} > 0$ and $x_i > 0$. For X it is known

We will use the poisson properties to calculate $\delta(Y)$:

$$\delta(Y) = \frac{\sum Y_i}{\sum x_i}$$

For the Y it is known that:

$$E[\bar{X}] = \theta\sum_{i=1}^{n} x_i \; and \; Var[\bar{X}] = \theta\sum_{i=1}^{n} x_i/n$$

Next, we will obtain the Cramer-Rao lower bound, to check whether the variability of the proposed estimate achieves it or not. For the numerator we have:

$$\left(\frac{\partial}{\partial\theta}E_{\lambda}[W(Y)]\right)^{2}=\left(\frac{\partial}{\partial\theta}E_{\lambda}[\bar{Y}]\right)^{2}=\left(\frac{\partial}{\partial\theta}\theta\sum_{i=1}^{n}x_i\right)^{2}=\left(\sum_{i=1}^{n}x_i\right)^{2}$$

By calculating the denominator and the the Variance we get that the calculated $\hat{\theta}$ achieves the C-R lower bound.

3. *Find the MLE of θ. Is it always defined? If not, calculate the probability of the event P(MLE is undefined).*

The MLE of θ was calculated in the previous question. It looks like it is always defined since $x_i > 0$.

4. *Assume that the number of incoming calls Y to a call center within x minutes is distributed according to the $Y \sim P(\theta x)$ distribution, where theta $\in \Theta = (0, \infty)$ denotes an uknown parameter. Estimate θ given the $n = 10$ independent observations in Table 1. What is the MLE of $g(\theta) = P_{\theta}($at least 2 calls within 30 seconds$)$?*

some texst here Based on the previous question we know that the best estimator for θ is:

$$\hat{\theta}=\frac{\sum Y_i}{\sum x_i}$$

i	1	2	3	4	5	6	7	8	9	10
x_i	0.12	2.42	.057	3.21	3.73	0.53	2.95	0.52	7.55	0.34
y_i	1	5	1	14	17	1	15	0	19	2

Table 1: Sample of the number of incoming calls (y) within x minutes.

Given the observations we have that $\hat{\theta} = 0.16$

For the MLE of $g(\theta)$ we will first calculate the requested probability. This will be given for x=0.5, thus:

$$\begin{aligned} g(\theta) &= P_{\theta}(Y >= 2) = 1 - P_{\theta}(Y = 1) \\ &= 1 - \frac{e^{-\theta 0.5}}{1} exp\left\{log(\theta 0.5)1\right\} \\ &\overset{\hat{\theta}=0.16}{=} 0.69 \end{aligned}$$

Exercise 2. A sample of 45 patients was selected in order to measure the response time to two different drugs A and B. Each patient was randomly assigned to one of the two drug types and the response time was recorded. The file ”`drug_response_time .txt`” contains the observed measurements (in hours).

1. *Inspect and visualize the data (you may use summary statistics, histograms, boxplots, etc. within each group).*

We ll start by loading the data:

```
### Ex.  2
# load file
> filename <- './drug_response_time.txt’
> df = read.table(filename, header=TRUE)
> head(df, n=3)
      time   drug
1 4.049562      B
2 3.998332      B
3 4.053662      B
> drug_a <- subset(df, drug == ’A’)
> dim(drug_a)
[1] 20 2
> sd(drug_a$time)
[1] 0.1243305
> drug_b <- subset(df, drug == ’B’)
> dim(drug_b)
[1] 25 2
> sd(drug_b$time)
[1] 0.1094215
```

We also fetch the summary table for each drug. The result may be found on Table 4:
`tapply(df$time, df$drug, summary)`

Drug_A	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	3.702	3.895	3.975	3.962	4.009	4.166
Drug_B	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	3.832	4.008	4.057	4.075	4.147	4.289

Table 2: `drug_response_time.txt` summary table

We will will now generate a frequency table along with a boxplot, comparing the response time for each group.

