Experimental Design and Data Analysis: Assignment 2

Andrew Bedard(2566978) & Simone van Gompel(2567525)
Group 19

April 29, 2015

1 Exercise 1

1

Using data from *peruvians.txt* we use the pairs command in R to produce a scatter plot of every variable against every other. The result can be seen in Fig:1. From this figure, the variables age and weight seem to have a linear relationship, which would suggest that they are correlated.

2

Using the Spearman rank correlation test for all variables against migration we obtain the following results:

The rank correlation between:

- age and migration is 0.47605753, this is visible in Fig1 in the form of an approximate linear relationship.
- weight and migration is 0.35069559, this is visible in Fig1 as an approximate linear relationship.
- Length and migration is 0.08458432, as can be seen in Fig 1, there is no discernible relationships between the two variables.
- wrist and migration is 0.21934983, this result is surprisingly high considering Fig1, as any relationship between these two variables is very difficult to see.

Table 1: Spearman Rank Correlation test between each variable and *migration* data

	age	weight	length	wrist	systolic	diastolic
migration	0.47605753	0.35069559	0.08458432	0.21934983	-0.16842856	0.07514098

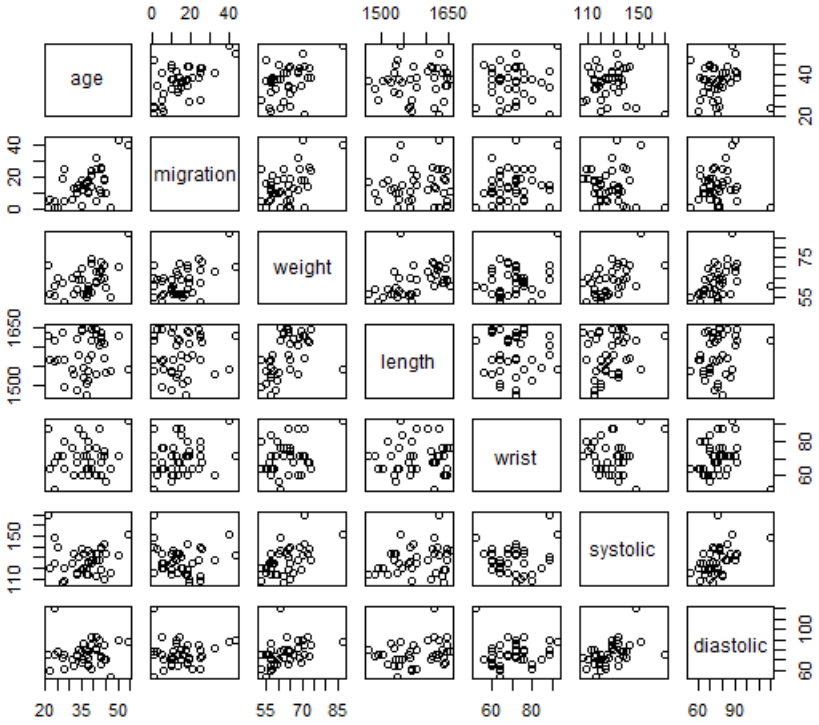

Figure 1: Pairwise scatterplots of the dataset peruvians

- systolic and migration is -0.16842856, which is obvious given the very small downward sloping structure seen in Fig1.
- diastolic and migration is 0.07514098, which is unsurprising considering the lack of structure seen in Fig1

2 Exercise 2

1

From the data in *clouds.txt*, we can perform various tests to determine whether the two independent samples: values of seeded clouds, and un-seeded clouds are equal.

- The two sample t-test produces a 95% confidence interval of [-4.7405,559.5859] with mean of x = 441.9846 mean of y = 164.5619 and a p-value of 0.05375, which suggests that we accept the null hypothesis, and thus our samples

are the same, this is misleading however. Consider Fig2, and Fig 5; The two sample t-test assumes that both our samples come from a normal population, Fig 5 represents a randomly generated normally distributed sample of equal size to both seeded and un-seeded cloud data, clearly Fig2 is very different, thus this assumption that our samples are from a normal population is not reasonable, and the two sample t-test is not suitable for this data as is.

