# assignment-2

## Ex 1

### 1)

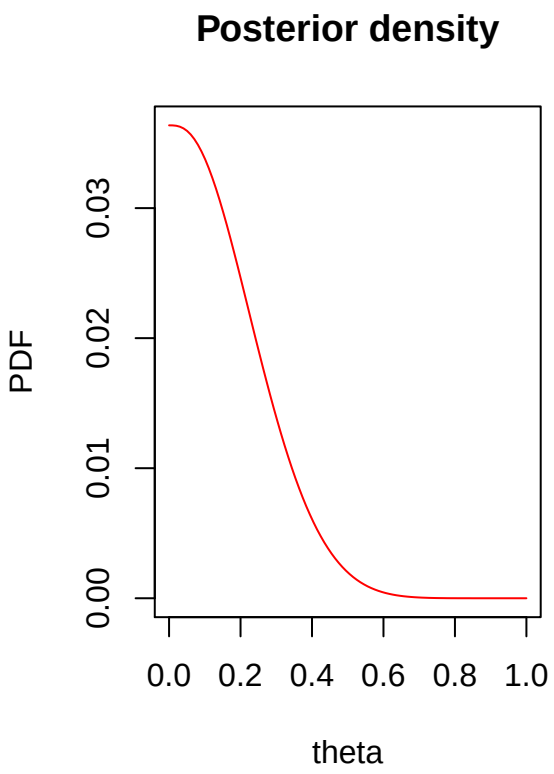
The posterior density function in case of censored observation is

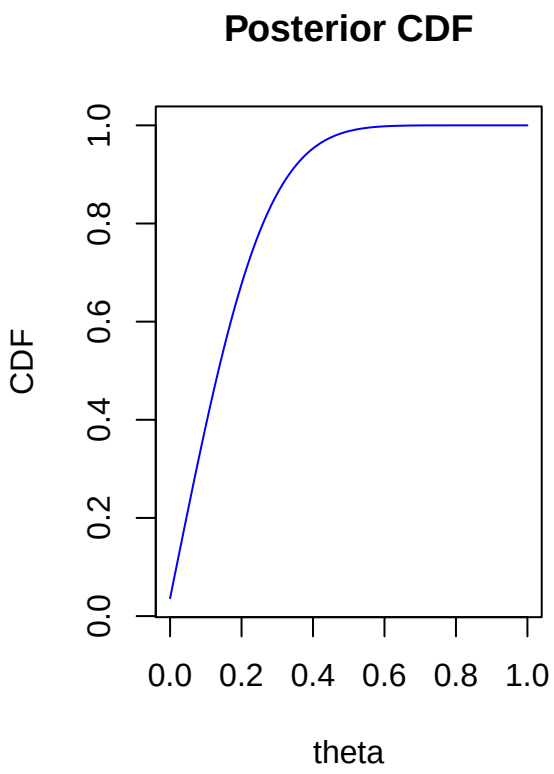
$p(\theta \mid y, n) \propto Beta(1, 11) + Beta(2, 10) + Beta(3, 9)$

When discretizing $\theta$ we can normalize the distribution as done below

### 2-3)

```r
# vectorize th into 100 bins
x <- seq(0, 1, length.out=100)

# calculate the unnormalized density at each bin
theta <- dbeta(x, 1, 11) + dbeta(x, 2, 10) + dbeta(x, 3, 9)

# normalize the discretized probability densities
theta = theta/sum(theta)

# calculate the cumulative distribution function
cdf <- cumsum(theta)

# divide plot into 2 subplots
par(mfrow=c(1,2))
# plot the posterior density
plot(x, theta, type='l', main='Posterior density', col='red', xlab='theta', ylab='PDF')
# plot the posterior cumulative distribution function
plot(x, cdf, type='l', main='Posterior CDF', col='blue', xlab='theta', ylab='CDF')
```




```r
# calculate the probability that theta < 0.3
sum(theta[which(x < 0.3)])
```

