Statistical Inference Part 1

Swee Yean

July 18, 2016

Statistical Inference Course Project 1 : A simulation exercise

In this project I will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda λ is the rate parameter. The mean of exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$. Set λ = 0.2 for all of the simulations. I will investigate the distribution of averages of 40 exponentials. Note that I will need to do a thousand simulations.

Simulations

load neccesary libraries

```
library(ggplot2)
```

set constants

```
lambda <- 0.2# lambda for rexp
sample_size <- 40 # 40 samples drawn from the exp distribution
sim_cnt <- 1000 # number of tests
```

Set the seed for reproducability

```
set.seed(567)
```

run the test resulting in n x sim matrix

```
sim <- matrix(data=rexp(sample_size * sim_cnt, lambda), nrow=sim_cnt)
```

Take the mean for each row

```
sim_means <- rowMeans(sim)
```

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

Sample Mean versus Theoretical Mean

The theoretical mean t_mean of a exponential distribution of rate λ is t_mean = $1/\lambda$

```
t_mean <- 1/lambda
t_mean
```

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

The sample mean. Let $\bar{X}$ be the average sample mean of 1000 simulations of 40 randomly sampled exponential distributions.

```
meanOfMeans <- mean(sim_means)
meanOfMeans
```

```
## [1] 4.9896
```

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

The theoretical standard deviation t_sd of a exponential distribution of rate λ is t_sd = 1/λ/sqrt(n)

```
t_sd <- 1/lambda/sqrt(sample_size) #The theoretical  standard deviation of the distribution
t_sd
```

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

The Variance of the theoretical standard deviation of the distribution Exp_Var = t_sd^2

```
Exp_Var = t_sd^2 #The theoretical  standard deviation  variance of the distribution
Exp_Var
```

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

The Sample standard deviation

```
ssd <- sd(sim_means)
ssd
```

```
## [1] 0.8015858
```

The variance Var of Sample standard deviation of the distribution t_sd is Var = $sd(sim\_means)^2$

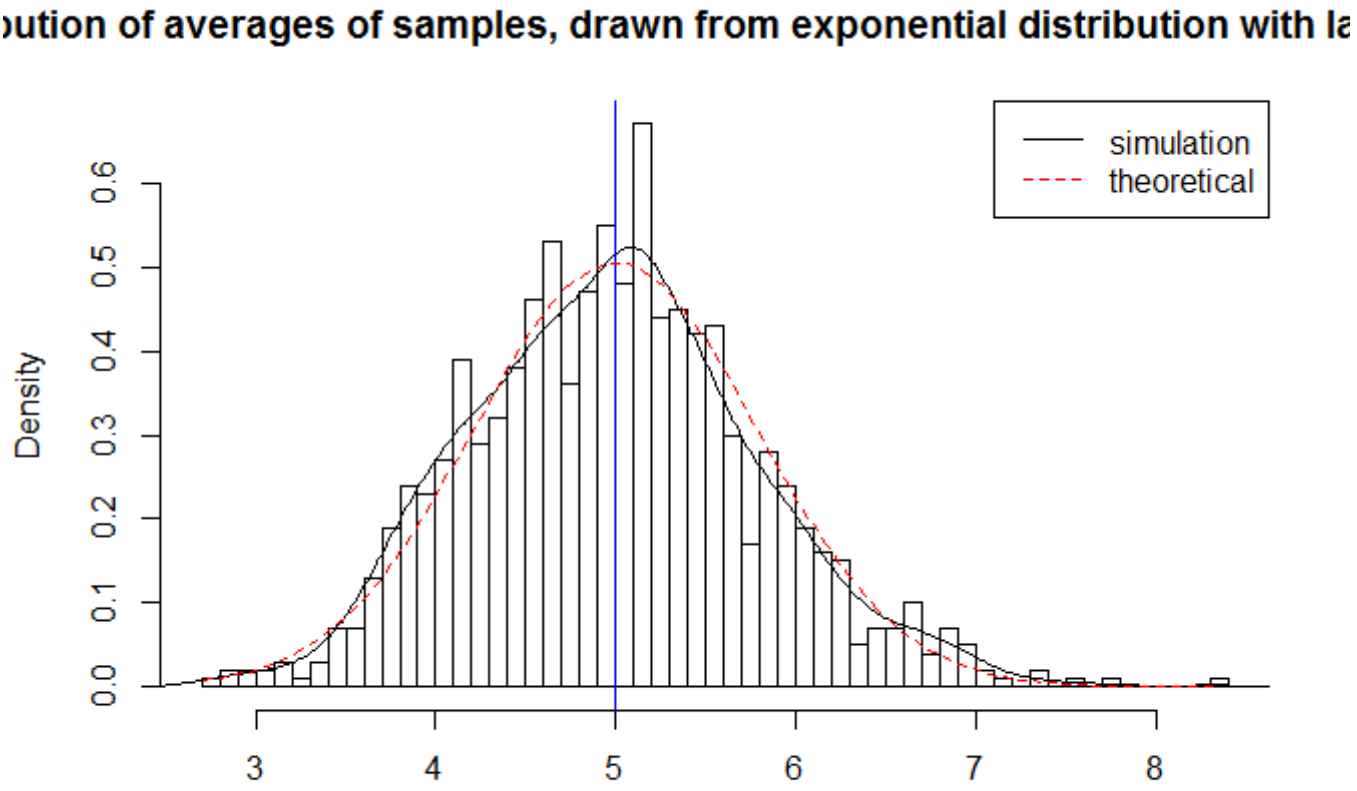
```
Var <- sd(sim_means)^2
Var
```

```
## [1] 0.6425399
```

