Does the Distribution of Means of 40 Exponentials Behave as Predicted by the Central Limit Theorem

Luke Coughlin

Synopsis

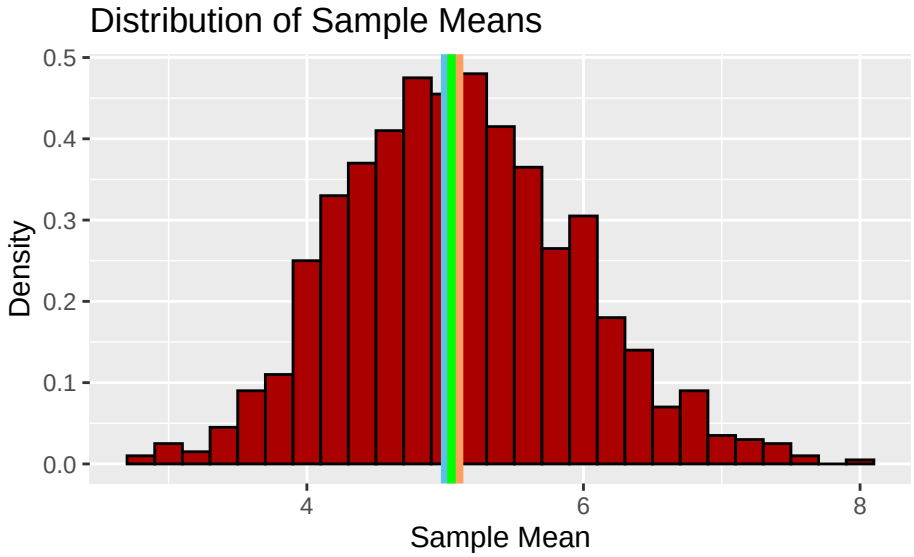
This simulation looks to test the central limit theorem on an exponential distribution. The central limit theorem states that the distribution of sample means of many large enough samples of a distribution will be approximately normal, even if the population is not normally distributed. This simulation upholds the central limit theorem by having a distribution of means that is approximately normal.

Simulations

We'll now run 1000 simulations of 40 random variables from an exponential distribution to compare the distribution of the sample means and standard distributions later. We'll set the seed so the simulation is reproducible, arbitrarily choosing to use the first phi-prime as our seed. After generating both sets of observations we'll store the means and standard deviations of each simulation within our large collection of simulations.

Sample Mean versus Theoretical Mean

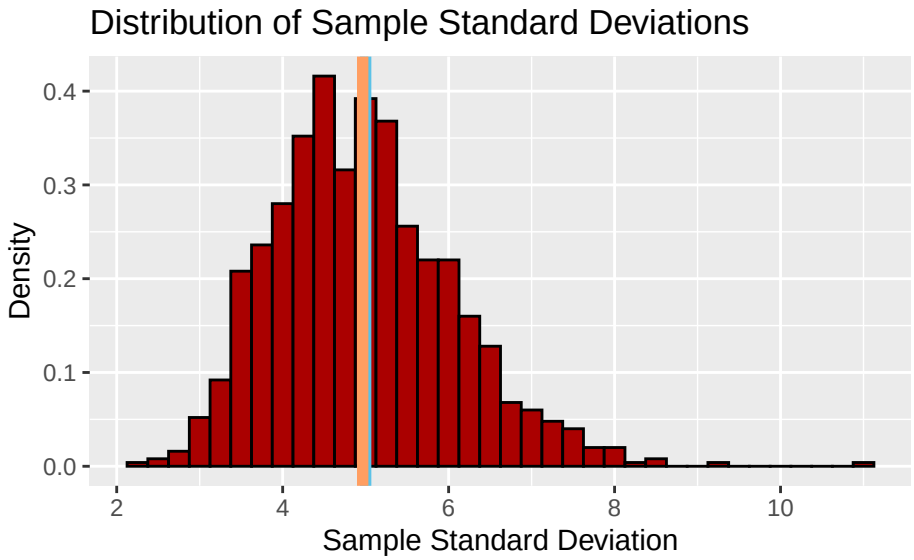
The theoretical mean of an exponential distribution is $\frac{1}{\lambda}$. In our simulation we set $\lambda = 0.2$. As such the theoretical mean is $\mu_T = 5$. We'll now look at how the distribution of the sample means compares to μ_T



In this histogram of the means the blue line represents μ_T, 5, the orange line represents the mean of the sample means, 5.1, and the green line represents the sample's median, 5.05. This shows that the mean of 1000 simulations of 40 samples from an exponential distribution closely resembles the true theoretical mean. While the sample median and mean are close, an attribute of normally distributed data.

