

Assignment 1

Tim But, Mehmet Bedirhan Gursoy, Vincentas Ryliskis, group 035

6 November 2023

Task 1

a)

Suppose the following:

- $P(D)$ is the probability of a positive cancer diagnosis
- $P(C)$ is the probability of an individual having cancer

Then, given the information given in exercise 1.3:

$$P(D|C) = 0.95$$

$$P(D|C') = 0.05$$

$$P(C) = 0.004$$

To calculate the probability of a random person being given a positive cancer diagnosis, the following must be calculated:

$$\begin{aligned} P(D) &= P(D|C) * P(C) + P(D|C') * P(C') \\ &= 0.95 * 0.004 + 0.05 * 0.996 \\ &= 0.0536 \end{aligned}$$

This differs from probability we are asked to calculate in the Exercise 1.3, as there we are asked to calculate the probability of someone having cancer given a cancer diagnosis, or: $P(C|D)$. These probabilities refer to different events, as the latter already assumes $P(D)$ as being true.

b)

As mentioned above, we need to calculate $P(C|D)$. We can calculate this value using Bayes' Theorem

$$P(C|D) = \frac{P(D|C) * P(C)}{P(D)} = \frac{0.95 * 0.004}{0.0536} \approx 0.071$$

c)

The two events that a person has cancer and that the test is positive are dependent. This is shown by the fact that probabilities $P(C|D)$ and $P(C)$ are different, as shown below (based on calculations done in parts a) and b)):

- $P(C|D) = 0.071$

- $p(C) = 0.004$

The fact that $P(C|D)$ is larger than $P(C)$ shows that a cancer diagnosis leads to an increased risk in cancer, when compared to the population average.

Task 2

a)

b)

c)

d)

e)

