Statistical Inference Project

Part 1

Introduction

The first part of this project consists in investigating the exponential distribution in R and comparing it with the Central Limit Theroem. We will take many observations of samples of the exponential distribution and confirm that the distribution of the means of the samples converges to the normal distribution.

We will first calculate the theorical values and show the convergence of the mean and the variance. Finally, we will show that the distribution of the means converges to a normal distribution.

The exponential distribution - theorical values

The exponential distribution has the following probability density function:

$$f(x;\lambda) = \left\{ \begin{array}{cc} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{array} \right.$$

For this exercise, we set $\lambda = 0.2$.

The expected value $E[X] = \frac{1}{\lambda}$ is 5, this is the theoretical mean.

And the variance $Var[X] = \frac{1}{\lambda^2}$ is 25, this is the theoretical variance.

Simulations

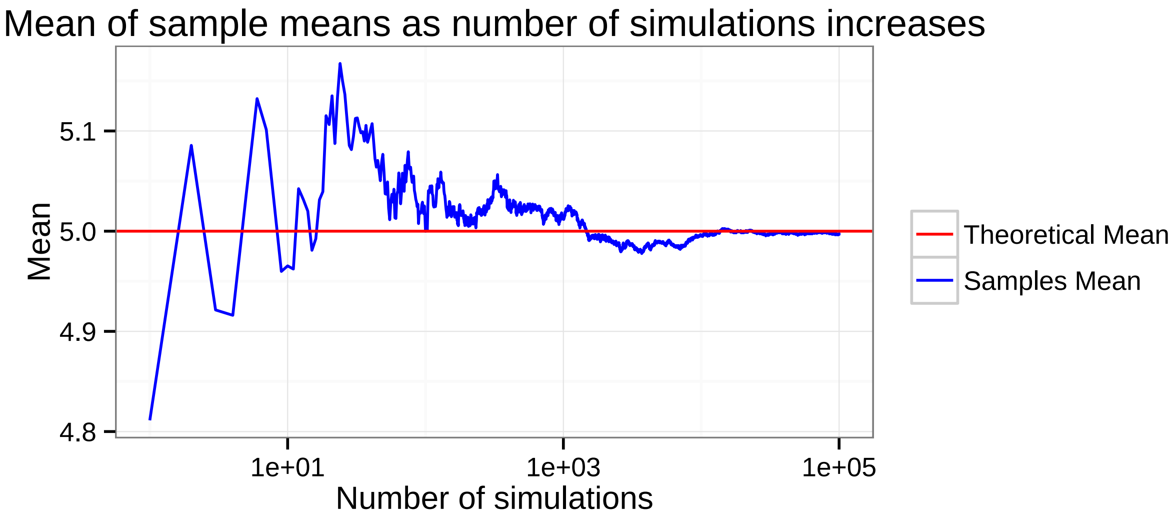
I will investigate the distribution of averages of 40 exponentials by simulating 100 000 distributions of 40 exponentials and storing their means and their variances to see if they converge to the theoretical values.

```r
set.seed(123)
nb_obs <- 100000
sample_means <- NULL
sample_vars <- NULL
for (i in 1 : nb_obs) {
    f <- rexp(40, lambda)
    sample_means <- c(sample_means, mean(f))
}
```

Sample mean versus theoretical mean

The mean of the 100 000 sample means is 4.9970167, which is pretty close to the theoretical mean of 5. This plot shows the mean of the sample means converging to the theoretical mean as the number of simulations increases.

```r
cum_means <- cumsum(sample_means) / (1:nb_obs)
```



Sample variance versus theoretical variance

The sample variance over the 100000 sample means is 0.6182664, which is far from the theoretical variance of 25.

The Central Limit Theorem tells us that with enough observations, the standard deviation of the means of the samples can be approximated by the sampling standard error (SE).

