Statistical Inference Part A

Suhas P K

2023-06-18

Introduction

This project is part of Coursera John Hopkins Data Science Course: Statistical Inference. This project consists of two parts: 1. A simulation exercise. 2. Basic inferential data analysis

Project 1 : A simulation exercise

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be 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. Set lambda = 0.2 for all of the simulations. We will investigate the distribution of averages of 40 exponentials. Note that we will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should

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.

In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

As a motivating example, compare the distribution of 1000 random uniforms.

```
hist(runif(1000))
```

and the distribution of 1000 averages of 40 random uniforms

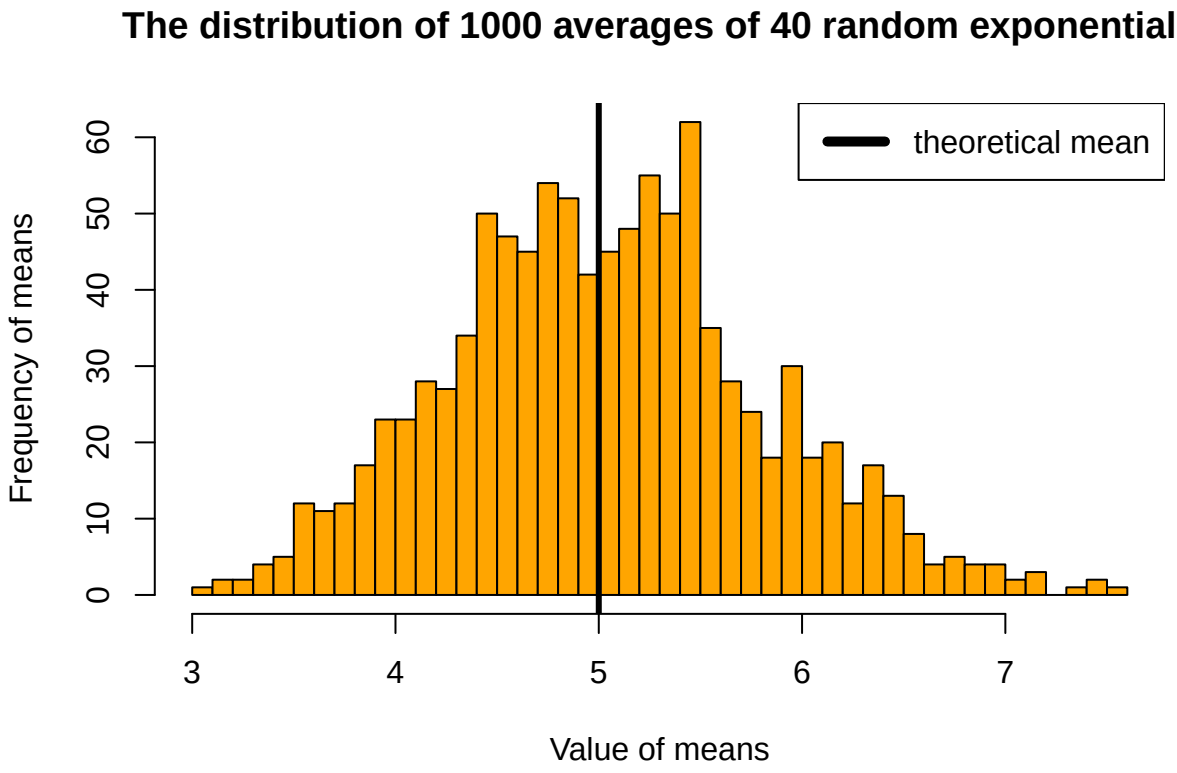
```
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(runif(40)))
hist(mns)
```

This distribution looks far more Gaussian than the original uniform distribution!

