Exponential Distribution and the Central Limit Theorem: A Simulation in R

by: Mohammed K. Barakat

July 1, 2015

Overview

This analysis investigates the exponential distribution and how it relates to the Central Limit Theorem. The analysis is performed using R-language where a simulation is done to illustrate the properties of the distribution of the means of 40 exponentials.

Simulations

The **Central Limit Theorem** (CLT) is one of the most important theorems in statistics. It states that the distribution of averages of iid (independent and identically distributed) variables becomes that of a standard normal as the sample size increases. So, the result of:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} = \frac{\text{Estimate} - \text{Meanofestimate}}{\text{Std. Err. ofestimate}}$$

has a distribution like that of a standard normal for large n. This will be applied later in this analysis to the exponential distribution and see the normality of the calculated results.

Hence, in order to apply the CLT to the exponential distribution we will investigate the distribution of averages of 40 exponentials. The average of 40 exponentials will be simulated 1000 times to the benefit of better CLT application results.

Properties of the Exponential Distribution

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of the exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Lambda is denoted by (λ).

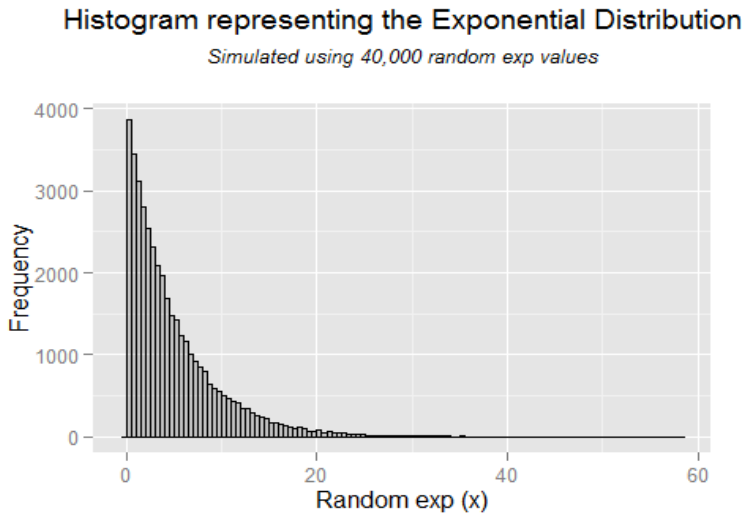
We will set λ **= 0.2** for all simulations. Hence,

```
lambda=0.2
mn=1/lambda
stdev=1/lambda
```

Mean (mn)= 5. And the **Standard deviation (stdev)= 5**

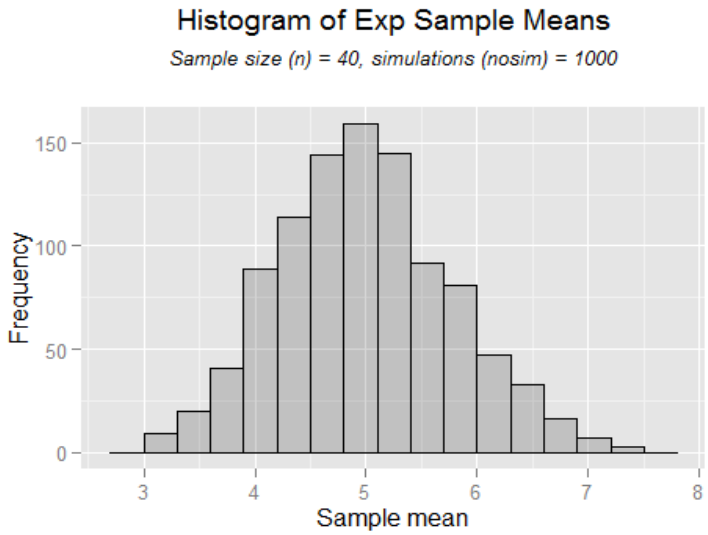
To build the exponential distribution we will simulate a total of 40,000 (nosim X n) exp randoms. Then, plot their distribution using Histogram. Where 'nosim' is the number of simulations, and n is the sample size. These variables will be used later when we study the

CLT. Graph.1 below shows the resulted exponential distribution. *(For R code, see Appendix: Graph.1-R Code)*



