Simulation Analysis of the Exponential Distribution

Shenay

June 9, 2019

Synopsis

In this report, I will conduct simulations on a particular exponential distribution (assuming that lambda = 0.2), and compare the generated results with properties underlying the Central Limit Theorem. Specifically, the CLT states that for a sufficiently large sample size, the distribution of the sample means are approximately normal. In this case, we will look at the averages of 40 exponentials that are randomly and independently drawn, and repeat the process 1000 times during the simulation.

Run Simulations

```r
library(ggplot2)
library(gridExtra)
```

STEP 0: Set up the global variables

```r
n <- 40
lambda = 0.2
sims <- 1000
avg <- 1/lambda
sd <- sqrt(1/lambda^2)
```

STEP 1: Conduct 1000 simulations of the exponential distribution

```r
set.seed(1234)
x1 <- rexp(n = sims, rate = lambda)

plot1 <- ggplot(data = as.data.frame(x1), aes(x = x1)) +
         geom_histogram(binwidth = 1, fill = "wheat", col = "1") +
         labs(title = "Histogram of 1000 Simulations for Y = 0.2*EXP(-0.2X)",
             x = "Y Values", y = "Frequency") +
         theme_classic()
```

STEP 2: Investigate the distribution of averages of 40 exponentials

```r
set.seed(4321)
x2 <- NULL
for (i in 1:sims) {
        x2 <- c(x2, mean(rexp(n = n, rate = lambda)))
}

plot2 <- ggplot(data = as.data.frame(x2), aes(x = x2)) +
         geom_histogram(fill = "plum", col = "1") +
         labs(title = "Histogram of 1000 Simulations Using the Average of 40",
              subtitle = "Function Y = 0.2*EXP(-0.2X)", x = "Y Values",
              y = "Frequency") + theme_classic()
```

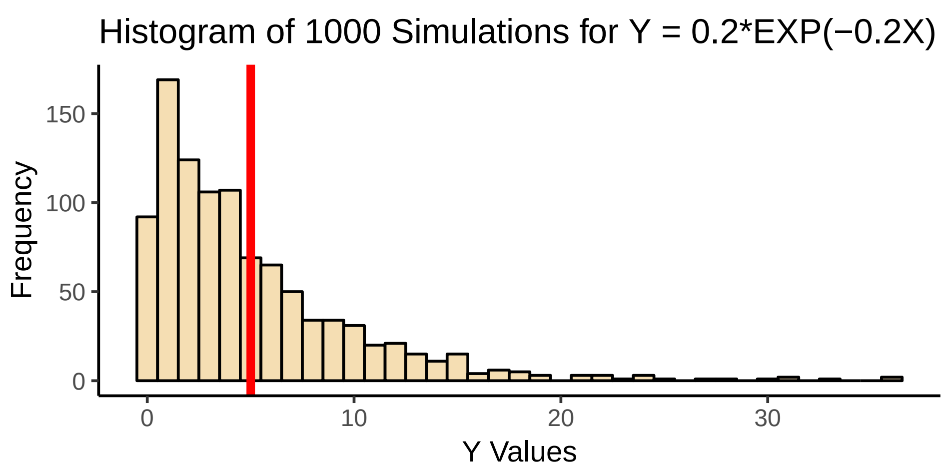
Comparison of Means

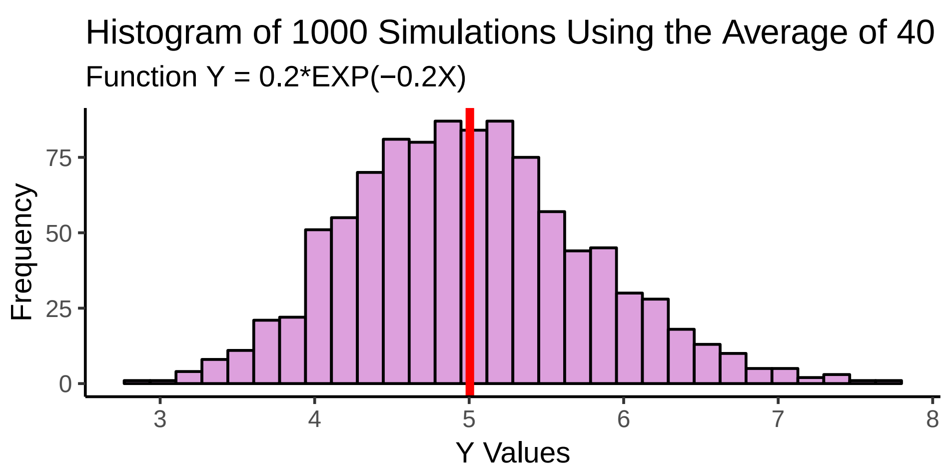
```
avg_mat <- matrix(0, nrow = 1, ncol = 2)
colnames(avg_mat) <- c("Sample", "Theoretical")
rownames(avg_mat) <- "Mean"

avg_mat[1, 2] <- avg
avg_mat[1, 1] <- mean(x2)
avg_mat
```

```
##         Sample Theoretical
## Mean 5.004574           5
```

```
grid.arrange(plot1 + geom_vline(xintercept = mean(x1), col = "red", size = 1.5),
             plot2 + geom_vline(xintercept = mean(x2), col = "red", size = 1.5), nrow = 2)
```