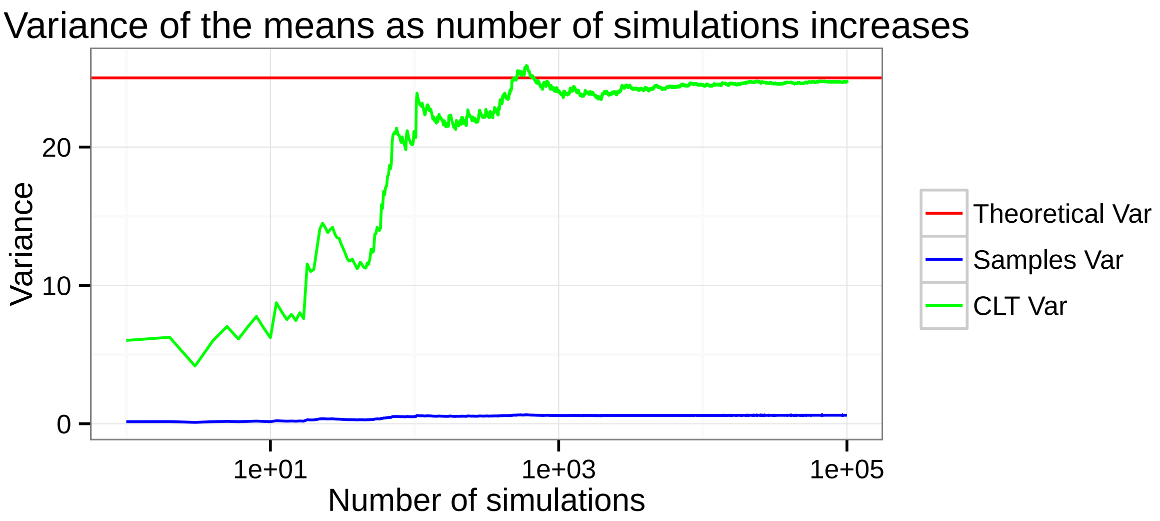
$$\sigma_{sample} \rightarrow SE = \sqrt{\frac{Var}{n}}$$

In other words, we should see that with enough simulations, the following should tend towards the theoretical variance.

$$\sigma^2_{sample} \times n \rightarrow Var$$

The plot below shows the variance of the samples and the variance os the samples adjusted as per the Central Limit Theorem, compared to the theoretical variance.

```r
clt_vars <- NULL
vars <- NULL
# we calculate the variance of the means, starting with the first 2 means
for (i in 2:nb_obs) {
    clt_vars <- c(clt_vars, var(sample_means[1:i]) * 40)
    vars <- c(vars, var(sample_means[1:i]))
}
```

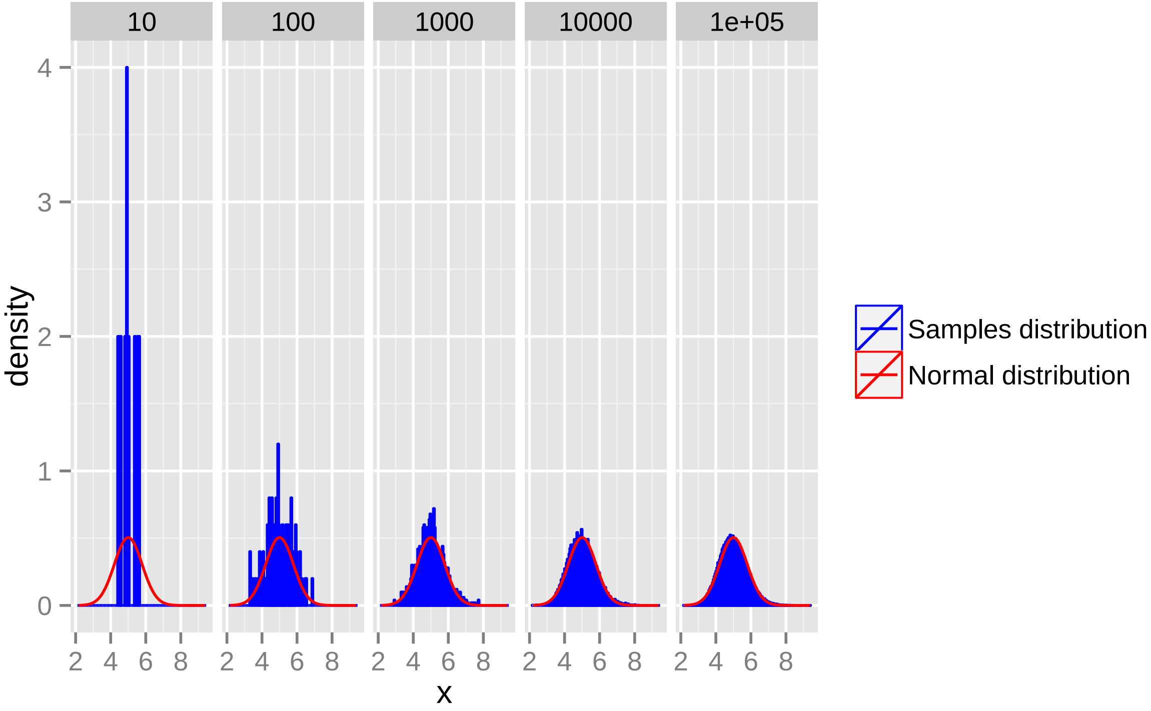


Distribution

The Central Limit Theorem also tells us that the distribution of the means of the samples should be approximately normal, given enough observations.

The normal distribution should be of the form $N(\mu, \sigma^2/n)$.

To show this, let's draw the distribution of the means with different number of observations, along with the theoretical normal distribution:



Conclusion

We can see that with enough observations, the distribution of the means of the samples is close to the normal distribution as given by the Central Limit Theorem.