Comparison of Sample Mean Vs Theoretical Mean for Exponential Distribution and the CLT Theorem

Sukdev Koner

Sept 22, 2018

Overview

In this project, we will explore the Exponential DistributioIn in R and compare it with the Central Limit Theorem (CLT).

The exponential distribution can be simulated in R with function `rexp(n, lambda)` where lambda (λ) represents the rate parameter. The mean (μ) is $\frac{1}{\lambda}$ and the standard deviation(σ) is also $\frac{1}{\lambda}$.

Simulations

In this exercise, I have run 1000 simulations, each with 40 samples from an exponential distribution. For each simulation, I have calculated the "sample mean". Then from the Distribution of those **sample means**, 'Mean' and 'Variance' will be calculated and compared to the theoretical Mean and Variances. Central limit theorem (CLT) states that as the number in the sample increases or the number of simulations increases, the sample mean and variance should more accurately represent the theoretical estimate(population parameter).

Here I will validate if the 'sampling distribution of the **Means**' of an exponential distribution with $n = 40$ observations and $\lambda = 0.2$ is approximately Normally distributed or not.

1. Comparison of Sample Mean and Theoretical Mean of the distribution

In this assignment, I will draw 1000 samples of size 40 from an $Exp(n = 40, lambda = 0.2)$ distribution. For each of the 1000 samples, I would calculate the **mean**. Theoretically, it is same as drawing a single sample of size 1000 from the Normal Dstribution with $\sim N(mean = \frac{1}{0.2}, var = \frac{\frac{1}{0.2^2}}{40})$.

According to the CLT, I should expect, my final Sample Mean of those '1000 Simulated Means' is approximately $\frac{1}{\lambda} = \frac{1}{0.2} = 5$. Since I am caluculating the **mean** of "1000 sampled means", I will expect the output to be very close to 5.

Let's check if this is the case. First, I have set the seed so the same random data can be reproduced.

```r
set.seed(12345)
samplemeans <- NULL
for(i in 1:1000) {
  samplemeans <- c(samplemeans, mean(rexp(40, 0.2)))
}
xbar <- mean(samplemeans)
xbar
```

```
## [1] 4.971972
```

Now let's calculate the Theoretical Mean.

```r
mu <- 1/ 0.2  # Theoretical mean is 1/lambda
mu
```

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

So Mean of the 'Sample Means' ($\bar{x}$) in our case is 4.972 which is very close to the mean of the theoretical distribution $\mu = \frac{1}{0.2} = 5$.

2. Comparison of the sample variance with theoretical variance

According to the CLT, Variance of distribution of 'Sample Means' is $\frac{\sigma^2}{n}$. So I would expect that the variance of the sample of the 1000 means is approximately $\frac{\frac{1}{0.2^2}}{40} = 0.625$.

Now let's see what Sample Variance is :

```
var(samplemeans)
```

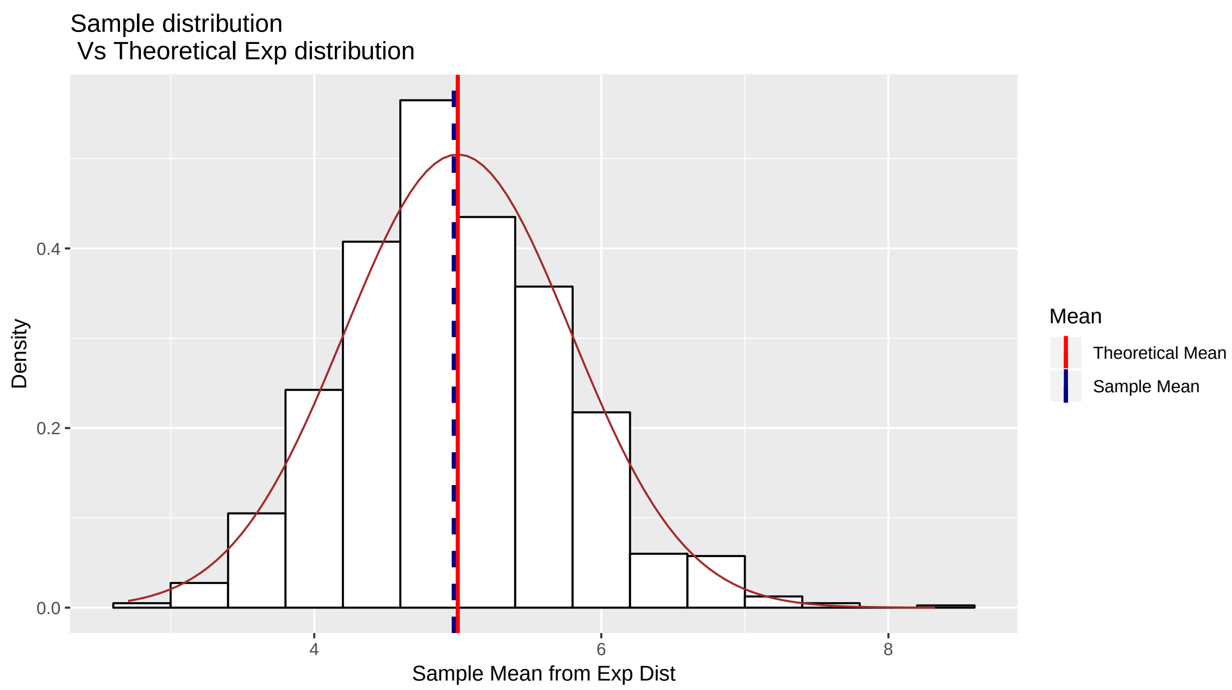
```
## [1] 0.5954369
```

So Variance of sample means in our case is 0.595 which is close to the variance of the theoretical distribution (0.625), as mentioned above.

3. Showing that the sample distribution is approximately normal

In order to validate, that the sample distribution of the 1000 sampled means is approximately normal, I will plot the histogram for the distribution of 'Sample Means' and then impose one red curve which is from Density function of Normal Distribution.

```
data <- as.data.frame(samplemeans)
ggplot(data, aes(x = samplemeans)) +
  geom_histogram(binwidth = 0.4, color = 'black', fill = 'white', aes(y = ..density..)) +
  stat_function(fun = dnorm, color = 'brown',
                args = list(mean = mu, sd = sqrt(0.625))) +
  geom_vline(aes(xintercept = mu, col = "mucol"), lwd = 1) +
  geom_vline(aes(xintercept = xbar, col = "xbarcol"), lwd = 1, lty = 2) +
  scale_colour_manual(name = "Mean", values = c(mucol = "red", xbarcol = "darkblue"), labels = c("Theore
  xlab('Sample Mean from Exp Dist') +
  ylab('Density') +
  ggtitle('Sample distribution\n Vs Theoretical Exp distribution')
```



Conclusion

In this project, we have validated that the sampling distribution of the mean of an exponential distribution with $n = 40$ observations and $\lambda = 0.2$ is approximately Normal with $\sim N(mean = \frac{1}{0.2}, var = \frac{\frac{1}{0.2^2}}{40})$ distributed. The complete assignment of this report can be found on GitHub.