- The Mann-Whitney test gives: W=473 with a p-value of 0.01383, suggesting that we reject the null hypothesis, and thus, our samples are not equal. The Mann-Whitney test makes no assumptions about the nature of the data but combine our samples to measure their differences, so it is suitable in this case.

- The Kolmogorov-Smirnov test gives: D= 0.4231 with a p-value of 0.01905, suggesting that we reject the null hypothesis and that our samples are not equal. The Kolmogorov-Smirnov test measures the differences in the histograms between the two sets of data, and with no constraints on the requirements for our data, this test is suitable.

Two sample t-test assumes that both samples come from a normal population. t-test: p-value =0.05375. The 95% confidence interval is [-4.7405,559.5859] with mean of x = 441.9846 mean of y = 164.5619. We accept the null hypothesis The Mann-Whitney test: W=473, p-value = 0.01383 we reject the null hypothesis Kolmogorov-Smirnov test: D = 0.4231, p-value = 0.01905 we reject the null hypothesis

2

We repeat the previous procedure after taking the square root of each data point from both seeded and un-seeded cloud samples.

- The two sample t-test produces a p-value of 0.01956, which suggests that we reject the null hypothesis and that our two samples are not equal. In this case the 95% confidence interval is [1.2021,13.0713] with mean x=17.068 and mean y= 9.9313. The transformed data, seen in Fig3 is still too far from normal for the t-test to be suitable in this case.

- The Mann-Whitney test gives: W=473 with a p-value of 0.01383, suggesting that we reject the null hypothesis, and thus, our samples are not equal. This is the same result as the original data, which is unsurprising considering the test measures the ranks of a combined sample, once again this test is suitable.

- The Kolmogorov-Smirnov test gives: D= 0.4231 with a p-value of 0.01905, suggesting that we reject the null hypothesis and that our samples are not equal. Again, this is the same result obtained with the original data, and as this test only measures the difference between histograms, it is again suitable.

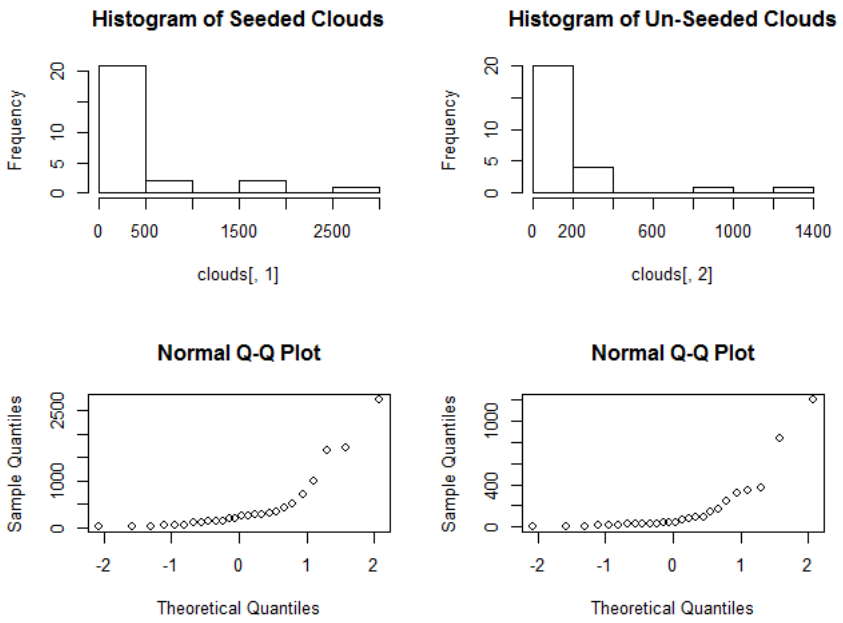

Figure 2: Histograms and QQ-Plots of Seeded Clouds and Un-Seeded Clouds respectively

t-test: t=2.4246, p-value = 0.01956, 95% confidence interval [1.2021,13.0713], mean x=17.068, mean y= 9.9313 Mann-Whitney test: The Mann-Whitney test: W=473, p-value = 0.01383 Kolmogorov-Smirnov test: D = 0.4231, p-value = 0.01905

3

We repeat the same procedure again, now obtaining the square root of the square root of the original data point in our samples seeded and un-seeded clouds.

