CLT theorem via exponential distribution

Kirill Panarin

Overview

Let's investigate the exponential distribution and validate the Central Limit Theorem on it. The exponential distribution has a rate parameter. I will use rate parameter equal to 0.2 and will assign it to the `lambda` variable. The mean of exponential distribution is `1/lambda` and the standard deviation is also `1/lambda`. In addition to `lambda` we need to define variables for the theoretical mean and standard deviation:

```
lambda <- 0.2
theoretical.mean <- 1/lambda
theoretical.sd <- 1/lambda
```

Simulations

The simulation process has the following structure: - Generate `n` random values. - Repeat the first procedure for `nsim` number of times in order to gather a reasonable amount of data for analysis.

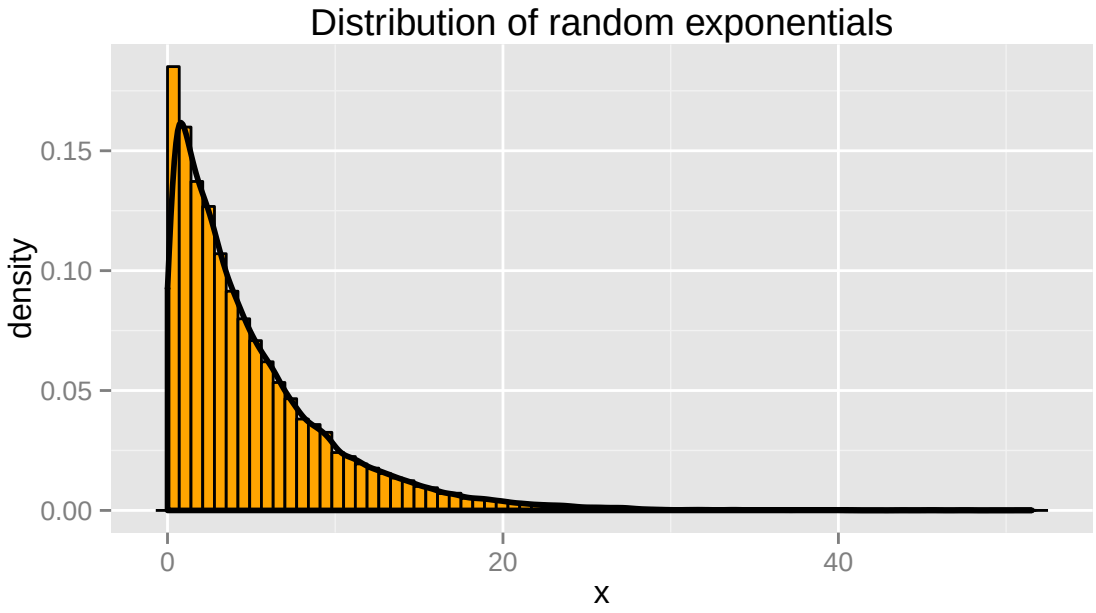
I will use the following values for `n` and `nsim` parameters:

```
n <- 40
nsim <- 1000
```

Let's start simulations from generating `n` times `nsim` random values with the exponential distribution. `simulation.data_raw` contains the array of random values and `simulation.data_matrix` contains the same data but restructured as a matrix with `n` columns:

