

Density Estimation

Bandwidth choice by leave-one-out maximum likelihood

Biel Caballero, Menzenbach Svenja and Reyes Illescas Kleber Enrique

2023-09-22

Histogram

1. At the slides we have seen the following relationship

$$\hat{f}_{h,(-i)}(x_i) = \frac{n}{n-1} \left(\hat{f}_h(x_i) - \frac{K(0)}{nh} \right)$$

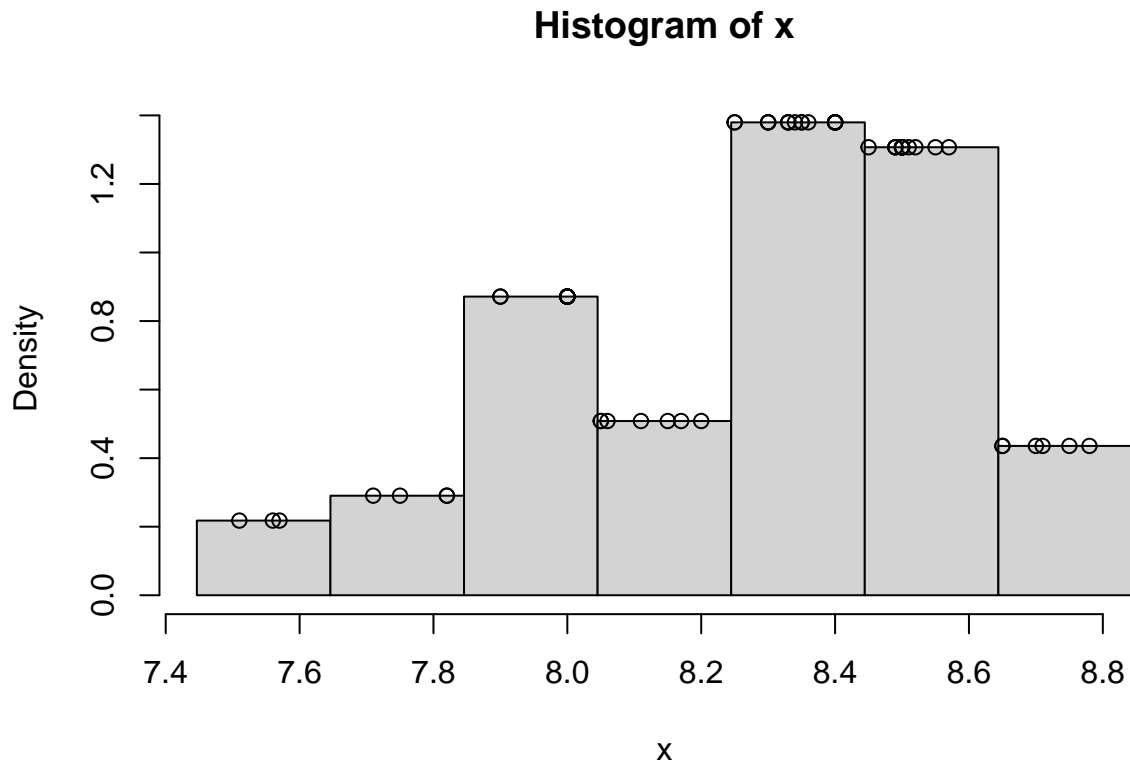
between the leave-one-out kernel density estimator $\hat{f}_{h,(-i)}(x)$ and the kernel density estimator using all the observations $\hat{f}_h(x)$, when both are evaluated at x_i , one of the observed data. Find a similar relationship between the histogram estimator of the density function $\hat{f}_{hist}(x)$ and its leave-one-out version, $\hat{f}_{hist,(-i)}(x)$, when both are evaluated at x_i .

2. Read the CD rate data set and call x the first column. Then define A, Z and nbr and plot the histogram of x

```
cdrate.df <- read.table("./cdrate.dat.txt")
# head(cdrate.df)
x <- cdrate.df[,1]
# sort(CDrate)
# # Stem-and-Leaf plot
# stem(CDrate)
```

```
A <- min(x) - .05*diff(range(x))
Z <- max(x) + .05*diff(range(x))
nbr <- 7

hx <- hist(x, breaks=seq(A,Z,length=nbr+1), freq=F)
hx_f <- stepfun(hx$breaks, c(0, hx$density, 0))
points(x, hx_f(x))
```



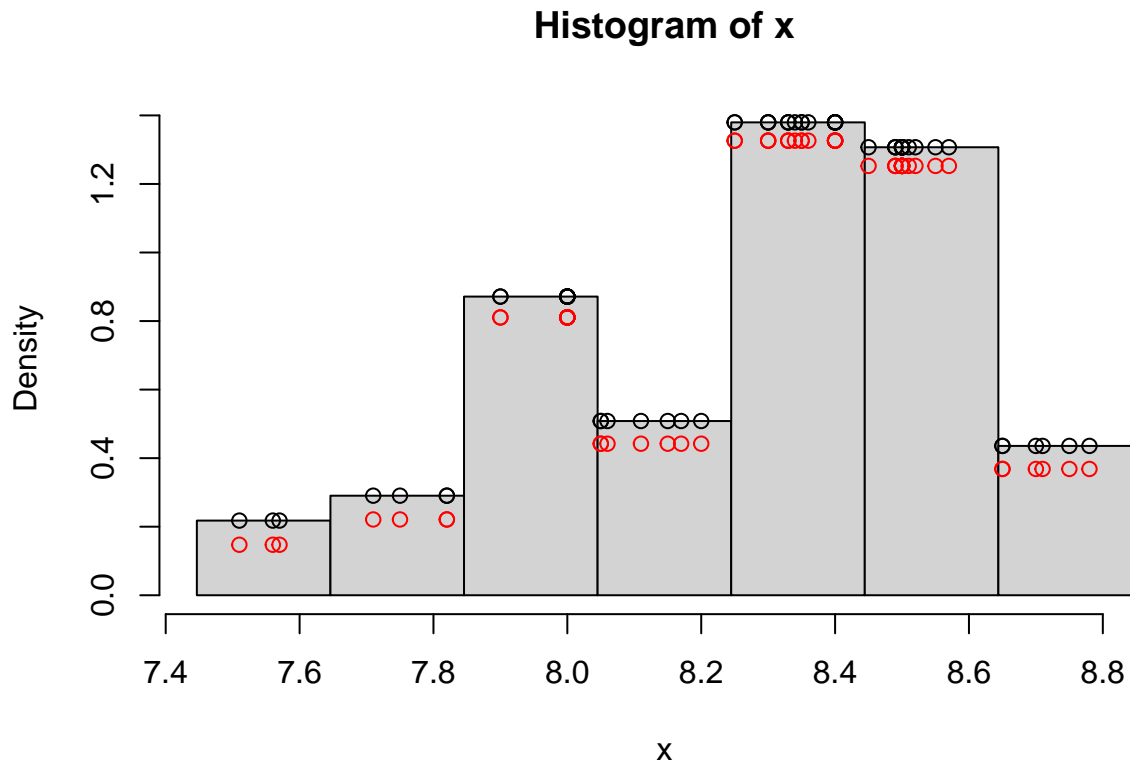
3. Use the formula you have found before relating $\hat{f}_{hist}(x_i)$ and $\hat{f}_{hist,(-i)}(x_i)$ to compute $\hat{f}_{hist,(-i)}(x)$, $i = 1, \dots, n$. Then add the points $(x_i, \hat{f}_{hist,(-i)}(x_i))$, $i = 1, \dots, n$, to the previous plot

```

hx_f2<-length(x)/(length(x)-1)*hx_f(x)-1/((length(x)-1)*(hx$breaks[2]-hx$breaks[1]))

A <- min(x)-.05*diff(range(x))
Z <- max(x)+.05*diff(range(x))
nbr <- 7
hx <- hist(x,breaks=seq(A,Z,length=nbr+1),freq=F)
hx_f <- stepfun(hx$breaks,c(0,hx$density,0))
points(x, hx_f(x))
points(x, hx_f2, col="red")

```



4. Compute the leave-one-out log-likelihood function corresponding to the previous histogram, at which $nbr=7$ has been used

```
looCV_log_lik_7 <- sum(log(hx_f2))
looCV_log_lik_7
```

```
## [1] -16.58432
```

5. **Choosing nbr by leave-one-out Cross Validation (looCV).** Consider now the set $seq(1,15)$ as possible values for nbr , the number of intervals of the histogram. For each of them compute the leave-one-out log-likelihood function ($looCv_log_lik$) for the corresponding histogram. Then plot the values of $looCv_log_lik$ against the values of nbr and select the optimal value of nbr as that at which $looCv_log_lik$ takes its maximum. Finally, plot the histogram of x using the optimal value of nbr

#sum of the product of the hx_f2 vector plot histograms for different number of breaks nbr

```
log_liks = list()
```

```
for (nbr in c(1:15)){
  #A <- min(hx_f2)-.05*diff(range(hx_f2))
  #Z <- max(hx_f2)+.05*diff(range(hx_f2))
  hx_i <- hist(x,breaks=seq(A,Z,length=nbr+1),freq=F, plot = FALSE)
  hx_f_i <- stepfun(hx_i$breaks,c(0,hx_i$density,0))
  n <- length(x)
```

```

b <- hx_i$breaks[2]-hx_i$breaks[1]
hx_f2_i <- n/(n-1)*hx_f_i(x) - 1/((n-1)*b)
print(min(hx_f2_i))

hx_i <- hist(x,breaks=seq(A,Z,length=nbr+1),freq=F)
points(x, hx_f_i(x))
points(x, hx_f2_i, col='red')

looCV_log_lik <- sum(log(hx_f2_i))
log_liks <- append(log_liks, looCV_log_lik)
}

```

```

## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :
## argument 'freq' is not made use of

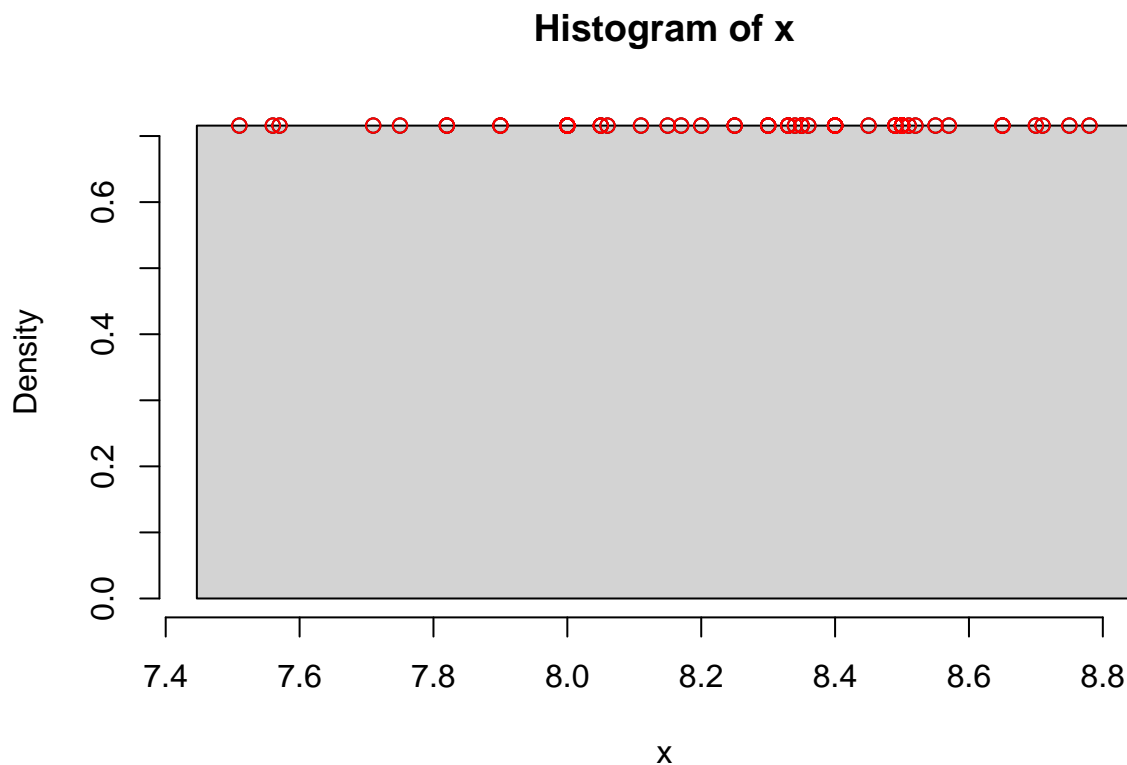
```

```
## [1] 0.7158196
```

```

## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :
## argument 'freq' is not made use of

```

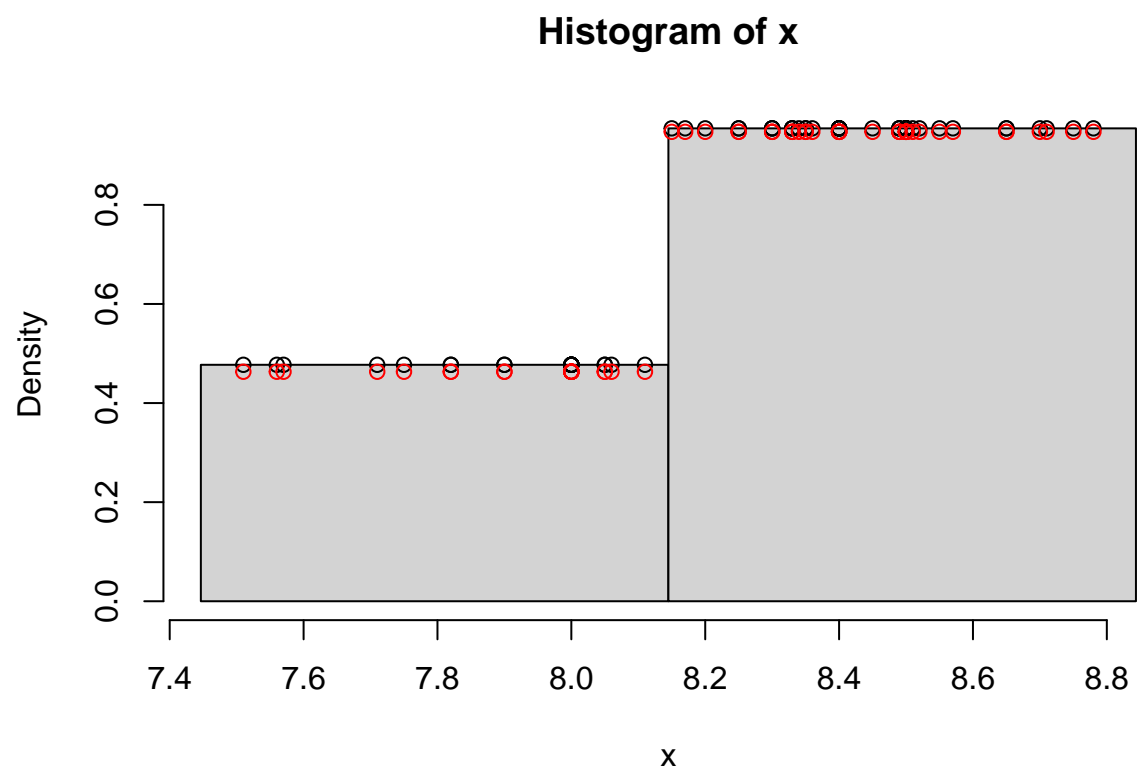


```
## [1] 0.4631774
```

```

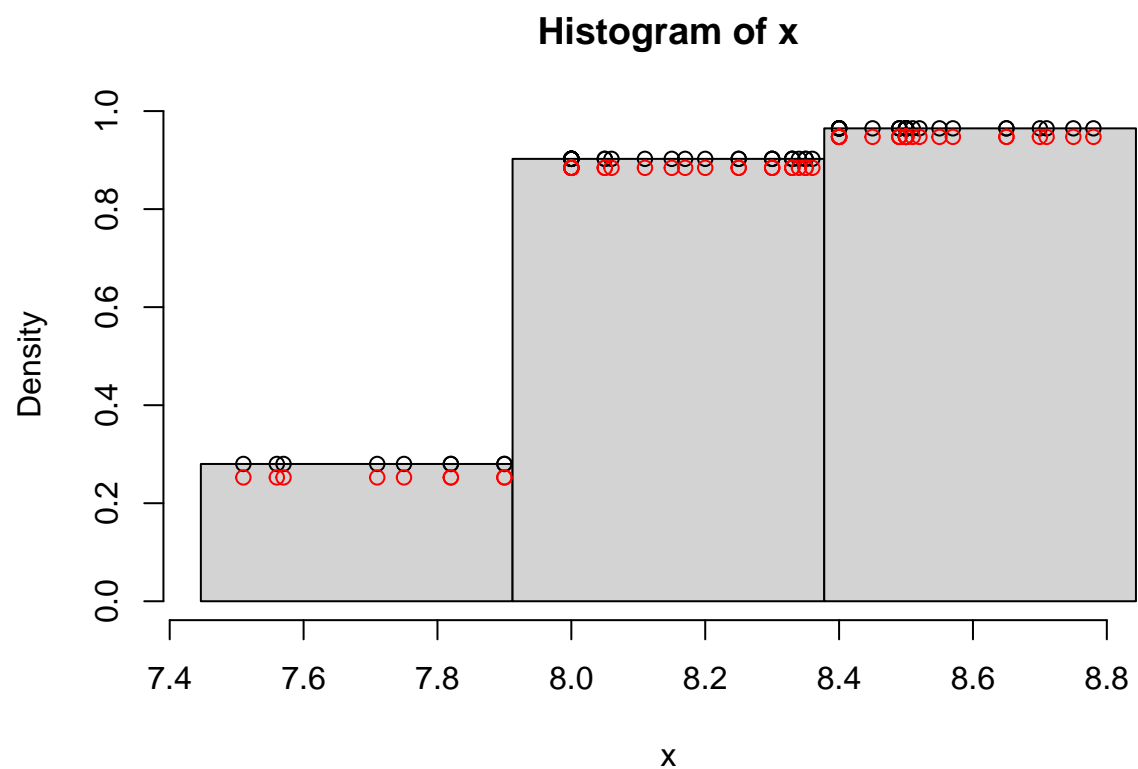
## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :
## argument 'freq' is not made use of

