Title: Statistical Inference Project – Assignment I

Author: Kashyap Kumar

Date: 28th Aug '16

Goal: To investigate the exponential distribution in R and compare it with the Central Limit Theorem.

Preliminary Setup

Simulate 1000 trials of 40 samples of the exponential distribution with rate (λ) 0.2.
We generate 40,000 exponentials using rexp(), create a 1000 by 40 matrix out of it.
Since we investigate averages of exponentials, we calculate average of each row of the matrix, and store the results in a vector.

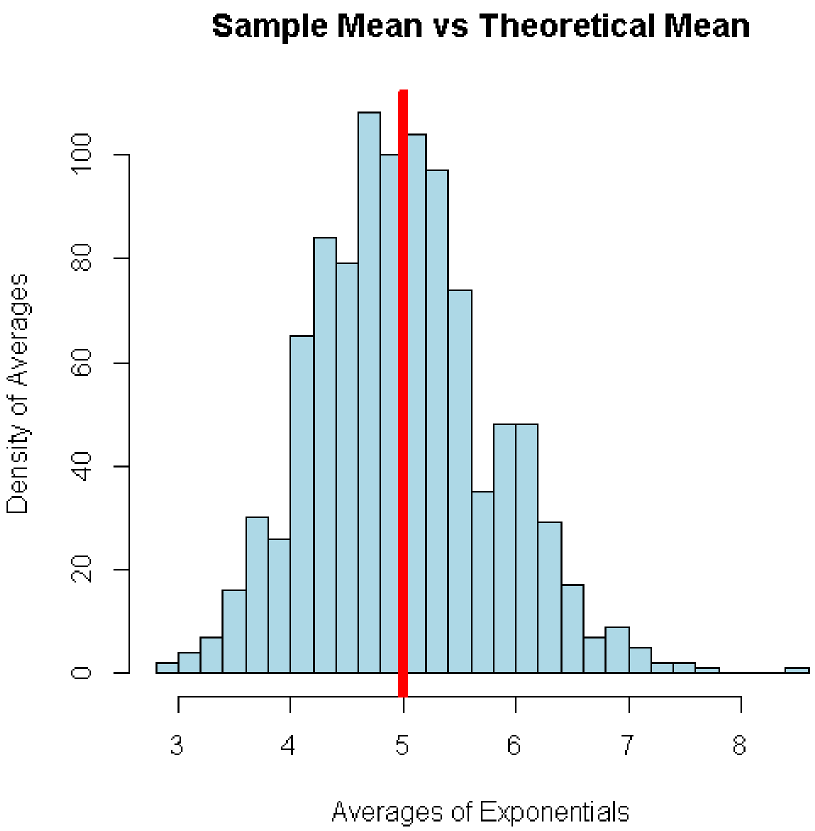
```
# Setting lambda as rate parameter, set variables for 1000 simulations and 40 exponentials to investigate

lambda <- 0.2              #rate parameter

num_sim <- 1000                  #number of simulations

exp_invest <- 40                 #exponentials

set.seed(100)

# Generate 40,000 exponentials and transfer to a Data frame

sim_exp  <- data.frame( x = rexp(num_sim * exp_invest, lambda))

# Create a 1000 by 40 matrix of exponentials

sim_mat <- matrix(sim_exp$x, num_sim)

# Create a vector to store Average of rows of sim_mat / vector has distribution of averages of 40 exponentials

avg_sim_mat_byrow <- rowMeans(sim_mat[, 1:exp_invest])
```

Question 1: Show the sample mean and compare it to the theoretical mean of the distribution.

```
# Calculate Sample mean and Theorectical mean

samp_mean <- mean(avg_sim_mat_byrow)             # Sample mean

theor_mean <- 1/lambda                           # Theoretical mean

# Compare Sample mean and Theoretical mean

samp_mean

[1] 4.97509
```

```
theor_mean
[1] 5
```

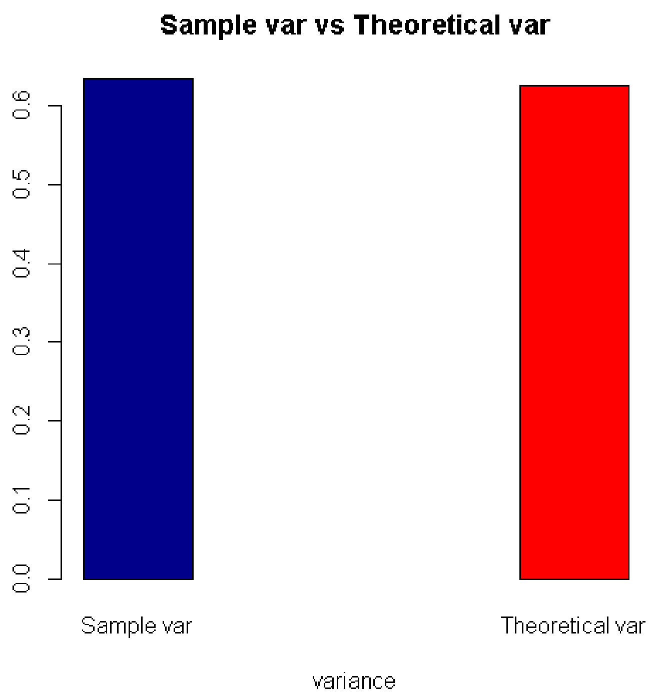
```
# Compare Sample mean and Theoretical mean via Histogram plot
hist(avg_sim_mat_byrow,  freq=TRUE, breaks=30,
    main="Sample Mean vs Theoretical Mean",
    xlab="Averages of Exponentials",
    ylab="Density of Averages",
    col='light blue')
abline(v=1/lambda,col='red',lwd=6)                 # Theoretical mean in Red
abline(v=samp_mean,col='green',lwd=6)          # Theoretical mean in Green
```



