Exponential distribution and Central Limit Theorem

Molina Rafidison

9 Mar 2016

Report

Introduction

Synopsis

Here we are going to investigate the exponential distribution in R and compare it with the Central Limit Theorem.
We will set lambda = 0.2 as the rate parameter for all of the simulations and explore the distribution of averages of 40 exponentials. This will be repeated a thousand times.

Consult the Appendix section for more information about the supporting code.

Preliminaries

Load the needed packages.

```
require("ggplot2")
```

Simulation

Properly build the environment to make sure the analysis is reproducible and the variables set.

```
set.seed(2016)
nSim <- 1:1000; n <- 40; lambda <- .2
```

Create a data frame that collects the mean of 40 exponentials over 1,000 simulations.

```
expSim <- data.frame(nSim = nSim,
                     means = sapply(nSim, function(x) {mean(rexp(n, lambda))}))

expMeans <- expSim$means
```

Plot a histogram of the thousand simulations on 40 exponentials *(cf. Appendix Fig. 1)*.

The shape of the histogram in Fig. 1 looks bell-shaped.

Sample vs. Theoretical

Means

The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Compute the mean of our sample distribution. And calculate the theoritical mean of this same sample.

```
## Sample mean
m <- mean(expMeans); print(m)
```

```
## [1] 4.979186
```

```
## Theoretical mean
mu <- 1/lambda; print(mu)
```

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

Both means are very close to each other.
Create a variable that is the cumulative average reviewed after each simulation.

```
expSim$cumMean <- cumsum(expMeans)/seq_along(expMeans)
```

Let us check how the sample mean behaves by plotting a graph that follows the trend of the sample mean *(cf. Appendix Fig. 2)*.

The curve approaches the theoretical mean although it would be interesting to have a bigger sample to double-check.

Variances

Compute the variance of our sample distribution.

```
## Sample standard deviation:
s <- sd(expMeans)

## Sample variance:
s2 <- var(expMeans)
print(s2)
```

```
## [1] 0.6384844
```

And calculate the theoretical variance and standard deviation. We know that the exponential distribution has a standard deviation of 1/lambda.

```
## Theoretical standard deviation:
sigma <- 1/lambda/sqrt(n)

## Theoretical variance:
sigma2 <- sigma^2
print(sigma2)
```

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

Just like the means, both variances are very close to each other.
Take a random sample of 40 means among the sample population, compute their variance and repeat this a thousand times.

```r
expSim$vars <- sapply(nSim, function(x) {var(sample(expMeans, 40))})
```

Plot the results with a linear regression (in green) smoothing the points and the theoretical variance (in orange) *(cf. Appendix Fig. 2)*.

In Fig. 3, the linear model representing the sample variance is very close to the theoretical variance.

Results

Plot the histogram for the sample distribution and draw its corresponding density line (both in blue). Add a line that draws the normal distribution (in red) and a vertical line representing the mean (in black) *(cf. Appendix Fig. 4)*.

In Fig. 4, we observe that the shape of the sample curve is very similar to the theoretical one, which is bell-shaped. The sample curve is symmetric around the mean and almost as variable the normal distribution. In order to confirm this trend, manually compute the slope and the intercept of this kind of `qqline`.

```r
qt75 <- quantile(expMeans, .75)
qt25 <- quantile(expMeans, .25)

qq <- data.frame(slope = (qt75 - qt25)/(qnorm(.75) - qnorm(.25)))
qq$intercept <- qt25 - qq$slope*qnorm(.25)
```

Now plot the quantile-quantile comparisons between the sample and the theoretical distribution *(cf. Appendix Fig. 5)*.

We can see that the tails of the sample distribution are diverging from the line.
However the points are rather close to it, which allows saying that the sample data approximately follows a normal distribution.

Conclusion

The sample distribution approximates the bell-shaped curve and the mean of 5. Its variability (variance) is also close to the theoritical one, 0.625. The sample is quite large with 40 x 1,000 = 40,000 elements.