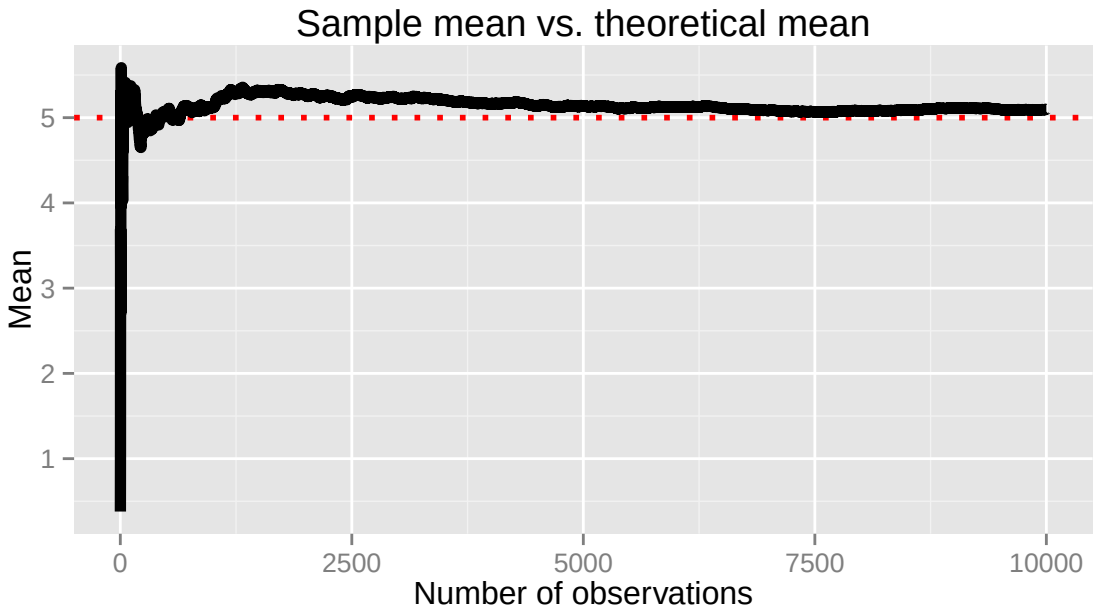
```
simulation.data_raw <- rexp(n * nsim, lambda)
simulation.data_matrix <- matrix(simulation.data_raw, ncol = n)
```

The first visualization will be the samples values distribution. It just shows that we are really dealing with the exponential distribution:



Sample Mean vs. Theoretical Mean

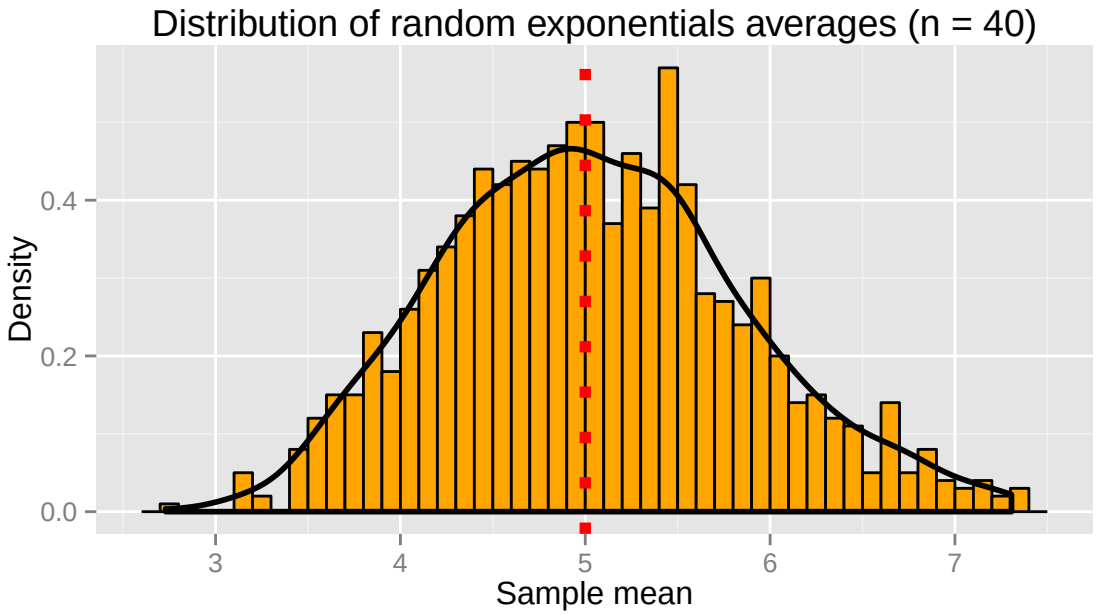
In the following simulation we can observe how sample mean approaches theoretical mean as number of observations increases. The red dotted line is the theoretical mean (`lambda`).



