STATISTICAL METHODS FOR DATA SCIENCE Mini-Project 1

Duo Group #23

Members: Hima Sri Tipirineni

Nithin Pingili

Contribution of each team member:

Hima Sri and Nithin worked together to complete both the questions. Collaborated to learn R and then worked on calculating the means and probability of the given questions using Monte Carlo simulations. Both worked together to answer the questions and report all the findings. Hima Sri wrote R code and annotated the code and Nithin worked to check the accuracy of the R code and annotations. Both worked efficiently to complete all sections of the project.

Question 1:

1. (10 points) Consider Exercise 4.11 from the textbook. In this exercise, let XA be the lifetime of block A, XB be the lifetime of block B, and T be the lifetime of the satellite. The lifetimes are in years. It is given that XA and XB follow independent exponential distributions with mean 10 years. One can follow the solution of Exercise 4.6 to show that the probability density function of T is

$$f_T(t) = \begin{cases} 0.2\exp(-0.1t) - 0.2\exp(-0.2t), & 0 \le t < \infty, \\ 0, & \text{otherwise}, \end{cases}$$

and E(T) = 15 years.

(a) Use the above density function to analytically compute the probability that the lifetime of the satellite exceeds 15 years.

Solution :

We need to calculate the probability of the lifetime of the satellite(T) exceeds 15 years. i.e P(T > 15). This can be also inferred as below :

$$P(T>15) = 1 - P(T<=15) = 1 - F(15)$$

This can be calculated with the CDF function of f_T (t).

$$F(15) = \int_0^{15} f(t)dt$$

$$= \int_0^{15} (0.2\exp(-0.1t) - 0.2\exp(-0.2t))\, dt$$

$= \int^{15}_0 0.2\exp(-0.1t)dt - \int^{15}_0 0.2\exp(-0.2t)dt$

$= -0.2/0.1 \exp(-0.1t) |_0^{15} + 0.2/0.2 \exp(-0.2t) |_0^{15}$

$= -2\exp(-0.1t) |_0^{15} + \exp(-0.2t) |_0^{15}$

$= -2\exp(-1.5) + 2 + \exp(-3) - 1$

$= -2\exp(-1.5) + \exp(-3) + 1$

$= - 0.44626 + 0.0497 + 1$

$= 0.60344$

Now,

$P(T > 15) = 1 - 0.60344 =$ **0.39656**

(b) Use the following steps to take a Monte Carlo approach to compute E(T) and P(T > 15).

Solution :

(i)

```
> ######################################
> ##Simulating 1 draw of XA , XB and T ##
> ######################################
>
> lambda = 0.1
> mc_XA = -1/lambda * (log(runif(1)))  # exponential distrution using Uniform distribution with lambda
> mc_XA
[1] 0.02101525
> mc_XB = -1/lambda * (log(runif(1)))  # exponential distrution using Uniform distribution with lambda
> mc_XB
[1] 11.61849

> mc_T = max(mc_XA , mc_XB)        # Since distributin of T is maximum of XA and XB
> mc_T
[1] 11.61849
```