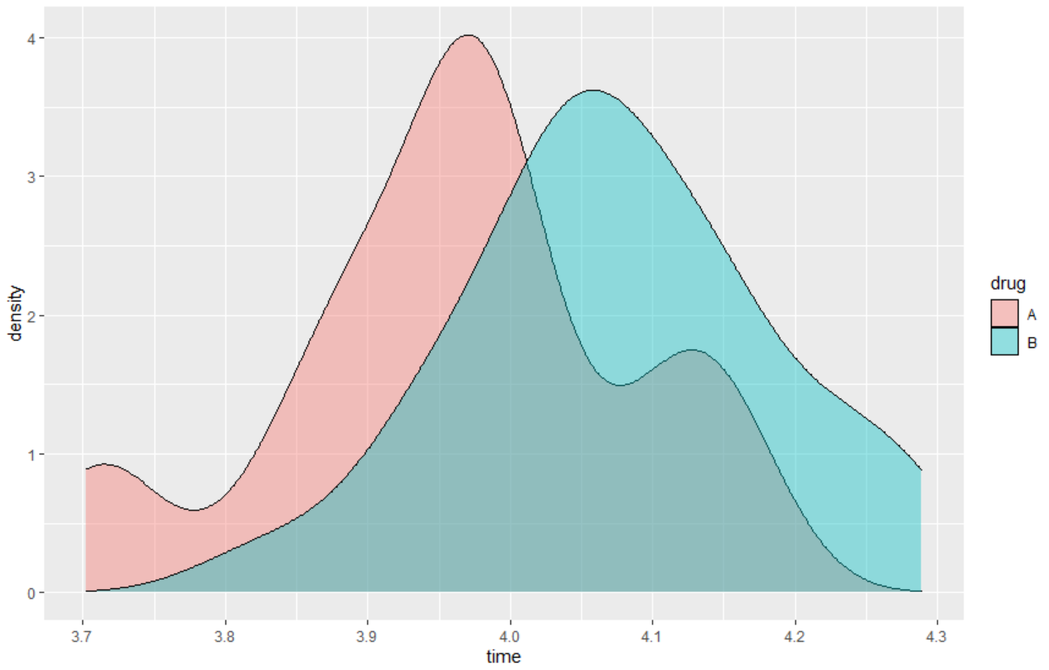

Figure 1: Drug A and B response time density plot

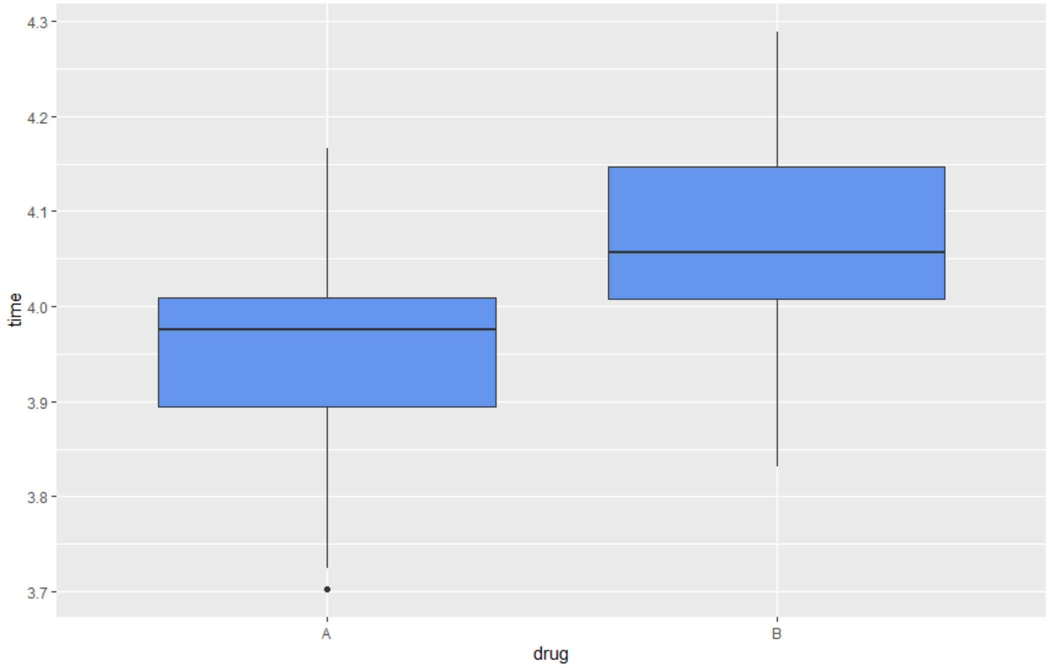

Figure 2: Drug A and B response time box plot

2. *Propose a sensible statistical model in order to describe the data and test all the distributional assumptions within each group ($\alpha = 5\%$).*

In order to describe the data we will assume that both responses follow a normal distribution (hypothesis H_0. We will first create a qqplot, to visually check the normality of the data (Figure 3). Judging from the graph we expect Drug B to be following a normal distribution, while Drug A might not. We will use the Shapiro-Wilk test to check for normality. The result will give a p-value, and if that value is greater that 0.05 then we can say that we failed to reject the hypothesis H_0. Using R we get:

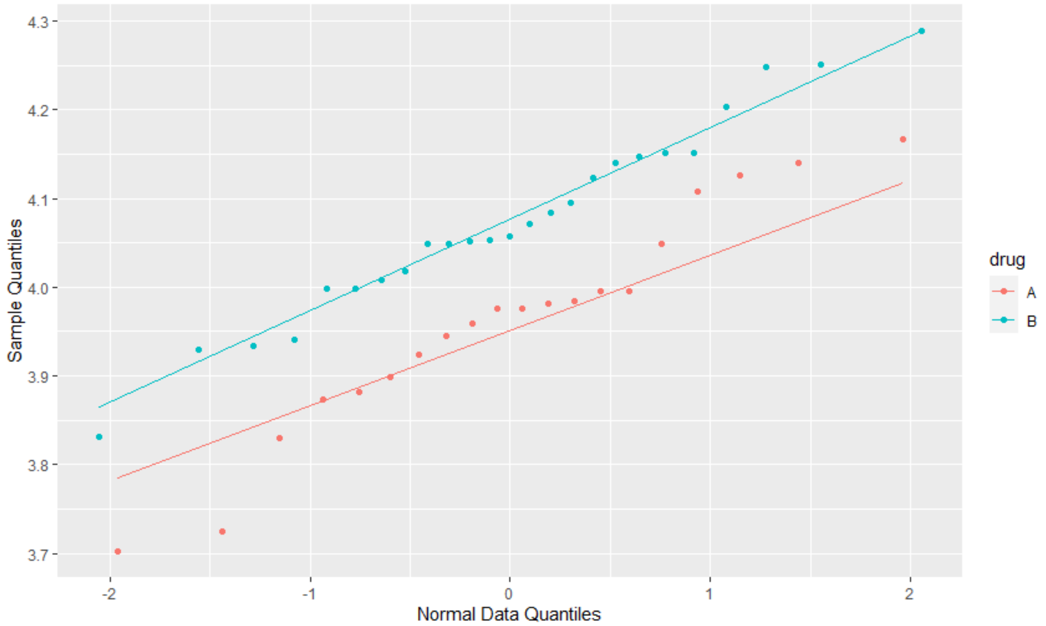

Figure 3: Sample Quantiles vs Normal Data Qantiles.

```
> shapiro.test(drug_a$time)
Shapiro-Wilk normality test
data:  drug_a$time
W = 0.95593, p-value = 0.4662
> shapiro.test(drug_$time)
Shapiro-Wilk normality test
data:  drug_b$time
W = 0.98129, p-value = 0.9094
```

As expected the p-value for drug B is much higher than the corresponding one of drug A. That said, in both cases we failed to reject the Null Hypothesis (H_0), since the p-value received from the Shapiro-Wilk nromality test is greater than 0.05.

3. *Calculate the MLE estimates of unknown parameters.*

We used the `R` library `bbmle` to estimate the mean and std of both drugs. The results we got may be found in Table 3

	Drug_A	**Drug_B**
mu	3.961834	4.074891
sigma	0.121178	0.107212

Table 3: `MLE estimates for drug A and B`

```
tapply(df$time, df$drug, summary)
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Drug_A	3.702	3.895	3.975	3.962	4.009	4.166
Drug_B	3.832	4.008	4.057	4.075	4.147	4.289

Table 4: `drug_response_time.txt` summary table

4. *Calculate manually the equally-tailed 95%, 99% and 99.9% confidence intervals for the difference of mean response time between groups.*

We want to test the following hypothesis:

$$\begin{aligned} H_0 &: \mu_X - \mu_Y = 0 \\ H_1 &: \mu_X - \mu_Y > 0 \end{aligned} \quad \textit{or} \quad \begin{aligned} H_0 &: \mu_X = \mu_Y \\ H_1 &: \mu_X > \mu_Y \end{aligned}$$

,where μ_X is the mean of drug's A response time and μ_Y the corresponding one for drug B. We, therefore, need to calculate the 95, 99 and 99.9 percent CI which in this case falls under the category of **CI for the difference of the means of two independent Normal populations with unknown unequal variances**. The $100(1-\alpha)$ confidence interval is given by:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\nu;a/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where,

$$\nu = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{s_1^4}{n_1^2(n_1-1)} + \dfrac{s_1^4}{n_1^2(n_1-1)}}$$

Using the estimates we calculated in the the previous section (Table 3) we can calculate the needed CI. For the calculation we will use R:

```
> x1 = 3.96183
> x2 = 4.074891
> s1 = 0.121178
> s2 = 0.107212
> n1 = length(drug_a$time)
> n2 = length(drug_b$time)
> nu = ((s1^2/n1) + (s2^2/n2))^2/((s1^4/(n1^2*(n1-1)))
+ (s2^4/(n2^2*(n2-1))))
> ci = c(0.05, 0.01, 0.001)

+ for (alpha in ci){
+ t <- qt(1-alpha/2, nu)
+ s <- sqrt((s1^2/n1) + (s2^2/n2))
+ result <- c(x1-x2-(t*s),x1-x2+(t*s))
+ cat('result for alpha of ', 100*(1-alpha), '% is
+ [', result[1],',',result[2],']')
+ cat('')
}
```

We get the following results:

```
result for alpha of 95 % is [ -0.1829914 , -0.04313057 ]
result for alpha of 99 % is [ -0.2067123 , -0.01940975 ]
result for alpha of 99.9 % is [ -0.2361783 , 0.01005625 ]
```

5. *Validate the confidence intervals computed at the previous question using R*

We will now use the student distribution in R to check the CI at 95, 99, 99.9 percent significance levels.