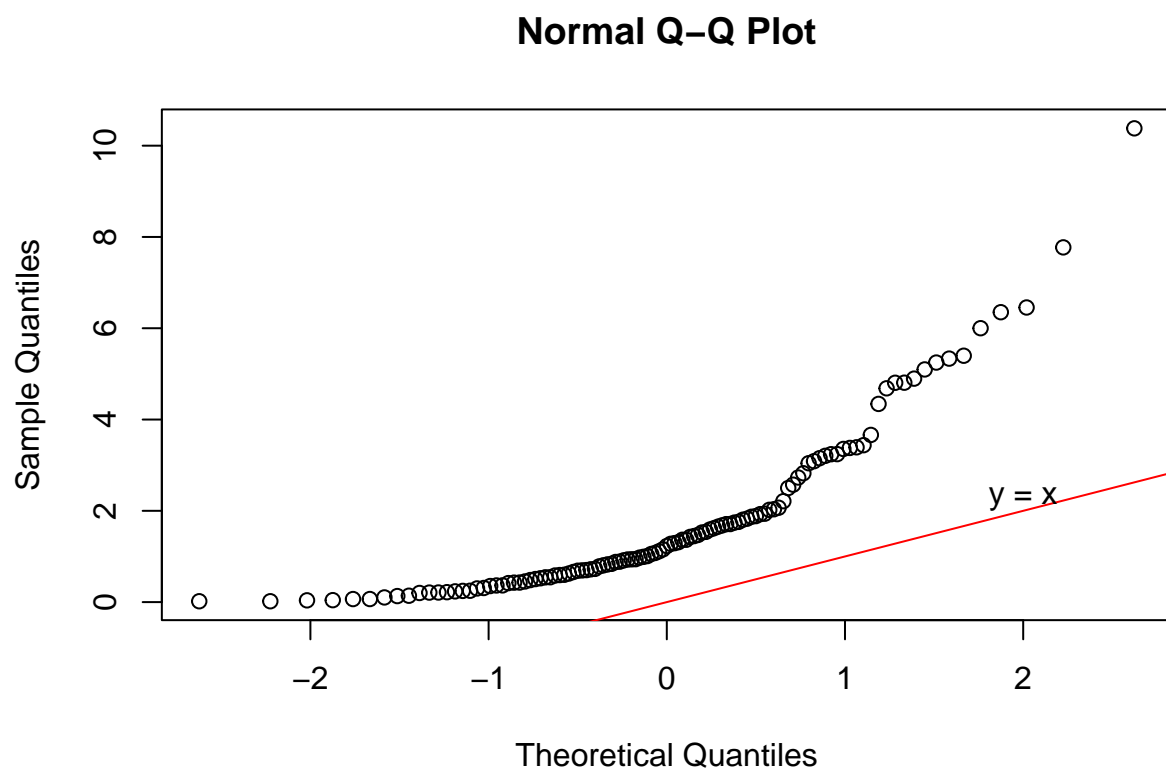
Task 3

a)

```
# Setting a seed
set.seed(123)

cn = 115
cdf=2
ci = sequence(cn)
cx = rchisq(ci, df=cdf)

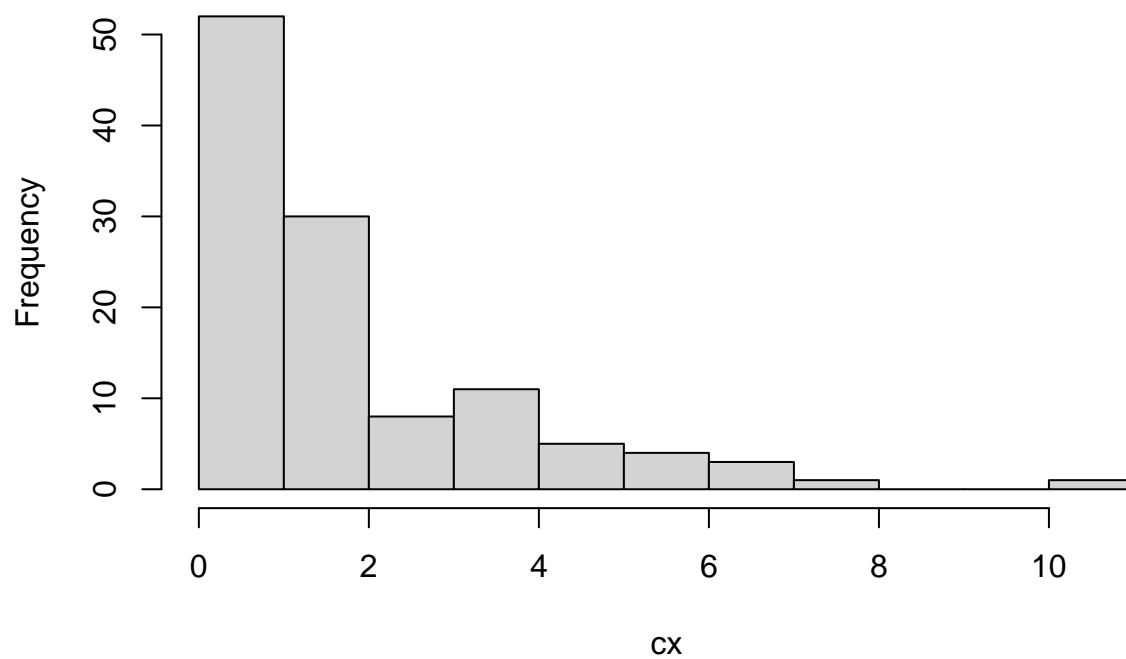
qqnorm(cx)
abline(0, 1, col = 'red')
text(2, 2.25, "y = x")
```



