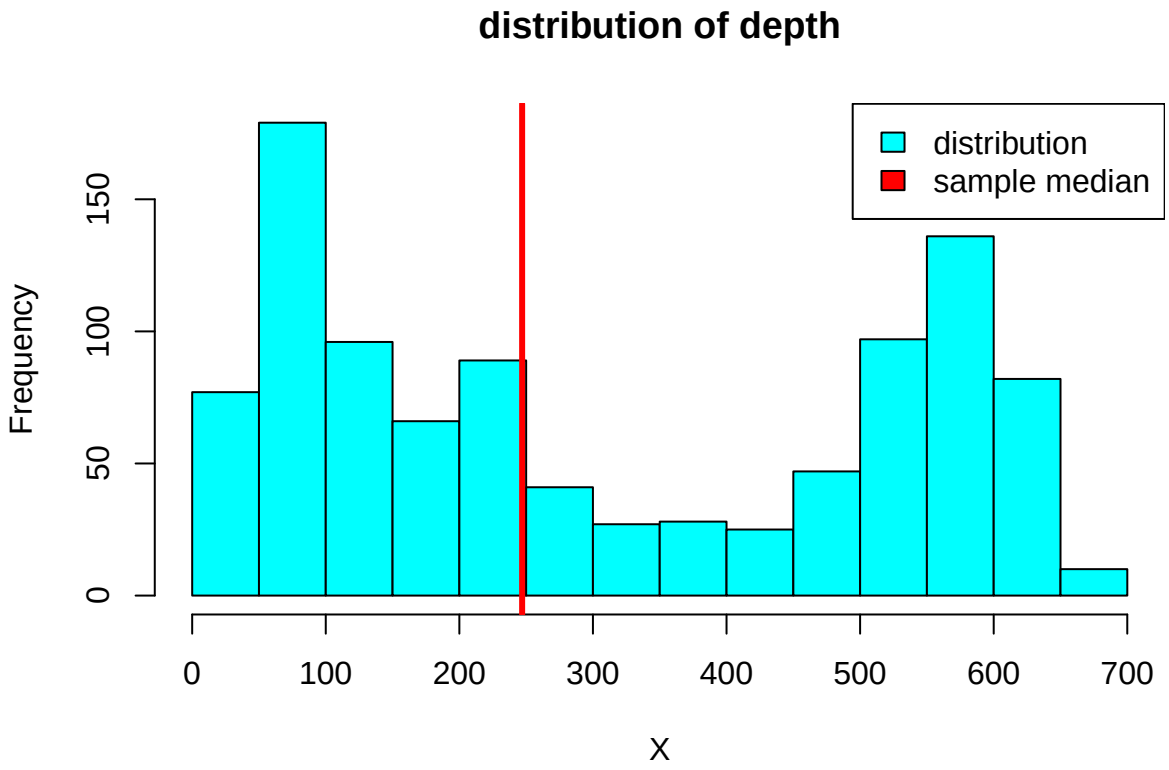
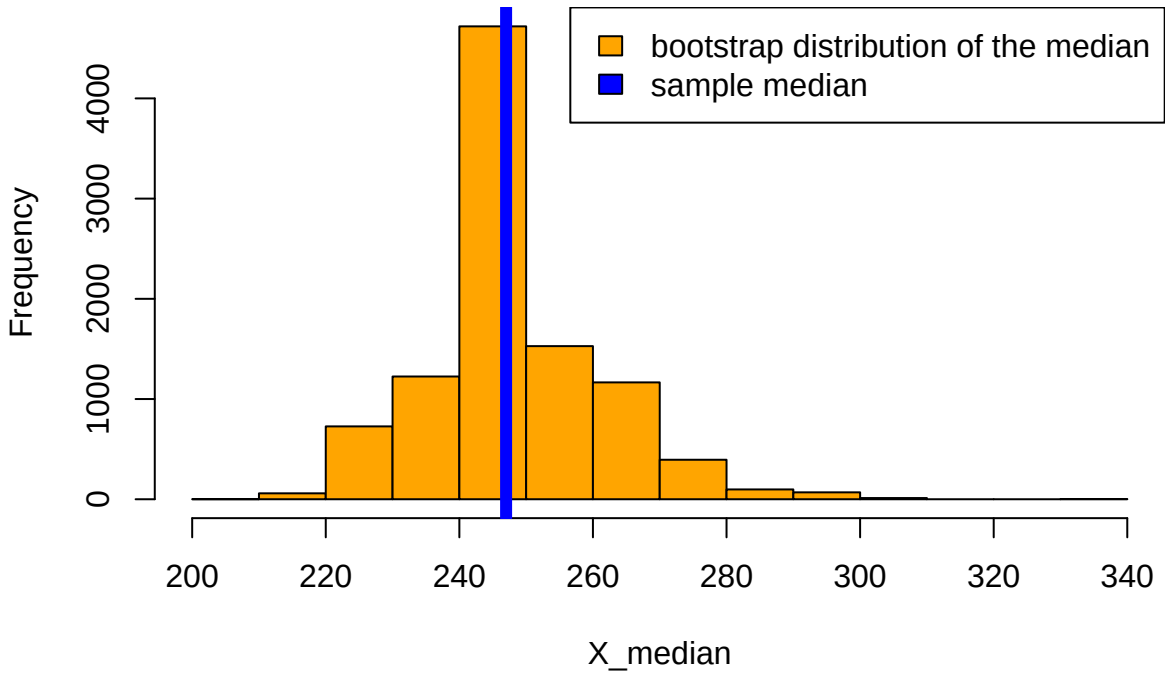
Stat403_HW7

```r
#1a
X <- quakes$depth
hist(X,col="cyan",main="distribution of depth")
abline(v=median(X), lwd=3, col="red")
legend("topright",c("distribution","sample median"),fill=c("cyan","red"))
```




```r
#1b
set.seed(1234)
n = length(X)
B = 10000
X_median = rep(NA, B)
for(i_BT in 1:B){
  w = sample(n,n,replace=T)
  X_BT <- X[w]
  X_median[i_BT] = median(X_BT)
}

hist(X_median, col="orange",main="bootstrap distribution of the median and the sample median")
abline(v=median(X),lwd=6, col="blue")
legend("topright",c("bootstrap distribution of the median","sample median"),fill=c("orange","blue"))
```




```
#1c
#The bootstrap estimate of the variance and MSE of the sample median do not agree.

var(X_median)
```

```
## [1] 175.5433
```

```
mean((X_median-median(X))^2)
```

```
## [1] 177.5564
```

```
#1d
#The IQR form both method match with the histogram from part b, and I expect to see no obvious differen
c(median(X) - qnorm(0.95)*sd(X_median), median(X) + qnorm(0.95)*sd(X_median))
```

```
## [1] 225.2069 268.7931
```

```
quantile(X_median, c(0.05,0.95))
```

```
##  5% 95%
## 227 271
```

```
#2a
X = rock$area
mean(X)
```

```
## [1] 7187.729
```

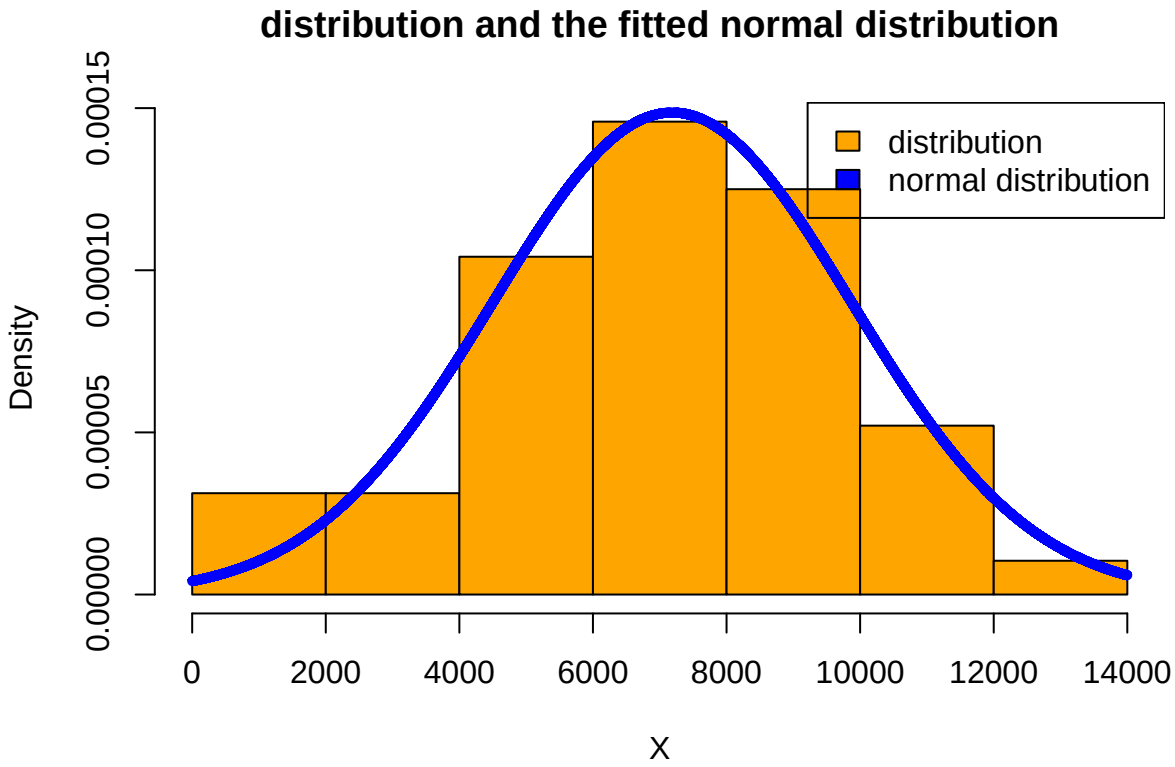
```
sd(X)
```

```
