Exam – R part

Stéphanie Allassonnière, Geneviève Robin, Elodie Vernet and Zoltán Szabó

November 13 2018

General instructions:

- Working in team is very useful, however this time please solve the following questions on your own.
- Please write an Rmarkdown report which answers the following questions and save it to pdf. Upload both files (call them **Firstname_Lastname.Rmd** and **Firstname_Lastname.pdf**) to moodle by 10pm (Tuesday November 13 2018).
- Include your codes and graphs in the report.

Exercise to solve:

A cholesterol level above $240 mg/dl$ of blood is a risk factor for heart disease. A lab has changed the flavour of a medicine used to diminish the cholesterol level. A study to test the effectiveness of the new version was conducted. Fifty people with cholesterol level between 240 and 260 were assigned at random to each of two treatments groups. Group A received the standard medicine and group B received the new medicine. The observed cholesterol level after the treatment are

- for group A: 275, 255, 247, 206, 245, 196, 229, 218, 238, 224, 217, 205, 252, 194, 203, 192, 232, 239, 223, 239, 217, 241, 230, 216, 227,
- for group B: 238, 216, 214, 247, 250, 223, 207, 254, 228, 219, 223, 247, 228, 207, 204, 225, 201, 253, 216, 223, 219, 224, 225, 228, 225.

The cholesterol levels of group A and B are independent and i.i.d. Gaussian random variables distributed as $\mathcal{N}(m_A, \sigma_A)$ and $\mathcal{N}(m_B, \sigma_B)$ respectively.

We don't expect to see a difference between these two medicines. We assume that $\sigma := \sigma_A = \sigma_B$ and we want to test $H_0 : \ m_A = m_B$ against $H_1 : \ m_A \neq m_B$.

1. Create a dataframe which contains the data.

```
A <- c(275, 255, 247, 206, 245, 196, 229, 218, 238, 224, 217, 205, 252, 194, 203, 192, 232,
 239, 223, 239, 217, 241, 230, 216, 227)
B <- c(238, 216, 214, 247, 250, 223, 207, 254, 228, 219, 223, 247, 228, 207, 204, 225, 201,
 253, 216, 223, 219, 224, 225, 228, 225)
data <- data.frame(A=sort(A),B=sort(B))
data
```

```
##      A   B
## 1  192 201
## 2  194 204
## 3  196 207
## 4  203 207
## 5  205 214
## 6  206 216
## 7  216 216
## 8  217 219
## 9  217 219
## 10 218 223
## 11 223 223
## 12 224 223
## 13 227 224
## 14 229 225
## 15 230 225
## 16 232 225
## 17 238 228
## 18 239 228
## 19 239 228
## 20 241 238
## 21 245 247
## 22 247 247
## 23 252 250
## 24 255 253
## 25 275 254
```

```
summary(data)
```

```
##        A               B
##  Min.   :192.0   Min.   :201.0
##  1st Qu.:216.0   1st Qu.:216.0
##  Median :227.0   Median :224.0
##  Mean   :226.4   Mean   :225.8
##  3rd Qu.:239.0   3rd Qu.:228.0
##  Max.   :275.0   Max.   :254.0
```

