# Statistical Inference Project - Part 1

*Jonathan Brant*

*Sunday, January 18, 2015*

## Synopsis

The goal of the first portion of this project is to compare the results of R's exponential distribution with the central limit theorem. Specifically, to determine whether the exponential distribution will exhibit a tendency toward normality regardless of the underlying distribution. In order to do this, R's rexp function is utilized, feeding it the population size and the rate. For the purposes of this study, the number of observations (population size) is 40 (n = 40) and the rate parameter is fixed at 0.2 ($\lambda = 0.2$) for all simulations. The mean is calculated for all exponentials in the populations over 1,000 simulations.

As a motivating example, differences between two distributions of random uniforms are compared: one with the density distribution of 1,000 random uniforms and the other containing the distribution of 1,000 averages of 40 random uniforms. Figure 1 below depicts the results.

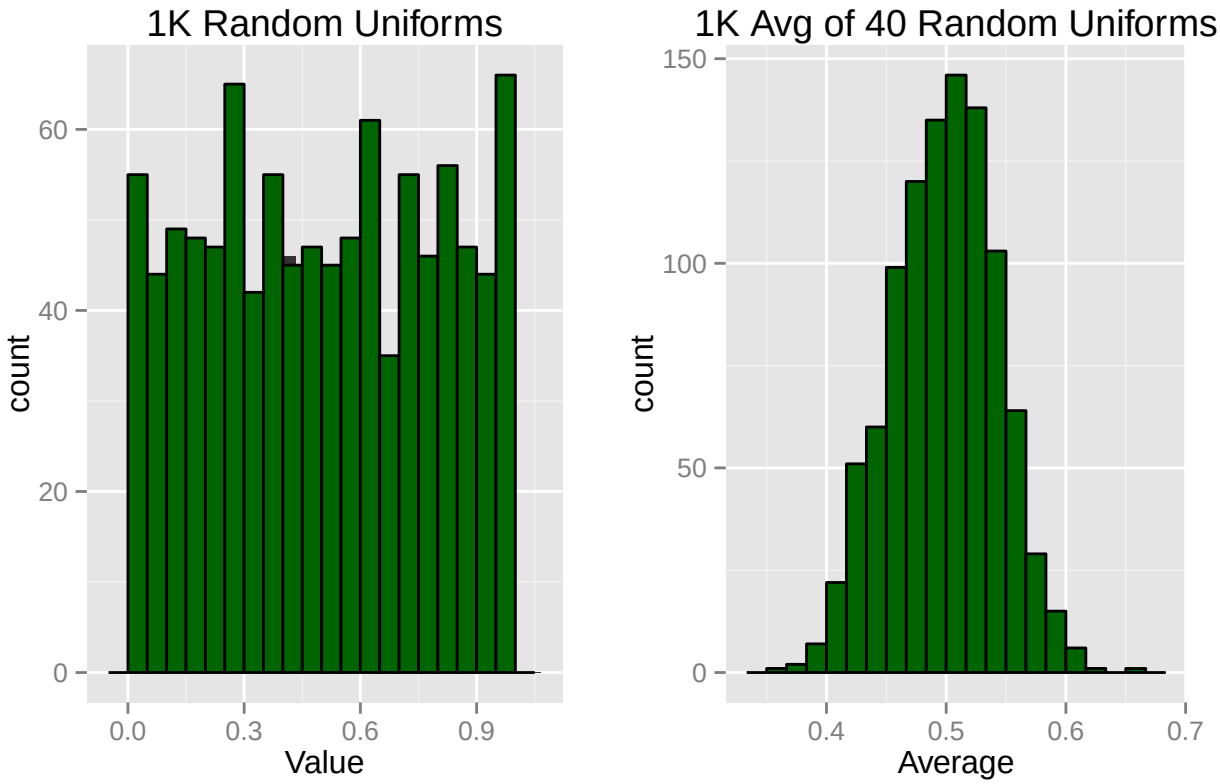


**Figure 1: Comparison of Random Uniform Distributions**

Notice how the distribution of 1,000 averages of 40 random uniforms is far more gaussian and symmetric about the mean compared to the more irregular distribution of 1,000 random uniforms. It's this distribution of averages of random normals that will be considered and compared to the results of theoretical values from the central limit theorem.

## Simulations

The first two tasks are to calculate the mean and variance of both the simulated exponential distribution as well as the theoretical (or expected) values produced by applying said distribution. The theoretical mean is given by $\frac{1}{\lambda}$ while the theoretical standard deviation and variation are given by $\frac{\frac{1}{\lambda}}{\sqrt{n}}$ and $(\frac{\frac{1}{\lambda}}{\sqrt{n}})^2$ respectively. As previously stated, the sample mean, standard deviation, and variance are all based on the 1,000 simulations of the average of 40 random uniforms.

