tutorial3

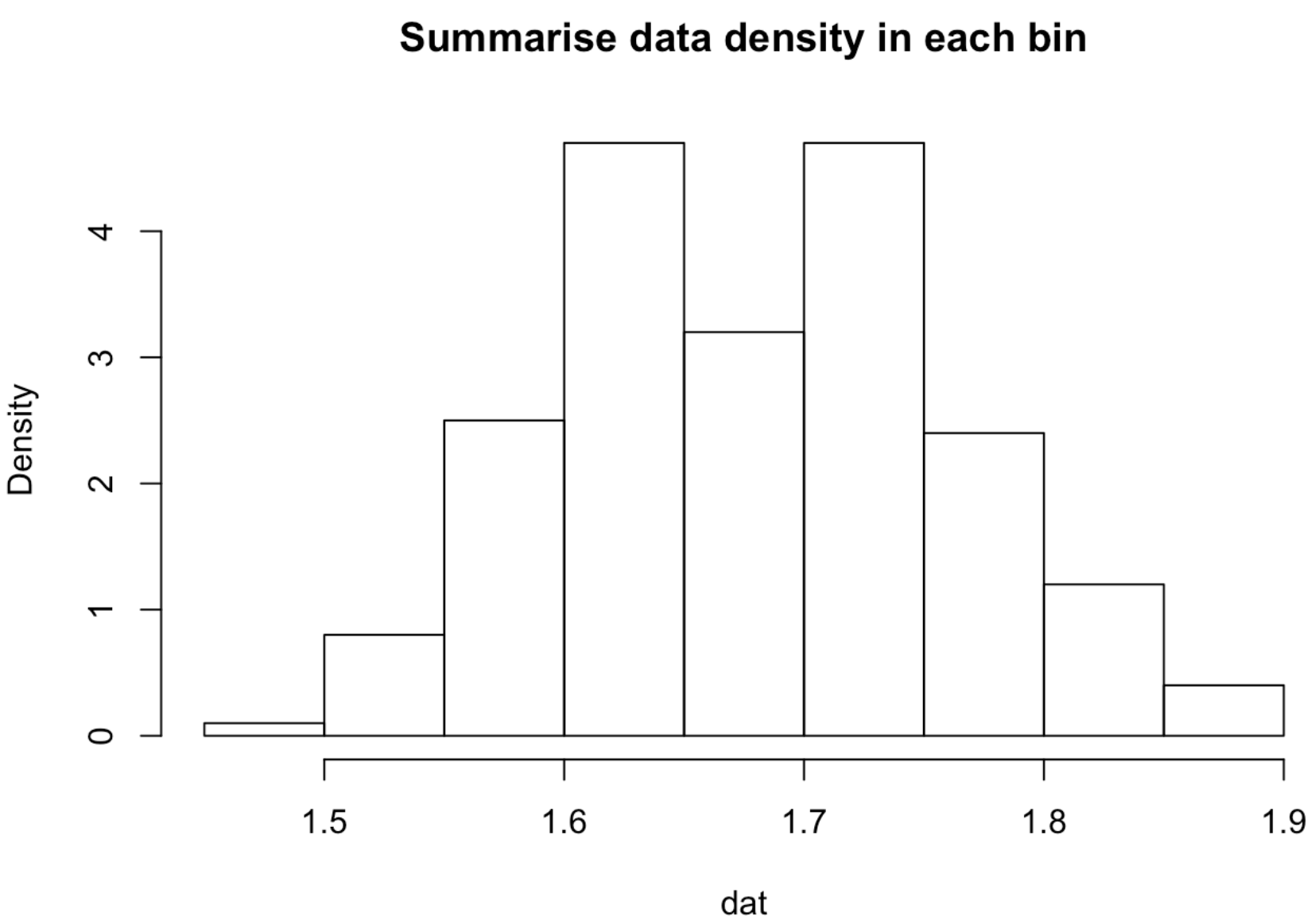
Likun Cui(41725041)

17 August, 2018

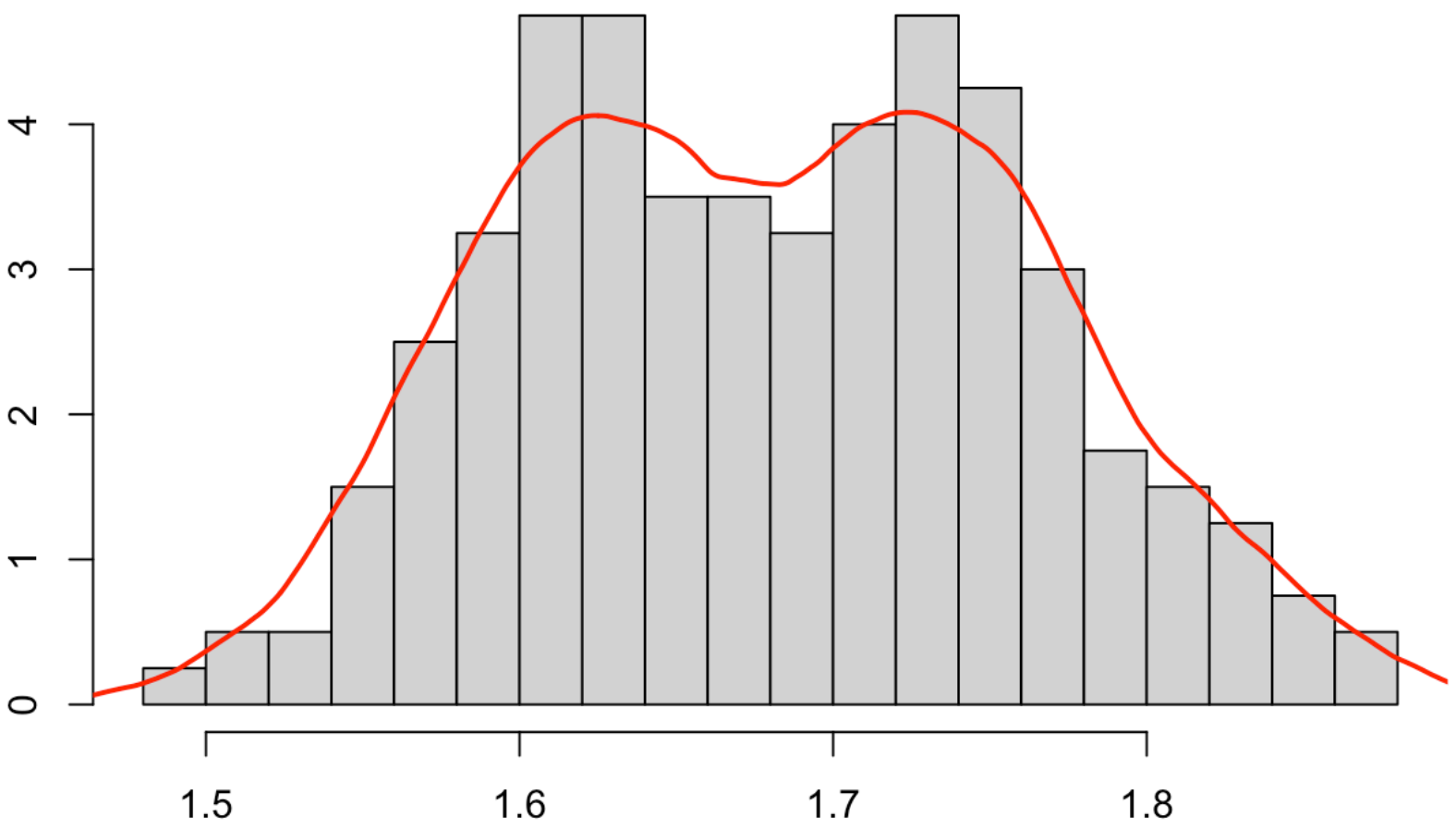
```
library(polspline)

#Task1.1 Read in height.txt data
