Stats 10 Lab 4

Name: Brandon Truong

Section 1

RStudio Version 1.2.5033

```
pawnee <-read.csv('~/UCLA Coursework/STATS 10/pawnee.csv', header=TRUE)
```

a)

```
> 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)
> rowsToSample <- sample(x=1:nrow(pawnee),size=30,replace = F)
> pawnee_sample <- pawnee[rowsToSample,]

> head(pawnee_sample)
     ID Latitude Longitude Arsenic Sulfur New_hlth_issue
147 147 41.03971 -85.72783       2    100              N
49   49 41.06113 -85.65553       0      0              Y
210 210 41.03178 -85.64253       0      0              N
356 356 41.01178 -85.66516       0      0              N
425 425 41.00096 -85.72899       0      0              N
239 239 41.02772 -85.72901       0      0              N
```

c)

The mean arsenic value of my sample of size 30 is 0.85 ppm

```
> mean(pawnee_sample$Arsenic)
[1] 0.85
```

The proportion of households with health issues can be found with p hat, which is 0.2 or 20% of my sample.

```
> p.hat <- mean(pawnee_sample$New_hlth_issue == "Y")
> print(p.hat)
[1] 0.2
```

d)

We would use x bar denoted by the symbol $\bar{x}$ for the mean arsenic level since it is the mean value in a sample and p hat denoted by $\hat{p}$ for the proportion of health issues in the sample

```
> #See Lab Manual
> #x_bar
> #p_hat
```

e)
We first calculate the standard error and the z scores for the sample. We then can use $\hat{p} \pm z * \sqrt{\hat{p}(1-\hat{p})/n}$ to calculate the confidence intervals, which we can use z1-z3 and se to calculate

We get these results then:

```
> se <- sqrt(p.hat*(1-p.hat)/30) # Standard Error
> #Critical values
> z1 <-qnorm(p=0.95) #95th Percentile
> z2 <-qnorm(p=0.975) #97.5 Percentile
> z3 <-qnorm(p=0.995) #99.5 Percentile
>
> p.hat+c(-1,1)*z1*se #90% Confidence Interval
[1] 0.07987688 0.32012312
> p.hat+c(-1,1)*z2*se #95% Confidence Interval
[1] 0.05686447 0.34313553
> p.hat+c(-1,1)*z3*se #99% Confidence Interval
[1] 0.01188802 0.38811198
```

We can interpret as:
We are 90% confident that the true population proportion of households with health issues is between (0.07987688, 0.32012312), or 7.9% and 32.0%

We are 95% confident that the true population proportion of households with health issues is between (0.05686447, 0.34313553), or 5.6% and 34.3%

We are 99% confident that the true population proportion of households with health issues is between (0.01188802, 0.38811198), or 1.1% and 38.8%

f)
The bounds for a 100% confidence interval would be [0,1] since it accounts for 0% to 100%. (All possible $\hat{p}$ values)