- The two sample t-test

t-test: t=2.5968, p-value=0.0124, 95% [0.2196,1.7236], mean x = 3.8789, mean y = 2.9073 Mann-Whitney test: The Mann-Whitney test: W=473, p-value = 0.01383 Kolmogorov-Smirnov test: D = 0.4231, p-value = 0.01905

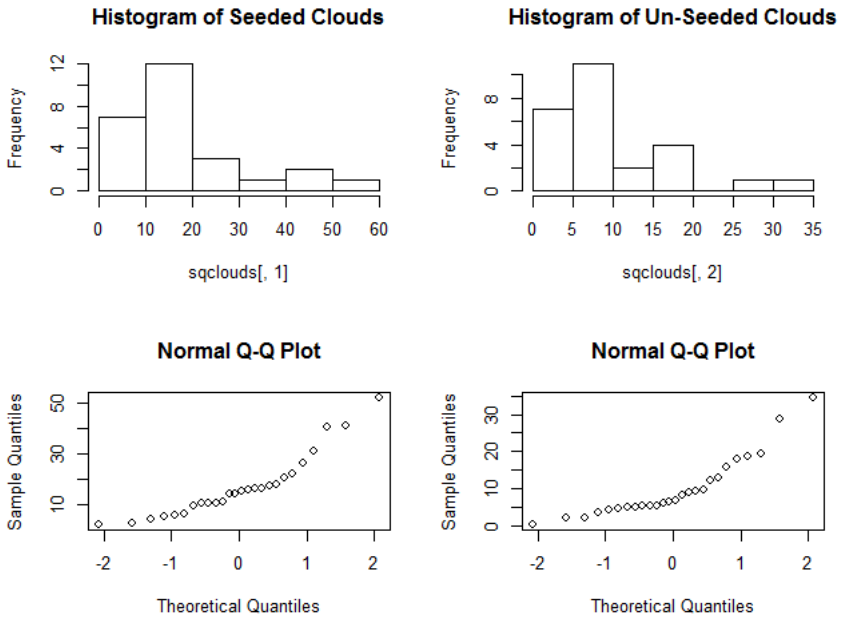

Figure 3: Histograms and QQ-Plots of Seeded Clouds and Un-Seeded Clouds after taking the square root of each data point

3 3

1

2

No

3

Yes

4

Using sq(genal) we find a p-value of 2.2e-16, therefore we reject the null hypothesis

5

Based on summary(sqaov) X=0.03 is the best, because 0.9033-0.2227 is the smallest value, and we are estimating the smallest value obtained from an opti-

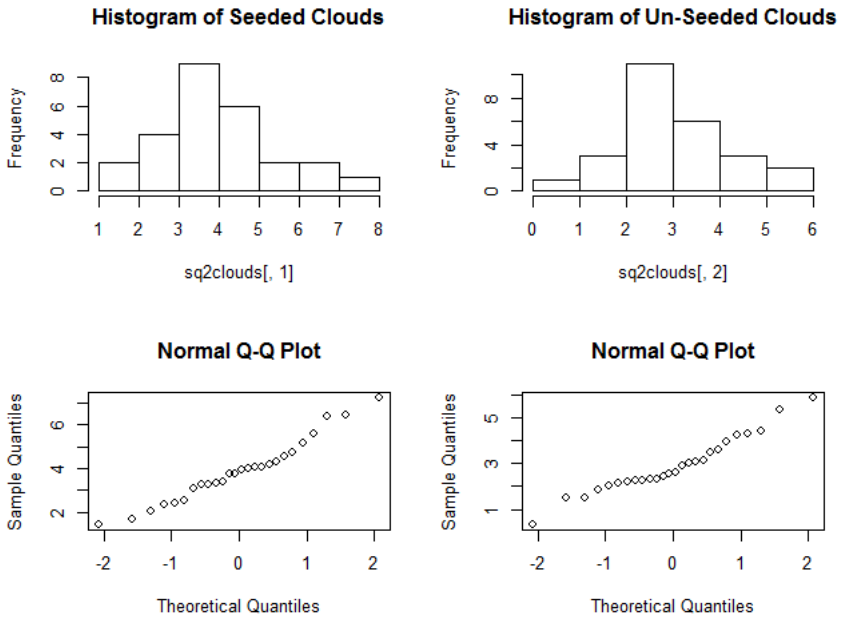

Figure 4: Histograms and QQ-Plots of Seeded Clouds and Un-Seeded Clouds after taking the square root of each data point twice

