Week 2 exercises

Exercise 1

The posterior density function in case of censored observation is

Prior:

$$P(\theta) \propto \theta^0(1-\theta)^0 \tag{1}$$

Likelihood:

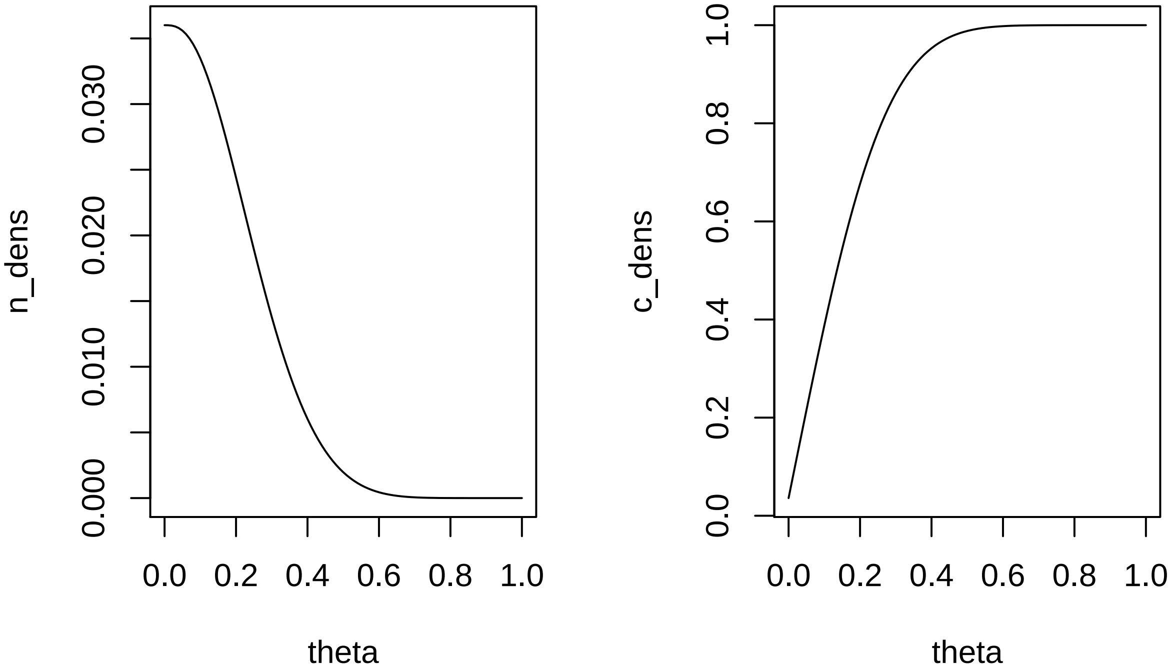
$$P(y|\theta) = \binom{10}{0}(1-\theta)^{10} + \binom{10}{1}\theta(1-\theta)^9 + \binom{10}{2}\theta(1-\theta)^8 \tag{2}$$

Posterior:

$$P(\theta|y) \propto \binom{10}{0}(1-\theta)^{10} + \binom{10}{1}\theta(1-\theta)^9 + \binom{10}{2}\theta^2(1-\theta)^8 \tag{3}$$

$$P(\theta|y) \propto (1-\theta)^{10} + 10\theta(1-\theta)^9 + 45\theta^2(1-\theta)^8 \tag{4}$$

```r
# vectorize th into 100 bins
theta = seq(0, 1, by=0.01)

# calculate the unnormalized density at each bin
u_dens = (1 - theta)^10 + 10 * theta * (1 - theta)^9 + 45 * theta^2 * (1 - theta)^8

# normalize the discretized probability densities
n_dens = u_dens / (sum(u_dens))

# calculate the cumulative distribution function
cdf = function(probs) {
  for (ind in 2:length(probs)) {
    probs[ind] = probs[ind - 1] + probs[ind]
  }
  probs
}


# plot the posterior density
par(mfrow=c(1,2))           # divide plot into 2 subplots
plot (theta, n_dens, type='l')
# plot the posterior cumulative distribution function

c_dens = cdf(n_dens)
plot (theta, c_dens, type='l')
```



