Part 1 - Overview

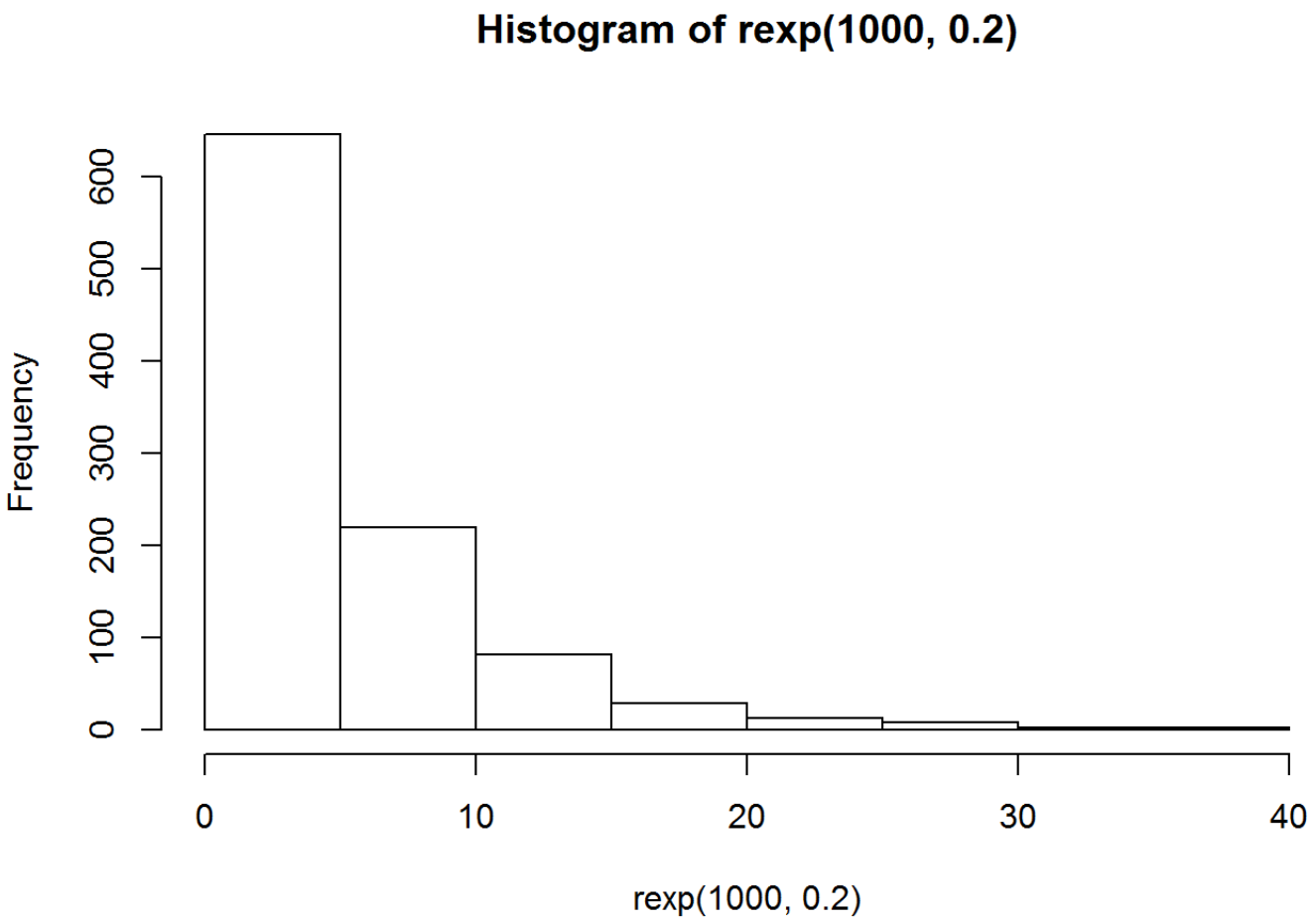
In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

1. Show the sample mean and compare it to the theoretical mean of the distribution.

Simulations

Here is an exploratory graph of the 1000 simulations:

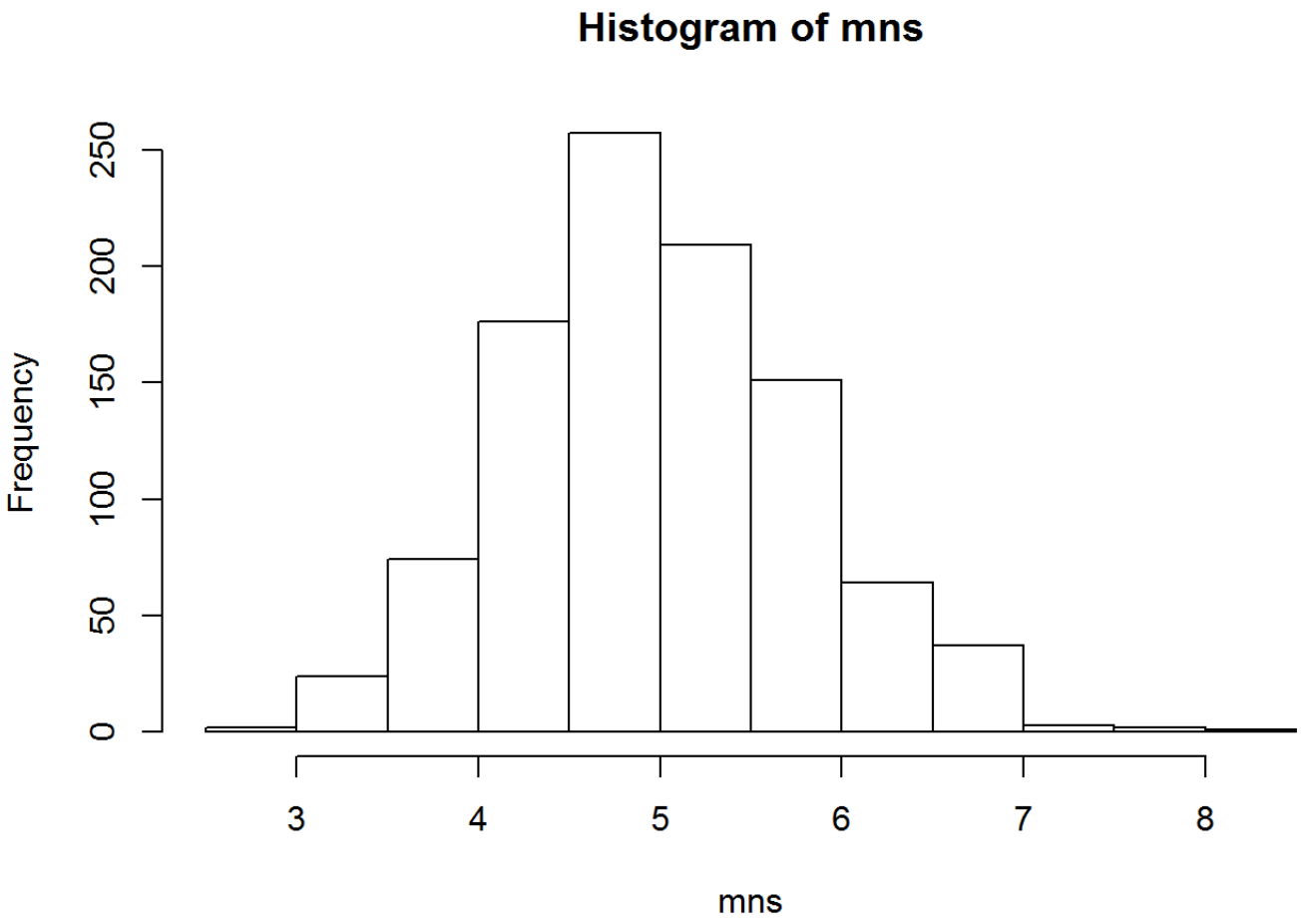
```
hist(rexp(1000,.2))
```



This follows the general expected shape of an exponential distribution.

Now, let's compare this to a plot of the means of the 40 exponentials run 1000 times to see if the resuls approaches the normal distribution as described in the inital project guidelines above.

```
mns=NULL
for (i in 1:1000) mns=c(mns,mean(rexp(40,.2)))
hist(mns)
```



We see that with 1000 simulations of 40 means, the shape of the curve is approaching that of the normal distribution.