Now, we will apply the CLT to the exponential distribution by calculating the average of 40 exponentials simulated over 1000 times. Then, we will plot the 1000 averages of 40 exponentials and see how the new distribution looks like. *(For R code, see Appendix: Graph.2-R Code)*

Graph.2: Distribution of Exponentials Sample Means



Sample Mean versus Theoretical Mean

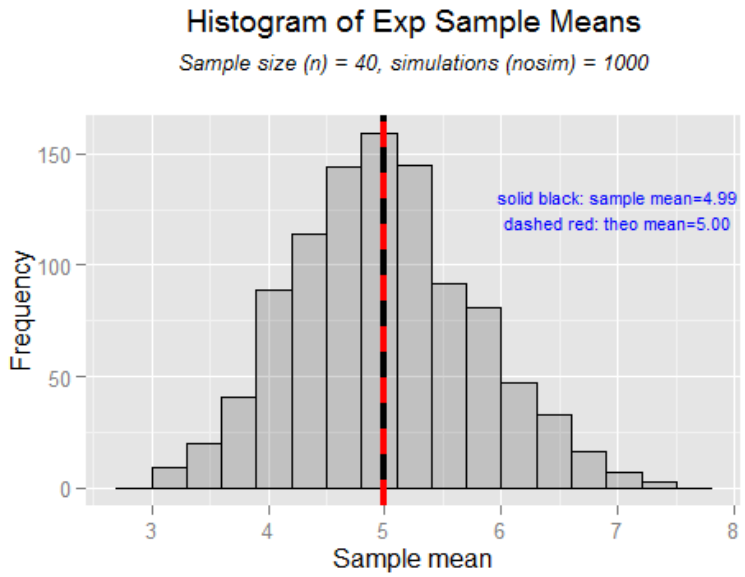
So far, the mean (the average value of the samples means with n = 40) can be derived in two ways:

1. Sample Mean (datamn): calculating the average of the 1000 means in the exponential dataset.
2. Theoretical Mean (theomean): calculated using the property of the exponential distribution that mean=1/λ

```
datamn=round(mean(expdatamns$xmns),2)
theomean=round(1/lambda,2)
```

Hence, **datamn= 4.99**, and **theomean= 5**

Graph.3: Distribution of Exponentials Sample Means (Theoretical Mean shown)



As shown in Graph.3 above the solid black line represents the *Sample Mean at 4.99*, while the red dashed line represents the *Theoretical Mean at 5.00*. Both lines are almost of the same value which validates the Central Limit Theorem and the LLN which state that averages of iid samples converge to the population mean they are estimating. Besides, the CLT states that averages are approximately normal with distribution centered at the population mean. *(For R code, see Appendix: Graph.3-R Code)*

Sample Variance versus Theoretical Variance

Similarly, the variance of sample means with n=40 can be derived in two ways:

1. Sample Variance (SampleVar): calculating the variance of the 1000 means in the exponential dataset.
2. Theoretical Variance (theoVar): since this distribution is a distribution of means, its variance is calculated using the formula σ^2/n. Where σ is the population standard deviation which is equal to 1/λ.

```
SampleVar<-round(var(expdatamns$xmns),2);
theoVar<-round(stdev^2/n,2)
```

Hence, **SampleVar= 0.61**, and **theoVar= 0.62**

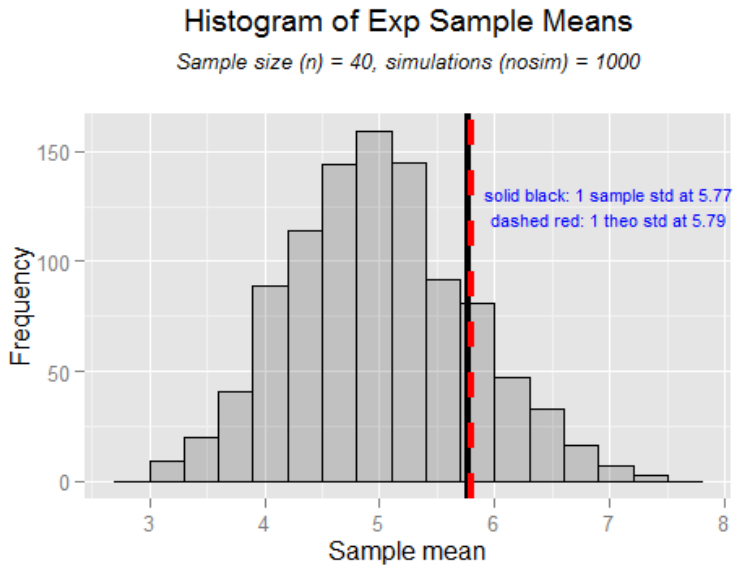
We can conclude that both the sample variance and the theoretical variance are very close. Since variances cannot be directly represented in graphs, we can use their respective standard deviation values which are their square roots. The standard deviation of the distribution of means is called the Standard Error ($\sigma/\sqrt{n}$).

```
SampleSTD<-round(sqrt(SampleVar),2)
theoSTD<-round(sqrt(theoVar),2)
```

