assignment

Sagnik Chakravarty (me)

Part 1: Simulation

Here we will first simulate an exponential distribution in R with function rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. We will set lambda = 0.2 for all of the simulations. We will investigate the distribution of averages of 40 exponentials and will to do this for thousand simulations.

Showing sample mean and comparing it to the theoretical mean of the distribution

Let's first show the sample mean and compare it to the theoretical mean. OUr theoretical mean here is 1/0.2= 5.

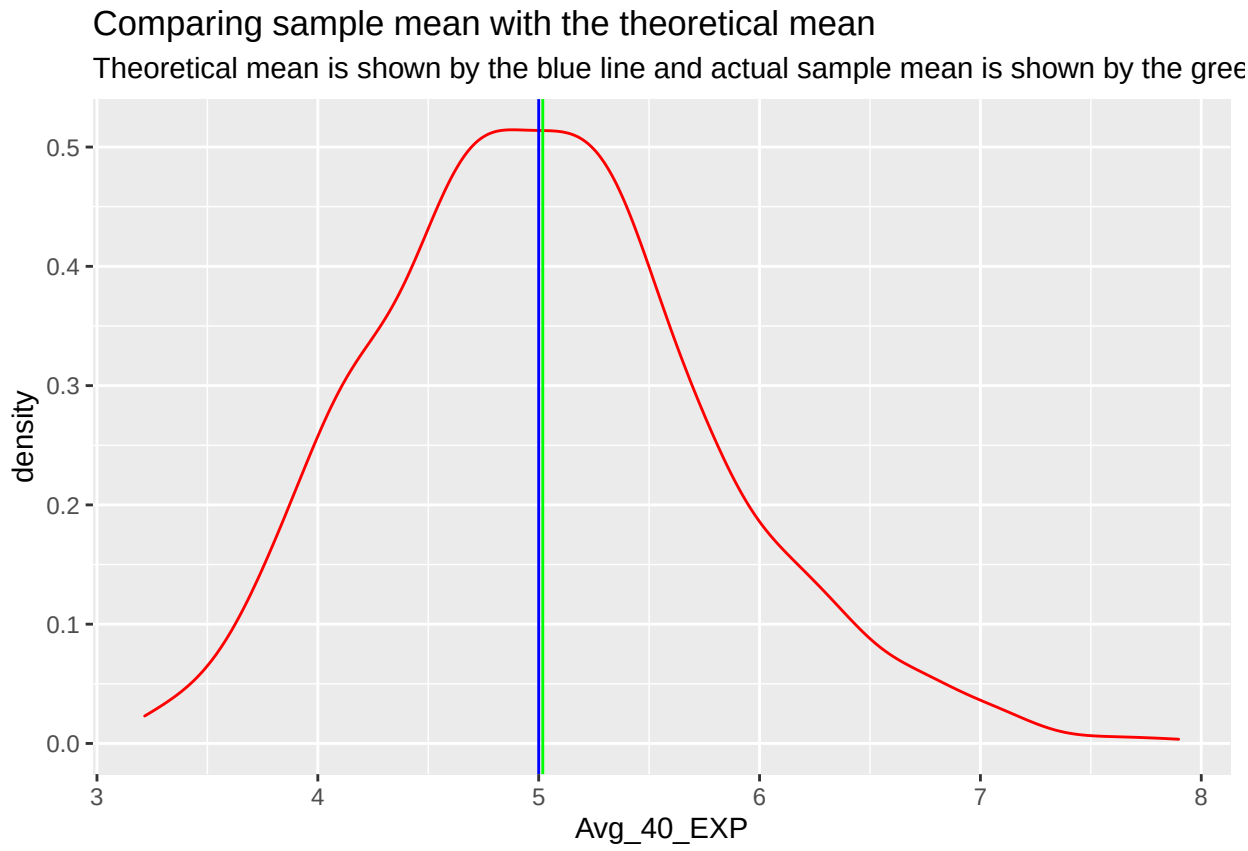
Let's create a function called mean_exp for handing the sample mean.

Our function mean_exp will first take argument no_sim, which stands for number of simulations. It will then simulate 40 exponentials with rate parameter 0.2, then find the mean (sample mean) of these 40 exponentials and store it in a dataframe called exp_df and repeat the same procedure as per the number of simulations entered by the user(value of no_sim). We will do 1000 simulations here so no_sim= 1000.

Finally it will plot density curve of this distribution of sample means and we will compare the theoretical mean i.e. 5, by plotting the two x intercepts i.e. theoretical mean (which is 5 denoted by blue line) and actual sample mean (green line).

```
knitr::opts_chunk$set(echo = TRUE)

library(ggplot2)

#Let's create a function called mean_exp for handing the sample mean.

#Our function mean_exp will first take argument no_sim, which stands for number of simulations. It will

#Finally it will plot density curve of this distribution of sample means and we will compare the theore

mean_exp<- function(no_sim){
  exp_df<- data.frame(x= c())
  for(i in 1:no_sim){

    exp<- rexp(n = 40, rate = 0.2)
    exp_mean<- mean(exp)
    exp_df<- rbind.data.frame(exp_df, exp_mean)
  }
  names(exp_df)<- 'Avg_40_EXP'
  x<- ggplot(exp_df, aes(Avg_40_EXP))+ geom_density(col='red')+ geom_vline(xintercept=c(1/0.2, mean(exp
```

```
  print(x)
  theoretical_mean= 5
  actual_mean= mean(exp_df$Avg_40_EXP)
  print('The theoretical mean is:')
  print(theoretical_mean)
  print('The actual sample mean is:')
  print(actual_mean)
}
mean_exp(1000)
```




```
## [1] "The theoretical mean is:"
## [1] 5
## [1] "The actual sample mean is:"
## [1] 5.018498
```

So as we can see from here, the blue line(theoretical mean) and the green line(actual sample mean) are very close to each other implying there is hardly any difference between the theoretical mean and the sample mean.

Showing sample variance and comparing it to the theoretical variance of the distribution

