Statistical Inference Course Project Part 1

Abhishek Kumar

8 August 2020

Overview

This is a project report for Statistical Inference course on Coursera. The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

In Part 1, the exponential distribution is investigated and compared with the Central Limit Theorem. This simulation is useed to illustrate the properties of the distribution of the mean of 40 exponentials. Specifically, I have

i. shown the sample mean and compare it to the theoretical mean of the distribution.
ii. shown how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
iii. shown that the distribution is approximately normal.

In Part 2, the `ToothGrowth` data in R is analysed and summarised. Then, some hypotheses were tested using Confidence intervals.

Part 1: Simulation Exercise

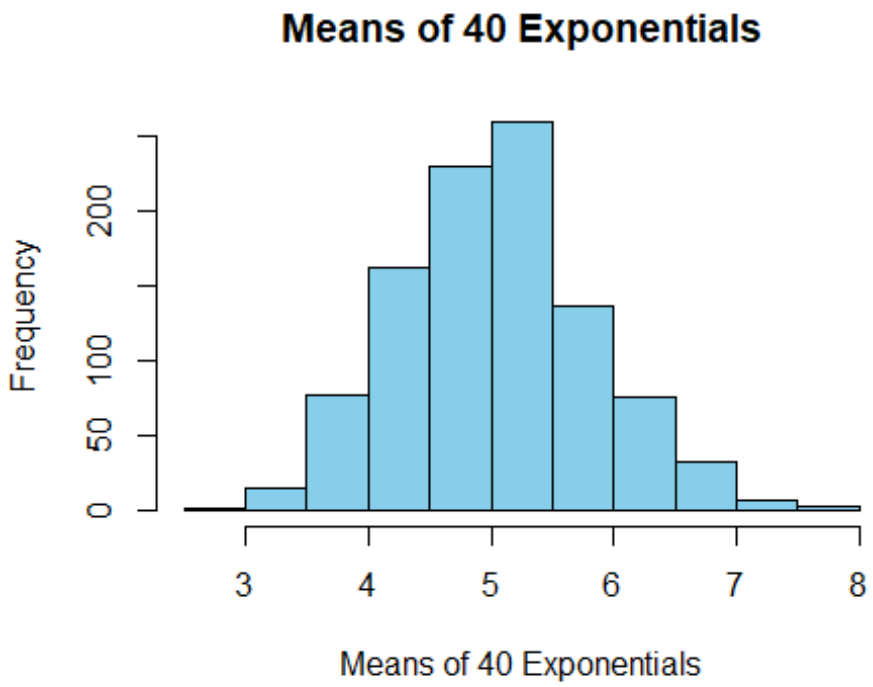
The exponential distribution is simulated in R with `rexp(n, lambda)` where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda.

For this simulation, I have defined parameters as instructed: $\lambda = 0.2; \quad n = 40; \quad Simulations = 1000$

Firstly, seed is set for reproducibility of results. Then, the parameters were assigned. The exponention simulatiion is achieved by `rexp()` function. The mean of these were calculated 1000 times by running a for loop. Finally, the results were plotted as histogram using the function `hist()`.

```
set.seed(2020) # for reproducibility
lambda <- 0.2; n <- 40; nosim <- 1000
mns <- NULL
for (i in 1 : 1000)
      mns = c(mns,
```

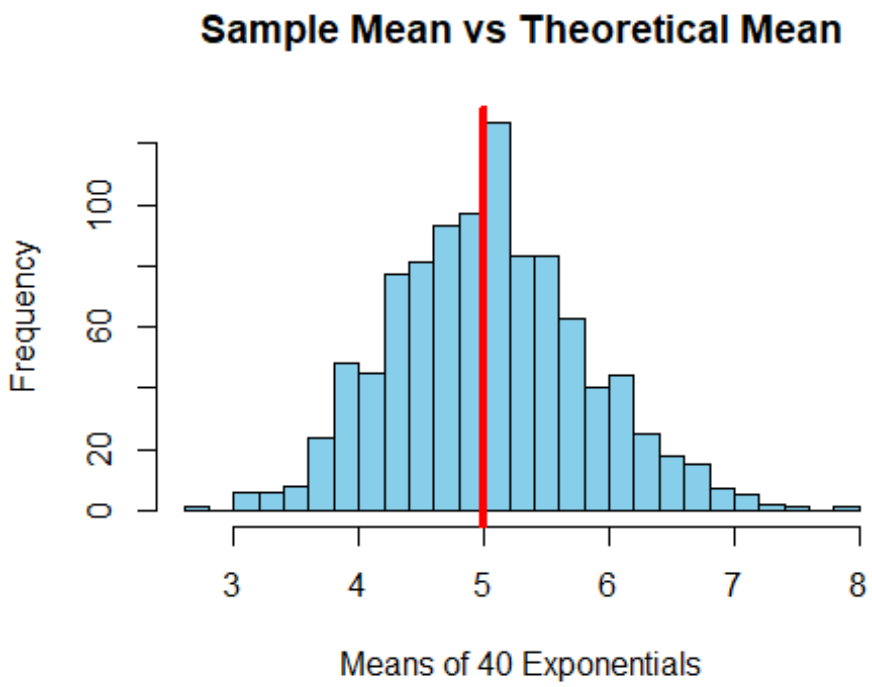
```
              mean(rexp(n, lambda))
              )
hist(mns, col = "skyblue", main = "Means of 40 Exponentials",
     xlab = "Means of 40 Exponentials")
```



