

Experimental Design and Data Analysis - Assignment 1

Group 75 - Sunny Soni, Maxine Palmen, Muska Neek

2022-02-28

Question 1

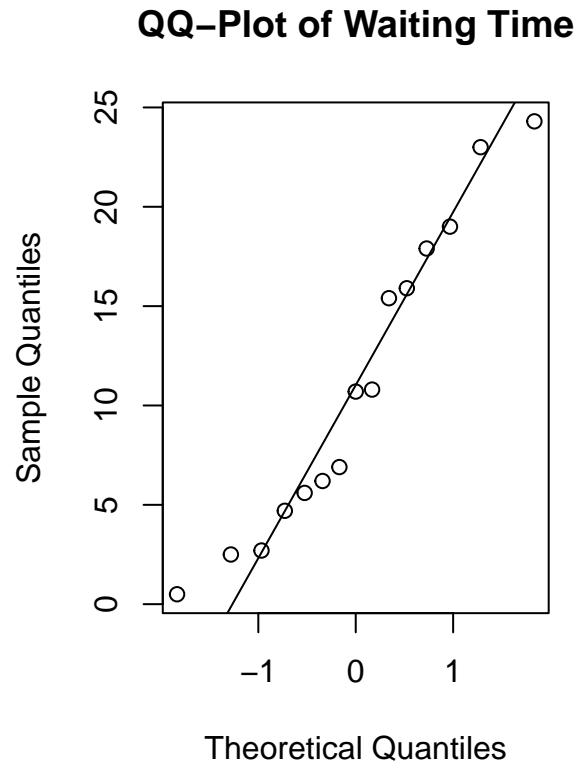
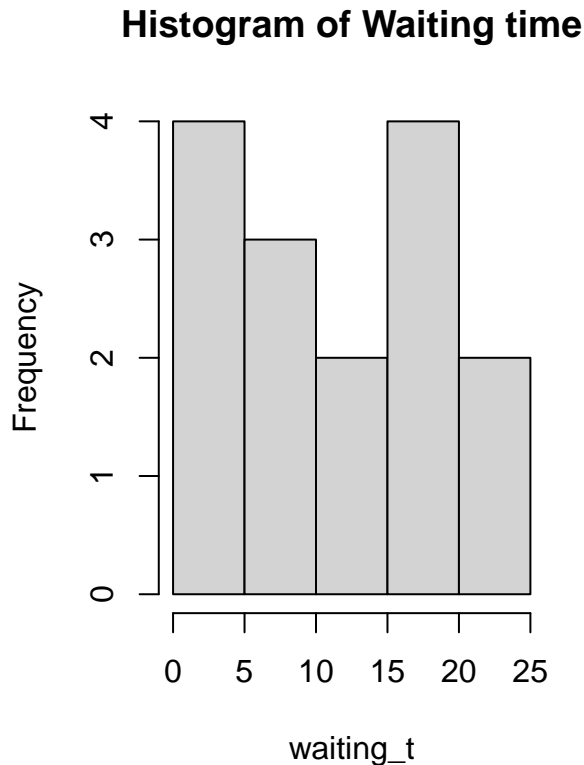
1A Below, the measured waiting time in the waiting room of doctor's office is given of 15 people. This waiting time is represented in a vector of length 15, where the mean waiting time is a little more than 11 minutes.

```
waiting_t <- c(15.4, 17.9, 19.0, 0.5, 15.9, 2.7, 6.2, 2.5, 4.7, 6.9, 10.8, 24.3, 5.6, 23.0, 10.7)
mu_t = mean(waiting_t)
format(round(mu_t, 3), nsmall = 3) #rounding off to 3 decimals places
```

```
## [1] "11.073"
```

There are several ways to check normality of the waiting time data, such as a histogram, QQ-plot, Shapiro-wilk Test. The Histogram and QQ-plot are graphical ways to check normality, and the Shapiro-wilk test is a statistical test to check normality. In the Shapiro-wilk test, the null-hypothesis claims that the sample data comes from a normal distribution, and the alternative hypothesis claims that the sample data does not come from normal distribution. This means that when the p-value is above the confidence level of $\alpha = 0.05$, that the sample data does come from normal distribution since the null hypothesis cannot be rejected.

Below, we will perform all of these normality checks.



According to the histogram, it seems that the waiting time data does not come from a normal distribution. However, the sample size is very small, which could influence this graphical test. The QQ-plot shows a quite straight line of the sample data and therefore this could mean that the sample data comes from normal distribution. To make the normality check final, the Shapiro-Wilk test is performed to accept this assumption that the sample data comes from normal distribution.

The result of the Shapiro-wilk test shows a p-value of 0.3. This means that the data of the waiting time is normally distributed. To construct a 97%-CI, the following formula needs to be computed: $\text{mean}(\text{sample}) \pm z^* \text{se}$ where se is the standard error which is computed by: sd/\sqrt{n} where n is the sample size.

The upper and lower bound of 97%-CI is [6.743, 15.403]

Bootstrap

```
B=1000
Tstar = numeric(B)
for(i in 1:B){
  Xstar=sample(waiting_t, replace=TRUE)
  Tstar[i]=mean(Xstar)}
Tstar15=quantile(Tstar, 0.015)
Tstar985=quantile(Tstar, 0.985)
sum(Tstar<Tstar15)
round(c(2*mu_t-Tstar985, 2*mu_t-Tstar15), 3)
```

The 97% bootstrap CI is [6.61, 14.86] The original CI was [6.743, 15.403], which means that the bootstrap 97%-CI and the original 97%-CI do not differ much.