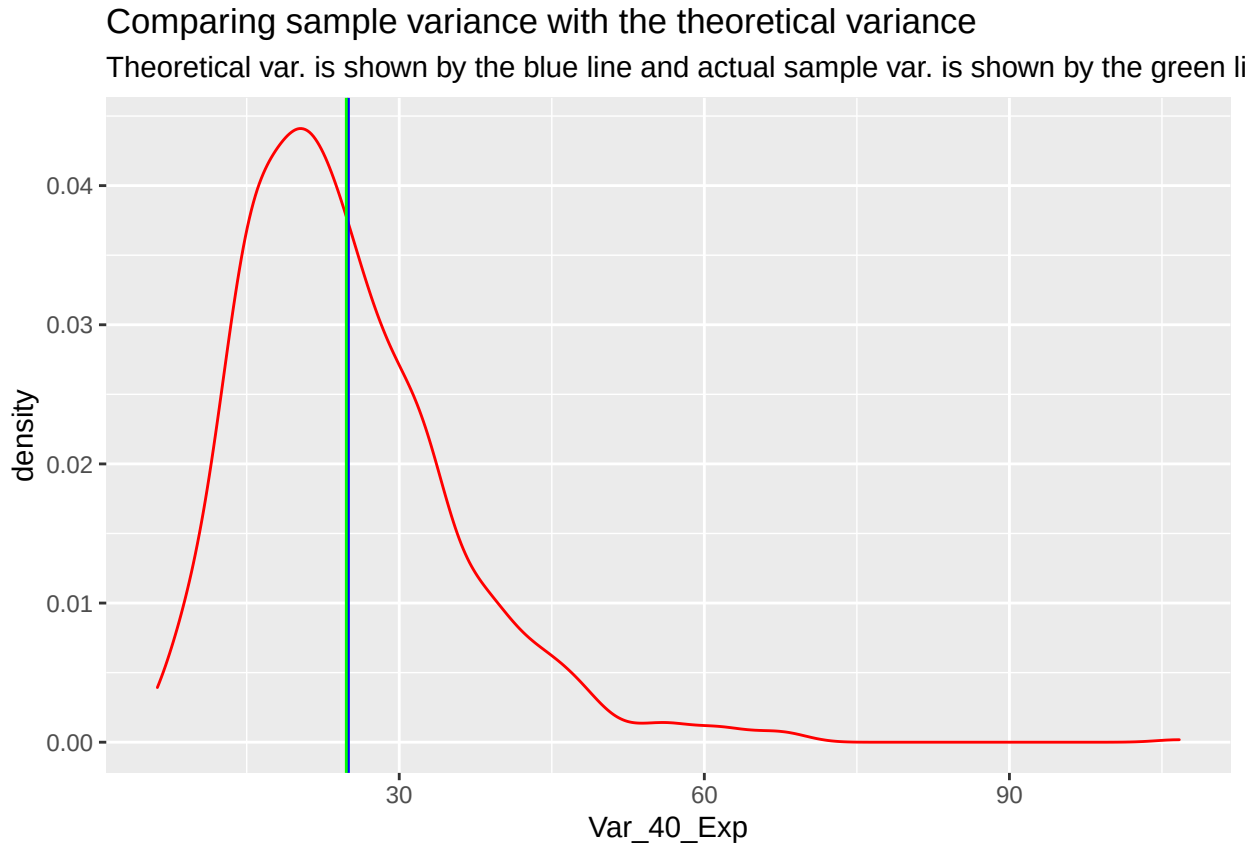
Now we will show the sample variance and compare it to the theoretical variance. OUr theoretical variance here is $(1/0.2)^2 = 25$.

Let's create a function called var_exp for handing the sample variance

Our function var_exp will first take argument no_sim, which stands for number of simulations. It will then simulate 40 exponentials with rate parameter 0.2, then find the variance (sample variance) of these 40 exponentials and store it in a dataframe called exp_df_var and repeat the same procedure as per the number of simulations entered by the user(value of no_sim). We will do 1000 simulations here so no_sim= 1000.

Finally it will plot density curve of this distribution of sample variances and we will compare the theoretical variance i.e. 25, by plotting the two x intercepts i.e. theoretical variance (which is 25 denoted by blue line) and actual sample variance (green line).

```
#Let's create a function called var_exp for handing the sample variance
#Our function var_exp will first take argument no_sim, which stands for number of simulations. It will

#Finally it will plot density curve of this distribution of sample variances and we will compare the th

var_exp<- function(no_sim){
  exp_df_var<- data.frame(x=c())

  for(i in 1:no_sim){
    exp<- rexp(n = 40, rate = 0.2)
    exp_var<- (sd(exp))^2
    exp_df_var<- rbind.data.frame(exp_df_var, exp_var)
  }

  names(exp_df_var)<- 'Var_40_Exp'
  y<-ggplot(exp_df_var, aes(Var_40_Exp))+ geom_density(col='red')+ geom_vline(xintercept=c((1/0.2)^2,  r

  print(y)

  theoretical_variance<- (1/0.2)^2
  actual_sample_variance<- mean(exp_df_var$Var_40_Exp)
  print('The theoretical variance is:')
  print(theoretical_variance)
  print('The actual sample variance is:')
  print(actual_sample_variance)

}
var_exp(1000)
```




```
## [1] "The theoretical variance is:"
## [1] 25
## [1] "The actual sample variance is:"
## [1] 24.79356
```

So as we can see the blue line(theoretical variance) and the green line(sample variance) are very close to each other implying there is very slight difference between the theoretical variance and the sample variance

Showing that the distribution of the sample means is approximately normal using Central Limit Theorem

The central limit theorem tells us that if we have a population with mean mu and variance sigma, as the sample size increases the distributiion of sample means becomes approximately normal.

So inorder to prove that our sample mean distribution here is approximately normal, we will increase the sample size and see if its density plot becomes more and more Gausian or not.

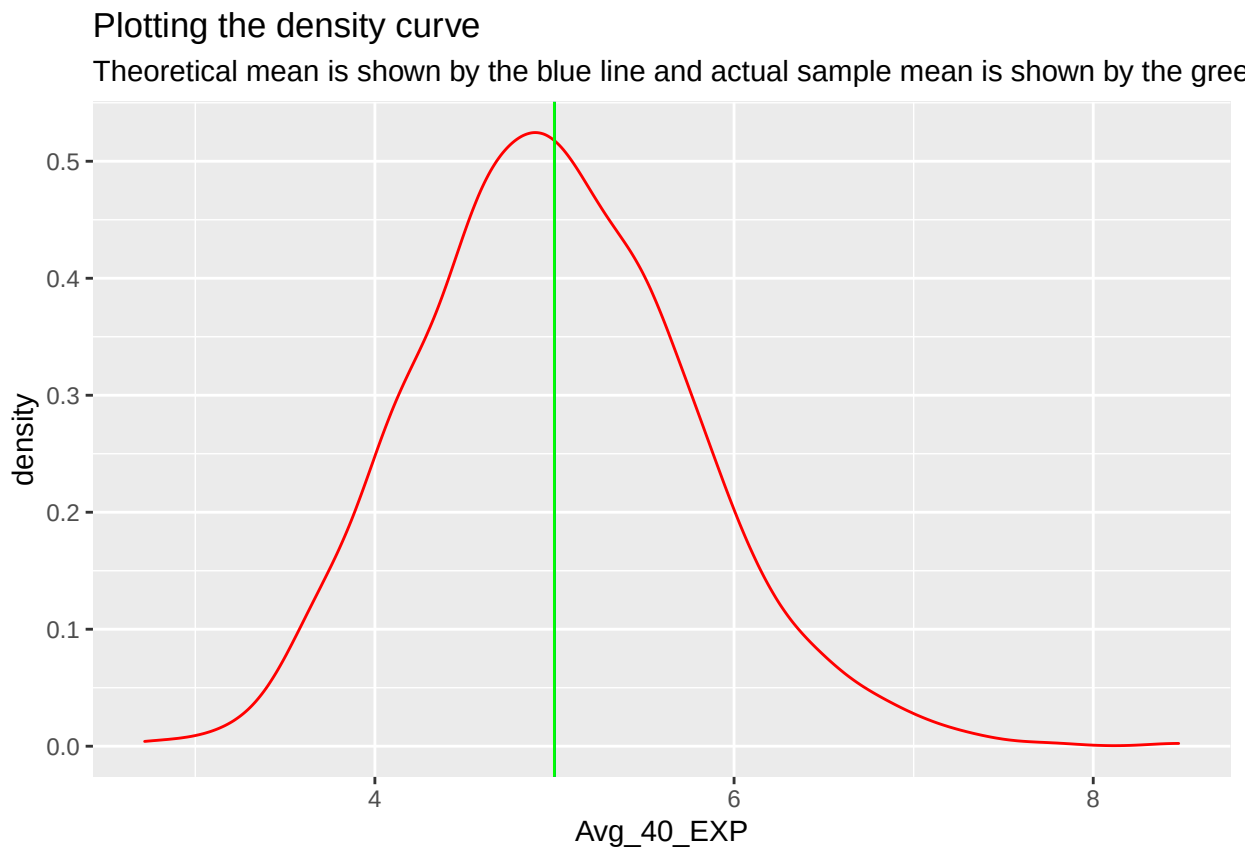
We will create a new function mean_exp2 for this exercise. Ir will be very similar to our previous mean_exp function but this time we will make sample size n an argument as well.

