Statistical Inference Course Project - Simulation Exercise

Gary Lu

January 7, 2021

Contents

Overview

The exponential distribution is the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate. The mean or expected value of an exponentially distributed random variable X with rate parameter λ is given by $E[X] = \frac{1}{\lambda}$. The variance of X is given by $Var[X] = \frac{1}{\lambda^2}$. [see https://en.wikipedia.org/wiki/Exponential_distribution]

In this part, we will investigate the exponential distribution in R and compare it with the Central Limit Theorem.

Simulations

As required in the instruction, we will carry out 1000 simulations of the exponential distribution that has $\lambda = 0.2$, in each simulation, we will investigate the mean and the variance of 40 samples.

```
# Define constants
lambda <- 0.2
num_of_sim <- 1000
sample_size <- 40
```

The population mean of the distribution $\mu = \frac{1}{0.2} = 5$:

```
pop_mean <- 1/lambda
pop_mean
```

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

The population variance of the distribution $\sigma^2 = \frac{1}{0.2^2} = 25$

```
pop_var <- 1/(lambda^2)
pop_var
```

```
## [1] 25
```

In the next step, we use `rexp()` to generate a matrix `matrix_data` of 1000 rows and 40 columns, in which each row stands for one simulation which has 40 samples.

```
sample_data <- rexp(sample_size * sample_size, rate=lambda)
matrix_data <- matrix(sample_data, nrow=num_of_sim, ncol=sample_size)
```

Sample Mean versus Theoretical Mean

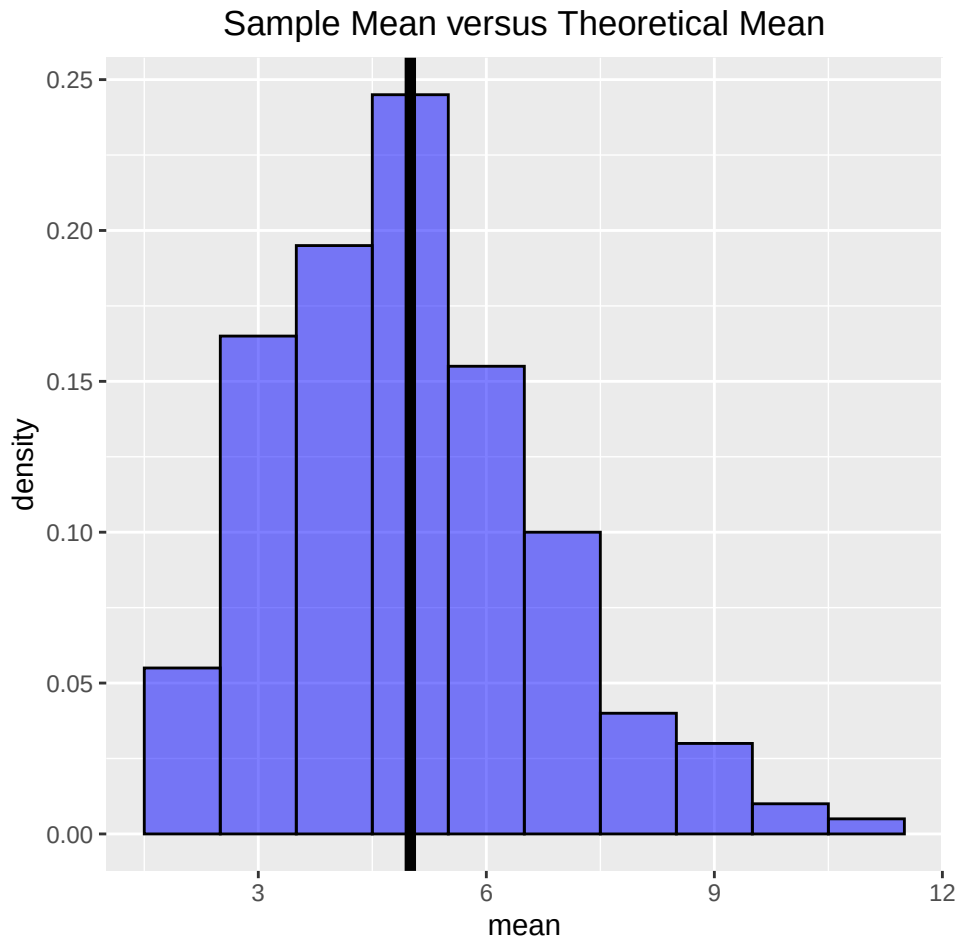
We can calculate the sample mean of each simulation, and we can see that the mean of the sample means is pretty close to the theoretical mean of the distribution 5.

```
means <- c(apply(matrix_data, 1, mean))
mean(means)
```

```
## [1] 4.98972
```