Sample Mean versus Theoretical Mean

The sample mean of the simulations is mean(mns) which gives:

```
mean(mns)
```

```
## [1] 4.995778
```

This is very close to the theoretical mean of the distibution which is: 1/lambda =1/.2 =5

2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

For our simulation, the variance is given by:

```
var(rexp(1000,.2))
```

```
## [1] 25.14632
```

Compare this to the theoretical value: (1/lambda^2) = (1/.2^2) = 25

For comparison, taking more samples brings the sample variance closer to the theoretical variance. Compare these 10000 and 100000 below:

```
var(rexp(10000,.2))
```

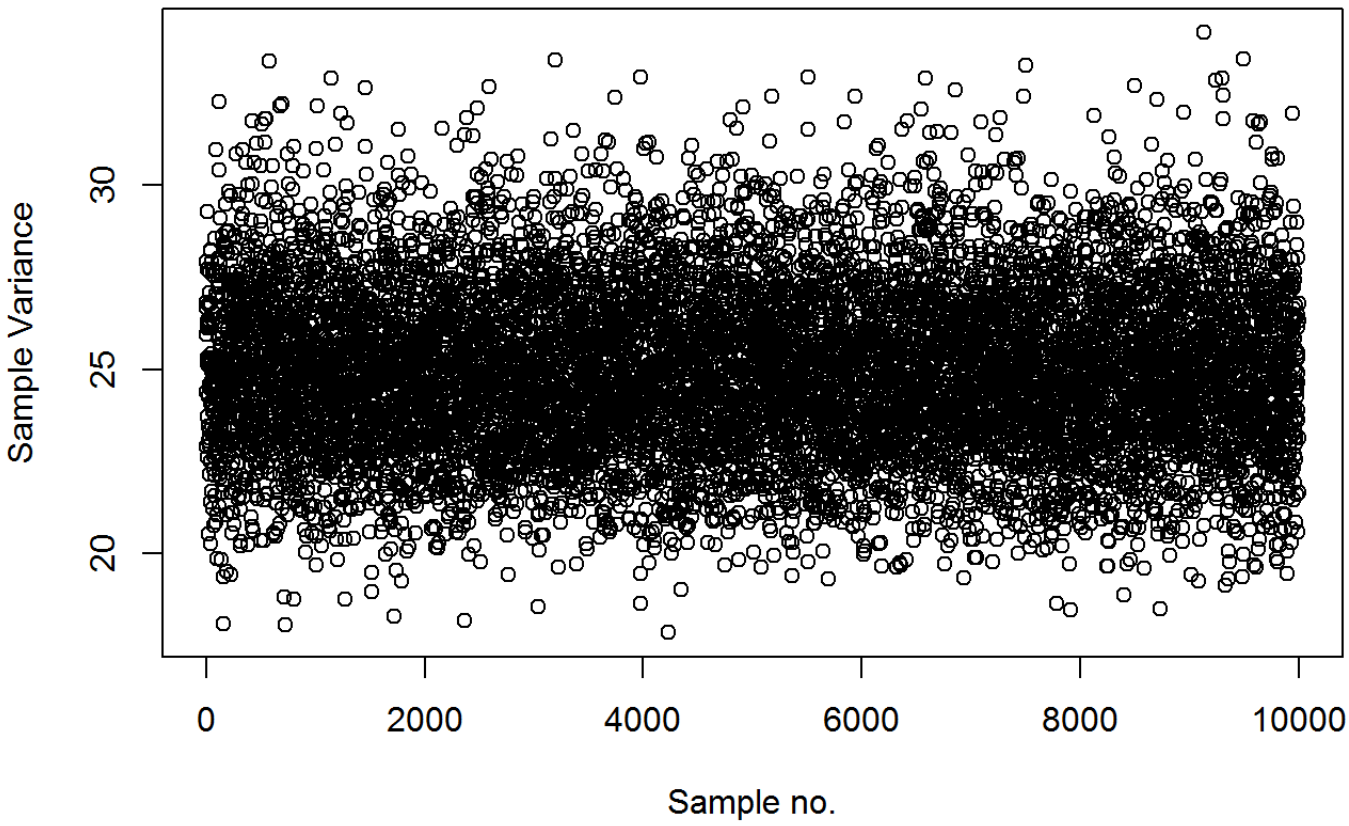
```
## [1] 25.66604
```

```
var(rexp(100000,.2))
```

