Simulation - Course Project - Statistical Inference

attila.toth86

22 Sep 2015

This is the project for the statistical inference class. In this, I use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

This document aims to fulfill requirements of first task.

Simulation

In this project I will investigate the exponential distribution in R and compare it with the Central Limit Theorem.

- *Mean* of exponential distribution is *1/lambda*
- Its *standard deviation* is also *1/lambda*

The exponential distribution can be simulated in R with `rexp(n, lambda)` function, where n represents the number of observations and lambda is the so-called rate parameter. For further analyses, I create a major population of 1 million observation with exponential distribution.

For all simulation I use lambda = 0.2.

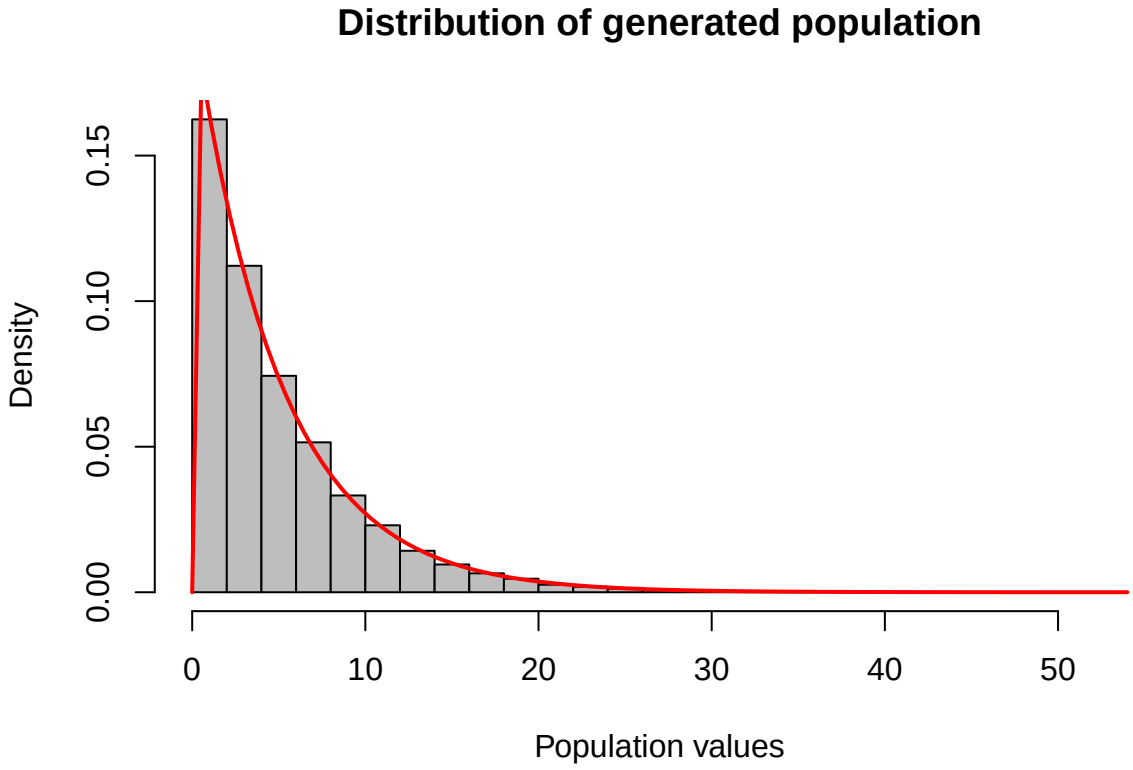
```
lambda <- 0.2
```

In order to ensure reproduciblilty, I set a seed.

```
set.seed(10000)
```

Then I create a random exponential population:

```
pop <- rexp(n = 10000, rate = lambda)
hist(pop,
     main="Distribution of generated population",
     freq=FALSE,
     xlab='Population values',
     breaks=30,
     col=8)
curve(dexp(x,lambda), add=TRUE, col=2, lwd=2)
```



Considering lambda equals to 0.2, literature states that

- **Theoretical mean of such distribution is 5,**
- **and theoretical standard deviation is also: 5** (In addition to that: **theoretical variance is then: 25**)

