Statistical Inference Course Project

Jimmy

6/6/2020

Statistical Inference Course Project s

Part 1 - A Simulation Exercise

Overview

In this project we 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. We will set lambda = 0.2 for all of the simulations. We will investigate the distribution of averages of 40 exponentials. We will do a thousand simulations.

Setting a seed for reproducibility:

```
set.seed(111)
```

Setting a rate parameter lambda:

```
lambda = 0.2
```

Setting a size of a sample:

```
n = 40
```

Setting a number of simulations:

```
nosim = 1000
```

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

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

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

```
class(r)
```

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

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.

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

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

Sample(empirical) mean:

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

```
## [1] 5.02562
```

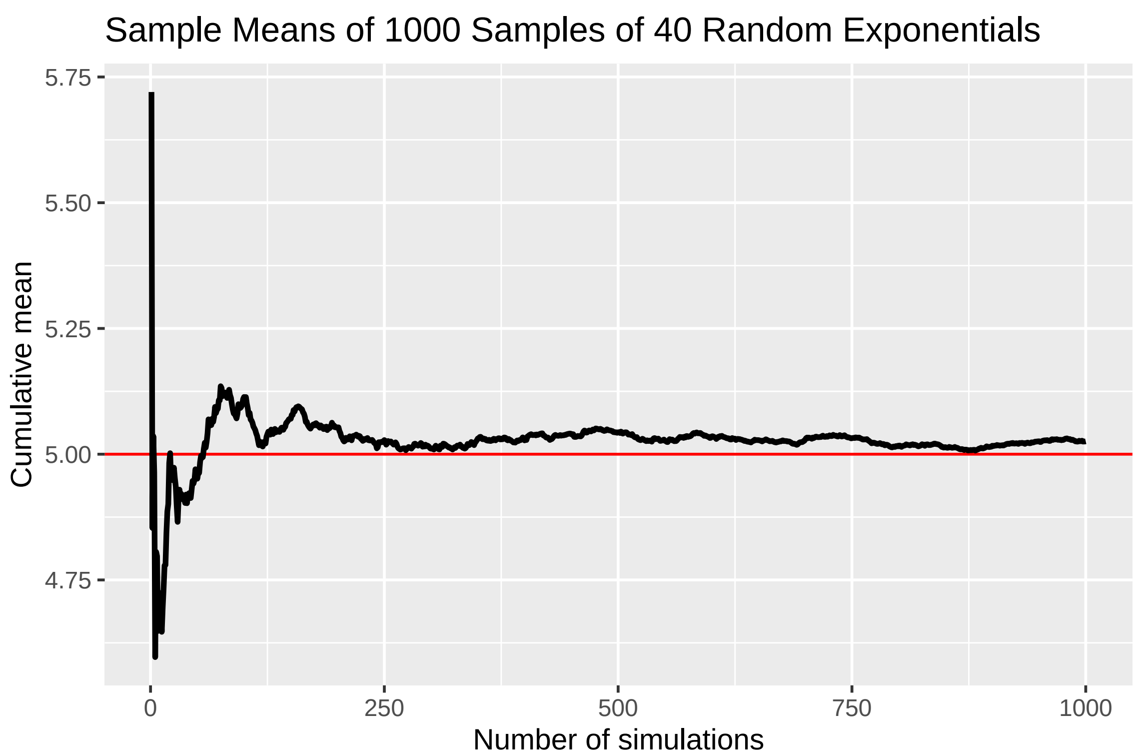
Theoretical mean:

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

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

We can see that the sample mean 5.02562 is good approximation of the theoretical mean t_mean = 5.

```
means <- cumsum(exp_means)/(1:nosim)
library(ggplot2)
g <- ggplot(data.frame(x = 1:nosim, 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.

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

Sample variance:

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

```
## [1] 0.6069798
```

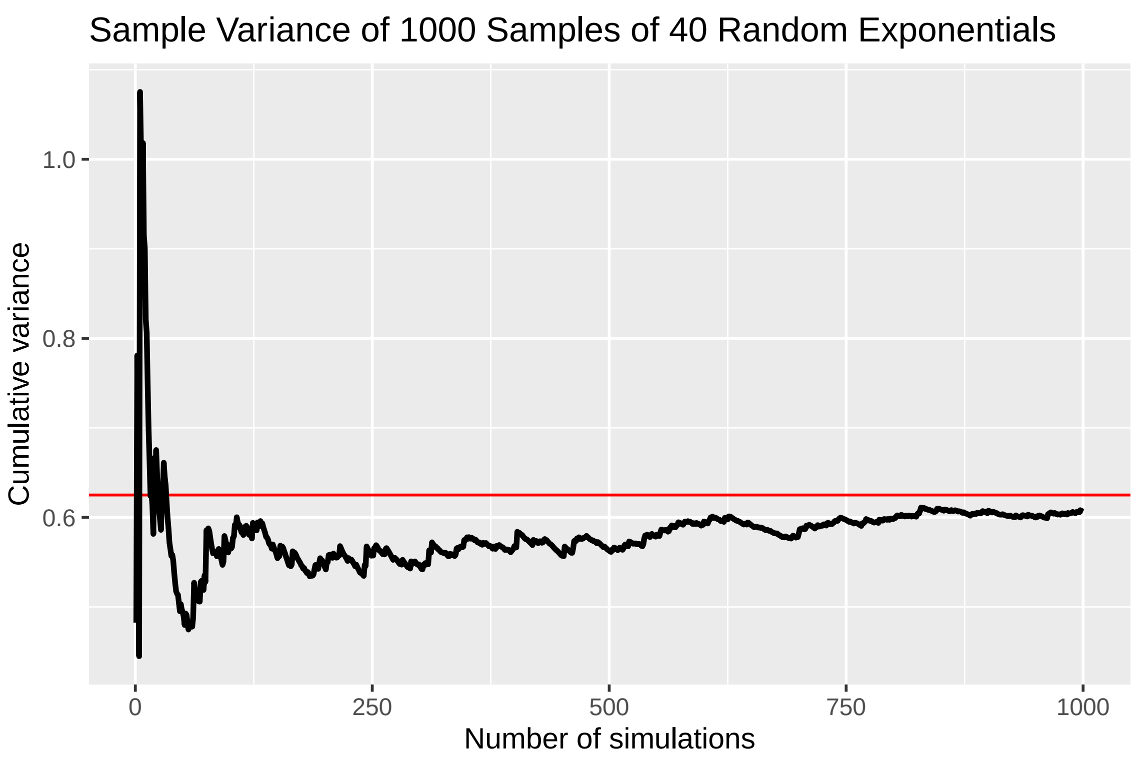
Theoretical variance of the distribution of sample means:

```
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:nosim, 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.

3. Show that the distribution is approximately normal.

```
g <- ggplot(data = 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))
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.