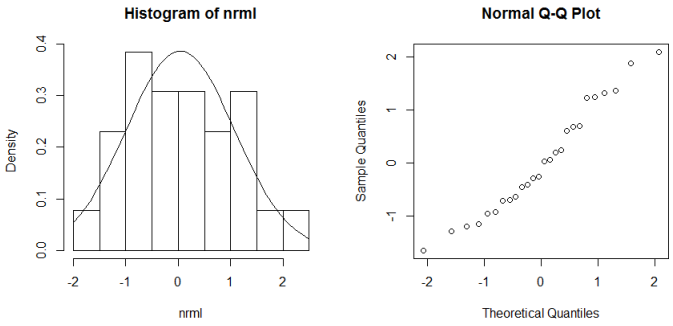

Figure 5: Histogram and QQ-Plot of randomly generated normally distributed data of same size as both Seeded and Un-Seeded Clouds

mization where smaller is better.

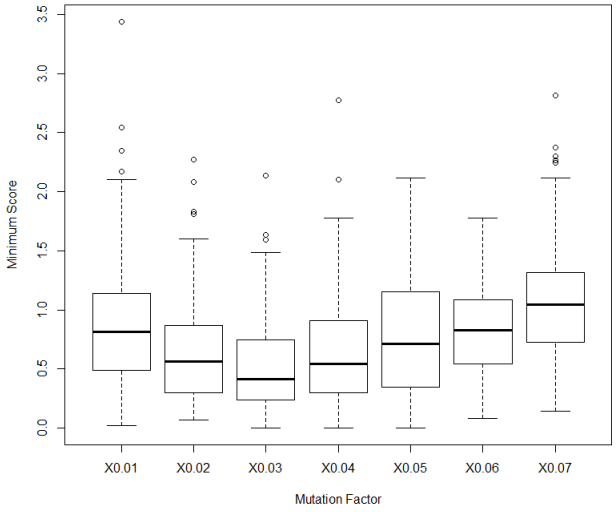

Figure 6: Box Plot of mutation factor and the resulting minimum result obtained

6

4 4

1

2

No, the sample sizes are far too small to reasonably judge whether or not they are normal populations.

3

Reject the null, p-value 0.011, estimated value of Isofluorane: 0.4340, of Halothane: 0.469 and of Cyclopropane: 0.853

4

The Kruskal-Wallis test for the same hypothesis gives a p-value of 0.05948, thus we do not reject the null hypothesis.

5

The differences between the 1-way Anova and the Kruskal-Wallis test could be explained by our initial assumption that the 3 samples represented in the data