1. Show the sample mean and compare it to the theoretical mean of distribution.

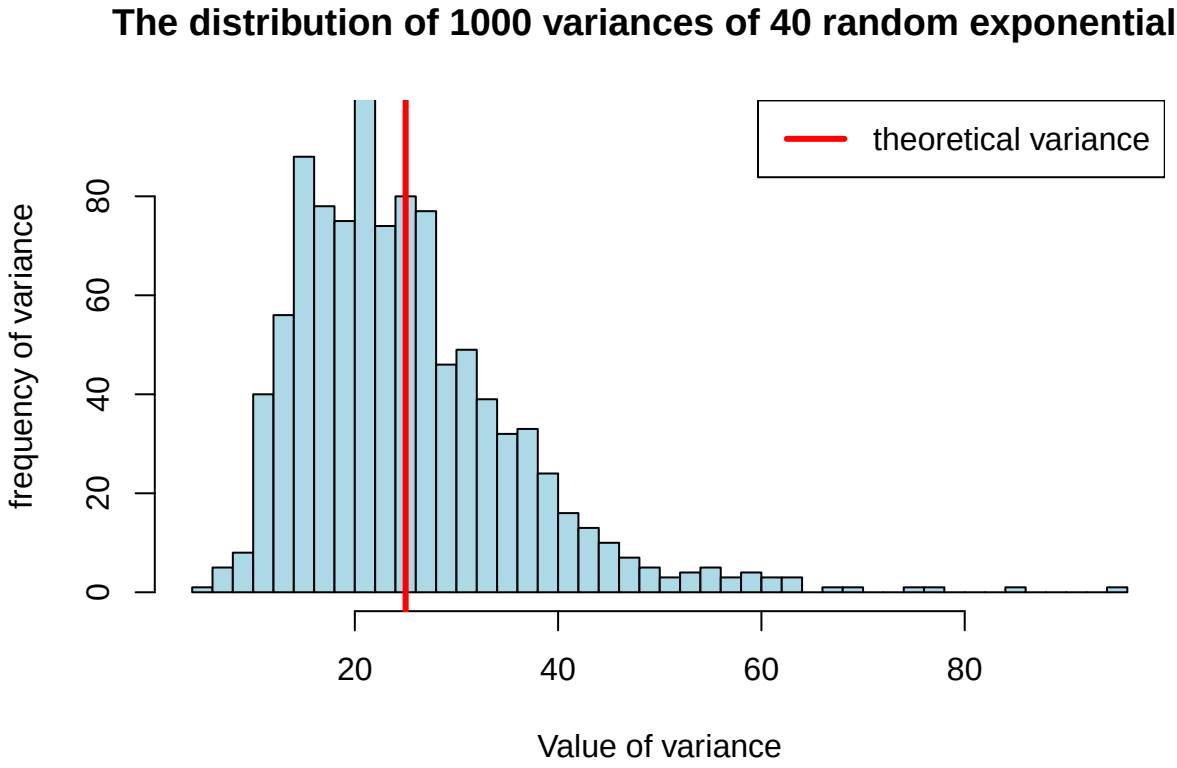
```
lambda <- 0.2
sim_data <- matrix(rexp(1000*40, lambda),
                   nrow = 1000, ncol = 40)

dist_mean <- apply(sim_data, 1, mean)

#Plotting
```

```r
hist(dist_mean, breaks = 50,
     main = " The distribution of 1000 averages of 40 random exponentials.",
     xlab = "Value of means",
     ylab = "Frequency of means", col = "orange")
abline(v=1/lambda, lty = 1,
       lwd=3, col = "black")
legend("topright", lty = 1,
       lwd = 5, col = "black", legend = "theoretical mean")
```