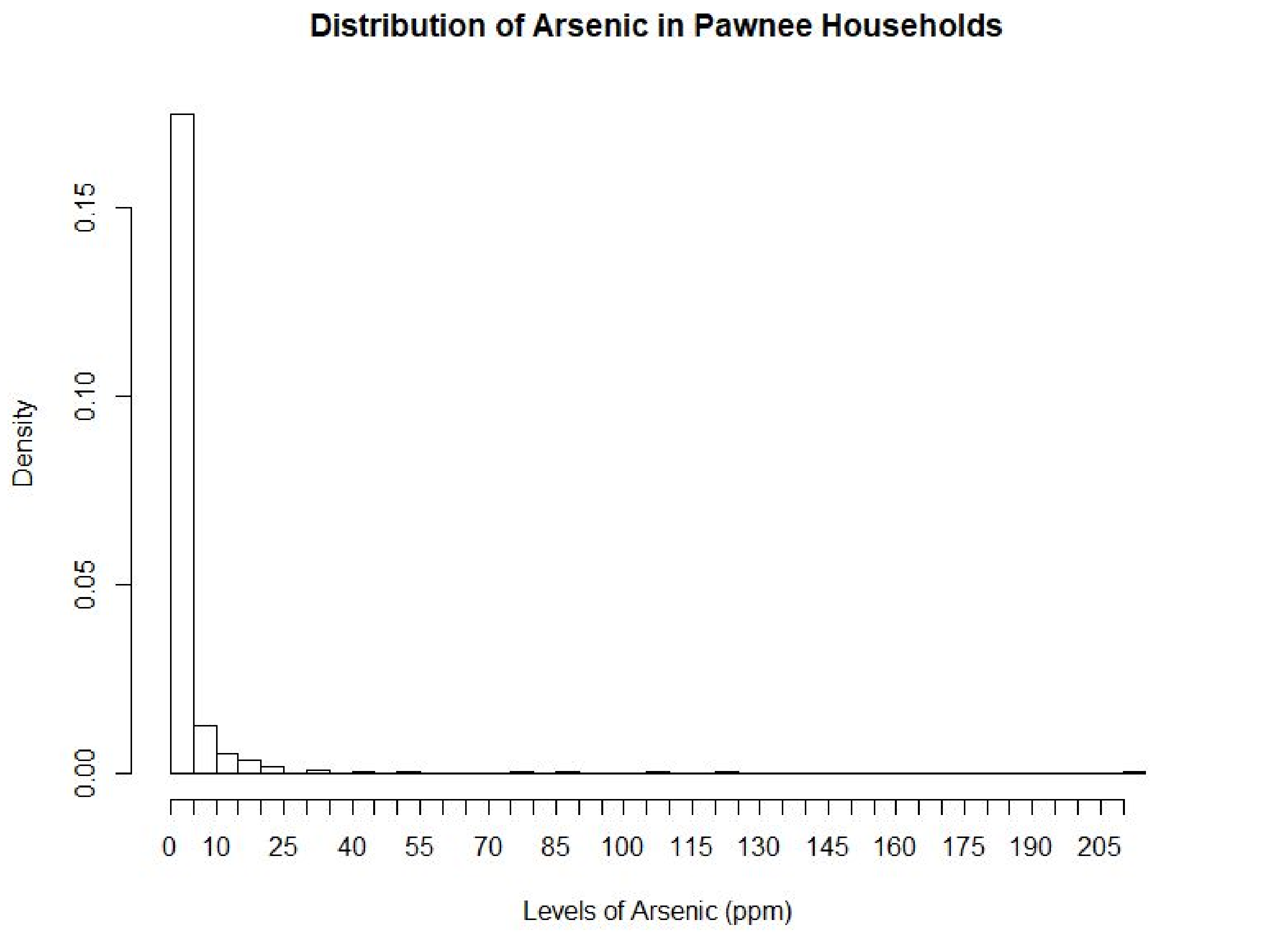
```
> # [0,1]
```

g)
The proportion of all households that experienced a health issue is 0.2920518, or 29.2%

