# STAT 210
# Applied Statistics and Data Analysis
# Problem List 4
# (Due on Week 5)

Fall 2022

The questions on this list come from previous exams. At the group meetings we will only address those items that directly relate to what we have covered this week. I have posted the complete questions so you can get an idea of how the test questions went in the past.

## Exercise 1

The file `theater.csv` has information on a survey conducted on visitors of a local Swiss theater in terms of age (`Age`), sex (`Sex`), annual income (`Income`), general expenditure on cultural activities (`Culture`), expenditure on theater visits (`Theatre`), and the estimated expenditure on theater visits in the year before the survey was done (`Theatre_ly`).

**Note:** Sex is coded 1 for females, and theater is spelled `Theatre`.

(a) Load the data into a data frame called `q1.df`. Divide the plotting window into two columns and draw boxplots for `Culture` and `Theatre` as a function of `Sex`. Comment on the results.

Loading the data

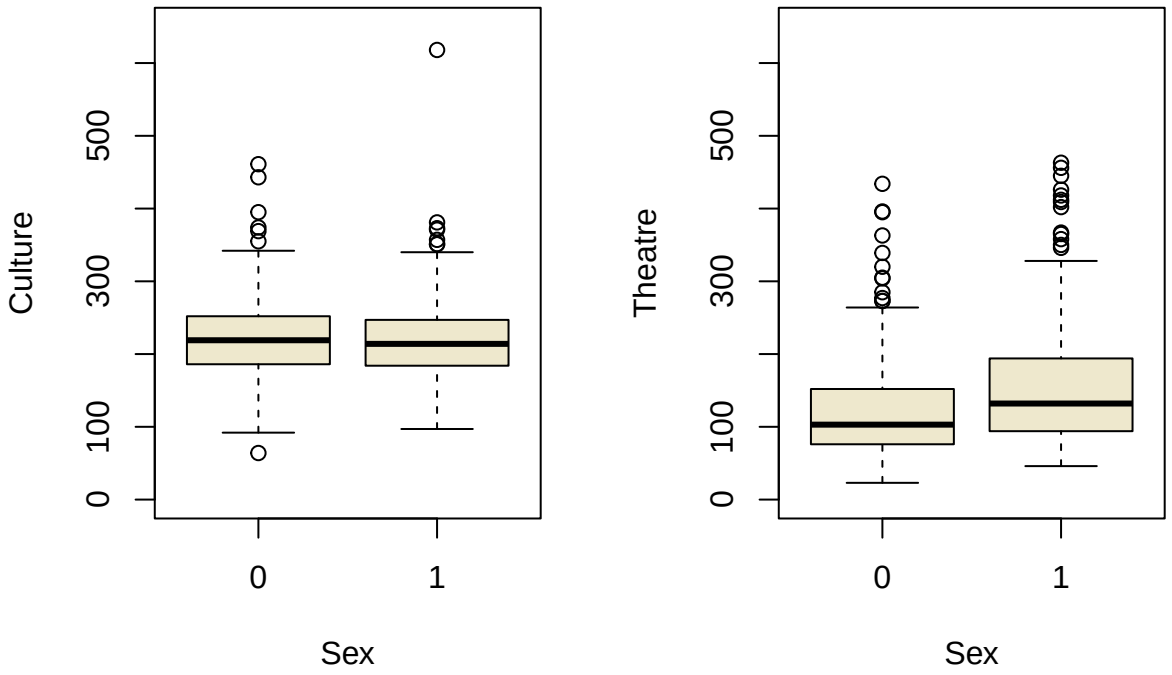
```
q1.df <- read.csv('theatre.csv')
str(q1.df)
```

```
## 'data.frame':    699 obs. of  7 variables:
##  $ X         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age       : int  31 54 56 36 24 25 61 50 53 52 ...
##  $ Sex       : int  1 0 1 1 1 0 1 0 1 0 ...
##  $ Income    : num  90.5 73 74.3 73.6 109 93.1 63.9 46.1 75 68.7 ...
##  $ Culture   : int  181 234 289 185 191 273 184 155 253 291 ...
##  $ Theatre   : int  104 116 276 75 172 168 119 97 152 166 ...
##  $ Theatre_ly: int  150 140 125 130 140 130 195 110 155 150 ...
```

```
attach(q1.df)
```

Graphs

```
par(mfrow=c(1,2))
boxplot(Culture ~ Sex, ylim = c(0,650), col = 'cornsilk2')
boxplot(Theatre ~ Sex, ylim = c(0,650), col = 'cornsilk2')
```




```
par(mfrow=c(1,1))
```

I set the same $y$ scales in both graphs, but this was not required. It just makes the comparisons easier. Since the expenditure in theater is part of the expenditure in culture, one would expect the values for theater to be lower than those for culture, which is indeed the case. The first boxplot (for culture) shows very similar distributions for both sexes but the second shows some differences, mainly in the expenditure level. Apparently, women spend more in theater visits than men do. We will test this later on in this question.

(b) The year before the survey was carried out, the average expenditure in culture was 216 Swiss francs. We want to test whether there is a change in the average behavior in this variable.

- What are the hypotheses you wish to contrast?
- What parametric test would be adequate for this?
- What are the assumptions on which this test is based, and why do you think they are satisfied?
- What is the test statistic? What is the corresponding sampling distribution?
- Carry out this test and discuss the results.

We want to compare whether the average culture expenditure $\hat{\mu}_n$ is equal to 216, so the hypotheses are

$$H_0 : \hat{\mu}_n = 216 \qquad \text{vs} \qquad H_1 : \hat{\mu}_n \neq 216$$

The adequate test for comparing the mean annual expenditure between the two years is the $t$-test, since we have to estimate the variance. In this case, there are 699 subjects in the sample. Since sample size is large, the Central Limit Theorem says that the normal approximation required for the $t$-test is reasonable. The test statistic for this test is the standardized sample mean

$$\frac{\hat{\mu}_n - 216}{s_n/\sqrt{699}}$$

where $s_n$ is the sample standard deviation. We calculate below the value for this statistic (this was not required)

```
(tn <- (mean(Culture)-216)/(sd(Culture)/sqrt(699)))
```

```
## [1] 1.967766
```

The sampling distribution is the $t$ distribution with 698 degrees of freedom. The following command carries out this test in R

```
t.test(Culture, mu=216)
```

```
##
##  One Sample t-test
##
## data:  Culture
## t = 1.9678, df = 698, p-value = 0.04949
## alternative hypothesis: true mean is not equal to 216
## 95 percent confidence interval:
##  216.0086 223.7024
## sample estimates:
## mean of x
##  219.8555
```

The $p$ value is just below 0.05 and the decision depends on our choice for $\alpha$. If we choose 0.05, the null hypothesis is rejected, while if we choose 0.02 or 0.01, we will not reject the null hypothesis.

(c) Test the hypothesis that women spend more on theater visits than men. What is your conclusion?

We now want to compare the average values for two populations (males and females) and the adequate test for this is the $t$-test. In this case it is justified because the sample sizes for each population are large enough:

```
sum(Sex==0); sum(Sex==1)
```

```
## [1] 309
```

```
## [1] 390
```

In the sample, 309 are males and 390 are females. Since sample sizes are large, the Central Limit Theorem says that the $t$-test is a reasonable choice. We want to test whether women spend more so we need a one-sided alternative hypothesis. The following command carries out this test in R

```
t.test(Theatre[Sex == 0], Theatre[Sex == 1], alternative = 'less')
```

```
##
##  Welch Two Sample t-test
##
## data:  Theatre[Sex == 0] and Theatre[Sex == 1]
## t = -5.6169, df = 694.15, p-value = 1.407e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -22.22746
## sample estimates:
## mean of x mean of y
##  122.1327  153.5821
```

The $p$-value is small and we reject the null hypothesis of equal expenditure. Our conclusion is that women spend more on theater visits than men.

(d) What parametric test would be adequate to compare the average expenditure in theater visits in the survey year and the preceding year for the whole population? Carry this test out and discuss your findings.

Since we have data for both years for the same subjects, the adequate test here is a paired test.

```
t.test(Theatre_ly, Theatre, paired = TRUE)
```

```
##
```

```
##  Paired t-test
##
## data:  Theatre_ly and Theatre
## t = -1.0925, df = 698, p-value = 0.275
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.707533  2.481496
## sample estimates:
## mean of the differences
##               -3.113019
```

The $p$-value is large, so we do not reject the null hypothesis of equal expenditure.

(e) What non-parametric test or tests would be adequate for (b) and (c)? Perform these tests and compare your results with what you obtained before.

For (b) we have the Wilcoxon one sample test

```
wilcox.test(Culture, mu=216)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Culture
## V = 124991, p-value = 0.527
## alternative hypothesis: true location is not equal to 216
```

We would reach the same conclusion with this test. For (c) we have Wilcoxon's test for two samples

```
wilcox.test(Theatre[Sex == 0],Theatre[Sex == 1], alternative = 'less')
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Theatre[Sex == 0] and Theatre[Sex == 1]
## W = 44669, p-value = 2.071e-09
## alternative hypothesis: true location shift is less than 0
```

Again, we reach the same conclusion.

## Exercise 2

The data for this question is stored in the file `data_q1`. Read the data and save it in a file with the same name.

The data corresponds to the sugar content per serving and in grams for two popular breakfast cereals. The manufacturer claims that cereal `c1` has 15 gr of sugar per serving. We want use the sample to test whether the cereal has more than this amount of sugar per serving. Assume that the population distribution is Gaussian.

Read the data

```
data_q1 <- read.table('data_q1', header = TRUE)
str(data_q1)
```

```
## 'data.frame':    20 obs. of  2 variables:
##  $ c1: num  17 17.4 16.5 18 12.9 ...
##  $ c2: num  7.67 15.31 11.61 17.12 14.6 ...
```

```
attach(data_q1)
```

(i) What is the sampling distribution for the empirical mean in this situation?

The empirical mean $\mu_n$ of a Gaussian sample has a Gaussian distribution $N(\mu, \sigma^2/n)$, where $\mu$ is the population mean, $\sigma^2$ the population variance, and $n$ is the sample size. However, since the variance is unknown, if we use the sample estimate $s^2$ instead of the true value, the sampling distribution of the normalized mean

$$\frac{\mu_n - \mu}{s/\sqrt{n}}$$

has a $t$ distribution with $n-1$ degrees of freedom.