Conclusion: The Sample mean is very close to the Theoretical mean.

Question 2: Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

```
# Calculate Sample variance and Theoretical variance

samp_var <- var(avg_sim_mat_byrow)

theor_var <- (1/lambda)^2/exp_invest

samp_var

[1] 0.6335302

theor_var

[1] 0.625

# Compare Sample variance and Theoretical variance via bar chart

var_diff <- c(samp_var, theor_var)

barplot(var_diff, main="Sample var vs Theoretical var",

        xlab="variance",names.arg=c("Sample var", "Theoretical var"),col=c("darkblue","red"),space = 3)
```



Conclusion: The Sample variance is very close to the Theoretical variance.

Question 3: Show that the distribution is approximately normal.

The distribution is plotted to visually compare against a Normal distribution.

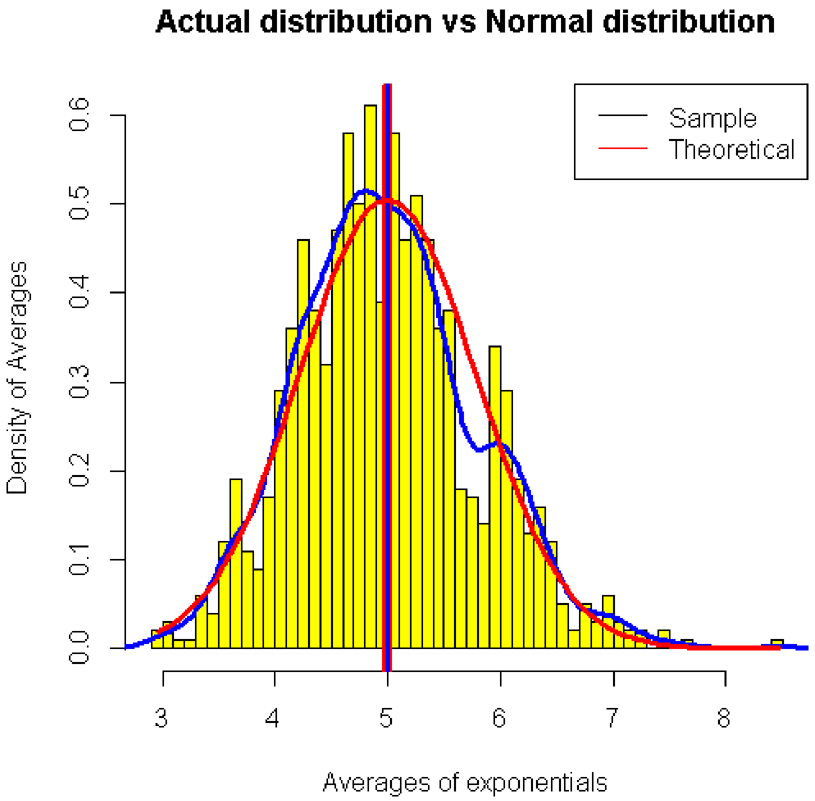
The *central limit theorem* states that the distribution of the standard score Zn converges to the standard normal distribution as n→∞.

And it gets clearer from the chart below.

Bars – represent distribution of sample values

Blue lines belong to sample; Red lines belong to Theoretical normal.

```
# Visually compare the actual distribution with Normal distribution
hist(avg_sim_mat_byrow,  prob=T,breaks=40,
     main="Actual distribution vs Normal distribution",
     xlab="Averages of exponentials",
     ylab="Density of Averages",
     col='yellow')
# Density of the means of sample exponentials
lines(density(avg_sim_mat_byrow),col="blue", lty=1, lwd =3)
# Theoretical Mean
abline(v=mean(avg_sim_mat_byrow), col='red',lty=1,lwd=6)
# Sample Mean
abline(v=samp_mean,col='blue',lty=1, lwd=3)
# Normal theoretical density of the exponential distribution
xfit <- seq(min(avg_sim_mat_byrow), max(avg_sim_mat_byrow), length=100)
yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(exp_invest)))
lines(xfit, yfit, col="red", lty=1, lwd =3)
legend('topright', c("Sample", "Theoretical"),
      col=c("black", "red"), lty=c(1,1))
```



Here, we can see the shape and bar (mean) of the sample distribution approximates a Normal distribution curve.