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

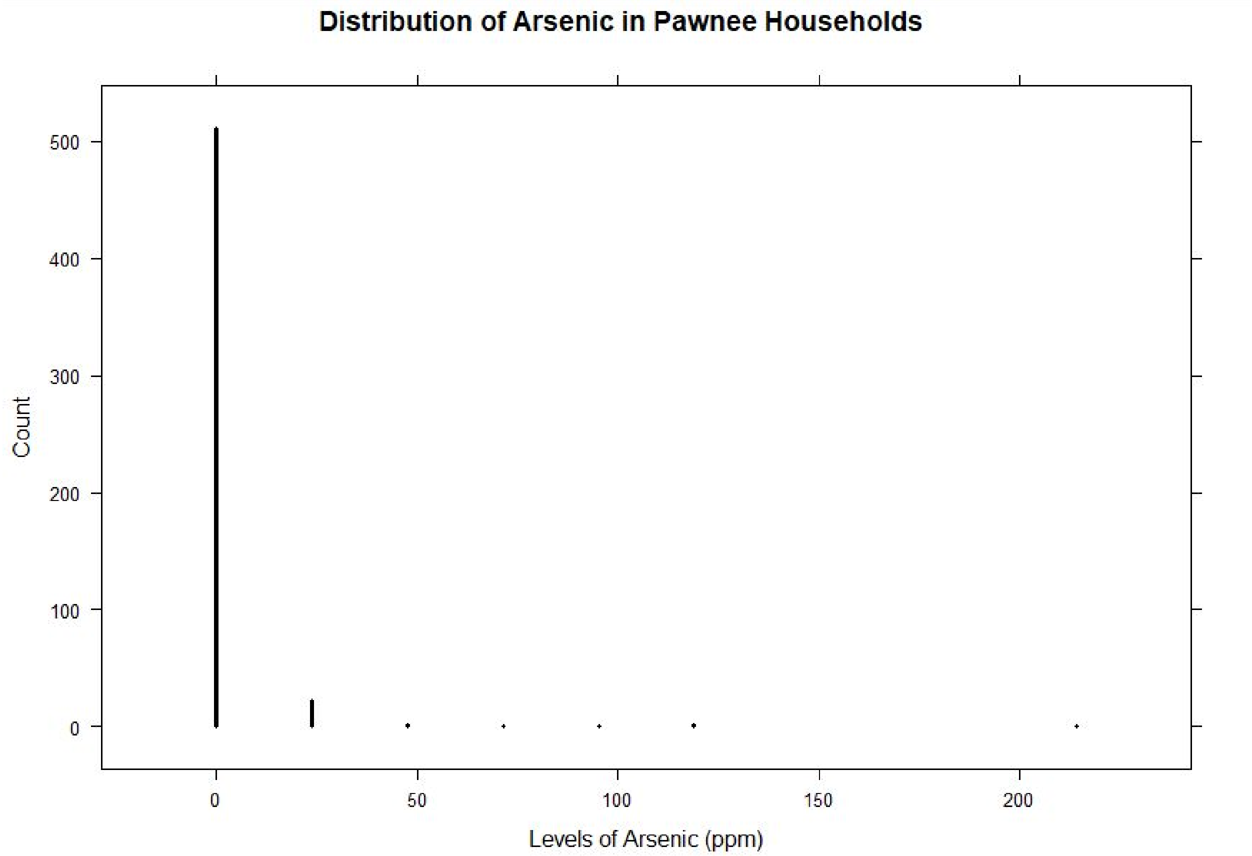
h)

We can get a graph displaying the distribution of arsenic levels in Pawnee houses by creating a histogram of the distribution



```
> hist(pawnee$Arsenic,breaks=42,xaxt='n',prob = T,xlab="Levels of Arsenic
(ppm)",main="Distribution of Arsenic in Pawnee Households")
> axis(side=1,at=seq(0,210,l=43),labels=seq(0,210,l=43))
```

We can also get a similar result by creating a dotPlot of the graph



```
dotPlot(pawnee$Arsenic,col="black",cex=5,xlab="Levels of Arsenic
(ppm)",main="Distribution of Arsenic in Pawnee Households")
```

