Stat 135 Lab1

Leomart Crisostomo

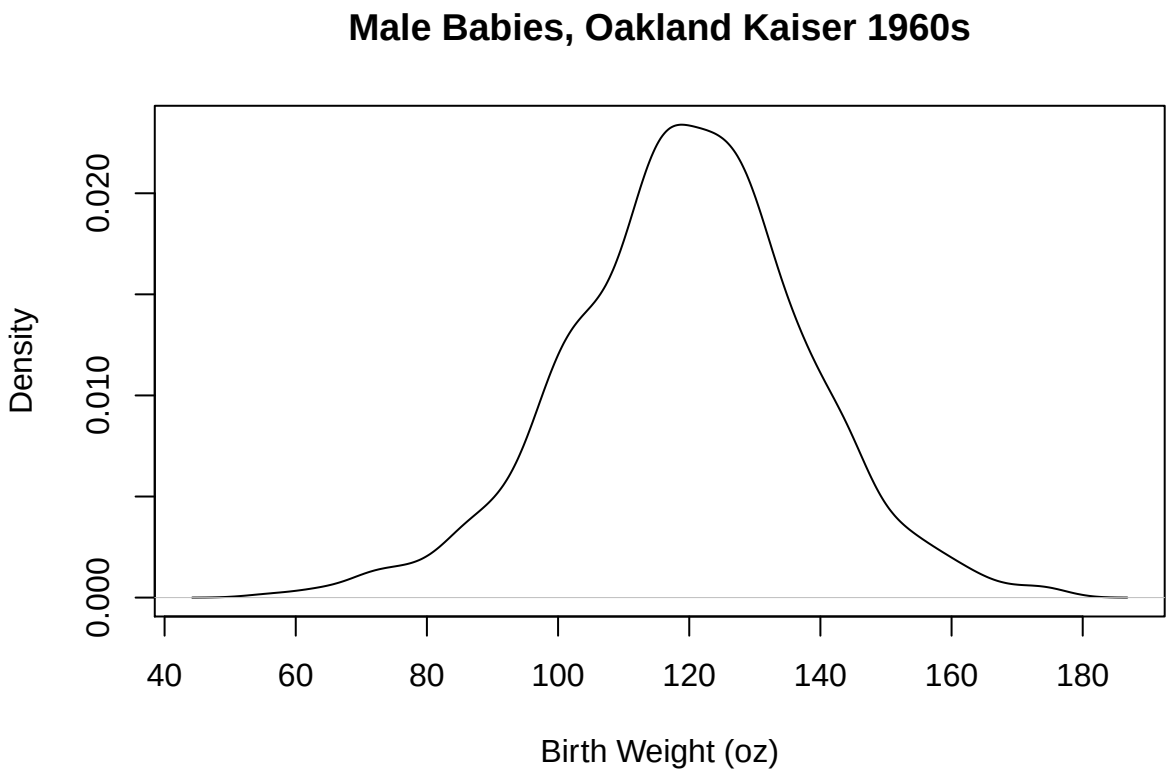
2/13/2018

R Markdown

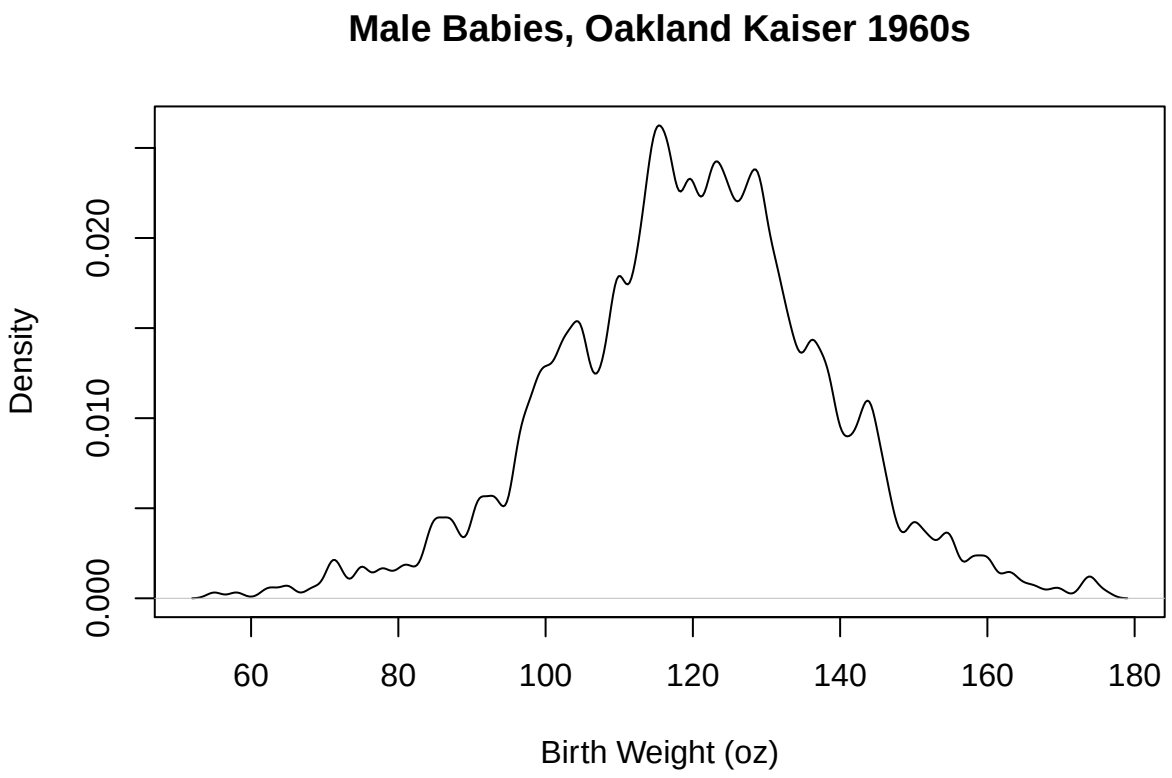
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