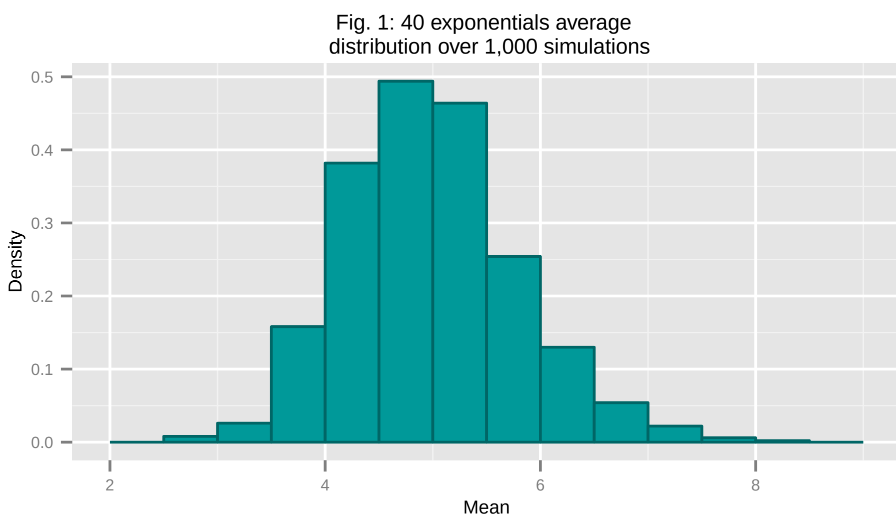
We then conclude that the sample average is approximately normally distributed with a mean given by the population mean 5, and a variance given by the standard error of the mean 0.625.

The Central Limit Theorem states that the distribution of averages of (indepedently and identically distributed) variables becomes that of a normal.

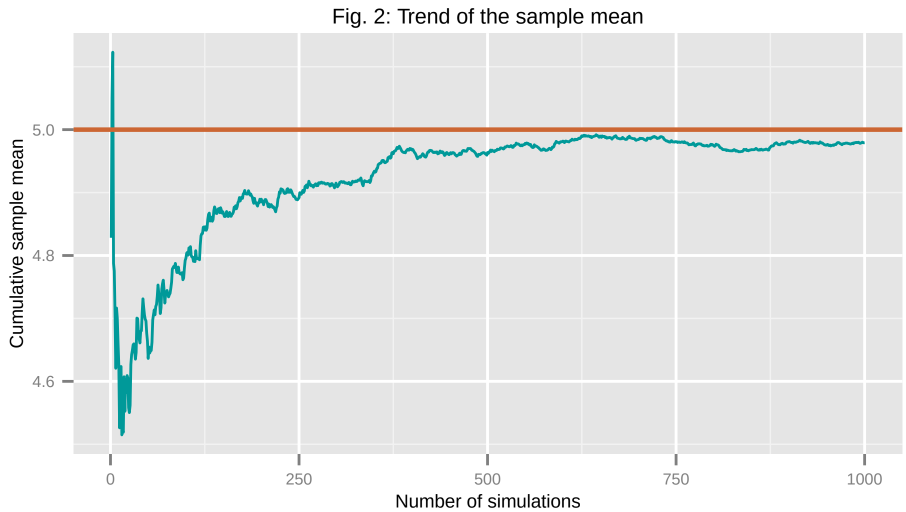
Appendix

Find below the supporting code and figures related to the Report part.

