

Experimental Design and Data Analysis:

Assignment 2

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

April 29, 2015

1 Exercise 1

1

Look at Fig:1 for the pairwise graph.... The variables that appear from inspection to correlate with migration are age and weight.

2

Using the Pearson (rank?) correlation test for all variables against migration we obtain the following results:

	age	migration	weight	length	wrist	systolic	diastolic
migration	0.47605753	1	0.35069559	0.08458432	0.21934983	-0.16842856	0.07514098

Age was the most obviously correlated variable between migration when the pairs scatter plots (Fig: 1), and age is the most correlated variable with migration. Weight is the second most correlated variable with migration, ...

2 Exercise 2

1

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

Two sample t-test

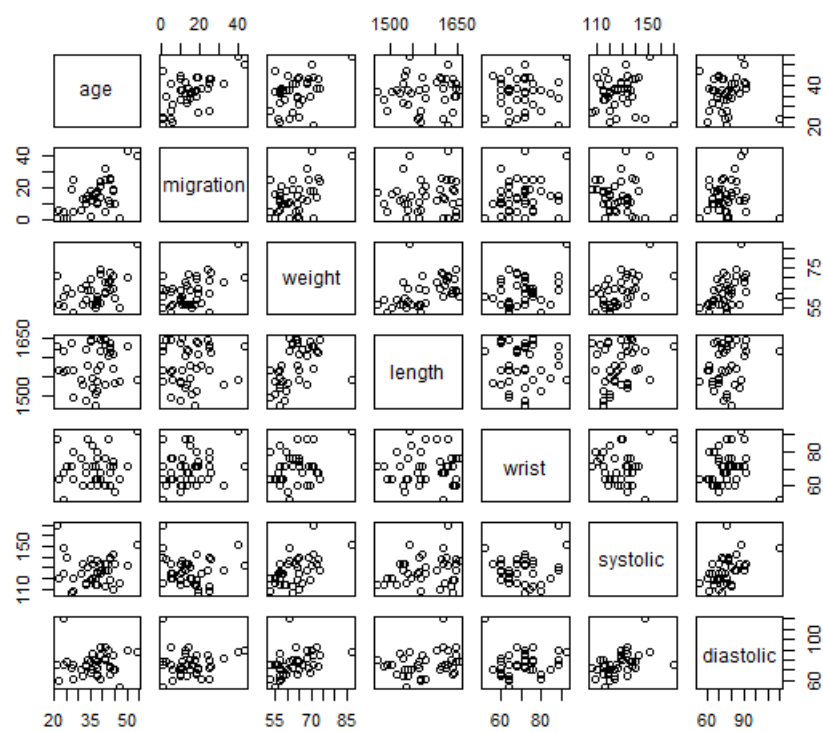


Figure 1: Pairwise scatterplots of the dataset peruvians

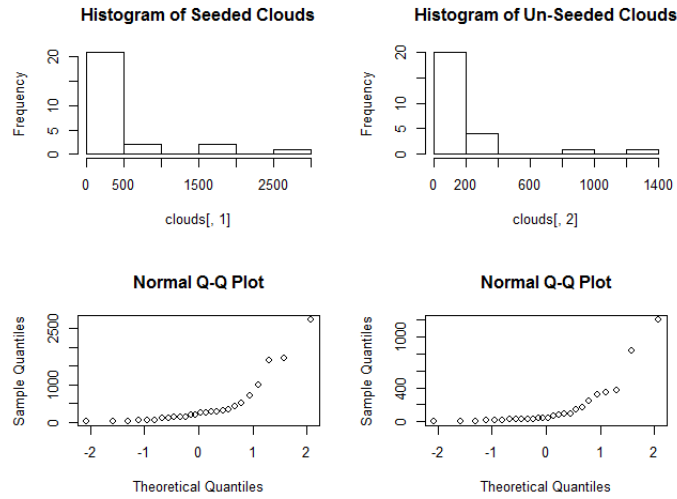


Figure 2: clouds

2

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

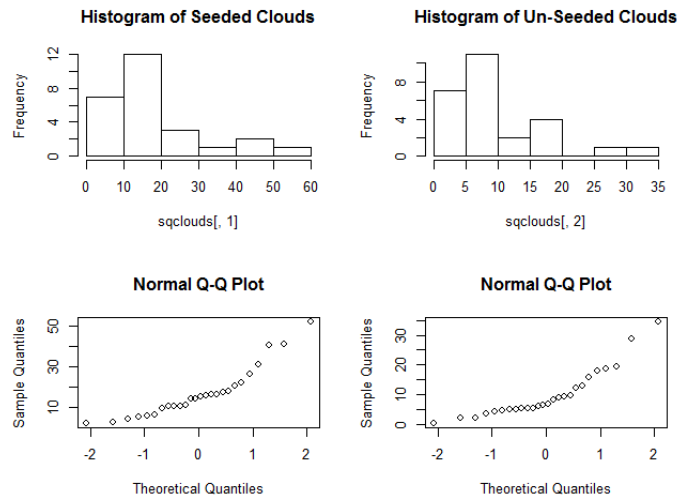


Figure 3: sq(clouds)

3

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

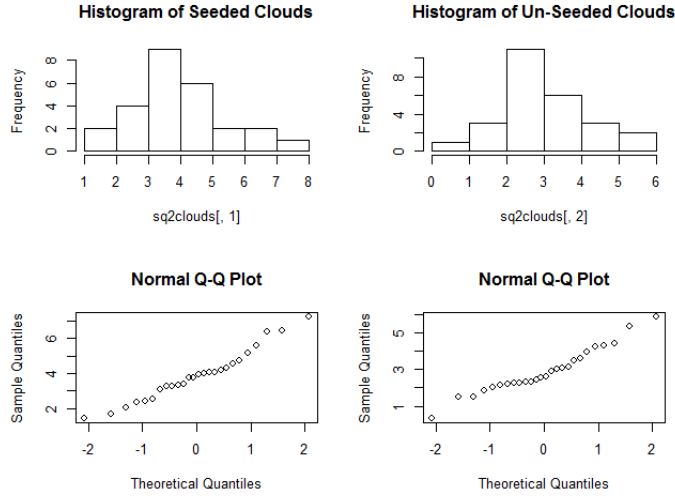


Figure 4: $\text{sq}(\text{sq}(\text{clouds}))$

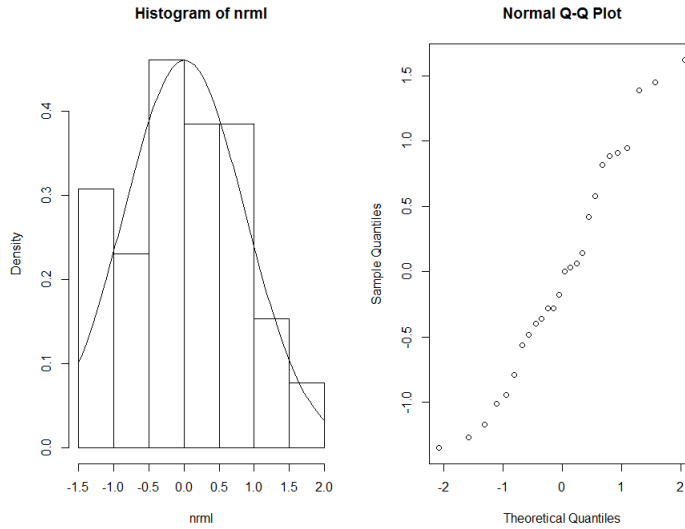


Figure 5: normal

3 3

1

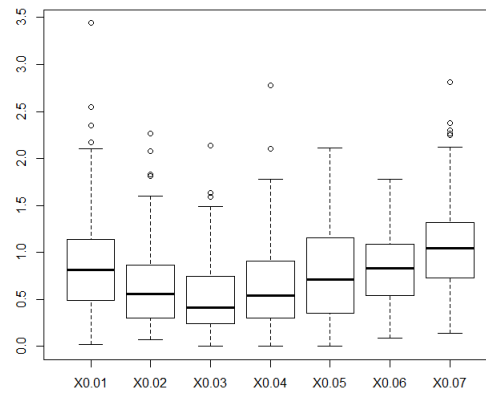


Figure 6: box

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 optimization where smaller is better.

