STA255: Assignment 1 Solution

Shahriar Shams

March 31, 2020

Few words: The purpose of this assignment was to give you a chance to see the idea of *repeated sampling* and *central limit theorem*.

For questions 1-2, those 125 sets of numbers (each of size 3) are the all possible ways we could have observed a sample of size 3 from the given population. In real life, we only observe one of these 125 sets and make all our conclusion based on that. A different way of looking at it would be to think that those 125 values of $\bar{X}$ (calculated in 1(d)) is the "population" of $\bar{X}$. And we only observe one of it's value. Similarly in Question 2, we have the "population" of S^2 and $\hat{\sigma}^2$.

Finally, in question (3), we looked at the idea of "how large n has to be" in order for $\bar{X}$ to have a Normal disitrubtion. From graphs produced in parts (a)-(e), our conclusion is that there isn't any fix n, rather it depends on the population distribution from which we are drawing our samples. If the population disitrubtion is skewed we need a large n, but if the populaiton disitribution is symmetric a small n is sufficient.

There are many ways of answering these questions (many ways of coding). I am just showing you one of the ways...

Question 1

Suppose you have a population of size 5 [i.e. N=5]. You measure some quantity (X) and the corresponding numbers are:

$$21, 22, 23, 24, 25$$

a) Calculate the population mean (μ)

b) Calculate the population variance (σ^2) using the formula $\sigma^2 = \frac{\sum_{j=1}^{N}(X_j - \mu)^2}{N}$

c) Imagine you are taking samples (of size n = 3) from this population with replacement. Write down **every possible** way that you could have a sample of size 3 **with replacement** from this population. (hint: there will $5 5 5 = 125$ possible combinations)

d) For each of these samples of size 3, calculate the sample mean and record it (either as a new object in R or as a new column if you are using excel). Lets call this new column *X_bar*. So you should have 125 values in this column.

e) You should have noticed that the values in the *X_bar* column are repetitive. For example, 21.3333333 will show up 3 times. Construct a frequency table based on the column *X_bar*. [i.e. write down which values showed up how many times]. Now using the frequencies (also known as counts) calculate proportion of each of those repeated values. [For example: proportion of 21.3333333 will be 3/125]

f) Plot these proportions against the values and connect the points using a non-linear line. Does the shape of this plot look like any known distribution? Name the distribution.

g) Using the table of proportions or otherwise, calculate the mean of these 125 numbers (values under *X_bar*) and compare it to your answer of 1(a).

h) Using the table of proportions or otherwise, calculate the variance of these 125 numbers. Use the population variance formula (i.e. divide by 125 not 124). What is the relationship of this answer to your answer of 1(b)?

i) Which theorem did you demonstrate empirically in part f, g and h?

```r
# saving the population numbers under the name "pop"
pop=c(21:25)

# Since we will calculate the population variance a number of times,
# lets define a function.
population.var=function(y){
  mean((y-mean(y))^2)
}
```

1(a)

Population mean,

```r
mean(pop)
```

```
## [1] 23
```

1(b)

Popualation variance,

```r
population.var(pop)
```

```
## [1] 2
```

1(c)

```r
# saving all possible combination of samples of size 3 and
# saving it under "all_sets"
all_sets=expand.grid(pop,pop,pop)
```

1(d)

```r
# calculating the mean for each row of all_sets
X_bar=apply(all_sets,1,mean)
```

