Coursera Statistical Inference Project Part 1

pankaj sharma

decemeber 24 2015

Synopsis

This is the project for the statistical inference coursera data science specialization class.This project consists of two parts :

- A simulation exercise.
- Basic inferential data analysis.

Task

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also also 1/lambda. Set lambda = 0.2 for all of the simulations. In this simulation, you will investigate the distribution of averages of 40 exponential(0.2)s. Note that you will need to do a thousand or so simulated averages of 40 exponentials.

```r
# set seed for reproducability
set.seed(31)

# set lambda to 0.2
lambda <- 0.2

# 40 samples
n <- 40

# 1000 simulations
simulations <- 1000

# simulate
simulated_exponentials <- replicate(simulations, rexp(n, lambda))

# calculate mean of exponentials
means_exponentials <- apply(simulated_exponentials, 2, mean)
```

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential(0.2)s. You should:

Question 1

Show where the distribution is centered at and compare it to the theoretical center of the distribution.

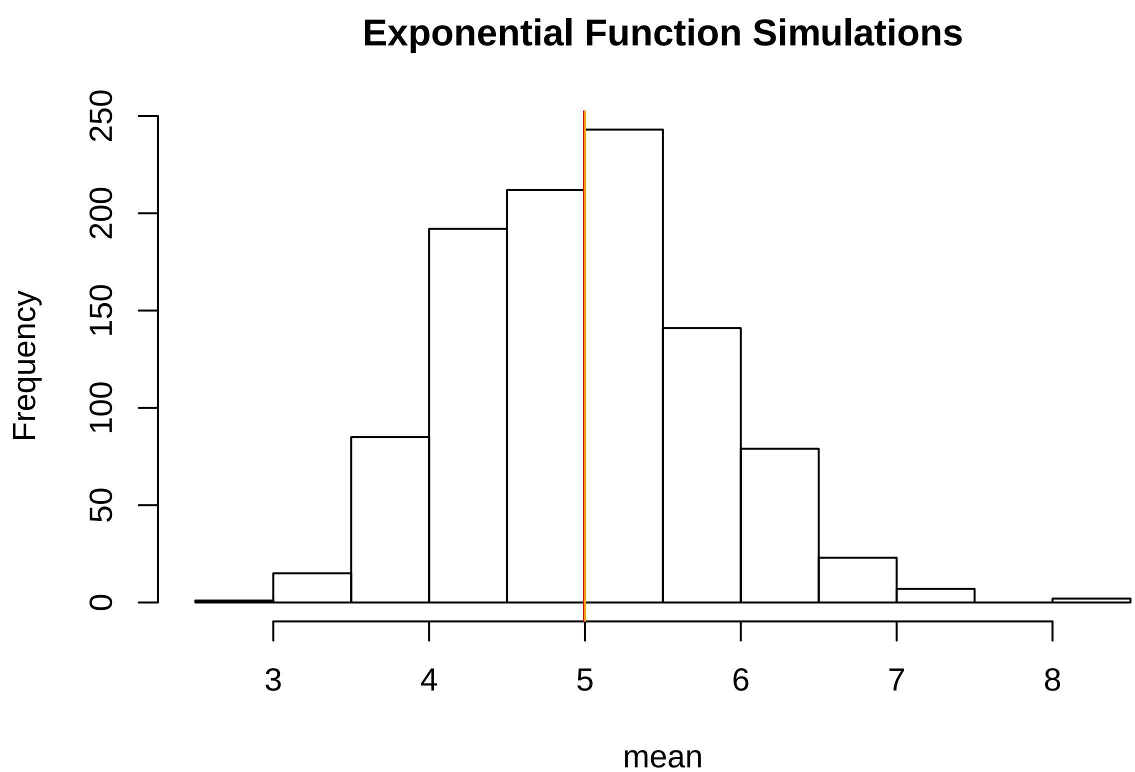
```r
# distribution mean
analytical_mean <- mean(means_exponentials)
analytical_mean
```

```
## [1] 4.993867
```

```r
# analytical mean
theory_mean <- 1/lambda
theory_mean
```

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

```r
# visualization
hist(means_exponentials, xlab = "mean", main = "Exponential Function Simulations")
abline(v = analytical_mean, col = "red")
abline(v = theory_mean, col = "orange")
```



The analytics mean is 4.993867 the theoretical mean 5. The center of distribution of averages of 40 exponentials is very close to the theoretical center of the distribution.