setwd("/Users/likuncui/Downloads/5003/tutorial3/")
dat<- as.numeric(read.table("height.txt",header=T)[,2])

#Task1.2 Generate histogram to summarise and visualise height variable
hist(dat, probability = TRUE, main = "Summarise data density in each bin")
```

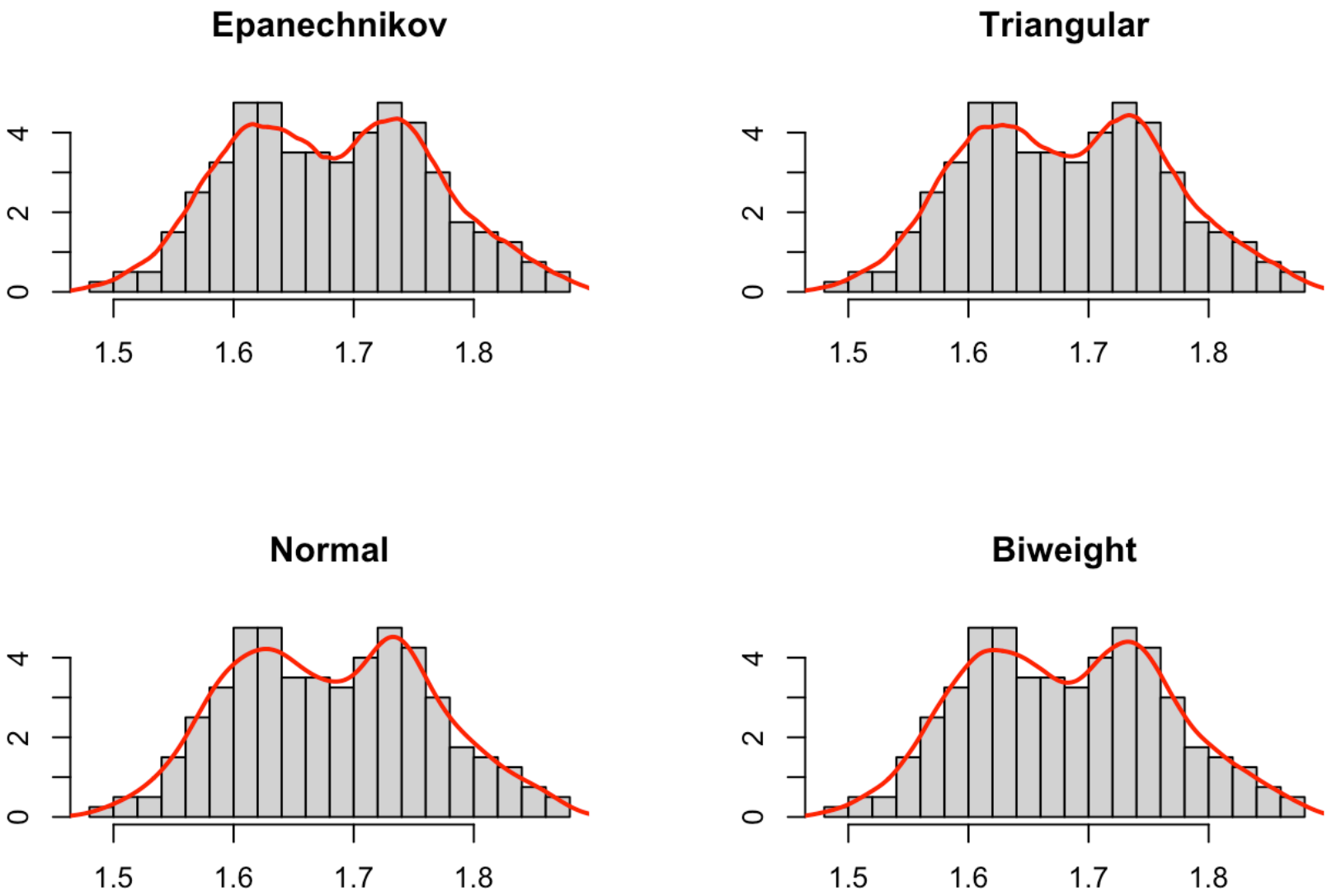



```
#Task2.1 Apply kernel density estimation methods to estimate height density.