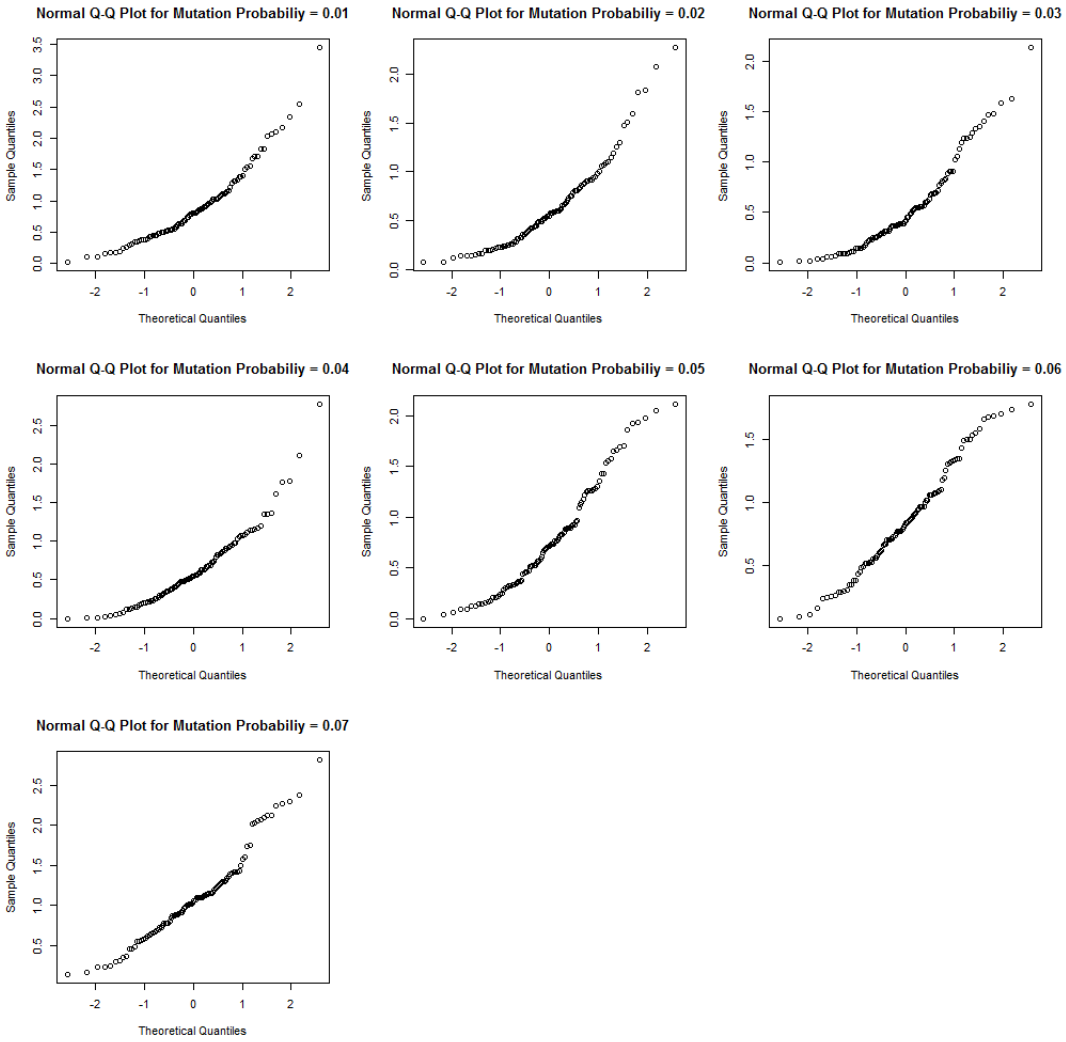

Figure 7: QQ-Plots for each mutation factor

do not in fact come from a normal population, because 1-way Anova assumes that we are sampling from normal populations this could produce conflicting results, though as we see in Figure 13 the residuals do not deviate from the normal significantly. It is perhaps more compelling that the differences between the 1-way Anova and the Kruskal-Wallis are due to the fact that 1-way Anova assumes our populations have equal population variances, and as is observed in Figure 11, they are obviously not equal.

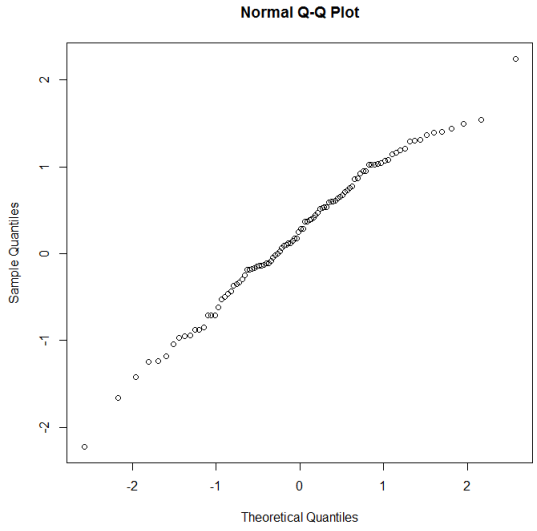

Figure 8: QQ-Plot of randomly generated normally distributed population of the same size as those in *genal.txt*

