Statistical Inference Course Project Part1

Naqeeb Asif

13 February 2018

Synopsis

This experiment explores the central limit theorem. The simulated data used in this experiment is from exponential distribution. 1000 simulations for The distribution of averages of 40 exponentials is investigated. We found out that sample mean and sample variance are nearly equal to the theoretical mean and theoretical variance respectively. We also found out that the distribution used in this experiment is nearly normal.

Simulations

In this section the data to be used is generated. number of samples to be averaged are set to 40, lambda to 0.2 and number of simulations to 1000. A data frame named `sim_data` is created in which replicate function is used to generate the averages of `40 samples` of exponential distribution with `lambda = 0.2` 1000 times.

```
library(ggplot2)
n <- 40
lambda <- 0.2
n_sims <- 1000
sim_data <- data.frame(x=replicate(n_sims,mean(rexp(n,lambda))))
```

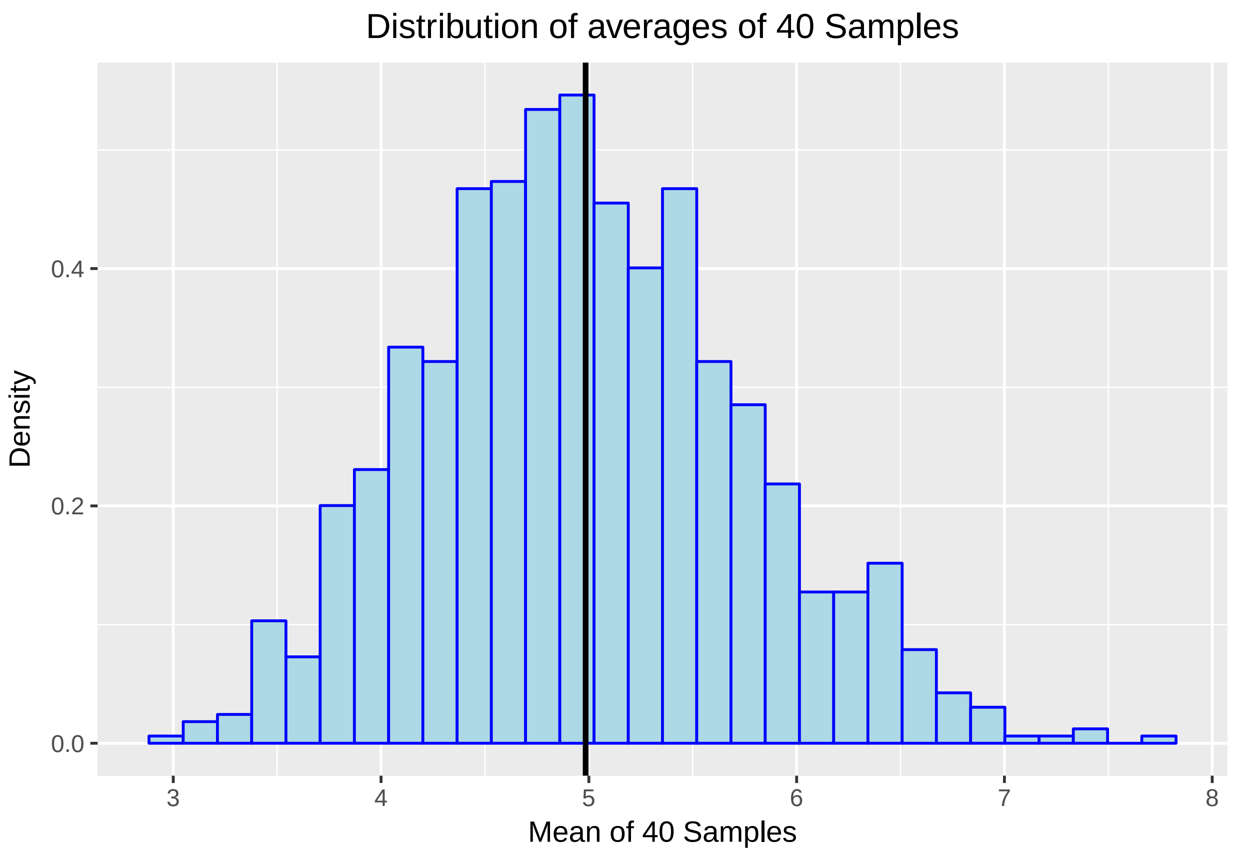
Results

This section contains the results.

