Statistical Inference, Simulations

Dorian Kcira

10/23/2015

Overview

This study ivestigates the exponential distribution in R and compares it with the Central Limit Theorem. A large number of exponentials are simulated. The distribution of the calculated means for each of the exponentials is then studied and compared to the theoretical expectations.

Simulations

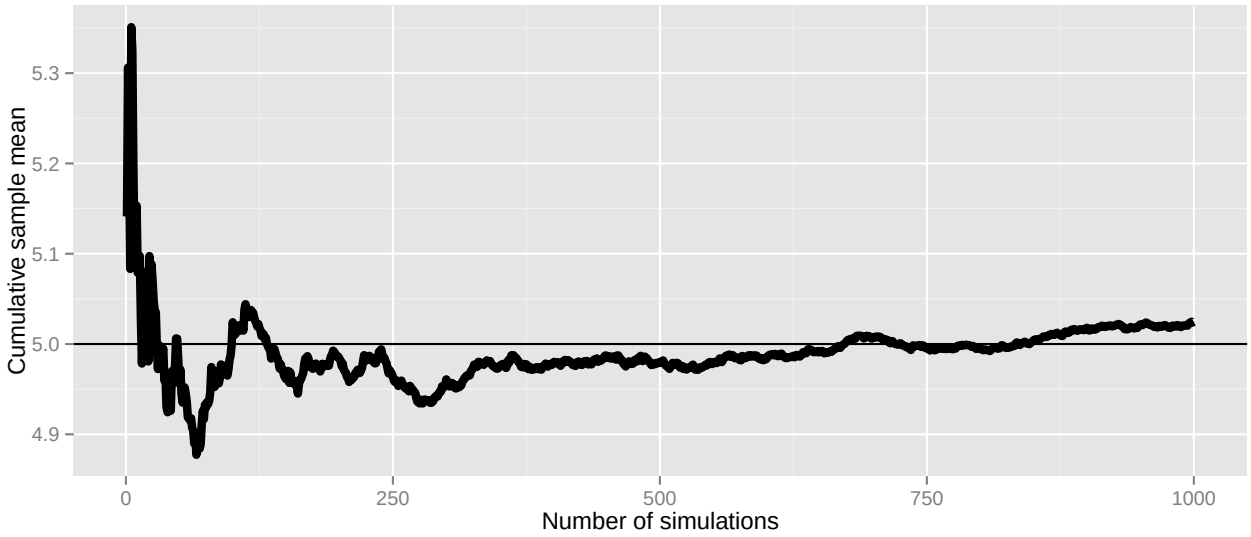
Simulate 40 exponentials, $f(x) = \lambda * \exp(-\lambda * x)$, with rate parameter $\lambda = 0.2$ and take the mean of them. Do this 1000 times. Keep each of the 1000 means in the vector *expmean*.

```
library(ggplot2)
nosim <- 1000 # number of simulations
lambda <- 0.2 # the rate parameter
nevts <- 40 # number of exponentials
# perform simulation
# set.seed(973422) # setting the seed would make this study reproducible
expmean = NULL; for (i in 1 : nosim) expmean = c(expmean, mean(rexp(nevts,lambda)))
```

Sample Mean versus Theoretical Mean

The theoretical mean for the exponential distribution is $1/\lambda = 1/0.2 = 5$. In the figure below the cumulative sample mean is plotted as the number of simulations increases from 1 to 1000. As it can be seen, the sample mean converges closer to the theoretical mean as the number of simulations increases. The theoretical value is plotted as a horizontal line. This is in accordance with what we expect from LLN.

```
sample_mean <- mean(expmean)
theoretical_mean <- 1 / lambda
means <- cumsum(expmean)/ (1  : nosim)
g <- ggplot(data.frame(x = 1 : nosim, y = means), aes(x = x, y = y))
g <- g + geom_hline(yintercept = theoretical_mean) + geom_line(size = 2)
g <- g + labs(x = "Number of simulations", y = "Cumulative sample mean")
g
```