We will do 1000 simulations and increase the sample size n from 40 to 60 to 80 and then to 100. We will observe how the density curve becomes more and more Gausian

```r
#We will create a new function mean_exp2 for this exercise. Ir will be very similar to our previous mea
#We will do 1000 simulations and increase the sample size n from 40 to 60 to 80 and then to 100. We wil

mean_exp2<- function(no_sim, n){
  exp_df<- data.frame(x= c())
  for(i in 1:no_sim){

    exp<- rexp(n, rate = 0.2)
    exp_mean<- mean(exp)
    exp_df<- rbind.data.frame(exp_df, exp_mean)
  }
  names(exp_df)<- 'Avg_40_EXP'
  x<- ggplot(exp_df, aes(Avg_40_EXP))+ geom_density(col='red')+ geom_vline(xintercept=c(1/0.2, mean(exp

  print(x)
  theoretical_mean= 5
  actual_mean= mean(exp_df$Avg_40_EXP)
  print('The theoretical mean is:')
  print(theoretical_mean)
  print('The actual sample mean is:')
  print(actual_mean)
}
```

Let's first do 1000 simulations for sample size 40

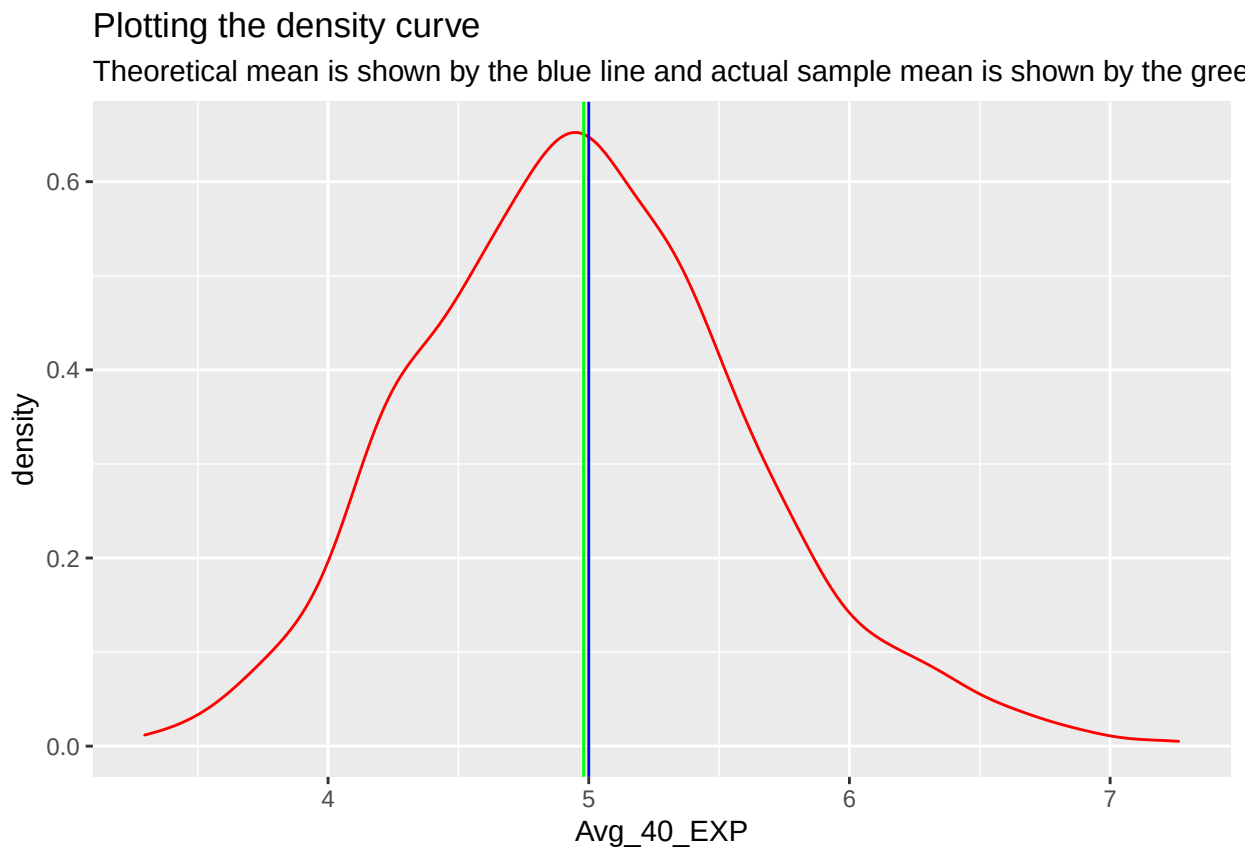
```r
mean_exp2(1000, 40)
```



```
## [1] "The theoretical mean is:"
## [1] 5
## [1] "The actual sample mean is:"
## [1] 4.999631
```

Now let's do 1000 simulations for sample size 60

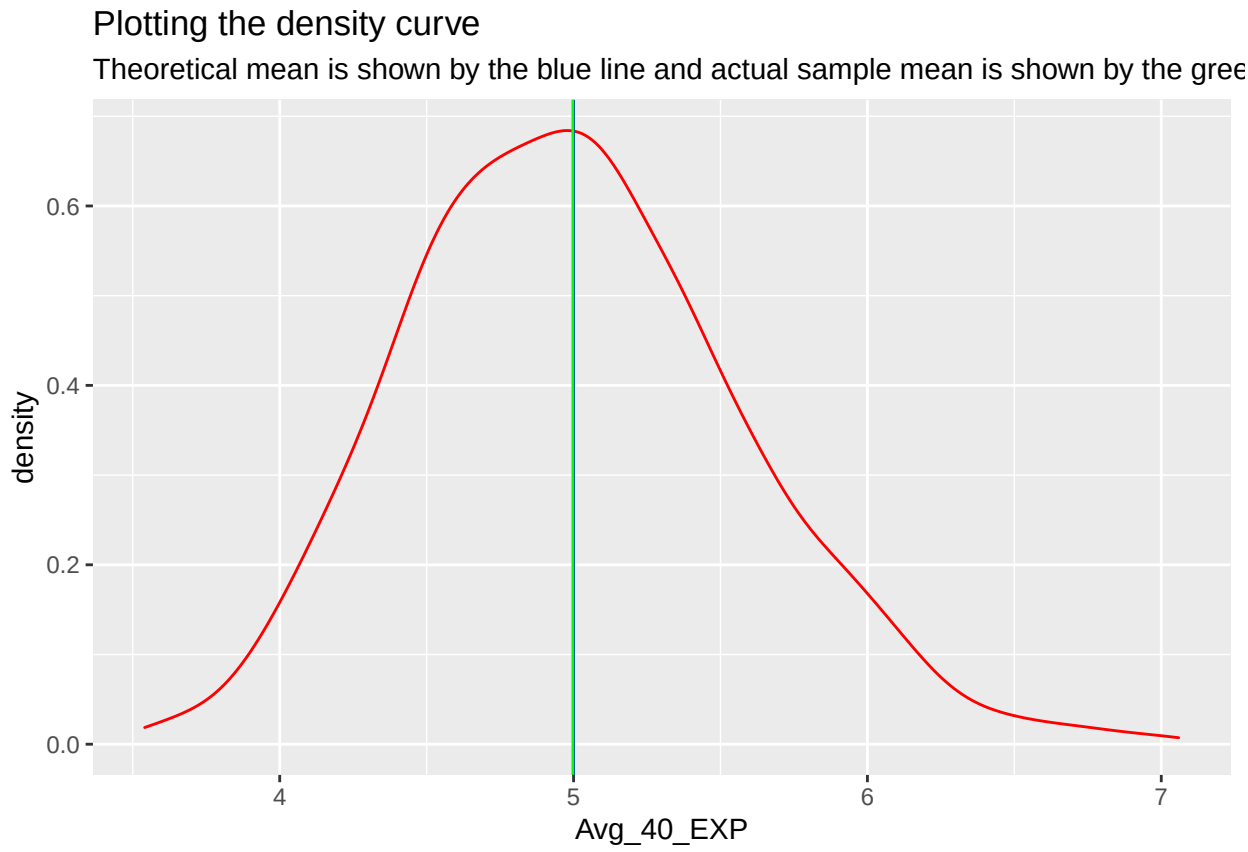
```
mean_exp2(1000, 60)
```




```
## [1] "The theoretical mean is:"
## [1] 5
## [1] "The actual sample mean is:"
## [1] 4.980246
```