Question 2

Show how variable it is and compare it to the theoretical variance of the distribution.

```r
# standard deviation of distribution
standard_deviation_dist <- sd(means_exponentials)
standard_deviation_dist
```

```
## [1] 0.7931608
```

```r
# standard deviation from analytical expression
standard_deviation_theory <- (1/lambda)/sqrt(n)
standard_deviation_theory
```

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

```r
# variance of distribution
variance_dist <- standard_deviation_dist^2
variance_dist
```

```
## [1] 0.6291041
```

```r
# variance from analytical expression
variance_theory <- ((1/lambda)*(1/sqrt(n)))^2
variance_theory
```

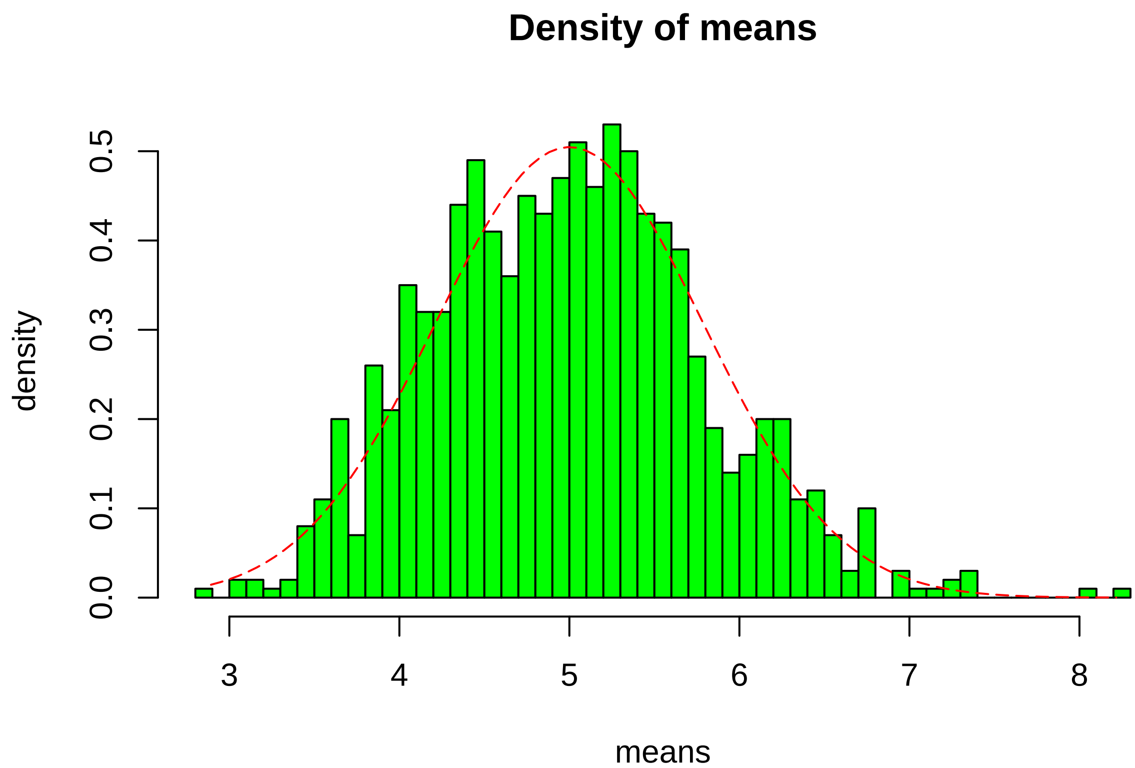
```
## [1] 0.625
```

Standard Deviation of the distribution is 0.7931608 with the theoretical SD calculated as 0.7905694. The Theoretical variance is calculated as 0.625. The actual variance of the distribution is 0.6291041