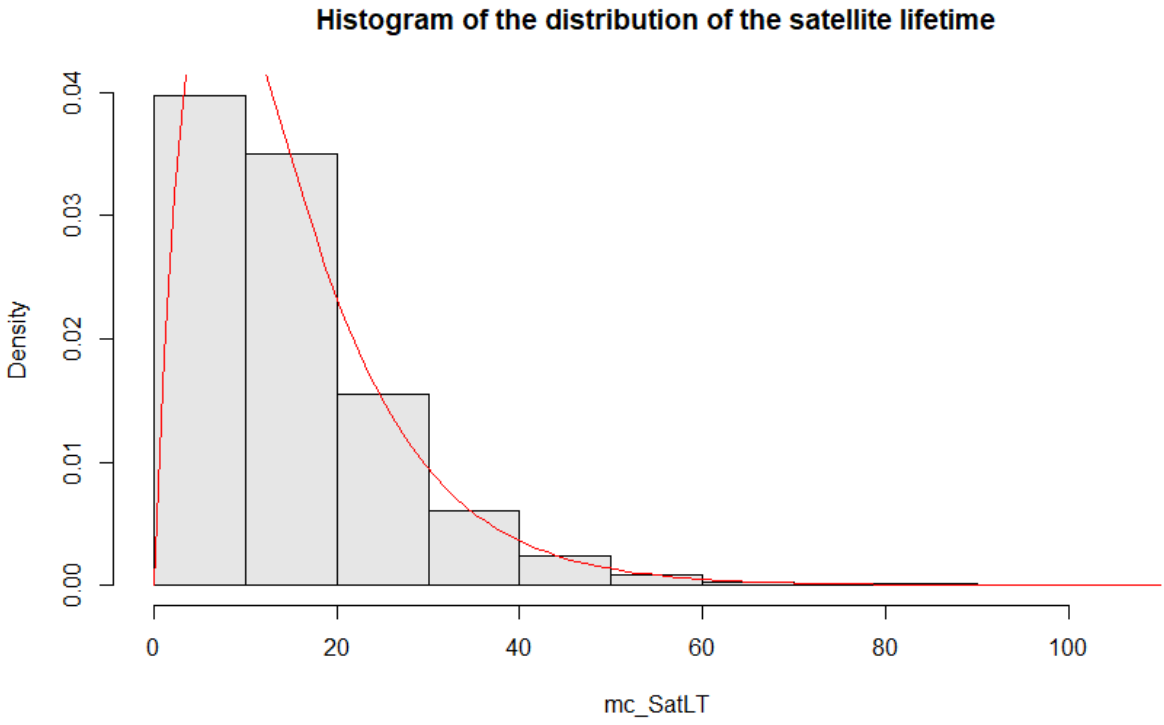
(ii)

```
>
> ##Replicating the above step 10,000 times in a single step is a below###
> ######################################################################
>
> mc_SatLT = replicate(10000, max(-10 *log(runif(1)), -10*log(runif(1))))  ##since 1/lambda = 1/0.1 = 10,sampleSize = 10,000
>
```

values	
lambda	0.1
mc_SatLT	num [1:10000] 6.92 22.05 22.93 18.75 11.84 ...

(iii) Histogram :

```
>
> ############################################################
> ########## Histogram of the draws of T using Hist function ####
> ############################################################
>
> hist(mc_SatLT,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
>
> densityFunc = function(t) {return (0.2*exp(-0.1*t) - 0.2*exp(-0.2*t))} # Declaring density function
>
> curve(densityFunc, add=TRUE, col="red")  #superimposing the density function curve on the histogram
>
>
>
```



Observation: The histogram has the probability of T on y-axis and the value of T on x-axis. The histogram is right skewed because the peak of the histogram is to the left. The histogram correlates with the density curve as both are right skewed.

(iv)

```
>
> ##########################################################################
> ######## To calculate E(T) using the saved draws ###################
> ##########################################################################
>
> EstimatedTime = mean(mc_SatLT)
> EstimatedTime
[1] 15.06925
>
>
```

The given value of mean is 15 years and the from the saved draws E(T) is 15.06925 which is pretty close to the given value.

(v)

Probability that the satellite lasts more than 15 years.

```
> ######################################################################
> ########### To calculate the probability that the satellite lasts more than 15 years ####
> ######################################################################
>
> Prob = mean(mc_SatLT > 15)    ## From Monte Carlo estimator of probablity i.e sum(draws > 15)/samplesize
> Prob
[1] 0.3942
>
>
```

The calculated value of the probability is P(T > 15) = **0.39656** and the value obtained from the simulations is **0.3942** which is approximately equal to the calculated value.

