Part 1 Simulation Exercise Instructions

Nelson Sánchez Sánchez

2/1/2021

Overview

In 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. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.

Setting a seed for reproducibility and set variables lambda, exponentials and number of simulations:

```
set.seed(321)
n <- 40
lambda <- 0.2
numSimulations <- 1000
```

Load the following R packages `ggplot2`.

```
library(ggplot2)
```

You should

Sample Mean versus Theoretical Mean:

Generating 1000 samples of 40 exponentials and calculating their mean values:

```
r <- replicate(numSimulations, rexp(n, lambda))
dim(r)
```

```
## [1]   40 1000
```

```
class(r)
```

```
## [1] "matrix" "array"
```

We can see that r is a matrix of 40 rows and 1000 columns. Since each column contains a sample of 40 random exponentials we'll apply mean() to columns to get 1000 sample means.

```r
exp_means <- apply(r, 2, mean)
```

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

Sample(empirical) mean:

```r
e_mean <- mean(exp_means)
e_mean
```

```
## [1] 4.974996
```

Theoretical mean:

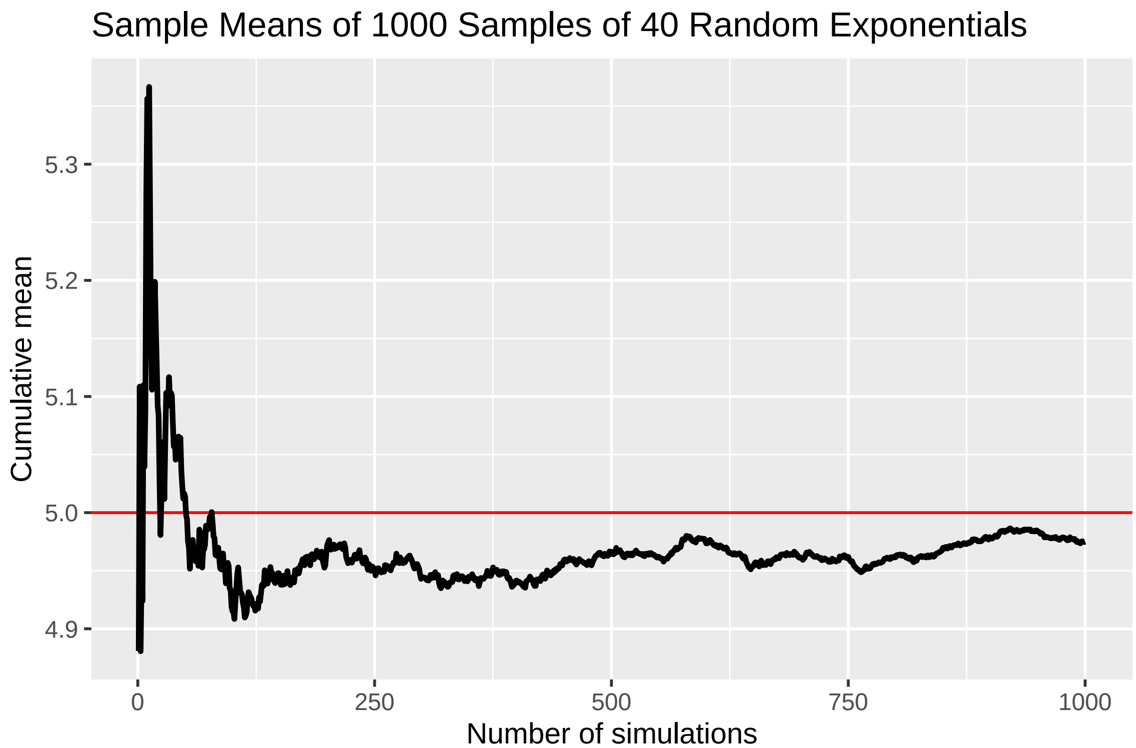
```r
t_mean<- 1/lambda
t_mean
```

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

Figures comparing.