```
> for (alpha in ci){
+ cat('result for alpha of ', 100*(1-alpha))
+ print(t.test(drug_a$time,drug_b$time,var.equal=TRUE,mu=0
,alternative="two.sided", conf.level = (1-alpha)))
}
```

We got the following results:

```
95 percent confidence interval:
-0.18338671 -0.04272819

99 percent confidence interval:
-0.20704528 -0.01906962

99.9 percent confidence interval:
-0.23621784 0.01010295
```

We observe that the CI calculated with both methods match.

6. *Is there any difference in mean response time between the two drugs at significance level 5%?*

Having calculate the 95% CI we see that the value of zero is outside the interval we have selected. The confidence interval tested the hypothesis that the means of the two drugs' response time is equal. Outside [-0.183, -0.043] this hypothesis is rejected, meaning that we rejected that the two means are equal, thus there is a difference.

Exercise 3. In a random sample of 200 voters of a given region it was found that 80 persons will vote a specific party. Compute a 99% asymptotic confidence interval for the party percentage at the given voting region. Clearly state all modelling assumptions.

We will tackle this problem considering that we have a random sample of $X_i, \ldots, X_n$ where $n = 200$ and X_i a Bernoulli Random Variable. We also consider that $X_i = 1$ if the i^{th} voter supports the candidate. Our goal is to build a 100(1 - α)% CI for the unknown population proportion, that we will call π. If we will call p the sample proportion (i.e. how many in the sample had the characteristic) it is easy to show that p is the MLE of π (i.e. $\hat{\pi}$ = p). Given that n is relatively large we estimate π:

$$\pi = \frac{80}{200} = 0.4 = p$$

We observe that π is neither close to 0 or 1, so we may proceed with the 99% asymptotic confidence interval calculation, which will be given by:

$$p \pm z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$$

where $p = \pi$, thus:

$$\begin{aligned} p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} &= \left[p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right] \\ &= \left[0.4 - 2.576\sqrt{\frac{0.4(1-0.4)}{200}}, p0.4 + 2.576\sqrt{\frac{0.4(1-0.4)}{200}}\right] \\ &= [0.311, 0.489] \end{aligned}$$

Exercise 4. *The joker lottery results are available online at*

https://www.opap.gr/en/joker-draw-results

We are focusing on the distribution of the winning joker number, that is, an integer between 1 and 20. Download the relevant data containing winning joker numbers (column H of the ⋆xlsx files) for the period 2017–2020.

1. *Import the data into R and produce a frequency table of the winning joker numbers for the period 2017–2020. Visualize the data.*

We, first, downloaded the data and combined it into a .csv file. Then, after reading the file, we produced the frequency table as shown in Figure 4.

```
> df2 = read.csv("./joker.csv", header=FALSE)
> joker <- unlist(df2[1])
>  plot the number freq
> ggplot() + aes(x=joker, after_stat(density)) +
+ geom_histogram(binwidth=1,color="lightblue"
+ ,fill="royalblue3") + xlab("Winning Number") +
+ ylab("Frequency")
```

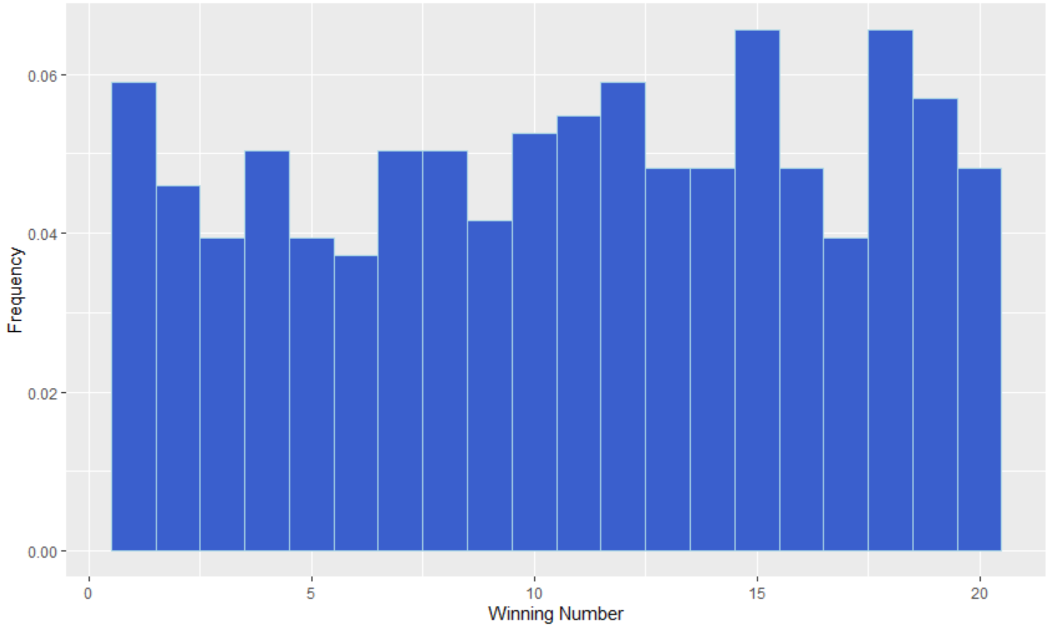

Figure 4: Winning Joker Numbers Freq Table

2. *At significance level 5%, is it a fair lottery?*

A fair lottery should follows a Uniform Distribution, in a way that all numbers have the same probability. We know that the Perason's chi-squared test in R can be used to examine whether an observed frequency distribution differs from a theoretical distribution or not. Using `R` we get the frequencies for each number and we can perform the Chi-squared test without specifying the distribution, since the default prob is the discrete uniform Chi-squared test for given probabilities.

```
> obs <- table(joker)
> chisq.test(obs)

Chi-squared test for given probabilities

data:  obs
X-squared = 12.19, df = 19, p-value = 0.8773
```

The p-value of 0.8773 indicates that the lottery is fair at a significance level of 5%.

Exercise 5. *A pharmaceutical company developed a new drug in order to reduce systolic blood pressure levels in hypertension patients. A random sample of 10 patients with similar lifestyle and medical background is collected. The following table shows the systolic blood pressure level (in mmHg) of each patient before and after receiving the new drug.*

	1	2	3	4	5	6	7	8	9	10
before	121.5	122.4	126.6	120.0	129.1	124.9	138.8	124.5	116.8	132.2
after	117.3	128.6	121.3	117.2	125.6	121.5	124.2	121.6	117.9	125.4

Table 5: Blood pressure measurements.

1. *Visualize the data and report summary statistics.*

We create a dataframe with the required data and run the summary statistics.

