Exponential Distributions vs the Central Limit Theorem

Nils Sandøy

August 19, 2015

This is a course project for the Coursera class in Statistical Inference. The goal of this project is to investigate the Exponential Distribution and make a comparison with the Central Limit Theorem (CLM).

Assignment Overview

> Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should
>
> 1. Show the sample mean and compare it to the theoretical mean of the distribution.
> 2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
> 3. Show that the distribution is approximately normal.

The **CLT** states that the distribution of averages of **independent** and **identically distributed** (**iid**) variables (properly normalized) becomes that of a standard normal as the sample size increases. The result is The result is that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}}$$

has a distribution like that of a standard normal for large n. A useful way to think about the CLT is that $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$

The **Exponental Distribution** is the probability distribution that describes the time between events in a **Poisson process**, in which events occur continuosly and independently at a constant average rate.

The **probability density function** (PDF) of an exponential function is

$$f(x;\lambda) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & x >= 0. \\ 0 & x < 0. \end{cases}$$

The mean or expected value of an exponential distribution is

$$E[X] = \frac{1}{\lambda} = \beta$$

The standard deviation (σ) is equal to the mean ($\frac{1}{\lambda}$), as the variance of X is given by

$$Var[X] = \frac{1}{\lambda^2}$$

The exponential distribution with rate λ has density

$$f(x) = \lambda \mathrm{e}^{-\lambda x}$$

λ is set to **0.2** for all simulations. I will investigate the distribution of averages of 40 exponentials and do a thousand simulations.

Example exponential distribution with $n = 40$

```r
lambda <- 0.2
n <- 40

# Example exponential distribution
ed1 <- rexp(n,lambda)
ed1
```

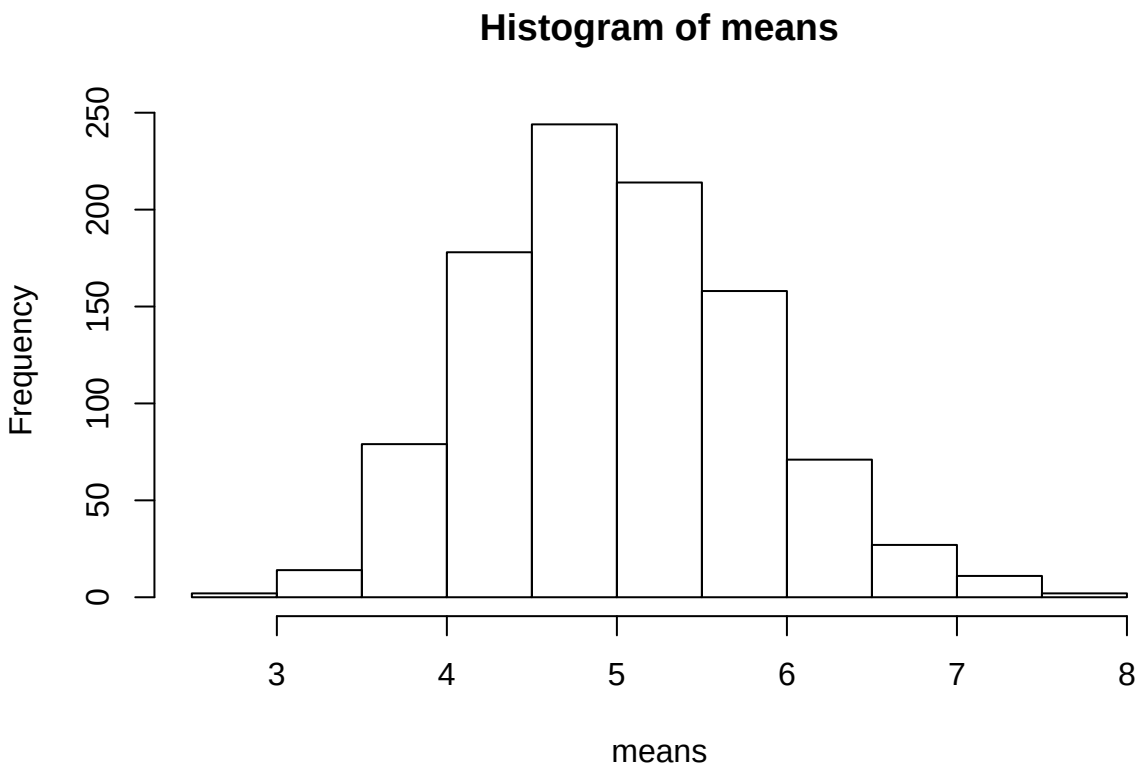
```
##  [1]  3.6883269  2.6196377 18.1053227 10.7138008 12.5190354  5.4318126
##  [7]  4.2014504  3.2931241  4.7532543  5.0375576  1.3406438  3.3090410
## [13]  1.9257257 19.9946317  3.0560778  3.5896887  4.1996815  2.9129789
## [19]  4.8614649  0.4536366  0.9947097  0.4142346 10.1041663  0.6498800
## [25]  2.4945593  2.6059373  2.4460382  8.0718811  0.2916472  0.4604749
## [31]  0.3637670  2.6747395  2.6599644  0.4346249  1.2965918  0.5306832
## [37]  8.2506314  9.0758392  0.2075193  0.1656208
```

