Exponential Distributions and Central Limit Theorem

Faye Gazave

February 20, 2015

Overview:

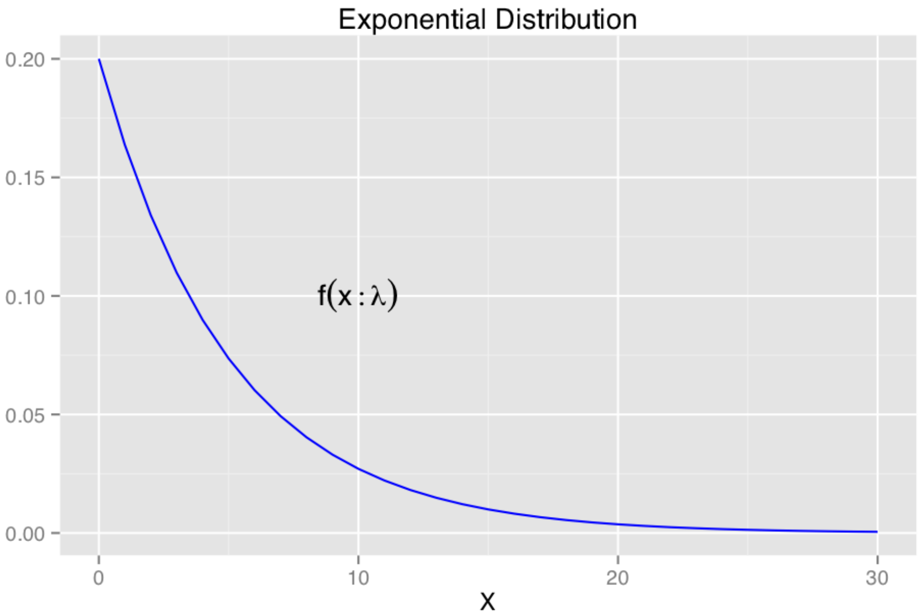
This paper presents a specific case study of the ‘Central Limit Theorem’. In particular it illustrates how the exponential distribution with probability density function $f(x;\lambda) = \lambda e^{-\lambda x}$ is subject to the CTL.

> 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]

Simulations:

We will first use the ‘R’ dexp function to simulate a ramdom sample, from the exponential density function $f(x;\lambda)$ with parameters: $\mu, \sigma = 1/\lambda where \lambda = .2$

```r
library(magrittr); library(ggplot2)
sim <- dexp(seq(0, 30), rate = .2)
ggplot(data.frame(peaks = sim , 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')
```



Now we can look at 40 simulations of random samples having a 1000 samples each.

```
#Pull 40000 samples from the probability distribution function.
simulations <- apply(matrix(rexp(40000, .2), nrow=1000), 1, mean)

head(simulations)
```

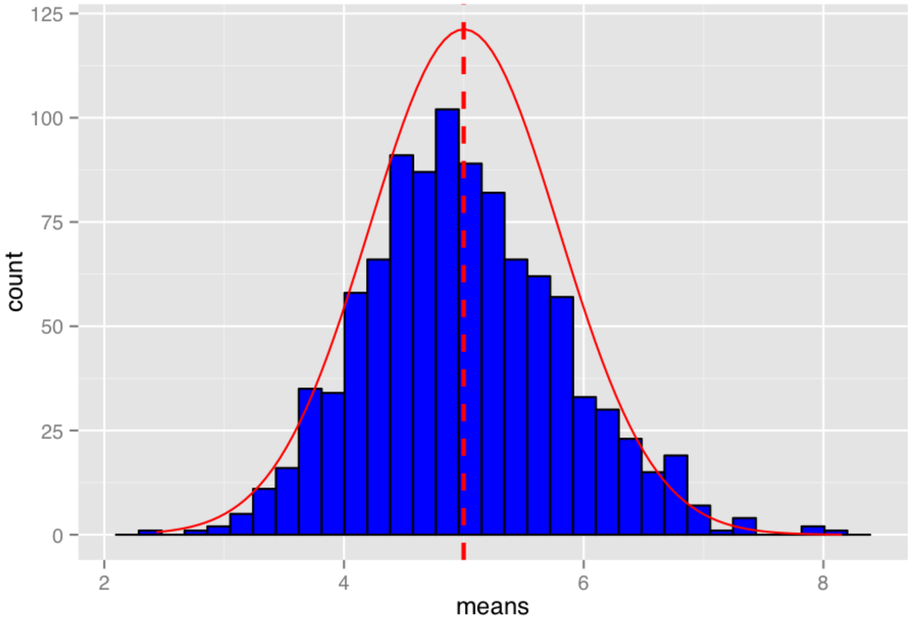
```
## [1] 6.441433 5.191165 4.505526 4.792823 4.048552 4.115367
```

The 40 simulation means above are centered around the theoretical $\mu = 1/.2 = 5$. Furthermore, taking the mean of the 40 means is 4.9983681, with sample standard deviation 0.8336931