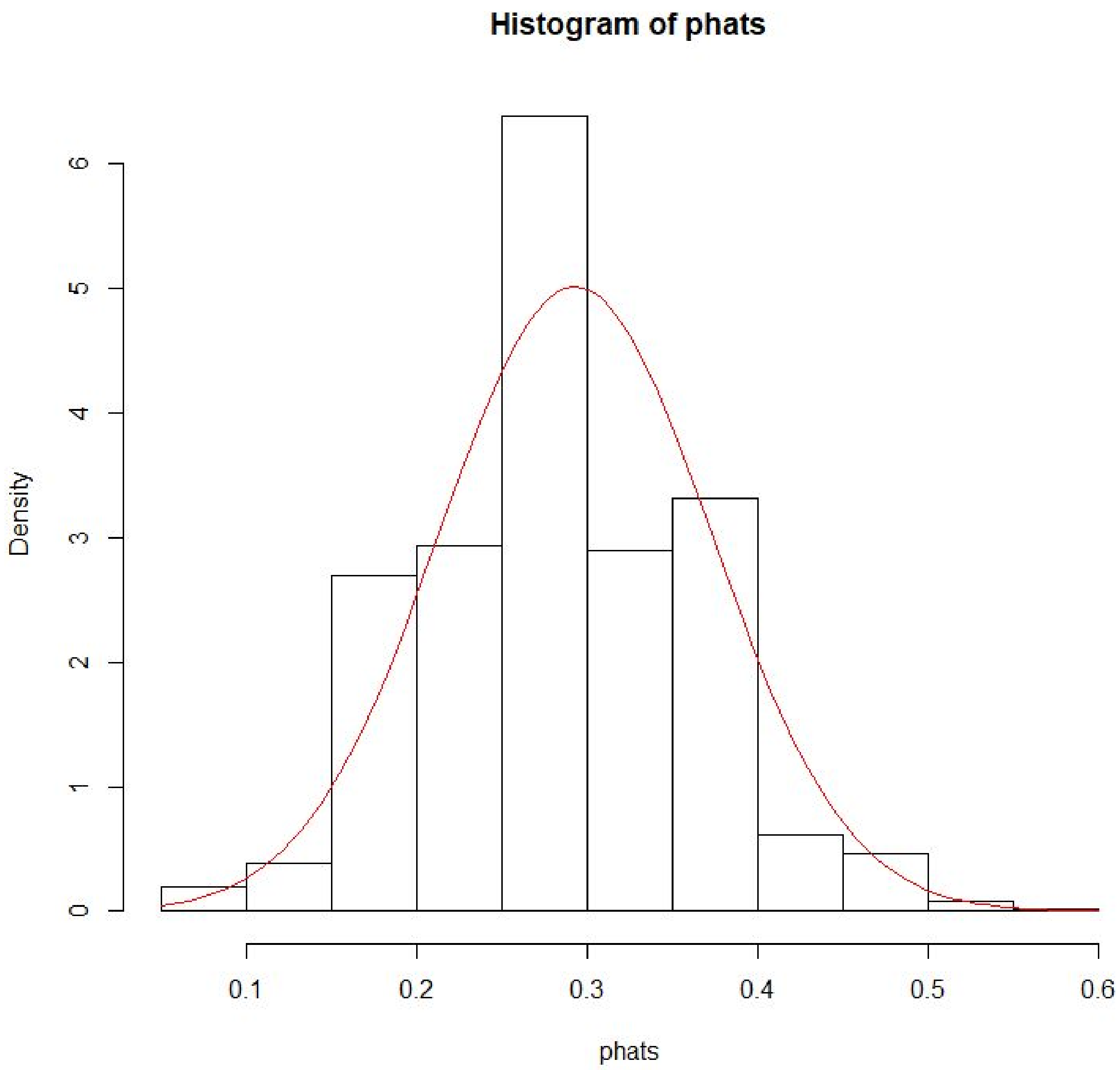
Section 2

a)

