Statistical Inference Course Project part 1

Alexandr Cherkashin

Thursday, February 19, 2015

Synopsis

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. We wil find how calculated mean and variance of this distribution similar to theoretical mean and variance. We will show that distribution of random samples means taken from exponential distribution is approximately normal.

Simulation

Let's make our samples from exponential distribution and calculate the means.

In the code below first we'll make a matrix with 1000 rows anf 40 columns filled with generated random values from exponential distribution. Each row will represent a sample, thus sample sizes is 40 values. We'll store this data in `samples` variable. Next we'll calculate mean value for each row by `apply`ing the function `mean` to each row of the matrix. We'll store all 1000 mean in `sample.means` variable. This will be the sample means distribution.

```
set.seed(42)
samples <- matrix(rexp(n = 40 * 1000, rate = 0.2), ncol = 40)
sample.means <- apply(samples, 1, "mean")
```

Basic inferential data analysis

Sample Mean versus Theoretical Mean

The theoretical mean of random samples from any distribution should be close to the mean of the original distribution. Let's look at theoretical and calculated actual samples mean. The theoretical mean would be `1 / labda` of exponentian distribution, and in this project lambda is equal to `0.2`.

```
#Theoretical mean is:
1 / 0.2
```

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

```
#Mean of 1000 samples of size 40 from exponential distribution is:
mean(sample.means)
```

```
## [1] 4.986508
```

The values are very close. The sample mean is estimating population mean.

Sample Variance versus Theoretical Variance

The variance of our exponential distribution equals `(1 / lambda)^2`. The theoretical variance of sample means should be equal to `(1 / lambda)^2 / n`, where and n is the size of samples. Let's compare theoretical variance and calculated on actual samples from `sample.means`.

```
#Theoretical sample variance is:
(1/0.2^2)/40
```

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

```
#Now compute it directly from sample.means
var(sample.means)
```

```
## [1] 0.6793521
```

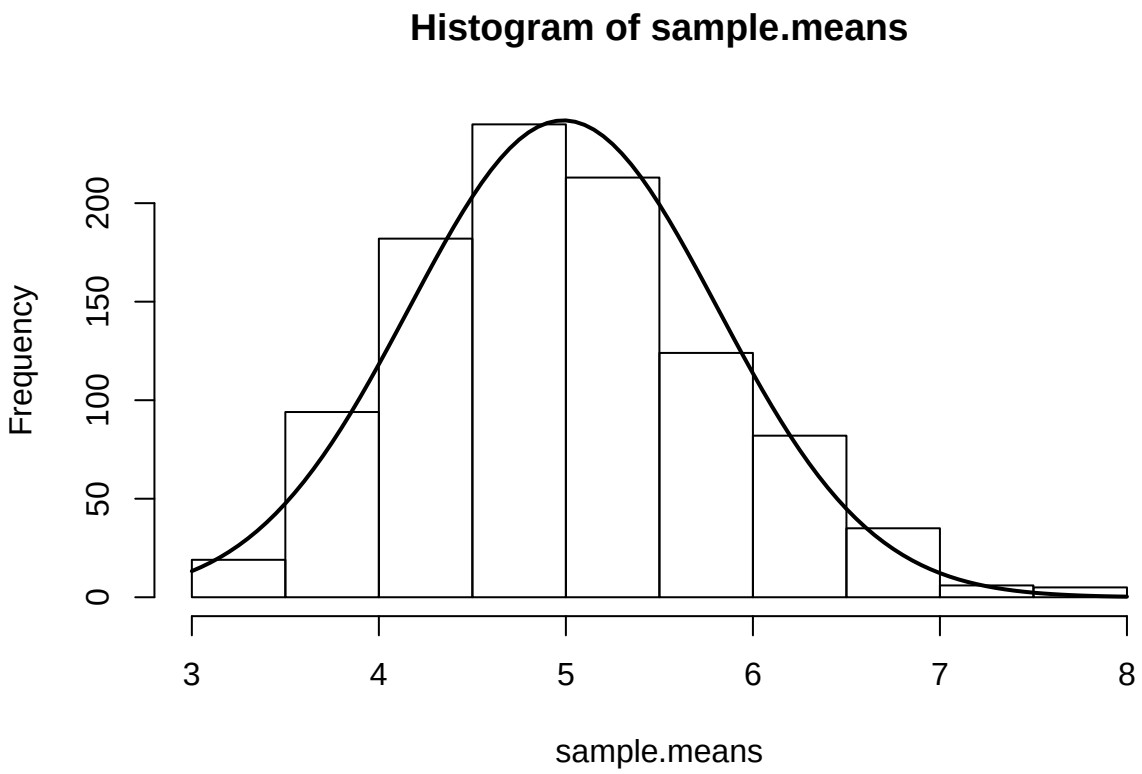
The values are close. If you'll take more samples, like 100 000 it would be even closer.

Distribution

We'll show that distribution of sample means is approximatly normal.

```
#Histgram of sample means
hist(sample.means)

#Normal distribution placed over
set.seed(42)
curve(dnorm(x, mean = mean(sample.means), sd = sd(sample.means))*500, add = TRUE, yaxt = "n", lwd=2)
```



Sample means distribution looks a lot like normal distribution.

About 68% of `sample.means` values lies withing one standart deviation above and below the mean. About 95% of `sample.means` values lies withing two standart deviation above and below the mean.

```
length(sample.means[sample.means > (mean(sample.means) - sd(sample.means)) &
                    sample.means < (mean(sample.means) + sd(sample.means))]) /
                    length(sample.means)
```

```
## [1] 0.678
```

```
length(sample.means[sample.means > (mean(sample.means) - 2 * sd(sample.means)) &
                    sample.means < (mean(sample.means) + 2 * sd(sample.means))]) /
                    length(sample.means)
```

```
## [1] 0.959
```

Defenately this distribution is close to noramal.