Let's calculate sample means and store the result array in the `simulation.means` variable:

```
simulation.means <- apply(simulation.data_matrix, 1, mean)
```

Now we can plot a histogram of sample means. The red dotted vertical line corresponds to the theoretical mean `1/lambda`.



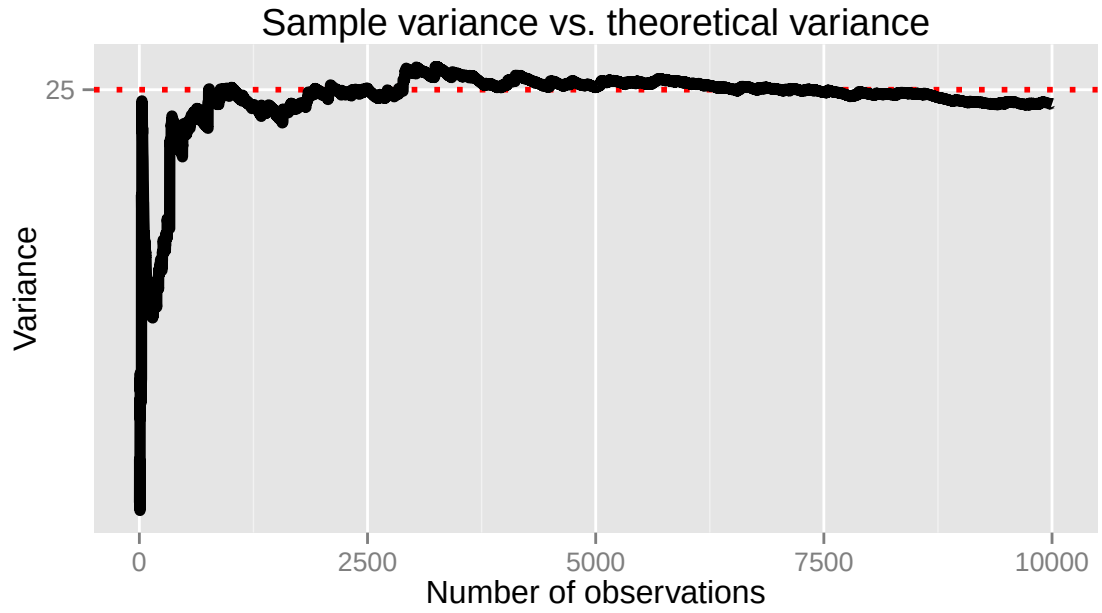
It's apparent from figures that sample mean approaches to `theoretical.mean` as number of observations `n` increases.

Sample Variance vs. Theoretical Variance

We can do the same simulations for sample variance as we did for sample mean. The theoretical variance is the square of the theoretical standard deviation:

```
theoretical.var <- theoretical.sd**2
```

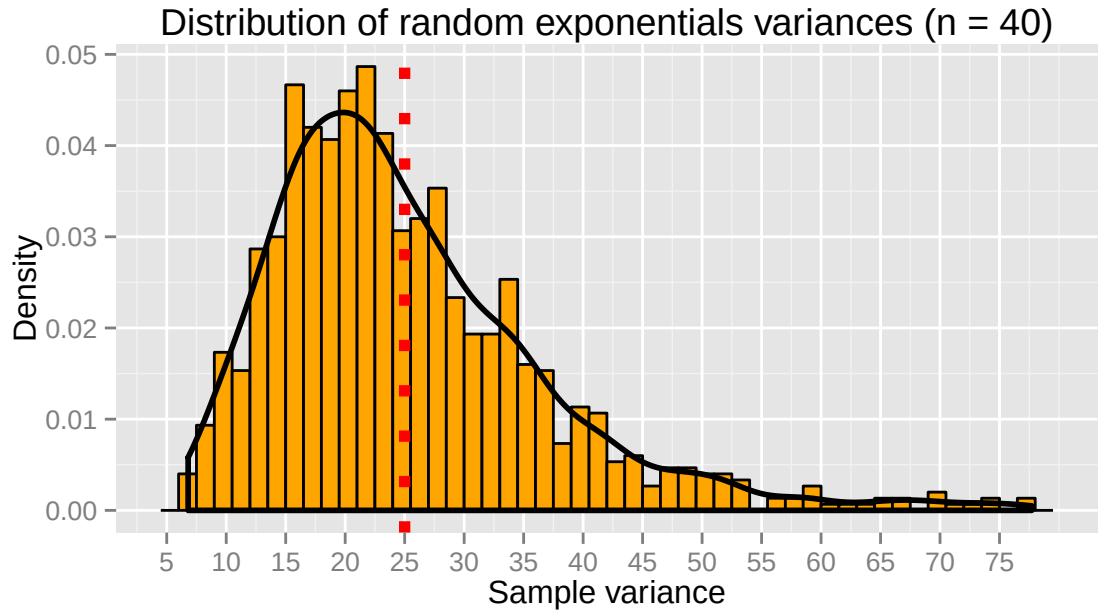
The plot below demonstrates that sample variance approaches to the theoretical variance as number of observation increases:



