Statistical Inference Course Project - Part 1: Simulation Exercise

geotsa

September 16, 2019

Overview:

In this report we're investigating the exponential distribution in R and compare it with the Central Limit Theorem. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. We are investigating the distribution of averages of 40 exponentials doing a thousand simulations.

Simulations:

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter.

To provide reproducability for exponentials generation the seed function establishes start position so every time generation gives the same values.

```r
# set seed
set.seed(123)
```

We set lambda = 0.2 for all of the simulations. Here we create the distribution of 1000 averages of 40 exponentials. Which means we do a thousand simulations.

```r
lambda <- 0.2
# number of observations following an exponential distributuion
n <- 40
# number of simulations, of different exponential distributions
simuls <- 1000
```

Sample Mean versus Theoretical Mean:

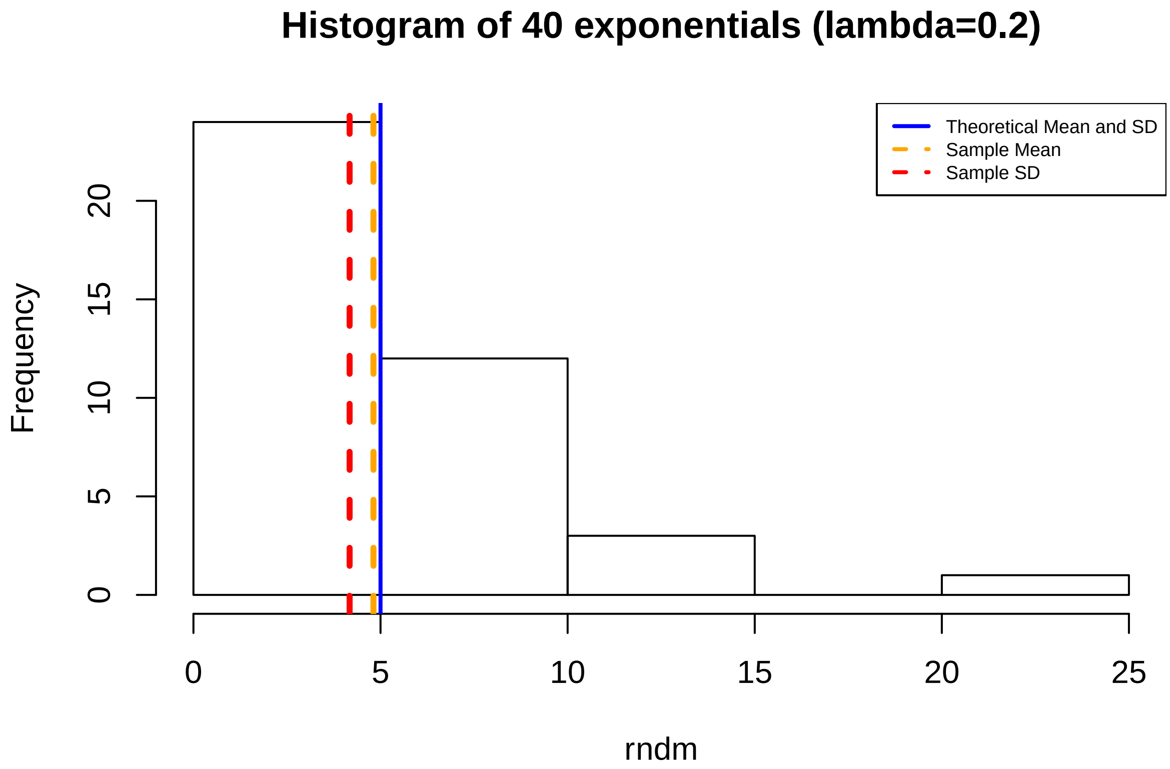
The theoretical Exponential Distribution Mean (lambda=0.2) is at 1/lambda= 5

