Statistical Inference Course Project - Part I

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. In this part, lambda is set as 0.2 for all of the simulations. The distribution of averages of 40 exponential(0.2)s will be investigated with 1000 simulations.

Show where the distribution is centered at and compare it to the theoretical center of the distribution

The simulation starts from creating 40 iid exponential distributions with lambda set as 0.2, length sets as 1000.

A simple implementation can be with replicate() function as like: replicate(40,rexp(1000,0.2)). This will create a 1000 rows 40 columns matrix. Every column is independent exponential distribution with lambda as 0.2.

With the rowMeans() function,the averages of these 40 iid exponential random sequence can be got.

The cumulative mean of this average distribution can be got by: cumsum(rmeans)/(1:1000). This will output the sequence of sample mean when the sample size increase from 1 to 1000. In one of a simulation, the head of sequence shows:

```
5.962585 5.135810 5.265281 4.908910 5.130389 5.092576
```

The tail of sequence is:

```
5.000642 4.999475 4.998772 5.000174 5.000207 4.999696
```

Comparing above results, we can see that with the samples number increasing, the average is centered to the theoretical center o distribution which is 5.

Show how variable it is and compare it to the theoretical variance of the distribution

Firstly, we need determine the theoretical variance of the distribution. Every sample point is the average of 40 exponential random as lambda =0.2. So the theoretical varianc of sample random is: 0.625

The variance simulation is similar to mean simulation. It also with the cumulative approach, which is produce a variance sequence of first n samples, where n is from 1 to 1000. Also similiar to mean simulation, we can compre the head and tail of this sequence:

Head: 0.0000000 1.3671152 0.7338457 0.9972314 0.9931886 0.8031297

Tail: 0.6110133 0.6117555 0.6116343 0.6129833 0.6123701 0.6120182

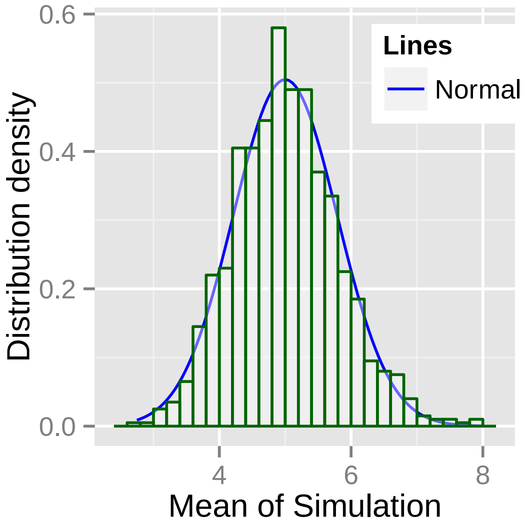
We can see that the value in tail of the sequence is much close to the theoretical variance.

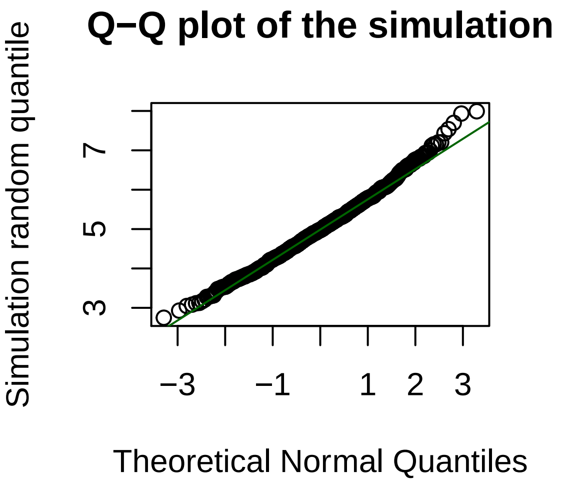
Show that the distribution is approximately normal

Firstly, let's determine the parameters of the normal distribution.From previous sections, we know the theoretical mean and variacn of the simlution is: 5 and 0.625. This mean and variance are also the mean and variance of the apporximation normal distribution.

Then we can use two approaches to make comparison. The first pproach is by comparing the simulation distribution'shist graph and the theoretical normal distribution. The second approach is compare the quantile of the simulation with the quantile of the normal distribution with Q-Q plot.

The comparison results are below:





From both above distribution graph and Q-Q plot, we can see that the simulation distribution is approximating to normal distribution.

Evaluate the coverage of the confidence interval for 1/lambda: $\hat{X} \pm 1.96S/\sqrt{n}$

The confidence interval for 1/lambda: $\hat{X} \pm 1.96S/\sqrt{n}$ corresponds to the 95% confidence interval.