## [1] 2683.849
```

```
median(X)
```

```
## [1] 7487
```

```r
hist(X,probability=T,col="orange",main="distribution and the fitted normal distribution")
lines(x=seq(0,14000,0.1),y=dnorm(x=seq(0,14000,0.1),mean=mean(X),sd=sd(X)),col="blue",lwd=5)
legend("topright",c("distribution","normal distribution"),fill=c("orange","blue"))
```




```r
#2b
#the variance and MSE of the median estimated from the bootstrap samples do not agree with each other.
set.seed(1234)
B = 10000
n <- length(X)
X_median = rep(NA, B)
for(i_BT in 1:B){
  X_BT <- rnorm(n,mean=mean(X),sd=sd(X))
  X_median[i_BT] = median(X_BT)
}
var(X_median)
```

```
## [1] 229103.8
```

```r
mean((X_median-median(X))^2)
```

```
## [1] 318747.4
```

```r
#2c
#The variance and MSE of the median estimated from the bootstrap samples agree with each other.
set.seed(1234)
B = 10000
n <- length(X)
X_median = rep(NA, B)
for(i_BT in 1:B){
  w <- sample(n,n,replace=T)
  X_BT <- X[w]
```

```
  X_median[i_BT] = median(X_BT)
}
var(X_median)
```

```
## [1] 211040.7
```

```
mean((X_median-median(X))^2)
```

```
## [1] 211035.2
```

```
#3a
data <- matrix(NA,nrow=10000,ncol=500)
for(i in 1:10000){
  for(j in 1:500) {
    data[i,j] <- runif(1,0,1)
  }
}
X_IQR <- rep(NA,10000)
for(i in 1:10000) {
  X_IQR[i] <- IQR(data[i,])
}
var(X_IQR)
```

```
## [1] 0.0005075596
```

