Examining the Central Limit Theorem using the Exponential Distribution

Jack Gidding

December 22, 2015

Overview

This project uses simulation to explore the exponential distribution in R and the Central Limit Theorem (CLT). There are two segments of the project. First, simulate randow draws from the exponential distribution. Second, perform inferential data analysis on the simulation data and compare to the Normal Distribution from the CLT. This report will look at the sample mean and sample variance versus the theoretical mean and theoretical variance for the exponential distribution.

The probability distribution function (PDF) for the exponential distribution is: $\mathrm{P(x)} = \lambda\ \mathrm{e}^{-\lambda \mathrm{t}}$. The first moment, mean, is $\mu = 1/\lambda$. The second moment, variance, is also $\sigma^2 = 1/\lambda$.

Libraries used in this experiment are loaded.

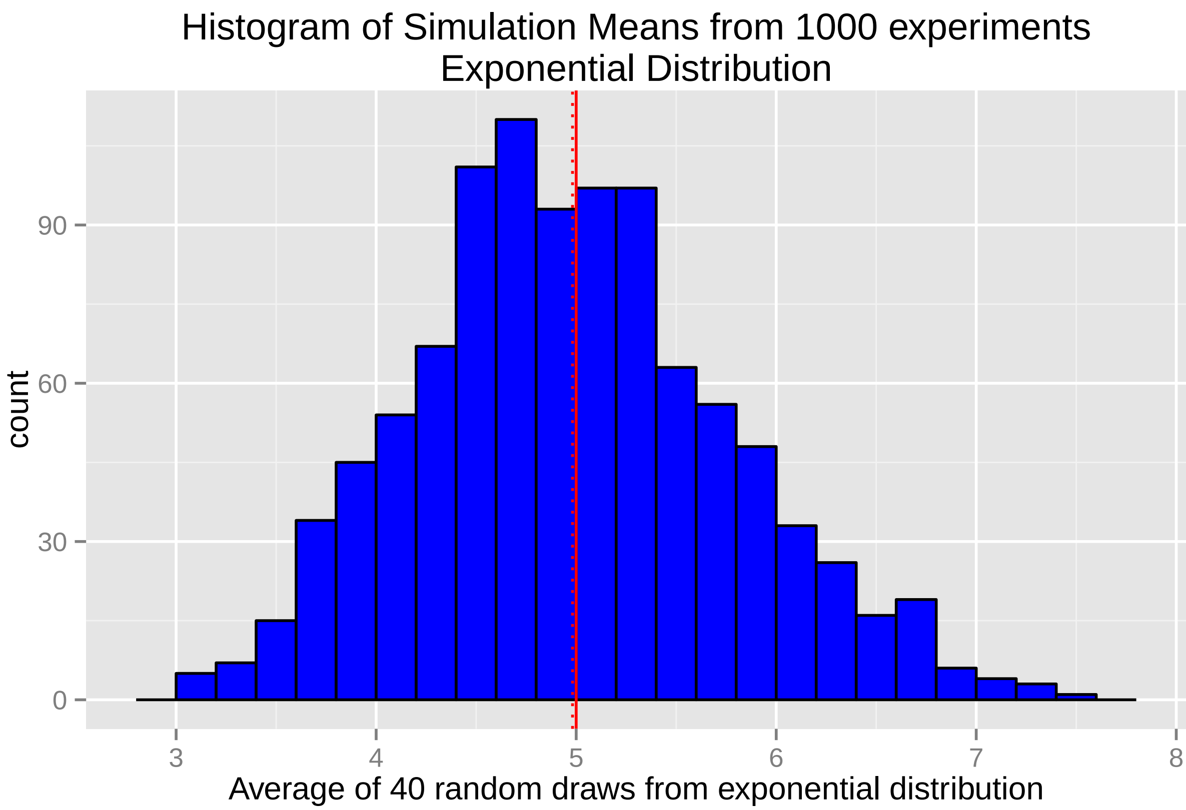
```
library(plyr)
library(ggplot2)
```

Simulations

```
simulations <- 1000
draws <- 40
lambda <- 0.2
```

For the experiment, we run 1000 simulations of 40 draws each from the exponential distribution using a lambda of 0.2 for the exponential distribution. The experiment is run. From each simulation, the mean of 40 draws is calculated and collected for all simulations. This is our sample population. A data frame is created with the means of each simulation. A histogram of the means is plotted to examine the distribution of the sample means.

```
simdata <- apply(matrix(rexp(simulations * draws, lambda), simulations), 1, mean)
simdata <- data.frame(simdata)
plotfn(simdata, lambda)
```



Sample Mean versus Theoretical Mean

```
# Calculate mean of all experiments (x bar)
simdata.mean <- mean(simdata$simdata)

# Calculate the theoretical mean (mu)
expdist.mean <- 1/lambda
```

In the diagram above, the red solid line shows the theoretical mean of the exponential distribution and the red dotted line shows the mean of the sample population. The theoretical mean of the exponential distributions, μ, is $\mu = 1/\lambda = 5$. The mean of the sample population, $\bar{x} = 4.9817912$. The difference between the two values is very small: μ - $\bar{x} = 0.0182088$. The percentage difference between the two values is 0.3641766%.

