Distribution of Averages of the Exponential Distribution

Pedro Magalhães Bernardo

Saturday, March 26, 2016

Overview

This report is part of a course project within the Statistical Inference course on the Data Science Specialization by Johns Hopkins University on Coursera.

On this report we will investigate the distribution of averages of the exponential distribution. Our goal is to investigate the exponential distribution in R and compare it with the Central Limit Theorem.

Simulations

First, to ensure reproducibility we set a seed, so the random numbers gerenated by the exponential distribution can be reproducible. Also, we set lambda (the rate parameter of the exponential distribution) to be 0.2.

```
set.seed(10232)
lambda <- 0.2
```

Then, we generate a vector with 1000 averages of 40 random exponentials, this vector will be saved on the variable **means**.

```
means <- NULL
for (i in 1:1000)
      means <- c(means, mean(rexp(40,lambda)))
set.seed(10232)
```

Sample Mean versus Theoretical Mean

Figure 1 shows the distribution of the 1000 averages of 40 random exponentials.

```
library(ggplot2)
histogram <- ggplot(data = data.frame(means),aes(x=means)) +
      geom_histogram(colour="black", fill="lightblue", binwidth = 0.3, aes(y = ..density..)) +
      xlab("Averages of 40 random exponentials") + ylab("Density") +
      ggtitle("Histogram of Averages of 40 random exponentials") +
      theme(plot.title = element_text(lineheight=.8, face="bold")) +
      stat_function(fun=dnorm, args=list(mean=5, sd=sd(means)), size=2)
```

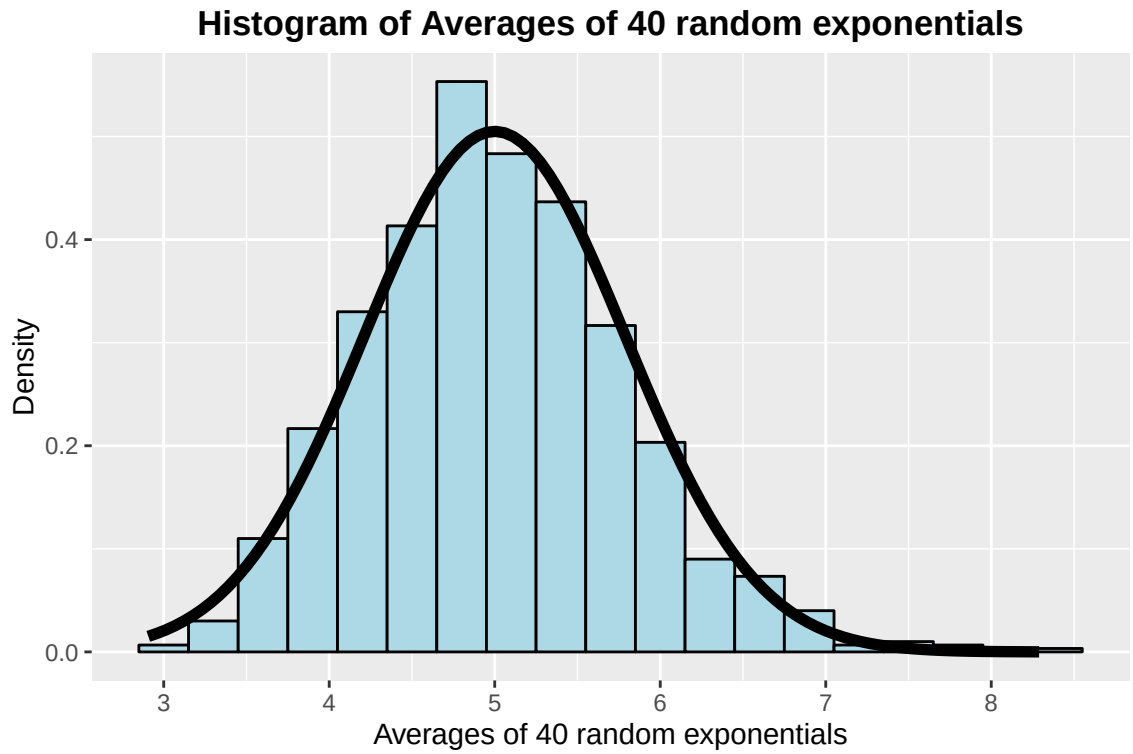

Figure 1

On top of the histogram we plot a normal density with mean 5 (the theoretical mean for the exponential distribution with lambda = 0.2)

We can see that the sample distribution is centered around 5.This was expected, since the mean of the distribution of averages should approximate the mean of the population.

