Statistical Inference Course Project (Part 1)

Jeffrey M. Hunter

10 May, 2019

Contents

Overview

The Central Limit Theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population (generally sample sizes greater than 30), then the distribution of the sample means will be approximately normally distributed about the population mean μ - no matter the shape of the population distribution.

This project explores the Central Limit Theorem using the exponential distribution in R. The theoretical normal distribution will be compared to the distribution of calculated means of samples from the exponential distribution.

Simulations

Perform 1000 simulations, each with 40 samples of an exponential distribution. The 40 samples will be used to calculate the arithmetic mean and variance and then compared to the theoretical estimates.

To make the data reproducible, a seed will be set. Also, set the control parameters $\lambda = 0.2$ (the rate) and $n = 40$ (number of samples). A histogram will be provided to show the averages of the 40 exponentials over 1000 simulations.

```
# load libraries
if (!require(ggplot2)) {
    install.packages("ggplot2", repos = "http://cran.us.r-project.org")
    library(ggplot2)
}
```

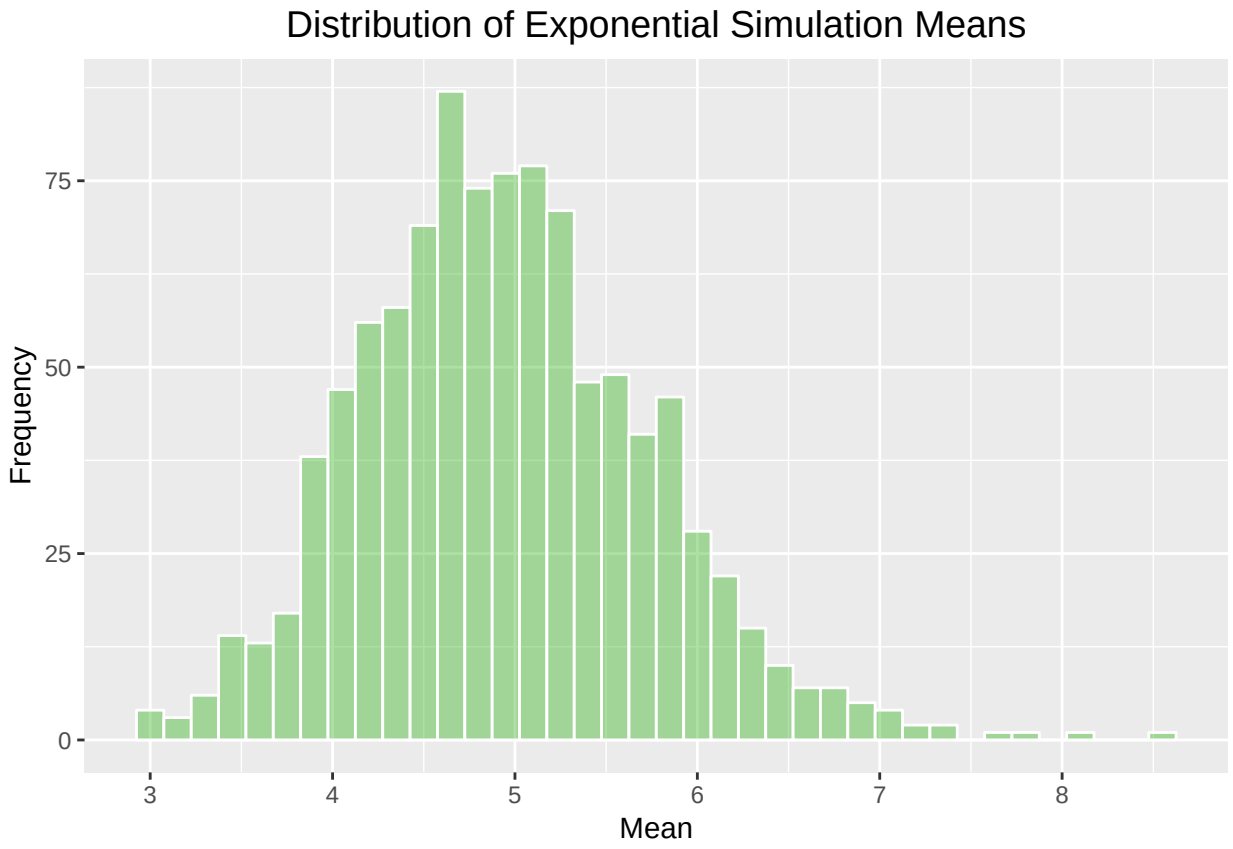
```
## Loading required package: ggplot2
```

```
# set seed for reproducability
set.seed(062000)

# set sampling values:
lambda <- 0.2             # rate parameter
n <- 40                   # number of samples (exponentials) in each simulation
numSimulations <- 1000    # number of simulations

# simulate the population
simMeans <- data.frame(expMean = sapply(1 : numSimulations, function(x) {mean(rexp(n, lambda))}))
```

```r
# plot the distribution
expSimulationMeansChart <- ggplot(simMeans, aes(x = expMean, y = ..count..)) +
    geom_histogram(binwidth = 0.15, color = "white", fill = rgb(0.2,0.7,0.1,0.4))  +
    xlab("Mean") +
    ylab("Frequency") +
    theme(plot.title = element_text(size = 14, hjust = 0.5)) +
    ggtitle("Distribution of Exponential Simulation Means")
print(expSimulationMeansChart)
```



