Stats 10 Lab 4 Submission
Name: Enmin Zhou
UID: 104756697

Exercise 1

1) a)

```
> pawnee <-read.csv("~/Desktop/stats10/lab4/pawnee.csv", header = TRUE)
> head(pawnee)
  ID Latitude Longitude Arsenic Sulfur New_hlth_issue
1  1 41.09414 -85.60974       0      0              N
2  2 41.09054 -85.70344       0    130              N
3  3 41.08601 -85.71996       4    170              N
4  4 41.08100 -85.75415       0      0              Y
5  5 41.07435 -85.70043       0      0              N
6  6 41.07399 -85.71788       0      0              N
> dim(pawnee)
[1] 541   6
```

b)

```
> set.seed(1337)
> sample_index <- sample(541, size=30)
> sample_pawnee <- pawnee[sample_index,]
> head(sample_pawnee)
     ID Latitude Longitude Arsenic Sulfur New_hlth_issue
312 312 41.01716 -85.66949     1.0      0              N
305 305 41.01742 -85.65858     0.5     40              N
40   40 41.06414 -85.72544     0.0      0              N
245 245 41.02714 -85.73328     0.0      0              N
201 201 41.03244 -85.63653     0.0      0              N
178 178 41.03568 -85.64353     0.0      0              Y
> dim(sample_pawnee)
[1] 30  6
```

c)

```
> mean(sample_pawnee$Arsenic)
[1] 5.566667
> p.hat<-mean(sample_pawnee$New_hlth_issue=="Y")
> print(p.hat)
[1] 0.2666667
```

d)

x bar for sample mean
p hat for sample proportion

e)

```
> se <- sqrt(p.hat*(1-p.hat)/30)
> z1<-qnorm(p=0.95)
> z2<-qnorm(p=0.975)
> z3<-qnorm(p=0.995)
> p.hat+c(-1,1)*z1*se
[1] 0.1338656 0.3994678
> p.hat+c(-1,1)*z2*se
[1] 0.1084244 0.4249090
> p.hat+c(-1,1)*z3*se
[1] 0.05870105 0.47463228
```

f)

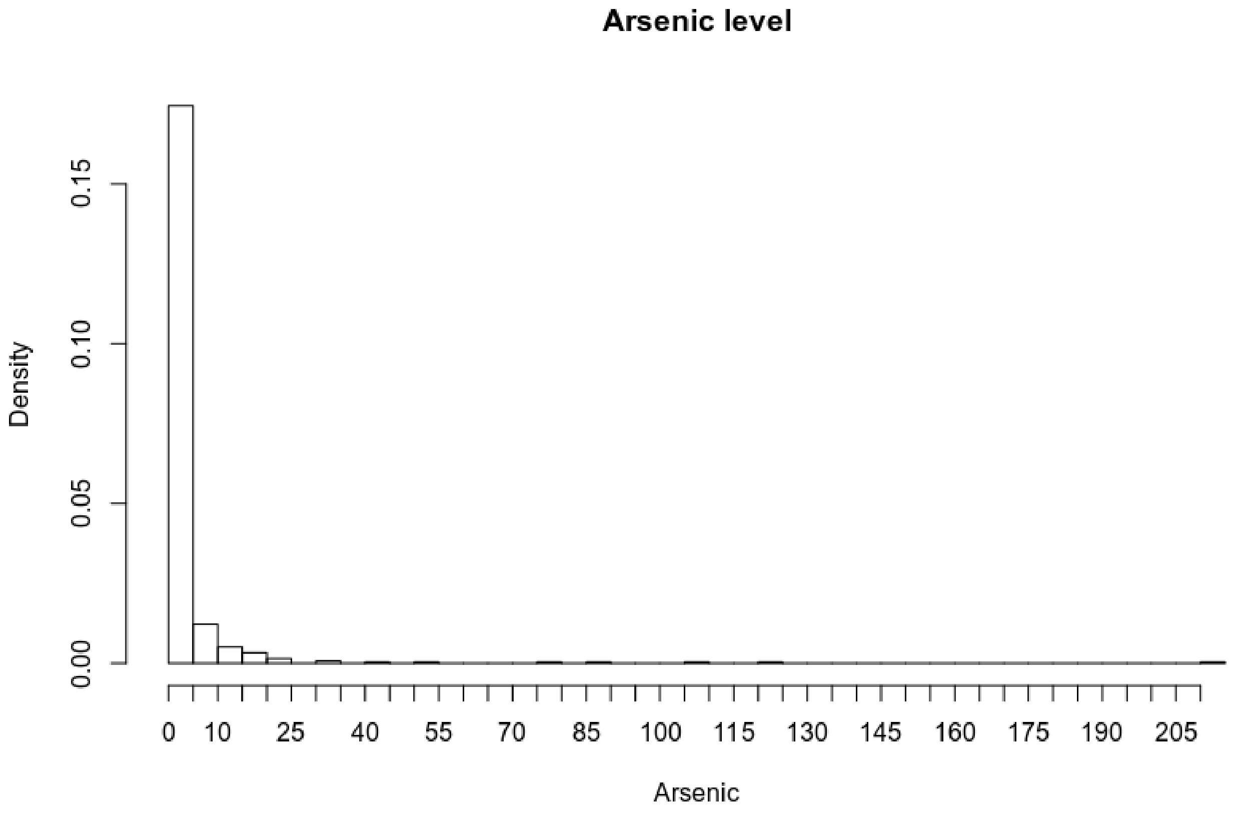
[0,1]

