Statistical Inference Course Project - Part 1

Orumwese Esosa Cyriaque

7/24/2021

PART 1: SIMULATION EXERCISE

Comparison Between the Exponential Distribution in R and the Central Limit Theorem

Overview

In this project I investigated the exponential distribution in R and compared 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 = 0.2 for all of the simulations. I investigated the distribution of averages of 40 exponentials and carried out a thousand simulations.

Simulations

Loading packages

```r
library(ggplot2)
```

Creating the sample exponential distribution

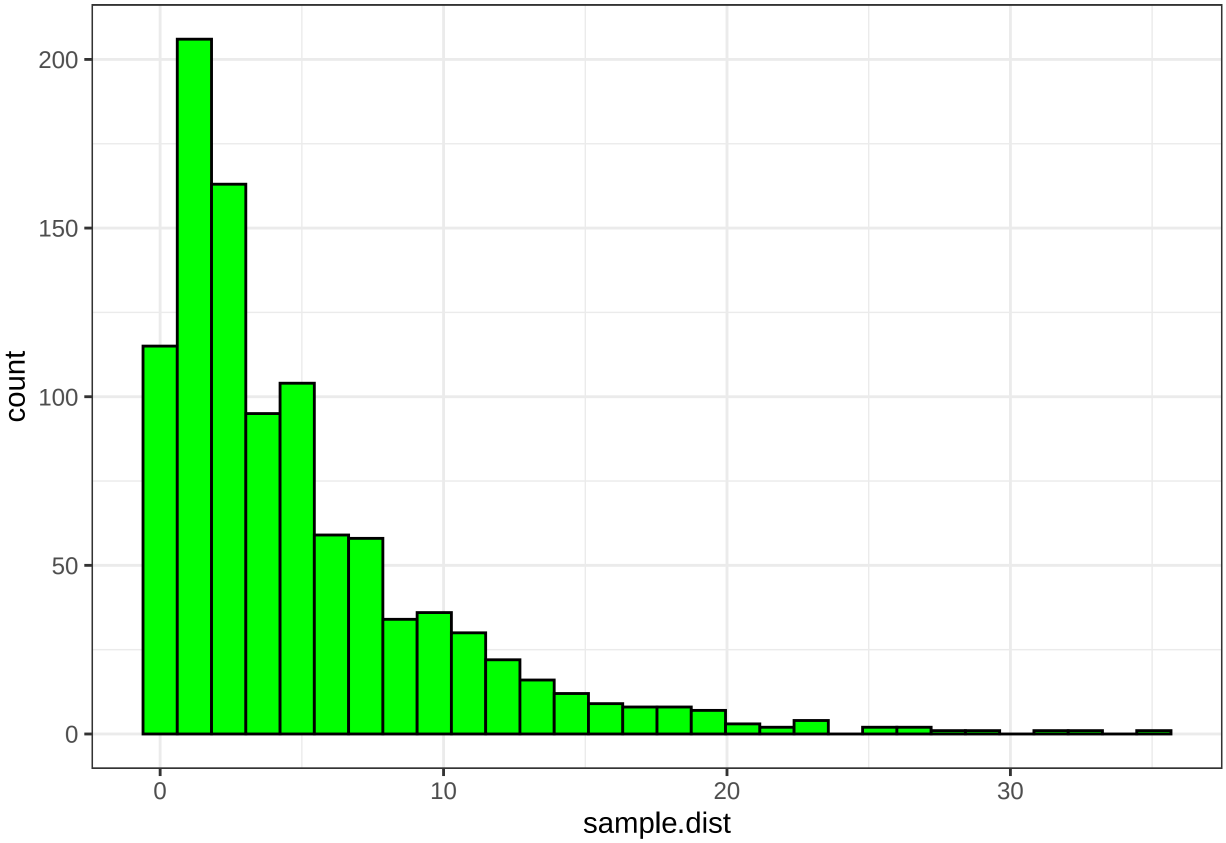
```r
# the rate parameter
lamda <- 0.2

# the number of observations
n <- 1000

# Generating 1000 random exponential variables with lamda = 0.2 and storing it in 'sample.dist'
sample.dist <- rexp(n, lamda)
head(sample.dist,30)
```

```
##  [1]  5.7406154  2.5789132  0.4632921  9.7612780  4.7497582  0.5153764
##  [7]  2.2243478  3.8122533  8.1775924  9.9512048  0.9533796  9.6665248
## [13]  3.1054832  1.6835307  4.0663885  2.9228195  0.8049378  0.5937818
## [19]  3.3539301 16.8474386  3.2968017  0.4393220  3.0168081  0.3752761
## [25]  0.7340362  8.1231128  1.9241638  2.7747012 10.3519098 13.4436461
```