Now let's do 1000 simulations for sample size 80

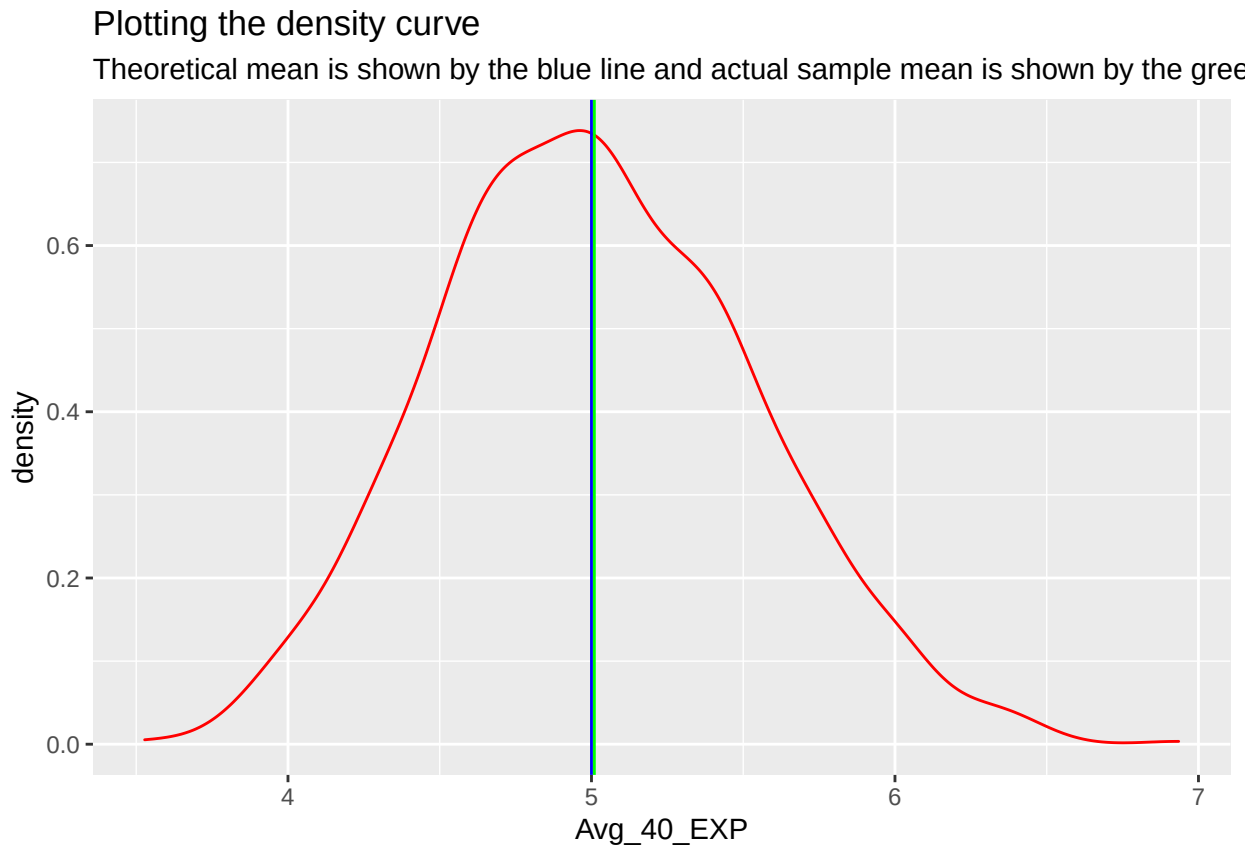
```
mean_exp2(1000, 80)
```




```
## [1] "The theoretical mean is:"
## [1] 5
## [1] "The actual sample mean is:"
## [1] 4.997783
```

Now let's do 1000 simulations for sample size 100

```
mean_exp2(1000, 100)
```




```
## [1] "The theoretical mean is:"
## [1] 5
## [1] "The actual sample mean is:"
## [1] 5.009445
```

So as we observe from the above 4 plots/simulations, the density curve becomes more and more Gausian as we increase the sample size 40 to 60 to 80 to 100. So the Central Limit Theorem is validated and our sample mean distribution is approximately normal

Part 2: Data Analysis of toothgrowth data

Here in this part, we are going to do some analysis on the ToothGrowth data in the R datasets package.

We will first load the ToothGrowth dataset into R and then do some exploratory analysis to get a good sense of the data. After that, we will run some appropriate tests to make some inferences about the data.

Exploratory analysis

Let's first load the dataset into R and do some exploratory analysis.

```
library(datasets) #loading the ToothGrowth datset
dataset<- ToothGrowth
head(dataset)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
summary(dataset)
```

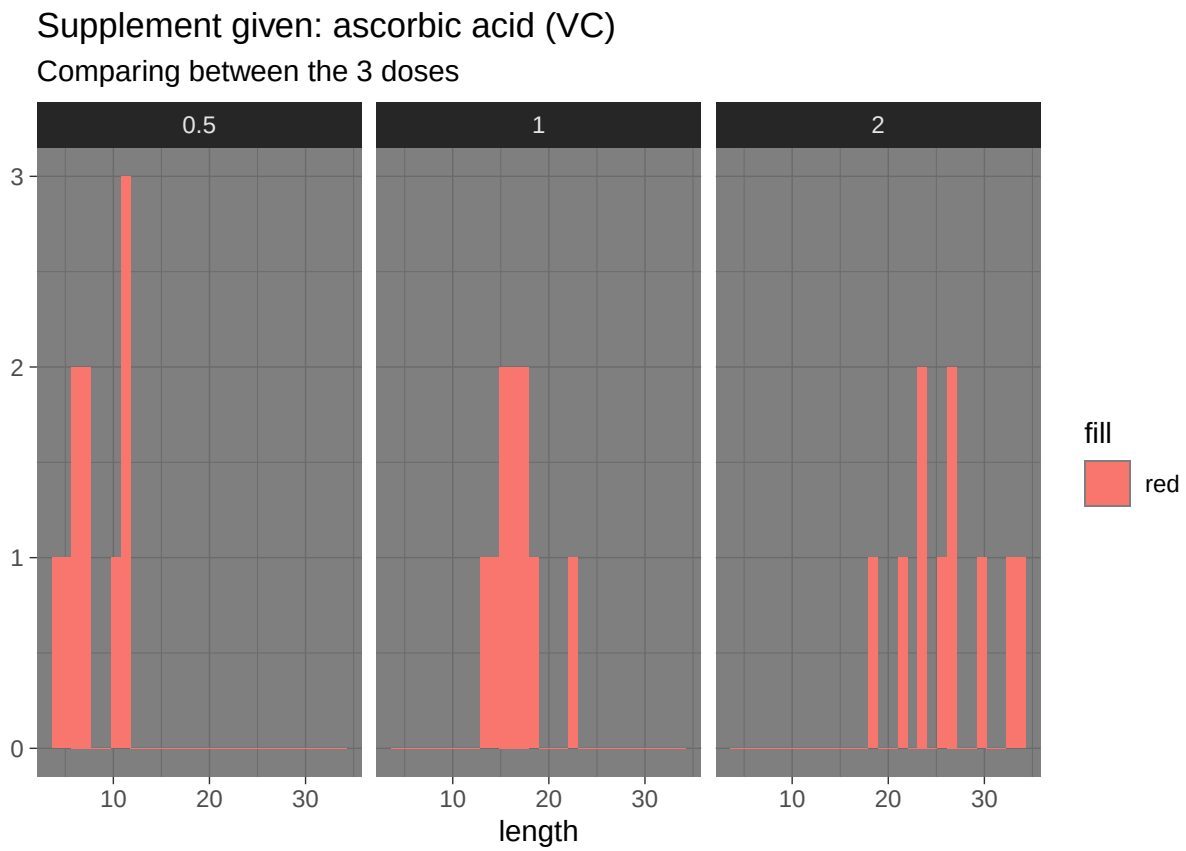
```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

