Assignment 3, EDDA 2017

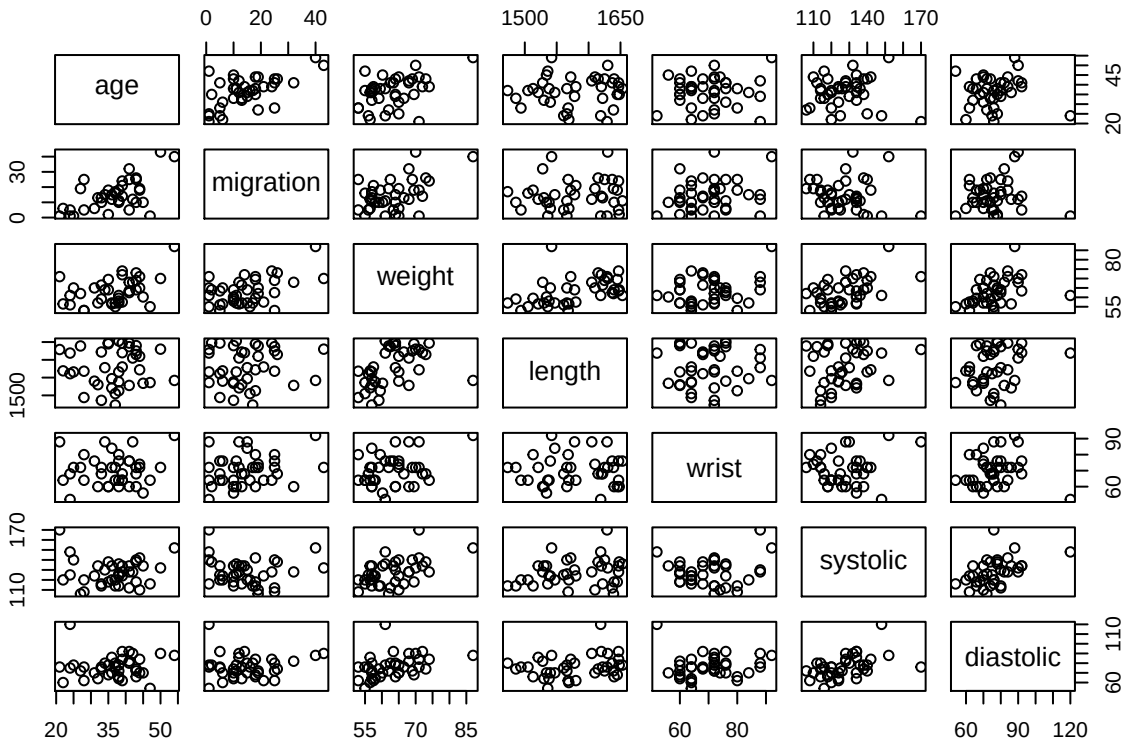
Martin de la Riva(11403799) and Kieran O'Driscoll(11426438), Group 23

