Exponential Distribution in R vs Central Limit Theorem (Part 1)

Jon Ting

18/08/2020

```
knitr::opts_chunk$set(warning=FALSE, fig.height=4, fig.width=7)
set.seed(77)
library(ggplot2)
```

Part 1

Overview

This project explores the exponential distribution in R with a lambda of 0.2 and compare it with the Central Limit Theorem. Part 1 of the project will investigate the distribution of averages of 40 exponentials over a thousand simulations.

Simulations

1000 simulations were performed on 40 exponentially distributed samples with lambda of 0.2. The means of the simulations were extracted.

```
# Setting variables
n <- 40
lambd <- 0.2

# Simulate 40 exponentials over 1000 simulations
simulations <- replicate(n=1000, expr=rexp(n, lambd))

# Extract the means of each simulation (down each column)
sim_means <- apply(X=simulations, MARGIN=2, FUN=mean)
```

Sample Statistics versus Theoretical Statistics

The sample and theoretical statistics are computed as below:

```
samp_mean <- mean(sim_means)
theo_mean <- 1 / lambd
samp_var <- var(sim_means)
theo_var <- (1/lambd)^2 / n
```

The theoretical and sample means are respectively

```r
c(theo_mean, samp_mean)
```

```
## [1] 5.000000 4.996833
```

which is rather close to the theoretical value. It is expected that plotting both values as lines would result in a physically-indistinguishable line if visualized (as shown in following section).

The theoretical and sample variances are respectively

```r
c(theo_var, samp_var)
```

```
## [1] 0.6250000 0.6162158
```

The two values also seem to be rather close to each other, indicating that the spreads of both curves are expected to be identical to each other if visualized (as shown below).

The upper and lower bounds of 95% confidence interval for sample mean are computed as below:

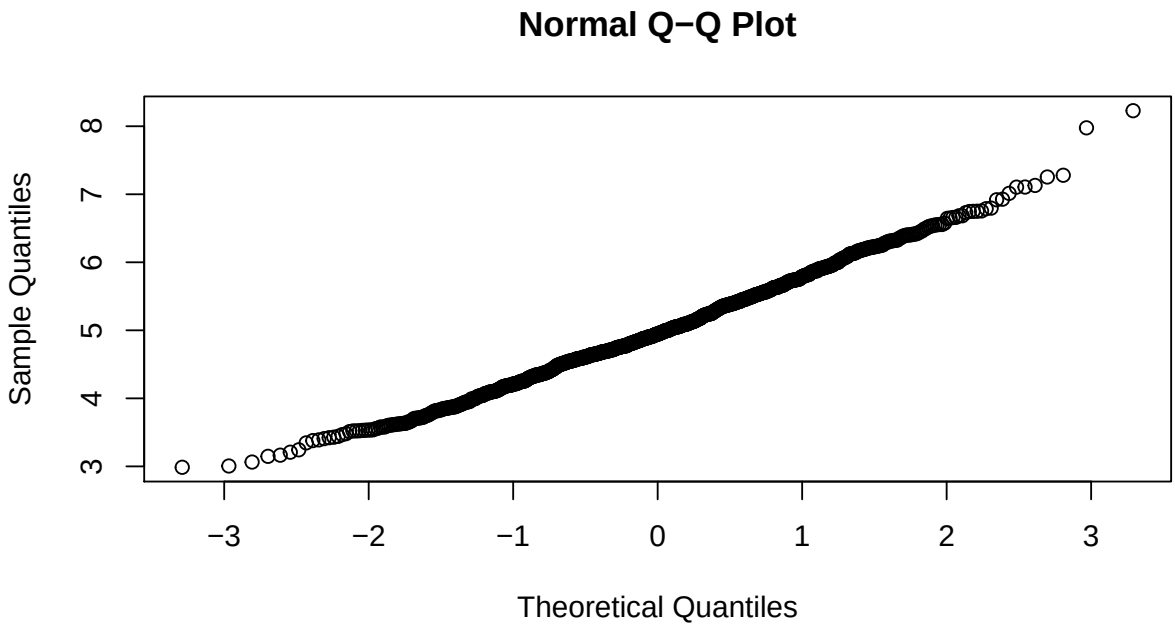
```r
c(t.test(sim_means)$conf.int[1], t.test(sim_means)$conf.int[2])
```

```
## [1] 4.948121 5.045546
```

Distribution and Comparisons between Sample and Theoretical Values

The normality of the distribution could be examined with a normal Q-Q plot as shown below:

```r
qqnorm(sim_means)
```



