Central limit theorem simulations

Akim van Eersel

2020-10-11

Synopsis

In this report we aim to describe one application of the law of larger numbers, the central limit theorem (CLT). CLT states that as you increase the sample size for a random variable, the distribution of the sample sums better approximates a normal distribution. Mathematical demonstrations prove it and here we will show it in practice by simulations.

Simulations : Central Limit Theorem

Exponential distribution

Exponential distribution is a highly left-skewed, far from being similar to a normal distribution. The probability density function is P(X=x;lambda) = lambda * exp(- lambda * x), with unique parameter lambda. Theory allows us to know about the exponential distribution :
- the mean is 1/lambda
- the standard deviation is also 1/lambda

Without understanding the CLT, it could be surprising to see a sampling sum distribution becoming normal from values of an abnormal distribution. However, law of larger numbers (so as CLT) requires a sampling big enough to function properly. For the CLT, one common threshold is 30 values/sample. This simulation will take 40 values/sample for 1000 samples.

Initialize simulation parameters

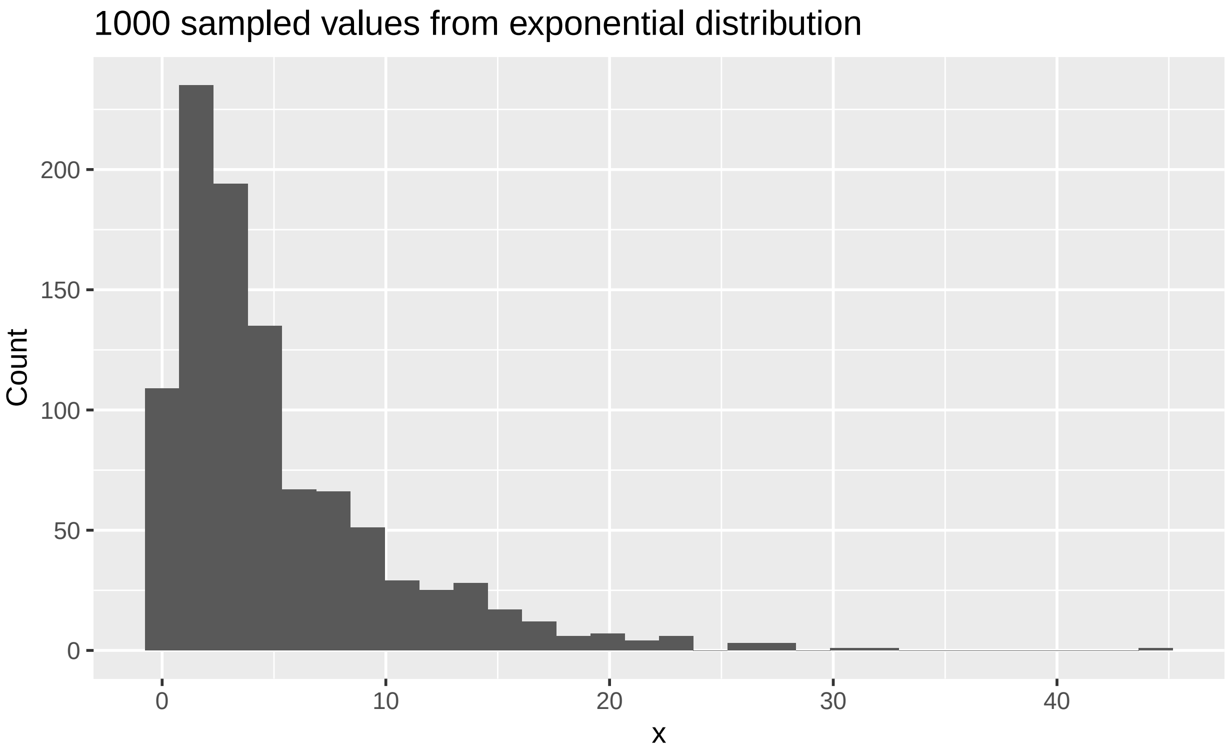
```r
lambda <- 0.2
nsamp <- 40
nsimu <- 1000
```

Initial distributions

In order to explicitly show the left-skewed exponential distribution, 1000 samples of exponential distribution function are shown below.

```r
library(ggplot2)

# Get 1000 random values of the exponential distribution
exp.sim <- rexp(nsimu, lambda)
```

```
# Plot the values
qplot(exp.sim,
      main= "1000 sampled values from exponential distribution",
      xlab = "x", ylab = "Count")
```



Sample mean vs Theoretical mean

Let's compare the theoretical mean and the mean of the 1000 samples distribution from the mean of 40 exponential distribution values.

