Examing the Central Limit Theorem Using an Exponential Distribution

## Overview

According to the CLT, the sample mean, , is approximately normal with mean and variance . In this report, I will use a simulation to demonstrate these characteristics of the CLT. Additionally, I will use another simulation to focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

## Simulations

The code below creates a data frame with two variables: sampleMeans, which contains the results of Simulation #1 (1,000 averages of 40 random variables selected from an exponential distribution ( = .2)), and largeSample, which contains the results from Simulation #2 (1,000 samples from the same exponential distribution).

noSim <- 1000; sampleSize <- 40; lambda <- .2  
rv <- rexp(noSim \* sampleSize, lambda)  
randomExps <- data.frame(  
 sampleMeans = rowMeans(matrix(rv, noSim)),   
 largeSample = sample(rv, noSim)   
)

## Sample Mean vs. Theoretical Mean

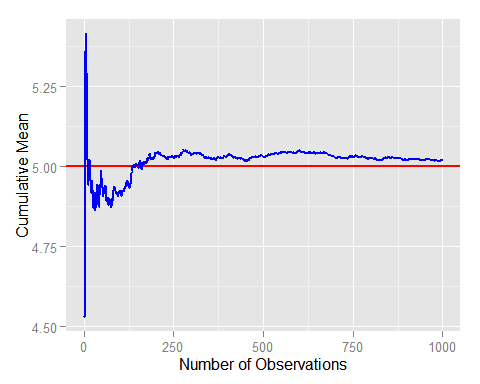
xbar <- mean(randomExps$sampleMeans)  
mu <- 1/lambda

The code above calculates , the sample mean, and , the theoretical mean. (5.02) is a good approximation of (5).

### Sample mean converging on theoretic mean as sample size increases

The code below calculates the cumulative mean, and plots the results, represented by the blue line, along with the theoretical mean, represented by the red line.

means <- cumsum(randomExps$sampleMeans) / (1 : noSim); library(ggplot2)  
g <- ggplot(data.frame(x = 1 : noSim, y = means), aes(x = x, y = y))  
g <- g + geom\_hline(yintercept = mu, size = 1, color = "red")  
g <- g + geom\_line(size = 1, color = "blue")  
g <- g + labs(x = "Number of Observations", y = "Cumulative Mean")  
g



## Sample Variance vs. Theoretical Variance

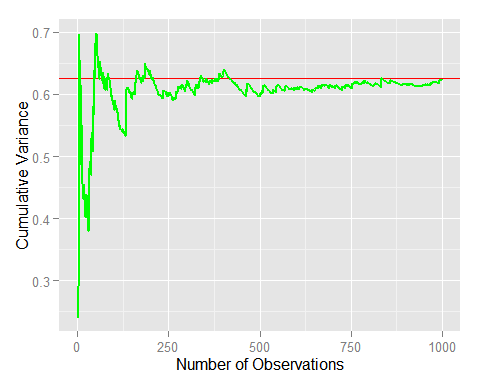
se <- 1/(lambda\*sqrt(sampleSize))  
sVar <- var(randomExps$sampleMeans)  
var <- se^2

The code above calculates the sample variance, and , the theoretical variance. The sample variance (0.623) is a good approximation of (0.625).

### Sample variance converging on theoretic variance as sample size increases

The code below calculates the cumulative variance, and plots the results, represented by the green line, along with the theoretical variance, represented by the red line.

vars <- cumsum((randomExps$sampleMeans-xbar)^2) / (1 : noSim)  
g <- ggplot(data.frame(x = 1 : noSim, y = vars), aes(x = x, y = y))  
g <- g + geom\_hline(yintercept = var, color = "red")  
g <- g + geom\_line(size = 1, color = "green")  
g <- g + labs(x = "Number of Observations", y = "Cumulative Variance")  
g



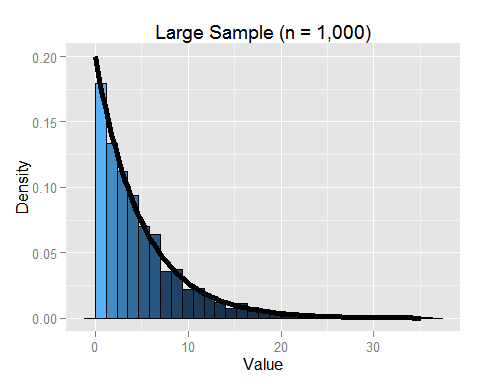
## Distributions

The figures below demonstrate the fact that even though the underlying distribution is not normal, as demonstrated by first figure, the sampling distribution of the sample means, as demonstrate by the second figure, is approximately normal.

### An Exponential Distribution

This figure is based on Simulation #2. The sample distribution (the histogram) is a good approximation for the exponential distribution with a = .2 (the black curve), from which the samples were taken.

g <- ggplot(randomExps, aes(x = largeSample))  
g <- g + geom\_histogram(color = "black", aes(y = ..density.., fill = ..density..))  
g <- g + guides(fill=FALSE)  
g <- g + stat\_function(fun = dexp, args = list(rate = lambda), size = 2)  
g + labs(x = "Value", y = "Density", title = "Large Sample (n = 1,000)")



### A Normal Distribution

This figure is based on Simulation #1. The sample distribution (the histogram) is a good approximation for a normal distribution (the black curve). Also shown in this figure is the sample mean (the red line), which is a good approximation of the theoretical mean (the green line).

g <- ggplot(randomExps, aes(x = sampleMeans))  
g <- g + geom\_histogram(colour = "black", aes(y = ..density.., fill = ..density..))  
g <- g + guides(fill=FALSE)  
g <- g + stat\_function(fun = dnorm, args = list(mean = mu, sd = se), size = 2)  
g <- g + geom\_vline(xintercept = mu, color = "green")  
g <- g + geom\_vline(xintercept = xbar, color = "red")  
g + labs(x = "Mean", y = "Density", title = "Sample Mean (1,000 samples of n = 40)")