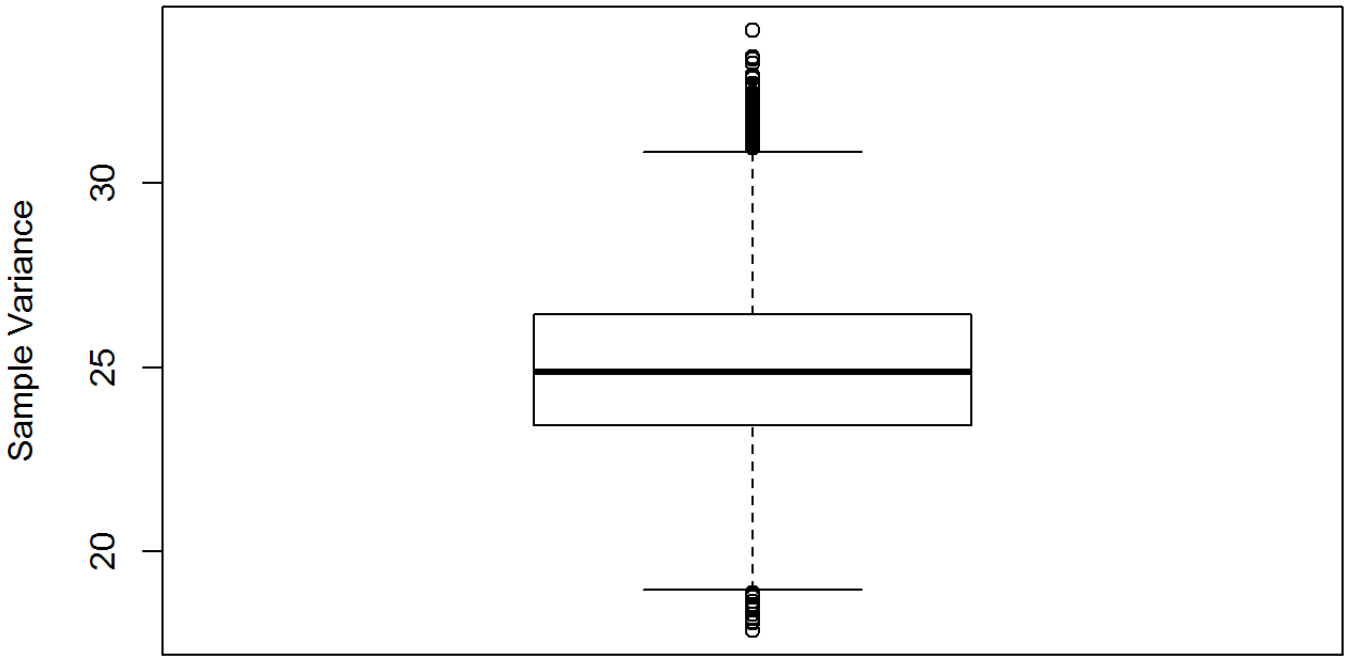
```
## [1] 24.85579
```

The results approaches the theoretical mean of 25 as the number of samples increases. We can plot this:

```
sample.vars=NULL
for (i in 1:10000) sample.vars=c(sample.vars,var(rexp(1000,.2)))
plot(sample.vars,xlab="Sample no.",ylab="Sample Variance")
```




```
boxplot(sample.vars,ylab="Sample Variance")
```

