Statistical Inference - Course Project - Part 1

Michael Elfellah

September 27, 2015

In this project, a simulation is used to explore inference and do some simple inferential data analysis. The project consists of two parts:

Part 1. A simulation exercise.

A report is created using knitr to produce a pdf document where the answer is presented.

In this project we 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. lambda is set to 0.2 for all of the simulations. We will investigate the distribution of averages of 40 exponentials. A thousand simulations will be performed.

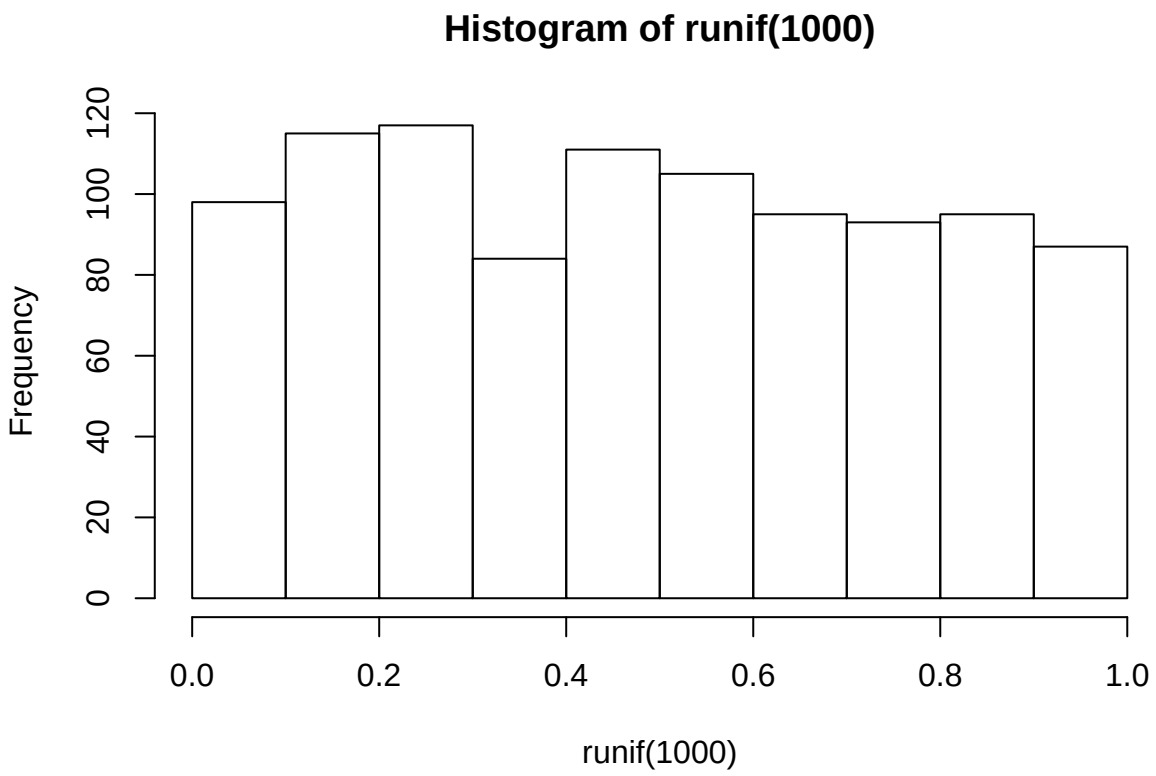
We illustrate via simulation the properties of the distribution of the mean of 40 exponentials and:

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, we 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.