1B Claim: mean waiting time < 15 H0: mean waiting time >= 15 H1: mean waiting time < 15 Since the data comes from a normal distribution (see normality check at top of document), a t-test can be used. We only have one data sample, and therefor perform a one sample t-test.

The p-value for this one-sample t-test is 0.03, which means that the null-hypothesis can be rejected and that the mean of the data sample indeed is less than 15. The CI-interval for this t-test is $[-\infty, 14.6]$, which means that we can be 95% sure that the mean of the data sample lies within the CI-interval.

In the sign test we care about the median instead of the mean. Our hypotheses H_0 and H_1 are still the same as in the one-sample t-test, but instead of mean we use median. The hypothesized median is 15.

For this problem, the Wilcoxon signed rank test could also be performed. This test needs a sample from a symmetric population and cares about the median.

1C Compute powers of t-test and sign test

By comparing the power of the t-test and the sign-test, we can see that the power of the t-test is higher than the power of the sign test. Therefore, the t-test is a better test for this application. A reason for this is that a t-test depends on normality and the data comes from a normal distribution. So, therefore the t-test is more reliable.

1D The recovered confidence interval of the probability that a patient has to wait longer than 15.5 minutes is $[0.137, 0.53]$, and this has a confidence level of 89.39%

1E Test of proportions H_0 : Proportion of long waiters of man is equal to proportion of long waiters of woman H_a : Proportions are different

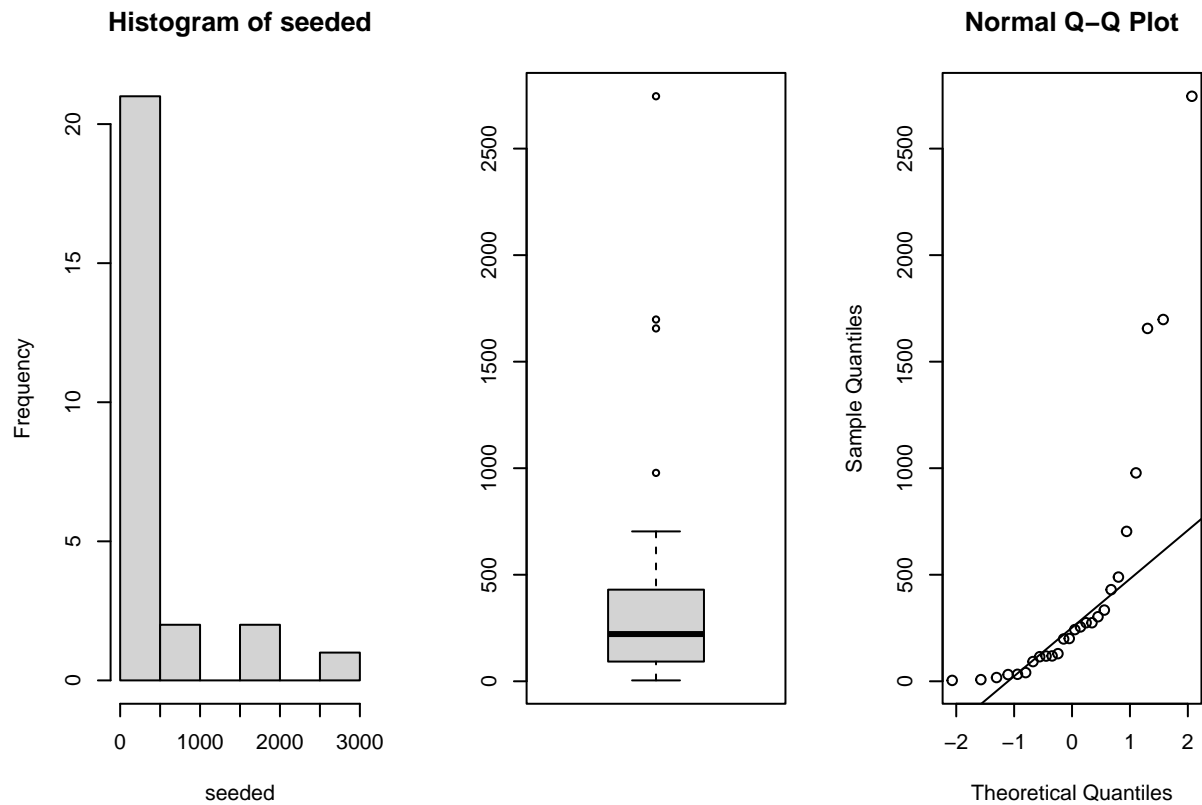
```
## Warning in prop.test(x = c(3, 2), n = c(7, 8)): Chi-squared approximation may be
## incorrect
```

The p-value is 0.853, which is much larger than 0.05. Therefore, the waiting time for man and woman are the same.

Question 2

2A The Kolmogorov-Smirnov normality test examines if variables are normally distributed, which it is not because the data is exponentially distributed. The Mann-Whitney U test is used to compare whether there is a difference in the dependent variable for two independent groups. It compares whether the distribution of the dependent variable is the same for the two groups and therefore from the same population. The setting needs to be symmetric, which our data is clearly not, therefore not applicable. A t-test can only be used when comparing the means of two groups (a.k.a. pairwise comparison), which is applicable.

