Exploring the Central Limit Theorem via the Exponential Distribution

Author: Lowell Ricketts Date: March 2016

Synopsis

The exponential distribution differs greatly from other distributions. However, when we take the mean of 40 random exponential variables, simulated 1,000 times. The distribution of means, after normalizing, is approximated quite well by the standard normal distribution. This is the fundamental finding of the Central Limit Theorem.

Introduction

This project was completed during the Statistical Inference course offered by Coursera in the Data Science specialization tract. The course was taught by Brian Caffo from Johns Hopkins Bloomberg School of Public Health.

In this project I was tasked with investigating the exponential distribution in R and to compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with `rexp(n, lambda)` where lambda is the rate parameter. The mean of the exponential distribution is `1/lambda` and the standard deviation is also `1/lambda`. Set `lambda = 0.2` for all of the simulations. I investigated the distribution of averages of 40 exponentials over 1,000 simulations.

Please see Appendix for supporting R code

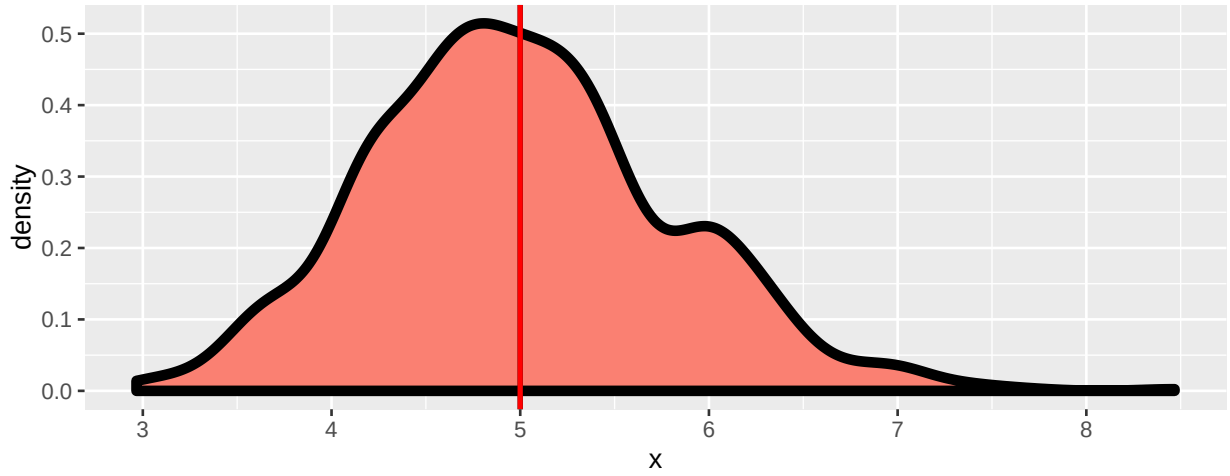
Sample Mean versus Theoretical Mean

As can be seen in Figure 1, the distribution of sample means is centered very close to the theoretical population mean of 5. I included two separate vertical lines to show the theoretical population (black) mean along with the sample mean (red). The values are so close to each other, the lines look almost indistinguishable. This shows that the sample size (1,000) is sufficient for a consistent estimator of the population mean according to the Law of Large Numbers.

```
## [1] "Sample mean: 4.9997019268744"
```

```
## [1] "Theoretical mean: 5"
```