Let's calculate sample variances and store the result array in the `simulation.vars` variable:

```
simulation.vars <- apply(simulation.data_matrix, 1, var)
```

Now we can plot a histogram of sample variances. The red dotted vertical line corresponds to the theoretical variance `(1/lambda)**2`.



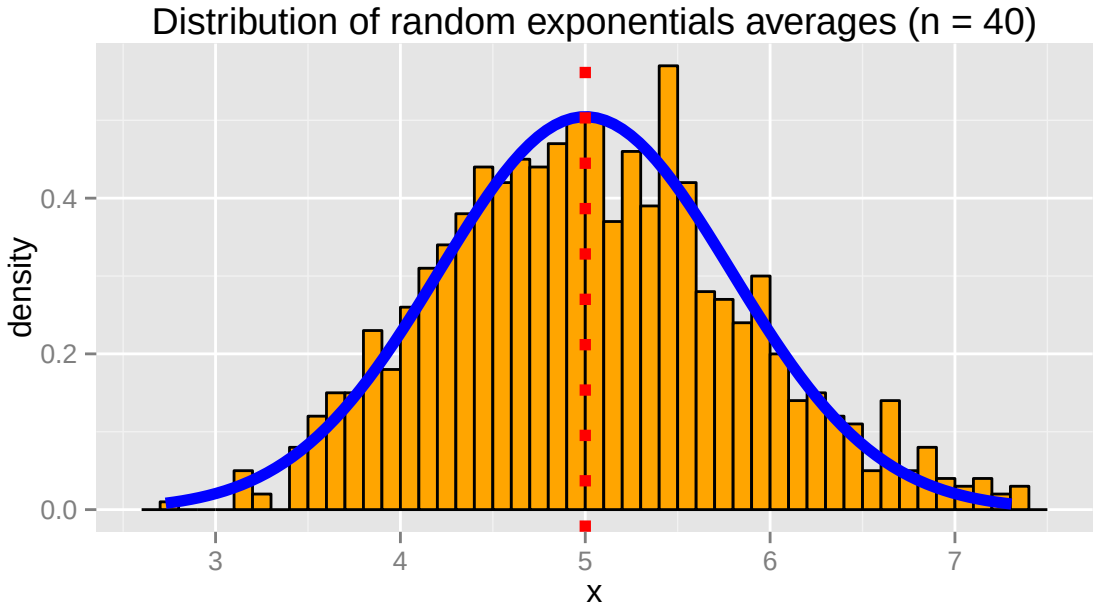
It's apparent from figures that sample variance approaches to `theoretical.var` as number of observations `n` increases.

