Part 1: The Exponential Distribution

Ivo Georgiev

July 28, 2016

Overview

This document demonstrates the **Central Limit Theorem** using the exponential distribution of a continuous random variable X

$$f_X(x|\lambda) = \begin{cases} \lambda x^{-\lambda x} & \text{for} \quad x > 0 \\ 0 & \text{for} \quad x \leq 0 \end{cases},$$

where $\lambda > 0$ is called the *rate* of the distribution.

The theoretical mean of the exponential distribution is

$$E[X] = 1/\lambda.$$

The theoretical variance of the exponential distribution is

$$Var[X] = 1/\lambda^2.$$

The Central Limit Theorem (CLT) states that, *given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined (finite) expected value and finite variance, will be approximately normally distributed, regardless of the underlying distribution* (Wikipedia).

So, our job will be to show that the mean of sufficiently large iterates of exponentials is approximately normally distributed.

Assumptions

Our assumptions for the following discussion are:

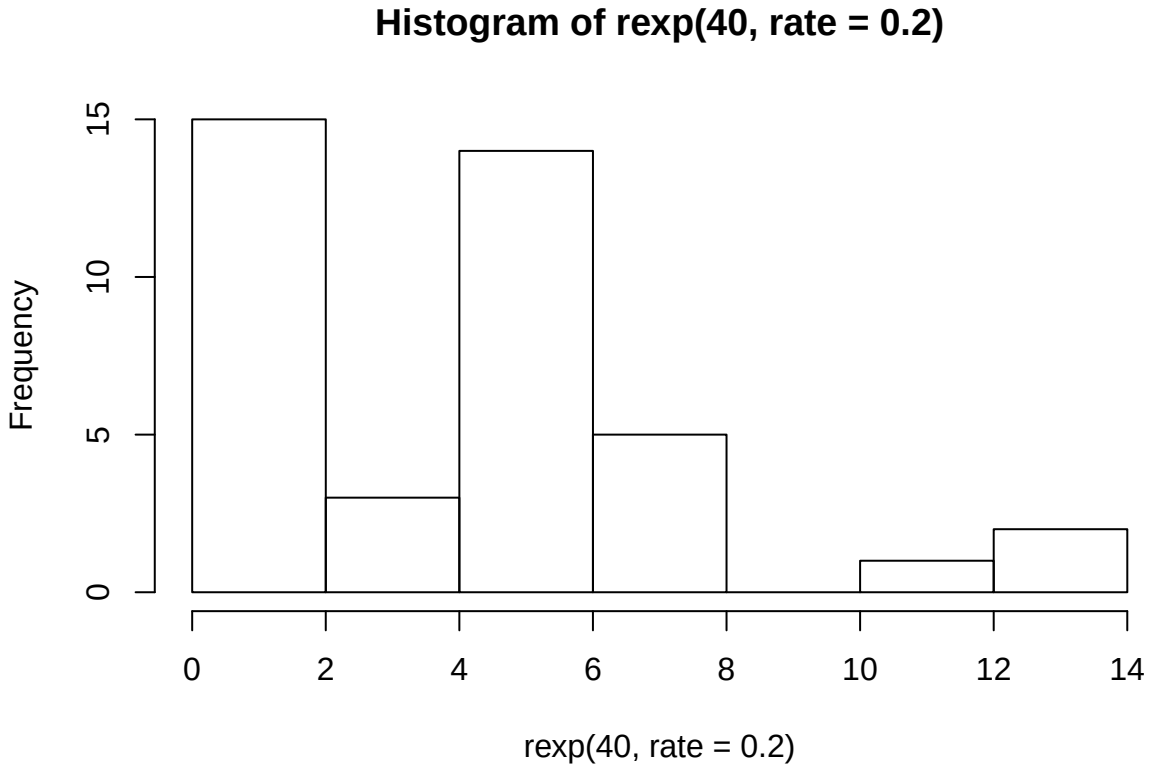
1. The samples are independent and identically distributed. Since we will be using the R function **rexp(n, lambda)**, we can be assured of that.
2. Our underlying distribution has well-defined finite mean and variance. We know that this is so. The values are shown above.