Histogram of data

In this section the histogram of the distribution is plotted . In the plot y-axis contains density.

```
g <- ggplot(data=sim_data,aes(x))
g + geom_histogram(fill="lightblue",bins = 30,color="blue",aes(y=..density..))+
  geom_vline(xintercept = mean(sim_data$x),color="black",size=1) +
  labs(title = "Distribution of averages of 40 Samples", x = "Mean of 40 Samples", y = "Density")+
  theme(plot.title = element_text(hjust = 0.5))
```



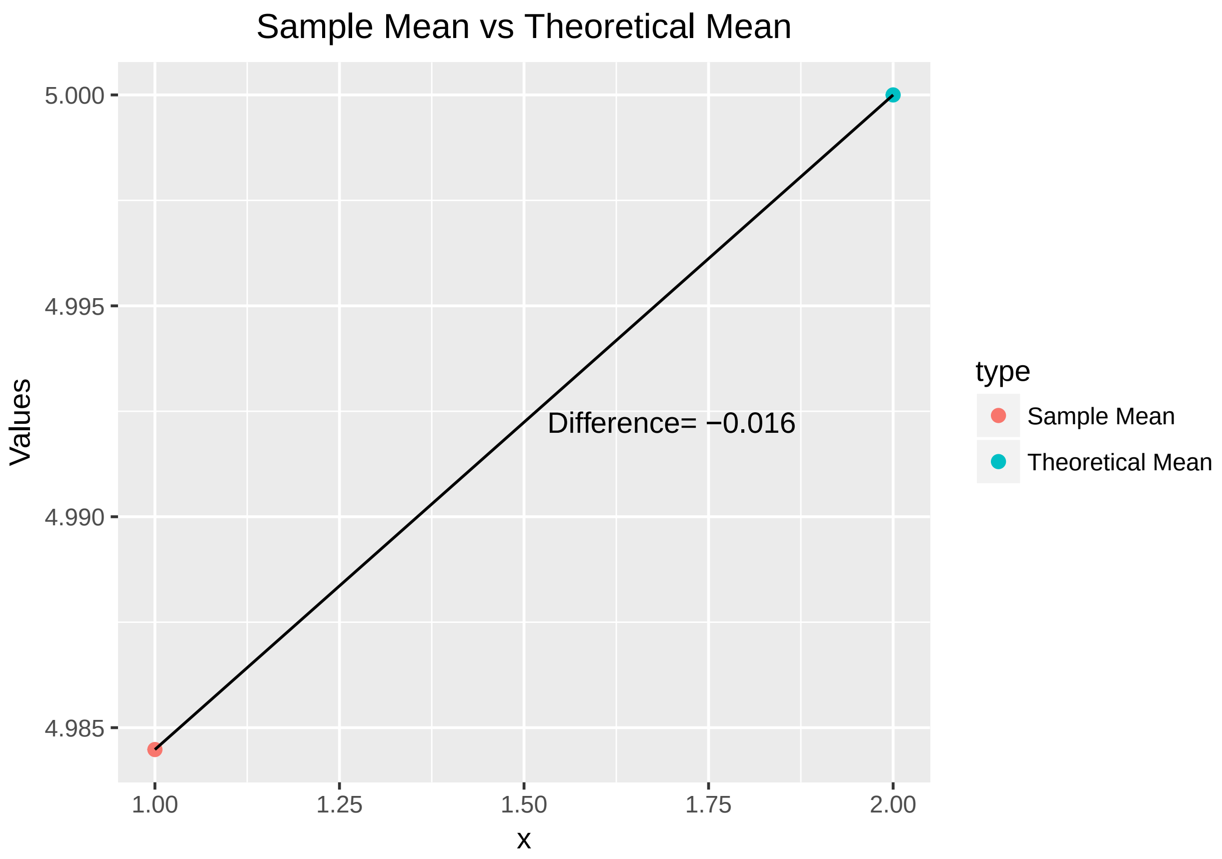
Sample Mean versus Theoreritcal mean

In this section sample mean is compared with the theoretical mean. Theoretical mean of exponential distribution can be calculated by taking reciprocal of lambda i.e. `1/lambda` is the theoretical mean.

```
sample_mean <- mean(sim_data$x)
theoretical_mean <- 1/lambda
```

Sample mean is `4.9844809` and theoretical mean is `5`. As it can be seen that sample mean is nearly equal to the theoretical mean. Lets see the difference between them in a plot.

```
ggplot(data=data.frame(x=c(1,2),Values=c(sample_mean,theoretical_mean),type=c("Sample Mean",
                                                                   "Theoretical Mean")),
       aes(x=x,y=Values))+ geom_point(size=2.0,aes(colour=type)) +geom_line(aes(y=Values,x=x))+
       annotate("text", x = 1.70, y = (sample_mean+theoretical_mean)/2, label = paste("Difference=",
                                                          round(sample_mean-theoretical_mean,3)))+
       labs(title="Sample Mean vs Theoretical Mean")+ theme(plot.title =
                                                              element_text(hjust = 0.5))
```