By mere visual observation of table 1, it's fairly clear that the sample mean, sample standard deviation, and sample variance all exhibit minimal deviation from their expected values (with the largest difference, unsurprisingly, being with the measure of variance). The sample mean, at 4.974, lies very close to the theoretical value of 5, indicating that the sample shares similar spatial characteristics. The sample variance, while not alarmingly far off from the theoretical variance, is a bit further off than one might initially expect. Given the greater similarity in means, this indicates that the sample distribution is likely shorter with a slightly greater width (as confirmed in figure 1 below).

| | Theoretical (Expected) | Sample |
|---|---|---|
| **Mean** | 5 | 4.974 |
| **Standard Deviation** | 0.791 | 0.755 |
| **Variance** | 0.626 | 0.57 |

Table 1: Mean, Standard Deviation, and Variance of Distributions

The last task involves verifying the normality of the sample distribution. The first step toward accomplishing is to plot the distribution of the sample mean of the 40 random exponentials. Figure 2 depicts the aforementioned distribution along with the density curve. Despite two peaks and a slightly leftward skew, the distribution appears to demonstrate salient gaussian features. In addition to the sample data, the theoretical density curve is overlayed, demonstrating that the sample density curve follows a similar path, albeit slightly broader, with a lower peak (confirming the above statement regarding the differences in variances). Finally, the sample mean and theoretical mean are plotted next to each other, demonstrating that the center of the distributions are almost exactly overlapping at the theoretical value of 5.

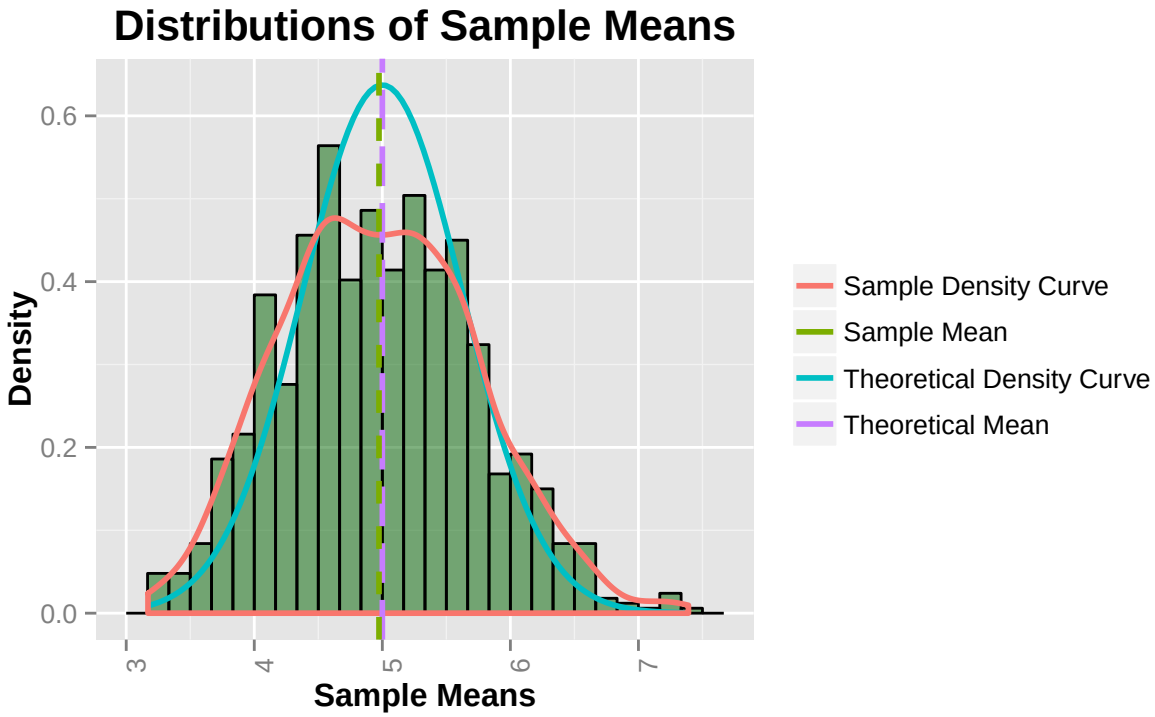


**Figure 2: Distribution of Sample data overlaved with theoretical distribution**

As a final sanity check, a Q-Q plot is utilized in order to chart the quantiles of the sample distribution against the quantiles of a perfectly normal distribution. Again, it's evident that the sample distribution is approximately normal with a very low margin of error.