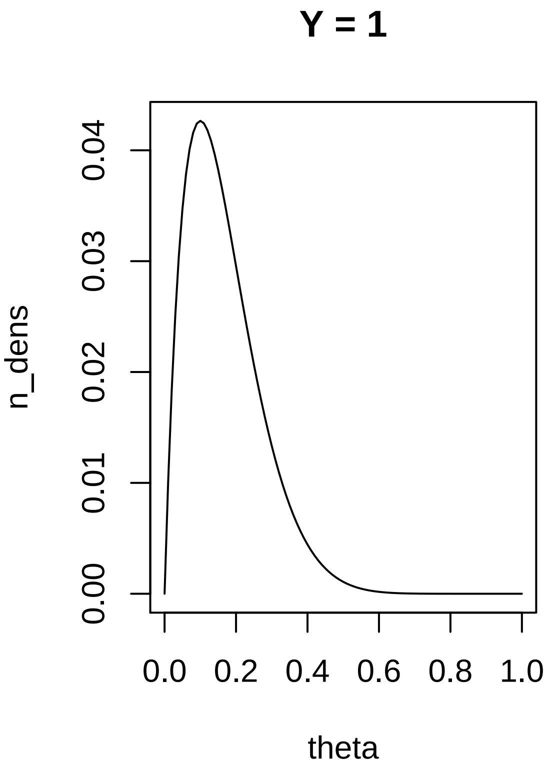
```
# calculate the probability that theta < 0.3
print(c_dens[length(theta[theta < 0.3])])
```

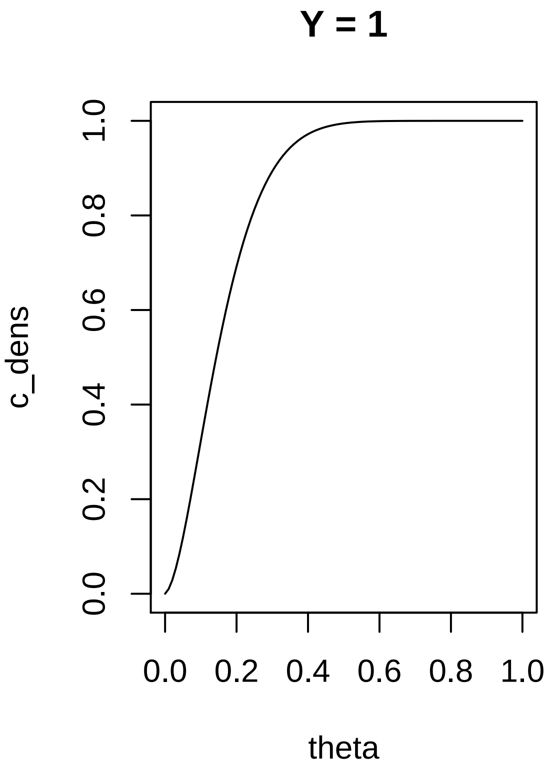
```
## [1] 0.8472008
```

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

$$P(\theta|y) \propto 10\theta(1-\theta)^9 \tag{5}$$

```
# calculate the density at each bin

# calculate the unnormalized density at each bin
u_dens = 10 * theta * (1 - theta)^9

# normalize the discretized probability densities
n_dens = u_dens / (sum(u_dens))

c_dens = cdf(n_dens)
# plot the unnormalized posterior
par(mfrow=c(1,2))           # divide plot into 2 subplots
plot (theta, n_dens, type='l', main='Y = 1')
# plot the posterior cumulative distribution function
plot (theta, c_dens, type='l', main='Y = 1')
```