(vi)

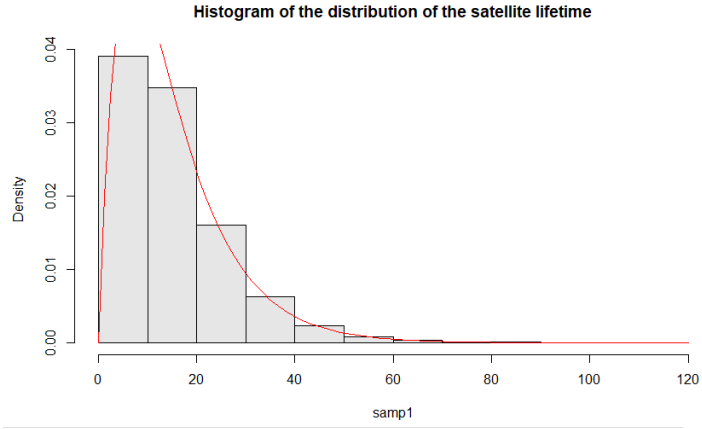
Repeating the process 4 times and calculating the mean and probability.

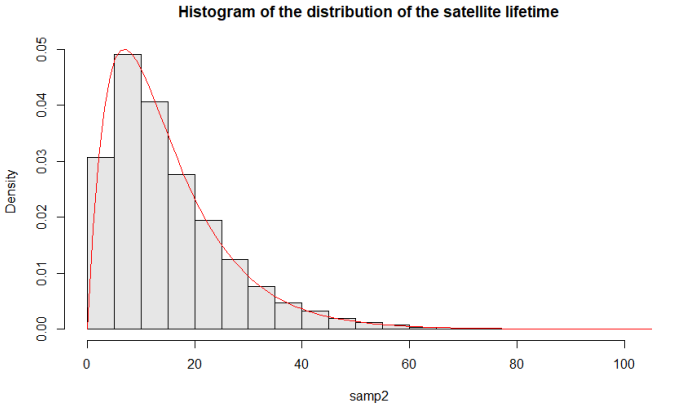
```
>
>
> ################################################################
> ####### Repeating the above steps 4 times to calculate mean and probability ######
> ################################################################
>
> samp1 = replicate(10000, max(-10 *log(runif(1)), -10*log(runif(1))))
> mean1 = mean(samp1)
> prob1 = mean(samp1 > 15)
> mean1
[1] 15.21964
> prob1
[1] 0.4057
>
>
> ##second test ##
> samp2 = replicate(10000, max(-10 *log(runif(1)), -10*log(runif(1))))
> mean2 = mean(samp2)
> prob2 = mean(samp2 > 15)
> mean2
[1] 15.11617
> prob2
[1] 0.3981
>
>
> ## third test ##
> samp3 = replicate(10000, max(-10 *log(runif(1)), -10*log(runif(1))))
> mean3 = mean(samp3)
> prob3 = mean(samp3 > 15)
> mean3
[1] 14.79931
> prob3
[1] 0.3956
>
>
> ## fourth test ##
> samp4 = replicate(10000, max(-10 *log(runif(1)), -10*log(runif(1))))
> mean4 = mean(samp4)
> prob4 = mean(samp4 > 15)
> mean4
[1] 15.03623
> prob4
[1] 0.4012
>
>
```

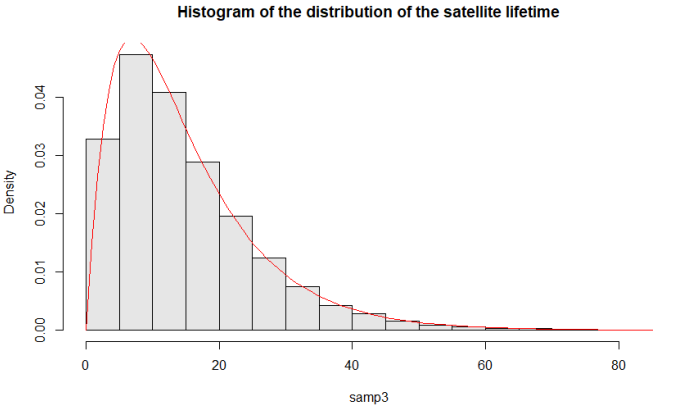
Histograms of the test conducted 4 times.