```
> id <- c(1:10)
> before <- c(121.5,122.4,126.6,120.0,129.1,124.9,138.8,
124.5, 116.8,132.2)
> after <- c(117.3, 128.6, 121.3, 117.2, 125.6, 121.5,
124.2, 121.6, 117.9, 125.4)
> b <- rep(’before’, 10)
> a <- rep(’after’, 10)
> df4 <- data.frame(id = rep(id, 2), pos = c(b, a),
+ val = c(before, after))
> head(df4,4)
  id    pos   val
1  1 before 121.5
2  2 before 122.4
3  3 before 126.6
4  4 before 120.0
> tapply(df4$val, df4$pos, summary)

$after
 Min.  1st Qu.  Median  Mean  3rd Qu.  Max.
117.2    118.8   121.5 122.1   125.1 128.6

$before
 Min.  1st Qu.  Median  Mean  3rd Qu.  Max.
116.8    121.7 124.7   125.7   128.5 138.8
```

```
> tapply(df4$val, df4$pos, sd)
   after   before
3.883355 6.399097
```

Based on the data calculated we create the following table, where we see that there is a drop in mean and standard deviation. This is also shown in Figure 6.

	Before	After
mu	125.7	122.1
sigma	6.399097	3.883355

Table 6: `Blood Pressure Mean and std (before and after).`

We will also visualize the data by comparing the before and after values for the blood pressure of each patient.

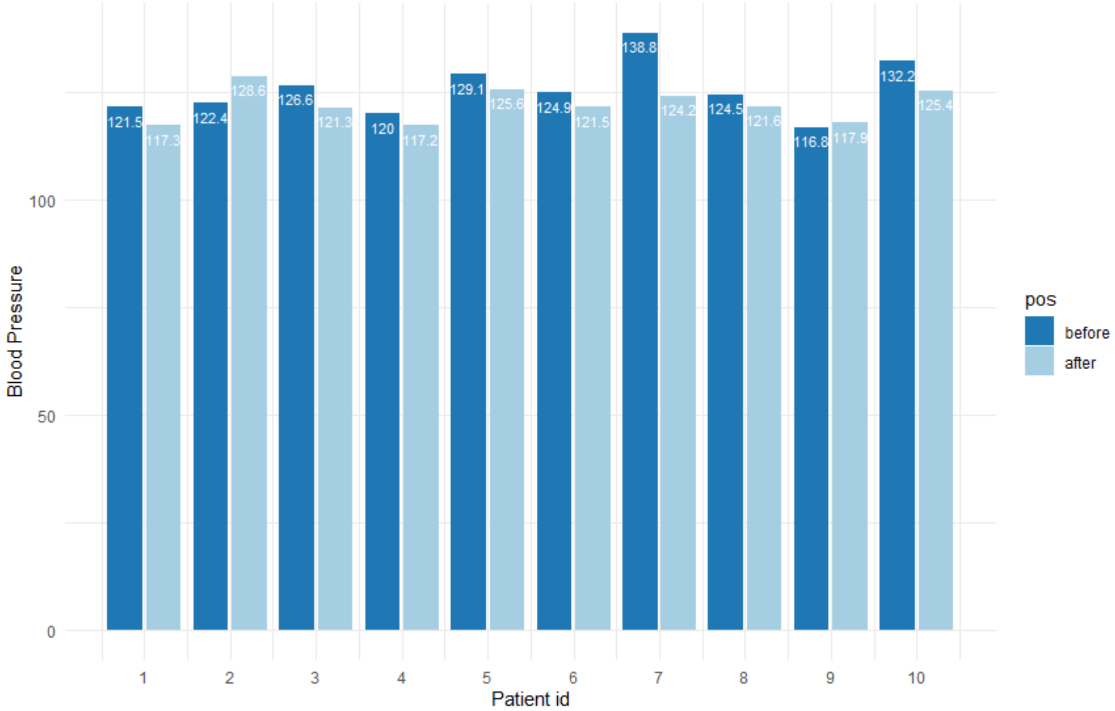

Figure 5: Blood Pressure levels (Before vs After).

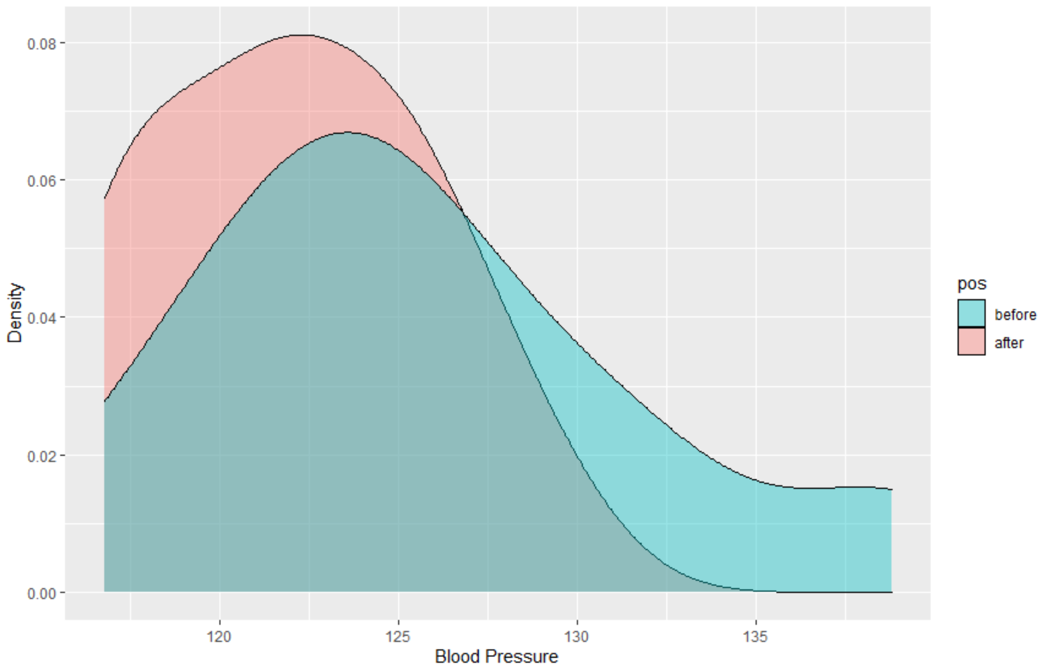

Figure 6: Blood Pressure Density plot (Before vs After).

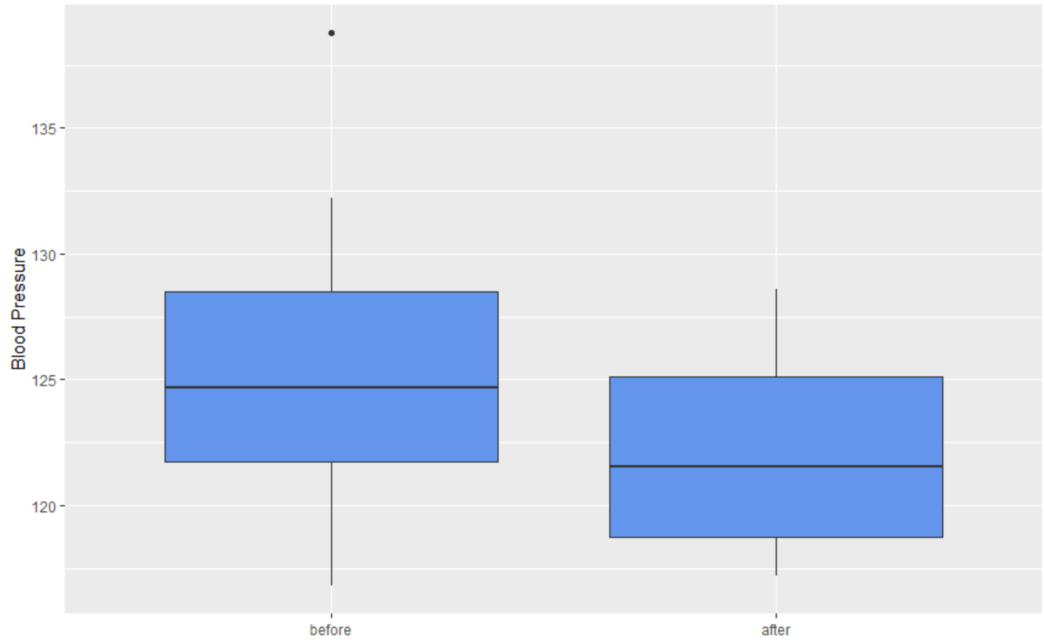

Figure 7: Blood Pressure Box plot (Before vs After).

2. *Test the claim of the pharmaceutical company at significance level 5%.*

In order to test the pharmacautical company's claim we will make use of the t.test command in R. The claim we will make is that the two means are equal and the alternative hypothesis is that the blood pressure mean before taking the drug is higher than the one after taking the drug. If we reject the Null Hypothesis it is an indication that the company's claim is valid, at a significance level of 5%.