```



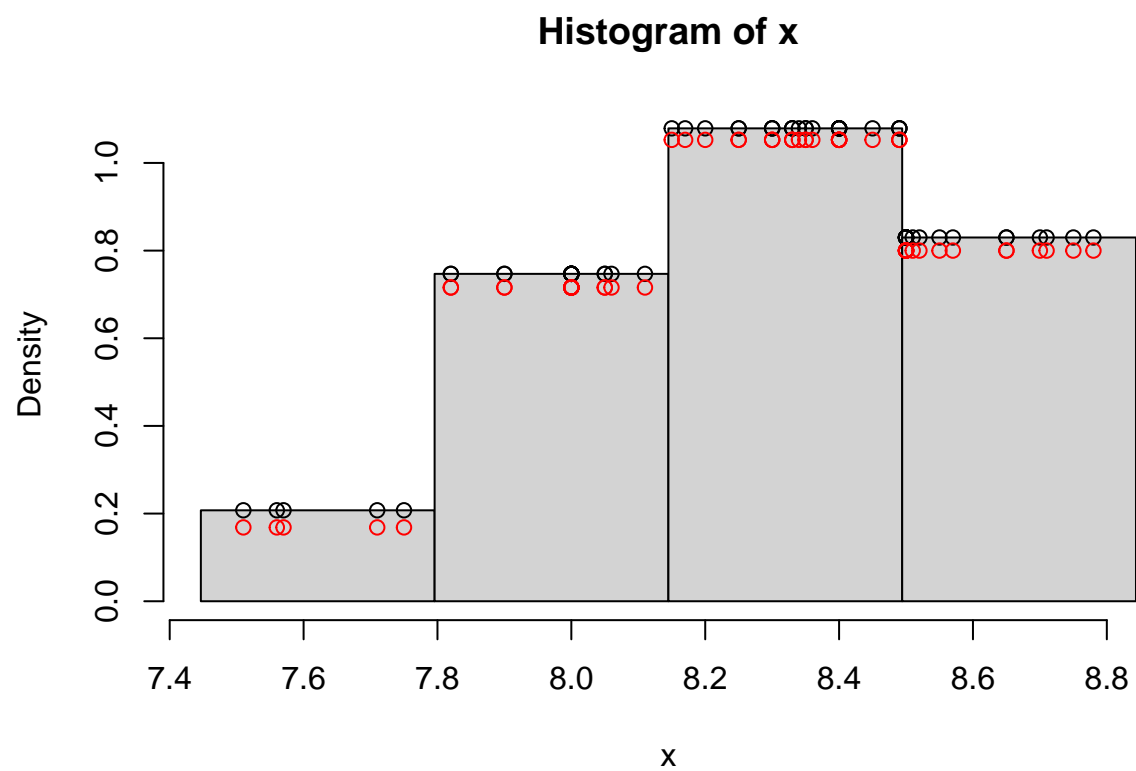
```
## [1] 0.2526422
```

```
## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :  
## argument 'freq' is not made use of
```



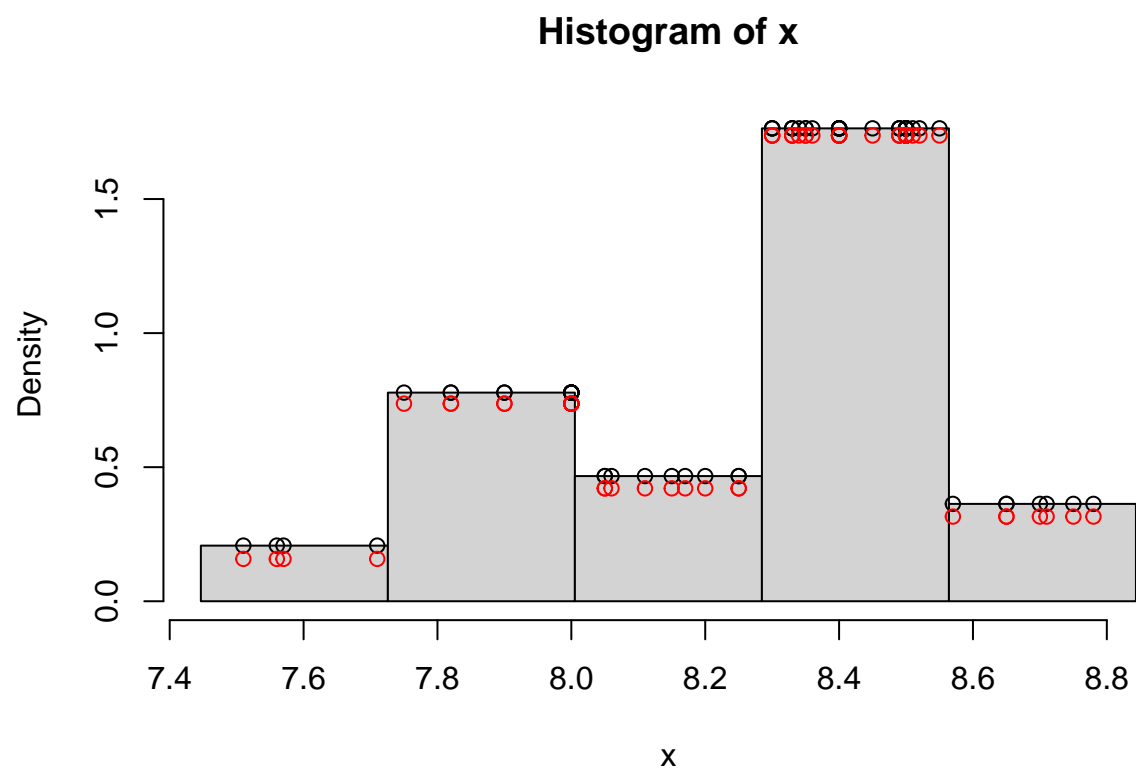
```
## [1] 0.1684281
```

```
## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :  
## argument 'freq' is not made use of
```



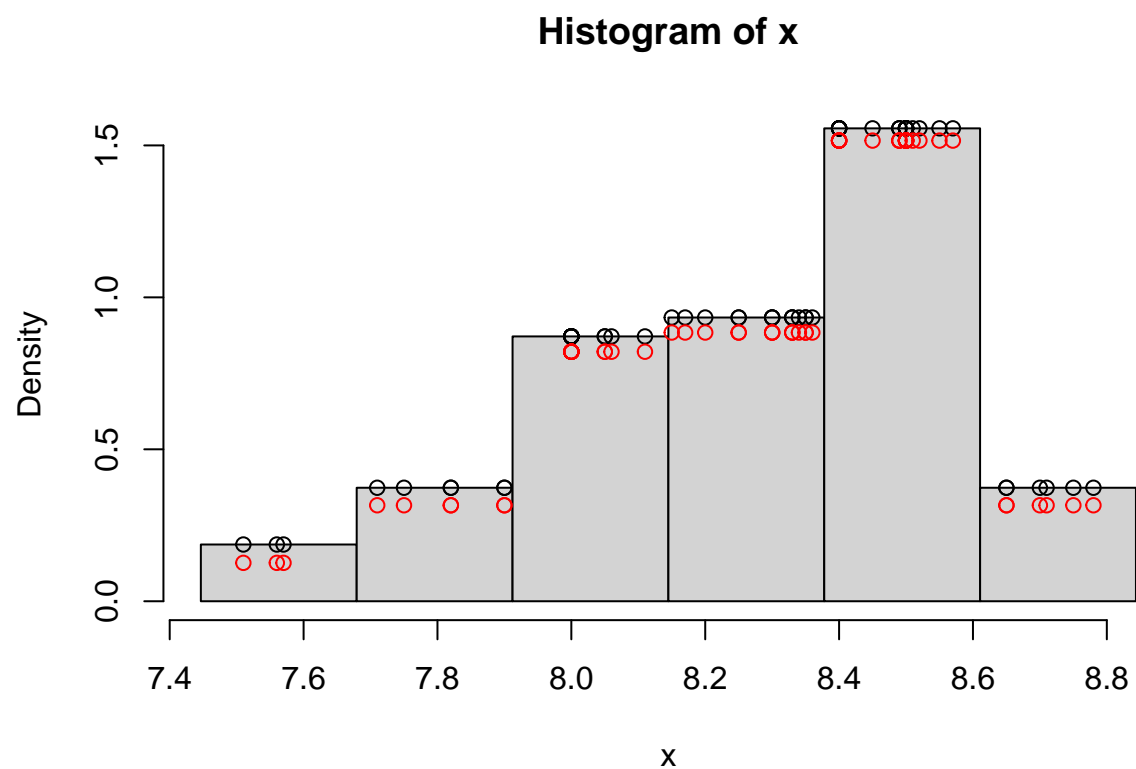
```
## [1] 0.1579014
```

```
## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :  
## argument 'freq' is not made use of
```



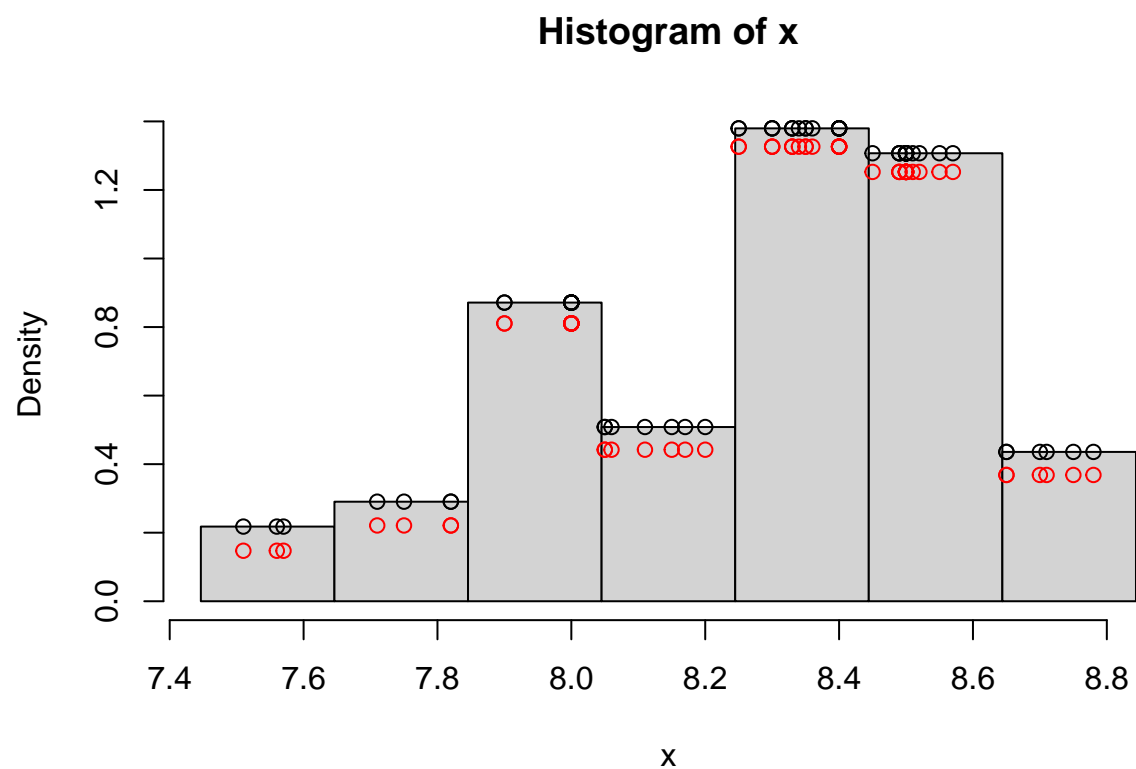
```
## [1] 0.1263211
```

```
## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :  
## argument 'freq' is not made use of
```

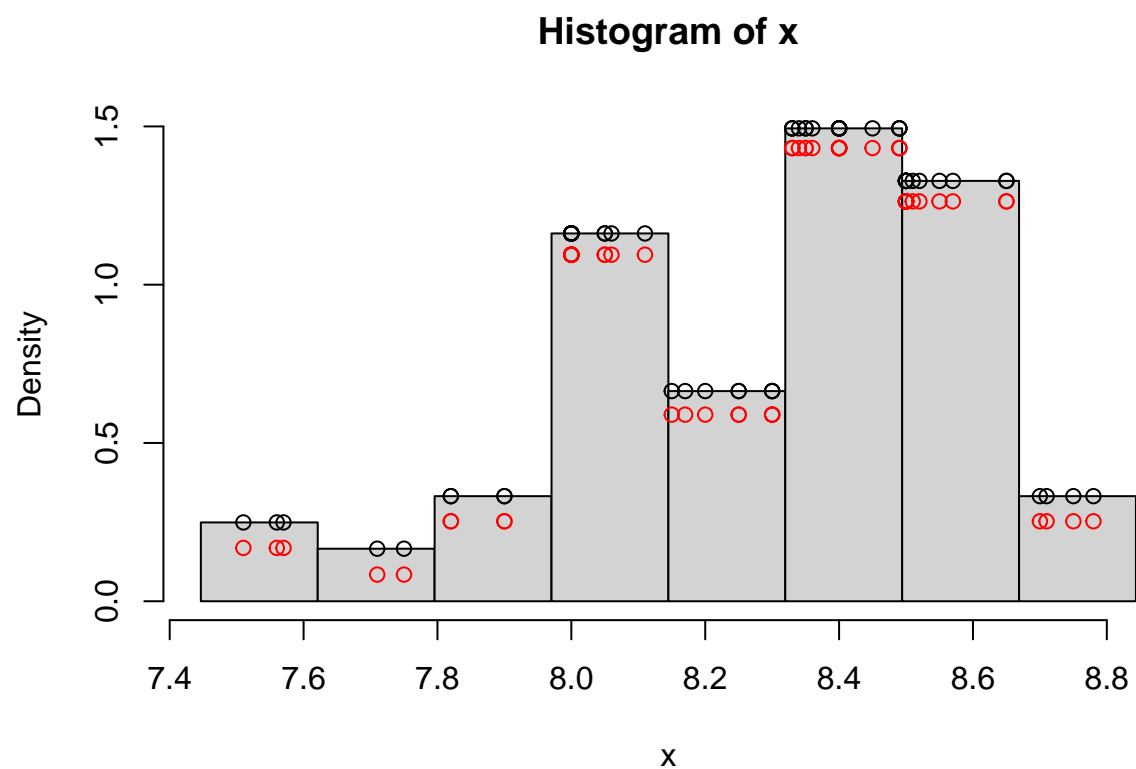
```
## [1] 0.1473746
```

```
## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :  
## argument 'freq' is not made use of
```



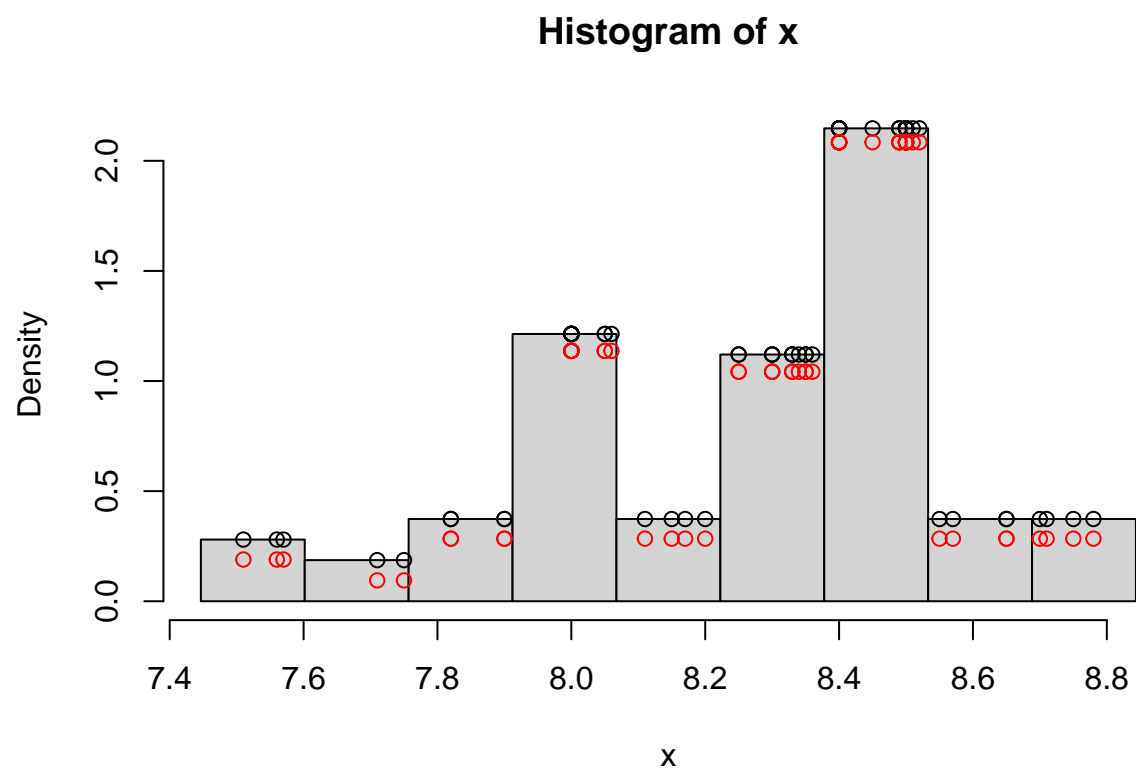
```
## [1] 0.08421407
```

```
## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :  
## argument 'freq' is not made use of
```