We are comparing the distribution of 40 random exponentials

```r
rndm <- rexp(40, lambda)
#40 Random exponentials mean
mean(rndm)
```

```
## [1] 4.811212
```

```
hist(rndm, main = "Histogram of 40 exponentials (lambda=0.2)")
abline(v=mean(rndm), lwd=3, col = "orange", lty=2)
abline(v=1/lambda, lwd=2, col = "blue", lty=1)
abline(v=sqrt(var(rndm)), lwd=3, col = "red", lty=2)
legend("topright", legend=c("Theoretical Mean and SD", "Sample Mean", "Sample SD"),
                             col=c("blue", "orange", "red"), lty=c(1,2,2), lwd = 2, cex = 0.6)
```



and the distribution of 1000 averages of 40 random uniforms

```
mns = NULL
for (i in 1 : simuls) mns = c(mns, mean(rexp(n, lambda)))
mean(mns)
```

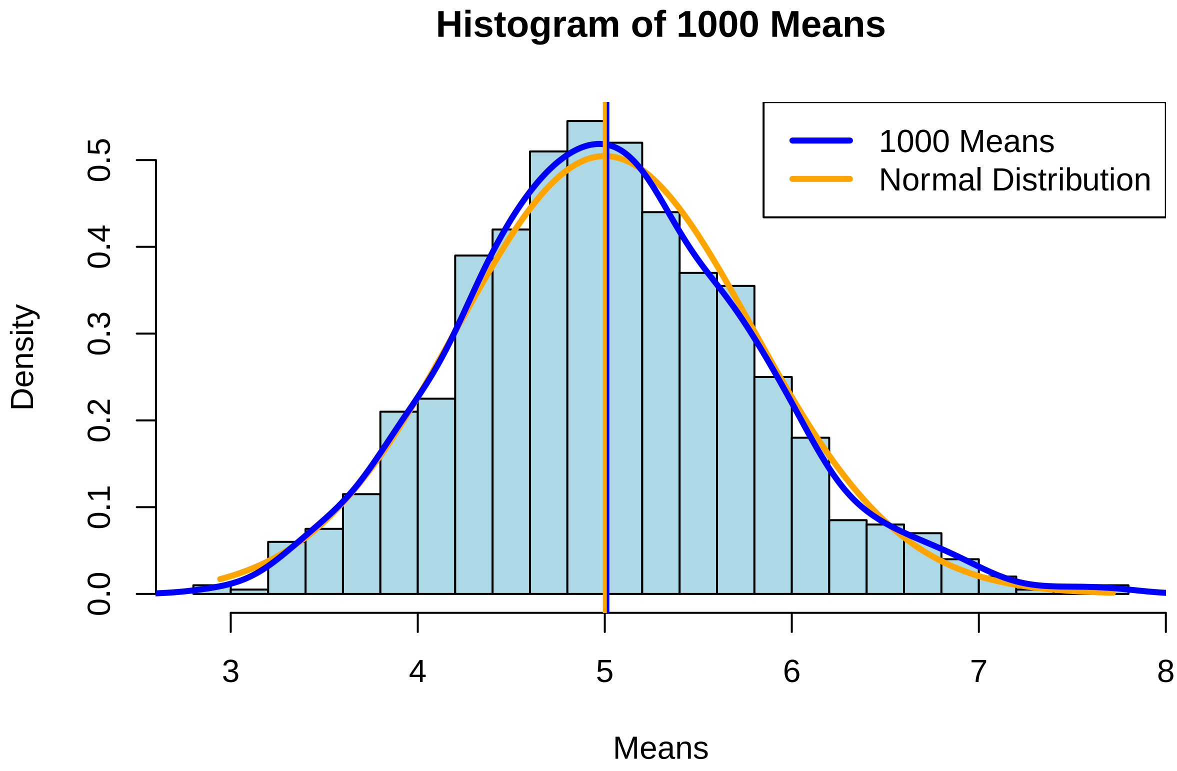
```
## [1] 5.013543
```

So, the theoretical distribution being centered around 5, the distribution of our 40 random exponentials is centered around 4.8112121 when the distribution of the means of 1000 simulations is centered around 5.0135427.

