Title: statistical_inference_project_part1

Author: Derek Chang

Overview

This is a project for statistical inference class on Coursera. There will two parts in this project: 1. A simulation exercise 2. Basic inferential data analysis. The first part is going to be reported on this PDF file.

Simulation Exercise

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

```
# load required package
library(ggplot2)
# set rate parameter to be 0.2
lambda = 0.2
# sample number 40
n = 40
# simulation times 1000
simulation = 1000
set.seed(123)
# run the simulation 1000 times and compute the mean of each 40 exponentials
simulated_exponents = replicate(simulation,rexp(n,lambda))
averages = apply(simulated_exponents,2,mean)
# compute the average of the distribution of the mean of 40 exponentials
sample_mean = mean(averages);sample_mean
```

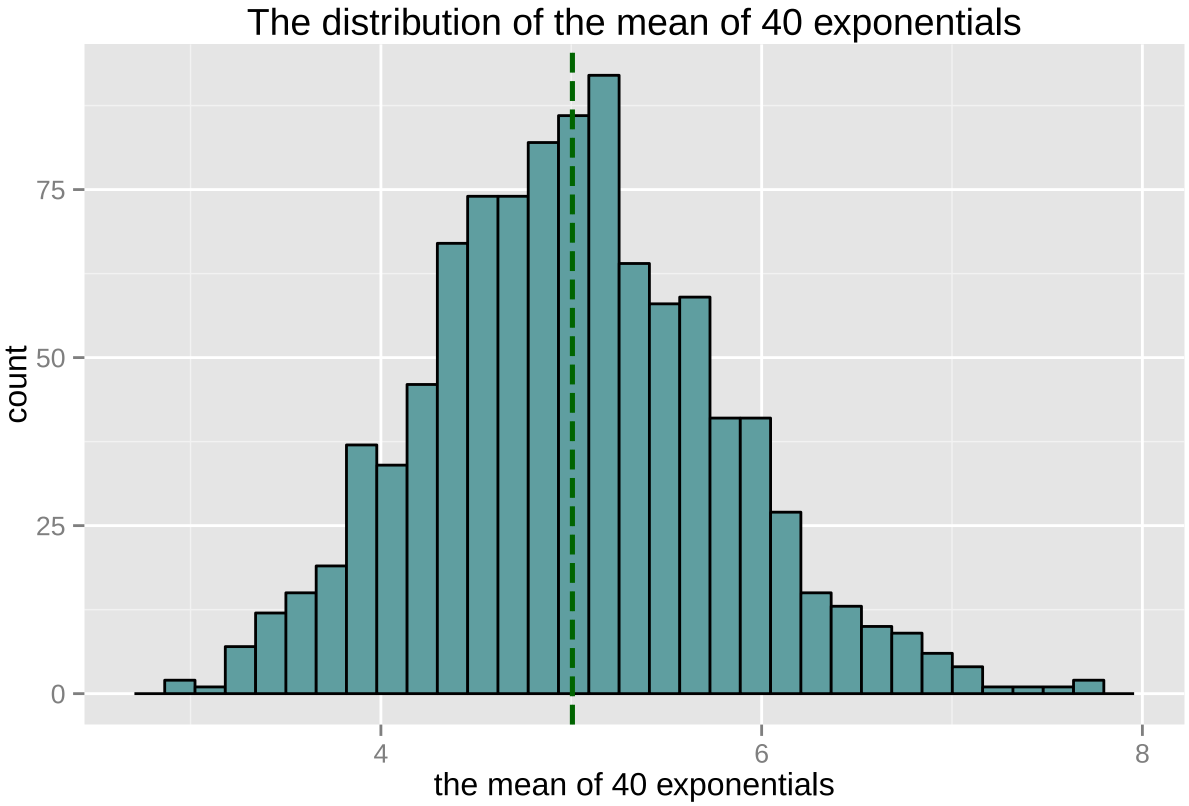