(ii) What hypothesis test should we carry out in this case? What are the assumptions for the test? What is the test statistic? Calculate the value for the test statistic based on the sample in `data_q1`.

We should do a one-sample $t$-test since we are comparing the average of the sample with a given (population) value. The test assumes that the data come from a normal distribution.

The sample mean and standard deviation are

```
(m1 <- mean(c1));(sd1 <- sd(c1))
```

```
## [1] 15.8115
```

```
## [1] 1.946139
```

The statistic for this test is

$$t_0 = \frac{\mu_n - \mu}{s/\sqrt{n}} = \frac{15.812 - 15}{1.9461/\sqrt{20}}$$

We calculate this in `R`:

```
(t_0 = (m1-15)/(sd1/sqrt(20)))
```

```
## [1] 1.864789
```

(iii) Carry out the test you proposed and discuss the results using a confidence level of 95%. Verify whether the assumptions for the test hold for this sample.

In this case, the null hypothesis is that the cereal has 15 grams of sugar per serving, so the alternative is that the mean value is greater than 15.

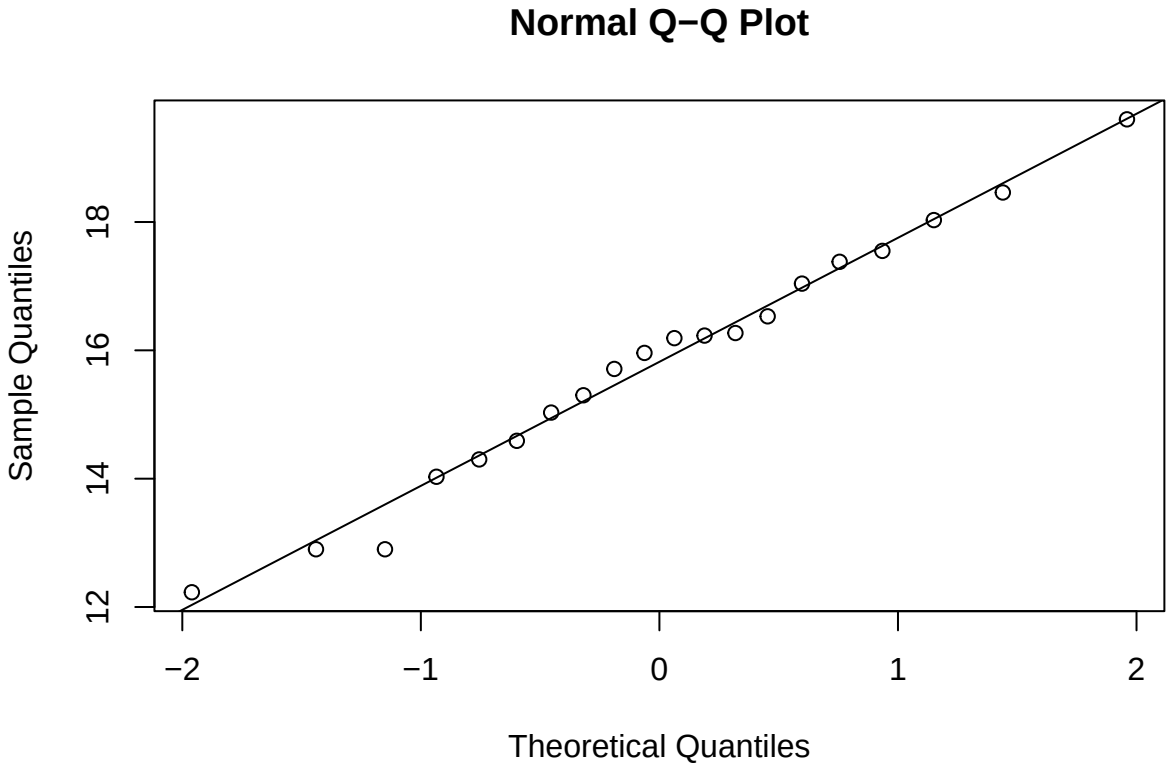
```
t.test(c1, mu = 15, alternative = 'greater')
```

```
##
##  One Sample t-test
##
## data:  c1
## t = 1.8648, df = 19, p-value = 0.03887
## alternative hypothesis: true mean is greater than 15
## 95 percent confidence interval:
##  15.05903      Inf
## sample estimates:
## mean of x
##   15.8115
```

Since the $p$-value is less than $\alpha = 0.05$, we reject the null hypothesis that $\mu = 15$.

The assumption for this test is that the distribution of the sample is normal. We do a qq-plot and a Shapiro-Wilk test

```
qqnorm(c1); qqline(c1)
```




```
shapiro.test(c1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  c1
## W = 0.98418, p-value = 0.9762
```

For this sample, both the quantile plot and the test support the assumption of a Gaussian distribution for the sample.

(iv) The data set `data_q1` also has a sample from another cereal, `c2`, produced by the same company. We want to compare whether this other cereal has the same average amount of sugar as the first. What test would be adequate for this problem? What are the assumptions? Carry out this test using the same confidence level as before and discuss your results.

The adequate test would be a two-sample $t$-test, which assumes that both samples come from Gaussian distributions.

```
t.test(c1, c2)
```

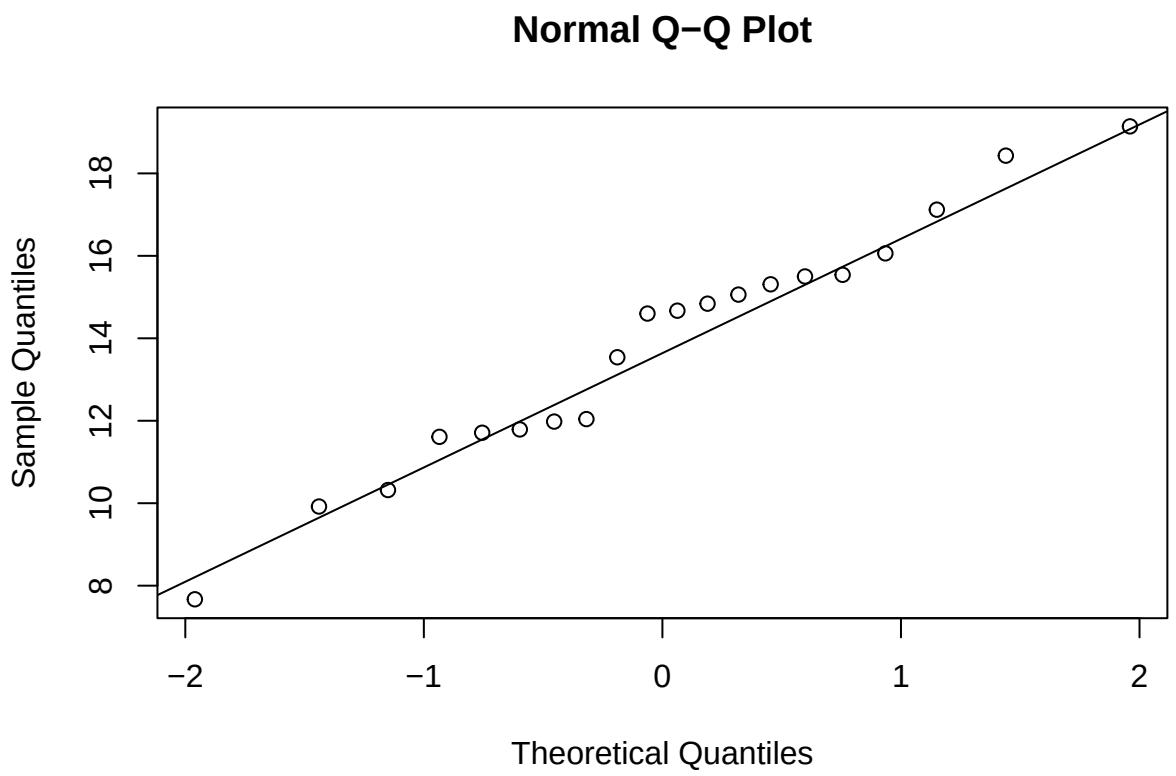
```
##
##  Welch Two Sample t-test
##
## data:  c1 and c2
## t = 2.5054, df = 33.055, p-value = 0.01733
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.370201 3.567799
## sample estimates:
## mean of x mean of y
##   15.8115   13.8425
```

The $p$ value is small, so we reject the null hypothesis that the two samples have the same mean.

(v) Check whether the assumptions needed for the test you carried out in (iv) hold in this case.

We have already checked that `c1` follows a normal distribution. Let's check this for `c2`.

```
qqnorm(c2); qqline(c2)
```




```
shapiro.test(c2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  c2
## W = 0.97306, p-value = 0.8177
```

Again, both the quantile plot and the test support the assumption of a Gaussian distribution for the sample.

(vi) What nonparametric test would make sense for comparing the two cereals? Carry out this test and compare it with your previous result.

We can do a Wilcoxon test, which compares the empirical means for two samples.

```
wilcox.test(c1,c2)
```

```
## Warning in wilcox.test.default(c1, c2): cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  c1 and c2
## W = 286, p-value = 0.02073
## alternative hypothesis: true location shift is not equal to 0
```

```
detach(data_q1)
```

The $p$-value is slightly larger than for the $t$-test but similar, and we would reach the same conclusion. Since the assumptions for the $t$-test seem to be satisfied, we would prefer the parametric test in this case.