1(e)

```r
table(X_bar)
```

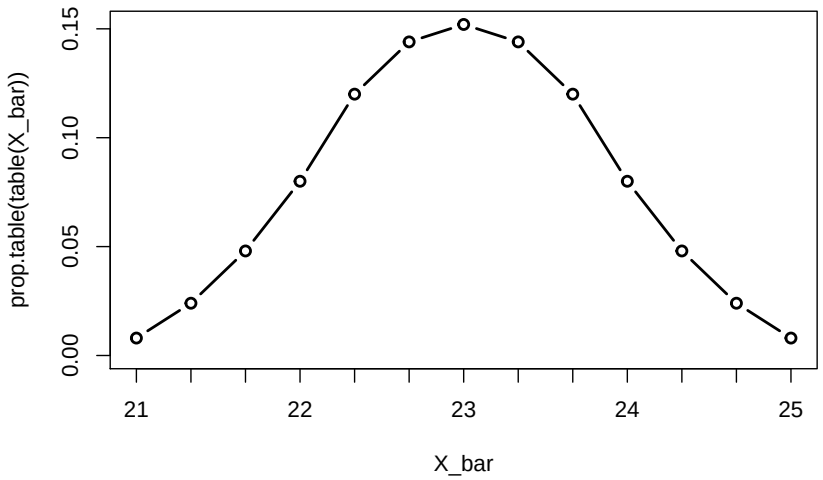
```
## X_bar
##                21 21.3333333333333 21.6666666666667                22
##                 1                 3                 6                10
## 22.3333333333333 22.6666666666667                23 23.3333333333333
##                15                18                19                18
## 23.6666666666667                24 24.3333333333333 24.6666666666667
##                15                10                 6                 3
##                25
##                 1
```

```
prop.table(table(X_bar))
```

```
## X_bar
##                21 21.3333333333333 21.6666666666667                22
##             0.008             0.024             0.048             0.080
## 22.3333333333333 22.6666666666667                23 23.3333333333333
##             0.120             0.144             0.152             0.144
## 23.6666666666667                24 24.3333333333333 24.6666666666667
##             0.120             0.080             0.048             0.024
##                25
##             0.008
```

1(f)

```
#this is the sampling distribution of X_bar
plot(prop.table(table(X_bar)),type="b")
```



1(g)

```
#this is the mean of X_bar
mean(X_bar)
```

```
## [1] 23
```

This is the same value we got in 1(a). This verifies the formula $E[\bar{X}] = \mu$

1(h)

```
#this is the variance of X_bar
population.var(X_bar)
```

```
## [1] 0.6666667
```

If we divide the population variance (calculated in part 1(b)) by the sample size which is 3, we will get the variance of $\bar{X}$. This verifies $var[\bar{X}] = \frac{\sigma^2}{n}$

1(i)

The plot in 1(f) looks roughly Normal, part 1(g) shows $E[\bar{X}] = \mu$ and part 1(h) shows $var[\bar{X}] = \frac{\sigma^2}{n}$. Putting all together we have $\bar{X} \xrightarrow{D} N(\mu, \frac{\sigma^2}{n})$. This is a (rather simple) demonstration of Central Limit Theorem (CLT) using empirical data.

Question 2

This question continues from question 1(c). For each of these sample of size 3, calculate the sample variance using the following two formulas

$$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$$

Assume the population variance, $\sigma^2 = 2$. (you should get 125 different values of S^2 and and 125 different values of $\hat{\sigma}^2$)

a. By calculating (numerically using the 125 different values) $Bias[S^2]$ and $Bias[\hat{\sigma}^2]$ check the unbiasedness of these two estimators.

b. By calculating all three components separately check the following identity

$$MSE[\hat{\sigma}^2] = var[\hat{\sigma}^2] + (Bias[\hat{\sigma}^2])^2$$

2(a)

```
# calculating S^2 for all 125 set of samples
S_sq=apply(all_sets,1,var)

# calculating sigma_hat^2 for all 125 set of samples,
# sigma_hat^2 is nothing but the population variance formula
sigma_hat_sq=apply(all_sets,1,FUN=population.var)

#here true paraemter value is 2.


#Bias of S^2
mean(S_sq) - 2
```

```
## [1] 0
```

```
#Bias of sigma_hat^2
mean(sigma_hat_sq) - 2
```

```
## [1] -0.6666667
```

S^2 is an unbiased estimator of σ^2 and $\hat{\sigma}^2$ is a biased estimator of σ^2.

2(b)

```r
# MSE of sigma_hat^2
mean((sigma_hat_sq-2)^2)
```

```
## [1] 1.451852
```

```r
# var of sigma_hat^2
population.var(sigma_hat_sq)
```

```
## [1] 1.007407
```

```r
# square of (bias of sigma_hat^2)
(mean(sigma_hat_sq) - 2)^2
```

```
## [1] 0.4444444
```

```r
# var of sigma_hat^2 + square of (bias of sigma_hat^2)
population.var(sigma_hat_sq)+(mean(sigma_hat_sq) - 2)^2
```

```
## [1] 1.451852
```

This shows that the identity is true.