The above shows a typical spread in values for a random set of 40 exponentials. The mean for this set is **4.2550101** while the theoretical mean ($\frac{1}{\lambda}$) is **5**.

Illustration of the properties of the distribution of the mean of exponentials.

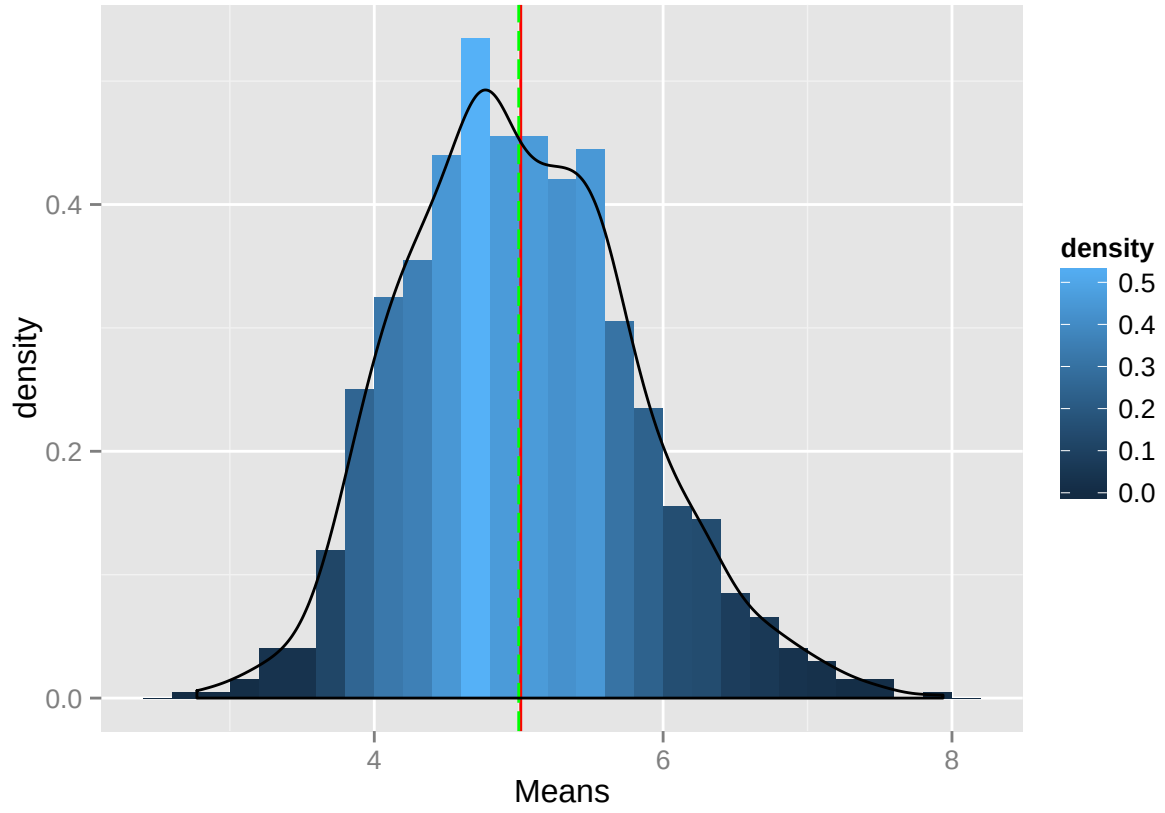
```r
lambda <- 0.2
n <- 40
theoretical_mean <- 1/lambda
theoretical_variance <- 1/(lambda^2)
nosim <- 1000

# Run the simulations
dist <- matrix(nrow=nosim, ncol=n)
for (i in 1:nosim) dist[i,] <- rexp(n, lambda)

# Find the mean and variance of each of the sample sets
means <-apply(dist, MARGIN=1, FUN=mean)
variances <- apply(dist, MARGIN=1, FUN=var)

# Take the average of the individual means and variations
sample_mean <- mean(means)
sample_variance <- mean(variances)
```