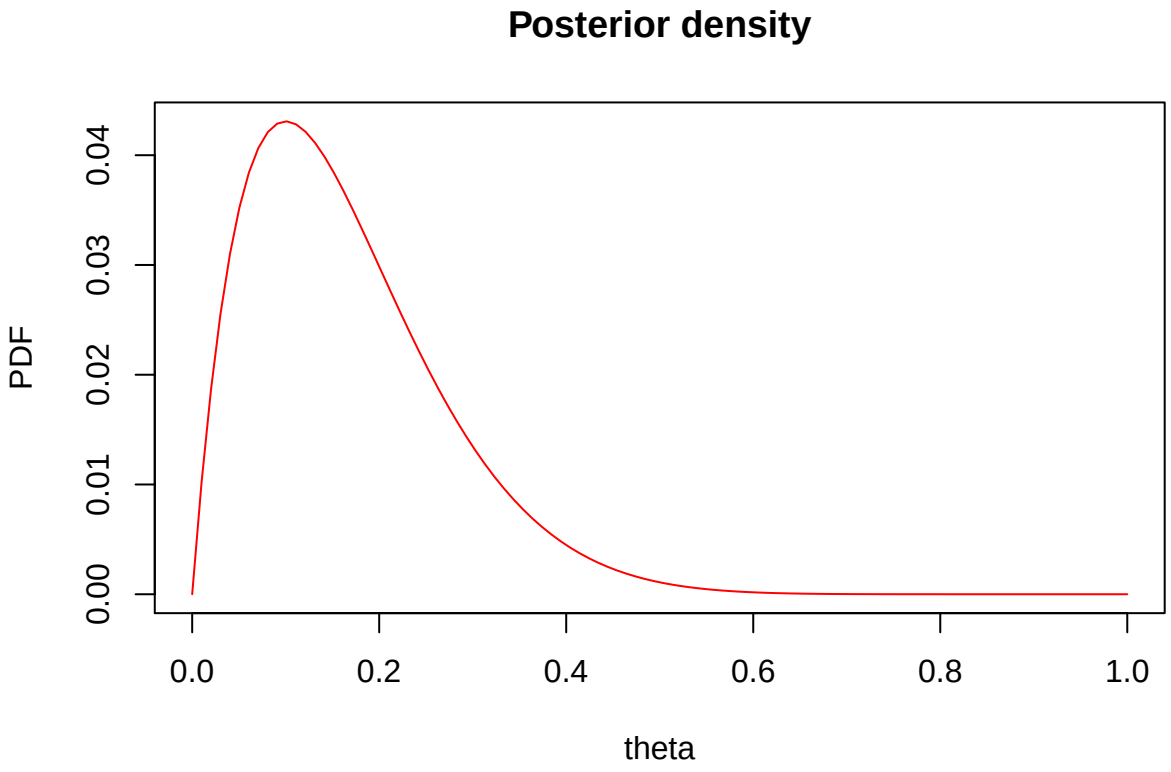
```
## [1] 0.8514363
```

## 4)

The posterior density funtion in case of $y = 1$ observation is

$p(\theta \mid y, n) \propto Beta(2, 10)$

```r
# calculate the density at each bin
theta <- dbeta(x, 2, 10)

# normalize
theta = theta/sum(theta)

# plot the unnormalized posterior
plot(x, theta, type='l', main='Posterior density', col='red', xlab='theta', ylab='PDF')
```




### What is the apparent differences between these two posterior densities?

The first distribution does not start from zero, the latter does. However, I'm not really sure what this really represents.

## Ex 2

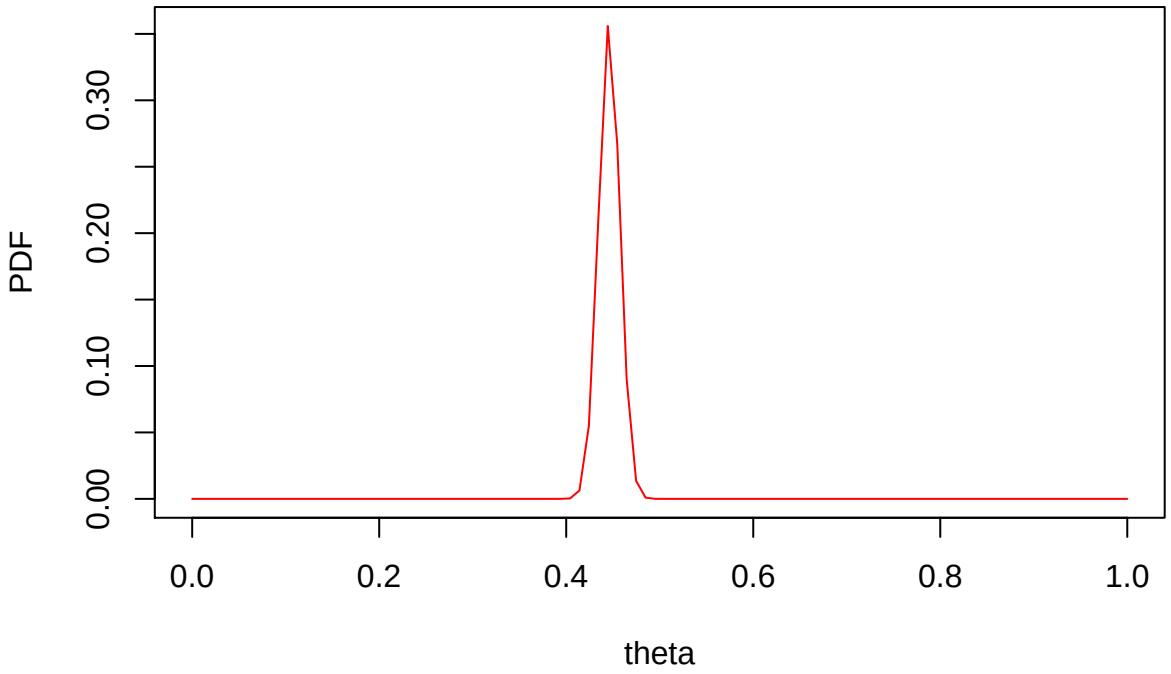
```
n <- 980
y <- 437
x <- seq(0, 1, length.out = 100)
```

### a)

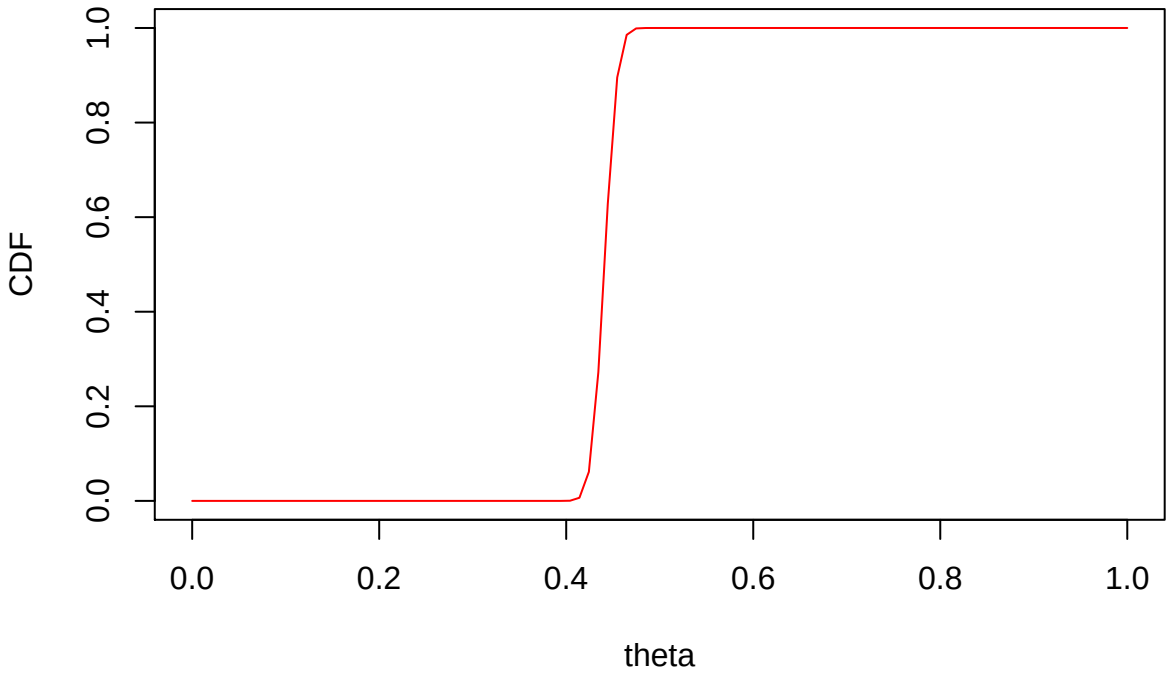
#### i) Uniform prior

```
alpha <- 1
alpha <- alpha + y

beta <- 1
beta <- beta + n - y

# 1. Visualize the posterior density for the proportion of female births
theta <- dbeta(x, alpha + y, beta + n - y)

# normalize
```

```r
theta = theta/sum(theta)

plot(x, theta, type='l', col='red', xlab='theta', ylab='PDF')
```




```r
# 2. Visualize cumulative density function
# This could also be calculated as:
#cdf <- pbeta(x, alpha, beta)
# But the results are same and we discretized it so cumsum seems natural.
cdf <- cumsum(theta)
plot(x, cdf, type='l', col='red', xlab='theta', ylab='CDF')
```



```r
# 3. Calculate posterior median and central 90% posterior interval for theta
qbeta(c(0.5,0.9), alpha+y, beta+n-y)
```

```
## [1] 0.4459551 0.4603656
```

```r
# 4. Calculate the posterior probability, p(theta < 0.485|y,n) and p(theta > 0.485|y,n)
sum(theta[which(x < 0.485)])
```

```
## [1] 0.9999712
```

```r
sum(theta[which(x > 0.485)])
```

```
## [1] 2.875266e-05
```

