Peer Graded Stat. Inf. Part 1

Dasarath S

22/10/2020

Brief

We are going to compare Exponential Distribution in R to the CLT(Central Limit Theorem) using a simulation, which is gonna make the part 1 of this project.

Setup

Loading the packs

We need `ggplot2` to do this analysis.

```
library(ggplot2)
```

Part 1: Simulating the Analysis

1.1 We take the sample mean and contrast it to the theoretical mean.

R has this function called rexp(n, lambda) with lambda being rate parameter, through which we can simulate the exponential distribution. Several iterations can be performed with the repetition function.

According to theory, the average exponential distribution is equal to standard deviation which is inturn equal to 1/lambda.

Taking the number of simulations be 1000, lambda is taken as 0.2 and n being the sample size to be equal to 40.

```
# Keeping seed so as to ensure reproducibility
set.seed(100)
# Initializing lambda as 0.2
lamb <- 0.2
# Having lambda=0.2 and n=40, we now try to compute sample mean per one simulation
my_exponent <- rexp(40, lamb)
mean(my_exponent)
```

```
## [1] 4.137412
```

```r
# Having lambda=0.2 and n=40, we now try to compute sample mean for 1000 simulations
my_exponent1000 <- as.data.frame(replicate(1000,mean(rexp(40, lamb))))
names(my_exponent1000) <- c("avg_of_samples")
# Calculate the mean of this simulation of sample means
my_avg1000 <- mean(my_exponent1000$avg_of_samples)
my_avg1000
```

```
## [1] 5.00122
```

```r
# The theoretical mean which is the centre of the distribution is
1/lamb
```

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

Looking at sample averages of 40 exponentials, its evident that the mid point of the distribution is nearer to the mid point of the distribution obtained from theory.

1.2 Trying to see how varied the taken sample is (via variance) and contrast it to the variance obtained through theory.

```r
# By having theta^2/n computing the varience for this simulation
my_variance1000 <- var(my_exponent1000$avg_of_samples)
my_variance1000
```

```
## [1] 0.6429278
```

```r
# Variance of the distribution according to theory is
((1/lamb)^2)/40
```

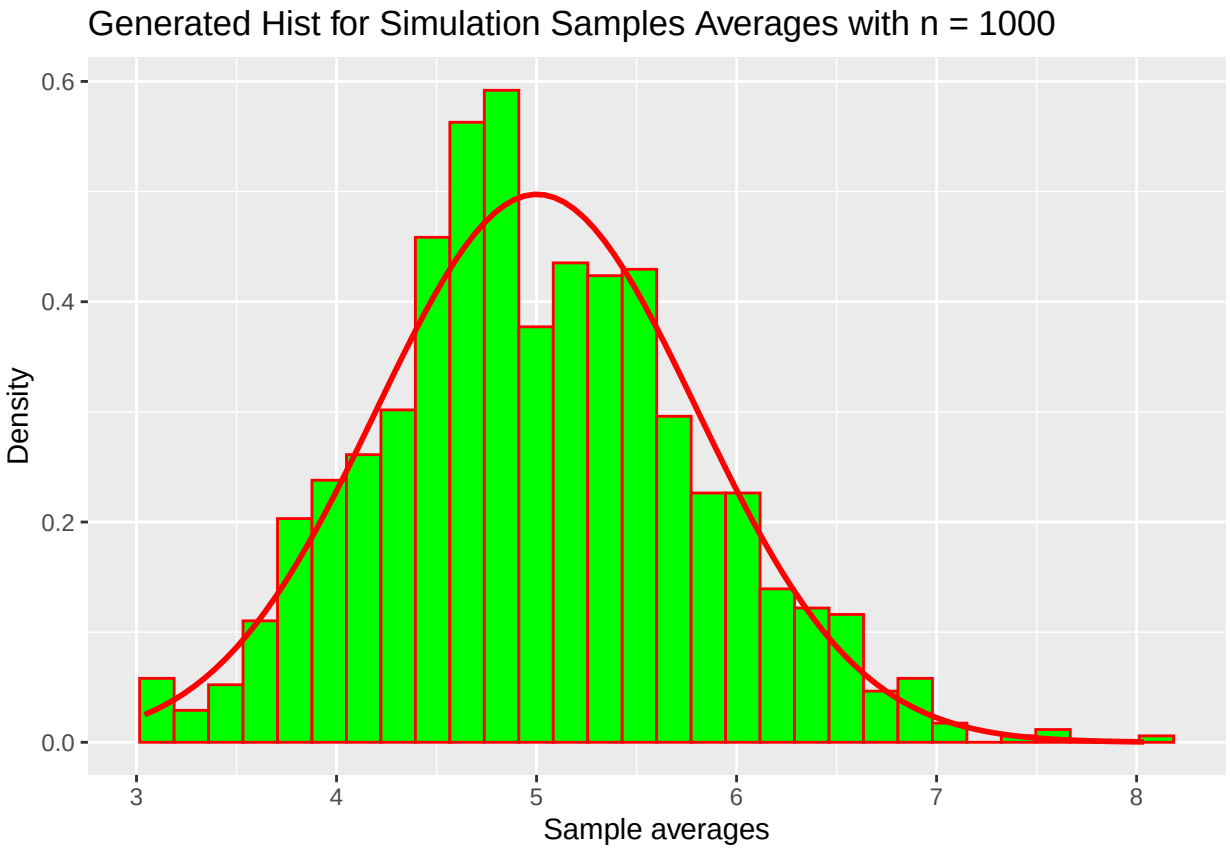
```
## [1] 0.625
```

Computing the variance of the distribution of average of 40 exponents is lower when compared to the variance calculated by theory.

1.3 Showing the distribution is roughly normal.

```r
# Using GGplot to plot the average of sims
ggplot(data = my_exponent1000, aes(x=avg_of_samples)) +
    geom_histogram(aes(y = ..density..),colour="red",fill="green")+
    stat_function(fun=dnorm,args=list( mean=my_avg1000, sd=sqrt(my_variance1000)),
    geom="line",color = "red", size = 1.0)+
    ggtitle("Generated Hist for Simulation Samples Averages with n = 1000") +
    scale_x_continuous("Sample averages")+
    ylab("Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We use the red line to mark the computed normal distribution which helps us contrast the normal distribution which is typically a bell curve with the histogram.

According to CLT, the averages of samples will converge to normal distribution if the size of sample goes high while being independent n<10% and normal, and in case of skewed distribution, n>30.