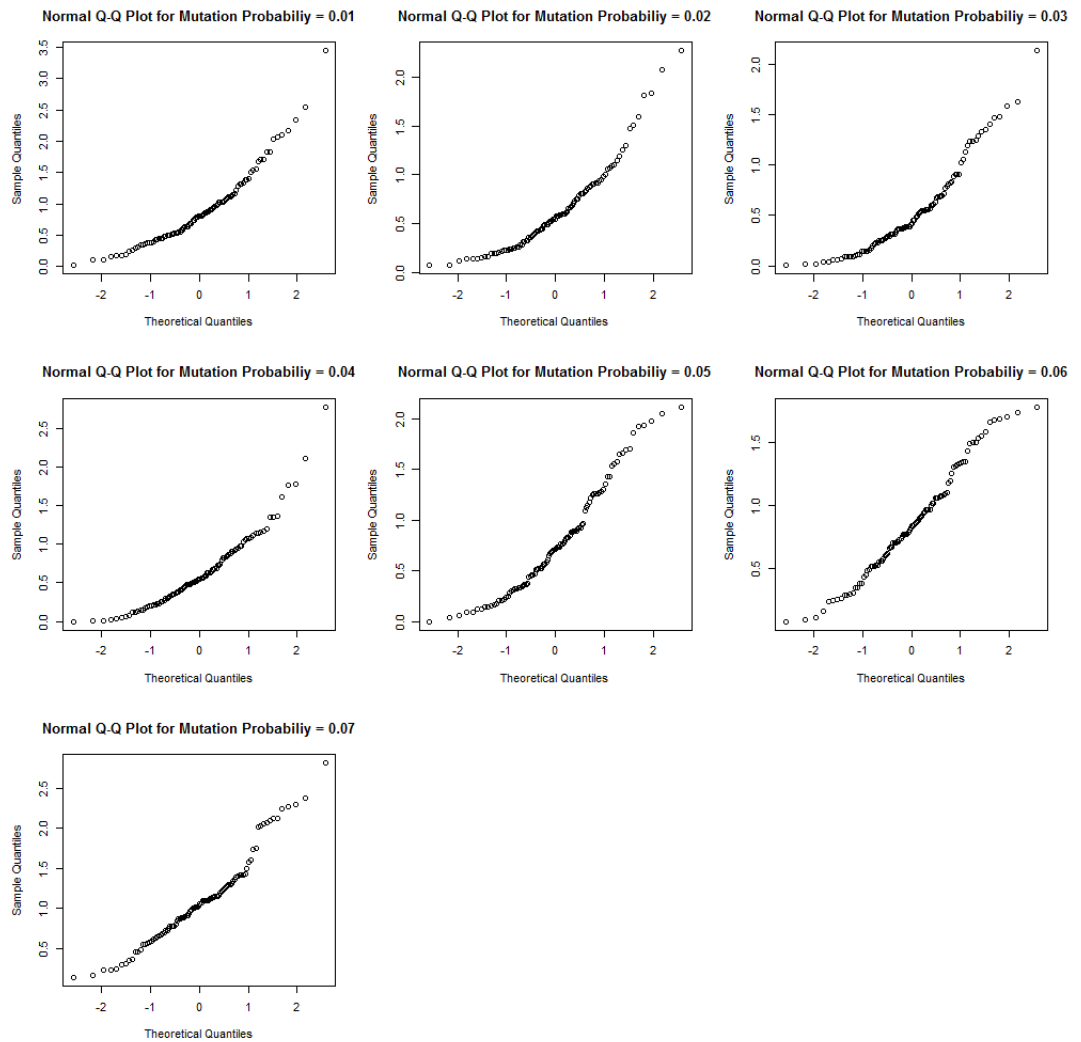


Figure 7: qq-genal

6

4 4

1

2

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

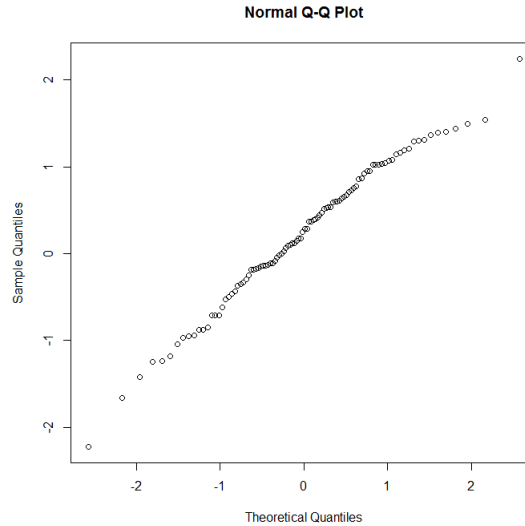


Figure 8: qq-nrml

3

Reject the null, p-value 0.011, estimated value of Isoflurane: 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 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.

