Comparison of exponential distributions with the Central Limit Theorem

David C Latshaw II

Overview

In this project I will investigate the exponential distribution 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. I use lambda = 0.2 for all of the simulations with the distribution of averages of 40 exponential and one thousand simulations.

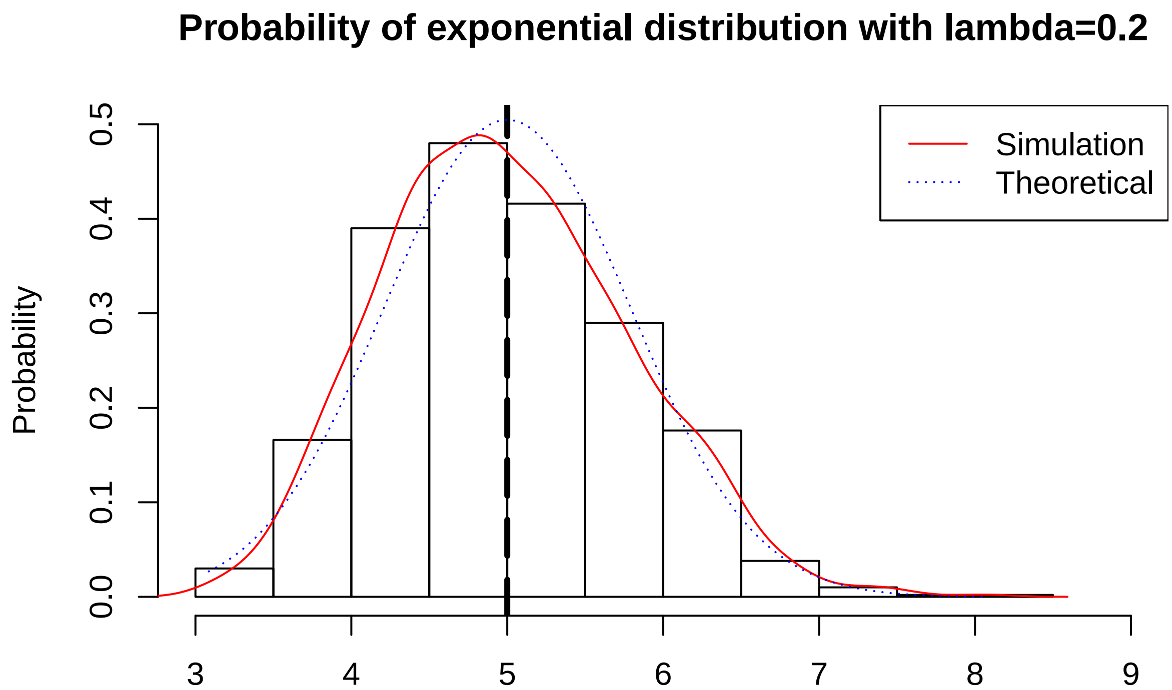
Simulations

First I set the seed for repeatability then load in all the necessary simulation parameters. A matrix is then loaded with all 1000 simulations. Means for each data point are then generated.

```
set.seed(1)
simNumber<-1000
numberSamples<-40
lambda<-0.2
simData<-matrix(rexp(simNumber*numberSamples,lambda),simNumber,numberSamples)
simMeans<-rowMeans(simData)
```

Next to visualize the distribution I have generated we create a histogram. The bars in the histogram represent the simulation data, the red curve is curve representing the same data, and the blue line is the theoretical normal distribution.The thick dotted line in the middle at x=5 is the theoretical mean.

```
hist(simMeans,
     breaks=15,
     probability=TRUE,
     main="Probability of exponential distribution with lambda=0.2",
     ylim=c(0,0.5),
     xlim=c(3,9),
     xlab="",
     ylab="Probability")
lines(density(simMeans), col="red", lty=1)
abline(v=1/lambda, col="black", lwd=3, lty=5)
xfit <- seq(min(simMeans), max(simMeans), length=1000)
ynorm<- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(numberSamples)))
lines(xfit, ynorm, col="blue", lty=3)
legend('topright', c("Simulation", "Theoretical"), lty=c(1,3), col=c("red", "blue"))
```



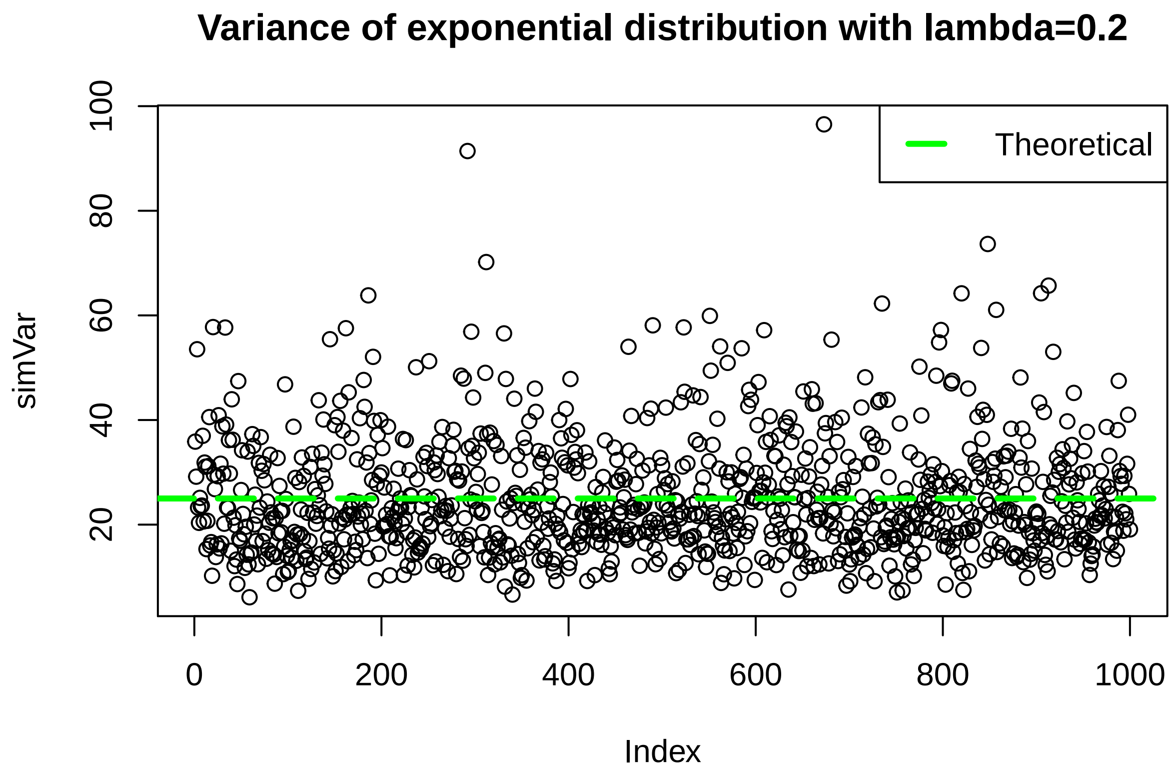

```r
median(simMeans)
```

```
## [1] 4.924455
```