The coverages simulation is based on the above cumulative mean and variance simulations. From these two sequence, we can get every smaple point's 95% confidence interval. Now we can check whether the theoretical mean which 1/lambda=5 is fall in this interval. The final step is calculating the cumulative rate of theoretical value falls into the simulated confidence interval. This gives us a 1000 values sequence, the tail of it in one simulation as:

```
0.9326633 0.9327309 0.9327984 0.9328657 0.9329329 0.9330000
```

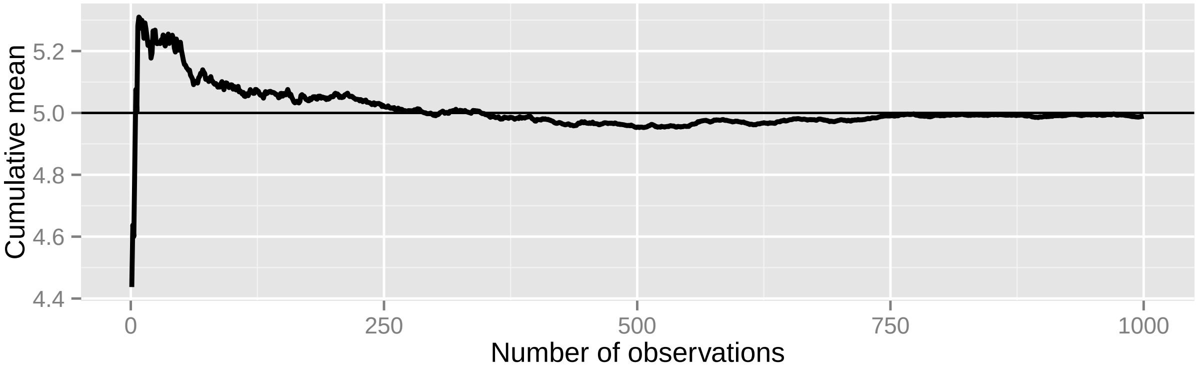
The last value 0.933 is the couverage of 1000 samples. This is below the theoretical value 0.95 for the smaple size is just 40, not big enough. The more detail discussion for this point is showed in the appendix.

Appendix

This appendix is show simulation code and more detail simulation result.

Show where the distribution is centered

```
rm(list=ls()) # remove all objects
sizes<-40; n<-1000;lambda<-0.2
## creating 40 iid exponential distributions with n=1000,lambda=0.2
rexpArr<-replicate(sizes,rexp(n,lambda))
## averaging 40 exponential distribution
rexpMean<-rowMeans(rexpArr)
cumsumMean <- cumsum(rexpMean)/(1:n)   ## cumulative mean
library(ggplot2)
g <- ggplot(data.frame(x = 1:n, y = cumsumMean), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 1/lambda) + geom_line(size = 1)
g + labs(x = "Number of observations", y = "Cumulative mean")
```



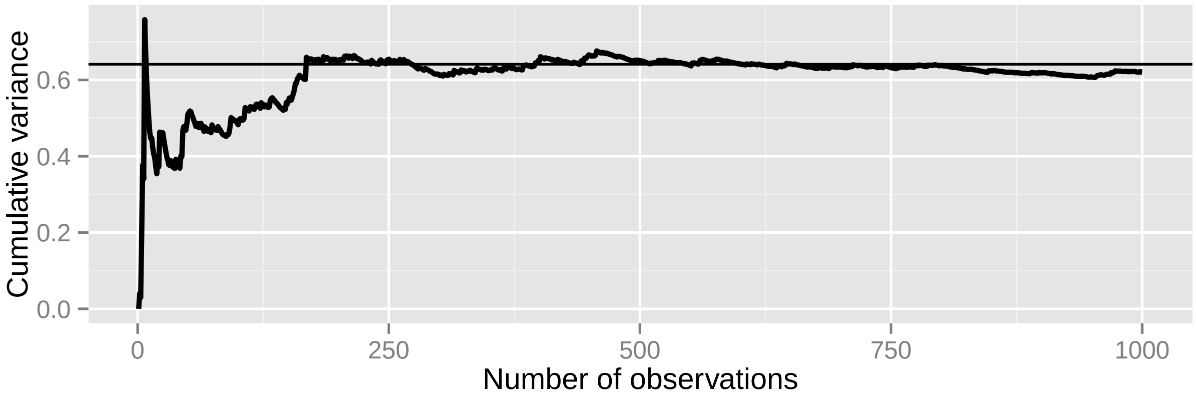
```
tail(cumsumMean)
```

```
## [1] 4.986 4.987 4.988 4.989 4.989 4.990
```

Compare simulation variance to the theoretical variance of the distribution

```
## Creating variance sequence
varSimul<-cumsum(rexpMean^2)/(1:n)-(cumsum(rexpMean)/(1:n))^2

## PLotting variance and comparing it with theoretical vale
g <- ggplot(data.frame(x = 1:n, y = varSimul), aes(x = x, y = y))
g <- g + geom_hline(yintercept = (1/lambda)^2/(sizes-1)) + geom_line(size = 1)
g + labs(x = "Number of observations", y = "Cumulative variance")
```



```
## Head and tail of variance
head(varSimul)
```

```
## [1] 0.00000 0.04012 0.02948 0.19480 0.37701 0.34048
```

```
tail(varSimul)
```

```
## [1] 0.6206 0.6207 0.6202 0.6216 0.6210 0.6208
```