2. Check visually that the distribution of the cholesterol levels of group A and B can be modeled as Gaussian.

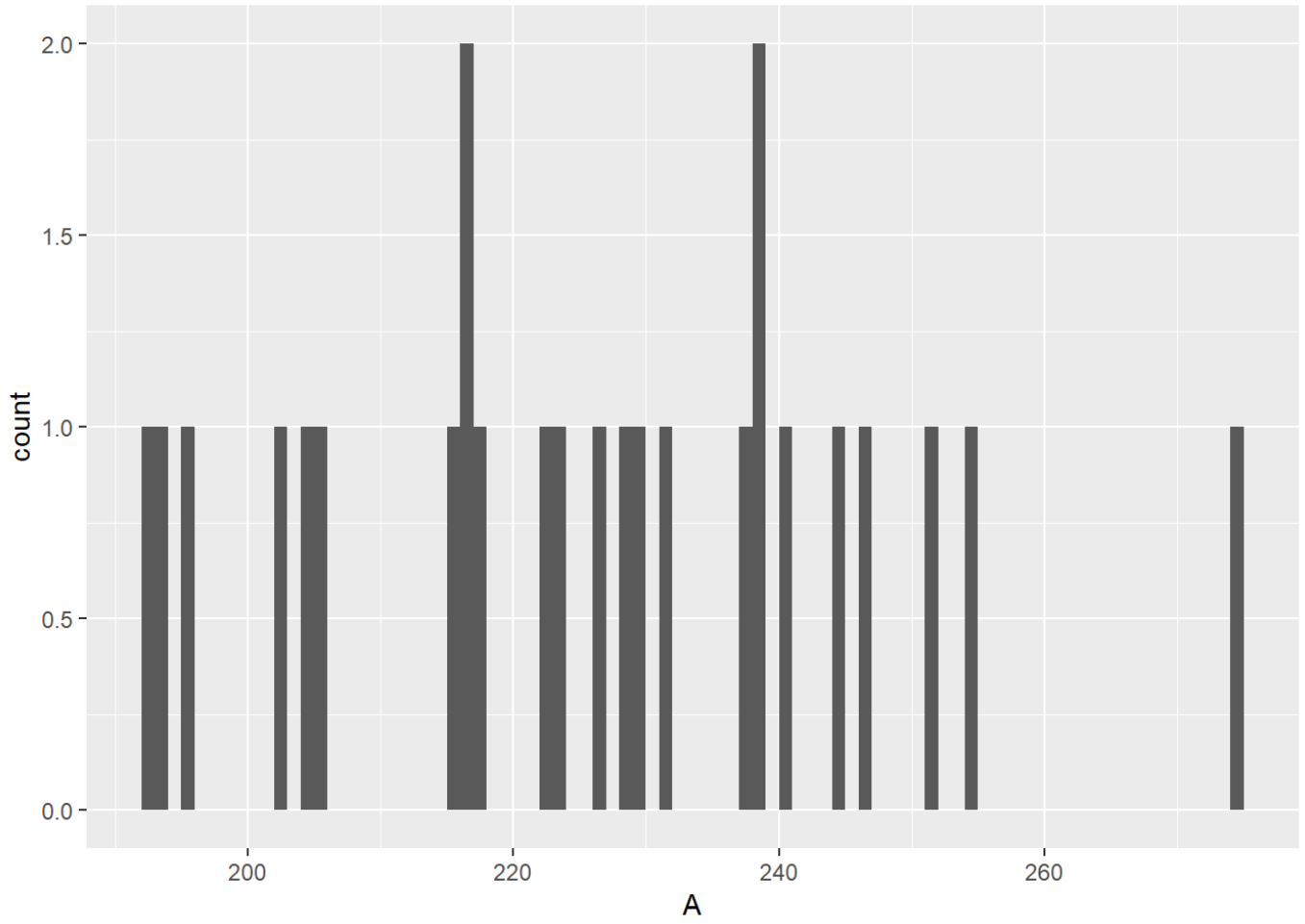