In the next figure we represent in the same histogram the distribution of means of our dataset along with a normal distribution with mean=1/lambda=1/0.2=5 and sd=1/lambda/sqrt(n)=0.7905694

```
hist(mns, breaks = 20, prob = T, col = "light blue", xlab = "Means", main="Histogram of 1000 Means")
x <- seq(min(mns), max(mns), length = 1000)
lines(x, dnorm(x, mean = 1/lambda, sd = (1/lambda/sqrt(n))), lwd=3, col = "orange")
lines(density(mns), lwd=3, col = "blue")
```

```
abline(v=mean(mns), lwd=2, col = "blue")
abline(v=1/lambda, lwd=2, col = "orange")
legend("topright", legend=c("1000 Means", "Normal Distribution"),
                            col=c("blue", "orange"), lty=1, lwd = 3)
```



Sample Variance versus Theoretical Variance:

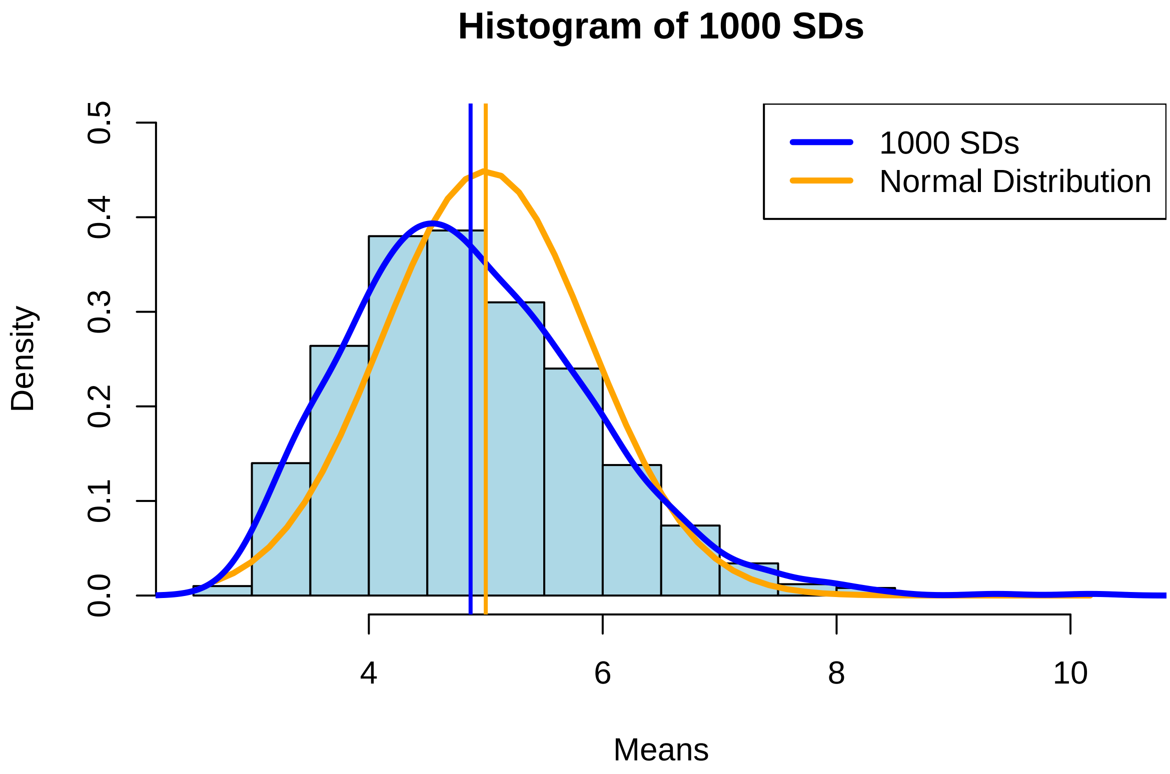
- 40 Random exponentials RD

but for the distribution of 1000 averages of 40 random uniforms the variance is:

```
sds = NULL
for (i in 1 : 1000) sds = c(sds, sqrt(var(rexp(n, lambda))))
mean(sds)
```

```
## [1] 4.870205
```

```
hist(sds, breaks = 20, prob = T, col = "light blue", xlab = "Means", main="Histogram of 1000 SDs", ylim=
x <- seq(min(sds), max(sds), length = 50)
lines(x, dnorm(x, mean = 1/lambda, sd = sqrt(1/lambda/sqrt(n))), lwd=3, col = "orange")
lines(density(sds), lwd=3, col = "blue")
abline(v=mean(sds), lwd=2, col = "blue")
abline(v=1/lambda, lwd=2, col = "orange")
legend("topright", legend=c("1000 SDs", "Normal Distribution"),
       col=c("blue", "orange"), lty=1, lwd = 3)
```

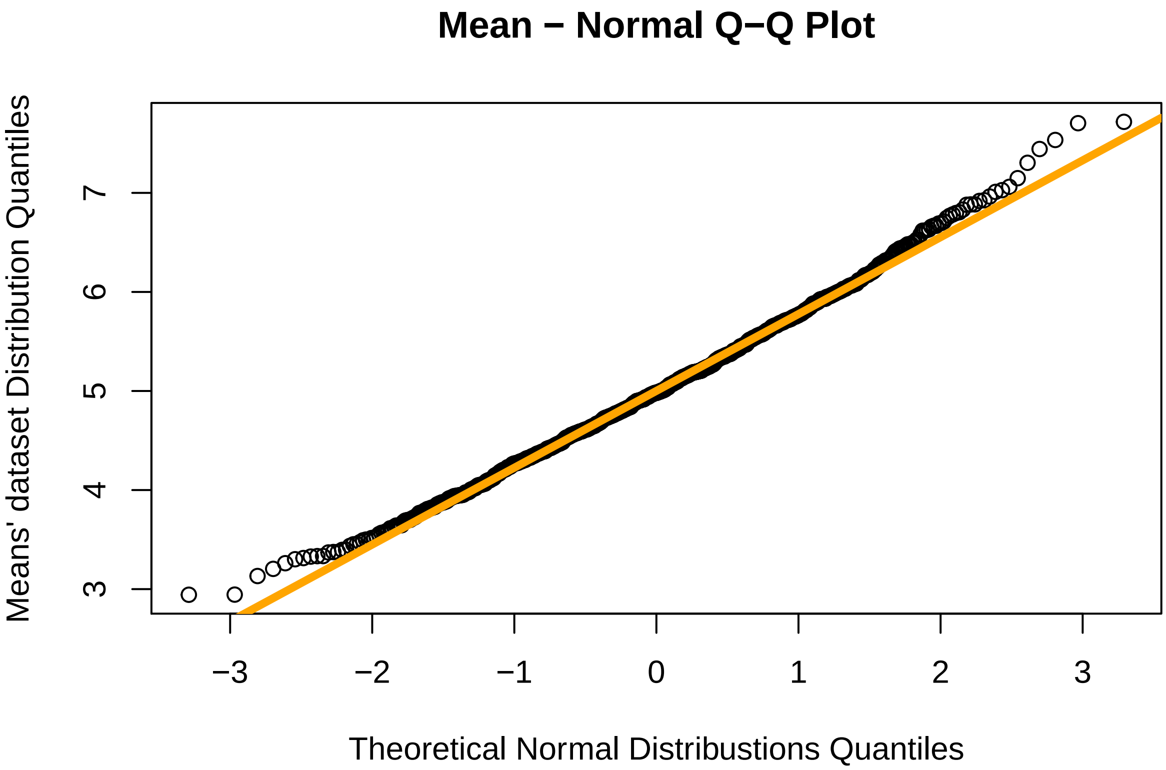


# Distribution: The convergence of distribution of the mean an the sd to the normal distribution, as we approaching big numbers of observations (e.g. simulations), has been already appeared.

