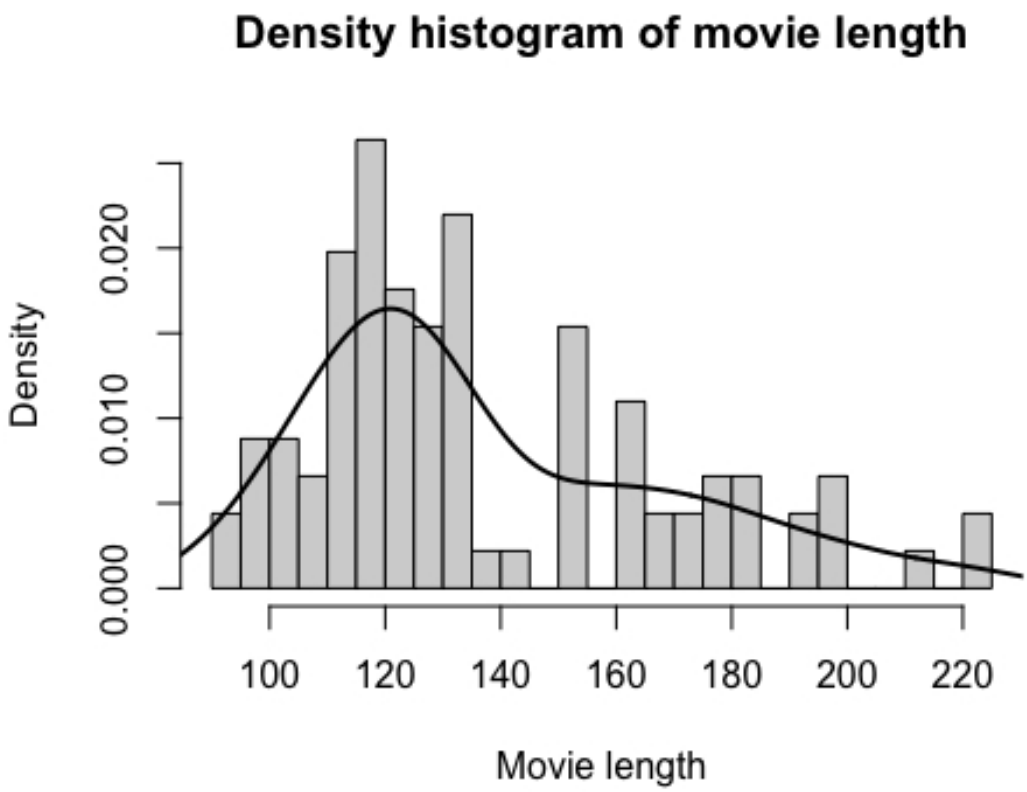
At the movies [8 marks]

In the file `movies.csv` (available on Learn together with this assignment) you can find a list of 91 Academy Award-winning movies, their length in minutes (`Length`) and their income at the box office in million US dollars (`Box_office`) [Source: Wikipedia].

Create a density-histogram for the length of the movies and add a line with an estimated kernel density; choose an adequate kernel function and bandwidth.

```
library(evmix)

movies <- read.csv("movies.csv")

x=movies$Length
hist(x, breaks=25, prob=T, xlab="Movie length", main="Density histogram of
movie length")
lines(density(x), lwd=2)
```



Calculate confidence limits for the average movie length based on a large sample approximation. Can we detect if Oscar-winning movies are significantly longer than two hours on average?

```r
N = 10000  # number of bootstrap resamples
alpha=0.05 # significance level

set.seed(31878039)
x = movies$Length

boot.means = numeric(N)
for(i in 1:N) {
  xstar = sample(x, replace=T)
  boot.means[i] = mean(xstar)
}
quantile(boot.means, probs=c(alpha/2, 1-alpha/2))
```

```
##     2.5%    97.5%
## 131.6154 144.1871
```

Yes we can detect them. The mean movie length is between 132 and 144 minutes. Oscar winning movies are longer than two hours on average.