We can see that the sample mean 4.974996 is good approximation of the theoretical mean t_mean = 5.

```r
means <- cumsum(exp_means)/(1:numSimulations)
library(ggplot2)
g <- ggplot(data.frame(x = 1:numSimulations, y = means), aes(x = x, y = y))
g <- g + geom_hline(yintercept = t_mean, colour = 'red') + geom_line(size = 1)
g <- g + labs(x = "Number of simulations", y = "Cumulative mean")
g <- g + ggtitle('Sample Means of 1000 Samples of 40 Random Exponentials ')
g
```



As we can see from the graph, empirical mean is a consistent estimator of theoretical mean, because it converges to the value of theoretical mean.

Sample Variance versus Theoretical Variance:

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

Figures comparing

Sample variance:

```
e_var <- var(exp_means)
e_var
```

```
## [1] 0.5900217
```

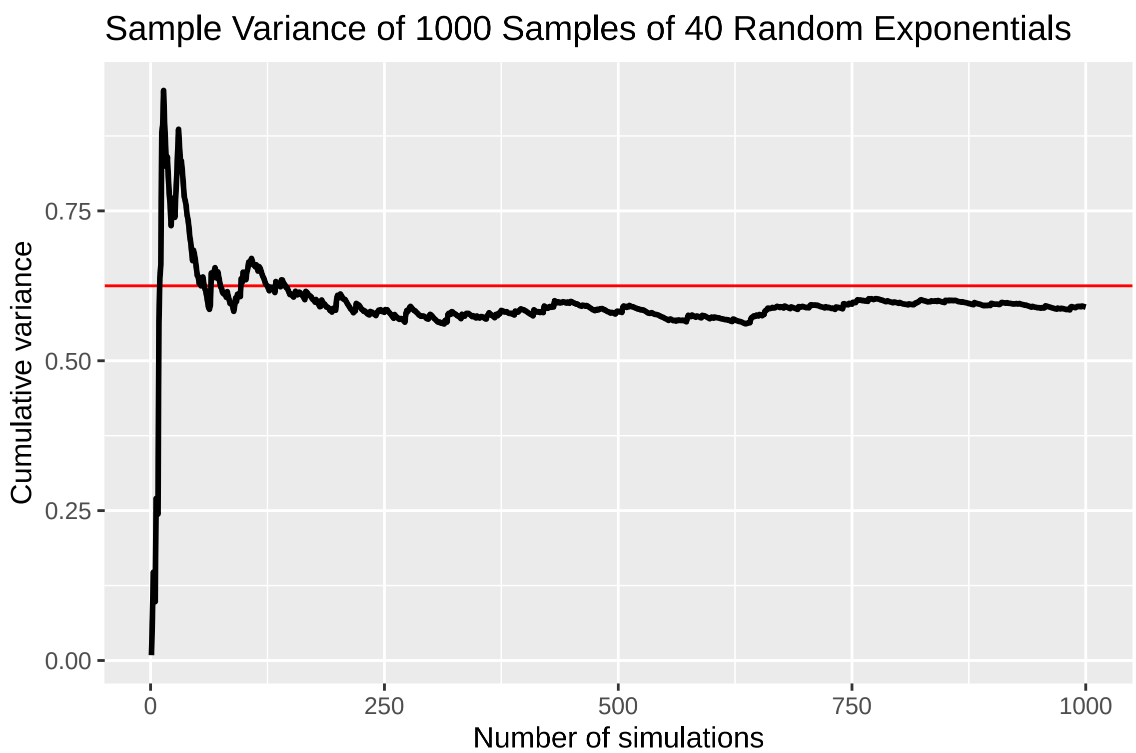
Theoretical variance of the distribution of sample means:

```
t_var <- (1/(lambda*sqrt(n)))^2
t_var
```

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

As we can see, both empirical and theorietical variance of the distribution of sample means have value close to 0.6.

```
cumvar <- cumsum((exp_means - e_mean)^2)/(seq_along(exp_means - 1))
g <- ggplot(data.frame(x = 1:numSimulations, y = cumvar), aes(x = x, y = y))
g <- g + geom_hline(yintercept = t_var, colour = 'red') + geom_line(size = 1)
g <- g + labs(x = "Number of simulations", y = "Cumulative variance")
g <- g + ggtitle('Sample Variance of 1000 Samples of 40 Random Exponentials ')
g
```