```
> d <- before - after
> t.test(d, mu=0, alternative = "greater")

       One Sample t-test

data:  d
t = 2.1557, df = 9, p-value = 0.02974
alternative hypothesis:  true mean is greater than 0
```

The p-value of 0.02974 is smaller than the value of = 0.05 that was needed in order not to reject the Null hypothersis, meaning that the Null hypothesis is rejected, at a significance level of 5%, thus the drug is effective.

Exercise 6. An experiment is conducted in order to assess differences in the ability of recalling previous cards among card players with varying experience levels. For this purpose three categories were selected: novice, advanced and proficient players. The observed data is given in the following table, where each number represents the score in ability.

	novice	advanced	proficient
1	22.10	32.50	40.10
2	22.30	37.10	45.60
3	26.20	39.10	51.20
4	29.60	40.50	56.40
5	31.70	45.50	58.10
6	33.50	51.30	71.10
7	38.90	52.60	74.90
8	39.70	55.70	75.90
9	43.20	55.90	80.30
10	43.20	57.70	85.30

Table 7: Player ability

1. *Import, inspect and visualize the data into* R.

We create the dataset and run the basic statistics as already done in the previous questions:

```
> novice <- c(22.10, 22.30, 26.20, 29.60, 31.70, 33.50,
+ 38.90, 39.70, 43.20, 43.20)
> advanced <- c(32.50, 37.10, 39.10, 40.50, 45.50, 51.30,
+ 52.60, 55.70, 55.90, 57.70)
> proficient <- c(40.10, 45.60, 51.20, 56.40, 58.10, 71.10,
+ 74.90, 75.90, 80.30, 85.30)
> names = c(rep('novice',10), rep('advanced',10),
+ rep('proficient',10))
> df6 <- data.frame(names=names, ability =c(novice, advanced,
+ proficient))
> head(df6)
   names ability
1 novice    22.1
2 novice    22.3
3 novice    26.2
4 novice    29.6
```

```
5 novice    31.7
6 novice    33.5
> df6$names <- factor(df6$names, levels = c("novice",
+ "advanced", "proficient"))
> tapply(df6$ability,df6$names, sd)
  novice advanced proficient
8.033292 9.030621 15.621456
> tapply(df6$ability,df6$names, summary)
```

The result of the last command may be found, aggregated, in Table 8:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
novice	22.10	27.05	32.60	33.04	39.50	43.20
advanced	32.50	39.45	48.40	46.79	54.92	57.70
proficient	40.10	52.50	64.60	63.89	75.65	85.30

Table 8: `Player Ability summary`

We will also visualize the data by group

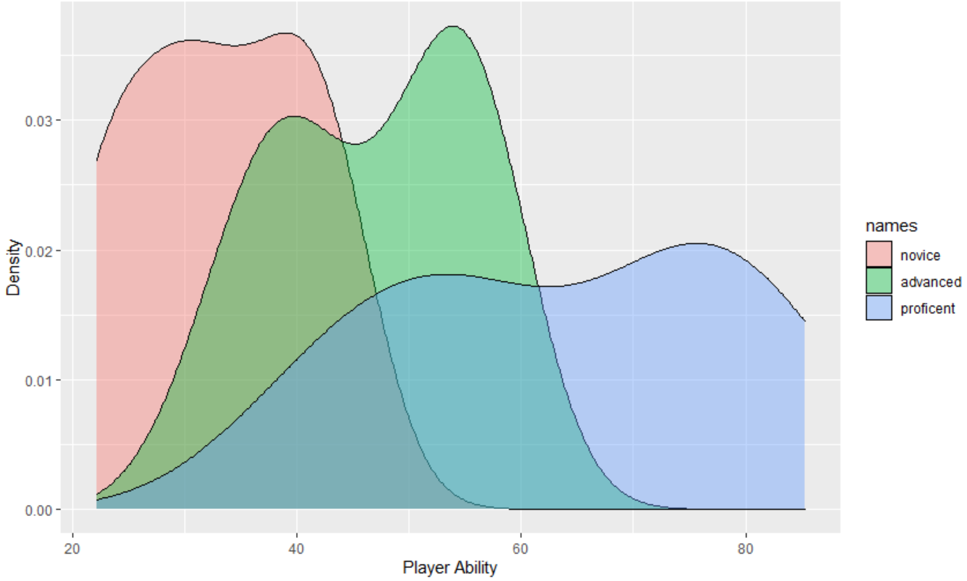

Figure 8: Player ability density function.

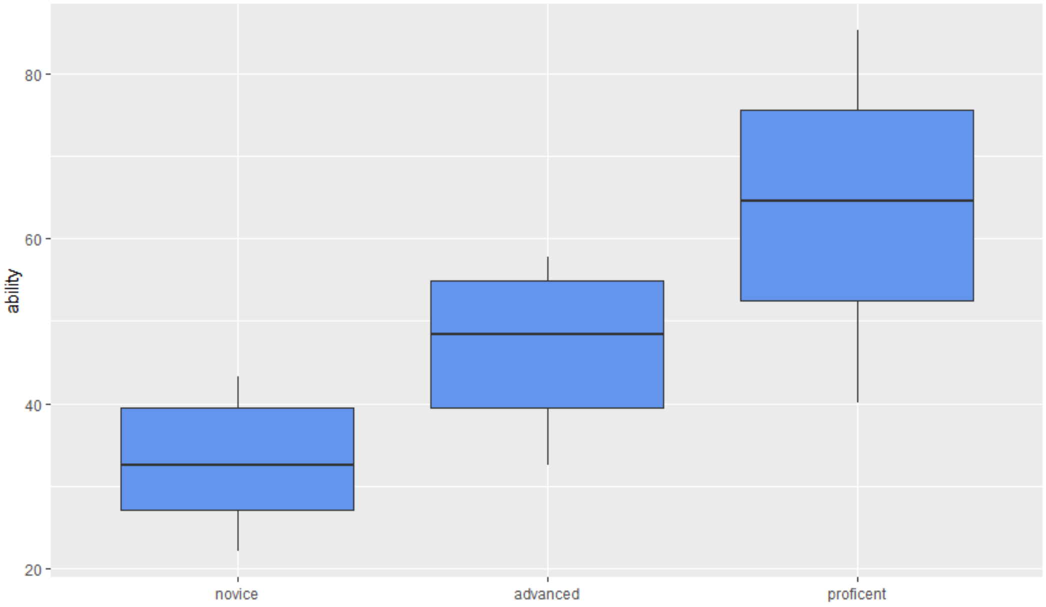

Figure 9: Player ability box plot.

2. *Is there any difference among the groups in the ability of recalling previous cards?*

Look at the graphical representation of the data, as well as the summary tables, we could say that there is a difference between the three groups. We will approach the problem by first assuming that the data comes from a normal distribution with a common Standard Deviation. In this case we may perform an ANOVA test to see if there is any significant difference between the mean of the three groups. The Null Hypothesis states that the mean of all groups is the same. After that we will test the normality and homogeneity of variances, to see if we will accept the normality hypothesis, else we will perform the Kryskal Test.