5 R-Code

5.1 Exercise 1

```
        #1.1
peru = read.table("peruvians.txt", header=T)
peru = peru[,-c(5,6,7)]
pairs(peru)

#1.2
#This shows the spearman rank correlation against all variables
correl_s = cor(peru, method="spearman")

correl_s[2, ]
```

5.2 Exercise 2

```
        #2.1
clouds = read.table("clouds.txt", header=T)
len = length(clouds[,1])
nrml = rnorm(len)

#Histograms, boxplots, qqplots to see if data is normal
par(mfrow=c(2,2))
```

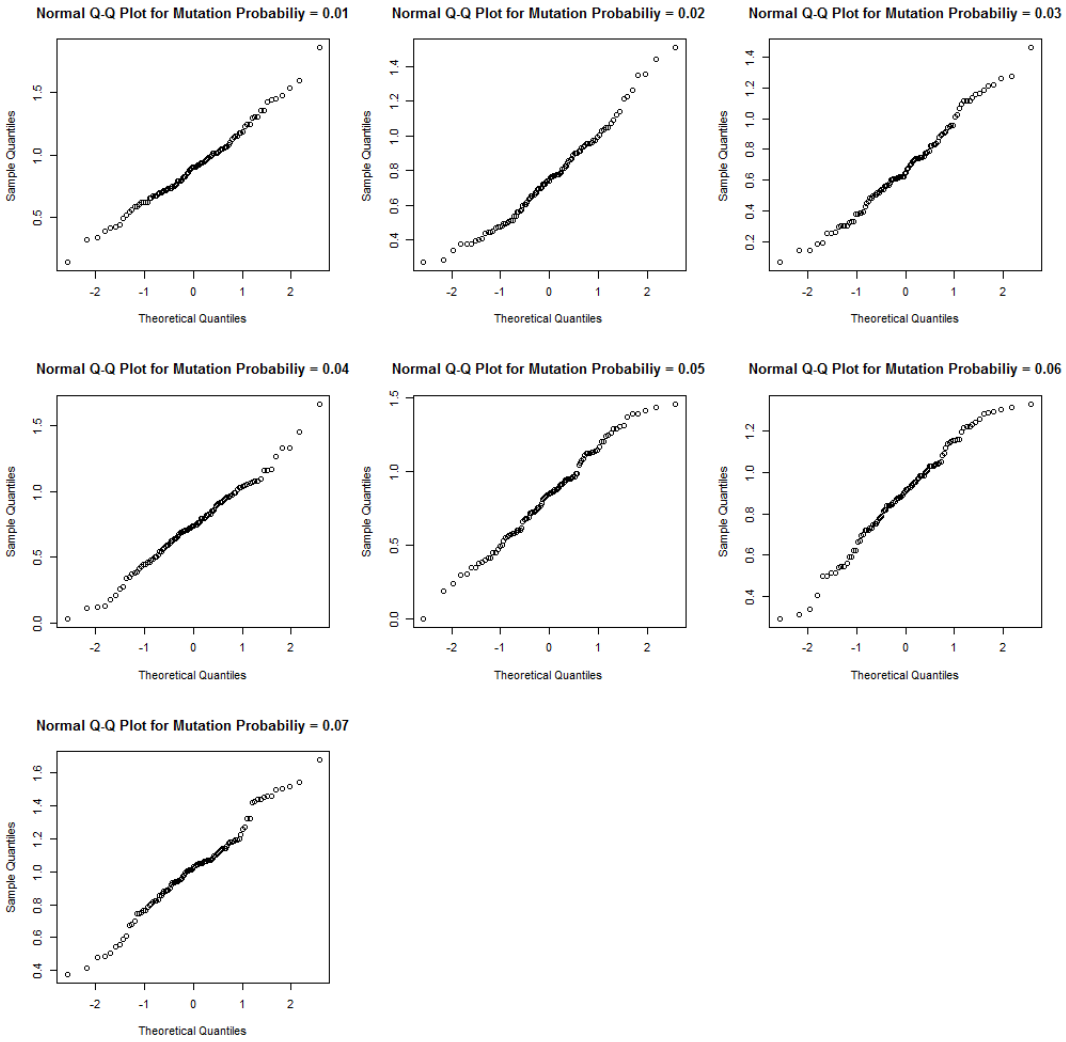

Figure 9: QQ-Plot for each mutation factor after take the square root of each data point

```
hist(clouds[,1])
hist(clouds[,2])
qqnorm(clouds[,1])
qqnorm(clouds[,2])
#These graphs should show the data is not normal

#Two sample t-test
t.test(clouds[  ,1],clouds[  ,2])
```