Plotting histogram of the Sample Exponential Distribution

```
# Plotting a histogram
sample.plot <- ggplot(NULL,aes(x=sample.dist)) + theme_bw() +
        geom_histogram(col="black",fill="green")
sample.plot
```

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



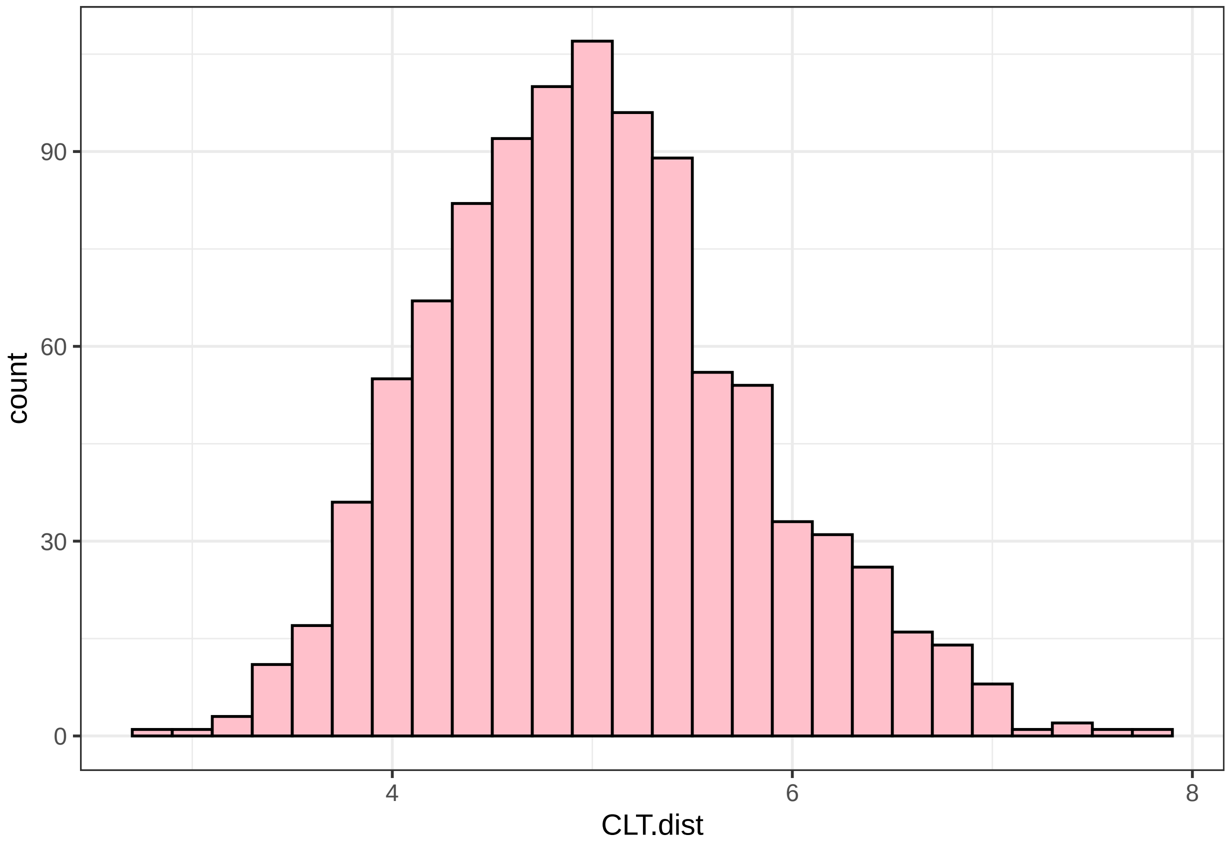
Creating sampling distribution of sample means

```
# Create an empty variable
CLT.dist <- NULL

# For each simulation, the mean of a new exponential distribution is calculated and appended to "CLT.di
# This is done 1000 times
for(i in 1:1000){
        CLT.dist = c(CLT.dist,mean(rexp(40,rate = 0.2)))
}
head(CLT.dist,30)
```

```
##  [1] 5.157911 3.331174 5.409467 6.522409 4.413108 3.810710 3.346892 4.888163
##  [9] 4.656344 4.687291 6.338663 5.302092 5.228709 4.531770 4.954855 5.560843
## [17] 4.921289 4.030952 4.652699 5.160496 5.773902 5.078252 7.308330 6.162492
## [25] 5.417022 3.449642 5.230139 4.712915 5.641428 6.190314
```