i. Sample Mean versus Theoretical Mean:

As stated earlier the mean and sd for this distribution is $\frac{1}{\lambda} = \frac{1}{0.2} = 5$

```
hist(mns, breaks = 20,
     col = "skyblue", main = "Sample Mean vs Theoretical Mean",
     xlab = "Means of 40 Exponentials")
theory_mean <- 1/lambda
abline(v = theory_mean, col = "red", lwd = 4)
```



This distribution is clearly showing that sample mean is very close to theoretical mean.

```
samp_mean <- mean(mns)
round(samp_mean, 2)

## [1] 5.03

pop_mean <- 1/lambda
pop_mean

## [1] 5
```

This suggests that at large number of samples, the sample mean closely follows the population mean.

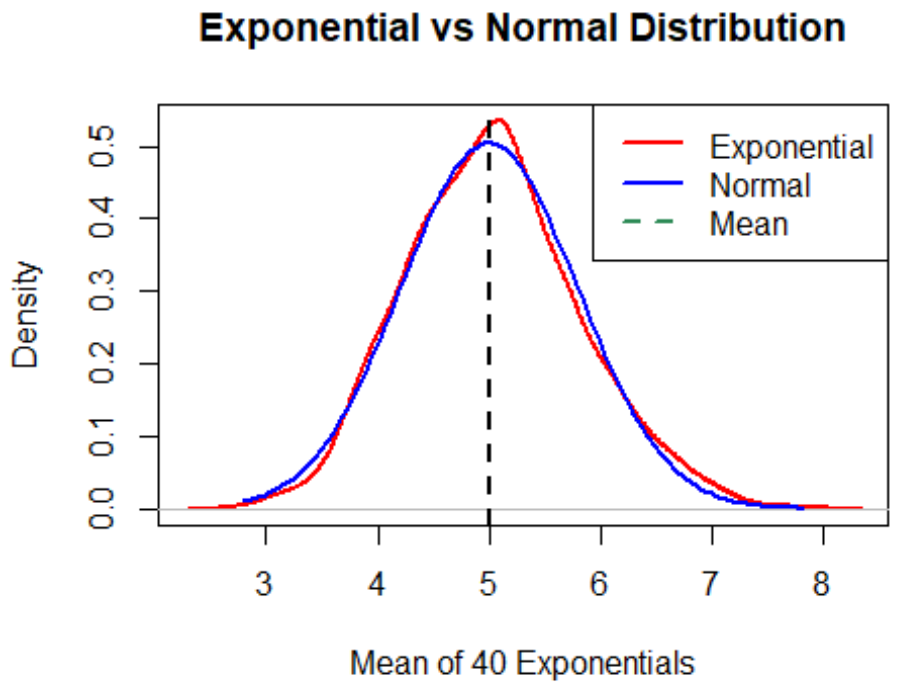
ii. Sample Variance versus Theoretical Variance:

```
samp_var <- var(mns)
round(samp_var, 2)

## [1] 0.61

theor_var <- 1/lambda^2/n
round(theor_var, 2)

## [1] 0.62
```

Again this suggests that at sufficiently larger number of samples the sample variances closely follows the population variances. So the sample variability is similar to the population variability.

iii. Distribution:

```
plot(density(mns), col = "red", lwd = 2.5, type = "l",
     main = "Exponential vs Normal Distribution",
     xlab = "Mean of 40 Exponentials")
x <- seq(min(mns), max(mns), length = 2*n)
y <- dnorm(x, mean = 1/lambda, sd = sqrt(((1/lambda)/sqrt(n))^2))
lines(x, y, pch = 20, col = "blue", lwd = 2.5)
abline(v = 1/lambda, lwd = 2.5, lty = 2.5)
legend("topright", legend = c("Exponential", "Normal", "Mean"), col =
c("red", "blue", "seagreen"),
       lty = c(1, 1, 2), lwd = 2.5)
```



This is also supporting the Central Limit Theorem. At sufficiently larger number of samples the The density distribution of Exponentials follows a Normal distribution.