Statistical Inference: Exponential Distribution Simulation

emeko

January 27, 2016

Overview

In this report, we will look at Exponential Distribution (ED) and it's relation to the the Central Limit Theorem (CLT). The ED has a theoretical mean and and standard distribution such that $\mu = \sigma = 1/\lambda$. The CLT states that the means of large number of samples of independent random variables from a distribution with a defined mean and variance will be approximately normally distributed.

The probability mass function for the Exponential Distribution is given by:

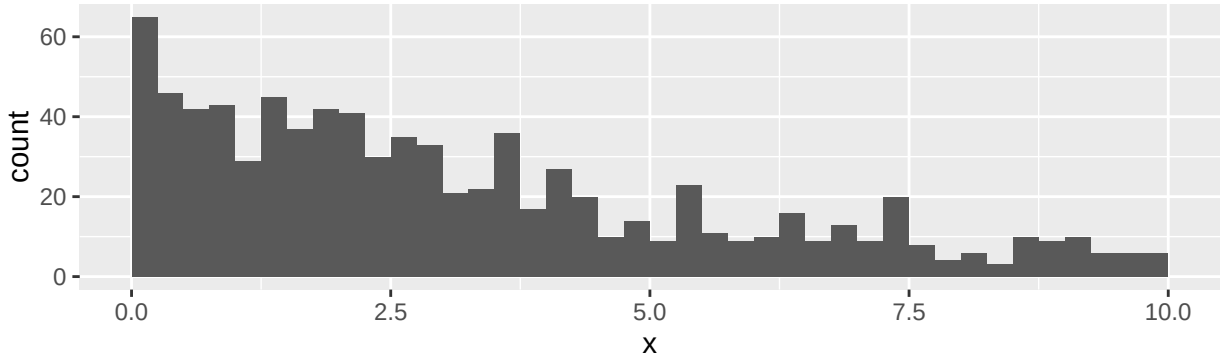
$$P(X = x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Simulations

rexp(n, rate) is an R function that returns n random variables from an Exponential Distribution with $rate = \lambda$. We can randomly choose 1000 values from the distribution using the following code chunk.

```
lambda = 0.2
iterations = 1000
random.exponential.values <- rexp( iterations, lambda )

qplot( random.exponential.values, xlim = c(0,10), xlab = 'x', binwidth = 0.25)
```



From the plot, the distribution of values look like they are distributed per the ED.

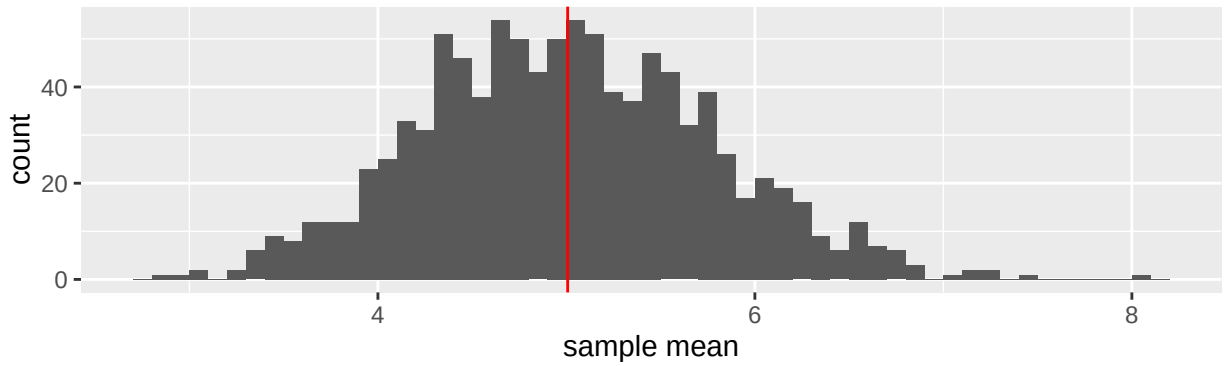
By using the *replicate()* function, we can create multiple samples (in this case, 40 values per sample) of random variables from an Exponential Distribution. Our aim is to take the mean of each of those samples to look at how the means of the samples are distributed:

```
observations = 40

simulation <- replicate(
```

```
    iterations,
    { mean( rexp( observations, lambda ) ) }
)

qplot( simulation, binwidth = 0.1, xlab = 'sample mean' ) +
    geom_vline( xintercept = mean( simulation ), color = 'red')
```



The distribution of means in the simuation is centered around 5.0066383, as indicated by the red line in the figure above.

Sample Mean versus Theoretical Mean

The aritmetic mean of the means of 1000 samples from the simulation is 5.0066383 which is quite close to the theoretical value of $1/\lambda$ (5), an error of just ϵ = -0.0066383 (0.133%):

```
simulation.mean <- mean( simulation )
simulation.mean
```

```
## [1] 5.006638
```

```
theoretical.mean <- 1 / lambda
theoretical.mean
```

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

```
theoretical.mean - simulation.mean
```

```
## [1] -0.006638315
```

Sample Variance versus Theoretical Variance

The theoretical standard deviation for the exponential distribution is $1/\lambda$ (5 for λ=0.2). Since $Var(x) = \sigma^2$, the theoretical variance would be 25 where λ=0.2.

