Exponential Distribution in R and the Central Limit Theorem Investigation

Martin Tapp

Sunday, January 25, 2015

Overview

The **Central Limit Theorem** (CLT) states that the distribution of averages of iid (independent and identically distributed) variables becomes that of a standard normal as the sample size increases. Therefore, we investigate this theorem using simulations in order to compare their results with the theoretical ones. The following sections present our findings which are conform to this theory.

Simulations

In order to investigate the CLT, we used the exponential distribution using R's `rexp` function with a rate of 0.2 (i.e. `lambda`). We simulated 1000 samples (i.e. `sim.count`) each one using 40 random points (i.e. `sample.size`) from this distribution. We used the `sapply` function to generate the `mean` of each sample which were aggregated into the `sim.data` data frame under the `means` variable.

The following R code demonstrates these various computations.

```
set.seed(1457)

lambda <- 0.2
sim.count <- 1000
sample.size <- 40

sim.data <- sapply(1:sim.count, function(q) mean(rexp(sample.size, lambda)))
sim.data <- data.frame(means=sim.data)
```

> Note the use of a specific seed in order to have reproducable results.

Sample versus Theoretical

In order to compare the theoretical and simulation results together, we used the following computations. We computed the expected, i.e. theoretical, and sample values using our simulations for both the mean and the variance. We also computed the standard deviation based on the number of samples (`1/lambda/sqrt(sample.size)`).

```
expected.sd <- 1/lambda/sqrt(sample.size)
expected.mean <- 1/lambda
expected.var <- 1 / (lambda^2 * sample.size)
sample.mean <- mean(sim.data$means)
sample.var <- var(sim.data$means)
```

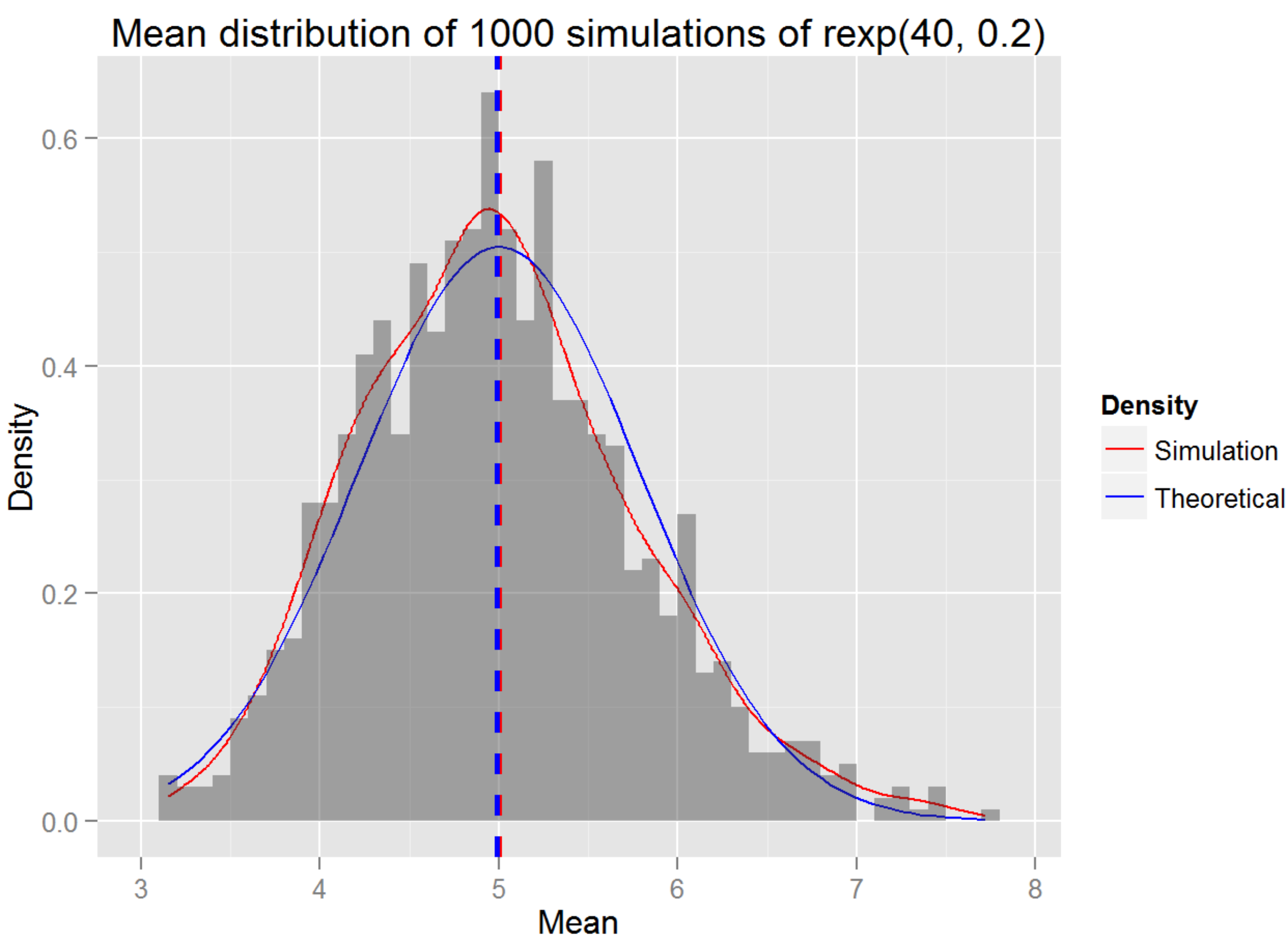
We can see that the expected mean is 5 (`1/lambda`) and that our simulations exhibit a mean of 4.990774 (`mean(sim.data$means)`). We also see that the expected variance is 0.625 (`1 / (lambda^2 * sample.size)`) and that our simulations' variance is 0.6048536 (`var(sim.data$means)`).

These results are therefore conform to the **Central Limit Theorem** being very close to one another. Using a higher number of simulations would further increase the convergence towards the theoretical mean and variance values. Refer to the following section for graphical illustrations of the theoretical versus simulation values.

Distribution

We have seen that the simulation results were very close to the theoretical ones for both the mean and the variance. Now, we will look at the shape of the distribution of the means which should be normal as stipulated by the CLT.

The following figure illustrates the density distribution of the simulated data.



The previous figure clearly shows that the results are centered around the theoretical mean which is seen by the dashed blue line (whereas the sample mean is illustrated by the dashed red line).

Moreover, note the shape of the simulation results in blue versus the theoretical ones in red. We can see from this figure that the distribution has the shape of a normal distribution. Therefore, these results clearly conform to the CLT.