Statistical Inference Course Project: 1st Part

Dineshkumar Murugan

Saturday, May 23, 2015

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 is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 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.

Simulation

Loading the Package and the input values

```
library(ggplot2)

n <- 40
numsim <- 1000
lamdaT <- 0.2
```

Getting the Theoretical Values

```
meanT <- 1/lamdaT

varT <- (1/lamdaT^2) * (1/n)

sdT <- ((1/lamdaT) * (1/sqrt(n)))
```

Creating a Dataset with 1000 Simulations

```
dataset<-matrix(rexp(n*numsim,rate = lamdaT),numsim)
```

Calculating the statistical values based on the Exponential Distribution formula's from the sample

```r
RowMeanfordataset <- apply(dataset,FUN = mean,MARGIN = 1)

CalculatedMean <- mean(RowMeanfordataset)
CalculatedVar <- var(RowMeanfordataset)
CalculatedSd <- sd(RowMeanfordataset)
```

Comparing Values

Comparing theoretical mean and sample mean

```r
c(meanT,CalculatedMean)
```

```
## [1] 5.000000 5.004092
```

Comparing theoretical mean and sample mean

```r
c(varT,CalculatedVar)
```

```
## [1] 0.6250000 0.6276036
```

Comparing theoretical Variance and sample Variance

```r
c(meanT,CalculatedMean)
```

```
## [1] 5.000000 5.004092
```

Comparing theoretical SD and sample SD

```r
c(sdT,CalculatedSd)
```

```
## [1] 0.7905694 0.7922144
```

Ploting the Distribution

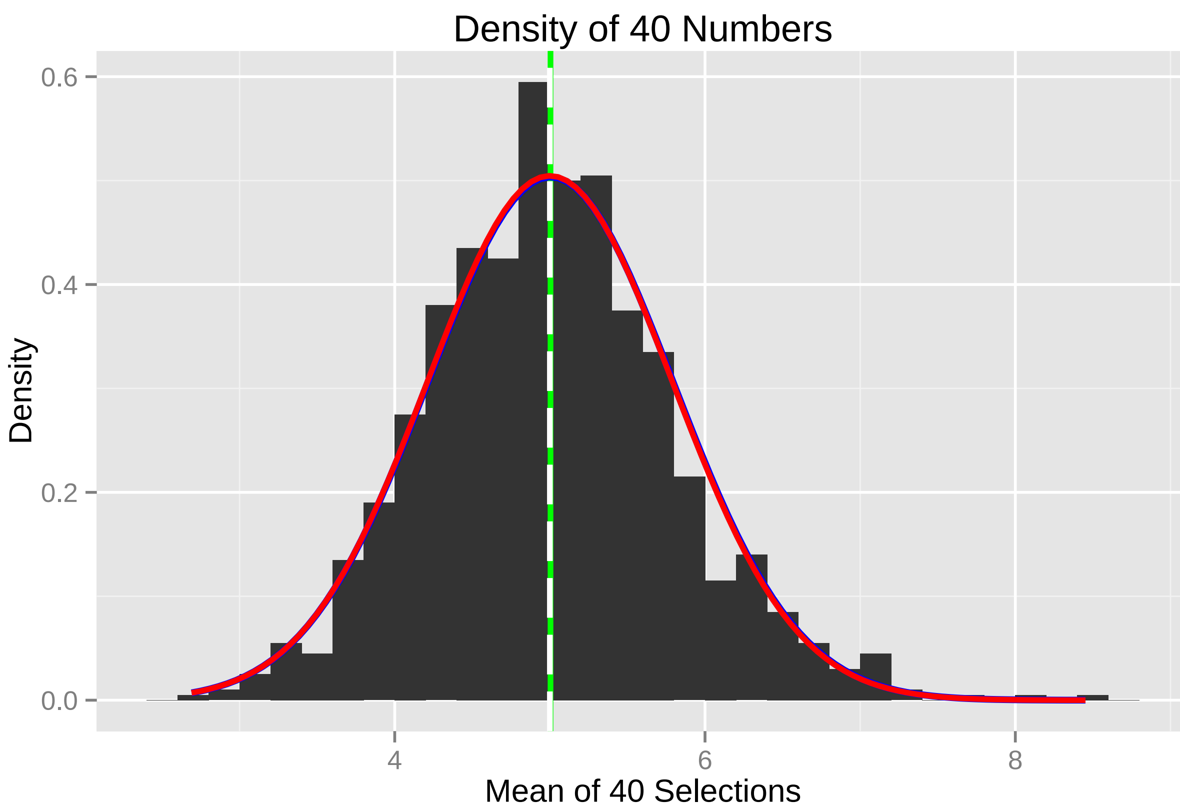
Theoreticaly distribution should be a normal distribution and the mean of the distribution should be same as the mean of the population - As per **Central Limit Theorem**

```r
dfRowMeans<-data.frame(RowMeanfordataset)

hist_cut <- ggplot(dfRowMeans,aes(x = RowMeanfordataset))

hist_cut <- hist_cut + geom_histogram(aes(y = ..density..),binwidth = lamdaT)
```

```r
hist_cut <- hist_cut + labs(title="Density of 40 Numbers", x="Mean of 40 Selections", y="Density")
hist_cut <- hist_cut + geom_vline(xintercept=CalculatedMean,size=1.0, color="green")
hist_cut <- hist_cut + stat_function(fun=dnorm,args=list(mean=CalculatedMean, sd=CalculatedSd),color =
hist_cut <- hist_cut + geom_vline(xintercept=meanT,size=1.0,color="white",linetype = "longdash")
hist_cut <- hist_cut + stat_function(fun=dnorm,args=list(mean=meanT, sd=sqrt(varT)),color = "red", size
hist_cut
```



Green Color Marks the Sample Mean, White Dotten Lines represent Theoretical Mean, Blue Line Represents Sample Normal Distribution and Red Line Represents Theoretical Normal Distribution.