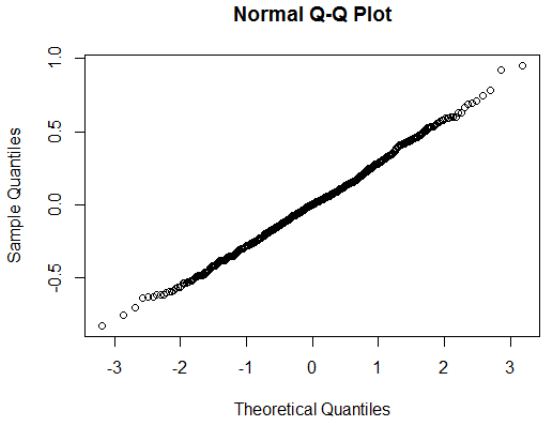

Figure 10: QQ-Plot of the residuals of 1-way Anova test

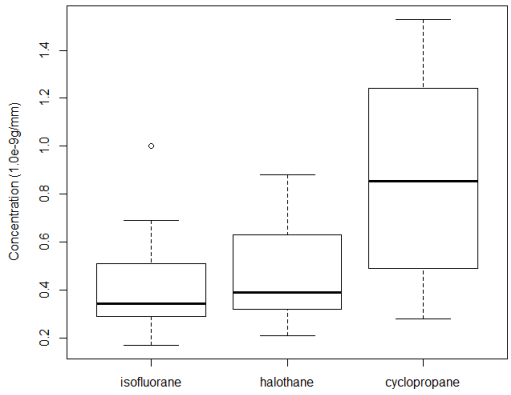

Figure 11: Box plot of plasma epinephrine concentrations under various anaesthesias

```
#Mann-Whitney test
wilcox.test(clouds[ ,1],clouds[ ,2])
#Kolmogorov-Smirnov test
ks.test(clouds[,1],clouds[,2])

#2.2
sqclouds = sqrt(clouds)
```