Simulations

The problem statement for this part of the assignment asks for the properties of the mean of **40** exponentials with a rate of **0.2**. This means that the theoretical *mean* and *standard deviation* of the distribution are both equal to **5**.

Let's take a look at the histogram of 40 exponentials.

```
hist(rexp(40, rate = 0.2))
```

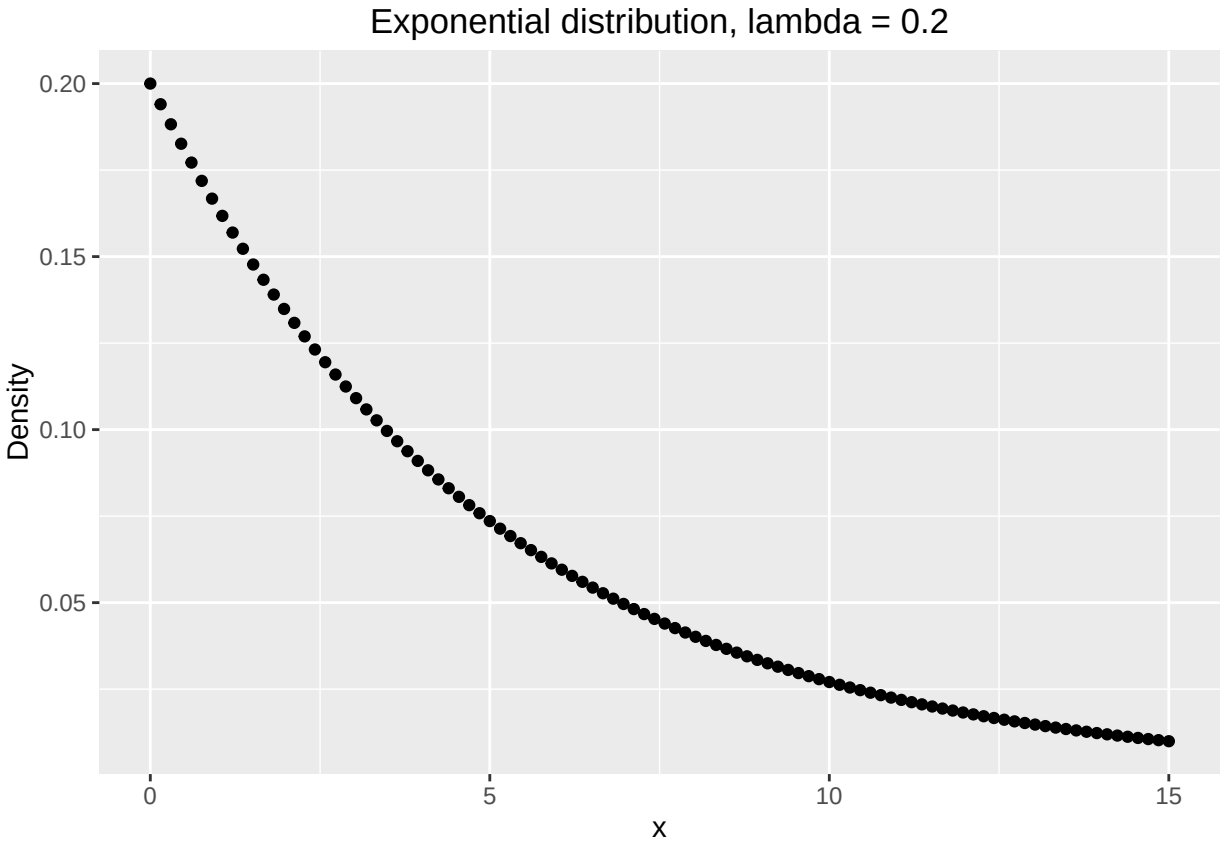


Every time this code is executed, the plot will look different, but it always looks like the values were drawn from the exponential distribution.

```
require(ggplot2)
```