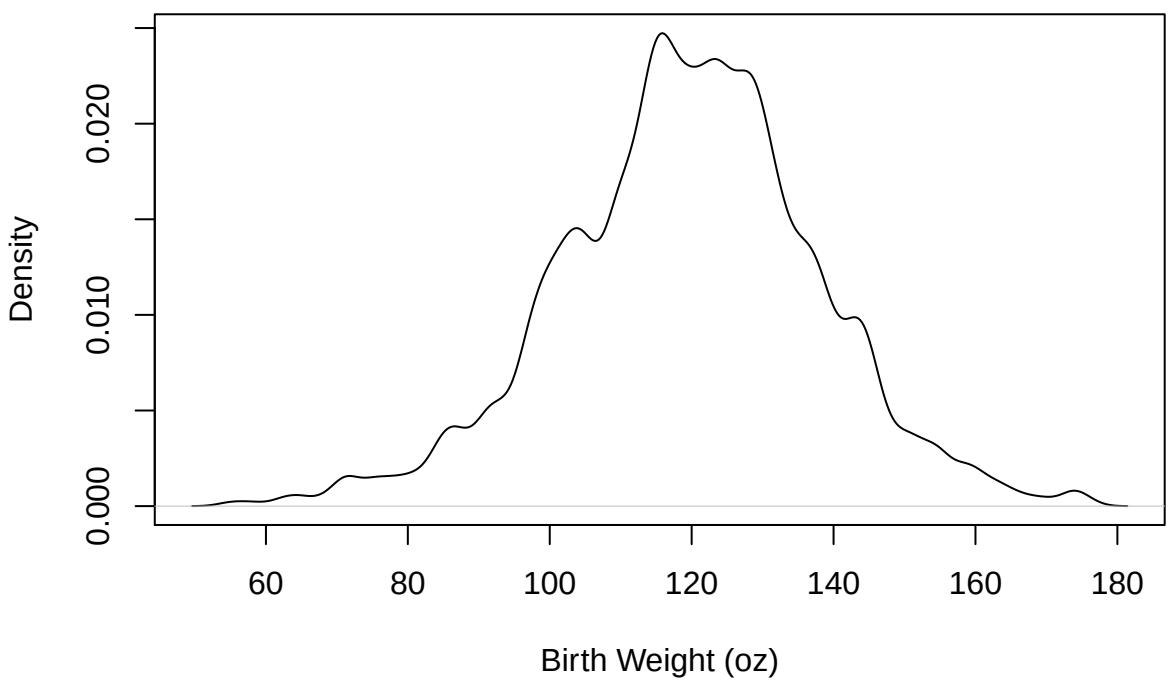
```
load("/Users/Leomart/Desktop/Stat135/KaiserBabies.rda")
plot(density(infants$bwt), xlab = "Birth Weight (oz)", main = "Male Babies, Oakland Kaiser 1960s")
```



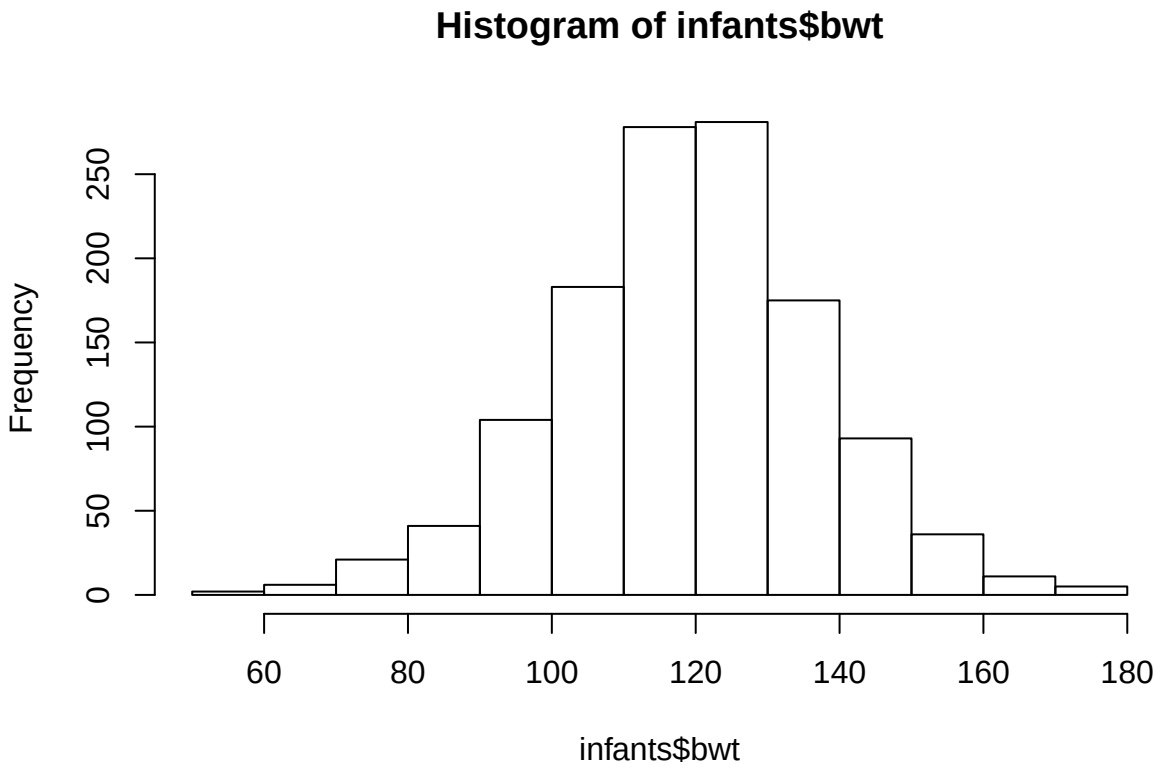

```
plot(density(infants$bwt,bw=1), xlab = "Birth Weight (oz)", main = "Male Babies, Oakland Kaiser 1960s")
```



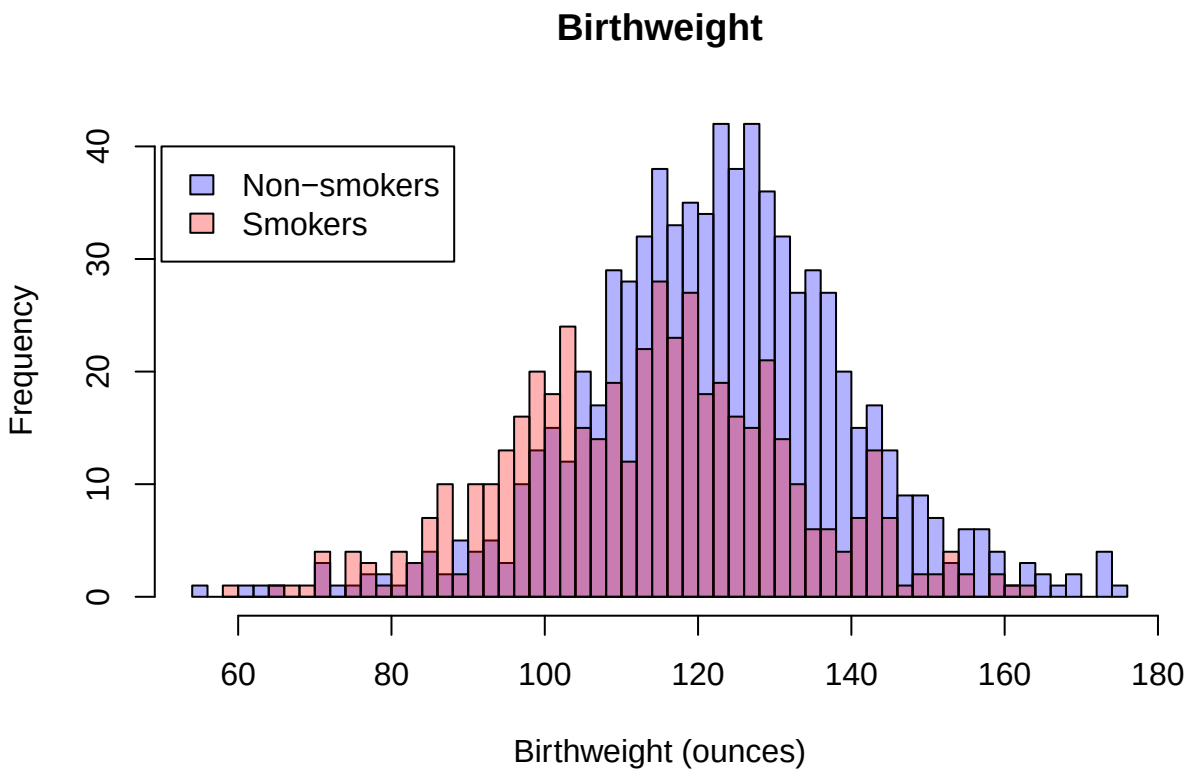

```
plot(density(infants$bwt,adjust=0.5), xlab = "Birth Weight (oz)", main = "Male Babies, Oakland Kaiser 19
```




```
hist(infants$bwt)
```




```
hist(infants$bwt[!infants$smoke=="Now"],breaks=50,col=rgb(0,0,1,.3),
xlab="Birthweight (ounces)",main="Birthweight")
hist(infants$bwt[infants$smoke=="Now"],breaks=50,col=rgb(1,0,0,.3),add=T)
legend(50,40,legend=c("Non-smokers","Smokers"),
fill=c(rgb(0,0,1,.3),rgb(1,0,0,.3)))
```



```r
mean(infants$bwt)
```

```
## [1] 119.5769
```

