Statistical Inference - final project

Wiktoria Urantowka

8/11/2017

The goal of this project is the applications of Central Limit THeorem to exponential distribution in R. The last one is simulated in R using command rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is the same as it's standard deviation and amounts 1/lambda.

A. Simulations

1. setting constants

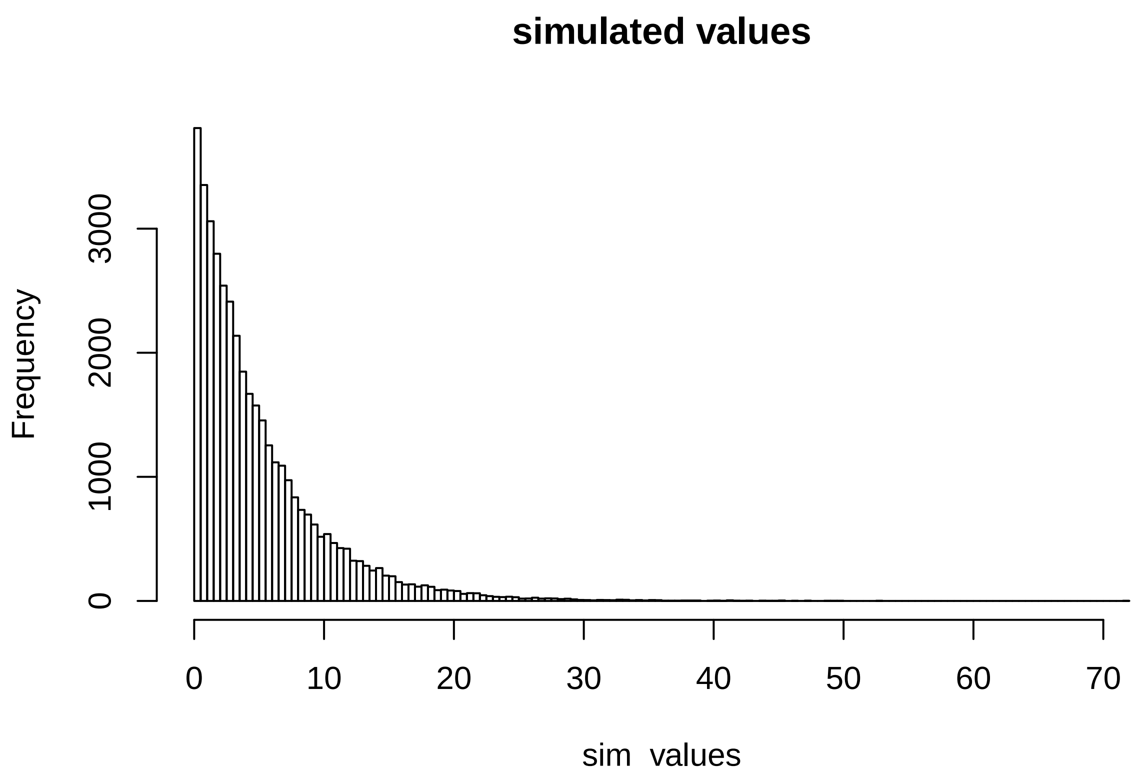
```
lambda <- 0.2 # lambda
n <- 40 # size of the sample
sim_num <- 1000 # number of drows
```

2. Setting seed for reproducibility

```
set.seed(67893)
```

3. Simulating values from exponential distribution

```
sim_values<- rexp(n = sim_num * n,lambda)
hist(sim_values, main = "simulated values", breaks =200)
```



4. Create a dataframe with means of 1000 simulation of size n=40

```
matrix <- matrix(rexp(n = sim_num * n,lambda), sim_num, n)
means <- rowMeans(matrix)
simData <- data.frame(cbind(matrix,means))
```

B. Comparaison of theoretical and sample means of the distribution.

a. Theoretical mean of exponential distribution = 1/lambda

```
theoretical_mean <- 1/lambda
theoretical_mean
```

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

b. Mean from one drow of n=40 is given by:

```
sample_mean <-mean(rexp(40, 0.2))
sample_mean
```

```
## [1] 3.995102
```

and the difference between the two is small

```
mean_difference<-abs(theoretical_mean-sample_mean)
mean_difference
```

```
## [1] 1.004898
```

c. However, when we repeat it 1000 times and compute mean,

```
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(rexp(40, 0.2)))
sample_mean1000<- mean(mns)
mean_difference2<-theoretical_mean-sample_mean1000
sample_mean1000
```

```
## [1] 4.954971
```