Sample Variance versus Theoretical Variance

The theoretical mean of an exponential distribution is $\frac{1}{\lambda}$. In our simulation we set $\lambda = 0.2$. As such the theoretical standard deviation is $\sigma_T = 5$. We'll now look at how the distribution of the standard deviations compares to σ_T.



In this histogram of the standard deviations the blue line represents σ_T, 5 and the orange line represents the mean of the standard deviations, 4.96. These values are so close that it reasonable to conclude the sample standard deviation accurately represents the true standard deviation.

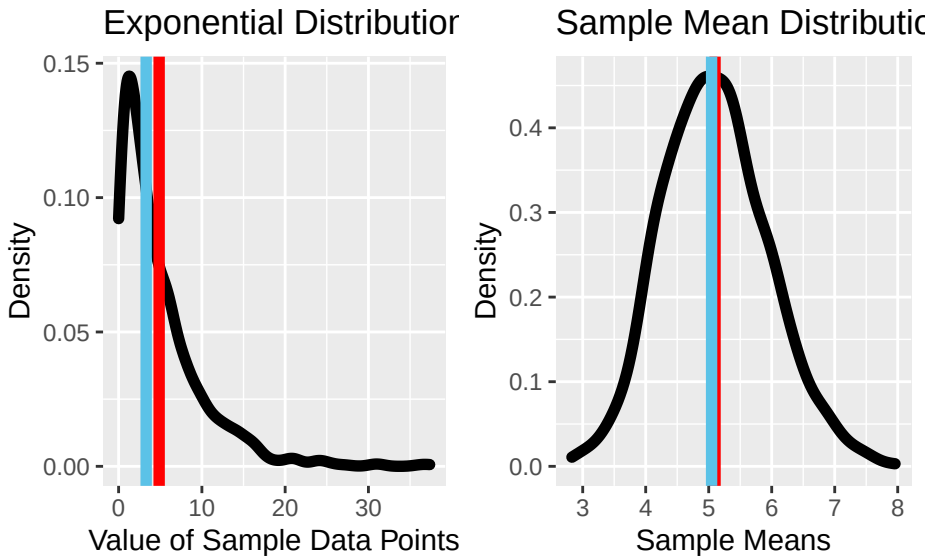
There seems to be an outlier that's greater than 10, this is likely due to an anomaly in sampling the exponential distribution. Let's look at that sample set now.

```
##  [1]  1.1115223 69.5594991  2.5845742  4.1456691  0.9881980  0.5764390
##  [7] 10.8931832 15.0100879  5.7394382  6.7351328  5.6811038  3.2551634
## [13]  2.9895914  1.7589532  1.1060825  1.1251588  0.2900541  1.2472671
## [19] 10.5095757  2.2223170  5.0132533  2.6577173  7.1168531  2.3240154
## [25]  1.9439511  3.7382340 10.2663309  2.2563026  3.0793852  1.1044089
## [31]  4.5625495  9.5708496  6.3844310 17.5178723  5.6246395  4.6515142
## [37]  2.4609859  0.9532821  0.2766272  0.2400065
```

It can be seen in this list that the large deviation likely came from having the large value, 69.559, in row 556's data. Since this is a simulation there is no reason to discard this data point.

Distribution

The final aspect we'll be looking at is how this distribution of means compares to a random distribution of many (1000) samples from an exponential distribution.