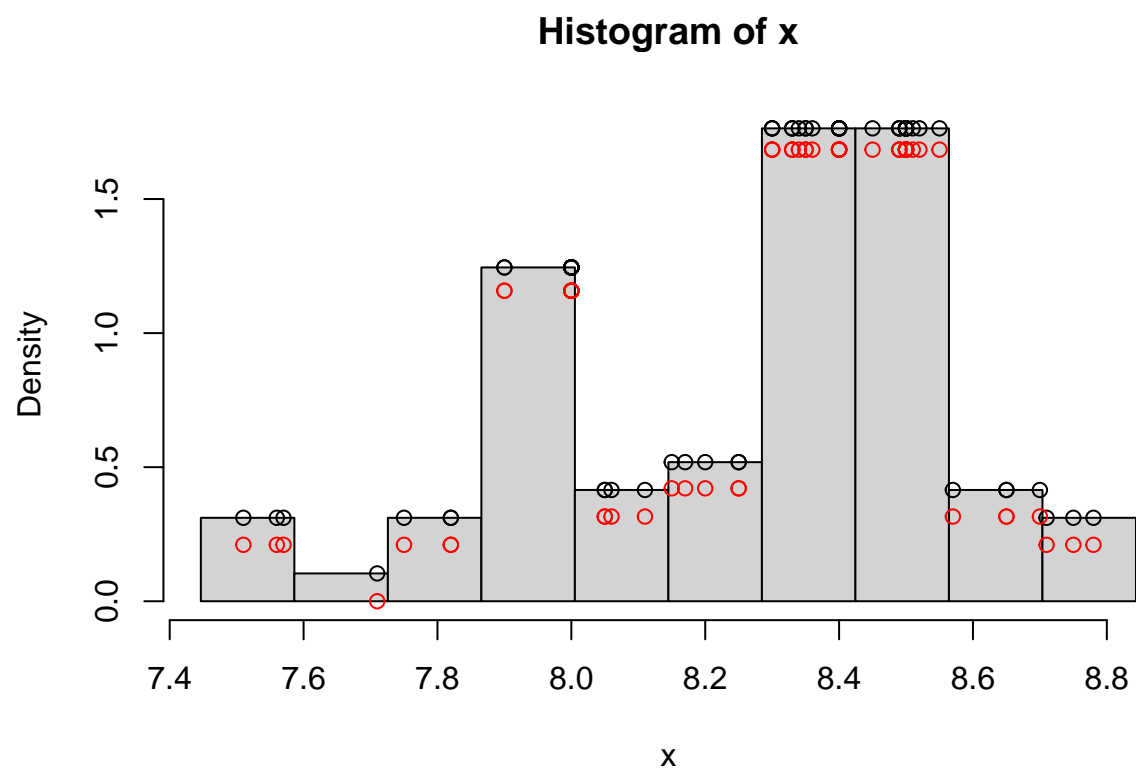
```
## [1] 0.09474083
```

```
## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :  
## argument 'freq' is not made use of
```



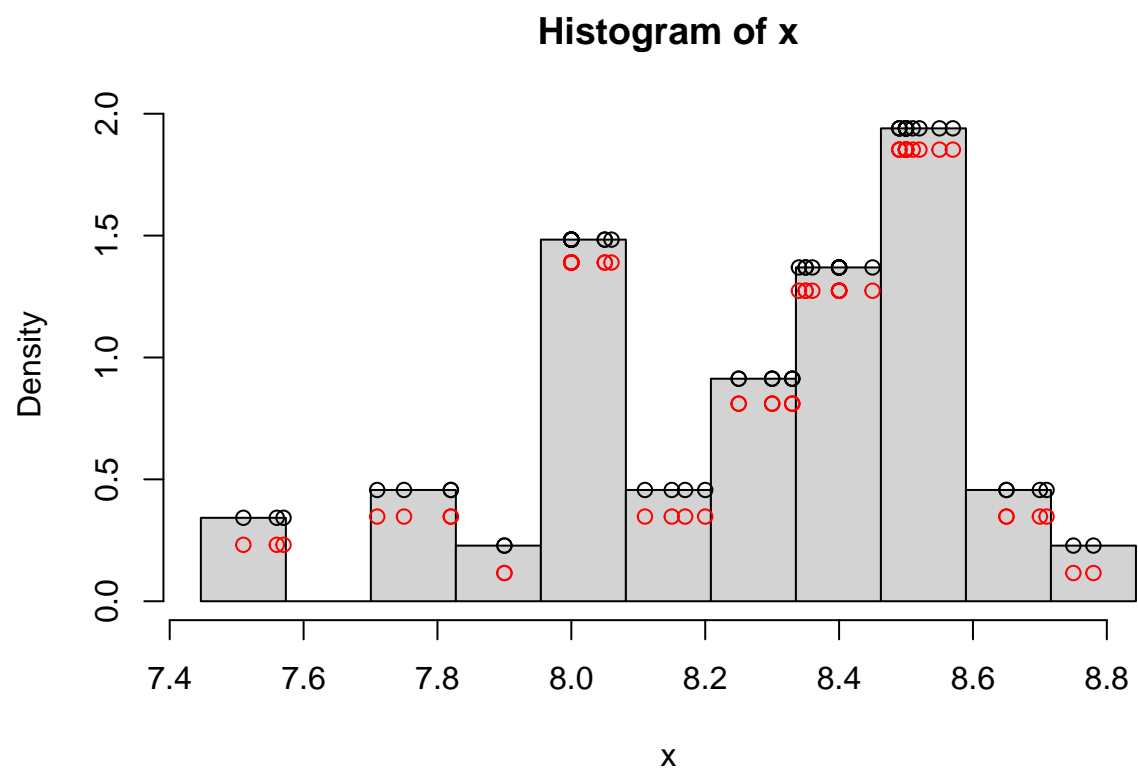
```
## [1] 6.661338e-16
```

```
## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :  
## argument 'freq' is not made use of
```



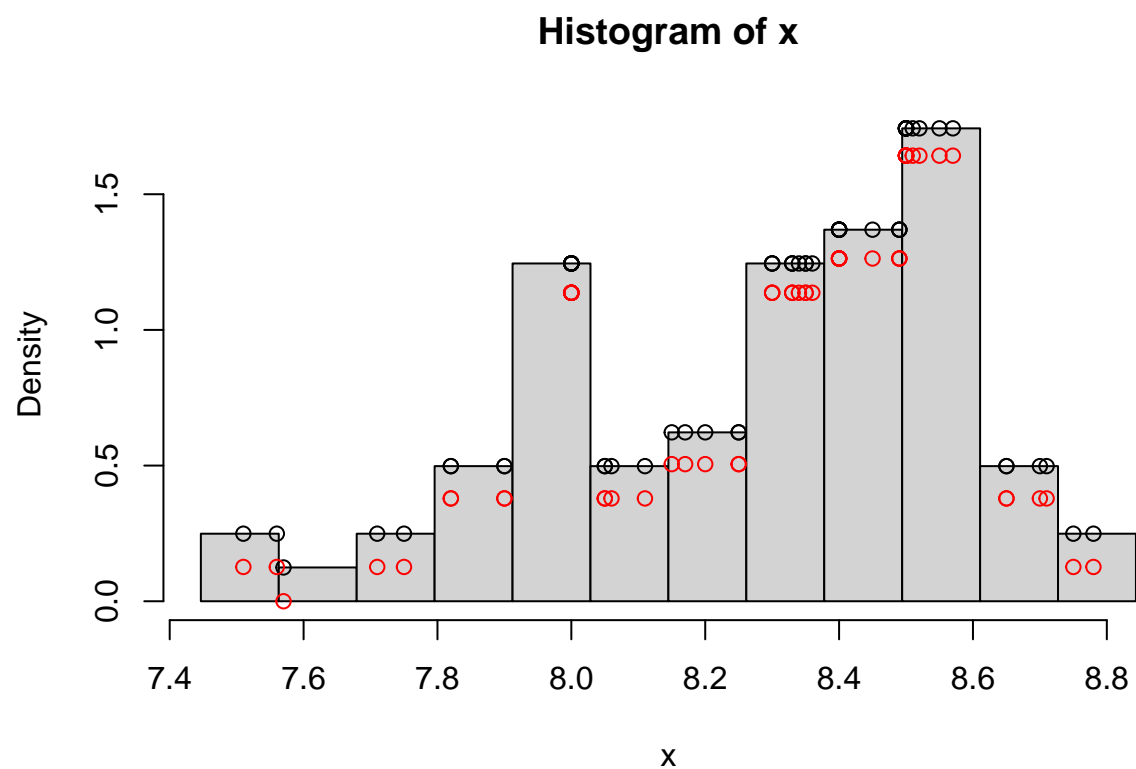
```
## [1] 0.1157943
```

```
## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :
## argument 'freq' is not made use of
```



```
## [1] 9.714451e-16
```

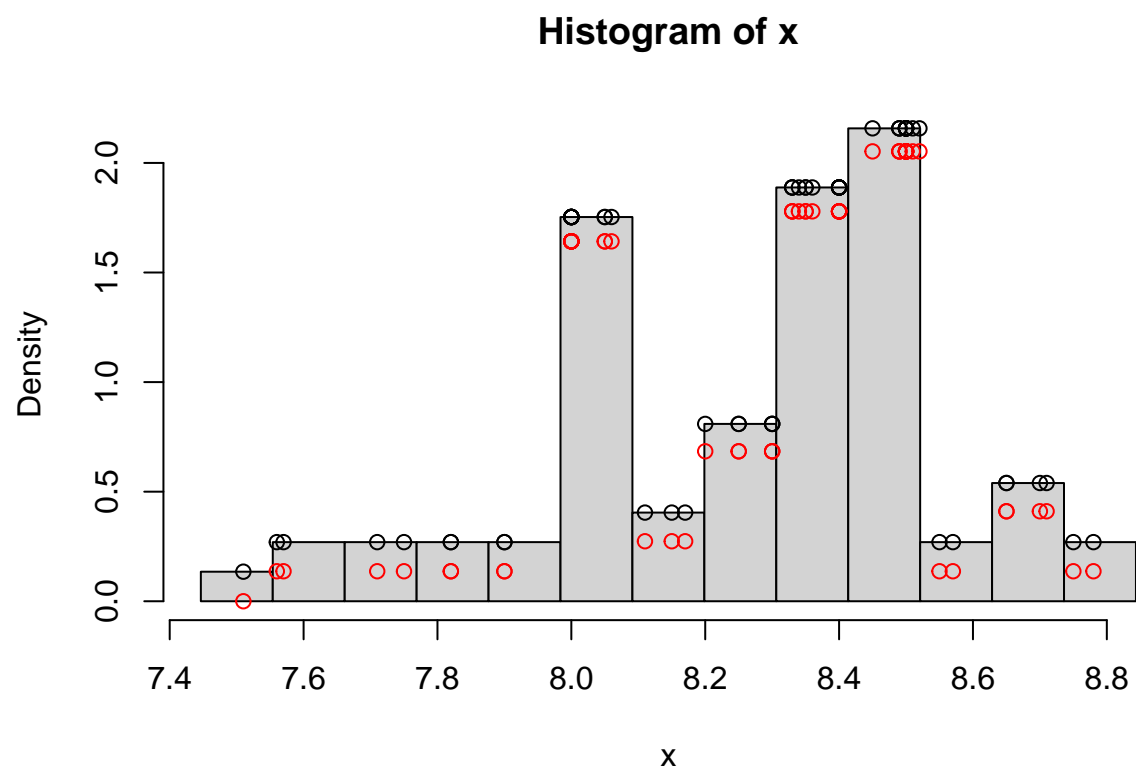
```
## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :  
## argument 'freq' is not made use of
```



```
## [1] -2.775558e-17
```

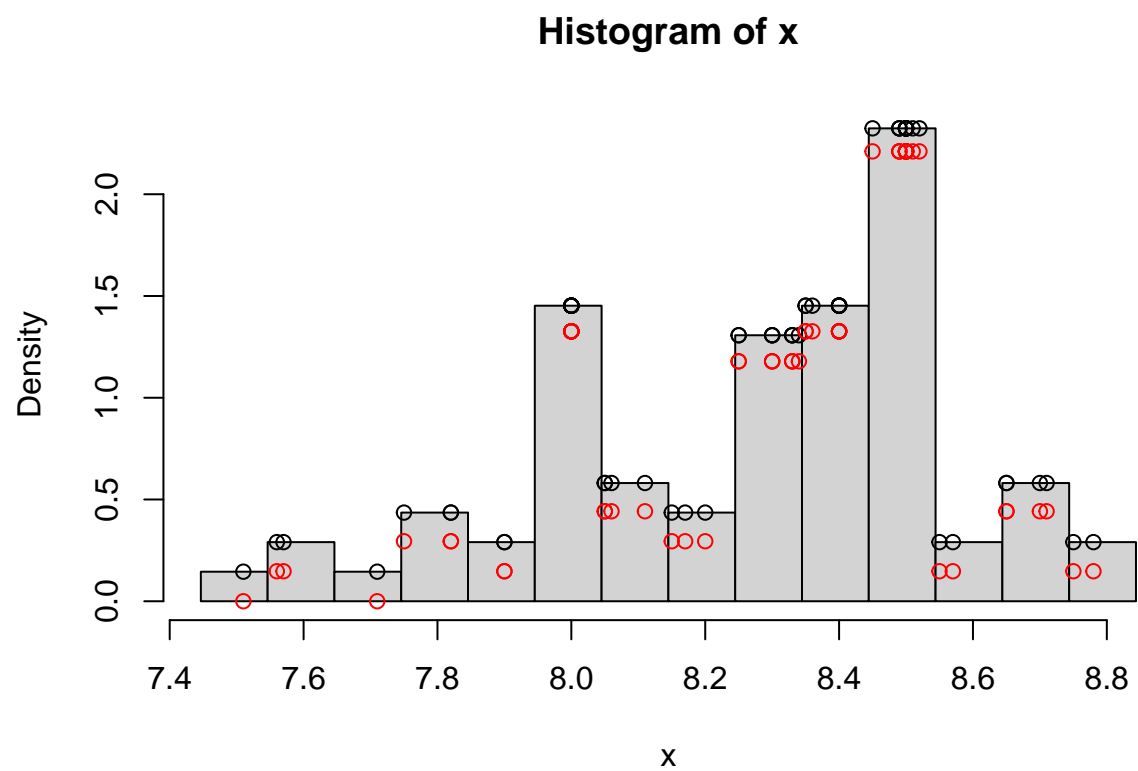
```
## Warning in log(hx_f2_i): Se han producido NaNs
```

```
## Warning in log(hx_f2_i): argument 'freq' is not made use of
```



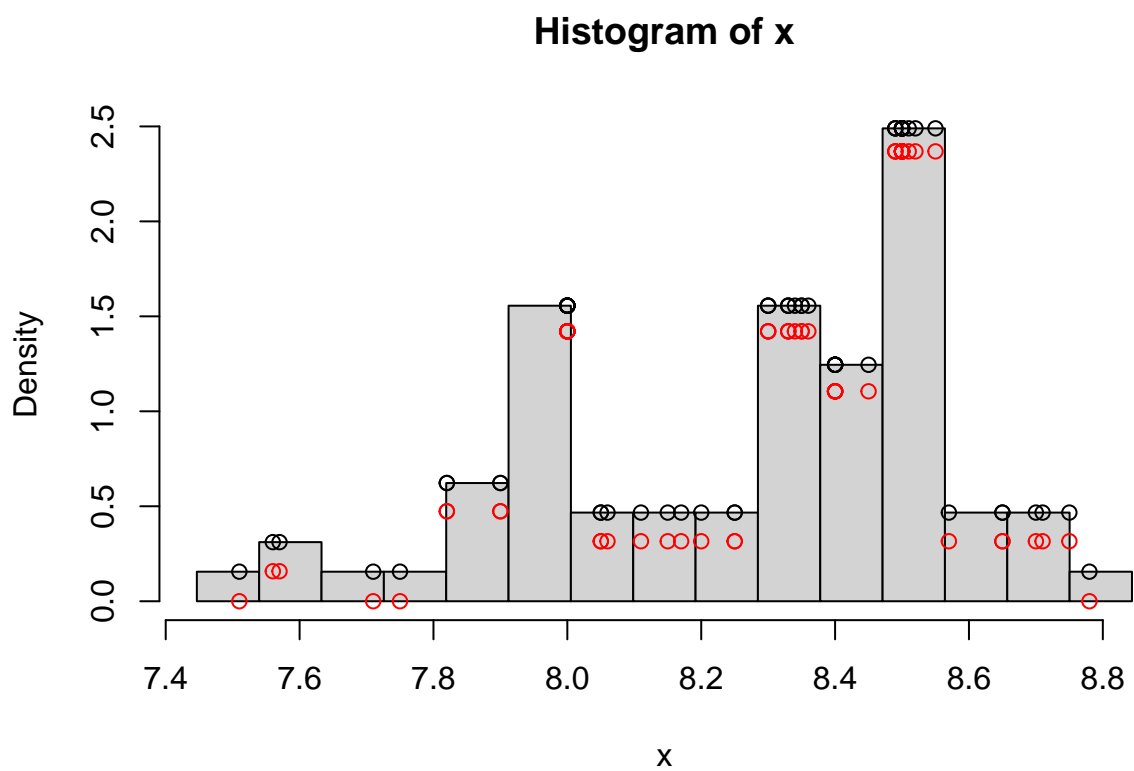
```
## [1] 0
```

```
## Warning in hist.default(x, breaks = seq(A, Z, length = nbr + 1), freq = F, :
## argument 'freq' is not made use of
```

