Exercises I2

Exercise 2.1

i) $S^2 \leq \frac{1}{n-1} \sum_i^n (X_i - a)^2$ for any $a \in \mathbb{R}$

$S^2 = \frac{1}{n-1} \sum_i^n (X_i - \overline{X})^2 \leq \frac{1}{n-1} \sum_i^n (X_i - a)^2$

$\sum_i^n (X_i - \overline{X})^2 \leq \sum_i^n (X_i - a)^2$

$\sum_i^n (-2X_i\overline{X} + \overline{X}^2) \leq \sum_i^n (-2X_i a + a^2)$

$\sum_i^n (-2X_i\overline{X}) + n\overline{X}^2 \leq -2a \sum_i^n (X_i) + \sum_i^n (a^2)$

$-2n\overline{X}^2 + n\overline{X}^2 \leq -2an\overline{X} + na^2$

$-n\overline{X}^2 \leq -2an\overline{X} + na^2$

$-\overline{X}^2 \leq -2a\overline{X} + n^2a^2$

$0 \leq \overline{X}^2 - 2a\overline{X} + n^2a^2 (\star)$

if a and $\overline{X}$ have the opposite sign the inequality $(\star)$ is true because a sum of positive numbers is ≥ 0.

otherwise:

$\overline{X}^2 - 2a\overline{X} + n^2a^2 \geq^{(n\geq 1)} \overline{X}^2 - 2an\overline{X} + n^2a^2 = (\overline{X} - na)^2 \geq 0$

$\square$

ii) $\frac{(n-1)S^2}{n} = \frac{1}{n}\sum_{i}^{n}(X_i - \overline{X})^2$
$= \frac{1}{n}\sum_{i}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2)$
$= \overline{X^2} - 2\overline{X}^2 + \overline{X}^2 = \overline{X^2} - \overline{X}^2$
$\square$

Exercise 2.2

$X_1, \ldots, X_n$ iid random sample with $X_i \sim F$ with continuous F
.
i)
$\min : \mathbb{R}^n \to \mathbb{R}$
I call Z the rv $X_{(1)}$
$B_z = \{x_1, \ldots, x_n : \min(x_1, \ldots, x_n) \leq z\}$
we have that B_Z is the set of points in which at least one component is less than z

${B_z}^C = (z, +\infty) \times \cdots \times (z, \infty)$
$F_Z(z) = 1 - \int_{B_z^C} \prod_{i=1}^{n} f(x_i)dx_1dx_2 \cdots dx_n = 1 - (1 - F(z))^n$
Now we can take the derivative:
$f_Z(z) = -n(1 - F(z))^{n-1}(-f(z)) = n(1 - F(z))^{n-1}f(z) \quad \square$
ii)
$\max : \mathbb{R}^n \to \mathbb{R}$
$B_z = \{x_1, \ldots, x_n : \max(x_1, \ldots, x_n) \leq z\}$

which means that B_z is the set of points in which all components are smaller than z (definition of minimum)
$B_z = (-\infty, z] \times \cdots \times (-\infty, z]$

we can find the df of Z,
$F_Z = \int\limits_{B_z} \prod\limits_{i=1}^{n} f(x_i) dx_1 dx_2 \cdots dx_n = (F(z))^n$
Now we can take the derivative:
$f(z) = n(F(z))^{n-1} f(z) \quad \square$
iii) if they are not independent we cannot factorize the pdfs and we would need the joint pdf to be able to calculate the integrals.
iv) Bz??????

Exercise 2.5

discrete

- Bernoulli: coin toss that either gives heads (1) with probability θ or tails (0) with probability $1 - \theta$
- Binomial: number of successes in n independent trials (that either succeed or fail) each with probability θ, for example number of heads when tossing a coin n times
- NegBin: it's a generalized version of geometric rv that models the number of failures until r successes happen (in n independent binary trials like the binomial). For

example the number of trials until we get r non consecutive heads when tossing a coin
- Poisson: number of events that happen independently from each other in a fixed amount of time. For example the number of connections to a server in a second.

continuous

- Gaussian: for example the height of humans.
- Exponential: for example the time a client waits in queue before being served by a server.
- Gamma: it's a generalization of the exponential distribution. It is also used to model waiting times.
- Weibull: for example it can model the time an electronic device lasts.
- Uniform: in telecommunications it can model the granular error of a symmetrical quantizer.