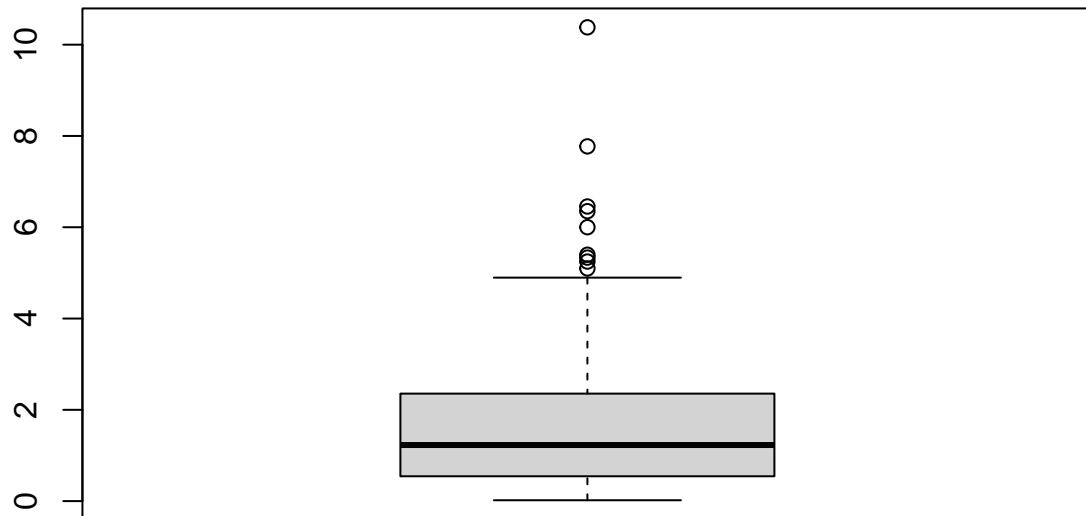
(i)

```
hist(cx)
```

Histogram of cx

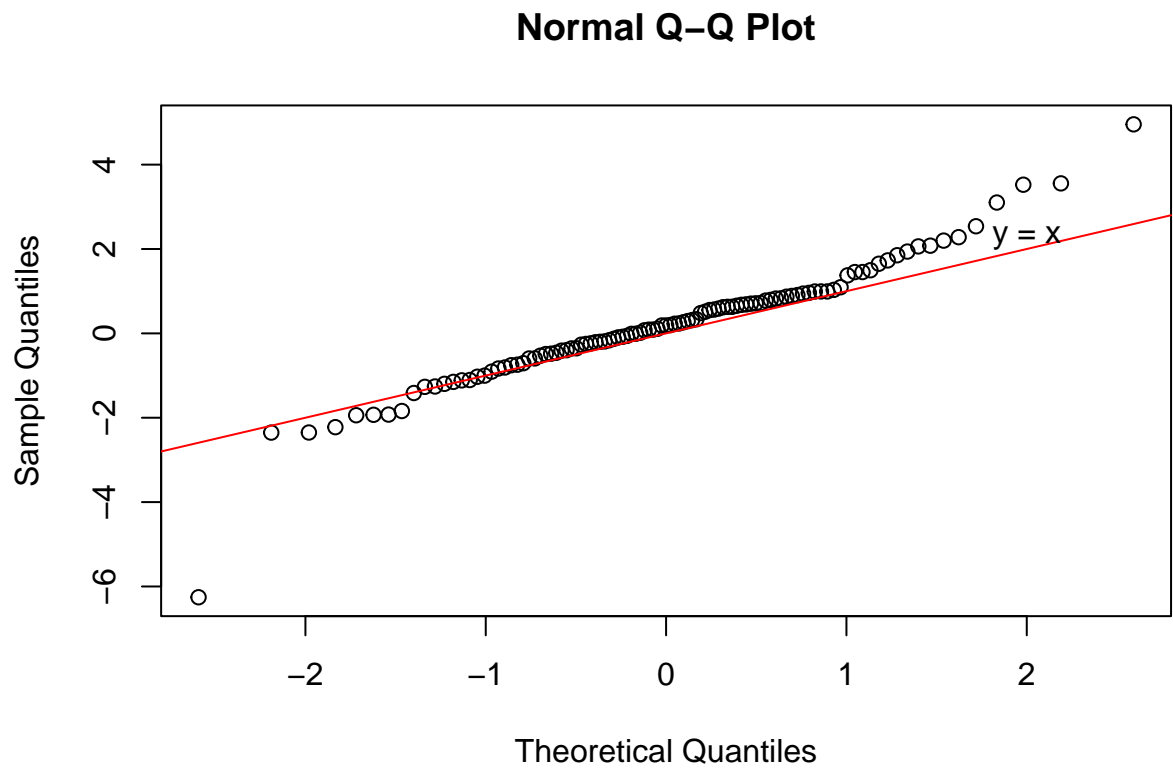


```
boxplot(cx)
```



```
tn = 105
tdf=4
ti = sequence(tn)
tx = rt(ti, df=tdf)

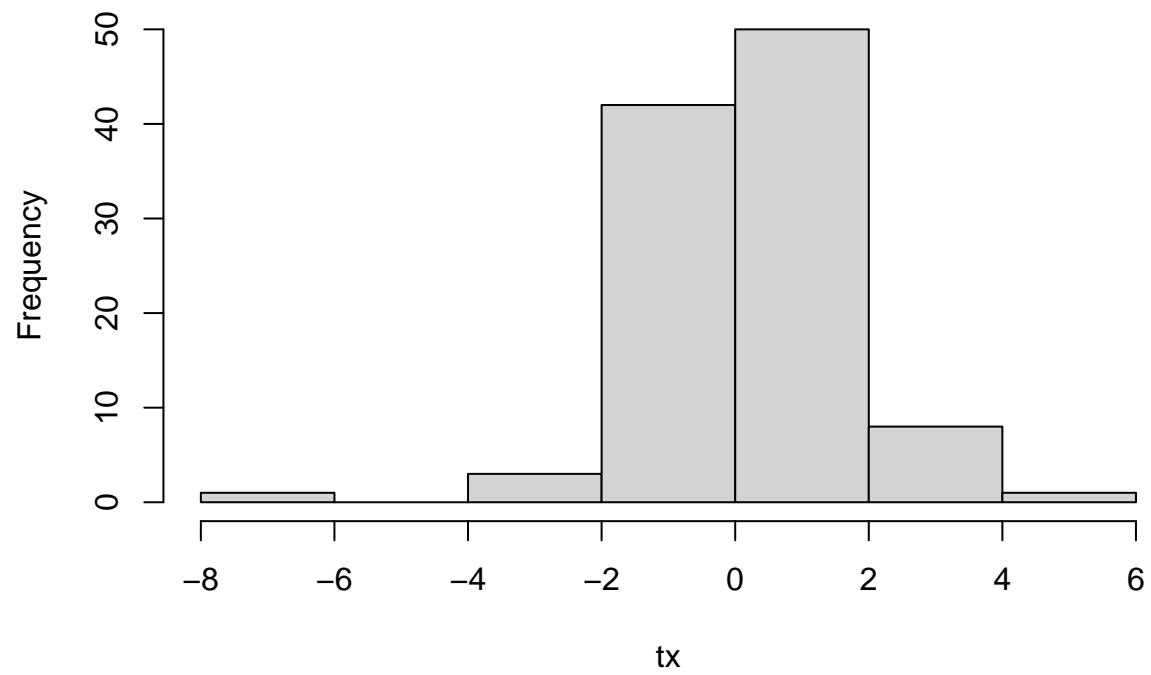
qqnorm(tx)
abline(0, 1, col = 'red')
text(2, 2.25, "y = x")
```