Calculate percentile confidence limits for the average movie length using a nonparametric bootstrap. Explain why there is a difference or why there is no large difference in the resulting confidence limits compared to the large sample approximation.

```r
mean((boot.means>=quantile(boot.means, alpha/2))
     & (boot.means<=quantile(boot.means, 1-alpha/2)))
```

```
## [1] 0.9502
```

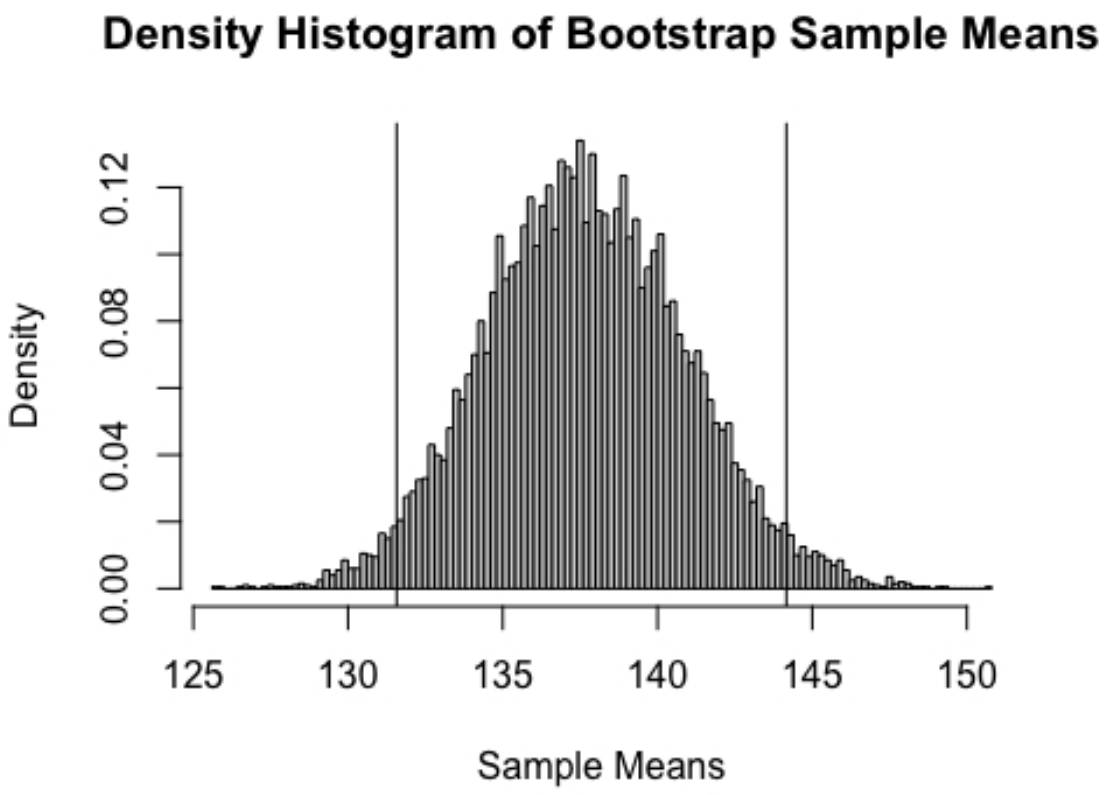
There is no large difference as it is already 95% confidence. This is because the sample size is large enough, roughly 100.

Calculate percentile confidence limits for the mean and the median of the box office income using a nonparametric bootstrap procedure. Compare the bootstrap distribution of the mean and the median by density histograms.