Figure 1: Sample Mean Consistent Estimator of Population Mean



Sample Variance versus Theoretical Variance

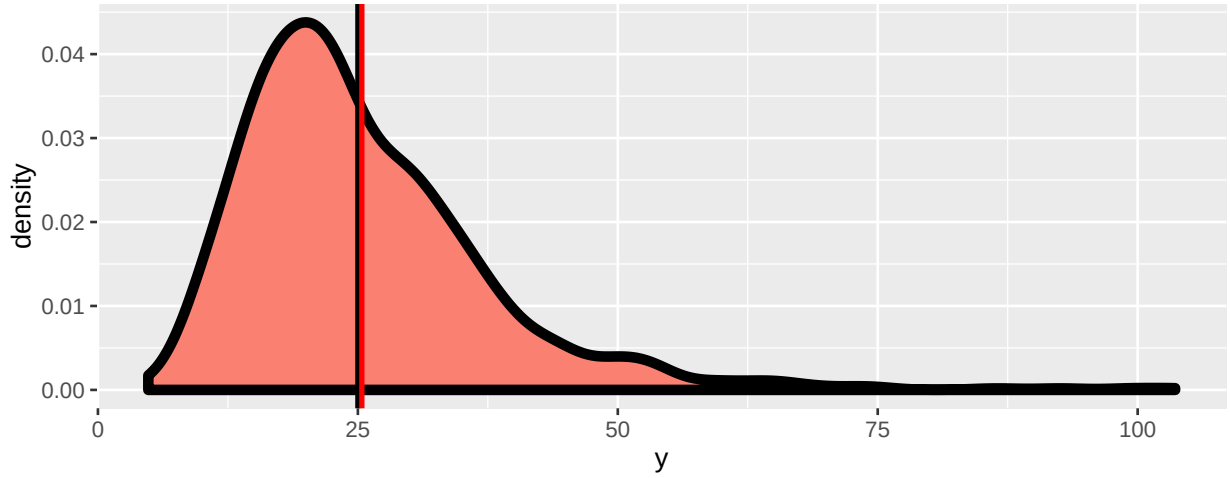
For Figure 2, I plot the distribution of 1,000 sample variances. The theoretical variance is `(1/lambda)^2 = 25`. As can be seen in Figure 2, the distribution of sample variances is centered very close to the population variance of 25. I included two separate vertical lines to show the theoretical population (black) variance along with the sample variance (red). The values are so close to each other, the lines look almost indistinguishable. This similarly shows that the sample size (1,000) is also sufficient for a consistent estimator of the population variance according to the Law of Large Numbers.

It is important to note that the variances reported here are the variances for each of the sets of 40 random exponentials. Not the variance of the means of those same sets.

```
## [1] "Sample variance: 25.3828234274576"
```

```
## [1] "Theoretical variance: 25"
```

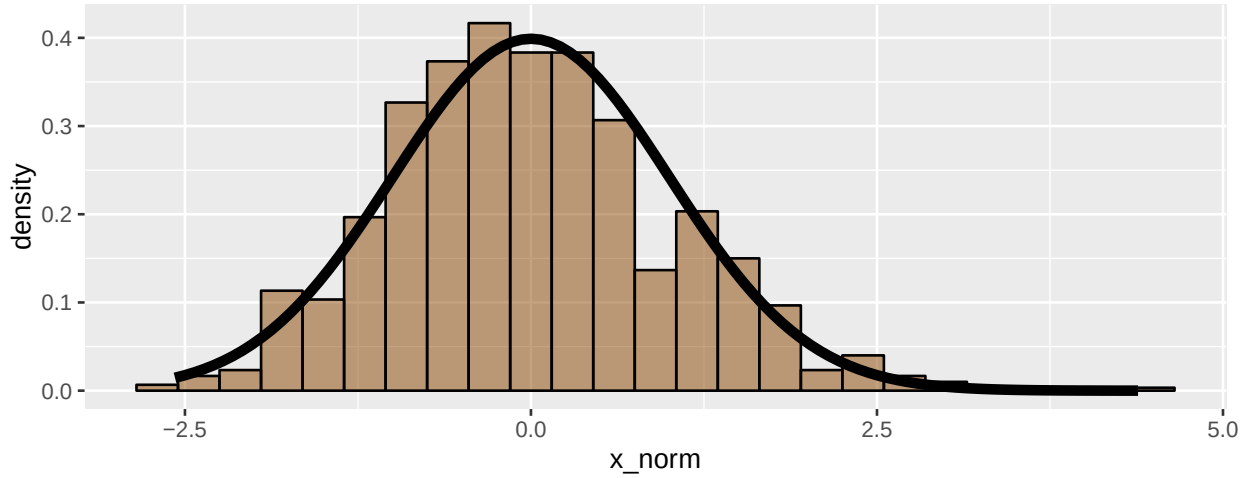
Figure 2: Sample Variance Consistent Estimator of Population Variance