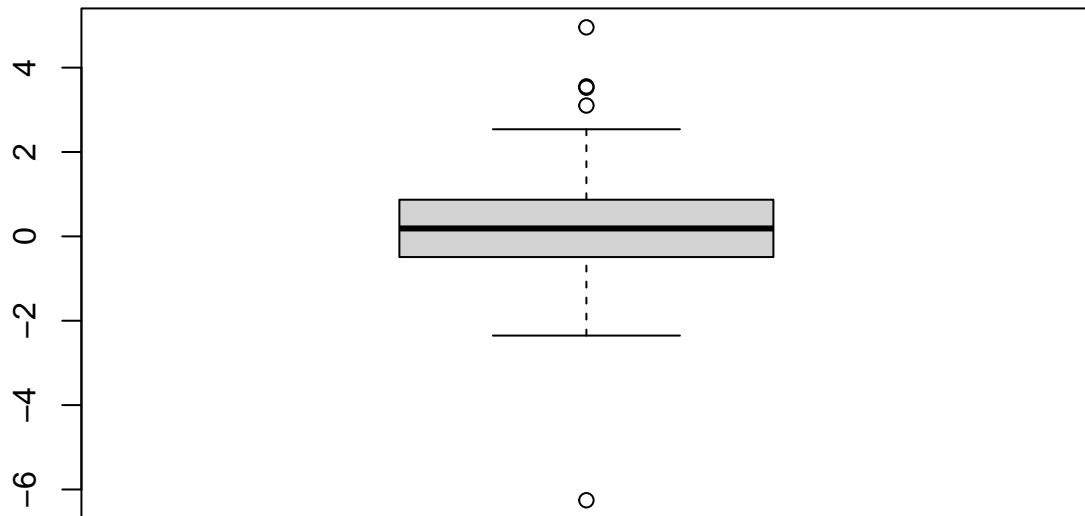
(ii)

```
hist(tx)
```

Histogram of tx



```
boxplot(tx)
```



b) ##### (i)

```
smallerThan3 = pnorm(3)
print(smallerThan3)
```

```
## [1] 0.9986501
```

```
biggerthanM0.5 = 1-pnorm(-0.5)
print(biggerthanM0.5)
```

```
## [1] 0.6914625
```

```
betweenM1and2 = pnorm(2) - pnorm(-1)
print(betweenM1and2)
```

```
## [1] 0.8185946
```

```
variance = 4
sd = sqrt(variance)
```



```
smallerThan3 = pnorm(3, mean = 3, sd = sd)
print(smallerThan3)
```