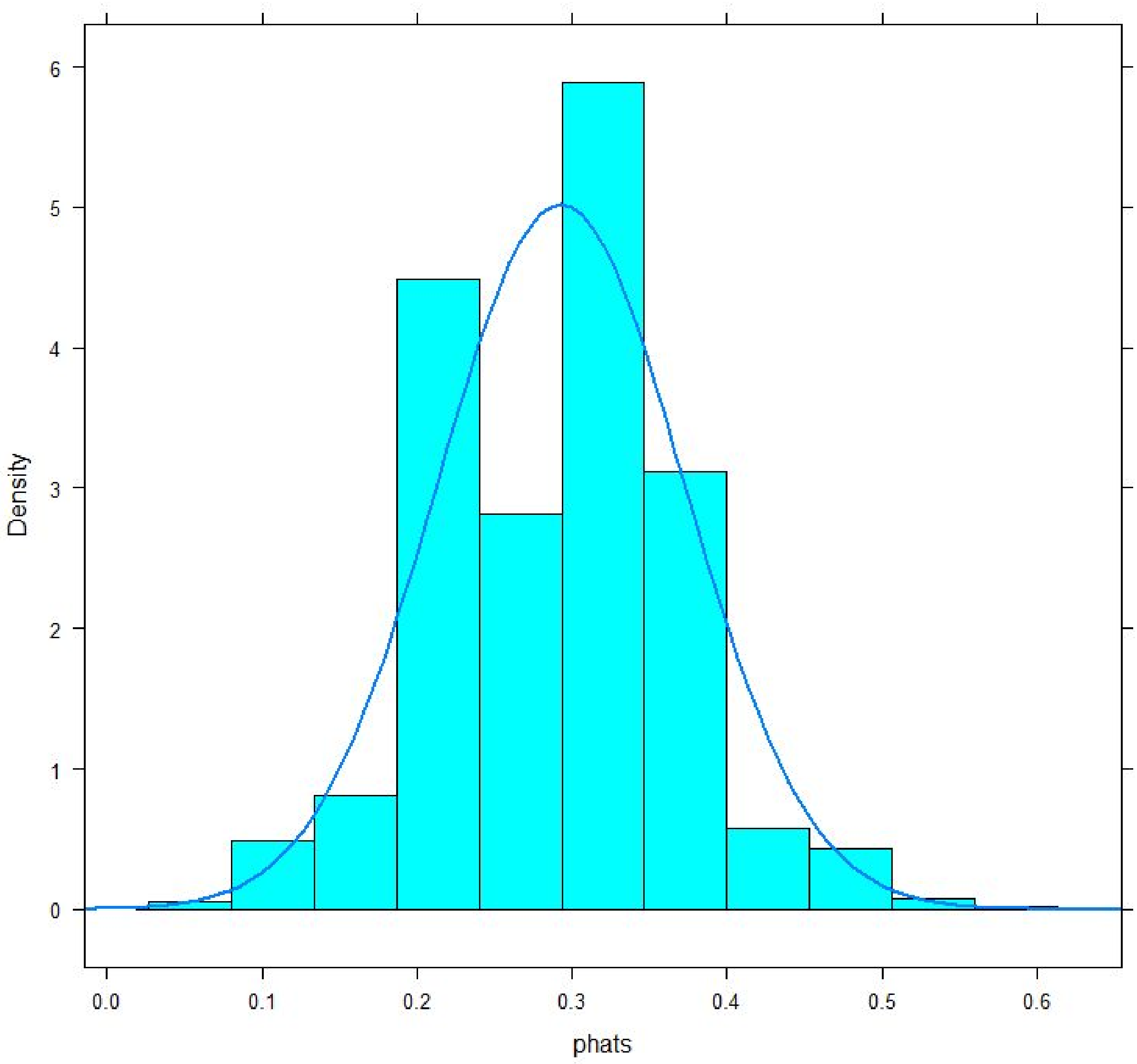
We first create the vector of sample proportions by running the following code:

```
> # We first create objects for common quantities we will use for
this exercise.
> n <-30
> # The sample size
> N <-541
> # The population size
> M <-1000 # Number of samples/repetitions
> # Create vectors to store the simulated proportions from each
repetition.
> phats <-numeric(M)
> # for sample proportions
> # Set the seed for reproducibility
> set.seed(123)
> # Always set the seed OUTSIDE the for loop.
> # Now we start the loop. Let i cycle over the numbers 1 to 1000
(i.e., iterate 1000 times).
> for(i in seq_len(M)){
+ # The i-th iteration of the for loop represents a single
repetition.
+  # Take a simple random sample of size n from the population of
size N.
+  index <-sample(N,size=n)
+  # Save the random sample in the sample_i vector.
+  sample_i <-pawnee[index,]
+  # Compute the proportion of the i-th sample of households with a
new health issue.
+  phats[i] <-mean(sample_i$New_hlth_issue == "Y")
+ }
```

