Exponential distribution analysis - Statistical Inference Course Project

Armando Güereca

September 23th, 2015

Overview

We will investigate the exponential distribution in R and compare it with the Central Limit Theorem. Our simulation will include the distribution of averages of 40 exponentials over thousand simulations.

Te objectives of this analysis are:

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Simulations

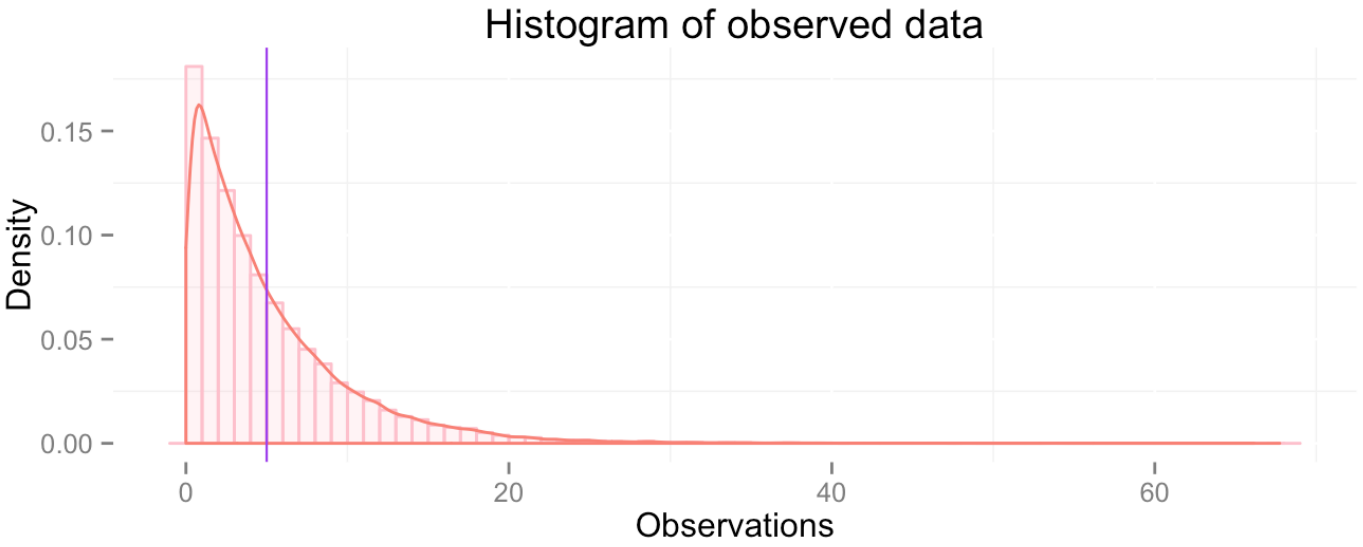
For our simulations we are going to use this parameters: (given as assignment instructions)

```
num_sim = 1000   # Number of simulations
num_obs = 40     # Observations per simulation
lambda = 0.2     # Rate of exponential distribution
set.seed(4242)
```

Data of all simulations is generated and placed on a table with an entry per observation.

```
# Random exponentially distributed data is generated and labeled by simulation num
ber
raw_data = data.frame(ObsVal=rexp(num_obs * num_sim, lambda), Simulation=rep(1:nu
m_sim, each=num_obs))
```

As expected, our data is exponentially distributed:



Sample Mean vs Theoretical Mean

One property of the exponential distribution is that its theoretical mean is: **1/lambda**

```
theo_mean <- 1/lambda
theo_mean
```