Summary: According to both the summary statistics and the histogram, the sample mean is centered at 5.004574 while the theoretical mean is 5. Therefore, we can conclude that the sample mean is approximately the same as the theoretical mean.

Comparison of Variances

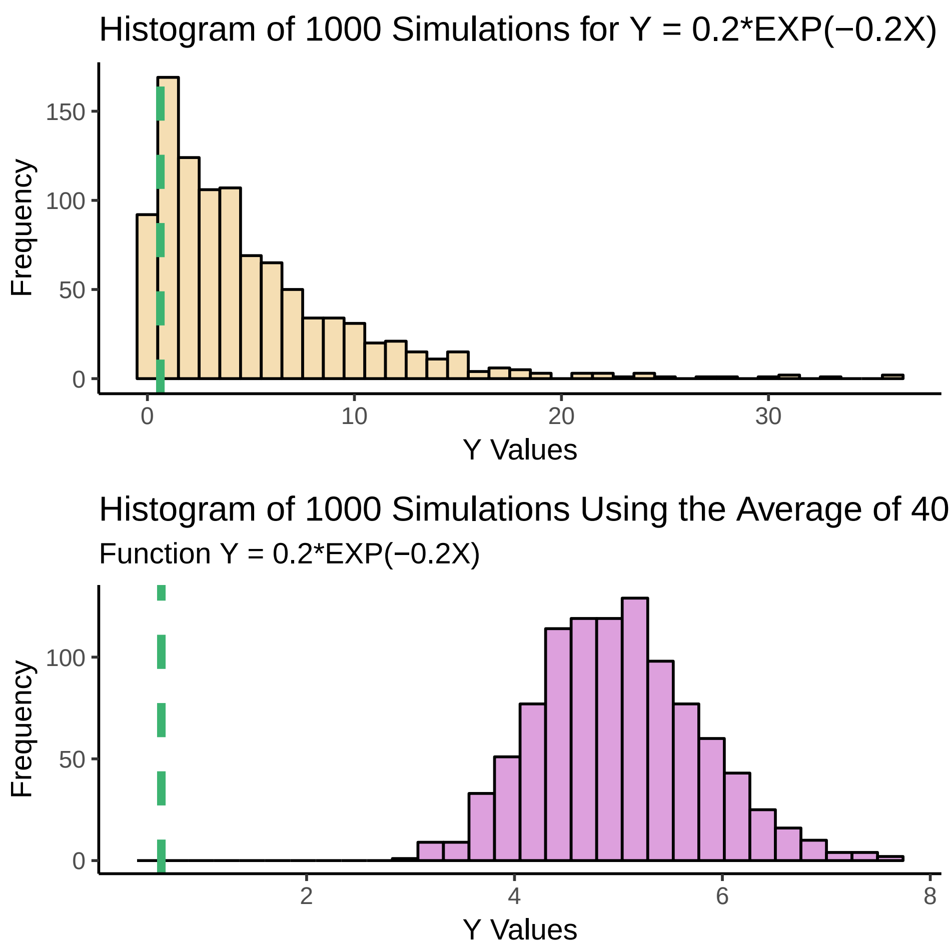
```
var_mat <- matrix(0, nrow = 1, ncol = 2)
colnames(var_mat) <- c("Sample", "Theoretical")
```

```r
rownames(var_mat) <- "Variance"

var_mat[1, 2] <- sd^2/n
var_mat[1, 1] <- var(x2)
var_mat
```

```
##              Sample Theoretical
## Variance 0.6021698       0.625
```