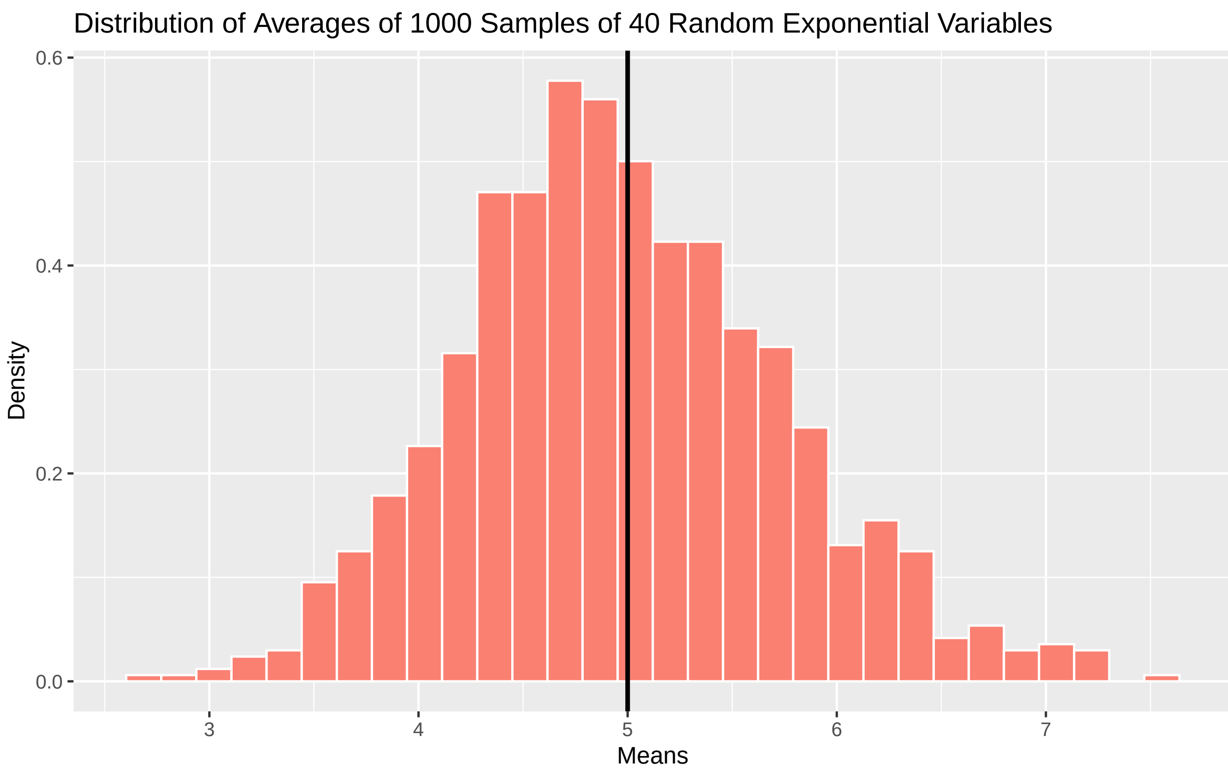
As we can see from the graph, sample variance is a consistent estimator of the theoretical variance, because it converges to the value of the theoretical variance.

Distribution: - Show that the distribution is approximately normal

In this point 3, focus is on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

Figures explaining how one can tell the distribution is approximately normal.

```
g <- ggplot(data.frame(x = exp_means), aes(x = x))
g <- g + geom_histogram(aes(y = ..density..), colour = 'white', fill = 'salmon')
g <- g + stat_function(fun = dnorm, colour = 'black', args = list(mean = t_mean, sd = sd))
g <- g + geom_vline(xintercept = t_mean, colour = 'black', size = 1)
g <- g + ggtitle('Distribution of Averages of 1000 Samples of 40 Random Exponential Variables')
g <- g + xlab('Means')
g <- g + ylab('Density')
g
```



As we can see from the graph the distribution of averages of 1000 samples of 40 iid exponentials is approximately normal.