```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

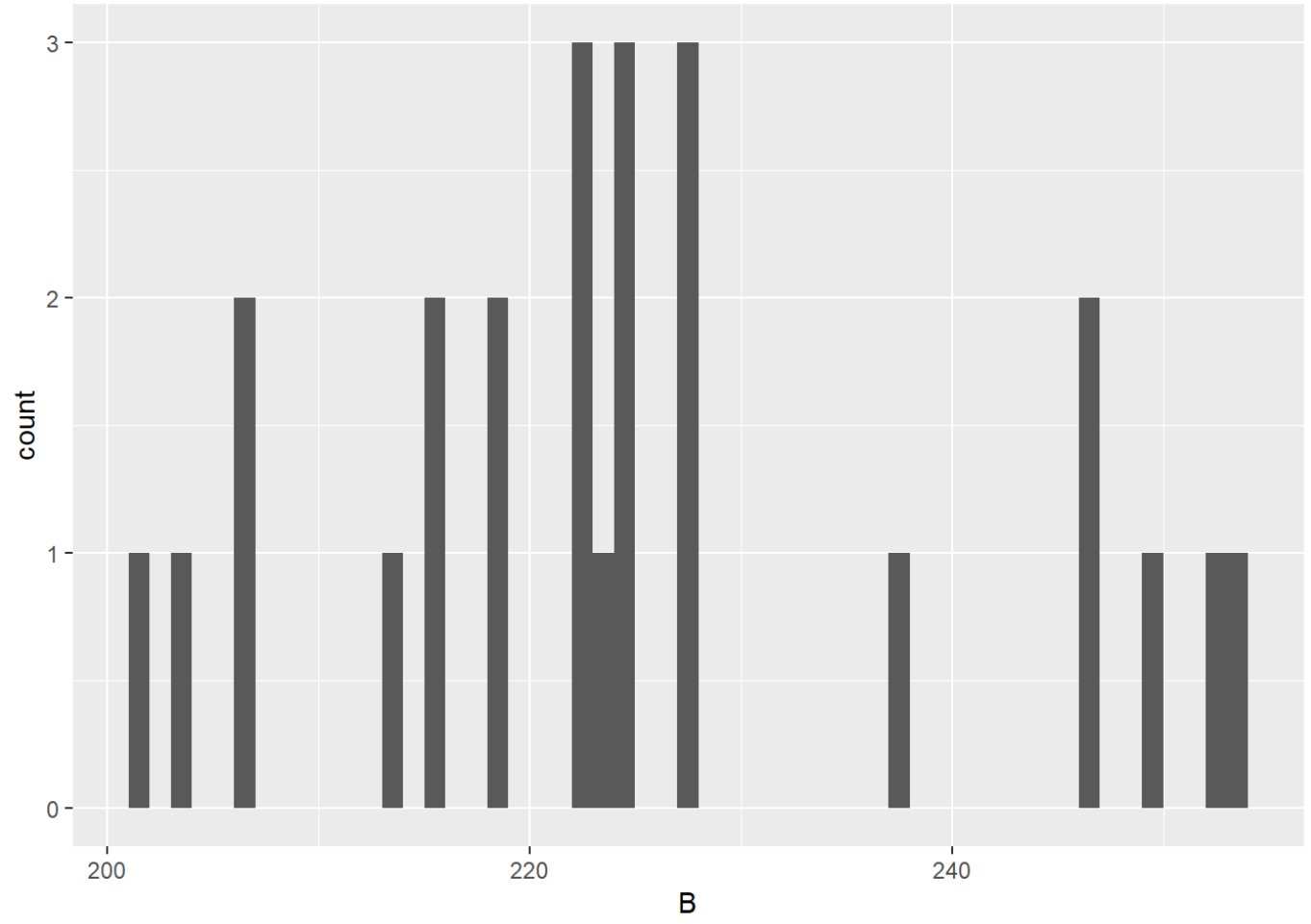
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

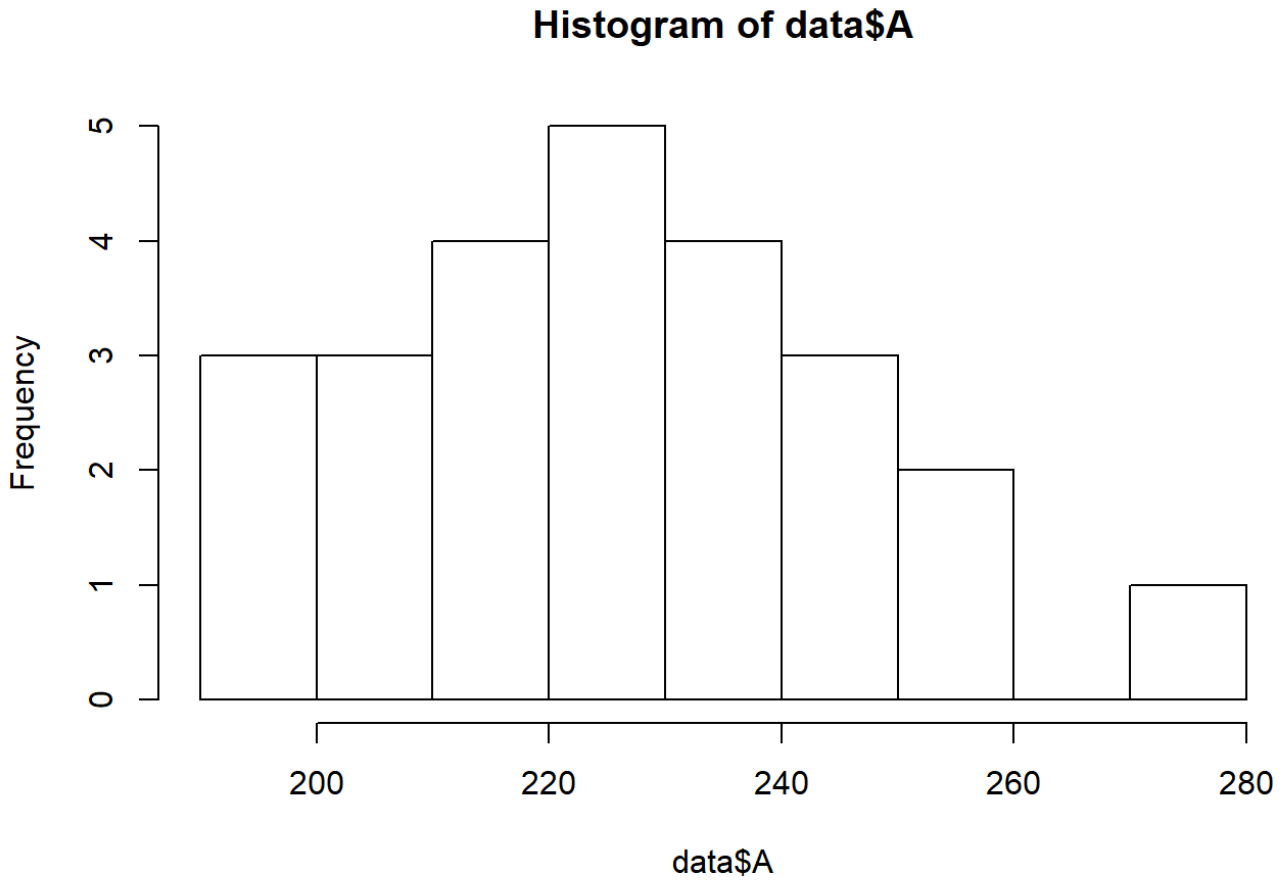
