Statistical Inference Project. Part 1.

Sergey Rakhmatullin

Overview

In this project we are going to study exponential distribution and simulate how the Central Limit Theorem works. We will perform simulations and study the behavior of the mean/variance of the distribution depending on the number of the simulations and show that when the number of simulations is large enough we are close to the normal distributed data.

Simulations

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.

Setting given parameters:

```
library(ggplot2)
library(reshape2)
require(dplyr, quietly=T, warn.conflicts = F)
set.seed (199) #seed for random data
lambda <- 0.2 # Setting lambda = 0.2 for all of the simulations.
n <- 40 # We are to investigate the distribution of averages of 40 exponentials.
sim_num <- 1000 # We are to do a thousand simulations.
Theory_Mean <- 1/lambda # Theoretical mean
Theory_Variance <- 1/(lambda^2*(n-1)) # Theoretical variance of the n exponentials sample
Theory_SD <- 1/(lambda*sqrt(n)) # Theor. standard deviation of the n exponentials sample
```

Simualting the distribution of 1000 averages of 40 random exponentials:

```
means = NULL
for (i in 1:sim_num) means = c(means, mean(rexp(n, lambda)))
```

Simulating collection of 1000 random exponentials:

```
set.seed (199) #seed for random data
exp_distr <- rexp(sim_num, lambda)
```

Sample Mean and Variance versus Theoretical Mean and Variance

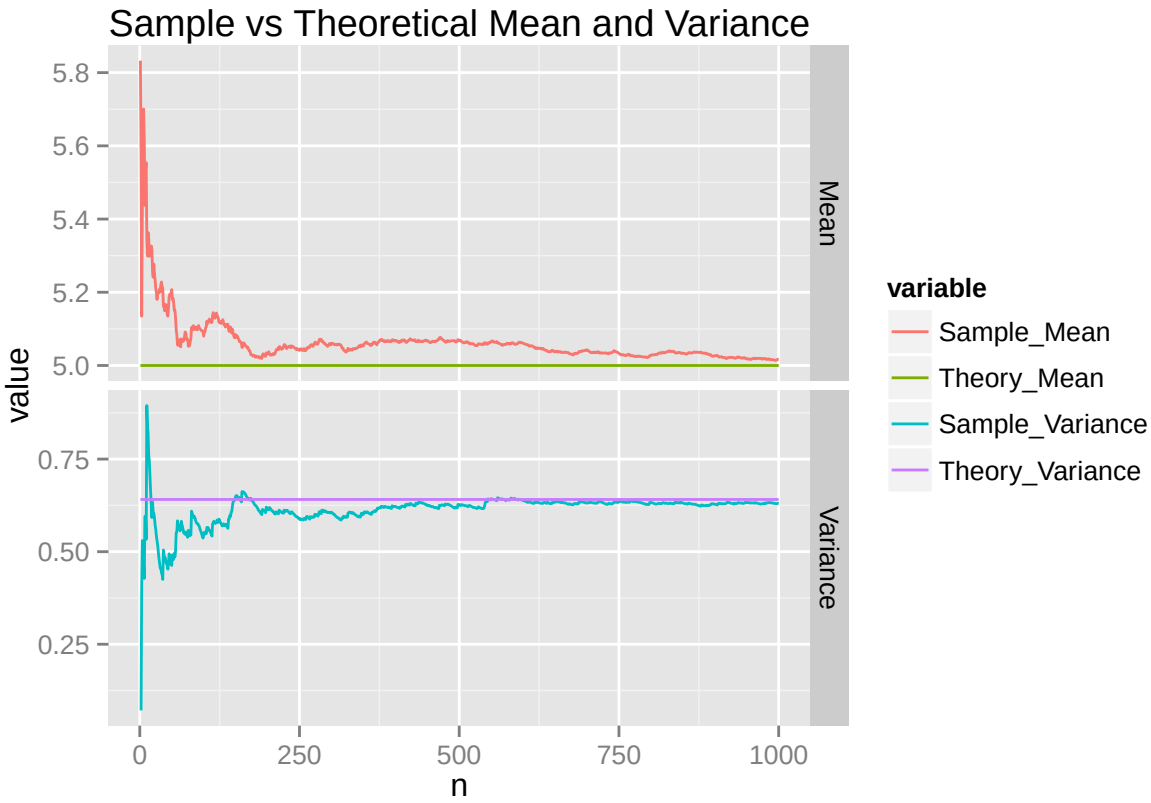
Let's analyze dependancy of average and variance of 40 exponentials on the number of the simulations (from 2 to 1000)

```
Sample_Mean <- cumsum(means)/seq_along(means) # Cumulative mean calculation
Sample_Variance <- cumsum((means-Sample_Mean)^2)/(seq_along(means)-1)
                                              # Cumulative variance calculation

data <- data.frame(cbind(n=seq_along(Sample_Mean), Sample_Mean,
                         Theory_Mean, Sample_Variance, Theory_Variance))
```

```
data1<-melt(data, id.vars=c('n'), na.rm=T) %>%
        mutate (var2=ifelse(variable=='Sample_Mean' |
                                    variable=='Theory_Mean', 'Mean', 'Variance'))

p1 <- ggplot(data1, aes(n, value, colour=variable))
p1 <- p1 + facet_grid(var2~., scales='free')
p1 + geom_line() + ggtitle('Sample vs Theoretical Mean and Variance')
```



Given graphs show that if n (number of simulations) is large enough sample mean and variance approach to the given theoretical values (horisontal lines on the graphs).