#range <- seq(min(dat)-sd(dat), max(dat)+sd(dat), length.out=1000)
# estimate density of x using different kernels
d1 <- density(dat, kernel="epanechnikov")
hist(dat, breaks=20, freq=FALSE, col="lightgray", xlab="", ylab="", main="Epanechn
ikov( Default h)")
lines(d1, lwd=2, col="red")
```




```
h<-0.02
d2<-density(dat,bw= h,kernel="epanechnikov")
d3<-density(dat,bw=h,kernel="triangular")
d4<-density(dat,bw=h,kernel="gaussian")
d5<-density(dat,bw=h,kernel="biweight")

par(mfrow=c(2,2))
hist(dat, breaks=20, freq=FALSE, col="lightgray", xlab="", ylab="", main="Epanechn
ikov")
lines(d2, lwd=2, col="red")
hist(dat, breaks=20, freq=FALSE, col="lightgray", xlab="", ylab="", main="Triangul
ar")
lines(d3, lwd=2, col="red")
hist(dat, breaks=20, freq=FALSE, col="lightgray", xlab="", ylab="", main="Normal")
lines(d4, lwd=2, col="red")
hist(dat, breaks=20, freq=FALSE, col="lightgray", xlab="", ylab="", main="Biweight
")
lines(d5, lwd=2, col="red")
```



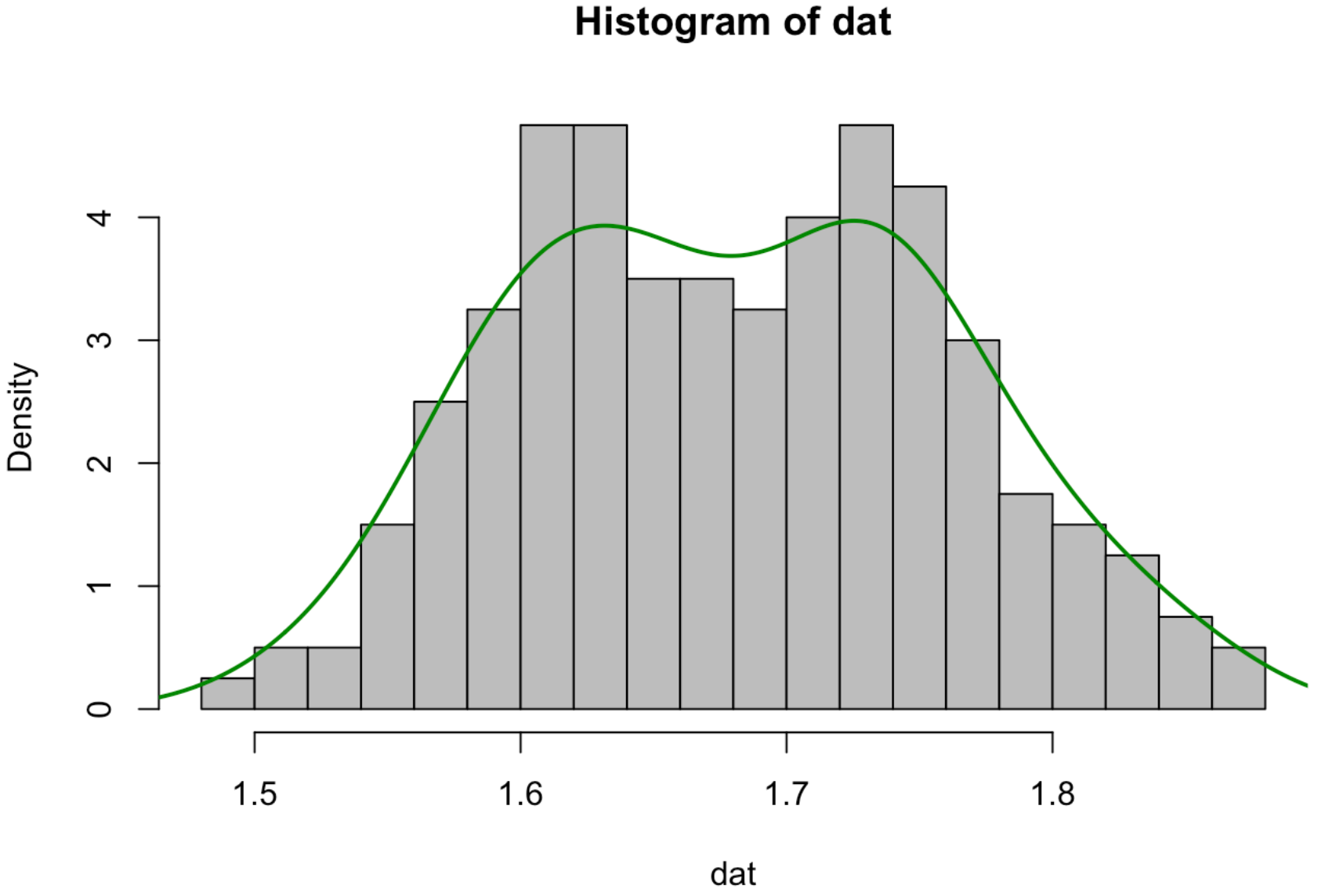

```
#Task3.1 Use BCV methods to select for optimal bandwidth for each kernel of choice
.
h <- c(0.3, 0.625, 1.875)
h.bcv <- bw.bcv(dat)
```

```
## Warning in bw.bcv(dat): minimum occurred at one end of the range
```

```
d.bcv <- density(dat, bw=h.bcv, kernel="gaussian")
h.bcv
```

```
## [1] 0.03196391
```

```
## Plot estimation results
par(mfrow=c(1,1))
hist(dat, breaks=20, freq=FALSE, col="gray")
lines(d.bcv, col="green4", lwd=2)
```

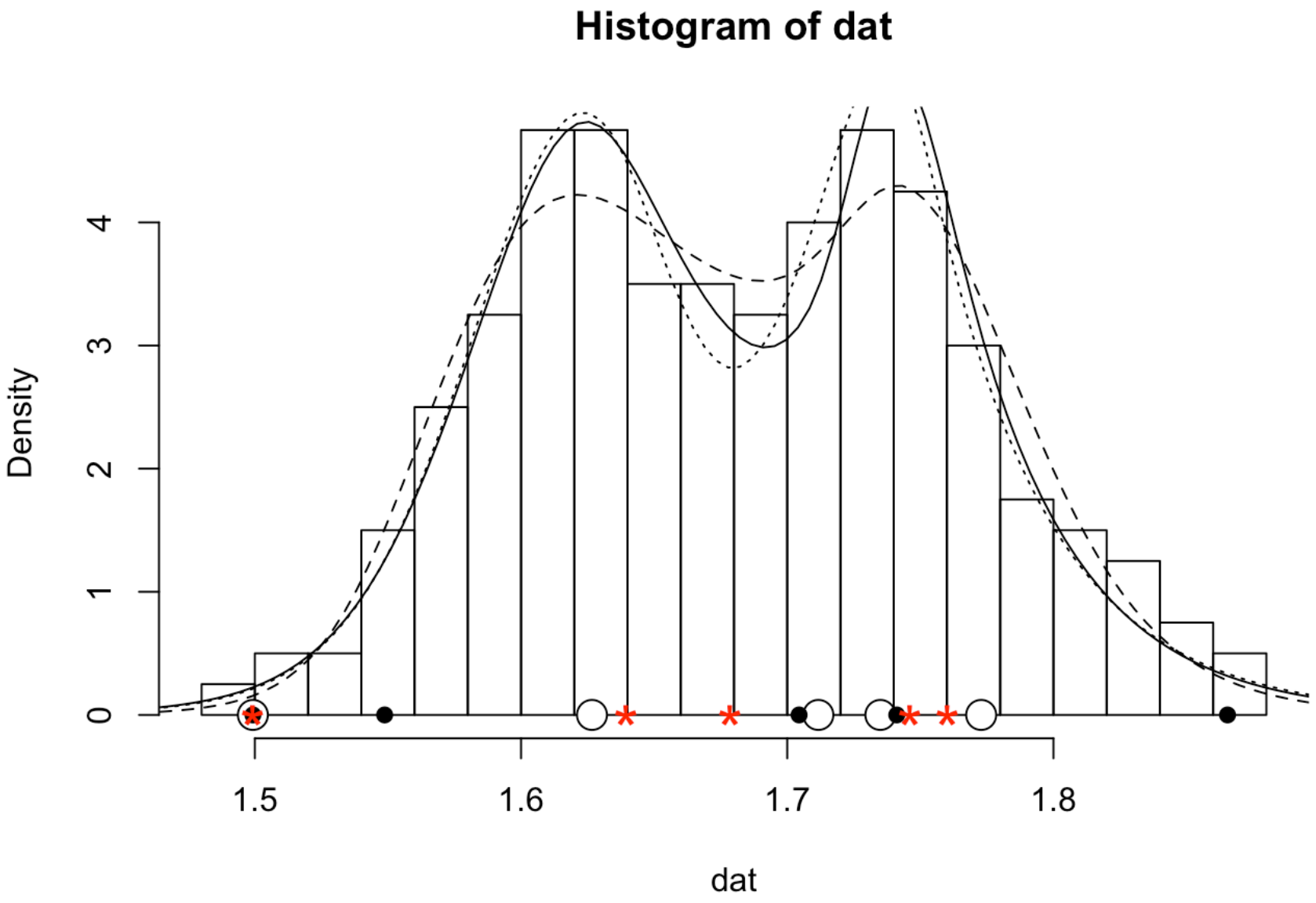



```