Comparison of Sample Distribution with Standard Normal Distribution

According to the Central Limit Theorem, the distribution of iid variables when normalized (substracting the mean of the estimate and dviding by the standard error of the estimate) becomes that of a standard normal as the sample size increases. In Figure 3, I plot the normalized distribution of sample means and overlay the standard normal distribution. Using means of 40 randomized exponentials appears to approximate the standard normal distribution quite well.

Figure 3: Normalized Sample Distribution Follows Normal Distribution



Conclusion

As shown in the previous analysis the Central Limit Theorem holds when simulating 1,000 means for 40 exponential random variables. After noramalization, the distribution of means is approximately identical to the standard normal distribution.

Appendix

```
# Load ggplot2 package for graphics
library(ggplot2)

# Set random number generation seed
set.seed(100)

# Set rate equal to 0.2 for all simulations per instructions
lambda <- 0.2

# Mean and standard deviation of exponential distribution are equal to 1/lambda
mu <- 1/lambda
sigma <- 1/lambda
var = sigma ^ 2
```

```r
# Number of simulations for this exercise will be 1,000
nosim <- 1000

# Create matrix of 1,000 sets of 40 random exponentials
sim <- matrix(rexp(nosim * 40, lambda), nosim)

# Calculate mean and variance of 40 random exponentials, 1,000 times
dat <- data.frame(x = apply(sim, 1, mean),
                  y = apply(sim, 1, var))

# Simulated population mean is very close to our theoretical
paste("Sample mean:", mean(dat$x))

paste("Theoretical mean:", mu)

# Plot the sample distribution of 1,000 means of 40 random exponentials
# vertical lines represent theoretical mean (black) and sample variance (red)
# The two lines almost sit directly on top of each other, the simulation confirms
# that the sample size is large enough to provide a consistent estimator of the
# population mean.
g1 = ggplot(dat, aes(x))
g1 = g1 + geom_density(size = 2, fill = "salmon")
g1 = g1 + geom_vline(xintercept = mu, colour = "black", size = 1)
g1 = g1 + geom_vline(xintercept = mean(dat$x), colour = "red", size = 1)
g1 + ggtitle("Figure 1: Sample Mean Consistent Estimator of Population Mean")

# Sample variance is very close to our theoretical
paste("Sample variance:", mean(dat$y))

paste("Theoretical variance:", var)

# Sample distribution of 1,000 variances of 40 random exponentials
# vertical lines represent theoretical variance (black) and sample variance (red)
# The two lines almost sit directly on top of each other, the simulation confirms
# Our suspected theoretical variance.
g2 = ggplot(dat, aes(y))
g2 = g2 + geom_density(size = 2, fill = "salmon")
g2 = g2 + geom_vline(xintercept = var, colour = "black", size = 1)
g2 = g2 + geom_vline(xintercept = mean(dat$y), colour = "red", size = 1)
g2 + ggtitle("Figure 2: Sample Variance Consistent Estimator of Population Variance")

# Normalize distribution of averages to compare with standard normal
normal_dist <- data.frame(x_norm = ((dat$x - mean(dat$x))/sigma) * sqrt(40))

# Normalized distribution of means fits the standard normal distribution quite
# well. This is a direct reflection of the Central Limit Theorem: the distribution
# of averages of iid variables (properly normalized) becomes that of a standard
# normal as the sample size increases.
g4 <- ggplot(normal_dist, aes(x = x_norm)) +
       geom_histogram(alpha = .50, binwidth = .3, fill = "darkorange4",
                      colour = "black", aes(y = ..density..))
g4 <- g4 + stat_function(fun = dnorm, size = 2)
g4 + ggtitle("Figure 3: Normalized Sample Distribution Follows Normal Distribution")
```