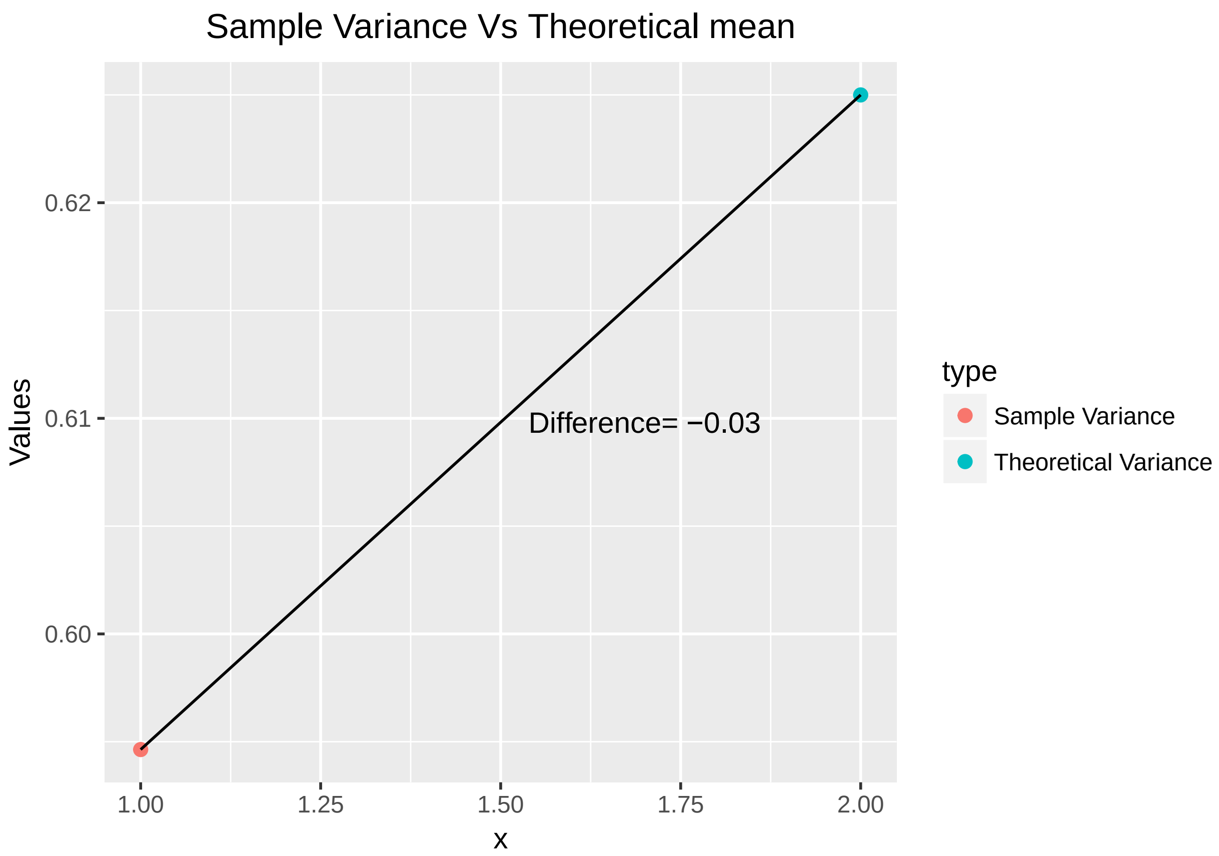
Sample Variance Vs Theoretical Variance

In this section sample variance is compared with the theoretical variance. Theoretical variance of exponential distribution can be calculated by taking the square of reciprocal of lambda and dividing it by n. i.e. `1/lambda^2 / n` is the theoretical mean.

```
sample_variance <- var(sim_data$x)
theoretical_variance <- (1/lambda)^2 /n
```

Sample variance is `0.5946373` and theoretical variance is `0.625`. As it can be seen that sample variance is nearly equal to the theoretical variance. Lets see the difference between them in a plot.

