A Simulation Exercise and the Central Limit Theorem

Jim White
November 17, 2015

Overview:

This project investigated the exponential distribution in R and compared it with the Central Limit Theorem. For the purposes of this demonstration lambda = 0.2 and the investigation includes the distribution of averages of 40 exonentials over 1000 simulations. Explanation is provided regarding the properties of the mean and variance of the 40 exponentials.

Sumulations

Exponential Distribution: "the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate." From Wikipedia (https://en.wikipedia.org/wiki/Exponential_distribution).

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The expected value (μ) of the exponential distribution is E[X] = $\frac{1}{\lambda}$. The variance of X is represented by Var[X] = $\frac{1}{\lambda^2}$. Consequently the standard deviation is $\frac{1}{\lambda}$.

To briefly examine a visual representation of the exponential distribution, let lambda be equal to the values of 1, 2, & 3. For the value of x we will create 100 random values between 0 and 5 for each of the λ values. The f(x) value will be expressed as follows: f(x;$\lambda) = \lambda e^{-\lambda x}$

The figure exponential distribution simluations can be found on the firts page of the Appendix.

Demonstration of the Central limit Theorem (CLT)

A definition of the **Central Limit Theorem:** "states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed." From Wikipedia (https://en.wikipedia.org /wiki/Central_limit_theorem).

Even though the exponential distribution does not represent a normally distributed curve, the following simulation(s) will show, through the use of the *Law of Large Numbers*, that taking a large number of random samples from the distribution will result in the averages (and standard deviations) of those samples approaching a normal distribution.

Or as stated within slide 7/31 from the *Asymptotics and LLN* lecture:

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma$$

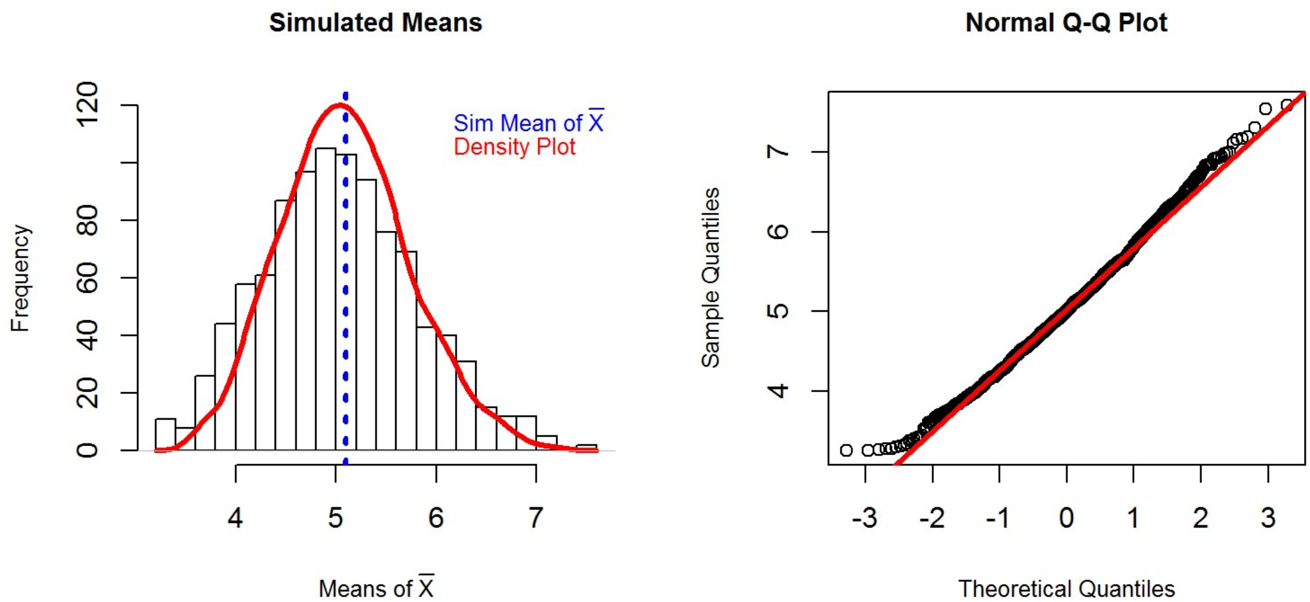
orr

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Simulation Samples Taken from the Exponential Distribution
As an indicator for the CTL, a simulation will be completed in which λ = 0.2, 1000 simulations are used representing 40 exponentials.

As shown, in Figure 1 (below), the mean of the simulated samples appear in the histogram to approach a normal distribution. To further emphasize the red line within the figure is a kernel density plot of the simulated means data. The blue line represents the mean of the simulation. In addition, a Q-Q Plot (a probability plot) was constructed to examine normal distribution. Departures from a straightline create via the qqline function indicates departures from normality. From Cookbook for R (https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test).

Figure 1: Simulated Mean Samples



The Q-Q Plot indicates some variation in the tails of the dataset.