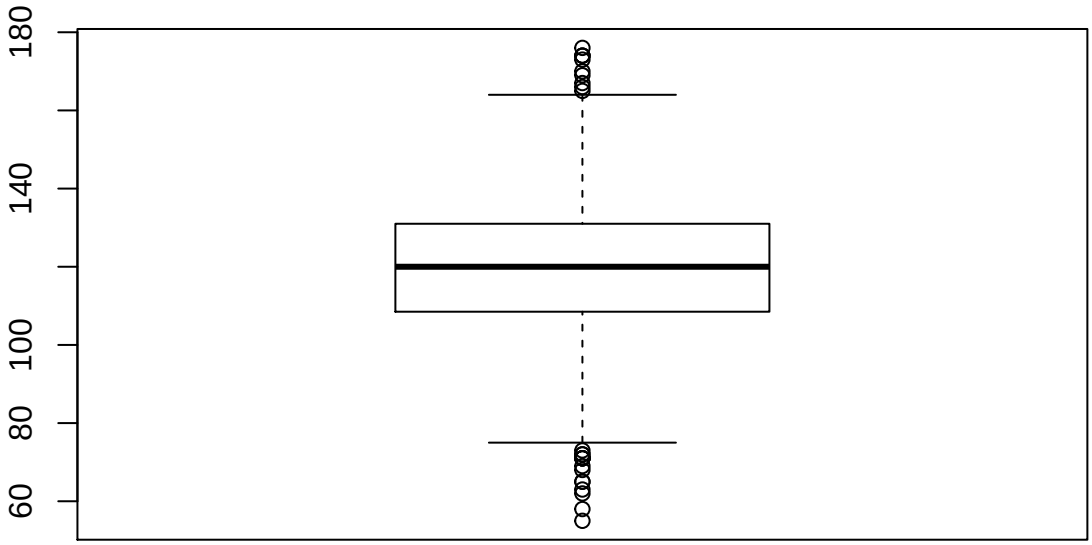
```r
sd(infants$bwt)
```

```
## [1] 18.23645
```

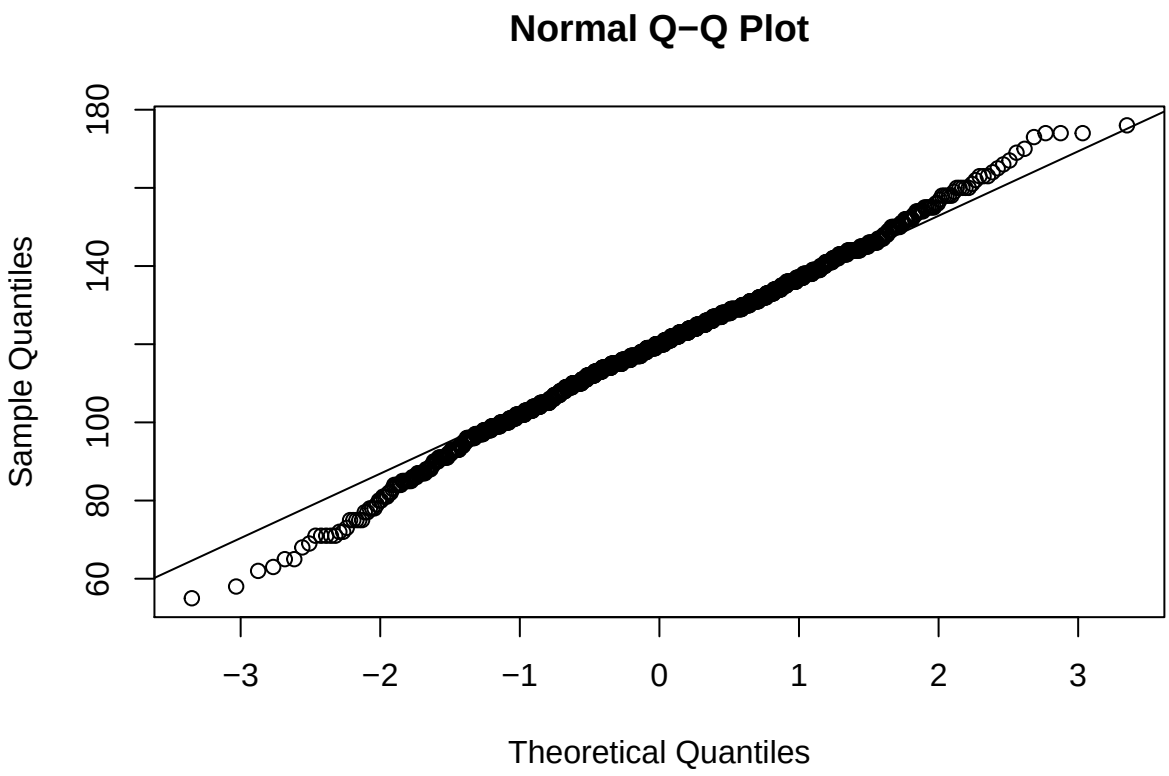
```r
summary(infants$bwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    55.0   108.8   120.0   119.6   131.0   176.0
```