```
library(ggplot2)
A <- data.frame(A)
B <- data.frame(B)
ggplot(A)+geom_histogram(aes(x=A),boundary=0,binwidth=1)
```



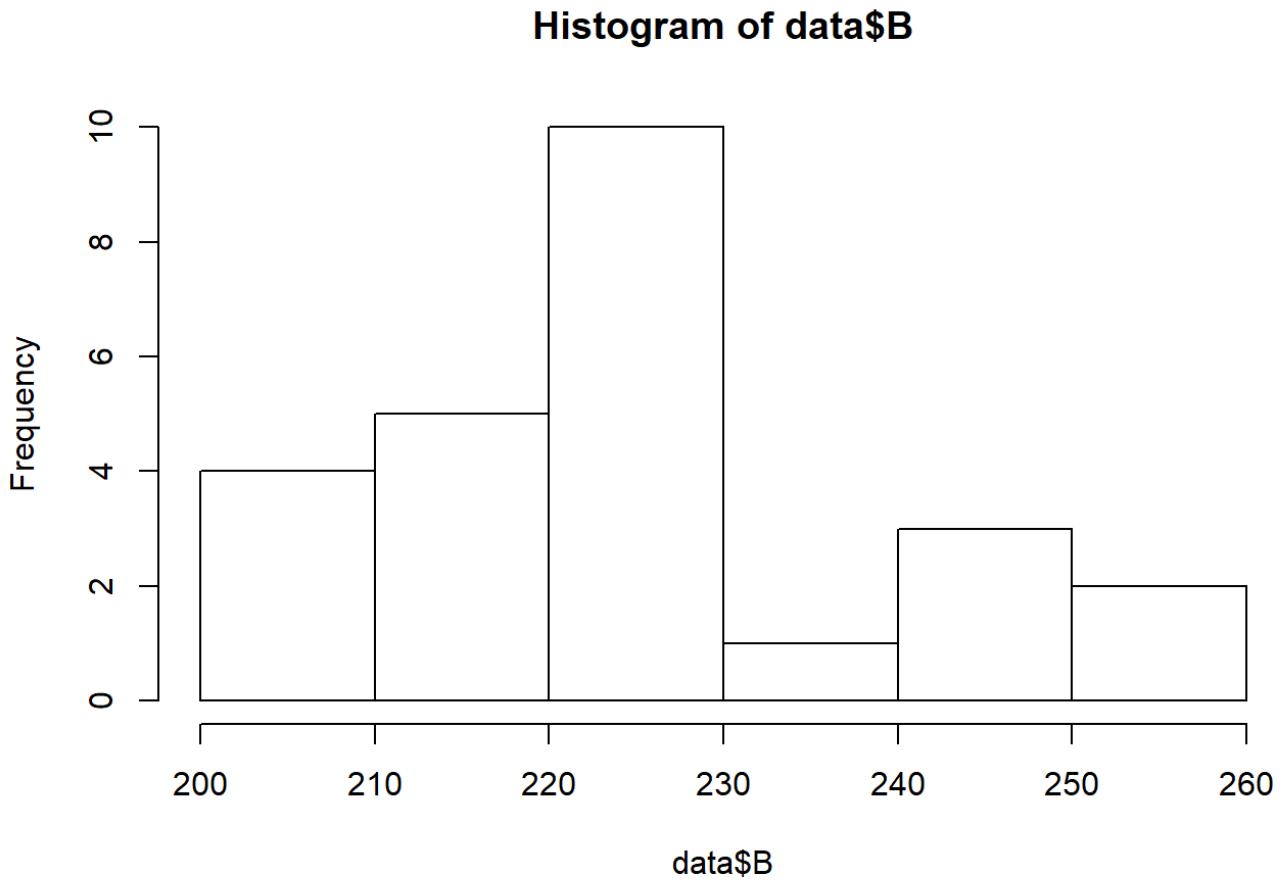