```
## [1] 5.011911
```

```
# compute theretical mean of the distribution
theoretical_mean = 1 / lambda;theoretical_mean
```

```
## [1] 5
```

```
# plot the distribution of the mean of 40 exponentials
p1 = ggplot(as.data.frame(averages),aes(x=averages))+
  geom_histogram(fill="cadetblue",colour="black")+
  geom_vline(xintercept = c(sample_mean,theoretical_mean),colour="darkgreen",linetype="longdash")+
labs(x="the mean of 40 exponentials",title="The distribution of the mean of 40 exponentials")
p1
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



In the plot, we can see that theoretical mean (5) and sample mean (5.011911) are very close to each other.

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

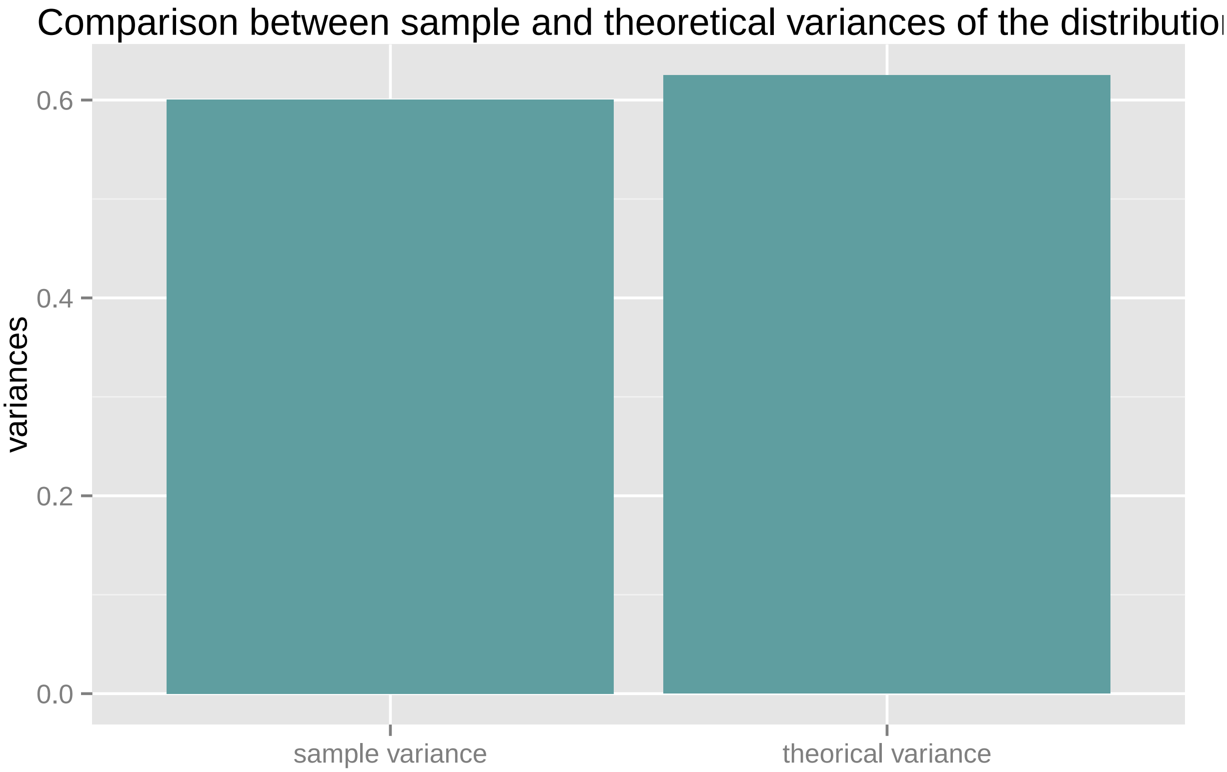
```
# sample variance and standard deviation
sample_variance = sd(averages)^2;sample_variance
```

