Statistical Inference - Project Part 1

N. Lakhani

17 January 2018

```
knitr::opts_chunk$set(echo=TRUE,cache=TRUE,warning=FALSE)
set.seed(1040)
```

```
library(ggplot2)
```

Synopsis

This report is part of the Coursera project on Statistical Inference. It provides:

- results of investigating a simulated exponential distribution and its comparison with CLT. The property of an exponential distribution is explored using simulated and theorical results for mean, variance and confidence interval

The simulation exercise shows a close approximation of the simulated mean and variance with CLT using 40 sample data points in each simulation and repeated for 1000 such simulations.

Part 1 - Simulation

1. Sample mean and comparison to theoretical mean of distribution

The code below does the following:

- Simulates rexp(n,lambda) distribution with lambda=.2,sampe size n=40 and 1000 such simulations.
- Compute the mean of the 1000 data sets (each with 40 simulations)
- Compare the distribution of the sample mean for normality and theoretical mean given by 1/Lambda
- A histogram and normal qq plot is used to investigate the simulated vs theoretical normal distribution.

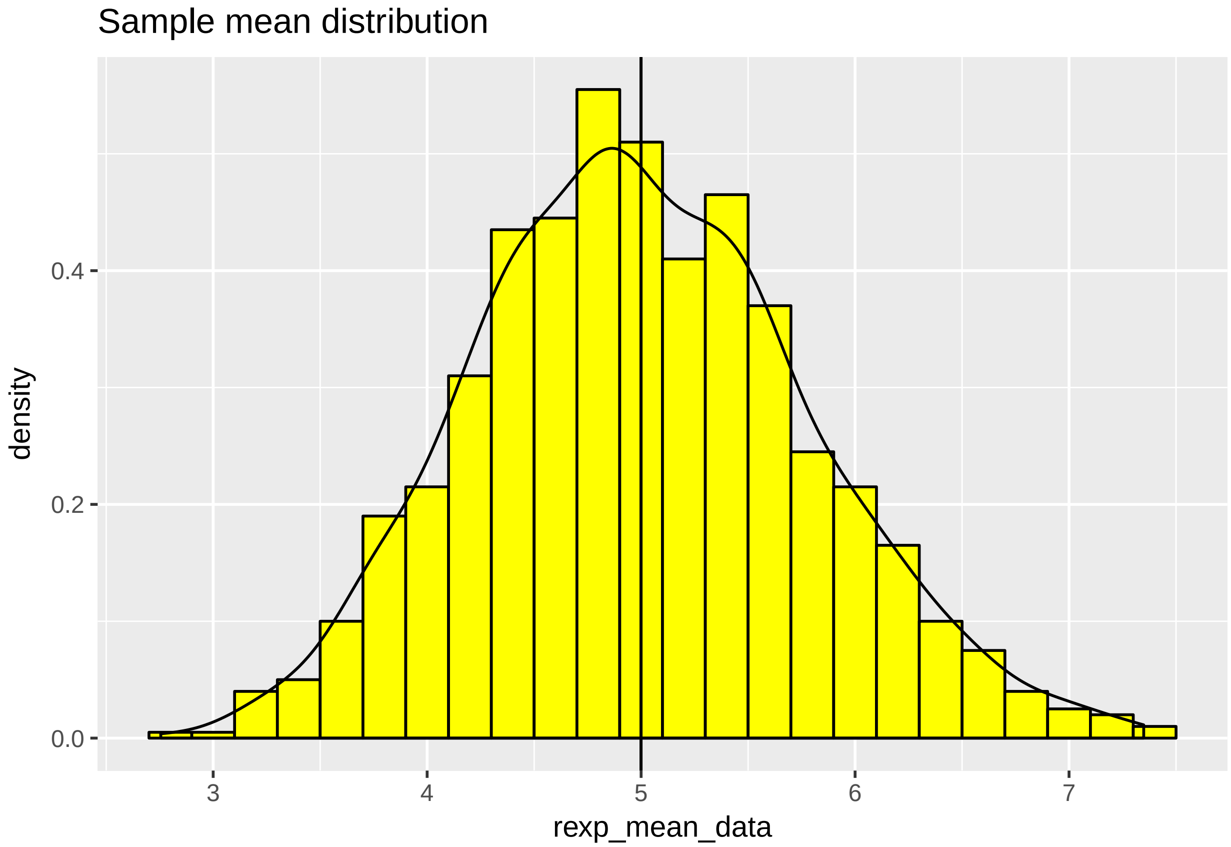
The histogram plot of the mean shows a normal distribution centered around 1/Lambda (or 5), which is also the theoretical mean. The QQ plot shows the sample quantiles vs normal quantiles, demonstratng a good approximation to a normal distribution except at the 2 extremes.

```
par(mfrow=c(1,2))
n1 <- 40
n2 <- 1000
lambda <- 0.2


rexp_mean_data <- apply(matrix(rexp(n1*n2,lambda),n2),1,mean)
rexp_mean <- mean(rexp_mean_data)

paste('Sample mean is: ',round(rexp_mean,3),' Theoretical mean is:',1/lambda)
```