g)

```
> mean(pawnee$New_hlth_issue=="Y")
[1] 0.2920518
```

h)

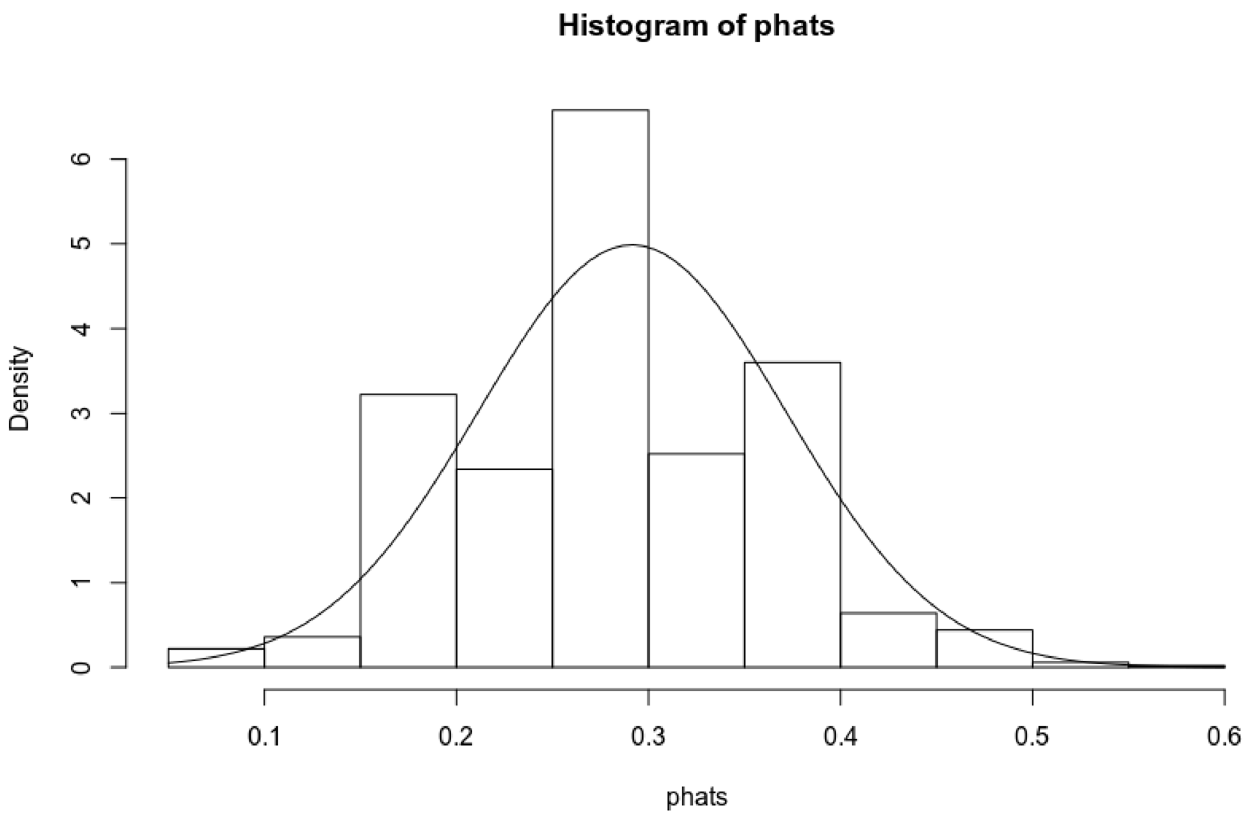
```
> hist(pawnee$Arsenic, breaks=42,xaxt='n',prob=T, xlab="Arsenic", main="Arsenic level")
> axis(side=1, at=seq(0,210,l=43), labels=seq(0,210,l=43))
```



Exercise 2

2)a)

```
> library(histogram)
> hist(phats, probability = TRUE)
> curve(dnorm(x,mean(phats),sd(phats)), add=TRUE)
```



b)

```
> mean(phats)
[1] 0.2914333
> sd(phats)
[1] 0.07997713
```

c)

It is approximately normal because the histogram is symmetric and it matches the curve of normal distribution.

d)

```
> p<-mean(pawnee$New_hlth_issue=="Y")
> sd<-sqrt(p*(1-p)/30)
> p
[1] 0.2920518
> sd
[1] 0.08301757
```

They are approximately the same.

