# Coursera - Statistical Inference Project - Part 1

### Introduction

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda ($\lambda$) is the rate parameter. The mean of the exponential distribution is $\frac{1}{\lambda}$ and the standard deviation is also also $\frac{1}{\lambda}$. For this assignment $\lambda = 0.2$ and the distribution of averages of $n = 40$ exponentials will be investigated.

*Note: For this assignment, two thousand simulated averages of 40 exponentials were done. All R-Code will be included in the Appendix.*

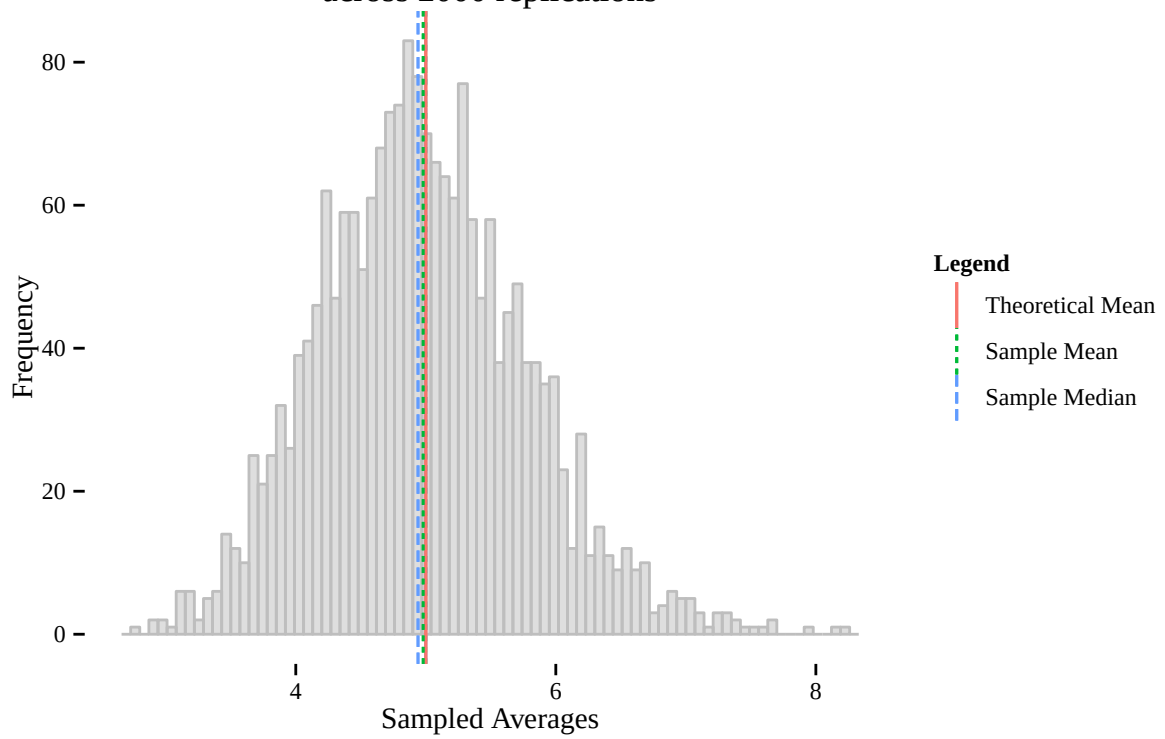
### Tasks to be addressed in this assignment:

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.
2. Show how variable it is and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.
4. Evaluate the coverage of the confidence interval.

### Results and Discussion:

The values of this distribution of means ($n = 40$) were calculated in R and stored in a data frame called *exp_means*. This data frame, consisting of the 2000 averages, was then plotted with a histogram (Figure 1):

Figure 1 − Distribution of Averages of exponentials (lambda = 0.2, n=40) across 2000 replications



As per the Central Limit Theorem (CLT), the distribution of averages of independent and identically distributed (iid) variables becomes that of a standard normal as the sample size increases. In other words, as sample size increases the average of the sample means converges to the population mean. For the exponential distribution, the mean($\mu$) is simply equal to the inverse of the rate parameter($\lambda$) : $\mu = \frac{1}{\lambda}$. As $\lambda = 0.2$ for this assignment, this gives a theoretical population mean of $\frac{1}{\lambda} = 5$ . This theoretical value, as well as the sample mean of averages and sample median of averages were plotted on the histogram above (Figure 1).

```
summary(exp_means)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.76    4.43    4.94    4.98    5.49    8.22
```

The sample mean (4.98) and median (4.94) approach very closely to the theoretical mean of 5.