```
ggplot(B)+geom_histogram(aes(x=B),boundary=0,binwidth=1)
```


3
2
1
0
count
200
220
240
B

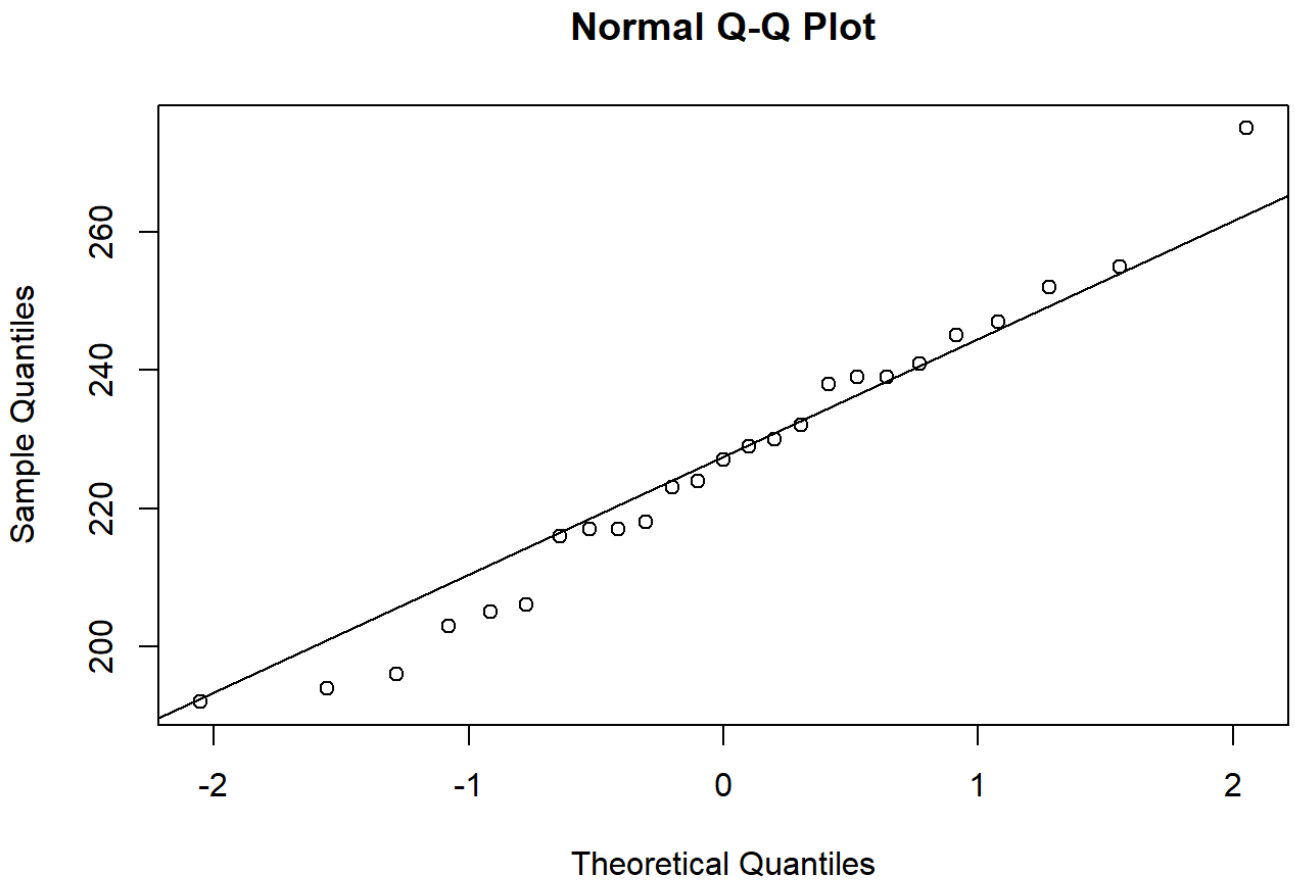
```
hist(data$A)