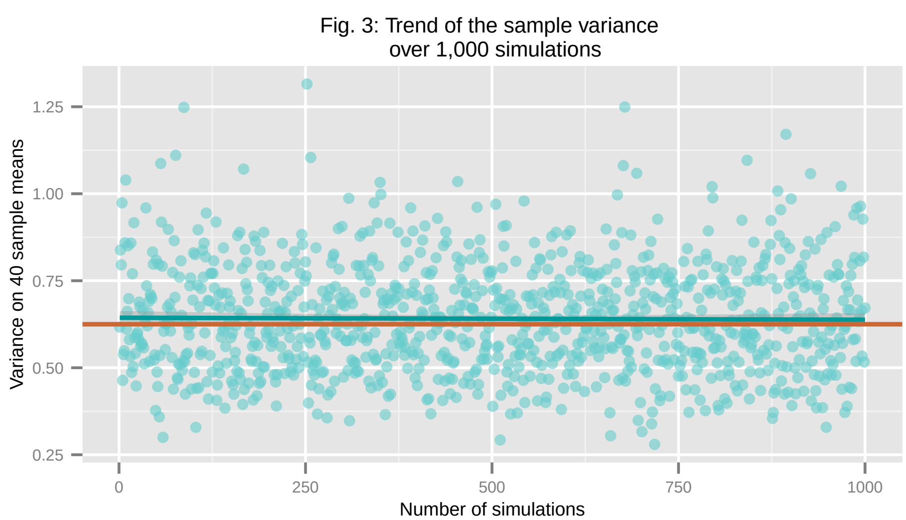
```
## Figure 1
ggplot(expSim, aes(x = means)) +
    geom_histogram(aes(y = ..density..), color = "#006666", fill = "#009999",
                   binwidth = .5) +
    labs(title = "Fig. 1: 40 exponentials average \n distribution over 1,000 simulations",
         x = "Mean",
         y = "Density") +
    theme(text = element_text(size = 7))
```




```
## Figure 2
ggplot(expSim, aes(x = nSim, y = cumMean)) +
    geom_line(color = "#009999") +
    geom_hline(yintercept = 5, color = "#CC6633", size = .8) +
    labs(x = "Number of simulations",
         y = "Cumulative sample mean",
         title = "Fig. 2: Trend of the sample mean") +
    theme(text = element_text(size = 7))
```

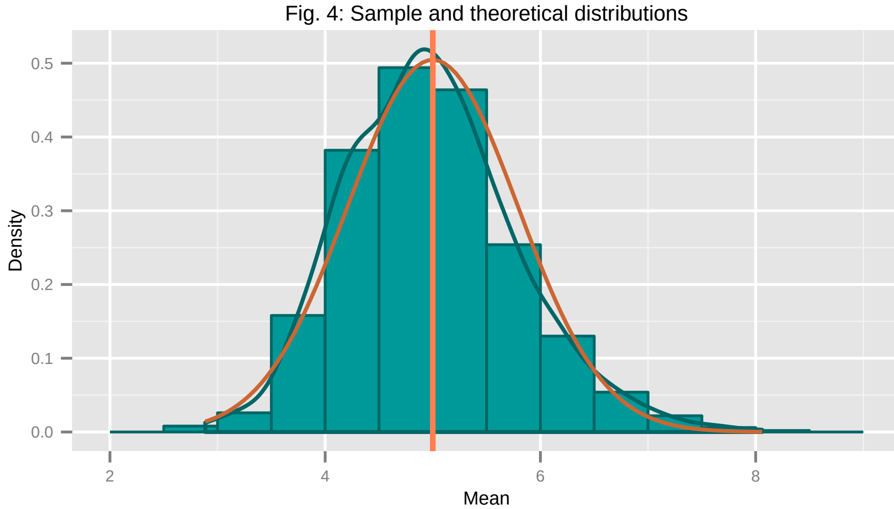



```
## Figure 3
ggplot(expSim, aes(x = nSim, y = vars)) +
    geom_point(color = "#66CCCC", alpha = .6) +
    geom_smooth(method = "lm", size = .8, color = "#009999") +
    geom_hline(yintercept = .625, size = .8, color = "#CC6633") +
    labs(x = "Number of simulations",
         y = "Variance on 40 sample means",
         title = "Fig. 3: Trend of the sample variance \n over 1,000 simulations") +
    theme(text = element_text(size = 7))
```




```
## Figure 4
ggplot(expSim, aes(x = means)) +
    geom_histogram(aes(y = ..density..), color = "#006666", fill = "#009999",
                   binwidth = .5) +
    geom_density(aes(y = ..density..), color = "#006666", size = .7) +
    stat_function(fun = dnorm, arg = list(mean = mu, sd = sigma), color = "#CC6633",
                  size = .7) +
    geom_vline(xintercept = 5, color = "coral", size = 1) +
    labs(x = "Mean",
```

```
        y = "Density",
        title = "Fig. 4: Sample and theoretical distributions") +
    theme(text = element_text(size = 7))
```




```
## Figure 5
ggplot(expSim, aes(sample = means)) +
    stat_qq(color = "darkcyan") +
    geom_abline(data = qq, aes(intercept = intercept, slope = slope),
                color = "indianred", size = .7) +
    labs(x = "Theoretical quantiles",
         y = "Sample quantiles",
         title = "Fig. 5: Comparing sample against theoretical quantiles") +
    theme(text = element_text(size = 7))
```