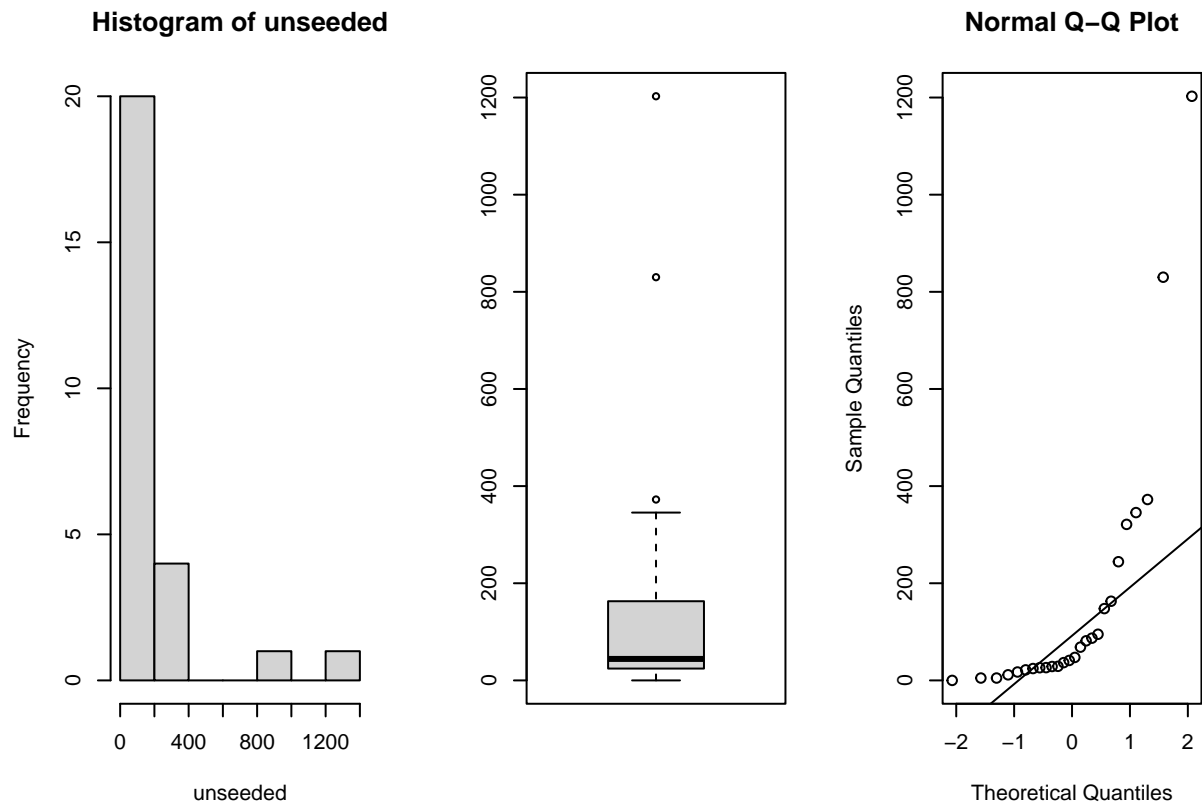
```
#check for normality, which they are not as seen in the following tests
clouds=read.table(file="clouds.txt",header=TRUE)
seeded = clouds[,1]
unseeded = clouds[,2]
par(mfrow=c(1,3))
hist(seeded)
boxplot(seeded)
qqnorm(seeded)
qqline(seeded)
```



```
shapiro.test(seeded) # extra test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  seeded
## W = 0.656, p-value = 1.4e-06
```

```
hist(unseeded)
boxplot(unseeded)
qqnorm(unseeded)
qqline(unseeded)
```



```
shapiro.test(unseeded) #extra test
```

```
##
## Shapiro-Wilk normality test
##
## data: unseeded
## W = 0.603, p-value = 3.2e-07
```

```
t.test(seeded, unseeded)
```

```
##
## Welch Two Sample t-test
##
## data: seeded and unseeded
## t = 2, df = 33.9, p-value = 0.054
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -4.7405 559.5859
## sample estimates:
## mean of x mean of y
## 441.98 164.56
```

```
wilcox.test(seeded,unseeded)
```

```
## Warning in wilcox.test.default(seeded, unseeded): cannot compute exact p-value
## with ties
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data:  seeded and unseeded
## W = 473, p-value = 0.014
## alternative hypothesis: true location shift is not equal to 0
```

```
ks.test(seeded,unseeded)
```

```
## Warning in ks.test(seeded, unseeded): cannot compute exact p-value with ties
```

```
##
## Two-sample Kolmogorov-Smirnov test
##
## data:  seeded and unseeded
## D = 0.423, p-value = 0.019
## alternative hypothesis: two-sided
```

2B Repeat the procedures from a) first on the square root of the values in clouds.txt, then on the square root of the square root of the values in clouds.txt. Comment on your findings.