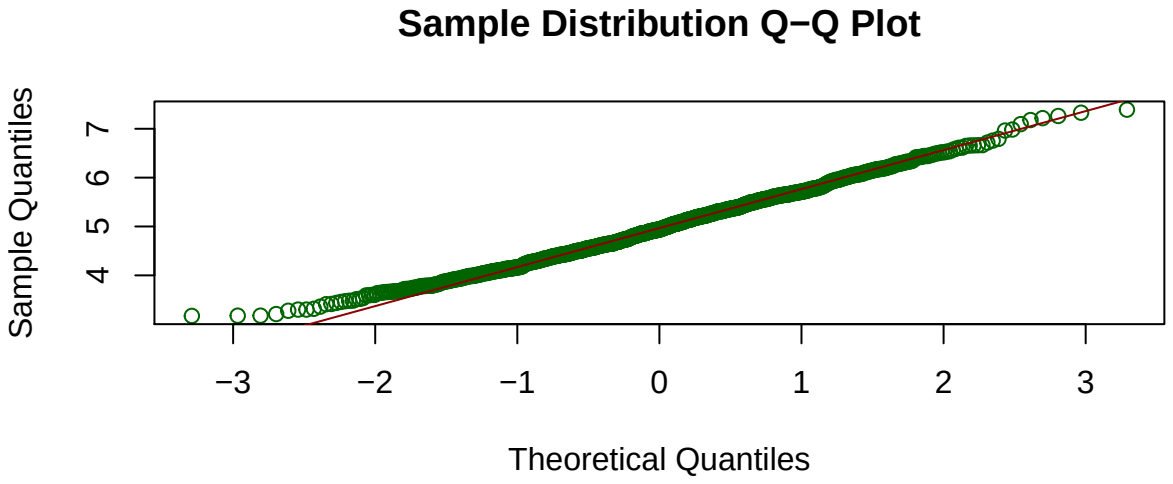


**Figure 3: Q−Q Plot showing sample distribution quantiles**

## Appendix

### Code for setting up and running simulation

```
## Setup simulation parameters
lambda <- 0.2                # Rate parameter
n <- 40                      # Population
total.simulations <- 1:1000  # Number of simulations

## Run the simulation
simulation.matrix <- data.frame(
  sim.means = sapply(
    total.simulations,
    function(x)
      mean(rexp(n, lambda))))
```

### Code for Calculating Mean, Standard Deviation, and Variance

```
## Calculate the sample mean (rounding to three decimals places)
sample.mean <- round(mean(simulation.matrix$sim.means), precision)

## Calculate the theoretical mean
theoretical.mean <- 1 / lambda

## Calculate the sample standard deviation
sample.sd <- round(sd(simulation.matrix$sim.means), precision)

## Calculate the theoretical standard deviation
theoretical.sd <- round(1/lambda/sqrt(n), precision)

## Calculate the sample variance
sample.variance <- round(sample.sd^2, precision)

## Calculate the theoretical variance
theoretical.variance <- round(theoretical.sd^2, precision)
```

```
## Import necessary plotting libraries
suppressMessages(library(ggplot2))
suppressMessages(require(gridExtra))

## Generate random uniform histogram
random.uniforms.distribution <- qplot(runif(1000), bindwidth = 1/20) +
  geom_histogram(color="black", fill="darkgreen", binwidth = 1/20) +
  labs(title = "1K Random Uniforms", x = "Value")

## Calculate the mean of 1,000 simulations of 40 random normals
mns = NULL
for (i in 1:1000)
  mns = c(mns, mean(runif(40)))

## Generate the average random normal histogram
```

```
avg.random.uniforms.distribution <- qplot(mns, binwidth = 1/60) +
  geom_histogram(color="black", fill="darkgreen", binwidth=1 / 60) +
  labs(title = "1K Avg of 40 Random Uniforms", x = "Average")

## Display the histograms
suppressMessages(grid.arrange(random.uniforms.distribution,
             avg.random.uniforms.distribution,
             ncol = 2,
             sub = textGrob(
               "Figure 1: Comparison of Random Uniform Distributions",
               gp = gpar(fontface = "bold", col = "grey20", fontsize = 12))))
```

**Code for Plotting Figure 2**

```
## Generate the plot of sample distributions
sample.distribution.plot <- ggplot(data = simulation.matrix, aes(x = sim.means)) +
  geom_histogram(
    aes(y = ..density..),
    fill = "darkgreen",
    binwidth = 1/6,
    color = "black",
    alpha=1/2) +
  stat_function(
    fun = dnorm,
    args = list(theoretical.mean, theoretical.variance),
    aes(color = "Theoretical Density Curve"),
    size = 1) +
  geom_density(
    aes(color = "Sample Density Curve"),
    size = 1,
    show_guide= FALSE) +
  geom_vline(
    aes(color = "Theoretical Mean", xintercept = theoretical.mean),
    size = 1,
    linetype = "longdash") +
  geom_vline(
    aes(color = "Sample Mean", xintercept = sample.mean),
    size = 1,
    linetype = "dashed") +
  labs(
    title = "Distributions of Sample Means",
    x = "Sample Means",
    y = "Density") +
  theme(plot.title = element_text(face = "bold", size = 16, vjust = 1),
      axis.title.x = element_text(face = "bold"),
      axis.title.y = element_text(face = "bold"),
      axis.text.x = element_text(angle = 90, hjust = 1),
      legend.title = element_blank())
```

**Code for Plotting Figure 3**

```
## Generate the Q-Q plot
par(omi = c(1,0,0,0))
qqnorm(
  simulation.matrix$sim.means,
  col="darkgreen",
  main="Sample Distribution Q-Q Plot")
qqline(simulation.matrix$sim.means, col="darkred")
```