Here we have run 1000 simulations where we generate 40 random exponentials. We have also calculated the the mean and variance of each of these simulations. Finally we have calculated the **sample mean**, 5.014332, and **sample variance**, 24.746751 accross all these 1000 simulations.



A simple histogram over the means of the 1000 simulations, each of 40 values, shows that the distribution of the mean of these exponentials does cluster around the theorectical mean ($\frac{1}{\lambda}$) of 5

Task 1: Show the sample mean and compare it to the theoretical mean of the distribution.

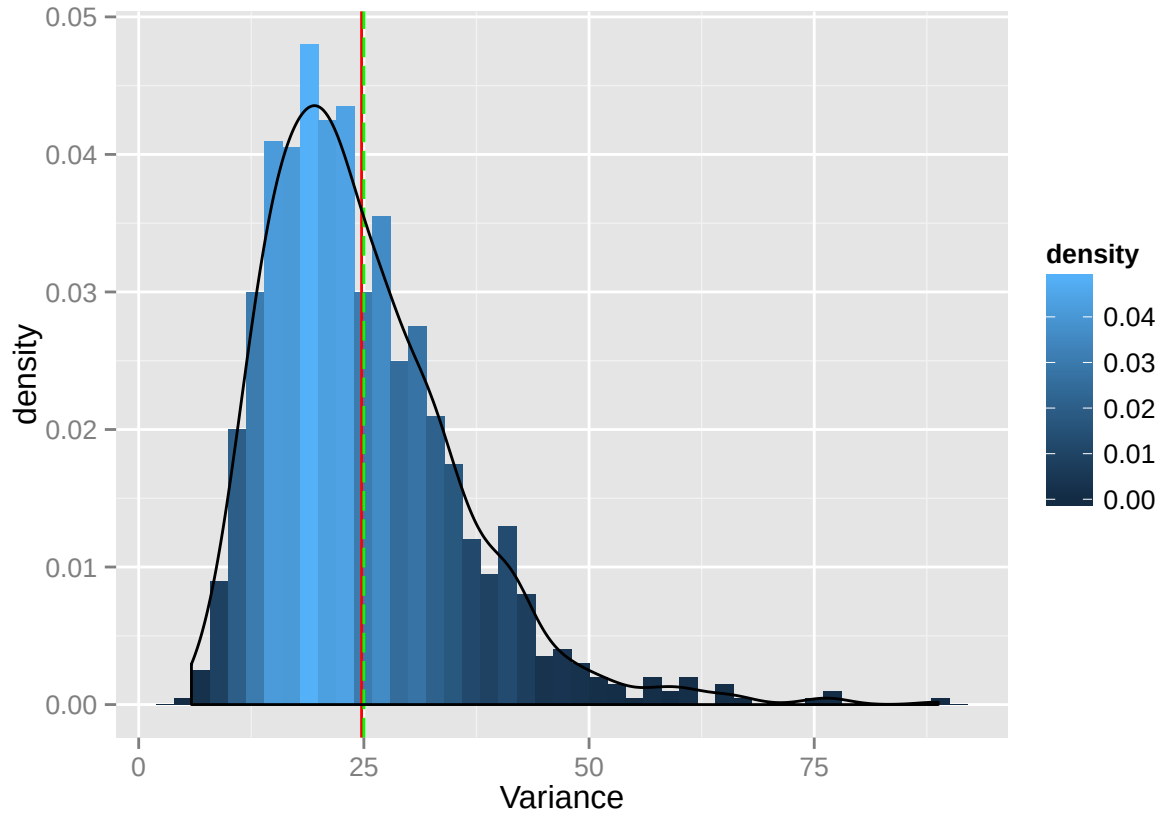


This is ahistogram over the means of each of the 1000 simulations with $n = 40$. The sample mean, **5.014332**, is here shown in red. This is the average of the individual means of each of the 1000 samples. It is very close to the theoretical mean ($\frac{1}{\lambda}$) of **5** shown in green.