(ii)

```
## [1] 0.5
```

```
biggerthanM0.5 = 1-pnorm(-0.5, mean = 3, sd = sd)
print(biggerthanM0.5)
```

```
## [1] 0.9599408
```

```
betweenM1and2 = pnorm(2, mean = 3, sd = sd) - pnorm(-1, mean = 3, sd = sd)
print(betweenM1and2)
```

```
## [1] 0.2857874
```

```
NFpercSmaller = qnorm(0.95, mean = 3, sd = sd)
print(NFpercSmaller)
```

```
## [1] 6.289707
```

```
mean = -1
sd = 5

samples = rnorm(1000)

corrected_samples = (samples * sd) + mean

sample_mean = mean(corrected_samples)
print(sample_mean)
```

(iii)

```
## [1] -0.9403835
```

```
sample_sd = sd(corrected_samples)
print(sample_sd)
```

```
## [1] 5.005643
```

```

verification_function <- function(min = -Inf, max = Inf){

  sample_100 = rnorm(100)
  sample_100k = rnorm(100000)

  prob_sample_100 = mean(sample_100 > min & sample_100 < max)
  prob_sample_100k = mean(sample_100k > min & sample_100k < max)

  cat("Probability (Sample 100):", prob_sample_100, "\n")
  cat("Probability (Sample 100k):", prob_sample_100k, "\n")
}

verification_function(max = 3)

```

(iv)

```

## Probability (Sample 100): 1
## Probability (Sample 100k): 0.99868

```

```

verification_function(min = -0.5)

```

```

## Probability (Sample 100): 0.66
## Probability (Sample 100k): 0.69379

```

```

verification_function(min = -1, max = 2)

