Investigation of Exponential Distribution

Course Project for **Statistical Inference** classes on Coursera

Igor Goltsov riversy@gmail.com

This is my own investigation of Exponential Distribution's behaviour created as Course Project for classes "**Statistical Inference**" on Coursera. In this work I will create a population of random values using R tools. I would like to investigate how the theory of Central Limited Theorem correlates with randomly created population and it's samples.

0. Prepare data

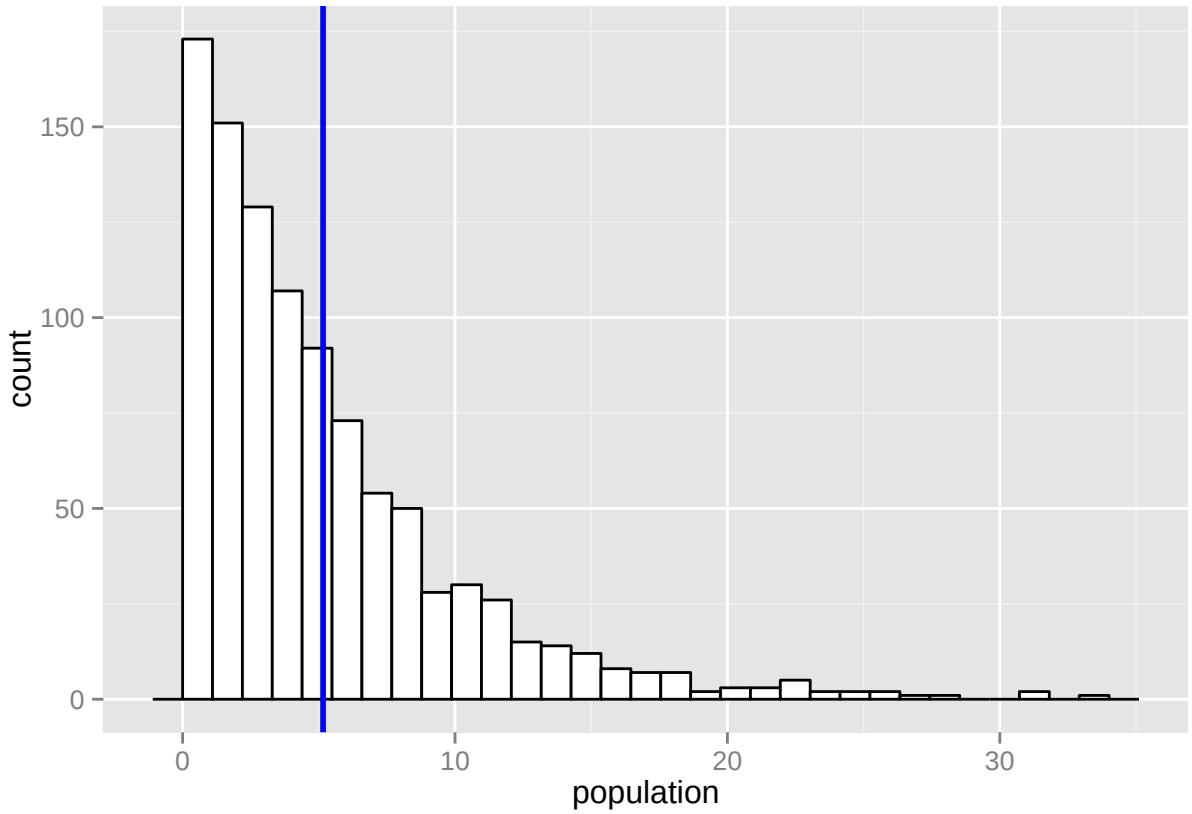
First of all I would like to generate a population of data. I will use R function *rexp()*. I will also use function *set.seed()* to create my random population reproducible. Rate **lambda** is defined in description of Course Project and should equals *0.2* and population size is defined as **n** that equals to *1000*.

```
lambda <- 0.2
n <- 1000
sample_n <- 40

set.seed(1)
population <- rexp(n, rate = lambda)
```

Let's have a look onto histogram of the theoretical population. I will use **ggplot2** library to build the plot. On that histogram I will also show the position of the **mean** for that population.

```
library(ggplot2)
ggplot(NULL,
       aes(x = population)
    ) +
    geom_histogram(fill = "white", color = "black") +
    geom_vline(xintercept = mean(population), color = "blue", size = 1)
```



I would like to define the *mean* of that theoretical population.

```
population_mean <- mean(population)
population_mean
```

