Part 1: Simulation Exercise - A Comparison of the Exponential Distribution in R with the Central Limit Theorem

Yohance Nicholas

3/15/2020

Overview

Asymptotics are an important topics in statistics. Asymptotics refers to the behaviour of estimators as the sample size goes to infinity. Our very notion of probability depends on the idea of asymptotics. Central to probability theory is the Central Limit Theorem (CLT). The central limit theorem establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed. This helps us create robust strategies for creating statistical inferences when we're not willing to assume much about the generating mechanism of our data.

The first part of this coursework project challenges candidates to put their skills learned into practice in comparing the Exponential Distribution in R with the Central Limit Theorem. In order to do this, candidates must perform simulations using a Lambda of 0.2 and comparing it to the distribution of averages of 40 exponentials with 1,000 simulations.

Simulations

In order to perform the required simulations, one must first set values for the following simulation variables:

- The number of simulations to be performed,
- The Seed,
- The Lambda value, and
- the number of exponentials.

```
set.seed(868)
lambda <-  0.2
exponentials <-  40
simulations <- 1:1000
```

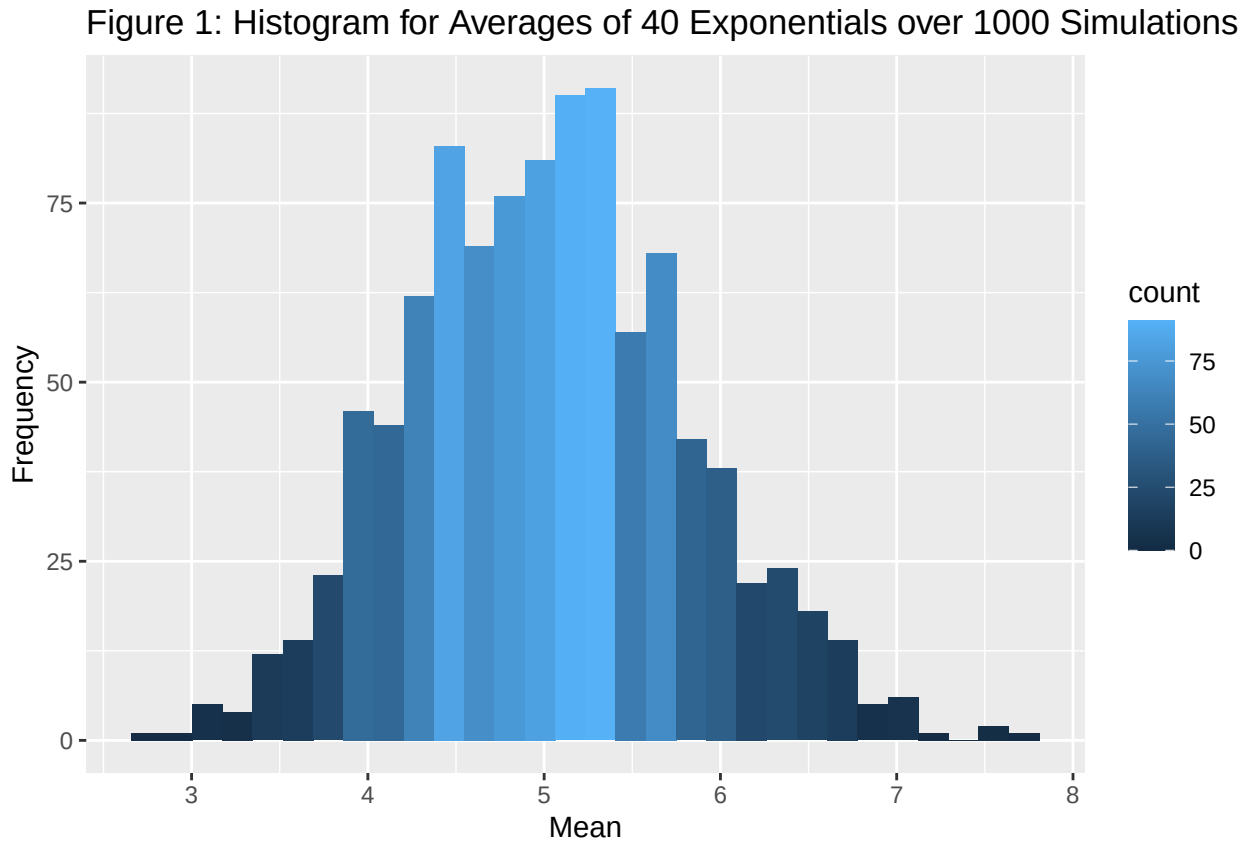
With these arguments defined, one can run the Simulations

```
simulated_population <- data.frame(x = sapply(simulations,
                                               function(x) {mean(rexp(exponentials, lambda))}))
```

