The relationship between Central Limit theorem and Exponential Simulations

Rebecca Elizabeth Kitching

May 2018

In this project, we want to investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT). Based on CLT, we expect that the simulated mean to approach the true mean of the distribution, noted as 1/lambda. We also want to examine the relaionship between the variance of the simulated data and the theoretical variance where the standard deviation is again noted as 1/lambda.

Defining Parameters

Here we define the known variables we want to use for the stimulations. We want to have 1000 simulations which will generate 40 numbers from an exponential distribution each time. Lambda will be set to 0.2.

```
# Add packages
library("ggplot2");library("grid")
# Force all out put to 5 s.f
options(digits=5)
# Set the seed to allow the simulation to be reproduciable
set.seed(263)
# Set distractibution parameters to sample from
lambda <- 0.2; stdv <- 1/lambda; mean <- 1/lambda
# Each simulations will produce 40 iems in the distribution
n <- 40
# Number of simulations
simulations <- 1000
```

Conducting the Simulations

We now want to run these 1000 simulations which we accomplish using the code below. For each simulation, the mean of the sample is calculated and stores it in a data frame.

```
# Create empty matrix to store simulations
data <- data.frame(Mean = double(),StDe = double())
names(data) <- c("Mean","StDv")
# Loops though 1000 simulations
for (sim in 1:simulations){
  # Sample 40 items from the distribution
  thissimulation <-  rexp(n, lambda)
  # Calulate the mean of the sample distribution
  data[sim,1] <- mean(thissimulation)
  # Calulate the standard deviation of the sample distribution
  data[sim,2] <- sd(thissimulation)
}
```

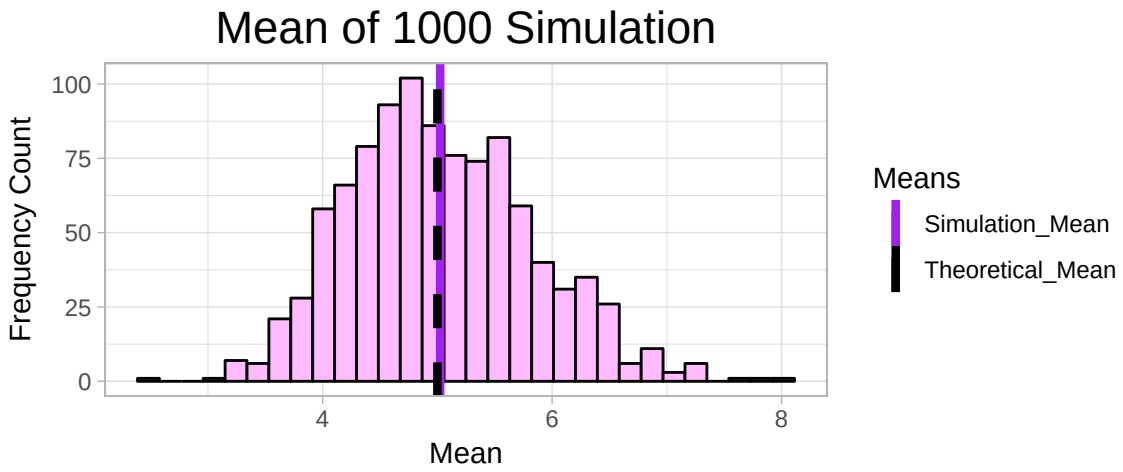
Sample Mean versus Theoretical Mean

After completeing 1000 simulations, we now have the corresponding 1000 mean values from the distribution. We now calculate the average mean for all these distributions and plot them with a histogram.

```
# Calculate average of the mean for the 1000 simulations
colMeans(data)[1]
```

```
##   Mean
## 5.0241
```

```
# Plot data using Histogram
ggplot(data = data, aes(x=Mean)) +
  geom_histogram(color="black", fill="plum1", bins=30)+
  ylab('Frequency Count')+
  theme_light()+
  geom_vline(aes(xintercept=colMeans(data)[1], colour='Simulation_Mean'),lwd=1.5,)+
  labs(title='Mean of 1000 Simulation')+
  theme(plot.title = element_text(hjust = 0.5,
                                  vjust=3.5,
                                  size = 17))+
  geom_vline(aes(xintercept=5,colour='Theoretical_Mean'), linetype='dashed',lwd=1.5)+
  scale_color_manual(name = "Means",
                     values = c(Simulation_Mean = "purple",
                                Theoretical_Mean = "black"))
```



From the above output, we can see from the 1000 simulations, the mean for this particular distribution is 5.02408. Remember that at the very start, we preicted that, using CLT, the mean would approach our theoretical mean of 1/lambda or 5. Since both the simulation mean and the theoretical mean are very similar, we can say that as we do 1000 simulations, the mean of all those simulations gets closer to the theoretical mean of the distribution sampled from, as predicted by CLT.

