Simulation Exercise

Remco Niggebrugge

01/20/2015

#### Overview

In this report I will simulate the exponential distribution using R. For a set **lambda** (rate parameter) of **0.2** and a set sample size of **40** I will see whether the mean and the standard deviation of the means of a sufficiently high number of such samples converges to the values predicted by the *Central Limit Theorem (CLT)*. This theorem predicts that the mean of sample means converges to the mean of the exponential distribution (1/lambda) and the standard deviation of the sample means converges to the standard deviation of the exponential distribution divided by square root of sample size (1/(lambda\*sqrt(40)). Furtermore the theorom predicts that the distribution of the means approaches a normal distribution.

#### Simulation

In the simulation 1000 samples, each containing 40 random values from the exponential distribution are created. This set of samples will than be compared to the theoretical predictions.

n <- 40 # sample size  
lambda <- 0.2 # lambda (rate parameter)  
means <- NULL # list containing all sample means  
sims <- 1000 # number of samples taken  
  
for (i in 1:sims)  
 {  
 sample <- rexp(n,lambda)  
 means <- c( means, mean(sample))  
 }

The set *means* will be used to test theoretical predications.

##### Simulation version theory: mean of the distribution

The mean of the means of the samples should be close to the distribution mean. Therefore we expect the mean of means to be around *1/lambda = 5*. And indeed this seems to be the case:

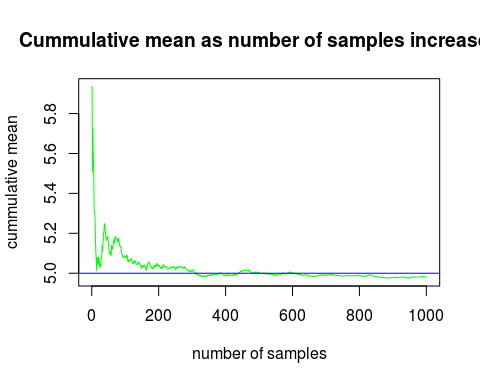
mean(means)

## [1] 4.98031

this is the case.

Futhermore, using the complete set of means, we can plot the cummulative mean value against the number of means considered. The resulting graph (green line) should get close to the predicted value (blue horizontal line).

mean\_means <- cumsum(means[1:sims])/(1:sims)  
plot(1:sims,mean\_means, type="l", col="green", xlab="number of samples",  
 ylab="cummulative mean", main="Cummulative mean as number of samples increases")  
abline(h=5, col="blue")



##### Simulation version theory: variance of the distribution

The variance of the distribution of means is predicted to be the variance of the distribution divided by the sample size. Therefore **Var(X')** should be close to *Var(dist)/N = 25/40 = 0.625*. And indeed this seems to be the case:

var(means)

## [1] 0.6239674

Futhermore, using the complete set of means, we can plot the variance value against the number of means considered. The resulting graph (green line) should get close to the predicted value (blue horizontal line).

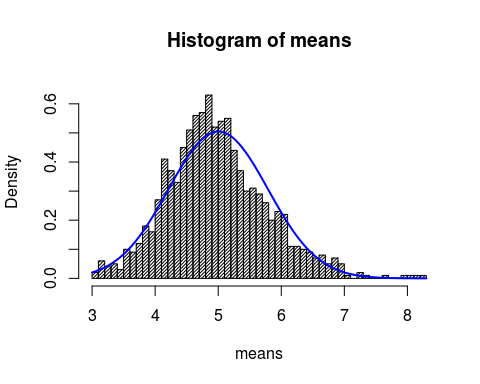
var\_means = NULL  
for(i in 1:sims) var\_means <- c( var\_means, var(means[1:i]))  
plot(1:sims,var\_means, type="l", col="green", xlab="number of samples",  
 ylab="cummulative variance", main="Cummulative variance as number of samples increases")  
abline(h=0.625, col="blue")



