Statistical Inference Project - Part 1: Simulation

Abdelaziz ALIAT

December 22, 2015

Overview

In this Statistical Inference project of Coursera, we investigate the exponential distribution to compare it with the Central Limit Theorem (CLT). We assume its mean and standard deviation to be to be both 1/lambda (lambda = 0.2 for all simulations). For this project, we consider a distribution of averages of 40 exponentials over 1000 simulations.

Goals

1. Show the sample mean and compare it to the theoretical mean of the distribution
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution
3. Show that the distribution is approximately normal

This work

1. **Show the sample mean and compare it to the theoretical mean of the distribution**

The sample mean is given by:

```
set.seed(123)
# Lambda
lambda<- 0.2
# Number of exponentials
n_exp<- 40
# Number of simulations
n_sim<- 1000

# Sample (40 exponentials, 1000 simulations)
sample <- matrix(rexp(n_sim*n_exp,rate=lambda),n_sim)
# Sample means
sample_means <- apply(sample, 1, mean)

# Sample mean
mean(sample_means)
```

```
## [1] 5.011911
```

The theoritical mean of the distribution is given by:

```
# Theoritical mean
1/lambda
```

```
## [1] 5
```

Answer : the sample mean is close to the theoretical mean of the distribution.

2. **Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution**

The sample variance is given by:

```
# Sample variance
var(sample_means)
```

```
## [1] 0.6088292
```

The theoritical variance of the distribution is given by:

```
# Theoritical variance
(1/lambda)^2/n_exp
```

```
## [1] 0.625
```

Answer: the sample variance is close to the theoretical variance of the distribution.

3. **Show that the distribution is approximately normal**

```
require(ggplot2)
```

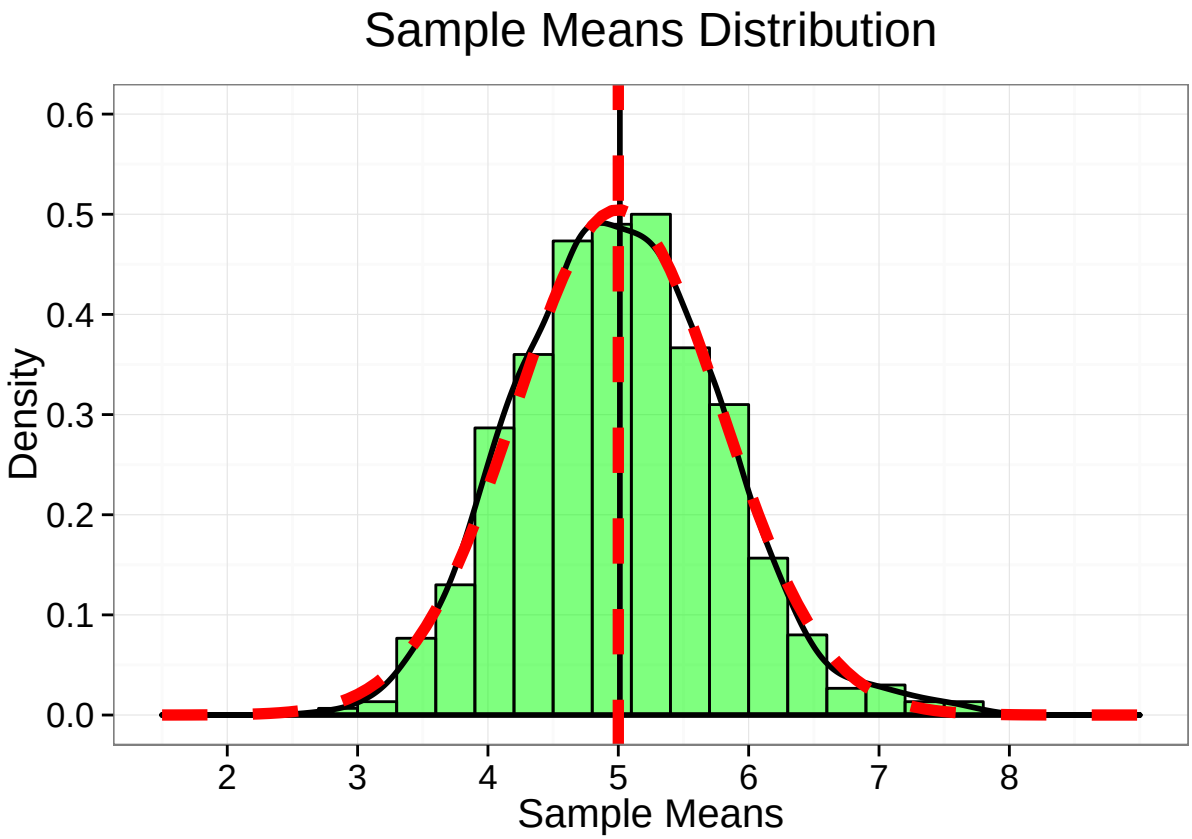
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.2.2
```

```
# Plot the sample means distribution
# to compare it with a normal shape

sample_means<-as.data.frame(sample_means)

ggplot(data=sample_means,aes(x=sample_means))+
  geom_histogram(aes(y=..density..), alpha = .5, fill ='green',binwidth = 0.3,col='black')+
  geom_density(lwd=1,col='black')+
  guides(fill=F)+theme_bw()+
  xlab('Sample Means')+ylab('Density')+
  ggtitle('Sample Means Distribution')+
  theme(axis.text= element_text(size=13),axis.title= element_text(size=15,vjust= 0.5)
        ,title=element_text(size=15,vjust = 2))+
  scale_x_continuous(limits=c(1.5,9),breaks=seq(0,8,1))+
  scale_y_continuous(limits=c(0,0.6),breaks=seq(0,0.6,0.1))+
  geom_vline(xintercept = mean(sample_means$sample_means), size = 1,col='black')+
  stat_function(fun = dnorm, arg = list(mean = 1/lambda,
                                        sd = sqrt((1/lambda)^2/n_exp)),col='red',lwd=2,lty=2)+
  geom_vline(xintercept = 1/lambda, size = 2,lty=2,col='red')
```



Answer: we can see from the above plot that the sample means density (black full line) is almost normal (red dashed line).

General conclusions: A sample from an exponential distribution (40 exponentials and 1000 simulations) is analysed in this project. We have shown that the mean and the variance of this sample match with the theoritical values. Moreover, the sample means distribution almost follows a normal shape as expected by the CLT.