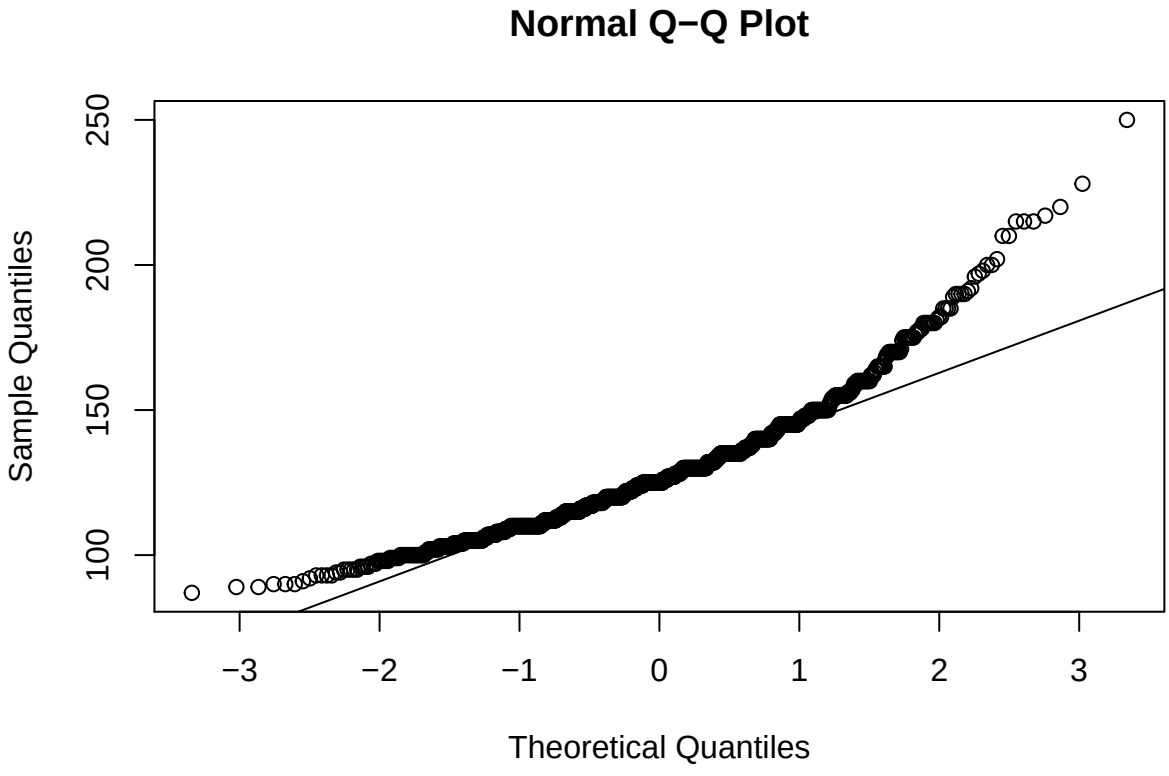
```r
boxplot(infants$bwt)
```




```r
qqnorm(infants$bwt)
qqline(infants$bwt)
```

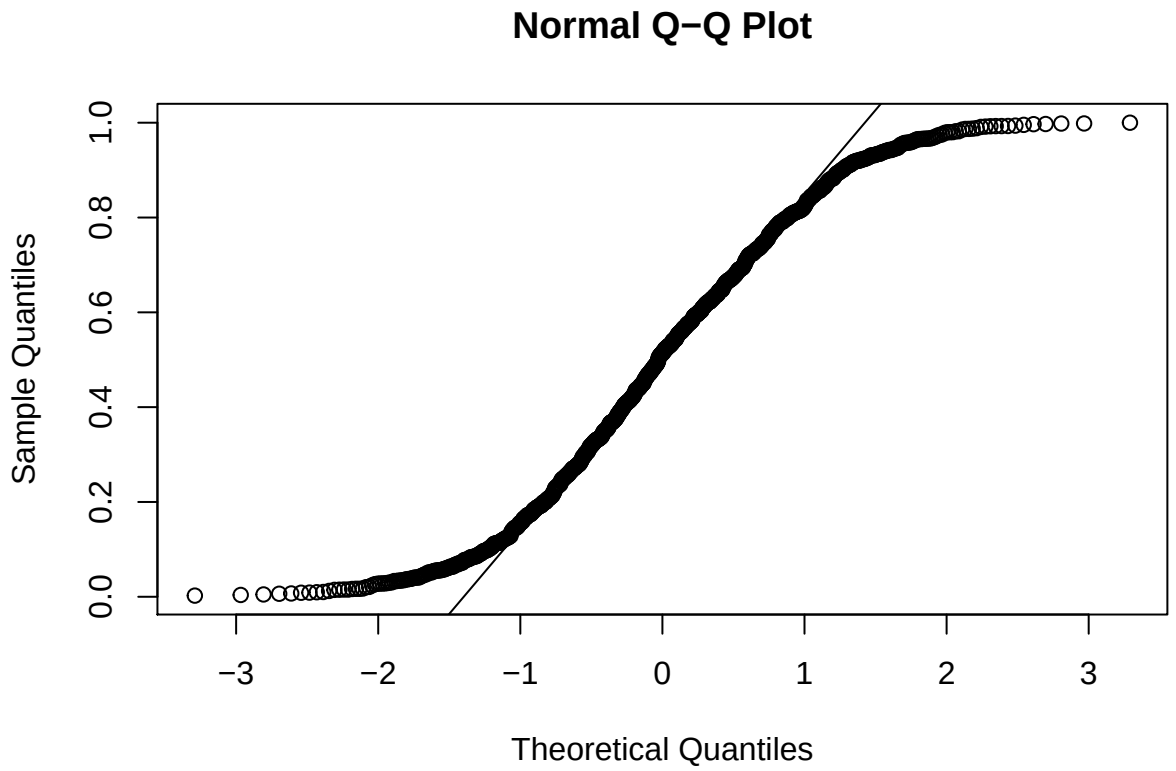



```
qqnorm(infants$wt)
qqline(infants$wt)
```