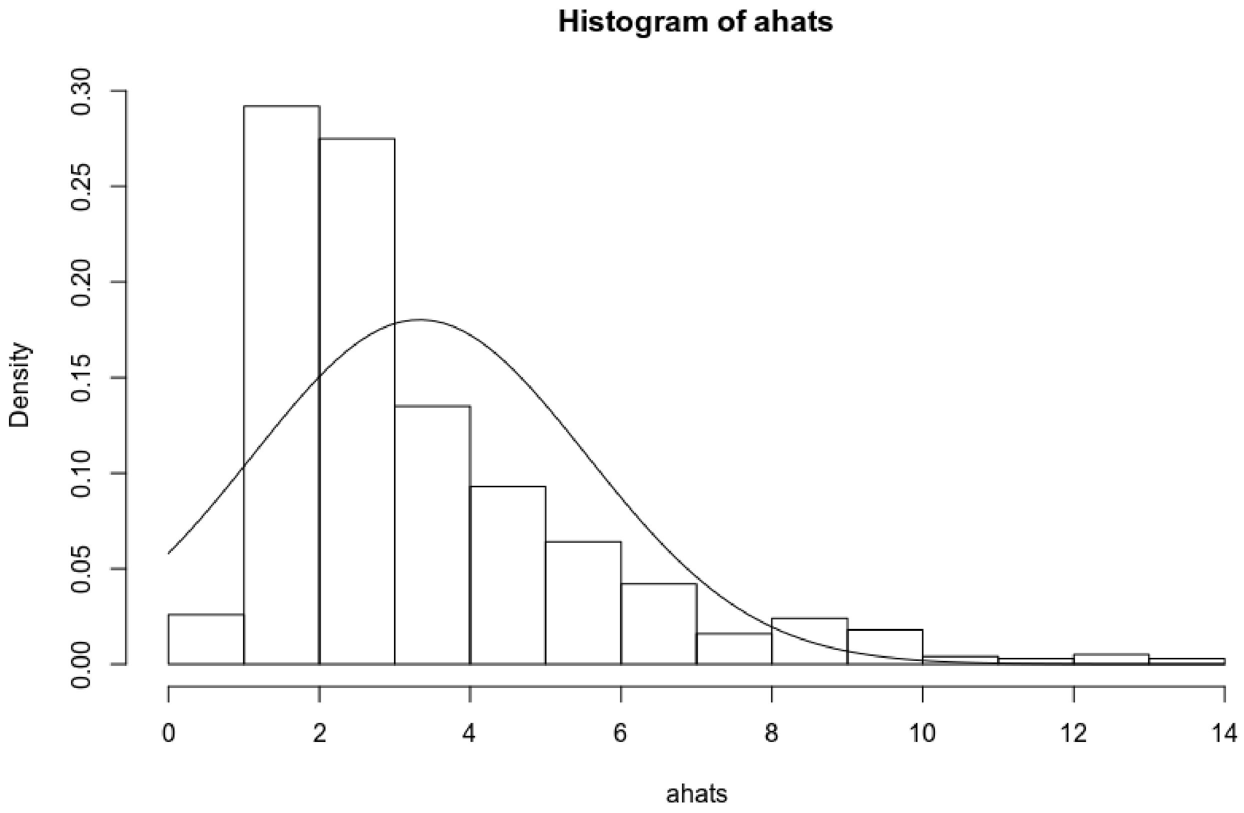
Exercise 3

3)a)

```
> n<-30
> N<-541
> M<-1000
> ahats <-numeric(M)
> set.seed(123)
> for(i in seq_len(M)){
+  index <- sample(N, size=n)
+  sample_i <- pawnee[index,]
+  ahats[i] <- mean(sample_i$Arsenic)
+ }
```

b)

```
> hist(ahats, probability = TRUE)
> curve(dnorm(x,mean(ahats),sd(ahats)), add=TRUE)
```



c)

It is not normal because the histogram is right skewed and not symmetric. The histogram graph does not match the normal distribution curve. The arsenic levels has more values close to 0, which makes the histogram right skewed so it is different from the result in exercise 2.