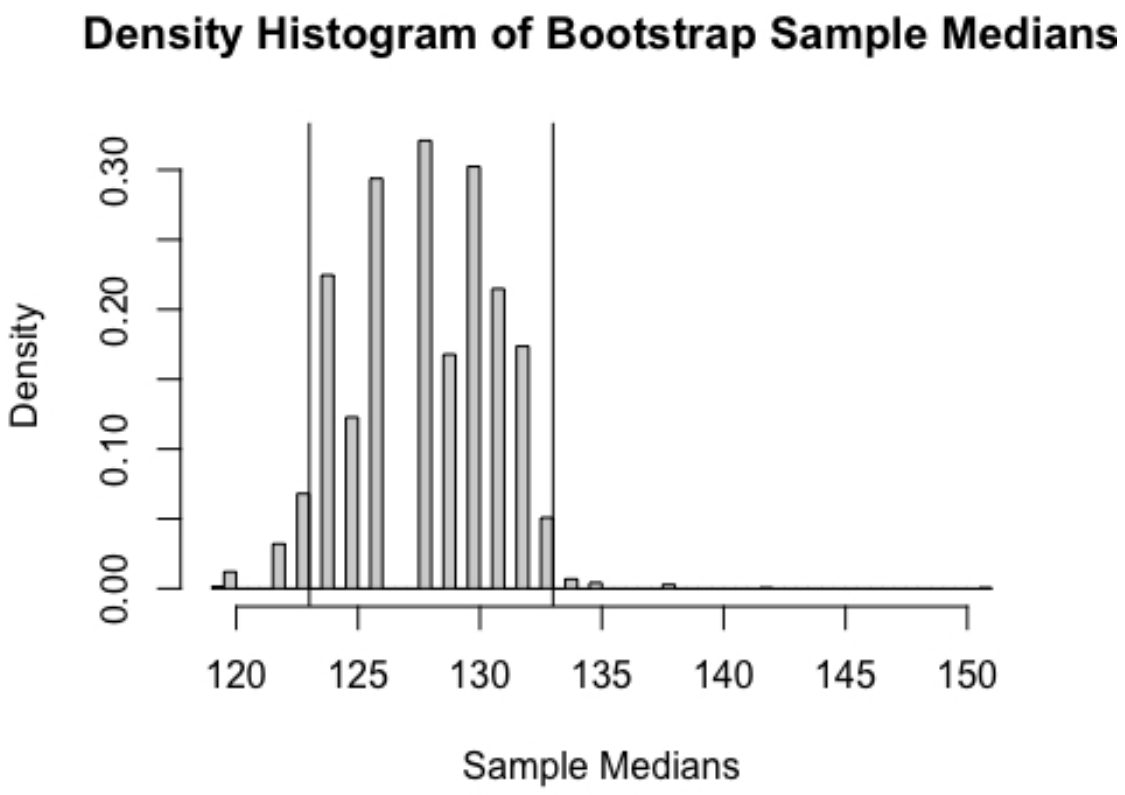
```r
library(boot)
boot.stat = function(x, indices) {
  xstar = x[indices]
  return(c(mean(xstar), median(xstar)))
}

N = 10000  # number of bootstrap resamples
alpha=0.05 # significance level

set.seed(31878039)
boot(movies$Box_office, statistic=boot.stat, R=N)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
```

```
## Call:
## boot(data = movies$Box_office, statistic = boot.stat, R = N)
##
##
## Bootstrap Statistics :
##     original     bias    std. error
## t1*  156.944 -0.3941378    28.91543
## t2*   61.800  5.2755500    18.65492

results = boot(x, statistic=boot.stat, R=N)
boot.ci(results, type="perc", index=1)

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, type = "perc", index = 1)
##
## Intervals :
## Level     Percentile
## 95%   (131.6, 144.2 )
## Calculations and Intervals on Original Scale

boot.ci(results, type="perc", index=2)

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, type = "perc", index = 2)
##
## Intervals :
## Level     Percentile
## 95%   (123, 133 )
## Calculations and Intervals on Original Scale

hist(results$t[,1], breaks=100, prob=T, xlab="Sample Means", main="Density
Histogram of Bootstrap Sample Means")
abline(v=boot.ci(results, type="perc", index=1)$percent[4:5])
```




```
hist(results$t[,2], breaks=100, prob=T, xlab="Sample Medians", main="Density
Histogram of Bootstrap Sample Medians")
abline(v=boot.ci(results, type="perc", index=2)$percent[4:5])
```




```
boot.ci(results, type="perc", index=1)$percent
```

```
##      conf
## [1,] 0.95 250.03 9750.98 131.5714 144.1648
```