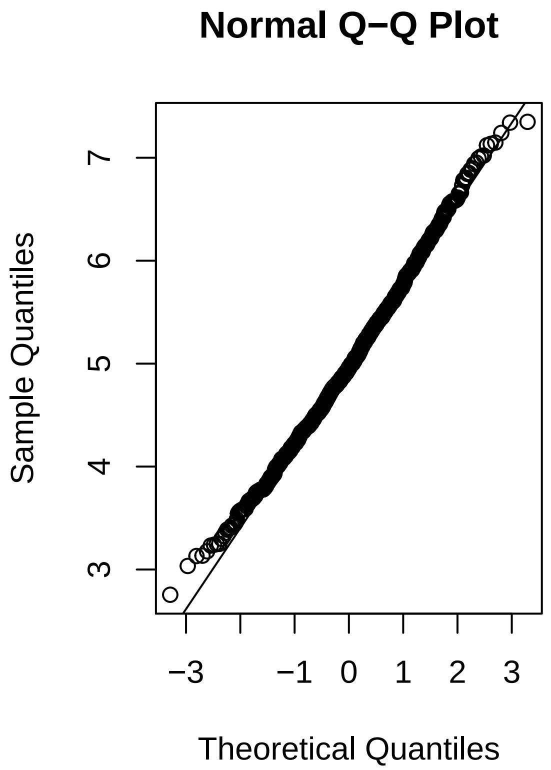
```
## [1] "Sample mean is:  4.999  Theoretical mean is: 5"
```

```
g1 <- ggplot(data.frame(rexp_mean_data),aes(rexp_mean_data)) +
      geom_histogram(aes(rexp_mean_data,..density..),binwidth=.2,fill='yellow',color='black') +
      labs(title='Sample mean distribution') + geom_density()
g1 <- g1 + geom_vline(aes(xintercept=rexp_mean))
```

```
g1 <- g1 + geom_vline(aes(xintercept=(1/lambda)))
g1
```



```
## QQ-plot to show normality of distribution

qqnorm(rexp_mean_data)
qqline(rexp_mean_data)
```



As can be seen from the figures above, the theoretical and the sample mean are very close and overlap.

2. Sample variance and comparison to theoretical variance for an rexp distribution

The code below does the following:

- Computes the variance of the simulated dataset
- Compare variance of simulated mean with the theoretical var given by (1/Lambda^2)/n1

The plot shows the normal distribution centered approximately around (1/Lambda^2)/n1 (or 0.625), which is the theoretical variance.

```
rexp_var <- var(rexp_mean_data)
paste('Sample Variance is: ',round(rexp_var,3),' Theoretical variance is:',(1/lambda^2)/n1)
```

```
## [1] "Sample Variance is:  0.606  Theoretical variance is: 0.625"
```