## Exercise 3

The data for this exercise is in the file `q1.data` and come from a study of adult patients in a detoxification unit in the USA. Read the data set and store it as a data frame named `dframe.ex1`.

```
dframe.ex1 <- read.table('q1_data.txt', header = TRUE)
#dframe.ex1 <- read.table("q1_data.txt", header = TRUE)
```

(i) Explore the structure of this data set. How many variables are there? How many are categorical?

```
str(dframe.ex1)
```

```
## 'data.frame':    453 obs. of  88 variables:
##  $ id          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ e2b1        : int  NA NA NA NA NA NA NA 1 1 1 ...
##  $ g1b1        : int  0 0 0 0 0 0 NA 0 0 0 ...
##  $ i11         : int  NA 8 NA NA 64 1 NA NA 12 13 ...
##  $ pcs1        : num  54.2 59.6 58.5 46.6 31.4 ...
##  $ mcs1        : num  52.2 41.7 56.8 14.7 40.7 ...
##  $ cesd1       : int  7 11 14 44 26 23 NA 18 33 37 ...
##  $ indtot1     : int  5 12 99 97 55 12 NA 82 92 104 ...
##  $ drugrisk1   : int  0 0 13 0 0 0 NA 0 8 14 ...
##  $ sexrisk1    : int  1 0 4 4 4 4 NA 4 4 5 ...
##  $ pcrec1      : int  1 0 0 0 0 0 NA 0 0 2 ...
##  $ e2b2        : int  NA NA NA 1 NA NA 1 NA NA 2 ...
##  $ g1b2        : int  NA NA NA NA 0 NA 0 NA NA 0 ...
##  $ i12         : int  NA NA NA NA NA NA 13 NA NA 22 ...
##  $ pcs2        : num  NA NA NA 57.6 44.8 ...
##  $ mcs2        : num  NA NA NA 23.9 42.4 ...
##  $ cesd2       : int  NA NA NA NA 27 NA 47 NA NA 47 ...
##  $ indtot2     : int  NA NA NA NA 34 NA 92 NA NA 100 ...
##  $ drugrisk2   : int  NA NA NA NA 0 NA 0 NA NA 19 ...
##  $ sexrisk2    : int  NA NA NA NA 2 NA 9 NA NA 0 ...
##  $ pcrec2      : int  NA NA NA 0 0 NA 2 NA NA 2 ...
##  $ e2b3        : int  NA NA NA NA 1 NA NA 2 4 NA ...
##  $ g1b3        : int  0 NA NA NA 0 NA 0 0 0 NA ...
##  $ i13         : int  NA NA NA NA 13 NA 13 12 27 NA ...
##  $ pcs3        : num  52.1 NA NA NA 25 ...
##  $ mcs3        : num  56.1 NA NA NA 50.9 ...
##  $ cesd3       : int  8 NA NA NA 15 NA 54 25 35 NA ...
##  $ indtot3     : int  0 NA NA NA 37 NA 104 92 88 NA ...
##  $ drugrisk3   : int  0 NA NA NA 0 NA 0 6 7 NA ...
##  $ sexrisk3    : int  1 NA NA NA 4 NA 7 6 8 NA ...
##  $ pcrec3      : int  2 NA NA NA 0 NA 2 0 2 NA ...
##  $ e2b4        : int  NA NA 2 NA NA NA 1 NA NA NA ...
##  $ g1b4        : int  0 NA 0 0 0 NA 0 NA 0 NA ...
##  $ i14         : int  8 NA 3 NA 6 NA NA NA 9 NA ...
##  $ pcs4        : num  52.3 NA 66.2 57.1 44.7 ...
##  $ mcs4        : num  58 NA 13.6 53.5 32.7 ...
##  $ cesd4       : int  5 NA 49 20 28 NA 52 NA 27 NA ...
```

```
##  $ indtot4     : int  34 NA 94 5 58 NA 113 NA 93 NA ...
##  $ drugrisk4   : int  0 NA 19 0 0 NA 0 NA 0 NA ...
##  $ sexrisk4    : int  3 NA 4 4 2 NA 7 NA 3 NA ...
##  $ pcrec4      : int  2 NA 0 0 0 NA 2 NA 2 NA ...
##  $ a15a        : int  0 2 0 0 15 0 0 4 1 4 ...
##  $ a15b        : int  0 3 0 0 0 0 0 1 134 20 ...
##  $ d1          : int  3 22 0 2 12 1 14 1 14 4 ...
##  $ e2b         : int  NA NA NA 1 1 NA 1 8 7 3 ...
##  $ f1a         : int  3 3 3 0 3 1 3 1 3 2 ...
##  $ f1b         : int  2 2 2 0 0 0 1 1 2 3 ...
##  $ f1c         : int  3 0 3 1 3 1 3 2 3 3 ...
##  $ f1d         : int  0 3 0 3 3 3 1 3 1 0 ...
##  $ f1e         : int  2 3 2 2 3 0 3 3 3 1 ...
##  $ f1f         : int  3 2 2 2 3 0 3 3 3 2 ...
##  $ f1g         : int  3 0 1 1 1 0 3 3 3 3 ...
##  $ f1h         : int  0 0 3 3 3 3 1 1 0 0 ...
##  $ f1i         : int  2 3 2 0 3 0 3 1 3 3 ...
##  $ f1j         : int  3 0 3 0 2 1 3 0 3 1 ...
##  $ f1k         : int  3 3 1 1 3 1 3 3 3 3 ...
##  $ f1l         : int  0 0 0 2 2 3 0 1 1 0 ...
##  $ f1m         : int  1 0 1 2 2 1 0 3 2 3 ...
##  $ f1n         : int  2 3 3 2 3 0 3 0 3 3 ...
##  $ f1o         : int  2 0 2 0 0 1 3 1 2 1 ...
##  $ f1p         : int  2 0 0 NA 3 3 1 0 0 0 ...
##  $ f1q         : int  2 0 0 2 3 0 3 2 1 0 ...
##  $ f1r         : int  3 2 3 0 3 0 3 2 3 3 ...
##  $ f1s         : int  3 0 2 0 3 0 3 0 1 0 ...
##  $ f1t         : int  2 0 0 1 3 0 3 0 2 3 ...
##  $ g1b         : int  1 1 0 0 0 0 1 1 0 0 ...
##  $ i1          : int  13 56 0 5 10 4 13 12 71 20 ...
##  $ i2          : int  26 62 0 5 13 4 20 24 129 27 ...
##  $ age         : int  37 37 26 39 32 47 49 28 50 39 ...
##  $ treat       : int  1 1 0 0 0 1 0 1 0 1 ...
##  $ homeless    : int  0 1 0 0 1 0 0 1 1 1 ...
##  $ pcs         : num  58.4 36 74.8 61.9 37.3 ...
##  $ mcs         : num  25.11 26.67 6.76 43.97 21.68 ...
##  $ cesd        : int  49 30 39 15 39 6 52 32 50 46 ...
##  $ indtot      : int  39 43 41 28 38 29 38 44 44 44 ...
##  $ pss_fr      : int  0 1 13 11 10 5 1 4 5 0 ...
##  $ drugrisk    : int  0 0 20 0 0 0 0 7 18 20 ...
##  $ sexrisk     : int  4 7 2 4 6 5 8 6 8 0 ...
##  $ satreat     : int  0 0 0 1 0 0 1 1 0 1 ...
##  $ drinkstatus : int  1 1 1 0 1 1 NA 1 1 1 ...
##  $ daysdrink   : int  177 2 3 196 2 31 NA 47 62 115 ...
##  $ anysubstatus: int  1 1 1 1 1 1 NA 1 1 1 ...
##  $ daysanysub  : int  177 2 3 189 2 31 NA 47 31 115 ...
##  $ linkstatus  : int  1 NA 0 0 1 0 0 0 0 0 ...
##  $ dayslink    : int  225 NA 365 343 57 365 334 365 365 382 ...
##  $ female      : int  0 0 0 1 0 1 1 0 1 0 ...
##  $ substance   : chr  "cocaine" "alcohol" "heroin" "heroin" ...
##  $ racegrp     : chr  "black" "white" "black" "white" ...
```

There are 88 variables of which only two are categorical (factors).

(ii) In what follows you will use only the variables `pcs, mcs` and `female`.

Create a new data frame named `df.q1` with these variable, using the same names.

```
attach(dframe.ex1)
df.q1 <- data.frame(pcs,mcs,female)
str(df.q1)
```

```
## 'data.frame':    453 obs. of  3 variables:
##  $ pcs   : num  58.4 36 74.8 61.9 37.3 ...
##  $ mcs   : num  25.11 26.67 6.76 43.97 21.68 ...
##  $ female: int  0 0 0 1 0 1 1 0 1 0 ...
```

```
detach(dframe.ex1)
```

(iii) Variables `mcs` and `pcs` stand for 'mental component score' and 'physical component score'. The variable `female` is the gender of the subject, with code 0 for male and 1 for female. Add a new categorical variable named `gender` with values `m` and `f` for male and female, respectively, to `df.q1`.

```
gender <- factor(df.q1$female, labels = c('m','f'))
df.q1$gender <- gender
str(df.q1)
```

```
## 'data.frame':    453 obs. of  4 variables:
##  $ pcs   : num  58.4 36 74.8 61.9 37.3 ...
##  $ mcs   : num  25.11 26.67 6.76 43.97 21.68 ...
##  $ female: int  0 0 0 1 0 1 1 0 1 0 ...
##  $ gender: Factor w/ 2 levels "m","f": 1 1 1 2 1 2 2 1 2 1 ...
```

(iv) We want now to explore the variables `mcs` and `pcs` and compare their values for both genders. Calculate means and standard deviations for both variables (`mcs` and `pcs`) according to `gender`.

As an example, we give two ways of calculating means.

Means and standard deviations for `mcs`:

```
attach(df.q1)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     gender
```

```
(mcs.mean <- tapply(mcs,list(gender),mean))
```

```
##        m        f
## 32.52331 28.93896
```

```
(mcs.sd <- tapply(mcs,list(gender),sd))
```

```
##        m        f
## 12.87024 12.40615
```

Means and standard deviations for `pcs`:

```
(pcs.mean <- c(mean(pcs[gender=='m']),
               mean(pcs[gender=='f'])))
```

```
## [1] 48.98624 45.01636
```

```
(pcs.sd <- c(sd(pcs[gender=='m']),
             sd(pcs[gender=='f'])))
```

```
## [1] 10.82679 10.11369
```

(v) Draw histograms for `mcs` according to `gender`. Recall that the purpose is to make comparisons between the two populations. Repeat for `pcs`. Comment on your findings.