Here is the plot comparing the sample means of each simulation and the population mean. The histogram shows the distribution of the sample means of 1000 simulations, and the vertical line shows the population mean $\mu = 5$.

```
dat <- data.frame(mean = means)
g <- ggplot(dat, aes(x = mean)) +
  geom_histogram(alpha =.5, binwidth=1, colour = "black", fill="blue", aes(y = ..density..)) +
  ggtitle("Sample Mean versus Theoretical Mean") +
  theme(plot.title = element_text(hjust = 0.5))
g + geom_vline(xintercept=pop_mean, color = "black", size=2)
```



Sample Variance versus Theoretical Variance

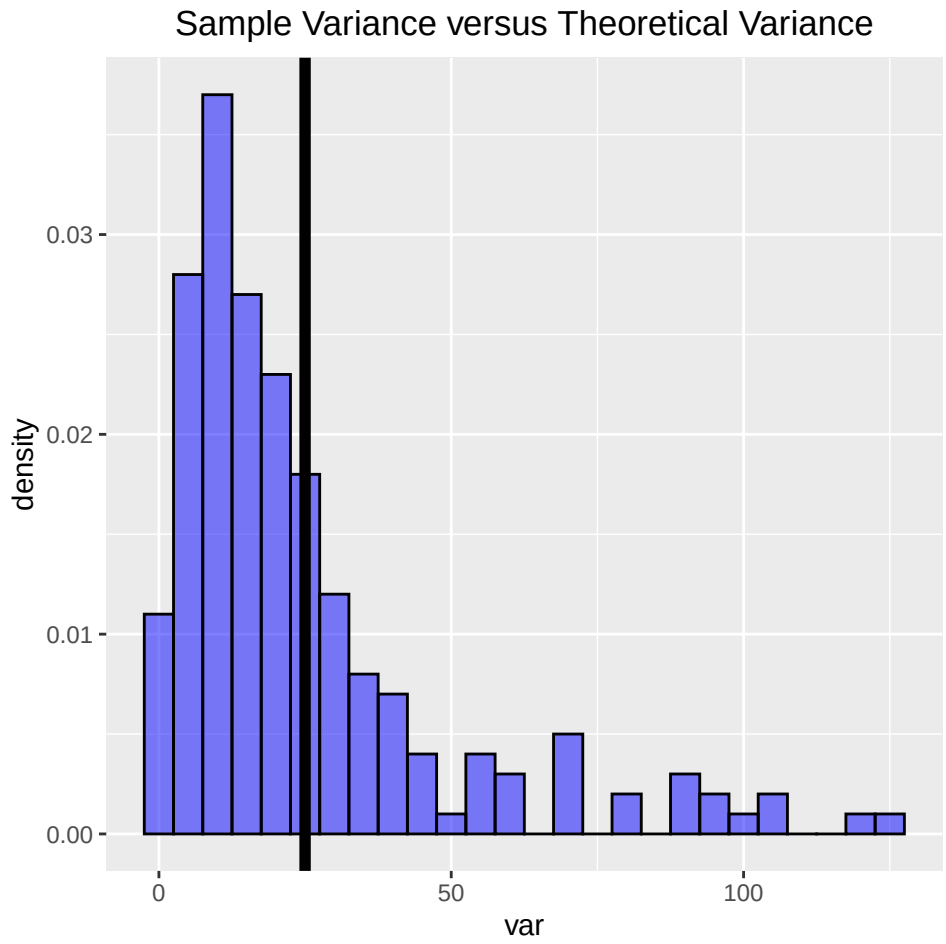
Also, we can calculate the sample variance of of each simulation. We can see that the mean of the sample variances is also not far away from to the theoretical variance of the distribution 25.

```
variances <- c(apply(matrix_data, 1, var))
mean(variances)
```

```
## [1] 24.61412
```

Here is the plot comparing the sample variances of each simulation and the population variance. The histogram shows the distribution of the sample variances of 1000 simulations, and the vertical line shows the population variance $\sigma^2 = 25$.