```
> fit<-aov(ability names,data=df6)
> fit
Call:
  aov(formula = ability   names, data = df6)

Terms:
                     names  Residuals
Sum of Squares    4777.317   3511.042
Deg.  of Freedom         2         27
```

```
Residual standard error:  11.40345
Estimated effects may be unbalanced
> summary(fit)
          Df  Sum Sq Mean Sq F value    Pr(>F)
names      2    4777    2389   18.37 9.21e-06 ***
Residuals 27    3511     130
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

We observe that the F value is high an the p-value is close to zero, thus the null hypothesis is rejected and we may assume that there is difference between the means of the three groups. We will also perform a Shapiro test to check whether the data comes from a normal distribution or not.

```
> qqnorm(fit$residuals,main="NPP for residuals")
> qqline(fit$residuals,col="red",lty=1,lwd=2)
> shapiro.test(fit$residuals)

  Shapiro-Wilk normality test

data:  fit$residuals
W = 0.97185, p-value = 0.5909
```

Looking at Figure 10, as well as the Shapiro test results, we may conclude that the Null Hypothesis is not Rejected, indicating that the data comes from a normal distribution. This conclusion is confirmed by the p-value of 0.5909. Next, we will test the homogeneity of variances, using Bartlett, Fligner-Killeen and Levene's Test for Homogeneity of Variances. The p-value for each individual test may be found in Table 9.

Test	p-value
Bartlett	0.09965
Fligner-Killeen	0.01544
Levene	0.007464

Table 9: Test for Homogeneity of Variances p-values

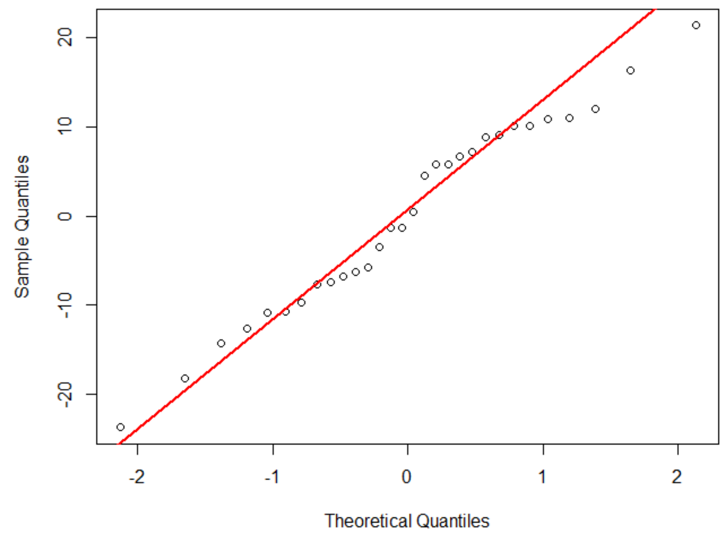

Figure 10: NPP for residuals.

We, therefore, cannot accept the ANOVA results and have to resort to the Kruskal-Wallis test.

```
> kruskal.test(ability names,data=df6)

        Kruskal-Wallis rank sum test

data:  ability by names
Kruskal-Wallis chi-squared = 17.387, df = 2, p-value =
0.0001677
```

The Kruskal Test resulted in a p-value smaller than 1% thus the initial assumption, that the means of the three groups are equal, is rejected, hypothesis is rejected and we may assume that there is difference between the means of the three groups.

3. *Perform all pairwise t-tests. Is it valid to decide which groups are different by performing all pairwise comparisons via a t-test with significance level 5%?.*

First we will attempt a graphical approach, providing the mean plot with their respective confidence interval, which will give visual evidence for the overlapping of mean confidence intervals among different groups. Looking at Table 11 we see that there is overlap between the means of the advanced and proficient groups, at the given significance level. We cannot make conclusions for mean based on the graph though, since the significance level of a pair is different than the one of the individuals.

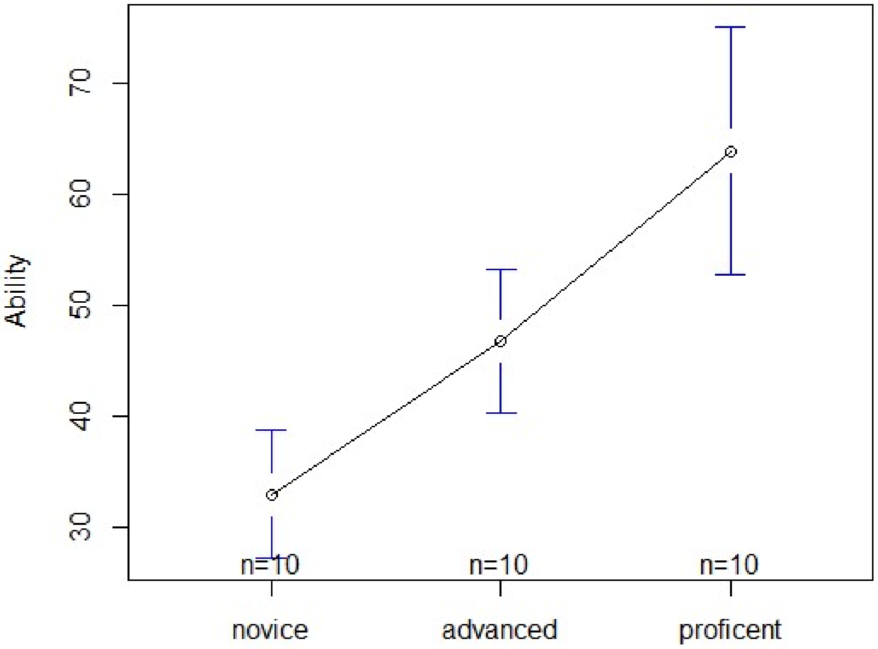

Figure 11: Mean plot with 95% CI

We will proceed with the pairwise t-tests, without adjusting the p-value. With this method we cannot decide if the groups are different based on the p-value reported, since it lacks adjustment. This will be done in the next step of this exercise.

```
> pairwise.t.test(df6$ability,df6$names,
+p.adjust.method = ’none’, pool.sd = FALSE)
           Pairwise comparisons using t tests with non-pooled
SD
data:  df6$ability and df6$names
           novice  advanced
advanced   0.0021        -
proficient 8.2e-05    0.0094

P value adjustment method:  none
```

4. *Properly identify which groups are different at significance level 5%.*

We will now perform various tests in order to check the difference of means at a significance level of 5%. The results may be found in Table 10, along with the `R` code provided with this report. In the following table the letter n, a, p stand for Novice, Advanced and Proficient and the corresponding p-value of each pair is reported for each method.