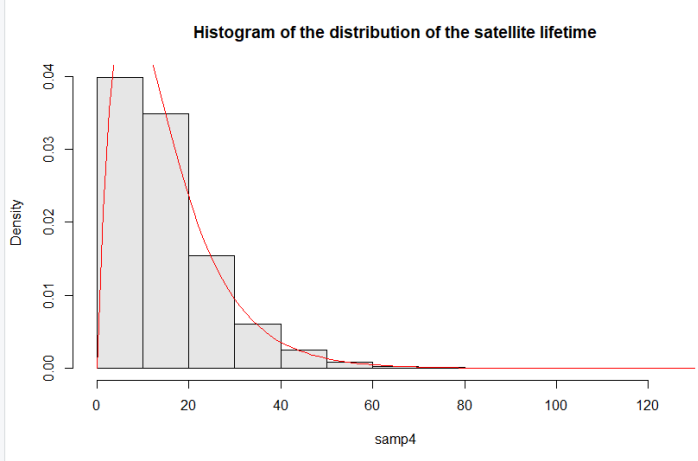
```
> hist(samp1,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
>
> hist(samp2,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
>
> hist(samp3,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")

> hist(samp4,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
>
```









S.No	E(T)	P(T > 15)
1	15.06925	0.3942
1	15.21964	0.4057
2	15.11617	0.3981
3	14.79931	0.3956
4	15.03623	0.4012

From the above test calculations, it is seen that for the sample size = 10,000, the mean[E(T)] and the probability are similar to the calculated values in 1a i.e 15 and 0.396 with very little variation. This shows that as n(sample size) approaches large values the mean value is almost normalized and is approaching a constant value. This shows the Central limit theorem and the law of large numbers is true.

(*c) Repeat part (vi) five times using 1,000 and 100,000 Monte Carlo replications instead of 10,000. Make a table of results. Comment on what you see and provide an explanation*

Solution :

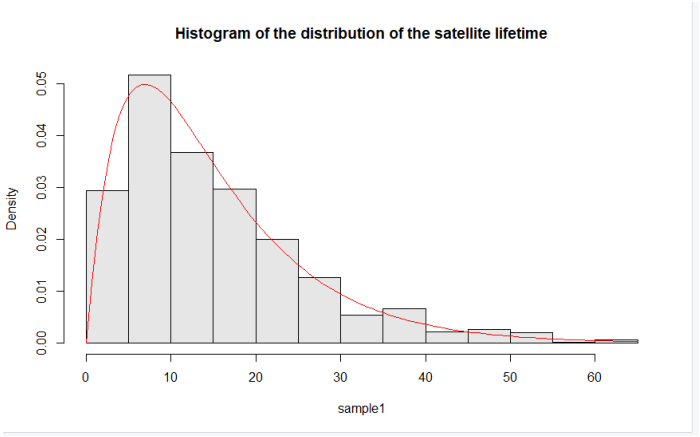
(i) 1000 times

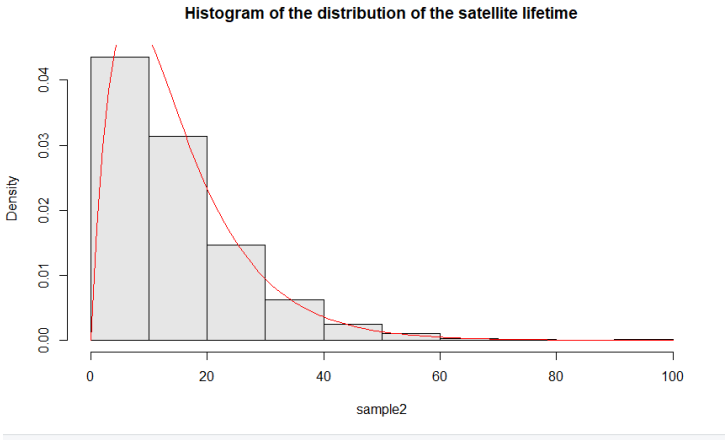
```
replicate(1000,max(-10*log(runif(1)),-10*log(runif(1))))
```