Two QQ plots of the two datasets make this fact even clearer. The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set (https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm).

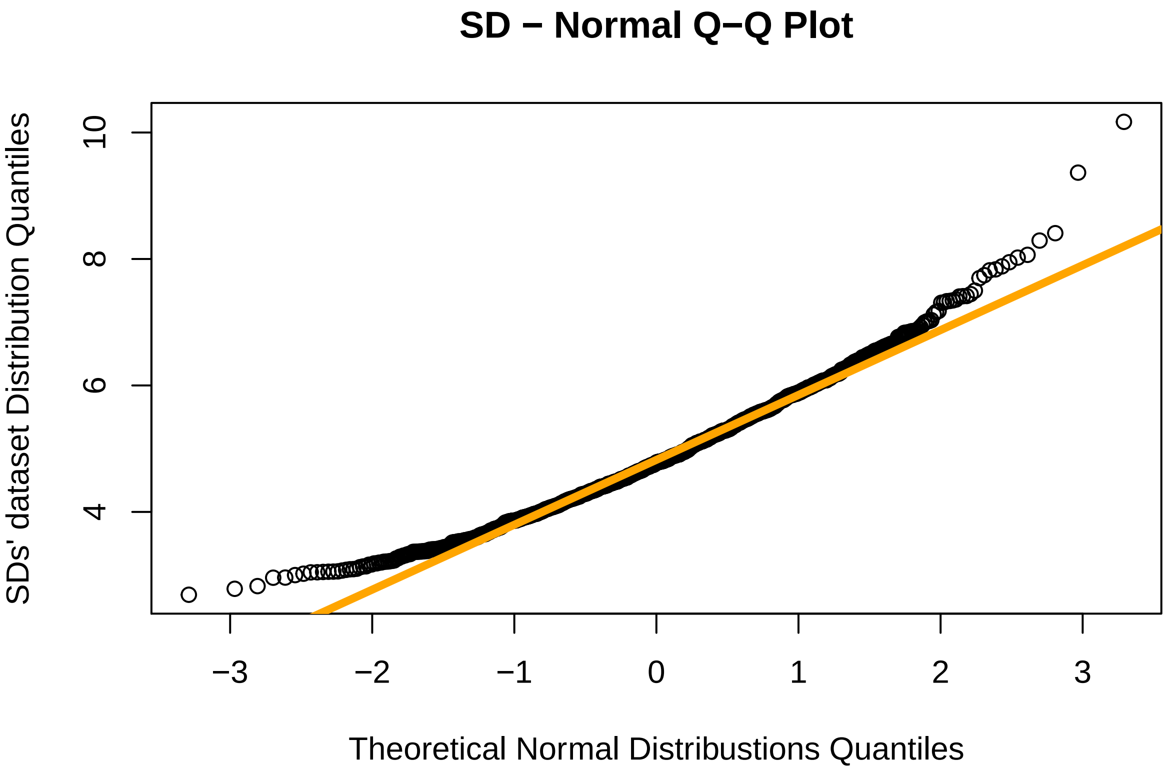
- Means' dataset and normal distribution:

```
qqnorm(mns, main = "Mean - Normal Q-Q Plot",
       xlab = "Theoretical Normal Distribustions Quantiles", ylab = "Means' dataset Distribution Quantil
qqline(mns, lwd=4, col = "orange")
```



- SDs’ dataset and normal distribution:

```
qqnorm(sds, main = "SD - Normal Q-Q Plot",
       xlab = "Theoretical Normal Distribustions Quantiles", ylab = "SDs’ dataset Distribution Quantile
qqline(sds, lwd=4, col = "orange")
```


SD − Normal Q−Q Plot
SDs' dataset Distribution Quantiles
4
6
8
10
−3
−2
−1
0
1
2
3
Theoretical Normal Distribustions Quantiles