Except for the t-test whereas the mean of both values are smaller, the p-values are unchanged in the Mann-whitney and Kolmogorov-Smirnov test, which is plausible and explainable when looking at their formulas, because it takes the proportions into consideration. Therefore leading to the same pvalue.

```
t.test(sqrt(seeded), sqrt(unseeded))
```

```
##
## Welch Two Sample t-test
##
## data:  sqrt(seeded) and sqrt(unseeded)
## t = 2.42, df = 43.4, p-value = 0.02
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.2021 13.0713
## sample estimates:
## mean of x mean of y
##  17.0680  9.9313
```

```
wilcox.test(sqrt(seeded),sqrt(unseeded))
```

```
## Warning in wilcox.test.default(sqrt(seeded), sqrt(unseeded)): cannot compute
## exact p-value with ties
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data:  sqrt(seeded) and sqrt(unseeded)
## W = 473, p-value = 0.014
## alternative hypothesis: true location shift is not equal to 0
```

```
ks.test(sqrt(seeded),sqrt(unseeded))
```

```
## Warning in ks.test(sqrt(seeded), sqrt(unseeded)): cannot compute exact p-value
## with ties
```

```
##
## Two-sample Kolmogorov-Smirnov test
##
## data: sqrt(seeded) and sqrt(unseeded)
## D = 0.423, p-value = 0.019
## alternative hypothesis: two-sided
```

2C H_0 : data is from a exponential distribution given that $p > 0.05$, also in KS-test, then the H_0 can not be rejected. Therefore, in conclusion of both tests the data is indeed from an exponential distribution.

```
lambda = 1 / mean(seeded)
T1 = mean(seeded)
z_a = qnorm(.975)
ci_r = mean(seeded)+z_a*(sd(seeded)/sqrt(26))
ci_l = mean(seeded)-z_a*(sd(seeded)/sqrt(26))
B=1000
Tstar=numeric(B)
for(i in 1:B) {
  Xstar=rexp(seeded,lambda)
  Tstar[i]=median(Xstar) }
pl=sum(Tstar<T1)/B; pr=sum(Tstar>T1)/B; p=2*min(pl,pr)
pl;pr;p
```

```
## [1] 0.92
```

```
## [1] 0.08
```

```
## [1] 0.16
```

```
ks.test(seeded, pexp(26, lambda))
```

```
## Warning in ks.test(seeded, pexp(26, lambda)): cannot compute exact p-value with
## ties
```

```
##
## Two-sample Kolmogorov-Smirnov test
##
## data: seeded and pexp(26, lambda)
## D = 1, p-value = 0.29
## alternative hypothesis: two-sided
```

2D Using an appropriate test, verify whether the median precipitation for seeded clouds is less than 300. Next, design and perform a test to check whether the fraction of the seeded clouds with the precipitation less than 30 is at most 25%.

```
sum(seeded<300)
```

```
## [1] 17
```

```
binom.test(17,26,p=0.5) #H0 is therefore not rejected, which means that at the median precipitation for
```

```
##  
## Exact binomial test  
##  
## data: 17 and 26  
## number of successes = 17, number of trials = 26, p-value = 0.17  
## alternative hypothesis: true probability of success is not equal to 0.5  
## 95 percent confidence interval:  
## 0.44333 0.82786  
## sample estimates:  
## probability of success  
## 0.65385
```

```
#there are different way to check this whereas the first method is:  
quantile(seeded)
```

```
##      0%      25%      50%      75%     100%  
##  4.100  98.125 221.600 406.025 2745.600
```

```
#the binomial method could also be used:  
binom.test(17,26,p=0.25, alternative = "l")
```

```
##  
## Exact binomial test  
##  
## data: 17 and 26  
## number of successes = 17, number of trials = 26, p-value = 1  
## alternative hypothesis: true probability of success is less than 0.25  
## 95 percent confidence interval:  
## 0.00000 0.80604  
## sample estimates:  
## probability of success  
## 0.65385
```

```
#just as the proportion test  
prop.test(17,26,p=0.25, alt="l")
```

```
##  
## 1-sample proportions test with continuity correction  
##  
## data: 17 out of 26, null probability 0.25  
## X-squared = 20.5, df = 1, p-value = 1  
## alternative hypothesis: true p is less than 0.25  
## 95 percent confidence interval:  
## 0.0000 0.8017  
## sample estimates:  
## p  
## 0.65385
```


Question 3

Loading the dataset

```
dogs_data = read.table("dogs.txt",header=TRUE)
Isofluorane = dogs_data$isofluorane
Halothane = dogs_data$halothane
Cyclopropane = dogs_data$cyclopropane
```