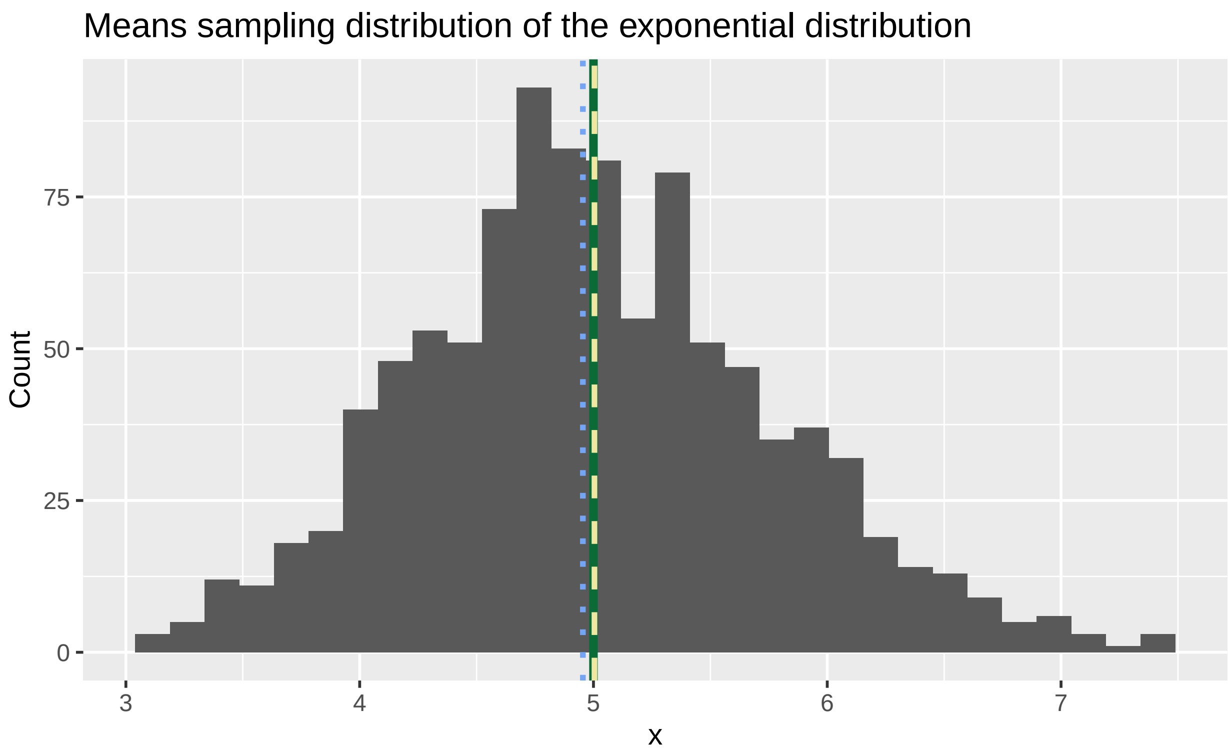
```
# Set random numbers generation reproducible
set.seed(126)

# From theory
mean.theo <- 1/lambda

# From simulation
mean.simu <- replicate(nsimu, expr = mean(rexp(nsamp, lambda)))

# Compare results
print(data.frame(Theorical.mean = mean.theo, Sample.mean = mean(mean.simu),
                 Sample.median = median(mean.simu), Sample.standard.deviation = sd(mean.simu)))
```

```
##   Theorical.mean Sample.mean Sample.median Sample.standard.deviation
## 1              5     5.00425       4.95466                 0.7662711
```

The two values are extremely close. It was expected regarding the CLT, since the mean is an unbiased statistic. Let's investigate more on the samples distribution.

Distribution

```
# Make a histogram of the distribution values and add vertical lines
qplot(mean.simu,
      main = "Means sampling distribution of the exponential distribution",
      xlab='x', ylab = 'Count') +
geom_vline(xintercept = 5, lwd = 1.5, col = '#0A6B37') +
geom_vline(xintercept = mean(mean.simu), lwd = 1, lty =2, col = '#EEE8A2') +
geom_vline(xintercept = median(mean.simu), lwd = 1, lty =3, col = '#74A5F5')
```



Above stands the distribution plot of the sampled means. Three vertical lines were added :
1. One filled in green intercepting the true value of the exponential distribution mean, 5.
2. Another dashed in yellow intercepting the mass center of the distribution, which in that case is the mean, ~ 5.00.
3. Last made out of points in blue intercepting the median of the distribution, ~ 4.95.

The distribution is almost bell-shaped, and values seem symmetrically distributed around the mean. Also the median is very close to the mean distribution, which both are equal in a normal distribution.

Another element reinforcing normality is the standard deviation of the sampling distribution, ~ 0.77, close to the theoretical value, 1, of a normal distribution.

All these insights on the means sampling distribution as a normal distribution make a solid assumption.

Sample variance vs Theoretical variance

Let's compare the theoretical variance and the mean of the 1000 samples distribution from the variance of 40 exponential distribution values.
Taking the average of the samples variance is relevant only if the distribution follows the CLT and approximates a normal distribution.

