The Exponential Distribution and The Central Limit Theorem

Michael Kroog

August 22, 2017

Overview

The following report will examine the exponential distribution in R and compare it with the Central Limit Theorem. In order to demonstate this comparison a simulation from the exponential distribution will be conducted and we will examine the relationship for the sample mean to the theoretical mean of the distribution. We will also examine the sample variance and the theoretical variance of the distribution.

```
set.seed(10)
## set lambda to 0.2, number of simulations to 1000 and number of random exponentials to 40
lambda = 0.2
n = 40
sim = 1000

## create an empty vector
meansimsamp <- NULL

## run 1000 simulations taking the mean of 40 random exponentials and assigning each to the vector
for (i in 1:sim) meansimsamp <- c(meansimsamp, mean(rexp(n, lambda)))

meansimsamp <- data.frame(meansimsamp)

1/0.2
```

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

```
meansimtheo <- sapply(meansimsamp, mean)
meansimtheo
```

```
## meansimsamp
##     5.04506
```

```
library(ggplot2)
```

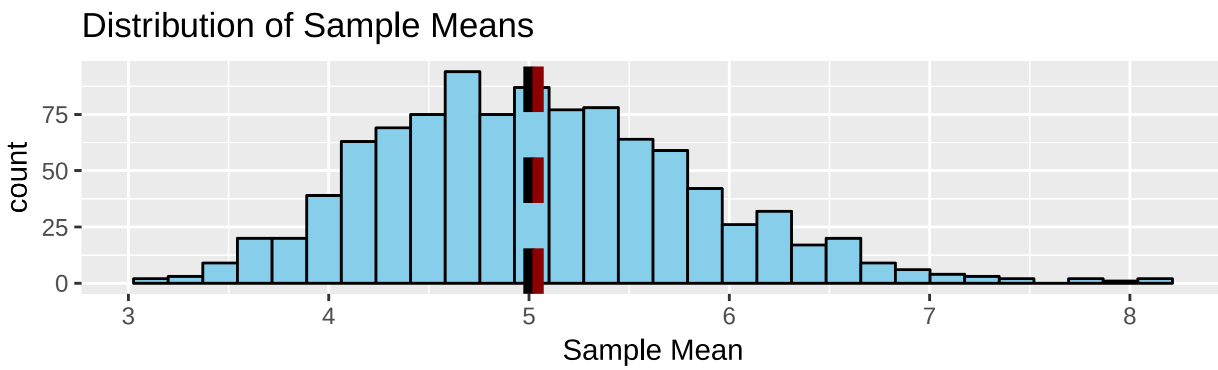
```
## Warning: package 'ggplot2' was built under R version 3.4.1
```

To demonstate how the CLT works for large enough sample sizes a simulation of 40 random samples from an exponential distribtuion will be taken 1000 times. These samples will then be averaged and plotted in a histogram. We will use a lambda of 0.2 during these simulations.

The R code above sets lamba to 0.2, the sample size to 40 and the number of simulations run to 1000. A for loop is run 1000 times of 40 random samples from the exponential distribution and averaged. The 1000 averages are then added as rows into a data frame.

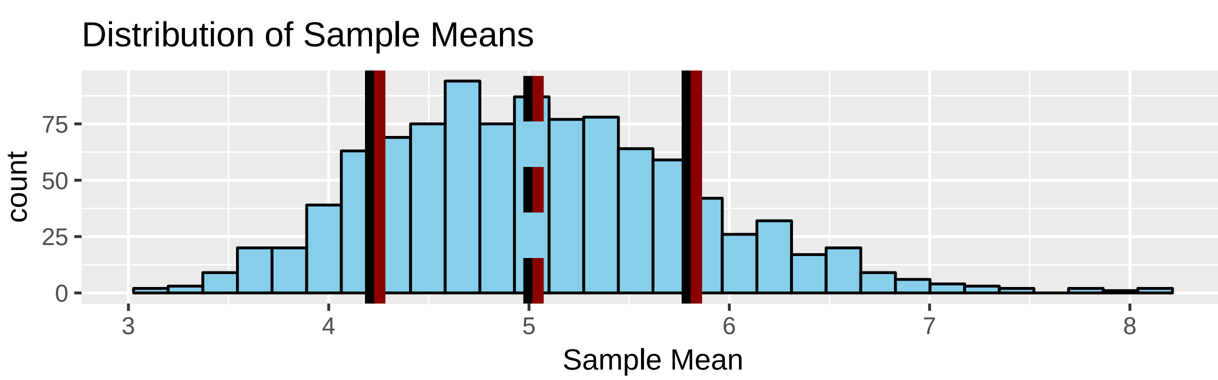
```
ggplot(meansimsamp, aes(x = meansimsamp)) + geom_histogram(bins = 30, fill = "sky blue",
                                            color = "black") +
                                            xlab("Sample Mean") +
                                            ggtitle("Distribution of Sample Means") +
                                            geom_vline(aes(xintercept = mean(5)), size = 2,
                                                       linetype = "dashed") +
```

```
                                                    geom_vline(color = "red4",
                                                    aes(xintercept = mean(mean(meansimsamp))), size = 2,
                                                    linetype = "dashed") +
                                                    scale_x_continuous(breaks = seq(1, 8, 1))
```