Question 3

In week 3, we demonstrated an R code that replicates the sample distribution of $\bar{X}$. Here is the code that was used in the lecture.

```
1  sample_4m_normal=function(x){
2    s=rnorm(30, mean=10,sd=2)
3    return(mean(s))
4  }
5
6  X_bar=replicate(100000, sample_4m_normal())
7
8  plot(density(X_bar))
```

Simply change the distribution and number of samples on line 2 of this code to do this question.

Produce the density of $\bar{X} = \frac{X_1+X_2+...+X_n}{n}$

a) when $n = 2$, $X \sim Unif[0, 1]$

b) when $n = 5$, $X \sim Unif[0, 1]$

c) when $n = 5$, $X \sim \chi^2_{df=2}$

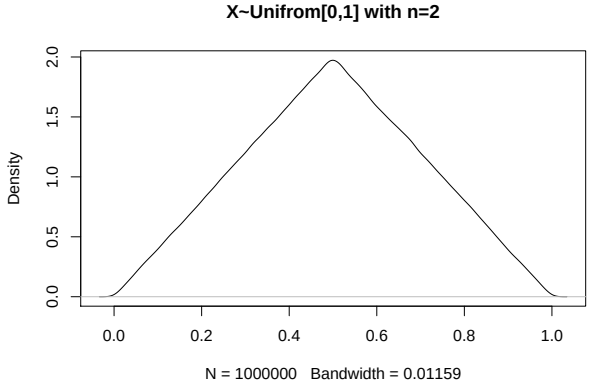
d) when $n = 30$, $X \sim \chi^2_{df=2}$

e) when $n = 5$, $X \sim \chi^2_{df=50}$

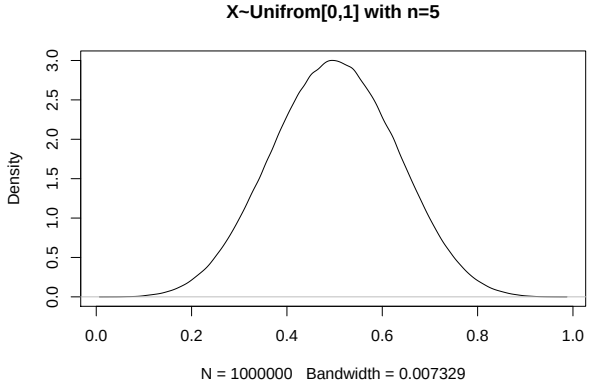
f) CLT says for large n, $\bar{X}$ converges(in distribution) to a Normal distribution. By comparing your graphs from parts (a) to (e), can you comment on how large n has to be in order for $\bar{X}$ to converge to a Normal distribution.

g) A quick way to plot any distribution in R is to draw a large sample from a distribution and plot its density. For example "plot(density(rnorm(100000,mean=10,sd=2)))" will produce a N(10,4) curve. Plot three separate density curves for $Unif[0, 1]$, $\chi^2_{df=2}$ and $\chi^2_{df=50}$. Looking at the skewness of these three curves, what comments can you make on the question asked in part(f)?

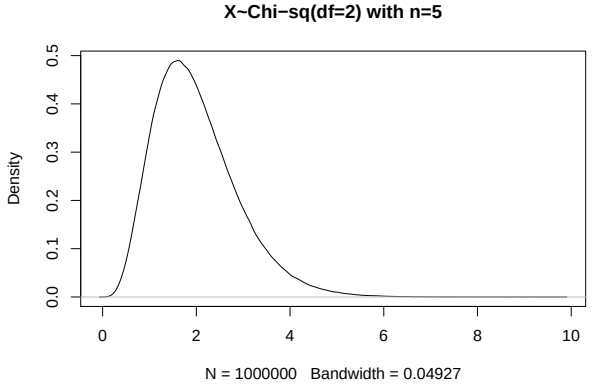
3(a)