Since the graphs are for comparing the distributions for the two `gender` categories, we should use the same scales in both axes. An appropriate $x$-axis scale can be determined by looking at the range of the variables:

```
range(mcs);  range(pcs)
```

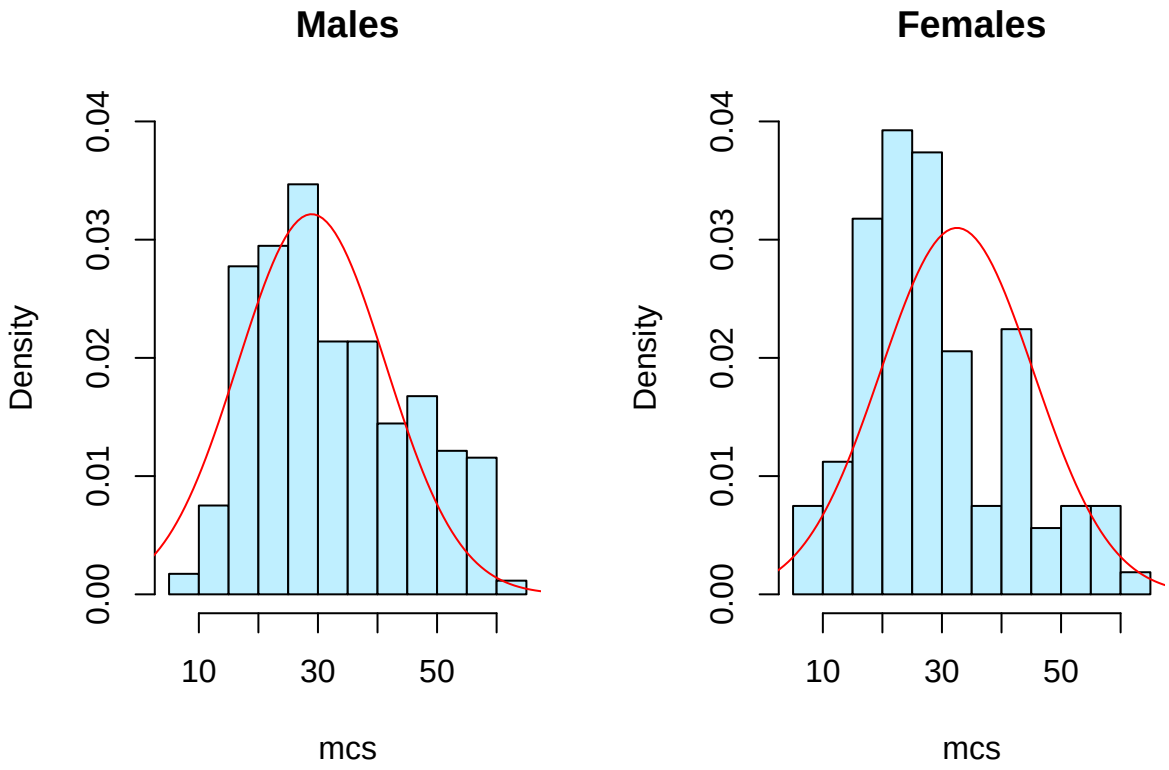
```
## [1]  6.762923 62.175503
```

```
## [1] 14.07429 74.80633
```

We may choose limits 5 and 65 for `mcs` and 10 and 80 for `pcs`. Other valid choices are possible. The scale for the $y$-axis was chosen after seeing some trial graphs. I chose to do relative frequency (`freq = FALSE`) histograms, but absolute frequency is also a valid choice. I have also added a normal density with the estimated mean and variance, but this was not required. I also present the graphs side by side and on a single column. Both alternatives are valid.

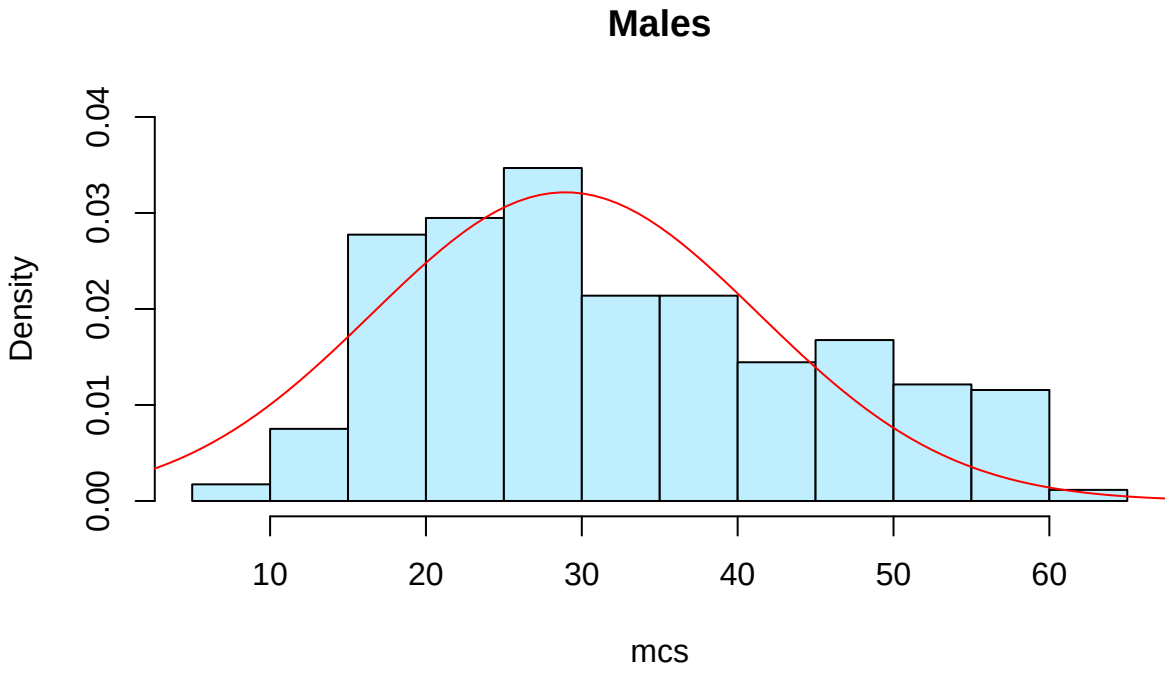
Graphs for `mcs`:

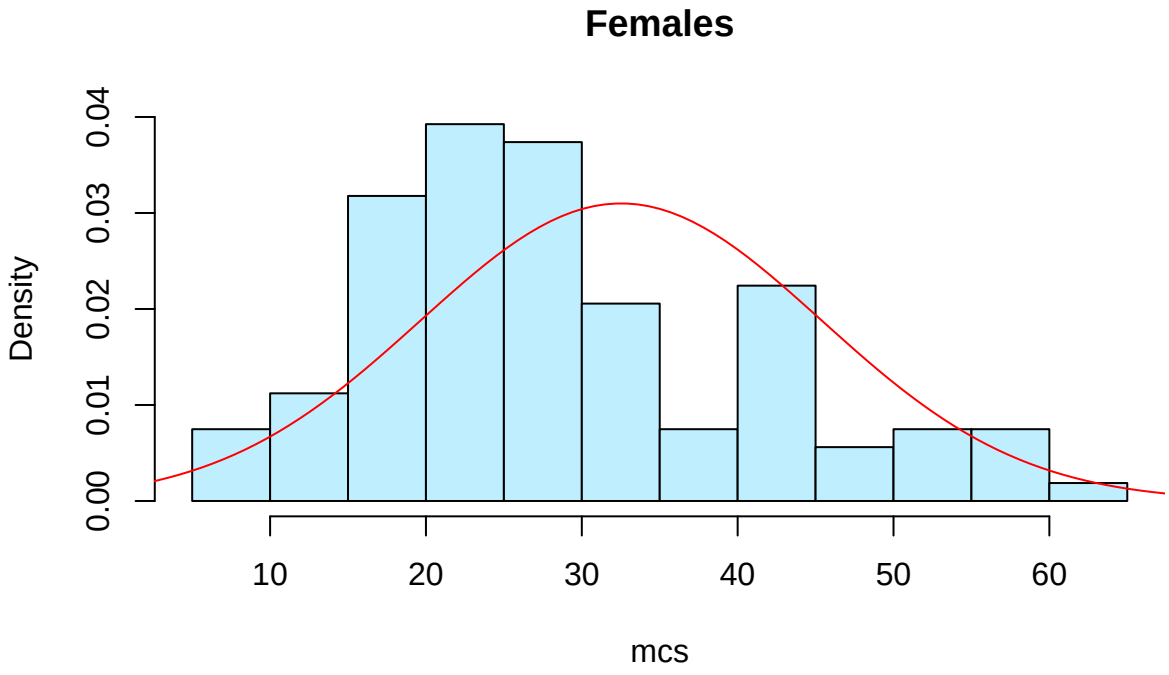
```
par(mfrow=c(1,2))
hist(mcs[gender=='m'], col='lightblue1',freq=FALSE, ylim=c(0,0.04),
     xlim = c(5,65), main = 'Males',xlab='mcs')
curve(dnorm(x,mcs.mean[2],mcs.sd[2]),0,70, col='red', add=TRUE)
hist(mcs[gender=='f'], col='lightblue1',freq=FALSE, ylim=c(0,0.04),
     xlim = c(5,65),main = 'Females',xlab='mcs')
curve(dnorm(x,mcs.mean[1],mcs.sd[1]),0,70,  col='red', add=TRUE)
```




```
par(mfrow=c(2,1))
hist(mcs[gender=='m'], col='lightblue1',freq=FALSE,
     xlim = c(5,65),main = 'Males',xlab='mcs', ylim=c(0,0.04))
curve(dnorm(x,mcs.mean[2],mcs.sd[2]),0,70, col='red', add=TRUE)
hist(mcs[gender=='f'], col='lightblue1',freq=FALSE,
```

```
     xlim = c(5,65),main = 'Females',xlab='mcs', ylim=c(0,0.04))
curve(dnorm(x,mcs.mean[1],mcs.sd[1]),0,70, col='red', add=TRUE)
```