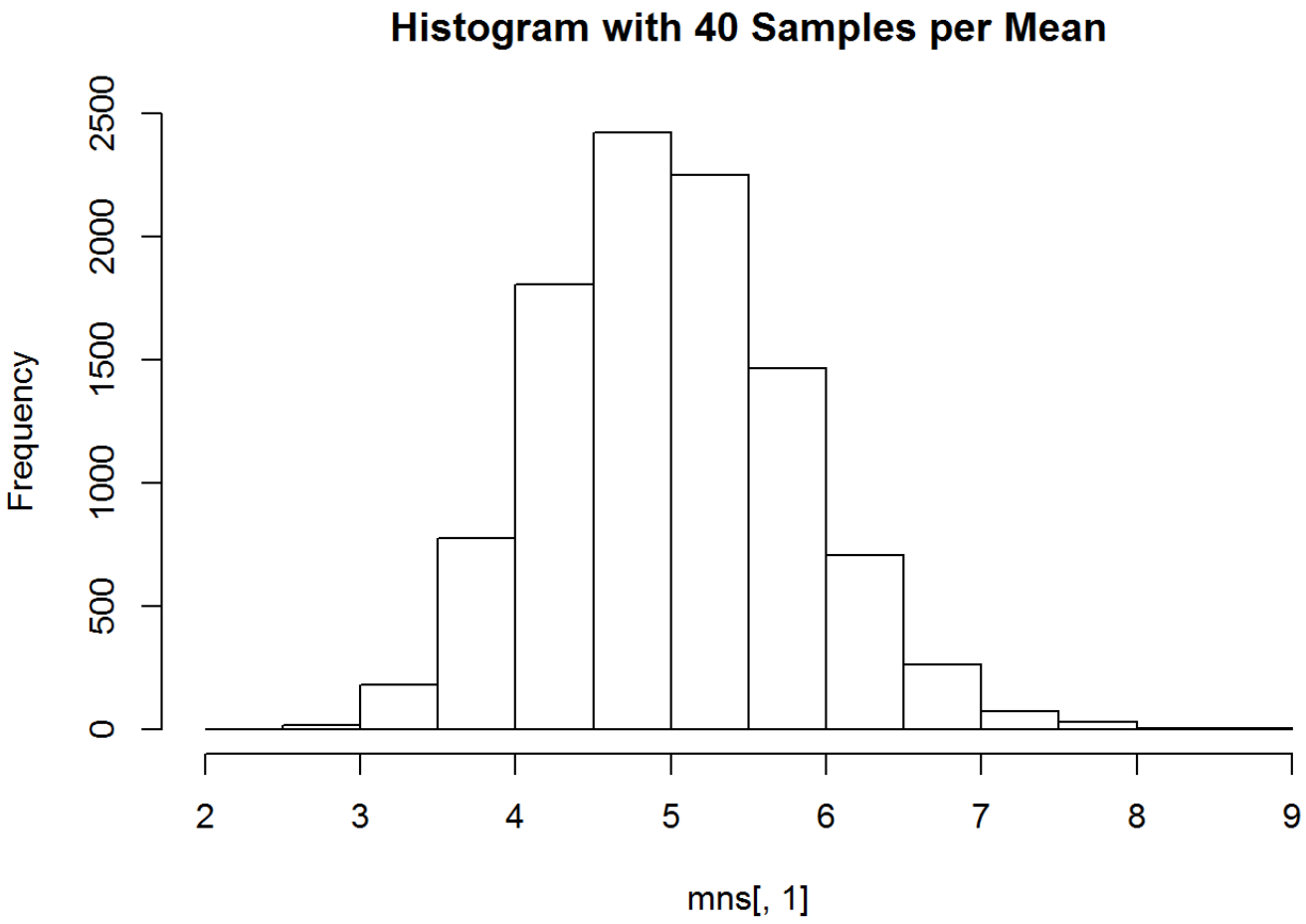


It can be observed that with many iterations, the variance clusters around the theoretical variance of 25 calculated further above.

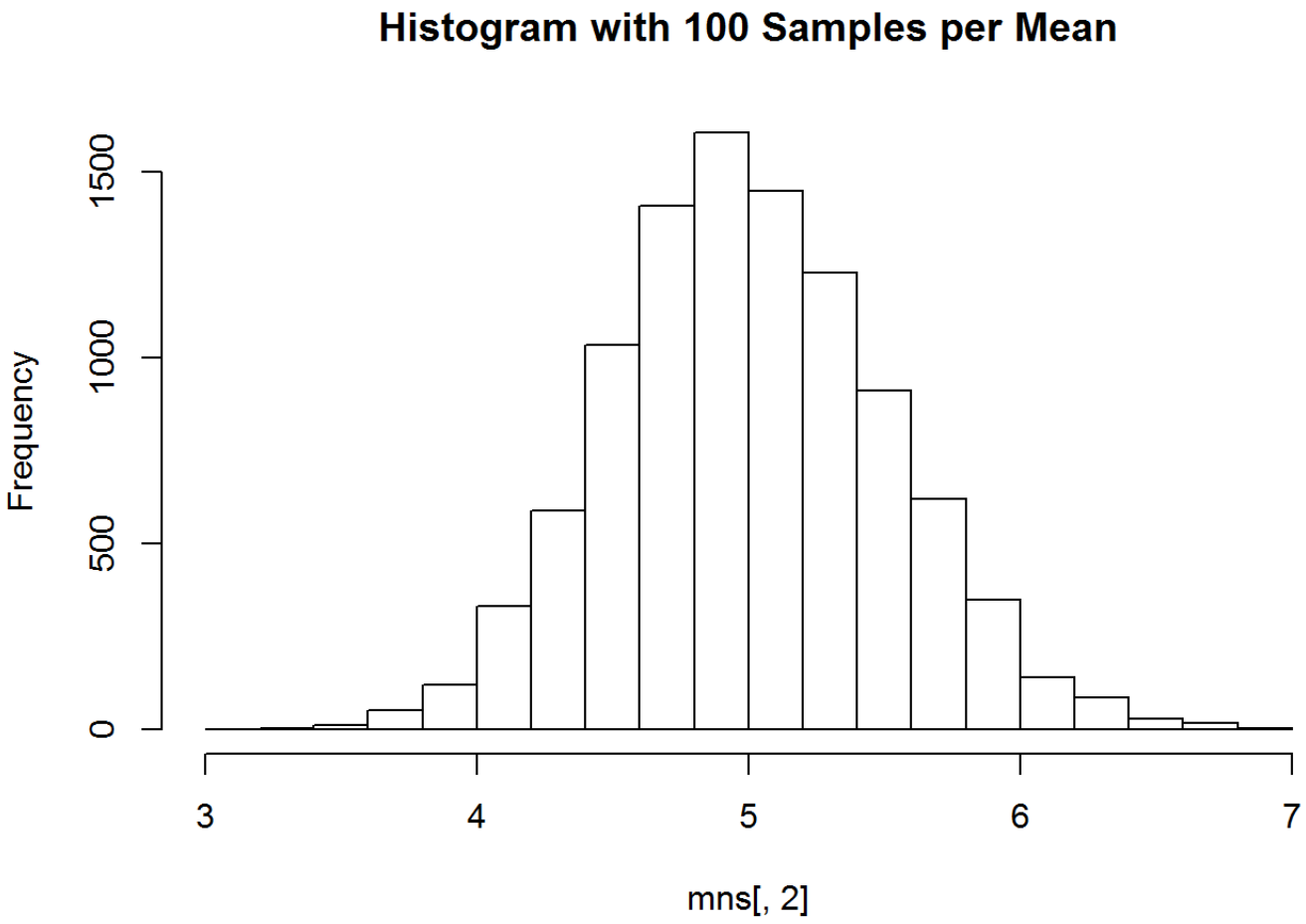
3. Show that the distribution is approximately normal.

