Statistical Inference Course Project - Part I

Evgeniia Golovina

12/03/2021

Overview

In this part of the project, we investigate the exponential distribution and compare it with the Central Limit Theorem. The Central Limit Theorem (CLT) is one of the most important theorems in statistics. The CLT states that the distribution of averages of independent and identically distributed (iid) variables becomes that of a standard normal as the sample size increases.

Via simulation and associated explanatory text we aim to show the following properties of the distribution of the mean of 40 exponentials:

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

We set lambda equal to 0.2, number of eponential equal to 40 and number of simulations equal to 1000.

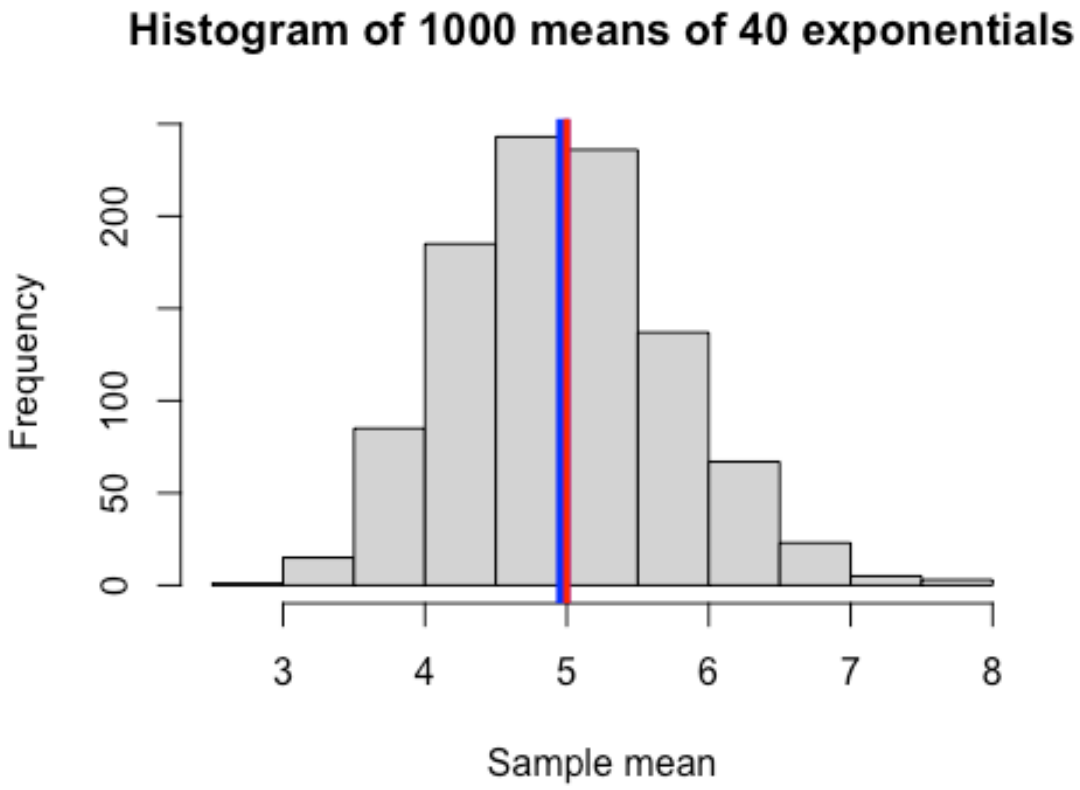
```
# setting lambda (0.2), number of exponentials (40) amd number of simulations
(1,000)
lambda <- 0.2; n <- 40; sim <- 1000

set.seed(1234)
# creating a data frame with 1000 X 40 from rexp(x, lambda)
sim_data <- matrix(rexp(n*sim, lambda), nrow=sim, ncol=n)
# calculating the mean of each row (40 exponentials): sample mean
sim_means <- apply(sim_data, 1, mean)
```

1. Sample mean vs theoretical mean

First, we compare sample mean to the theoretical mean of the distribution. The theoretical mean of the distribution is equal to 1/lambda.

```
# calculating sample mean
sample_mean <- round(mean(sim_means), 3)
# calculating theoretical mean
theoretical_mean <- round(1/lambda, 3)
# plotting mean distribution of 1000 simulations
```

```r
hist(sim_means, main="Histogram of 1000 means of 40 exponentials",
xlab="Sample mean",
     ylab="Frequency")
abline(v=sample_mean, col="blue", lwd=6)
abline(v=theoretical_mean, col="red", lwd=3)
```



As we can see, sample mean (blue) of *4.974* is very close to theoretical mean (red) of *5*.