```


Histogram of data$A
Frequency
0
1
2
3
4
5
200
220
240
260
280
data$A

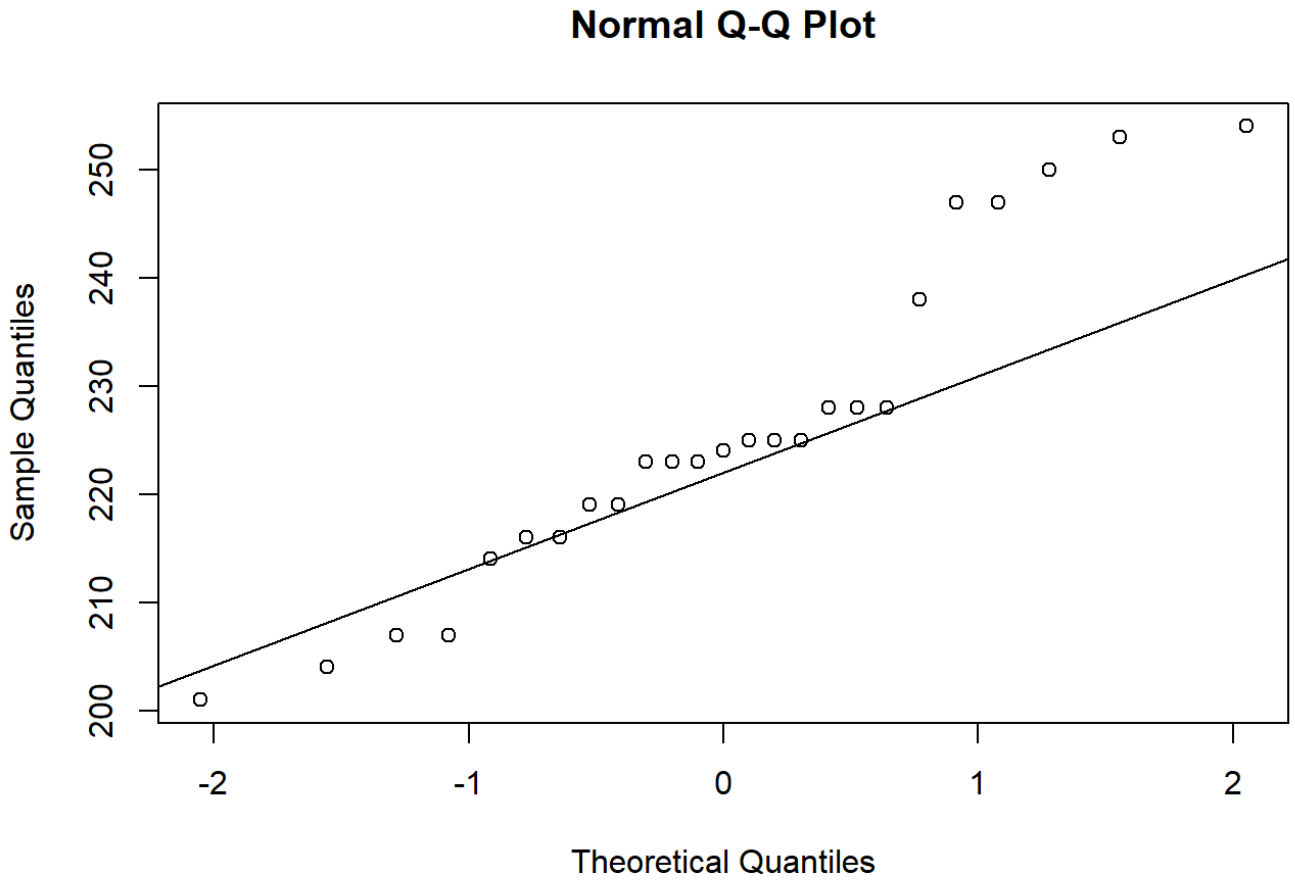
```
hist(data$B)
```




