Statistical Inference Project Part 1

Wednesday, April 22, 2015

Synopsis

Simulation to explore inference and some simple inferential data analysis on the following parts will be showed in this report:

1. A simulation exercise.
2. Basic inferential data analysis.

Simulation of exponential distribution in R will be used to compare with the Central Limit Theorem. The distribution of averages of 40 exponentials will be studied in this project: 1. The sample mean of the distribution of average of 40 exponentials will be compared with the theoretical mean of the distribution. 2. The variance of the sample as compared with the theoretical variance of the distribution. 3. Showing the distribution is approximately normal.

Result

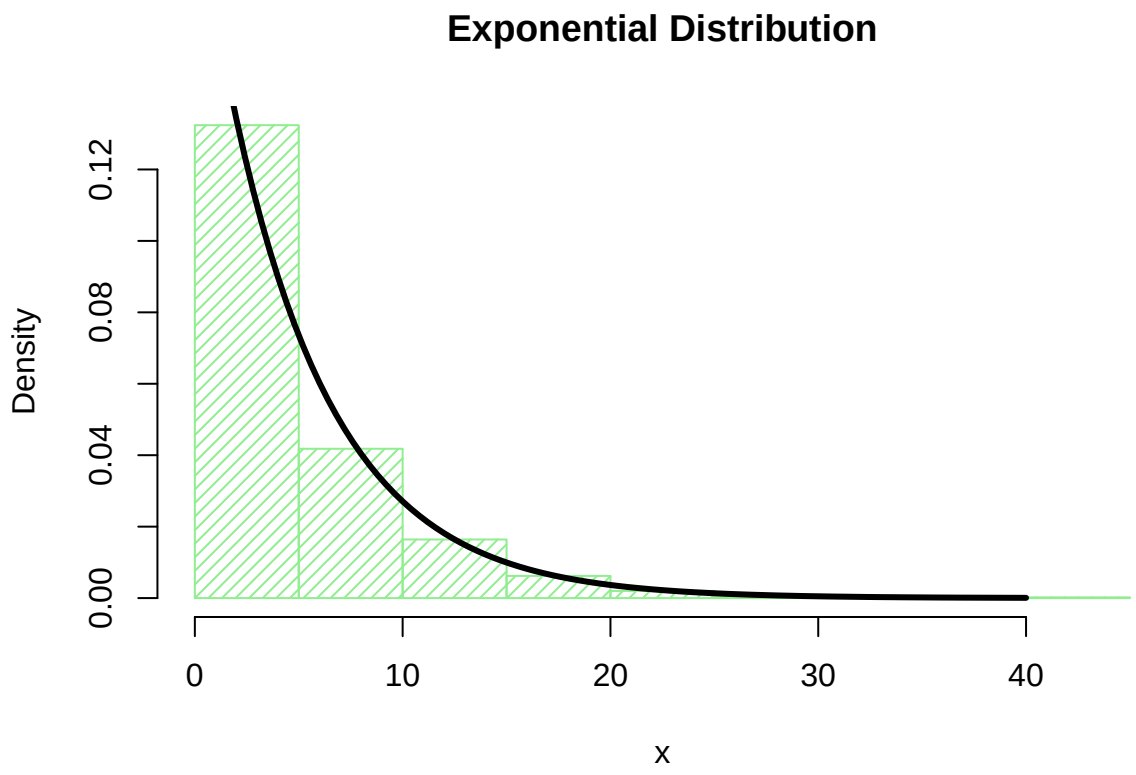
First, let create a thousand simulated averages of 40 exponentials, i.e. rexp(40,0.2)

```
# Load necessary libaries.
library(ggplot2)



# Set Lambda to 0.2.
lambda <- 0.2

# Set number of exponentials to 40.
num_of_exponentials <- 40

# Set number of Simulation to 1000.
num_of_simulation <- 1000

# set the seed create random for create reproducability.
set.seed(33)

# Start to simulate.
simulated_exponentials <- replicate(num_of_simulation, rexp(num_of_exponentials, lambda))

# Find the mean of exponentials.
simulation_means_exponentials <- apply(simulated_exponentials, 2, mean)

# Find the analytical sample mean.
analytical_sample_mean <- mean(simulation_means_exponentials)

# Theoretical mean is 1/lambda.
theoretical_mean <- 1/lambda
```

Question 1: Compare the sample mean with the theoretical mean of the distribution.

Answer :

```
# Show a Histogram of 1000 exponential distributions is asymmetrical around a central point as in the N
# Actual curve line in black colour shows the actual exponential distributions.
num   <- 1000
x     <- rexp(num, 0.2)
hist(x, probability = T, col = 'light green', density = 20, main ='Exponential Distribution')
curve(dexp(x, 0.2), xlim = c(0,40), col = 'black', lwd = 3, add = T)
```




```
# Show the Analytical Sample mean.
analytical_sample_mean
```

```
## [1] 4.964431
```

```
# Show the Theoretical mean.
theoretical_mean
```

