Part1: Simulation Exercise

Endri Raco

4/10/2020

Overview

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

We will illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. We will:

1. Show the **sample mean** and compare it to the **theoretical mean** of the distribution.
2. Show how variable the **sample** is (via variance) and compare it to the **theoretical variance** of the distribution.
3. Show that the distribution is **approximately normal**.

Simulations

According to William Feller is a continuous distribution of a random variable defined by the density

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{për } x \geq 0 \\ 0,, & \text{për } x < 0 \end{cases} \tag{1}$$

where $\lambda > 0$ - the rate parameter. For the exponential distribution we know that:

$$\mu(X) = \frac{1}{\lambda} \qquad \sigma^2(X) = \frac{1}{\lambda^2} \quad \sigma_x = \frac{1}{\lambda} \tag{2}$$

where $\mu(x)$ is the **mean** (expectation), $\sigma^2(X)$ is the **variance** and σ_x is the **standart deviation**.

If we want for tha random variable $X \sim Exp(X)$ to have probability inside interval (a, b), we know that:

$$\begin{gathered} P\{a < X < b\} \\ = F(b) - F(a) \\ = (1 - e^{-\lambda b}) - (1 - e^{-\lambda a}) \\ = e^{-\lambda a} - e^{-\lambda b}, \text{pra } P\{a < X < b\} \\ = e^{-\lambda a} - e^{-\lambda b} \end{gathered}$$

Below there is **Central Limit Theorem** statement adapted for exponential distribution:

Let $X_i : i > 1$ be independent random variables having exponential distribution. Let μ be their mean and $\sigma^2(X)$ be their variance. Then

$$Z_n = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$$

the standardized scores , converges in distribution to $Z \sim N(0, 1)$ a standard normal random variable.

We will use $\lambda = 0.2$ so for exponential distribution: we have

theoretical mean:

$$\mu = \frac{1}{\lambda} = \frac{1}{0.2} = 5$$

theoretical variance:

$$\sigma^2 = \frac{1}{\lambda^2 n} = \frac{1}{(0.2)^2 n} = \frac{25}{n}$$

theoretical standart deviation

$$\sigma = \frac{1}{\lambda} = \frac{1}{0.2} = 5$$

Now lets start simulations. We will create a vector of length 1000. Every element of this vector will represent the sample means out of each 1000 samples. Remember that each sample has size 40 meaning that we draw random 40 elements out of population. The population distribution is exponential with parameter $\lambda = 0.2$. The exponential distribution will be simulated using R function **rexp(n, lambda)** where **lambda = 0.2** is the λ rate parameter.

```
## Initiate an empty vector s_means:
s_means = NULL
## Populate vector with mean
for (i in 1:1000) s_means = c(s_means, mean(rexp(40, rate = 0.2)))
## Let's see some of generated means
head(s_means, 10)
```

```
##  [1] 5.350569 4.770701 5.352900 5.653520 3.589472 4.995938 4.014643 4.855047
##  [9] 4.874788 6.898431
```

Sample Mean versus Theoretical Mean

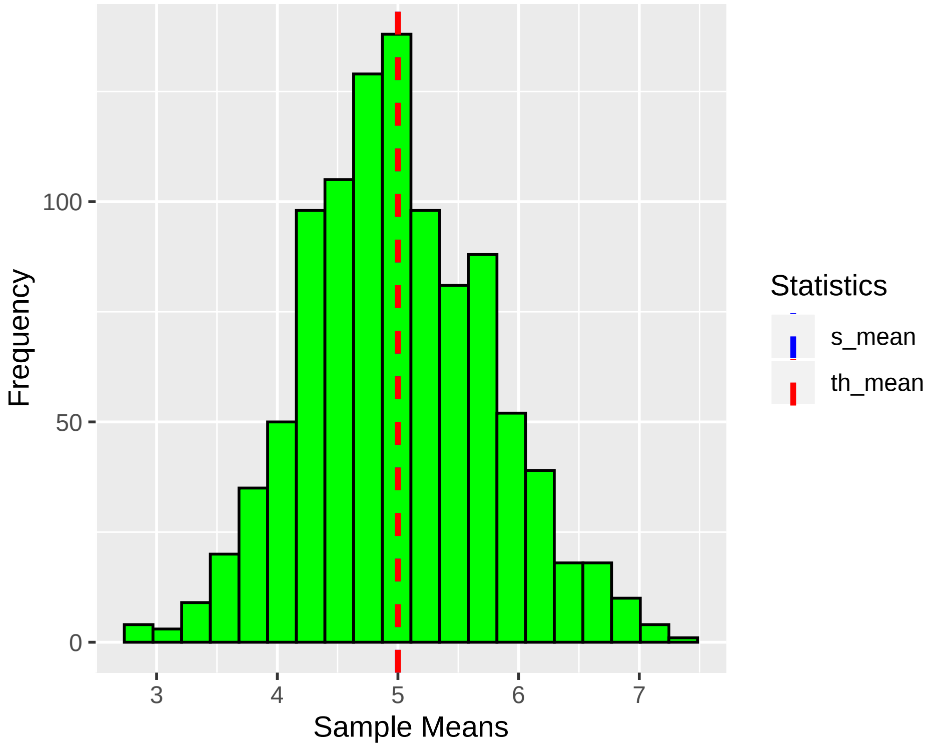
The next step is to compare **sample mean** with **_theoretical mean_. We stated before that our theoretical mean is