2. Show how variable the sample is (using variance) and compare it to the theoretical variance of the distribution.

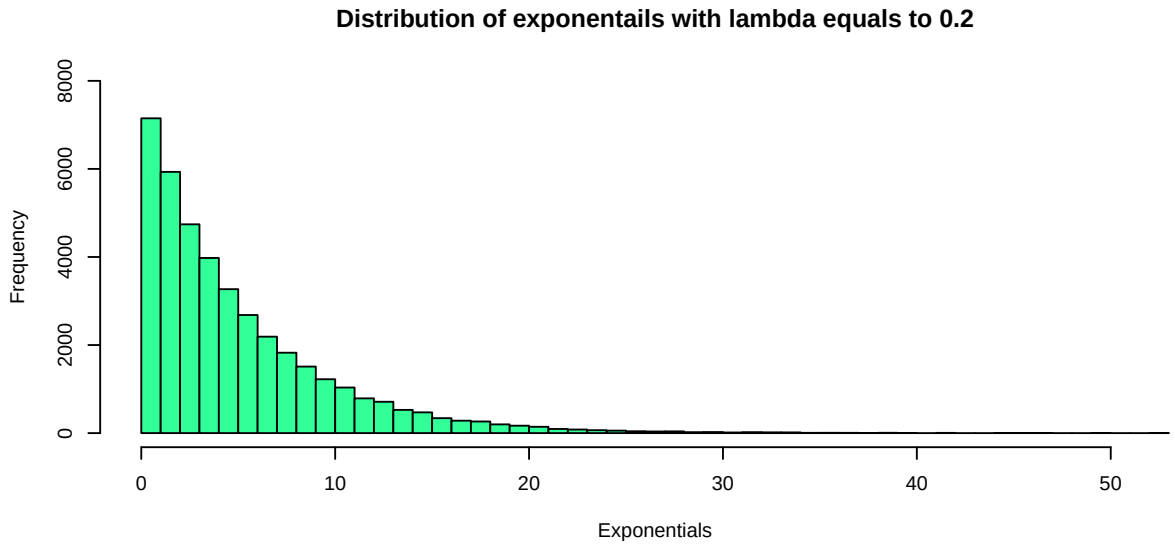
```r
variance_dist <- apply(sim_data, 1, var)

#plotting
hist(variance_dist, breaks = 50,
     main = "The distribution of 1000 variances of 40 random exponentials.",
     xlab = "Value of variance",
     ylim = c(0,max(variance_dist)),
     ylab = "frequency of variance",
     col = "lightblue")
abline(v = (1/lambda)^2, lty = 1,
       lwd = 3, col = "red")
legend("topright",lty = 1,
       lwd = 3,
       col = "red",
       legend = "theoretical variance",)
```

