Statistical Inference Project Part 1

Matt Turner

Monday, January 12, 2015

The Question

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 (λ) is the rate parameter. The mean of exponential distribution is $\frac{1}{\lambda}$ and the standard deviation is also $\frac{1}{\lambda}$. 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.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.
You should:
1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

The Solution

Taking the information given in the question, the first step is to run 1000 simulations collecting the mean value of 40 exponentials with λ=0.2. This is done with the following code:

```
set.seed(1000)
Lambda <- 0.2
Size <- 40
SimCount <- 1000

ExpDist <- matrix(rexp(Size*SimCount, Lambda),SimCount)
ExpDistMeans <- rowMeans(ExpDist)
```

Here, we initially calculate a 1000 row by 40 column matrix and store it in the variable **ExpDist**. Next, the mean of each row is calculated resulting in a matrix **ExpDistMeans** that contains the mean of each group of 40 expontential. It therefore has 1000 data points.

The question has stated that the *theoretical mean* of an exponential distribution is $\frac{1}{\lambda}$. In our case this results in the theoretical mean being 5.
In addition, we are told that the theoretical standard deviation is also $\frac{1}{\lambda}$. Now, using our knowledge gained from the lecture videos, when taking the average of a sample distribution the standard deviation becomes $\sigma = \frac{Pop.Std.Dev}{\sqrt{SampleSize}}$. This means that in our case, the theoretical standard deviation for the distribution of sample averages will be $\sigma = \frac{\frac{1}{\lambda}}{\sqrt{40}}$.

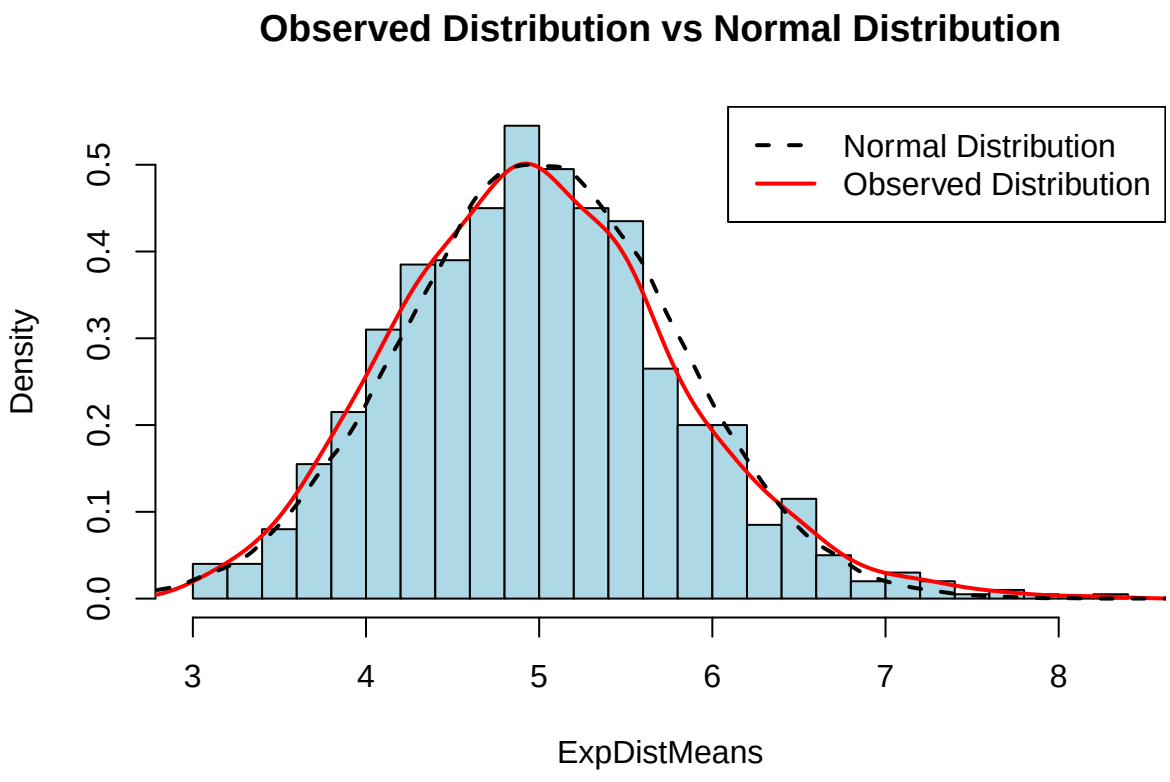
We can now put this information into R to give us some data to work with.
Note: Th. represents the theoretical values, Ob. represents the observed values.

```r
Th.mean <- 1/Lambda
Th.sd <- (1/Lambda)/sqrt(Size)
Th.var <- Th.sd^2

Ob.mean <- mean(ExpDistMeans)
Ob.sd <- sd(ExpDistMeans)
Ob.var <- Ob.sd^2

Result <- cbind(c(Th.mean, Th.sd, Th.var),c(Ob.mean, Ob.sd, Ob.var))
rownames(Result) <- c("Mean","Std.Dev","Variance")
colnames(Result) <- c("Theoretical","Observed")
Result
```

```
##          Theoretical  Observed
## Mean       5.0000000 4.9869634
## Std.Dev    0.7905694 0.8113908
## Variance   0.6250000 0.6583551
```