```
#separating the dataset by the supp (supplement) variable i.e. VC (Ascorbic acid) and OJ (Orange juice)
dataset_VC<- subset(dataset, supp=='VC')
dataset_OJ<- subset(dataset, supp=='OJ')

library(ggplot2)
```

Comparing the toothgrowth i.e. len variable across the 3 doses for animals who were given ascorbic acid

```
qplot(len, data = dataset_VC, facets = .~dose, xlab = 'length', fill= 'red')+ theme_dark()+ggtitle('Supp
```

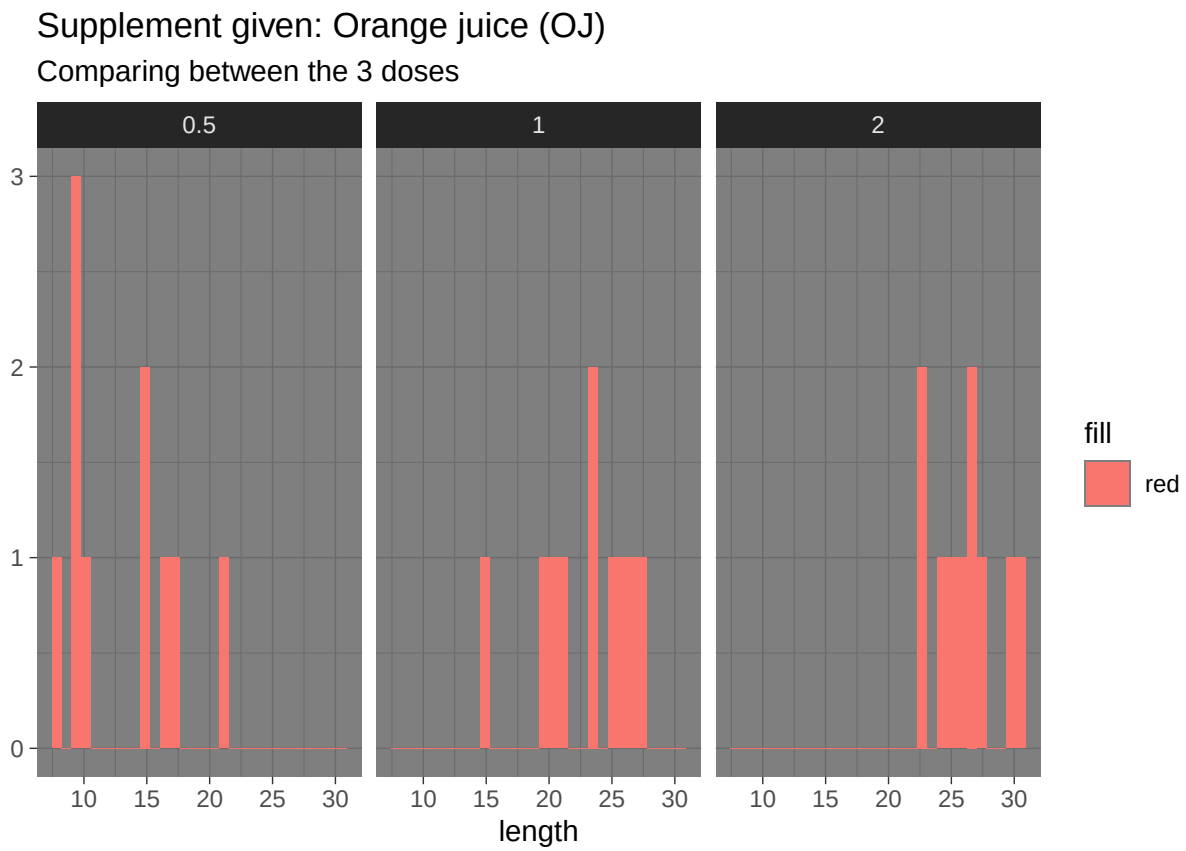
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



comparing the toothgrowth i.e. len variable across the 3 doses for animals who were given orange juice

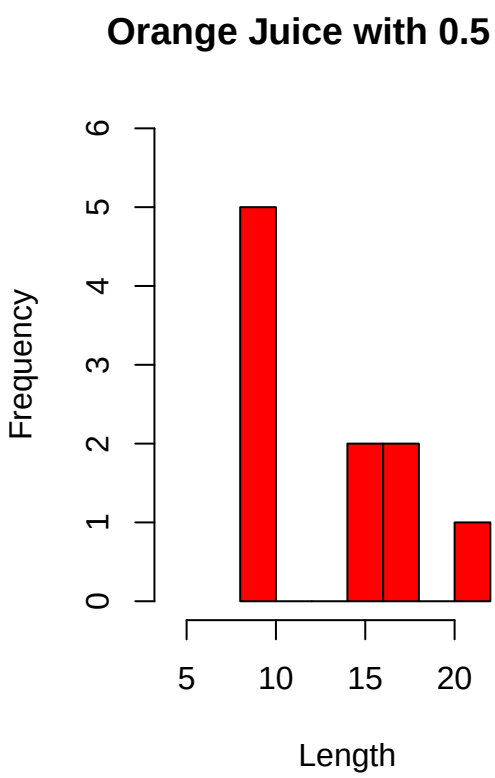
```
qplot(len, data = dataset_OJ, facets = .~dose, xlab = 'length', fill= 'red')+ theme_dark()+ggtitle('Supp
```

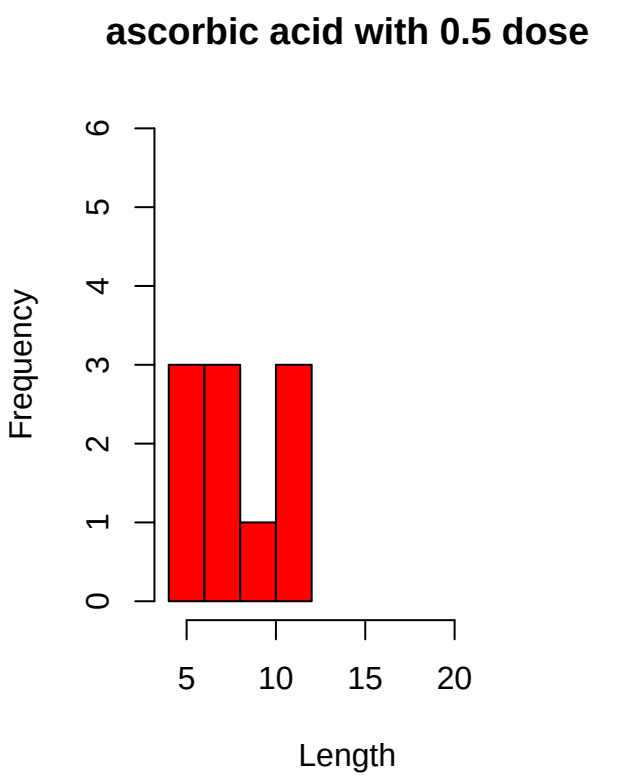
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Comparing toothgrowth between animals who were given Orange juice and those who were given ascorbic acid for dose 0.5 mg/day

