Project Statistical Inference

JOHN HOPKINS UNIVERSTY THROUGH COURSERA. January 5th - February 2nd, 2015 Oscar Portillo

Project Description:

The project consists of two parts: a) simulation to explore inference and b) basic inferential data analysis.

A) Simulation Excercise

In this part, the exponential distribution is studied in R and compared with the Central Limit Theorem. The exponential probability distribution describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate [Ref.1]. f(x;lambda) = lambda * exp ^{-lambda *x} for lambda >= 0 and f(x;lambda) = 0 for lambda < 0. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda.In this work, all simulations are performed with lambda = 0.2. A thousand simulations are run to investigate the distribution of averages of 40 exponentials.

Tasks:

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Task 1: First, the parameters required to conduct the simulation are defined:

```
# 40 exponentials, 1000 simulations and lambda is set to 0.2
sample_size<-40
number_simulations<-1000
lambda<- 0.2
```

The average of 40 exponentials are computed using the exponential distribution in R, rexp(n, lambda):

```
simulation = NULL
for (i in 1 : number_simulations) simulation = c(simulation, mean(rexp(sample_size,
lambda)))

# Print summary of the data
summary(simulation)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.123   4.408   4.941   5.011   5.527   7.986
```

The sample mean and theoretical mean of the distrubution are calculated. We observe that the theoretical mean is equal to 5 and the sample mean of the distribution averages of the 40 exponentials is also very close to 5:

```
theoretical_mean <- 1/lambda
simulation_mean<- mean(simulation)
# Print theoretical and sample mean
theoretical_mean; simulation_mean
```

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

```
## [1] 5.010943
```

Task 2: The variance of the sample simulations is also calculated as well as the theoretical variance of the exponential distribution. theoretical mean of the distrubution are calculated. We observed that variance of the sample is approximately equal to the theoretical variance of the distribution, 0.625.

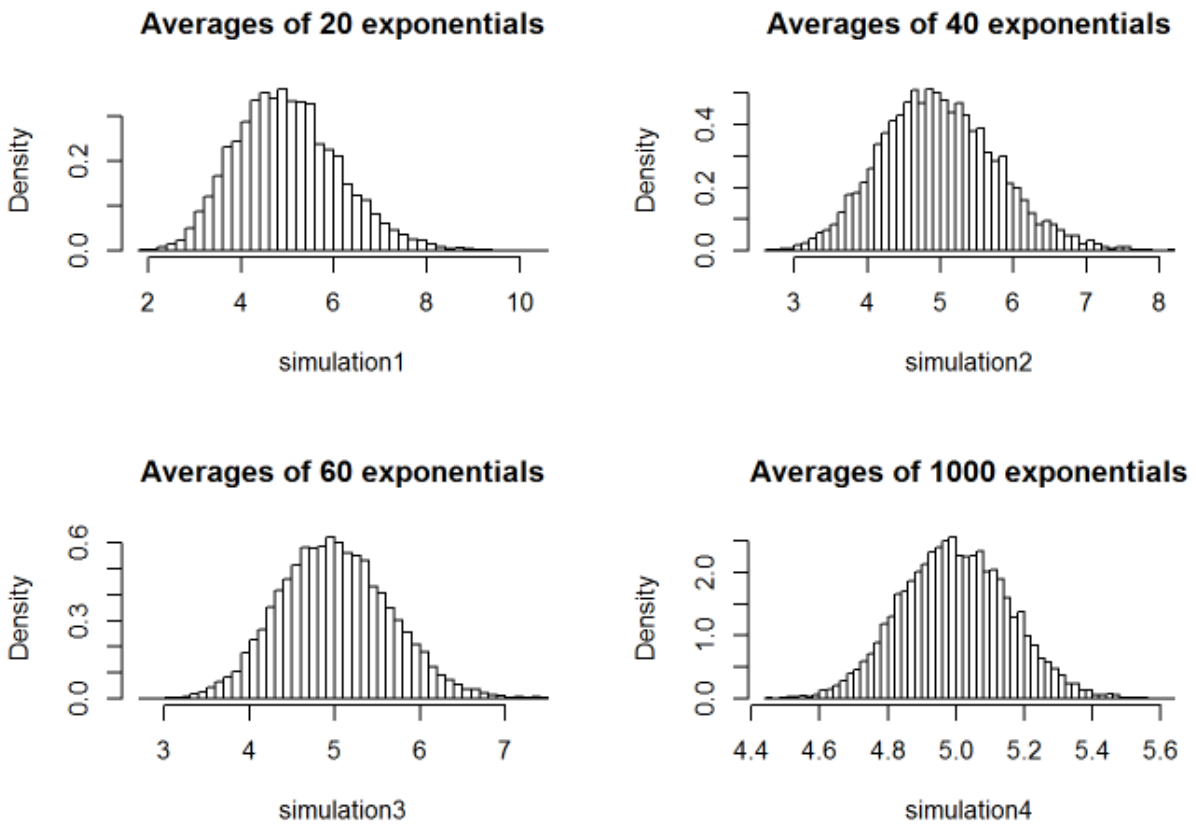
```
theoretical_variance <- ((1/lambda)^2)*(1/sample_size)
simulation_var<-var(simulation)
# Print theoretical and sample variance
theoretical_variance; simulation_var
```