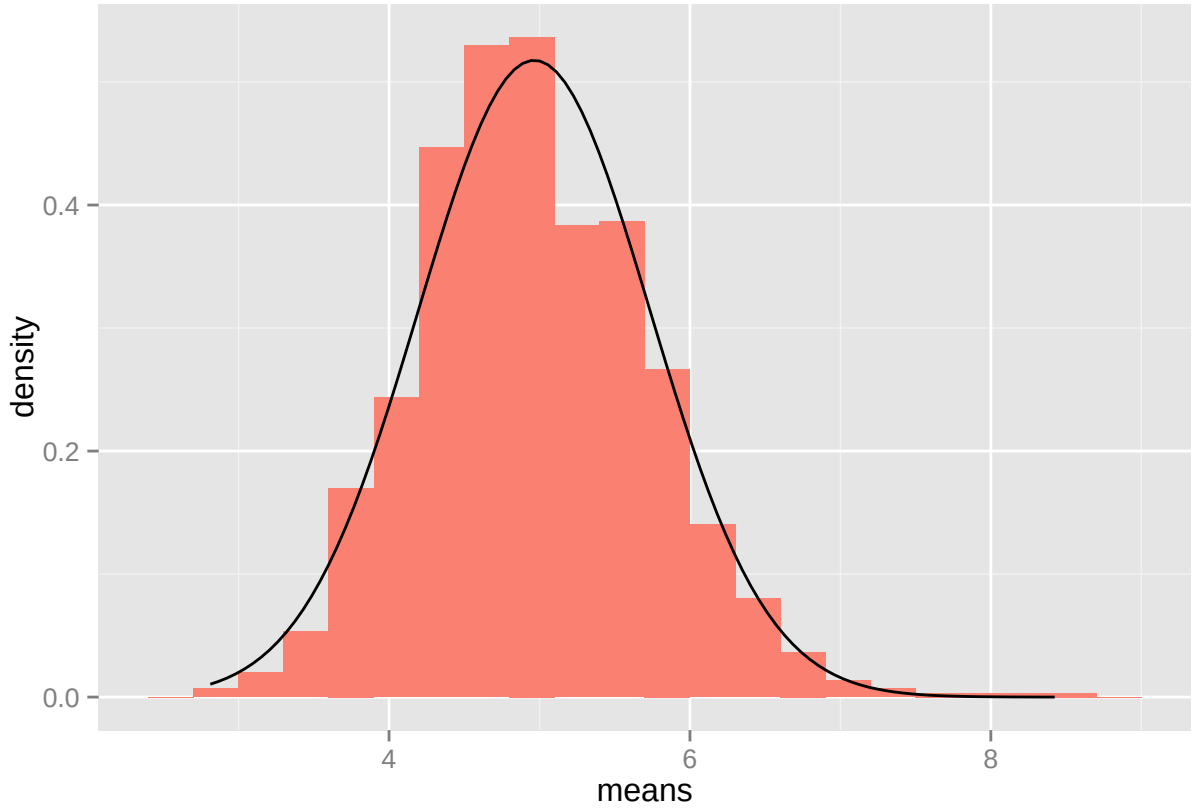
I investigate the distribution of averages of 40 exponentials by generating 1000 simulation. The code snippet below generates samples of 40 exponentials 1000 times, stores the samples into a 1000x40 data frame called `sample_matrix`. Then for each row (sample) I calculate the mean and store it into the 41th column.

```
no_of_exp <- 40
no_of_iteration <- 1000
sample_matrix <- data.frame(
                            matrix(
                                sample(pop, no_of_exp*no_of_iteration, replace = TRUE),
                                nrow = no_of_iteration
                                )
                            )
sample_matrix <- cbind(sample_matrix, means=rowMeans(sample_matrix))
```

The distribution of averages of the observed exponentials looks like this:

```
library(ggplot2)
ggplot(sample_matrix, aes(x=means)) +
geom_histogram(binwidth=0.3,
               fill="salmon",
               aes(y=..density..)) +
stat_function(fun=dnorm,
```

```
            args=list(mean=mean(sample_matrix$means),
                      sd=sd(sample_matrix$means))
)
```




```
## add legend!
## rename X axis
```

Comparing this new distribution to the generated & theoretical population:

Population	Mean	Standard Deviation	Variance
Theoretical	5	5	25
Simulated population	4.98	4.95	24.49
Average of 40 randomly selected set	4.96	0.77	0.59

So, the sample mean estimates the original population's mean quite well. As it was expected. The variance of the sample mean is significantly smaller as Central Limit Theorem states it should centralize around the mean in a narrow range.

The figure above presents that the new distribution is approximately normal.