Exercise 2.4

i) The following script:

```
get_theta <- function(x_1,x_2,x_3){
  beta0=0.5
  beta1=-1
```

```
  beta2=1
  beta3=0.1
  mu=beta0+beta1*x_1+beta2*x_2+beta3*x_3
  return(1/(1+exp(-mu)))
}


tabella=read.table('C:\\Users\\matteo\\Desktop\\eggs.
txt')



number_of_pred_M=0
for(i in (2:101)){

theta_i=get_theta(as.double(tabella[i,"V1"]),as.doubl
e(tabella[i,"V2"]),as.double(tabella[i,"V3"]))
  predictedsex=rbinom(1,1,prob=theta_i)
  cat(predictedsex)
  cat(" ")
  number_of_pred_M=number_of_pred_M+predictedsex
}
print("number of males")
print(number_of_pred_M)
```

produces:

```
1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1
1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[1] "number of males"
[1] 96
```

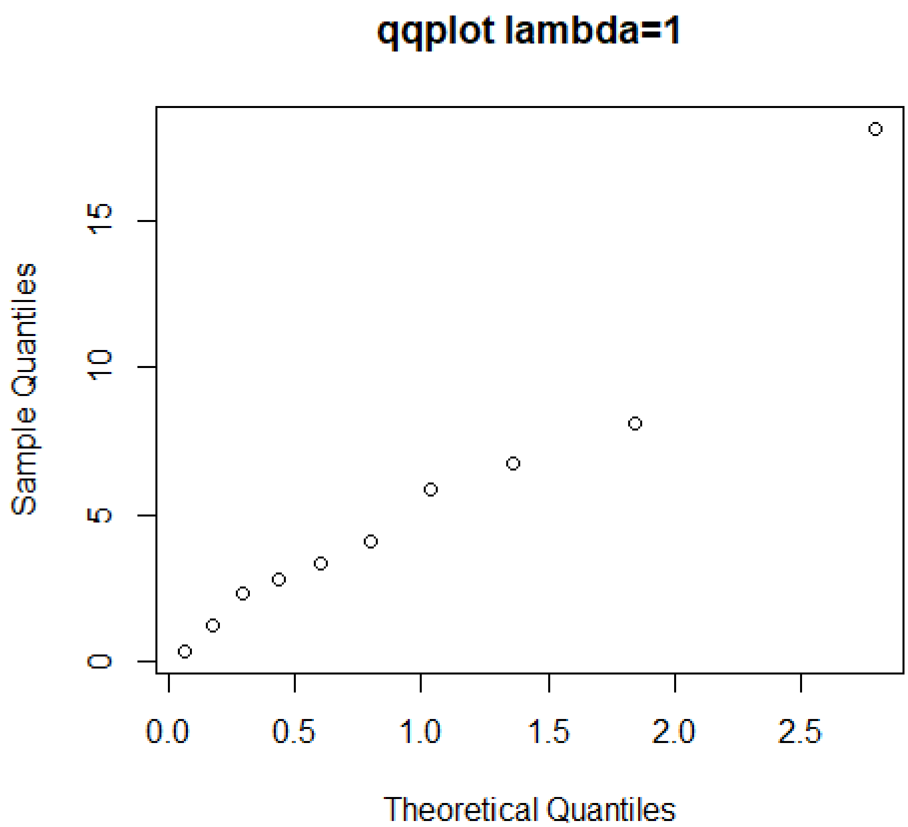
ii) with the following script

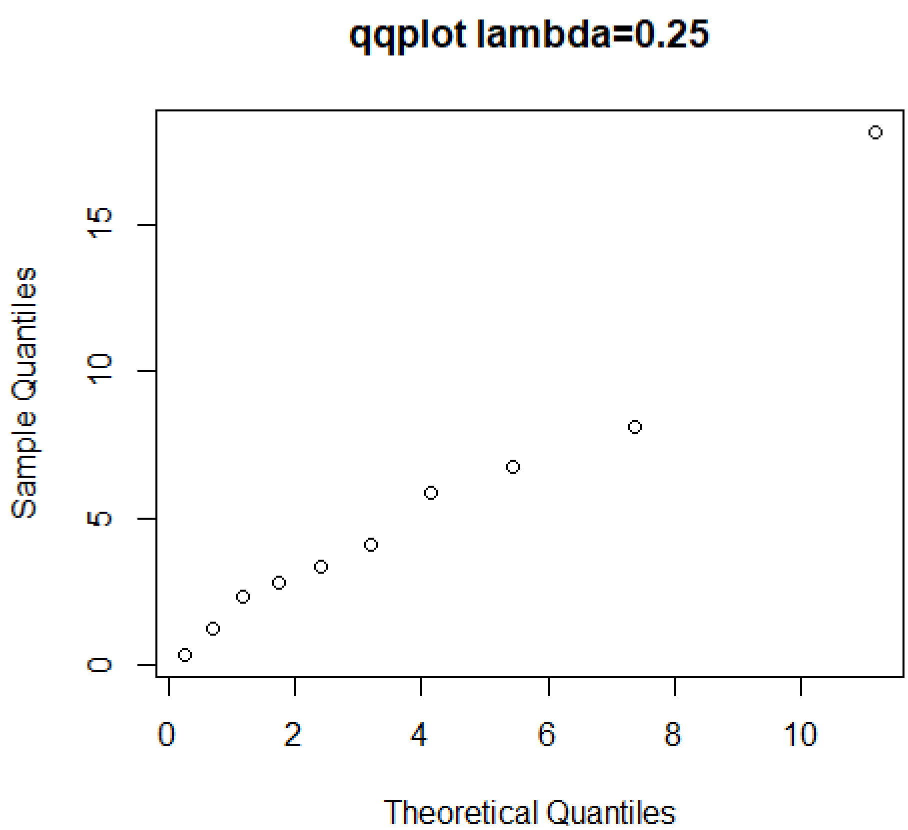
```
qvolume=as.double(quantile(as.double(tabella[2:101,1]
))[2])
qheight=as.double(quantile(as.double(tabella[2:101,2]
)))[2]
qrough=as.double(quantile(as.double(tabella[2:101,3])
))[2]
rbinom(1,1,prob=get_theta(qvolume,qheight,qrough))
```

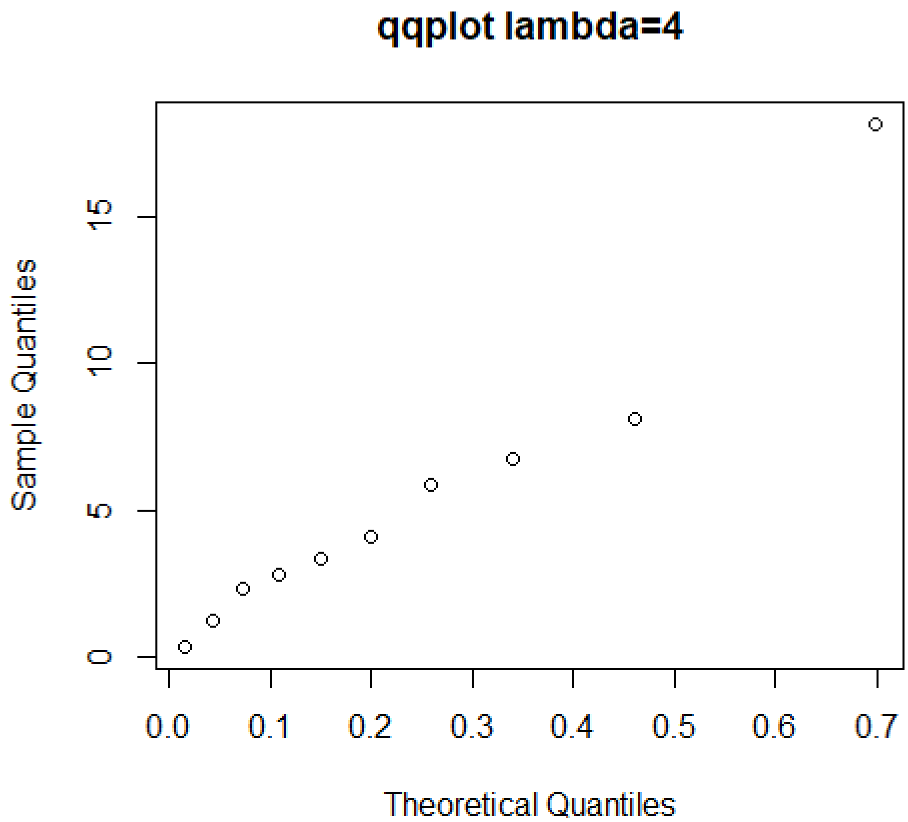
we get very often male as the result with $\theta = 0.9441$.