```
mean((X_IQR-0.5)^2)
```

```
## [1] 0.0005094214
```

```
#3b
set.seed(1234)
X <- runif(500,0,1)
X_IQR = rep(NA, B)
for(i_BT in 1:10000){
  w <- sample(n,n,replace=T)
  X_BT <- X[w]
  X_IQR[i_BT] = IQR(X_BT)
}
var(X_IQR)
```

```
## [1] 0.0027987
```

```
mean((X_IQR-IQR(X))^2)
```

```
## [1] 0.007809219
```

```
#3c
#I'm surprised that answers in 3a and 3b are so close, and the MSE and variance from both methods are a
```

```
xx = chickwts[chickwts$feed=="meatmeal",1]
yy = chickwts[chickwts$feed=="casein",1]

mean1 = mean(xx)
mean2 = mean(yy)

diff_med = abs(mean1-mean2)
n = length(xx) + length(yy)
```

```r
data_pull = c(xx, yy)

N_per = 10000
diff_med_per = rep(NA, N_per)
for(i_per in 1:N_per){
  w_per = sample(n, n, replace=F)
  data_per = data_pull[w_per]
  # data after permutation
  data1_new = data_per[1:length(xx)]
  data2_new = data_per[(length(xx)+1):n]
  # first n_M are new group M; the others are new group F
  diff_new = abs(mean(data1_new)-mean(data2_new))
  # compute the difference
  diff_med_per[i_per] = diff_new
}

(length(which(diff_med_per >= diff_med))+1)/(N_per+1)
```

```
## [1] 0.1035896
```

```r
par(mfrow=c(1,1))
hist(diff_med_per, col="cyan")
abline(v = diff_med, col="red",lwd=6)
legend("topright",c("distribution","actual sample value"),fill=c("cyan","red"))
```