We can then create a relative frequency histogram by either creating a histogram and then attaching a curve of normal to it, or using the mosaic function to create it for us.



```
> hist(phats,prob=T)
> curve(dnorm(x,mean(phats),sd(phats)),add=TRUE,col="red")
```



```
> library(mosaic)
> histogram(phats, fit="normal")
```

b)
We can find the mean and standard deviation of phat by running the following code

```
> mean(phats)
[1] 0.2928
> sd(phats)
[1] 0.07951963
```

Which gives us 0.2928 for the mean and 0.07951963 for the standard deviation respectively

c)
I believe that the simulated distribution of the sample proportions are approximately normal due to the fact that visually, the histograms seem to follow the normal curve attached to it, suggesting a normal distribution. We can also see that the distribution is roughly symmetric and that it also follows the empirical rule (which we can see with the histogram’s density). Supporting evidence includes the fact that two of the three conditions of CLT are met, since we did random and independent sampling, and the population is at least 10 times the sample. The large sample size is nearly met but falls short, so CLT is best used as supporting evidence.

d)
We can predict the mean and standard deviation of the sampling distribution to be 0.2920518 for the mean and 0.08301757 for the SD respectively by using the theory based method of CLT.

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

We can then compare them to the answers of part B, which shows that the predictions are very close to the actual proportions, suggesting normal distribution

Actual vs predicted:
[mean] 0.2928 vs 0.2920518
[standard deviation] 0.08301757 vs 0.07951963

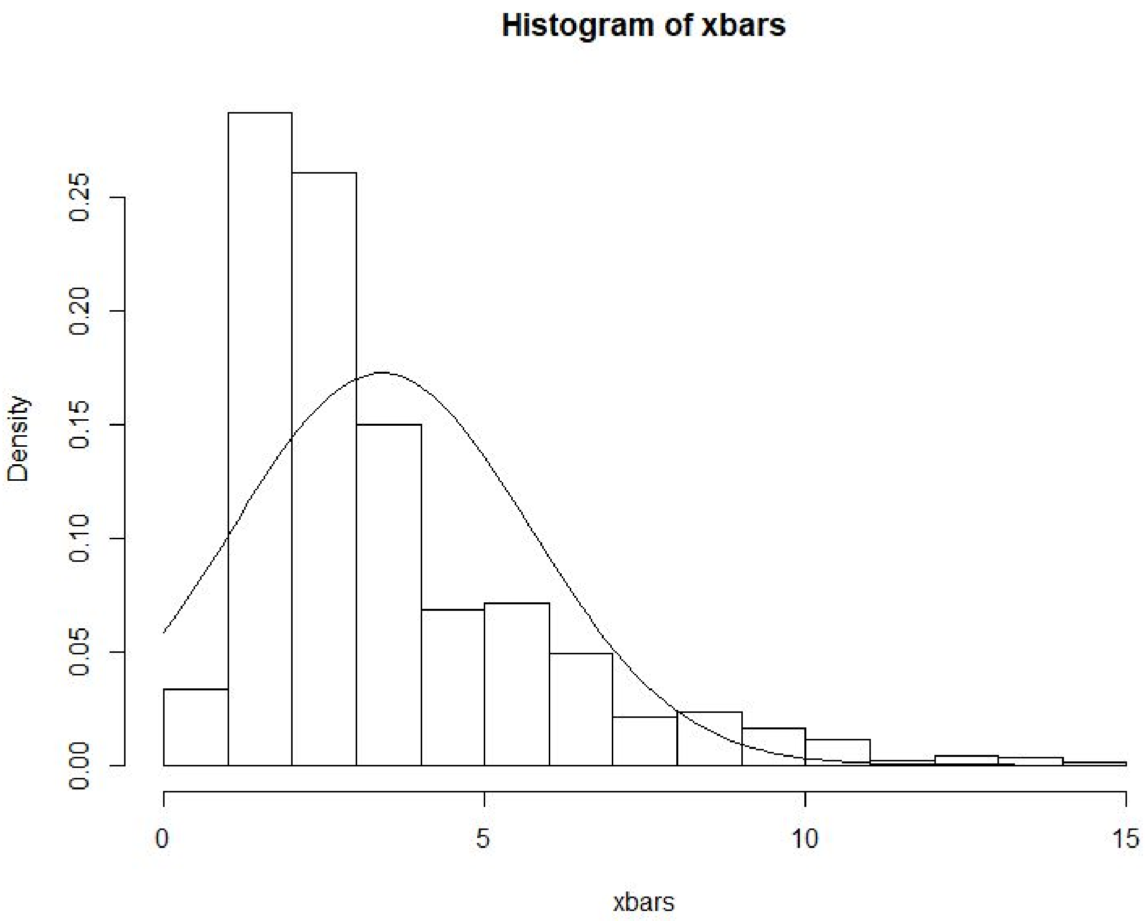
```
> (p_true = mean(pawnee$New_hlth_issue=="Y")) #theory
[1] 0.2920518
> (se_true = sqrt(p_true*(1-p_true)/n))
[1] 0.08301757

> p_true = mean(phats) #actual
> se_true = sd(phats)
```