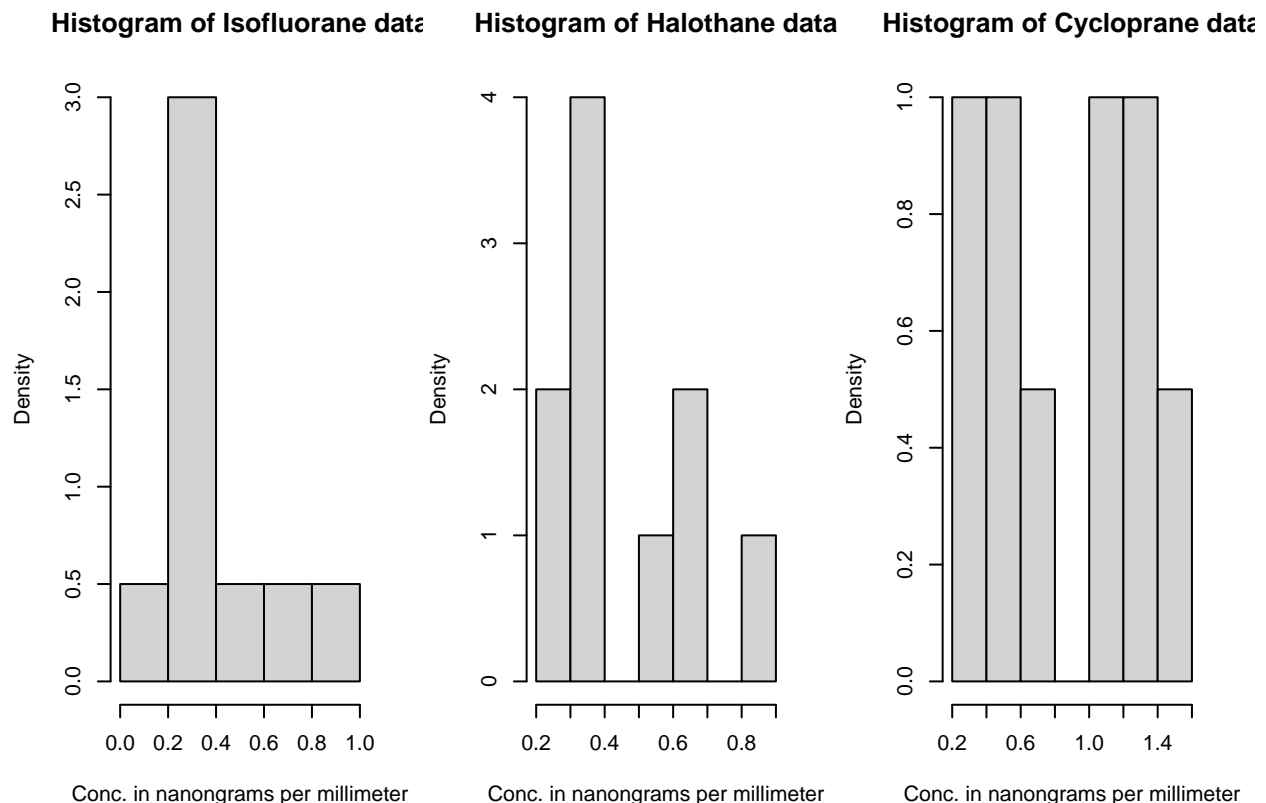
3A To check for normality of data distribution across the 3 variables in each record:

Visually, we plot the histogram and qq-plot for each variable to inspect for normality.

Mathematically, we can go through Shapiro-Wilk normality test to see if the

Histograms

```
par(mfrow=c(1,3))
hist(Isofluorane,main="Histogram of Isofluorane data",probability = TRUE,xlab="Conc. in nanograms per milliliter")
hist(Halothane,main="Histogram of Halothane data", probability = TRUE, xlab="Conc. in nanograms per milliliter")
#par(mfrow=c(1,1))
hist(Cyclopropane,main="Histogram of Cyclopropane data", probability = TRUE, xlab="Conc. in nanograms per milliliter")
```



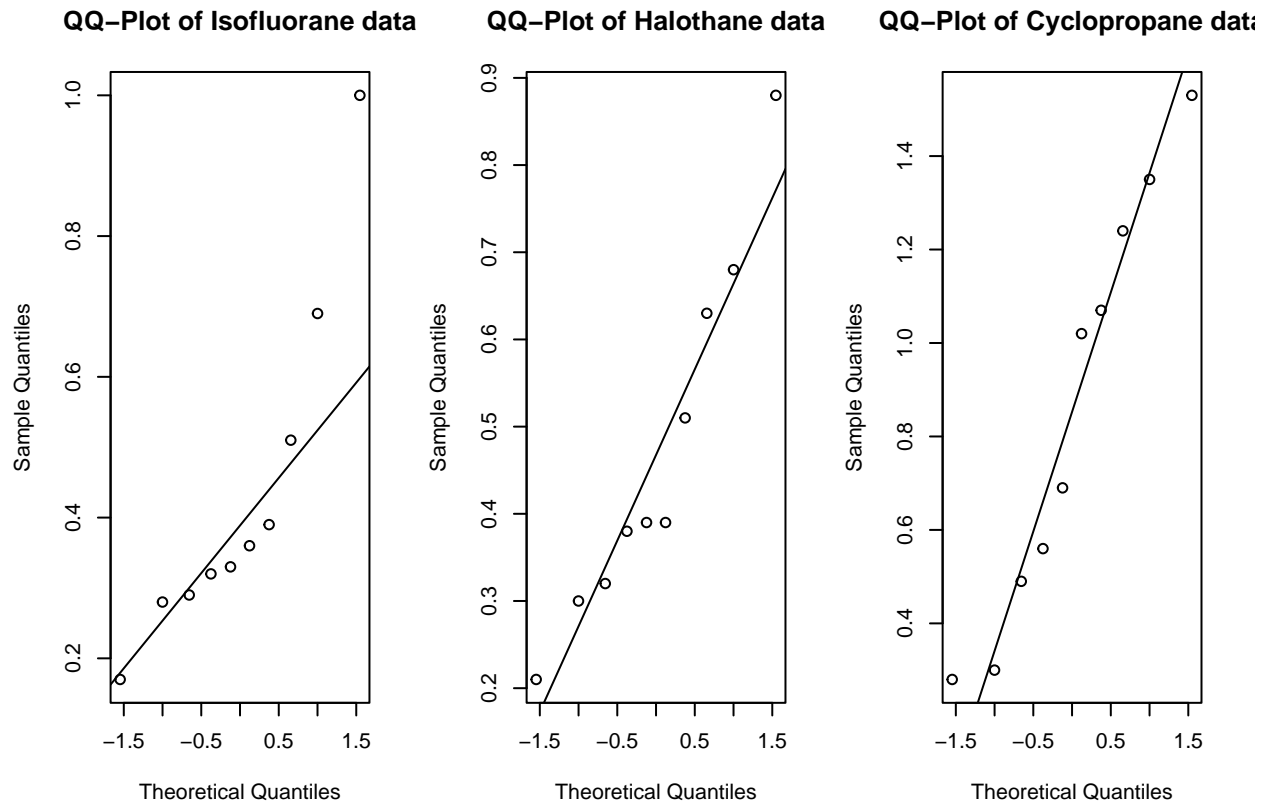
Looking visually at histograms, the data doesn't look to be distributed normally across any variable.