Sample Variance versus Theoretical Variance

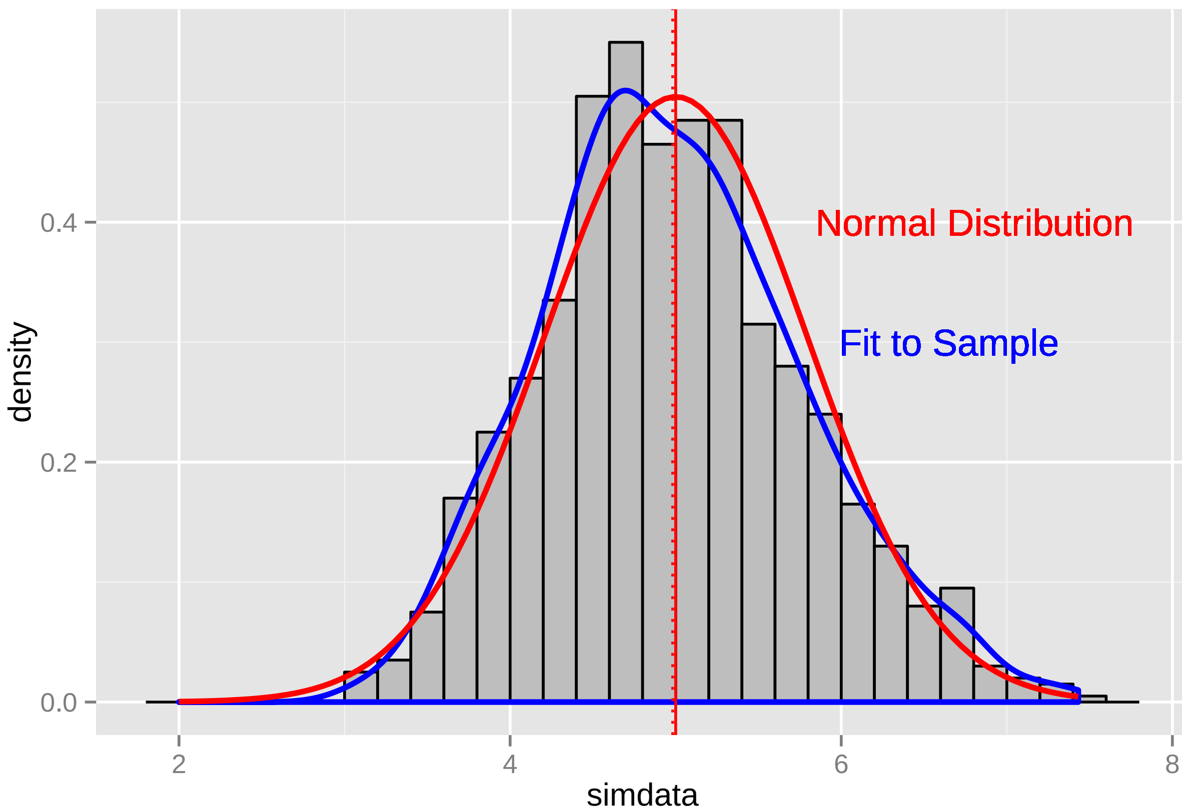
```
# Calculate mean of all experiments (s^2)
simdata.var <- var(simdata$simdata)

# Calculate the theoretical variance (E[x bar] = sigma^2/n)
expdist.var <- expdist.mean^2/draws
```

The theoretical variance of the means of draws from the exponential distribution, $\mathrm{Var}(\bar{x}) = (1/\mathrm{n})\ \lambda^{-2} = 0.625$. The variance of the sample population, $\mathrm{s}^2 = 0.6189741$. The difference between the two values is very small: $\mathrm{Var}(\bar{x})$ - $\mathrm{s}^2 = 0.0060259$. The percentage difference between the two values is 0.96415%.

Distribution

A histogram of the simlation means is again plotted to examine the distribution. In the plot below, the histogram was adjusted for y to be between 0 and 1. Scaling the y-axis enabled an overlay of the Normal Distribution N(5,0.625) in red as defined by the Central Limit Theorem. A curve is fit to the sample distribution and overlayed in blue. This plot shows that the distribution is approximately normal as stated by the CLT.

```
plotdist(simdata, lambda, expdist.mean, expdist.var)
```



Appendix

The code for the two plotting functions was hidden in the document above to keep the length of the core analysis to three pages (minus appendix). The code for the two functions is presented here.

Function to plot a histogram of the samples with the theoretical exponential distribution mean and sample mean.

```
plotfn <- function(vals, lmd) {
    plot1 <- ggplot(simdata, aes(x = simdata)) +
        geom_histogram(binwidth = 0.2, color="black", fill="blue") +
        labs(title=paste("Histogram of Simulation Means from 1000 experiments\n",
            "Exponential Distribution",sep="")) +
        labs(x="Average of 40 random draws from exponential distribution") +
    geom_vline(xintercept=mean(vals$simdata), linetype="dotted", color="red") +
    geom_vline(xintercept=1/lmd, color="red")
    print(plot1)
}
```

Function to plot a histogram of the samples with an overlayed curve fit to the histogram and Normal Distribution.

```
plotdist <- function(dst, lmd, nmean, nvar) {
    plot2 <- ggplot(dst) +
        geom_histogram(aes(x = simdata, y=..density..), binwidth = 0.2,
                       color="black", fill="grey") +
        geom_density(aes(x = simdata), color="blue", size=1) +
        stat_function( fun=dnorm ,arg=list(mean=nmean, sd=sqrt(nvar)),
                                            color = "red", size=1) +
        geom_vline(xintercept=mean(dst$simdata), linetype="dotted",
                   color="red") +
        geom_vline(xintercept=1/lmd, color="red") +
        geom_text(aes(2,0.4,label = "Normal Distribution", hjust = -2), color="red") +
        geom_text(aes(2,0.3,label = "Fit to Sample", hjust = -3),
                  color="blue")


    print(plot2)
}
```