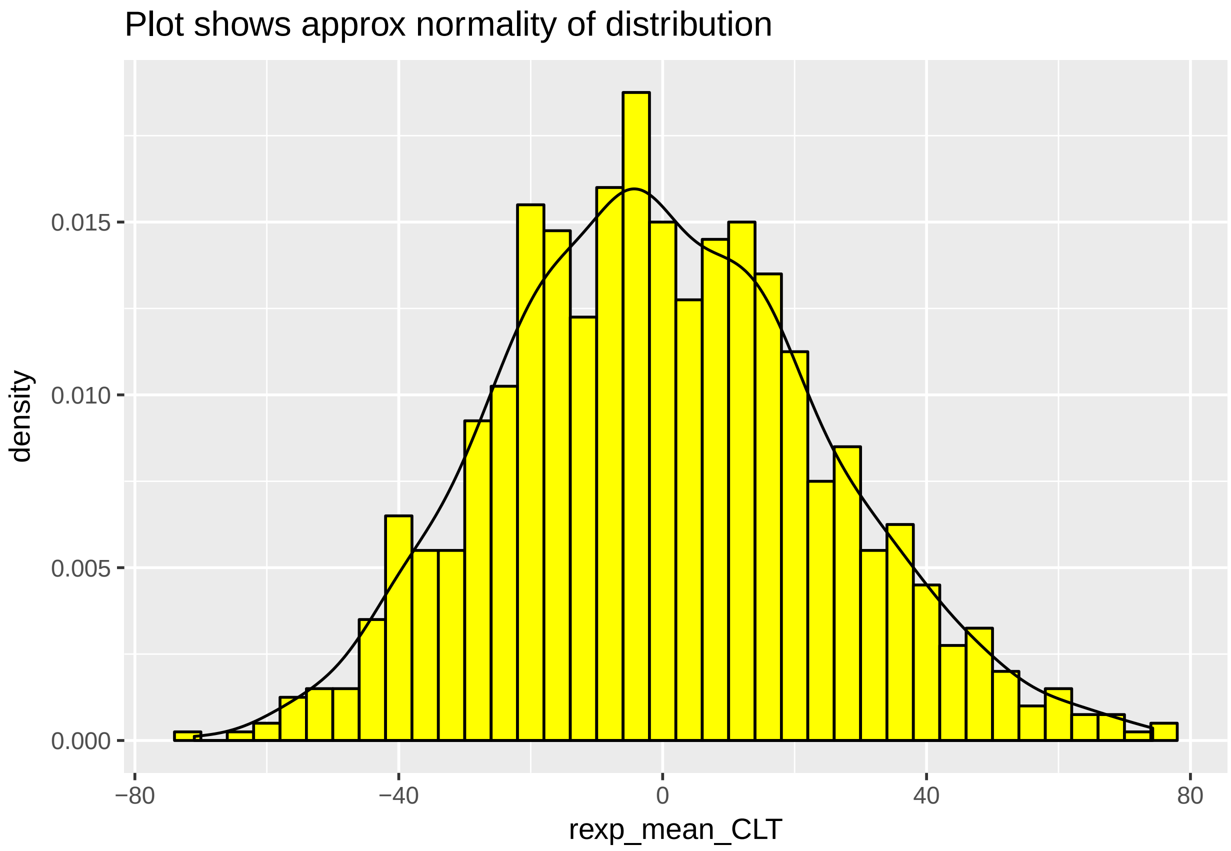
3. Sample mean and comparison to theoretical mean of distribution

The code below does the following:

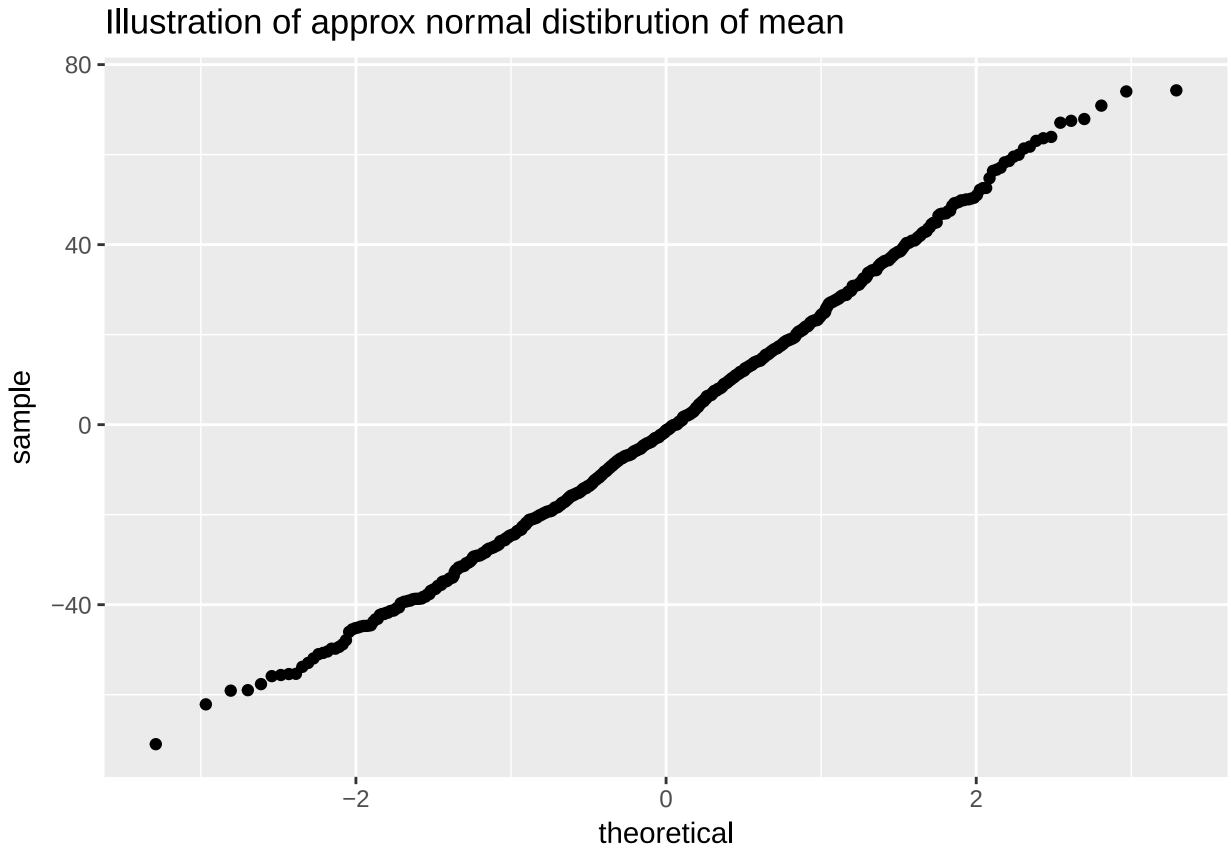
- Normalizes the rexp data and stores result in rexp_mean_CLT
- Histogram of the nromalized mean data. The plot shows an approx normal distribution of the mean as per CLT
- Normal qq plot showing the simulated vs theoretical normal distribution

The plot below shows the normal distribution centered around 0 with 40 data points for each simulation. The distribution becomes more normal as the number of data points in each simulation is increased.

```
rexp_mean_CLT <- (rexp_mean_data-5)*sqrt(n1)/lambda

g2 <- ggplot(data.frame(rexp_mean_CLT),aes(rexp_mean_CLT)) +
      geom_histogram(aes(rexp_mean_CLT,..density..),binwidth=4,fill='yellow',color='black') +
      labs(title='Plot shows approx normality of distribution') + geom_density()
s2 <- ggplot(data.frame(rexp_mean_CLT),aes(sample=rexp_mean_CLT)) + stat_qq() +
      labs(title='Illustration of approx normal distibrution of mean')
g2
```



```
s2
```



Conclusion

Investigated simulated exponential distribution for a normal distribution compared to CLT and found that the normality assumption is approximately valid for simulated data of 40 data points when done across 1000 simulations. The normality test is conducted for mean and variance including comparison to the theoretical mean and variance.