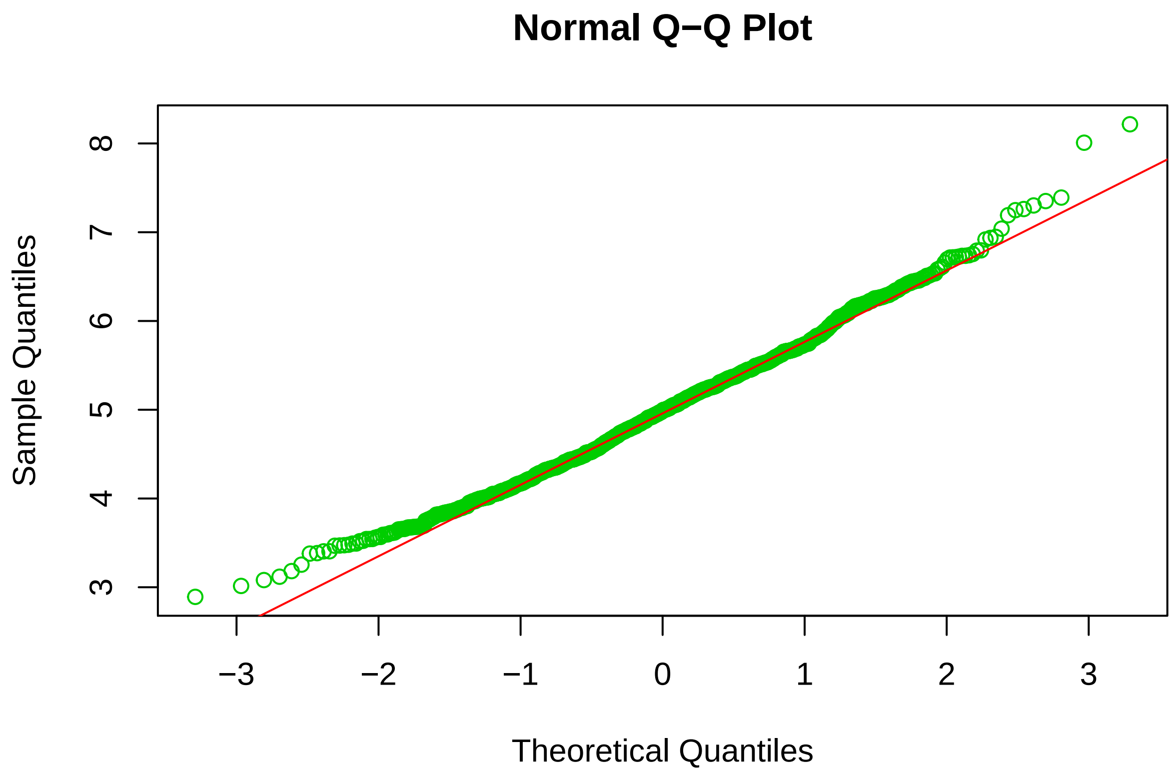
Question3

Show that the distribution is approximately normal.

```r
xfit <- seq(min(means_exponentials), max(means_exponentials), length=100)
yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(n)))
hist(means_exponentials,breaks=n,prob=T,col="green",xlab = "means",main="Density of means",ylab="densit
lines(xfit, yfit, pch=22, col="red", lty=5)
```




```r
# compare the distribution of averages of 40 exponentials to a normal distribution
qqnorm(means_exponentials,col=3)
qqline(means_exponentials, col = 2)
```



Due to Due to the central limit theorem (CLT), the distribution of averages of 40 exponentials is very close to a normal distribution.