Method	**advanced-novice**	**proficient-novice**	**proficient-advanced**
TukeyHSD	0.0310169	0.0000054	0.0065079
pairwise - bonferroni	0.0358	5.6e-06	0.0071
pairwise - holm	0.0119	5.6e-06	0.0048

Table 10: `Pairwise adj. p-values.`

We observe that all the values p-values are smaller than 0.05, which is the specified significance level we have defined, thus the Null Hypothesis, that the means of the three groups are equal, is rejected. It is worth noting that based on the mean plot with 95%CI (Table 11) the means of the proficient and advanced group seemed to have an overlap, which after performing the pairwise tests was rejected.

Exercise 7. A researcher wishes to assess whether alcohol consumption is related to the way we choose partners and whether this also depends on gender. For this purpose, the following experiment is conducted: 48 students (24 male and 24 female) were divided into three groups with 8 persons per gender. All groups went into a club for a night out. The students in the first group consumed non-alcoholic drinks. Each student in the second group consumed 2 pints of beer. Each student in the third group consumed 4 pints of beer. During the night, each participant met other people (previously unknown to them) and started flirting. The researcher took a photo of each student's partner and the photos was shown to them in a random order, one month after the night out. Each participant rated the subject on each photo using a scale from 0 to 100. The results are illustrated below.

Alcohol	None		2 pints		4 pints	
Gender	Female	Male	Female	Male	Female	Male
	65	50	70	45	55	30
	70	55	65	60	65	30
	60	80	60	85	70	30
	60	65	70	65	55	55
	60	70	65	70	55	35
	55	75	60	70	60	20
	60	75	60	80	50	45
	55	65	50	60	50	40

Table 11: Partner's attractiveness per gender and alcohol level.

1. *Import, inspect (using descriptive statistics) and visualize the data into* R.

We create the corresponding database in R and visualize the data. For the summary statistics we subset the dataframe per Number of Pints and get the summary per Gender as shown in Table 12.
Furthermore we will plot the density and box plots, as done in the previous questions, but we will also plot the interaction plot, in order to see if Alcohol and Gender interact with each-other. For the sake of simplicity, we will only show the data import part of the R process, since the next steps (summary and plots) have been shown in the previous quesitons.

```
> alcohol <- c(c(rep('None', 16)), c(rep('2 pints', 16)),
+ c(rep('4 pints', 16)))
> gender <- rep(c(c(rep('Female', 8)), c(rep('Male', 8))), 3)
> rating <- c(65, 70, 60, 60, 60, 55, 60,
             55, 50, 55, 80, 65, 70, 75,
             75, 65, 70, 65, 60, 70, 65,
             60, 60, 50, 45, 60, 85, 65,
             70, 70, 80, 60, 55, 65, 70,
             55, 55, 60, 50, 50, 30, 30,
             30, 55, 35, 20, 45, 40)
> df7 <- data.frame(alcohol=alcohol, gender=gender,
+ rating=rating)
```

Alcohol	Gender	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std
None	**Female**	55.00	58.75	60.00	60.62	61.25	70.00	4.96
	Male	50.00	62.50	67.50	66.88	75.00	80.00	10.33
2 Pints	**Female**	50.00	60.00	62.50	62.50	66.25	70.00	6.55
	Male	45.00	60.00	67.50	66.88	72.50	85.00	12.52
4 Pints	**Female**	50.00	53.75	55.00	57.50	61.25	70.00	7.07
	Male	20.00	30.00	32.50	35.62	41.25	55.00	10.84

Table 12: `Partner's attractiveness summary table.`

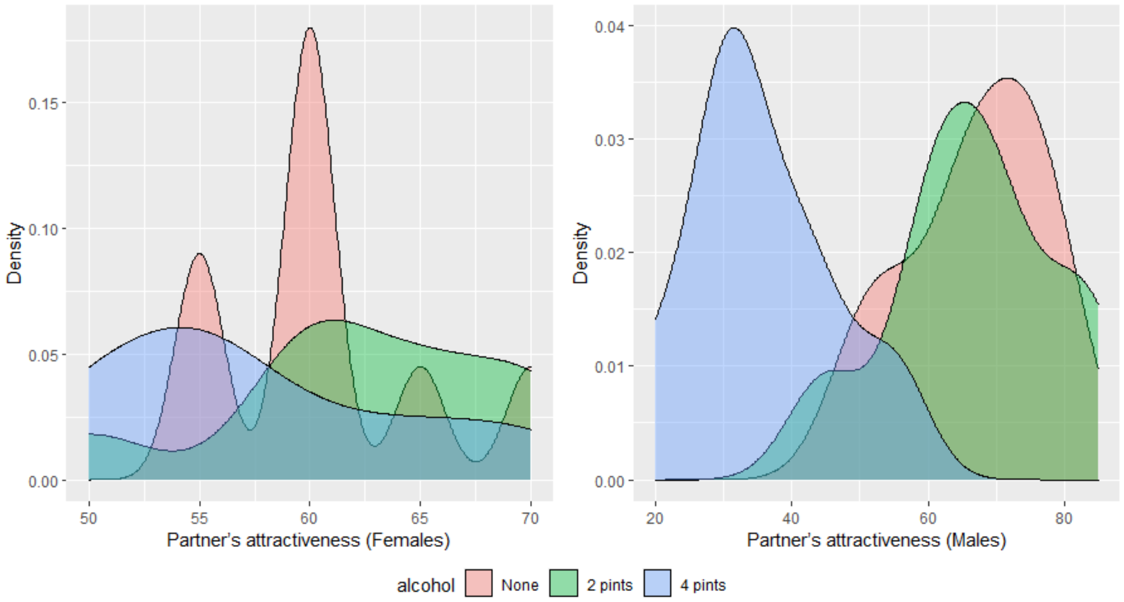

Figure 12: Attractiveness Density plot per Gender and Alcohol.

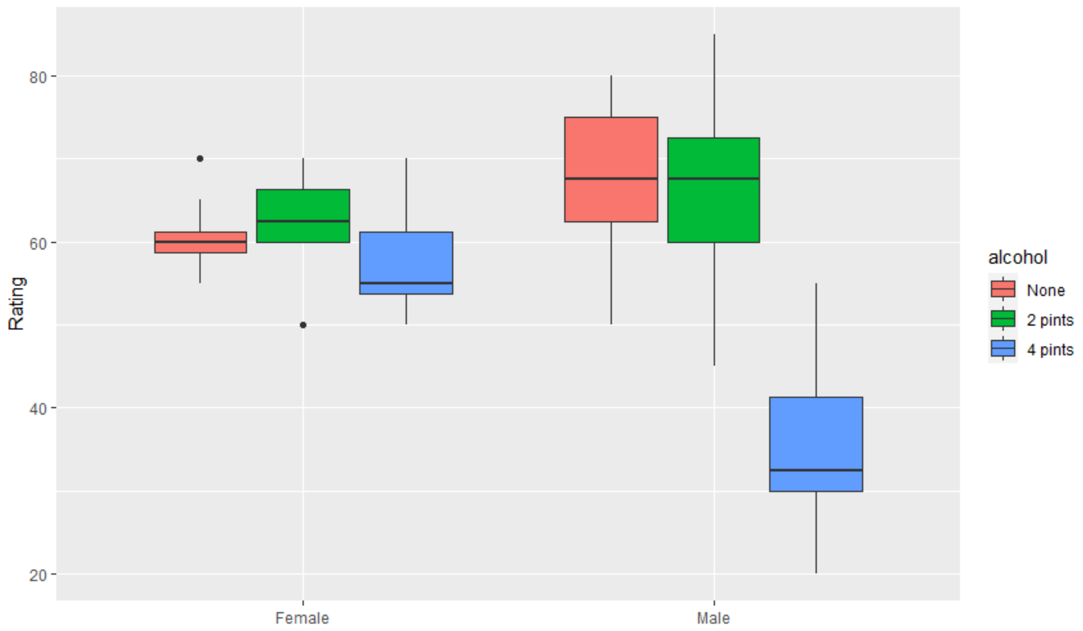

Figure 13: Boxplot per Gender and Alcohol.