One point to note, is that as expected the sample mean and median are also approaching each other as the distribution converges to a normal distribution (in a normal distribution, the median = mean). Contrast this with the mean and median of an exponential distribution, which do not usually equal each other. In an exponential distribution, the mean = $\frac{1}{\lambda}$, but the median = $\frac{ln2}{\lambda}$ .

Per the CLT, the theoretical variance of the sample distribution of means is: $\sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n}$. For an exponential distribution, $\sigma = \frac{1}{\lambda}$ . With $n = 40$, and $\lambda = 0.2$, this gives a theoretical variance ($\sigma^2_{\bar{X}}$) of 0.625.
This compares with the actual sample variance calculated for $n = 40$ (across 2000 replications) of:

```
var(exp_means)
```

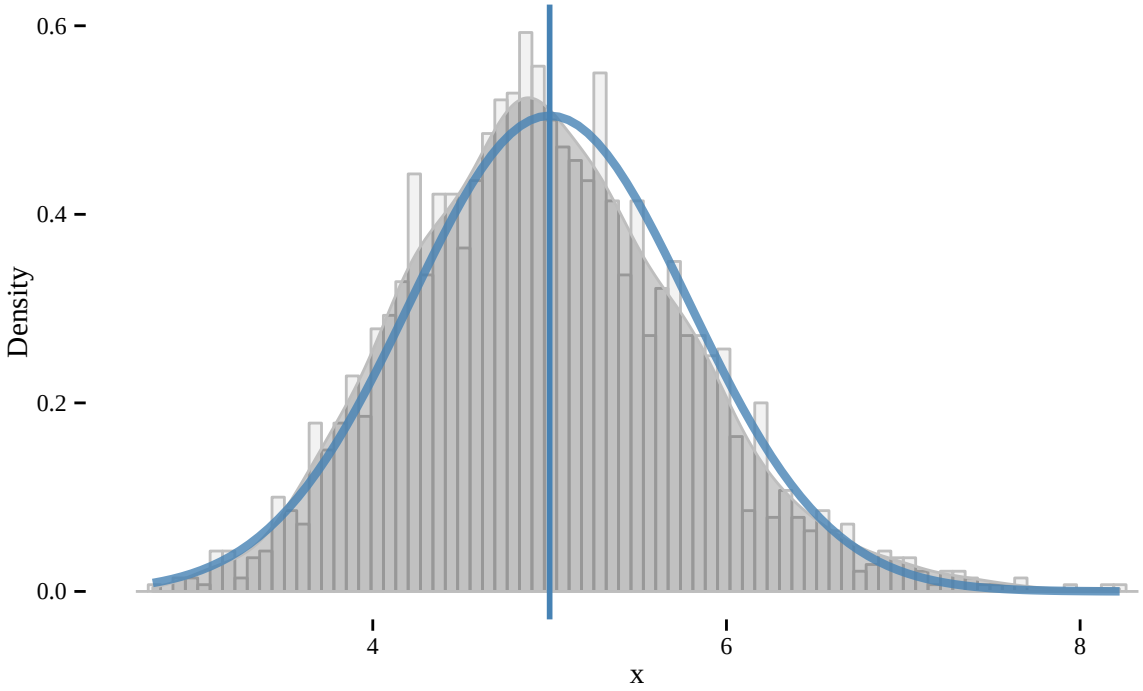
```
## [1] 0.6331
```

In both instances, the Standard Deviation is simply the square root of the variance.