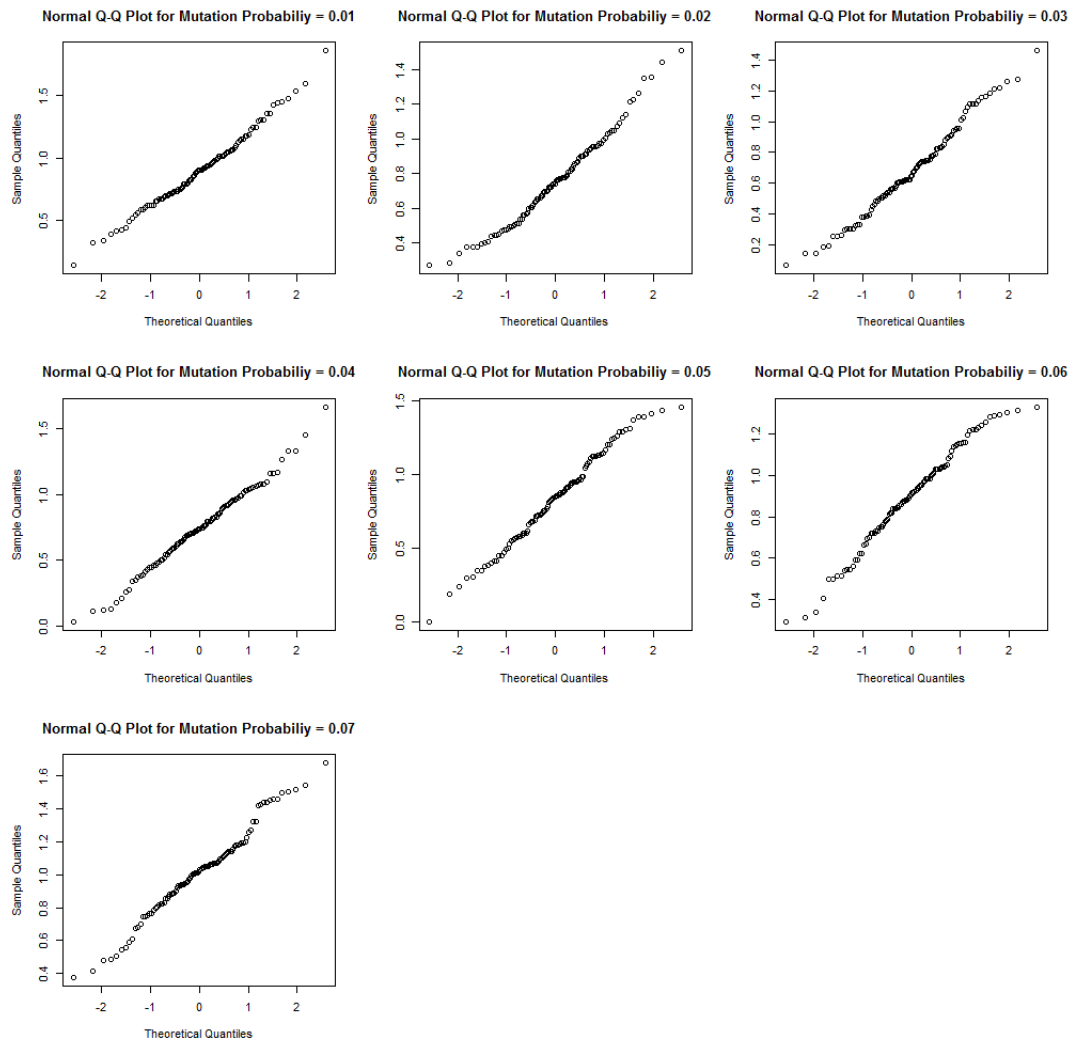


Figure 9: qq-sq(genal)