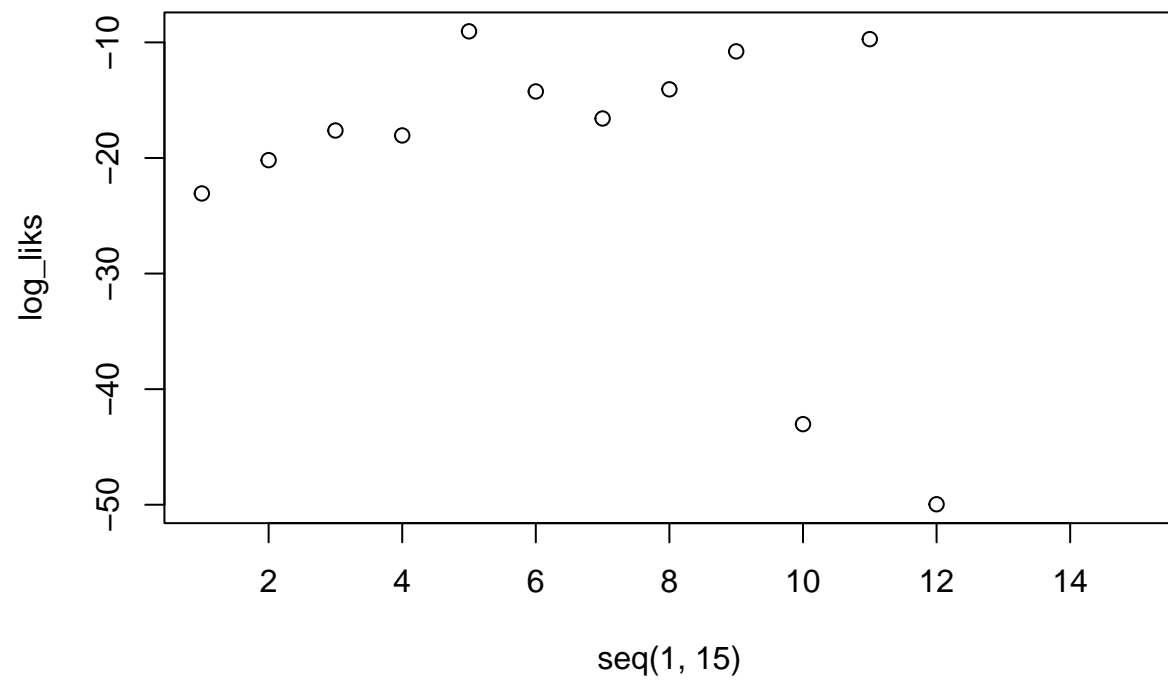



```
## [1] -1.498801e-15
```

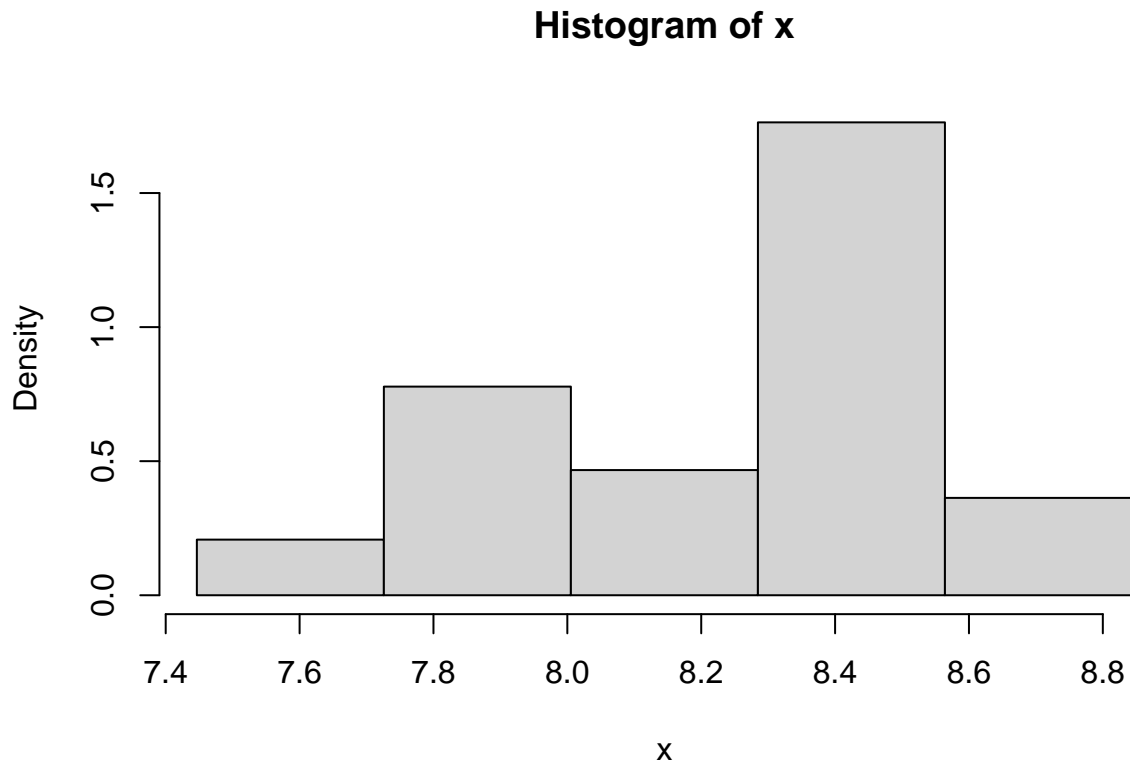
```
## Warning in log(hx_f2_i): Se han producido NaNs
```



```
plot(seq(1, 15), log_lik)
```



```
nbr_opt <- which.max(log_lik)  
hist(x,breaks=seq(A,Z,length=nbr_opt+1),freq=F)
```

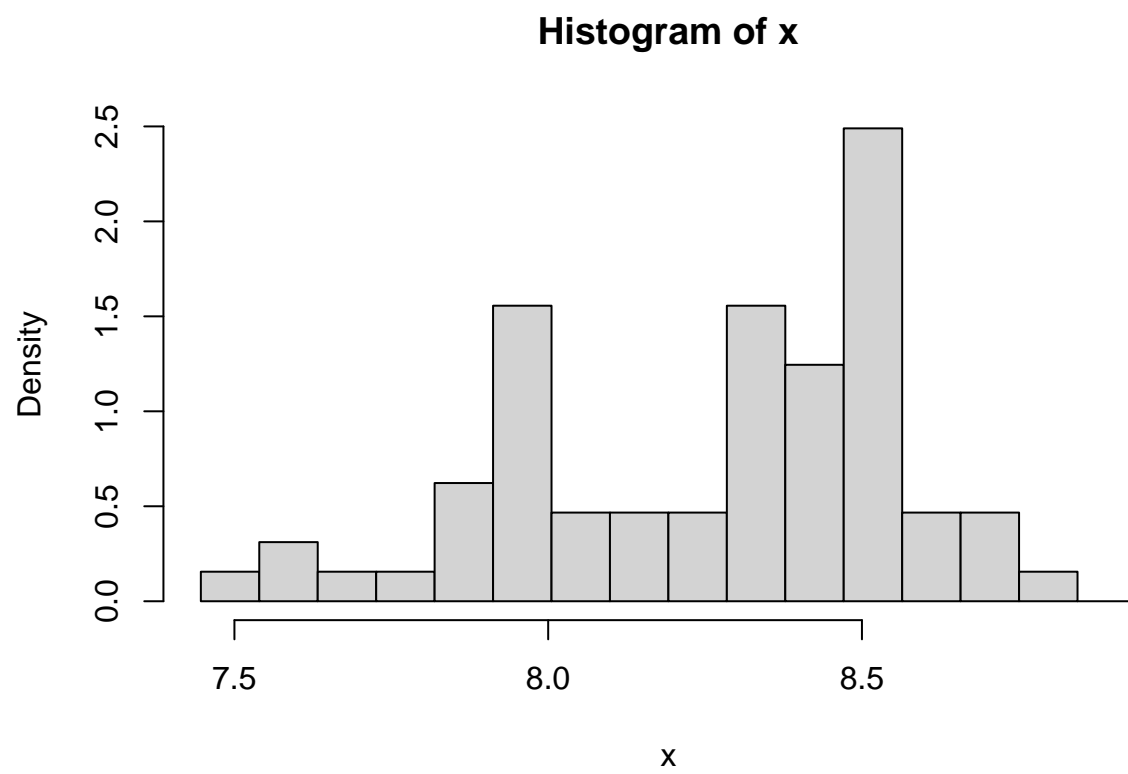


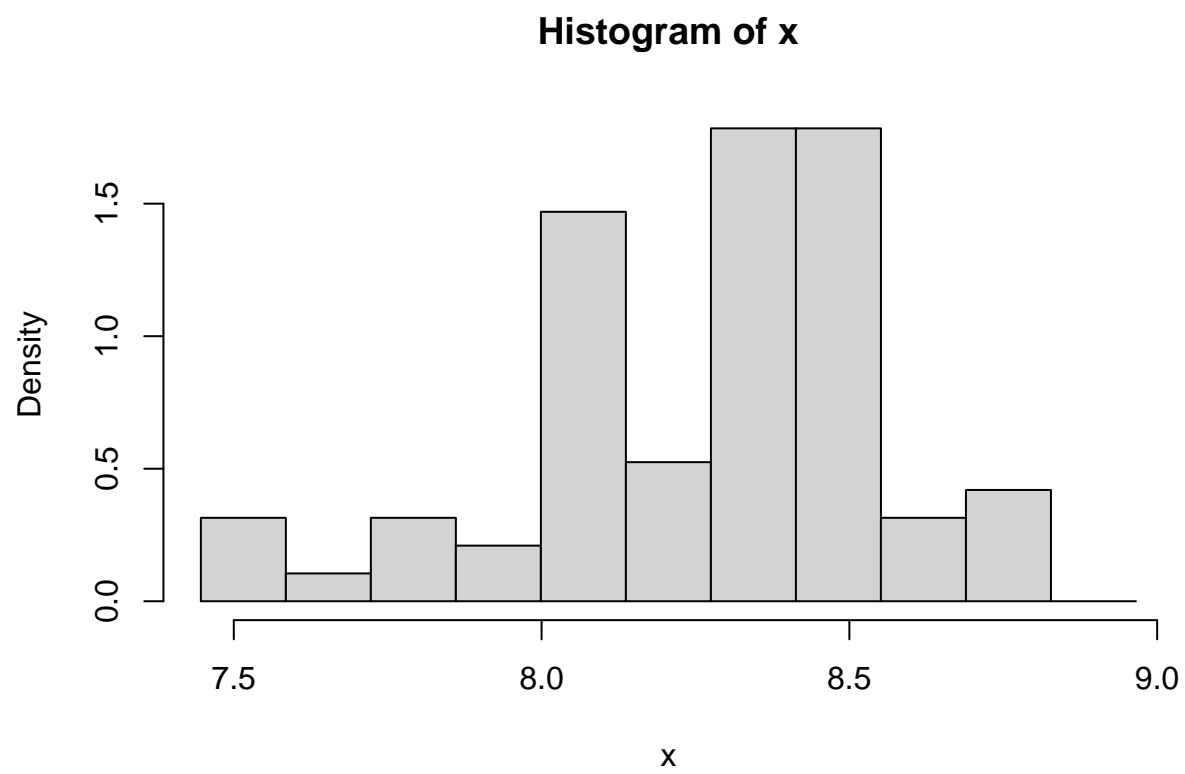
```
# TODO: - avoid production of NaNs,
#        - is it a problem that 'freq' is not used for some cases?
```

6. **Choosing b by looCV.** Let b be the common width of the bins of a histogram. Consider the set $\text{seq}((Z - A)/15, (Z - A)/1, \text{length} = 30)$ as possible values for b . Select the value of b maximizing the leave-one-out log-likelihood function, and plot the corresponding histogram

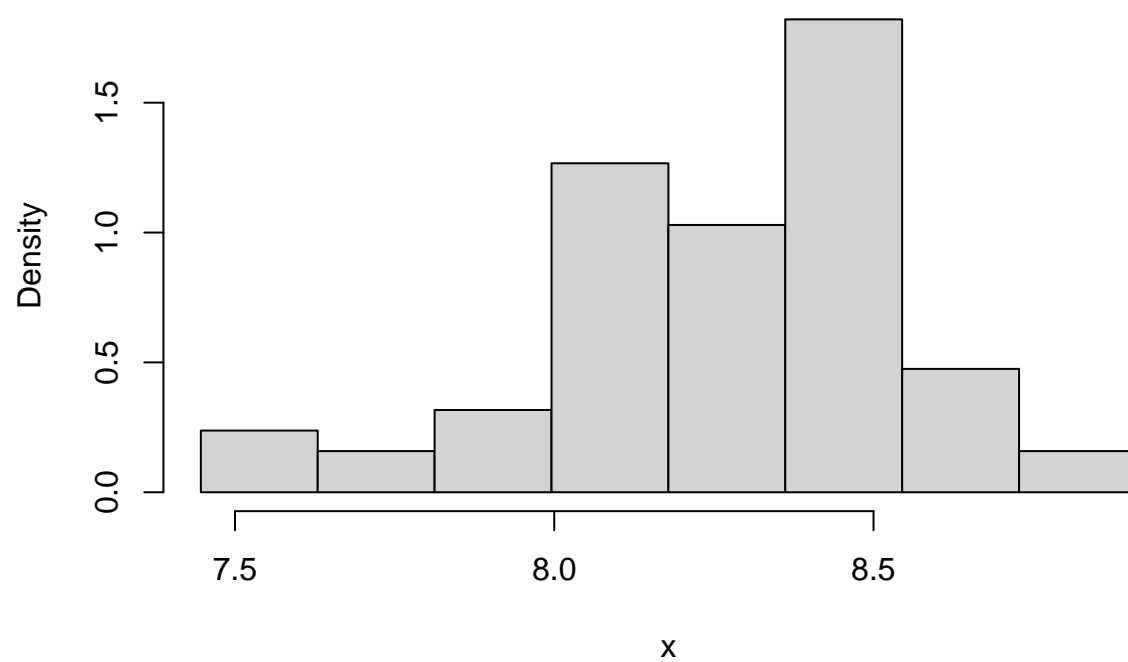
```
b_set <- seq((Z-A)/15, (Z-A), length=30)

for(b in b_set){
  hx <- hist(x, breaks=seq(A, Z+b, by=b), plot=F)
  plot(hx, freq = FALSE)
}
```