This corresponds with the CLT which predicts that the sample mean will approximate the theoretical mean when we do many simulations.

Task 2: Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

```
library(ggplot2)

dat <- data.frame(Variance = c(variances))

g <- ggplot(dat, aes(x = Variance)) +
    geom_histogram(binwidth=2, aes(y = ..density.., fill=..density..))
g <- g + geom_vline(xintercept=sample_variance, color="red")
g <- g + geom_vline(xintercept=theoretical_variance, color="green", linetype="longdash")
g <- g + geom_density()
plot(g)
```



The average variance of the 1000 simulations (plotted in red) is $Var(\bar{X}) = \frac{\sigma^2}{n} =$ **24.746751**, which is close to the theoretical variance (green) of $Var(X) = \frac{1}{\lambda^2} =$ **25**.

This confirms again the CLT by showing that, when running many simulations, the sample variance estimates the population variance.

Task 3: Show that the distribution is approximately normal.

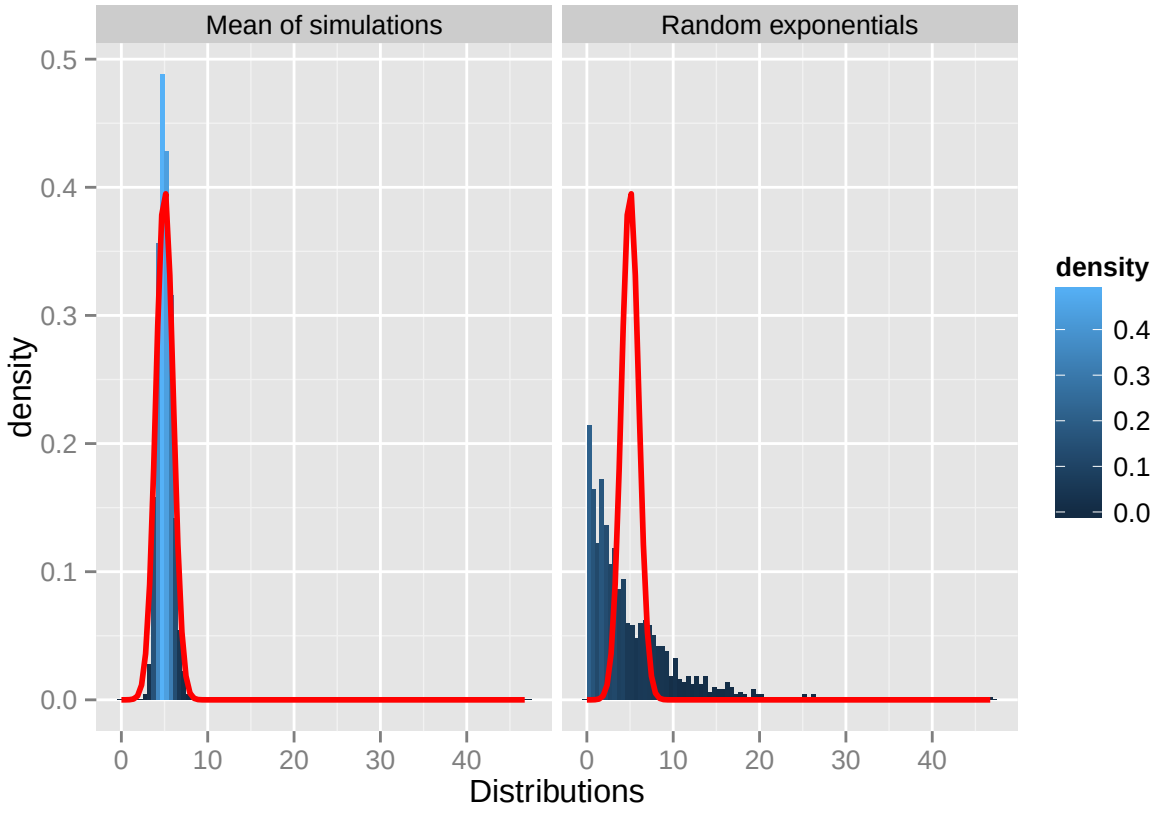
A random variable is said to follow a **normal** or **Gausian** distribution with mean μ and variance σ^2 if the associated density is:

$$(2\pi\sigma^2)^{-\frac{1}{2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