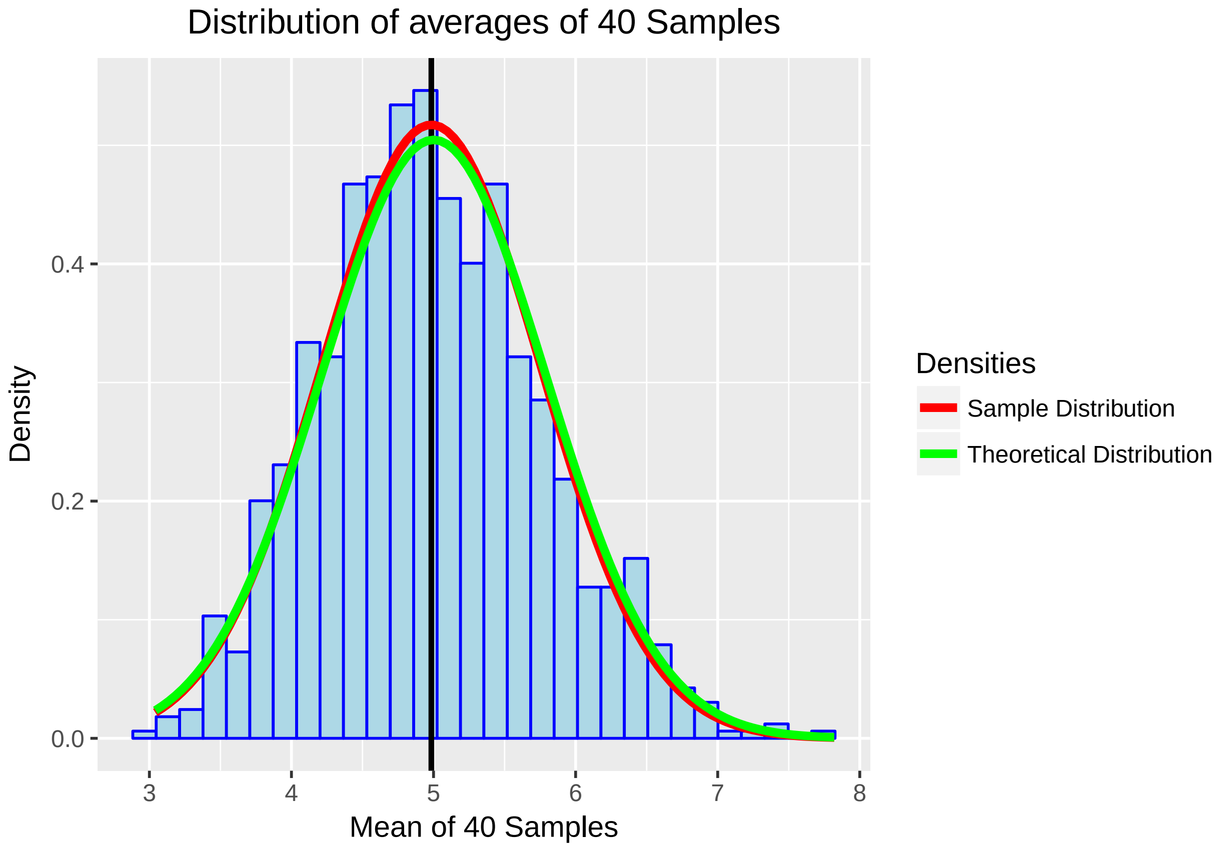
```
ggplot(data=data.frame(x=c(1,2),Values=c(sample_variance,theoretical_variance),type=c("Sample Variance"
                                                                   "Theoretical Variance")),
       aes(x=x,y=Values))+ geom_point(size=2.0,aes(colour=type)) +geom_line(aes(y=Values,x=x))+
  annotate("text", x = 1.70, y = (sample_variance+theoretical_variance)/2,
           label = paste("Difference=",round(sample_variance-theoretical_variance,3))) +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(title="Sample Variance Vs Theoretical mean")
```



Sample and Theoerical Normal Plot

This section contains histogram with two normal density plots . One density plot is with variance = sample variance and mean= sample mean and other density plot is with variance = theoretical variance and mean = theoretical mean.

```
g <- ggplot(data=sim_data,aes(x))
g + geom_histogram(fill="lightblue",bins = 30,color="blue",aes(y=..density..))+
  geom_vline(xintercept = mean(sim_data$x),color="black",size=1) +
  labs(title = "Distribution of averages of 40 Samples", x = "Mean of 40 Samples",
       y = "Density")+ theme(plot.title = element_text(hjust = 0.5)) +
  stat_function(fun = dnorm,args = list(mean = sample_mean, sd = sample_variance^0.5),
                aes(color="Sample Distribution"), size = 1.5) +
  stat_function(fun = dnorm,args = list(mean = theoretical_mean, sd = theoretical_variance^0.5),
                aes(color="Theoretical Distribution"), size = 1.5)+
  scale_colour_manual(values=c("Theoretical Distribution"="green",
                               "Sample Distribution"="red"), name="Densities")
```