Question 3. Show that the distribution is approximately normal.

plot the histogram of averages

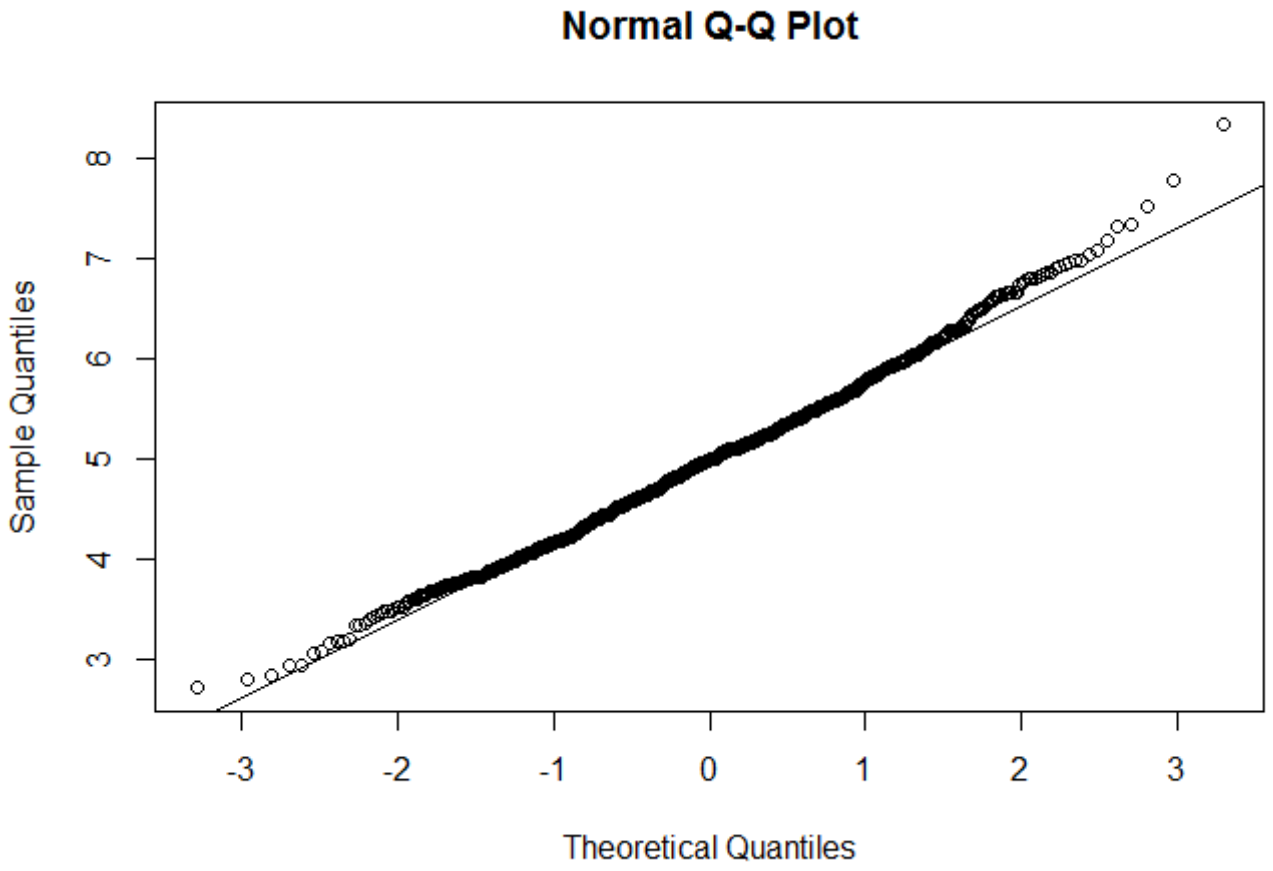
```
hist(sim_means, breaks=50, prob=TRUE, main="Distribution of averages of samples, drawn from e
xponential distribution with lambda=0.2", xlab="")

lines(density(sim_means))

abline(v=1/lambda, col="blue")

xfit <- seq(min(sim_means), max(sim_means), length=100)
yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(sample_size)))
lines(xfit, yfit, pch=22, col="red", lty=2)

legend('topright', c("simulation", "theoretical"), lty=c(1,2), col=c("black", "red"))
```