24 April 2017

Exercise 1

Q1

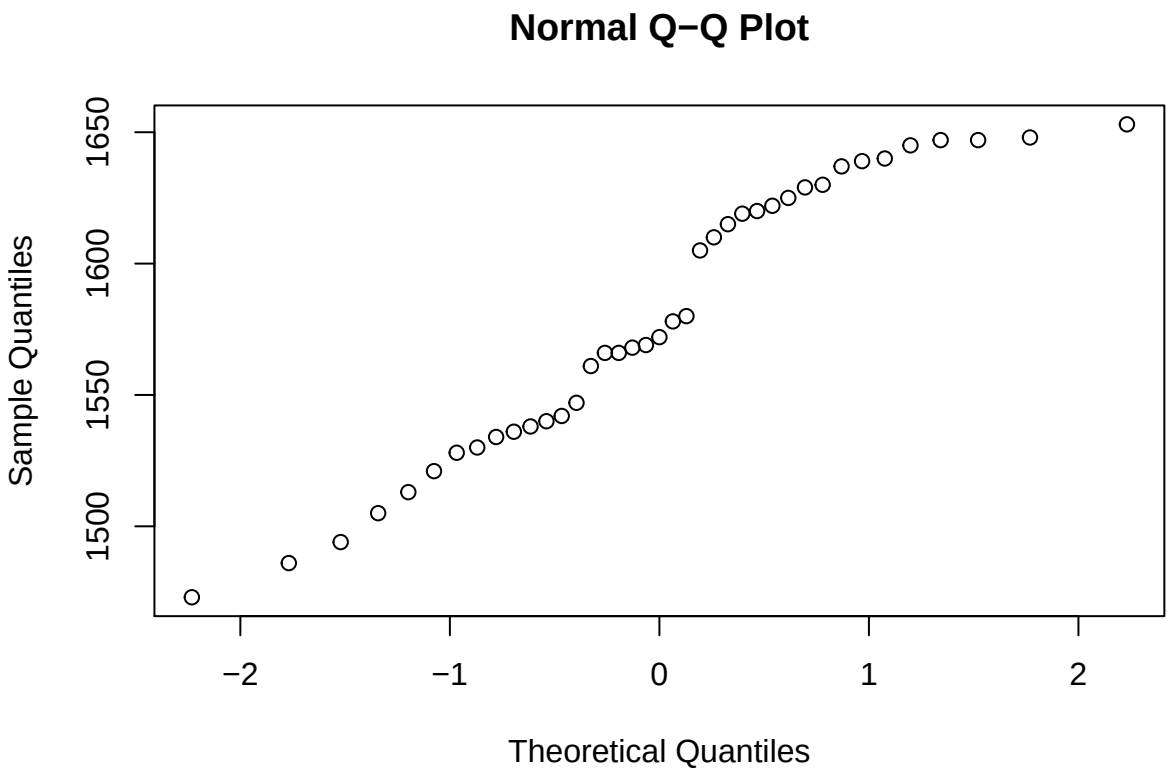
```
data = read.table("data/peruvians.txt", header=TRUE)
pairs(data[,-c(5,6,7)])
```



Based on the diagram above, age, weight and perhaps a case can be made for diastolic. These were chosen because their scatter plots show a cluster that shows the values are in porportion with migration whereas the othe graphs have a scatter plot that don't show any connection with migration.

Q2

```
qqnorm(data[['length']])
```




```
attach(data)
```

Peruvians is not drawn from a normal distribution so therefore will use Spearmen correlation test to test for correlation.

```
cor.test(age,migration,method="spearman")
```

```
## Warning in cor.test.default(age, migration, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  age and migration
## S = 5176.6, p-value = 0.002189
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4760575
```

The p-value recived is 0.002189 which falls under the 0.05 signicant level. Therefore the value is correlated with migration. Since the value is quite low, it can be said there is a high positive correlation as the correlation increases as values get bigger.

```r
cor.test(length,migration,method="spearman")
```

```
## Warning in cor.test.default(length, migration, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  length and migration
## S = 9044.3, p-value = 0.6087
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.08458432
```

With a P-value of 0.6087 it's clear that lenght and migration do not have a correlation.

```r
cor.test(wrist,migration,method="spearman")
```

```
## Warning in cor.test.default(wrist, migration, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  wrist and migration
## S = 7712.8, p-value = 0.1797
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2193498
```

Wrist gets a p-value of 0.1797 which means there is no correlation.

```r
cor.test(systolic,migration,method="spearman")
```

```
## Warning in cor.test.default(systolic, migration, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  systolic and migration
## S = 11544, p-value = 0.3054
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.1684286
```

Systolic has no correlation with migration, p-value= 0.3054

```r
cor.test(diastolic,migration,method="spearman")
```

```
## Warning in cor.test.default(diastolic, migration, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  diastolic and migration
## S = 9137.6, p-value = 0.6494
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.07514098
```