Section 2

a)
```
> # We first create objects for common quantities we will use for this
exercise.
```

```
> n <-30
> # The sample size
> N <-541
> # The population size
> M <-1000 # Number of samples/repetitions
> # Create vectors to store the simulated proportions from each repetition.
> xbars <- c()
> # for sample proportions
> # Set the seed for reproducibility
> set.seed(123)
> # Always set the seed OUTSIDE the for loop.
> # Now we start the loop. Let i cycle over the numbers 1 to 1000 (i.e.,
iterate 1000 times).
> for(i in 1:M){
+  # The i-th iteration of the for loop represents a single repetition.
+  # Take a simple random sample of size n from the population of size N.
+  index <-sample(N,size=n)
+  # Save the random sample in the sample_i vector.
+  sample_i <- pawnee[index,]
+  # Compute the proportion of the i-th sample of households with a new health
issue.
+  xbars[i] <- mean(sample_i$Arsenic)
+ }
```

b)
We can create a histogram by using the same statements used previously in 2a, replacing phats with xbars



```
> hist(xbars,prob=T)
> curve(dnorm(x,mean(xbars),sd(xbars)),add=TRUE)
```

c)
We do not believe that the distribution is approximately normal. We can look at the created histogram and notice that it is not symmetric in data and that the distribution leans toward skewed right rather than a unimodal, symmetric distribution. We can also tell that the proportion of xbars more than 3 standard deviations above the mean is 13 times higher than what normal should be, and that the maximum xbar in the distribution is 5.19 standard deviations away, which is a value that should be seen only in extremely high trial size. CLT does not apply here since the data is clearly not normal.

Our answer is different from our answer in 2c since the data holds roughly 55% of all arsenic data points at 0, and that there are extreme outliers in the distribution. We can see that there are 3 data points over 100, thus causing such a skew and adding roughly 7.13 to the non-resistant mean. If we excluded those points, the data would be less skewed and the chances of an xbar over 7.13 would be drastically reduced to 1/500 times. This data could possibly be fixed with a larger sample size, which would make the data less affected by outliers and more likely to be normal.