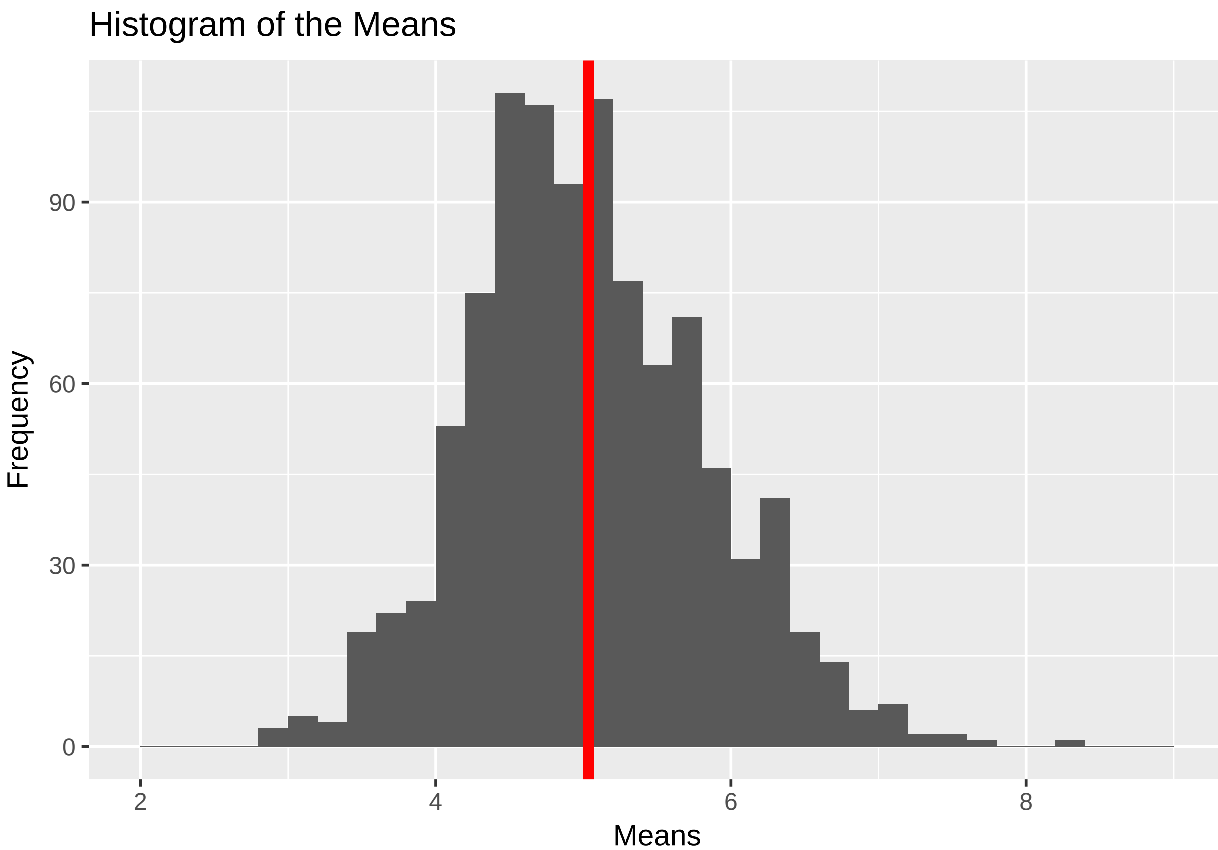
```
mean_difference2
```

```
## [1] 0.04502878
```

One can see, that the difference between sample mean and theoretical mean shrinks further

d. Plotting the results and finding their mean value

```
library("ggplot2")
ggplot(data = simData, aes(simData$means)) +
  geom_histogram(breaks = seq(2, 9, by = 0.2)) +
  labs(title = "Histogram of the Means", x = "Means", y = "Frequency") +
  geom_vline(aes(xintercept=mean(simData$means)), color="red",
 size=2)
```



The histogram proves that the distribution of the means is centered around the mean of the simulated distribution, very very close to the theoretical mean = 5

2. Comparaison of theoretical and sample variance of the distribution.

a. Theoretical variance of exponential distribution is given by:

```
theoretical_variance = ((1 / lambda) ^ 2) / n
theoretical_variance
```

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

b. sample variance is given by

```
sample_variance<-var(means)
sample_variance
```

```
## [1] 0.6412434
```

```
var_difference <- theoretical_variance-sample_variance
var_difference
```