```

```

## Probability (Sample 100): 0.8
## Probability (Sample 100k): 0.81713

```

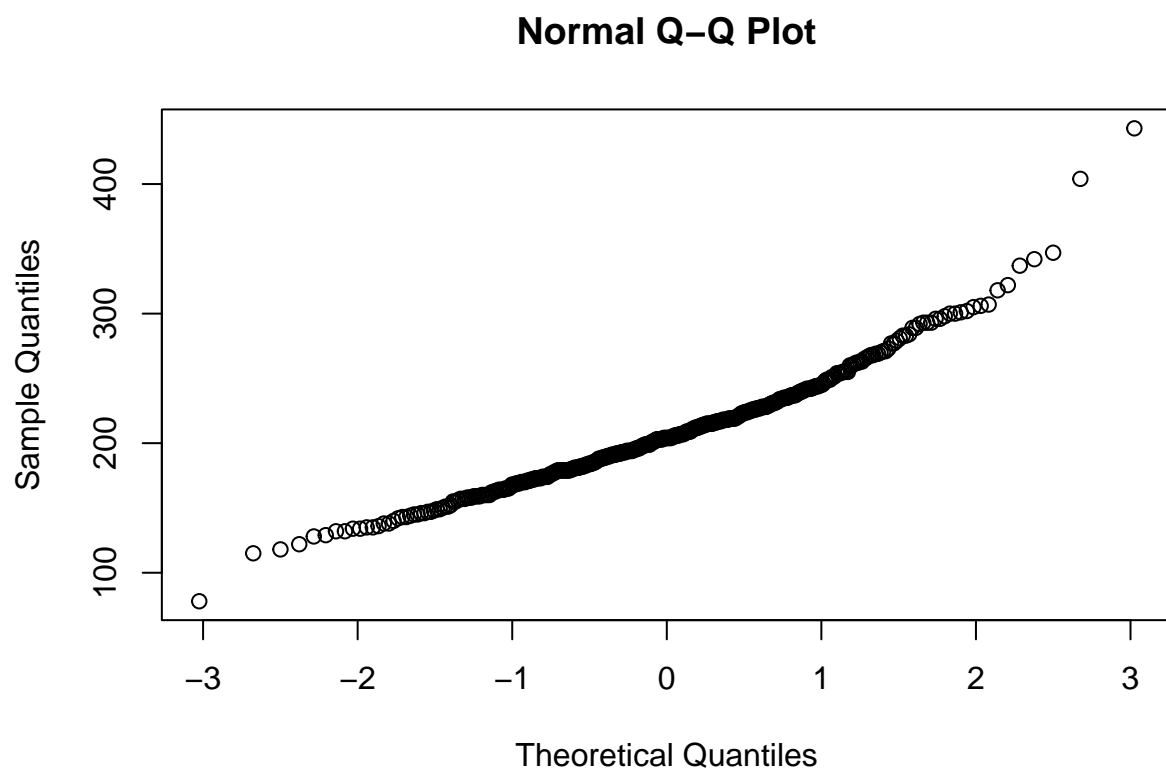
c)

```

diabetes <- read.csv("diabetes.csv")
chol = diabetes$chol

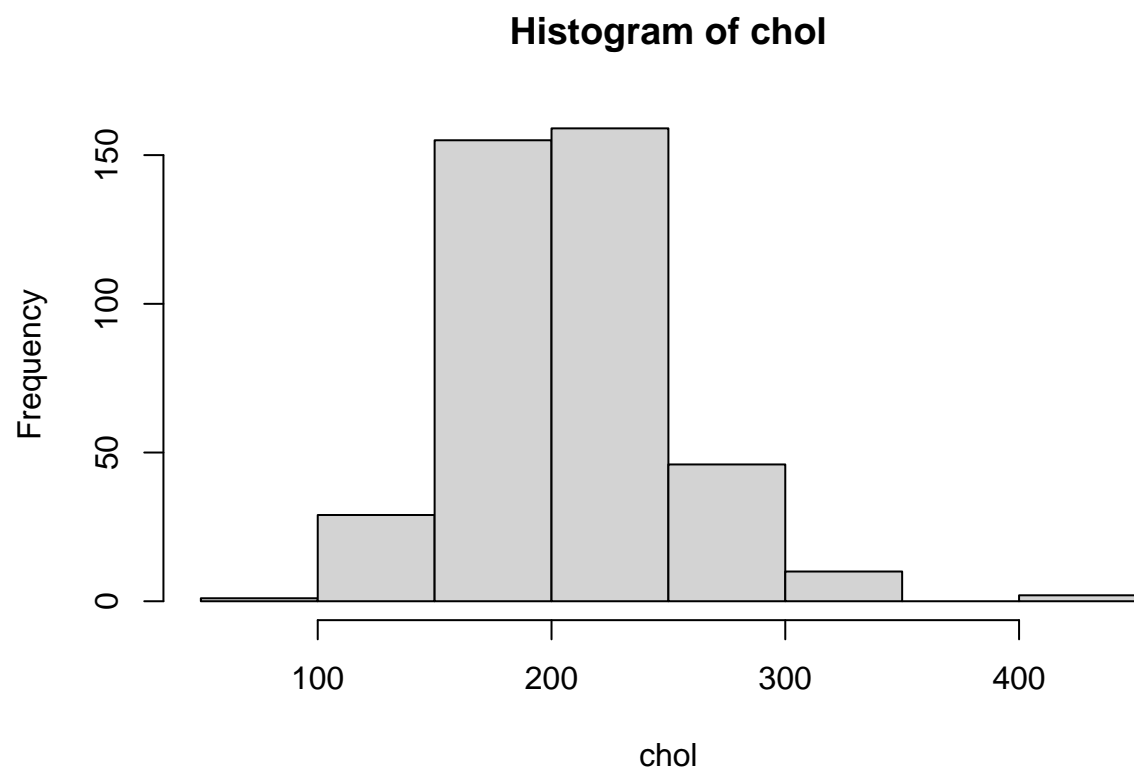
qqnorm(chol)

