Exploration of Central Limit Theorem (CLT) using Exponential Distribution

Kumar Chandrakant

December 23, 2015

Overview

Through this analysis we want to validate the Central Limit Theorem using a population which is exponentially distributed. As par the CLT irrespective of the distribution of underlaying population the distribution of means of sample drawn from it is normal. This obviously given that the variable is independent random and sample size is large enough.

We will be running a 1000 simulations on sample size of 40 drawn from an exponential distribution with rate as 0.2.

Simulations

Let us start by setting some of the basic parameters for the analysis like number of simulations that we would be running, sample size choosen and the rate of exponential distribution.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

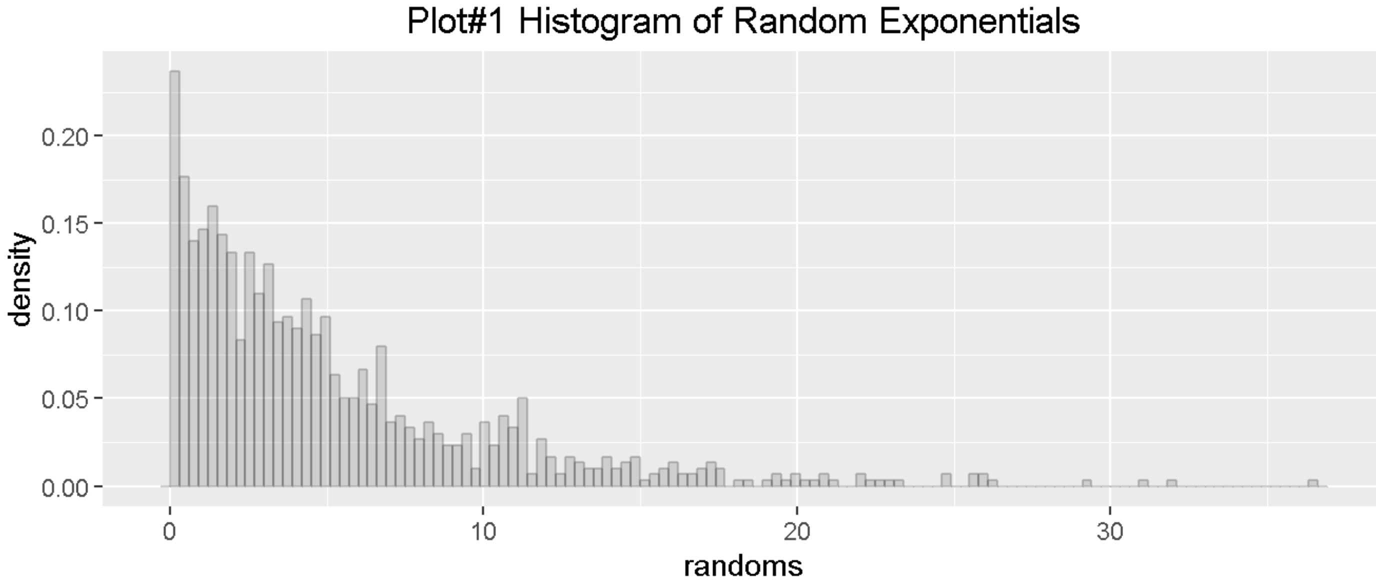
```
library(grid)
nosim <- 1000
sample_size <- 40
rate <- 0.2
```

The theoretical mean, standard deviation and variance for the exponential distribution used are presented below:

```
theo_mean <- 1/rate
theo_sd <- 1/rate
theo_var <- theo_sd^2
```

- Theoretical Mean: 5
- Theoretical Standard Deviation: 5
- Theoretical Variance: 25

Before we begin with our simulations lets see how does a sample picked from the expoential distribution look like.