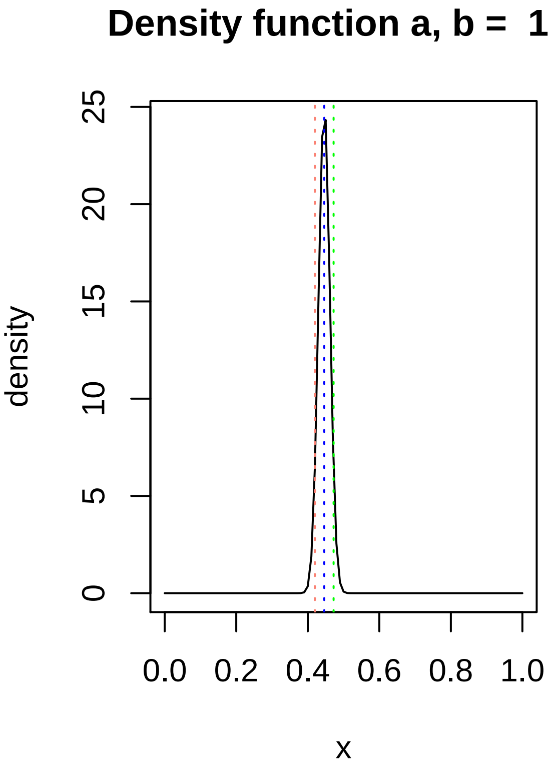


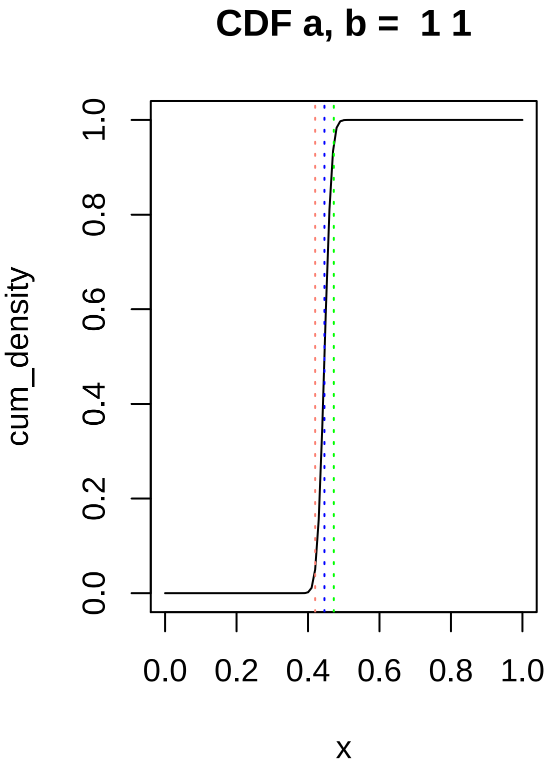
```r
# calculate the probability that theta < 0.3
```

Clearly when Y = 1, we are slightly more certain about the most probable value, which can be seen as the focus on the probability of theta having value around 0.1. In the case Y = 0-2, we still have the possibility that theta = 0, whereas when we have observed one toss with heads, we know that the probability can't be zero.