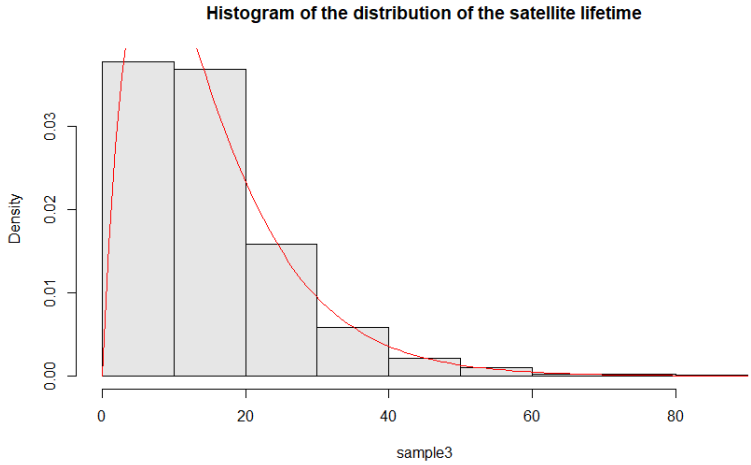
```
>
>
> ########################################################################
> ############## Calculating mean and probability with sample size = 1000 ###
> ########################################################################
>
> #1
> sample1 = replicate(1000,max(-10*log(runif(1)),-10*log(runif(1))))
> m1 = mean(sample1)
> p1 = mean(sample1 > 15)
> m1
[1] 15.16828
> p1
[1] 0.4
>
> #2
> sample2 = replicate(1000,max(-10*log(runif(1)),-10*log(runif(1))))
> m2 = mean(sample2)
> p2 = mean(sample2 > 15)
> m2
[1] 14.71538
> p2
[1] 0.382
>
>
> #3
> sample3 = replicate(1000,max(-10*log(runif(1)),-10*log(runif(1))))
> m3 = mean(sample3)
> p3 = mean(sample3 > 15)
> m3
[1] 15.08626
> p3
[1] 0.409
>
>
> #4
> sample4 = replicate(1000,max(-10*log(runif(1)),-10*log(runif(1))))
> m4 = mean(sample4)
> p4 = mean(sample4 > 15)
> m4
[1] 15.52209
> p4
[1] 0.415
>
>
> #5
> sample5 = replicate(1000,max(-10*log(runif(1)),-10*log(runif(1))))
> m5 = mean(sample5)
> p5 = mean(sample5 > 15)
> m5
[1] 15.2225
> p5
[1] 0.392
>
>
```

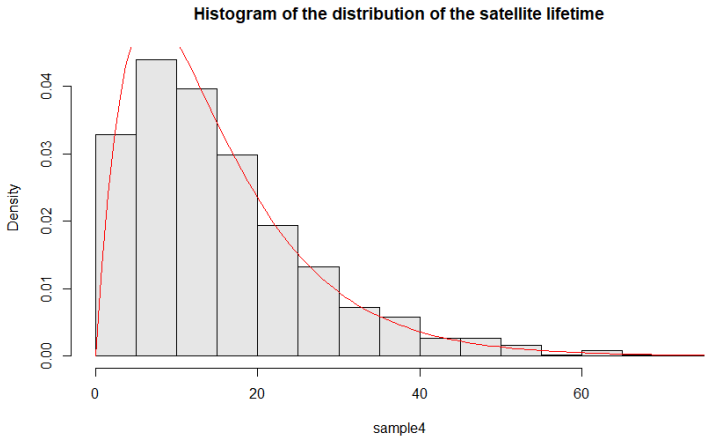
Histograms :