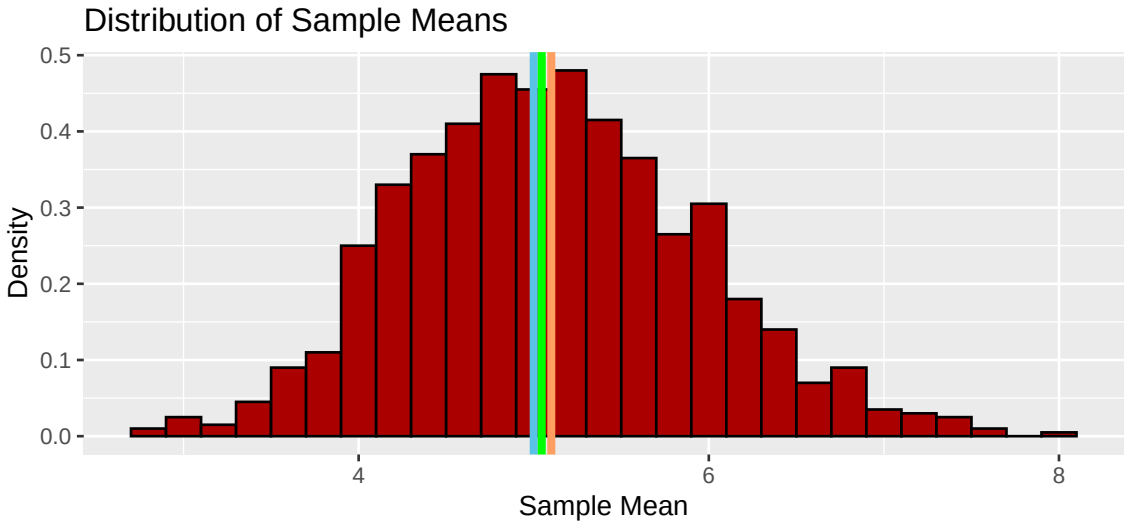
In the above graphs the blue line represents the median and the red line is the mean of each respective data set. In the exponential distribution it can be seen the values are pretty far off (1.6), indicating that the data is skewed right. On the contrary, the sample means have nearly overlapping means and medians, suggesting the data is approximately normal. This is what the central limit theorem proposes, that given a large enough sample size, the distribution of the sample means will be approximately normal. Therefore this simulation has upheld the central limit theorem.

