The Central Limit Theorem by Simulation for the Exponential Distribution

Lou Marvin Caraig

Overview

The purpose of this simulation is to compare the theoretical mean and the theoretical variance of the exponential distribution with the sample mean and the sample variance of the sampling distribution of the mean. Additionally it will be shown that this sampling distribution is approximately normal thanks to the Central Limit Theorem.

Simulations

The figures that appear in this work are made using `ggplot2`, `gridExtra` and `RColorBrewer`. See the Appendix for details.

The first thing to do is to set the parameters that will be used during the simulation of the exponential distribution:

```r
set.seed(100)
lambda <- 0.2; m <- 1000; n <- 40
```

where `lambda` is the rate parameter of the exponential distribution, `m` is the number of simulations that will be performed and `n` is the size of each sample generated with a simulation. Given these variables the simulation can be run with the following code:

```r
expSims <- sapply(1:m, function(i) {rexp(n, lambda)})
```

where `rexp(n, lambda)` generates 40 random numbers from an exponential distribution with `lambda` equal to 0.2. The result is `expSims` which is a matrix with 40 rows and 1000 columns. The sampling distribution of the mean can be now obtained as follows:

```r
samplingDistribution <- apply(expSims, 2, mean)
```

Sample Mean versus Theoretical Mean

By plotting the sampling distribution of the mean we can notice that its shape clearly reminds a normal distribution with mean 5 which is also the value of the theoretical mean $1/\lambda$ of the exponential distribution with $\lambda = 0.2$. See Figure 1 generated by Code 1.

In fact the Central Limit Theorem states that the distribution of means is approximately normal regardless of whether the underlying distribution is normal, with an expected value equal to the mean of the underlying distribution.

The mean of the sampling distribution is equal to 4.9997019, but it's interesting to see how the mean of the sampling distribution converges to $1/\lambda$ after each simulation. To do this we calculate the mean of the sampling distribution of the mean after each simulation. See Figure 2 generated by Code 2.

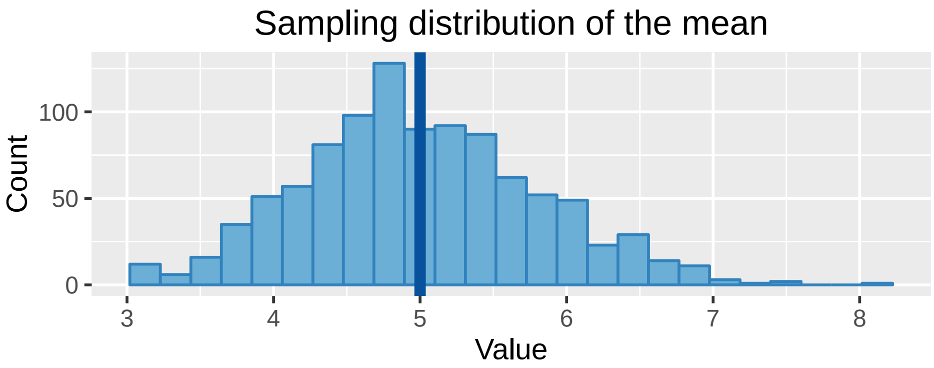

Figure 1: Sampling distribution of the mean. The blue vertical line corresponds to the theoretical mean.

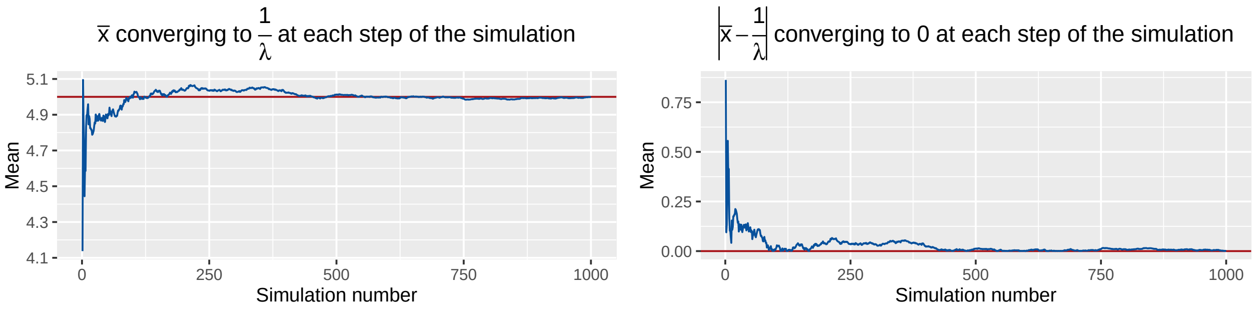

Figure 2: In the plot on left we can see the mean which converges to $1/\lambda$ while in the right plot we can see the absolute difference between the mean and $1/\lambda$ that consequently converges to 0.