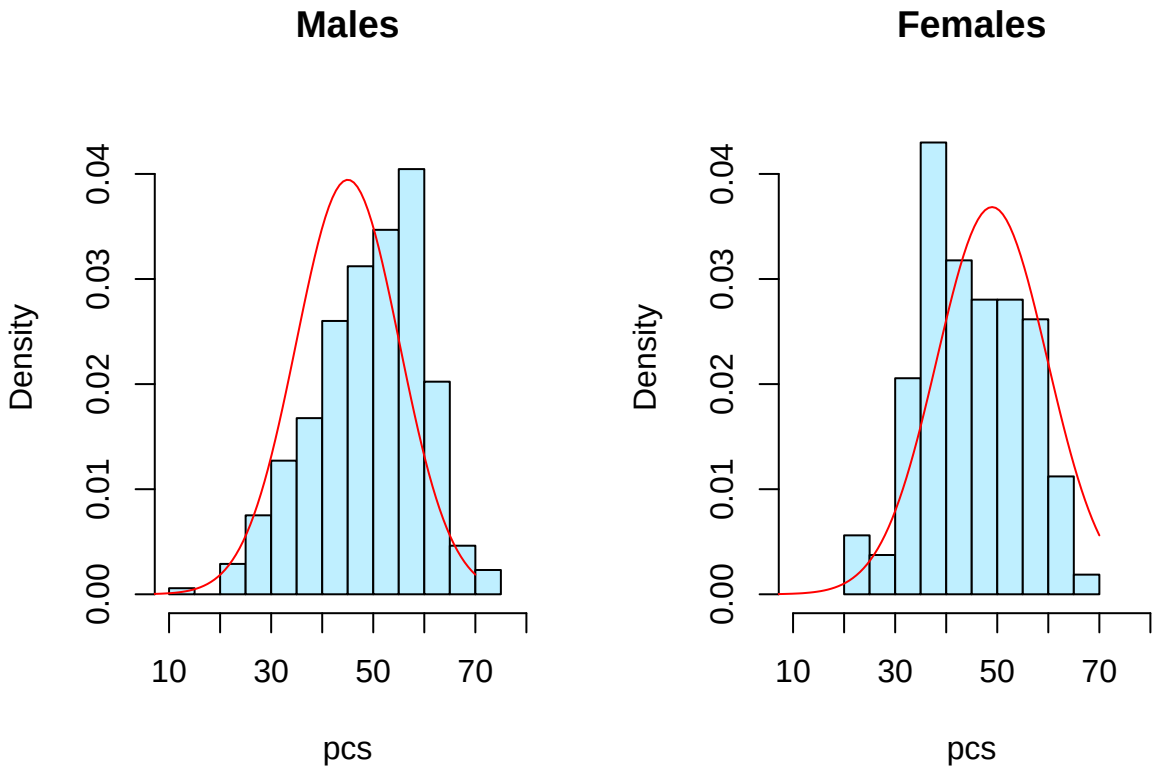




```
par(mfrow=c(1,1))
```

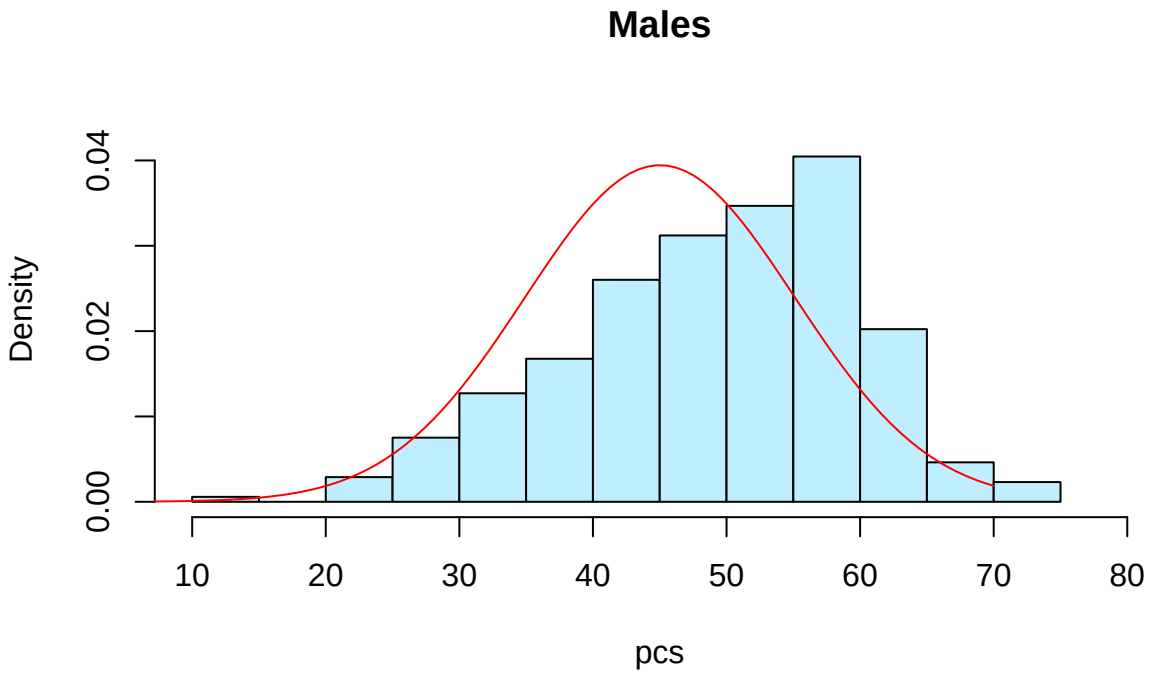
Graphs for pcs:

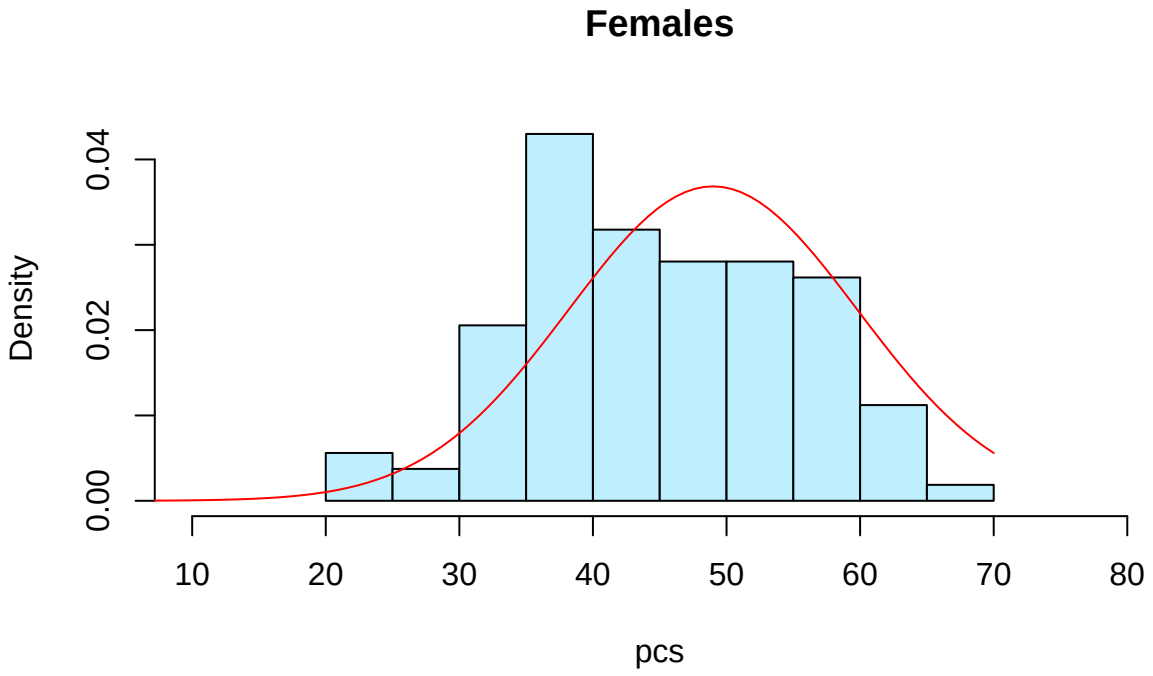
```
par(mfrow=c(1,2))
hist(pcs[gender=='m'], col='lightblue1',freq=FALSE,
```

```
     xlim = c(10,80),main = 'Males',xlab='pcs', ylim=c(0,0.045))
curve(dnorm(x,pcs.mean[2],pcs.sd[2]),0,70,   col='red', add=TRUE)
hist(pcs[gender=='f'], col='lightblue1',freq=FALSE,
     xlim = c(10,80),main = 'Females',xlab='pcs', ylim=c(0,0.045))
curve(dnorm(x,pcs.mean[1],pcs.sd[1]),0,70, col='red', add=TRUE)
```




```
par(mfrow=c(2,1))
hist(pcs[gender=='m'], col='lightblue1',freq=FALSE,
     xlim = c(10,80),main = 'Males',xlab='pcs', ylim=c(0,0.045))
curve(dnorm(x,pcs.mean[2],pcs.sd[2]),0,70, col='red', add=TRUE)
hist(pcs[gender=='f'], col='lightblue1',freq=FALSE,
     xlim = c(10,80),main = 'Females',xlab='pcs', ylim=c(0,0.045))
curve(dnorm(x,pcs.mean[1],pcs.sd[1]),0,70, col='red', add=TRUE)
```