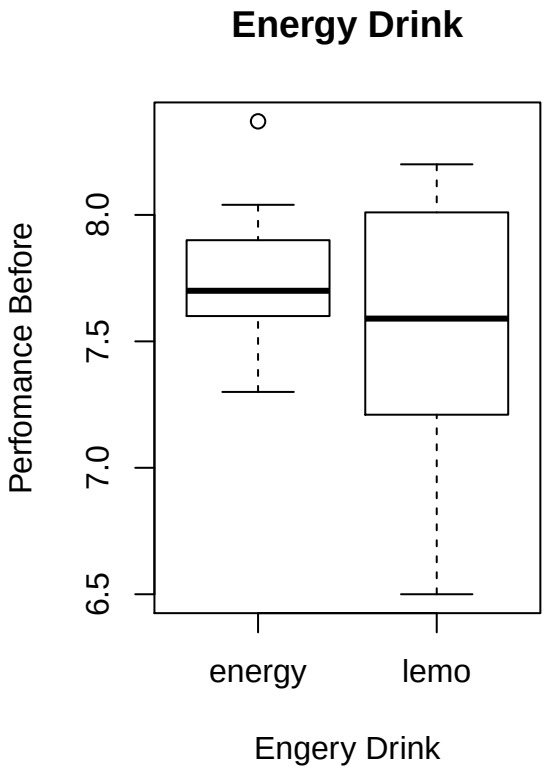
Diastolic has no correlation, the p-value was 0.6494

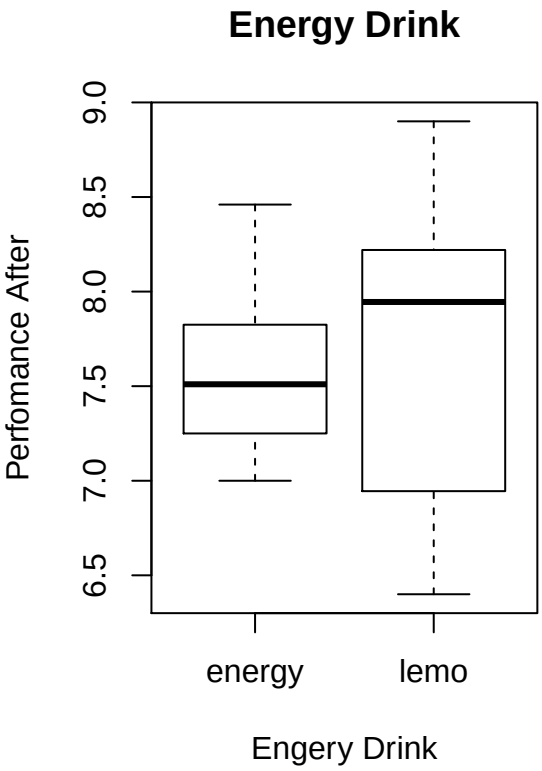
```r
cor.test(weight,migration,method="spearman")
```

```
## Warning in cor.test.default(weight, migration, method = "spearman"): Cannot
## compute exact p-value with ties
```

