Exponential distribution and Central Limit Theorem

Javier Angoy

Jun 4th, 2017

Synopsis

This document contains the first part of the assignment for the Statistical Inference Course of Coursera. The objective is try to answer the question "Does de distribution of means of 40 exponentials behave as predicted by the Central Limit Theorem?". It consists of an analysis of the Central Limit Theorem applied to the exponential distribution. A simulation of this distribution will be carried out using R and the results obtained will be compared with the theoretical equivalents: mean and variance. It is also shown how the distribution approximates to normal.

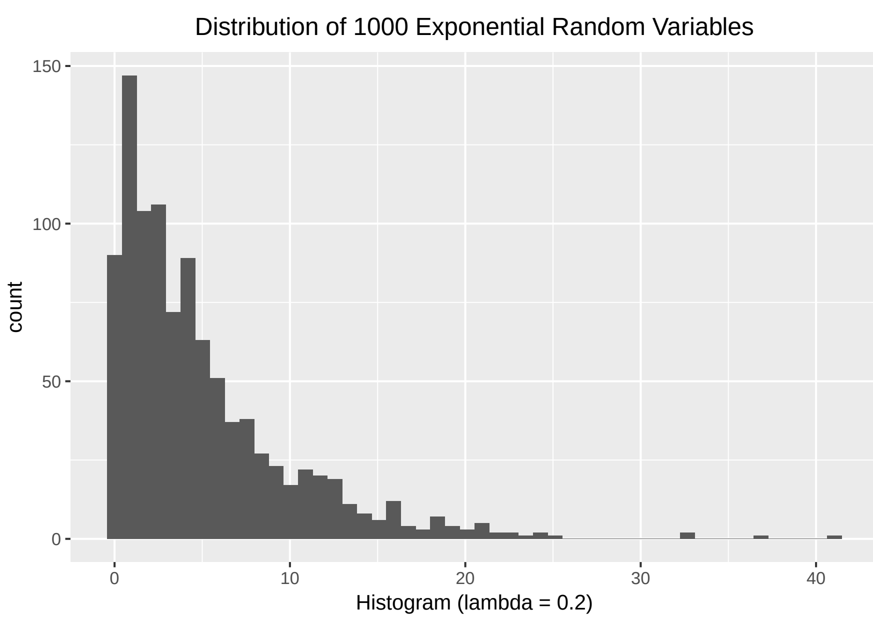
Simulation

The R function rexp(n, lambda) generates random variables with an exponential distribution. First we set seed and lambda = 0.2 for all of the simulations. We later generate a sample of 1000 exponentials.

```r
set.seed(1000)
lambda <- 0.2 #Set lambda value
ns <- 1000 #Set number exponential dists.
sample.exp <- rexp(ns,lambda)
```

So we have simulated 1000 random exponentials. Histogram for the simulation.

```r
qplot(sample.exp, bins=50, geom="histogram",
      main = "Distribution of 1000 Exponential Random Variables",
      xlab = "Histogram (lambda = 0.2)")+
      theme(plot.title = element_text(hjust = 0.5))
```



As our first approach to Central Limit theorem, we will simulate the distribution of 1000 means for 40 random exponentials.

```r
ns <-1000 ; n <- 40 #Set number of means and number of exponential dists.
means = NULL; for (i in 1 : ns) means = c(means, mean(rexp(n, lambda)))
```

Sample mean versus Theoretical mean

Both mean and standard deviation for a exponential distribution are equal to $1/\lambda$. So since lambda is 0.2, the theoretical mean should be:

```r
th.mean <- 1/lambda ; th.mean          #theoretical mean
```