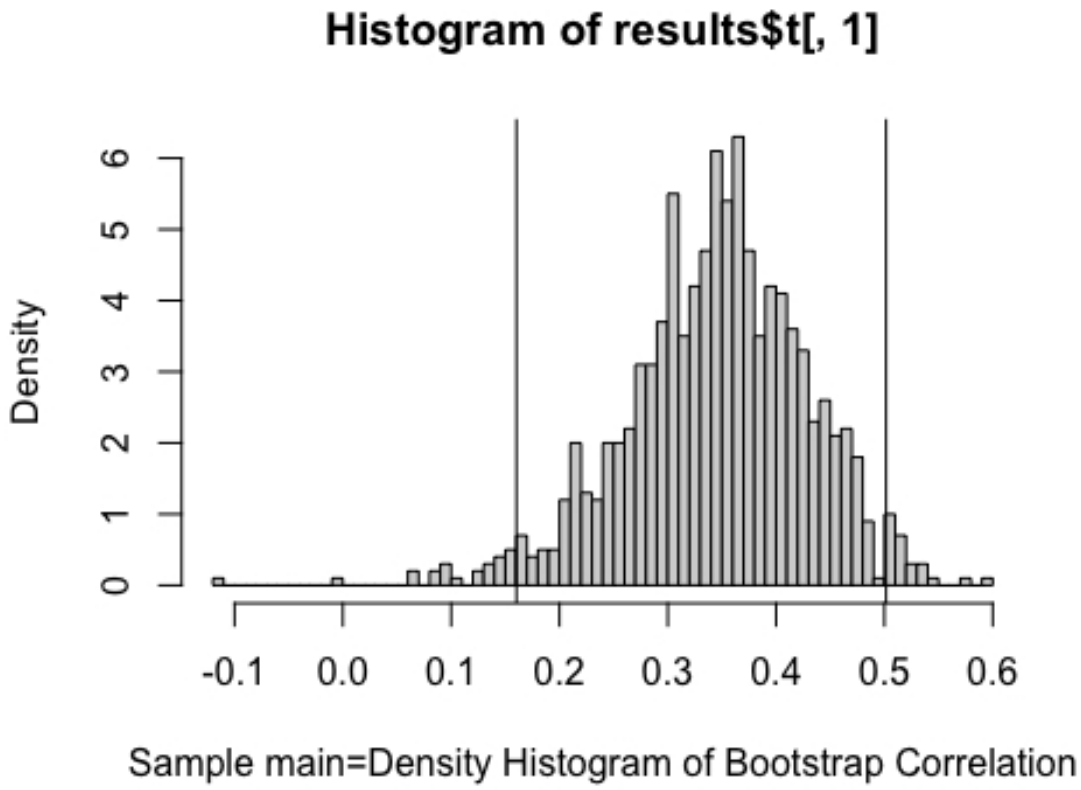
The confidence limits for the mean of the box office of movies is between 132 and 144 million US dollars. The confidence limits for the median of the box office of movies is between 123 and 133 million US dollars.

The mean distribution looks more complete/continuous and resembles a normal distribution as every samples have its own mean whereas the median distribution does not look complete as every sample can have the same median.

Calculate percentile confidence limits for the Pearson correlation coefficient between length and income of a movie using a nonparametric bootstrap procedure. Visualise the bootstrap distribution of the correlation coefficient with a density histogram.

```
set.seed(31878039)
x <- movies$Length
y <- movies$Box_office
xy <- cbind(x,y)
boot.stat = function(xy, indices){
  xystar <- xy[indices,]
  xstar <- xystar[,1]
```

```
  ystar <- xystar[,2]
  return(cor(xstar, ystar, method="pearson"))
}

results <- boot(xy, statistic = boot.stat, R=1000)
quantile(results$t, probs=c(0.025, 0.975))

##      2.5%     97.5%
## 0.1606954 0.5006836

hist(results$t[,1], breaks=100, prob=T, xlab="Sample main=Density Histogram
of Bootstrap Correlation")
abline(v=boot.ci(results, type="perc", index=1)$percent)
```



Jelly Beans [2 marks]

A package of jellybeans contains 76 beans in total. In a randomly chosen package, we count 16 red jellybeans. Estimate percentile confidence limits for the probability of picking a red jellybean out of a package using a nonparametric bootstrap.

```
set.seed(31878039)
red <- rep(c(1, 0), times=c(16, 76))
```

```
jbsample <- replicate(1000, mean(sample(red, replace=T)))
quantile(jbsample, probs=c(alpha/2, 1-alpha/2))

##       2.5%      97.5%
## 0.09782609 0.25000000
```

The confidence limits is between 9.7% and 25% for picking a red jellybean out of a package using a nonparametric bootstrap.