QQ-Plots

```

par(mfrow=c(1,3))
qqnorm(Isofluorane,main="QQ-Plot of Isofluorane data")
qqline(Isofluorane)
qqnorm(Halothane,main="QQ-Plot of Halothane data")
qqline(Halothane)
qqnorm(Cyclopropane,main="QQ-Plot of Cyclopropane data")
qqline(Cyclopropane)

```



QQ-Plots reveal a better picture of the normality and looking at it, Halothane and Cyclopropane data look promising candidates for normality.

To confirm our inference, let's go through Shapiro-Wilk normality test.

Shapiro-Wilk Normality Test

```
print(shapiro.test(Isofluorane))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Isofluorane
## W = 0.831, p-value = 0.034
```

```
print(shapiro.test(Halothane))
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  Halothane
## W = 0.923, p-value = 0.39

print(shapiro.test(Cyclopropane))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Cyclopropane
## W = 0.933, p-value = 0.48
```

Halothane and Cyclopropane data have p-values > 0.05 , which means we can accept the Null Hypothesis H_0 that data is normally distributed. On the other hand, Isofluorane data isn't normally distributed as p-value ≤ 0.05 .

Final Answer

On visual inspection, all 3 variables in dataset didn't look visually normal using histogram inspection, but Halothane and Cyclopropane had somewhat straight qq-plots (variation due to low number of observations).

On closer inspection through Shapiro-Wilk Normality test, Isofluorane proved to be not normally distributed at all but Halothane and Cyclopropane concentrations were indeed normally distributed through the dataset.

3B To check correlation between 2 normally distributed variables, we can apply Pearson's test for correlation. From 3-A, we realise that variable Isofluorane is not normally distributed. Therefore, we cannot use this test to check for its correlation with Halothane.

In this case, we will use Spearman's rank test for correlation.

```
print(cor.test(Isofluorane,Halothane,method=c("spearman")))

## Warning in cor.test.default(Isofluorane, Halothane, method = c("spearman")):
## Cannot compute exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  Isofluorane and Halothane
## S = 129, p-value = 0.54
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.21885
```

Since obtained $\rho \neq 0$, there might be some degree of positive correlation between both variables in our data. But p-value of 0.54 indicates that Null-Hypothesis is significantly supported, hence true $\rho = 0$ and it indicates no correlation between values of Isofluorane and Halothane. We need to note that p-value might not be correct due to presence of ties (same-values) in a fairly small sample size (10).

In 3-A, using Shapiro-Wilk Normality Test, it has already been concluded that variable Isofluorane doesn't come from a normal distribution while variable Halothane comes from a normal distribution.

```
print(ks.test(Isofluorane,Halothane))
```

```
## Warning in ks.test(Isofluorane, Halothane): cannot compute exact p-value with
## ties
```

```
##
## Two-sample Kolmogorov-Smirnov test
##
## data: Isofluorane and Halothane
## D = 0.3, p-value = 0.76
## alternative hypothesis: two-sided
```

Using Kolmogorov-Smirnov Tests, we get a P-value of 0.76 indicating that 2 variable distributions might be similar. Again, due to low sample size and presence of ties, the calculated value of p-value can have some inaccuracies.

```
print(paste("Length of datasamples : ",length(Isofluorane),length(Halothane),length(Cyclopropane)))
```

```
## [1] "Length of datasamples : 10 10 10"
```

Since size of data sample is small ($10 < 30$) and from 3-A we know 1st column does not follow normal distribution, we can say that Permutation tests are applicable in this case.

3C Performing ANOVA test,

```
#Creating DataFrame
```

```
df = data.frame(Concentration=as.vector(as.matrix(dogs_data)),An_Type=factor(rep(1:3,each=10)))
```

```
#Creating Linear Model
```

```
mod_1 = lm(Concentration ~ An_Type, data=df)
```

```
#Performing ANOVA Test
```

```
anova(mod_1)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: Concentration
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## An_Type    2   1.08   0.540    5.35 0.011 *
## Residuals 27   2.73   0.101
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Printing output parameters of our Linear Model
```

```
summary(mod_1)
```

```
##
```

```
## Call:
```

```
## lm(formula = Concentration ~ An_Type, data = df)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.573 -0.161 -0.079  0.200  0.677
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.434      0.101    4.32 0.00019 ***
## An_Type2       0.035      0.142    0.25 0.80727
## An_Type3       0.419      0.142    2.95 0.00650 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.318 on 27 degrees of freedom
## Multiple R-squared:  0.284, Adjusted R-squared:  0.231
## F-statistic: 5.35 on 2 and 27 DF, p-value: 0.011
```