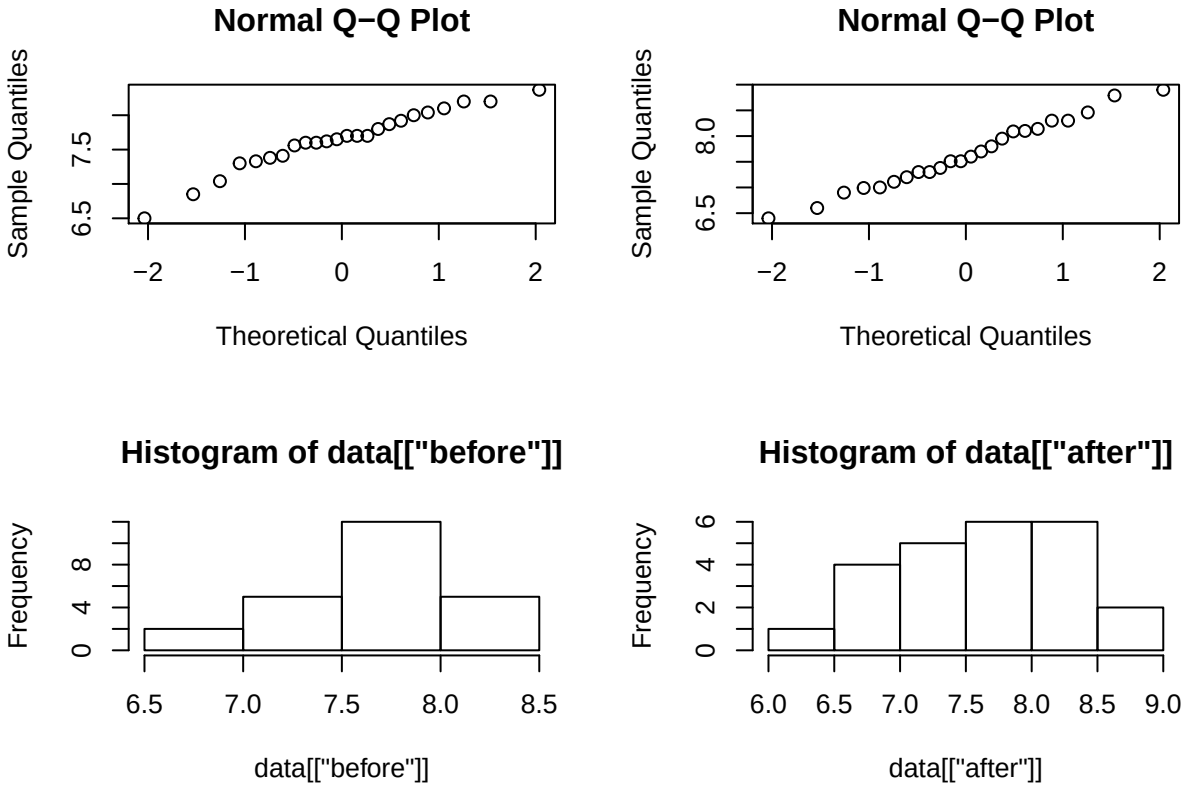
```
##
##  Spearman's rank correlation rho
##
## data:  weight and migration
## S = 6415.1, p-value = 0.02861
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3506956
```

Lastly, weight is correlated with migration having a p-value of 0.02861. Two of the tree values predicted in question one were correlated, bith weight and age. ##Exercise 1 ##Q1





From the boxplots it indicates that the engery actaully makes students worse after 30 mins and the lemonade makes students better.



Data looks like it was drawn from a normal distribution. With the exception of the histogram for "after" but there are only 12 entries per drink type which is a small amount to guage whether this is drawn from a normal distribution

Q2

Using a two paired sample test

```
##
##  Paired t-test
##
## data:  data["before"][filter] and data["after"][filter]
## t = 1.6538, df = 11, p-value = 0.1264
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05101059  0.35934392
## sample estimates:
## mean of the differences
##               0.1541667
```

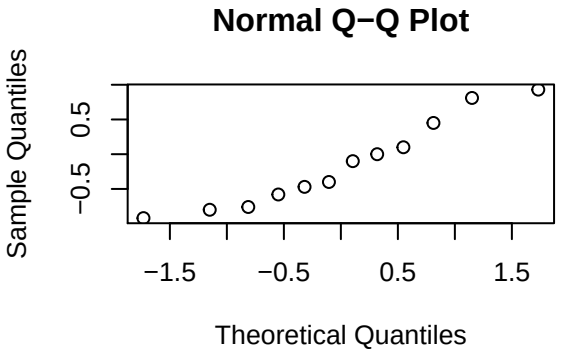
```
##
##  Paired t-test
##
## data:  data["before"][filter] and data["after"][filter]
```

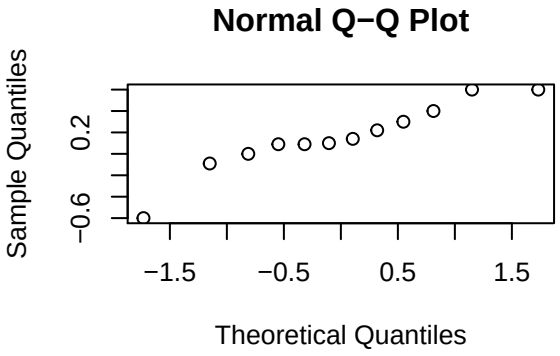
```
## t = -0.80596, df = 11, p-value = 0.4373
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5409781  0.2509781
## sample estimates:
## mean of the differences
##                  -0.145
```

