Statistical Inference Course Project - A simulation exercise

Yuan Liao

8/26/2020

Overview

This is an R Markdown document for Statistical Inference Course Project - part 1. Here, we investigate the exponential distribution in R and compare it with the Central Limit Theorem.

```
# Load necessary libs
library(ggplot2)
```

Simulations

Do the simulation to calculate the mean value of 1000 randomly generated exponential distributions.

```
# Define the distribution
lambda <- 0.2
n <- 40
sim_num <- 1:1000
set.seed(5)

# Simulate for 1000 times
df_sim <- data.frame(
  t(
    sapply(
      1:1000,
      function(x){c(x, mean(rexp(n, lambda)), var(rexp(n, lambda)))}
      )
    )
  )
colnames(df_sim) <- c('sim_#', 'mean', 'var')
```

Sample Mean versus Theoretical Mean

Calculate the sample mean and theoretical mean. They are close.

```
# Get theoretical mean
theo.mean <- 1/lambda

# Get sample mean
sample.mean <- mean(df_sim$mean)

print(cbind(theo.mean, sample.mean))
```

```
##      theo.mean sample.mean
## [1,]         5    5.022343
```

Sample Variance versus Theoretical Variance

Calculate the sample variance and theoretical variance. They are close.

```
# Get theoretical variance
theo.var <- (1/lambda)^2

# Get sample variance mean
sample.var <- mean(df_sim$var)

print(cbind(theo.var, sample.var))
```

```
##      theo.var sample.var
## [1,]       25   25.51103
```

Distribution

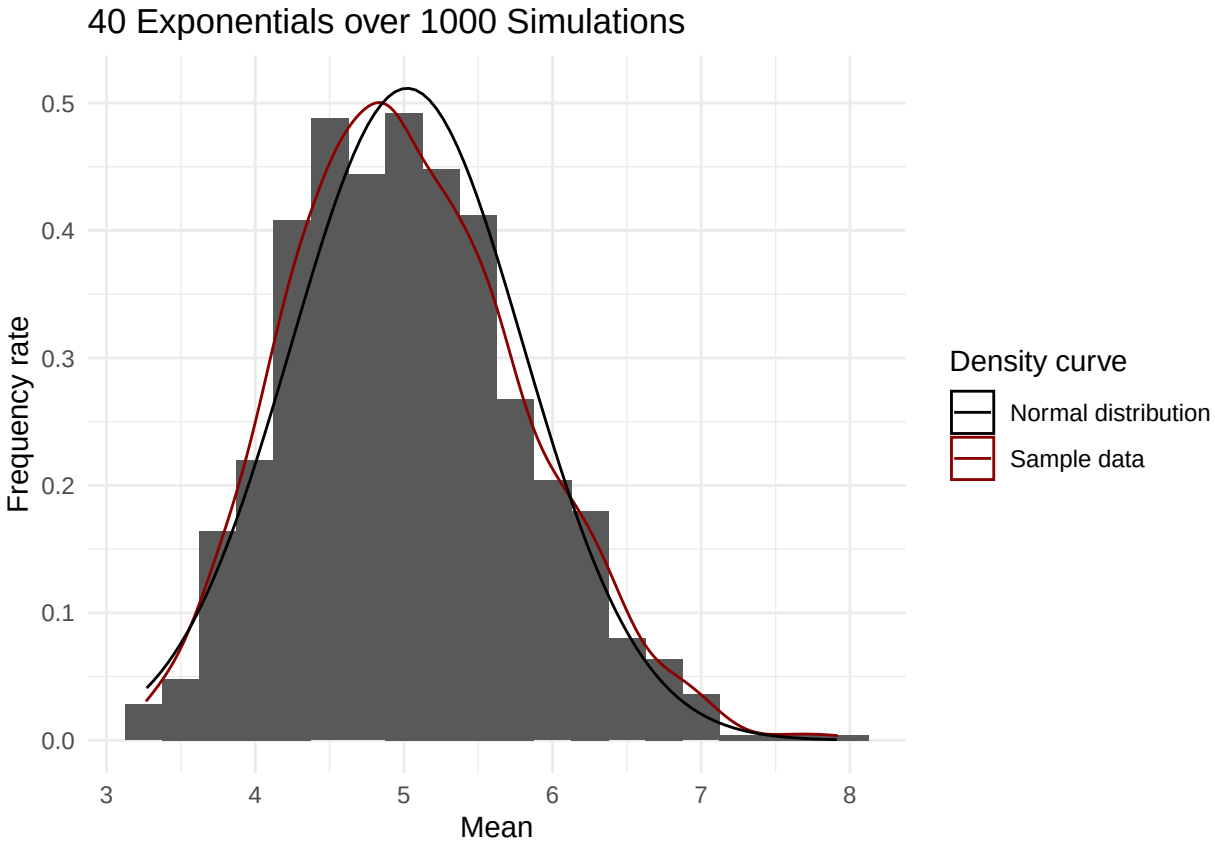
One can tell the distribution is approximately normal. We first generate a Gaussian distribution with the same mean and variance as the simulated mean results.

```
# Get sample mean
sample.mean <- mean(df_sim$mean)

# Get variance of sample mean
sample.mean.var <- var(df_sim$mean)
```

