Statistical Inference Course Project

Author: Rustam Mansyrov

According to the questions stated, this project has several concerns regarding the comparisons of sample and theoretical means, sample and theoretical variances and the practical implementation of Central Limit Theorem. The distribution upon which the computations are made is defined to be exponential with fixed rate lambda = 0.2.

Simulations

To start with, the empty vector having length of 1000 was created and used for 'for-loop' in order to obtain averages of 40 exponentials 1000 times. In other words, here we simulated our sample mean:

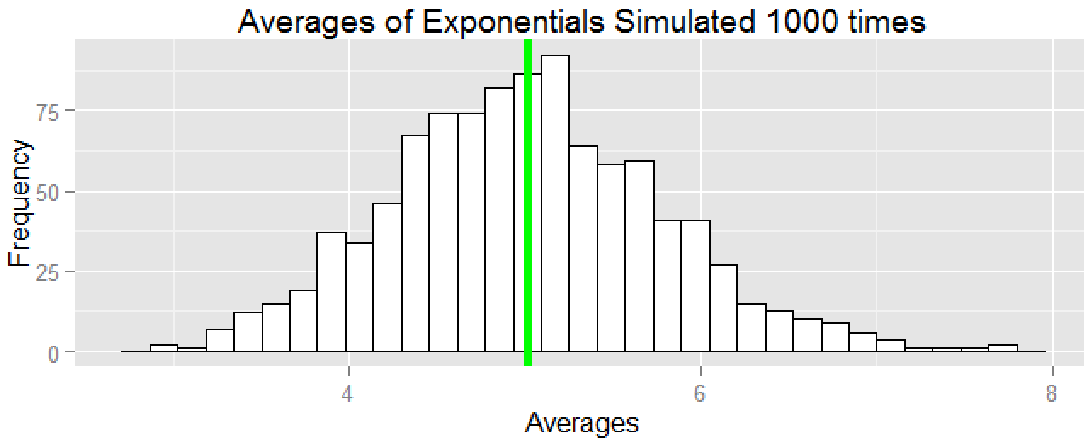
```
set.seed(123)
averages <- vector("numeric", length = 1000)
for(i in 1:1000){
        temp <- NULL
        temp <- rexp(40, rate = 0.2)
        averages[i] <- mean(temp)
}
sample_mean <- mean(averages)
```

Next, to be able to compare sample and theoretical variances of exponential distribution, there are simulated 1000 variances of exponentially distributed random numbers and found the average of these variances, which is further stated to be sample variance:

```
set.seed(123)
variances <- vector("numeric",1000)
for(i in 1:1000){
        temp <- NULL
        temp <- rexp(40, rate = 0.2)
        variances[i] <- var(temp)
}
sample_variance = mean(variances)
```

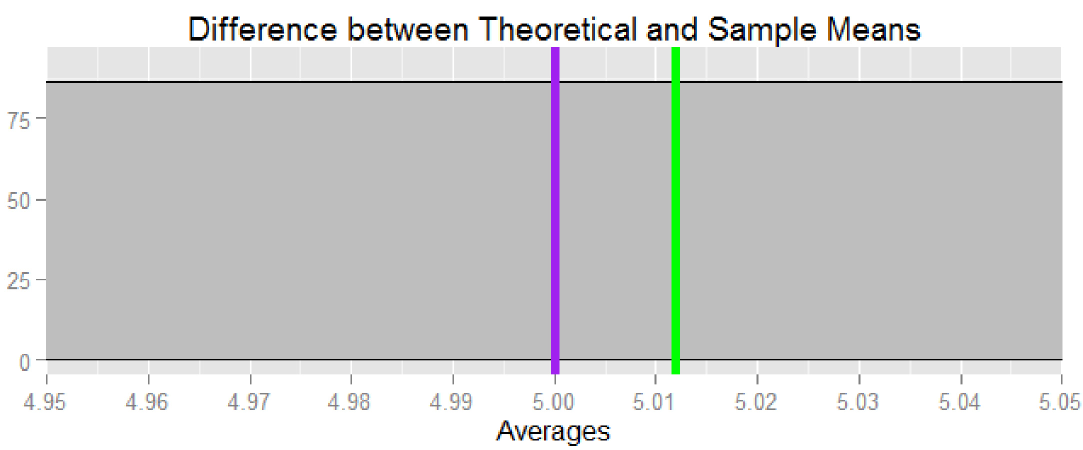
Sample Mean versus Theoretical Mean

After simulating 1000 averages of exponential distribution, the shape of distribution along with the sample mean is presented in a graph as follows:



Apparently from the histrogram above, it is seen that the green line represents the sample mean and it is quite close to theoretical mean which is a reciprocal of 0.2, that is, it is equal to 5.