#Task4.1  Apply cubic spline density estimation with different number of knots
fit1 <- logspline(dat)
hist(dat,breaks=20,freq=FALSE)
fit1
```

```
##  knots A(1)/D(2) loglik     AIC minimum penalty maximum penalty
##      4         2 211.95 -408.00           26.78             Inf
##      5         2 225.34 -429.48            2.76           26.78
##      6         2 225.36 -424.23              NA              NA
##      7         2 228.09 -424.40            0.43            2.76
##      8         2 228.12 -419.15              NA              NA
##      9         2 228.53 -414.67            0.31            0.43
##     10         2 228.68 -409.68            0.26            0.31
##     11         2 228.81 -404.64            0.02            0.26
##     12         2 228.82 -399.37            0.01            0.02
##     13         1 228.83 -394.08            0.00            0.01
## the present optimal number of knots is  5
## penalty(AIC) was the default: BIC=log(samplesize): log( 200 )= 5.3
```

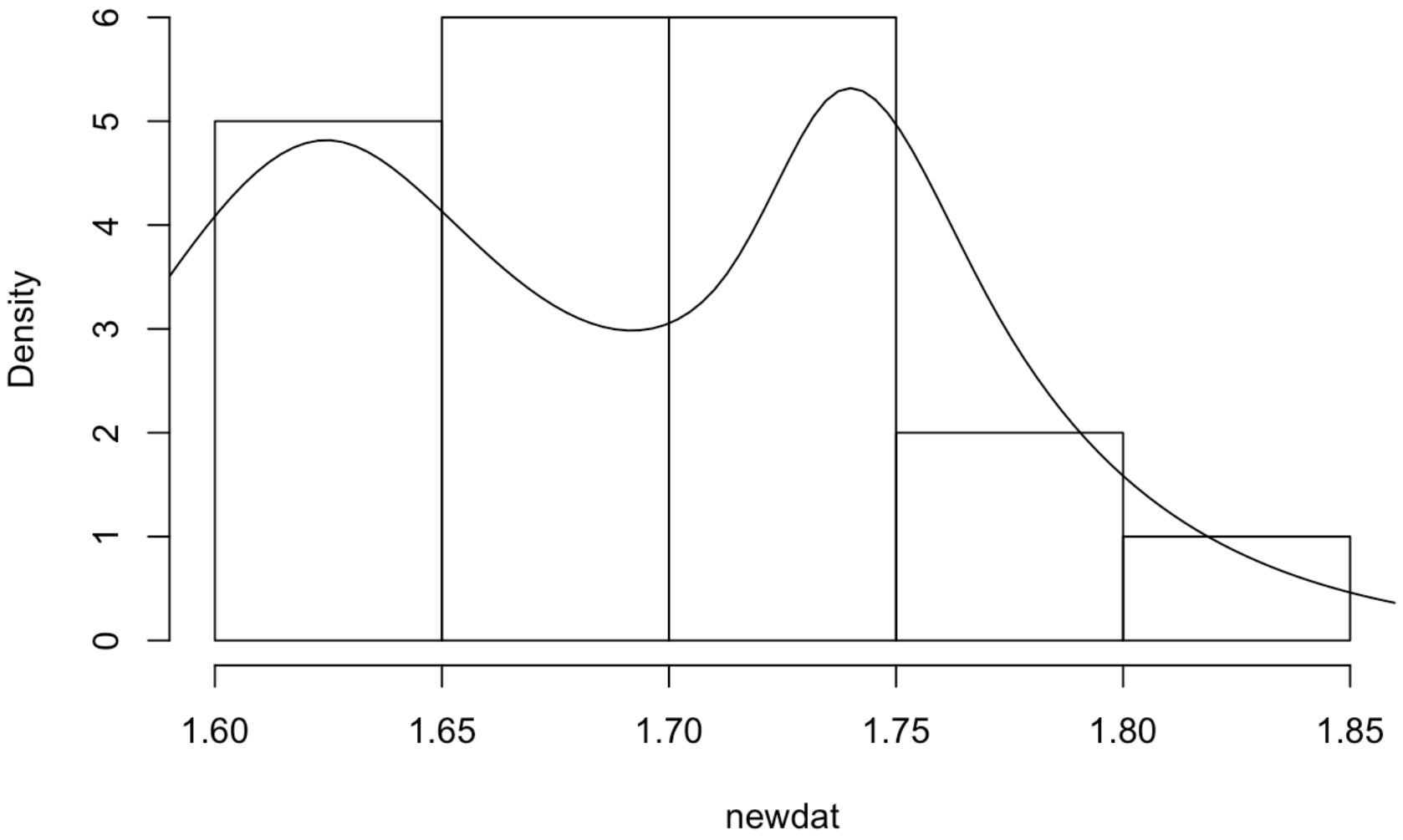
```
plot(fit1, add=T)
points(fit1$knots,rep(0,5),pch=21,cex=2,bg="white")

fit2 <- logspline(dat, nknots=3)
fit3 <- logspline(dat, nknots=7)
plot(fit2, add=T,lty=2)
plot(fit3, add=T,lty=3)
points(fit2$knots,rep(0,5),pch=21,cex=1,bg="black")
points(fit3$knots,rep(0,5),pch="*",cex=2, col="red")
```



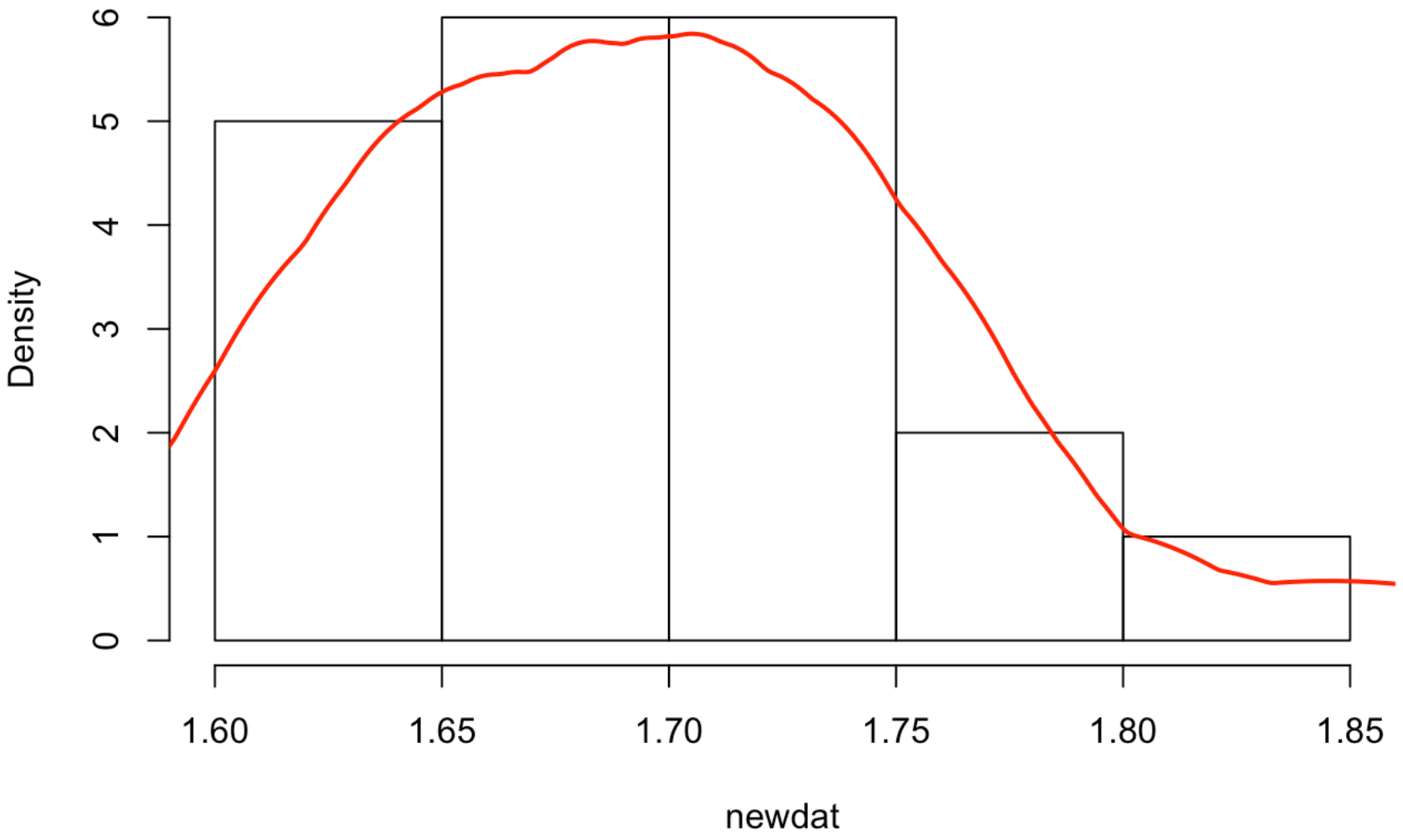

```
#Task4.2  Compare these results to those from using kernel functions.

#Task5.1 Read in newHight.txt data.
newdat<- as.numeric(read.table("newHeight.txt",header=T)[,2])
hist(newdat, probability = TRUE, main = "Summarise data density using previous mod
el")
#classify each of these new samples based on density estimation results from heigh
t.txt
plot(fit1, add=T)
```




```
#the perfect prediction should be the following plot:
hist(newdat, probability = TRUE, main = "Summarise data density by default functio
n")

d7 <- density(newdat, kernel="epanechnikov")
lines(d7, lwd=2, col="red")
```




```
print("since the are totally deffences between the two curves, although it could b
e used as classification, it not a proper result.")
```

```
## [1] "since the are totally deffences between the two curves, although it could
be used as classification, it not a proper result."
```