If we increase the number of samples used in taking the average mean, it can be seen that the distribution approaches an increasingly normal distribution shape. Here, we increase the number of samples in the mean calculations from 40 to 100 to 1000:

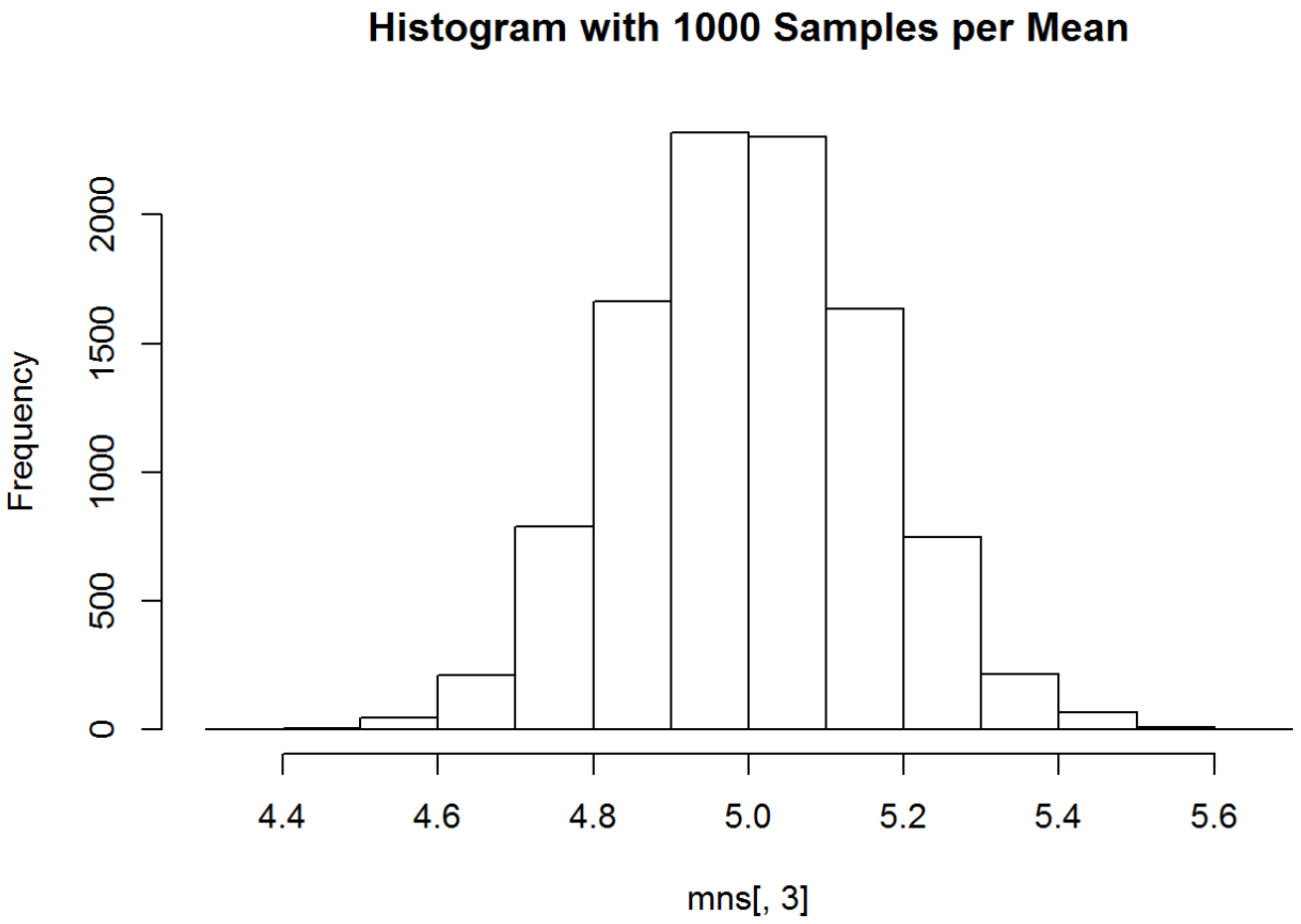
```
mns=NULL
mns2=NULL
mns3=NULL

for (i in 1:10000) mns=c(mns,mean(rexp(40,.2)))
for (i in 1:10000) mns2=c(mns2,mean(rexp(100,.2)))
for (i in 1:10000) mns3=c(mns3,mean(rexp(1000,.2)))

mns=cbind(mns,mns2,mns3)

hist(mns[,1],main="Histogram with 40 Samples per Mean")
```




```
hist(mns[,2],main="Histogram with 100 Samples per Mean")
```




```
hist(mns[,3],main="Histogram with 1000 Samples per Mean")
```