```
## [1] 0.6004928
```

```
sample_std = sqrt(sample_variance);
# theoretical variance
theoretical_variance = (1/lambda)^2/n;theoretical_variance
```

```
## [1] 0.625
```

```
# plot for comparing the sample variance to the theoretical variance of the distribution
variance_data = data.frame(variances=c(theoretical_variance,sample_variance))
rownames(variance_data) = c("theorical variance","sample variance")
p2 = ggplot(variance_data,aes(x=rownames(variance_data),y=variances))+
  geom_bar(stat="identity",fill="cadetblue")+
  labs(x="",title="Comparison between sample and theoretical variances of the distribution")
p2
```

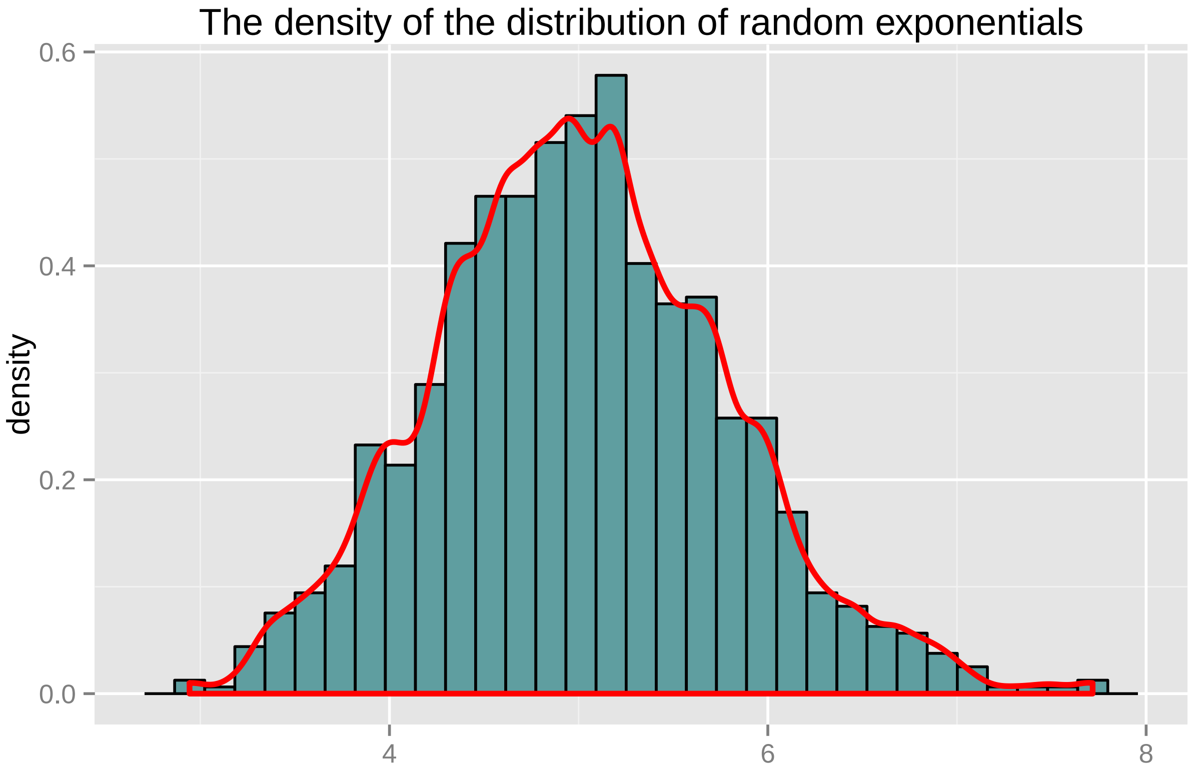


In this figure, sample variance (0.6004928) and theoretical variance (0.625) are very close.

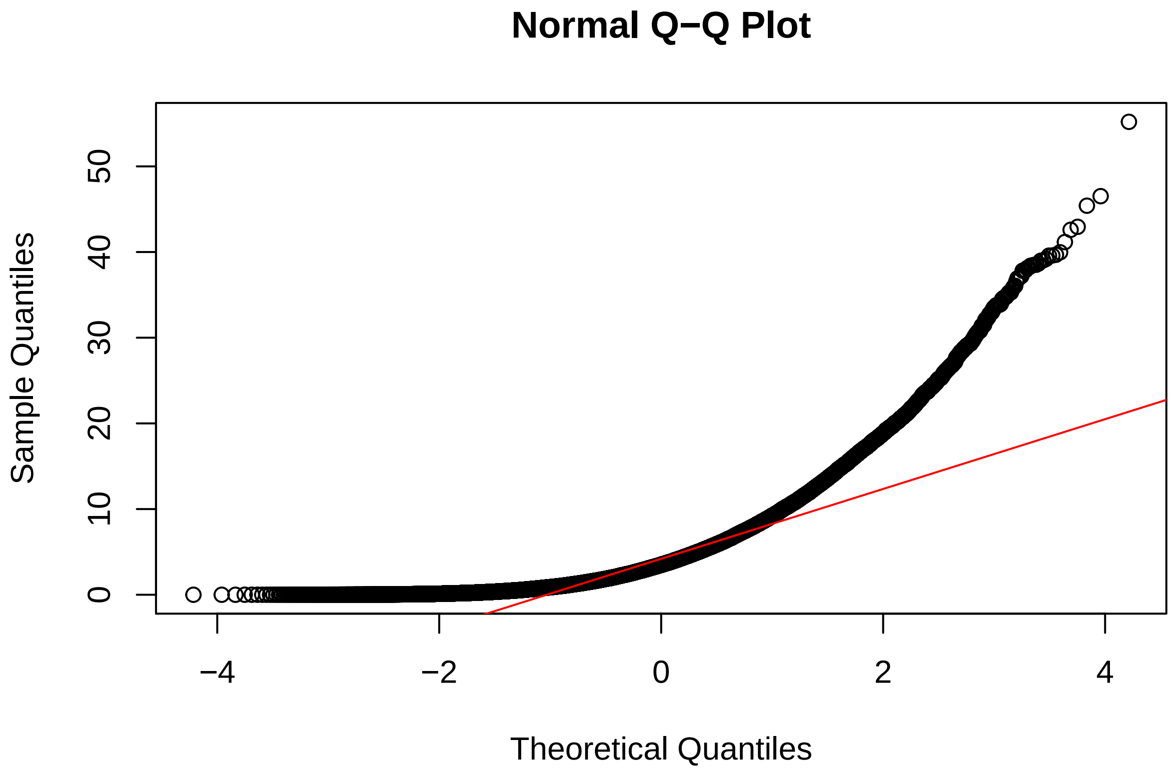
Show that the distribution is approximately normal