The variance for the distribution of sample means is 0.5936371. Per the Central Limit Theorem, we expect the sample means to be normally distributed around the population mean ($1/\lambda$; also 5 where λ=0.2) where the variance of the means is equal to σ^2/n.

```
simulation.var  <- var( simulation )
simulation.var
```

```
## [1] 0.5936371
```

```
theoretical.sd  <- 1 / lambda
theoretical.var <- theoretical.sd ^ 2
theoretical.var
```

```
## [1] 25
```

We can estimate the variance of the population from the variance seen in the sample means by dividing by the number of observations:

```
estimated.var = simulation.var * observations
estimated.var
```

```
## [1] 23.74548
```

```
theoretical.var - estimated.var
```

```
## [1] 1.254517
```

The difference between the observed variance and the theoretical variance is $\epsilon = 1.2545172$ (5.018%) – close but not quite as close as the means. This variance would be expected to decrease dramatically along with decreasing the number of observations.

Distribution

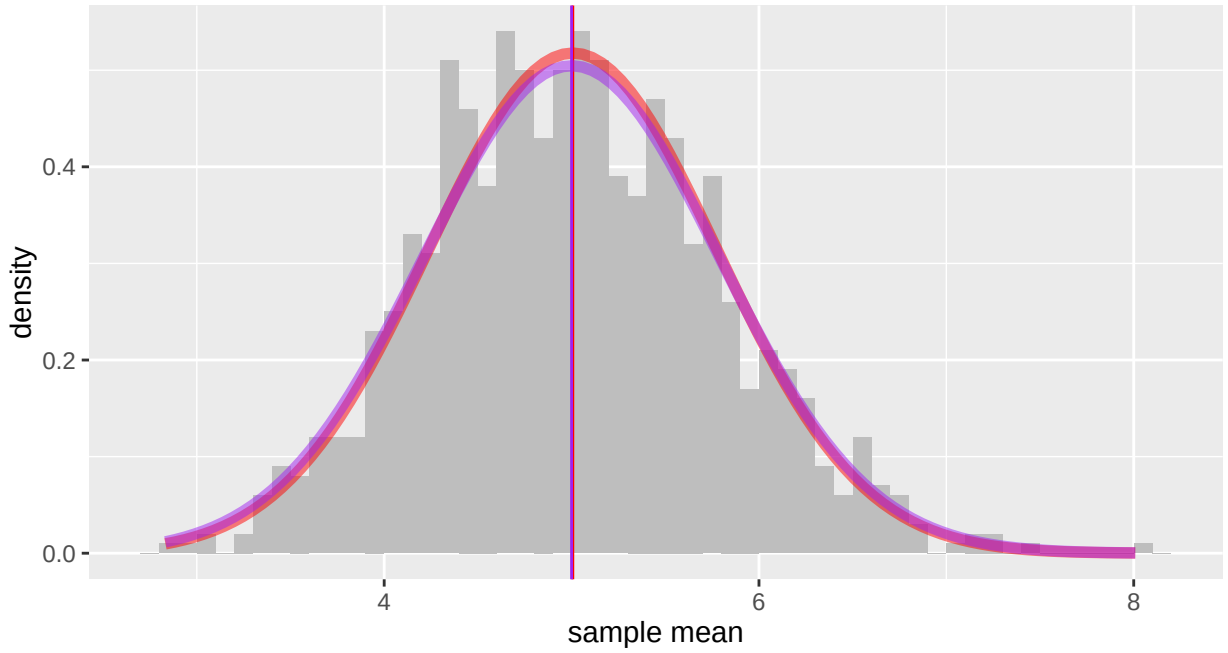
The most effective illustration of how the arithmatic means of the samples have a distribution that approximates the normal distribution is to superimpose a normal distribution on top of the histogram of sample means from the simulation. The following illustration shows a density histogram of the sample means (where each bar's heigh is divided by the sample count so that the sum of their heights equals 1), a normal distribution with the theoretical mean and standard deviation of an exponential distribution with λ=0.2 (in red), and the observed mean and standard deviation of the samples (in purple).

```
simulation.mean     <- mean( simulation )
simulation.sd       <- sd( simulation )

theoretical.mean    <- 1 / lambda
theoretical.sd      <- 1 / lambda

qplot( simulation, xlab = 'sample mean', geom = 'blank' ) +
    theme( legend.position = 'none' ) +
    geom_histogram( aes( y=..density..), binwidth=0.1, fill='grey' ) +
    geom_vline( xintercept = simulation.mean, color = 'red' ) +
    stat_function(
        fun = dnorm,
        args = list( mean = simulation.mean, sd = sqrt(simulation.var) ),
```

```
        color = 'red', size=2, alpha=0.5
    ) +
    geom_vline( xintercept = theoretical.mean , color = 'purple' ) +
    stat_function(
        fun = dnorm,
        args = list( mean = theoretical.mean, sd = theoretical.sd / sqrt(observations) ),
        color = 'purple', size=2, alpha=0.5
    )
```



Appendix

1. R libraries used - ggplot2
2. TeX is required to create the PDF output - You should install a recommended TeX distribution for your platform
3. Source code - https://github.com/emeko/StatInfProject