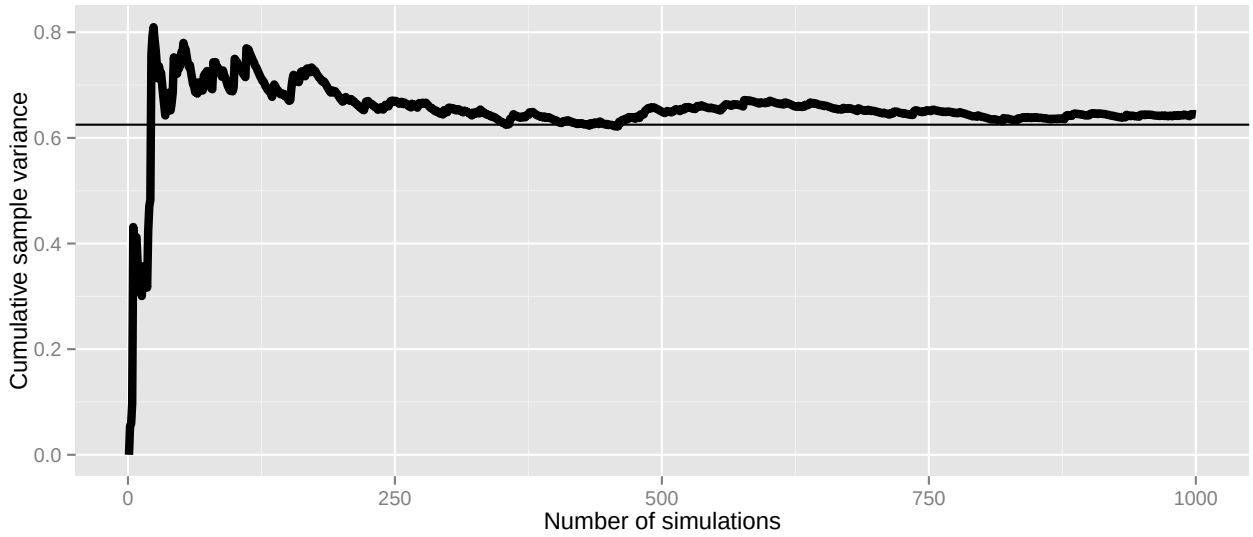


The sample mean for 1000 simulations is 5.0245778, which is close to the theoretical mean.

Sample Variance versus Theoretical Variance

The theoretical standard deviation for one exponential distribution is $1/\lambda = 1/0.2 = 5$. For a sample of exponential distributions, we need to take the variance on the mean, which is the squared of the standard deviation, divided by the sample size $5 * 5/40 = 0.625$. In the figure below the cumulative sample variance is plotted as the number of simulations increases from 1 to 1000. As it can be seen, the sample variance converges closer to the theoretical mean as the number of simulations increases. The theoretical value is plotted as a horizontal line. Again, this is in accordance with what we expect from LLN.

```
variances = NULL; variances <- c(variances, 0); # no variance with just 1 element
for (i in 2 : nosim) variances = c(variances, var(expmean[1:i]))
theoretical_variance <- (1 / lambda)^2 / nevts
sample_variance <- var(expmean)
g <- ggplot(data.frame(x = 1 : nosim, y = variances), aes(x = x, y = y))
g <- g + geom_hline(yintercept = theoretical_variance) + geom_line(size = 2)
g <- g + labs(x = "Number of simulations", y = "Cumulative sample variance")
g
```



The sample variance for 1000 simulations is 0.6445543, which is close to the theoretical variance.

Distribution

In the figure below the distributions of the means for each simulation is plotted as a histogram. A normal distribution is overlayed as a red line with the theoretical mean and standard deviation (as square root of the variance) with the values described above. As we can clearly see, the simulated distribution is very close to a normal distribution, which is what we expect from the CLT. This distribution is much more different than the exponential falling distribution we would observe if we picked single exponentials instead of 40.

```
g <- ggplot(data.frame(x = 1 : nosim, y = expmean), aes(x = expmean)) +
  geom_histogram(alpha = .20, binwidth=.2, colour = "black", aes(y = ..density..))
g <- g + labs(x = "Sample mean", y = "density")
g <- g + stat_function(fun = dnorm, colour = "red", size=2,
  arg = list(mean=theoretical_mean, sd=sqrt(theoretical_variance)))
g
```