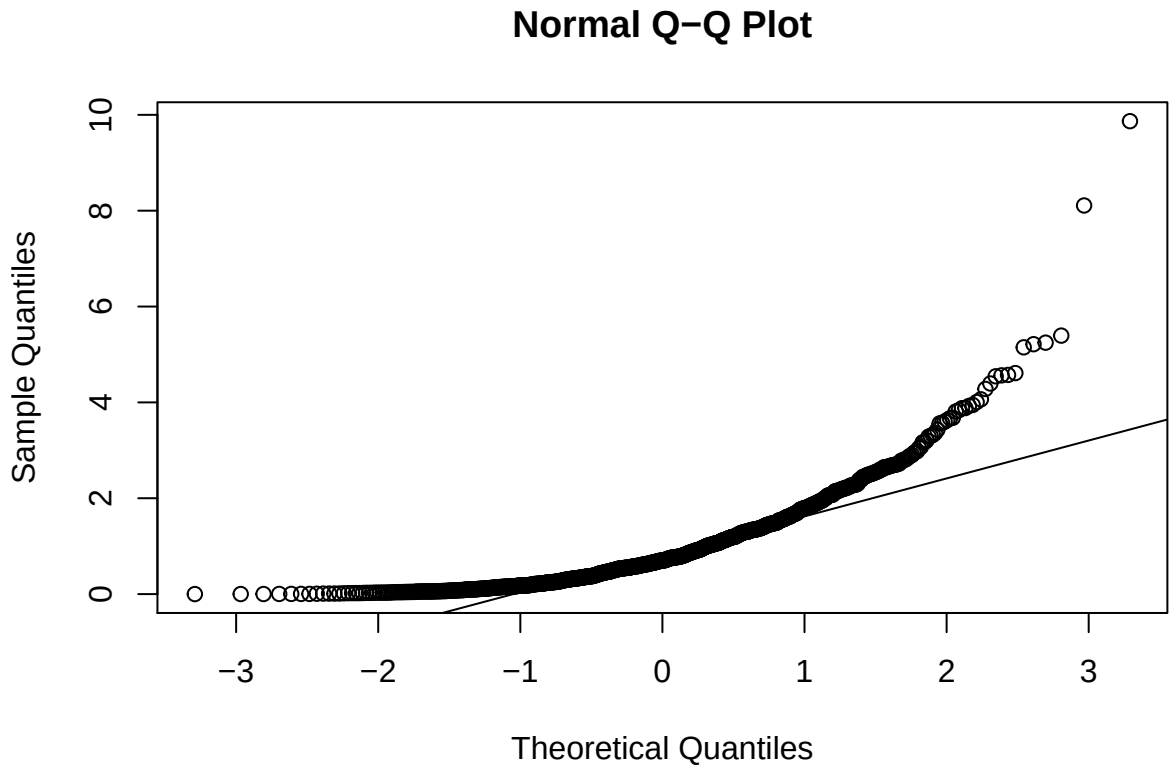



```
X=runif(1000)
qqnorm(X)
```

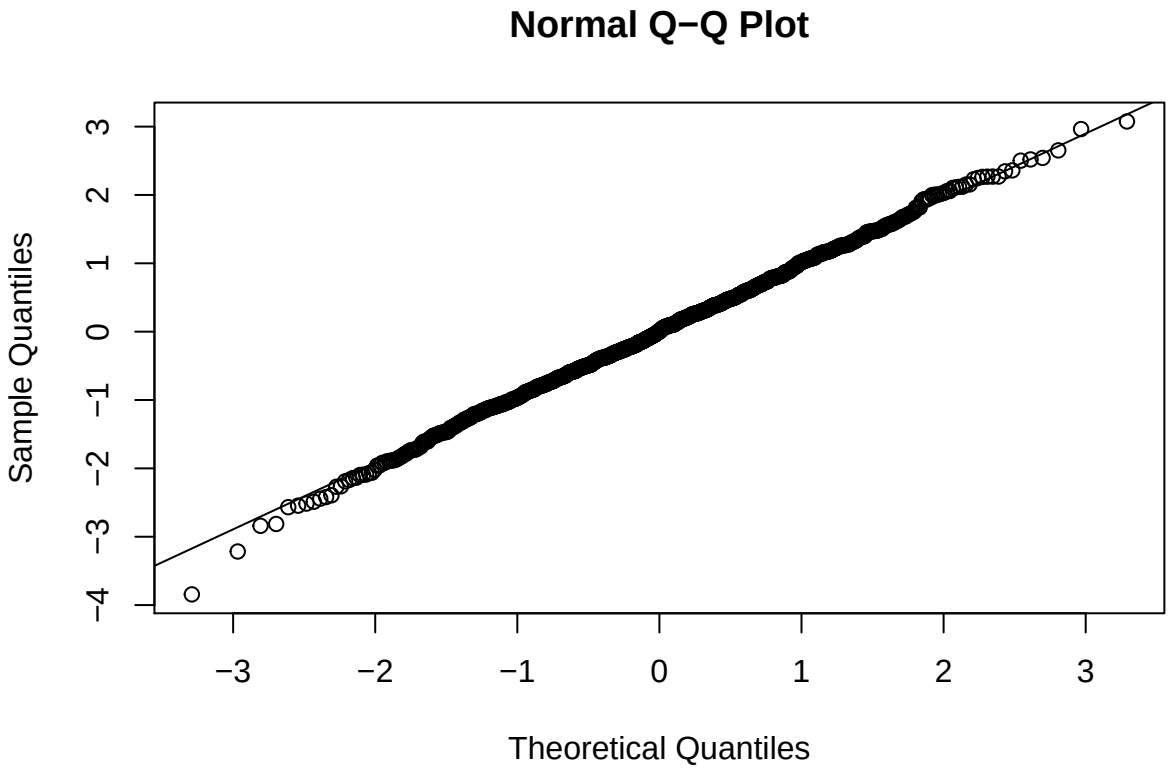
```
qqline(X)
```




```
X=rexp(1000)
qqnorm(X)
qqline(X)
```



```r
X=rnorm(1000)
qqnorm(X)
qqline(X)
```




```r
set.seed(7)
mysample=sample(na.omit(infants$wt),10)
# Part 1
# 1a
true_average = mean(infants$bwt)
x_bar = mean(mysample)
estimated_se = sd(mysample) / sqrt(length(mysample))
# 95% CI Interval
interval = c(x_bar - 1.96*estimated_se, x_bar + 1.96*estimated_se)
interval
```

```
## [1] 125.0711 144.3289
```

```r
# 1b
# creates 1000 95% Confidence Interval
thousand_interval = c()
thousand_averages = c()
num_interval = 0
for (i in 1:1000)
{
sample = sample(na.omit(infants$wt),10)
std_error = sd(sample) / sqrt(length(sample))
thousand_averages = c(thousand_averages, mean(sample))
ci_interval = c(mean(sample) - 1.96*std_error, mean(sample) + 1.96*std_error )
thousand_interval = c(thousand_interval, ci_interval)
if(ci_interval[1] <= true_average & ci_interval[2] >= true_average){
```

```
num_interval = num_interval +1
}
}
# I expect 95% (950 intervals) of the intervals cover the true average
cat("I expect 95% (950 intervals) of the intervals cover the true average")
```

```
## I expect 95% (950 intervals) of the intervals cover the true average
```

```
# The number of 95% CI that has true average is in num_interval
cat("The number of 95% CI that has true average is", num_interval)
```

```
## The number of 95% CI that has true average is 736
```

```
# 1c
sd_averages = sd(thousand_averages)
sd_averages
```

```
## [1] 6.740372
```

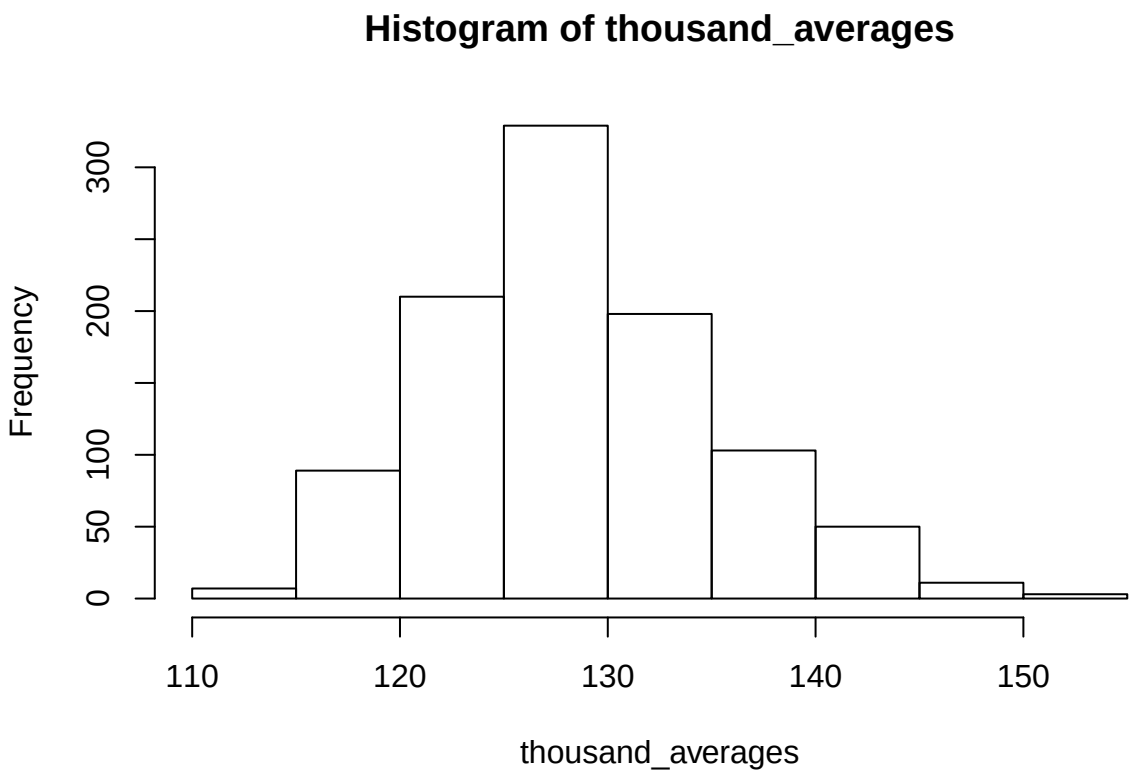
```
estimated_se
```

```
## [1] 4.912682
```