So, the respective standard deviations are: **SampleSTD= 0.78**, and the **theoSTD= 0.79**

Now we can represent both the sample and the theoretical standard deviations on the distribution of means and see how close they are.

Graph.4: Distribution of Exponentials Sample Means (Theoretical std shown)



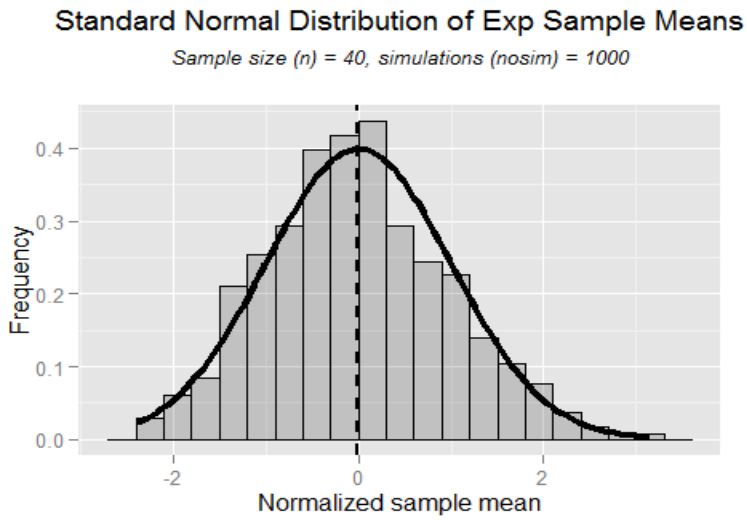
As shown in Graph.4 above the standard deviations (and, hence, variances) of both sample and theoretical are very close to each other. *(For R code, see Appendix: Graph.4-R Code)*

Normality of the Distribution of Exponentials Sample Means

Since the histogram shown in Graph.2 is almost bell-shaped (Gaussian) distribution we can conclude that the distribution of means follows the Normal Distribution with mean = 4.99.

Moreover, as explained in the Simulations section, by applying the CLT and plotting the values of: $\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}$ to all calculated means we should get a Standard Normal Distribution with $\mu = 0$ and $\sigma = 1$ as shown in Graph.5 below.

Graph.5: Standard Normal Distribution of the Exponentials Sample Means



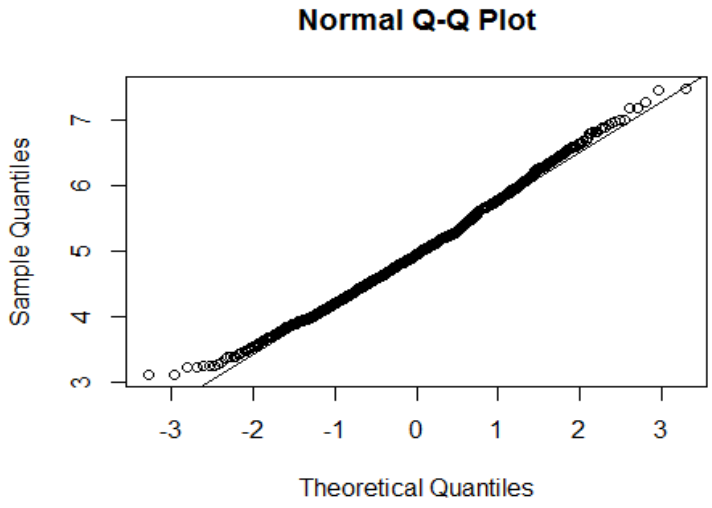
(For R code, see Appendix: Graph.5-R Code)

We can prove Normality of the distribution of means using the Normal Probability plot (QQ Plot) which is a graphical technique to assess whether or not a dataset is approximately normally distributed.

