Statistical Inference Course Project - Part 1: Simulation Exercise Instructions

Omer Shechter

October 13, 2018

Overview

This project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials.

Simulations

This part includes the simulation itself. Create a simulation vector of 1000 simulation, and calculating the mean of 40 samples in each simulation.

Load required libraries

```
library(ggplot2)
library(gridExtra)
```

Create the actual simulation : Using the apply function to create 1000 simulation of 40 sampling and calculating the mean.

```
set.seed(56)
nosim<-1000
n<-40
lambada<-0.2
sim_vect = c(apply(matrix(rexp(nosim * n,lambada), nosim), 1, mean))
```

Sample Mean versus Theoretical Mean

Distribution for the mean of random samples The expected value of the mean of the distribution of means = expected value of the sample mean = population mean * E[¯X] = µ

In this case Expected mean = µ = 1/lambada : 5

Calculate the simulated sample mean:

```
exp_mean<-mean(sim_vect)
print(exp_mean)
```

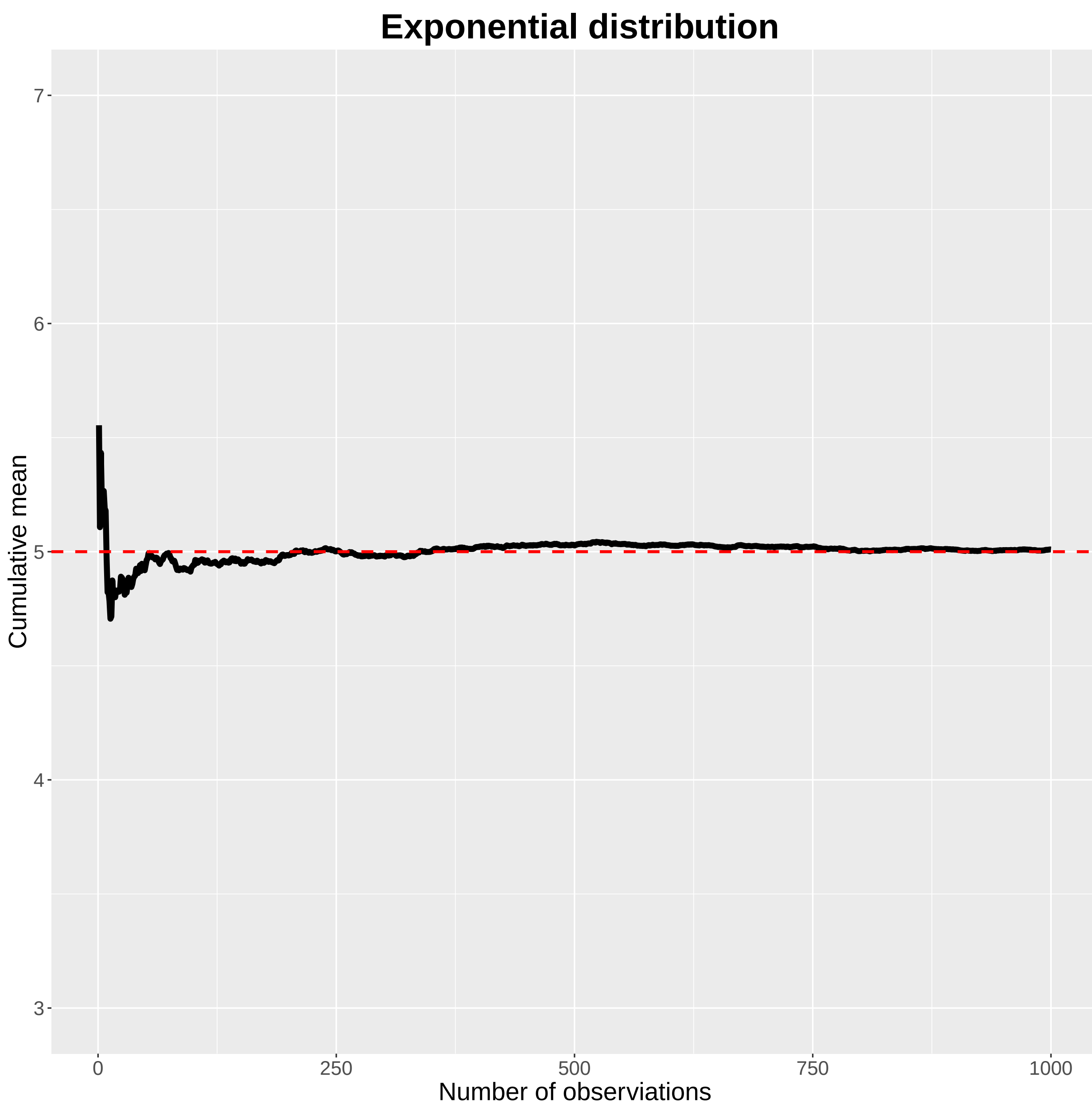
```
## [1] 5.009432
```

plot the distribution of sample means as n increases, and compare it to the population mean.

```
#calculate sample  means for different size of n
means <- cumsum(sim_vect) / (1 : nosim)
#plot sample size vs. sample mean as a function of the number of samples
theme_update(plot.title = element_text(hjust = 0.5))
#plot using ggplot
g <- ggplot(data.frame(x = 1 : nosim, y = means), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 0) + geom_line(size = 2)
g <- g + labs(x = "Number of observiations", y = "Cumulative mean")
```

```r
g <- g + ggtitle(" Exponential distribution")
g <- g + geom_hline(yintercept=5, linetype="dashed", color = "red",size=1)
g <- g + ylim(3,7)
#g <- g + scale_y_continuous(breaks=c(3,4,5,6,7))
g <- g +  theme(axis.text = element_text(size = 14),
        axis.title = element_text(size = 18),
        plot.title = element_text(size = 25, face = "bold"))
g
```

```
## Warning: Removed 1 rows containing missing values (geom_hline).
```



Summary: As it can be seen from the results above The sample mean provides an accurate estimator to the general population mean.