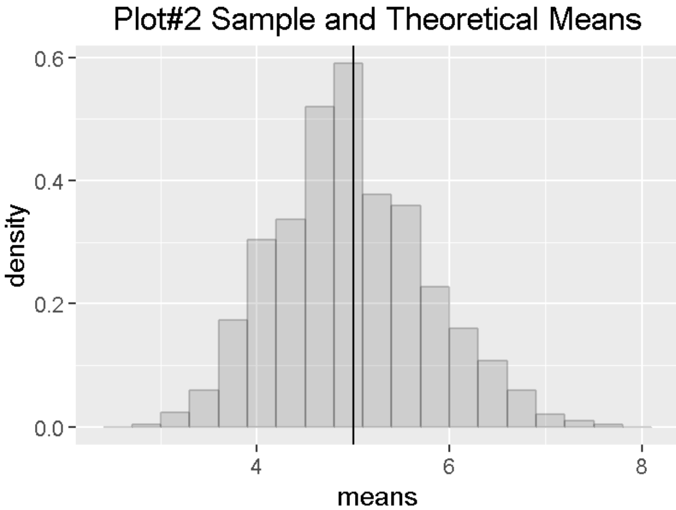
As expected this looks like an exponetial distribution. Ofcourse!

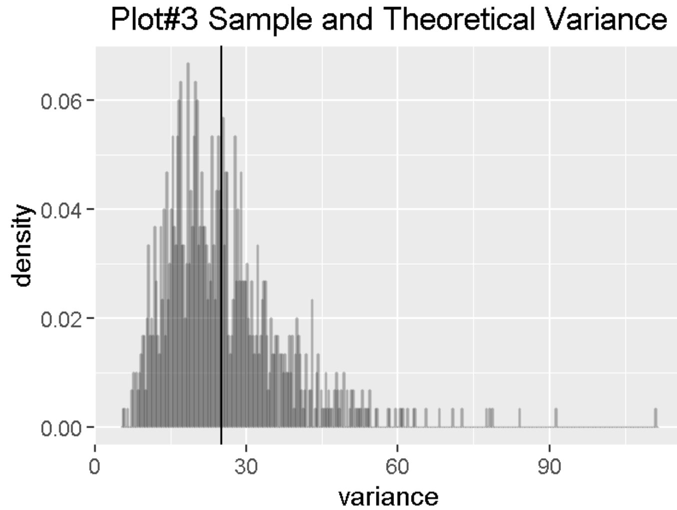
Let us start with the simulations now. We pick a sample of 40 exponetials and calculate their mean and standard deviation. This will be repeated a 1000 times.

```
df <- data.frame(means=double(),
                 means_norm=double(),
                 std_dev=double(),
                 variance=double())
for(i in 1:nosim){
  x <- rexp(sample_size, rate = rate)
  df[i,1] <- mean(x)
  df[i,2] <- ((mean(x)-theo_mean)/(theo_sd/sqrt(sample_size)))
  df[i,3] <- sd(x)
  df[i,4] <- df[i,3]^2
}
```

Sample Mean & Variance versus Theoretical Mean & Variance

Let us see how does the sample means and variance compare against the theoretical mean and standard deviation of the population.