Sample Variance versus Theoretical Variance

After completeing the 1000 simulations, we also examine the variance of their means and compare them to the theoretical variance of the original sampled distribution. Given the standard deviation of the popultion is 1/lambda or 5, the theoretical standard error of the distribution (giving the variability of sampled averages) is found using (1/lambda)/sqrt(40) giving 0.79057. Now looking back at our means from the 1000 simulations, we can work out the actual variance of the simulations.

```r
# Calculate variance of the simuated data set
variance <- var(data[,1]);variance
```

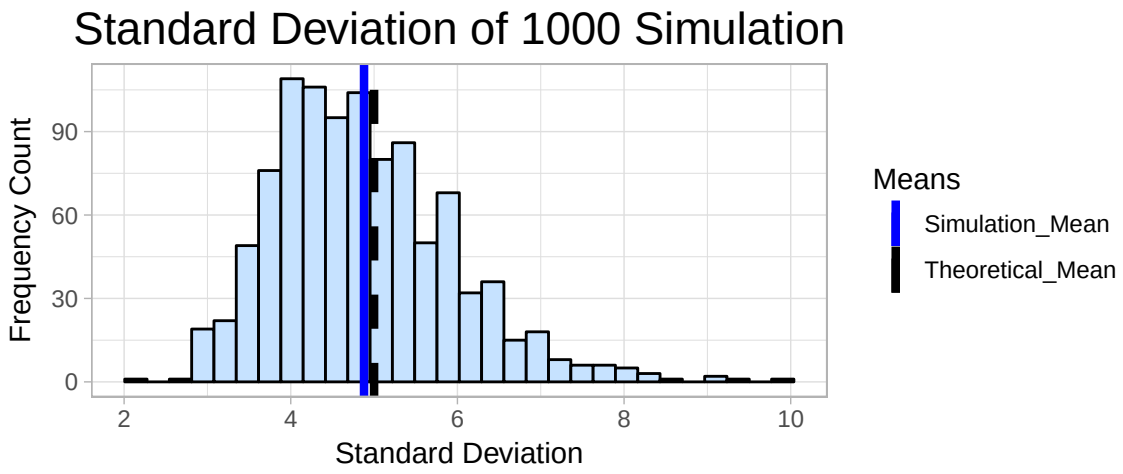
```
## [1] 0.65079
```

We can therefore see that the variance of the simulations is 0.65079 which is close to the theoretical variance predicted from the samping population of 0.79057.To further investigate this, we can also compare the standard deviation of each individual simulation to the standard deviation expected from the population distribution which is 1/lambda or 5.

```r
# Calculate mean of the standard deviations for the 1000 simulations
colMeans(data)[2]
```

```
##   StDv
## 4.8793
```

```r
# Plot data using Histogram
ggplot(data = data, aes(x=StDv)) +
  geom_histogram(color="black", fill="slategray1", bins=30)+
  ylab('Frequency Count')+xlab("Standard Deviation")+
  theme_light()+
  geom_vline(aes(xintercept=colMeans(data)[2],
                 color='Simulation_Mean'), lwd=1.5)+
  labs(title='Standard Deviation of 1000 Simulation')+
  theme(plot.title = element_text(hjust = 0.5,
                                  vjust=3.5,
                                  size = 17))+
  geom_vline(aes(xintercept=5, color='Theoretical_Mean'),
             lwd=1.5, linetype='dashed')+
  scale_color_manual(name = "Means",
                     values = c(Simulation_Mean = "blue",
                                Theoretical_Mean = "black"))
```



From the above output, we can see that from the 1000 simulations, the mean standard deviation for this particular distribution is 4.87935. Remember that, we again predicted, using CLT, the mean standard deviation would approach our theoretical standard deviation of 1/lambda or 5. Similar to the above mean example, we can say that in our 1000 simuations, the average standard deviation has approached the theoretical prediction proposed by CLT.

Normal Distribution

Here we can also examine whether our the distribution of our simulations means fits within a normal distribution which, accoridng to CLT, it should do. To do this, we compare the distribution of the 1000 data points relative to each quantile to the theoretical quantiles predicted by a normal distribution. From the Q-Q plot below (left), we can see a strong positive linear relationship between the theoretical quantiles and sample quantiles. This suggests the simulation data are distributed within a normal distribution with the majority of the data points in the middle of the line and fewer towards the tails. This is particually apparant when comparing to a uniform distribution plot (right) where there are increased frequencies of values at the top and bottom of the distribution.

```
# Plot exponential data simulations on a QQ plot with line of best fit
qqnorm(data$Mean,main="Normal Distribution Q-Q Plot");qqline(data$Mean,col = "plum",lwd=3)
qqnorm((runif(1000)),main="Uniform Distribution Q-Q Plot");qqline((runif(1000)),col="blue",lwd=3)
```