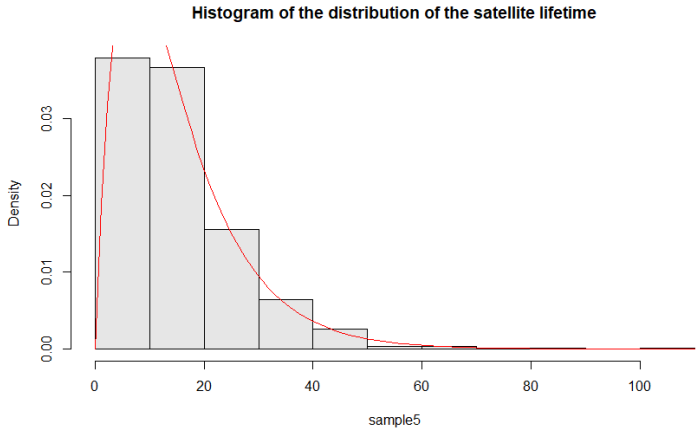
```
>
> hist(sample1,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
>
> hist(sample2,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
>
> hist(sample3,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
>
> hist(sample4,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
>
> hist(sample5,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
```


Histogram of the distribution of the satellite lifetime
Density
0.00
0.01
0.02
0.03
0.04
0.05
0
10
20
30
40
50
60
sample1


Histogram of the distribution of the satellite lifetime
Density
0.00
0.01
0.02
0.03
0.04
0
20
40
60
80
100
sample2


Histogram of the distribution of the satellite lifetime
Density
0.00
0.01
0.02
0.03
0
20
40
60
80
sample3





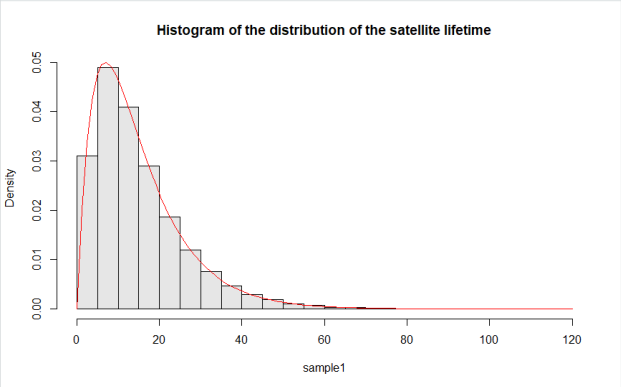
(ii) 100,000

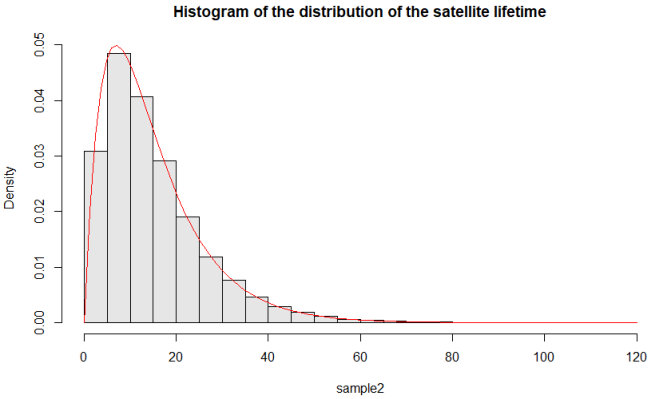
```
replicate(100000,max(-10*log(runif(1)),-10*log(runif(1))))
```