```
p3 = ggplot(as.data.frame(as.numeric(simulated_exponents)),aes(x=averages))+
  geom_histogram(aes(y=..density..), fill="cadetblue",colour="black")+
  geom_density(colour="red",size=1)+
  labs(x="",title="The density of the distribution of random exponentials")
p3
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

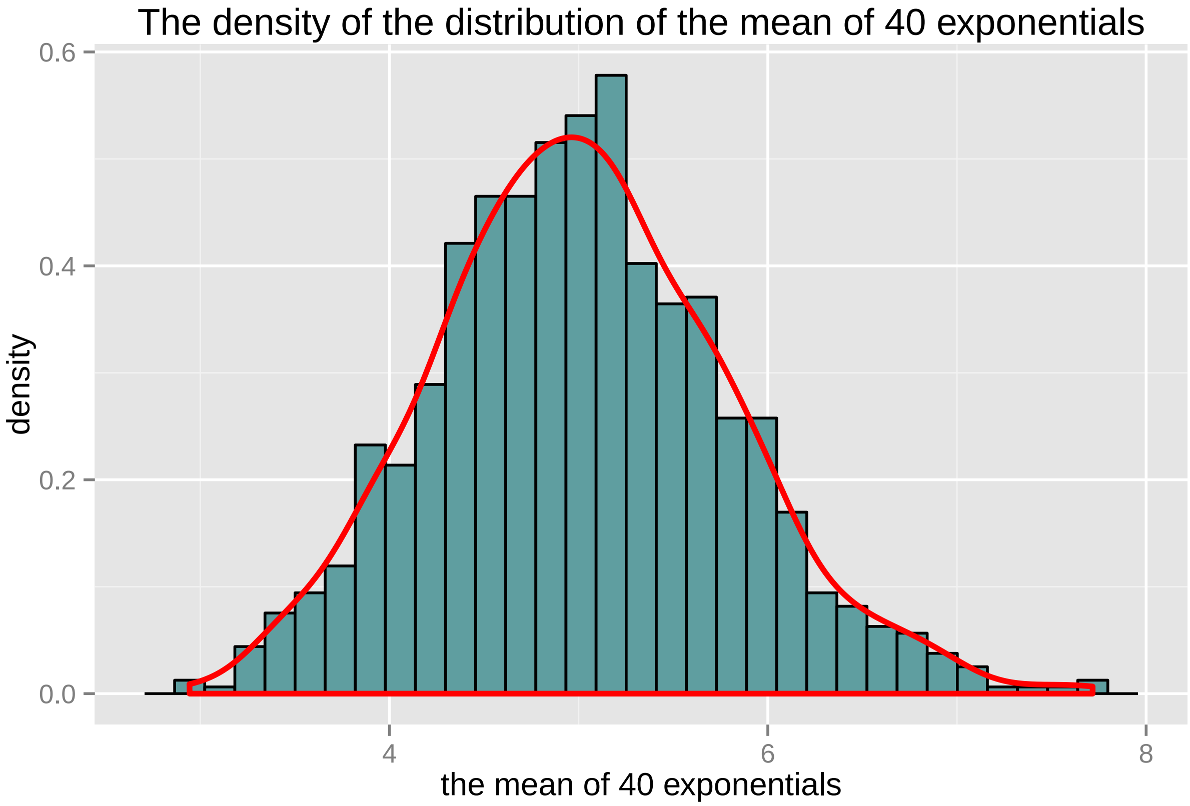


```
qqnorm(simulated_exponents);qqline(simulated_exponents,col=2)
```