The histogram above shows the distribution of our simulation. The black dashed line represents the theoretical mean of 5 (1/0.2) and the red dashed line represents the sample mean of 5.045.

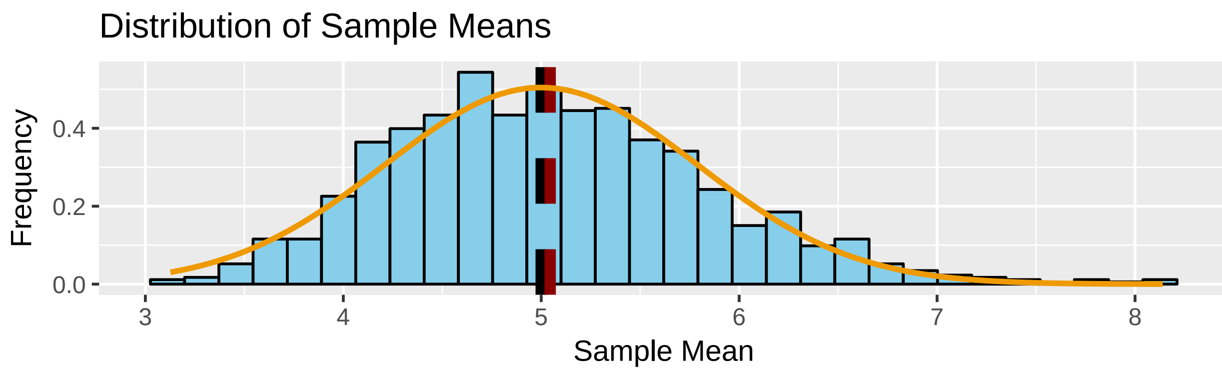
```
## sample standard deviation rounded to 3 decimal places
sdsimsamp <- round(sapply(meansimsamp, sd), 3)

## theoretical standard deviation
sdsimtheo <- sqrt((1/lambda^2)/n)

ggplot(meansimsamp, aes(x = meansimsamp)) + geom_histogram(bins = 30, fill = "sky blue",
                                                    color = "black") +
                                                    xlab("Sample Mean") +
                                                    ggtitle("Distribution of Sample Means") +
                                                    geom_vline(aes(xintercept = mean(5)), size = 2,
                                                              linetype = "dashed") +
                                                    geom_vline(xintercept = c(5 - sdsimtheo,
                                                    geom_vline(aes(xintercept = meansimtheo), size = 2,
                                                              linetype = "dashed", color = "red4") +
                                                    geom_vline(xintercept = c(meansimtheo - sdsimtheo,
                                                             color = "red4") +
                                                    scale_x_continuous(breaks = seq(1, 8, 1))
```



Continuing with the color theme the solid black bars represent the theoretical standard deviation and the solid red bars represent the sample standard deviation

The variance of all 1000 of the samples is calculated. The theoretical variance is also calculated. We can see from this particular experiment that the sample variance is 0.637 and the theoretical variance is 0.625.

```
ggplot(meansimsamp, aes(x = meansimsamp)) + geom_histogram(aes(y = ..density..), bins = 30,
                                                       fill = "sky blue", color = "black") +
                                            xlab("Sample Mean") +
                                            ylab("Frequency") +
                                            ggtitle("Distribution of Sample Means") +
                                            geom_vline(aes(xintercept = mean(5)), size = 2,
                                                      linetype = "dashed") +
                                            geom_vline(color = "red4",
                                            aes(xintercept = mean(mean(meansimsamp))), size = 2,
                                                       linetype = "dashed") +
                                            scale_x_continuous(breaks = seq(1, 8, 1)) +
                                            stat_function(fun = dnorm, args = list(mean = 5,
                                                       sd = sdsimtheo), color = "orange 2",
                                                       size = 1)
```



The histogram above now contains an orange curve representing a normal distribution. We can already see from the histogram that the distribution looks approximatley normal, but the overlay of the orange cruve helps exemplify our visualization of the CLT and how for large enough sample sizes the distribution becomes approximatley normal.