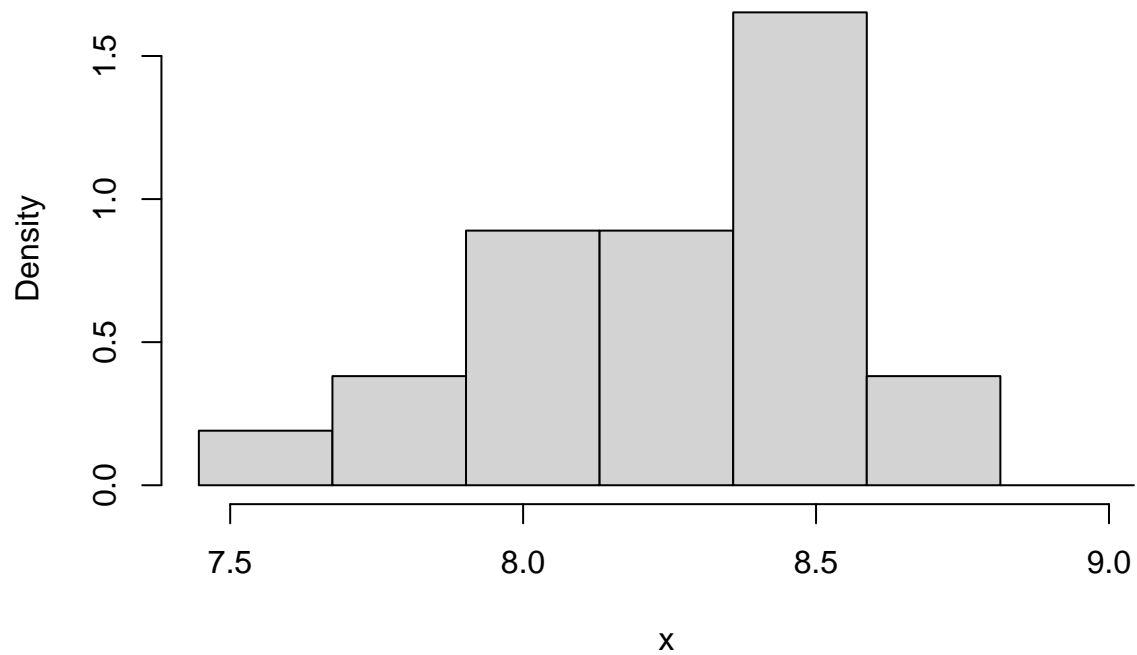




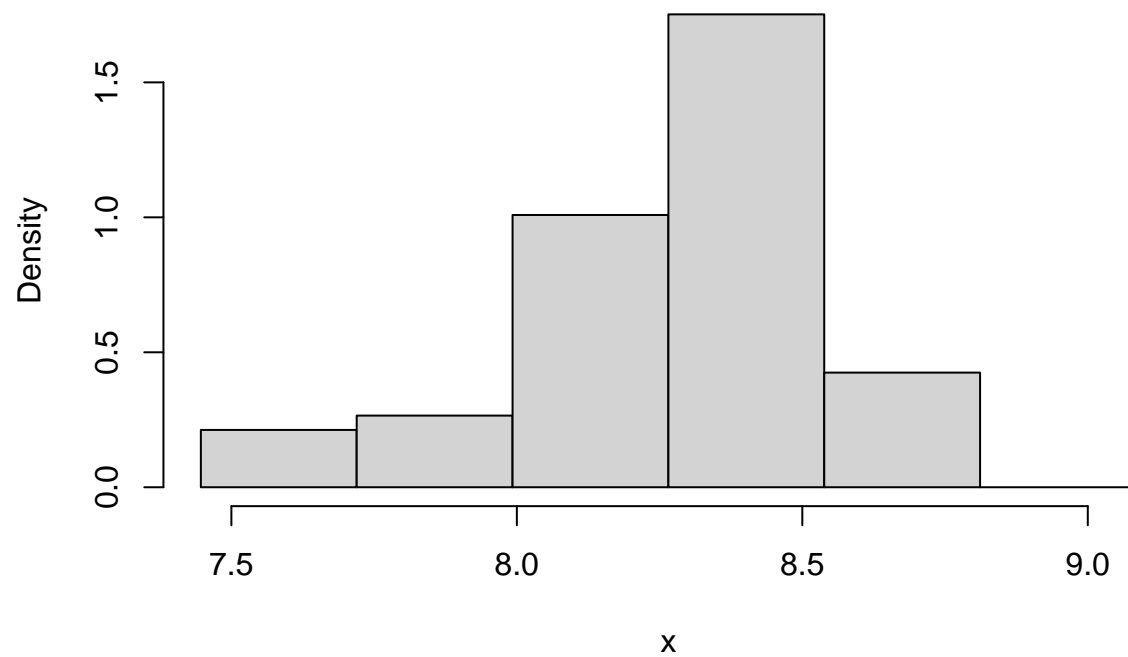
Histogram of x

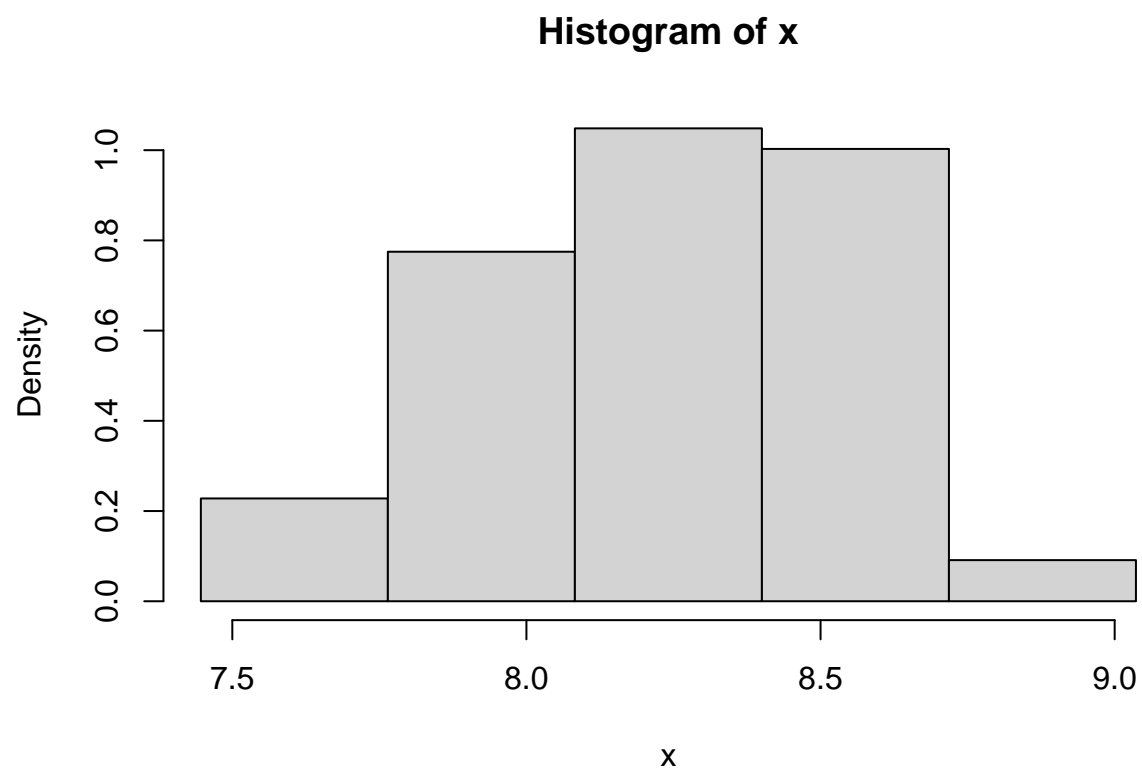


Histogram of x

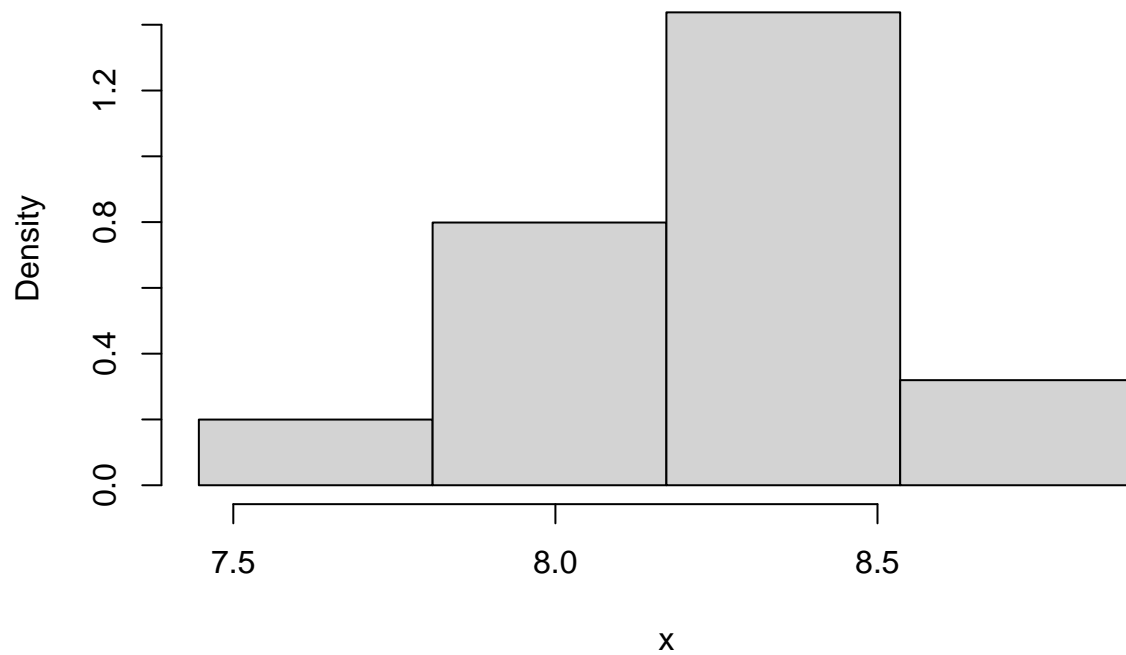


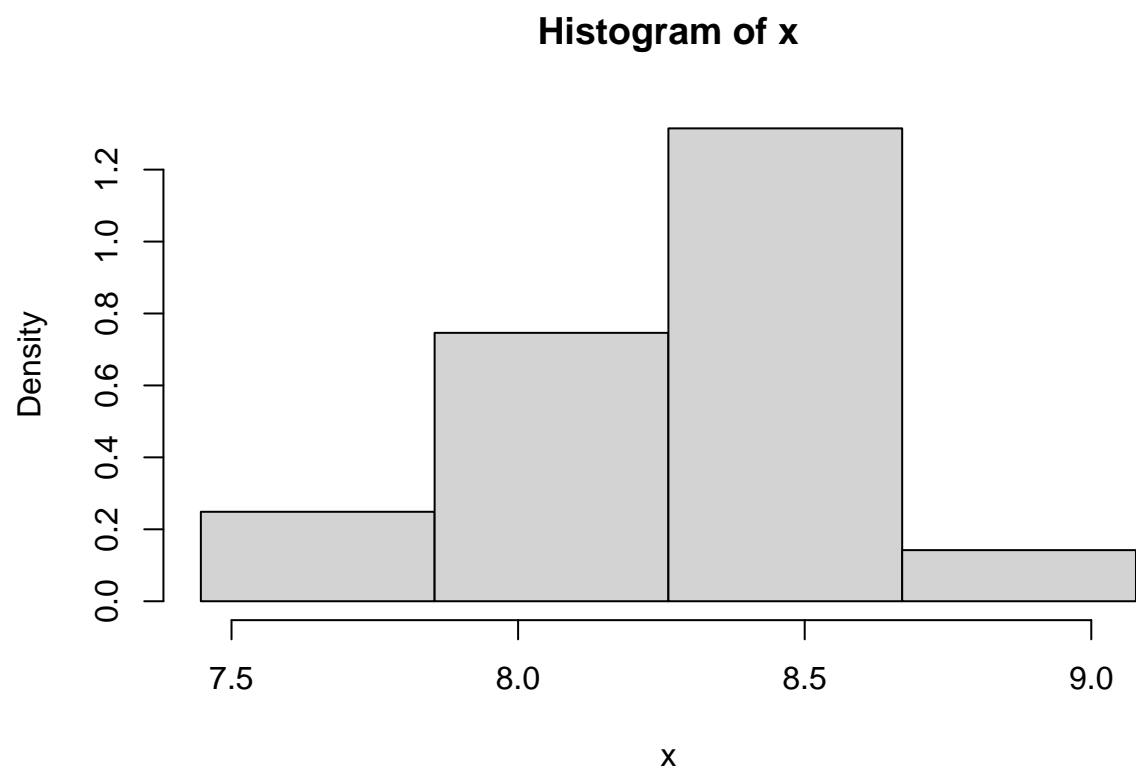
Histogram of x

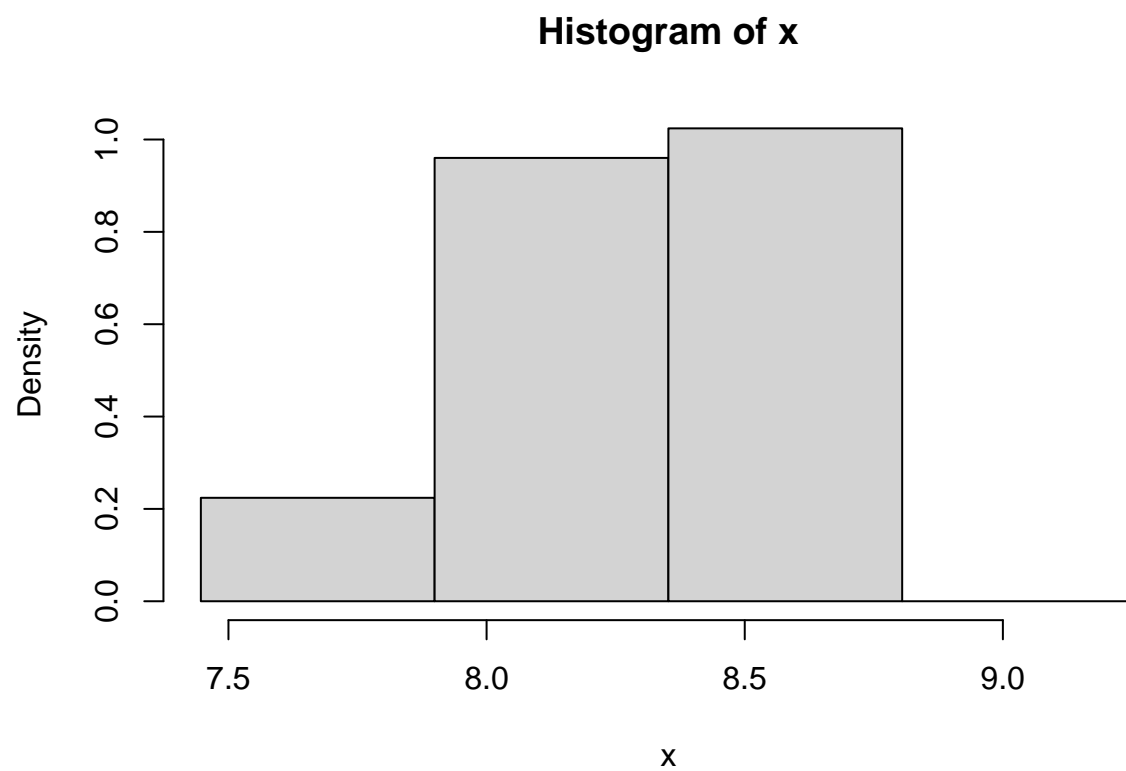


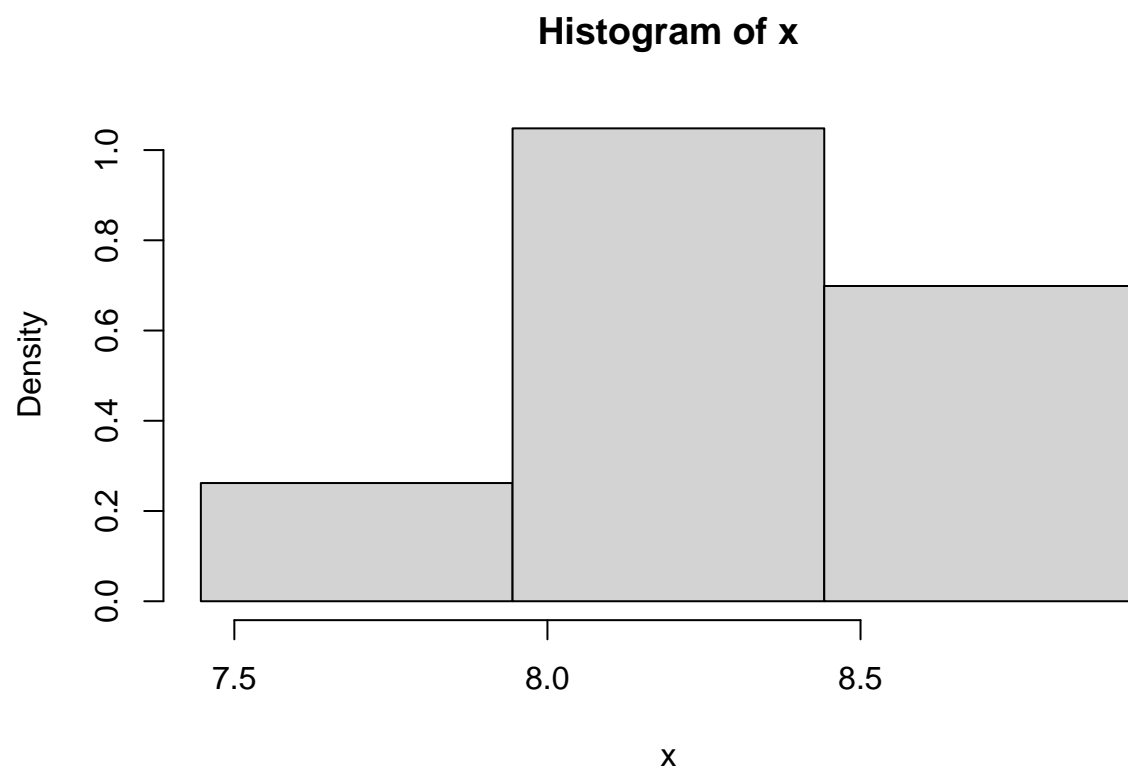


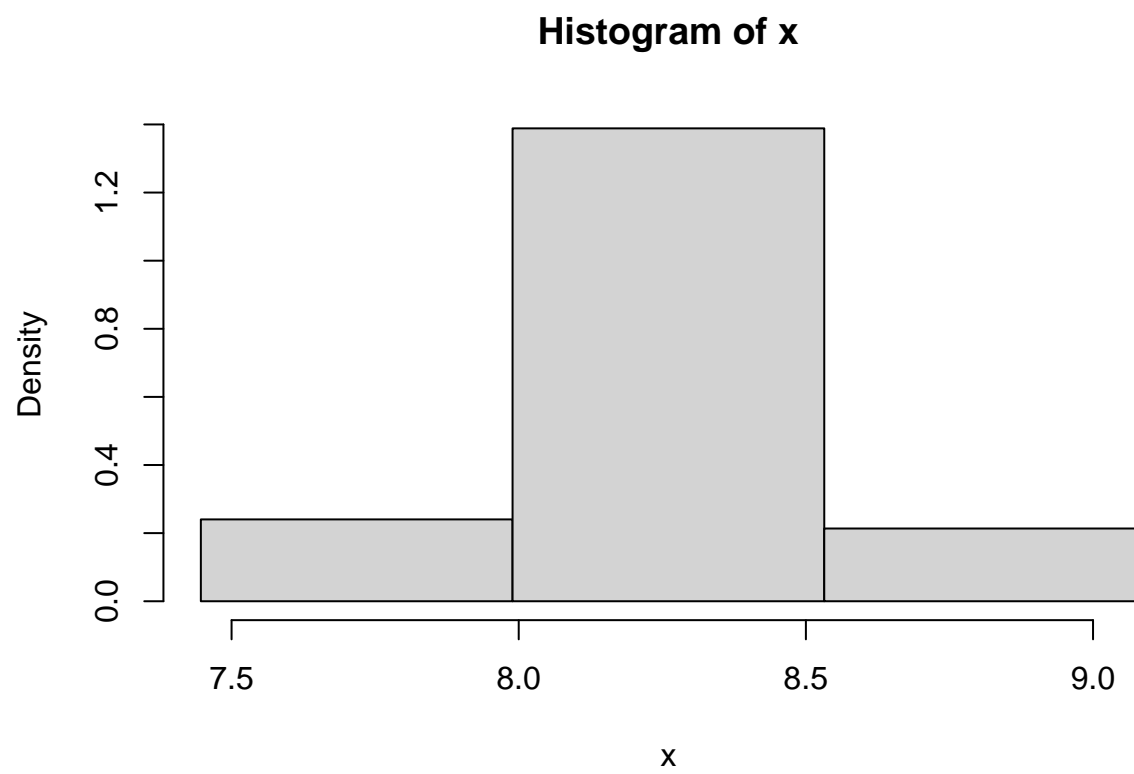
Histogram of x

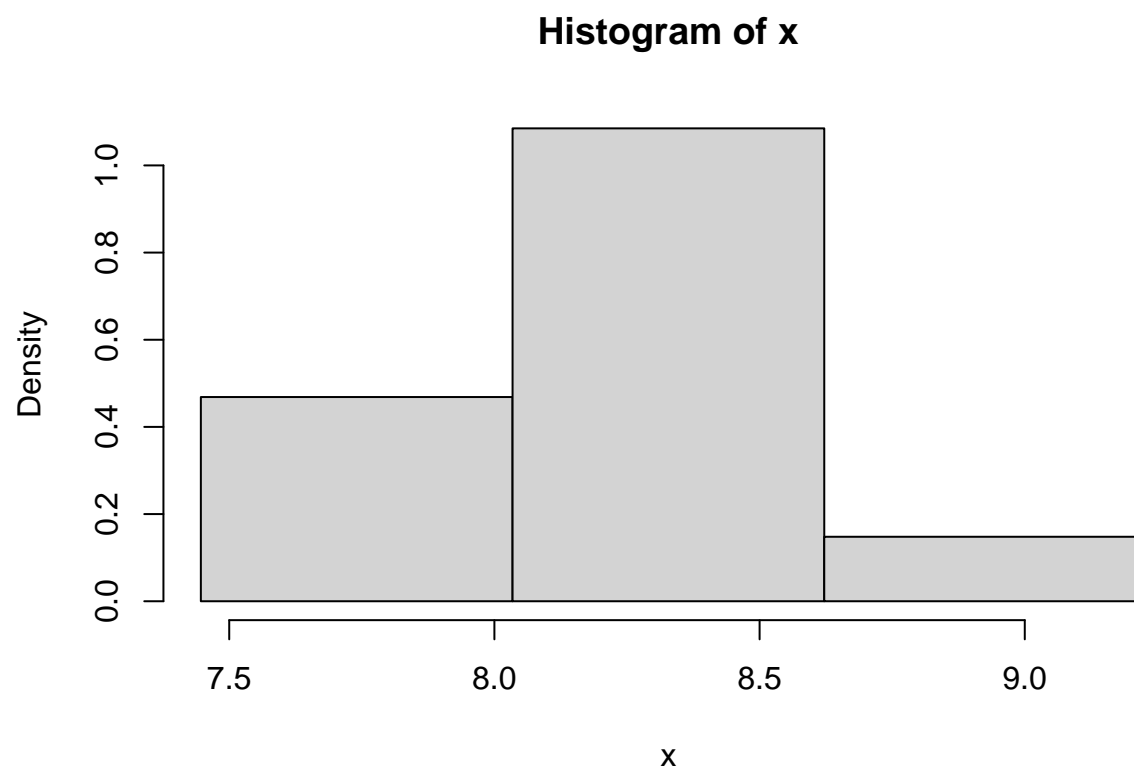


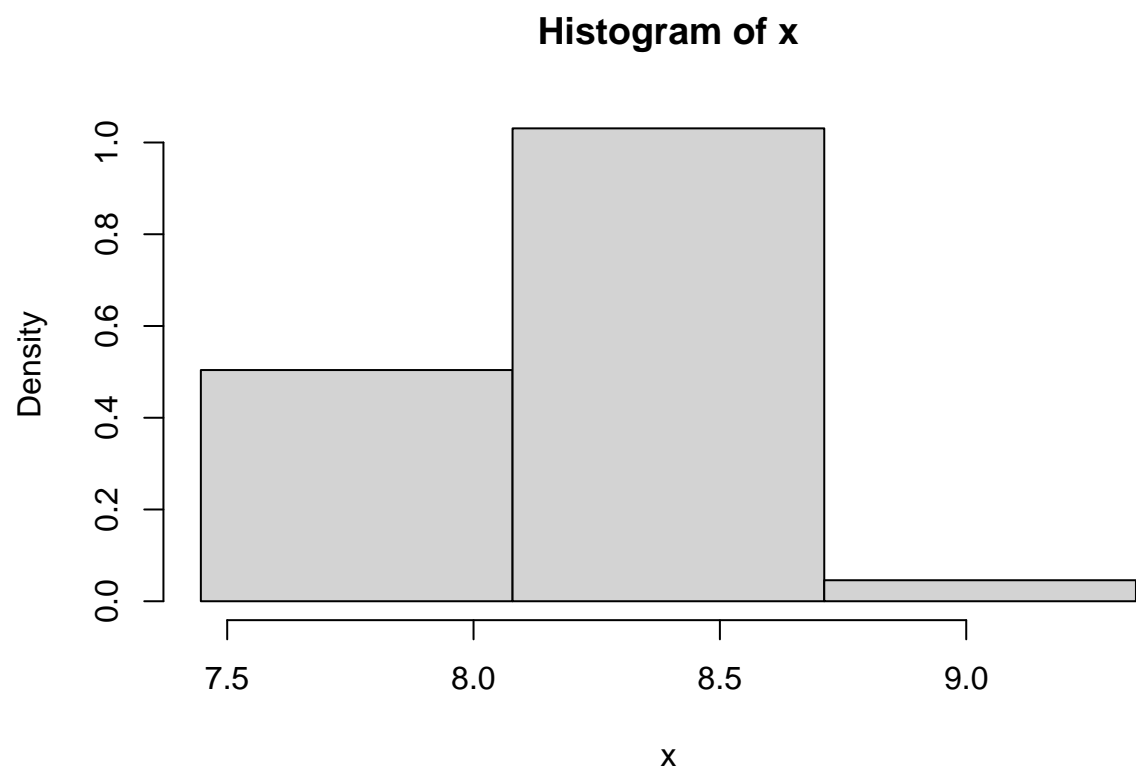