Plotting histogram of the Sampling Distribution of sample means

```
# Plotting a histogram
CLT.plot <- ggplot(NULL, aes(CLT.dist)) + theme_bw() +
        geom_histogram(col="black",fill="pink",binwidth = 0.20)
CLT.plot
```



Sample Mean versus Theoretical Mean

The sample mean is the mean of the exponential distribution which is;

```
sample.mean <- mean(sample.dist)
sample.mean
```

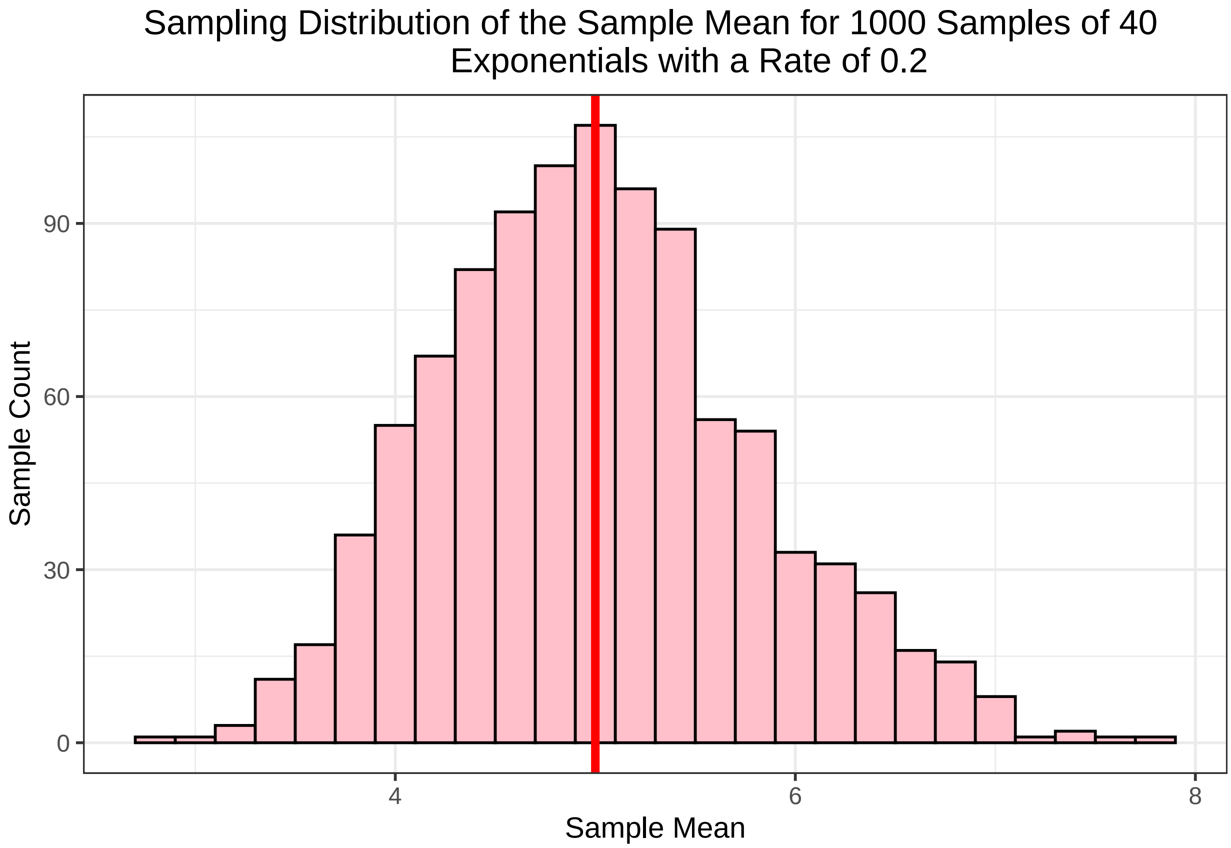
```
## [1] 4.915572
```

The theoretical mean is expected to be at the center of the distribution of the distribution of the sampling means. The theoretical mean is;

```
theoretical.mean <- 1/lamda
theoretical.mean
```

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

```
CLT.plot <- CLT.plot + geom_vline(xintercept=1/lamda,color="red",size=1.5) +
        labs(title="Sampling Distribution of the Sample Mean for 1000 Samples of 40
        Exponentials with a Rate of 0.2 ", x="Sample Mean", y="Sample Count") +
        theme(plot.title = element_text(hjust = 0.5))
CLT.plot
```



From the plot, the theoretical mean, which is at 5 is shown by the red vertical line. By comparison, it is noticed that the sample mean (4.9155723) is similar to the theoretical mean (5)

Sample Variance versus Theoretical Variance

According to the Central Limit Theorem, the sampling distribution of the sample means should have a standard deviation equivalent to the standard error of the mean (sigma/sqrt(N)).

The sample variance;

```
sample.variance <- var(CLT.dist)
sample.variance
```

```
## [1] 0.6224175
```

The theoretical variance;

```
theoretical.variance <- ((1/lamda)/sqrt(40))^2
theoretical.variance
```

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

It can be observed that the sample variance of 0.6224175 is equivalent to the theoretical variance of 0.625.

