Application of the Central Limit Theorem to Means of Exponentially Distributed Random Variables

Dominik Sudwischer

5 September 2017

Motivation

The Central Limit Theorem (CLT) states that the average of n independent and identically distributed ('iid') random variables $X_1, \ldots, X_n$ with a common finite expected value μ and a common finite standard deviation σ approximates a normally distributed random variable with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as n approaches infinity.

In order to get a better understanding of the assumptions and consequences of the CLT, we will consider an example in which we will run a simulation to approximate the distribution of the mean of 40 exponentially distributed random variables with parameter $\lambda = .2$.

We will then compare the resulting distribution of the aforementioned random variable, i.e. $X = \frac{1}{n}\sum_{i=1}^{n} X_i$, to the normal distribution that X should be close to accourding to the CLT.

Generating the Required Data

We will start by using a deterministic seed to ensure our results can be reproduced.

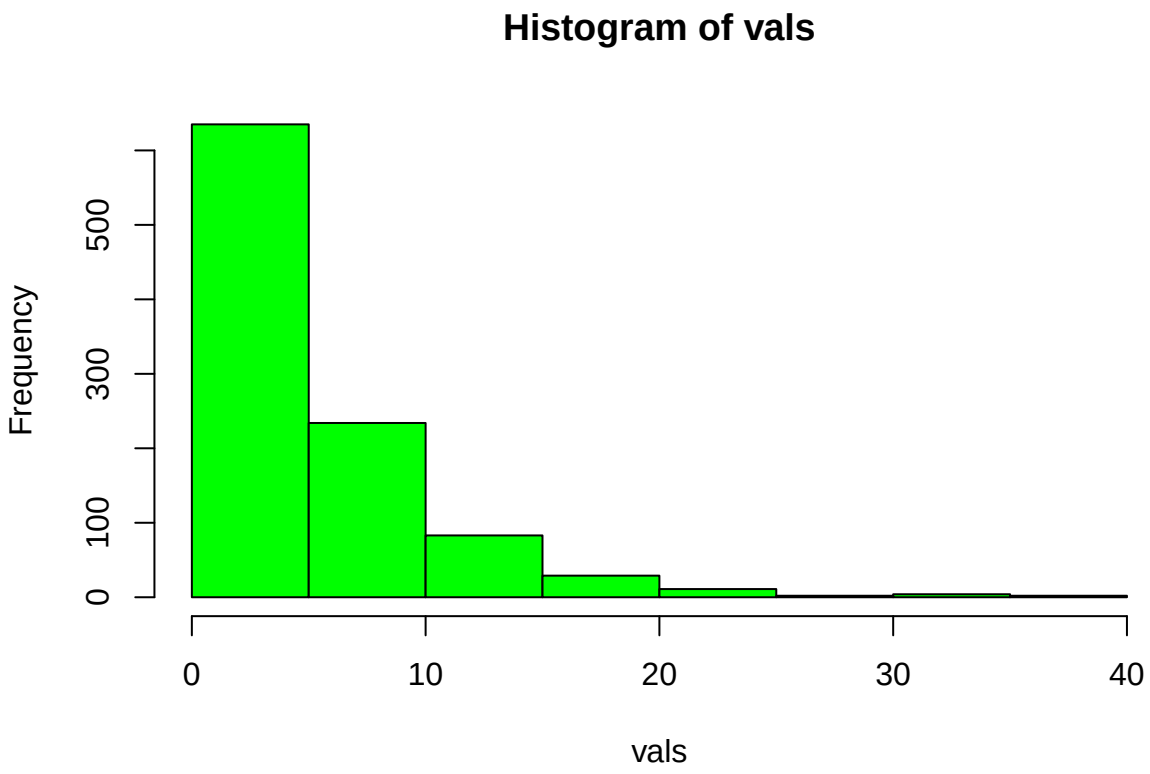
```
set.seed(1234)
```

To get a good grasp of what the distribution of the sample mean of $n = 40$ independent exponentially distributed random variables looks like, we will draw this mean a $num.iterations = 1000$ times in total. The distribution parameter we will use is $\lambda = .2$, By definition of the exponential distribution, each of those random variables will have mean $\frac{1}{\lambda} = 5$ and variance $\frac{1}{\lambda^2} = 25$.

```
n <- 40
num.iterations <- 1000
lambda <- .2
```

We can now proceed to use R's 'rexp' function to generate 40 values of independent exponentially distributed random variables with paramter λ. Before we step into the analysis of the CLT, let us begin with a simple draw of *num.iterations* values from such an exponentially distributed random variable:

```
vals <- rexp(num.iterations, lambda)
hist(vals, col = "green")
```



Note that it suffices to plot the non-negative half axis since exponentially distributed random variables do not attain negative values. Clearly, the exponential distribution has a very different shape than any normal distribution. We will continue our analysis by repeatedly calculating sample means of 40 iid exponentially distributed random variables.In total, we will draw 1000 means of 40 iid random variables each.

```
means <- NULL
for (i in 1 : num.iterations) means <- c(means, mean(rexp(n, lambda)))
```

Let us have a look at the first 10 means. Since the exponential distribution has mean 5 and standard deviation 5, we would expect our means to be centered around 5.

```
print(head(means, n = 10))
```

```
##  [1] 5.164118 4.739591 5.701339 4.912194 3.909602 5.928937 5.219741
##  [8] 5.221456 4.717802 4.325011
```