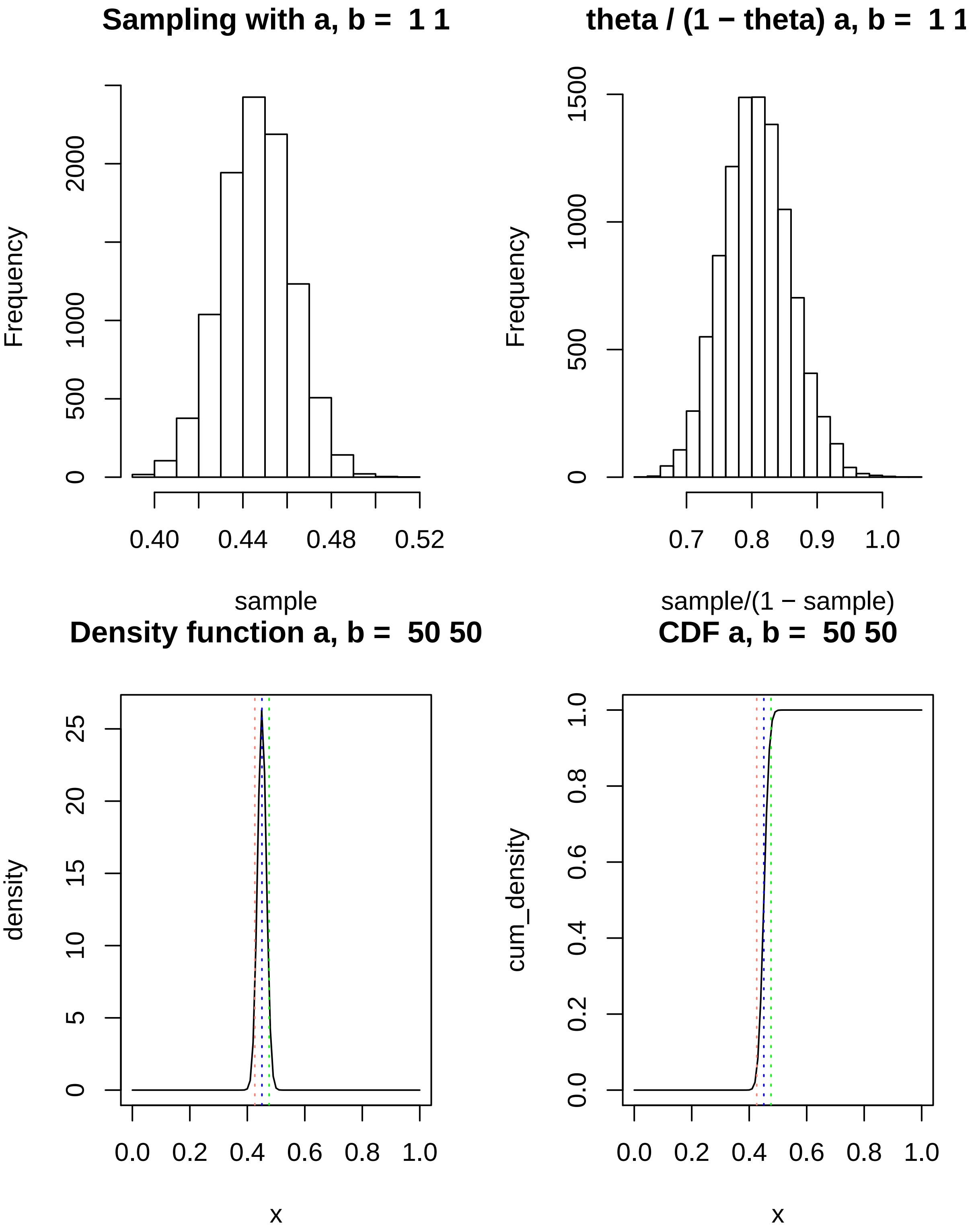
Exercise 2

```r
for (prior in list(c(1,1), c(50,50), c(200, 200))) {
  p_a = prior[1]
  p_b = prior[2]

  n = 980
  y = 437

  x = seq(0, 1, .01)

  cum_density = pbeta(x, p_a + y, p_b + n - y)
  density = dbeta(x, p_a + y, p_b + n - y)
  median = qbeta(.5, p_a + y, p_b + n - y)
  lower_90 = qbeta(.05, p_a + y, p_b + n - y)
  upper_90 = qbeta(.95, p_a + y, p_b + n - y)

  par(mfrow=c(1,2))
  plot (x, density, type='l', main=paste('Density function a, b = ', p_a, p_b))
  abline(v=median, lty='24', col='blue')
  abline(v=lower_90, lty='24', col='salmon')
  abline(v=upper_90, lty='24', col='green')
  plot (x, cum_density, type='l', main=paste('CDF a, b = ', p_a, p_b))
  abline(v=median, lty='24', col='blue')
  abline(v=lower_90, lty='24', col='salmon')
  abline(v=upper_90, lty='24', col='green')

  print(paste(c('P(theta < 0.485) =',  pbeta(0.485, p_a + y, p_b + n - y))))
  print(paste(c('P(theta > 0.485) =', 1 - pbeta(0.485, p_a + y, p_b + n - y))))


  sample = rbeta(10000, p_a + y, p_b + n - y)
  sd = sqrt(var(sample))
  med = median(sample)
  avg = mean(sample)
  coef_var = (sd / avg) * 100
  hist(sample, main=paste('Sampling with a, b = ', p_a, p_b))
  print(paste('Median ', med))
  print(paste('Standard deviation ', sd))
  print(paste('Coefficient of variation ',coef_var))
  print(paste('Prob(theta / (theta -1) < 0.9) = ', sum( (sample/(1-sample)) < 0.9)/length(sample)))
  hist(sample/(1-sample), main=paste('theta / (1 - theta) a, b = ', p_a, p_b))
}
```



