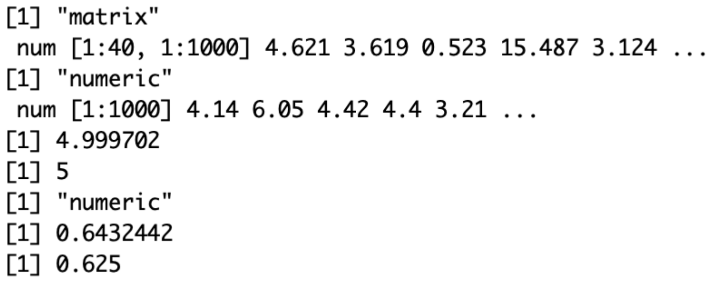
Part 1

1. Run 1000 simulations, each with 40 samples of an exponential distribution, calculate the mean and variance and compare them to the theoretical estimates. Control limit therom: as the number in the sample increases or the number of simulations increase, the sample mean and variance should more acurately represent the theoretical estimate

R codes:

```
set.seed(100)
lamba <- 0.2
n <- 40
simulations <- 1000


sim.exp <- replicate(simulations,rexp(n,lamba))
class(sim.exp)
str(sim.exp)
#it is the matrix of 40 rows and 1000 columns
#use apply to calculate the mean of this matrix

sim.exp_mean <- apply(sim.exp, 2, mean)
class(sim.exp_mean)
str(sim.exp_mean)
```

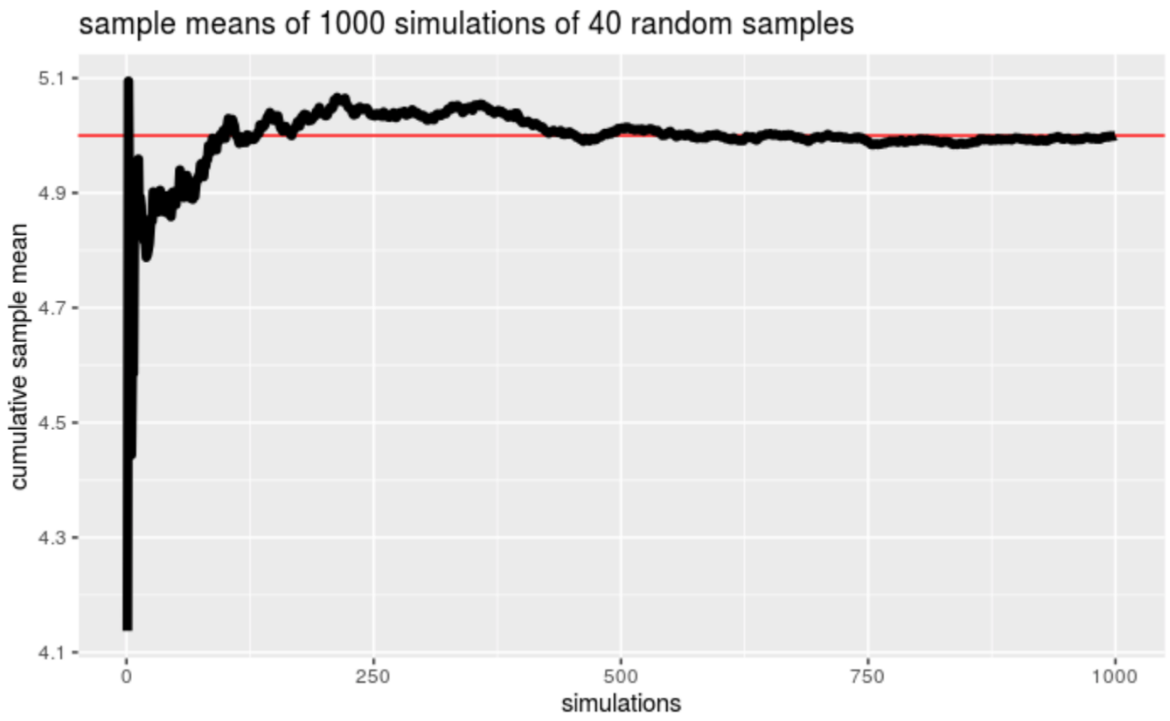


```
[1] "matrix"
 num [1:40, 1:1000] 4.621 3.619 0.523 15.487 3.124 ...
[1] "numeric"
 num [1:1000] 4.14 6.05 4.42 4.4 3.21 ...
[1] 4.999702
[1] 5
[1] "numeric"
[1] 0.6432442
[1] 0.625
```