The distribution of the simulation means is 4.924455 while the theoretical center of the normal distribution is 5. With this large number of simulations the exponential distribution has approached the theoretical mean.

Next calculate the variance of the simulation data and plot it against the theoretical variance. The dotted green line is the theoretical variance while the individual dots are the variance of each simulation.

```r
simVar<-apply(simData,1,var)
plot(simVar,
     main="Variance of exponential distribution with lambda=0.2")
abline(h=1/lambda^2, col="green", lwd=3, lty=5)
legend('topright', "Theoretical", lwd=3, lty=5, col="green")
```



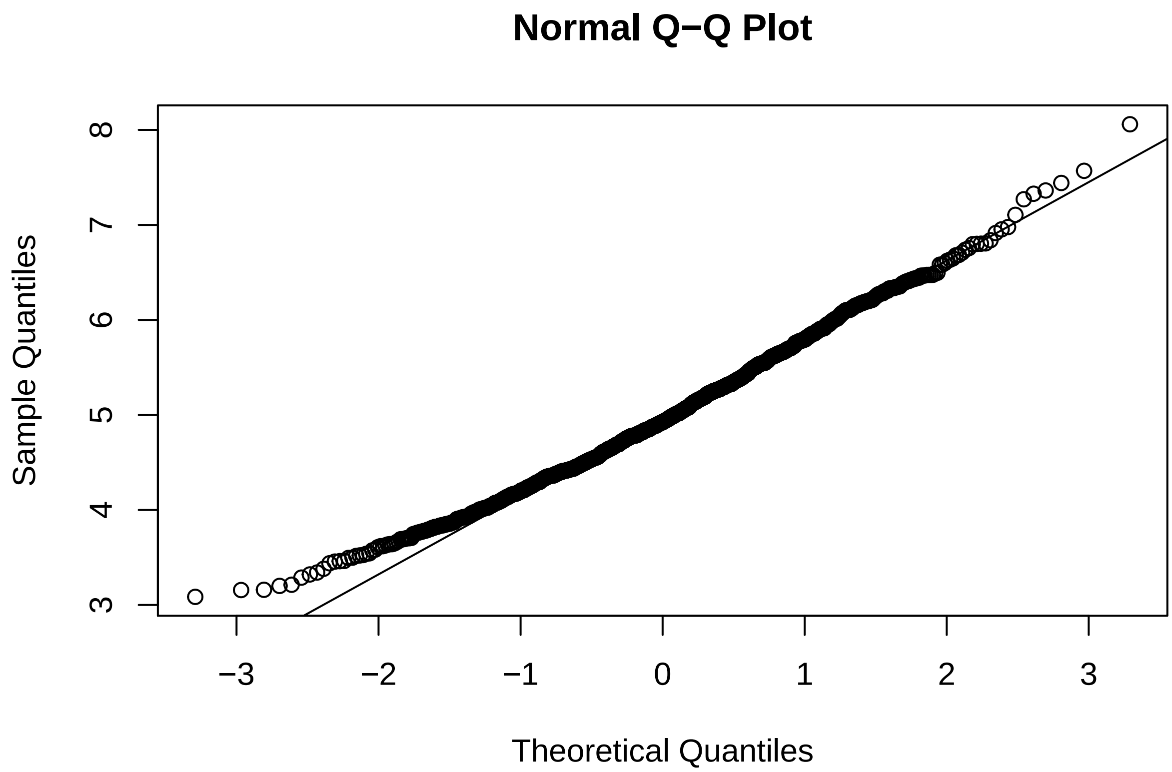

```r
mean(simVar)
```

```
## [1] 25.05783
```

The mean variance for the simulated exponential is 25.05783 whereas the theoretical variance of a normal distribution would be 25. Again with a large number of simulations the variance of the exponential distribution as almost the same as the normal distribution.

One final check we can perform is a QQ plot to show that the simulated exponential data is almost exactly the same as a normal distribution when the number of simulations is large:

```r
qqnorm(simMeans)
qqline(simMeans)
```



The majority of the data falls along the x-y line indicating that the data is essentially normal.

Conclusions

The central limit theorem (CLT) states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution. In this case 1000 simulations of an exponential distribution have been shown to agree with the theorem.