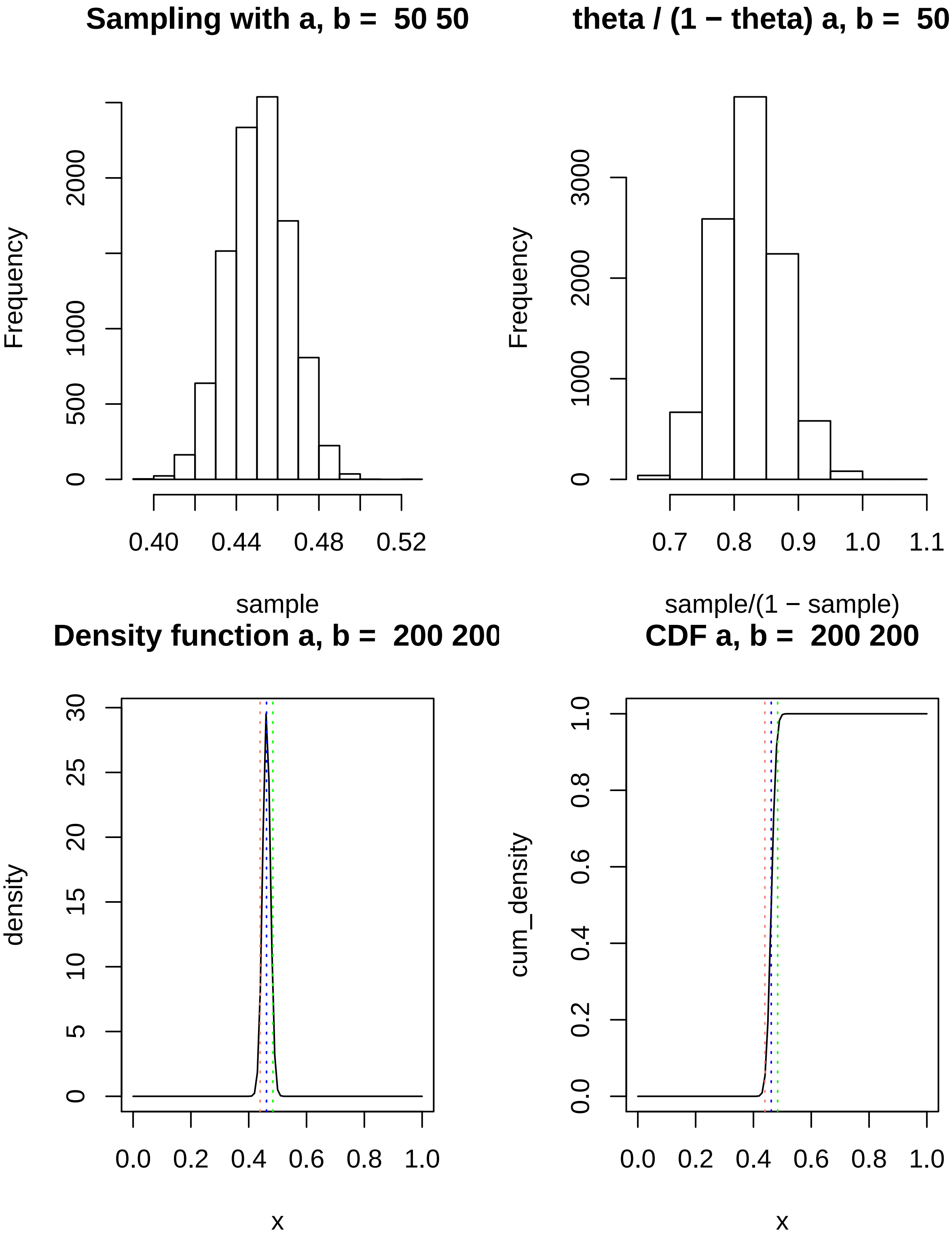


```
## [1] "P(theta < 0.485) =" "0.992825988560652"
## [1] "P(theta > 0.485) ="  "0.00717401143934815"

## [1] "Median  0.446340173571819"
## [1] "Standard deviation  0.0158948056106084"
## [1] "Coefficient of variation  3.56192141443147"
## [1] "Prob(theta / (theta -1) < 0.9) =  0.9568"
```