Sample Variance versus Theoretical Variance

Variance of distribution of $\bar{X}$ = Sigma^2/n estimate of variance = S2/n

In our example the Variance estimator is (1/Lambada/sqrt(n))^2 = 25/40 = .625

Calculate the simulated variance

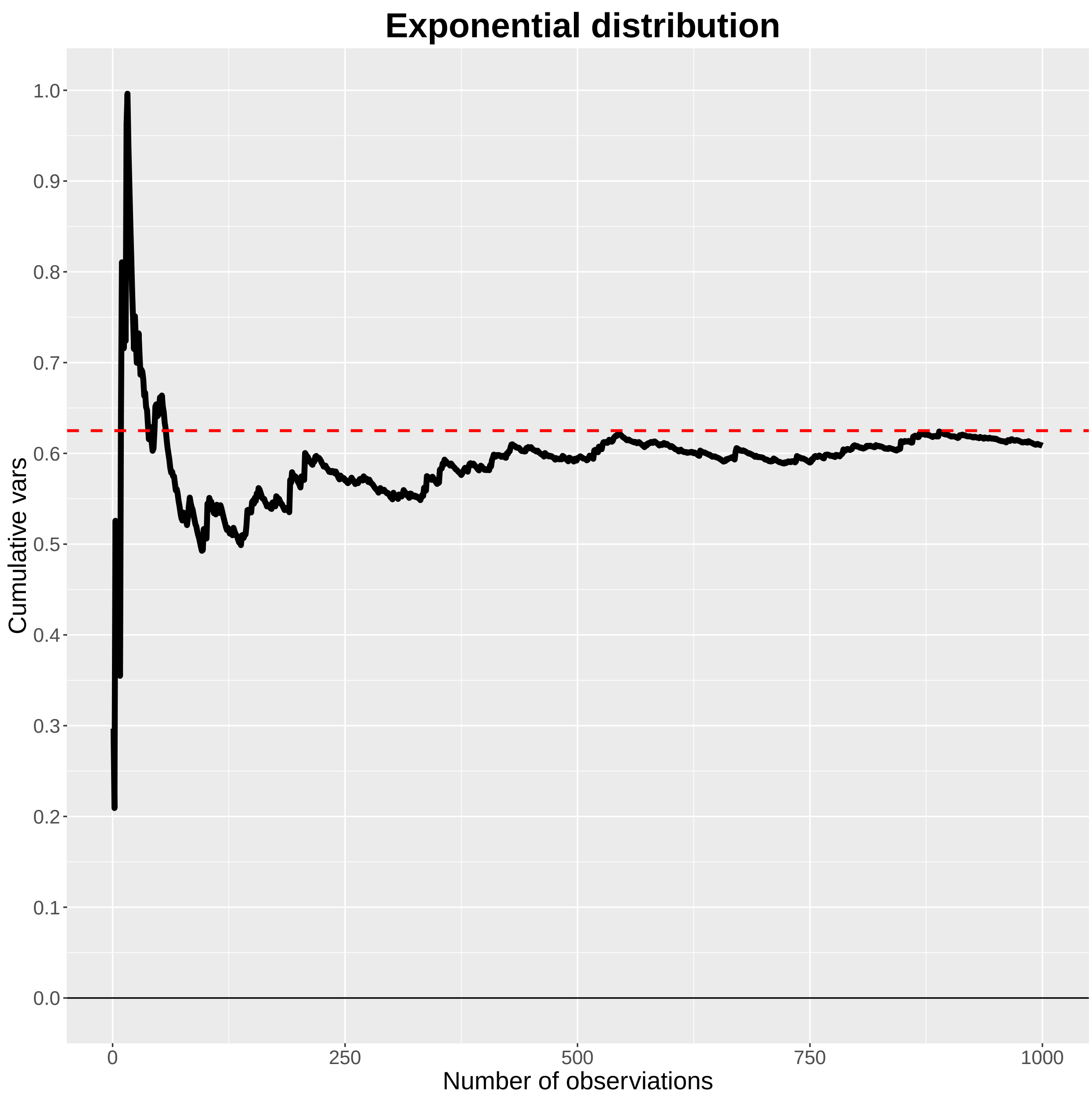
```r
e_var<-(1/lambada/sqrt(n))^2
var(sim_vect)
```

```
## [1] 0.6090837
```

Plot the sample variance as n increases and compare it to the population vars

```r
#calculate sample  vars for different size of n
e_mean<-mean(sim_vect)
cvars <- cumsum((sim_vect-e_mean)^2)/seq_along((sim_vect-1))
#plot sample size vs. sample's var
theme_update(plot.title = element_text(hjust = 0.5))

g <- ggplot(data.frame(x = 1 : nosim, y = cvars), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 0) + geom_line(size = 2)
g <- g + labs(x = "Number of observiations", y = "Cumulative vars")
g <- g + ggtitle(" Exponential distribution")
g <- g + geom_hline(yintercept=e_var, linetype="dashed", color = "red",size=1)
g <- g + scale_y_continuous(breaks=seq(0, 1, length.out = 11))
g <- g +  theme(axis.text = element_text(size = 14),
        axis.title = element_text(size = 18),
        plot.title = element_text(size = 25, face = "bold"))
g
```

