Exponential Distributions and Central Limit Theorem

Faye Gazave

February 20, 2015

Overview:

This paper presents a specific case study of the 'Central Limit Theorem'. In particular it illustrates how the exponential distribution with probability density function $f(x;\lambda) = \lambda e^{-\lambda x}$ is subject to the CTL.

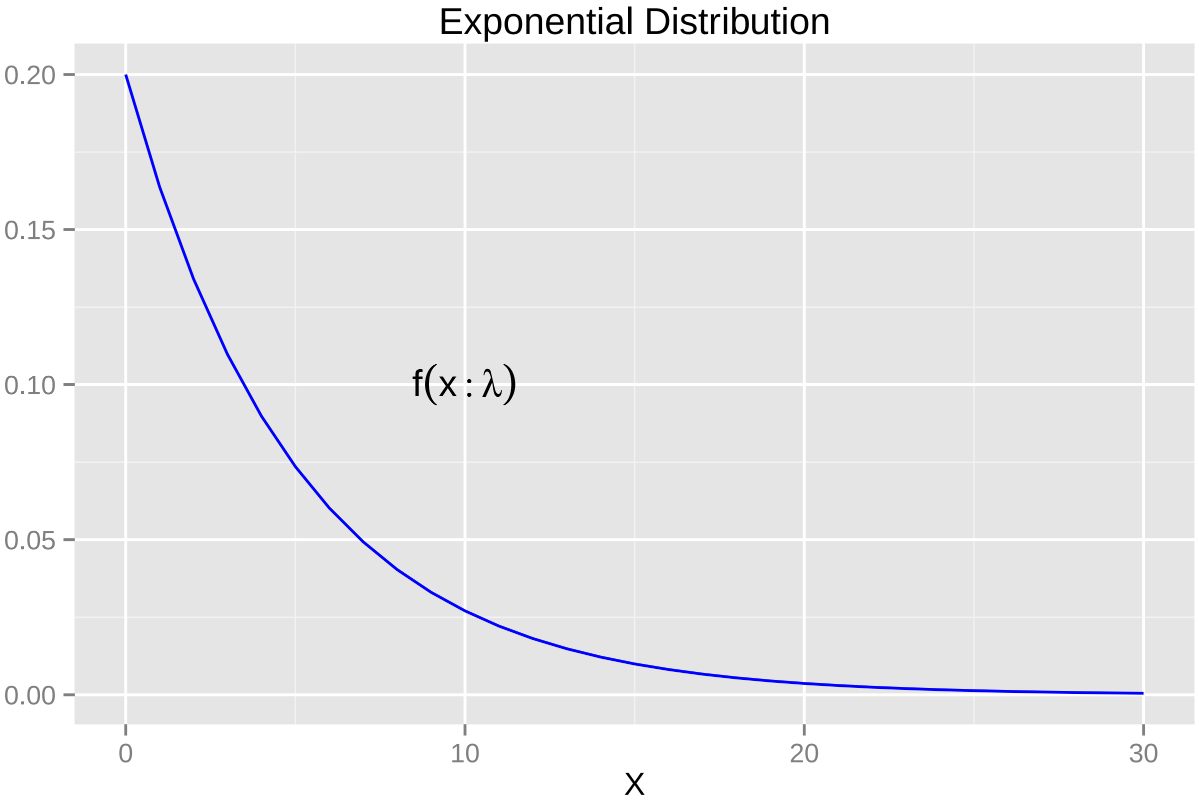
> Roughly, the central limit theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution. [1][2]

Simulations:

The exponential density function $f(x;\lambda)$ with parameters: $\mu, \sigma = 1/\lambda where \lambda = .2$ can be generated and plotted with the 'R' dexp function. It is shown here for reference.

