Statistical Inference - Basic Inference Simulation

Dexter Wang

21 May 2016

Synopsis

This report involves a practice of basic data simulation and inferential data analysis.

The goal is to 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 is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. In the simulation, set lambda = 0.2 and run the distribution of averages of 40 exponentials for 1000 times.

The questions to answer includes:

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.

Setup simulation

```
l <- 0.2    #set lambda=0.2
n<- 40      #40 simulations per set
i<- 1000    #run simulation 1000 times

set.seed(220516)
sim<- replicate(i,rexp(n,l))
```

Task One

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

```
sMean <- round(mean(colMeans(sim)),3) #Sample Mean
tMean <- 1/l #Theoretical Mean
data.frame(Type=c("Sample","Theoretical"),Mean=c(sMean,tMean))
```

```
##          Type  Mean
## 1      Sample 4.985
## 2 Theoretical 5.000
```

It can be seen that the sample mean (4.985) is very close to the theoretical Mean (5)

Task Two

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

```
sVar <- round(var(colMeans(sim)), 3)
sSd <- round(sd(colMeans(sim)), 3)
tSd <- round((1/l * 1/sqrt(n)), 3)  #Theoretical Standard Deviation
tVar <- round(tSd^2, 3)  #Theoretical Variance
diff_Var <- paste("", round((sVar - tVar)/tVar, 3) * 100, "%")
diff_Sd <- paste("", round((sSd - tSd)/tSd, 3) * 100, "%")
data.frame(Calculation = c("Standard Deviation", "Variance"),
    Sample = c(sSd, sVar), Theoretical = c(tSd, tVar), Difference = c(diff_Sd,
        diff_Var))
```

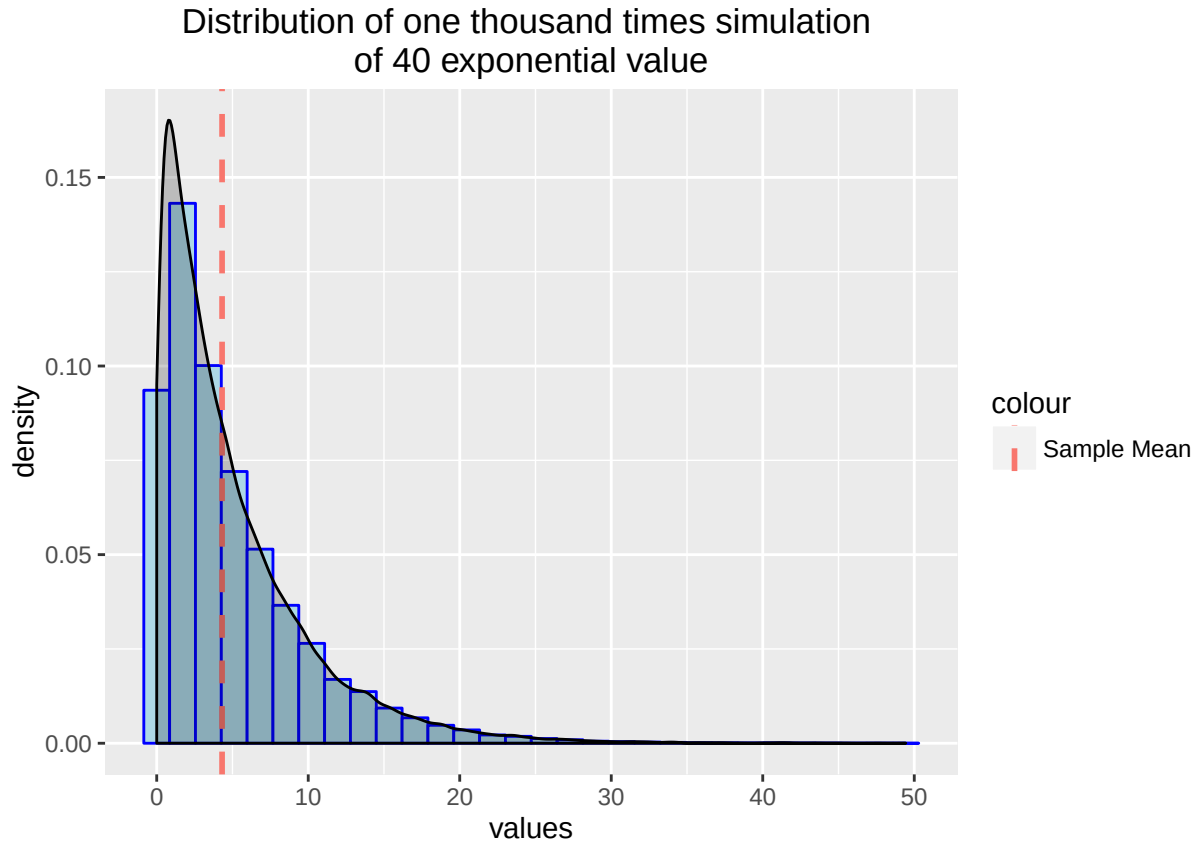
```
##          Calculation Sample Theoretical Difference
## 1 Standard Deviation  0.794       0.791      0.4 %
## 2           Variance  0.630       0.626      0.6 %
```

The result shows that both Sample Variance (0.63) and Sample Standard Deviation (0.794) are perfectly matching Theoretical Variance (0.626) and Theoretical Standard Deviation (0.791) with only 0.6 % and 0.4 % difference compare to Theoretical value respectively.