Exercise 2.3

i) plots in R:


qqplot lambda=1
Sample Quantiles
0
5
10
15
0.0
0.5
1.0
1.5
2.0
2.5
Theoretical Quantiles


qqplot lambda=0.25
Sample Quantiles
0
5
10
15
0
2
4
6
8
10
Theoretical Quantiles



ii)

boxplot by hand:

sort the samples: 0.111 0.492 2.120 2.699 3.255 4.102 6.254 6.951 8.935 29.389

$q_1 = X_{\lfloor 0.25 \cdot 11 \rfloor} = X_{(2)} = 0.492$

$q_2 = \frac{3.255+4.102}{2} = 3.6785$

$q_3 = X_{\lfloor 0.75 \cdot 11 \rfloor} = X_{(8)} = 6.951$

$iqr = q_3 - q_1 = 6.459$

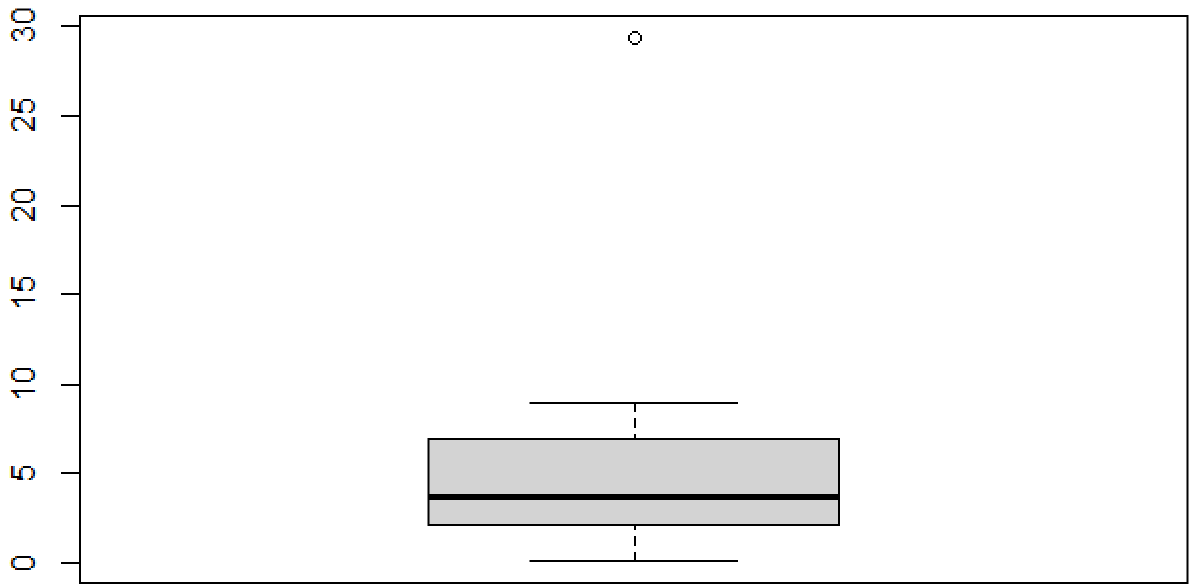
$q_1 - 1.5iqr = -9.1965$

$q_3 + 1.5iqr = 16.6395$

bottom whisker = 0.11

upper whisker = 8.935

we have an outlier 29.389 outside the whiskers boxplot in R:



the difference is that the boxplot in R is not using the first and third quartiles as the boundaries of the box, from the R documentation we can find that R is plotting the hinges instead of q_1 and q_3.
iii) in R

```
fn = ecdf(x)
fn(5.25)
```

returns 0.6

by hand we can compute the ecdf: there are six sample points smaller than 5.25

$$\hat{F}_n(5.25) = \frac{1}{10}6 = 0.6$$