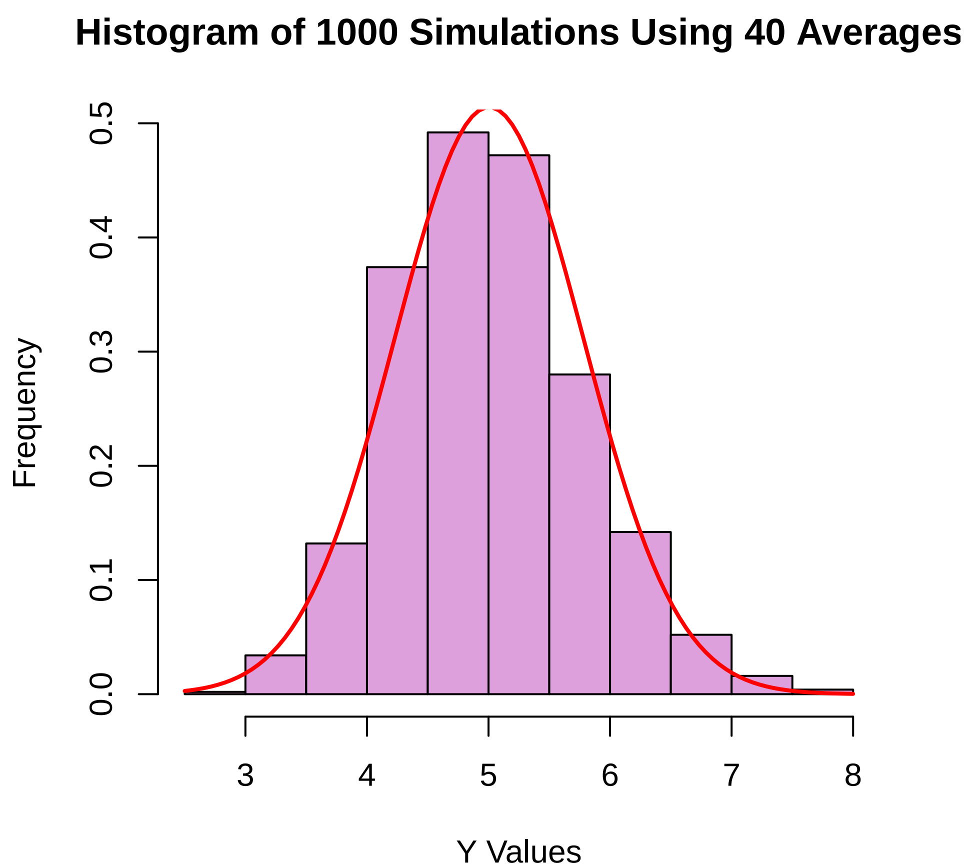
```r
grid.arrange(plot1 + geom_vline(xintercept = sd^2/n, linetype = "dashed",
                               col = "mediumseagreen",size = 1.5),
             plot2 + geom_vline(xintercept = var(x2), linetype = "dashed",
                               col = "mediumseagreen", size = 1.5), nrow = 2)
```



Summary: As shown from both the summary statistics and the histogram, the sample variance is 0.6021698 and the theoretical variance is 0.625. Hence the sample variance is also roughly equal to the theoretical variance of the distribution.

Comparison with the Normal Distribution

```r
hist(x2, freq = FALSE, col = "plum", xlab = "Y Values", ylab = "Frequency",
     main = "Histogram of 1000 Simulations Using 40 Averages")
curve(dnorm(x, mean = mean(x2), sd = sd(x2)), add = TRUE, col = "red", lwd = 2)
```



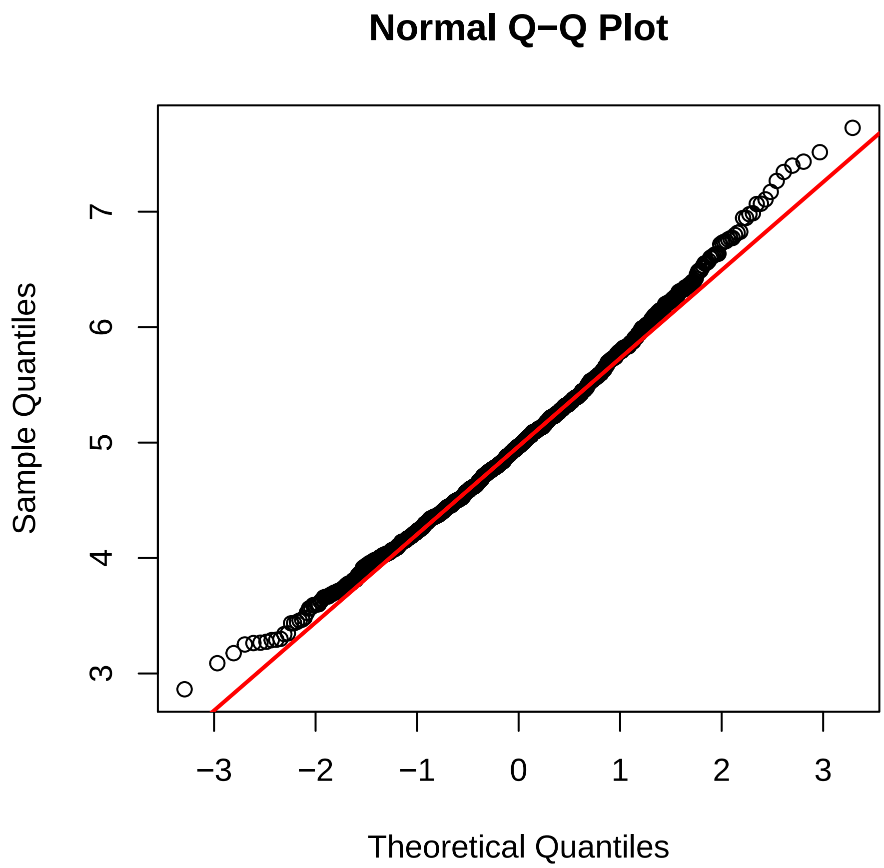

```r
shapiro.test(x1)  #p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x1
## W = 0.80427, p-value < 2.2e-16
```

```r
shapiro.test(x2)  #p-value is 0.0001751, therefore reject null
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x2
## W = 0.99328, p-value = 0.0001751
```

```r
qqnorm(x2)
qqline(x2, col = "red", lwd = 2)  #vast majority aligns with theoretical line
```