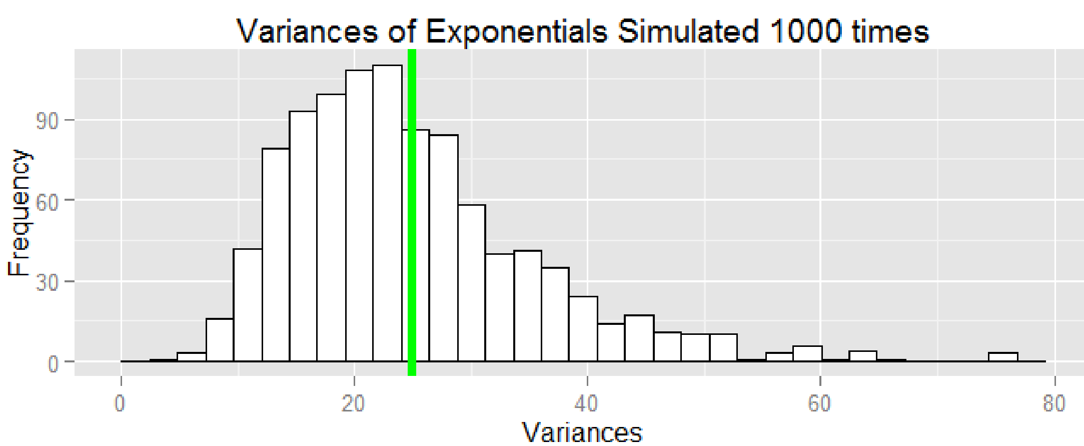
To observe the precise difference, it is useful to look at the following plot:



On the plot, the purple line obviously stands for the theoretical mean and the green - sample mean.

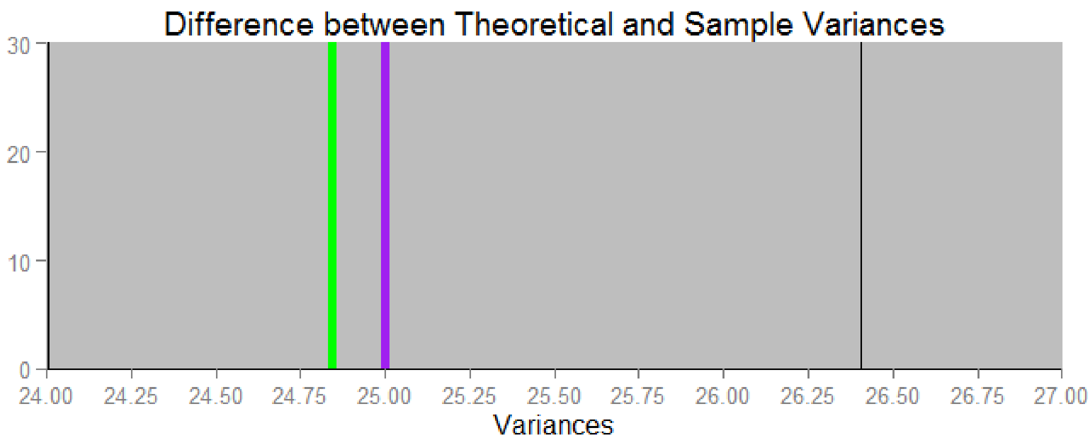
The error term(small deviation of sample mean from theoretical one) is insignificant enough to state that 1000 simulations yield a very good approximation of mean.

Next, aforementioned simulation of variances is also analyzed in a following plot:



It can be seen that the average of all variances is somwhere around the theoretical variance which is equal to squared reciprocal of rate lambda, i.e, 25.

