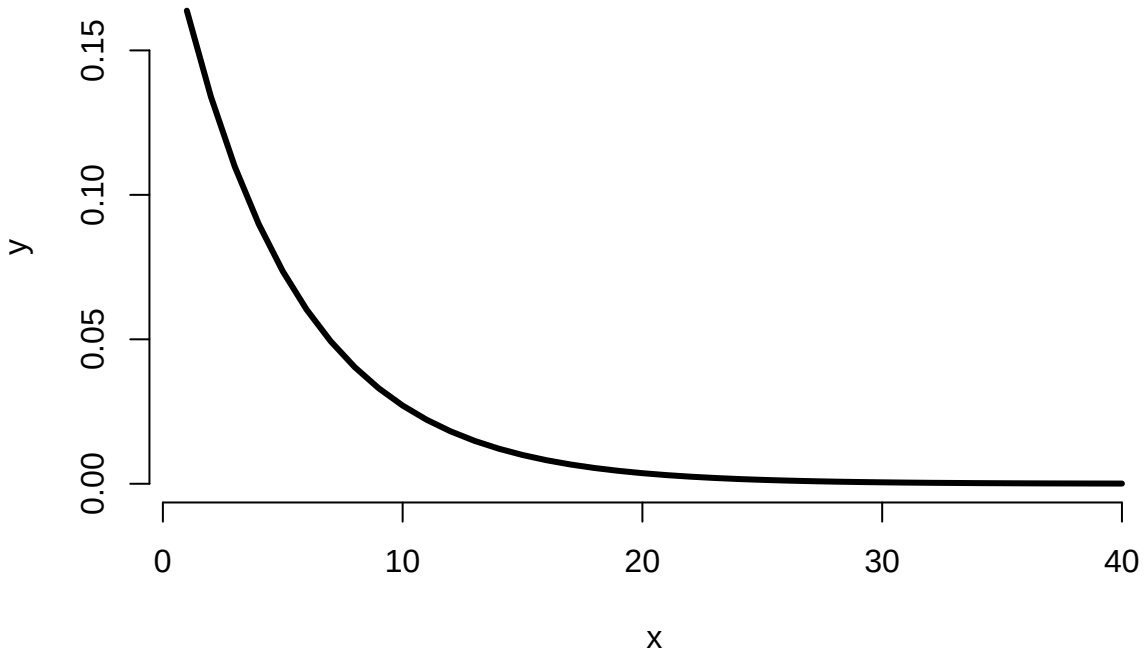
Statistical Inference, Course Project, Part 1

Simulation of exponential distribution

Part 1 of the course project is about analyzing exponential distribution using Central Limit Theorem. For this analysis 40 (n = 40) exponentials will be simulated 1000 times and cample mean and variance collected from this simulations will be compated to theoretical mean and variance of distribution. Exponential distribution is the probability distribution that describes the time between events in a process in which events occur continuously and independently at a constant average rate [1] (in all of our simulations this rate (lambda) will be set to 0.2). Its PDF is $f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$ which on the graph looks like this:



Theoretical mean of this distribution is $1/\lambda$, which, in case of $\lambda = 0.2$ has the value of 5. After running 1000 simulations (nosim = 1000)

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

We can see, that sample means vary a lot

```
## [1] 3.685888 4.441841 5.310328 3.544193 3.570054 6.022497
```