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

As the result obtained for sample mean is close to that of theoretical mean, it can concluded that the sample mean is similar to the theoretical mean of the distribution.

```
# Compute Theoretical Standard Deviation, Actual Standard Deviation, Theortical Variance and Actual Var
theoretical_sd    <- ((1/lambda) * (1/sqrt(num_of_exponentials)) )
actual_sd         <- sd(simulation_means_exponentials)
theoretical_var   <- theoretical_sd^2
actual_var        <- var(simulation_means_exponentials)
```

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

Answer :

```
# Show the Theoretical Standard Deviation.
theoretical_sd
```

```
## [1] 0.7905694
```

```
# Show the Actual Standard Deviation.
actual_sd
```

```
## [1] 0.8035309
```

```
# Show the Theortical Variance.
theoretical_var
```

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

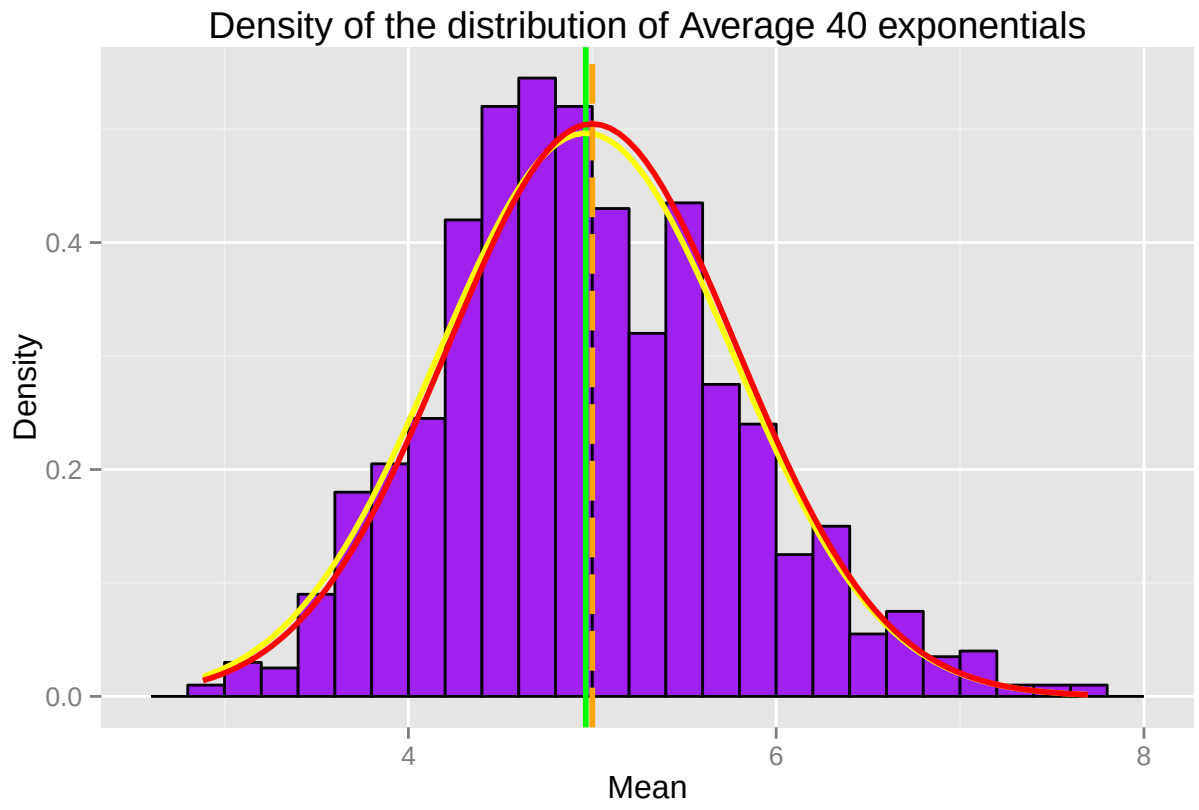
```
# Show the Actual Variance.
actual_var
```

```
## [1] 0.6456619
```

Question 3: Show that the distribution is approximately normal.

```
df_forty_simulation_means  <- data.frame(simulation_means_exponentials)

plot1  <- ggplot(df_forty_simulation_means, aes(x = simulation_means_exponentials))

# Draw lambda in purple colour histogram.
plot1  <- plot1 + geom_histogram (binwidth = lambda, fill = "purple", color = "black", aes(y = ..density

# Indicate the plot title, the title for x axis and y axis.
plot1  <- plot1 + labs (title = "Density of the distribution of Average 40 exponentials", x = "Mean", y

# Draw the continous green colour line to indicate the anaytical sample mean.
plot1  <- plot1 + geom_vline (xintercept = analytical_sample_mean, size = 1.0, color = "green" )

plot1  <- plot1 + stat_function( fun = dnorm, args = list( mean = analytical_sample_mean, sd = actual_s
```

```r
# Draw the longdash orange line to indicate the theoretical mean.
plot1  <- plot1 + geom_vline( xintercept = theoretical_mean, size = 1.0, color = "orange", linetype = "

plot1    <- plot1 + stat_function( fun = dnorm, args = list( mean = theoretical_mean, sd = theoretical_
print(plot1)
```



Answer : As seen in distribution of average 40 exponentials, actual data with yellow colour is close to the normal distribution. Others like actual mean in continous green line and theoretical mean in orange longdash lone are also displayed in this plot.