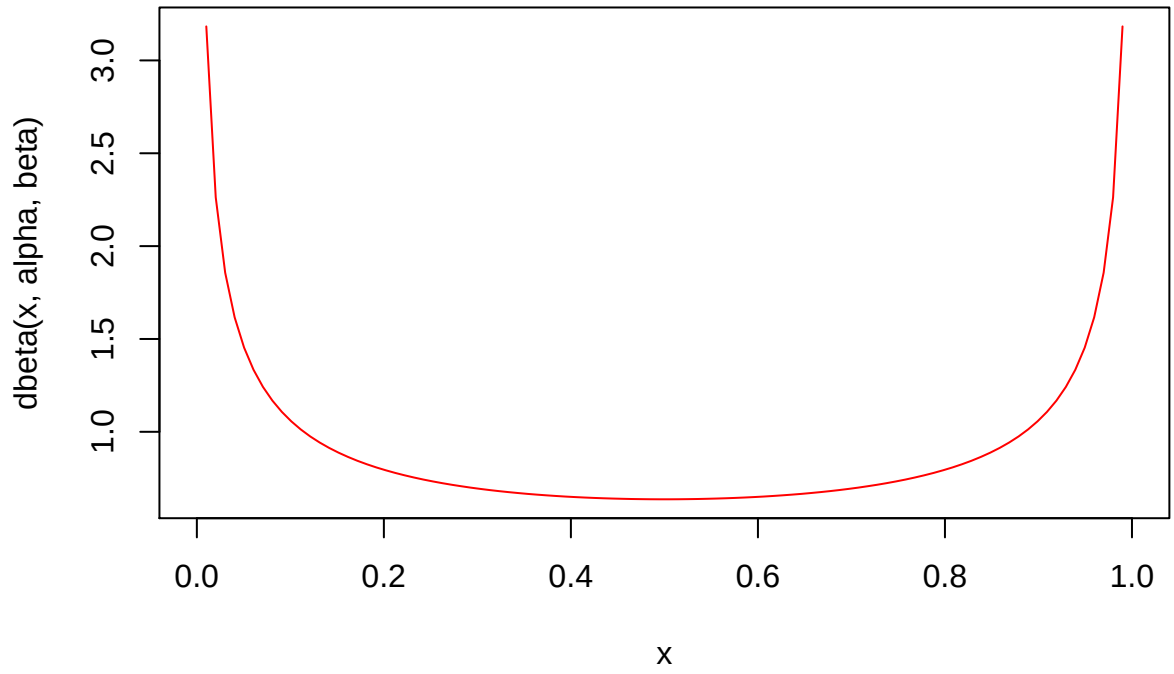
**ii)**

**1st Informative distribution with mean 0.5**

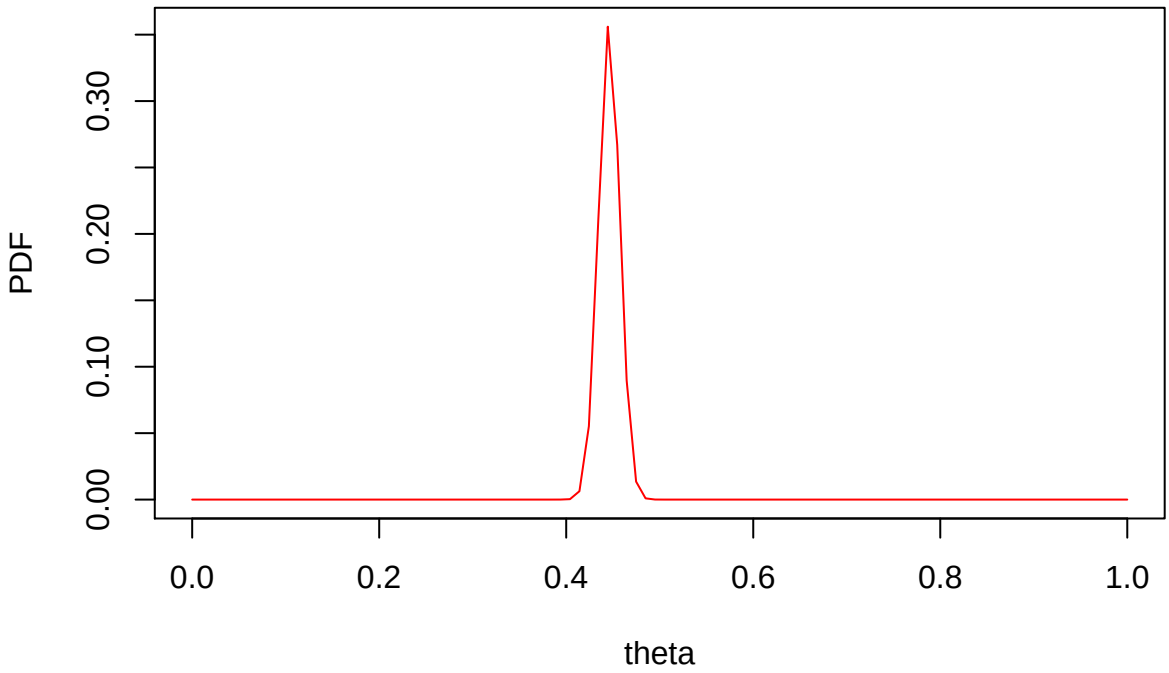
$\alpha = \beta = \frac{1}{2}$

$E(X) = \frac{\alpha}{\alpha+\beta} \implies \frac{1/2}{1/2+1/2} = \frac{1}{2}$

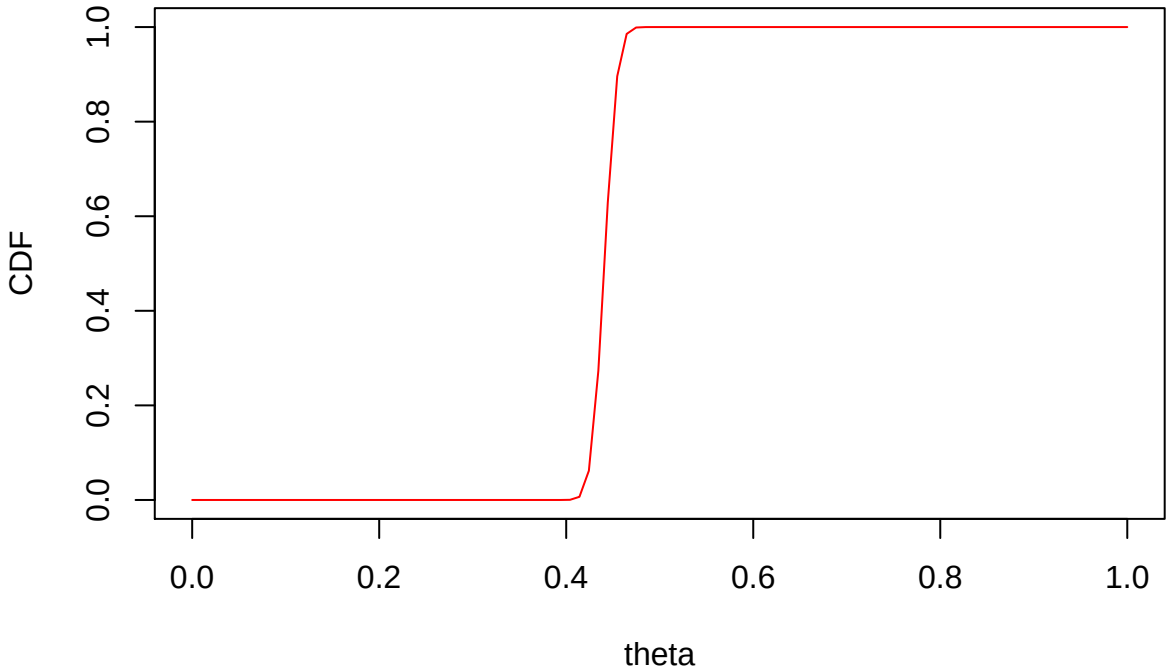
```r
alpha <- 1/2
beta <- 1/2

# Plot of the prior
plot(x, dbeta(x, alpha, beta), type='l', col='red')
```




```r
alpha <- alpha + y
beta <- beta + n - y

# 1. Visualize the posterior density for the proportion of female births
theta <- dbeta(x, alpha + y, beta + n - y)
theta = theta/sum(theta)