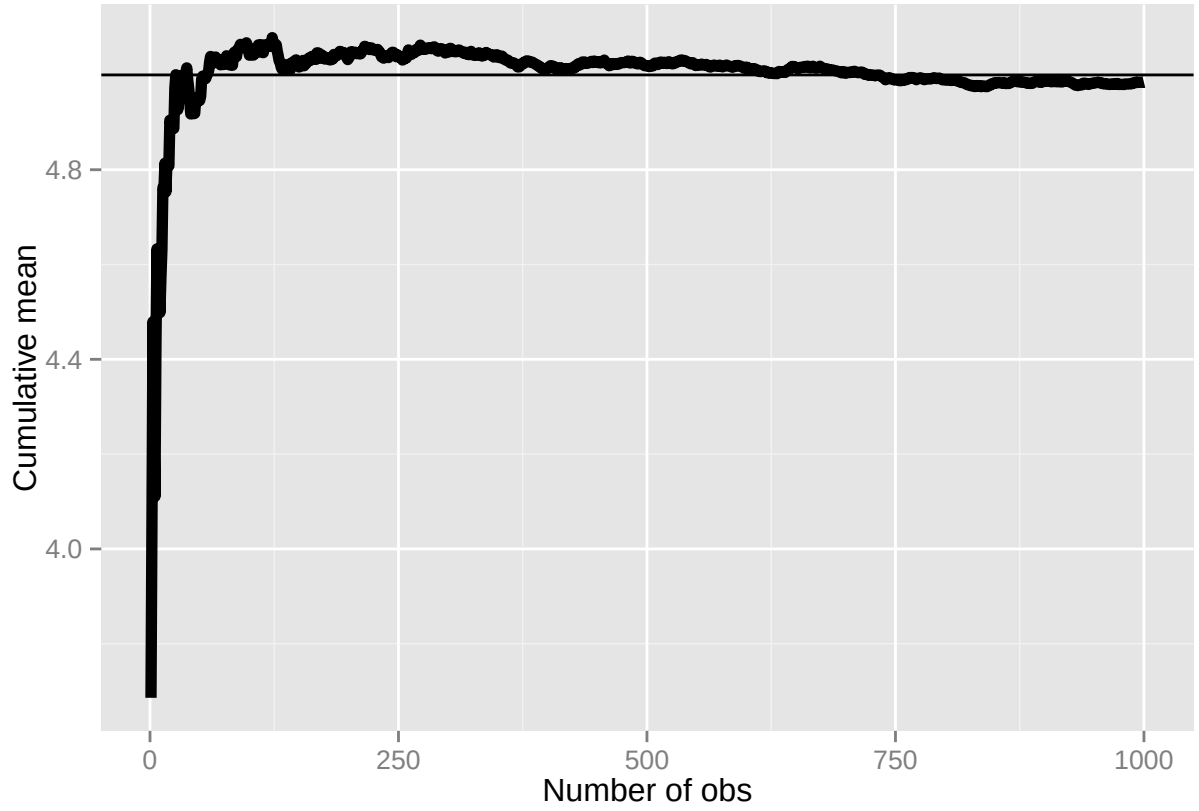
ranging from 3.0668726 to 7.6274949, which is obvious, as the sample mean depends on the sample. But if we now take mean of this vector

```
mean(means)
```

```
## [1] 4.984611
```

this will give us the number which is quite close to our theoretical mean value of 5. If we increase the number of simulations we will get even better estimation of sample mean.

For example after 100000 simulations we will get 4.9979513. And this makes sense as our sample mean is trying to estimate the population mean and more data we collect (or simulate in this case) more precise this estimation become. We can illustrate this using cumulative mean