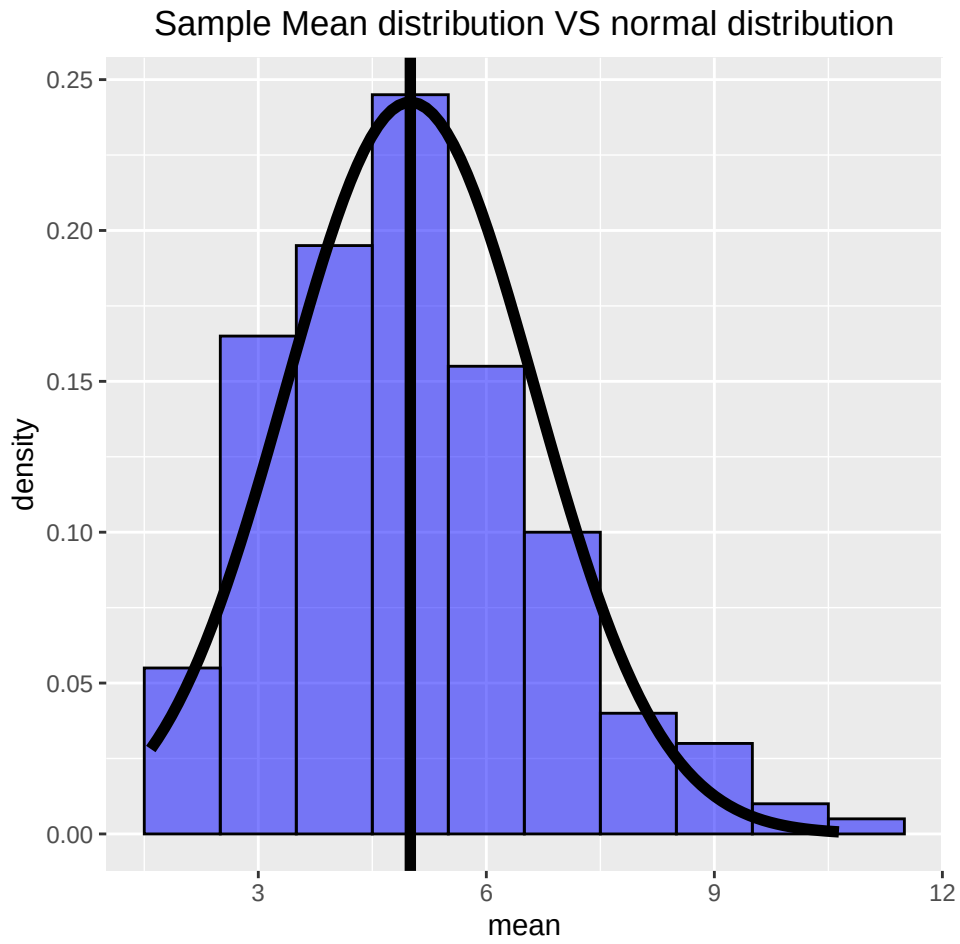
```
dat <- data.frame(var = variances)
g <- ggplot(dat, aes(x = var)) +
  geom_histogram(alpha =.5, binwidth=5, colour = "black", fill="blue", aes(y = ..density..)) +
  ggtitle("Sample Variance versus Theoretical Variance") +
  theme(plot.title = element_text(hjust = 0.5))
g + geom_vline(xintercept=pop_var, color = "black", size=2)
```