```
#dose of 0.5 mg/day
par(mfrow=c(1,2))
hist(dataset_OJ[dataset_OJ$dose==0.5,]$len, xlim = c(4,24), ylim = c(0, 6), main = 'Orange Juice with 0
hist(dataset_VC[dataset_VC$dose==0.5,]$len, xlim = c(4,24), ylim = c(0, 6), main = 'ascorbic acid with
```






```
#comparing the means
mean(dataset_OJ[dataset_OJ$dose==0.5,]$len) #Orange juice
```

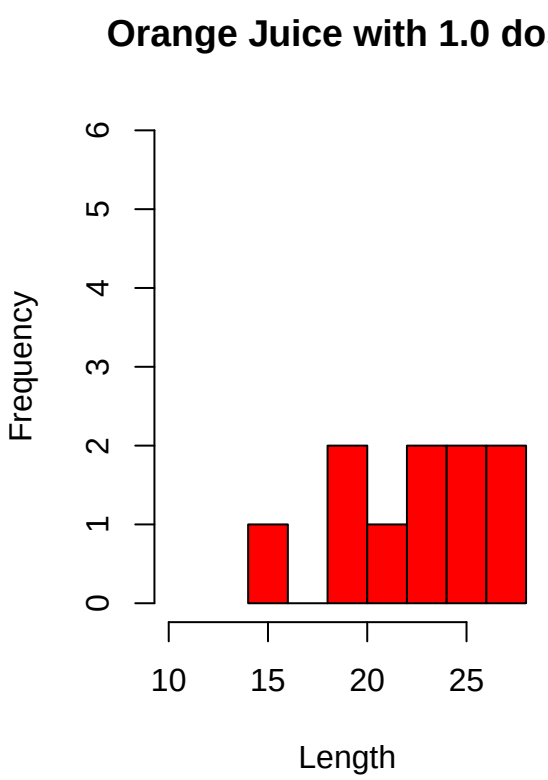
```
## [1] 13.23
```

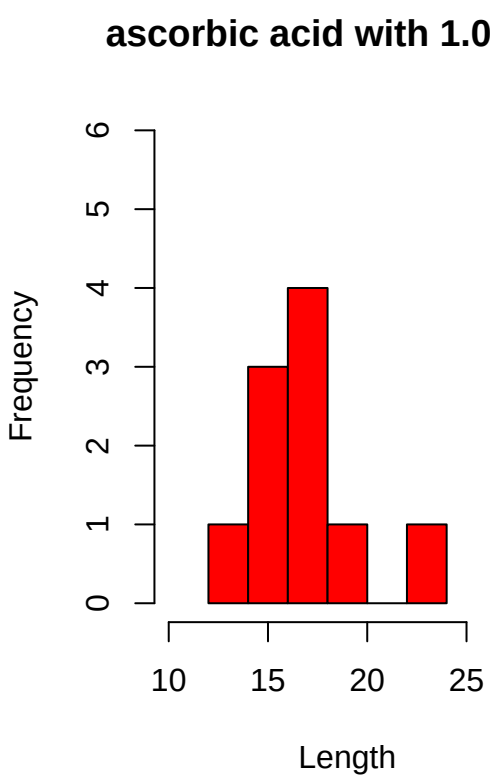
```
mean(dataset_VC[dataset_VC$dose==0.5,]$len) #Ascorbic acid
```

```
## [1] 7.98
```

Comparing toothgrowth between animals who were given Orange juice and those who were given ascorbic acid for dose 1.0 mg/day

```
#dose of 1.0 mg/day
par(mfrow=c(1,2))
hist(dataset_OJ[dataset_OJ$dose==1.0,]$len, xlim = c(10,28), ylim = c(0, 6), main = 'Orange Juice with
hist(dataset_VC[dataset_VC$dose==1.0,]$len, xlim = c(10,28), ylim = c(0, 6), main = 'ascorbic acid with
```






```
#Comparing means
mean(dataset_OJ[dataset_OJ$dose==1.0,]$len) #Orange juice
```

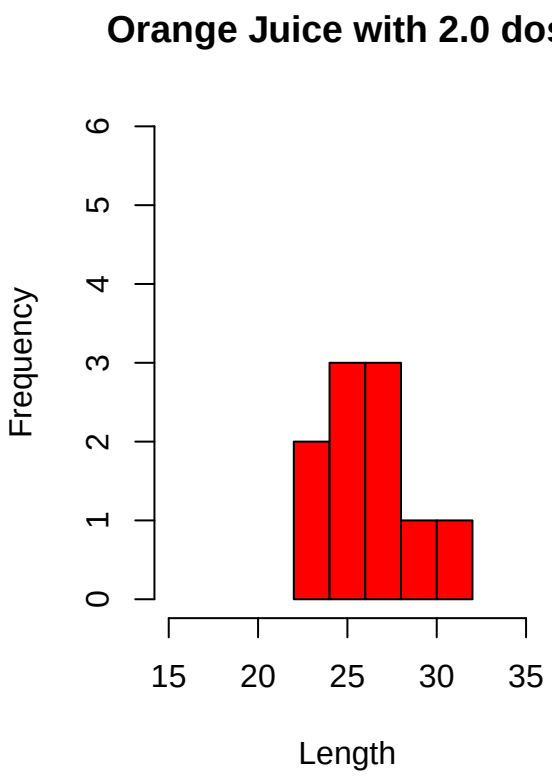
```
## [1] 22.7
```

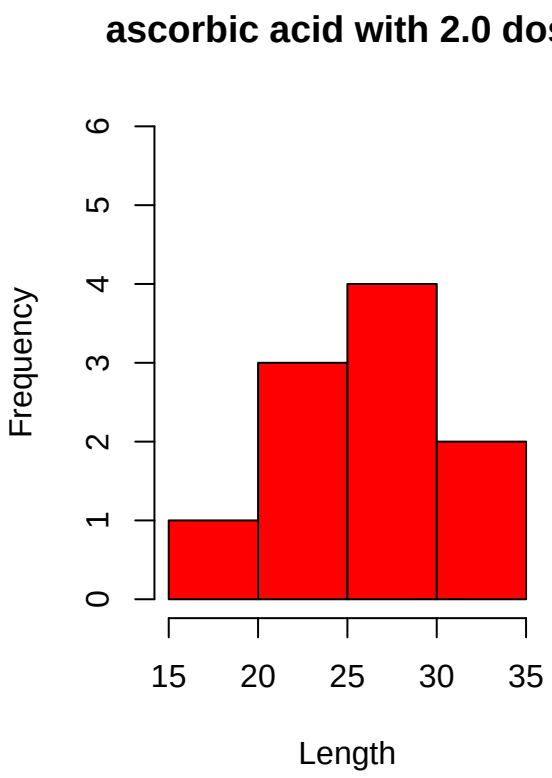
```
mean(dataset_VC[dataset_VC$dose==1.0,]$len) #Ascorbic acid
```

```
## [1] 16.77
```

Comparing toothgrowth between animals who were given Orange juice and those who were given ascorbic acid for dose 2.0 mg/day

```
#dose of 2.0 mg/day
par(mfrow=c(1,2))
hist(dataset_OJ[dataset_OJ$dose==2.0,]$len, xlim = c(15,35), ylim = c(0, 6), main = 'Orange Juice with 1
hist(dataset_VC[dataset_VC$dose==2.0,]$len, xlim = c(15,35), ylim = c(0, 6), main = 'ascorbic acid with
```