```

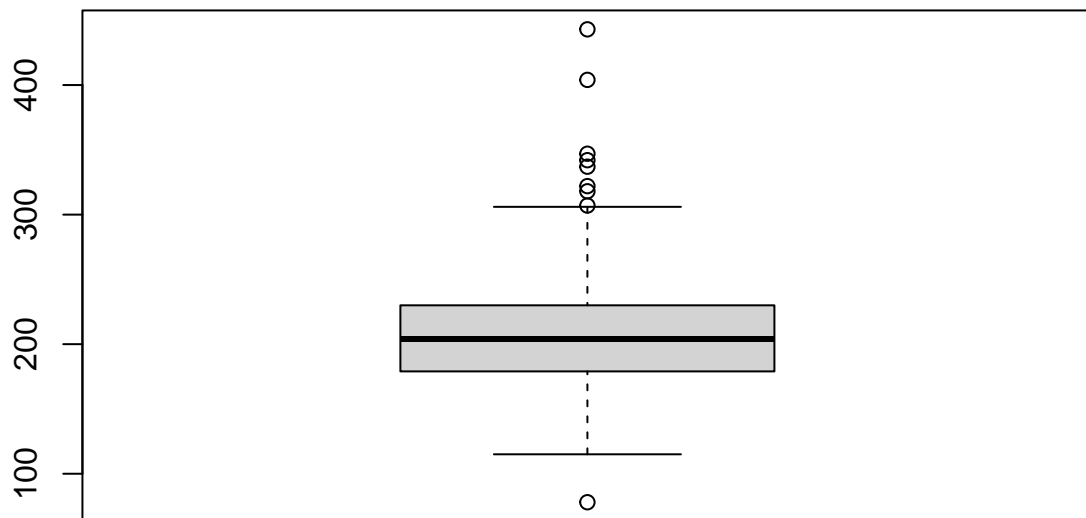


(i)

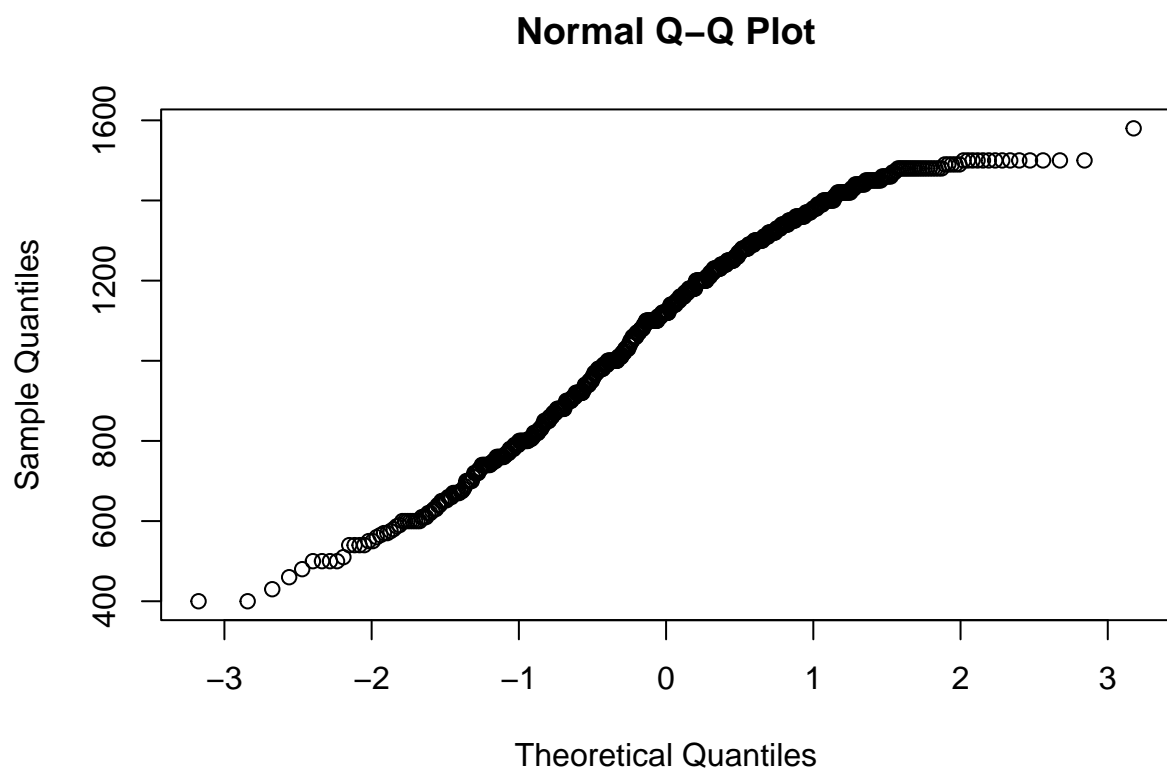
```
hist(chol)
```



```
boxplot(chol)
```