Q1:

R codes:

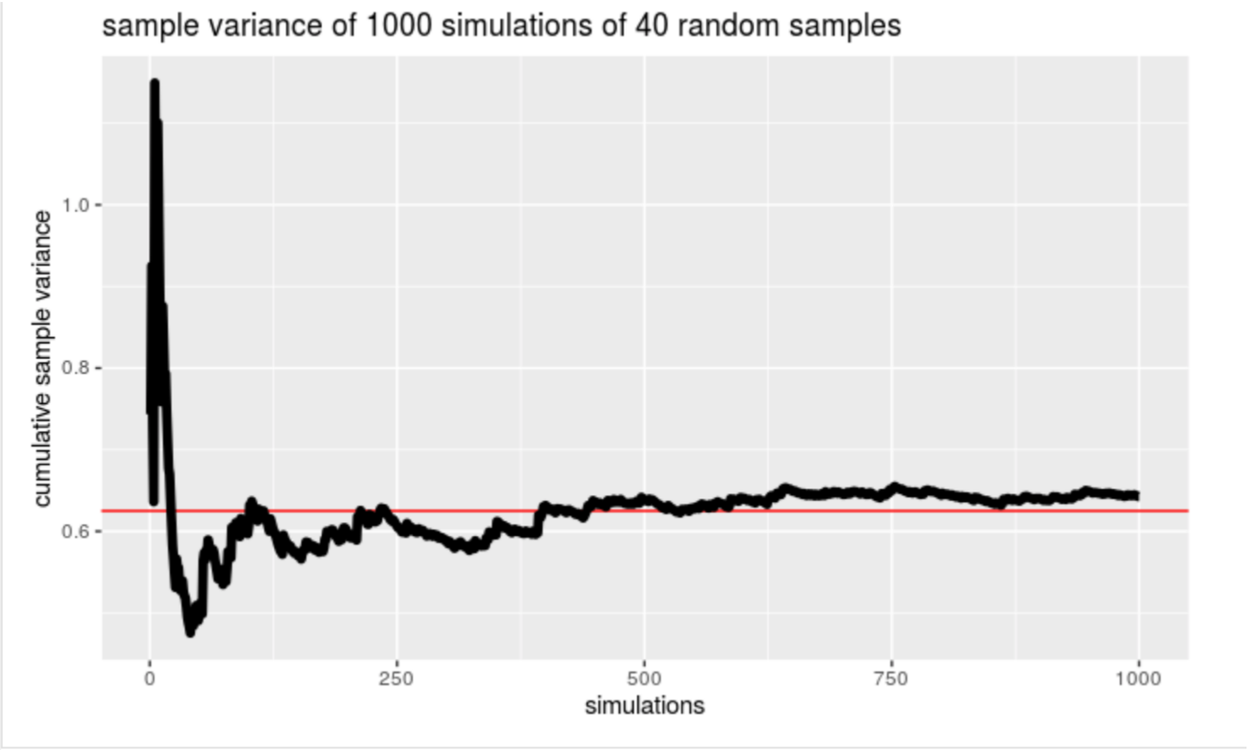
```
sample_mean <- mean(sim.exp_mean)
sample_mean

theore_mean <- 1/lamba
theore_mean

#use the ggplot to compare the mean
cum_mean <- cumsum(sim.exp_mean)/(1:simulations)
class(cum_mean)
g <- ggplot(data.frame(x=1:simulations, y=cum_mean), aes(x=x,y=y))+geom_hline(yintercept = theore_mean, color='red')+geom_line(size=2)
g <- g+labs(x='simulations', y='cumulative sample mean')+ggtitle('sample means of 1000 simulations of 40 random samples')
g
```



Sample mean=4.999
Theoretical mean=5

Q2:
R codes:

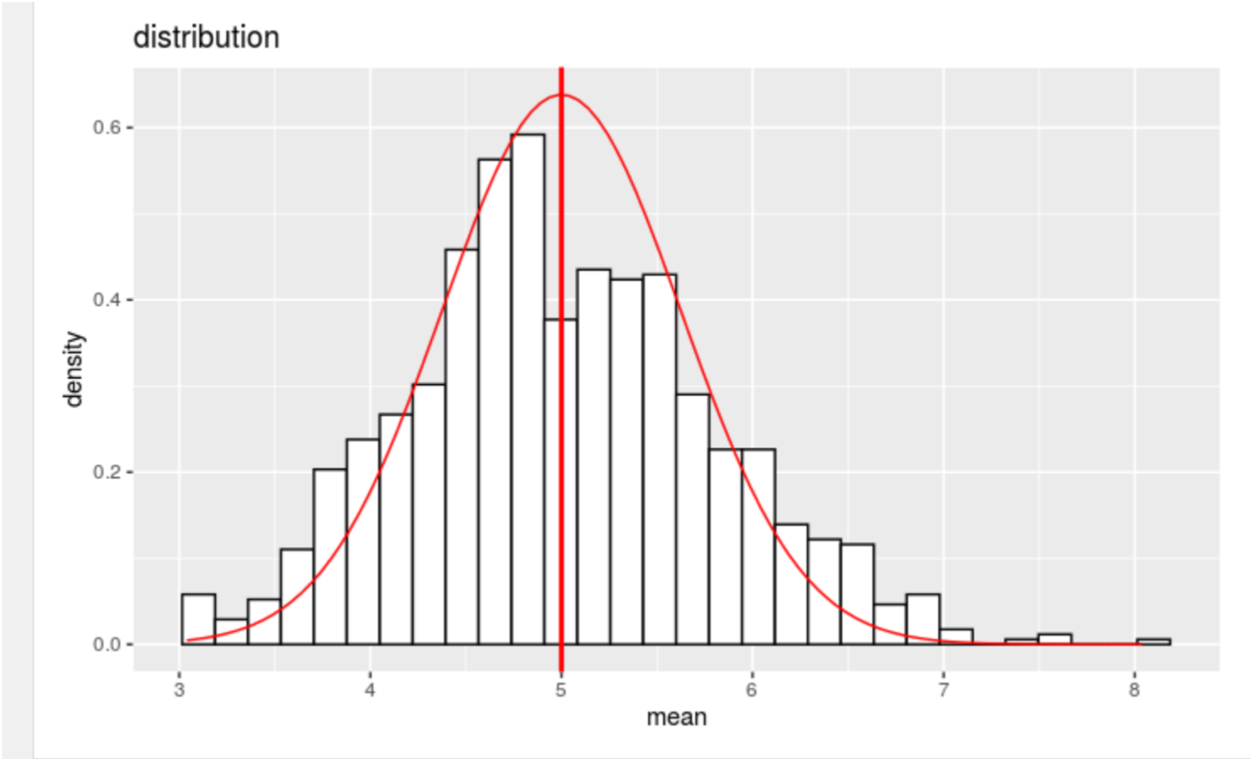
```
 sample_var <- var(sim.exp_mean)
sample_var

theore_var <- ((1/lamba)^2)/n
theore_var

#use the ggplot to comapre the variance
cum_var <- cumsum((sim.exp_mean-sample_mean)^2)/(seq_along(sim.exp_mean-1))
g <- ggplot(data.frame(x=1:simulations, y=cum_var), aes(x=x,y=y))+geom_hline(yintercept = theore_var, color='red')+geom_line(size=2)
g <- g+labs(x='simulations', y='cumulative sample variance')+ggtitle('sample variance of 1000 simulations of 40 random samples')
g
```



Q3 :Show that the distribution is approximately normal.

