Coursera Statistical Inference Project Part 1 - Check CLT with Simulation of Exponential distribution

Zhenkun Guo

February 27, 2016

1. Project Introduction

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 also 1/lambda. To investigate the distribution and check its aggrement with centrel limit theory, we will run 1000 simulations. In each simulation, 40 samples will be drawn from the exponential distribution with lambda=0.2.

2. Simulate Exponential Distribution

2.1. Load Necessary Library and Set Parameters

We load ggplot2 libary for drawing graphs. Number of simulations is 1000, each simuluation contains 40 samples. The lambda in the exponential distribution is set to be 0.2. A random seed 9527 is assigned to guarantee reproducibility.

```
library(ggplot2)
set.seed(9527)
nsamples<-40
nsim<-1000
lambda<-0.2
```

2.2. Run the Simulation

Results of the simulation are stored in a data frame

```
sim_data<-matrix(data=0,nrow=nsim,ncol=nsamples)
for (i in 1:nsim)
        sim_data[i,]<-rexp(nsamples,lambda)
exp_mean<-apply(sim_data,1,mean)
```

3. Compare Simulated Parameters with Theoretical

3.1. Mean

```
sample_mean<-mean(exp_mean)
theoretical_mean<-1/lambda
mean_text<-paste(paste("Simulated Mean is",round(sample_mean,3)),paste("Comparing to Theoretical Mean",t
mean_text
```

```
## [1] "Simulated Mean is 5.025 Comparing to Theoretical Mean 5"
```

The means of simulation and thoery are pretty close.

3.2. Variance

```
sample_sd<-sd(exp_mean)
theoretical_sd<-1/lambda/sqrt(nsamples)
mean_var<-paste(paste("Simulated Var is",round(sample_sd^2,3)),paste("Comparing to Theoretical Var",rou
mean_var
```

```
## [1] "Simulated Var is 0.66 Comparing to Theoretical Var 0.625"
```

The variances of simulation and thoery are pretty close.

3.3. Double-Sided 95% Confidence Interval

```
conf_sim<-sample_mean+c(-1,1)*qnorm(0.975,0,sample_sd)
conf_the<-theoretical_mean+c(-1,1)*qnorm(0.975,0,theoretical_sd)
conf_text_low<-paste(paste("Simulated 95% Confidence Interval Lower Bound is",round(conf_sim,3)[1]),
                 paste("Comparing to Theoretical 95% Confidence Interval Lower Bound",round(conf_the,3)
conf_text_high<-paste(paste("Simulated 95% Confidence Interval Upper Bound is",round(conf_sim,3)[2]),
                      paste("Comparing to Theoretical 95% Confidence Interval Upper Bound",round(conf_th
conf_text_low
conf_text_high
```

```
## [1] "Simulated 95% Confidence Interval Lower Bound is 3.433 Comparing to Theoretical 95% Confidence I
## [1] "Simulated 95% Confidence Interval Upper Bound is 6.617 Comparing to Theoretical 95% Confidence I
```

The 95% confidence intervals of simulation and thoery are pretty close.