plot(x, theta, type='l', col='red', xlab='theta', ylab='PDF')
```




```r
# 2. Visualize cumulative density function
cdf <- cumsum(theta)
plot(x, cdf, type='l', col='red', xlab='theta', ylab='CDF')
```




```r
# 3. Calculate posterior median and central 90% posterior interval for theta
qbeta(c(0.5,0.9), alpha+y, beta+n-y)
```

```
## [1] 0.4459276 0.4603417
```

```r
# 4. Calculate the posterior probability, p(theta < 0.485|y,n) and p(theta > 0.485|y,n)
theta = theta/sum(theta)
sum(theta[which(x < 0.485)])
```

```
## [1] 0.9999714
```

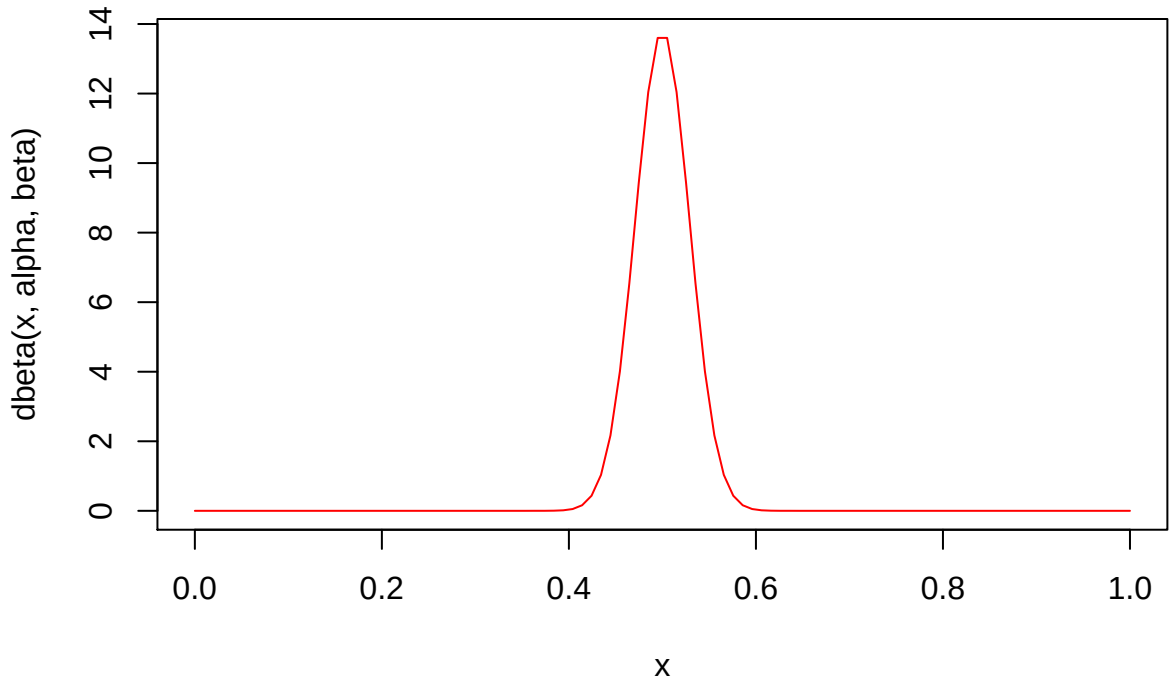
```
sum(theta[which(x > 0.485)])
```

```
## [1] 2.857823e-05
```