R codes:

```
# Use the ggplot to plot the exponential distribution and the normal distribution
g <- ggplot(data.frame(x=sim.exp_mean),
aes(x=x))+geom_histogram(aes(y=..density..),color='black',fill='white',bins=30)
g <- g+stat_function(fun=dnorm, color='red', args=list(mean=theore_mean, sd=theore_var))+geom_vline(xintercept = theore_mean,color='red',size=1)
g <- g+xlab('mean')+ylab('density')+ggtitle('distribution')
g
```



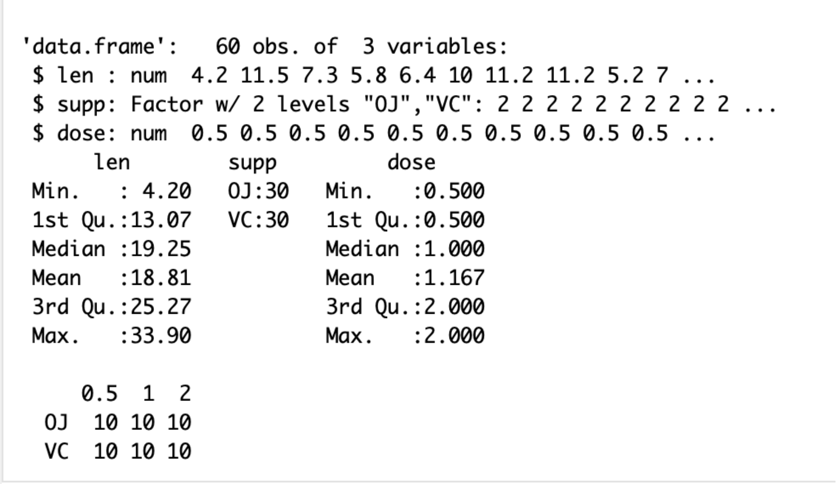
Conclusion
From the graph above, this exponential distribution is approximately normal.

Part 2 Analyze the ToothGrowth data in the R datasets package

Q1. Load the ToothGrowth data and perform some basic exploratory data analyses

R codes:

```
data("ToothGrowth")
str(ToothGrowth)
summary(ToothGrowth)
table(ToothGrowth$supp,ToothGrowth$dose)
```



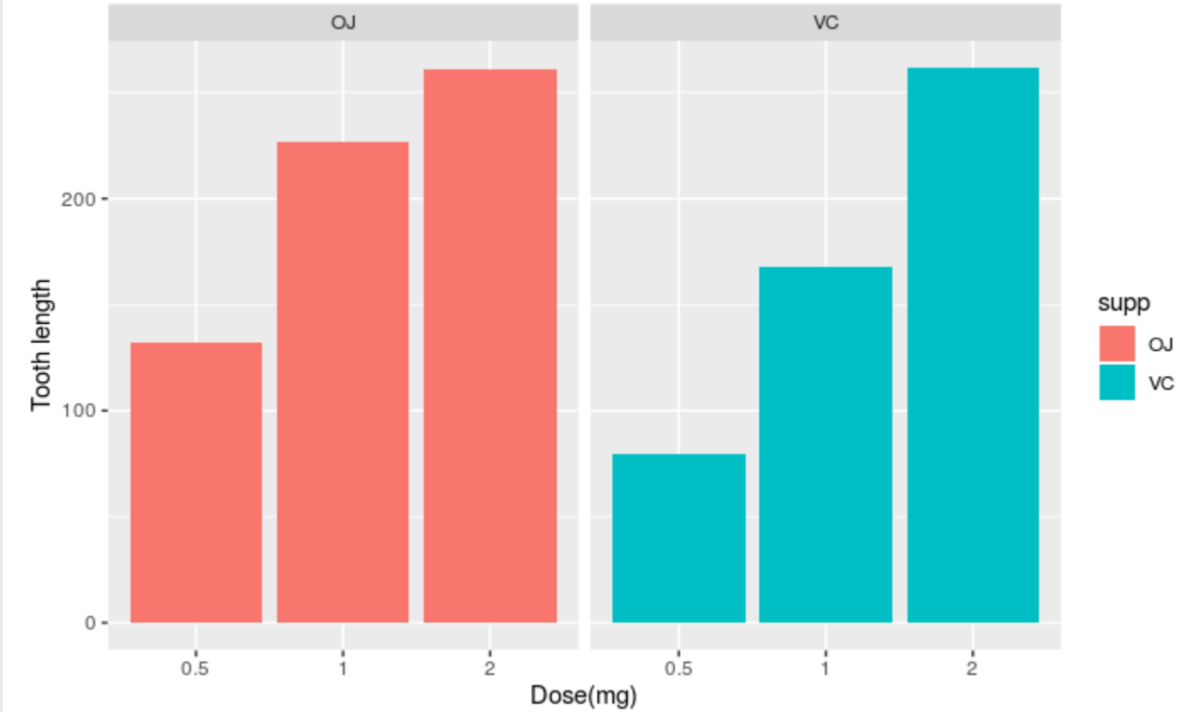
```
'data.frame':	60 obs. of  3 variables:
 $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
 $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
 $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
      len        supp         dose      
 Min.   : 4.20   OJ:30   Min.   :0.500  
 1st Qu.:13.07   VC:30   1st Qu.:0.500  
 Median :19.25           Median :1.000  
 Mean   :18.81           Mean   :1.167  
 3rd Qu.:25.27           3rd Qu.:2.000  
 Max.   :33.90           Max.   :2.000  

     0.5  1  2
  OJ  10 10 10
  VC  10 10 10
```

