Central Limit Theorem

Alyssa Goldberg

2015-11-21

# Introduction

The Central Limit Theorem (CLT)states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the *sample size increases*. This means that it is possible to get an approximation of mean , standard deviation , and variance , for the whole distribution with only one observed average () and without knowing the population distribution.

# Test the Central Limit Theorem by simulation:

## Some Initial Assumptions

The theoretical mean () and variance of exponential distribution with parameter are respectively and ,

If the CLT is true, then:

* The mean of our simulation should approach
* The variance of our simulation, should approach
* The variance of our sample mean, should approach
* The standard deviation of our sample mean should approach

Let's find out if this is true by using simulation of **random exponentials** with =40 drawn 1,000 times.

## Set Constants

* The rate parameter, , as prescribed by the assignment, is 0.2
* The number of random exponentials to means test, is 40.
* The number of simulations of 40 random exponents is 1000
* The total population, \* 1000 = 40,000
* Theoretical constants:  
   - Theoretical means = = 5  
   - Standard deviation = =0.7906  
   - Variance of sample mean = = 0.625

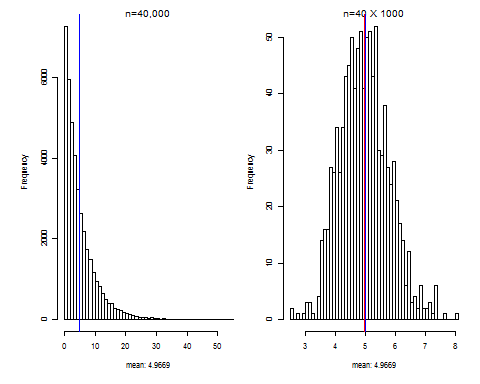
## Create the Population, Draw Samples

Draw 40 samples of exponentials, 1,000 times, producing a matrix with 1000 rows and 40 columns and create a vector of means of each of those 1000 rows

smatrix <- matrix(rexp(ex \* sims, .2), nrow = sims, ncol=ex) #produce a 1000 x 40 matrix  
  
smatrixmeans<-data.frame(value=rowMeans(smatrix)) #produce a data frame of the means of the 1000 rows in smatrix

## Compare Means:

The theoretical for a population this size= = 5.  
The calculated = 4.9669



## Compare Variance of the Sample Mean:

Let R calculate the sample mean, sample standard deviation and variance of the sample mean:

smean<-mean(smatrixmeans$value) #sample mean  
ssd<-sd(smatrixmeans$value) #standard deviation of sample means  
svar<-var(smatrixmeans$value) #variance of sample means

Producing:

|  |  |  |
| --- | --- | --- |
| Theoretical Values |  | Sample Values |
| = 5 |  | = 4.9669 |
| = 0.625 |  | = 0.6509 |
| = 0.7906 |  | = 0.8068 |

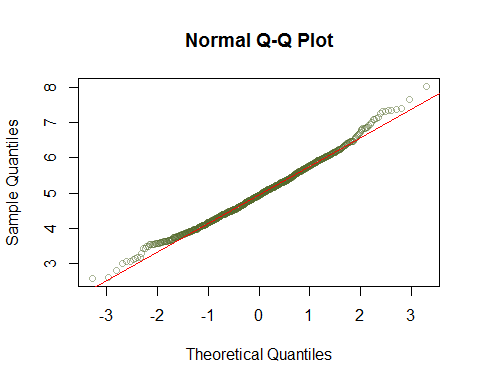
## Compare Density Distribution:

As the number of means of sample means ( times the number of simulations) increases, the density distribution should more closely resemble the normal density distribution:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 25% | 50% | 75% |
| sample | 4.399 | 4.931 | 5.492 |
| normal | 4.467 | 5.000 | 5.533 |

A Quantile-Quantile plot, displaying both the sorted theoretical normal distribution of means of a large sample (straight line) vs the sorted distribution of calculated means (plot points) shows a fairly tight fit, though it varies a bit more at the tails, yet more evidence that the Central Limit Theorem is useful for working with very large data sets.

qqnorm(smatrixmeans$value, col=rgb(.333, 0.42, .18, 0.5))  
qqline(smatrixmeans$value, col="red", lwd=1.5)



We can see that the behavior of large samples approaches that of the theoretical normal for Mean, Variance and Density, the closer we get to the theoretical Mean, Variance and Density.

#### APPENDIX

## Libraries:

library(knitr)  
library(dplyr)  
library(stats)  
library(ggplot2)

## Density Plot Theoretical vs. Sample

To compare densities, we can overlay a density plot for the distribution of the calculated sample means with a theoretical normal distribution density (the classic curve) to test this.

# for calculated  
sdx <- svar^2 #density standard deviation  
x.dens <- density(smatrixmeans$value) #create list of densities  
df.dens <- data.frame(x = x.dens$x, y = x.dens$y) #create a dataframe with densities  
varplot <- df.dens[df.dens$x >= smean - svar & df.dens$x <= smean + svar, ] #subset the density data to the area of the variance  
  
dnorm\_limit <- function(x) {  
 y <- dnorm(x, mean = mean(smatrixmeans$value), sd = sqrt((tmean^2)/ex))  
 y[x < tmean - tvar | x > tmean + tvar] <- NA  
 return(y)  
}  
  
# Normal vs Sample Plot  
p <- ggplot(data.frame(x = c(min(smatrixmeans$value), max(smatrixmeans$value))),   
 aes(x = x))  
  
p + stat\_function(fun = dnorm\_limit, geom = "area", fill = "blue", alpha = 0.2) +   
 stat\_function(fun = dnorm, args = list(mean(smatrixmeans$value), sd = sqrt((tmean^2)/ex)),   
 col = "blue", lwd = 1, lty = 2) + geom\_density(data = smatrixmeans,   
 aes(value), alpha = 0.25, col = "red", lwd = 1) + geom\_area(data = varplot,   
 aes(x = x, y = y), fill = "red", alpha = 0.25) + geom\_vline(xintercept = tmean,   
 stat = "vline", col = "blue") + geom\_vline(xintercept = smean, stat = "vline",   
 col = "red") + xlab("Theoretical Normal vs Sample Densities") + theme\_bw()