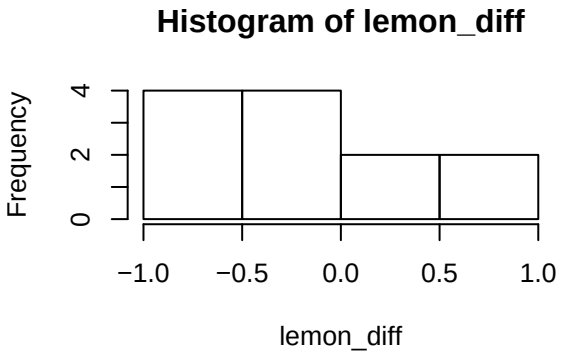
For both drinks the p-values obtained are 0.1264 and 0.4373 for "energy" and "lemo" respectively. These p-values fall above the 0.05 range so therefore we cannot reject the hypothesis nor say that there is increase from the before to the after.

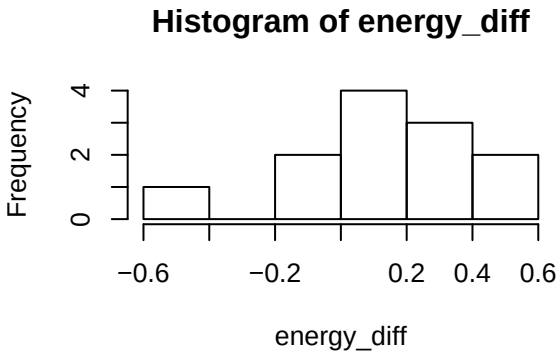
Q3

Using permuation test permutation for independent samples.










```
##
##  Welch Two Sample t-test
##
## data:  lemon_diff and energy_diff
## t = -1.4764, df = 16.509, p-value = 0.1586
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.7276409  0.1293076
## sample estimates:
##  mean of x  mean of y
## -0.1450000  0.1541667
```

With a high p-value of 0.1586 the null hypothesis can not be rejected therfore there in no meaningful difference between both energy and lemonade.

Q4

There is only a small sample size, each drink gets allocated 12 people, it could be the case that one group were faster on average therefore would be faster before and after whereas some people might need more than 30 minutes to recover. Also, perhaps a bigger margin then 30 minutes for the gap between before and after.

Q5

It would have affected the time difference if people were still tired after the first run.

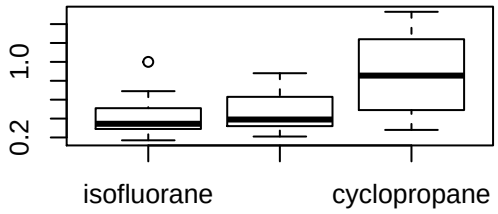
Q6

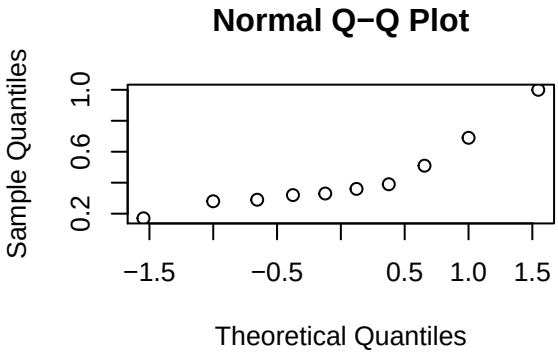
The conclusion is that both the differences for lemonade and energy are drawn from a normal distrubution. The qq plots can be seen in question 3.