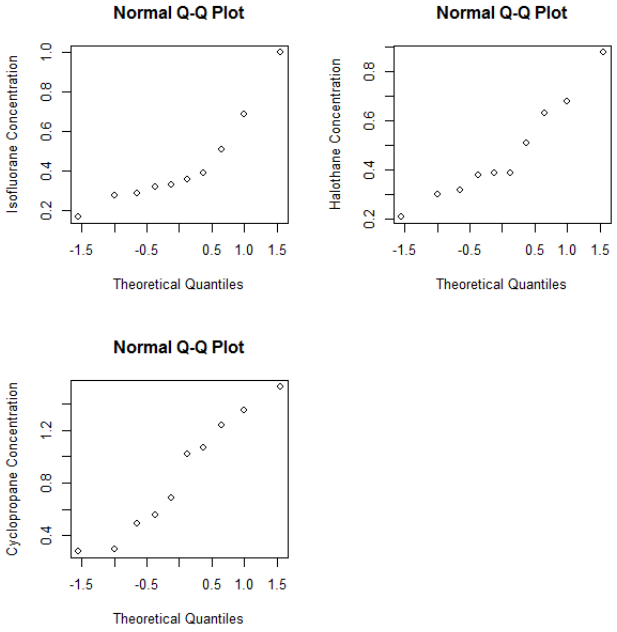

Figure 12: QQ-Plots for data from each type of anaesthesia

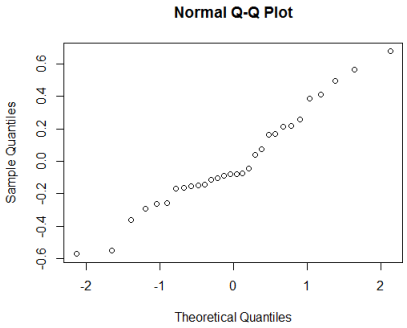

Figure 13: QQ-Plot of residuals of 1-way Anova analysis

```
#Histograms, boxplots, qqplots to see if data is normal
par(mfrow=c(2,2))
hist(sqclouds[,1])
hist(sqclouds[,2])
qqnorm(sqclouds[,1])
qqnorm(sqclouds[,2])
#These graphs should show the data is not normal
```

```
#Two sample t-test
t.test(sqclouds[ ,1],sqclouds[ ,2])
#Mann-Whitney test
wilcox.test(sqclouds[ ,1],sqclouds[ ,2])
#Kolmogorov-Smirnov test
ks.test(sqclouds[,1],sqclouds[,2])

#2.3
sq2clouds = sqrt(sqclouds)

#Histograms, boxplots, qqplots to see if data is normal
par(mfrow=c(2,2))
hist(sq2clouds[,1])
hist(sq2clouds[,2])
qqnorm(sq2clouds[,1])
qqnorm(sq2clouds[,2])
#These graphs should show the data is not normal

#Two sample t-test
t.test(sq2clouds[ ,1],sq2clouds[ ,2])
#Mann-Whitney test
wilcox.test(sq2clouds[ ,1],sq2clouds[ ,2])
#Kolmogorov-Smirnov test
ks.test(sq2clouds[,1],sq2clouds[,2])

par(mfrow=c(1,2))
hist(nrml, prob=TRUE)
curve(dnorm(x,mean=mean(nrml),sd=sd(nrml)), add=TRUE)
qqnorm(nrml)
```

5.3 Exercise 3

```
#3.1
genal = read.table("genal.txt", header=T)
len=length(genal[,1])
nrml=rnorm(len)

par(mfrow=c(1,1));boxplot(genal)

#3.2
#Loops through, creating QQ-plot for each mutation probability
par(mfrow=c(3,3))
for (i in 1:7) {
  qqnorm(genal[,i],
         main =paste("Normal Q-Q Plot for Mutation Probabiliy =",
```

```
                        i/100))
}
#qqplot of normal random sample of same size as genal[,i]
par(mfrow=c(1,1));qqnorm(nrml)
#3.3
sqgenal=sqrt(genal)

#Loops through, creating QQ-plot for each mutation probability
par(mfrow=c(3,3))
for (i in 1:7) {
  qqnorm(sqgenal[,i],
         main =paste("Normal Q-Q Plot for Mutation Probabiliy =",
                     i/10))
}

#3.4

#Create data-frames for the origional genal data,
#and the squareroot genal data
sqframe=data.frame(yield=as.vector(as.matrix(sqgenal)),
                           variety=factor(rep(1:7,each=100)))

sqaov=lm(yield~variety,data=sqframe)
anova(sqaov)

#3.5
summary(sqaov)

#3.6
par(mfrow=c(1,1));qqnorm(residuals(sqaov))
```

5.4 Exercise 4

```
      #4.1

dogs = read.table("dogs.txt", header=T)
len=length(dogs[,1])
boxplot(dogs)

#4.2
par(mfrow=c(2,2))
qqnorm(dogs[,1],ylab="Isofluorane Concentration")
qqnorm(dogs[,2],ylab="Halothane Concentration")
qqnorm(dogs[,3],ylab="Cyclopropane Concentration")

#4.3
```

```
dogframe = data.frame(yield=as.vector(as.matrix(dogs)),
                      variety=factor(rep(1:3,each=10)))
dogaov=lm(yield~variety,data=dogframe)
anova(dogaov)
summary(dogaov)

#Calculate expected value
u1=0.4340;
u2=0.0350+u1
u3=0.4190+u1

print(u1,u2,u3)
#4.4

attach(dogframe)
kruskal.test(yield,variety)
par(mfrow=c(1,1));qqnorm(dogaov$residuals)

#Calculate and print population variances
for (i in 1:3){
  print(sum((dogs[,i]-mean(dogs[,i]))^2)/(len-1))
}
```