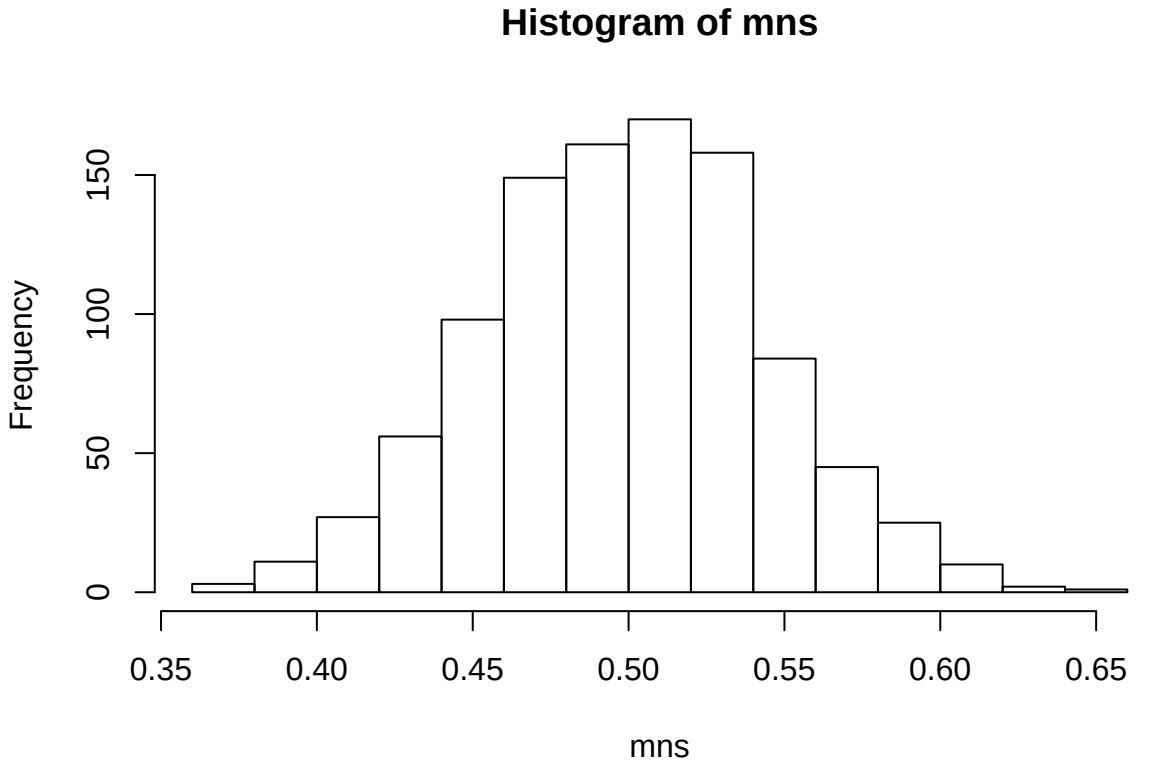
As a motivating example, we compare the distribution of 1000 random uniforms

```
hist(runif(1000))
```



The distribution of 1000 averages of 40 random uniforms:

```
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(runif(40)))
hist(mns)
```



To relate the two distributions, We Start by running a 1000 simulations of 40 exponentials:

```
set.seed(555)
lambda <- 0.2
nosim <- 1:1000
```

Number of Simulations/rows:

```
n <- 40
```

Matrix of simulated values:

```
expMatrix <- data.frame(x = sapply(nosim, function(x) {mean(rexp(n, lambda))}))
head(expMatrix)
```

```
##          x
## 1 5.228362
## 2 3.714709
## 3 4.100575
## 4 5.187063
## 5 5.515633
## 6 4.592376
```

Showing where the distribution is centered at and comparing it to the theoretical center of the distribution.

Center of simulated distribution is:

```
simMeanExp <- apply(expMatrix, 2, mean)
simMeanExp
```

```
##        x
## 4.991502
```

Which is very close to the expected theoretical center of the distribution:

```
theoMeanExp <- 1/lambda
theoMeanExp
```

```
## [1] 5
```

Showing how variable it is and compare it to the theoretical variance of the distribution.