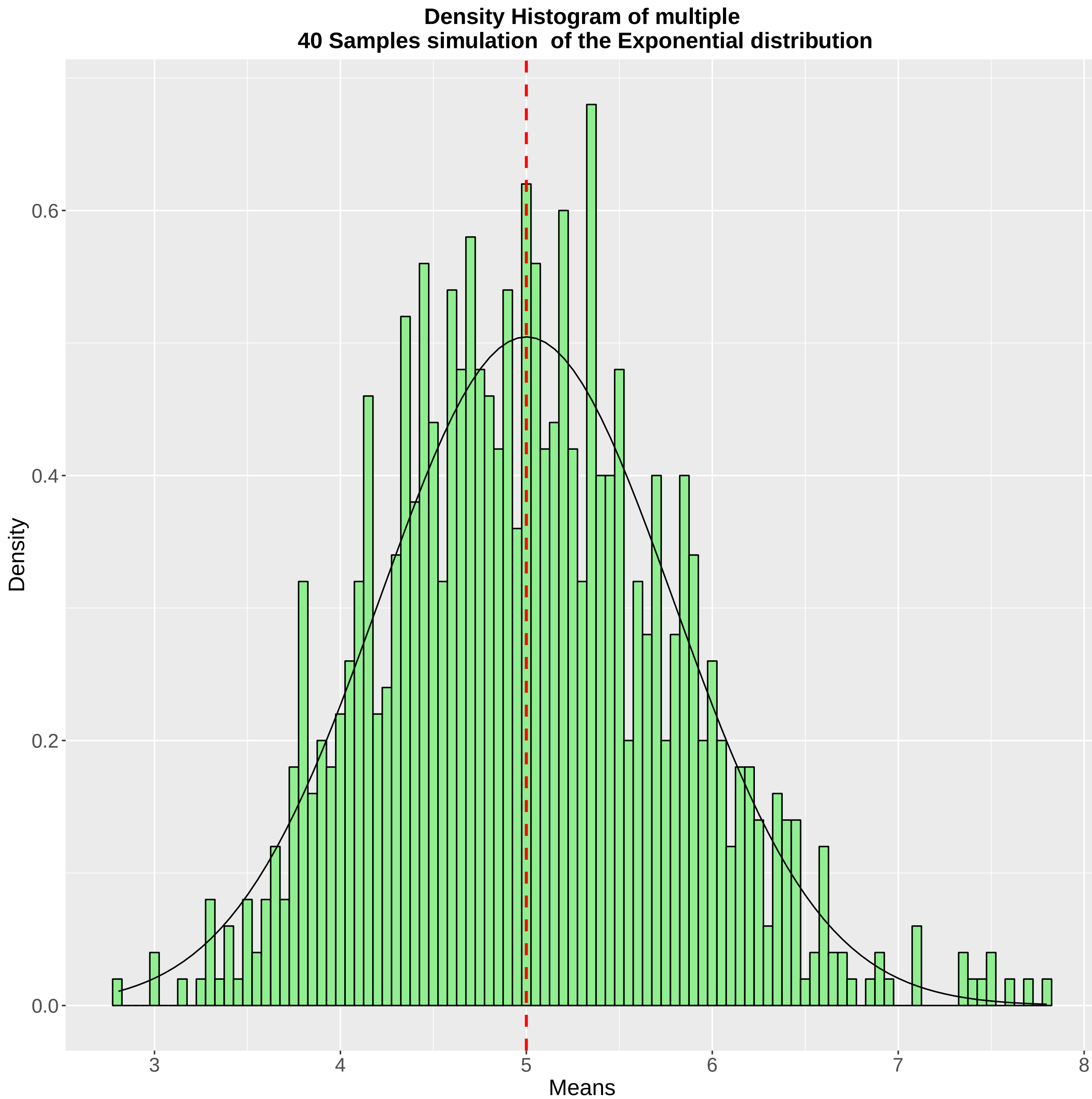


Distribution

Plot a histogram of the density of the sample's mean Show that it is close to the normal distribution.

```
g <- ggplot(data.frame(mu = sim_vect),aes(x=mu))
g <- g + geom_histogram(aes(y=..density..),color="black",fill="lightgreen",binwidth = .05)
g <- g + stat_function(fun=dnorm , color="black",args=list(mean=5,sd=sqrt(1/lambada/sqrt(n))^2))
g <- g + labs(x = "Means", y = "Density")
g <- g + ggtitle(" Density Histogram of multiple \n 40 Samples simulation  of the Exponential distributi
g <- g + geom_vline(xintercept=1/lambada, linetype="dashed", color = "red",size=1)
g <- g +  theme(axis.text = element_text(size = 14),
        axis.title = element_text(size = 16),
        plot.title = element_text(size = 16, face = "bold"))
```

```
g
```



Conclusion: The above three sections give a good indication that the mean of samples (using R simulation) of a population that is IID with exponential distribution behave like a normal distribution. The expected mean is 1/lambada, and the variance is (1/Lambada/sqrt(n))^2.