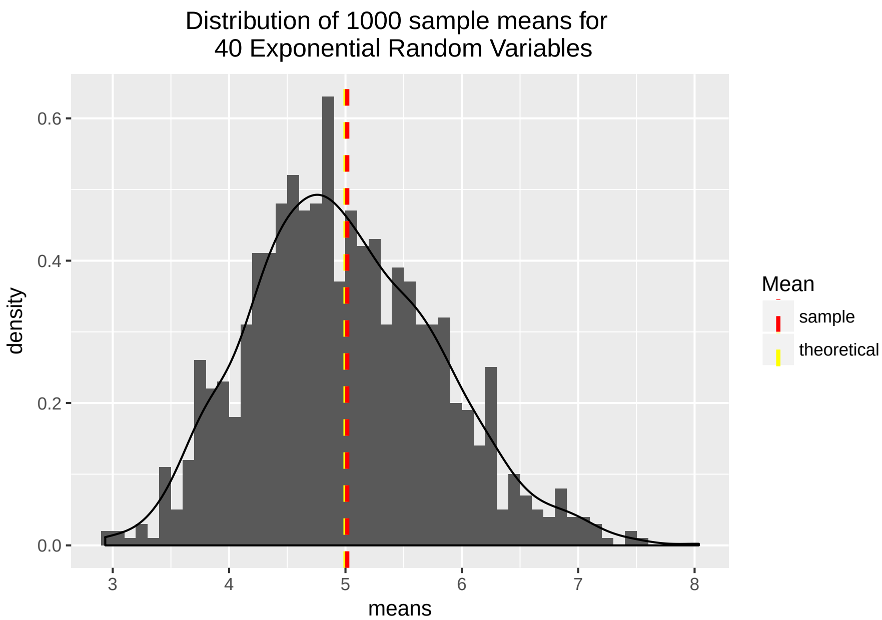
```
## [1] 5
```

```r
sa.mean <- mean(sample.exp) ; sa.mean #sample mean
```

```
## [1] 5.015616
```

Histogram overlaid with density curve. Theoretical and sample mean also shown. As we see, theoretical mean is 5, whilst sample mean is 5.0156161. The difference is small and in some cases it may be difficult to distinguish when they are presented together in the graph.

```r
ggplot(as.data.frame(means), aes(x=means)) +
    geom_histogram(aes(y=..density..),
    breaks=seq(round(min(means),1), round(max(means),1), by=0.1)) +
    ggtitle("Distribution of 1000 sample means for \n 40 Exponential Random Variables") +
    geom_density()+
    geom_vline(aes(xintercept=th.mean, color="theoretical"),
               linetype="dashed", size=1, show.legend =TRUE)+
    geom_vline(aes(xintercept=sa.mean, color="sample"),
               linetype="dashed", size=1, show.legend = TRUE)+
    scale_color_manual(name="Mean", values=c("red","yellow"))+
    theme(plot.title = element_text(hjust = 0.5))
```



Another thing we can see is that the distribution of means of the samples from the exponential is centered around the population mean.

Sample Variance versus Theorectical Variance

According to the Central Limit Theorem, the sample means should have a (theoretical) variance of $(\frac{1/\lambda}{\sqrt(n)})^2$.

```r
th.variance <- (((1/lambda))/sqrt(n))^2 ; th.variance  #theoretical variance
```

```
## [1] 0.625
```

```r
sa.variance <- var(means) ; sa.variance                #sample variance
```

```
## [1] 0.6562298
```

So as we calculate the variance of the sample means we get 0.6562298, which is very close to the theoretical variance of 0.625.

Distribution

The Center Limit Theorem states that our distribution of 1000 samples of 40 exponentials, properly normalized should become a standard normal. So we need to normalize (center) the distribution. We will do so by subtracting the population mean ($\mu = 1/\lambda$) to our sample mean and then divide the result by the standard error of the sample mean $\frac{\bar{X}-\mu}{\sigma/\sqrt(n)}$. The histogram for our distribution standardized means. Unlike the generated exponential distribution, we can see that its bell shape looks quite similar to the normal distribution.

```r
beta <- 1/lambda
centered = (means - beta)/(beta/sqrt(n)) # calculate centered means
g <- ggplot(as.data.frame(centered), aes(x = centered))
g <- g + geom_histogram(breaks=seq(round(min(centered),1), round(max(centered),1), by=0.1),
                        color="black", aes(y = ..density..))
g <- g + stat_function(fun = dnorm, size = 2)
g <- g + ggtitle("Distribution of 1000 sample means for \n 40 Exponential Random Variables") +
    labs(x="Normalized Sample Means (centered)") +
    theme(plot.title = element_text(hjust = 0.5))
g
```