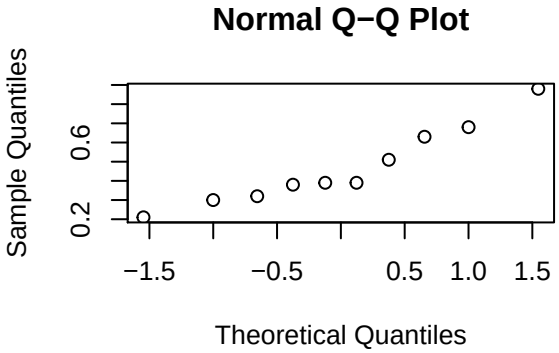
Exercise 3

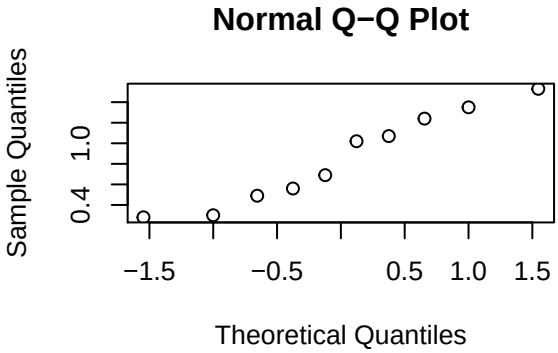
Q1

```
data = read.table("data/dogs.txt", header=TRUE)
par(mfrow=c(2,2))
boxplot(data,data=data)
qqnorm(data[['isofluorane']])
qqnorm(data[['halothane']])
qqnorm(data[['cyclopropane']])
```









Each drug type has 10 examples. All qqplots are close to a normal distribution with the exception of isofluorane which is displayed in top right. It can be presummed that by adding more data that the qqplots would more closely resembles a qqplot of a normal distubution. In this case it is reasonable to assume sample were taken from normal distribution.

Q2

Using anova.

```
treats = data.frame(dog=as.vector(as.matrix(data)),treatment=factor(rep(1:3,each=10)))
attach(treats)
pvcaov=lm(dog~treatment,data=treats)
anova(pvcaov)
```

```
## Analysis of Variance Table
##
## Response: dog
##           Df Sum Sq Mean Sq F value Pr(>F)
## treatment  2 1.0808 0.54040   5.355  0.011 *
## Residuals 27 2.7247 0.10092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The hypothesis is rejected as the p-value is 0.011 and therefore there is a difference between the

treatments.

```
summary(pvcaov)
```

```
##
## Call:
## lm(formula = dog ~ treatment, data = treats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5730 -0.1608 -0.0790  0.2000  0.6770
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4340     0.1005   4.320 0.000189 ***
## treatment2    0.0350     0.1421   0.246 0.807266
## treatment3    0.4190     0.1421   2.949 0.006504 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3177 on 27 degrees of freedom
## Multiple R-squared:  0.284,  Adjusted R-squared:  0.231
## F-statistic: 5.355 on 2 and 27 DF,  p-value: 0.011
```

The mean estimation for $\mu_i sofluorane$ is 0.434 and with a p-value of 0.000189. The estimation for $\mu_h alo - \mu_i so$ is 0.035 and with a p-value of 0.807266. The estimation for $\mu_c yclo - \mu_i so$ is 0.419 and with a p-value of 0.006504.

```
confint(pvcaov)
```

```
##                 2.5 %   97.5 %
## (Intercept)  0.227879 0.640121
## treatment2  -0.256499 0.326499
## treatment3   0.127501 0.710499
```

Those are the 95% confidence intervals for $\mu_i sofluorane$, $\mu_h alo - \mu_i so$ and $\mu_c yclo - \mu_i so$, respectively.

Q3

```
kruskal.test(dog,treatment,data=treats)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dog and treatment
## Kruskal-Wallis chi-squared = 5.6442, df = 2, p-value = 0.05948
```

The p-value is really close to 0.05, but it is slightly above it, so we cannot reject the hypothesis and therefore we cannot assume that the samples come from different populations. This contrasts with the findings we found with anova test.

```
treats = data.frame(dog=as.vector(as.matrix(data)),treatment=factor(rep(1:3,each=10)))
attach(treats)
```

```
## The following objects are masked from treats (pos = 3):
##
##     dog, treatment
```

```
pvcaov=lm(dog~treatment,data=treats)
anova(pvcaov)
```

```
## Analysis of Variance Table
##
## Response: dog
##           Df Sum Sq Mean Sq F value Pr(>F)
## treatment  2 1.0808 0.54040   5.355  0.011 *
## Residuals 27 2.7247 0.10092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qqnorm(pvcaov$residuals)
```