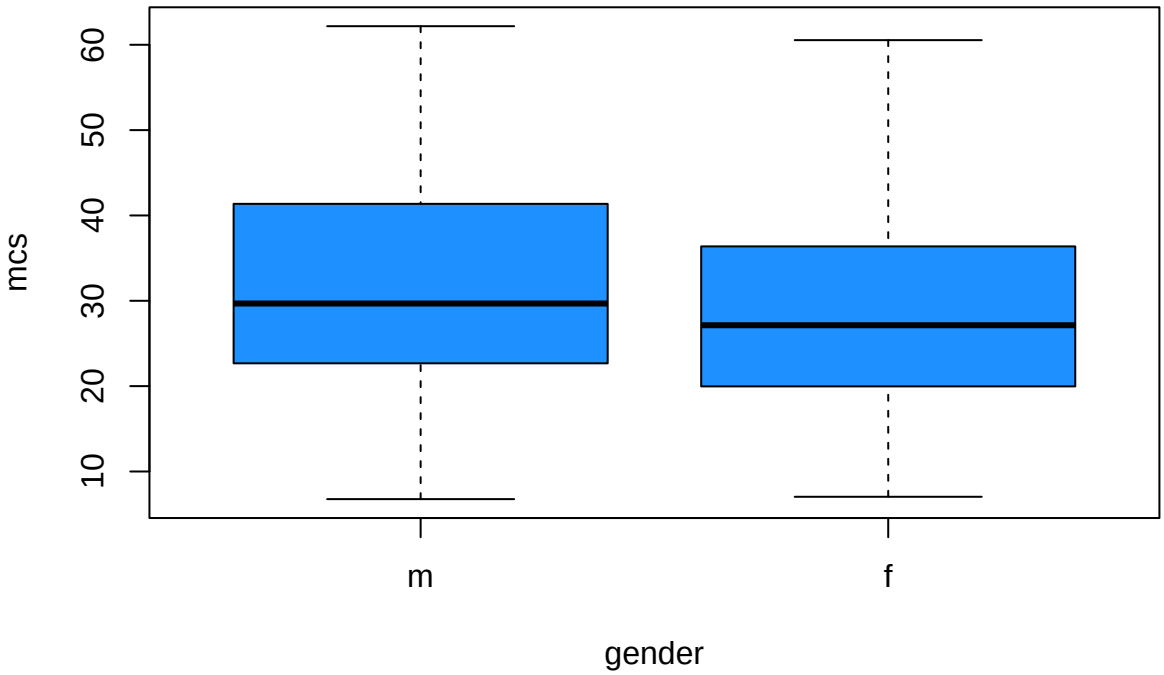




```
par(mfrow=c(1,1))
```

(vi) Draw boxplots for `mcs` according to `gender`. Both boxplots should appear in the same graph window. Repeat for `pcs`.

Boxplots for `mcs`:

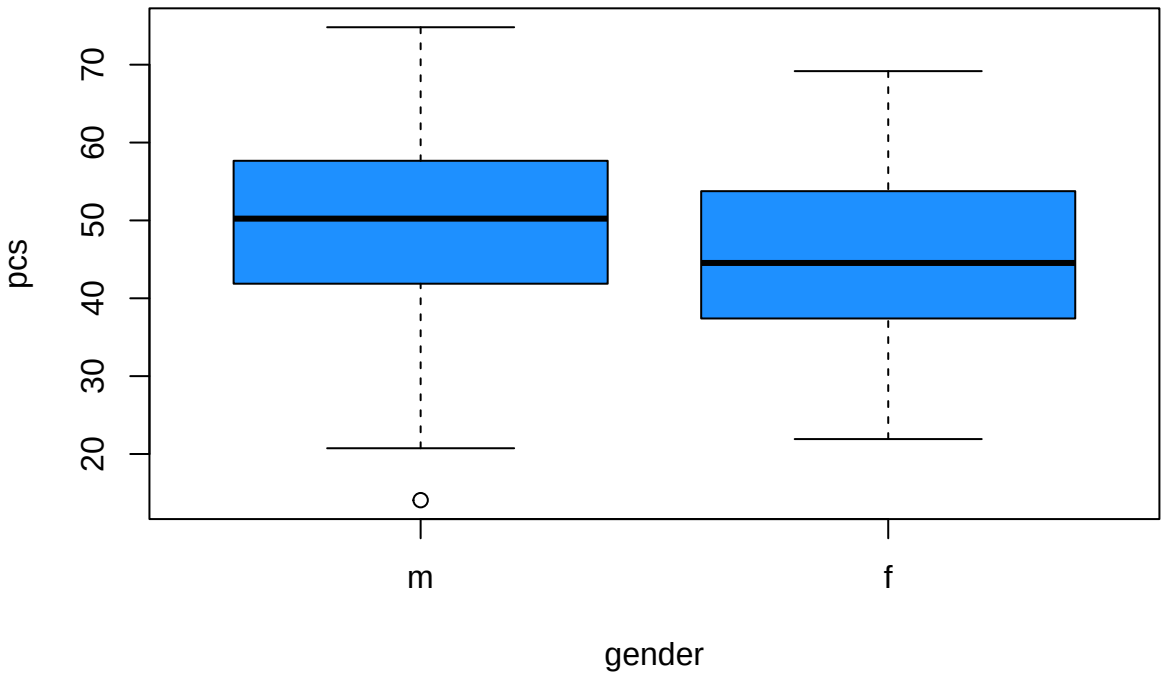
```
boxplot(mcs ~ gender, col='dodgerblue1')
```




The boxplots show that there is more dispersion in the male distribution and a slightly larger average value. The distribution for males is shifted to higher values in comparison with the female distribution.

Boxplots for `pcs`:

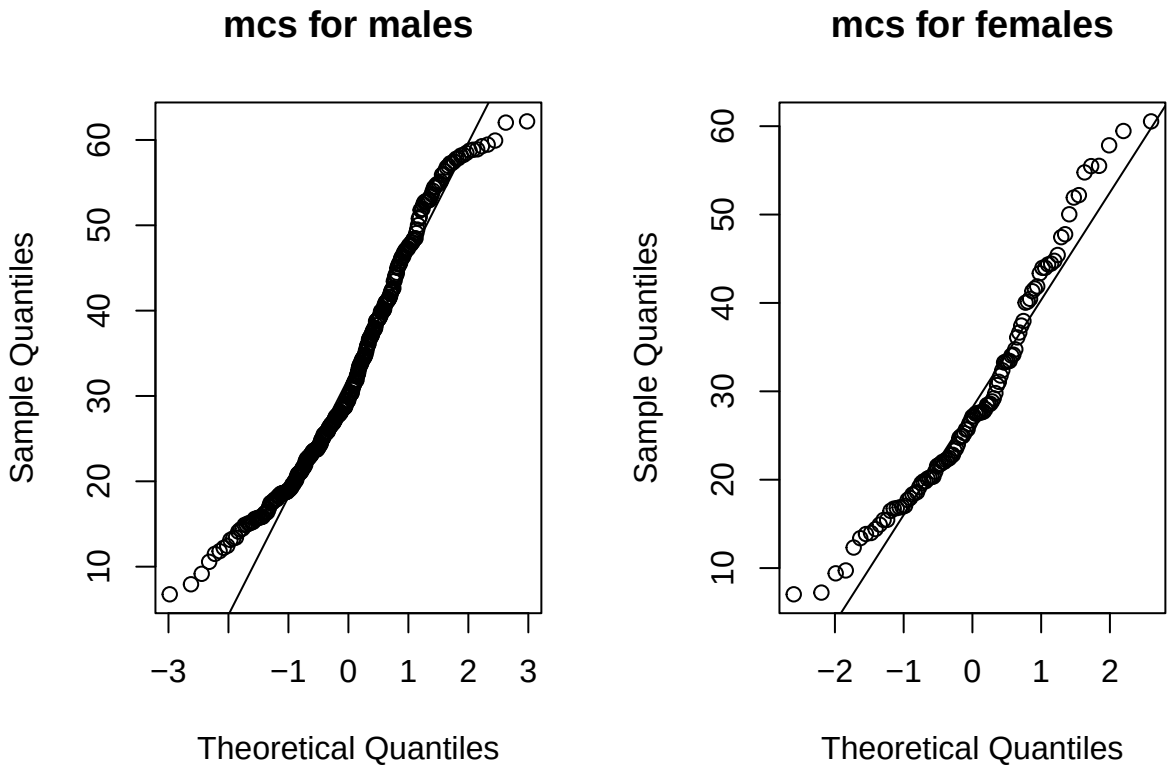
```
boxplot(pcs ~ gender, col='dodgerblue1')
```




(vii) Draw normal quantile plots for `mcs` according to `gender` and discuss the results. Repeat for `pcs`.

For `mcs`:

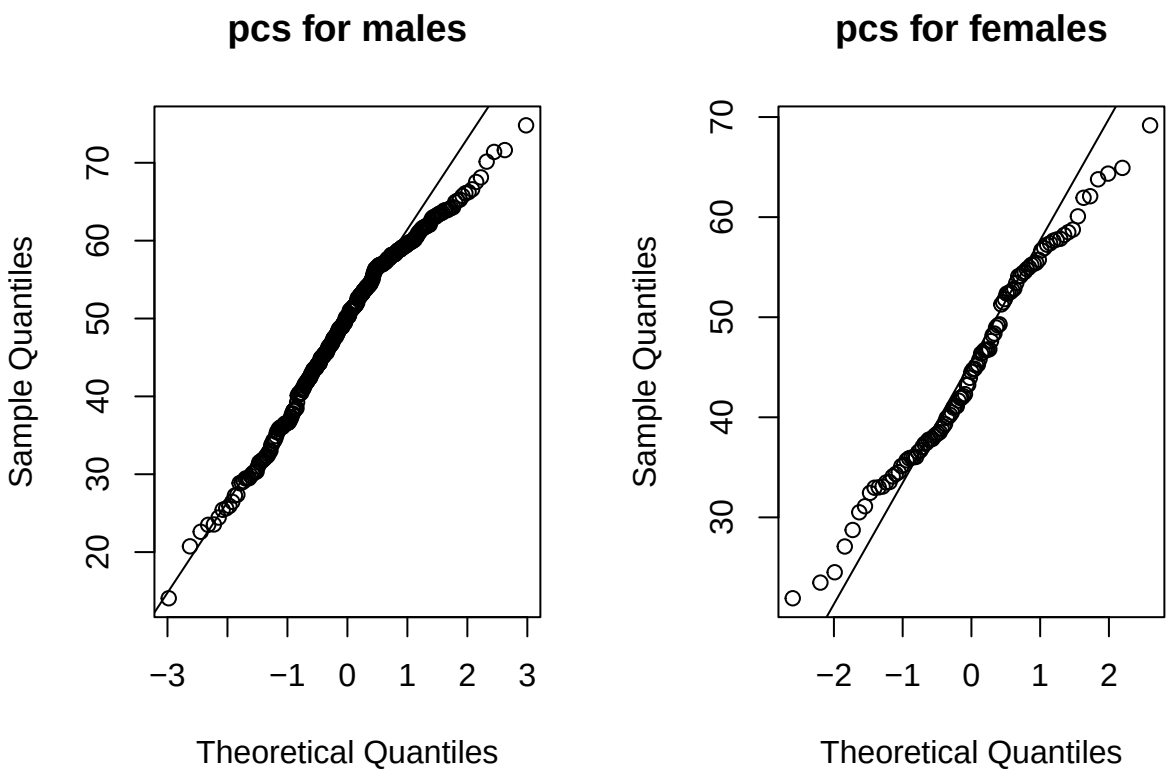
```
par(mfrow=c(1,2))
qqnorm(mcs[gender=='m'], main='mcs for males')
qqline(mcs[gender=='m'])
qqnorm(mcs[gender=='f'], main='mcs for females')
qqline(mcs[gender=='f'])
```




The fit is good on the central part but not at the tails, particularly for males.

