Statistical Inference Course Project

## Overview

This project investigates the simulated exponential function and compare it with the central limit theorem. Here we use R to simulate data (rexp(n,lambda)) with exponential distribution that has rate parameter(lambda) = 0.2, mean & standard deviation of 1/lambda. We will investigate the distribution of averages of 40 exponentials. In this project we will use no. of simulations = 1000.

Load R libraries

library(ggplot2)  
library(knitr)

## Data Processing

Set initial values

lambda <- 0.2 # rate parameter  
mn <- 1/lambda # mean of the distribution  
nsample <- 40 # number of samples  
sd\_th <- 1/lambda # standard deviation  
var\_th <- sd\_th^2 # variance  
nsim <- 1000 # Number of simulations

Simulate the data

set.seed(12345) # For reproducibilty  
simdata <- matrix(replicate(nsim, rexp(nsample, lambda)), nrow=nsim) # Simulated data  
mean\_simdata\_distribution <- apply(simdata, 1, mean) # mean of each simulated data  
mean\_simdata <- mean(mean\_simdata\_distribution) # overall mean of the simulated data  
sd\_simdata <- sd(mean\_simdata\_distribution) # overall standard deviation  
var\_simdata <- sd\_simdata^2

## Results

Q1. Show the sample mean and compare it to the theoretical mean of the distribution.

Theoritical mean: 5  
   
 Mean of simulated data : 4.971972

Q2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

Theoretical standard deviaton: 5  
   
 Theoretical variance : 25  
   
 Standard deviation of simulated data: 0.7847246  
   
 Variance of simulated data : 0.6157926

Q3. Show that the distribution is approximately normal.

Figure below shows the histogram of mean of simulated data. Here red line outlines the distribution and mean of the distribution with theoritical lamda=0.2 and blue line represents the distribution of the means of the simulated data. The red and blue horizontal line represents the mean of theoritical data and simulated data respectively. The plot here shows that the simulated data closely matches the data with given theoritical values.

R code to plot histogram of the means of the simulated data.

library(ggplot2)  
df <- data.frame(mean\_simdata\_distribution)  
plot <- ggplot(df, aes(x = mean\_simdata\_distribution)) + geom\_histogram(binwidth = 0.05,aes(y = ..density..))  
plot <- plot + geom\_vline(xintercept=mean\_simdata,size=1.0, color="blue", linetype="dotdash") +  
 stat\_function(fun=dnorm,args=list( mean=mean\_simdata, sd=sd\_simdata),color = "blue", size = 1.0)  
plot <- plot +geom\_vline(xintercept=mn,size=1.0,color="red",linetype = "twodash") +  
 stat\_function(fun=dnorm,args=list( mean=mn, sd=sd\_th),color = "red", size = 1.0) +   
 ggtitle("Histogram of mean of simulated data") +xlab("Mean of simulated data") +  
 ylab("Density") + annotate("text", x = 7, y = 0.5, label = "Red : Represent theoritical data", colour="red") +  
 annotate("text", x = 7, y = 0.45, label = "Blue : Represent simulated data", colour="blue")  
print(plot)