Code for this plot and all other plots in the simulations are included in the references.

```
library(magrittr); library(ggplot2); suppressMessages(library(gridExtra))
exp_points <- dexp(seq(0, 30), rate = .2)
```

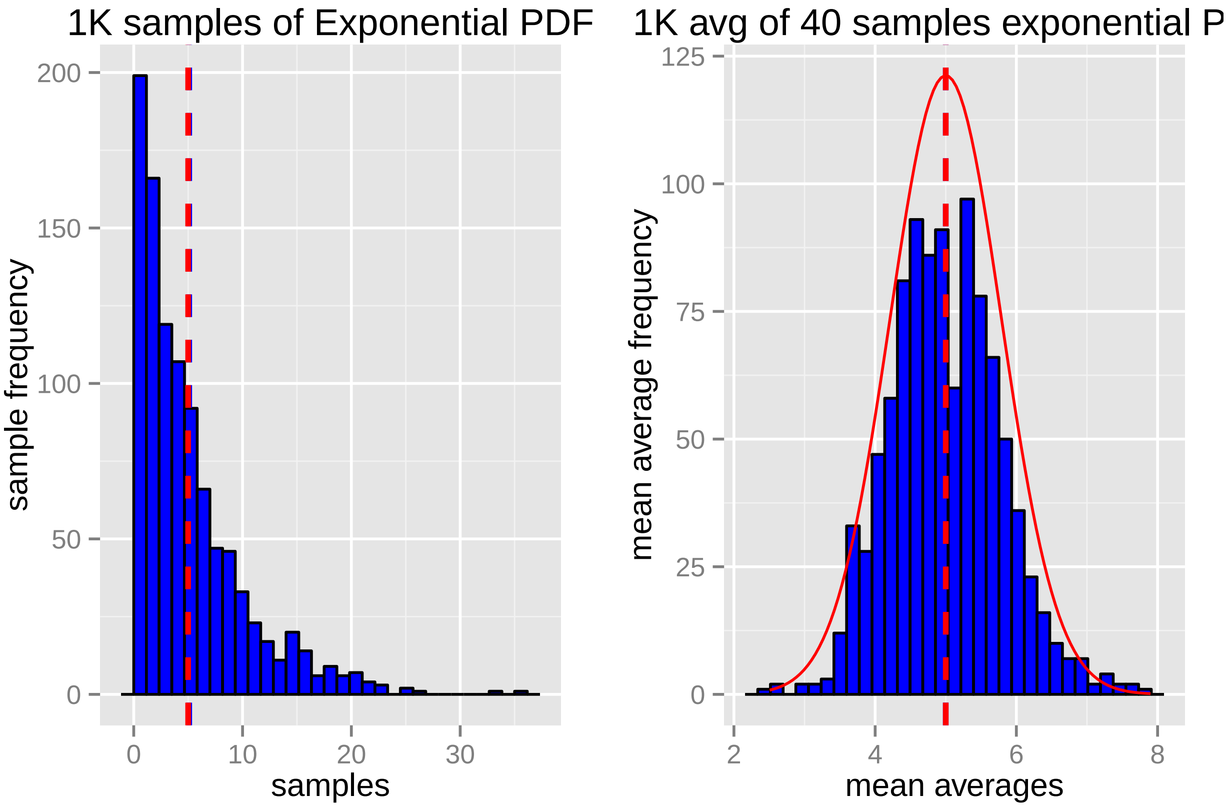


Let us now do some visual exploratory analysis to see how the distribution of the means follows the Central Limit Theorem.