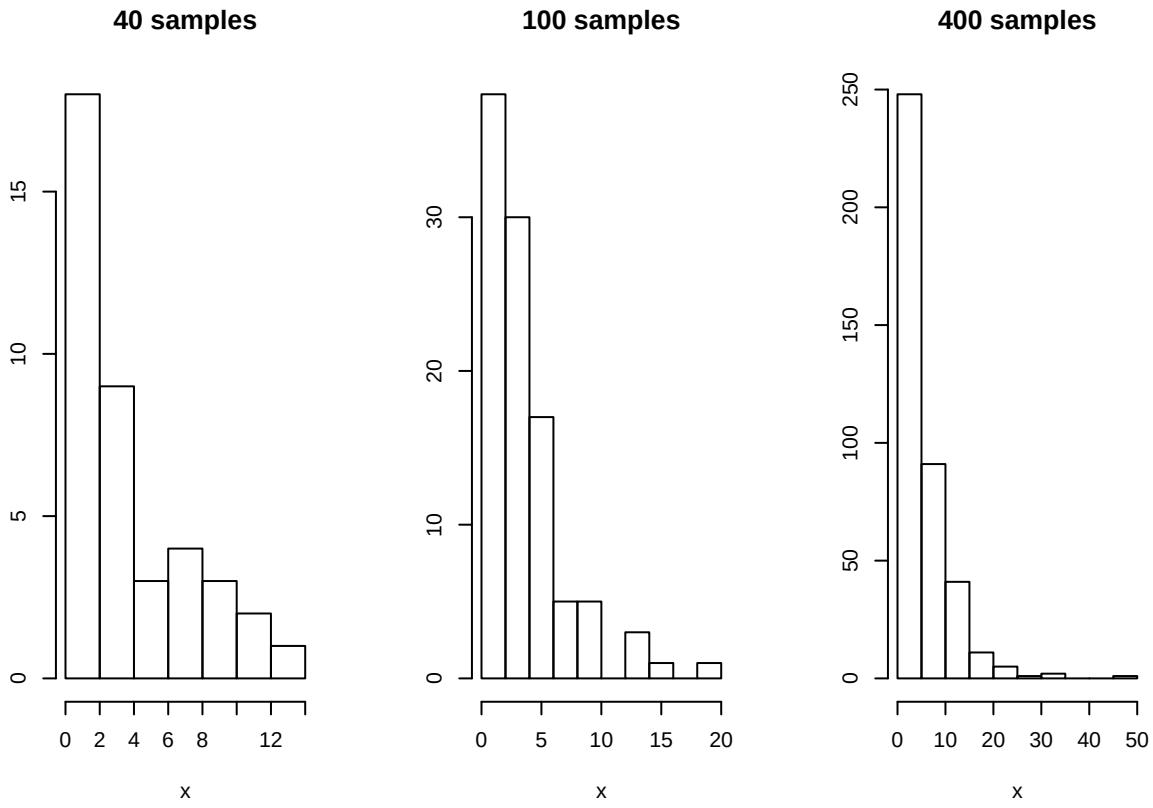
```
## Loading required package: ggplot2
```

```
x <- seq(0, 15, length = 100)
qplot(x, dexp(x, rate = 0.2), main="Exponential distribution, lambda = 0.2",
      ylab = "Density")
```



For larger sample values, the plot should look more closely like the density function above.

```
par(mfrow = c(1, 3))
hist(rexp(40, rate = 0.2), main = "40 samples", xlab = "x", ylab = NULL)
hist(rexp(100, rate = 0.2), main = "100 samples", xlab = "x", ylab = NULL)
hist(rexp(400, rate = 0.2), main = "400 samples", xlab = "x", ylab = NULL)
```

