Project Part 1

Statistical Inference: Simulation Exercise

Submitted By: Shivam Singh Baghel

Overview & Objectives:

- Exponential distribution in R and comparison with the Central Limit Theorem.
- A thousand simulations of the distribution of 40 exponentials would be investigated.

Project-Work:

Part-A:

For the exponential simulation in R we will be using R with rexp(n,λ). Here n is the number of observation and λ is the rate parameter. Here we will be taking λ=0.2 for our work.

Following steps are adopted for the simulations:

Step-1: Loading the ggplot2 plotting library in R.

Step-2: Initializing the simulation controlling variables.

Step-3: Making the analysis is reproducible: Setting the seed of the Random Number Generator.

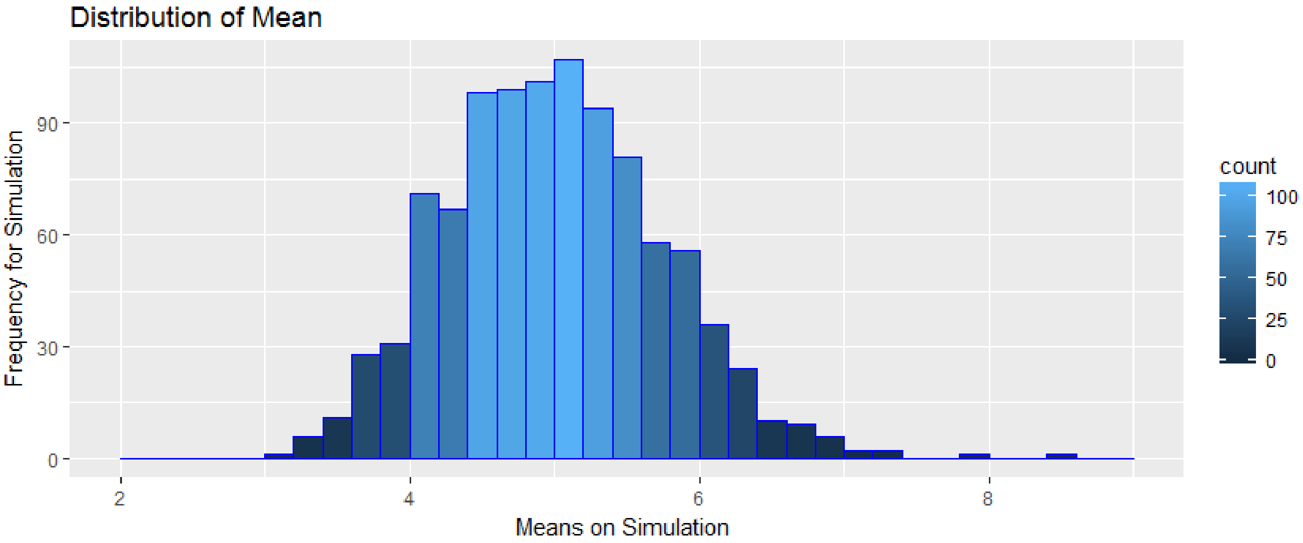
Step-4: Creating a matrix with forty columns corresponding to each of 40 random simulations and thousand rows corresponding to 1000 simulations.

Step-5: Creating a vector of thousand rows containing the mean of each row of the sim_matrix.

Step-6: Creating a data frame containing the data from step4 and step-5.

Step-7: Plotting the simulation data to visualize it.

Sample Mean Versus Theoretical Mean:



The actual mean of the simulated mean sample data is 4.9866197 and the theoretical mean is 5. Thus, we can see that the actual mean of the simulated mean sample data is very close to the theoretical mean of original data distribution.

Sample Variance Versus Theoretical Variance:

The actual variance of the simulated mean sample data is 0.6257575 and the theoretical variance is 0.625. Thus, we can see that the actual variance of the simulated mean sample data is very close to the theoretical variance of original data distribution.