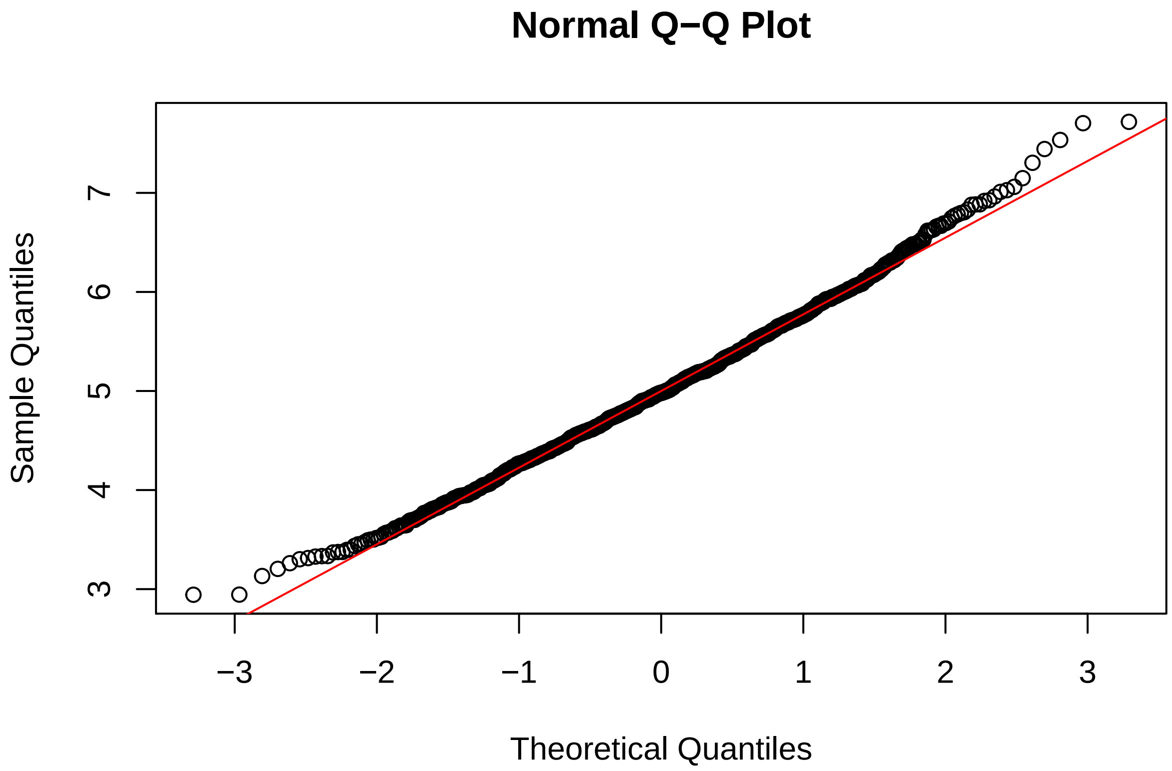


```
p4 = ggplot(as.data.frame(averages),aes(x=averages))+
  geom_histogram(aes(y=..density..), fill="cadetblue",colour="black")+
  geom_density(colour="red",size=1)+
  labs(x="the mean of 40 exponentials",title="The density of the distribution of the mean of 40 exponen
p4
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



```
qqnorm(averages);qqline(averages,col=2)
```



From qqplot and the density of the distribution, the distribution of the mean of 40 exponentials is more like a normal distribution than that of a large collection of random exponentials.