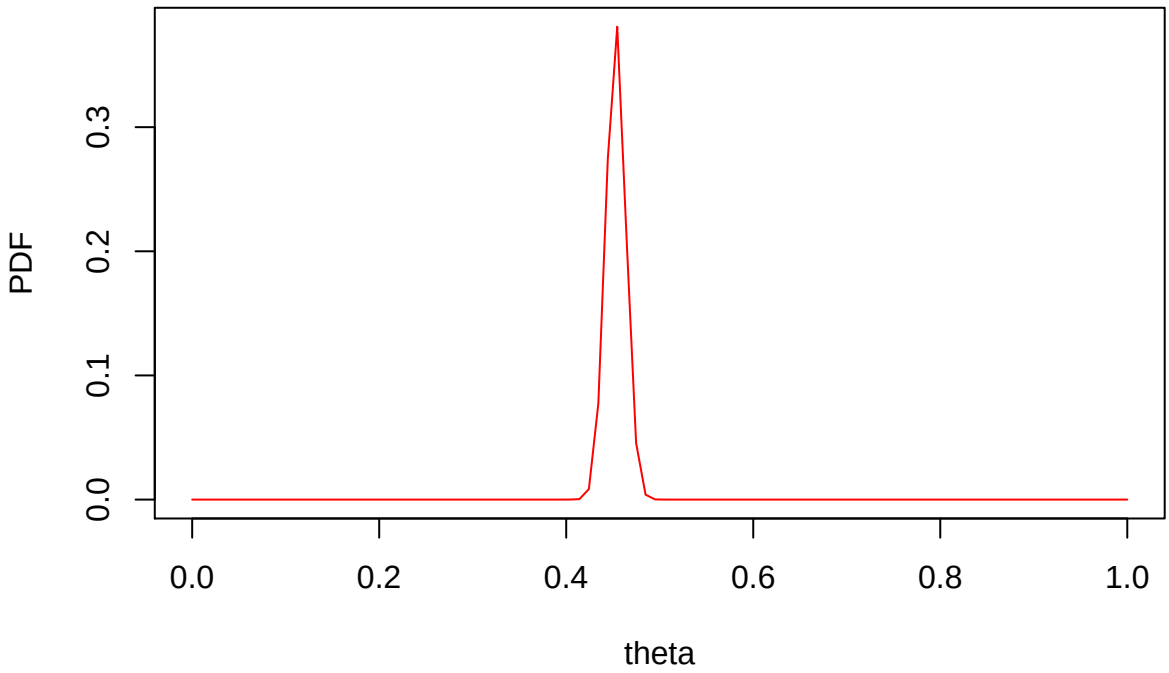
### 2nd Informative distribution with mean 0.5

$E(X) = \frac{\alpha}{\alpha+\beta} \implies \frac{150}{150+150} = \frac{1}{2}$

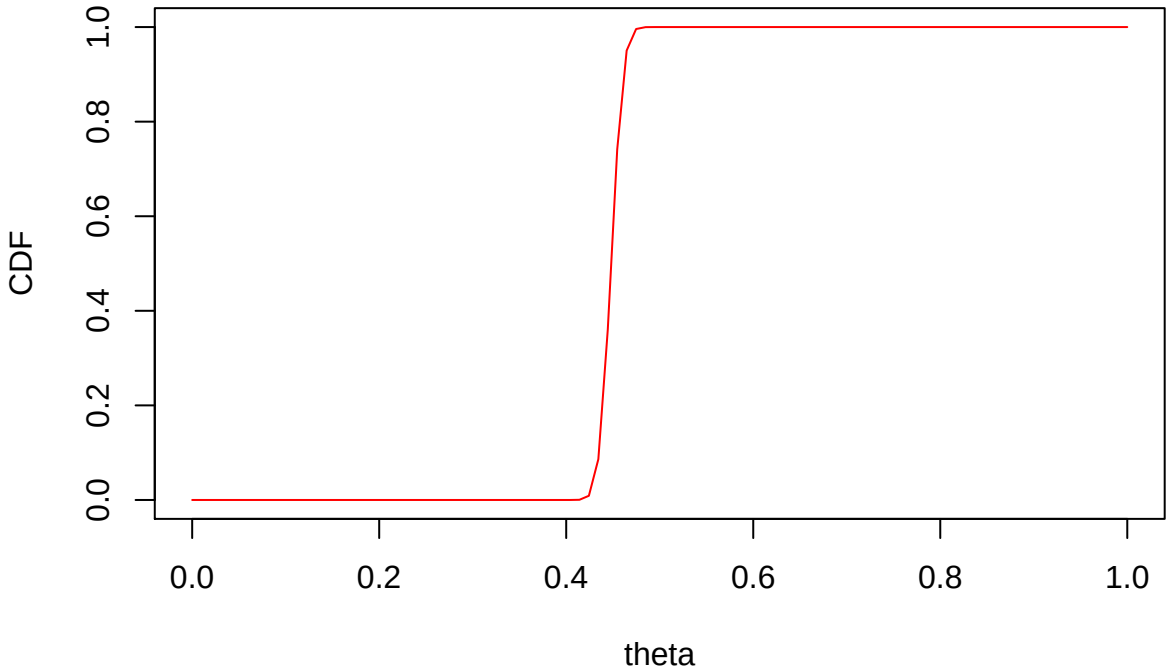
```
alpha <- 150
beta <- 150

# Plot of the prior
plot(x, dbeta(x, alpha, beta), type='l', col='red')
```




```
alpha <- alpha + y
beta <- beta + n - y

# 1. Visualize the posterior density for the proportion of female births
theta <- dbeta(x, alpha + y, beta + n - y)

# normalize
theta = theta/sum(theta)
plot(x, theta, type='l', col='red', xlab='theta', ylab='PDF')
```




```
# 2. Visualize cumulative density function
cdf <- cumsum(theta)
plot(x, cdf, type='l', col='red', xlab='theta', ylab='CDF')
```




```
# 3. Calculate posterior median and central 90% posterior interval for theta
qbeta(c(0.5,0.9), alpha+y, beta+n-y)
```

```
## [1] 0.4530835 0.4665246
```

```
# 4. Calculate the posterior probability, p(theta < 0.485|y,n) and p(theta > 0.485|y,n)
sum(theta[which(x < 0.485)])
```

```
## [1] 0.9998611
```