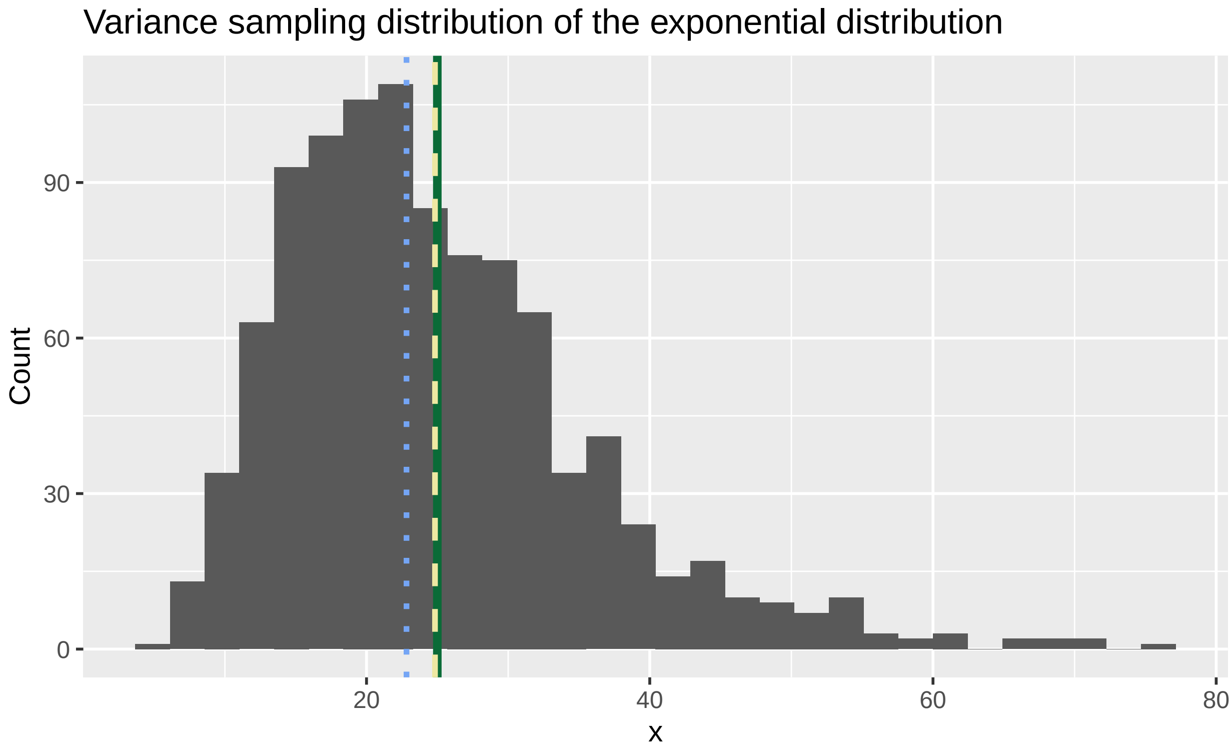
```r
# Set random numbers generation reproducible
set.seed(126)

# From theory
var.theo <- (1/lambda)^2

# From simulation
var.simu <- replicate(nsimu, expr = var(rexp(nsamp, lambda)))

# Compare results
print(data.frame(Theorical.variance = var.theo, Sample.variance = mean(var.simu),
                 Sample.median = median(var.simu)))
```

```
##   Theorical.variance Sample.variance Sample.median
## 1                 25        24.83586       22.8235
```

The two values are very close. It was also expected regarding the CLT. Let's investigate more on the samples distribution.

Distribution

```r
# Make a histogram of the distribution values and add vertical lines
qplot(var.simu,
      main = "Variance sampling distribution of the exponential distribution",
      xlab='x', ylab = 'Count') +
geom_vline(xintercept = 25, lwd = 1.5, col = '#0A6B37') +
geom_vline(xintercept = mean(var.simu), lwd = 1, lty =2, col = '#EEE8A2') +
geom_vline(xintercept = median(var.simu), lwd = 1, lty =3, col = '#74A5F5')
```



As previously, above stands the distribution plot of the sampled variance. Three vertical lines were added :
1. One filled in green intercepting the true value of the exponential distribution variance, 25 (standard deviation squared : 1/lambda²).
2. Another dashed in yellow intercepting the mass center of the distribution, which in that case is the mean, ~ 24.83.
3. Last made out of points in blue intercepting the median of the distribution, ~ 22.82.

The distribution is roughly bell-shaped, and values seem roughly symmetrically distributed around the mean. Even if the median is quite close to the mean distribution, the 2 units of difference between them are explained by the skewed right-tail of the distribution. Nevertheless, normality assumption stands.

Conclusion

Our results show that CLT is respected since the average and variance are extremely close between theory and simulation values. Moreover, their distribution approximate with success normal distribution.