Distribution

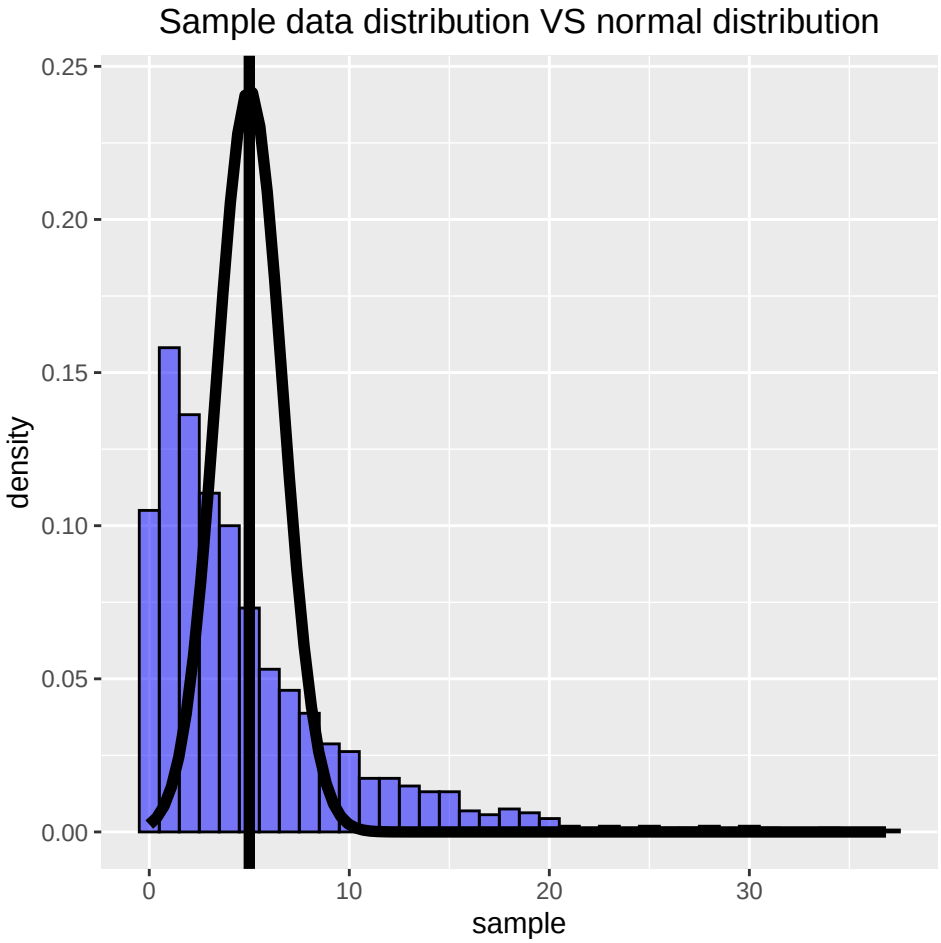
Now, let's discuss the distribution of the sample means. According to Central Limit Theorem (CLT), as we carry out a large number of simulations, the distribution of the sample means will closely approximate a normal distribution. We can observe this from the previous plot that the distribution of the sample means is in a bell shape. And in the following plot, we also draw a normal distribution line over the histogram so that we can see it more obviously.

```
dat <- data.frame(mean = means)
g <- ggplot(dat, aes(x = mean)) +
  geom_histogram(alpha =.5, binwidth=1, colour = "black", fill="blue", aes(y = ..density..)) +
  ggtitle("Sample Mean distribution VS normal distribution") +
  theme(plot.title = element_text(hjust = 0.5))
g +
  geom_vline(xintercept=pop_mean, color = "black", size=2) +
  stat_function(fun = dnorm, size = 2, args=list(mean=pop_mean, sd=1.645))
```



As a contrast, if we draw the distribution of the original sample data, we can see that it is skewed and nowhere close to a normal distribution.

```
dat <- data.frame(sample = sample_data)
g <- ggplot(dat, aes(x = sample)) +
  geom_histogram(alpha =.5, binwidth=1, colour = "black", fill="blue", aes(y = ..density..)) +
  ggtitle("Sample data distribution VS normal distribution") +
  theme(plot.title = element_text(hjust = 0.5))
g +
  geom_vline(xintercept=pop_mean, color = "black", size=2) +
  stat_function(fun = dnorm, size = 2, args=list(mean=pop_mean, sd=1.645))
```



In conclusion, we can see from the 1000 simulations of the exponential distribution:

- The mean of the sample means is close to the theoretical mean;
- The mean of the sample variances is close to the theoretical variance;
- The distribution of the sample means approximates to a normal distribution, despite that the sample data are not normally distributed.