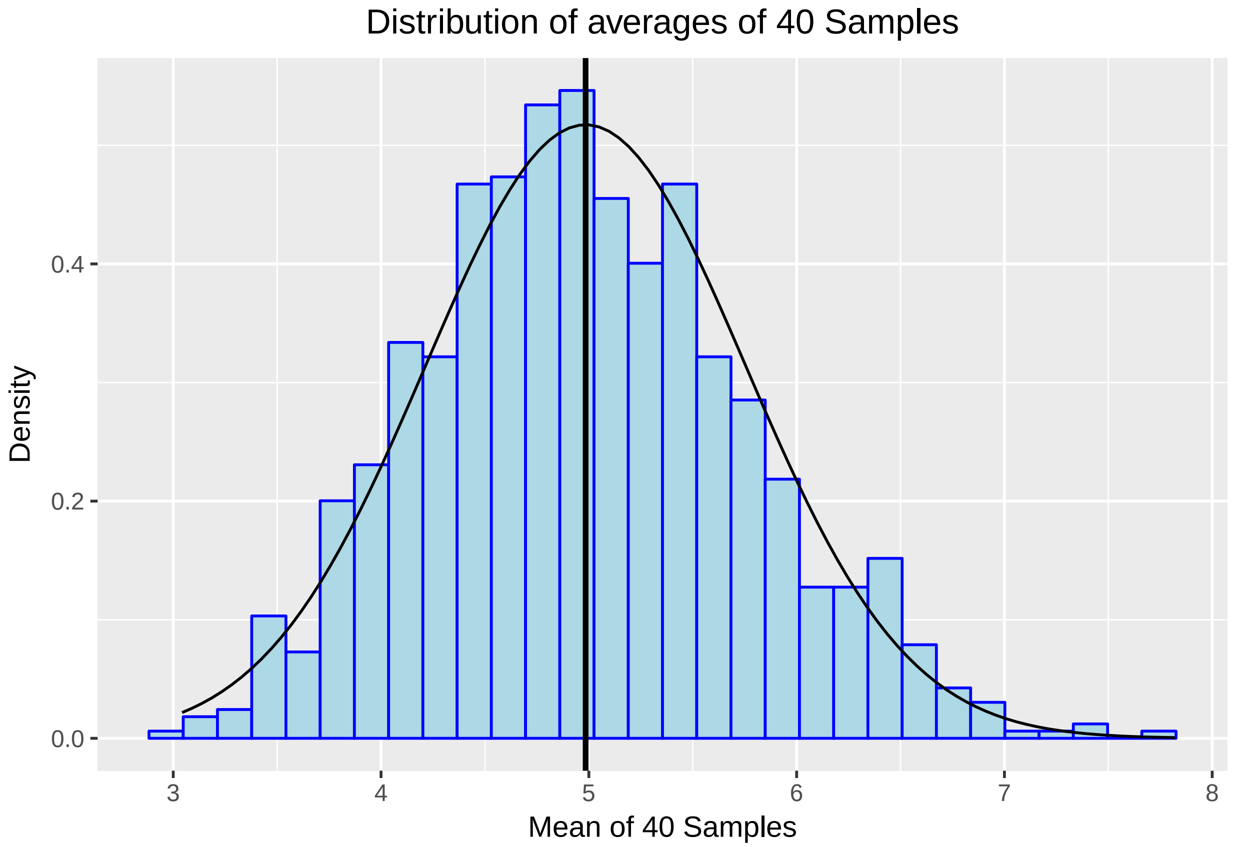


As it can be seen from the plot that two distributions are nearly the same.

Distribution

In this section , we will try to compare the distributiion of 40 sample averages to the normal distribution.

```
g <- ggplot(data=sim_data,aes(x))
g + geom_histogram(fill="lightblue",bins = 30,color="blue",aes(y=..density..))+
  geom_vline(xintercept = mean(sim_data$x),color="black",size=1) +
  labs(title = "Distribution of averages of 40 Samples", x = "Mean of 40 Samples",
      y = "Density")+ theme(plot.title = element_text(hjust = 0.5)) +
  stat_function(fun = dnorm,args = list(mean = sample_mean, sd = sample_variance^0.5))
```



As we can see that distribution of averages of 40 samples is very similar to that of normal distribution with mean =4.984 and standard deviation = 0.771

Conclusion

We can conclude from this experiment that distribution of averages of 40 samples follow normal distribution and mean and variances of the sample are very close to the theoretical values.