5 R-Code

5.1 Exercise 1

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

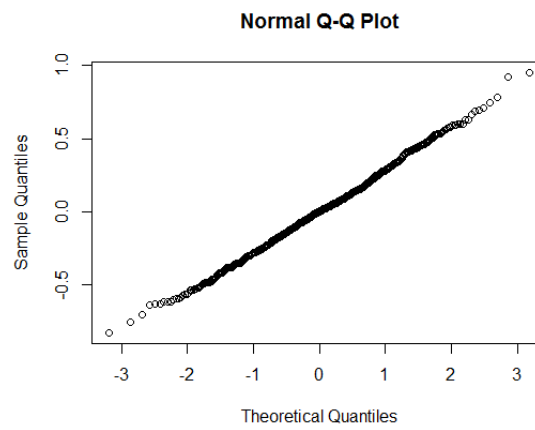



Figure 10: qq-resid

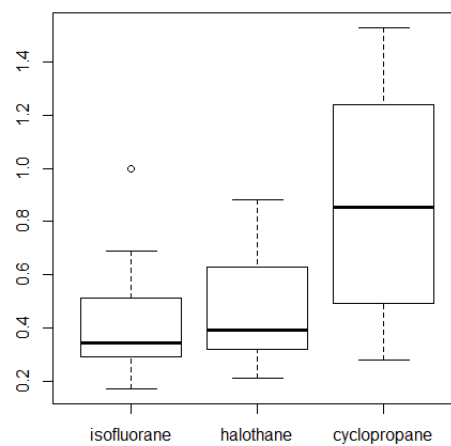


Figure 11: dog box

#1.2

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

`correl_s[2,]`

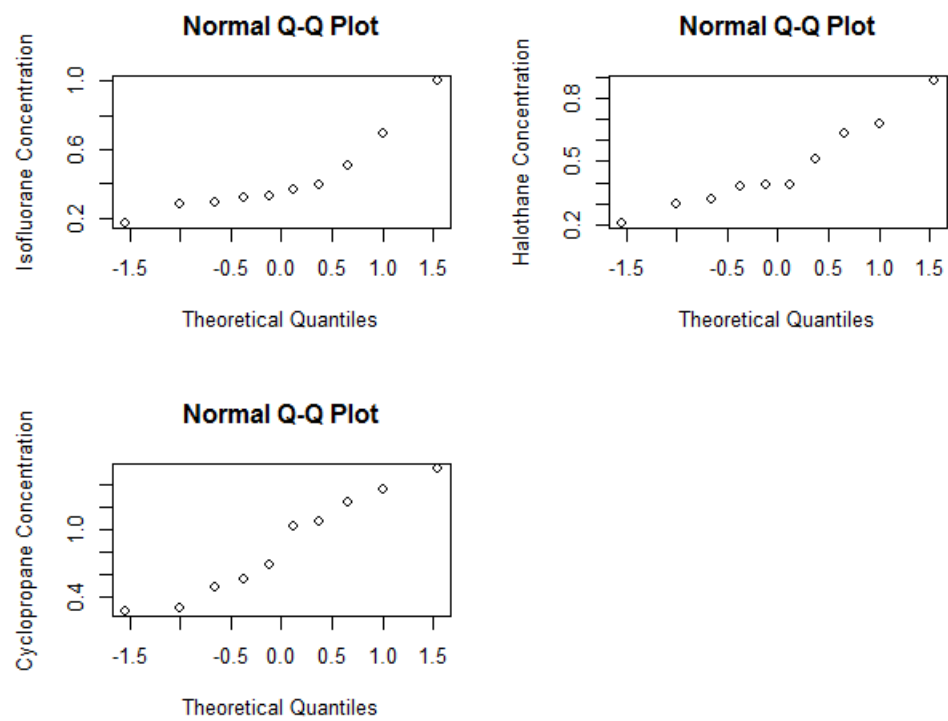


Figure 12: qq dog

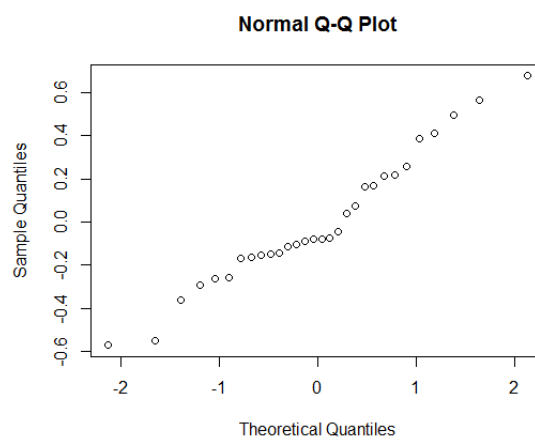


Figure 13: dog-resid

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

#2.2
sqclouds = sqrt(clouds)

#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 original 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))
}

```