Part 2 - Overview

This portion of the project performs some basic analyses of the ToothGrowth dataset included with R. Here is the description of the data set from the R documentation:

Description The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

Format A data frame with 60 observations on 3 variables.

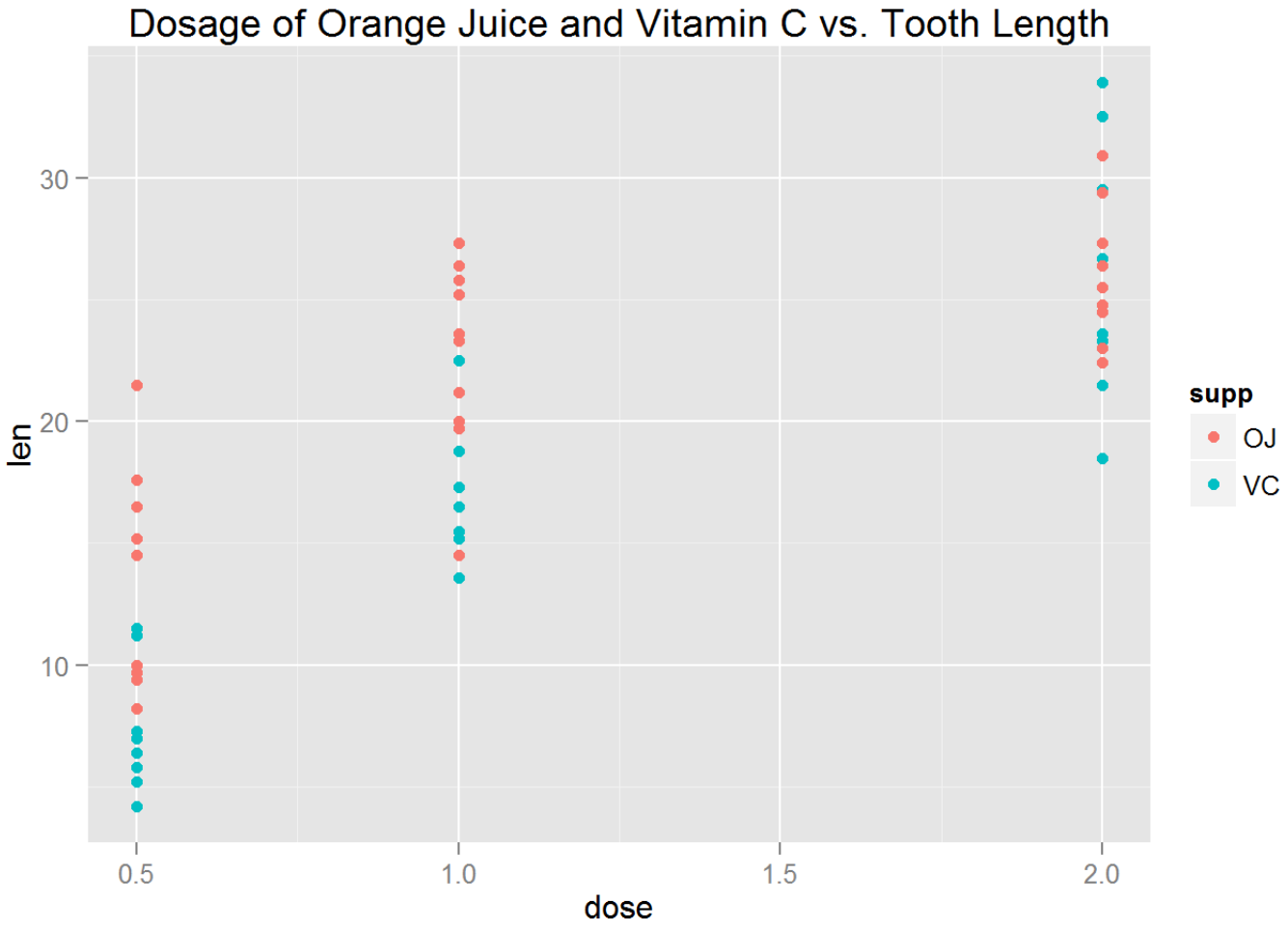
[,1] len numeric Tooth length [,2] supp factor Supplement type (VC or OJ). [,3] dose numeric Dose in milligrams.