```
#Comparing means
mean(dataset_OJ[dataset_OJ$dose==2.0,]$len) #Orange juice
```

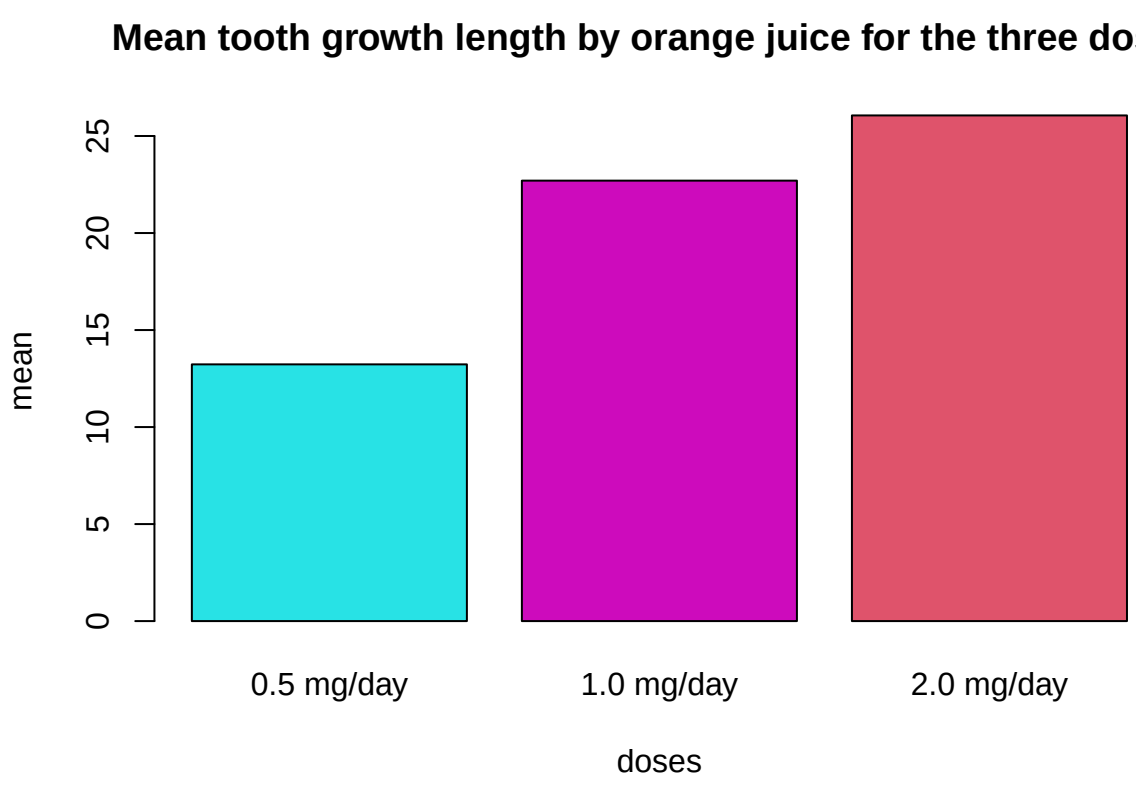
```
## [1] 26.06
```

```
mean(dataset_VC[dataset_VC$dose==2.0,]$len) #Ascorbic acid
```

```
## [1] 26.14
```

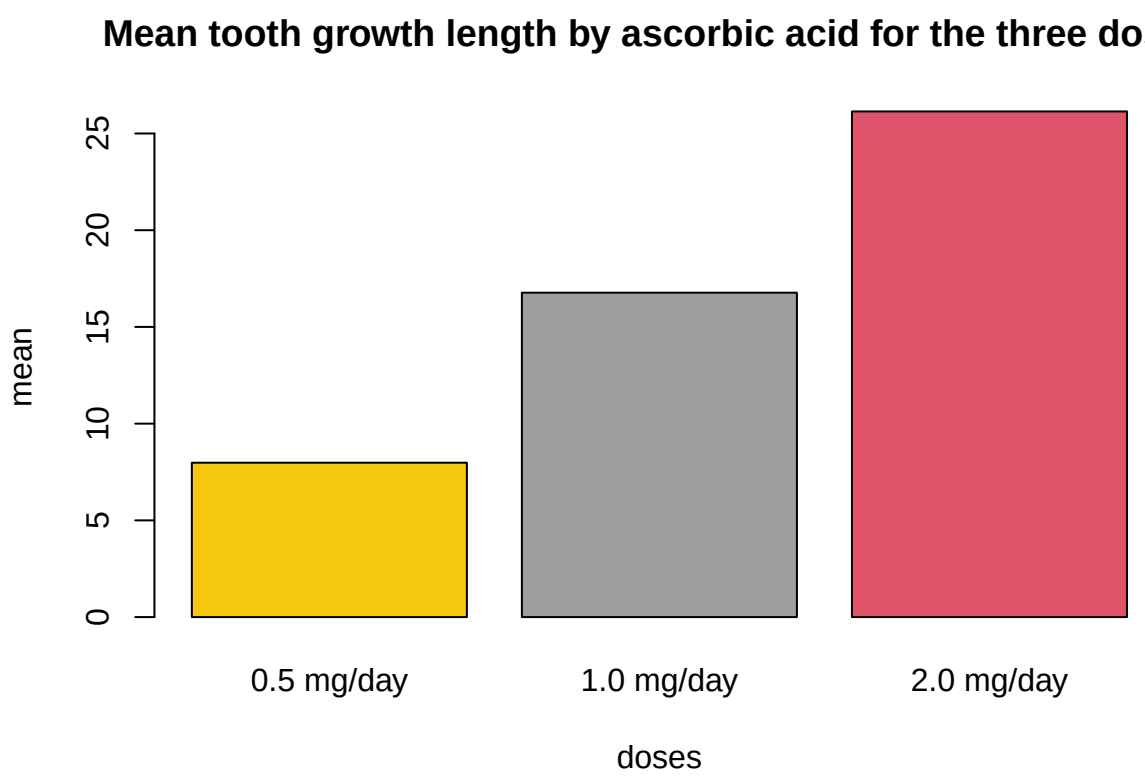
Comparing mean tooth growth for animals who were given orange juice across the 3 doses: 0.5 mg/day, 1.0 mg/day and 2.0 mg/day

```
par(mfrow= c(1,1), mar=c(5.1, 4.1, 4.1, 2.1))
OJ_mean_0.5<- mean(dataset_OJ[dataset_OJ$dose==0.5,]$len) #calculating mean toothgrowth for dose 0.5 mg
OJ_mean_1.0<- mean(dataset_OJ[dataset_OJ$dose==1.0,]$len) #calculating mean toothgrowth for dose 1.0 mg
OJ_mean_2.0<- mean(dataset_OJ[dataset_OJ$dose==2.0,]$len) #calculating mean toothgrowth for dose 2.0 mg
barplot(height = c(OJ_mean_0.5, OJ_mean_1.0, OJ_mean_2.0), names.arg = c('0.5 mg/day', '1.0 mg/day', '2
```



Comparing mean tooth growth for animals who were given ascorbic acid across the 3 doses: 0.5 mg/day, 1.0 mg/day and 2.0 mg/day