#### Normal distribution

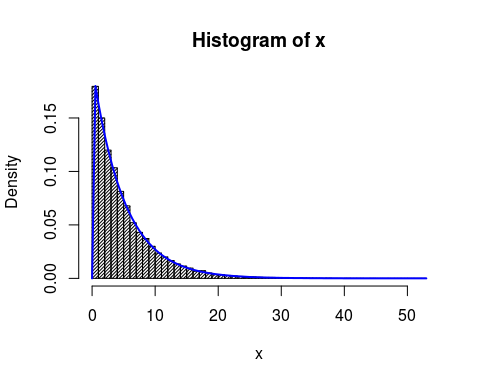
The distibution of the means, according to the CLT, should approximate a normal distribution. In order to check this, the 1000 means are plotted in a histogram. By comparison the real normal distrution (mean=5, sd=sqrt(var)=5/sqrt(40)=0.79) is plotted over the histogram (blue line). As can be seen the histogram indeed resembles the normal distribution.

hist(means, density=40, breaks=40, prob=TRUE, ylim=c(0,0.66))  
curve(dnorm(x, mean=5, sd=0.79), col="blue", lwd=2, add=TRUE, yaxt="n")



To show the difference between the distribution of random exponentials and the distribution of means of a large number of samples of exponentials, we will make another simulation. In order to compare the two sets, we again take 40,000 exponentials, but instead of dividing these over 1,000 samples this time they will be treated as one large collection. The resulting histogram (below) is very different. It does resample closely the exponential distribution function (blue line), as it to be expected.

x <-rexp(n\*sims,lambda)  
hist(x, density=40, breaks=40, prob=TRUE)  
curve(dexp(x, lambda), col="blue", lwd=2, add=TRUE, yaxt="n")



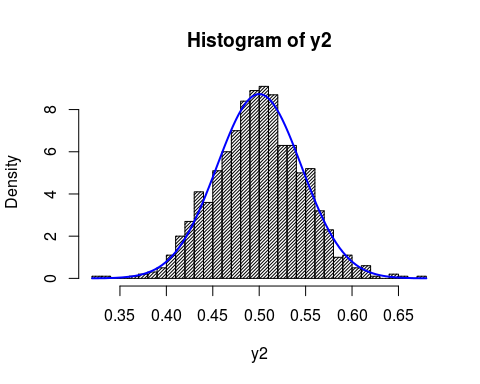
In order to show the CLT does apply to all distrubition functions, not only the exponential distribution, we will repeat the previous steps with another function, the uniform distribution. Again the distribution of means of a large number of samples should resemble a normal distribution, whereas the distribution of a large number of random uniforms should resemble closely the uniform distribution.

y1 <- runif(n\*sims)  
y2 <- NULL  
for (i in 1:sims)  
 {  
 sample <- runif(n)  
 y2 <- c( y2, mean(sample))  
 }

The *y1* object contains a large (40,000) set of randon uniforms, *y2* contains the means of 1,000 samples of 40 random uniforms.

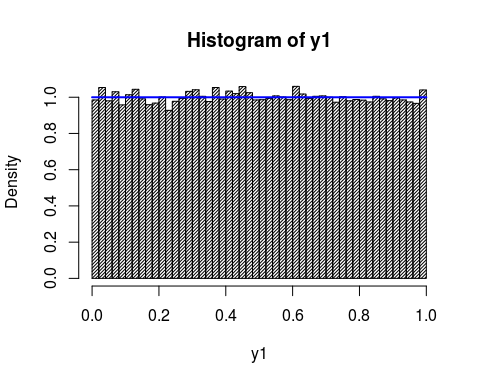
Plotting the histogram of 1,000 samples' means and the predicted normal distribution (with mean=0.5 and sd=1/sqrt(12\*40):

hist(y2, density=40, breaks=40, prob=TRUE)  
curve(dnorm(x, mean=.5, sd=1/sqrt(12\*40)), col="blue", lwd=2, add=TRUE, yaxt="n")



Plotting histogram of the large set of random uniforms, and comparing to uniform distribution function (blue line):

hist(y1, density=40, breaks=40, prob=TRUE)  
curve(dunif(x), col="blue", lwd=2, add=TRUE, yaxt="n")



#### Conclusion

From this short simulation exercise it is clear the Central Limit Theorem accurately predicts the behaviour of sample randoms taken from two different distribution functions (uniform and exponential). The match between predictions and simulations can be improved by increasing the number of simulations, which was set to 1,000 for this report.