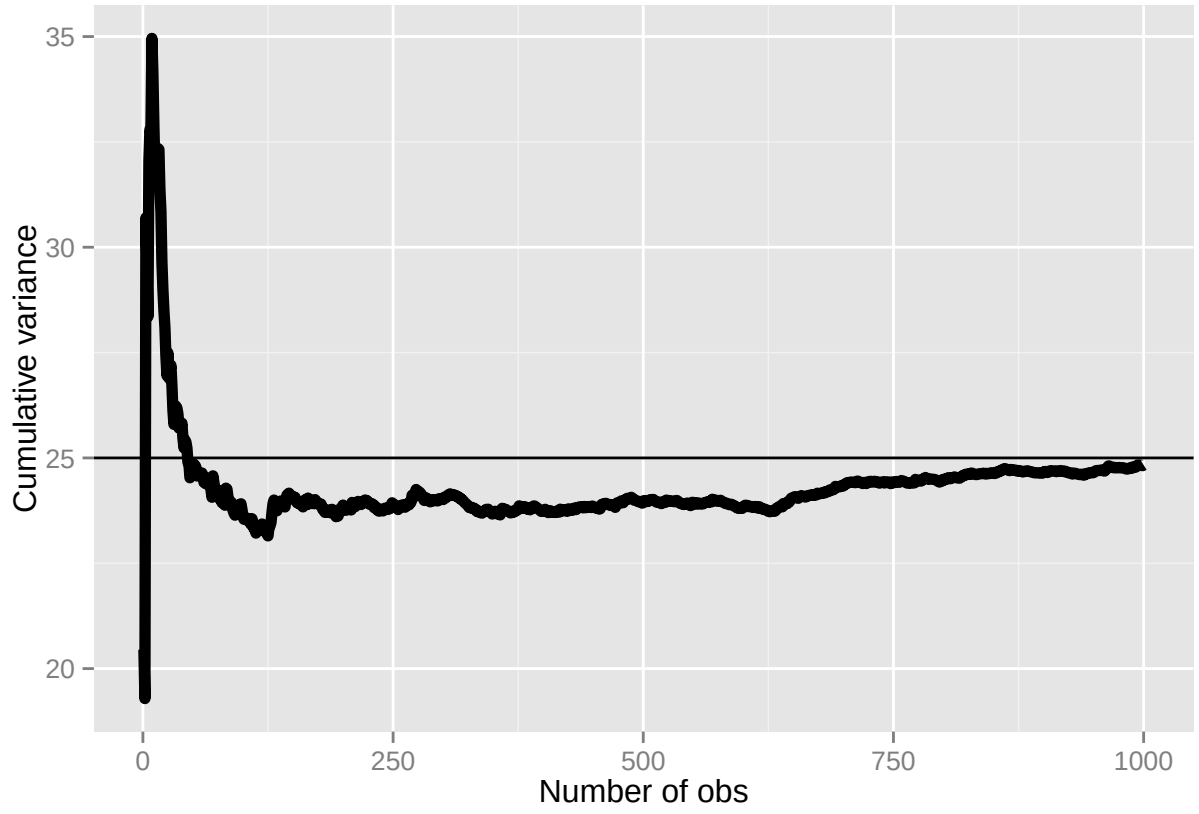
As can be seen from this graph, as we include more and more simulations, cumulative mean is approaching theoretical mean.

Theoretical variance of exponential distribution is $1/\lambda^2$, so in case of $\lambda = 0.2$ variance $= 25$

```
## [1] 20.44703 18.13000 53.51853 26.26685 23.40014 50.57341
```

Now if we do basically the same calcualtions as we did previously for the mean, we will get the vector of variances, which will vary from 5.8898303 to 89.0077318. But if we calculate the average of this vector, we will get 24.8316377, which is pretty close to our theoretical variance.

Again, if we enhance the number of simulations to 100000 we will get even closer in our estimations: var $=$ 25.0066529. And this can be further illustrated with the help of cumulative variance:



To show that distribution of averages of 40 exponentials is approxiamtely normal, we first need to normalize the data by subtracting the mean, dividing by standard deviation and multiplying by the square root of sample size:

```
means_norm <- NULL
for (i in 1 : nosim) means_norm = c(means_norm, sqrt(n)*(means[i] - mean(means))/(1/lambda))
```

If we did this right, mean of this new vector should be 0, and standard deviation should be 1. We can check this:

```
mean(means_norm)
```

