Investigation into the Exponential Distribution and the Central Limit Theorem

Overview

This report is a brief study intended to investigate how the exponential distribution adheres to the **Central Limit Theorem(CLT)**. The CLT states that the distribution of the sum (or mean) of a large number of independent, identically distributed variables will be approximately normal irrespective of the underlying distribution for a sufficiently large sample size.

Simulations

We now proceed to perform simulations to investigate how the sample mean and variance approximate the population variance and mean. We take mutltiple samples of the mean of 40 exponentials. From the central limit theorem, we would expect:

$$E[\overline{X}] = \mu_{Pop}$$

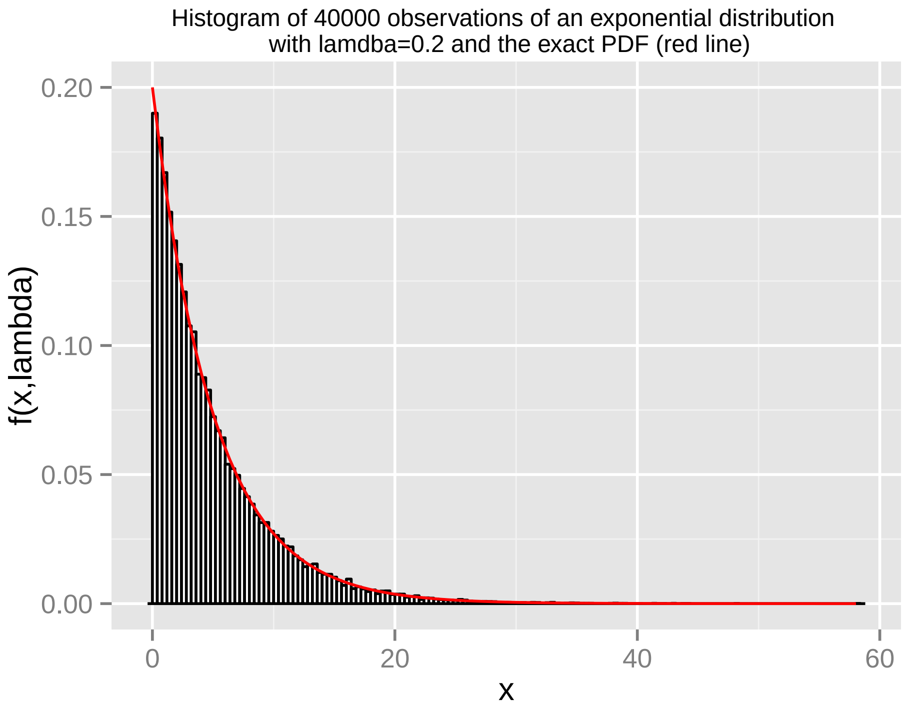
$$Var[\overline{X}] = \sigma^2_{Pop}/n$$

For the Exponential distribution

$$E[X] = 1/\lambda$$
$$Var[X] = 1/\lambda^2$$

In this analysis we calculate the exponentials using the R function $rexp(n, \lambda)$ where we fix $\lambda = 0.2$. Firstly we generate the data set.

```
numberSims <-  1000
numberexps <- 1000
numberMeans <- 40
numberMeans2 <- 200
lambda <- 0.2
mns <- NULL
exps <- NULL
for (i in 1 : numberSims) {mns <- c(mns, mean(rexp(numberMeans,lambda)))}
for (i in 1 : numberexps) {exps <- c(exps, rexp(numberMeans,lambda))}
```



Sample Mean versus Theoretical Mean

Initially we consider comparison between the mean calculated from the data set and what the mean should be according to the CLT. So according to the CLT, combining the equations defined on page 1.

```
theoretical_mean <- 1/lambda
theoretical_mean
```

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

```
sample_mean <- mean(mns)
sample_mean
```

```
## [1] 5.026616
```

```
meandeltapct <- (theoretical_mean - sample_mean) * 100 / theoretical_mean
meandeltapct
```

```
## [1] -0.5323277
```

The mean calculated in the simulation is very close to the theoretical mean.