```
qqnorm(data$A)
qqline(data$A)
```




```
qqnorm(data$B)
qqline(data$B)
```



With these histograms and qqplot, we can say that the two datasets can be modeled as Gaussian.

3. Propose a test at level α for $H_0 : m_A = m_B$ against $H_0 : m_A \neq m_B$. Give the statistical model, the statistic that you use, its distribution under H_0, the test function and its p-value. Hint: You could look at PC8 as a source of inspiration.

ANSWER: Assuming $\overline{Y_n}$ is the mean of the first dataset, and $\overline{X_n}$ the mean of the second one. The statistical model is $\left((\mathbb{R}^{2n}, \mathcal{B}(\mathbb{R}^{2n}), \{\mathcal{N}^{\otimes 2n}((m_A)(m_B), \sigma^2 I_n); m_A, m_B \in \mathbb{R}\}\right)$.

The test statistic can be $Z = \overline{Y_n} - \overline{X_n}$ Its distribution is $Z \sim \mathcal{N}(m_A - m_B, \frac{2\sigma^2}{n})$

Under H0, its distribution is $Z \sim \mathcal{N}(0, \frac{2\sigma^2}{n})$ Using the student-gosset theorem, we have (as the two dataset are independant and the variance is unknown) : $K = \sqrt{n}\frac{\overline{Y_n} - \overline{X_n}}{\sqrt{\sigma_X^2 + \sigma_Y^2}} \sim t_{2(n-1)}$ with

$\sigma_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})$ and $\sigma_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y_n})$

The test function is then:

$$\phi(X_1, \ldots, X_n) = \begin{cases} 0 & \text{if } K \in \left[-t_{1-\alpha/2}(n-1)\frac{Sn^2}{\sqrt{n}}, t_{1-\alpha/2}(n-1)\frac{Sn^2}{\sqrt{n}}\right] \\ 1 & \text{otherwise} \end{cases}.$$

The p-value is then
$\hat{\rho}(x_1, \ldots, x_n, y_1, \ldots, y_n) = \mathbb{P}_{H_0}(k_{obs} > K) + \mathbb{P}_{H_0}(-k_{obs} > K) = 2(1 - \mathcal{F}_{t_{2(n-1)}}(k_{obs}))$

4. Write an R function **test** with inputs $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, α and output of a vector of size five containing the empirical mean of x, the empirical mean of y, the observed value of the statistics, the critical value of the rejection region, the observed value of the test function (0 or 1) and the p-value.