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

```
## [1] 0.6324995
```

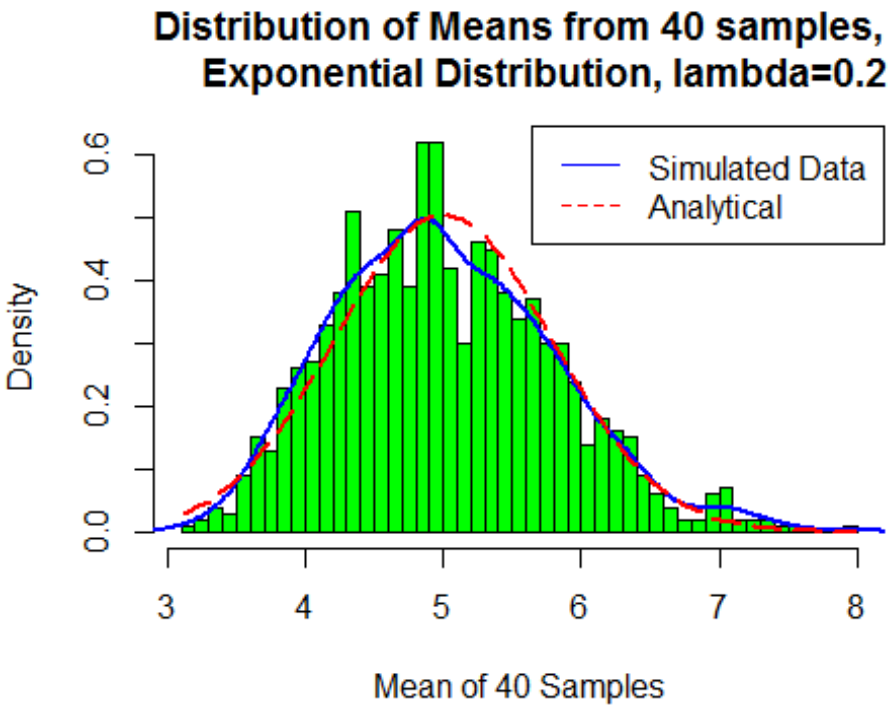
The following plot shows the kernel density distribution of averages of 20, 40, 60 and 1000 exponentials (10 000 simulations, lambda =0.2). We can observe that as the number of samples increases, the more concentrated its density fuction is around the theoretical mean (5.0). The mean of the 4 distributions are about the same, but the spread of the distribution is smaller as the sample size increases. This is behaviour is in agreement with the CLT.

```
par(mfrow=c(2,2))

simulation1 = NULL
for (i in 1 : 10000) simulation1 = c(simulation1, mean(rexp(20, lambda)))
hist(simulation1, breaks=50, prob=TRUE, main="Averages of 20 exponentials")

simulation2 = NULL
for (i in 1 : 10000) simulation2 = c(simulation2, mean(rexp(40, lambda)))
hist(simulation2,breaks=50, prob=TRUE, main="Averages of 40 exponentials")

simulation3 = NULL
for (i in 1 : 10000) simulation3 = c(simulation3, mean(rexp(60, lambda)))
hist(simulation3, breaks=50, prob=TRUE, main="Averages of 60 exponentials")

simulation4 = NULL
for (i in 1 : 10000) simulation4 = c(simulation4, mean(rexp(1000, lambda)))
hist(simulation4, breaks=50, prob=TRUE, main="Averages of 1000 exponentials")
```