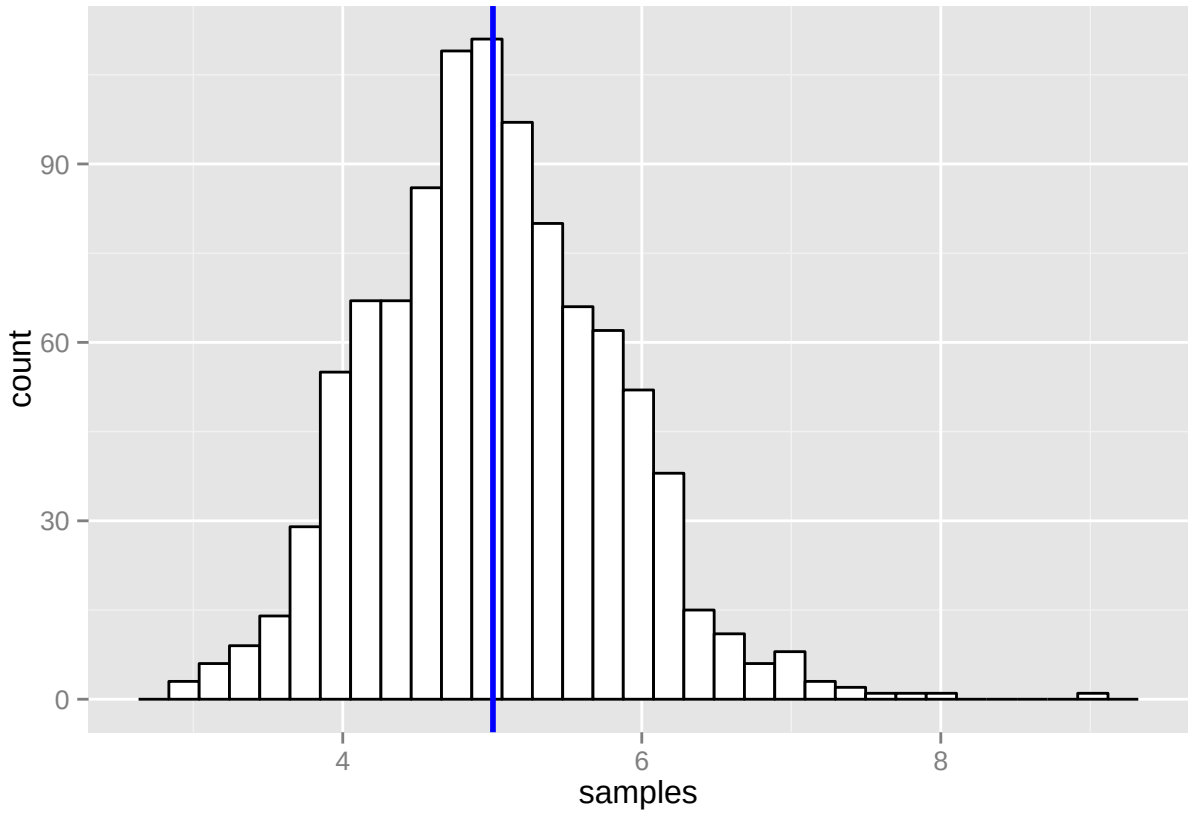
```
## [1] 5.156513
```

1. Sample Means

I would like to generate 40 samples by 1000 random variables now using same **lambda** value. And check it. I will use *set.seed()* again to make the example reproducible.

```
samples = NULL
for (i in 1:n){
    set.seed(i + 1)
    sample <- rexp(sample_n, rate = lambda)
    samples <- c(samples, mean(sample))
}

ggplot(NULL,
       aes(x = samples)
    ) +
    geom_histogram(fill = "white", color = "black") +
    geom_vline(xintercept = mean(samples), color = "blue", size = 1)
```



The *mean* of that samples will be defined as:

```
samples_mean <- mean(samples)
samples_mean
```

```
## [1] 5.00444
```

If we compare *mean* of theoretical population and sample's mean, we may see it quite different but it pretty close. It's different because we used samples of random variables. And it close due to Central Limit Theorem works here.

```
rbind(
    population_mean,
    samples_mean
)
```

```
##                     [,1]
## population_mean 5.156513
## samples_mean    5.004440
```

2. Sample Variance

In this section I would like to compare the variance of the population of the data versus variance of the populations of samples by 40 objects in the sample.

```r
population_variance <- var(population)
sample_variance <- var(samples)

rbind(
    population_variance,
    sample_variance
)
```

```
##                           [,1]
## population_variance 24.4658326
## sample_variance      0.6346866
```

The variance of the samples was significantly reduced if compate to the theoretical population of data values.

3. Normality of the Distrubution

Large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials has different variances. But in the same time these collections has a same **mean** values. So the distribution is normally distributed according to Cenral Limit Theorem.