```
test <- function(x, y , a ){
meanx <- mean(x)
meany <- mean(y)

tot <-length(x)+ length(y)
obsvalue = meanx - meany

criticalvalue <- sd(x-y)*qt(c(a/2,1-a/2),df=tot-2)/sqrt(length(x))

testresult <- as.numeric(abs(obsvalue)>criticalvalue[2])

pvalue <- t.test(x,y, var.equal = TRUE)$p.value

std <- sqrt(tot)*abs(meanx-meany)/sqrt(var(x)+var(y))

pvalue <- (1-pt(std, tot-2, log = FALSE))*2

result <- list(meanx, meany, obsvalue, criticalvalue, testresult, pvalue)
return(result)
}

test(data$A,data$B, 0.05)
```

```
## [[1]]
## [1] 226.4
##
## [[2]]
## [1] 225.76
##
## [[3]]
## [1] 0.64
##
## [[4]]
## [1] -3.099422  3.099422
##
## [[5]]
## [1] 0
##
## [[6]]
## [1] 0.8603306
```

For the critical values, as the region is bounded, the two critical values are returned.

5. Check that significance level of your test through $I = 1000$ simulations when $n = 25$, $\alpha = 0.05$, $m_A = m_B = 220$ and $\sigma = 20$.

```
results <- 0
for (i in 1:1000){
  xa <- rnorm(n=25,mean=220,sd=20)
  xb <- rnorm(n=25,mean=220,sd=20)
  results[i] <- test(xa,xb,0.05)[[5]]
}
sum(results)/1000
```

```
## [1] 0.045
```

The significance seems to be close to alpha = 0.05 when running the function 1000 times.

6. Approximate the probability that your test appropriately rejects H_0 when $n = 25$, $\alpha = 0.05$, $m_A = 220$, $m_B = 230$ and $\sigma = 20$ using the Monte-Carlo method (using $I = 1000$ simulations).

```
mte <- 0
for(i in 1:1000){
  uniform1 <- runif(25)
  uniform2 <- runif(25)

  x<- sqrt(-2*log(uniform2))*cos(2*pi*uniform1)
  y<- sqrt(-2*log(uniform2))*sin(2*pi*uniform1)
  mte[i] <- (test(x,y,0.05)[[5]])
}
sum(mte)/1000
```

```
## [1] 0.053
```

With the monte carlo methos, we can say again that that the test appropriately rejects H_0 with the correct significance alpha.

7. Apply your test to the cholesterol data.

```
test(data$A,data$B,0.05)
```

```
## [[1]]
## [1] 226.4
##
## [[2]]
## [1] 225.76
##
## [[3]]
## [1] 0.64
##
## [[4]]
## [1] -3.099422  3.099422
##
## [[5]]
## [1] 0
##
## [[6]]
## [1] 0.8603306
```

The test does not reject H0, the two samples follow the same law.