Summary: The super-imposed normal distribution curve roughly matches that of our simulated distribution. In order to examine further, we will run the Shapiro-Wilk test where the null hypothesis states that the population is normally distributed. Based on the test results, the p-values for the sample and theoretical distributions are less than 0.05, which suggests that the null hypothesis should be rejected.

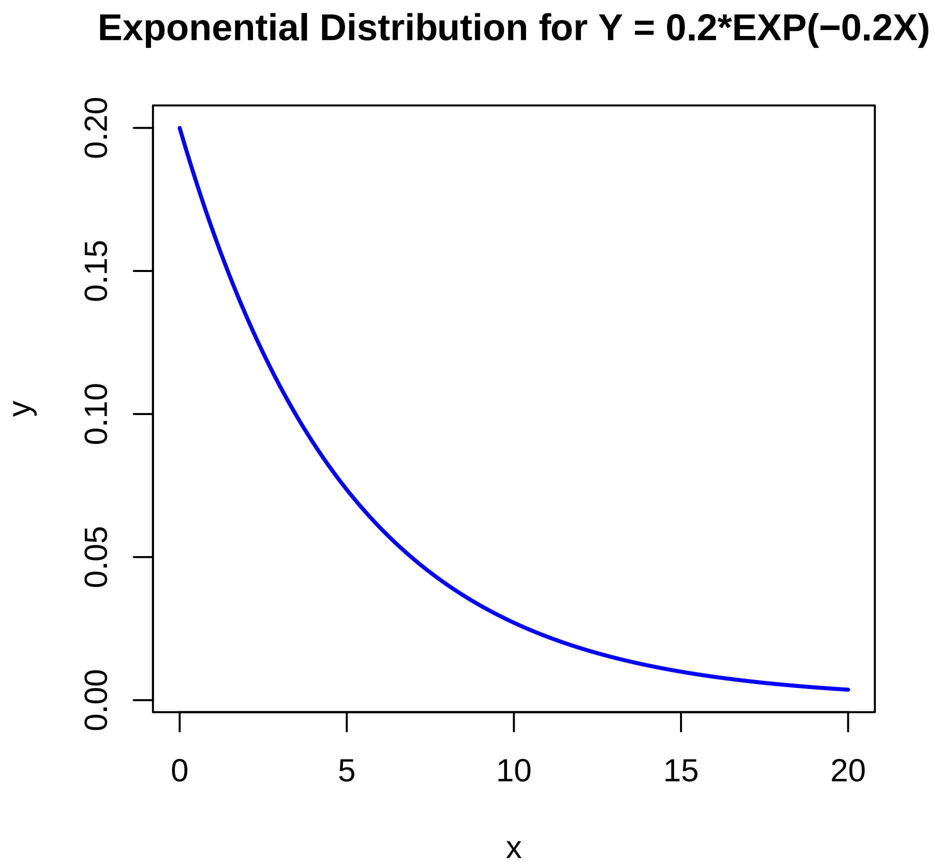
Additionally, the quantile-quantile plot further enables us to see whether there is a deviation from the normal distribution. According to the Q-Q plot, the sample distribution of the average of 40 exponentials approximately matches the normal distribution.

Appendix

- n = number of observations/exponentials
- lambda = rate parameter
- sims = number of simulations
- avg = mean of exponential distribution
- sd = standard deviation of exponential distribution

Plot of the exponential distribution:

```
x <- seq(0, avg + 3 * sd, 0.01)
y <- lambda * exp(-lambda * x)
plot(x, y, type = "l", col = "blue", lwd = 2,
     main = "Exponential Distribution for Y = 0.2*EXP(-0.2X)")
```


Exponential Distribution for Y = 0.2*EXP(−0.2X)
0.20
0.15
0.10
0.05
0.00
y
0
5
10
15
20
x