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

```
ggplot(data.frame(means=simulations), 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 * 40 * dnorm(x, mean=5, sd = .79)}, color='red')
```

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



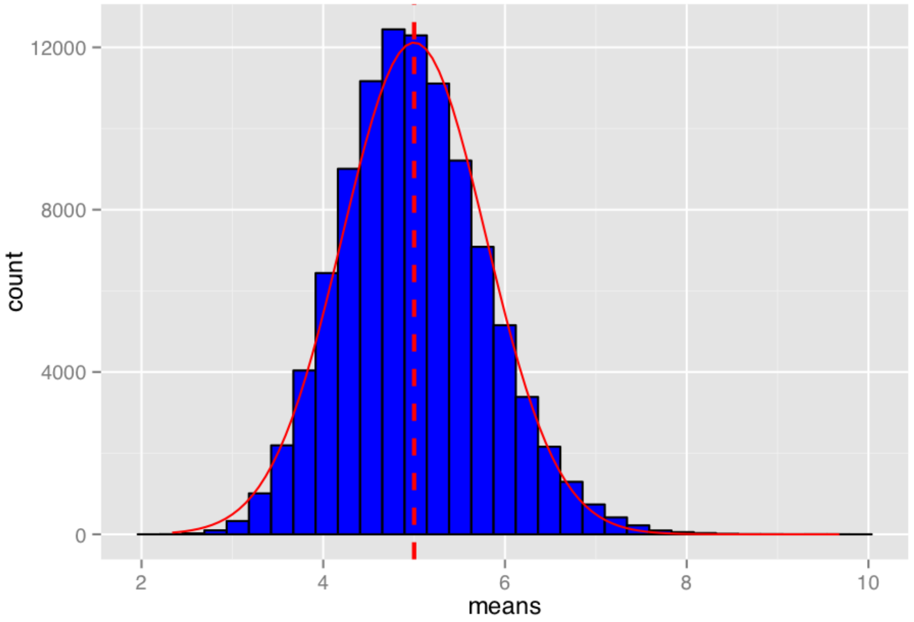
Our dashed vertical red line is at the sample mean 4.9983681 and approximates μ with standard error $\sigma^2/n = \sigma\sqrt{(n)} = 5\sqrt{(40)} = 0.7905694$

Recalling our standard deviation of an exponential to be $1/\lambda = 1/.2 = 5$ it is trivial to see our sample standard deviation of $0.8336931 \approxeq 5/\sqrt{(40)} = 0.7905694$

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

ggplot(data.frame(means=simulations), 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')
```

```
## 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 with slightly better mean 5.000086, and standard deviation 0.7903241 approximately as before; extremely close to the standard error of a normal distribution with sample size = 40.

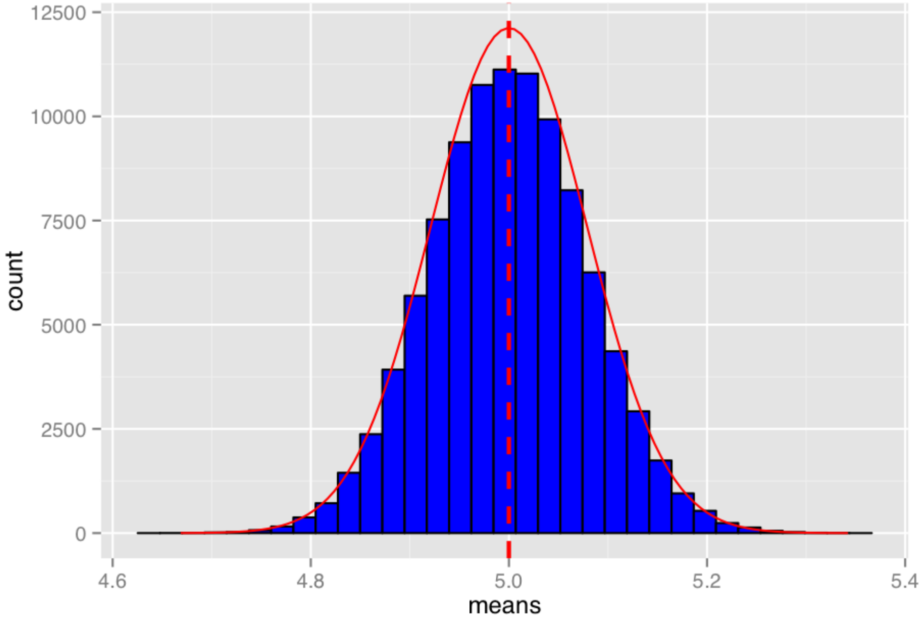
If we hold simulations to 100,000 and increase the sample size to 4000, we expect the standard error to be smaller i.e $\approxeq 5/\sqrt{(4000)} = 0.0790569$

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

Taking the mean of the 100,000 simulations now has mean
4.9999661, with sample standard deviation 0.0790784 $\approxeq 5/\sqrt{(4000)}$ 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.

```
ggplot(data.frame(means=simulations), 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')
```

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


count
means