Looking at the p-value from the summary of the linear model and ANOVA test table, $P\text{-value} = 0.011 < 0.05$, we can definitely say that One Way ANOVA test determines that the type of drug has a significant effect on the concentration of plasma epinephrine.

```
confint(mod_1)
```

```
##              2.5 %  97.5 %
## (Intercept) 0.22788 0.64012
## An_Type2    -0.25650 0.32650
## An_Type3     0.12750 0.71050
```

In the confidence level of 95%,

Estimated concentration of each type of drug is :

Isofluorane : 22.79 % to 64.01%

Halothane : 0.0% to 32.65 % (ignoring the negative Concentration value as it is not semantically possible)

Cyclopropane : 12.75% to 71.05%

3D Performing the Kruskal-Wallis test :

```
Concentration = df$Concentration
Anesthesia_Type = df$An_Type
kruskal.test(Concentration ~ Anesthesia_Type, data = df)
```

```
##
## Kruskal-Wallis rank sum test
##
## data:  Concentration by Anesthesia_Type
## Kruskal-Wallis chi-squared = 5.64, df = 2, p-value = 0.059
```

The p-value given using Kruskal-Wallis Test comes out to be 0.059 (> 0.05). Hence the Null Hypothesis of this test is accepted, meaning it considers the effect of type of drug insignificant on the concentration of plasma epinephrine in the blood.

This is in contrast with the inference deduced using ANOVA test, although not by a far off margin since the p-value is very close 0.05 in case of Kruskal-Wallis test as well. This might be due to the nature of distribution as Kruskal-Wallis is a rank-based test, while ANOVA uses the nature of distribution such as Mean to compute the effect of the independent variable on the dependent variable. In our case, the first variable (Isoflurane) doesn't come from a standard normal distribution, while the other 2 variables do. Hence, Kruskal-Wallis might be a better representation of the effect of drugs on the plasma epinephrine concentration in our case since it is a non-parametric test.

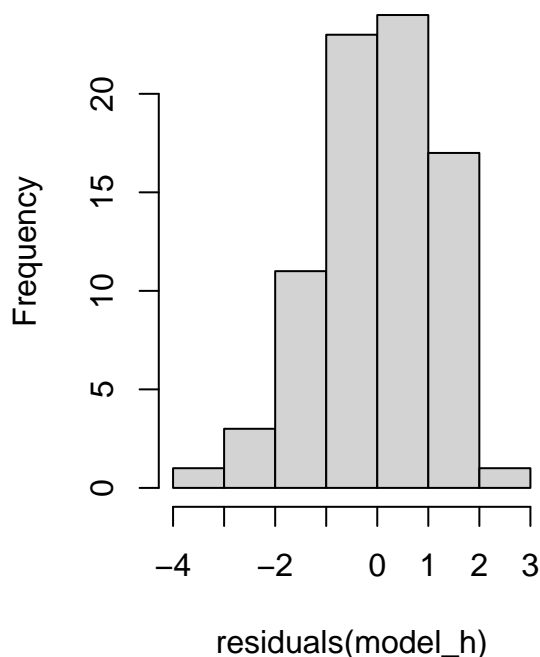
Question 4

Hemoglobin in trout

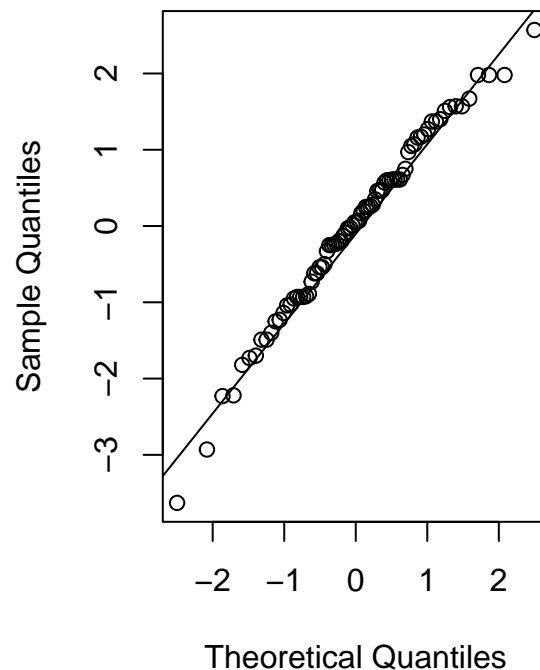
4A

4B Check for normality before using ANOVA-test. The data comes from normal distributions, since both the histogram and QQ-plot show a normally distributed graph. Besides, the shapiro-wilk test gives a p-value of 0.526 which is above 0.05 and therefore suggests that the data comes from normal distribution.

Histogram of residuals of mode



QQ-Plot of residuals of model



From the two-way ANOVA test, we can see that the treatment variable 'rate' has a significant effect on hemoglobin, since the p-value is below 0.05. However, the block variable 'method' does not have a significant effect on hemoglobin since its p-value is 0.216, which is above 0.05. The p-value for rate:method is 0.377, which means that there is no evidence for interaction.

4C The factor 'rate' has the greatest influence on hemoglobin, since rate has a significant effect on hemoglobin and method does not have a significant effect on hemoglobin. However, since there is no interaction between rate and method, we have to remove the interaction term from the model and create an

additive model where there is no interaction. Still, only the variable rate has a main effect on hemoglobin, and not the variable 'method'.