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

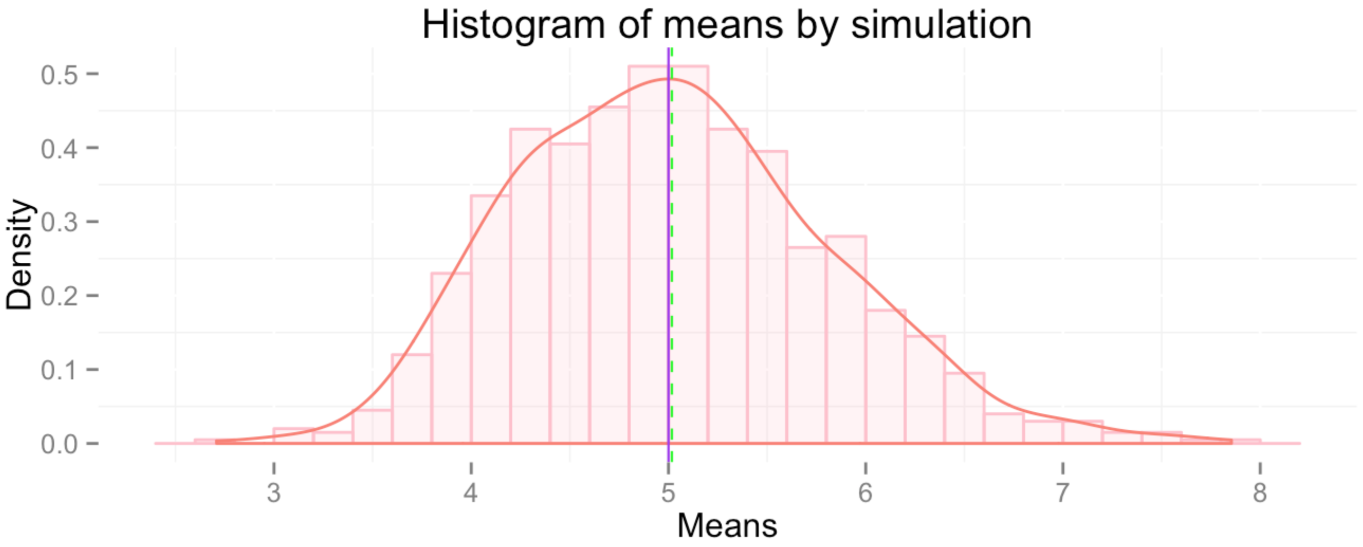
The mean of data observed on our simulations is:

```
# Mean of observed data is computed by simulation
means = ddply(raw_data, .(Simulation), summarise, SimMean=mean(ObsVal))
# Mean of data across all simulations
observed_mean <- mean(means$SimMean)
observed_mean
```

```
## [1] 5.017097
```

Given that a big number of simulations is performed, the mean of our sample (observed data), is closely related to the theoretical mean of our generated distribution.

This close approximation is explained by the Central Limit Theorem, evidence of it is that means of observed data per simulation follows a normal distribution centered at the theoretical mean **5** (purple line) and the mean of our sample **5.02** (dashed green line) converges to the expected value.



Sample Variance vs Theoretical Variance

Another property of the exponential distribution is that its theoretical standard deviation is also: **1/lambda**, the expected theoretical variance is given by:

```
theo_var <- (1/lambda)^2  #  Variance = SD^2
# Population theoretical variance
theo_var
```

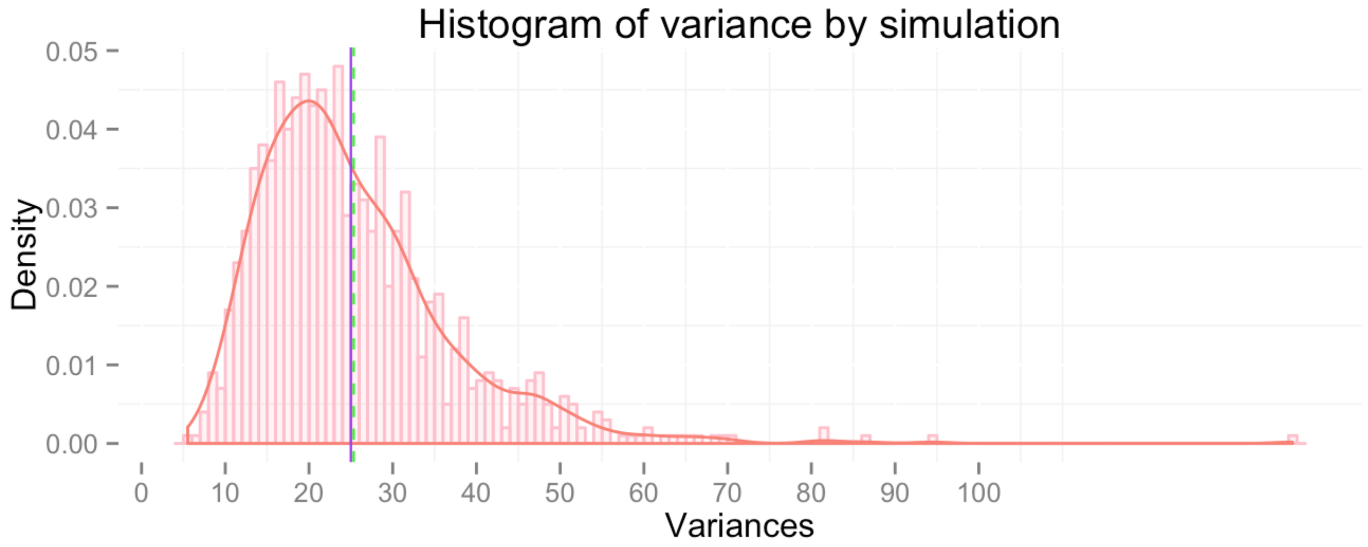
```
## [1] 25
```

The variance of data observed on each simulation is:

```
# Variance of observed data is computed for each simulation
varians = ddply(raw_data, .(Simulation), summarise, SimVar=var(ObsVal))
# Mean variance of sample across all simulations
mean_sample_var <- mean(varians$SimVar)
mean_sample_var
```

```
## [1] 25.33956
```

Similar to the mean of our sample, the mean variance of our sample **25.34** (dashed green line) has converged towards the true expected theoretical value **25** (purple line), also the variances of our simulations follow a normal distribution.



