Statistical Inference Course Project - CTL Simulation

Daniel Pont

2018-09-30

Synopsis

This report is the first part of the Statistical Inference Course Project. It's a simulation exercise, that investigates the distribution of averages of 40 exponentials in R and shows it behaves as predicted by the Central Limit Theorem.

Simulation of the Central Limit Theorem with exponential distribution averages

1) Let's simulate 1000 exponential samples of n=40 values with parameter lambda=0.2

For each one we calculate its average value. Hence we get a distribution of 1000 averages "rexpAvgsDist".

```
lambda <- 0.2
n  <- 40
rexpAvgsDist = NULL
for (i in 1 : 1000) rexpAvgsDist = c(rexpAvgsDist, mean(rexp(n = n, rate=lambda)))
```

2) Now let's compare theoritical and simulation values for both mean and variance :

- **The theoritical mean for the distribution of averages**
  as predicted by the Central Limit Theorem is the mean of the exponential :

```
1/lambda
```

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

- **The simulation mean is** :

```
mean(rexpAvgsDist)
```

```
## [1] 4.990025
```

- **The theoritical variance for the distribution of averages**
  as predicted by the Central Limit Theorem is the variance of the exponential / n :

```
1/lambda^2/n
```

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

- **The simulation variance is** :

```
var(rexpAvgsDist)
```

```
## [1] 0.6111165
```

Value	Theoritical	Simulation
Mean	5	4.99
Variance	0.625	0.611

As summed up in the table above :

- Theoritical and simulation values for both mean and variance are pretty close
- The mean of the averages distribution is the mean of the exponential
- The variance of the averages distribution is much smaller than the variance of the exponential (1/lambda^2/n = 0.625 vs 1/lambda^2 = 25).

3) The distibution of averages is normal

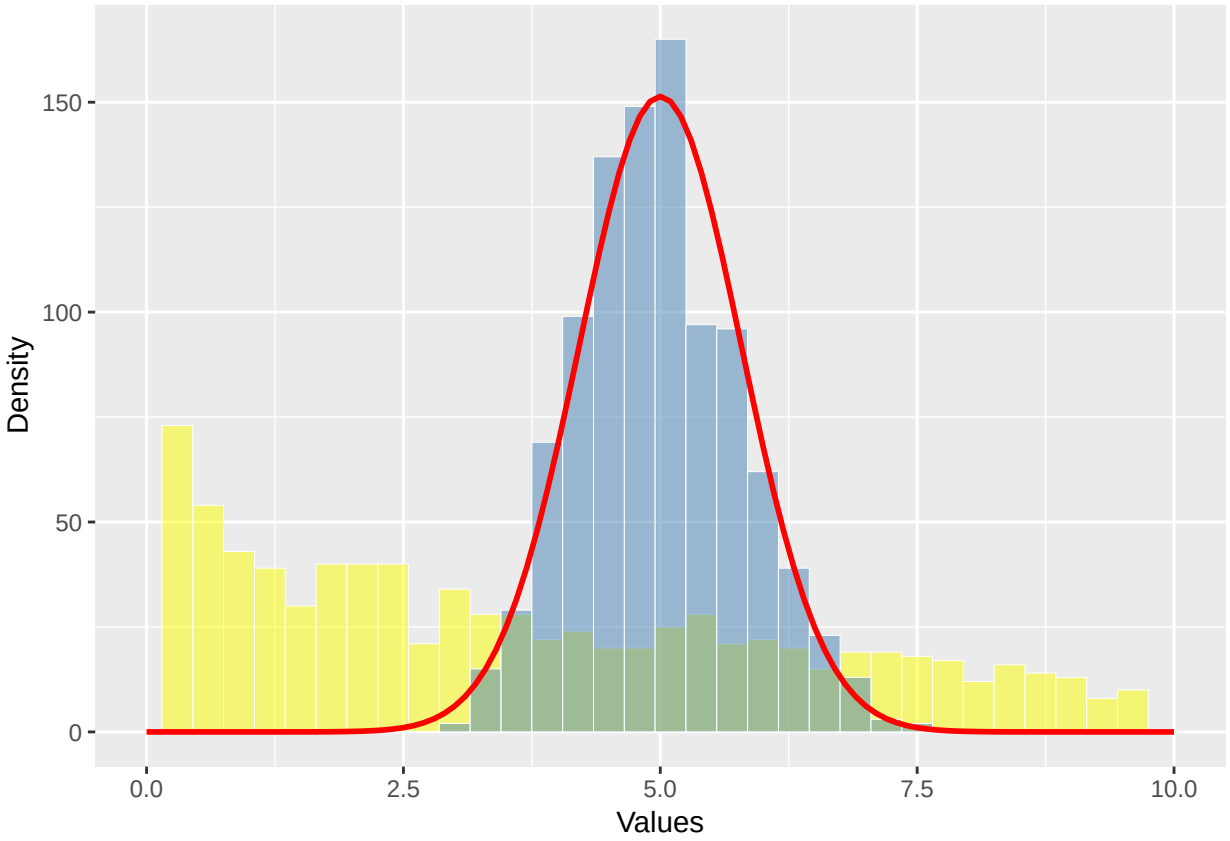
To show this, we simply plot the normal distribution over the distribution of averages with the mean and variance parameters calculated in point 2). We compare this plot with the histogram of large collection of random exponentials.

```
nbPoints <- length(rexpAvgsDist) # 1000

rexpRandomDist <- rexp(n = nbPoints, rate=lambda)

df <- data_frame(rexpAvgsDist,rexpRandomDist)

binwidth <- 0.3
ggplot(df) +
        scale_x_continuous(limits = c(0, 10)) +
        geom_histogram(binwidth = binwidth, aes(x=rexpRandomDist),
                       colour = "white", fill = "yellow",  size = 0.1,alpha=0.5)+
        geom_histogram(binwidth = binwidth, aes(x=rexpAvgsDist),
                       colour = "white", fill = "steelblue", size = 0.1,alpha=0.5)  +
        stat_function(fun = function(averages)
            dnorm(averages, mean = 5, sd = sqrt(0.625)) * nbPoints*binwidth,
            color= "red",size=1 ) +
        labs(x = "Values",y ="Density")
```



- The histogram in yellow represents the distribution of 1000 random exponentials
- The histogram in blue represents the distribution of 1000 averages of 40 exponentials
- The red curve is the normal distribution :
  - centered on the mean of the exponentials : 1/lambda = 5
  - with standard deviation = sqrt(1/lambdaˆ2/n) = sqrt(0.625)

Visually the distribution of averages clearly matches the normal distribution. On the other hand the distribution of random exponential doen't look at all like a normal distribution since it has not a bell shape.