X~Unifrom[0,1] with n=2
Density
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
N = 1000000 Bandwidth = 0.01159

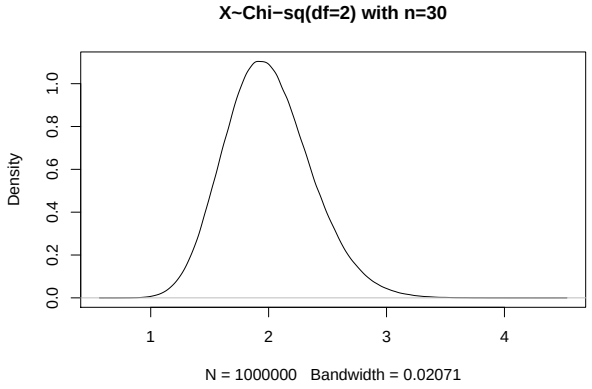
3(b)


X~Unifrom[0,1] with n=5
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0
N = 1000000 Bandwidth = 0.007329

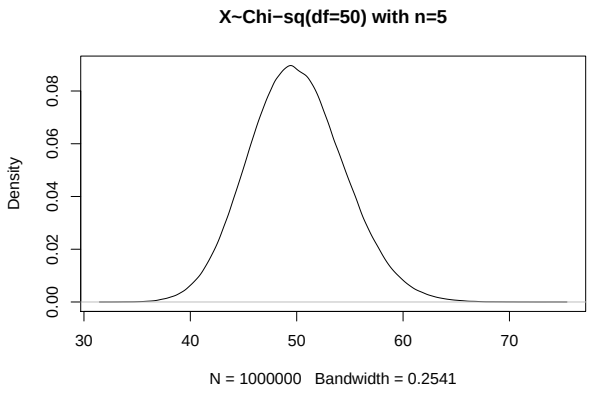
3(c)


X~Chi−sq(df=2) with n=5
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
N = 1000000 Bandwidth = 0.04927

3(d)



3(e)



3(f)

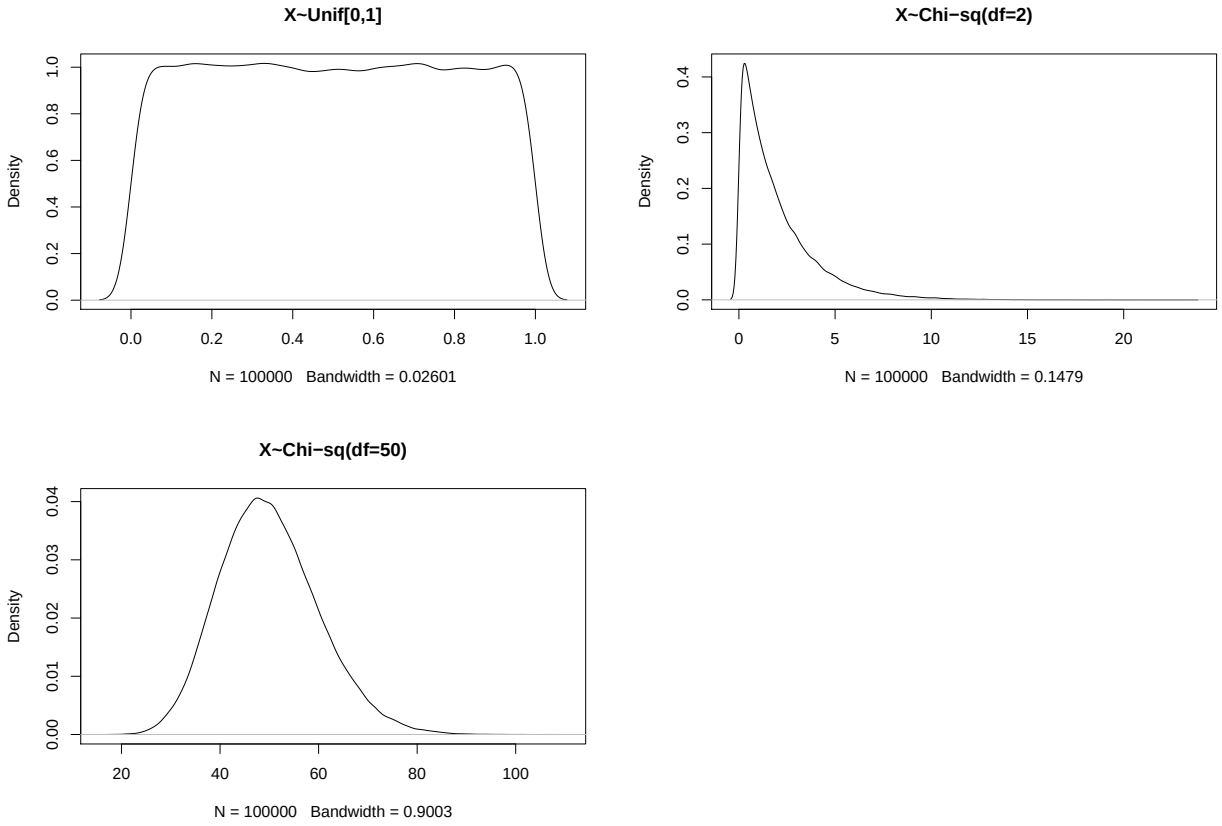
Based on the densities of $\bar{X}$ in part (a)-(e) we can see, how large n has to be, for the denisty of $\bar{X}$ to converge in distribution to a Normal distribution, depends on the actual population that we are drawing the samples from.