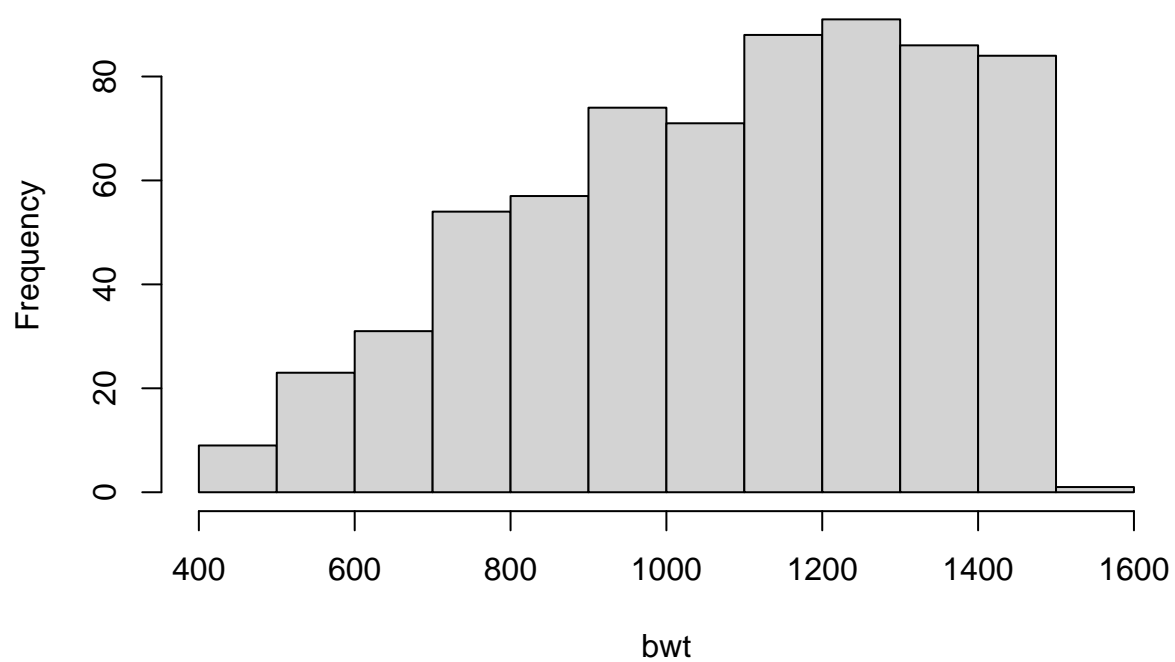
```
vlbw <- read.csv("vlbw.csv")  
bwt = vlbw$bwt  
  
qqnorm(bwt)
```



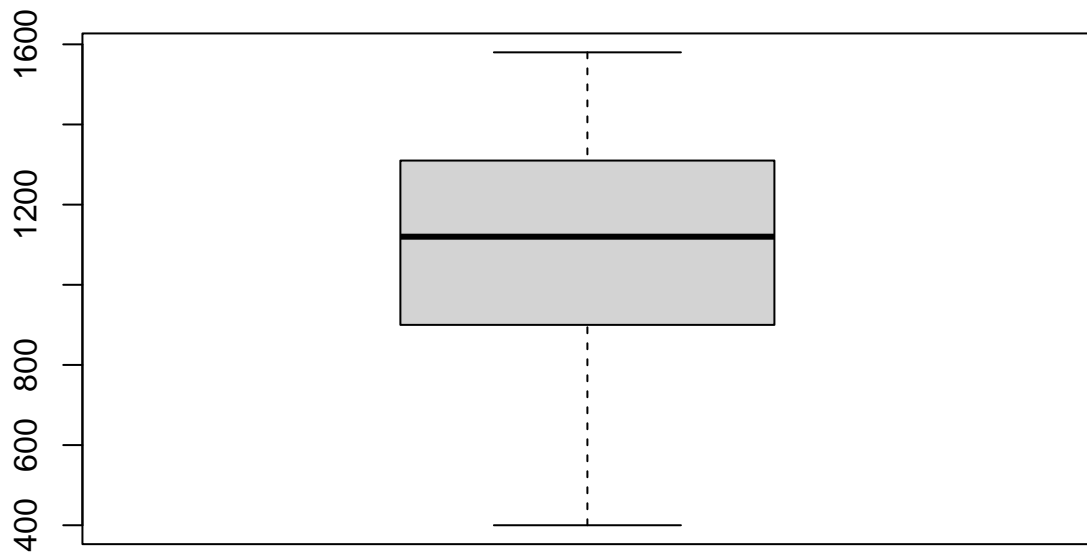
(ii)

```
hist(bwt)
```

Histogram of bwt



```
boxplot(bwt)
```



Task 4 ## a) ## b) ## c) ## d)