Distribution

Central Limit Theorem says that as number of observation increases the sample average becomes normally distributed with the expected value equal to `theoretical.mean` and the following standard deviation:

```
theoretical.means_sd <- theoretical.sd/sqrt(n)
```

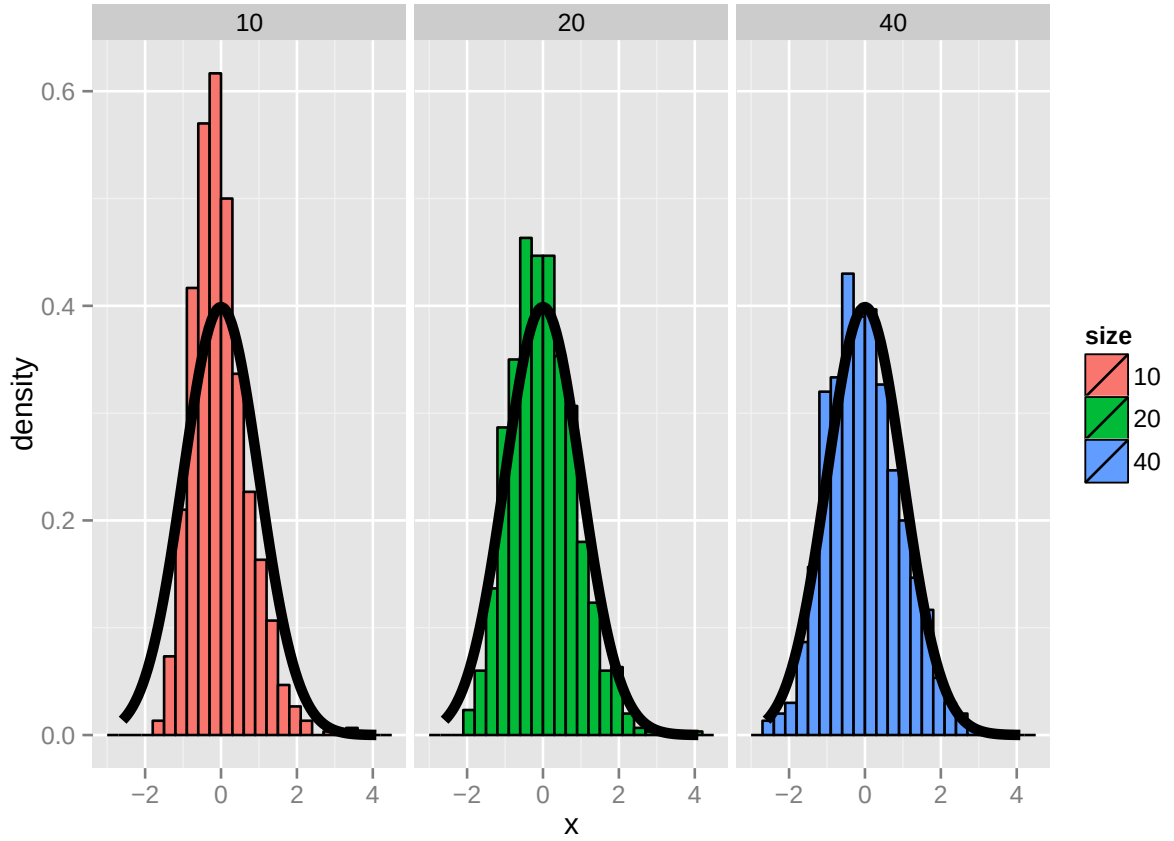
Let's double check this assumption on the exponential distribution. The blue line corresponds to the normal distribution with the parameters `theoretical.mean` and `theoretical.sd/sqrt(n)`. The red dotted line shows the theoretical mean of the normal distribution.



Let's also plot sample mean distributions for three numbers of observations: 5, 15 and 40. In this case sample mean is normalized using the following formula:

```
norm <- function(x, n) sqrt(n) * (mean(x) - theoretical.mean) / theoretical.sd
```

Since the sample mean is normalized its distribution should converge to the standard normal distribution. The blue lines correspind to the standard normal distributions on the following plots:



Based on the figures above and the exponential distribution plotted at the beggining we can deduct that CLT works on practise even for relatively small number of observations. Note that for small number of observation the distribution of means is close to the exponential distribution itself.