For pcs:

```
par(mfrow=c(1,2))
qqnorm(pcs[gender=='m'], main='pcs for males')
qqline(pcs[gender=='m'])
qqnorm(pcs[gender=='f'], main='pcs for females')
qqline(pcs[gender=='f'])
```




(viii) Do a test to compare whether the means for `mcs` are equal in the two populations (male and female). Explain carefully the assumptions you are making. Do you think they are justified by what you have seen in your exploratory analysis? Discuss your results. Do you think a paired test would make sense in this context? (why or why not).

To compare means, a natural choice is a t-test. Since we are comparing averages and the sizes of the populations are reasonably large, the normal approximation to the sampling distribution of the mean seems justified. The distribution of the data do not seem to follow a Gaussian distribution but the Central Limit Theorem allows us to use the normal approximation.

A paired test makes no sense in this context.

```
t.test(mcs[gender=='m'],mcs[gender=='f'])
```

```
##
##  Welch Two Sample t-test
##
## data:  mcs[gender == "m"] and mcs[gender == "f"]
## t = 2.5887, df = 182.1, p-value = 0.01041
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8523895 6.3163135
## sample estimates:
## mean of x mean of y
##  32.52331  28.93896
```

The $p$ value is exactly 0.01. Whether this is significant is up to the experimenter (you!)

(ix) Do now a non-parametric test for the same variables. Compare with your previous result and comment.

```
wilcox.test(mcs[gender=='m'],mcs[gender=='f'])
```

```
##
```

```
##  Wilcoxon rank sum test with continuity correction
##
## data:  mcs[gender == "m"] and mcs[gender == "f"]
## W = 21553, p-value = 0.01017
## alternative hypothesis: true location shift is not equal to 0
```

We get exactly the same result in this case.

(x) Repeat (viii) and (ix) for `pcs`. Comment on your findings.

```
t.test(pcs[gender=='m'],pcs[gender=='f'])
```

```
##
##  Welch Two Sample t-test
##
## data:  pcs[gender == "m"] and pcs[gender == "f"]
## t = 3.4889, df = 187.22, p-value = 0.0006046
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.725196 6.214562
## sample estimates:
## mean of x mean of y
##  48.98624  45.01636
```

```
wilcox.test(pcs[gender=='m'],pcs[gender=='f'])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pcs[gender == "m"] and pcs[gender == "f"]
## W = 22755, p-value = 0.0003364
## alternative hypothesis: true location shift is not equal to 0
```

Now the p-values are really small so we reject the null hypothesis with both tests.

# Exercise 4

Consider the data set

```
a = c(13.9, 14.5, 13.8, 16.6, 18.2, 20.2, 13.6, 16.3, 15.4, 12.3)
b = c(13.7, 14.6, 13.0, 16.2, 17.8, 21.0, 13.0, 16.9, 17.0, 12.1)
q2.df <- data.frame(before = a, after = b)
```

You can copy and paste the previous instructions to create `q2.df`.

These values represent the time taken for 10 student athletes to perform the same physical task, `a` or `before` are times before training while `b` or `after` corresponds to time after training. Training was supposed to improve the athletes performance at this task.

(i) Calculate the summary statistics for the two vectors, `before` and `after` and compare them. Do they seem different?

```
apply(q2.df,2,summary)
```

```
##         before  after
## Min.    12.300 12.100
## 1st Qu. 13.825 13.175
## Median  14.950 15.400
## Mean    15.480 15.530
```

```
## 3rd Qu. 16.525 16.975
## Max.    20.200 21.000
```

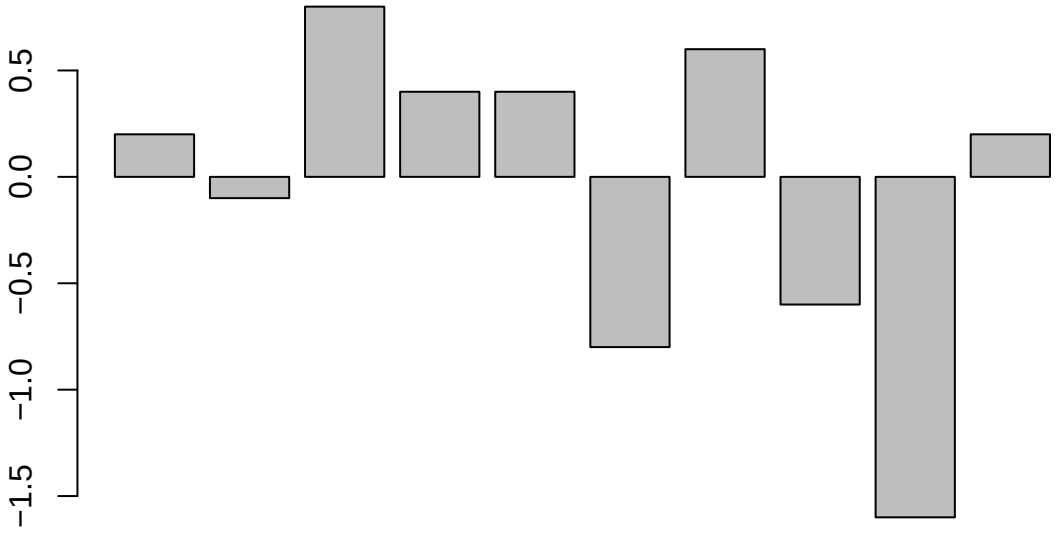
No, the values are similar.

(ii) Calculate the differences between the times before and after training. Do a barplot of your results and comment.

```
((q2.dif <- q2.df[,1] - q2.df[,2]))
```

```
##  [1]  0.2 -0.1  0.8  0.4  0.4 -0.8  0.6 -0.6 -1.6  0.2
```

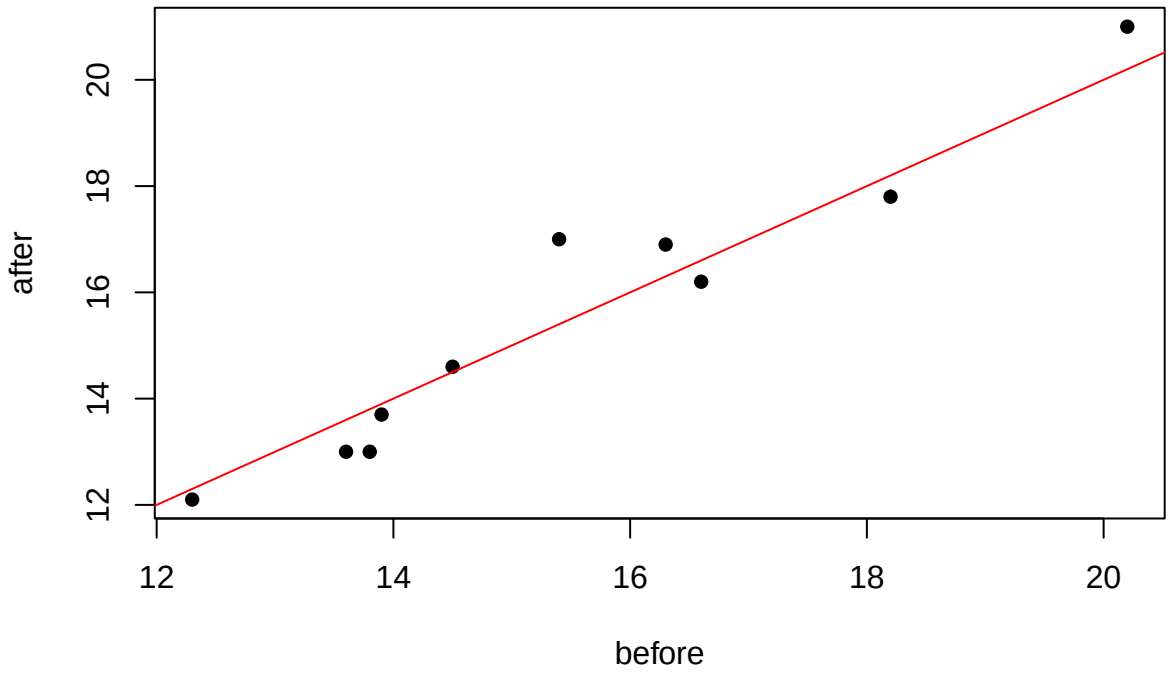
```
barplot(q2.dif)
```




If training has an effect on the task, one would expect the time taken to execute the task to change after training. In this case this does not seem to happen. Out of 10 results, 6 have positive differences and 4 have negative differences. The negative differences look bigger.

(iii) Do a scatterplot of `before` in the $x$ axis against `after` in the $y$ axis. Add the line $y = x$. What does this graph show? Can you conclude anything about the effectiveness of the training method?