Distribution:

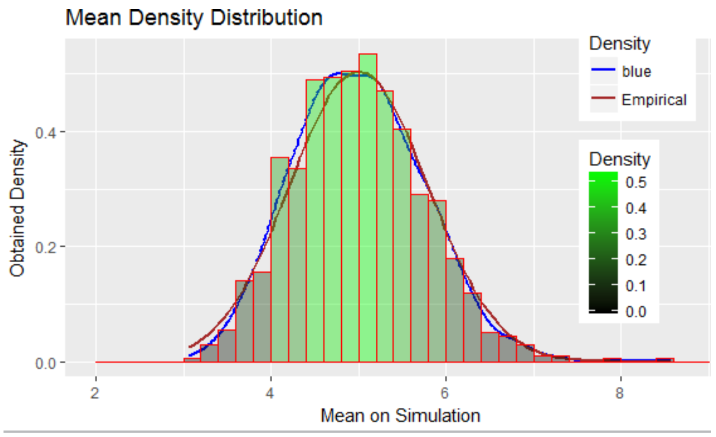
To prove that the simulated mean sample data approximately follows the Normal distribution, we perform the following three steps:

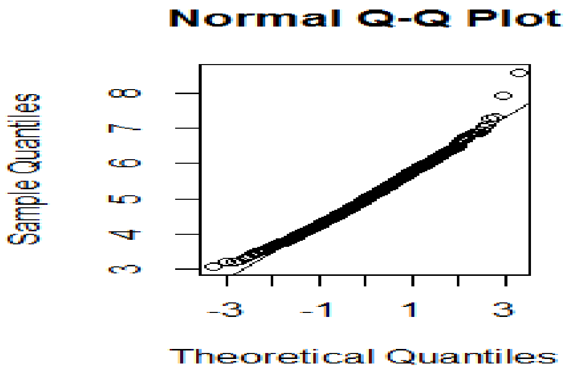
Part-B:

Step 1: Create an approximate normal distribution and see how the sample data alligns with it and from the obtained histogram, the simulated mean sample data can be adequately approximated with the normal distribution.

Step 2: Compare the 95% confidence intervals of the simulated mean sample data and the theoretical normally distributed data. From the obtained result we can see that actual 95% confidence interval is [4.7414712, 5.2317681] and theoretical 95% confidence interval is [4.755, 5.245] and we see that both of them are approximately same.

Step 3: q-q Plot for Qunatiles: The actual quantiles also closely match the theoretical quantiles, hence the above three steps prove that the distribution is approximately normal.





Used R Programming Codes:

Used R-Programming Codes (Part-A)

```
library(ggplot2)
no_sim<-1000
sample_size<-40
lambda <- 0.2
set.seed(3)
sim_matrix <- matrix(rexp(n = no_sim * sample_size,
rate = lambda), no_sim, sample_size)
sim_mean <- rowMeans(sim_matrix)
sim_data <- data.frame(cbind(sim_matrix, sim_mean))
ggplot(data = sim_data, aes(sim_data$sim_mean)) +
geom_histogram(breaks = seq(2, 9, by = 0.2), col =
"blue", aes(fill = ..count..)) +labs(title = "Mean
Distribution", x = "Simulation Means", y =
"Frequency") +
geom_vline(aes(xintercept=mean(sim_data$simulatio
```

Used R-Programming Codes (Part-B)

```
actual_mean <- mean(sim_mean)
theoretical_mean <- (1 / lambda)
actual_variance <- var(sim_mean)
theoretical_variance <- ((1 / lambda) ^ 2) / sample_size
qplot(sim_mean, geom = 'blank') + geom_line(aes(y=..density..,
colour='blue'), stat='density', size=1) + stat_function(fun=dnorm,
args=list(mean=(1/lambda),
sd=((1/lambda)/sqrt(sample_size))),aes(colour='Empirical'),
size=1) + geom_histogram(aes(y=..density.., fill=..density..),
alpha=0.4,breaks = seq(2, 9, by = 0.2), col='red') +
scale_fill_gradient("Density", low = "black", high = "green") +
scale_color_manual(name='Density', values=c('blue', 'brown')) +
theme(legend.position = c(0.85, 0.60)) +labs(title = "Mean Density
Distribution", x = "Mean on Simulation", y = "Obtained Density")
qqnorm(sim_mean)
qqline(sim_mean)
```