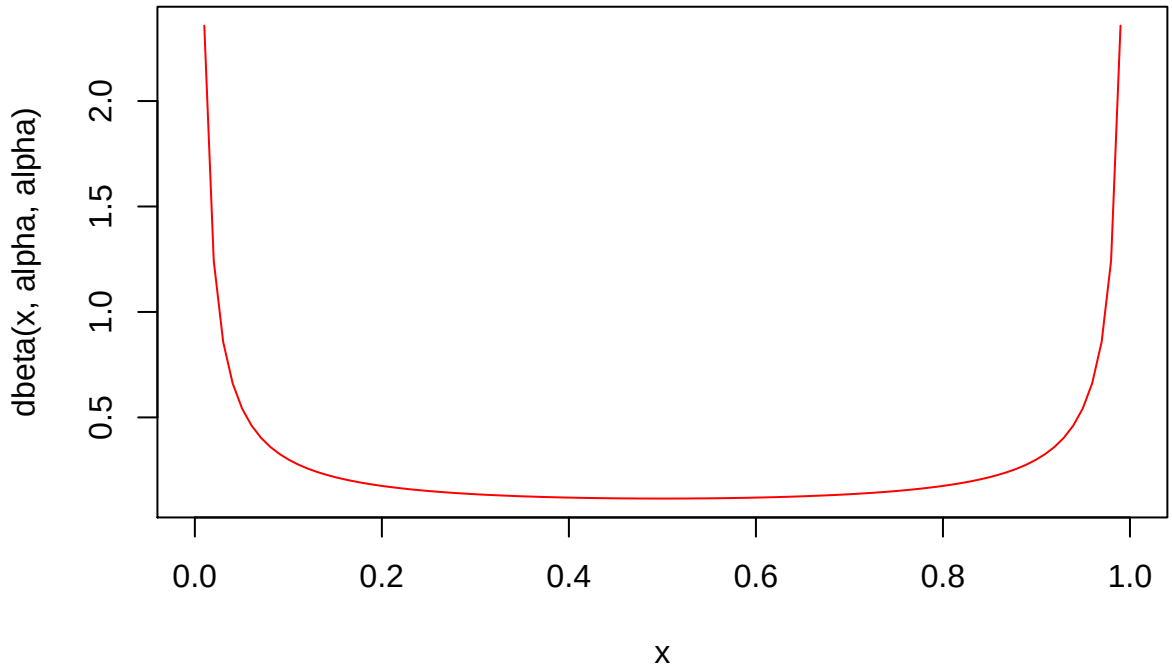
```r
sum(theta[which(x > 0.485)])
```

```
## [1] 0.0001388927
```

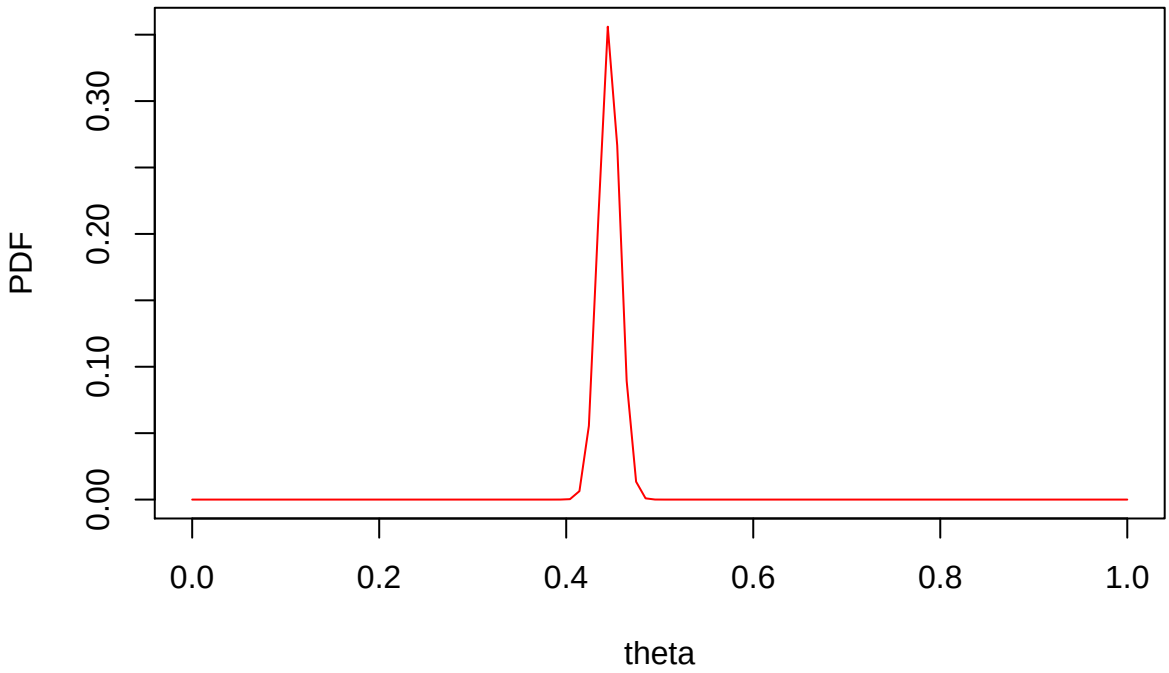
### 3rd Informative distribution with mean 0.5