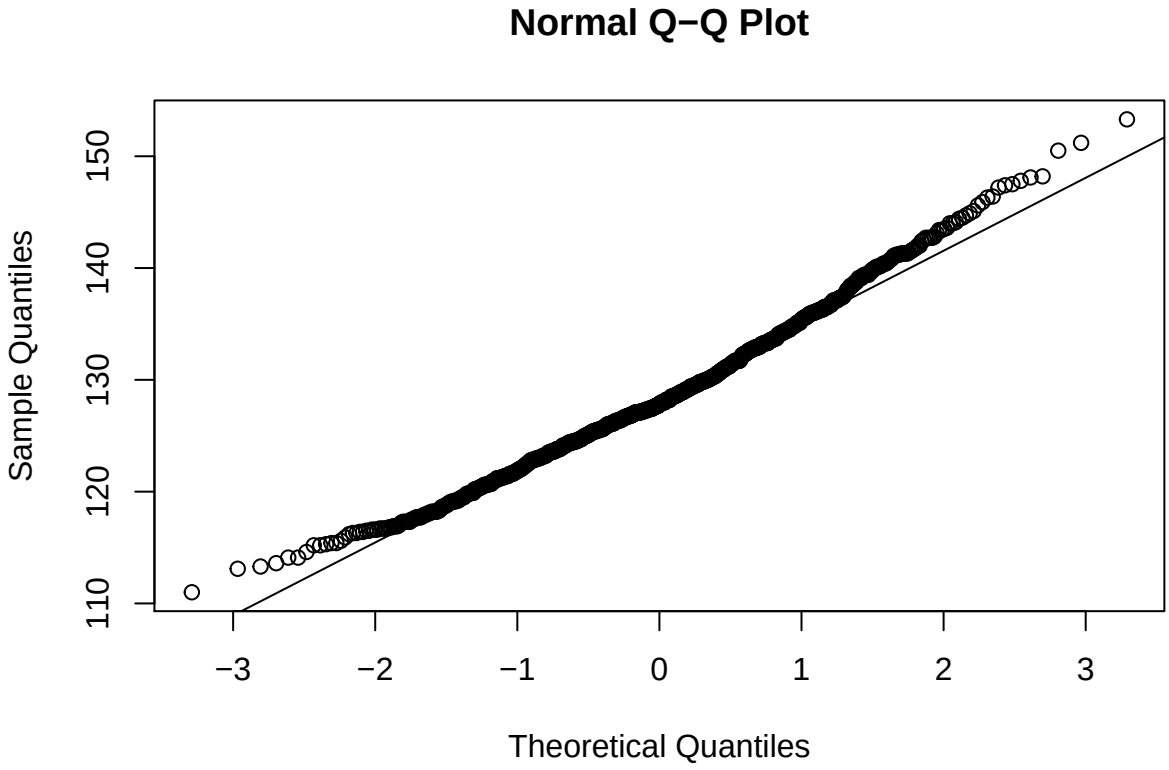
```
cat('The SD of the sample averages, ' , sd_averages, ', is not very close to the estimated standard err
estimated_se)
```

```
## The SD of the sample averages,  6.740372 , is not very close to the estimated standard error,  4.912
```

```
hist(thousand_averages)
```



