Statistical Inference Course Project (part 1)

Overview

In this project we investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT). The exponential distribution can be simulated in R with `rexp(n, lambda)` where lambda λ is the rate parameter. The mean of exponential distribution is $\mu = \frac{1}{\lambda}$ and the standard deviation is also $\mu = \frac{1}{\lambda}$. Set $\lambda = 0.2$ for all of the simulations. We will investigate the distribution of averages of 40 exponentials with 1000 simulations.

The Central Limit Theorem is among the most important theorems in statistics. The CLT states that the distribution of averages of iid properly normalized variables becomes that of a standard normal as the sample size increases.

A useful way to think about the CLT is that $\bar{x}_n$ is approximately $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Simulations comparing the sample mean vs. the theoretical mean

We will run 1000 simulations of 40 exponentials from a $Exp(\frac{1}{0.2}, \frac{1}{0.2})$ distribution, and we will calculate the mean of each of the 1000 samples. This will give us the same theoretical result as if we drew an actual sample size of 1000 with $N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$.

Taking into account the parameters of our 1000 simulations, CLT suggests that we will find that the means of each simulation are approximately $\frac{1}{\lambda} = \frac{1}{0.2} = 5$. Thus, we expect our sample mean to be approximately 5.

First, we create 1000 simulations of 40 exponentials and calculate their means.

```r
set.seed(123)
lambda <- 0.2
n <- 1000
sample <- 40
simulation <- matrix(rexp(n * sample, rate = lambda), n, sample)
Row_Means <- rowMeans(simulation)
round(mean(Row_Means), 2)
```

```
## [1] 5.01
```

We find that $\bar{x}$ equals `[1] 5.01`. This is consistent with the theoretical mean, which is $\mu = \frac{1}{0.2} = 5$.

Comparison of the sample variance and the theoretical variance

The CLT suggests that that the variance of our sample of 1000 simulations of 40 exponentials should be $\frac{\frac{1}{0.2^2}}{40} = 0.625$.

```r
round(var(Row_Means), 2)
```

```
## [1] 0.61
```

Our result, `[1] 0.61` is very close to the expected variance, so we may conclude that our simulation is in line with the CLT.

Distribution of sample mean vs. theoretical mean

We now look at the distribution of sample means.

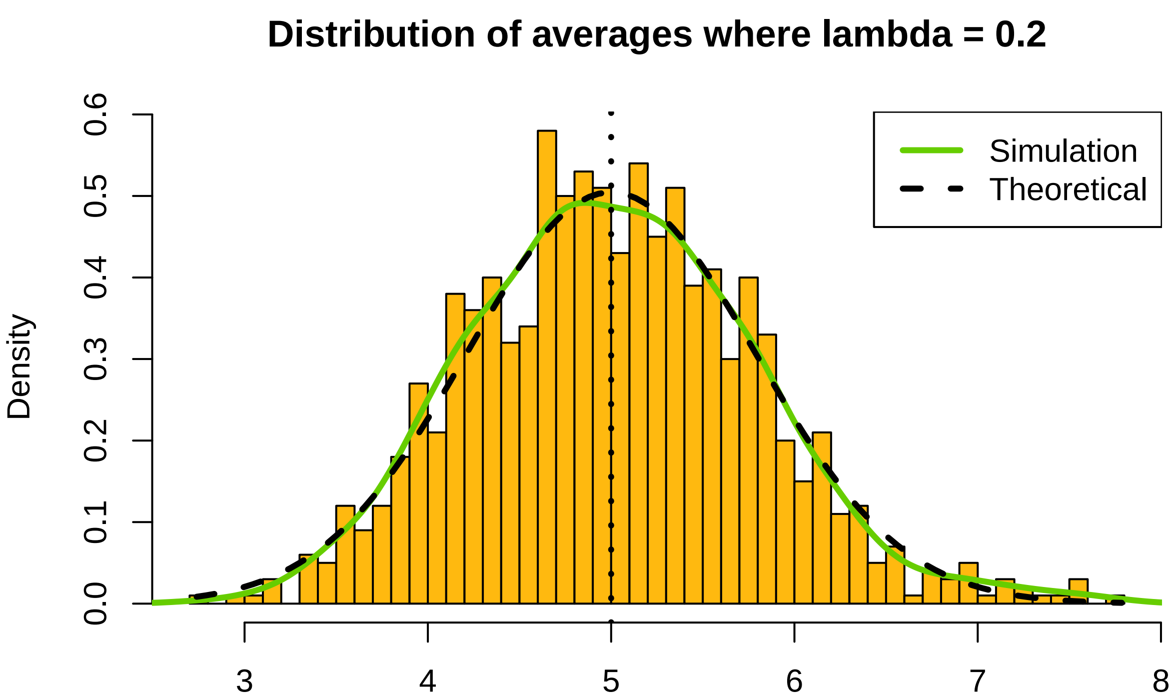
```
# Plot a histogram of the averages
hist(Row_Means, breaks = 50, prob = TRUE,
     main = "Distribution of averages where lambda = 0.2",
     xlab = "", col = "darkgoldenrod1")

# Find the density of the averages of the sample
lines(density(Row_Means), lwd = 3, col = "chartreuse3")

# Include the theoretical center of distribution
abline(v = 1/lambda, col = "black", lty = 3, lwd = 3)

# Find the theoretical density of averages
xfit <- seq(min(Row_Means), max(Row_Means), length=100)
yfit <- dnorm(xfit, mean = 1/lambda, sd = (1/lambda/sqrt(sample)))
lines(xfit, yfit, pch=22, col="black", lwd = 3, lty = 2)

legend("topright", c("Simulation", "Theoretical"), lty = c(1, 2),
       lwd = 3, col = c("chartreuse3", "black"))
```



This figure reflects that the mean of the simulation is equal to 5 and, thus, that it parallels the theoretical mean consistent with the CLT. We also find that the distribution of the sample means approximates that of the CLT.

Conclusion

Our analysis demonstrated that the sample distribution of the mean where exponential values $n = 40$ and $\lambda = 0.2$ with 100 simulations approximates that as determined by the CLT, or $N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$, with a variance approximately equal to that expected according to the CLT.