Sample Mean versus Theoretical Mean

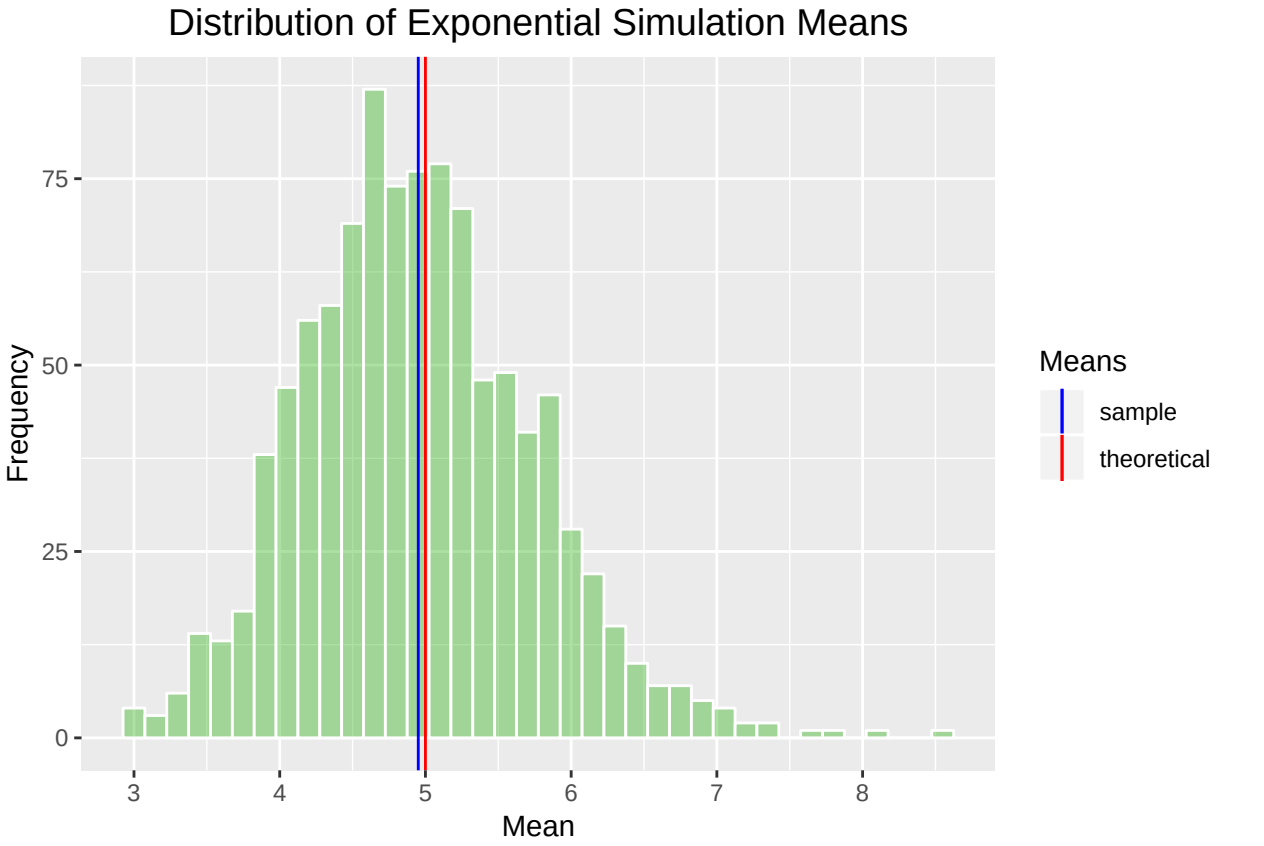
According to the Central Limit Theorem, the distribution of the sample means will be approximately normally distributed with a mean equal to the population mean μ of the underlying distribution. Because the underlying distribution in this simulation is exponential, the theoretical mean of the exponential distribution will be compared to the corresponding sample mean of the simulation. For an exponential distribution, the theoretical mean is equal to $\frac{1}{\lambda}$.