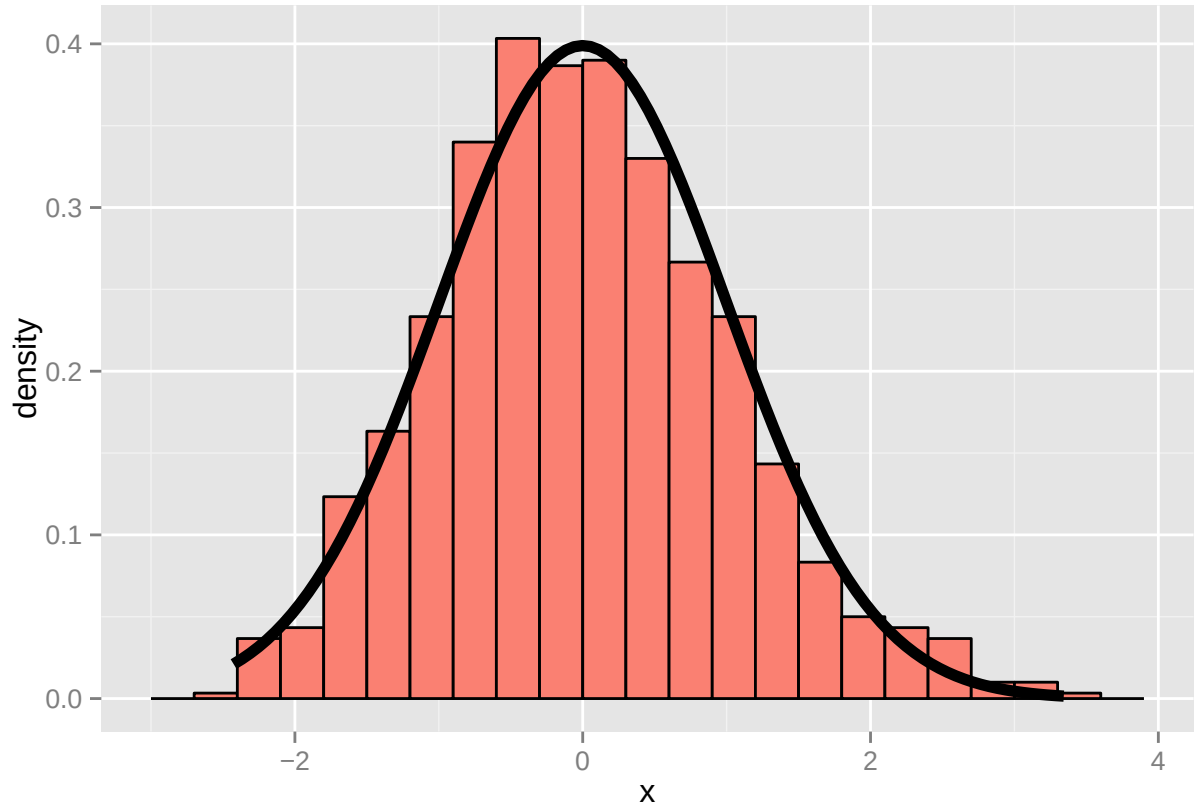
```
## [1] -1.887515e-16
```

```
sd(means_norm)
```

```
## [1] 1.003617
```

Now we can plot a histogram of the averages and a normal distribution on top of it:

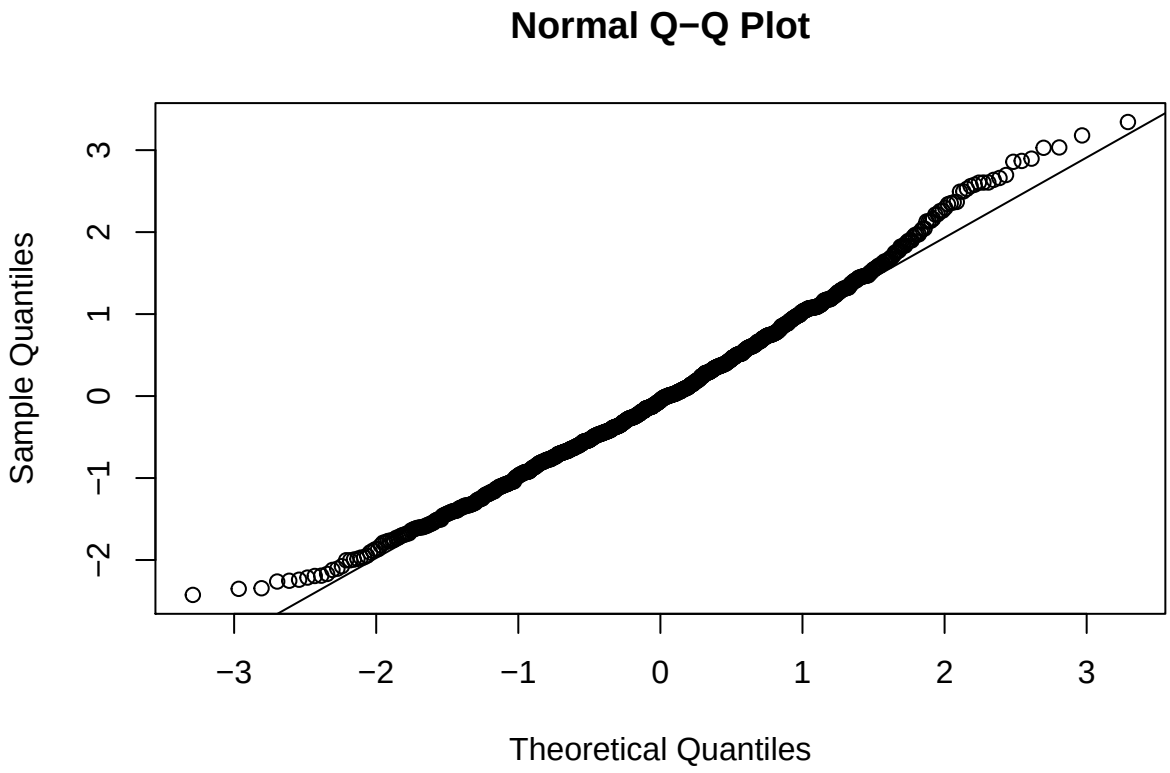
```
g <- ggplot(data.frame(x = means_norm), aes(x = x))
g <- g + geom_histogram(fill = 'salmon', colour = 'black', binwidth = 0.3, aes(y = ..density..))
g <- g + stat_function(fun = dnorm, size = 2)
g
```