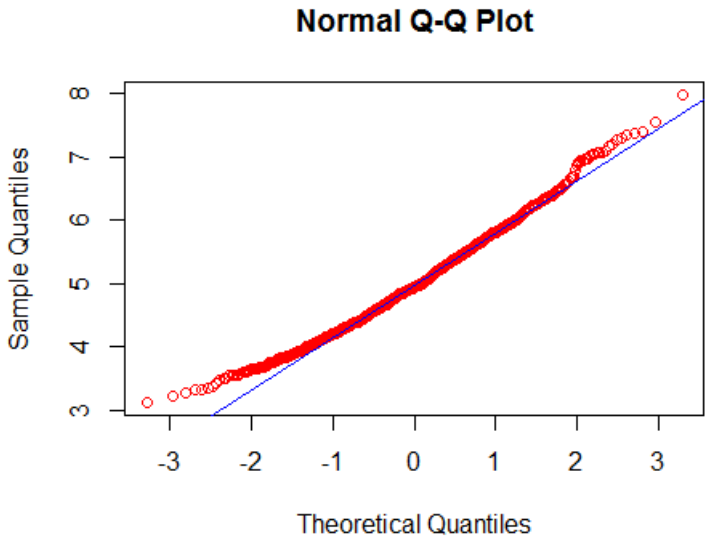
Task 3. A plot of density distribution is used to asses the relationship between the distribution of means. The density of the sample data is shown in green bars. The blue line corresponds to the simulated points and the red line corresponds to the normal distribution (curved constructed by taking the theoretical mean and standard deviation). The Central Limit Theorem statates that the distribution of averages of iid variables becomes that of a standard normal as the sample size increases. We have simulated 40 averages of the exponential distribution, the graph below demostrates the validity if the CLT since the curves do match each other well.

```
hist(simulation, breaks=50, prob=TRUE, col="green",
     main="Distribution of Means from 40 samples,
     Exponential Distribution, lambda=0.2",
     xlab="Mean of 40 Samples")
lines(density(simulation), lwd=2, col="blue")
xVariable <- seq(min(simulation), max(simulation), length=100)
yVariable <- dnorm(xVariable, mean=1/lambda, sd=(1/lambda/sqrt(sample_size)))
lines(xVariable, yVariable, pch=22, col="red", lty=5, lwd=2)
legend('topright', c("Simulated Data", "Analytical"), lty=c(1,2), col=c("blue", "red"))
```



Another way to show that the distribution is approximately normal is by looking at the normal Q-Q plot. The figure below suggests that the averages of samples follow normal distribution.

```
qqnorm(simulation, col=10); qqline(simulation, col=12)
```



CONCLUSIONS: It was shown that the average of the iid exponential distributions (40 samples) is approximately normal. Both, sample mean and sample variance compared well with theoretical mean and theoretical variance (CLT). Looking closely at the simulations with 20, 40, 60 and 1000 sample size, we can see that the sampling distribution of small size (i.e., 20 and 40) do have a slight skew. The larger the sample size (1000), the closer the sampling distribution of the mean is to a normal distribution.

Note: This report consist of 4 pages, this fulfill the project requirement page length .