Sample Variance versus Theoretical Variance

The theoretical variance of the exponential distribution is equal to $1/\lambda^2$ while thanks to the Central Limit Theorem we also know that the distribution of means is approximately normal regardless of whether the underlying distribution is normal, with a variance equal to the variance of the underlying distribution divided by the sample size $(1/(\lambda^2 n))$. In this case the variance should be approximately equal to 0.625 as in our case $\lambda = 0.2$ and $n = 40$.

The sample variance results to be equal to 0.6432442 and as we have done for the sample mean we can see how the sample variance converges to the value that we expect. See Figure 3 generated by Code 3.

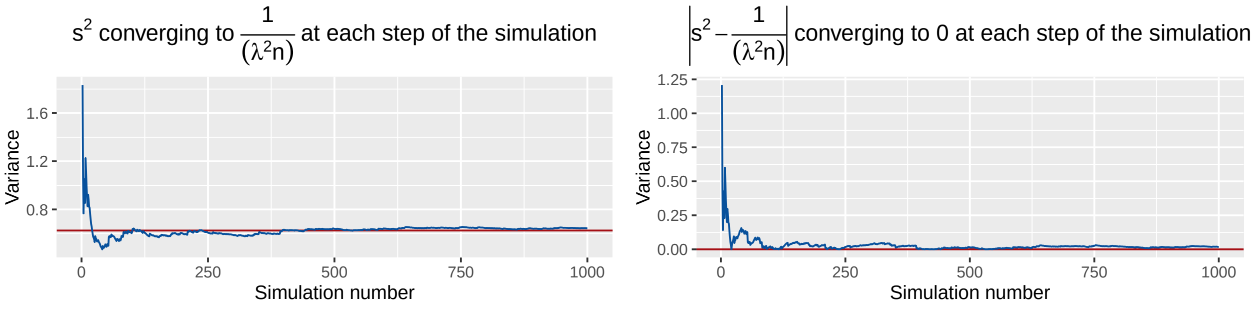

Figure 3: In the plot on left we can see the variance which converges to $1/(\lambda^2 n)$ while in the right plot we can see the absolute difference between the mean and $1/(\lambda^2 n)$ that consequently converges to 0.

Distribution

In summary thanks to the Central Limit Theorem we can say that the distribution of means is approximately normal regardless of whether the underlying distribution is normal, with an expected value equal to the mean of the underlying distribution with a variance equal to the variance of the underlying distribution divided by the sample size.

This can be seen by superimposing the plot of a normal distriubtion with mean $1/\lambda$ and standard deviation $1/(\lambda\sqrt{n})$ to the plot of the density of the sampling distribution. See Figure 4 generated by Code 4.

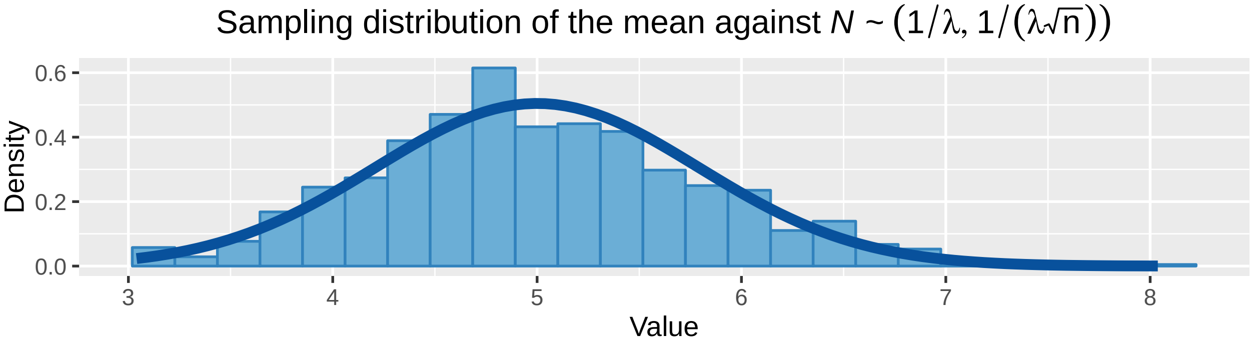

Figure 4: Sampling distribution vs. Normal distribution

A more efficient way to check the normality of the distribution is to generate the Q-Q plot in order to compare the quantiles of the sampling distribution against those of the normal distribution. See Figure 5 generated by Code 5. As we can see the Q-Q plot stands over the diagonal pretty well especially if we exclude the tails where the deviation of the sampling distributions from normality is higher.