Calculate the sample mean and theoretical mean across all 1000 simulations of 40 samples from an exponential distribution where $\lambda = 0.2$.

```r
# calculate sample mean and theoretical mean
sampleMean <- mean(simMeans$expMean)
theoMean <- 1/lambda
compMeans <- data.frame(sampleMean, theoMean)
names(compMeans) <- c("Sample Mean", "Theoretical Mean")
print(compMeans)
```

```
##   Sample Mean Theoretical Mean
## 1    4.950877                5
```

```r
# plot the distribution (sample mean versus theoretical mean)
expSimulationMeansChart <- ggplot(simMeans, aes(x = expMean, y = ..count..)) +
    geom_histogram(binwidth = 0.15, color = "white", fill = rgb(0.2,0.7,0.1,0.4))  +
    geom_vline(aes(xintercept = sampleMean, color = "sample"), size = 0.50) +
    geom_vline(aes(xintercept = theoMean, color = "theoretical"), size = 0.50) +
    xlab("Mean") +
    ylab("Frequency") +
    theme(plot.title = element_text(size = 14, hjust = 0.5)) +
    scale_color_manual(name = "Means", values = c(sample = "blue", theoretical = "red")) +
    ggtitle("Distribution of Exponential Simulation Means")
print(expSimulationMeansChart)
```



The sample mean came out to be 4.9508767 while the theoretical mean is 5. As shown in the above chart, the mean of the sample means of exponentials (blue vertical line) is very close to the theoretical mean of an exponential distribution (red vertical line).

Sample Variance versus Theoretical Variance

In the same manner used to compare the Sample Mean and Theoretical Mean, the Sample Variance will be compared to the Theoretical Variance.

The theoretical variance is $\frac{(\frac{1}{\lambda})^2}{n}$.

```r
# calculate sample variance and theoretical variance
sampleVariance <- var(simMeans$expMean)
```

```r
theoVariance <- ((1/lambda)^2)/n
compVariance <- data.frame(sampleVariance, theoVariance)
names(compVariance) <- c("Sample Variance", "Theoretical Variance")
print(compVariance)
```

```
##   Sample Variance Theoretical Variance
## 1       0.6222257                0.625
```

The sample variance came out to be 0.6222257 which is very close to the theoretical variance 0.625.

Distribution