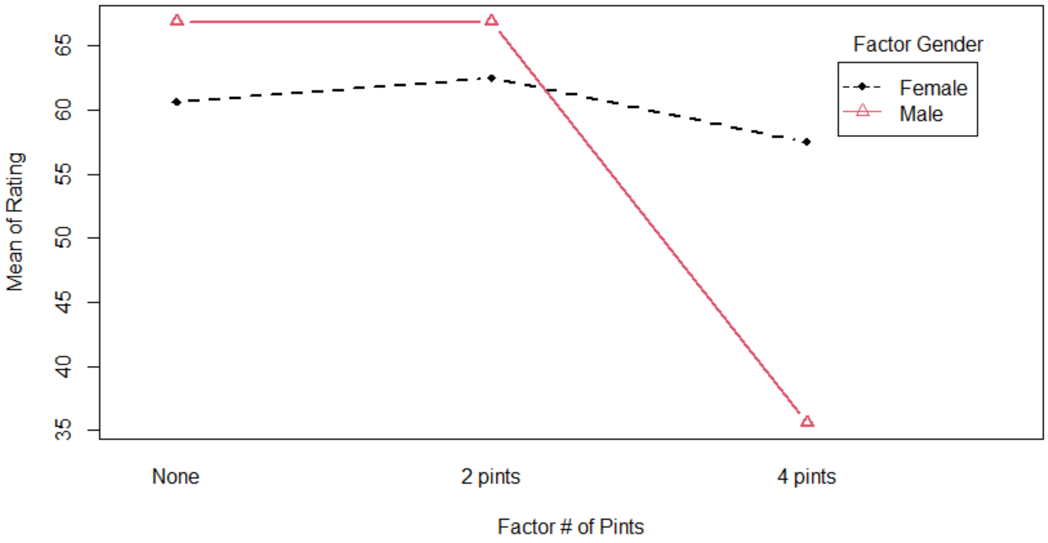

Figure 14: Gender and Alcohol interaction plot.

2. *Propose a statistical model in order to describe the data.*

The initial suggestion is to use the two way ANOVA approach. In order to decide if we can use this approach, we will first test the normality of the data, along with the homogeneity of variances. For the normality test we will use the Shapiro test and for the Homogeneity of Variances we will perform multiple tests. Finally, we assume that there is independence within each factor and across different factors.

```
> qqnorm(fit$residuals,main="NPP for residuals")
> qqline(fit$residuals,col="red",lty=1,lwd=2)
> shapiro.test(fit$residuals)

  Shapiro-Wilk normality test

data:  fit$residuals
W = 0.98201, p-value = 0.6643
```

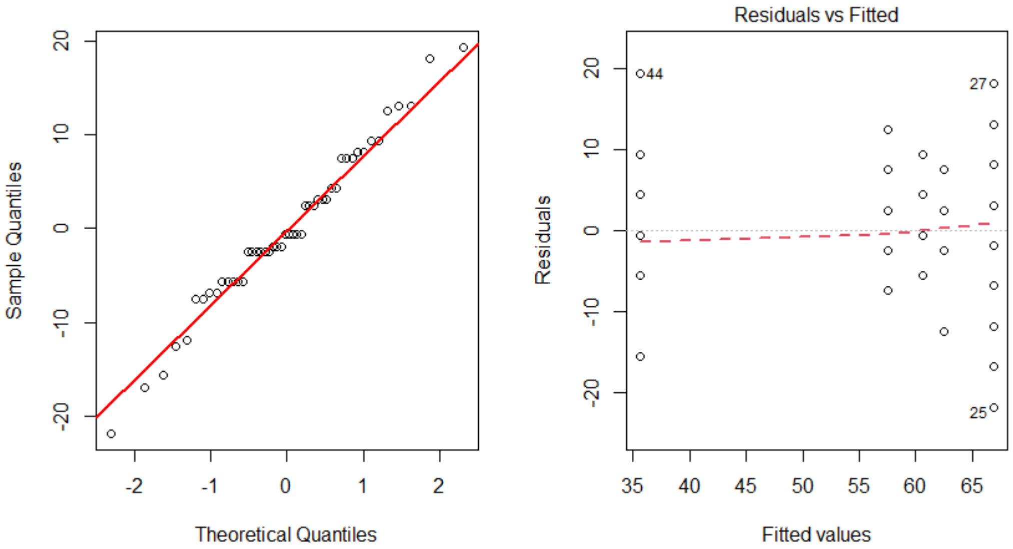

Figure 15: NPP for residuals. and Residuals vs Fitted.

Looking at Figure 15, as well as the Shapiro test results, we may conclude that the Null Hypothesis, for the normality of the data, is not Rejected, indicating that the data comes from a normal distribution. This conclusion is confirmed by the p-value of 0.6643. Next, we will test the homogeneity of variances, using Bartlett, Fligner-Killeen and Levene's Test for Homogeneity of Variances. The p-value for each individual test may be found in Table 13.

Test	p-value
Bartlett	0.1761
Fligner-Killeen	0.2478
Levene	0.2351

Table 13: Test for Homogeneity of Variances p-values

From the output above we can see that the p-value is not less than the significance level of 0.05. This means that there is no evidence to suggest that the variance across groups is statistically significantly different. Therefore, we can assume the homogeneity of variances in the different groups.

Based on the assumptions above we may proceed with the ANOVA test.

3. *Test whether the partner selection depends on alcohol and/or gender (α = 0.05).*

Having established our model above, we may run the summary for the two way ANOVA test.

```
> summary(fit)
               Df  Sum Sq Mean Sq F value    Pr(>F)
alcohol         2    3332  1666.1  20.065 7.65e-07 ***
gender          1     169   168.7   2.032    0.161
alcohol:gender 2    1978   989.1  11.911 7.99e-05 ***
Residuals      42    3487    83.0
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1
```

The Null hypothesis of the two way ANOVA test states that each individual factor as well as their combination is not affecting the results. Based on the above we observe that the Null hypothesis is rejected for the Alcohol parameter as well as the interaction of Alcohol and Gender (also shown in Figure 14), meaning that Alcohol and the Interaction of Gender and Alcohol are affecting the results, but not Gender, at a significance level of 5%.

4. *Test all modelling assumptions.*

All modelling assumptions have been tested in Question 2, where we proposed the statistical model.

5. *Briefly explain your findings into someone who doesn't know anything about statistics.*

The purpose of this experiment was to check whether Alcohol and/or Gender affect the way we flirt and see other people. In order to test that we picked three groups of 8 Males and 8 Females and split them into three groups. Each group consumed different levels of Alcohol. After a month the participants were requested to rate the people of the group based on the attractiveness.
The results of this test revealed that the way we see people is not affected by the Gender of the person that is rating but Alcohol has an impact on the way we see other people.