Assignment Solution

Shahriar Shams

April 7, 2019

There are many ways of getting these numbers. This is just one of them…

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 $\sigma^2 = \frac{\sum_{j=1}^{N}(X_j-\mu)^2}{N}$

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

#1(a)
mean(pop)
```

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

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


#1(b) population variance
population.var(pop)
```

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

Question 2. Imagine you are taking samples (of size n = 3) from this population with replacement. Recall: “sampling WITH replacement” ensures independence.

a. 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)

b. 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 in excel). Lets call this new column “X_bar”. So you should have 125 values in this column.

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

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

Question 3. You should have noticed that the values in the “X_bar” column are repetitive. For example, 21.3333333 will show up 3 times.

a. 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]

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

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

d. 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)?

e. Which theorem did you demonstrate empirically in part b, c and d?

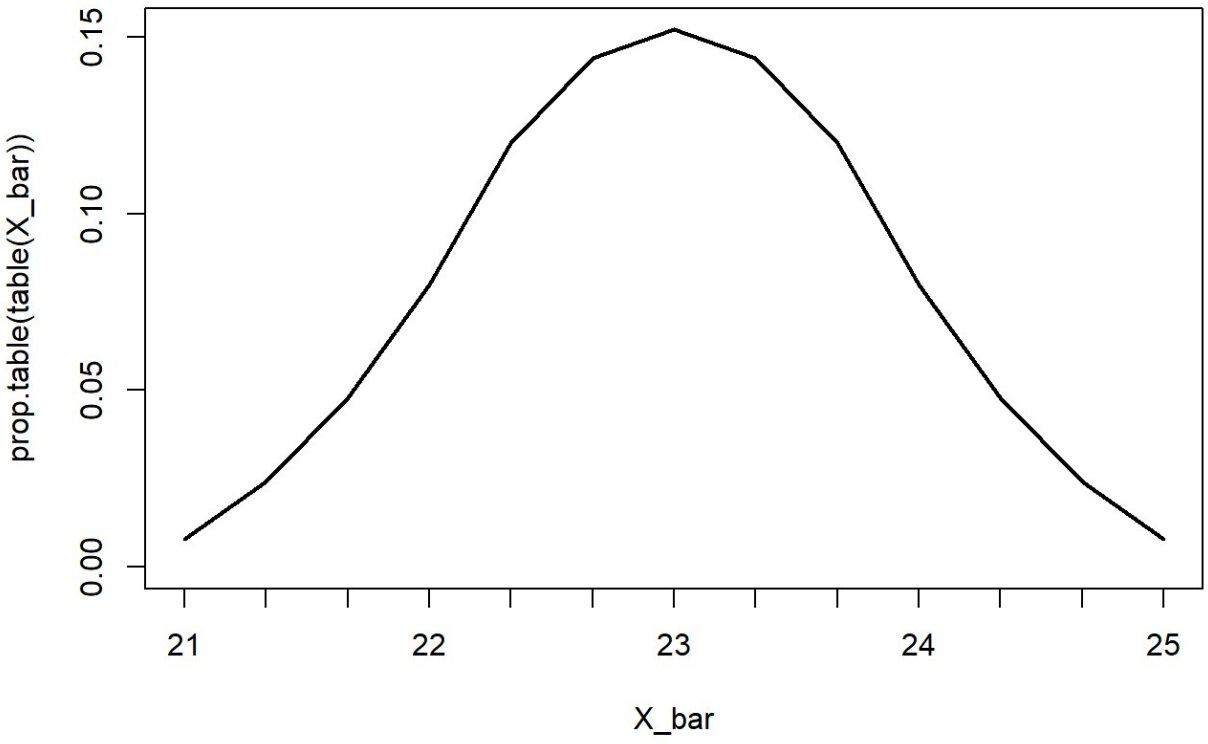
```
#3(a)
table(X_bar)
```

```
## 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
```

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




```
#3(c)
#this is the mean of the sampling distribution of X_bar
mean(X_bar)
```

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

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

```
#3(d)
#this is the variance of the sampling distribution 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 $var[\bar{X}] = \frac{\sigma^2}{n}$

3(e) The plot in 3(b) looks roughly Normal, part 3(c) shows $E[\bar{X}] = \mu$ and part 3(d) shows $var[\bar{X}] = \frac{\sigma^2}{n}$. Putting all together we have $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. This demonstrates Central Limit Theorem (CLT) empirically.

Question 4. 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, $\sigma^2 = 2$.

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

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

```
# 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 po
pulation variance formula
sigma_hat_sq=apply(all_sets,1,FUN=population.var)

#here true paraemter value is 2.

