Statistical Inference - Course Project - Part 1

Philip Mateescu

November 15, 2015

Investigation of Exponential Distribution in R

Synopsis

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem.
We will show that the mean variance of a large number of samples from this distribution is distributed approximately normal.

Simulations

We will execute a 1,000 simulations of 40 exponentials and will capture the average of each such simulation.

```
# setting a seed for reproducibility
set.seed(42)
n.sim <- 1000
n.exp <- 40
lambda = 0.2

data.sim <- replicate(n.sim, rexp(n.exp, rate = lambda))
# data.sim is a matrix of n.sim columns and n.exp rows
# in other words, each column contains 40 random exponentials
# let's calculate the means and the variance of each experiment
data.means <- colMeans(data.sim)
data.vars <- apply(data.sim, MARGIN = 2, FUN = var)
# let's also keep around a 1,000 exponentials for comparisons
exp1k <- rexp(n.sim, rate = lambda)
```

Sample Mean versus Theoretical Mean

The **Central Limit Theorem** states that:

> the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution [CLT]

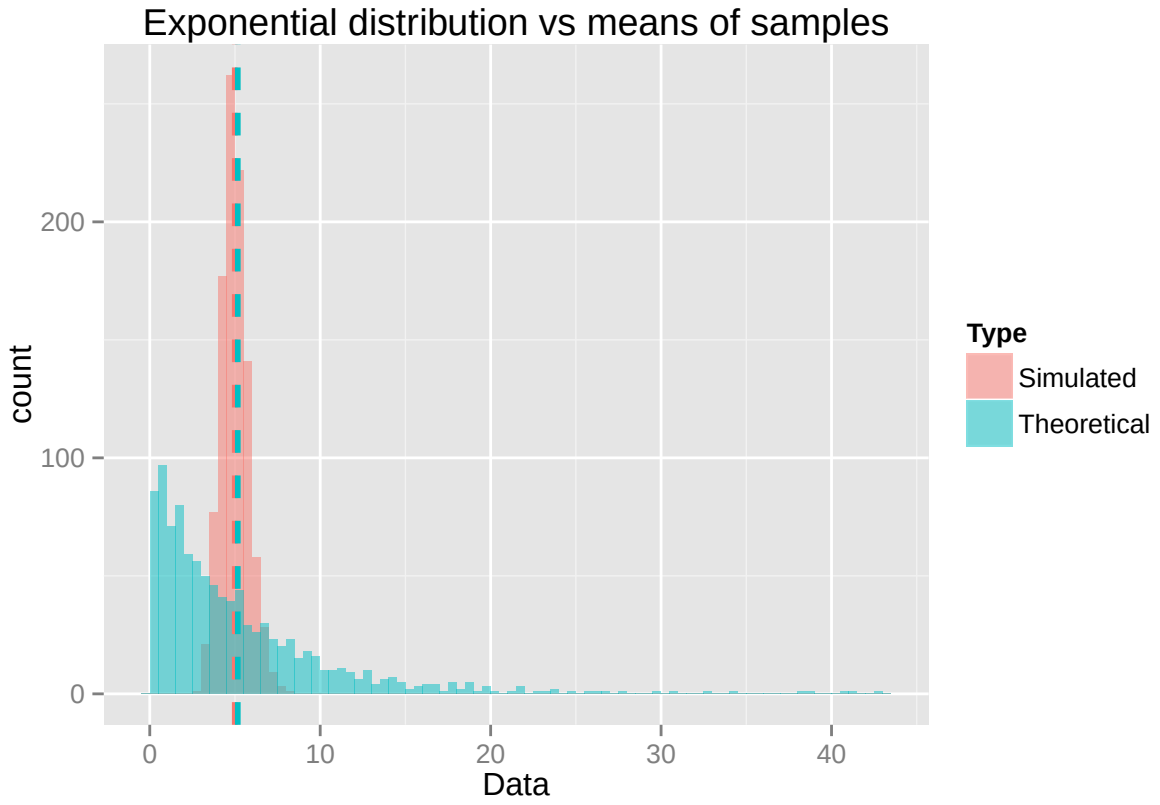
We have a large number of iterates – 1,000. Let's calculate the means of the samples in order to investigate whether they follow a normal distribution.

```
sim.mean <- mean(data.means)
sim.var <- mean(data.vars)
act.mean <- mean(exp1k)
act.var <- var(exp1k)
```

```r
theor.mean <- 1/lambda
theor.var <- 1/(lambda^2)

types <- c('Simulated', 'Theoretical')
df <- data.frame(Type = factor(rep(types, each=n.sim)),
    Data=c(data.means, Theoretical=exp1k))
means <- data.frame(Type = factor(types),
            Means = c(sim.mean, act.mean),
            Variance = c(sim.var, act.var))

ggplot(df, aes(x = Data, fill=Type)) +
    geom_histogram(binwidth=0.5, alpha=.5, position='identity') +
    geom_vline(data=means, aes(xintercept=Means, color=Type), linetype='dashed', size=1) +
    labs(title='Exponential distribution vs means of samples')
```



It's easy to notice that the exponential distribution is not normal, while the means of our simulations look positively Gaussian!

Indeed, notice how close the theoretical and simulated means and variances are, in other words the distribution is centered around the theoretical center of the distribution:

Type	Theoretical (`theor.?`)	Simulated (`sim.?`)	Difference from Theoretical	Actual (`act.?`)
Mean	5	4.9865083	-0.0134917	5.1653808
Variance	25	25.165914	0.165914	31.1432777

For illustration puposes, let's compare the distribution of these means to the normal distribution.

```
density <- (data.means - sim.mean)/sd(data.means)
qplot(density, geom = "blank") +
    geom_line(aes(y = ..density.., color = 'Simulated'), stat = 'density') +
    stat_function(fun = dnorm, aes(color = 'Normal')) +
    geom_histogram(aes(y = ..density..), alpha = 0.1, binwidth=.5) +
    geom_vline(xintercept=0, colour="black", linetype='dashed') +
    labs(title='Distribution of simulated mean values vs normal', y="Density", x="z") +
    theme_bw() + theme(legend.position = c(0.85, 0.85))
```