4. Compare the Distrbution of Sample Means to Theoretical Normal Distribution

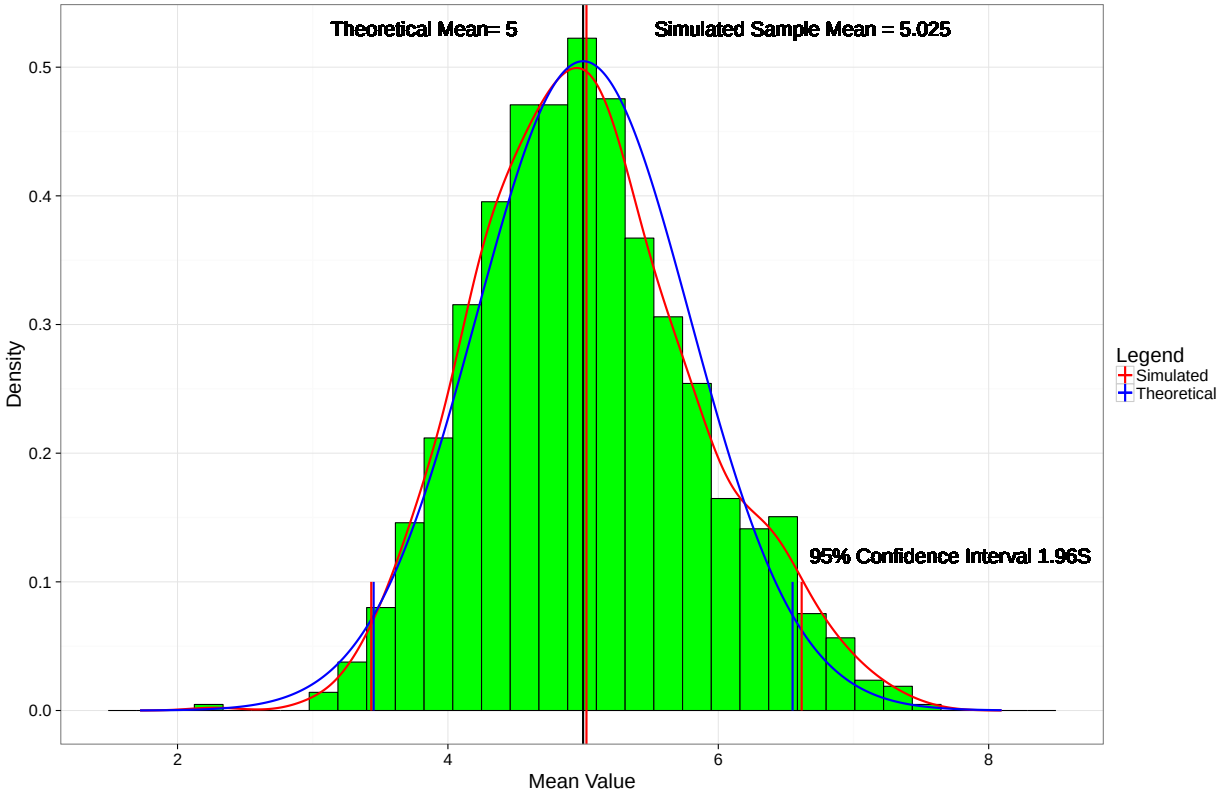
4.1. Prepare Data for Plots

To compare the simulation and theory, density function of the simulation results needs to be calculated. Also a normal distribution function is calculated. These data are stored in a data frame to be plotted.

```
density_f<-density(exp_mean)
sample <- seq(min(density_f$x), max(density_f$x), length=length(exp_mean))
sim_density<-approx(density_f$x,density_f$y,sample)
theoretical_density <- dnorm(sample, mean=1/lambda, sd=(1/lambda/sqrt(nsamples)))
plotdata<-data.frame(sim=exp_mean)
plotdata$x<-sample
plotdata$sim_y<-sim_density$y
plotdata$the_y<-theoretical_density
```

4.2. Making the Plot with ggplot2

```
figure<-ggplot(plotdata)
figure<-figure+geom_histogram(aes(x=sim,y=..density..), bins=30, colour="black",fill = "green")
figure<-figure+geom_vline(aes(xintercept = mean(sim),colour="sim"),size=1,linetype=1)
figure<-figure+geom_text(aes(x=mean(sim)+1.6,label=paste("Simulated Sample Mean =",round(mean(sim),3)))
figure<-figure+geom_vline(xintercept = 5, aes(colour="the"),size=1,linetype=1)
figure<-figure+geom_text(aes(x=mean(sim)-1.2),label="Theoretical Mean= 5",y=0.53,size=7)
figure<-figure+geom_line(aes(x=x,y=sim_y,color='sim'),linetype=1,size=1)
figure<-figure+geom_line(aes(x=x,y=the_y,color='the'),linetype=1,size=1)
figure<-figure+scale_colour_manual(name="Legend",values=c("sim"="red", "the"="blue"),labels=c("Simulate
figure<-figure+theme_bw(base_size = 20)+xlab("Mean Value")+ylab('Density')
figure<-figure+annotate("segment", x = conf_the[1], xend = conf_the[1], y = 0, yend = 0.1,
                        colour = "blue",size=1)
figure<-figure+annotate("segment", x = conf_the[2], xend = conf_the[2], y = 0, yend = 0.1,
                        colour = "blue",size=1)
figure<-figure+annotate("segment", x = conf_sim[1], xend = conf_sim[1], y = 0, yend = 0.1,
                        colour = "red",size=1)
figure<-figure+annotate("segment", x = conf_sim[2], xend = conf_sim[2], y = 0, yend = 0.1,
                        colour = "red",size=1)
figure<-figure+geom_text(x = conf_sim[2]+1.1,y=0.12,size=7,label="95% Confidence Interval 1.96S")
print(figure)
```