#4(a)
#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 .

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

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

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

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

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

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

```
#Numerically MSE is equal to the sum of the var and bias^2
```

Question 5. Even though we need sample size n to be large to apply central limit theorem, but let's apply it anyway. Suppose you know that the population variance, $\sigma^2 = 2$.

a. For each of these 125 cases, calculate a 95% confidence interval and finally calculate the proportion of the intervals that includes $\mu = 23$.

b. Suppose someone observes only one of these 125 combinations (23,24,25). If that person is testing the null hypothesis $H_0 : \mu = 23$, based on this observed sample calculate the p-value that the person will get using central limit theorem.

c. Calculate the p-value numerically using the 125 $\bar{X}$ values that you calculated in part 2(b) (do not use CLT here).

```
#5(a)
#Lower and Upper bounds of each of these 125 confidence intervals
lower_bound=X_bar - qnorm(0.975)*sqrt(2/3)
upper_bound=X_bar + qnorm(0.975)*sqrt(2/3)

#true mean is 23. The proportion of intervals containing the true mean
mean( (lower_bound<= 23)&(23<=upper_bound) )
```

```
## [1] 0.936
```

```
#5(b)
#The observed sample mean is 24.
#p-value = 2*P[X_bar >=24|mu=23]
#using CLT, p-value = 2*P[Z >= (24-23)/sqrt(2/3)]
2*(1-pnorm(24,mean=23,sd=sqrt(2/3)))
```

```
## [1] 0.2206714
```

```
#5(c)
#without CLT, we look at the 125 possible values of X_bar literally calculate the P[X_
bar >=24]
# by checking how many of the X_bar's are equal or abover 24.
# we multiply it by 2, since just like 24 and above, 22 and below also defines "as or
more extreme"
#keeping the hypothetical mean (in this case we know it's true since we have the popul
ation) 23 in the middle.
# I did it in the following way.
mean(abs(X_bar-23)>=1)
```

```
## [1] 0.32
```

Comments: The purpose of this assignment was to give you a chance to see the idea of repeated sampling and an application of central limit theorem and few other identities. Those 125 sets of numbers 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 the 125 values in the X_bar column (calculated in 1(b)) is the population of $\bar{X}$. And we only observe on of it's value.

We calculated 95% CI for μ and in part 5(a) checked the true coverage probability and found it to be 93.6%. For a bigger sample size we expect this coverage probability to be more closer to 95%.

In calculating p-value, the one we calculated in par 5(c) is the true p-value. But using CLT (in 5(b)) we get an approximate value. Eventhough in this case we find these two values to be different, but for a bigger sample size the difference will get smaller.

Some of you asked me during lectures, how large n has to be for the sampling distribution of $\bar{X}$ to approach Normal distribution. Here is a little exercise for you.

a. Generate 2 random numbers from a Unifrom[0,1] distribution and calculate their mean.
b. Repeat the task in part(a) ten thousand times and save these means.
c. Plot a histogram of these means and add a normal denisity curve on this histogram.

Do you think the histogram and the normal density matches perfectly? Repeat the task by generating 3 random number now. Repeat it for n=4, 5 etc… Check at which point it's really a Normal.

Now do the same task but instead of Uniform, use a Chi-sq(df=2) distribution. Check at which n-value you get a Normal curve. Repeat the task for a Chi-sq(df=30).

Hoping that you will do these tasks, here is what to expect: at which n, distribution of $\bar{X}$ converges to Normal depends on the skewness of the actual distribution.