First we generate 1000 samples from the exponential distribution and compare its histogram to the histogram of generating 1000 averages of 40 samples from the exponential distribution.

```
# Take 1000 samples from the exponential PDF.
large_sample <- rexp(1000, .2)
# Create 1000 simulations of 40 samples from exponential
# PDF and take the mean of each simulation i.e mean of the matrix row.
simulations <- apply(matrix(rexp(40000, .2), nrow=1000), 1, mean)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Sample Mean versus Theoretical Mean:

In both plots a blue vertical line is drawn at the respective sample means of 5.1000209 and 4.9995687 A dashed vertical red line at drawn at the theoretical mean $\mu = 1/\lambda = 1/.2 = 5$.

In both plots the blue line for the sample mean is hidden behind the red theoretical mean.

Sample Variance versus Theoretical Variance:

Variance for our first plot is var(large_sample) = 24.3762566 which is very close to our theoretical of $\sigma^2 = (1/\lambda)^2 = 25$.

Variance of our second plot is 0.6156852 which is **not** at all near the theoretical variance $(1/\lambda)^2$ of the given exponential, **nor is it expected to be.**

The important difference of the two plots is how the averages of the 40 samples in the second plot follows a normal distribution centered at the theoretical mean. Taking the square root of the variance = sd(simulations) = 0.7846561 it is shown to approximate the theoretical standard error of $\sigma^2/n = \sigma\sqrt(n) = 5\sqrt(40) =$ 0.7905694

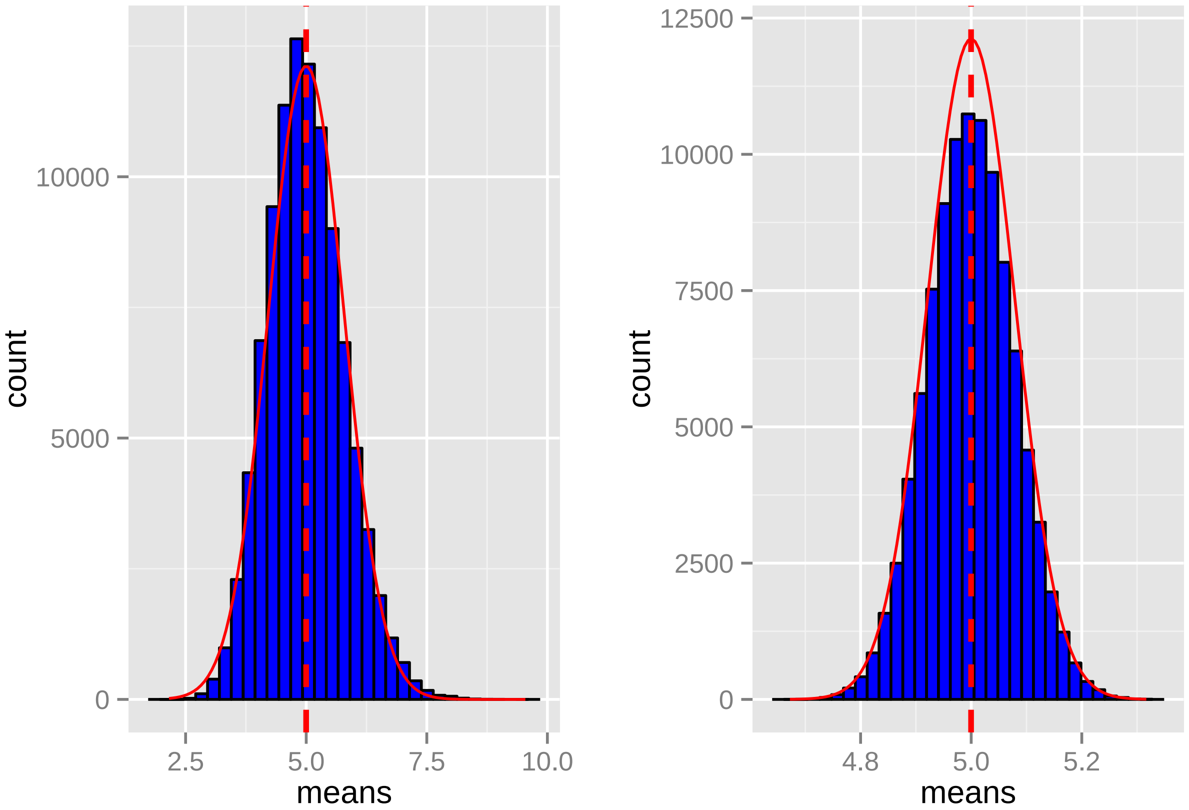
Let's explore the variation of the second plot a little more and see how the Central Limit Theorem applies to its variance.

The next two plots simulate taking even more averages and sample sizes. Our first plot below which will be called plot 3 simulates 100,000 averages of sample size 40. Note how the sample size of 40 is the same as in the plot above with 1000 averages.

The second plot below which we shall call plot 4 is also 100,000 averages, but the sample size is 4000 instead of 40.