As shown in Graph.6 below the data points follow the straight line of the theoretical normal distribution which implies normality of the means distribution.

Graph.6: Normal Probability Plot of the distribution of means

```
qqnorm(expdatamns$xmns)
qqline(expdatamns$xmns)
```



APPENDIX

Loading necessary R packages

The analysis uses some packages that need to be installed before running the code. The code below loads these packages. Yet, you need to make sure they are installed in your R version before loading.

```
library(ggplot2);library(dplyr)
```

The Exponential Distribution

Graph.1-R Code

```
nosim<-1000
n<-40
lmda=0.2
mn=1/lmda
stdev=1/lmda

set.seed(1)
expdata<-data.frame(x=rexp(n*nosim,rate=lmda))
a<-ggplot(expdata,aes(x=x))+
        geom_histogram(alpha = .20, binwidth=.5, colour = "black")+
        labs(x="Random exp (x)",y="Frequency")+
        theme(plot.title = element_text(size = 14, face = "bold", colour = "black", v
just = +1))+
        ggtitle(expression(atop("Histogram representing the Exponential Distribution"
,
                                atop(italic("Simulated using 40,000 random exp values
")))))
a
```

Distribution of exponentials means

Graph.2- R Code

```
set.seed(1)
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(rexp(40,rate=lmda)))
expdatamns=data.frame(xmns=mns)
datamn=mean(expdatamns$xmns) # simulated sample mean


b <- ggplot(expdatamns, aes(x = xmns)) +
        geom_histogram(alpha = .20, binwidth=.3, colour = "black")+
        scale_x_continuous(breaks=2:8)+
        labs(x="Sample mean",y="Frequency")+
        theme(plot.title = element_text(size = 14, face = "bold", colour = "black", v
just = +1))+
        ggtitle(expression(atop("Histogram of Exp Sample Means",
                                atop(italic("Sample size (n) = 40, simulations (nosim
) = 1000")))))

b
```

Sample Mean versus Theoretical Mean

Graph.3-R Code

```
c<-b+geom_vline(xintercept=datamn,size=1.5,linetype=1)+
        geom_vline(xintercept=mn,size=1.5,linetype=2,colour="red")+
        annotate("text",x=7,y=125,label= "solid black: sample mean = 4.99\ndashed red
: theoretical mean = 5.00",size=3,color="blue")
c
```

Sample Variance versus Theoretical Variance

Graph.4-R Code

```
d <- ggplot(expdatamns, aes(x = xmns)) +
        geom_histogram(alpha = .20, binwidth=.3, colour = "black")+
        geom_vline(xintercept=SampleSTD+datamn,size=1.5,linetype=1)+
        geom_vline(xintercept=theoSTD+mn,size=1.5,linetype=2,colour="red")+
        scale_x_continuous(breaks=2:8)+
        labs(x="Sample mean",y="Frequency")+
        annotate("text",x=7,y=125,label= "solid black: 1 sample std at 5.77\ndashed r
ed: 1 theoretical std at 5.79",size=3,color="blue")+
        theme(plot.title = element_text(size = 14, face = "bold", colour = "black", v
just = +1))+
        ggtitle(expression(atop("Histogram of Exp Sample Means",
                                atop(italic("Sample size (n) = 40, simulations (nosim
) = 1000")))))
d
```

Normality of the Distribution of Exponentials Sample Means

Graph.5-R Code

```
library(dplyr)
expdatamns<-mutate(expdatamns,nval=sqrt(n)*(expdatamns$xmns-mn)/stdev)

e <- ggplot(expdatamns, aes(x = nval)) + geom_histogram(alpha = .20, binwidth=.3, col
our = "black", aes(y = ..density..)) +
        stat_function(fun = dnorm, size = 2)+
        geom_vline(xintercept=mean(expdatamns$nval),size=1)+
        labs(x="Normalized sample mean",y="Frequency")+
        theme(plot.title = element_text(size = 14, face = "bold", colour = "black", v
just = +1))+
        ggtitle(expression(atop("Standard Normal Distribution of Exp Sample Means",
                                atop(italic("Sample size (n) = 40, simulations (nosim
) = 1000")))))
e
```