# Statistical Inference Course Project - Part 1

Alexander M Fisher

November 17, 2020

**Introduction:** This is part 1 of the course project that goes along with the statistical inference course a part of the data science specialization that is run by John Hopkins University on Coursera. The project consists of two parts:

- A simulation exercise.
- Basic inferential data analysis.

This part of the project will investigate the exponential distribution in R and compare it with the Central Limit Theorem. There are three main instructions relating to this part listed below.

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Note for this part of the assesment the rate paramater **lamba = 0.2**. This implies the mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda, i.e. a numerical value of 5.

---

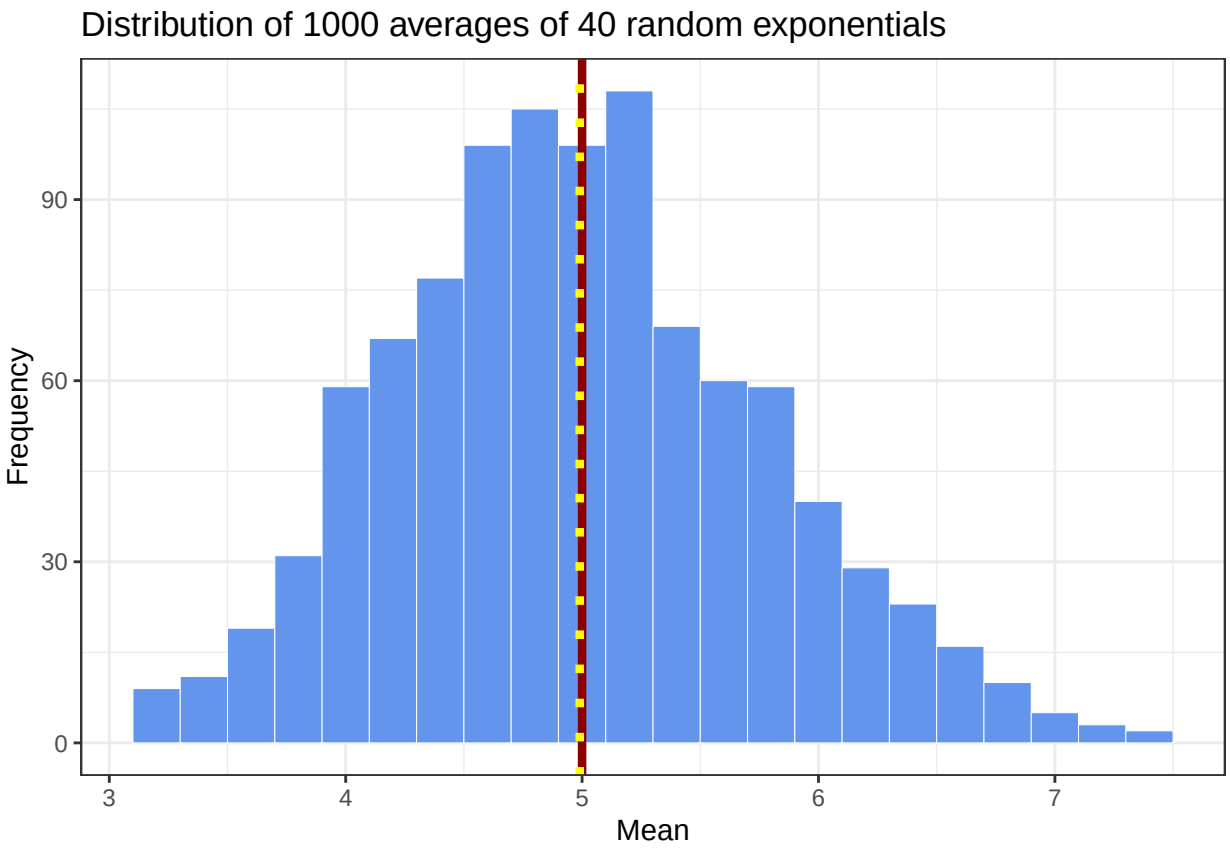
**Part 1: A simulation exercise** Lets generate some 40 exponentials with rate parameter = 0.2 and calculate the average. This simulation will be done 1000 times. The results is then displayed in a histogram.

```
# set seed for reproducability:
set.seed(1)

# generating simulated data:
n_obs <- 40      #number of observations (i.e. per simulation)
n_sim <- 1000     # number of simulations
lambda <- 0.2    # rate paramater lambda
simulated_exponentials <- replicate(n_sim, rexp(n_obs, lambda))
simulated_means <- apply(simulated_exponentials, MARGIN = 2, FUN = mean)

# calculating sample mean and theoretical mean:
theoretical_mean <- 1/lambda
sample_mean <- mean(simulated_means)
```

**Generating Simulated Data and Looking at Means:** Now to plot the results:

```
library(ggplot2)
plot <- ggplot(data = data.frame(simulated_means), aes(x = simulated_means)) +
        geom_histogram(col = "white", fill = "cornflowerblue", binwidth = 0.2, size = 0.1) +
        scale_x_continuous(breaks = scales::pretty_breaks(n = 5)) +
```

```
        geom_vline(aes(xintercept = theoretical_mean), color = "darkred", linetype="solid", size=1.5) +
        geom_vline(aes(xintercept = sample_mean), color = "yellow", linetype="dotted", size=1.5) +
        labs(title = "Distribution of 1000 averages of 40 random exponentials", x = "Mean", y = "Freque
        theme(plot.title = element_text(hjust = 0.5),text = element_text(size=10)) +
        theme_bw()
print(plot)
```




So from this it can be seen that the distribution of averages is not the same as the original distribution (in this instance exponential) that the samples are taken from. The theoretical mean of the distribution is equal to 5 and the sample mean is 4.99 approximately equal.