Sample Variance versus Theoretical Variance

Again combining the equations defined on page 1.

```
theoretical_variance <- 1/(lambda^2*numberMeans)
theoretical_variance
```

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

```
sample_variance <- var(mns)
sample_variance
```

```
## [1] 0.6289329
```

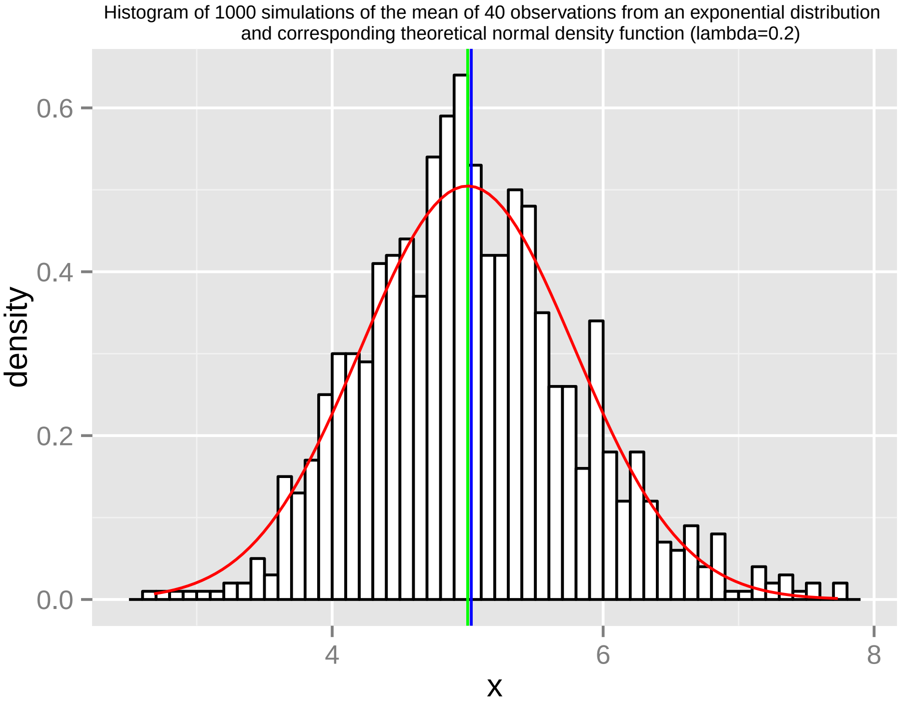
```
variancedeltapct <- (theoretical_variance - sample_variance) * 100 / theoretical_mean
variancedeltapct
```

```
## [1] -0.07865865
```

Again we see that the theoretical variance is very close to that obtained from the simulation.

Comparison of sample data with the corresponding Normal Distribution

We now proceed to investigate how close to the theoretically predicted Normal distribution the simulated data is.



This shows that the distribution of the means certainly appears to have a Gaussian form and as we have seen above, the mean and variance obtained from the simulated data are very close to that predicted by the CLT (green(theoretical) and blue(sample) vertical lines). Now we can see that the distribution is still a little bit skewed, this comes from the combination of the exponential distribution being skewed and the relatively small number of samples used in the simulation. If we up the number of samples then we will see an improved fit. Appendix B shows the results of repeating the experiment with 200 observations used for each simulation. We see that the simulated data is much closer to the theoretical normal distribution and the difference between the theoretical and sample means and variances is also reduced.

Conclusions

- The theoretical mean and distribution for sample means derived from the Central Limit Theorem agrees with simulated sample mean values for the Exponential distribution.
- Increasing the number of observations used to calculate the mean improves the approximation.