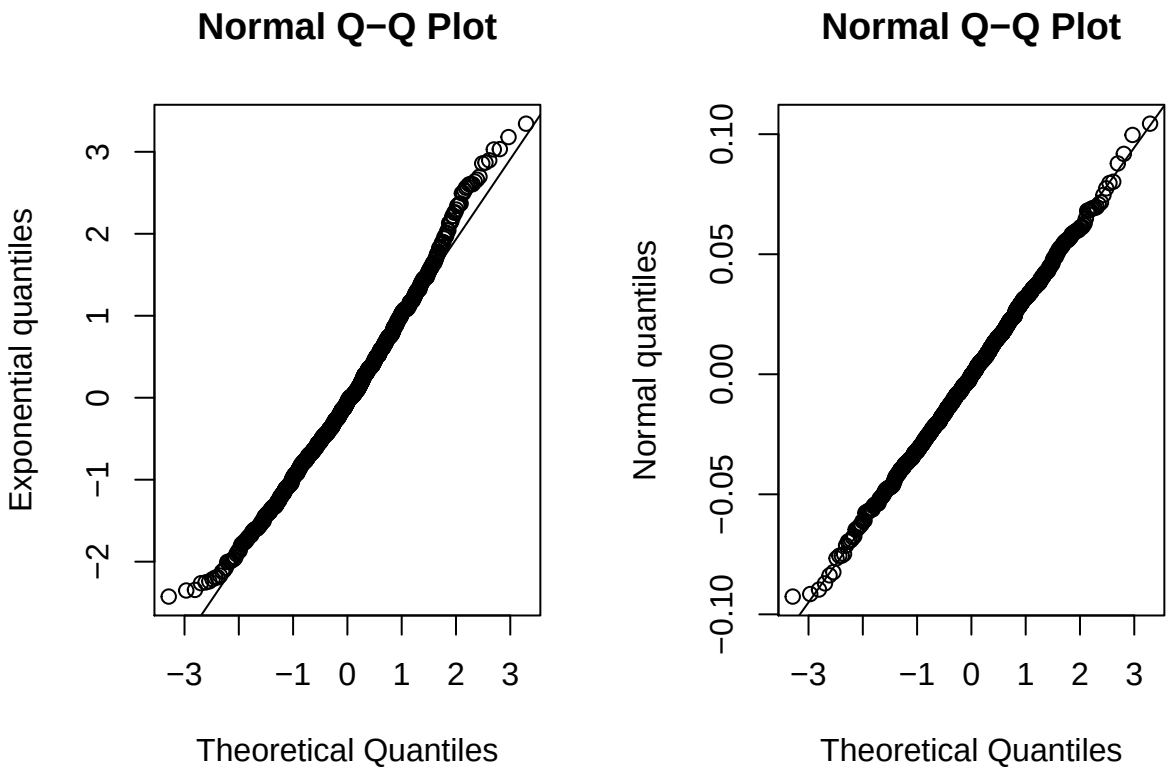
As we can see, it fits quite well.

However, we can go a little bit further and make a Q-Q plot (this is a method of comparing two probability distributions by plotting their quantiles against each other [2]):

```
qqnorm(means_norm)
qqline(means_norm) #this will add a line for normal distribution
```



If we compare it to normal vs normal:



It is obvious from this comparing, that distribution of averages of 40 exponentials is approximately normal.

References

[1] http://en.wikipedia.org/wiki/Exponential_distribution

[2] http://en.wikipedia.org/wiki/Q%E2%80%93Q_plot