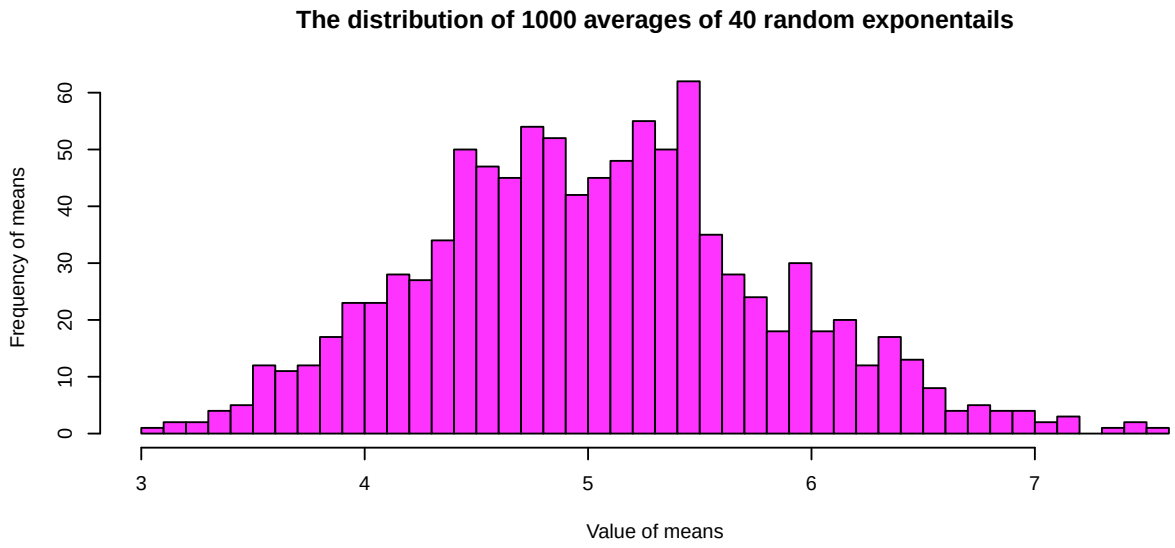


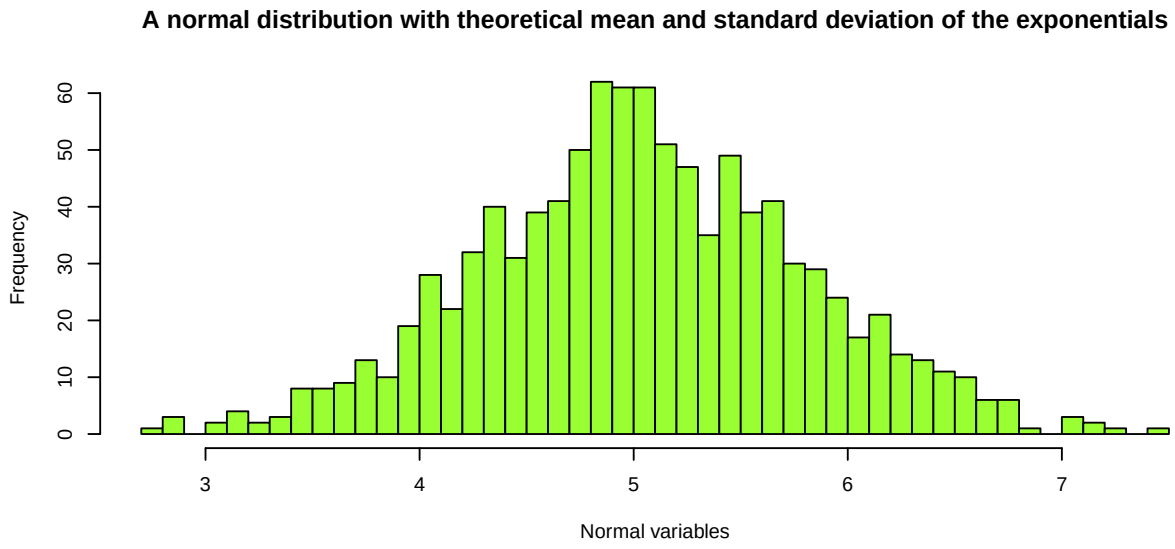
The simulated sample variances are almost normally distributed with a center near the theoretical values.

3. Show that the distribution is approximately normal.

```
# plotting
par(mfrow = c(3,1))

hist(sim_data, breaks = 50,
     main = "Distribution of exponentails with lambda equals to 0.2",
     xlab = "Exponentials",
     ylim = c(0,8000),
     col = "#33ff99")

hist(dist_mean, breaks = 50,
     main = "The distribution of 1000 averages of 40 random exponentails",
     xlab = "Value of means", ylab = "Frequency of means", col="#ff33ff")

normal_sim <- rnorm(1000, mean = mean(dist_mean), sd = sd(dist_mean))
hist(normal_sim, breaks = 50,
     main = "A normal distribution with theoretical mean and standard deviation of the exponentials",
     xlab = "Normal variables",
     col = "#99ff33")
```


Distribution of exponentails with lambda equals to 0.2
Frequency
0
2000
4000
6000
8000
0
10
20
30
40
50
Exponentials


The distribution of 1000 averages of 40 random exponentails
Frequency of means
0
10
20
30
40
50
60
3
4
5
6
7
Value of means


A normal distribution with theoretical mean and standard deviation of the exponentials
Frequency
0
10
20
30
40
50
60
3
4
5
6
7
Normal variables

The first histogram is the distribution of exponential with lambda equals to 0.2. The second histogram is the distribution of 100 averages of 40 random exponential. The third histogram is a real normal distribution with a mean and standard deviation equals to the second histograms. Comparing the first with the second histogram, we can see the distribution becomes normal as the means were taken from each groups. It is a result of the **Central Limit Theorem**.Comparing the second and the third histogram, we can see the distribution with the same mean and standard deviation.