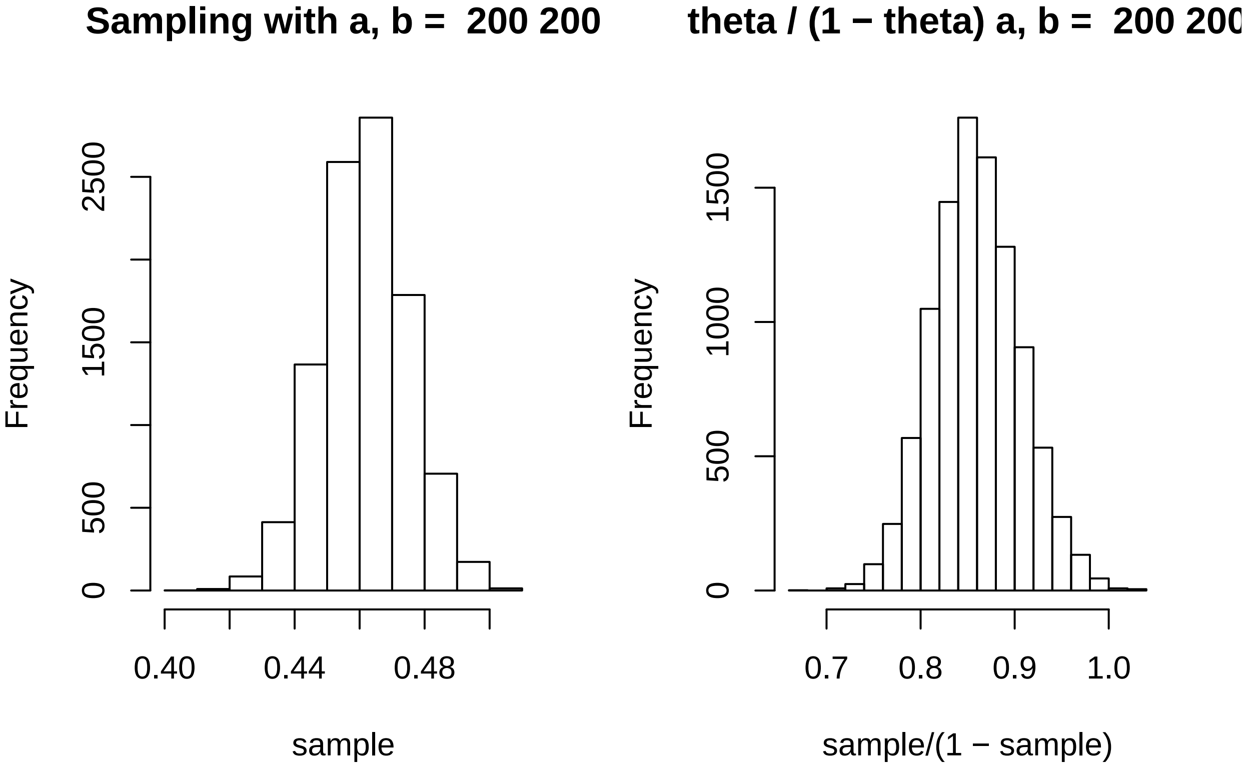
```
## [1] "P(theta < 0.485) =" "0.987598861905189"
## [1] "P(theta > 0.485) =" "0.012401138094811"
```

```
## [1] "Median  0.451287772666244"
## [1] "Standard deviation  0.0150878575712592"
## [1] "Coefficient of variation  3.34467611878597"
## [1] "Prob(theta / (theta -1) < 0.9) =  0.9336"
```



```
## [1] "P(theta < 0.485) =" "0.959242128934935"
## [1] "P(theta > 0.485) =" "0.0407578710650647"
```

```
## [1] "Median  0.461729604988279"
## [1] "Standard deviation  0.0133900570471774"
## [1] "Coefficient of variation  2.89933834659304"
## [1] "Prob(theta / (theta -1) < 0.9) =  0.8097"
```



Exercise 3

1) The superpopulation is 5 year old Finns. The data is not completely representative of the superpopulation as firstly not all 5 year olds go to a day care and this could mostly leave out some particular group of people from the sample. Also sampling daycares could lead to not representative samples as there might be different quantities of 5 year olds at each of the daycares of daycares that may not be uniformly distributed and thus some populations might get overrepresented.

2) Pretty much the same problems as in the previous part except now we are likely counting people who are not 5 year olds since we don't know their ages so that would make it even worse of a representation. If the question to these would have been what is the average height of 5 year olds in daycares then these both would have been more representative samples.

3) The superpopulation would be the sandy shores of Gulf of Bothnia. Assuming that we have a good portion of the shores sampled and there are not massive regional differences in the fish spawning between northern/southern shores etc. this would be a representative sample of the superpopulation, if there were some spatial differences then we might have to integrate the location of the shore into the model.

4) Here the superpopulation would be people of Puumala. Most likely the data from Helsinki is not very representative here since the differences in how people move in a big city vs a smaller city are most likely quite large and likely the demographies are quite different.

5) Here the superpopulation is the people of Tampere. Now the user statistics would be more useful since likely the populations and their preferred mobility types are likely kind of similar. Still though there is likely more differences so it would not be a very precise estimate. For example driving a car in Tampere might be easier than in Helsinki so there might be less need for electronic scooters. Also it could be that in Tampere people make shorter trips and the electronic scooters could be more useful but these are the differences that can't really be predicted from statistics gathered in Helsinki.