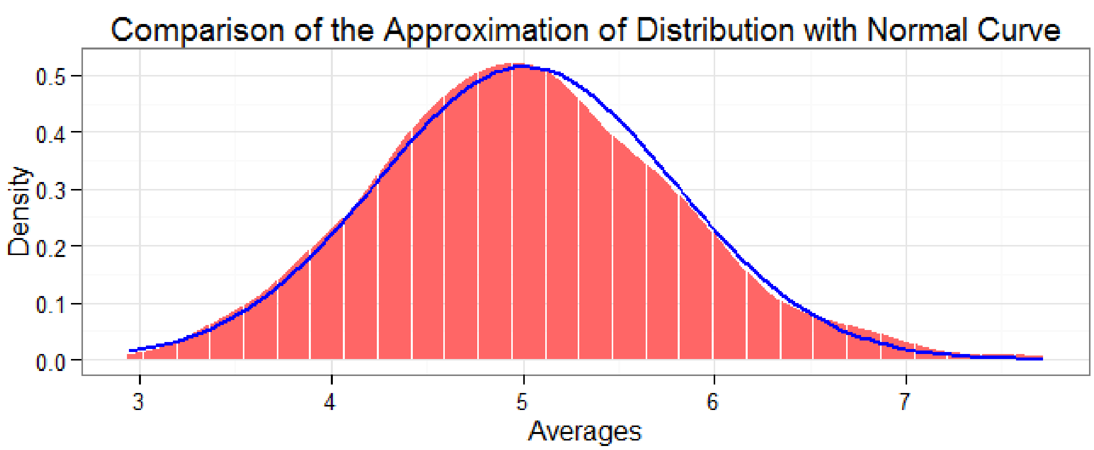
Again, to see the precise difference, as for mean, the same was applied to variances:



On the plot, the purple line obviously stands for the theoretical variance and the green - sample variance.

As concluded for the sample mean, the same might be said for the sample variance. That is, 1000 simulations yield very good approximation of sample variance and deviation of it from the theoretical variance is very small.

Finally, checking the normality of distribution might be accomplished using several techniques.



Apparently, what we can see is the bell - shaped distribution indicating the normality and small deviations imply that the approximation is quite successful.

Appendix

R codes:

```
set.seed(123)
averages <- vector("numeric", length = 1000)
for(i in 1:1000){
        temp <- NULL
        temp <- rexp(40, rate = 0.2)
        averages[i] <- mean(temp)
}
sample_mean <- mean(averages)
set.seed(123)
variances <- vector("numeric",1000)
for(i in 1:1000){
        temp <- NULL
        temp <- rexp(40, rate = 0.2)
        variances[i] <- var(temp)
}
sample_variance = mean(variances)
require(ggplot2)
sample_mean <- mean(averages)
theoretical_mean <- 1/0.2
qplot(averages) + geom_histogram(color = "black", fill = "white") + geom_vline(xintercept
= mean(averages), col = "green", linetype = 1, size = 2)   + labs(title = "Averages of
Exponentials Simulated 1000 times", x = "Averages", y = "Frequency")

qplot(averages) + geom_histogram(color = "black",fill = "grey") + coord_cartesian(xlim =
c(4.950,5.05)) + geom_vline(xintercept = mean(averages), col = "green", size = 2) +
geom_vline(xintercept = theoretical_mean, col = "purple", size = 2) + scale_x_continuous(
breaks = seq(0,6, by = 0.01))+labs(title = "Difference between Theoretical and Sample
Means",x = "Averages", y = "")

theoretical_variance <- (1/0.2)^2
qplot(variances) + geom_histogram(color = "black", fill = "white")+geom_vline(xintercept
= mean(variances),col = "green", linetype = 1, size = 2)+labs(title = "Variances of
Exponentials Simulated 1000 times", x = "Variances", y = "Frequency")

qplot(variances) + geom_histogram(color = "black",fill = "grey") + coord_cartesian(xlim =
c(24,27), ylim = c(0,30)) + geom_vline(xintercept = mean(variances), col = "green", size
= 2) + geom_vline(xintercept = theoretical_variance,col = "purple", size = 2) +
scale_x_continuous( breaks = seq(20,30, by = 0.25)) +  labs(title = "Difference between
Theoretical and Sample Variances", x = "Variances", y = "")
options(warn = -1)
g <- ggplot(, aes(averages))
g + geom_histogram(fill = "#FF6666", stat = "density") + stat_function(fun = dnorm, args
= list(mean = mean(averages), sd = sd(averages)), col = "blue", size = 1) + theme_bw() +
labs(x = "Averages", y = "Density") + ggtitle("Comparison of the Approximation of
Distribution with Normal Curve")
```