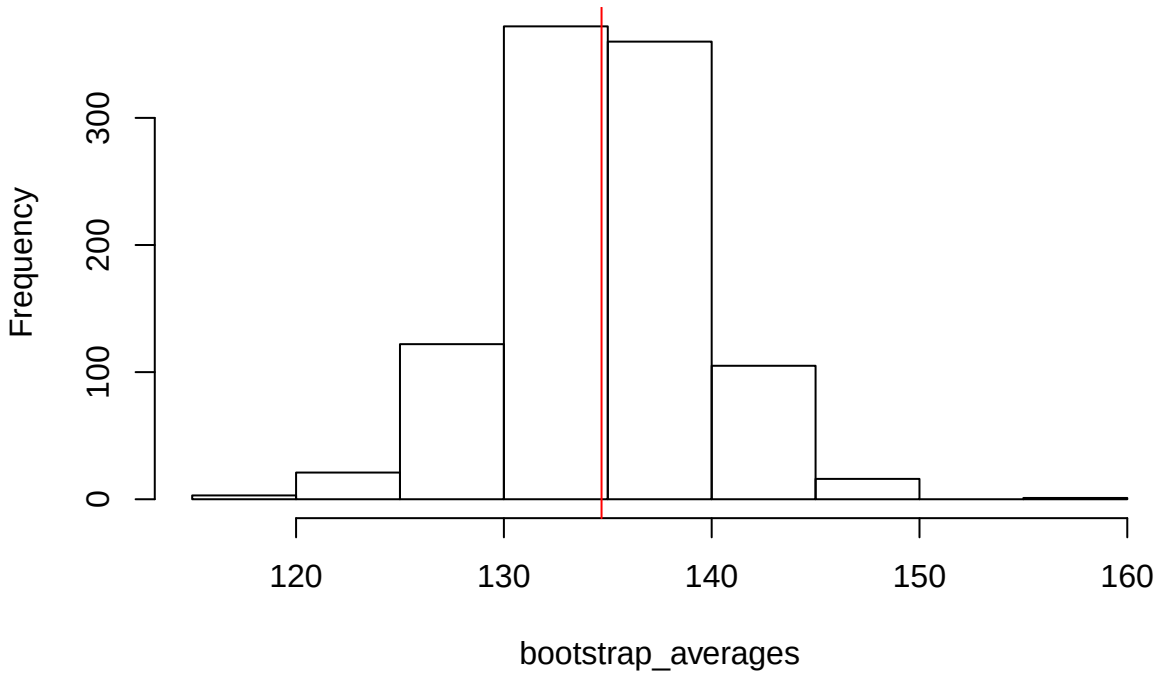

```
qqnorm(thousand_averages)
qqline(thousand_averages)
```




```
cat('As we can see in the histogram, the shape looks very close to a bell shaped curve, normal distribu
The qq plot shows that most of the points are in the line, so the sample average follows the normal cur
')
```

```
## As we can see in the histogram, the shape looks very close to a bell shaped curve, normal distributi
## The qq plot shows that most of the points are in the line, so the sample average follows the normal
```

```
# the confidence interval seems not valid
# Part 2
# 2a
bootStrap = function(mySample, popSize = NULL, B = 1000, repl = FALSE){
if (repl) {
# Bootstrap should be done the same way as original sample, usually without rep
return(replicate(B, mean(sample(mySample, length(mySample), TRUE))))
} else {
vals = sort(unique(mySample))
counts = table(mySample)
# makes the bootstrap pop as rounded version of sample, not quite right
bootPop = rep(vals, round(counts * popSize / length(mySample)))
return(list(bootPop,
bootSamps = replicate(B,mean(sample(bootPop, length(mySample), FALSE))))
)
}
}
bootstrap_averages =  bootStrap(mysample, 10, repl = TRUE)
hist(bootstrap_averages)
abline(v=x_bar,col="red")
```




```
bootstrap_sd = sd(bootstrap_averages)
cat('The SD of the sample averages from using bootstrap, ' , bootstrap_sd, ', is very close to the esti
estimated_se)
```

```
## The SD of the sample averages from using bootstrap,  4.721071 , is very close to the estimated standa
```

```
# 2b
quantile(bootstrap_averages, probs = c(0.025, 0.975))
```

```
##  2.5% 97.5%
## 125.1 144.1
```

```
cat("The 95% confidence interval from the bootstrap is closer to the 95% confidence interval of the boo
```

```
## The 95% confidence interval from the bootstrap is closer to the 95% confidence interval of the boots
```

```
# Part 3
set.seed(7)
mysample=sample(na.omit(infants$wt),100)
mysample
```

```
##   [1] 116 162 115 117 135 140 132 145 150 135 145 157 122 103 170 108 160
##  [18] 120 145 145 113 147 130 146 150 103 147 155 115 138 112 136 106 115
##  [35] 149 127 110 122 127 135 124 145 175 120 137 115 110 135 130 106 102
##  [52] 107 175 122 191 158 147 128 125 120 117 107 116 108 112 111 155 134
##  [69] 115 130 157 116 110 165 111 110 200 110 140 135 110 115 103 113 135
##  [86] 112 154 110 117 135 103 136 128 115 107 135 115 139 115 145
```

```
# 1a
true_average = mean(infants$bwt)
x_bar = mean(mysample)
estimated_se = sd(mysample) / sqrt(length(mysample))
# 95% CI Interval
```

```r
interval = c(x_bar - 1.96*estimated_se, x_bar + 1.96*estimated_se)
interval
```

```
## [1] 125.7439 133.8161
```

```r
# 1b
# creates 1000 95% Confidence Interval
thousand_interval = c()
thousand_averages = c()
num_interval = 0
for (i in 1:1000)
{
sample = sample(na.omit(infants$wt),100)
std_error = sd(sample) / sqrt(length(sample))
thousand_averages = c(thousand_averages, mean(sample))
ci_interval = c(mean(sample) - 1.96*std_error, mean(sample) + 1.96*std_error )
thousand_interval = c(thousand_interval, ci_interval)
if(ci_interval[1] <= true_average & ci_interval[2] >= true_average){
num_interval = num_interval +1
}
}
# I expect 95% (950 intervals) of the intervals cover the true average
cat("I expect 95% (950 intervals) of the intervals cover the true average")
```

```
## I expect 95% (950 intervals) of the intervals cover the true average
```

```r
# The number of 95% CI that has true average is in num_interval
cat("The number of 95% CI that has true average is", num_interval)
```

```
## The number of 95% CI that has true average is 2
```

```r
# 1c
sd_averages = sd(thousand_averages)
sd_averages
```

```
## [1] 2.000529
```

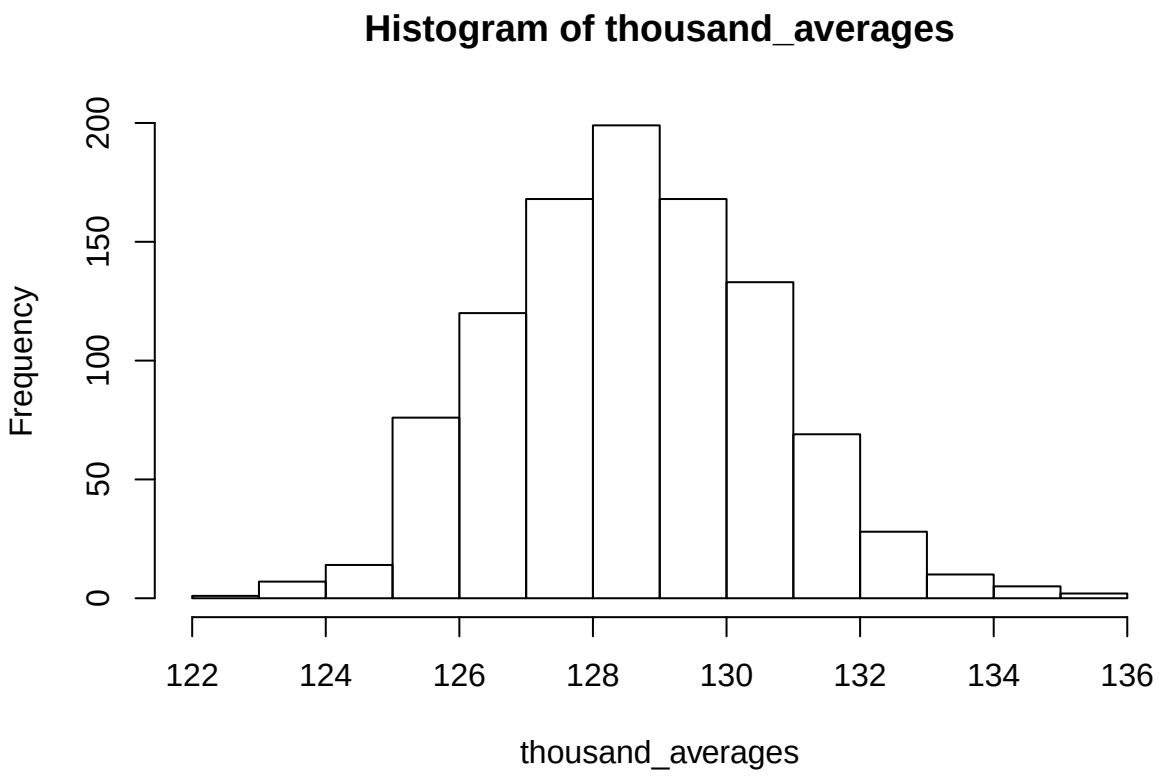
```r
estimated_se
```

```
## [1] 2.059253
```

