# Statistical Inference Course Project Part One: Simulation

*John Higgins*

*November 30, 2017*

## Overview

The purpose of this report will be to demonstrate the Central Limit Theorem. We will be looking at 1,000 averages of 40 exponential random variables, independent and identically distributed with $\lambda = 0.2$.

## Simulations

We set a seed for purposes of repeatability; for personal reasons, I have chosen 1990. We are going to be looking at averages of 40 random variables, so n is set to 40.

We generate 1,000 experiments by generating 40,000 values and organizing them in a matrix with 1,000 rows (one for each experiment) and 40 columns (the 40 values to be averaged for each experiment).

We take the average value of each row to obtain our 1,000 averages of 40 exponential random variables.

```
set.seed(1990)
lambda <- 0.2
n <- 40
simulation.values <- rexp(1000*n, lambda)
simulation.data <- matrix(simulation.values, nrow = 1000, ncol = 40)
simulation.means <- apply(simulation.data, 1, mean)
```

## Mean: Theoretical vs. Observed

Our theoretical mean is $\frac{1}{\lambda} = \frac{1}{0.2} = 5$. Let us compare this to our observed mean of the averages of 40 random variables.

```
mean(simulation.means)
```

```
## [1] 5.035035
```

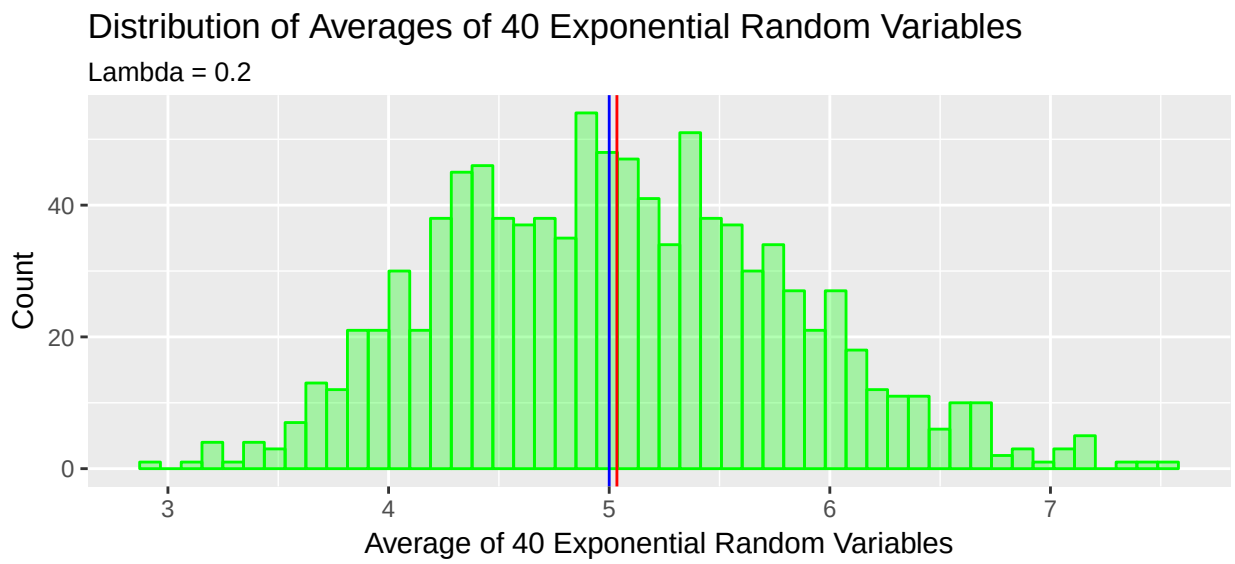
Very close! Let us visualize this result. We will look at a histogram of the data. The blue line represents the theoretical mean, and the red line represents the observed mean.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.1
```

```
ggplot() + aes(simulation.means) +
  geom_histogram(bins = 50,
                 col = "green",
                 fill = "green",
                 alpha = .3) +
  geom_vline(xintercept = mean(simulation.values),
             color = "red") +
  geom_vline(xintercept = 5,
             color = "blue") +
```

```
  labs(title = "Distribution of Averages of 40 Exponential Random Variables",
       subtitle = "Lambda = 0.2",
       x = "Average of 40 Exponential Random Variables",
       y = "Count")
```




## Variance: Theoretical vs. Observed

The theoretical standard deviation of a single exponential random variable with $\lambda = 0.2$ is $\frac{1}{\lambda} = \frac{1}{0.2} = 5$. Thus, the variance is $5^2 = 25$.

The theoretical standard deviation of an average of 40 exponential random variables with $\lambda = 0.2$ is $\frac{5}{\sqrt{40}}$. Thus, the theoretical variance is $\frac{25}{40} = 0.625$.