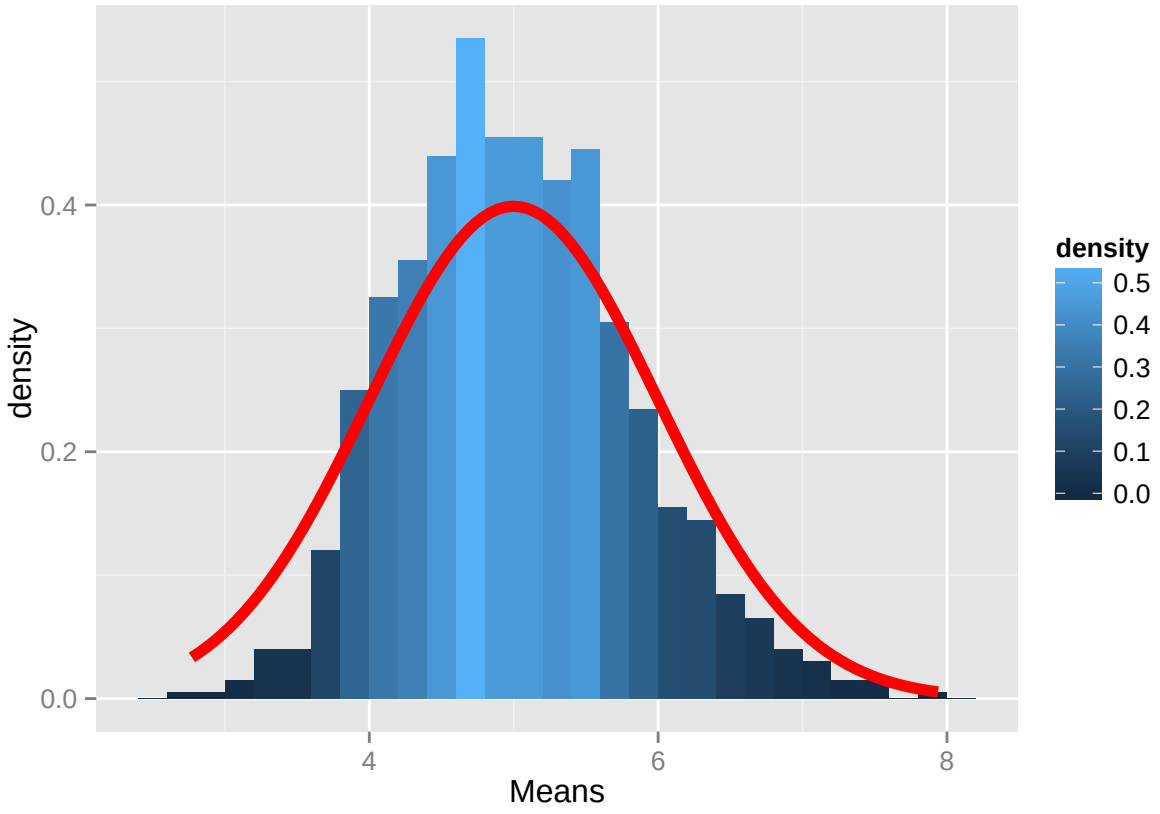
To make our argument, we focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials. I.e. we look at the distribution of 1000 random exponentials, and compare that to the average of 1000 simulations of sice n=40.

```
library(ggplot2)

sample_dist <- c(rexp(nosim, lambda))
dat <- data.frame(Distributions=c(sample_dist, means),
                  type = factor(rep(c("Random exponentials",
                                      "Mean of simulations"),
                                    rep(nosim, 2))))

ggplot(dat, aes(x = Distributions)) +
```

```
    geom_histogram(binwidth=.5, aes(y = ..density.., fill=..density..)) +
    stat_function(fun = dnorm, size = 1, color="red", args=list(mean=theoretical_mean)) +
    facet_grid(. ~ type)
```



It is clear that the distribution of the means of 1000 distribution with $n = 40$, is more "bell shaped" than the straight exponential distribution of 1000 elements. This is even clearer when we just compare the distribution of the means against the normal distribution:



This comparison between the means of 1000 distributions of size 40 and the normal distribution with σ=5 shows that **the distribution of means is approximately normal.**