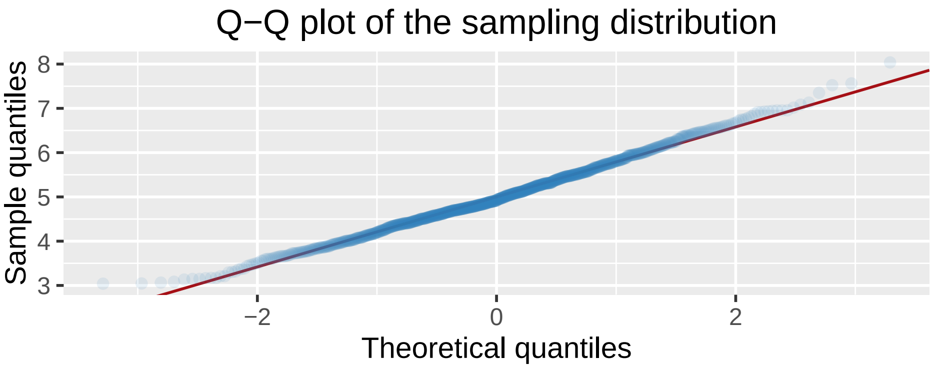

Figure 5: Q-Q plot of the sampling distribution

Appendix

This appendix contains the code that has been used to generate the above presented plots. Here's the required libraries loaded:

```r
library('ggplot2')
library('gridExtra')
library('RColorBrewer')
```

and here's the colors used:

```r
blues <- brewer.pal(5, 'Blues')[3:5]
reds <- brewer.pal(5, 'Reds')[3:5]
```

Code 1:

```r
ggplot(mapping=aes(x=samplingDistribution)) +
    geom_histogram(bins=25, col=blues[2], fill=blues[1]) +
    geom_vline(xintercept=1 / lambda, col=blues[3], size=2) +
    labs(title='Sampling distribution of the mean', x='Value', y='Count')
```

Code 2:

```r
cumMeans <- cumsum(samplingDistribution) / seq_along(samplingDistribution)
diff <- abs(cumMeans - (1 / lambda))

title1 <- expression(paste(bar(x), ' converging to ', frac(1, lambda),
                           ' at each step of the simulation'))
p1 <- ggplot(mapping=aes(x=seq_along(cumMeans), y=cumMeans)) +
    geom_hline(yintercept=1 / lambda, col=reds[3]) +
    geom_line(col=blues[3]) +
    labs(title=title1, x='Simulation number', y='Mean')

title2 <- expression(paste(abs(bar(x) - frac(1, lambda)), ' converging to 0 ',
                           'at each step of the simulation'))
p2 <- ggplot(mapping=aes(x=seq_along(diff), y=diff)) +
    geom_hline(yintercept=0, col=reds[3]) +
    geom_line(col=blues[3]) +
    labs(title=title2, x='Simulation number', y='Mean')

grid.arrange(p1, p2, ncol=2)
```

Code 3:

```r
cumVars <- sapply(seq_along(samplingDistribution),
                  function(x) {
                      var(samplingDistribution[1:x])
                  })
diff <- abs(cumVars - (1 / (lambda^2 * n)))

title1 <- expression(paste(s^2, ' converging to ', frac(1, (lambda^2 * n)),
                           ' at each step of the simulation'))
```

```
p1 <- ggplot(mapping=aes(x=seq_along(cumVars), y=cumVars)) +
    geom_hline(yintercept=(1 / (lambda^2 * n)), col=reds[3]) +
    geom_line(na.rm=T, col=blues[3]) +
    labs(title=title1, x='Simulation number', y='Variance')

title2 <- expression(paste(abs(s^2 - frac(1, (lambda^2 * n))), ' converging to 0 ',
                           'at each step of the simulation'))
p2 <- ggplot(mapping=aes(x=seq_along(diff), y=diff)) +
    geom_hline(yintercept=0, col=reds[3]) +
    geom_line(na.rm=T, col=blues[3]) +
    labs(title=title2, x='Simulation number', y='Variance')

grid.arrange(p1, p2, ncol=2)
```

Code 4:

```
title <- expression(paste('Sampling distribution of the mean against ',
                          italic(N) %~% (list(1 / lambda, 1 / (lambda * sqrt(n))))))
ggplot(mapping=aes(x=samplingDistribution)) +
    geom_histogram(aes(y=..density..), bins=25, col=blues[2], fill=blues[1]) +
    stat_function(fun=dnorm,
                  args=list(mean=1 / lambda, sd=1 / (lambda * sqrt(n))),
                  col=blues[3], size=2) +
    labs(title=title, x='Value', y='Density')
```

Code 5:

```
ggplot(mapping=aes(sample=samplingDistribution)) +
    geom_abline(intercept=1 / lambda, slope=1 / (lambda * sqrt(n)),
                col=reds[3]) +
    stat_qq(col=blues[2], fill=blues[1], alpha=0.1) +
    labs(title='Q-Q plot of the sampling distribution',
         x='Theoretical quantiles', y='Sample quantiles')
```