As a normal distribution is determined by it's mean and standard deviation, the fairly close approximation of the sample means and variance (and hence sample standard deviation) to the expected normal values imply the sample distribution is approximately normal.

See Figure 2 (below) for a visual depiction. It overlays the original histogram with the a sample density plot (taken from the same *exp_means* data frame) and a normal curve with $\mu = \frac{1}{\lambda}$ and standard deviation (sd) = $\frac{1}{\lambda\sqrt{n}}$ . As can be seen, the 2000 averages sample distribution does approximate a normal distribution with the theoretical mean and sd parameters (as expected from the CLT).

Figure 2 − Histogram of averages of exponentials overlayed with density sample distribution & normal curve




**Coverage of Confidence Interval**

The quantity $\bar{X} \pm \frac{1.96\sigma}{\sqrt{n}}$ is the coverage for the 95% confidence interval for $\mu = \frac{1}{\lambda}$. For the samples of averages computed, this interval consists of:

```
coverage_interval
```

```
## [1] 3.430 6.529
```

This 95% confidence interval refers to the fact that if one were to repeatedly get samples of size $n = 40$, about 95% of the intervals obtained (noted in the coverage range[3.43 , 6.529 ]) would contain $\mu = \frac{1}{\lambda}$, for $\lambda = 0.2$.

## APPENDIX - R Code

This Appendix consists of all the R-Code used throughout this document:

```
library(ggplot2)
library(ggthemes)
library(knitr)

set.seed(42)

lambda <- 0.2
no_exps <- 40
no_sims <- 2000
norm_vars <- data.frame(rnorm(2000,1/lambda,sqrt(0.625)))

set.seed(42)

exp_means <- mean(rexp(no_exps,lambda))
for(i in 1:(no_sims-1)){exp_means <- c(exp_means,
                                       mean(rexp(no_exps,lambda)))}

summary_stats <- data.frame(Legend = "Theoretical Mean", vals = 1/lambda)
summary_stats <- rbind(summary_stats,
                       data.frame(Legend = "Sample Mean", vals = mean(exp_means)))
summary_stats <- rbind(summary_stats,
                       data.frame(Legend = "Sample Median", vals = median(exp_means)))

summary(exp_means)

ggplot() + geom_histogram(data = data.frame(x=exp_means), aes(x=x),binwidth = 0.07
                          ,  fill = "grey",alpha = 0.5, colour = "grey") +
        geom_vline(data = summary_stats,aes(xintercept = 1/lambda, color = Legend
                                            , linetype = Legend), show_guide = TRUE) +
        geom_vline(data = summary_stats,aes(xintercept = vals, color = Legend,
                                            linetype = Legend),show_guide = TRUE) +
        geom_vline(data = summary_stats,aes(xintercept = vals, color = Legend,
                                            linetype = Legend),show_guide = TRUE) +
        ggtitle("Figure 1 - Distribution of Averages of exponentials (lambda = 0.2
                , n=40)\nacross 2000 replications") +
        xlab("Sampled Averages") + ylab("Frequency") + theme_tufte()

var(exp_means)

ggplot() + geom_histogram(data = data.frame(x=exp_means), aes(x=x, y = ..density..),
                          binwidth = 0.07,  fill = "grey",alpha = 0.2, colour = "grey") +
        geom_density(data = data.frame(x=exp_means), aes(x=x), colour = "grey"
                     , fill ="black", alpha = 0.2) +
        geom_vline(xintercept = 1/lambda, color = "steelblue", linetype = 1, lwd = 1) +
        ggtitle("Figure 2 - Histogram of averages of exponentials
                \noverlayed with density sample distribution &\nnormal curve") +
        xlab("x") + ylab("Density") +
        stat_function(data = norm_vars, fun = dnorm, colour = "steelblue", lwd = 1.5,
                      alpha = 0.8,args = list(mean =5, sd = sqrt(0.625)))
        + theme_tufte()
```

```r
set.seed(42)
coverage_interval <- mean(exp_means) + c(-1,1) * qnorm(0.975)*(1/lambda)*sqrt(1/no_exps)
```