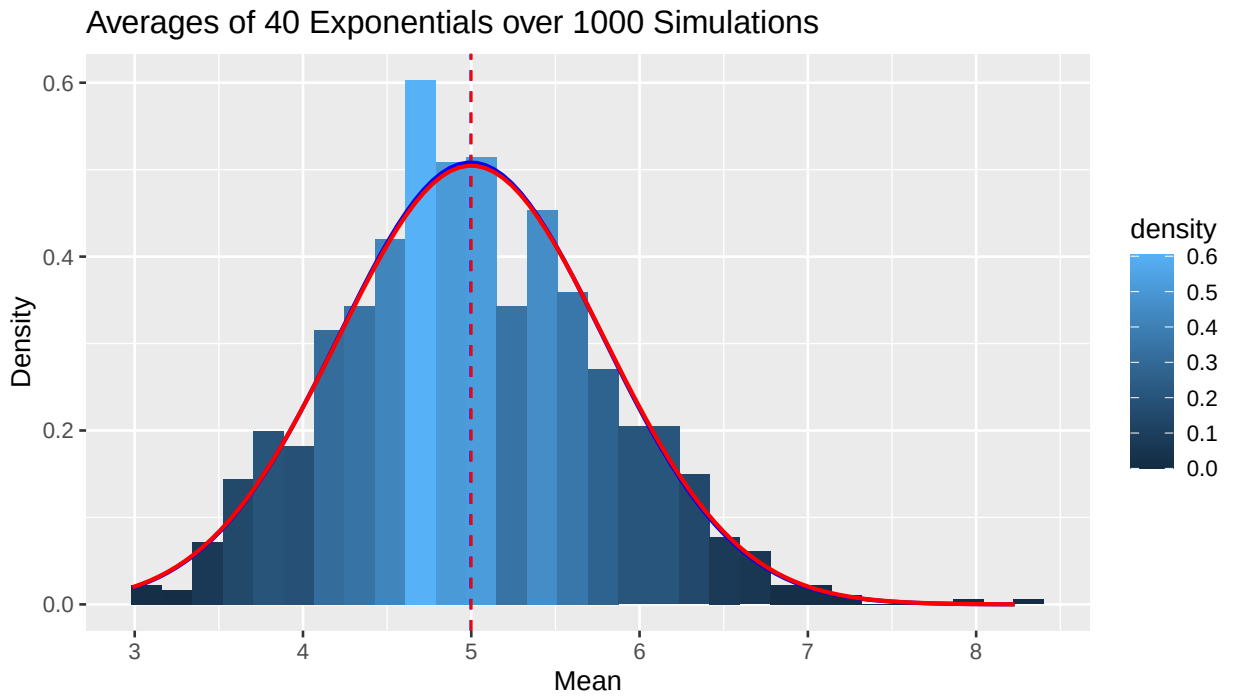
It is seen that the line is approximately straight along the diagonal line, which indicates that the means are indeed approximately normally distributed.

The probability density functions of the sample and theoretical distributions are plotted as below:

```
# Need to first turn the data into a dataFrame for ggplot to work
sim_means <- data.frame(sim_means)

# Compare the sample and theoretical statistics
ggplot(data=sim_means, mapping=aes(x=sim_means)) +
  geom_histogram(mapping=aes(y=..density.., fill=..density..)) +
  geom_vline(xintercept=samp_mean, linetype="dashed", col="blue", size=0.5) +
  geom_vline(xintercept=theo_mean, linetype="dashed", col="red", size=0.5) +
  stat_function(fun=dnorm, args=list(mean=samp_mean, sd=sqrt(samp_var)), size=0.8, col="blue") +
  stat_function(fun=dnorm, args=list(mean=theo_mean, sd=sqrt(theo_var)), size=0.8, col="red") +
  labs(title="Averages of 40 Exponentials over 1000 Simulations", y="Density", x="Mean")
```

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



The blue and red lines refer to sample and theoretical distributions. The proximity between the theoretical and the sample means could be witnessed from the nearly **overlapped** red and blue **dashed lines** at around 5 along with the 2 density curves that resemble each other in terms of **spread** and **shape**.

In conclusion, the distribution of means of 40 exponentials closely approximates the normal distribution with the expected theoretical statistics. This is in accordance to the **Central Limit Theorem** which states that the distribution of sample means approximates a normal distribution as the sample size increases.This is under the assumption that the samples are **identically and independently distributed**.