$E(X) = \frac{\alpha}{\alpha+\beta} \implies \frac{1/16}{1/16+1/16} = \frac{1}{2}$

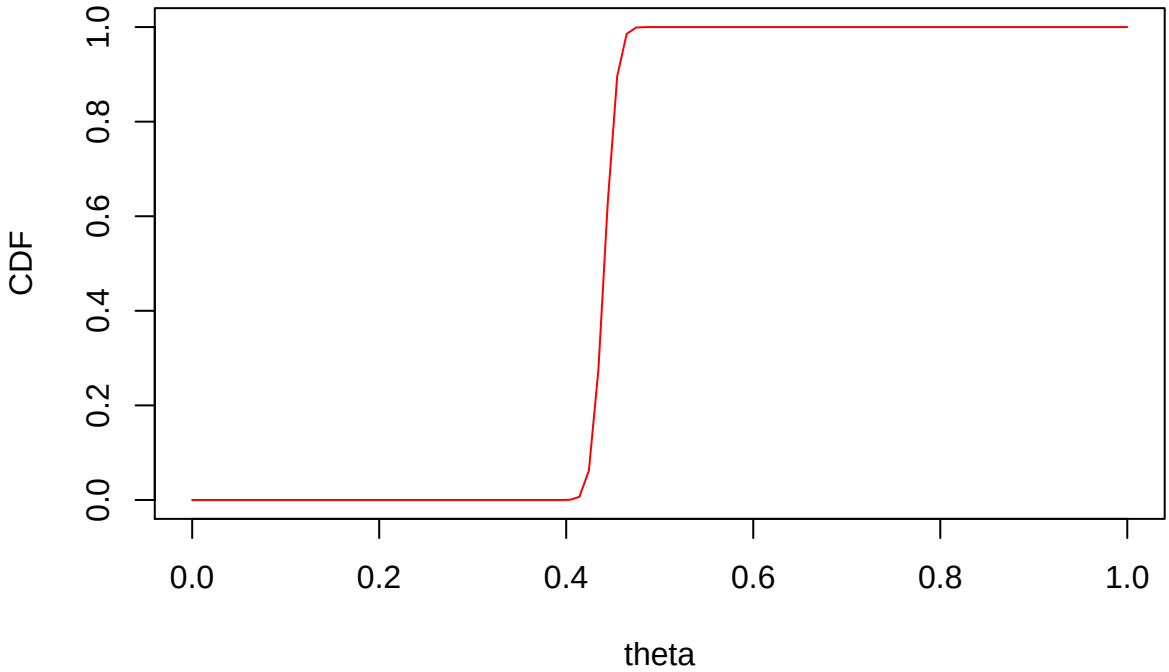
```r
alpha <- 1/16
beta <- 1/16

# Plot of the prior
plot(x, dbeta(x, alpha, alpha), type='l', col='red')
```




```r
alpha <- alpha + y
beta <- beta + n - y

# 1. Visualize the posterior density for the proportion of female births
theta <- dbeta(x, alpha + y, beta + n - y)
theta = theta/sum(theta)
plot(x, theta, type='l', col='red', xlab='theta', ylab='PDF')
```




```r
# 2. Visualize cumulative density function
cdf <- cumsum(theta)
plot(x, cdf, type='l', col='red', xlab='theta', ylab='CDF')
```




```r
# 3. Calculate posterior median and central 90% posterior interval for theta
qbeta(c(0.5,0.9), alpha+y, beta+n-y)
```

```
## [1] 0.4459034 0.4603207
```

```r
# 4. Calculate the posterior probability, p(theta < 0.485|y,n) and p(theta > 0.485|y,n)
theta = theta/sum(theta)
sum(theta[which(x < 0.485)])
```

```
## [1] 0.9999716
```

```
sum(theta[which(x > 0.485)])
```