Same conclusion can be reached by comparing the expected standard distribution for our sample size against the standard distribution of the simulation means obtained on previous section:

```
# Standard deviation of exponential distributions
theo_sd <- (1/lambda)
# Expected standard deviation of sample
theo_sample_sd <- theo_sd/sqrt(num_obs)
# Standard deviation of simulation means
sd_sim_means <- sd(means$SimMean)
# Both values have converged
c(theo_sample_sd, sd_sim_means)
```

```
## [1] 0.7905694 0.7920459
```

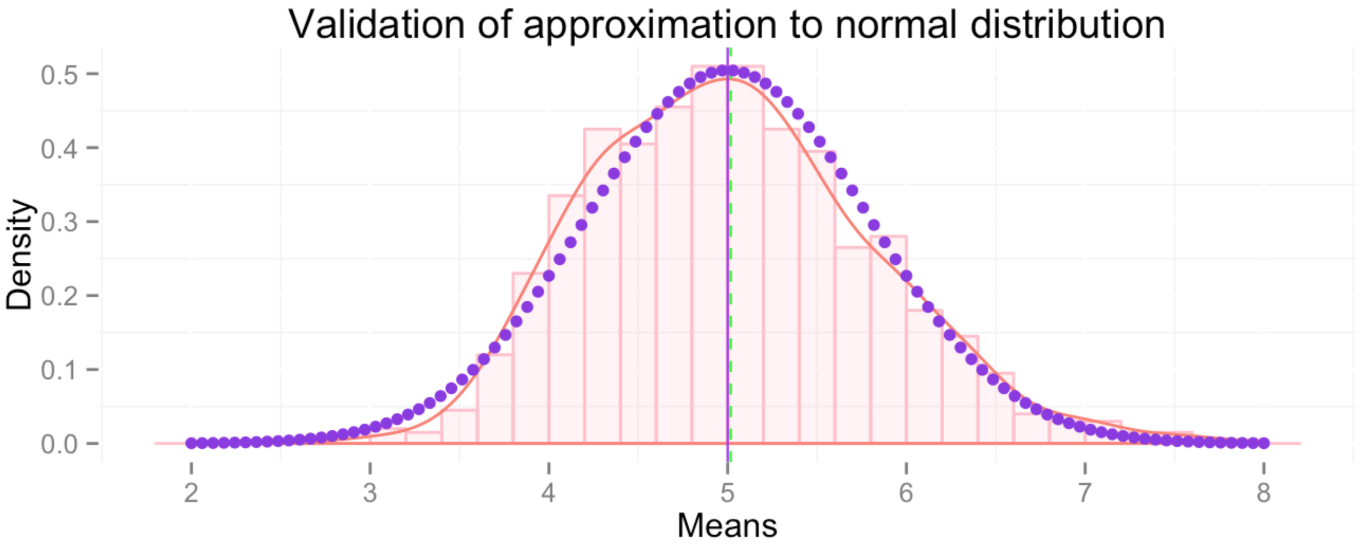
Distribution - Normality test

Our first approach to validate the hypotesis that means of our simulations are normally distributed, will be to overlay into our previous graph how such normal distribution will look.

To achieve this we start by generating a normally distributed dataset with statistics similar to our simulation data.

```
# Generate a normally distributed dataset with statistics matching our simulation
data
XVals <- seq(2, 8, length=100)
theo_data = data.frame(XVal=XVals, YVal=dnorm(XVals, mean=theo_mean, sd=theo_sampl
e_sd))
```

Then we overlay this theoretical data over our previous graph of means distribution:



By empirical observation we can conclude that efectivelly the mean values of simulation data obtained from exponential distributions are normally distributed.

Aditionally and as statistical complement of our validation; We can assess how well the dataset matches a 95% confidence level to the population.

Taking the sample mean and +/- 1.96 * stderror we can then determine the number of times this result includes the population mean. This will be expected to be approximately 95% of the time.

Note: increasing the sample size eg to 400, should improve this figure, whereas reducing it is likely to lead to a lesser proportion.

```
# Take the absolute difference of the means, and simply subtract the Zvalue * stde
rror
means_diff = abs(means$SimMean - theo_mean) - qnorm(.975) * theo_sample_sd
# Convert the number that have a range that covers 0 (ie 95% interval) to a percen
tage
prop = 100 * length(means_diff[means_diff <= 0])/num_sim
prop
```

```
## [1] 95.6
```

Once more, from the sample data we conclude that proportion of sample means that are within a 95% interval of the population mean is 95.6%