Visualize the results where the sample density curve looks similar to the normal distribution with the sample mean and standard deviation.

```
# Visualize the results
ggplot(data = df_sim, aes(x=mean)) +
  geom_histogram(aes(y=..density..), binwidth = 0.25) +
  geom_density(aes(color = 'Sample data')) +
  stat_function(fun=dnorm,
                args=list(mean=sample.mean,
                          sd=sqrt(sample.mean.var)),
                aes(colour = "Normal distribution")) +
  scale_colour_manual('Density curve', values=c('black', 'darkred')) +
  labs(title='40 Exponentials over 1000 Simulations',
       x='Mean',
       y='Frequency rate') +
  theme_minimal()
```



Conclusions

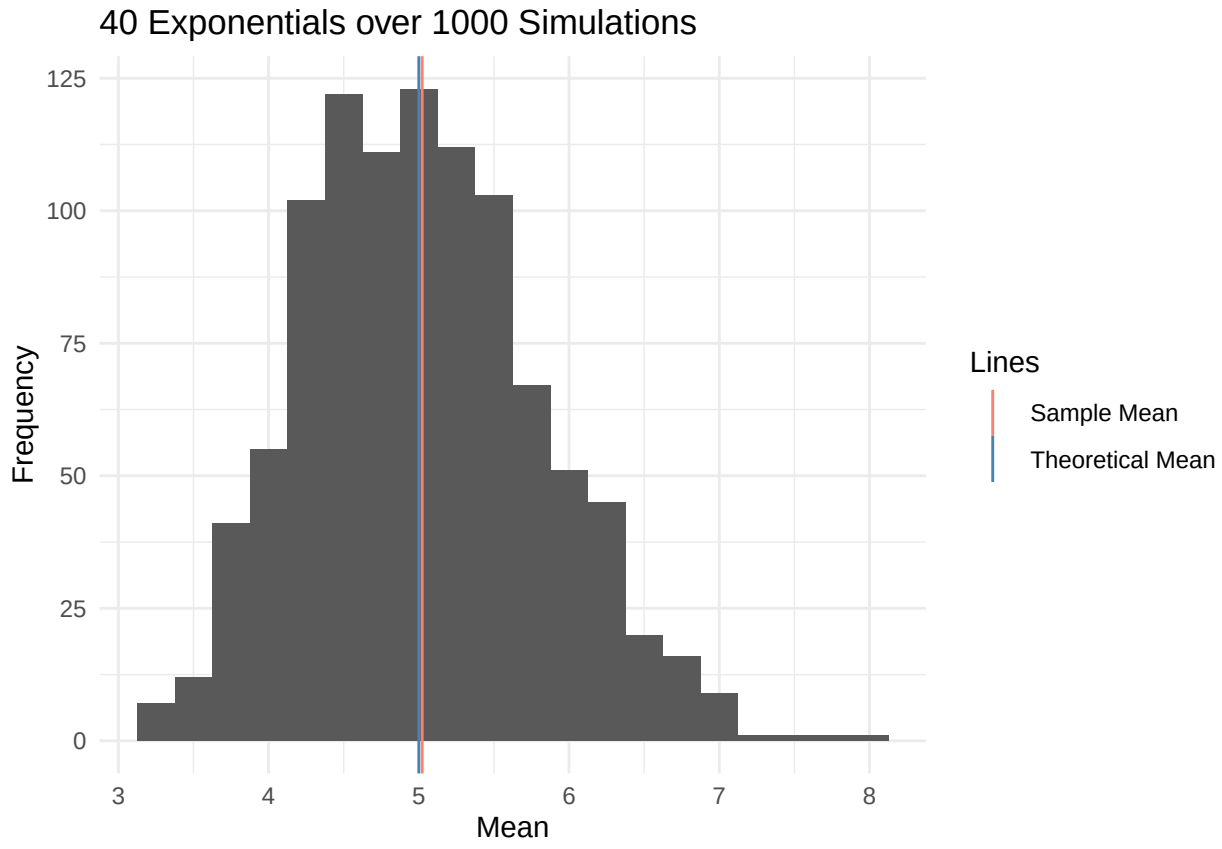
Given the assumption that the sample represents the population, the results confirm the Central Limit Theorem.

Appendix

Visualize the results where the sample mean and theoretical mean are illustrated.

```
# Visualize the results
line.data <- data.frame(xintercept = c(theo.mean, sample.mean),
                        Lines = c("Theoretical Mean", "Sample Mean"),
                        color = c("salmon", "steelblue"),
                        stringsAsFactors = FALSE)

ggplot(data = df_sim, aes(x=mean)) +
  geom_histogram(binwidth = 0.25) +
  geom_vline(aes(xintercept = xintercept, color=Lines), line.data) +
  scale_colour_manual(values = line.data$color) +
  labs(title='40 Exponentials over 1000 Simulations',
       x='Mean',
       y='Frequency') +
  theme_minimal()
```



Visualize the results where the sample variance and theoretical variance are illustrated.

```
# Visualize the results
line.data <- data.frame(xintercept = c(theo.var, sample.var),
                        Lines = c("Theoretical Variance", "Sample Variance"),
                        color = c("salmon", "steelblue"),
                        stringsAsFactors = FALSE)
```

```r
ggplot(data = df_sim, aes(x=var)) +
  geom_histogram(binwidth = 5) +
  geom_vline(aes(xintercept = xintercept, color=Lines), line.data) +
  scale_colour_manual(values = line.data$color) +
  labs(title='40 Exponentials over 1000 Simulations',
       x='Variance',
       y='Frequency') +
  theme_minimal()
```