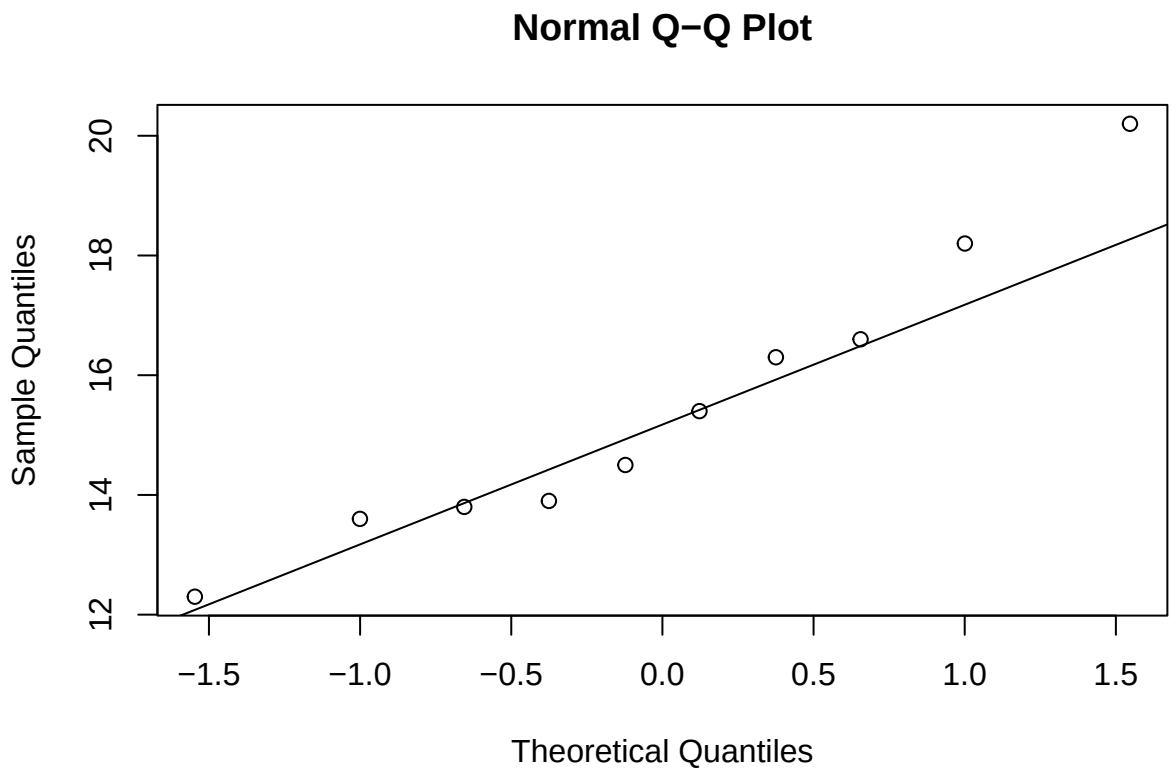
```
plot(q2.df, pch=16)
abline(coef=c(0,1), col='red')
```



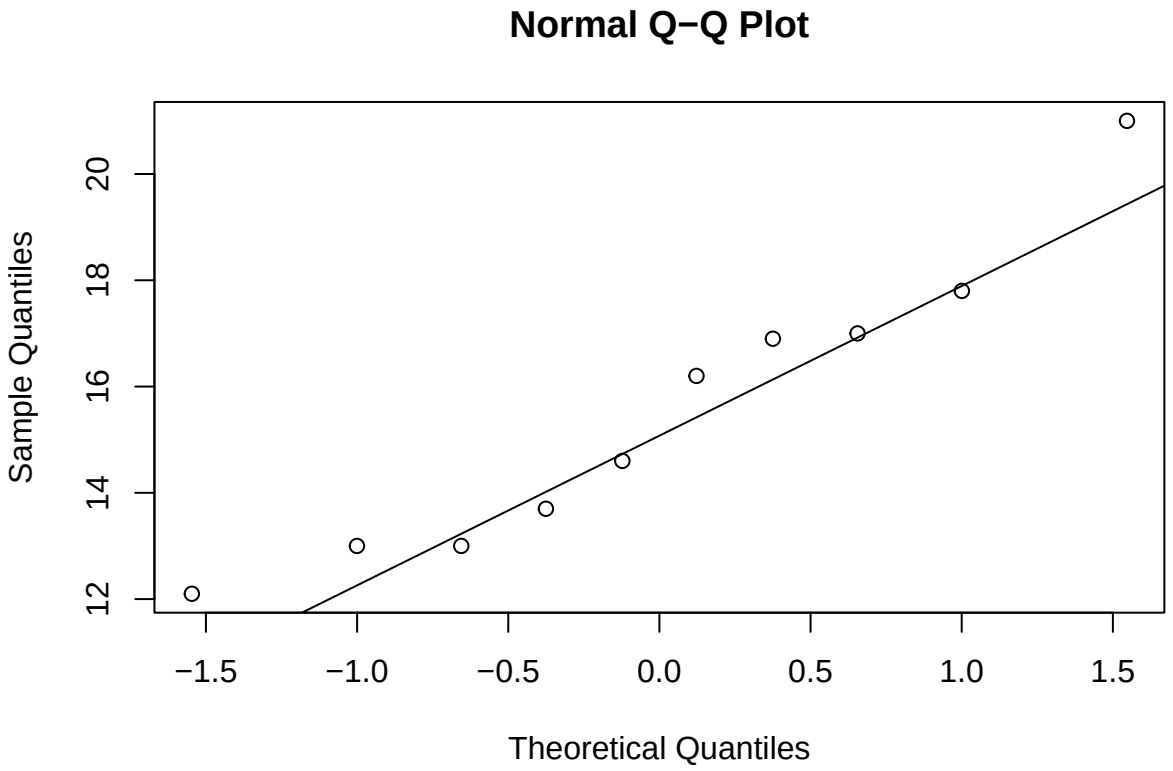

The scatterplot confirms the previous comment. If a change had been observed, most points should be on one side of the line $y = x$. As it is, they are close to the line and 6 are below while 4 are above. There does not seem to be significant changes after training.

(iv) Do normal quantile plots for `before`, `after` and the difference `after-before`. Comment on your results.

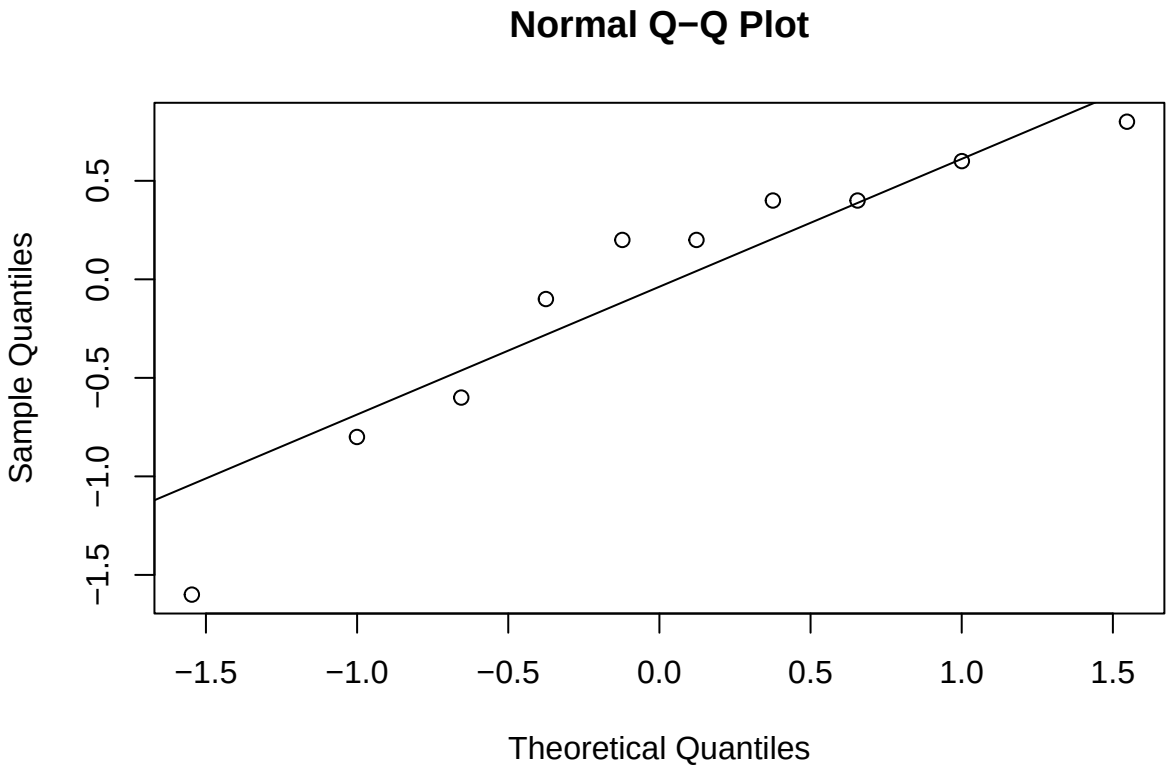
```
qqnorm(a); qqline(a)
```



```
qqnorm(b); qqline(b)
```




```
qqnorm(a-b); qqline(a-b)
```




Taking into account that there are only 10 points in each graph, they seem to support the normality of the

data.

(v) You are asked to compare the average time taken by the students before and after training to assess if there has been an improvement. Which test would you perform? What are the underlying assumptions? Why do you think they are satisfied? What hypotheses are you testing? Carry out the test and comment on the results.

Since we have paired data (the performance of each student athlete is measured before and after training) one should do a paired t-test. The basic assumption is that the data follow a Gaussian distribution, which seems justified by the quantile plots. If you understand improvement to mean a faster time, you should carry out a one sided test, otherwise you should do a two-sided test.

```
t.test(a,b,'greater', paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  a and b
## t = -0.21331, df = 9, p-value = 0.5821
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.4796859        Inf
## sample estimates:
## mean of the differences
##                   -0.05
```

```
t.test(a,b, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  a and b
## t = -0.21331, df = 9, p-value = 0.8358
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5802549  0.4802549
## sample estimates:
## mean of the differences
##                   -0.05
```

In both cases the p-value is too big to reject the null hypothesis of equal means.

(vi) The Wilcoxon non-parametric test can be used with paired and non-paired data. Which option would you choose for this data set? Carry out this test and comment on your results. (NB: If you want to do a Wilcoxon paired test you need to add the option `paired = TRUE` when you call the test).

One should use a paired test, for the reason explained previously.

```
wilcox.test(a,b,'greater',paired = TRUE)
```

```
## Warning in wilcox.test.default(a, b, "greater", paired = TRUE): cannot compute
## exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  a and b
## V = 29.5, p-value = 0.4392
## alternative hypothesis: true location shift is greater than 0
```

```r
wilcox.test(a,b, paired = TRUE)
```

```
## Warning in wilcox.test.default(a, b, paired = TRUE): cannot compute exact p-
## value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  a and b
## V = 29.5, p-value = 0.8783
## alternative hypothesis: true location shift is not equal to 0
```

This test also has a large p-value, so we do not reject the null hypothesis of equal means.