**Looking at Standard Deviations/Variance:** The distribution of averages of iid variables becomes that of a standard normal as the sample size increases. We expect the mean to stay the same, and the variance to become sd/sqrt(n). This theoretical knowledge will be used to make the variance calculation. See results below. Notice for all intents and purposes the sample and theoretical variances are equal as well as the means in the above section. This is because we took enough samples for CLT to apply well.

```
# calculating sd/variances for distribution of averages:
theoretical_sd <- (1/lambda)/sqrt(n_obs); theoretical_sd
```

```
## [1] 0.7905694
```

```
sample_sd <- sd(simulated_means); sample_sd
```

```
## [1] 0.7817394
```

```
theoretical_var <- theoretical_sd^2; theoretical_var
```

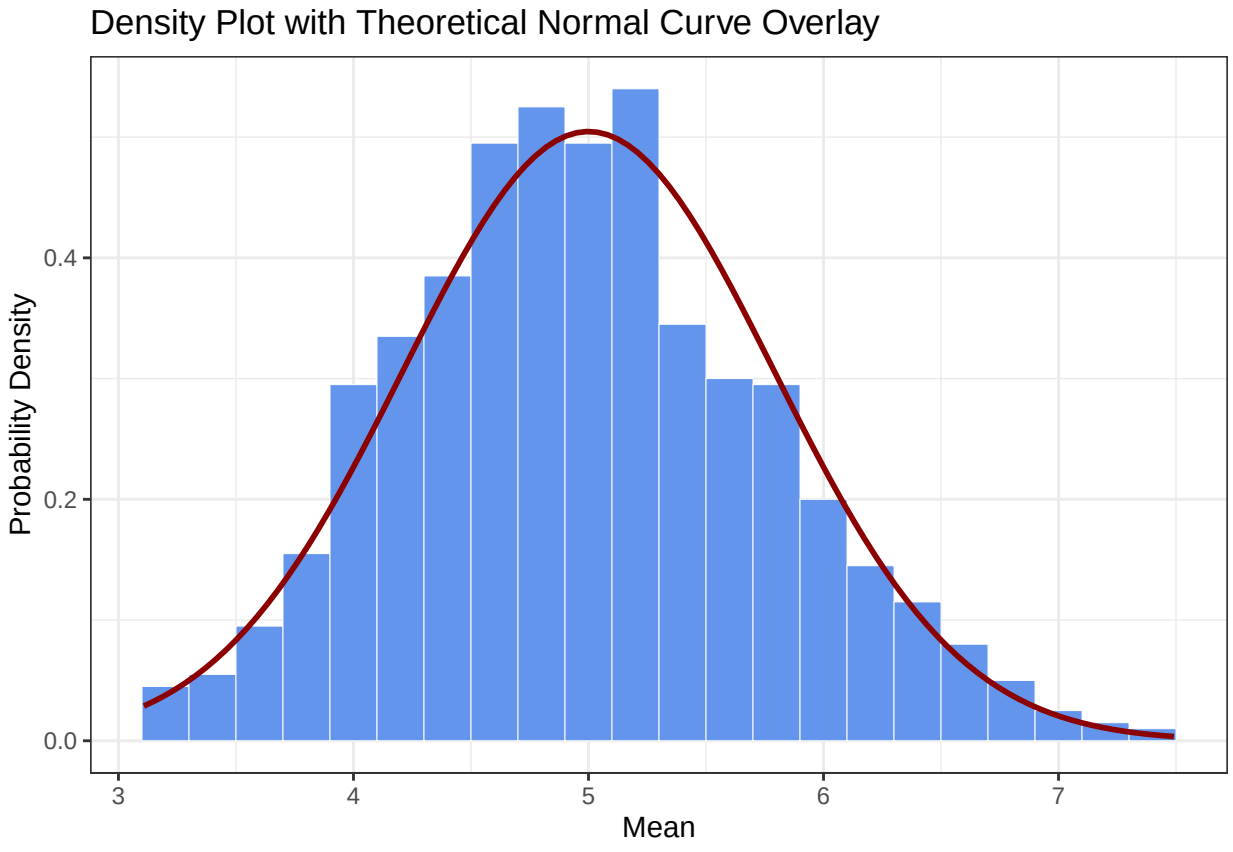
```
## [1] 0.625
```

```
sample_var <- sample_sd^2 ; sample_var
```

```
## [1] 0.6111165
```

**Is the Distribution Normal?:** As discussed above, yes, the distribution of averages becomes normal as sample sizes increase. We can roughly see that in our previous histogram plot, where the samples were pulled from an exponential distribution. Lest have a deeper look at this idea and over lay a normal bell curve with theoretical mean and sd over the plot to see this visually.

```
plot <- ggplot(data = data.frame(simulated_means), aes(x = simulated_means)) +
        geom_histogram(aes(y=..density..),col = "white", fill = "cornflowerblue", binwidth = 0.2, size =
        stat_function(fun = dnorm, args = list(mean = theoretical_mean, sd = theoretical_sd),color = "da
        labs(title = "Density Plot with Theoretical Normal Curve Overlay", x = "Mean", y = "Probability
        theme(plot.title = element_text(hjust = 0.5),text = element_text(size=10)) +
        theme_bw()
print(plot)
```




The distribution quite clearly approximates a normal distribution. As more samples are taken the resulting density plot would get better due to central limit theorem. Other tests such as Q-Q plots can be done, to demonstrate the normality of the resulting distribution, but for now that will complete this analysis and report.

---