```
var(simulation.means)
```

```
## [1] 0.6171751
```

Again, very close.

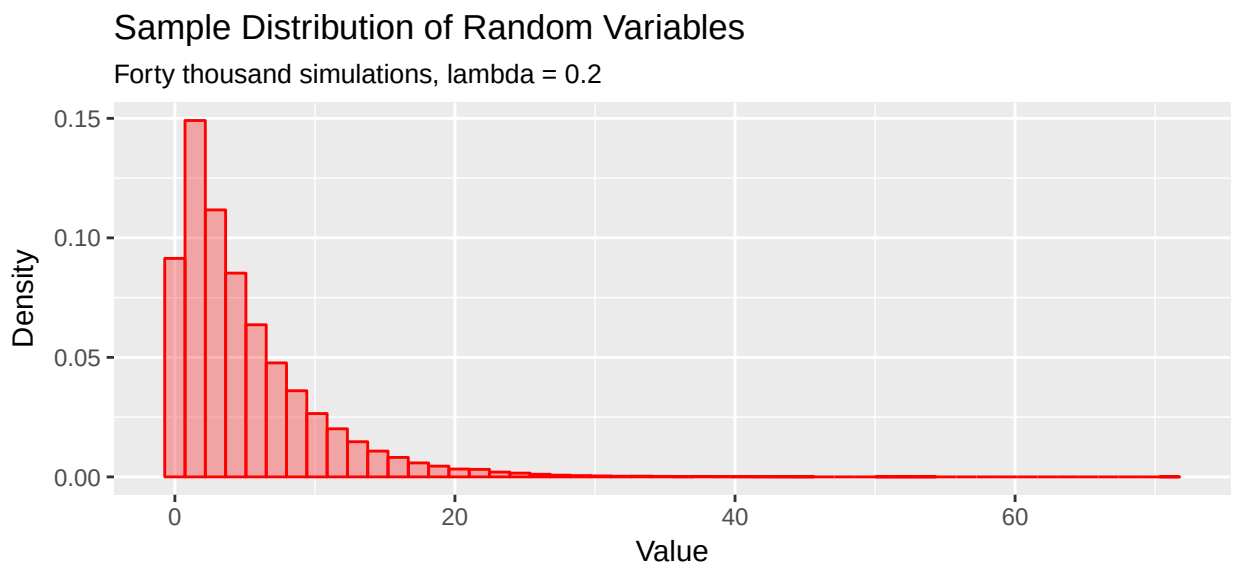
## Distribution

We observe density instead of count – that is, we look at relative frequency. This normalizes the area under the histogram, so it can be compared to a normal distribution curve.

First, the original simulated values.

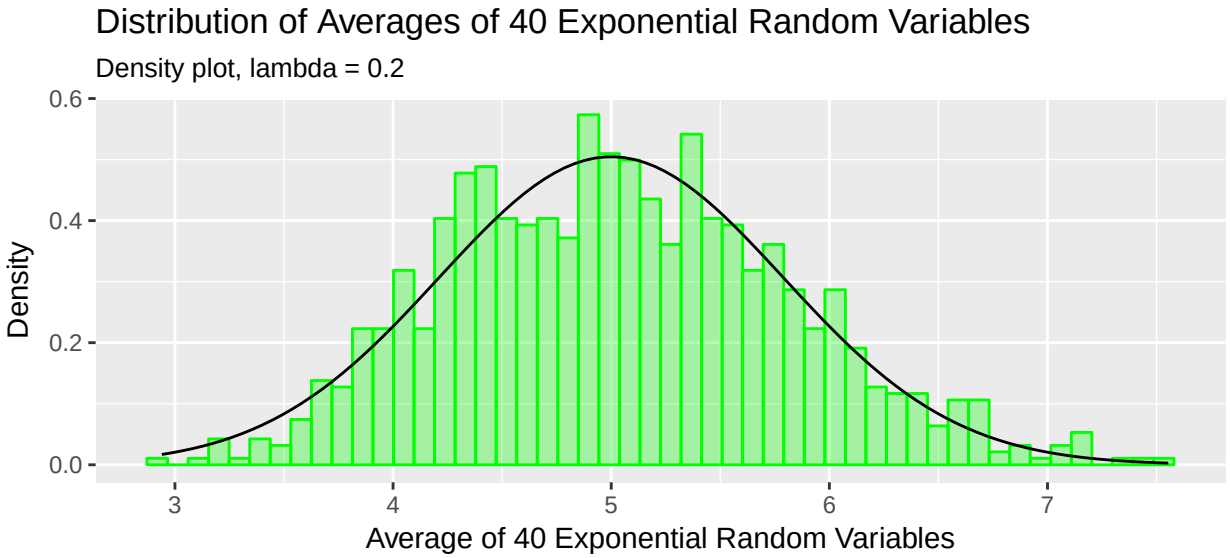
```
ggplot() + aes(simulation.values) +
  geom_histogram(aes(y = ..density..),
                 bins = 50,
                 col = "red",
                 fill = "red",
                 alpha = .3) +
  labs(title = "Sample Distribution of Random Variables",
       subtitle = "Forty thousand simulations, lambda = 0.2",
```

```
      x = "Value",
      y = "Density")
```




Now we look again at the simulated averages of 40 values. We overlay a curve showing a normal distribution to illustrate similarity. This normal curve has mean of 5 (our theoretical mean) and standard deviation of $\frac{5}{\sqrt{40}}$, as we are looking at the distribution of an average of 40 samples.

```
ggplot() + aes(simulation.means) +
  geom_histogram(aes(y = ..density..),
                 bins = 50,
                 col = "green",
                 fill = "green",
                 alpha = .3) +
  stat_function(fun = dnorm,
                args = c(mean = 5, sd = 5/sqrt(40))) +
  labs(title = "Distribution of Averages of 40 Exponential Random Variables",
       subtitle = "Density plot, lambda = 0.2",
       x = "Average of 40 Exponential Random Variables",
       y = "Density")
```


Distribution of Averages of 40 Exponential Random Variables
Density plot, lambda = 0.2
Density
0.6
0.4
0.2
0.0
3
4
5
6
7
Average of 40 Exponential Random Variables