An additional test that can be used to test normality is the Shapiro-Wilk test. This is the most common test for normality ([Wikipedia][4]).

```
## 
##  Shapiro-Wilk normality test
## 
## data:  sim_means
## W = 0.99398, p-value = 0.0004781
```

From the Shapiro-Wilk test, the p-value indicates that there is only a small chance (0.0004781) that the data is normal. But $\bar{X} \approx \mu$ and $s \approx \sigma$ may still be true.

For the exponential distribution, the theoretical (or expected) mean (μ) = $\frac{1}{\lambda}$. If λ = 0.2, then the μ = 5.

To determine if the sample means approximate the theoretical (or expected) mean, a confidence interval is taken at the 95% interval for the sample:

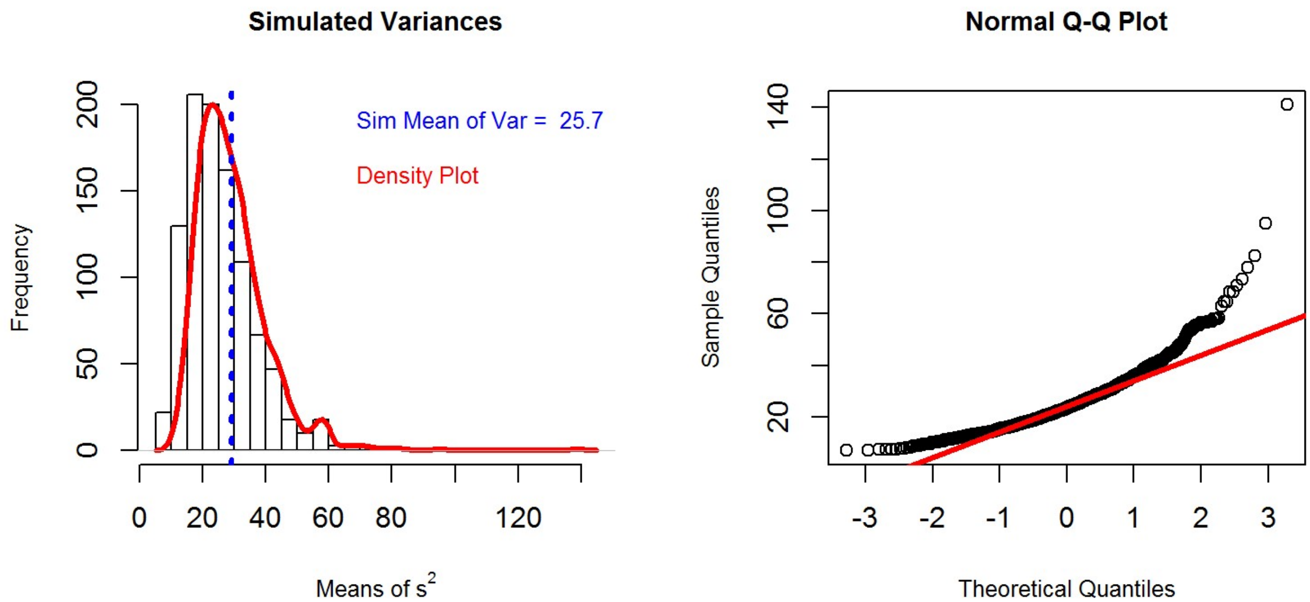
```
## [1] 4.994940 5.091166
```

Thus since the theoretical (or expected) mean (of 5) falls within the confidence interval, the hypothesis that $\bar{X} \neq \mu$ can be rejected.

Simulation of the variance:

Next to analyze the variance of the simulation. The theoretical or expected variance for the exponential distribution is equal to $\frac{1}{\lambda^2}$ or $(1/.2^2 = 25)$.

Figure 2: Simulated Variance Samples



As with the mean, the variance indicates a variation (though somewhat larger) from normal within the tails. This is also evident when examining Figure 2. Next to examine the Shapiro - Wilk test.

```
##
##  Shapiro-Wilk normality test
##
## data:  sim_var
## W = 0.87442, p-value < 2.2e-16
```

The Shapiro-Wilk test indicates the distribution is not normal.

To determine if the sample mean of the variances approximate the theoretical (or expected) variance, a confidence interval is taken at the 95% interval for the sample:

```
## [1] 24.94734 26.42501
```

Thus since the theoretical (Or expected) variance (of 25) falls within the confidence interval, the hypothesis that $s^2 \neq \sigma^2$ can be rejected.