2. Sample variance vs theoretical variance

Next, we show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. The theoretical standard deviation of the distribution is equal to 1/lambda.

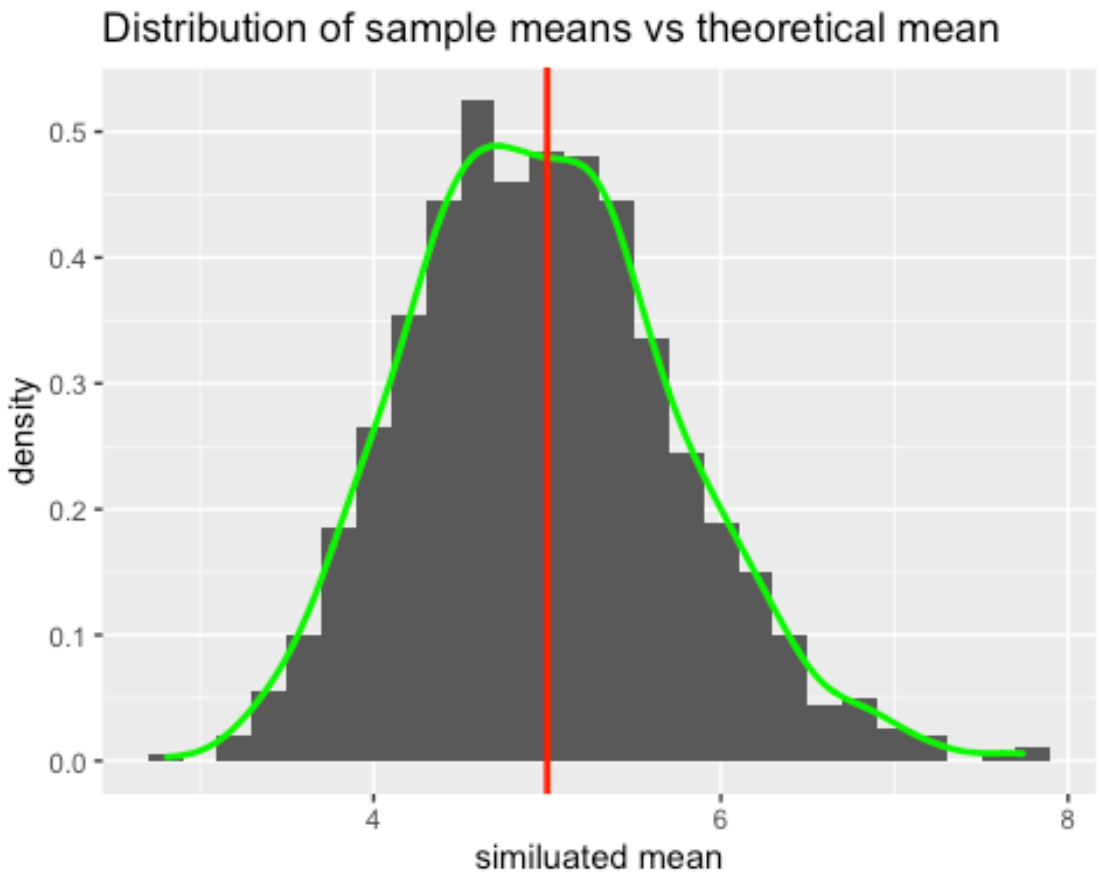
```r
# calculating the variance of the distribution
sample_var <- round(var(sim_means), 3)
# calculating the theoretical variance of the distribution
theoretical_var <- round((1/lambda)^2/n, 3)

# calculating sd of the sample mean
sample_sd <- round(sd(sim_means), 3)
# calculating the theoretical sd
theoretical_sd <- round((1/lambda)/sqrt(n), 3)
```

We can see that the sample variance of *0.595* is very close to the theoretical variance of *0.625*. Sample standard deviation of *0.771* is also very close to the theoretical standard deviation of *0.791*.

3. The distribution of the simulated means

Lastly, we want to show that the distribution is approximately normal.

```
df <- data.frame(sim_means); names(df) <- c("sim_mean")
ggplot(df, aes(x=sim_mean)) +
       labs(x="similuated mean", title="Distribution of sample means vs
theoretical mean") +
       geom_histogram(aes(y=..density..), size=1, binwidth=0.2) +
       geom_density(color="green", size=1) +
       geom_vline(xintercept=theoretical_mean, color="red", size=1)
```



The green line depicts the distribution of means of the simulated samples. The red line is the theoretical mean. This figure shows that the distribution of means of the simulated samples is very close to normal distribution.