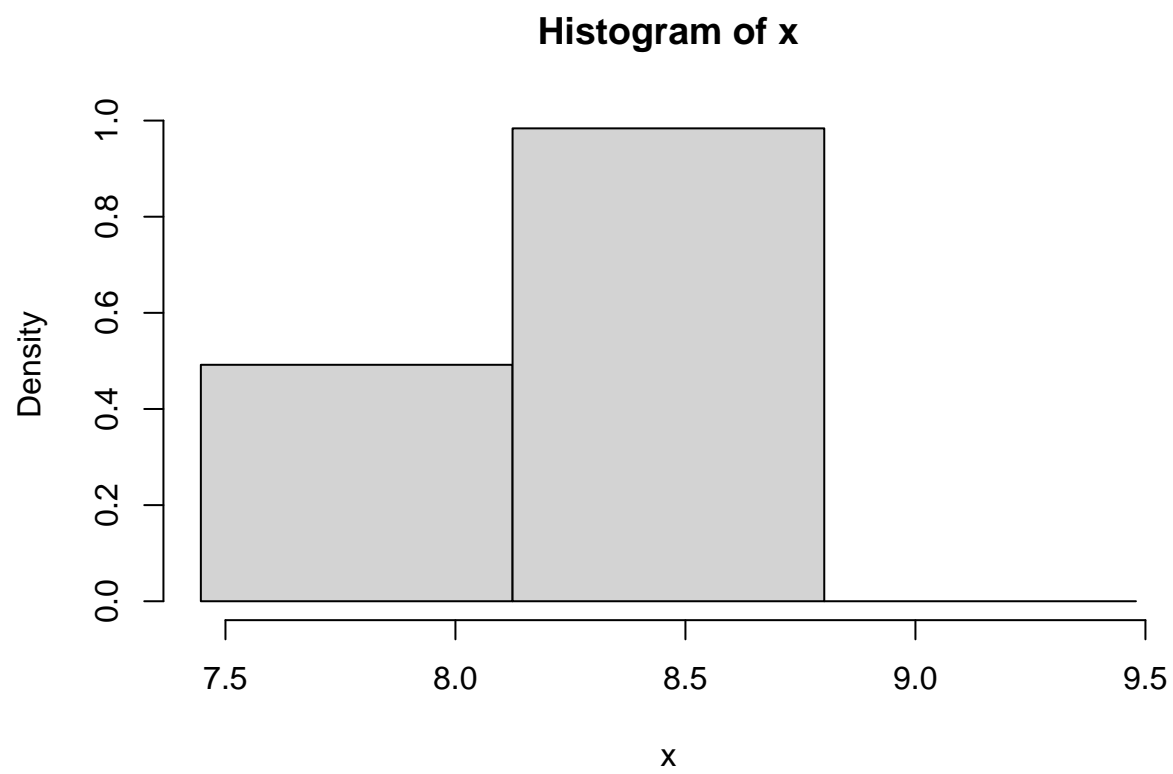




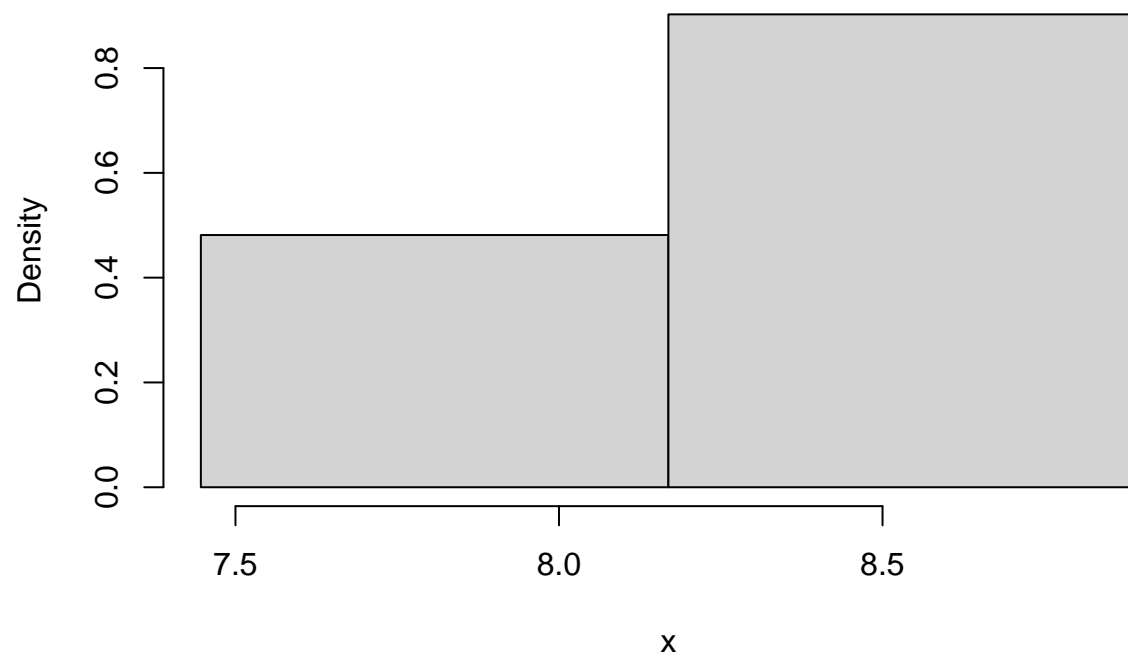




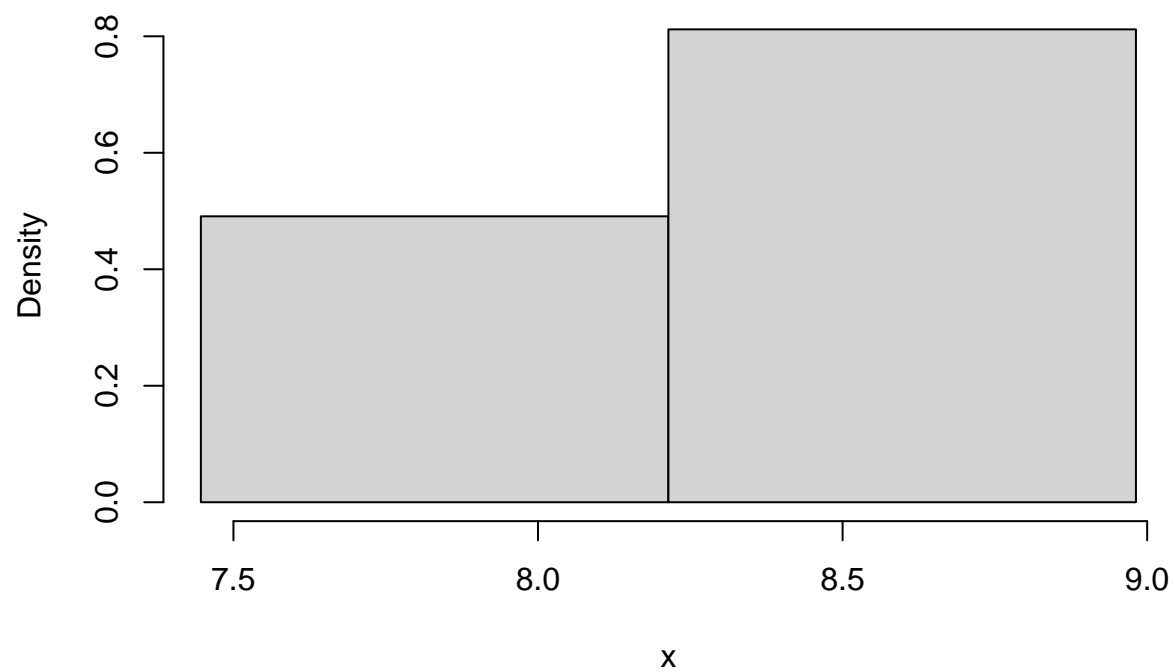




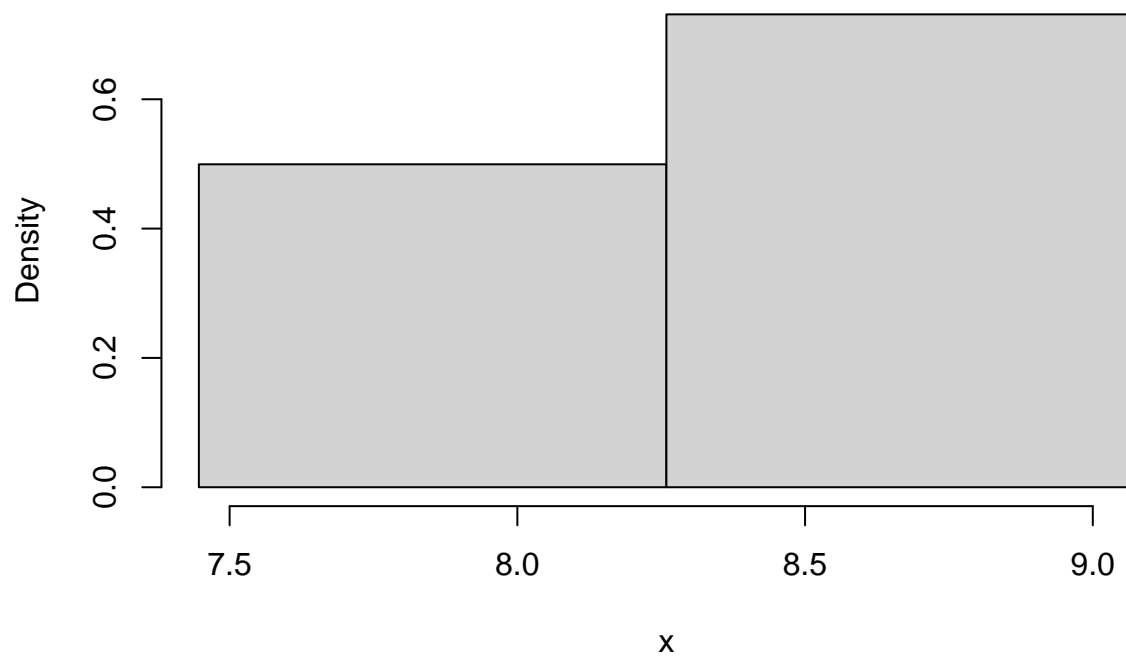
Histogram of x

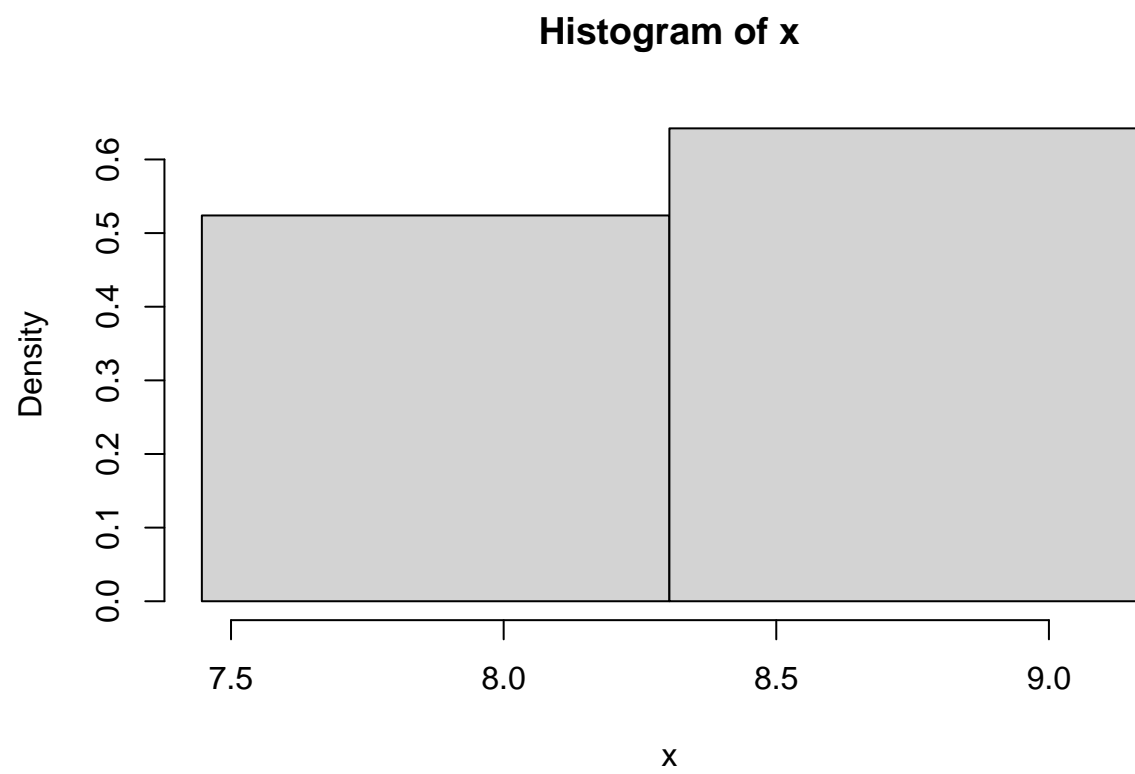


Histogram of x

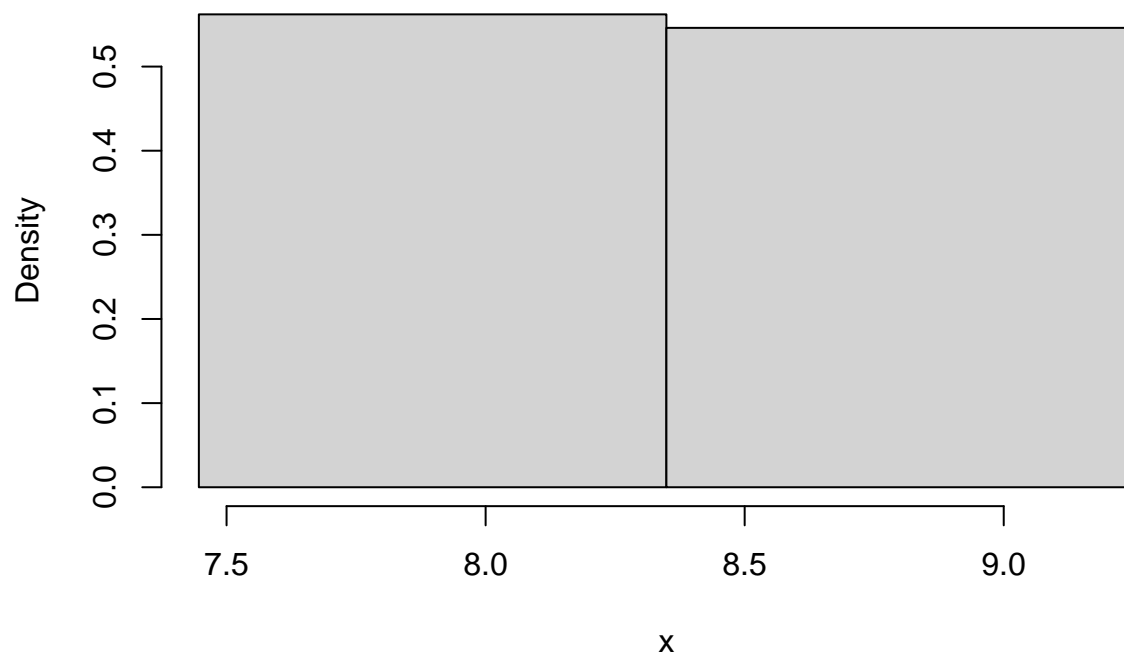


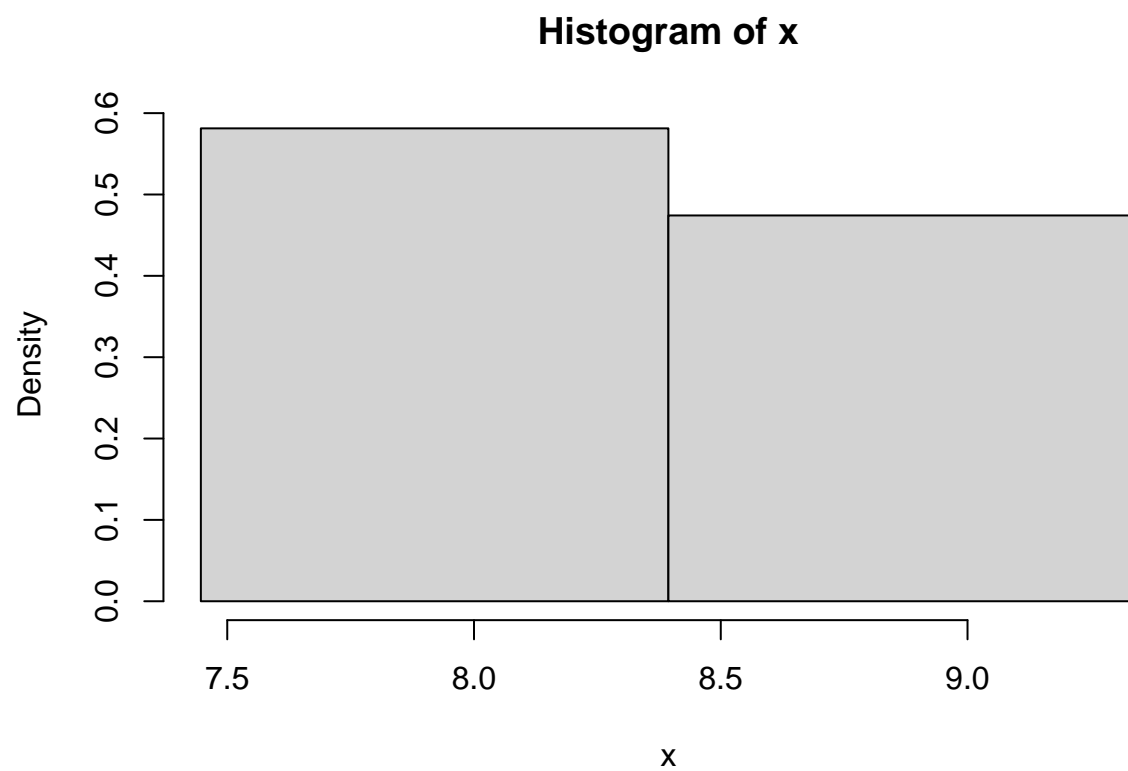
Histogram of x



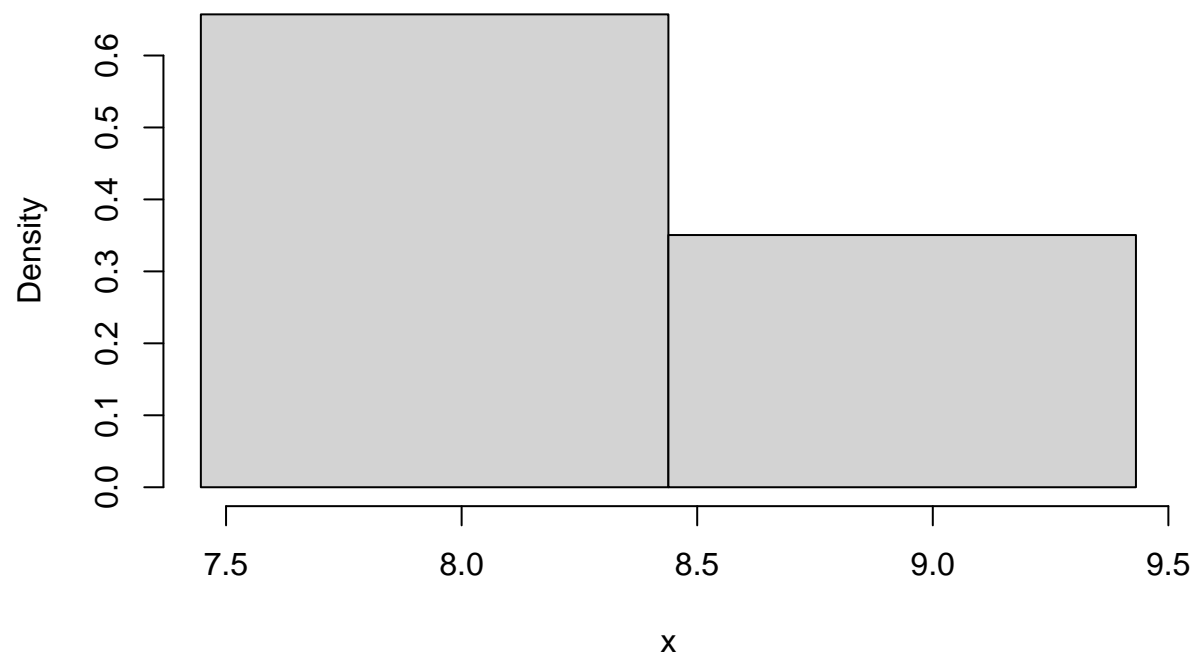


Histogram of x

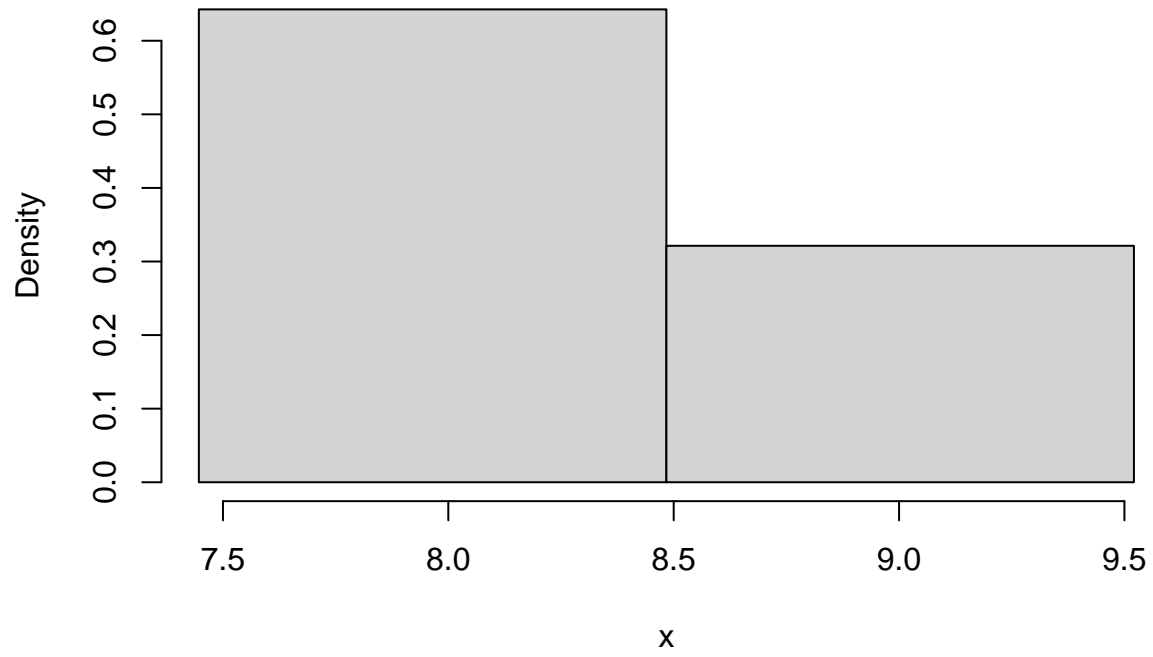




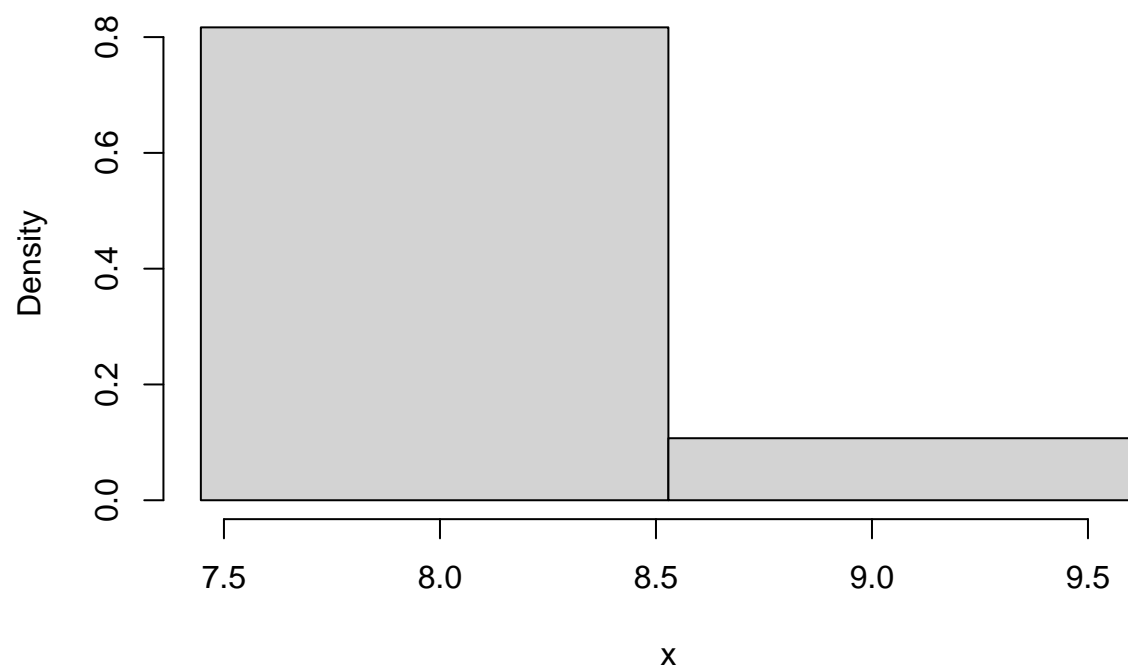
Histogram of x