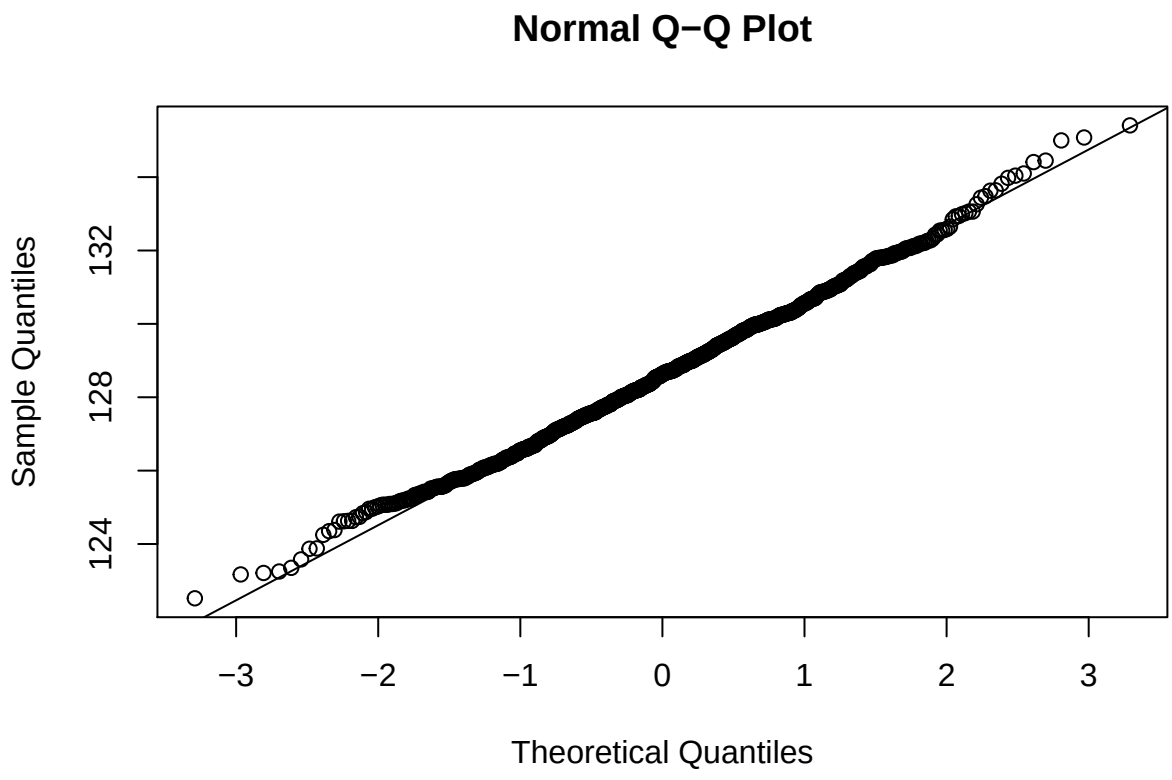
```r
cat('The SD of the sample averages, ' , sd_averages, ', is very close to the estimated standard error,
estimated_se)
```

```
## The SD of the sample averages,  2.000529 , is very close to the estimated standard error,  2.059253
```

```r
hist(thousand_averages)
```



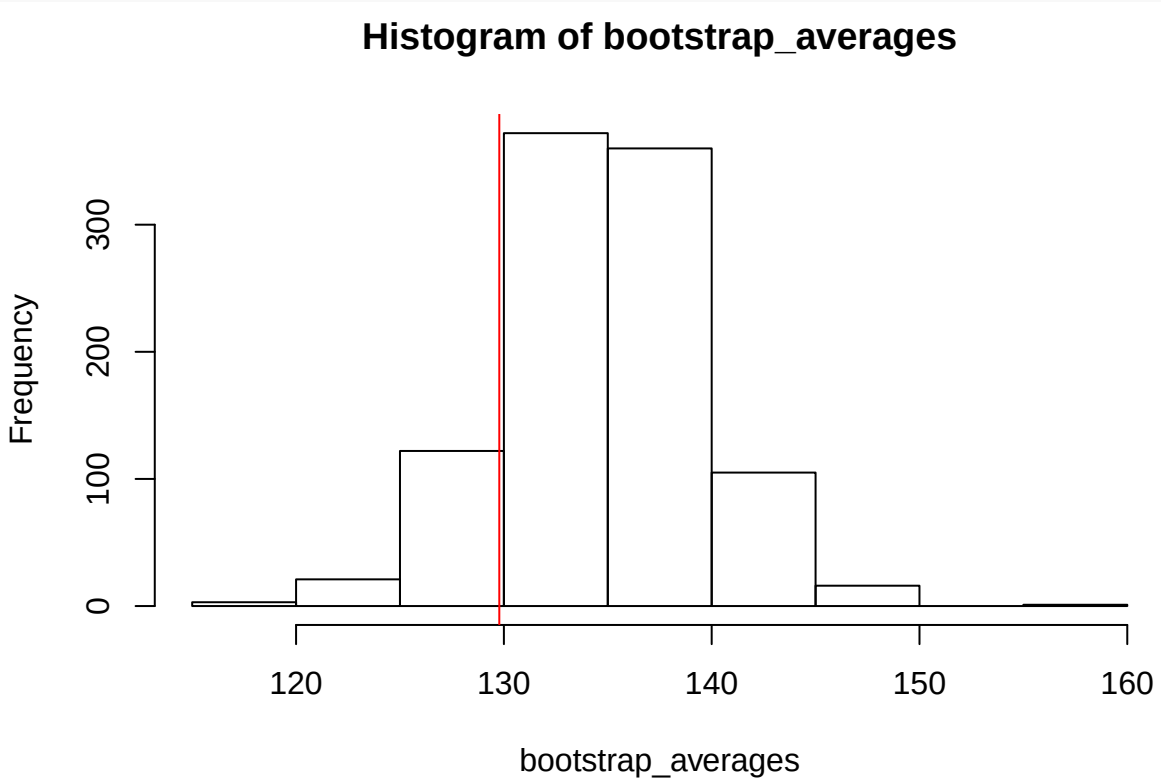

```
qqnorm(thousand_averages)
qqline(thousand_averages)
```




```
cat('As we can see in the histogram, the shape looks very close to a bell shaped curve, normal distribu
The qq plot shows that most of the points are in the line, so the sample average follows the normal curv
```

```
## As we can see in the histogram, the shape looks very close to a bell shaped curve, normal distributi
## The qq plot shows that most of the points are in the line, so the sample average follows the normal
```

```
# The confidence interval is valid
# Part 2
# 2a
hist(bootstrap_averages)
abline(v=x_bar,col="red")
```




```
bootstrap_sd = sd(bootstrap_averages)
cat('The SD of the sample averages from using bootstrap, ' , bootstrap_sd, ', is very close to the esti
estimated_se)
```

```
## The SD of the sample averages from using bootstrap,  4.721071 , is very close to the estimated standa
```

```
# 2b
quantile(bootstrap_averages, probs = c(0.025, 0.975))
```

```
##  2.5% 97.5%
## 125.1 144.1
```

```
ci_interval
```

```
## [1] 125.1747 132.2253
```

```
cat("The 95% confidence interval from the bootstrap is closer to the 95% confidence interval of the boo
```

```
## The 95% confidence interval from the bootstrap is closer to the 95% confidence interval of the boots
```