We can clearly see here that the observed values for the mean and variance are very close to the theoretical values. This is because both the sample size (40) and the number of simulations (1000) were large enough to produce an accurate estimation. ***Therefore, parts 1 and 2 of the question have been addressed.***

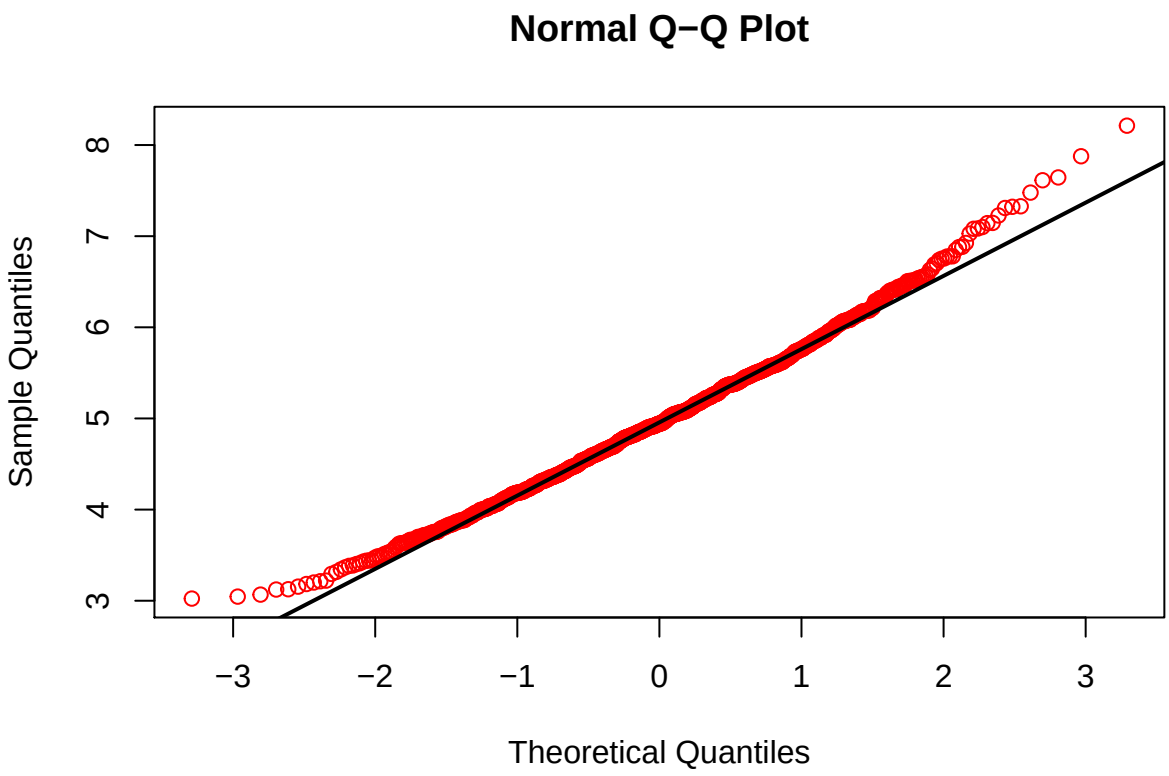
The following graph compares the theoretical and observed distributions:



Code shwon in Appendix A

Again, this shows that the centres, shown by the maximum point, of each distribution are very close. This is a demonstration of the central limit theorem.

To further assess normality, a simple QQ-Plot will display how close our observed values are to normal:



Code shown in Appendix B

The above graph shows that our observed sample means distribution is approximately equal to a normal distribution. The *tails* are the only section that do not fall on the normal line as they represent the extreme values. Increasing the sample size will go some way to resolve this. ***The two graphs have therefore addressed part 3 of the question.***

Appendix

All chunks set with "eval=FALSE" to prevent unnecessary re-running of code

Appendix A Code:

```
hist(ExpDistMeans, prob=T, breaks=20, col="lightblue",
     main="Observed Distribution vs Normal Distribution")
lines(density(ExpDistMeans), col="red", lwd=2)
lines(density(rnorm(100000,mean=Th.mean,sd=Th.sd)), col="black",lty=2,lwd=2)
legend("topright",c("Normal Distribution","Observed Distribution"),lty=c(2,1),
     lwd=2, col=c("black","red"))
```

Appendix B Code:

```
qqnorm(ExpDistMeans, col="red")
qqline(ExpDistMeans, col="black", lwd=2)
```