In fact the thoretical mean is 1/lambda, which in our case equals to 5.

Also, the code below shows the exact mean of the sample means (which is really close to 5, as expected.)

```
mean(means)
```

```
## [1] 5.003511
```

Sample Variance versus Theoretical Variance

We can calculate the sample variance using the variance of the sample mean.

We know that the variance of the sample mean is equal to the sample variance divided by the sample size.

Thefore the sample variance is equal to the variance of the sample mean times the sample size.

```
##Sample Variance
var(means)*40
```

```
## [1] 24.9628
```

We get a value close to 25.

In fact the variance of the population is (1/lambda)^2, which in our case is 25.

Distribution

Figure 1 already shows that the distribution of 1000 averages of 40 random exponentials is approximately normal. But we can go further.

The code below normalizes this averages, therefore when plotting the histogram we expect to have a standard normal distribution.

```
set.seed(10232)
normmeans <- NULL
for (i in 1:1000)
      normmeans <- c(normmeans, (mean(rexp(40,lambda))-5)/(5/sqrt(40)))
```

```
library(ggplot2)
histogram <- ggplot(data = data.frame(normmeans),aes(x=normmeans)) +
      geom_histogram(colour="black", fill="lightblue", binwidth = 0.3, aes(y = ..density..)) +
      xlab("Averages of 40 random exponentials") +
      ylab("Density") +
      ggtitle("Histogram of Averages of 40 random exponentials")  +
      theme(plot.title = element_text(lineheight=.8, face="bold")) +
      stat_function(fun=dnorm, size=2)
```

Figure 2 shows exactly that.

Now we have approximately a standard normal (mean 0 and variance 1). We can see that by plotting a standard normal density curve on top of the histogram.

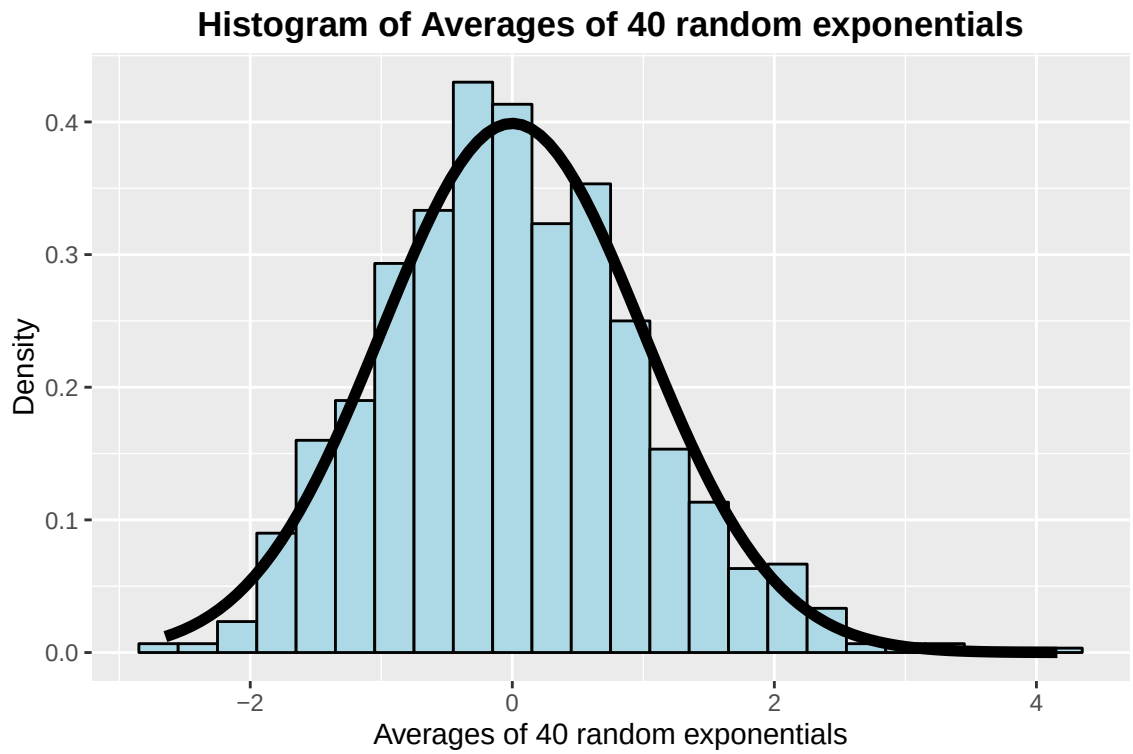

Figure 2