```
VC_mean_0.5<- mean(dataset_VC[dataset_VC$dose==0.5,]$len) #calculating mean toothgrowth for dose 0.5 mg
VC_mean_1.0<- mean(dataset_VC[dataset_VC$dose==1.0,]$len) #calculating mean toothgrowth for dose 1.0 mg
VC_mean_2.0<- mean(dataset_VC[dataset_VC$dose==2.0,]$len) #calculating mean toothgrowth for dose 2.0 mg

par(mfrow= c(1,1), mar=c(5.1, 4.1, 4.1, 2.1))
barplot(height = c(VC_mean_0.5, VC_mean_1.0, VC_mean_2.0), names.arg = c('0.5 mg/day', '1.0 mg/day', '2
```



So we can see from these above analysis clearly that for both cases of orange juice and ascorbic acid, the toothgrowth seems to increase as we increase the doses for each of them. Now lets conduct some tests to validate this observation

Conducting tests

As we saw from the above exploratory analysis, it appears that for both orange juice and ascorbic acid the toothgrowth seems to increase if we increase the doses given to the animals.

So let's test this observation by conducting tests in R.

Tests for the animals who were given Orange juice (OJ).

dose 1.0 mg/day vs 0.5 mg/day (orange juice) **Null hypothesis**: Our null hypothesis will be that the true mean toothgrowth or population mean toothgrowth for animals given the dose 1.0 mg/day of orange juice is equal to that of the animals given 0.5 mg/day of orange juice. In other words, the difference between the true means for dose 1.0 mg/day orange juice and dose 0.5 mg/day orange juice is 0

Altenative hypothesis: Our alternative hypothesis is that the true mean toothgrowth for animals who were given the dose 1.0 mg/day of orange juice, is greater than that of the animals who were given 0.5 mg/day of orange juice. In other words, the difference between the true means for dose 1.0 mg/day orange juice and dose 0.5 mg/day orange juice is greater than 0.

So we are conducting a greater than test. Our alpha level is 5 % or alpha= 0.05

```
t.test(dataset_OJ[dataset_OJ$dose==1.0,]$len, dataset_OJ[dataset_OJ$dose==0.5,]$len, paired = FALSE, al
```

```
##
##  Welch Two Sample t-test
##
## data:  dataset_OJ[dataset_OJ$dose == 1, ]$len and dataset_OJ[dataset_OJ$dose == 0.5, ]$len
## t = 5.0486, df = 17.698, p-value = 4.392e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.214316      Inf
## sample estimates:
## mean of x mean of y
##     22.70     13.23
```

Result of test: so we reject null hypo and accept alternative hypo as the p value is lower than our alpha= 0.05. SO the true mean toothgrowth for animals who were given 1.0 mg/day dose of orange juice is greater than that of animals who were given dose of 0.5 mg/day of orange juice

dose 2.0 mg/day vs 1.0 mg/day (orange juice) **Null hypothesis**: Our null hypothesis will be that the true mean toothgrowth or population mean toothgrowth for animals given the dose 2.0 mg/day of orange juice is equal to that of the animals given 1.0 mg/day of orange juice. In other words, the difference between the true means for dose 2.0 mg/day orange juice and dose 1.0 mg/day orange juice is 0

Altenative hypothesis: Our alternative hypothesis is that the true mean toothgrowth for animals who were given the dose 2.0 mg/day of orange juice, is greater than that of the animals who were given 1.0 mg/day of orange juice. In other words, the difference between the true means for dose 2.0 mg/day orange juice and dose 1.0 mg/day orange juice is greater than 0.

So we are conducting a greater than test. Our alpha level is 5 % or alpha= 0.05

```
t.test(dataset_OJ[dataset_OJ$dose==2.0,]$len, dataset_OJ[dataset_OJ$dose==1.0,]$len, paired = FALSE, al
```

```
##
##  Welch Two Sample t-test
##
## data:  dataset_OJ[dataset_OJ$dose == 2, ]$len and dataset_OJ[dataset_OJ$dose == 1, ]$len
## t = 2.2478, df = 15.842, p-value = 0.0196
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7486236       Inf
## sample estimates:
## mean of x mean of y
##     26.06     22.70
```

Result of test: so we reject null hypo and accept alternative hypo as the p value is lower than our alpha= 0.05. SO the true mean toothgrowth for animals who were given 2.0 mg/day dose of orange juice is greater than that of animals who were given dose of 1.0 mg/day of orange juice

Tests for the animals who were given Ascorbic acid (VC).

dose 1.0 mg/day vs 0.5 mg/day (Ascorbic acid) Null hypothesis: Our null hypothesis will be that the true mean toothgrowth or population mean toothgrowth for animals given the dose 1.0 mg/day of Ascorbic acid is equal to that of the animals given 0.5 mg/day of ascorbic acid. In other words, the difference between the true means for dose 1.0 mg/day Ascorbic acid and dose 0.5 mg/day Ascorbic acid is 0

Altenative hypothesis: Our alternative hypothesis is that the true mean toothgrowth for animals who were given the dose 1.0 mg/day of Ascorbic acid, is greater than that of the animals who were given 0.5 mg/day of Ascorbic acid. In other words, the difference between the true means for dose 1.0 mg/day Ascorbic acid and dose 0.5 mg/day Ascorbic acid is greater than 0.

So we are conducting a greater than test. Our alpha level is 5 % or alpha= 0.05

```
t.test(dataset_VC[dataset_VC$dose==1.0,]$len, dataset_VC[dataset_VC$dose==0.5,]$len, paired = FALSE, al
```

```
##
##  Welch Two Sample t-test
##
## data:  dataset_VC[dataset_VC$dose == 1, ]$len and dataset_VC[dataset_VC$dose == 0.5, ]$len
## t = 7.4634, df = 17.862, p-value = 3.406e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.746867      Inf
## sample estimates:
## mean of x mean of y
##     16.77      7.98
```

Result of test: so we reject null hypo and accept alternative hypo as the p value is lower than our alpha= 0.05. SO the true mean toothgrowth for animals who were given 1.0 mg/day dose of Ascorbic acid is greater than that of animals who were given dose of 0.5 mg/day of Ascorbic acid

dose 2.0 mg/day vs 1.0 mg/day (Ascorbic acid) Null hypothesis: Our null hypothesis will be that the true mean toothgrowth or population mean toothgrowth for animals given the dose 2.0 mg/day of Ascorbic acid is equal to that of the animals given 1.0 mg/day of ascorbic acid. In other words, the difference between the true means for dose 2.0 mg/day Ascorbic acid and dose 1.0 mg/day Ascorbic acid is 0

Altenative hypothesis: Our alternative hypothesis is that the true mean toothgrowth for animals who were given the dose 2.0 mg/day of Ascorbic acid, is greater than that of the animals who were given 1.0 mg/day of Ascorbic acid. In other words, the difference between the true means for dose 2.0 mg/day Ascorbic acid and dose 1.0 mg/day Ascorbic acid is greater than 0.

So we are conducting a greater than test. Our alpha level is 5 % or alpha= 0.05

```
t.test(dataset_VC[dataset_VC$dose==2.0,]$len, dataset_VC[dataset_VC$dose==1.0,]$len, paired = FALSE, al
```

```
##
##  Welch Two Sample t-test
##
## data:  dataset_VC[dataset_VC$dose == 2, ]$len and dataset_VC[dataset_VC$dose == 1, ]$len
## t = 5.4698, df = 13.6, p-value = 4.578e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
```

```
##  6.346525      Inf
## sample estimates:
## mean of x mean of y
##     26.14     16.77
```

Result of test: so we reject null hypo and accept alternative hypo as the p value is lower than our alpha= 0.05. SO the true mean toothgrowth for animals who were given 2.0 mg/day dose of Ascorbic acid is greater than that of animals who were given dose of 1.0 mg/day of Ascorbic acid

Conclusion:

So as observed from the exploratory analysis and validated with our tests, we can conclude that for both cases of orange juice and ascorbic acid, the toothgrowth increases as we increase the doses for each of the 2 supplements.

Assumptions used: 1. we assumed the distributions were not paired 2. we assumed alpha level to be 5% or 0.05 3.We used greater than test as per our assumption that mean or average toothgrowth for a group of animals given a higher dose of either of the 2 supplements: ascorbic acid or orange juice, is greater than that of a group given a lower dose of the same supplement. We formed this assumption based on our observations from our exploratory graphs and analysis.