Source C. I. Bliss (1952) The Statistics of Bioassay. Academic Press.

Exploratory data analysis

Let's quickly plot the data for the Vitamin C vs Orange Juice to see an overview of the data:

```
library(ggplot2)
qplot(data=ToothGrowth,x=dose,y=len,color=supp,main="Dosage of Orange Juice and Vitamin C vs. Tooth Length")
```



Data preparation

To simplify computations, the data was broken into 6 sets:

```
toothvc.5<-ToothGrowth[1:10,]
toothvc.1<-ToothGrowth[11:20,]
toothvc.2<-ToothGrowth[21:30,]

toothoj.5<-ToothGrowth[31:40,]
toothoj.1<-ToothGrowth[41:50,]
toothoj.2<-ToothGrowth[51:60,]
```

Confidence intervals

Let's compare confidence intervals for the results using the unequal variances method at each dosage to see if we can conclude which supplement is better according to the data.

Dosage analyses Let's take the mean and standard deviations of the 0.5 dosages:

```
ybar<-mean(toothoj.5$len)
vary<-var(toothoj.5$len)
ny<-10

xbar<-mean(toothvc.5$len)
varx<-var(toothvc.5$len)
nx<-10

df<-(varx/nx + vary/ny)^2/((varx/nx)^2/(nx-1)+(vary/ny)^2/(ny-1))
ybar-xbar*c(-1,1)*qt(0.975,df)*sqrt(varx/nx+vary/ny)
```

```
## [1]  41.40692 -14.94692
```

At 0.5 mg, since the confidence interval contains zero, Vitamin C works better.

Let's perform the same tests for 1 mg and 2 mg:

```
ybar<-mean(toothoj.1$len)
vary<-var(toothoj.1$len)
ny<-10

xbar<-mean(toothvc.1$len)
varx<-var(toothvc.1$len)
nx<-10

df<-(varx/nx + vary/ny)^2/((varx/nx)^2/(nx-1)+(vary/ny)^2/(ny-1))
ybar-xbar*c(-1,1)*qt(0.975,df)*sqrt(varx/nx+vary/ny)
```

```
## [1]  75.15407 -29.75407
```

At 1 mg, since the confidence interval contains zero, Vitamin C works better.

```
ybar<-mean(toothoj.2$len)
vary<-var(toothoj.2$len)
ny<-10

xbar<-mean(toothvc.2$len)
varx<-var(toothvc.2$len)
nx<-10

df<-(varx/nx + vary/ny)^2/((varx/nx)^2/(nx-1)+(vary/ny)^2/(ny-1))
ybar-xbar*c(-1,1)*qt(0.975,df)*sqrt(varx/nx+vary/ny)
```

```
## [1] 123.25036 -71.13036
```

At 2 mg, since the confidence interval contains zero, Vitamin C works better.

Conclusion

These results do not appear to be supported by the initial exploratory plot which suggests that the orange juice works better at 0.5 and 1 mg and Vitamin C at 2 mg. Unfortunately, due to time constraints the author could not explore this further.