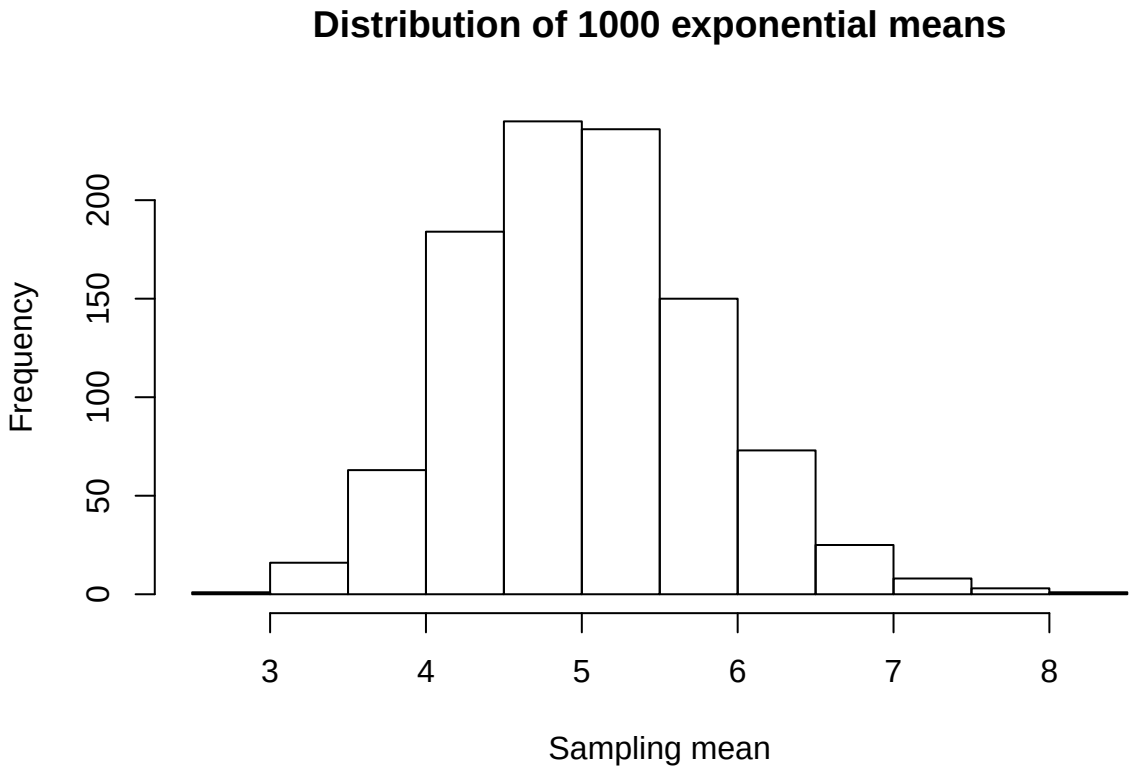


As expected, with larger samples, the histogram of the sampling closer resembles the underlying distribution.

```
## null device
##           1
```

Now let's take a look at the histogram of the means of a large number of exponential samplings, using our 40-sample sampling.

```
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(rexp(40, rate = 0.2)))
hist(mns, xlab = "Sampling mean", main = "Distribution of 1000 exponential means")
```



There are two things to observe here:

1. The distribution looks *normal*.
2. It seems to be centered at *5* which is the mean of Exponential(λ) with rate=0.2.

We are starting to glimpse the *Central Limit Theorem* in action!