Appendix

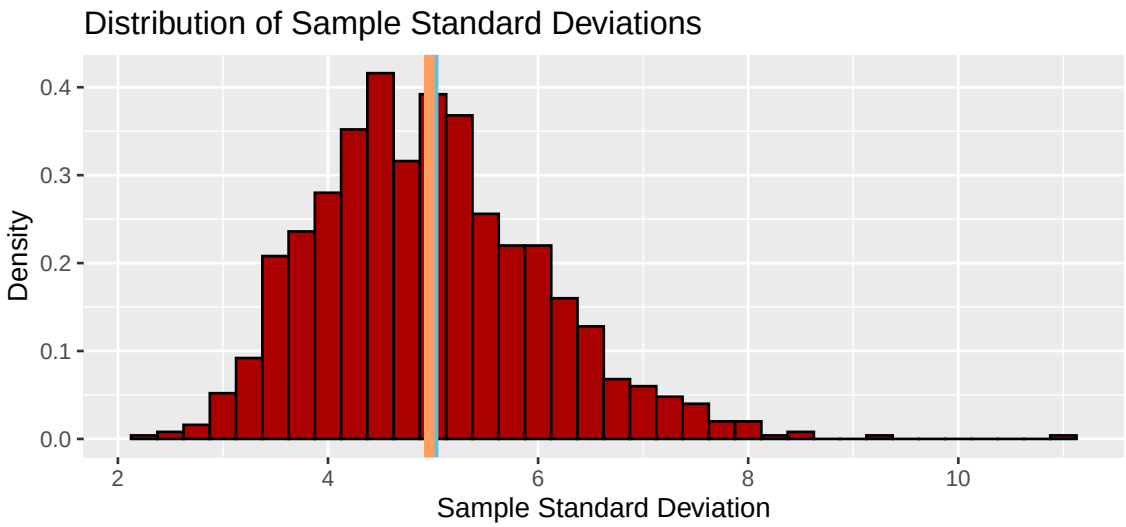
Luke Coughlin

5/30/2020

```
## ----setup, include=FALSE-------------------------------------------------------
knitr::opts_chunk$set(echo = FALSE)


## ----Load_Libraries, message = FALSE--------------------------------------------
library(tidyverse); library(gridExtra)
phi <- (1+sqrt(5))/2 #For dimensions


## ----Simulation-----------------------------------------------------------------
set.seed(1618033)
n <- 40
sims <- 1000
l <- 0.2 #lambda
multipleObs <- matrix(rexp(n*sims, l), sims, n)
means <- apply(multipleObs, 1, mean)
sds <- apply(multipleObs, 1, sd)


## ----Histogram_of_Sample_Means, fig.height= 3, fig.width = 3*phi----------------
muplot <- ggplot(data.frame(x = means), aes(x)) +
        geom_histogram(binwidth = 0.2, aes(y = ..density..),
                       fill = "#AA0000", colour = "#000000") +
        labs(x = "Sample Mean", y = "Density") +
        ggtitle("Distribution of Sample Means")
muplot + geom_vline(xintercept = 1/l, lwd = 1.5, colour = "#5BC2E7") +
        geom_vline(xintercept = mean(means), lwd = 1.5, colour = "#FF9E61") +
        geom_vline(xintercept = median(means), lwd = 1.5, colour = "#00FF00")
```



```
## ----Histogram_of_sds, fig.height= 3, fig.width = 3*phi-----------------------
plot <- ggplot(data.frame(x = sds), aes(x)) +
        geom_histogram(binwidth = 0.25, aes(y = ..density..),
                       fill = "#AA0000", colour = "#000000") +
        geom_vline(xintercept = 1/l, lwd = 2, colour = "#5BC2E7") +
        geom_vline(xintercept = mean(sds), lwd = 2, colour = "#FF9E61") +
        labs(x = "Sample Standard Deviation", y = "Density") +
        ggtitle("Distribution of Sample Standard Deviations")
plot
```




```
## ----Outlier_Row--------------------------------------------------------------
multipleObs[sds > 10,]
```

```
##  [1]  1.1115223 69.5594991  2.5845742  4.1456691  0.9881980  0.5764390
##  [7] 10.8931832 15.0100879  5.7394382  6.7351328  5.6811038  3.2551634
## [13]  2.9895914  1.7589532  1.1060825  1.1251588  0.2900541  1.2472671
## [19] 10.5095757  2.2223170  5.0132533  2.6577173  7.1168531  2.3240154
## [25]  1.9439511  3.7382340 10.2663309  2.2563026  3.0793852  1.1044089
## [31]  4.5625495  9.5708496  6.3844310 17.5178723  5.6246395  4.6515142
## [37]  2.4609859  0.9532821  0.2766272  0.2400065
```

```
## ----Exp_Dist-----------------------------------------------------------------
exsam <- rexp(sims,l)

## ----Plot_Densities, fig.height= 3, fig.width = 3*phi-------------------------
explot <- ggplot(data.frame(x = exsam), aes(x))+
        geom_density(lwd = 2) +
        labs(x = "Value of Sample Data Points", y = "Density") +
        ggtitle("Exponential Distribution")
g1 <- explot +
        geom_vline(xintercept = mean(exsam),
              colour = "#FF0000", lwd = 2) +
        geom_vline(xintercept = median(exsam),
              colour = "#5BC2E7", lwd = 2)

CLTplot <- ggplot(data.frame(x = means), aes(x)) +
        geom_density(lwd = 2) +
        labs(x = "Sample Means", y = "Density") +
        ggtitle("Sample Mean Distribution")
g2 <- CLTplot +
        geom_vline(xintercept = mean(means),
```

```
            colour = "#FF0000", lwd = 2) +
        geom_vline(xintercept = median(means),
            colour = "#5BC2E7", lwd = 2)

grid.arrange(g1, g2, ncol = 2)
```