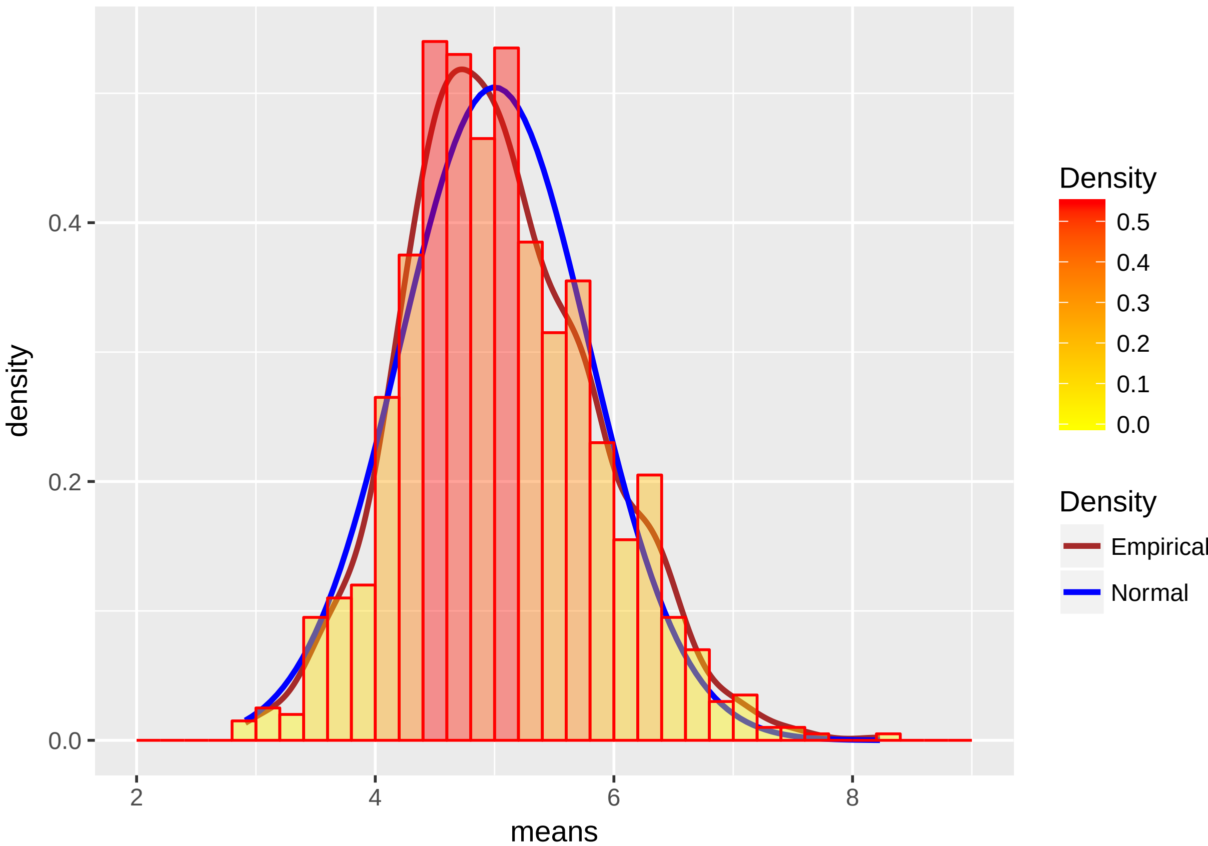
```
## [1] -0.01624341
```

One can see, that the variance difference between the theoretical and sample variance is very small

Show that the distribution is approximately normal.

Showing that the simulated mean sample data approximately follows the Normal distribution when n is big (Central Limit Theorem) can be illustrated below: (Comparison of the plotted sample data with the plot of normal distribution)

```
qplot(means, geom = 'blank') +
  geom_line(aes(y=..density.., colour='Empirical'), stat='density', size=1) +
  stat_function(fun=dnorm, args=list(mean=(1/lambda), sd=((1/lambda)/sqrt(n))),
                aes(colour='Normal'), size=1) +
  geom_histogram(aes(y=..density.., fill=..density..), alpha=0.4,
                 breaks = seq(2, 9, by = 0.2), col='red') +
  scale_fill_gradient("Density", low = "yellow", high = "red") +
  scale_color_manual(name='Density', values=c('brown', 'blue'))
```



Comparing the “Empirical” (derived from exponential distribution) and “Normal” distributions clearly illustrates, that the sample data can be approximated with the normal distribution.