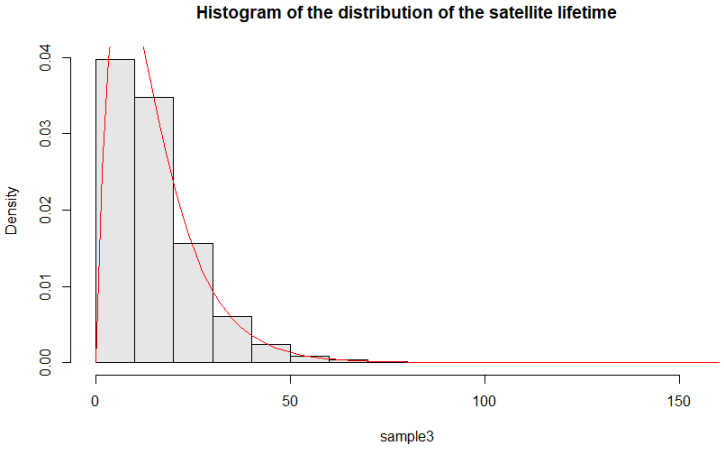
```
>
> ###############################################################################
> ############# Calculating mean and probability with sample size = 100000 ###
> ###############################################################################
>
> #1
> sample1 = replicate(100000,max(-10*log(runif(1)),-10*log(runif(1))))
> m1 = mean(sample1)
> p1 = mean(sample1 > 15)
> m1
[1] 15.00806
> p1
[1] 0.39592
>
>
> #2
> sample2 = replicate(100000,max(-10*log(runif(1)),-10*log(runif(1))))
> m2 = mean(sample2)
> p2 = mean(sample2 > 15)
> m2
[1] 15.04716
> p2
[1] 0.39919
>
>
> #3
> sample3 = replicate(100000,max(-10*log(runif(1)),-10*log(runif(1))))
> m3 = mean(sample3)
> p3 = mean(sample3 > 15)
> m3
[1] 15.06432
> p3
[1] 0.3992
>
>
> #4
> sample4 = replicate(100000,max(-10*log(runif(1)),-10*log(runif(1))))
> m4 = mean(sample4)
> p4 = mean(sample4 > 15)
> m4
[1] 14.947
> p4
[1] 0.39491
>
>
> #5
> sample5 = replicate(100000,max(-10*log(runif(1)),-10*log(runif(1))))
> m5 = mean(sample5)
> p5 = mean(sample5 > 15)
> m5
[1] 15.08632
> p5
[1] 0.39974
>
>
```

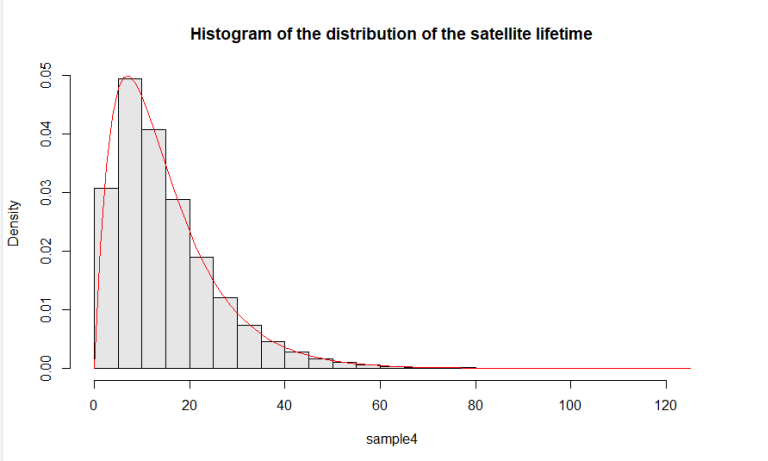
Histograms :

```
> hist(sample1,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
>
> hist(sample2,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
>
> hist(sample3,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
>
> hist(sample4,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
>
> hist(sample5,probability=TRUE, col=grey(0.9), main="Histogram of the distribution of the satellite lifetime")
> curve(densityFunc, add=TRUE, col = "red")
```


Histogram of the distribution of the satellite lifetime
Density
0.00
0.01
0.02
0.03
0.04
0.05
0
20
40
60
80
100
120
sample1


Histogram of the distribution of the satellite lifetime
Density
0.00
0.01
0.02
0.03
0.04
0.05
0
20
40
60
80
100
120
sample2


Histogram of the distribution of the satellite lifetime
Density
0.00
0.01
0.02
0.03
0.04
0
50
100
150
sample3