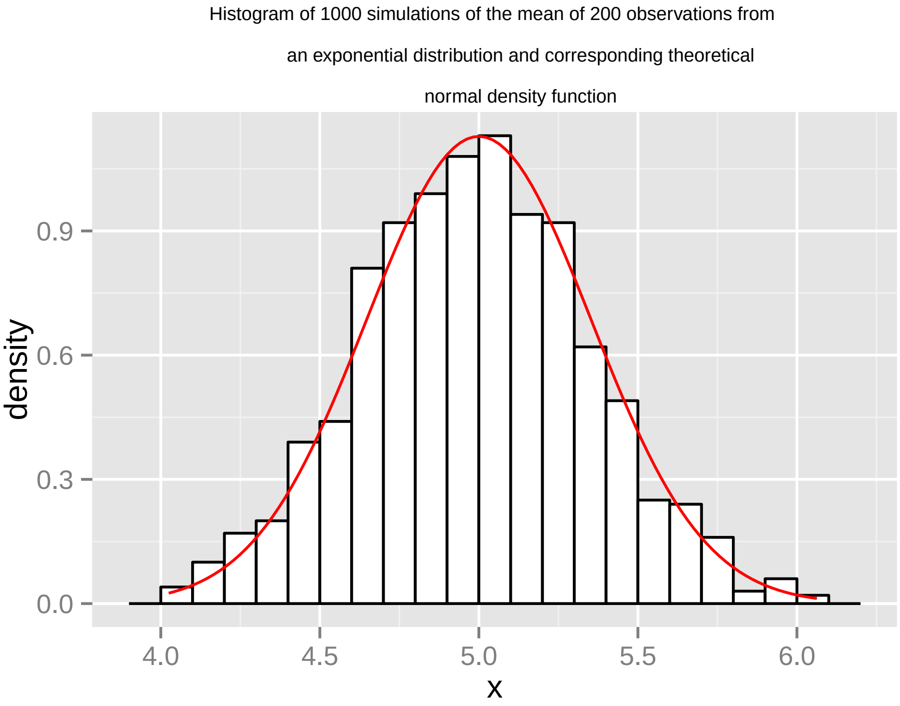
Appendix A - ggplot code for Graphs

```
expdist <- function(x, lbd = 0) {lbd * exp(-1 * lbd *x)}
expsdf <- as.data.frame(exps)
ggplot(data=expsdf, aes(x=exps)) + geom_histogram(aes(y = ..density..),
        colour="black", fill="white", binwidth = 0.4) +
   stat_function(fun=expdist, arg=list(lbd=lambda ), colour="red" ) +
  ggtitle("Histogram of 40000 observations of an exponential distribution \n
          with lamdba=0.2 and the exact PDF (red line)") + xlab("x") + ylab ("f(x,lambda)") + theme(plot.t
```

```
mnsdf <- as.data.frame(mns)
ggplot(data=mnsdf, aes(x=mns)) + geom_histogram(aes(y = ..density..),
          colour="black", fill="white", binwidth = 0.1) +
  geom_vline(xintercept = (1/lambda), colour="green") +
  geom_vline(xintercept = mean(mns), colour="blue") +
     stat_function(fun=dnorm, arg=list(mean=1/lambda,
    sd=sqrt(1/(lambda^2*numberMeans)) ), colour="red" ) +
  ggtitle("Histogram of 1000 simulations of the mean of 40 observations from an exponential distribution
          and corresponding theoretical normal density function (lambda=0.2)") +
  xlab("x") + theme(plot.title = element_text(size = rel(0.6)))
```

Appendix B - Investigation of Increasing number of observations used in calculating the mean.

I'm not necessarily expecting any credit for this appendix but I did the analysis, couldnt fit it in and couldnt bear to throw it away!!!! Hopefully at least you wont mark me down for it :-)



This is also reflected in the means and variance

```
theoretical_mean <- 1/lambda
theoretical_mean
```

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

```
sample_mean <- mean(mns2)
sample_mean
```

```
## [1] 4.987728
```

```
theoretical_variance <- 1/(lambda^2*numberMeans2)
theoretical_variance
```

```
## [1] 0.125
```

```
sample_variance <- var(mns2)
sample_variance
```

```
## [1] 0.1285631
```

```
meandeltapct <- (theoretical_mean - sample_mean) * 100 / theoretical_mean
meandeltapct
```

```
## [1] 0.2454435
```

```
variancedeltapct <- (theoretical_variance - sample_variance) * 100 / theoretical_mean
variancedeltapct
```

```
## [1] -0.07126159
```