Sample mean vs theoretical mean

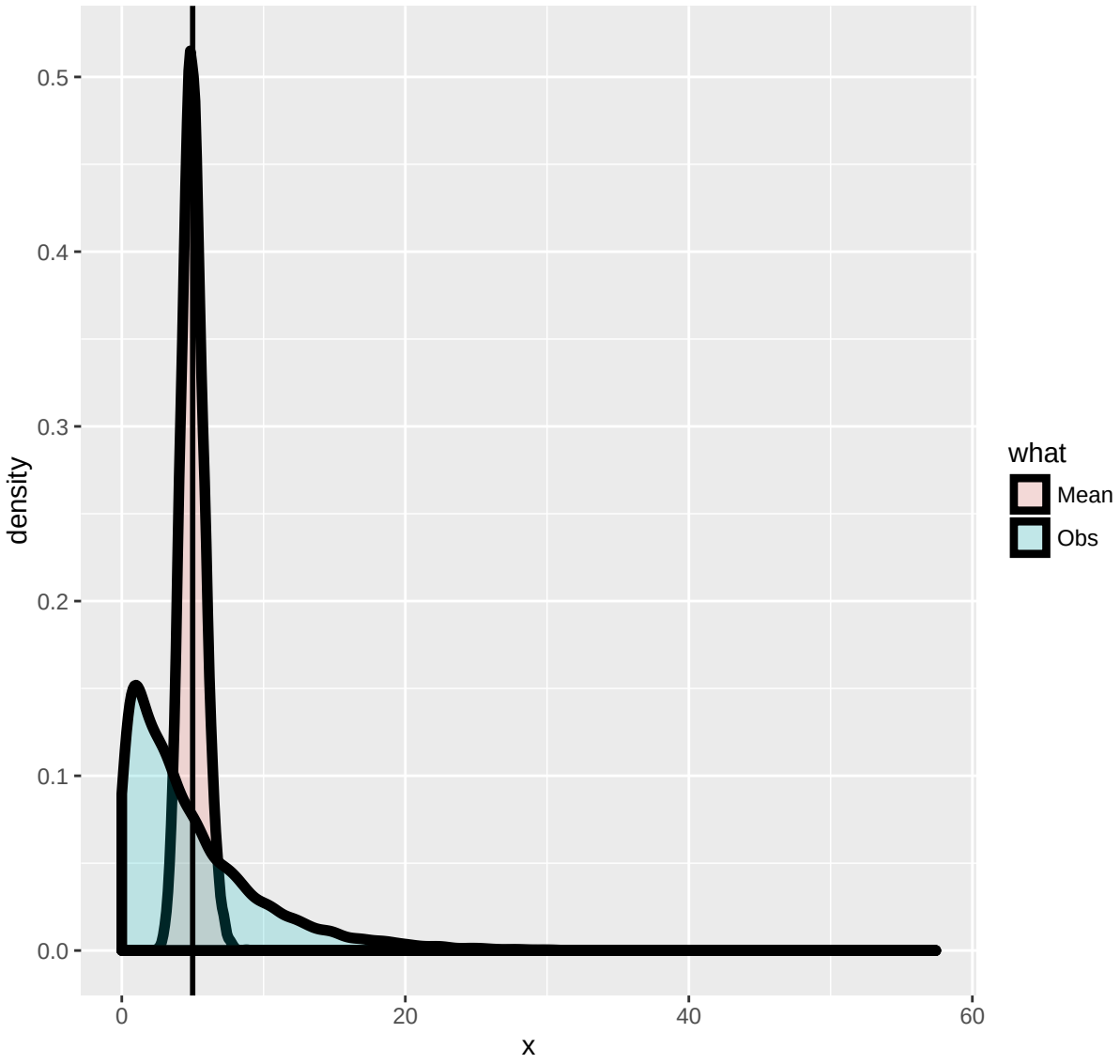
Now, what is the mean of one of our samplings?

```
mean(rexp(40, rate = 0.2))
```

```
## [1] 5.352012
```

Note that this value will be different every time the code is executed but it will be close to **5**, the expected long-term value for the sampling. The sample mean is an *unbiased estimator* of the population mean, so we are getting the correct values.

The mean of a random variable is also a random variable. We saw earlier that its distribution looks normal. Let's look more closely at the distribution of 40 exponentials vs the distribution of means of 40 exponentials, again with rate (lambda) 0.2.



Bearing in mind that the center of mass of the exponential with lambda = 0.2 has an x-intercept of 5, we see that the distributions of exponentials and means of exponentials are **both** centerd at 5, which is the theoretical mean of the exponential distribution. So, *the mean of exponentials estmates the population mean.*

Sample variance vs theoretical variance

Another thing to notice is that the curve of the means (salmon color in the above plot) is a lot more tightly centered around the mean of the population, which it estimates. We can show that the sample variance approximates the theoretical variance σ^2/n, where n = 40, or $Var[\overline{X}] = 1/40\lambda^2 = 0.625$.