```
#Pull 400000 samples from the probability distribution function.
#to create 100000 simulations of sammple size 40
lg_simulation <- apply(matrix(rexp(4000000, .2), nrow=100000), 1, mean)
#Pull 400000000 samples from the probability distribution function.
#to create 100000 simulations of sammple size 4000
ctl_simulation <- apply(matrix(rexp(400000000, .2), nrow=100000), 1, mean)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



By increasing the number of simulations from 1000 to 100,000, we see a distribution having shape much closer to normal distribution. In plot 3 there is slightly better mean 4.9986027, but **little change to standard deviation** 0.7892179 from plot 2 with only 1000 averages.

If we hold simulations to 100,000 and increase the sample size from 40 to 4000, we expect the standard error to be smaller i.e $\approxeq 5/\sqrt(4000) = 0.0790569$

Taking the mean of the 100,000 simulations with 4000 samples now has mean
4.9997636, with sample standard deviation $0.0787222 \approxeq 5/\sqrt(4000)$ Notice how much smaller our variance has become when we increase the number of samples n in a simulation. **As n goes to ∞ the variance will go to 0. Hence, convergence upon the true or theoretical μ of our exponential distribution.**

We see this born out in the 4th plot, by observing just how much smaller the variance is. Notice how at two standard deviations away we are still very close to the mean. Hence, the variance is diminishing.

Reference

Code plot of exponential distribution with given $\lambda = .2$.

```
exp_plot <- ggplot(data.frame(peaks = exp_points , divs=0:30),
                   aes(x=divs, y=peaks)) +
    labs(y=NULL, x='X') + geom_line(color='blue' ) +
    annotate("text", ,x = 10, y = .10, label=paste(
        'f(', expression(x), ':',  expression(lambda), ')' ), parse=TRUE ) +
    ggtitle('Exponential Distribution')
```

Code for exploratory plots 1 and 2.

```
plot1 <- ggplot(data.frame(means=large_sample), aes(x=means)) +
    geom_histogram(color='black', fill='blue') +
    # Indicate the sample mean with vertical line.
    geom_vline(aes(xintercept=mean(means)),
               color="blue", linetype="dashed", size=1) +
    # Theoretical mean with vertical line
    geom_vline(aes(xintercept=5),
               color="red", linetype="dashed", size=1) +
    ggtitle("1K samples of Exponential PDF") +
    labs(x = 'samples', y = 'sample frequency' )

plot2 <- ggplot(data.frame(means=simulations), aes(x=means)) +
    geom_histogram(color='black', fill='blue') +
    # Indicate the mean with vertical line.
    # Indicate the sample mean with vertical line.
    geom_vline(aes(xintercept=mean(means)),
               color="blue", linetype="dashed", size=1) +
    # Theoretical mean with vertical line
    geom_vline(aes(xintercept=5),
               color="red", linetype="dashed", size=1) +
    stat_function( fun = function(x)
        {6 * 40 * dnorm(x, mean=5, sd = .79)}, color='red') +
    labs(x = 'mean averages', y = 'mean average frequency' ) +
    ggtitle("1K avg of 40 samples exponential PDF")

#grid.arrange(plot1, plot2, ncol=2)
```

Code for exploratory plots 3 and 4.

```
plot3 <- ggplot(data.frame(means=lg_simulation), aes(x=means)) +
    geom_histogram(color='black', fill='blue') +
    # Indicate the mean with vertical line.
    geom_vline(
        aes(xintercept=mean(means)), color="red", linetype="dashed", size=1) +
    stat_function( fun = function(x){6 * 4000 * dnorm(x, mean=5, sd = .79)}, color='red')

plot4 <- ggplot(data.frame(means=ctl_simulation), aes(x=means)) +
    geom_histogram(color='black', fill='blue') +
    # Indicate the mean with vertical line.
    geom_vline(aes(xintercept=mean(means)), color="red",
               linetype="dashed", size=1) +
    stat_function( fun = function(x)
        {.6 * 4000 * dnorm(x, mean=5, sd = .079)}, color='red')
#Don't show in the reference section.
#grid.arrange(plot3, plot4, ncol=2)
```