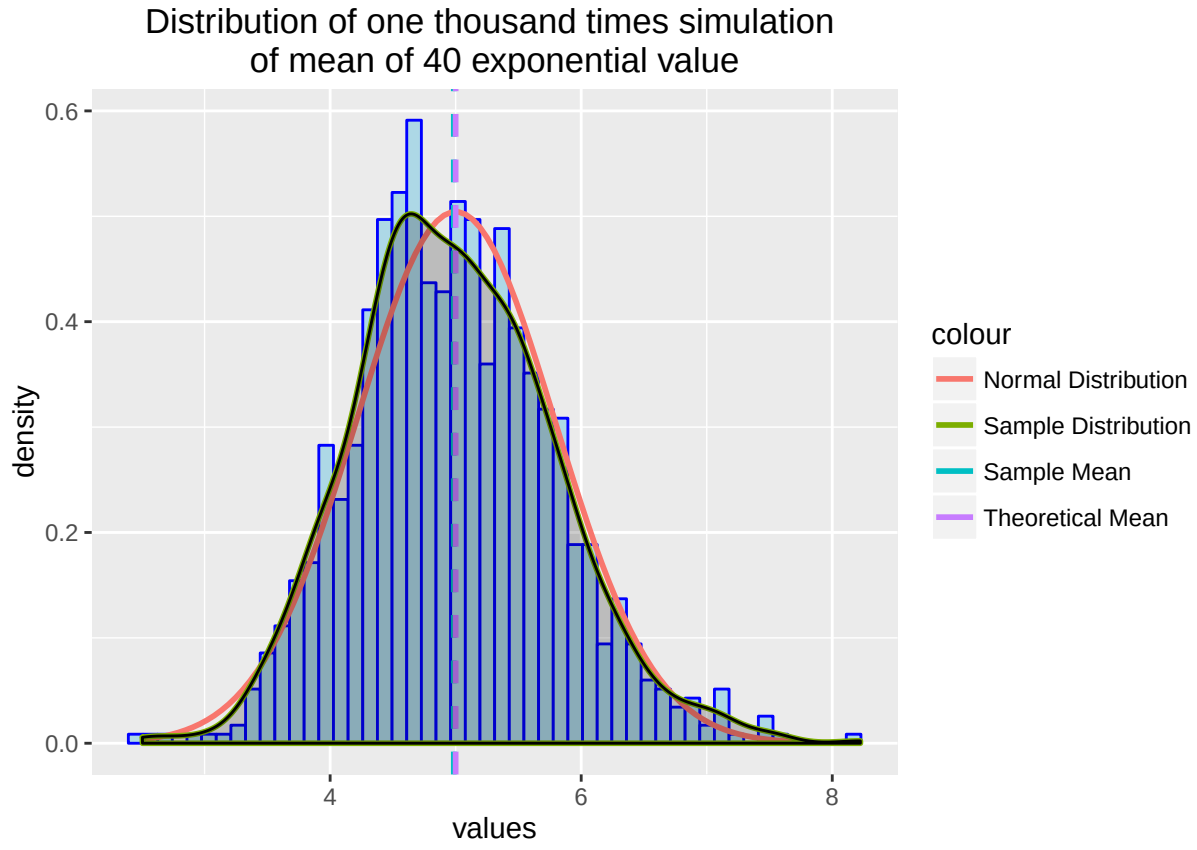
Task Three

```
library(ggplot2)

# plot 1
sim.plot <- data.frame(as.vector(sim))
title <- "Distribution of one thousand times simulation \nof 40 exponential value"
g <- ggplot(sim.plot, aes(sim.plot))
g <- g + geom_histogram(aes(y = ..density..), bins = 30, color = "blue",
    fill = "lightblue")
g <- g + geom_vline(aes(color = "Sample Mean", xintercept = mean(sim[,
    1])), linetype = "dashed", size = 1)
g <- g + geom_density(alpha = 0.2, fill = "black") + labs(x = "values",
    y = "density", title = title)
g
```




```
# plot 2
simMean.plot <- data.frame(mean = colMeans(sim))
title <- "Distribution of one thousand times simulation \nof mean of 40 exponential value"
g2 <- ggplot(simMean.plot, aes(simMean.plot))
g2 <- g2 + geom_histogram(aes(y = ..density..), bins = 50, color = "blue",
    fill = "lightblue")
g2 <- g2 + stat_function(fun = dnorm, args = list(mean = tMean, sd = tSd),
    aes(color = "Normal Distribution"), size = 1)
g2 <- g2 + geom_density(aes(color = "Sample Distribution"), size = 1,
    show_guide = FALSE)
g2 <- g2 + geom_vline(aes(color = "Sample Mean", xintercept = sMean),
    linetype = "dashed", size = 1, show_guide = FALSE)
g2 <- g2 + geom_vline(aes(color = "Theoretical Mean", xintercept = tMean),
    linetype = "dashed", size = 1, show_guide = FALSE)
g2 <- g2 + geom_density(alpha = 0.2, fill = "black")
g2 <- g2 + labs(title = title, x = "values", y = "density")
g2
```



From the figures, it can be seen that the simulation is exponentially distributed. The mean of 40 exponential values in one thousand simulation almost overlap with normal distribution with slight left skew. The sample and theoretical mean are nearly same.

Thanks for viewing :)