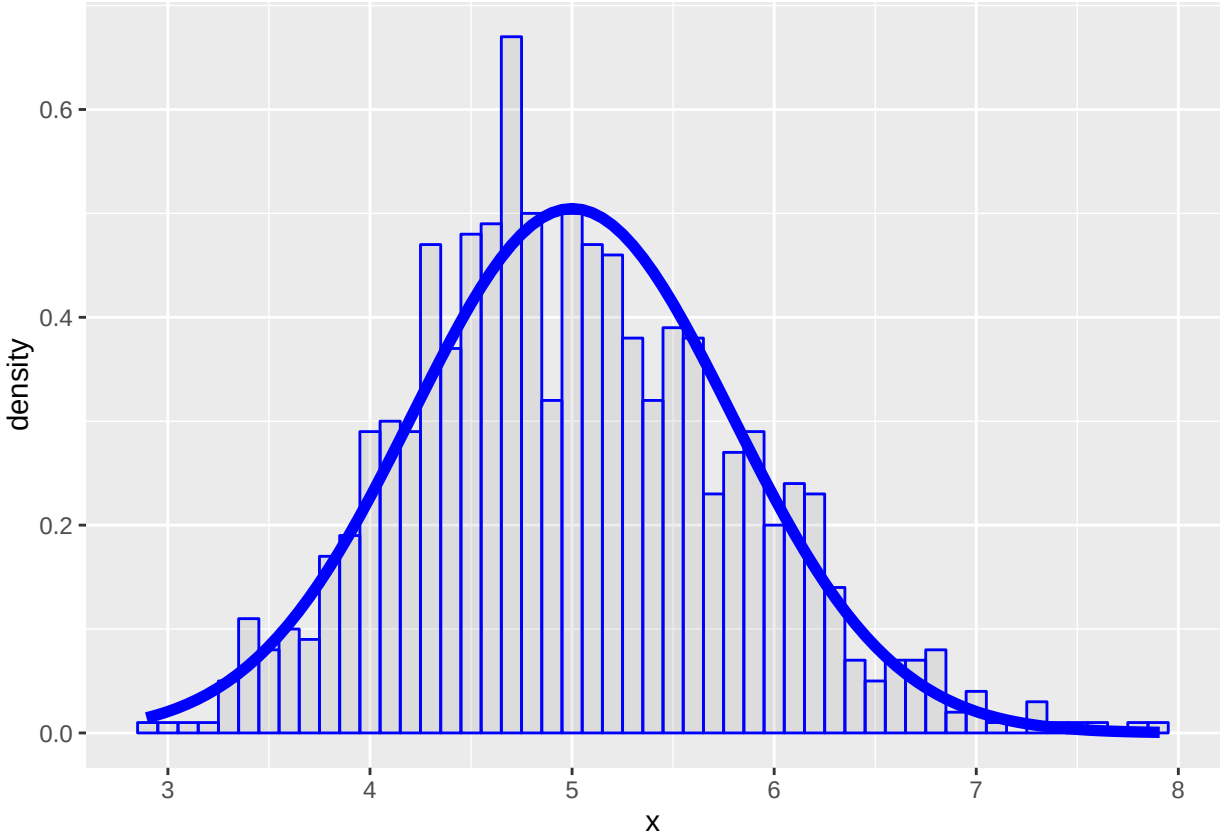
```
set.seed(4995) # seed the random number generator for reproducibility
means <- data.frame(x = apply(matrix(rexp(1000 * 40, rate = 0.2), 1000), 1, mean))
var(means$x)
```

```
## [1] 0.6530999
```

Indeed the empirical variance of the averages closely approximates the theoretical value.

Distribution

Finally, to illustrate the *Central Limit Theorem*, we can plot the means distribution along with a stardard normal with mean = **5.0** and standard deviation = **5/sqrt(40)**.



Indeed, our distribution closely fits approximates a normal distribution, showing good evidence for the CLT.