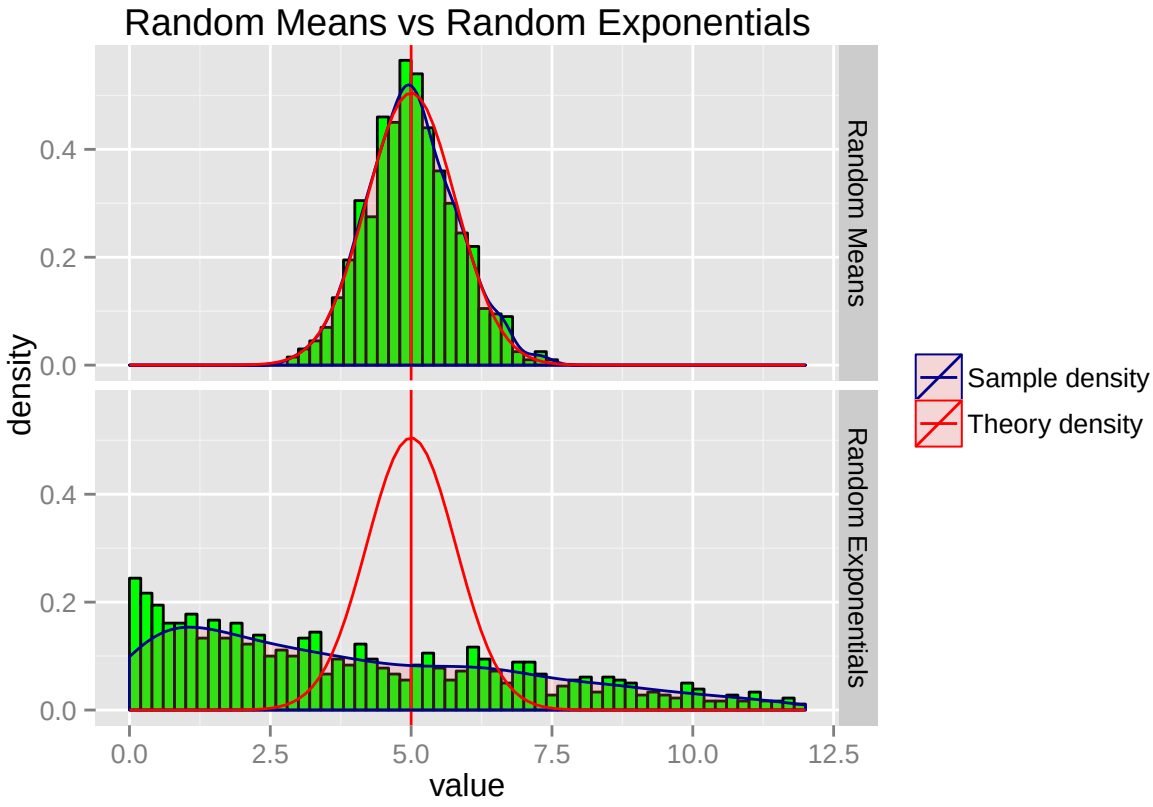
Distribution

Comparing the distribution of a large collection of random exponentials, having it's mean and standard deviation, and the distribution of a large collection of averages of 40 exponentials,it's mean and standars deviation.

```
data<-data.frame(n=seq_along(means), means, exp_distr)
data2<-melt(data, id.vars=c('n'))
levels(data2$variable) <-c('Random Means','Random Exponentials')
d1<-ggplot(data2, aes(x=value))+xlim(0,12)
d1<-d1+geom_histogram(aes(y=..density..), binwidth=0.2, fill='green', colour='black')
d1<-d1+facet_grid(variable~.)
d1<-d1+geom_vline(xintercept=Theory_Mean, colour='red')
d1<-d1+geom_density(alpha=.2, fill="#FF6666", aes(colour='Sample density'), showGuide=T)
```

```
d1<-d1+stat_function(fun=dnorm, args=list(mean=Theory_Mean, sd=Theory_SD),
                     aes(colour='Theory density'), showGuide=T)
d1<-d1+scale_colour_manual("",values = c("darkblue", "red"))
d1 + ggtitle('Random Means vs Random Exponentials')
```

```
## Warning: Removed 100 rows containing non-finite values (stat_density).
```



On the graphs shown you can see that random means density function quite similar to the normal distribution density function (red curve), while the random exponetials density function is totally different. This fact demonstrates how the Central Limit Theorem works on 1000 simulations.