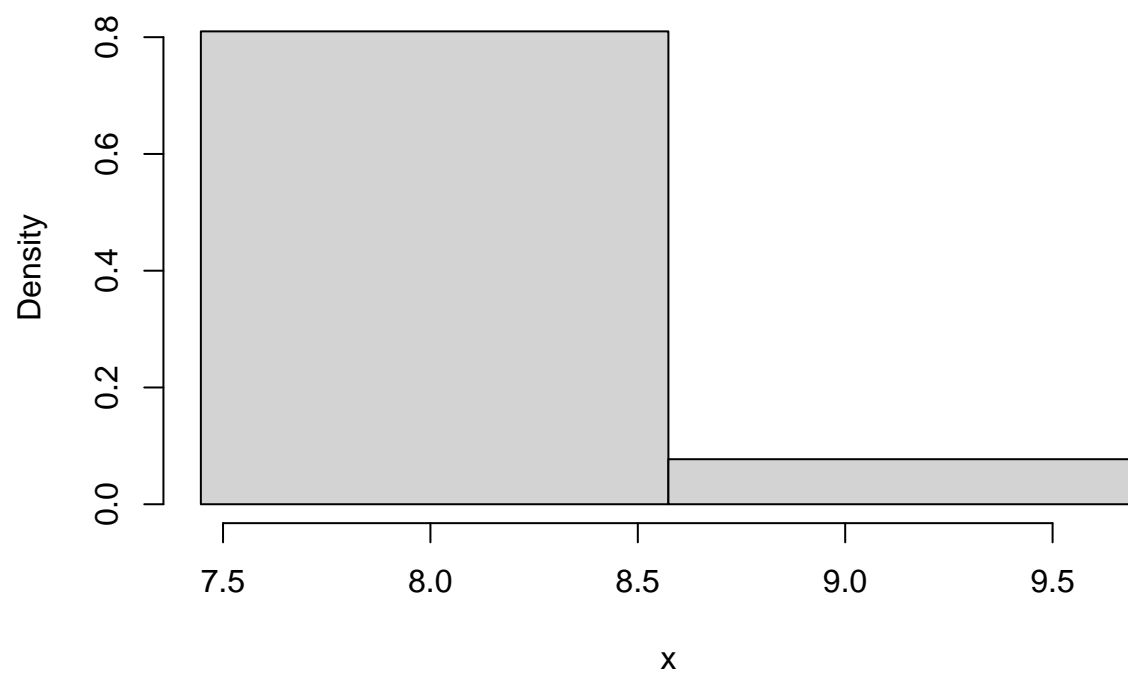
Histogram of x

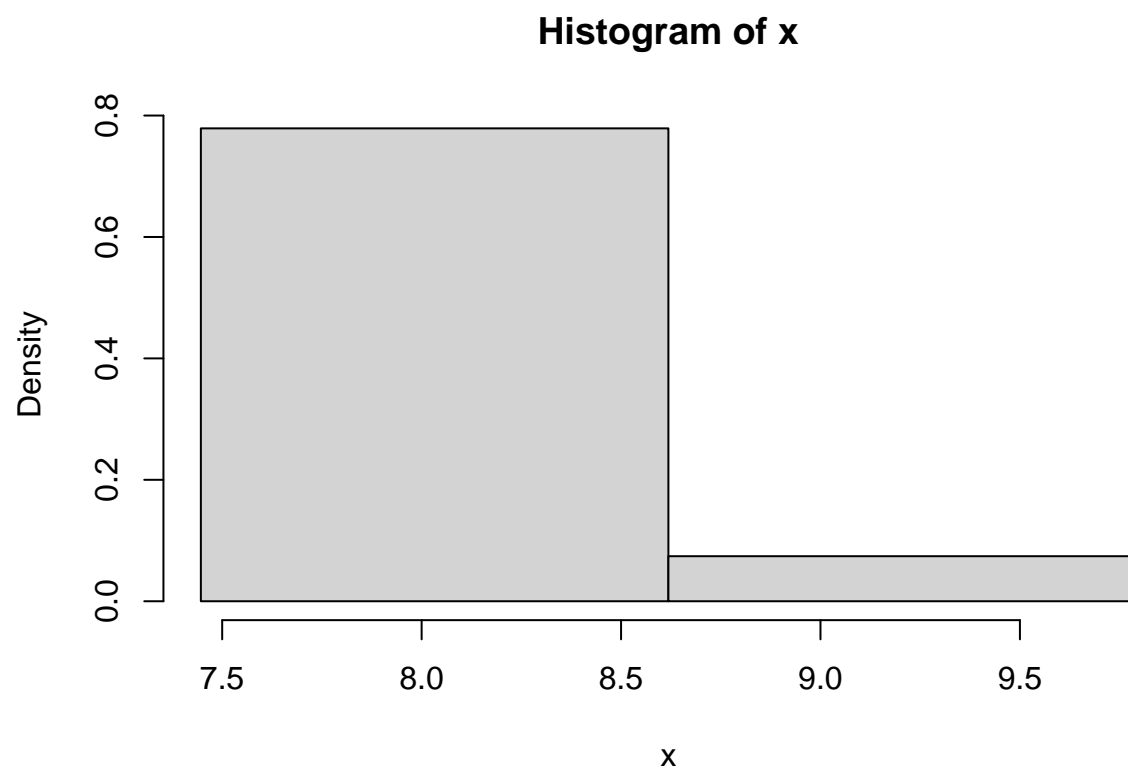


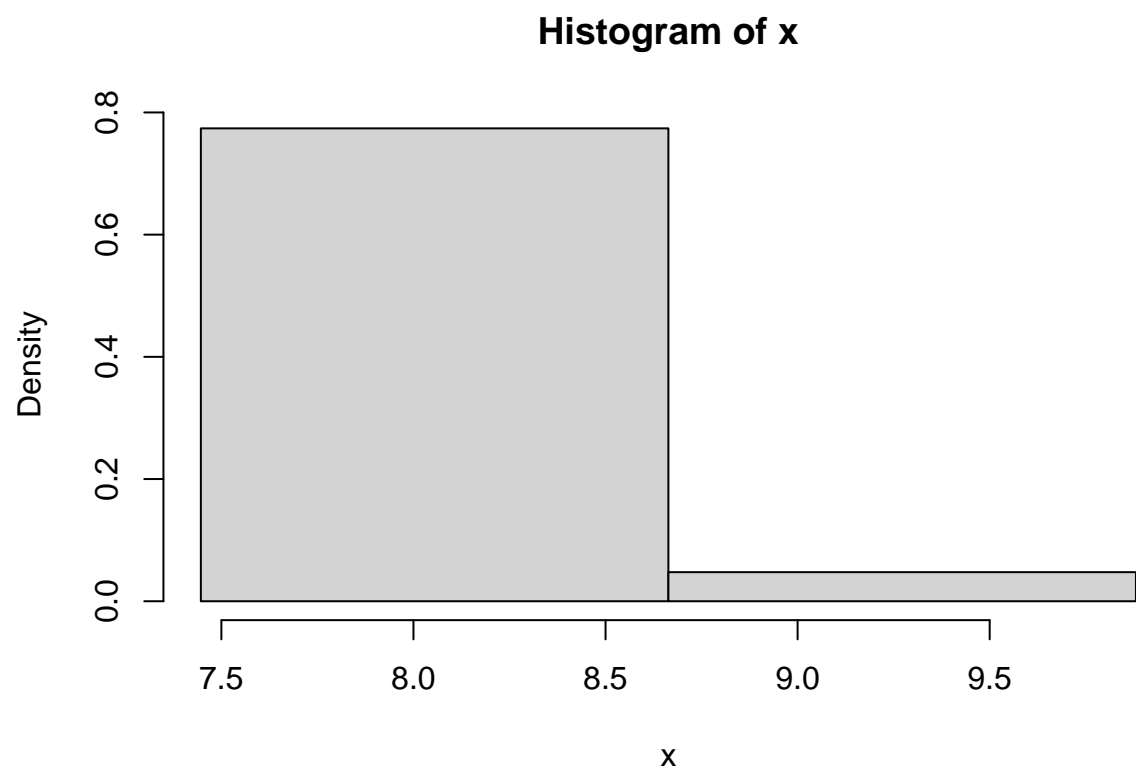
Histogram of x



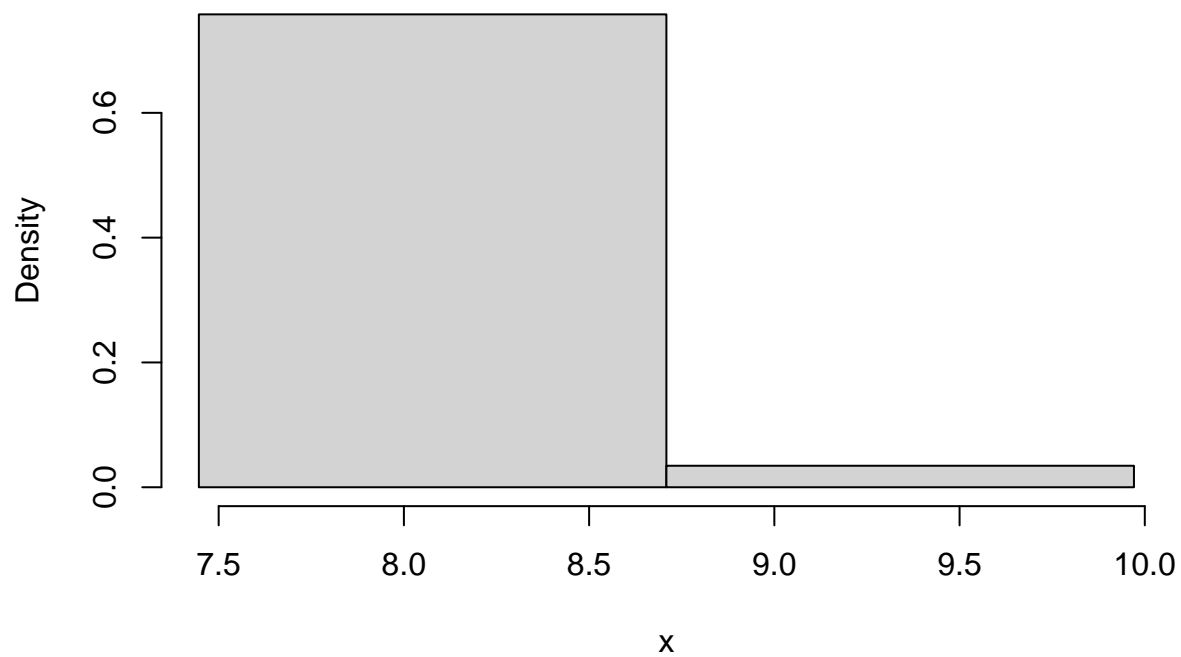
Histogram of x



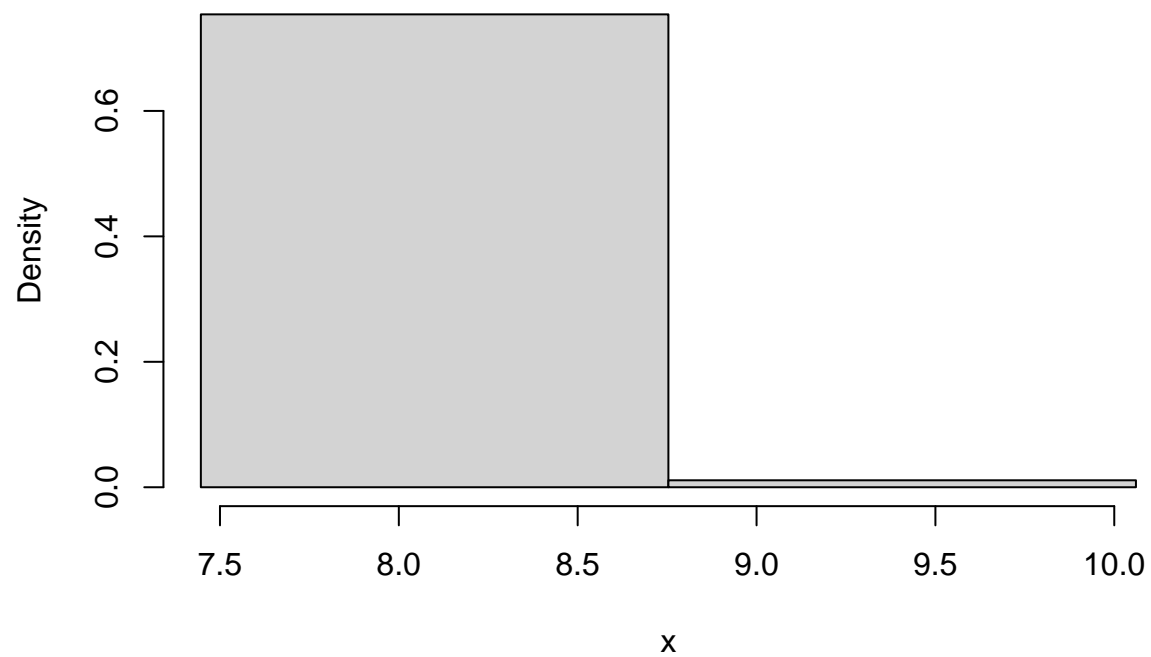




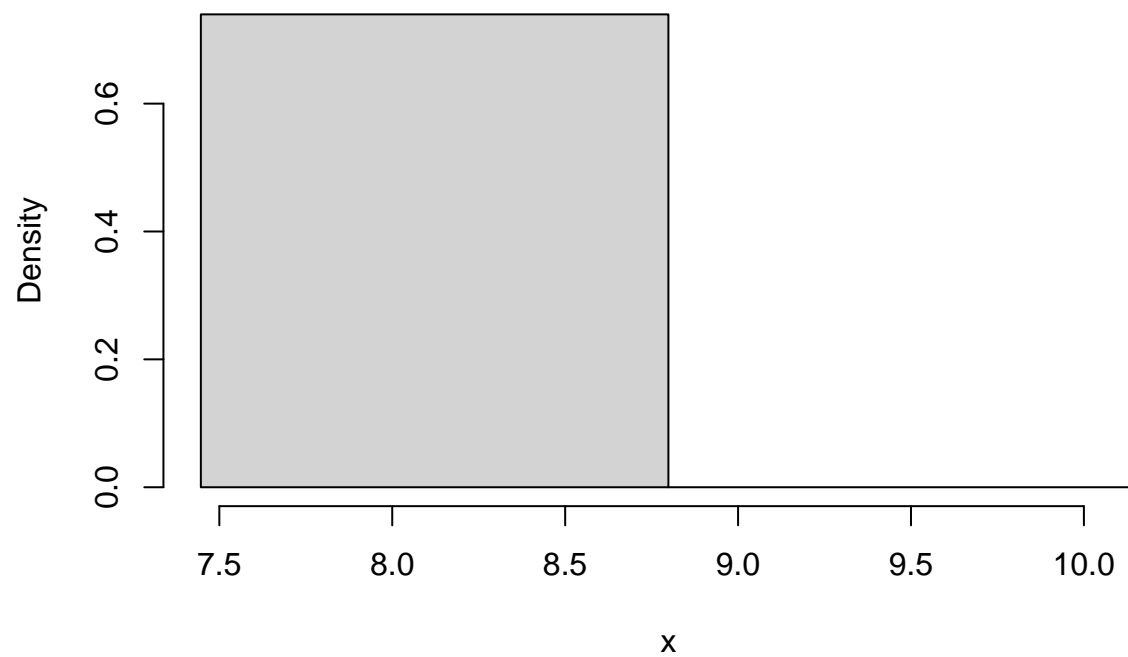
Histogram of x

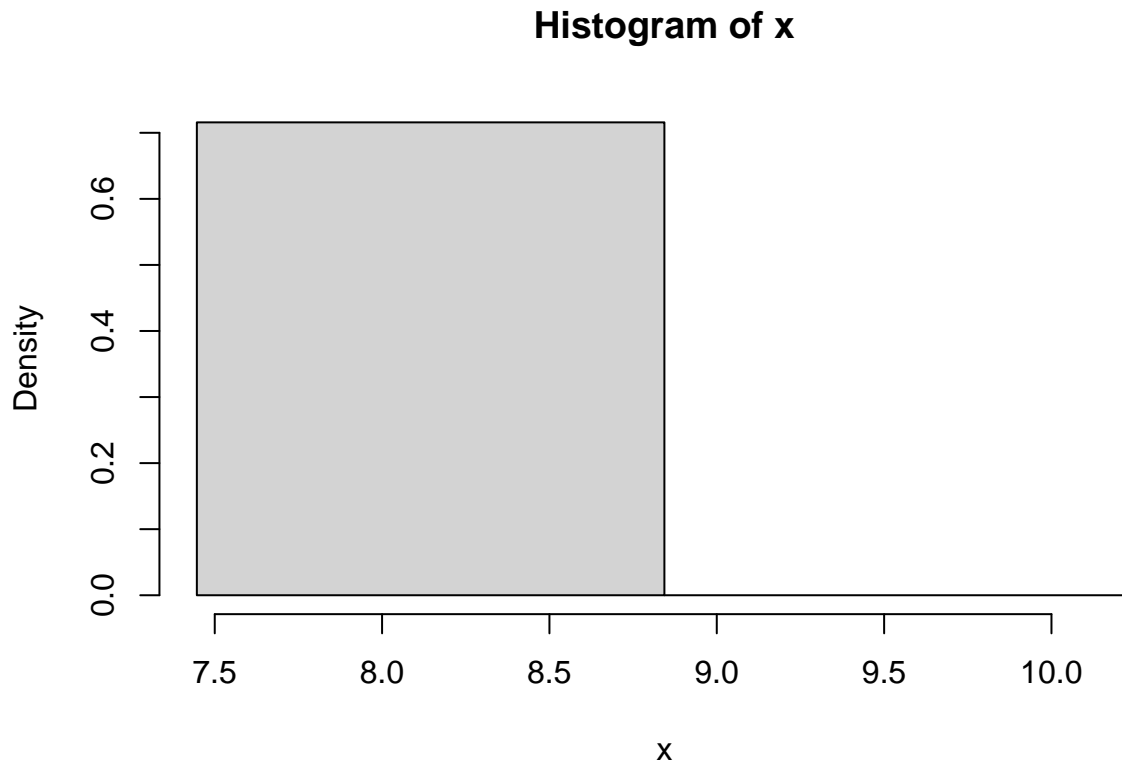


Histogram of x



Histogram of x





7. Recycle the functions *graph.mixt* and *sim.mixt* defined at *density_estimation.Rmd* to generate $n = 100$ data from

$$f(x) = (3/4)N(x; m = 0, s = 1) + (1/4)N(x; m = 3/2, s = 1/3)$$

Let b be the bin width of a histogram estimator of $f(x)$ using the generated data. Select the value of b maximizing the leave-one-out log-likelihood function, and plot the corresponding histogram. Compare with the results obtained using the Scott's formula:

$$b_{Scott} = 3.49 St.Dev(X)_n^{-1/3}$$

.

```
# TODO
```

Kernel density estimator

8.

$$\hat{f}_{h,(-i)}(x_i) = \frac{n}{n-1} \left(\hat{f}_h(x_i) - \frac{K(0)}{nh} \right)$$