$$\mu = \frac{1}{\lambda} = \frac{1}{0.2} = 5$$

. We will plot the histogram of our generated sample means and see how the mean of generated values (sample mean) stands vs theoretical mean equal to 5.

```
# Create dataframe with sample means and their indexes
index <- c(1:1000)
df <- data_frame(index, s_means)
# Initiate values for theoretical and mean of generated
# values
th_mean = 5
```

```
s_mean = mean(s_means)
# Plot histogram of generated sample means
ggplot(aes(x = s_means), data = df) + geom_histogram(color = "black",
    fill = "green", bins = 20) + geom_vline(aes(xintercept = s_mean,
    color = "s_mean"), linetype = "dashed", size = 1) + geom_vline(aes(xintercept = th_mean,
    color = "th_mean"), linetype = "dashed", size = 1) + scale_color_manual(name = "Statistics",
    values = c(s_mean = "blue", th_mean = "red")) + xlab("Sample Means") +
    ylab("Frequency") + labs("Sample Mean versus Theoretical Mean")
```



From the plot we can see clearly that the mean of simulated values **s_mean = 5.034666** is close to the theoretical value **th_mean = 5**.

Sample Variance versus Theoretical Variance

Now let's compare sample variance and theoretical variance. Above we have stated that **theoretical variance** is:

$$\sigma^2 = \frac{1}{\lambda^2 n} = \frac{1}{(0.2)^2 n} = \frac{25}{n}$$

For $n = 40$ theoretical variance will be $\sigma^2 = \frac{25}{40} = 0.625$.

The sample variance is calculated below:

```
## Variance of sample mean
sample_vars <- var(s_means)
sample_vars
```

```
## [1] 0.5842821
```

We see that theoretical variance **0.625** is close to variance of sample 0.5842821. Now for the sake of comparision let's compare again using standart deviations.

Theoretical standart deviation will be $\sigma = \sqrt{\sigma^2} = \sqrt{0.625} = 0.7905694$.

The sample standart deviation is calculated below:

```r
## Standart deviation of sample mean
sample_sds <- sd(s_means)
sample_sds
```

```
## [1] 0.7643834
```

We see that theoretical standart deviation **0.790** is close to standart deviation of sample 0.7643834.

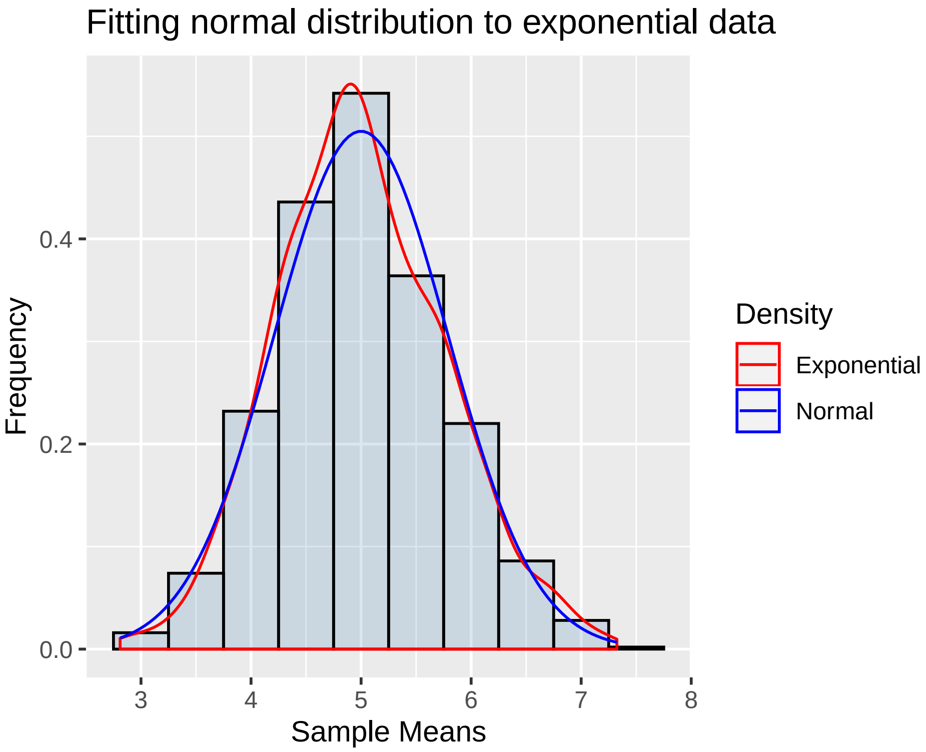
Show that the distribution is approximately normal

To show that our exponential distribution is approximately normal, We will create a vector of length 1000. Every element of this vector will represent the sample means out of each 1000 samples. Remember that each sample has size 40 meaning that we draw random 40 elements out of population.

The population distribution will be normal. Normal distribution requires two parameters for simulation, mean and variance. We will use the same values for these two parameters to mean and variance of exponential distribution.

The normal distribution will be simulated using R function **dnorm(n, mean, sd)**.

```r
# Plot histogram of generated sample means
fit_normal <- ggplot(df, aes(x = s_means)) + geom_histogram(aes(y = ..density..),
    color = "black", fill = "steelblue", binwidth = 0.5, alpha = 0.2) +
    labs(title = "Fitting normal distribution to exponential data") +
    geom_density(aes(color = "Exponential")) + stat_function(aes(color = "Normal"),
    fun = dnorm, args = list(mean = th_mean, sd = 0.79)) + xlab("Sample Means") +
    ylab("Frequency") + labs("Sample Mean versus Theoretical Mean") +
    scale_colour_manual("Density", values = c("red", "blue"))
fit_normal
```



From the plot is very clear that **sample mean** of exponential distribution simulated 1000 times is very close to **theoretical mean** for a normal distribution.