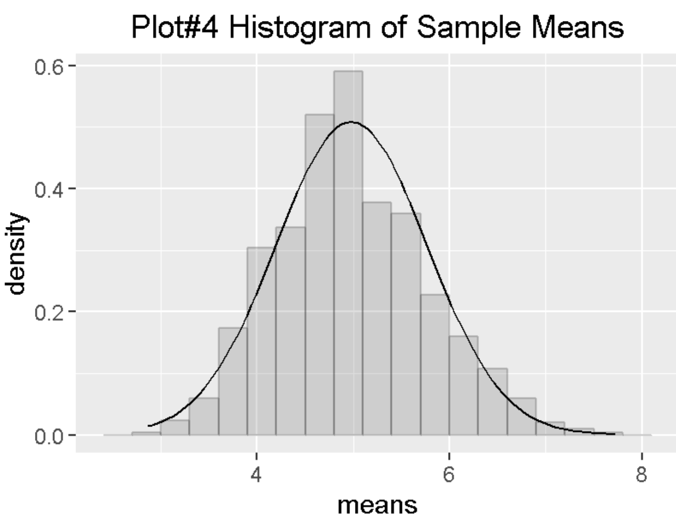


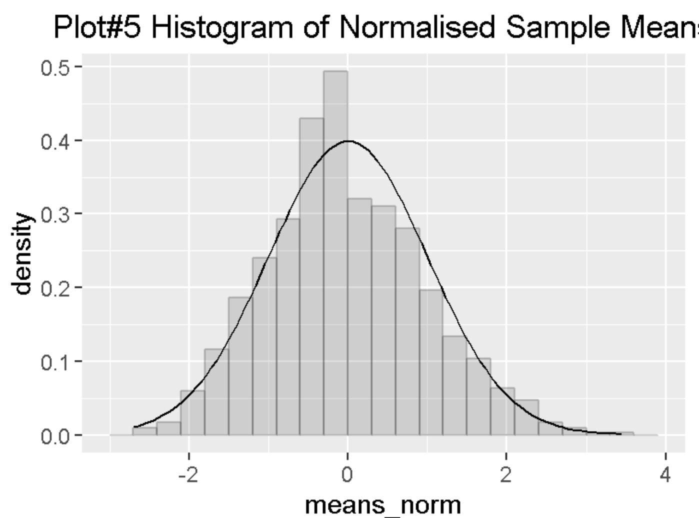


As expected the distribution of sample means and standard deviations is centered around its theroretical poulation mean and variance respectively shown by veritcal lines.

Distribution

Let us now see how does the means and normalised means of samples behave. We expect this to be normal distribution for means and standard normal distribution for normalised means.





As evident from the plot of sample means above the distribution is very similar to a normal distribution. Also the plot of normalised sample means is similar to a standard normal distribution. This holds true the much famous Central Limit Theorem!

Annexure

R code to generate Plot#1

```
randoms <- rexp(nosim, rate = rate)
df <- data.frame(randoms = randoms)
g <- ggplot(df, aes(x = randoms))
g <- g + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))
g <- g + ggtitle("Plot#1 Histogram of Random Exponentials")
g
```

R code to generate Plot#2 & Plot#3

```
g1 <- ggplot(df, aes(x = means))
g1 <- g1 + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))
g1 <- g1 + geom_vline(xintercept=theo_mean)
g1 <- g1 + ggtitle("Plot#2 Sample and Theoretical Means")
g2 <- ggplot(df, aes(x = sds))
g2 <- g2 + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))
g2 <- g2 + geom_vline(xintercept=theo_sd)
g2 <- g2 + ggtitle("Plot#3 Sample and Theoretical Std. Deviations")
pushViewport(viewport(layout = grid.layout(1, 2)))
print(g1, vp = viewport(layout.pos.row = 1, layout.pos.col = 1))
print(g2, vp = viewport(layout.pos.row = 1, layout.pos.col = 2))
```

R code to generate Plot#4 & Plot#5

```
g1 <- ggplot(df, aes(x = means))
g1 <- g1 + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))
g1 <- g1 + stat_function(fun = dnorm, args=list(mean=mean(df$means), sd=sd(df$means)))
g1 <- g1 + ggtitle("Plot#4 Histogram of Sample Means")
g2 <- ggplot(df, aes(x = means_norm))
g2 <- g2 + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))
g2 <- g2 + stat_function(fun = dnorm)
g2 <- g2 + ggtitle("Plot#5 Histogram of Normalised Sample Means")
pushViewport(viewport(layout = grid.layout(1, 2)))
print(g1, vp = viewport(layout.pos.row = 1, layout.pos.col = 1))
print(g2, vp = viewport(layout.pos.row = 1, layout.pos.col = 2))
```