Show that the distribution is approximately normal

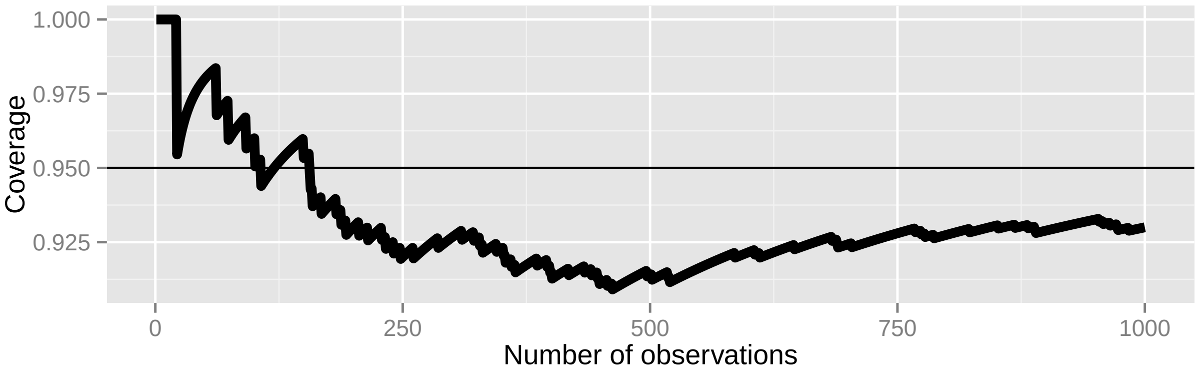
Below, just show the normal approximation code, the graph outputs aleady showed in the report.

```
meanSimul<-rexpMean
## distribution comparising
qplot(meanSimul, geom = 'blank') +
  stat_function(fun = dnorm, args=list(mean=1/lambda,sd=sqrt((1/lambda)^2/sizes)), aes(colour = 'Normal
  geom_histogram(aes(y = ..density..), colour = "darkgreen", fill = "white", binwidth = 0.2,alpha = 0.4
  scale_colour_manual(name = 'Lines', values = c('blue')) +
  theme(legend.position = c(0.85, 0.85))+
  labs(y = "Distribution density", x = "Mean of Simulation")
```

```
## Q-Q plot
qqnorm(meanSimul,main="Q-Q plot of the simulation",
       xlab="Theoretical Normal Quantiles",
       ylab="Simulation random quantile")
qqline(meanSimul, col = "darkgreen")
```

Coverage simulation

```
varSimul <-apply(rexpArr,1,var)
meanSimul<-rexpMean
## Calculating coverage sequence
coverage <- mapply(function(mean,var){
  lhats <- mean
  ll <- lhats - qnorm(.975) * sqrt(var/sizes)
  ul <- lhats + qnorm(.975) * sqrt(var/sizes)
  ll < 1/lambda & ul > 1/lambda
}, meanSimul,varSimul)
cumsumCover<-cumsum(coverage)/(1:n)
## Plotting coverage
ggplot(data.frame(x=1:n, cumsumCover), aes(x = x, y = cumsumCover)) +
  geom_line(size = 2) + geom_hline(yintercept = 0.95) +
  labs(x = "Number of observations", y = "Coverage")
```

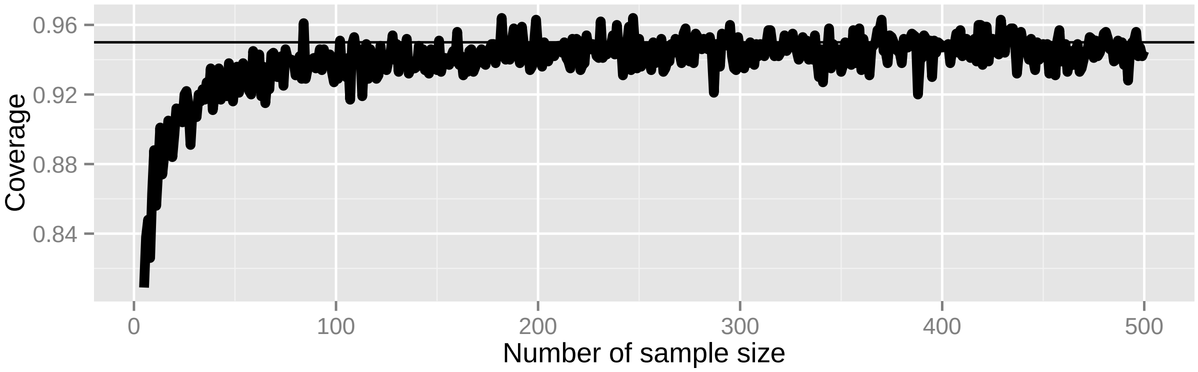


```
tail(cumsumCover)
```

```
## [1] 0.9296 0.9297 0.9298 0.9299 0.9299 0.9300
```

Coverage variance with sample size

The coverage variance with sample size was also studied, for the length limits, below just show the simulation result, the simulation code was hide. The simulation is simulating the average exponential variable from size=5 to size=500:



We can see that at the small sample size,the coverage is noticeable below the theoretical coverage value 95%. With smaple size increasing, the coverage also increasing. At about avearge more 150 exponential, the coverage is very close to the theoretical value 95%.