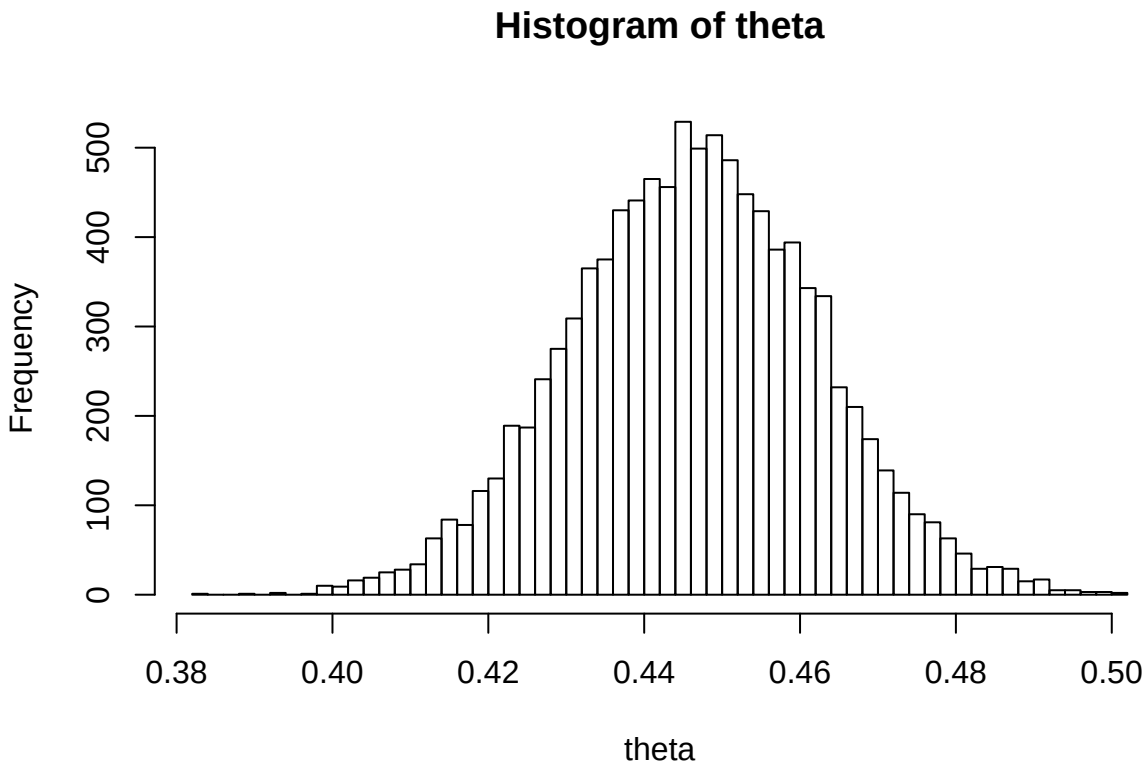
```
## [1] 2.842633e-05
```

## b)

```
alpha <- 1
alpha <- alpha + y

beta <- 1
beta <- beta + n - y
```

**1. Visualize the distribution with histogram**

```
theta <- rbeta(10000, alpha, beta)
hist(theta, breaks=50)
```




**2. Calculate the posterior median, standard deviation and coefficient of variation of theta**

```
sprintf('Posterior median: %f', median(theta))
```

```
## [1] "Posterior median: 0.446442"
```

```
sprintf('Posterior standard deviation: %f', sd(theta))
```

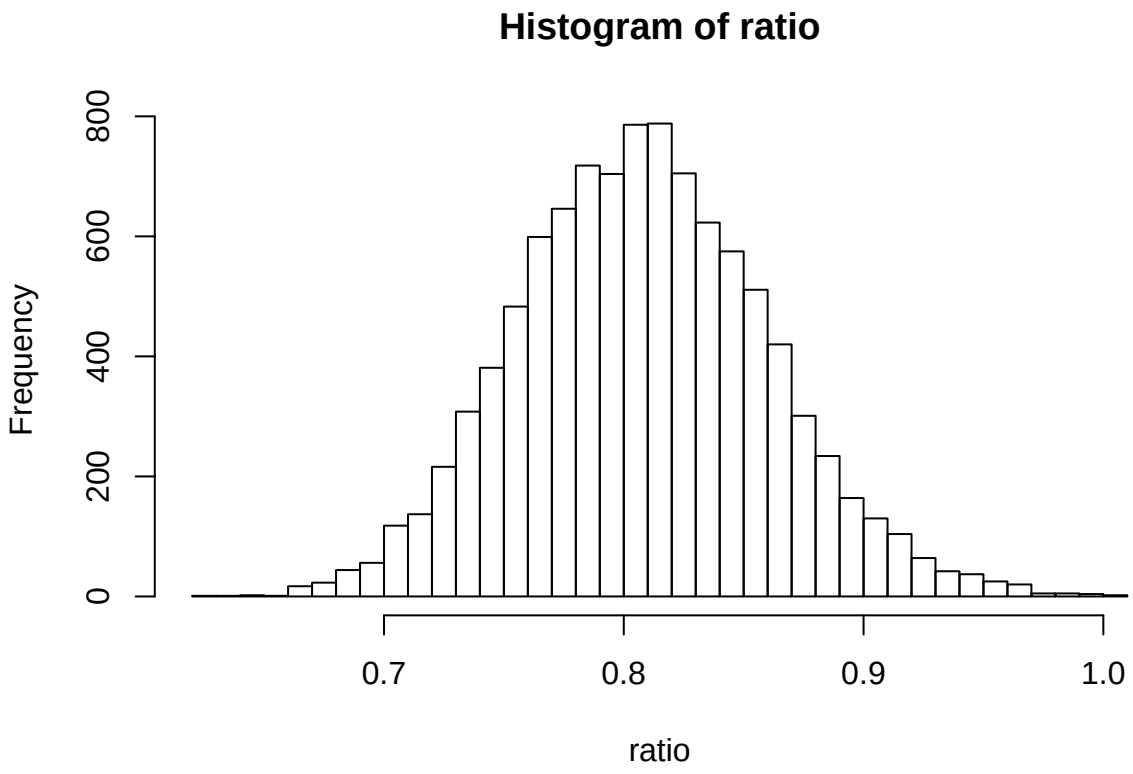
```
## [1] "Posterior standard deviation: 0.015899"
```

```r
sprintf('Posterior variance: %f', var(theta))
```

```
## [1] "Posterior variance: 0.000253"
```

**3. Visualize the female-to-male sex ratio with histogram**

```r
ratio = theta/(1 - theta)
hist(ratio, breaks=50)
```




### 4.

Calculate

```r
ratio = theta/(1 - theta)
sum(ratio < 0.9)/length(ratio)
```

```
## [1] 0.9562
```

# Ex 3

## 1.

### a)

The superpopulation is all the 5 year old Finns.

### b)

Yes. All the information we need is age and height, and both are available.

## 2.

**a)**

Again, the superpopulation is all the 5 year old Finns.

**b)**

No, the age information is missing from the random sample.

## 3.

**a)**

The superpopulation is all the sandy shores of Gulf of Bothnia.

**b)**

Yes. The sample contains information about the shores and the fish spawning there.

## 4.

**a)**

The superpopulation is all the people in Puumala.

**b)**

This could work if there is a correlation between different age groups and their usage of electronic scooter. But probably not a very good estimate since this is missing information about number of roads and their types in Puumala. One could guess that electronic scooters are not so usable on rocky roads in the woods.

## 5.

**a)**

The superpopulation is all the people in Tampere.

**b)**

Yes. Helsinki and Tampere are quite similar cities so probably the same type of people that use electronic scooters in Helsinki probably would use them in Tampere also.