Comparing (a) and (b), when we are sampling from $Unif[0,1]$ looks like $n = 5$ is enough.

Comparing (c) and (d), $n = 5$ is not enough when we are sampling from $\chi^2_{(df=2)}$. Looks like $n = 30$ is enough (though one might say you need more than 30).

Compring (d) and (e), though we need at least $n = 30$ when sampling from $\chi^2_{(df=2)}$, $n = 5$ is enough when sampling from $\chi^2_{(df=50)}$.

3(g)



If we look at the actual population denisties(density of X) of $Unif[0,1]$, $\chi^2_{(df=2)}$ and $\chi^2_{(df=50)}$ we can see

- $Unif[0,1]$ is completely symmetric and $\chi^2_{(df=50)}$ almost looks like a symmetric distribution (the tail is a bit longer on the right).
- $\chi^2_{(df=2)}$ is clearly a highly skewed distribution.

Combining all of these, how large n should in order to apply CLT depends on the skewness of the popultion distribution from which samples are drawn. Cases when population disitrbution is symmetric we don't need a large n, but if the population distribution is skewed we need a large value of n.

Codes for Question 3

```
# 3(a)
sample_fn=function(x){
  s=runif(2,0,1 )
  return(mean(s))
}
X_bar=replicate(1000000, sample_fn())
plot(density(X_bar),main="X~Unifrom[0,1] with n=2")

# 3(b)
sample_fn=function(x){
  s=runif(5,0,1 )
  return(mean(s))
}
X_bar=replicate(1000000, sample_fn())
plot(density(X_bar),main="X~Unifrom[0,1] with n=5")

# 3(c)
sample_fn=function(x){
  s=rchisq(5,df=2)
  return(mean(s))
}
X_bar=replicate(1000000, sample_fn())
plot(density(X_bar),main="X~Chi-sq(df=2) with n=5")

# 3(d)
sample_fn=function(x){
  s=rchisq(30,df=2)
  return(mean(s))
}
X_bar=replicate(1000000, sample_fn())
plot(density(X_bar),main="X~Chi-sq(df=2) with n=30")

# 3(e)
sample_fn=function(x){
  s=rchisq(5,df=50)
  return(mean(s))
}
X_bar=replicate(1000000, sample_fn())
plot(density(X_bar),main="X~Chi-sq(df=50) with n=5")
```

```r
# 3(g)
plot(density(runif(100000,0,1)),main="X~Unif[0,1]")
plot(density(rchisq(100000,df=2)),main="X~Chi-sq(df=2)")
plot(density(rchisq(100000,df=50)),main="X~Chi-sq(df=50)")
```