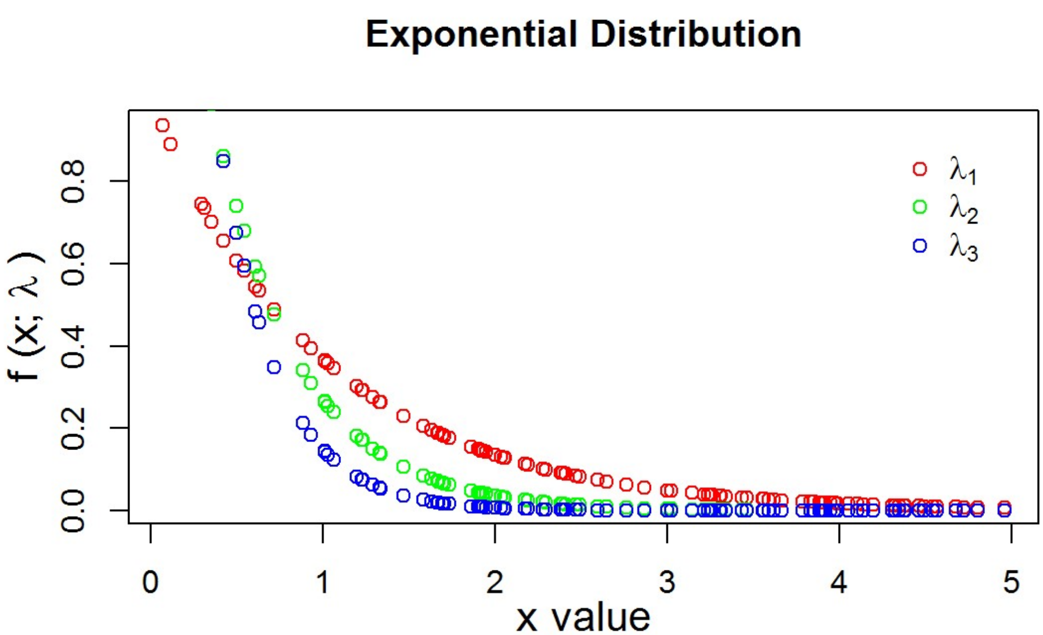
Conclusion: Consequently, it can be concluded that the Central Limit Theorem appears to be true.

Appendix

Code and Figure: The Exponential Distribution (an example)

```
set.seed(1) # to ensure reproducibility
# randomly select values for x
x <- runif(100, 0, 5)
# set lambda values and calculate respective y values
lambda <- 1
while(lambda <= 3){
    #Calculate the y values
    assign(paste("y", lambda, sep = ""),
           lambda*exp(-lambda*x))
    lambda <- lambda + 1
    }
# plot the three x.y groups as points
par(cex.lab=1.3, mgp = c(2, 1, 0))
plot(x, y1, type = "p", col = "red", main = "Exponential Distribution",
     xlab = "x value", ylab = expression(paste("f (x; ", lambda," )")))

points(x, y2, col = "green")
points(x, y3, col = "blue")
legend("topright", c(expression(lambda[1]), expression(lambda[2]), expression(lambda[3
])), col = c("red", "green", "blue"), bty = "n", inset = .05, pch = 1, cex = 1.0)
```



As can be seen from the figure, the exponential ditribution is basically an asymptote. In addition, as the value of $lambda$ -> 0, the line approaches a straight-line model.

R Code for creating figures within the document

For Figure 1 Simulated Mean Samples

For Figure 2 Simulated Variance Samples

```
# run simulation
set.seed(5)  # set seed for randomization for reproducibility
lambda <- 0.2  # set value of constant lambda
sample_size <- 40  # set sample size
sim <- 1000  # set number of simulations to run
sim_var <- replicate(sim, var(rexp(sample_size, lambda))) # run simulation

# create plots
par(mfrow = c(1, 2)) # set parameters
# histogram of simulation
hist(sim_var, breaks = 30, main = "Simulated Variances",
    xlab = expression(paste("Means of ", s^{2})), ylim = c(0, 200), cex.main = 0.9,
    cex.lab = 0.8)
par(new = TRUE)
# density plot of simulation
plot(density(sim_var), axes = FALSE, bty = "n", xlab="", ylab="", col = "red", lwd = 3
,
     main = "")
# line representing mean of simulation
abline(v = mean(sim_var), col = "blue", lwd = 3, lty = 3, xpd = FALSE)
# create legend
x <- toString(round(mean(sim_var), 1))
legend("topright", legend = c(paste("Sim Mean of var = ", x), " ", "Density Plot"),
     text.col = c("blue", "white", "red"), bty = "n", cex = .8)
par(new = FALSE)

## create Q-Q Plot
par(cex.main = 0.9, cex.lab = .8)
qqnorm(sim_var)
qqline(sim_var, col = "red", lwd = 3, xpd = FALSE)
```

Code for running the Shapiro Tests

```
shapiro.test(sim_means)
```

```
shapiro.test(sim_var)
```

Code for the confidence intervals

```
mean(sim_means) + c(-1, 1)*qnorm(0.975)*sd(sim_means)/sqrt(length(sim_means))
```

```
mean(sim_var) + c(-1, 1)*qnorm(0.975)*sd(sim_var)/sqrt(length(sim_var))
```