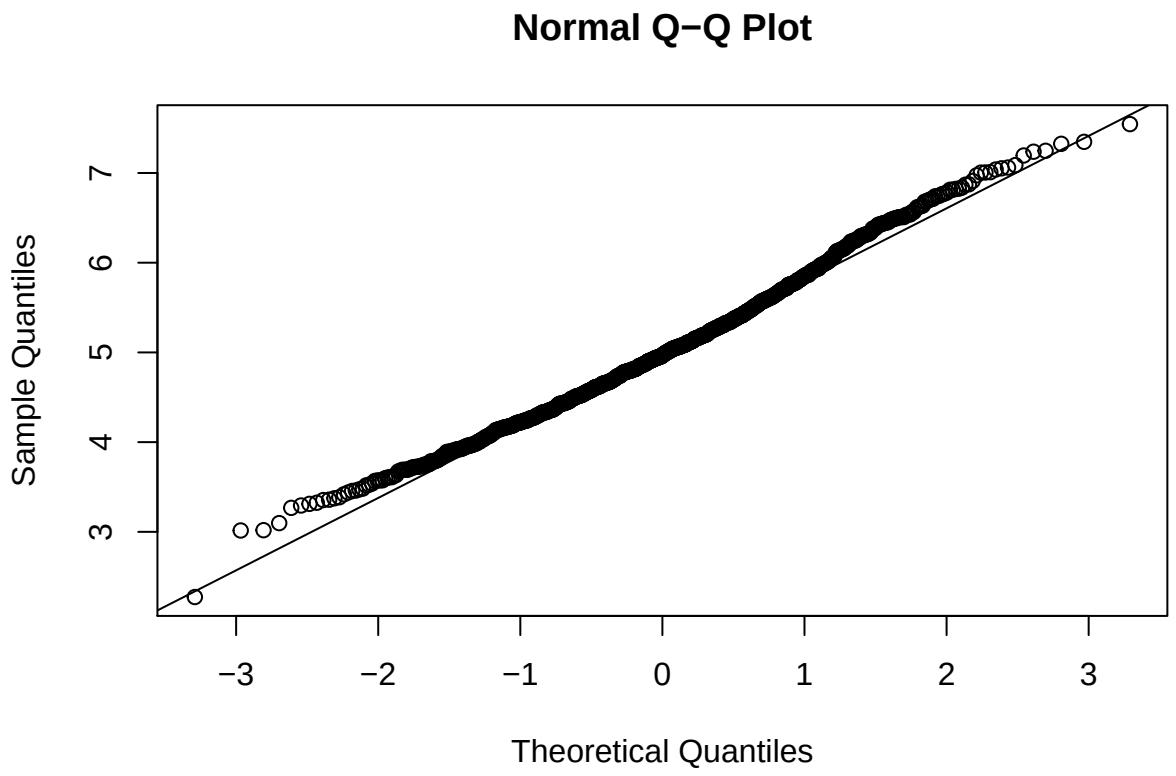


By Comparing the red simulated line and the blue theoretical normal distribution line, we can see the distribution of means is pretty close to normal distribution. Also the 95% confidence intervals for both simulation and theory are very close.

4.3. Comparing the Simulation to Theory through QQ plot

A qq plot is a good way to evaluate the normalarity of the simulation

```
qqnorm(exp_mean)
qqline(exp_mean)
```



QQ plot also shows the similarity between the simulated distribution and the theoretical normal distribution