The combination of rate 2 and method B leads to the highest mean of hemoglobin

Estimate the mean hemoglobin value for rate 3 by using method A.

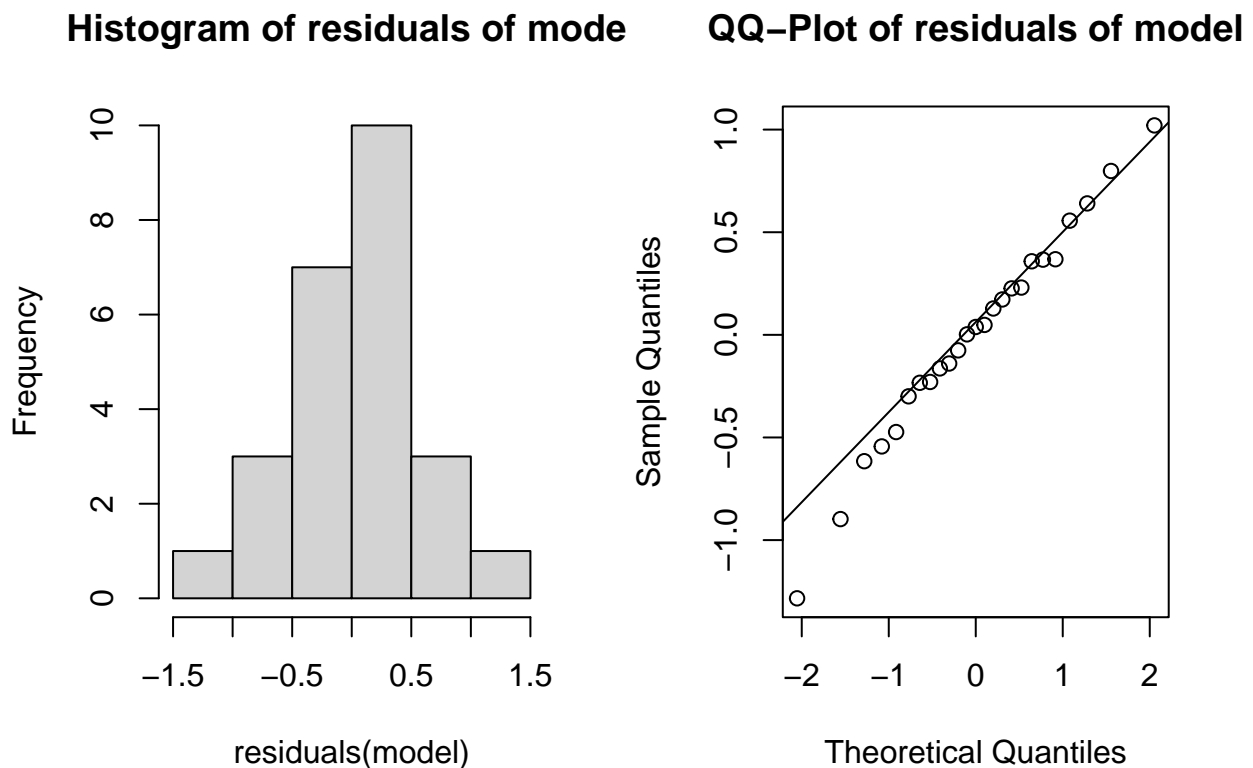
The mean of rate 3 and method A is 9.03

Rate 2 leads to the highest mean of hemoglobin.

4D The results from the one-way ANOVA are exactly equal to the results of the two-way ANOVA. In both tests the effect of each rate on hemoglobin is tested and therefore, it is unnecessary to do the one-way ANOVA.

Question 5

5A To perform a three-way ANOVA test, we need to check normality. This is done by creating a QQ-plot and histogram of the residuals of the data. Additionally, the Shapiro-Wilk test was performed. The line in the QQ-plot looks linear and the distribution of the histogram can come from a normal distribution. Besides, the Shapiro-Wilk test had a p-value larger than 0.05 which means that the data comes from normal distribution. Therefore, a three-way ANOVA can be performed.



Starter 1 is the intercept, which has a p-value of less than 0.05. Therefore there is a significant effect of starter 1 on acidity. While starter 2 has a p-value of 0.754, and therefore has no significant effect on acidity.

5B Insignificant block variables, are variables that have a p-value larger than 0.05. So this is the block variable position, which has a p-value of 0.411

After deleting the block variable position, starter 4 leads to a significantly different acidity.

5C In the Friedman test the data does not need to come from a normal distribution, but it can also be used when the data comes from a normal distribution, like in this case. Instead of using the mean of the groups in an ANOVA test the Friedman makes use of ranks. However, this difference does not make any difference in the use of this test. Therefore, the Friedman test can also be used in this application to test whether there is an effect of starter on acidity. However, the Friedman test is not necessary when the ANOVA test has already been performed.

5D Formula mixed effects is $\text{acidity} \sim \text{starter} + (1|\text{batch})$. Where the $1|\text{batch}$ is used to make from the block variable 'batch' a random effect variable.

When using the fixed effects model the variables 1 and starter 4 had a significant effect on acidity. In the mixed effects model variables starter 1, starter 3, starter 4 had a significant effect on acidity. So, with the mixed effect model, more starters have a significant effect on acidity.