Comparing Mean and Standard Deviation

To continue our analysis, let us calculate the sample mean and the sample standard deviation of our 1000 simulated draws from X:

```
mean(means)
```

```
## [1] 4.972319
```

```
sd(means)
```

```
## [1] 0.7602225
```

According to the CLT, the normally distributed random variable approximated by our sample mean distribution has the following mean and standard deviation:

```
1 / lambda
```

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

```
1 / lambda / sqrt(n)
```

```
## [1] 0.7905694
```

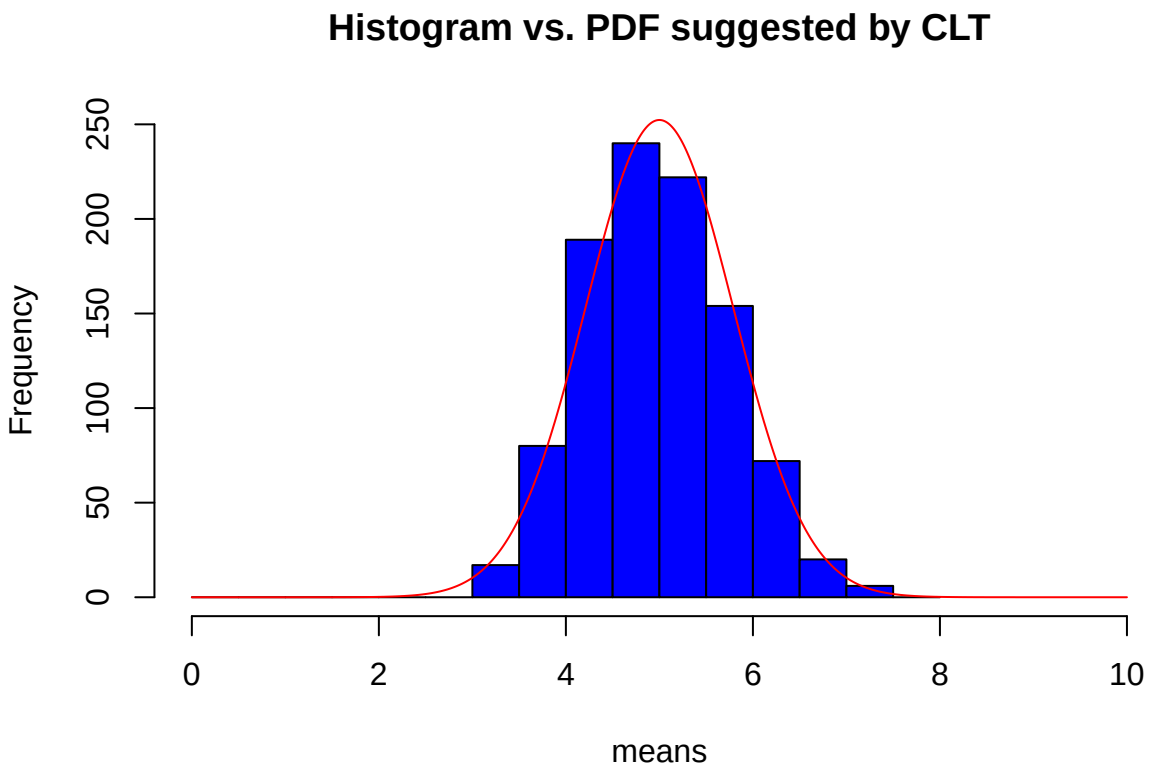
Creating Plots to Visualize Results

Our results seem reasonably close to what the CLT suggests. Let us create a histogram for 1000 draws from independent copies of X to visualize the result of our simulation and to compare its shape to a Gaussian probability density function. We will set the size of the disjoint bins to .5 for a reasonable granularity.

```
bin.size <- 0.5
h <- hist(means, plot = TRUE, col = "blue",
          breaks = seq(0, ceiling(max(means)), bin.size))
```

We recall that we would expect our histogram to be a discrete approximation of a normal distribution. More specifically, the CLT suggests it should be close to $N(5, \frac{25}{40})$. Let us compare the proposed density to our histogram in a plot. Note that we scale our density by a factor corresponding to the product of the bin size and the number of draws to match the bins of the histogram and the absolute frequencies. Without that scaling, our density function would not be clearly visible in the graph.

```
plot(h, xlim = c(0, 10), ylim = c(0, 250), col = "blue",
     main = "Histogram vs. PDF suggested by CLT")
x = seq(0, 10, 0.01)
lines(x, num.iterations * bin.size * dnorm(x, mean = 1/lambda, sd = 1 / lambda / sqrt(n)),
      xlim = c(0, 10), col = "red")
```



As visible in the figure, the blue histogram resembles the characteristic bell shape of the normal distribution whose (scaled) density is shown as a red line.

Conclusion

We conclude our analysis with the insight that our results match the statement of the CLT. We ran a simulation to approximate the distribution of $X = \frac{1}{40}\sum_{i=1}^{40} X_i$ where the X_i are independent exponentially distributed random variables with parameter $\lambda = .2$. We then compared the sample mean and standard deviation of our sample of 1000 draws of X to a normal distribution with mean 5 and standard deviation $\frac{5}{\sqrt{40}}$ and found that the statement of the CLT explains our observation. By increasing the number of draws per average from 40 to a higher value, we would expect the distribution of X to come even closer to a normal distribution.