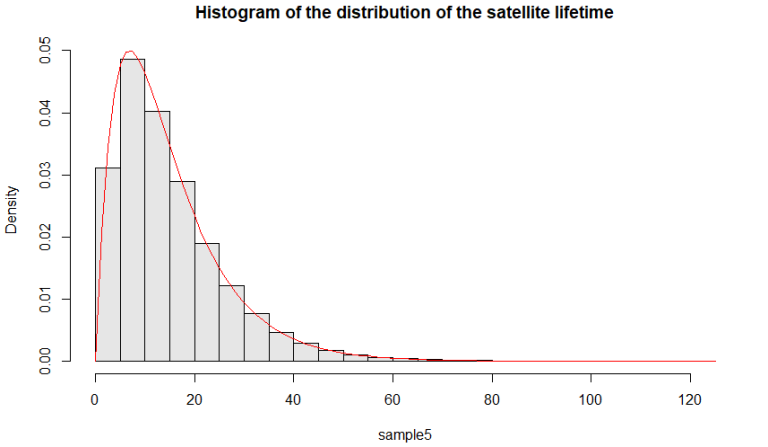


TABLE FOR SAMPLE SIZE = 1000

S.No	MEAN	P(T > 15)
1	15.16828	0.4
2	14.71538	0.382
3	15.08626	0.409
4	15.52209	0.415
5	15.2225	0.392

TABLE FOR SAMPLE SIZE = 10000

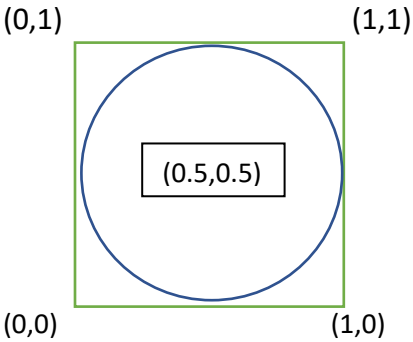
S.No	Mean	P(T>15)
1	15.00806	0.39592
2	15.04716	0.39919
3	15.06432	0.3992
4	14.947	0.39491
5	15.08632	0.39974

Observation: As the sample size increases, the variation of the mean to the calculated numerical value of 15 means has decreased. They have become normally distributed and are approaching a constant value which corresponds with the Central Limit Theorem and the Law of large numbers. Similar behavior is observed for the probability as well.

2. (10 points) Use a Monte Carlo approach estimate the value of pi based on 10,000 replications. [Ignorable hint: First, get a relation between _ and the probability that a randomly selected point in a unit square with coordinates | (0; 0), (0; 1), (1; 0), and (1; 1) | falls in a circle with center (0:5; 0:5) inscribed in the square. Then, estimate this probability, and go from there.]

Solution :

Consider a unit square with coordinates { (0,0) , (0,1) , (1,0) and (1,1) }. There is a circle inscribed in the square with center at (0.5, 0.5)



If a choose an arbitrary point whose coordinates lie between 0 and 1, then

P(point lying in the inscribed circle) = Area of circle / Area of Square

$$= \pi (0.5)^2 / 1^2$$

$$= \pi (0.25) = \pi / 4$$

Hence, choose a point and find its probability that it lies inside the inscribed circle. The probability value multiplied by 4 gives an estimate of pie (π) value.

Sample size = 10,000

```
> ########################################################
> ######Question 2 : Estimating pi value ###################
> ########################################################
>
> sq_sample_size = 10000
> sq_x_coord = runif(1) # single x-coordinate
> sq_y_coord = runif(1) # single y-coordinate
> sq_Distance_from_circle_center = replicate(sq_sample_size , (runif(1) - 0.5)^2 + (runif(1) - 0.5)^2) #Calculating distance square from the circle center
> ## If the point lies in the circle then the square of distance is less than or equal to radius square i.e (0.5)^2
> sq_MC_pi = mean(sq_Distance_from_circle_center <= 0.5^2) * 4  #calculating the probability that the point lies inside the inscribed circle

> sq_MC_pi
[1] 3.1492
>
```

sq_Distance_from_circle_center	num [1:10000] 0.24 0.314 0.011 0.162 0.425 ...
sq_MC_pi	3.1492
sq_sample_size	10000
sq_x_coord	0.152509116800502
sq_y_coord	0.180913888150826

Using a Monte Carlo approach based on 10,000 replications, the estimated value of pi is **3.1492** which is close to the actual value of pi of 3.1415