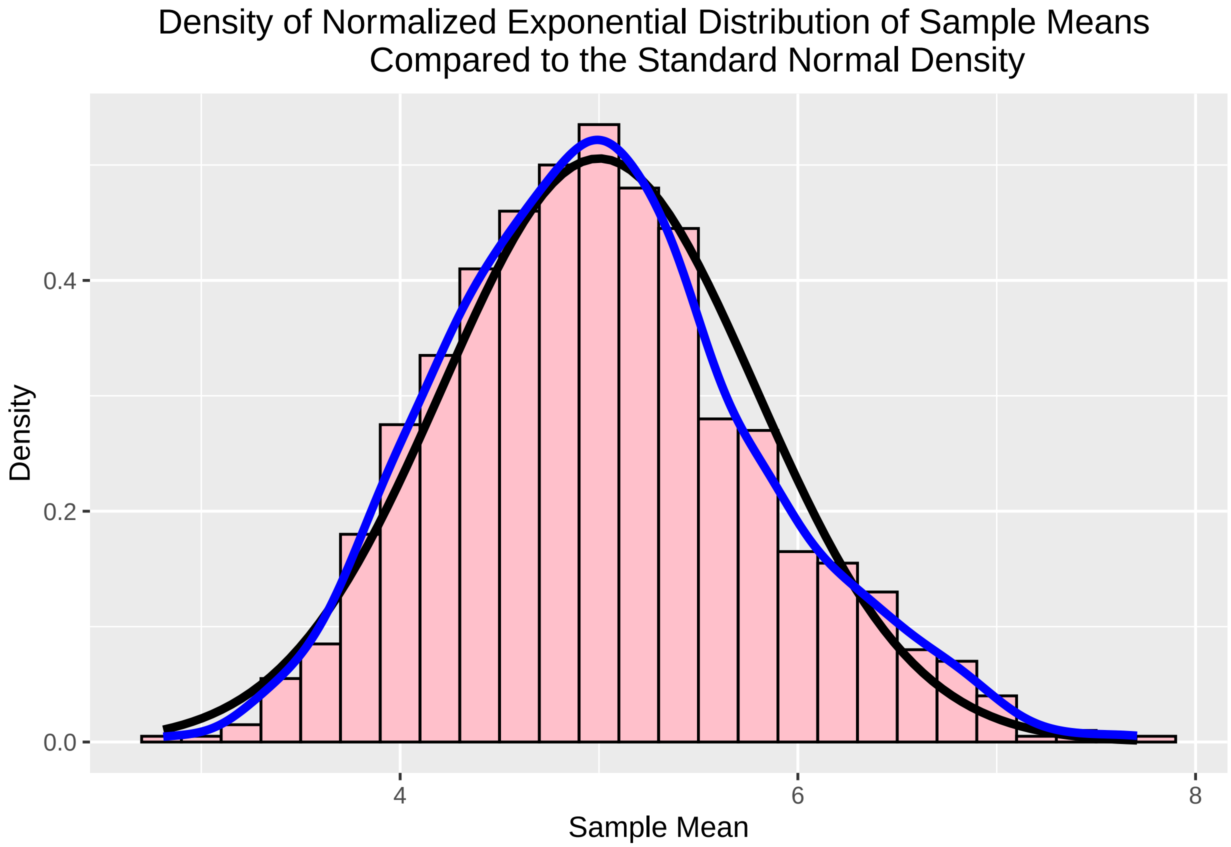
Distribution

According to the Central Limit Theorem, the sample distribution of the mean is approximately a normal distribution if the sample size is large enough (30 or more). The sample size of the sampling distribution of the mean is 40, therefore it is expected to be approximately normal.

```
# Creating a data frame of the theoretical distribution data
CLT.df <- as.data.frame(CLT.dist)
# changing column name
colnames(CLT.df) = "x"
```

Plotting the distribution and comparing it to a standard normal distribution

```
# The data frame CLT.df will be plotted with using the distribution values as its aesthetic
g <- ggplot(data = CLT.df, aes(x=CLT.dist))

# adding the histogram function with black borders and a pink fill
g <- g + geom_histogram(col="black", fill="pink", aes(y=..density..), binwidth=0.20)

# adding the function for plotting the density of a standard function with mean equal to the theoretica
# standard deviation equal to the standard error of the mean.
g <- g + stat_function(fun=dnorm, args=list(mean=theoretical.mean, sd=sd(CLT.dist)), color="black", siz

# adding the function for plotting the density of the sample exponential distribution
g <- g + stat_density(geom="line",color="blue",size=1.5)

# adding the function for labels
g <- g + labs(x="Sample Mean", y="Density", title="Density of Normalized Exponential Distribution of Sa
       Compared to the Standard Normal Density") +
       theme(plot.title = element_text(hjust=0.5)) # positions the plot title at the centre
g
```



It is observed that the normalized exponential distribution of the sample means (blue curve) is approximately normal seeing as it is very similar to the normal distribution curve (black curve)

Conclusion

Upon comparison of the exponential distribution with the Central Limit Theorem, it was concluded that the function is approximately normal.