The simulated Standard Deviation and Variance are:

```
simSDExp <- sd((expMatrix$x))
simSDExp
```

```
## [1] 0.7979096
```

```r
simVarExp <- var(expMatrix$x)
simVarExp
```

```
## [1] 0.6366598
```

In comparison, the expected theretical SD and Variance are:

```r
theoSDExp <- (1/lambda)/sqrt(n)
theoSDExp
```

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

```r
theoVarExp <- theoSDExp^2
theoVarExp
```

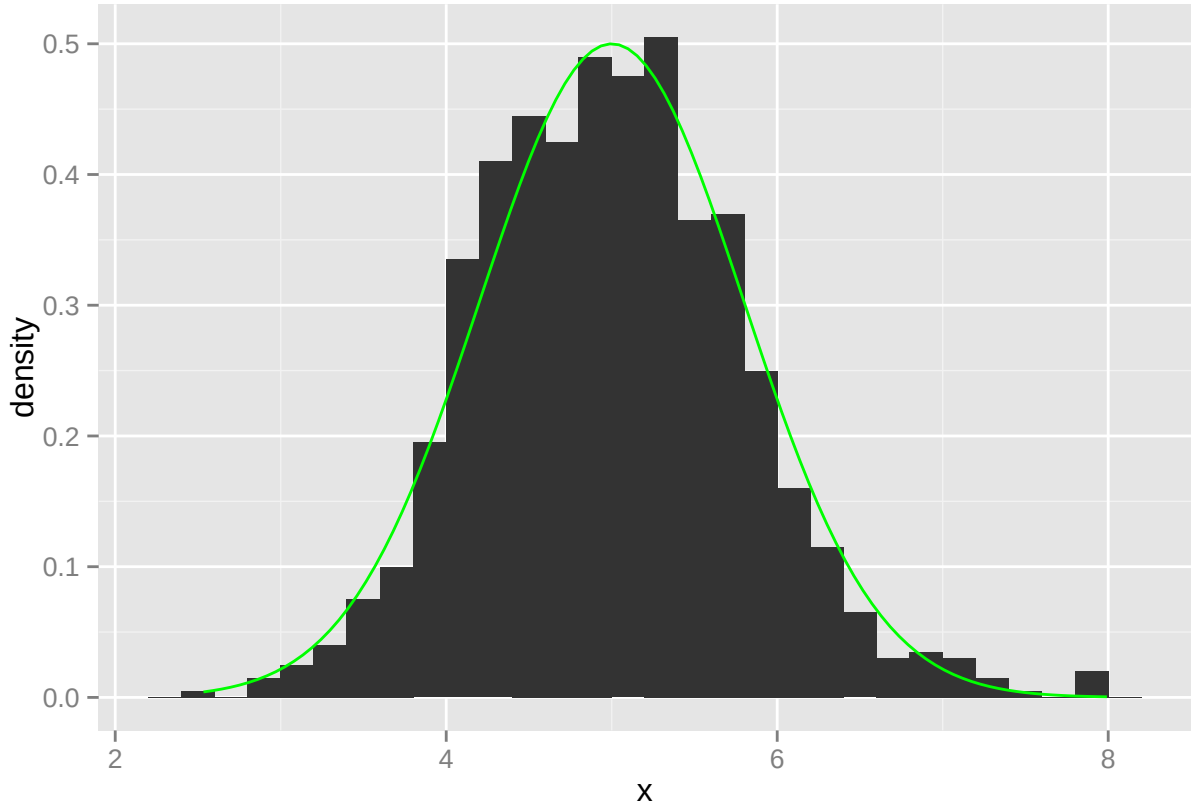
```
## [1] 0.625
```

We can safely conclude that the differences are minimal, as expected.

Showing that the distribution is approximately normal.

In order to understand if distribution is approximately normal, we perform a plot.

```r
library(ggplot2)
ggplot(data = expMatrix, aes(x = x)) +
    geom_histogram(aes(y=..density..), binwidth = 0.20) +
    stat_function(fun = dnorm, color = "green", arg = list(mean = theoMeanExp, sd = sd(expMatrix$x)))
```



We conclude that the function appears approximately Normal.