The averages of 40 Exponentials over 1,000 Simulations are depicted with the assistance of the histogram in Figure 1 below.

```
library(ggplot2)
ggplot(simulated_population, aes(x=x)) +
  geom_histogram(aes(y=..count.., fill=..count..)) +
  labs(title="Figure 1: Histogram for Averages of 40 Exponentials over 1000 Simulations",
       y="Frequency",
       x="Mean")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Sample Mean versus Theoretical Mean

Sample Mean

Calculating the mean from the simulations with give the sample mean.

```
sample_mean <- mean(simulated_population$x)
sample_mean
```

```
## [1] 5.025526
```

Theoretical Mean

The theoretical mean of an exponential distribution is lambda^-1.

```
theoretical_mean <- lambda^-1
theoretical_mean
```

```
## [1] 5
```

Comparison

Only marginal differences exist between the simulations sample mean and the exponential distribution theoretical mean.

```
cbind(sample_mean, theoretical_mean)
```

```
##      sample_mean theoretical_mean
## [1,]    5.025526                5
```

```
t.test(simulated_population$x)[4]
```

```
## $conf.int
## [1] 4.976442 5.074610
## attr(,"conf.level")
## [1] 0.95
```

Sample Variance versus Theoretical Variance

Sample Variance

Calculating the variance from the simulation means with give the sample variance.

```
sample_variance <- var(simulated_population$x)
sample_variance
```

```
## [1] 0.625648
```

Theoretical Variance

The theoretical variance of an exponential distribution is (lambda * sqrt(n))^-2)

```
theoretical_variance <- (lambda * sqrt(exponentials))^-2
theoretical_variance
```

```
## [1] 0.625
```

Comparison

There is only a slight difference between the simulations sample variance and the exponential distribution theoretical variance.

```
cbind(sample_variance, theoretical_variance)
```

```
##      sample_variance theoretical_variance
## [1,]        0.625648                0.625
```

```
abs(var(simulated_population$x)-(lambda * sqrt(exponentials))^-2)
```

```
## [1] 0.000647969
```

Distribution

With a view to compare the theoretical and sample distributions, Figure 2 depicts the density histogram of 1,000 simulations with an overlay of the theoretical with an overlay of the normal distribution that has a mean of lambda^-1 and standard deviation of lambda*sqrt(n))^-1, the theoretical normal distribution for the simulations.

```
distribution_plot <- ggplot(simulated_population, aes(x=x)) +
  geom_histogram(aes(y=..density.., fill=..density..)) +
  labs(title="Figure 2: Histogram of Averages of 40 Exponentials over 1,000 Simulations",
       y="Density",
       x="Mean") +
  geom_density(colour="blue") +
  geom_vline(xintercept = sample_mean,
             colour="blue",
             linetype="dashed") +
  geom_vline(xintercept = theoretical_mean,
             colour="red",
             linetype="dashed") +
  stat_function(fun = dnorm,
                args=list(mean = 1/lambda,
                          sd= sqrt(theoretical_variance)),
                color = "red")
distribution_plot
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Figure 2: Histogram of Averages of 40 Exponentials over 1,000 Simulations