This dataframe contains 60 observations of 3 variables.Variables supp of type factors has two levels (VC ad OJ). From the table command, there are three dose levels: 0.5,1,and 2 mg/day.

Q2. Provide a basic summary of the data.
From Q1, this command:summary(ToothGrowth); provides the basic summary of this data
We can see the min, 1^{st} qu, median, 3 rd qu and max of len and dose. Also, we can sue ggplot to apply more detailed analysis.

R codes:

```
g <- ggplot(data=ToothGrowth, aes(x=as.factor(dose), y=len, fill=supp)) + geom_bar(stat="identity") +
facet_grid(. ~ supp) +xlab("Dose(mg)") +ylab("Tooth length")
g
```

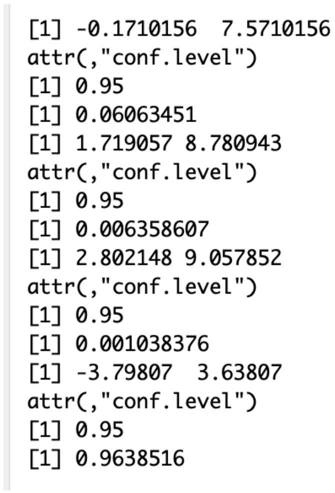


Q3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

Test the null hypothesis that mean length of odontoblasts for ascorbic acid (VC) and orange juice (OJ) does not differ significantly

R codes:

```
hypoth1 <- t.test(len ~ supp, data = ToothGrowth)
 hypoth1$conf.int
 hypoth1$p.value
 #alpha=0.05

hypoth2<-t.test(len ~ supp, data = subset(ToothGrowth, dose == 0.5))
hypoth2$conf.int
hypoth2$p.value

hypoth3<-t.test(len ~ supp, data = subset(ToothGrowth, dose == 1))
hypoth3$conf.int
hypoth3$p.value

hypoth4<-t.test(len ~ supp, data = subset(ToothGrowth, dose == 2))
hypoth4$conf.int
hypoth4$p.value
```



```
[1] -0.1710156  7.5710156
attr(,"conf.level")
[1] 0.95
[1] 0.06063451
[1] 1.719057 8.780943
attr(,"conf.level")
[1] 0.95
[1] 0.006358607
[1] 2.802148 9.057852
attr(,"conf.level")
[1] 0.95
[1] 0.001038376
[1] -3.79807  3.63807
attr(,"conf.level")
[1] 0.95
[1] 0.9638516
```

Q4. State your conclusions and the assumptions needed for your conclusions.

OJ ensures more tooth growth than VC for dosages 0.5 & 1.0. Because p-value is smaller than alpha, so we reject the null hypothesis and conclude that a significant difference does exis.
OJ and VC gives the same amount of tooth growth for dose amount 2.0 mg/day. Because p-value is larger than alpha, so we cannot conclude that a significant difference exists
Therefore, for the entire trail we cannot conclude OJ is more effective that VC for all situations.