The distribution of sample means is centered at 4.9896 and the theoretical center of the distribution is λ^{-1} = 5. The variance of sample means is 0.6425399 where the theoretical variance of the distribution is $\sigma^2/n = 1/(\lambda^2 n) = 1/(0.04 \times 40)$ = 0.625.

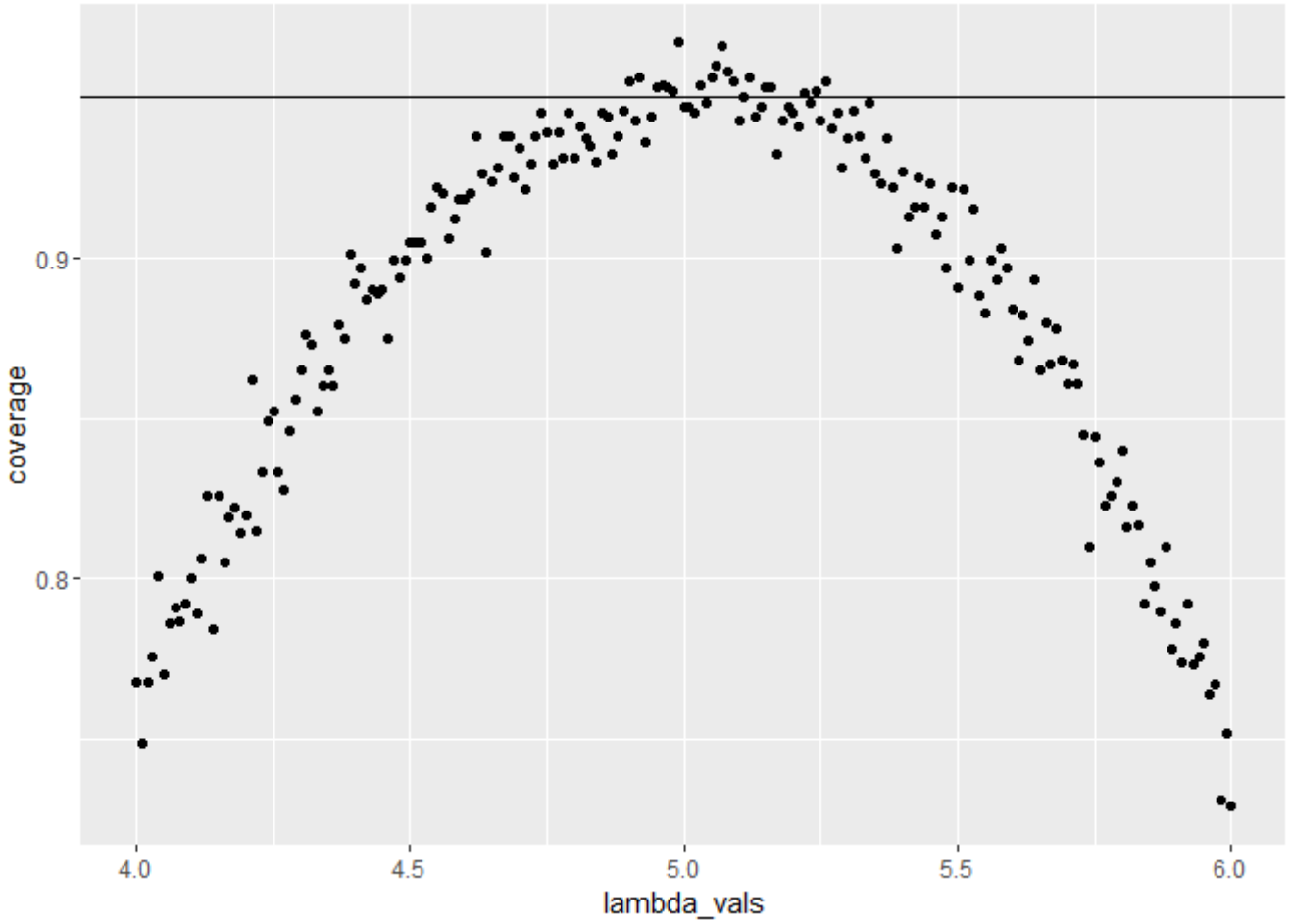
Due to the central limit theorem, the averages of samples follow normal distribution. The figure above also shows the density computed using the histogram and the normal density plotted with theoretical mean and variance values. Also, the q-q plot below suggests the normality.

```
qqnorm(sim_means);qqline(sim_means)
```



Finally, let's evaluate the coverage of the confidence interval for $1/\lambda = \bar{X} \pm 1.96\frac{S}{\sqrt{n}}$

```
lambda_vals <- seq(4, 6, by=0.01)
coverage <- sapply(lambda_vals, function(lamb) {
    mu_hats <- rowMeans(matrix(rexp(sample_size*sim_cnt, rate=0.2),
                               sim_cnt, sample_size))
    ll <- mu_hats - qnorm(0.975) * sqrt(1/lambda**2/sample_size)
    ul <- mu_hats + qnorm(0.975) * sqrt(1/lambda**2/sample_size)
    mean(ll < lamb & ul > lamb)
})


qplot(lambda_vals, coverage) + geom_hline(yintercept=0.95)
```



The 95% confidence intervals for the rate parameter (λ) to be estimated ($\hat{\lambda}$) are $\hat{\lambda}low = \hat{\lambda}(1 - \frac{1.96}{\sqrt{n}})$ agnd $\hat{\lambda}upp = \hat{\lambda}(1 + \frac{1.96}{\sqrt{n}})$. As can be seen from the plot above, for selection of $\hat{\lambda}$ around 5, the average of the sample mean falls within the confidence interval at least 95% of the time. Note that the true rate, λ is 5.