Statistical Inference Course Project

nsamples <- 1000  
lambdas <- rep(0.2, nsamples)  
sample\_size <- 40

## Assignment

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of the exponential distribution is:

and the standard deviation is

.

Set for all of the simulations. We will investigate the distribution of averages of 40 exponentials. Note that we will do 1000 simulations.

We will illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. We will

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

In point 3, we focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 1000 exponentials.

## Overview

This task demonstrates the central limit theorem, which states the distribution of the sample means follows the central limit theorem, which states:

* mean of sample means =
* sd of sample means =
* where n is the sample size
* and sample means are normally (gaussian) distributed regardless of the population distribution, exponential in this case

We do this by obtaining 1000 samples of sample size 40 from a population exponentially distributed with lambda (rate) of 0.2.

If we define the exponential random variables as , and the random variables representing the sample means as ,

According to the central limit theorem, the distribution of the means of the samples should be normally distributed with mean and standard deviation .

With , this translates to and .

## Simulations

*Instructions: Include English explanations of the simulations you ran, with the accompanying R code. Your explanations should make clear what the R code accomplishes.*

The following code generates 1000 samples of from an exponential distribution with

samples <- lapply(lambda, function(x) rexp(sample\_size, x))

The following code calculates the means and standard deviations of the sample means, , and .

means <- sapply(samples, mean) # vector of length 1,000 (length(means))  
sample\_means = mean(means) # approx 1/lambda = 1/0.2 = 5  
print(sample\_means)

## [1] 4.976754

sample\_sds = sd(means) # approx 5/sqrt(40) = 5/6.25 = 0.8  
print(sample\_sds)

## [1] 0.7823584

These are close to the theoretical mean and standard deviation, respectively, which are calculated in the next section.

## Sample Mean versus Theoretical Mean

*Instructions: Include figures with titles. In the figures, highlight the means you are comparing. Include text that explains the figures and what is shown on them, and provides appropriate numbers.*

# Function factory. DRY.  
# This function allows us to create histogram/distribution overlay plots  
# for any distribution. Used here for the Normal and Exponential.  
plotFunction <- function(fun, pl=seq(2,8)) {  
   
 function(histx, argslist, plotlimits=pl, col="red") {  
 df <- data.frame(vals=histx)  
 ggplot() +   
 geom\_histogram(data = df,   
 mapping = aes(x=vals, y=..density..)) +   
 stat\_function(data = data.frame(x=plotlimits),   
 mapping = aes(x),   
 fun = fun, # dnorm, dchisq, dexp  
 args = argslist, # for different dist'ns  
 color = col)  
 }  
}

The following code generates a histogram of the sample means overlaid with a normal distribution with the theoretical mean and sd of the sample means.

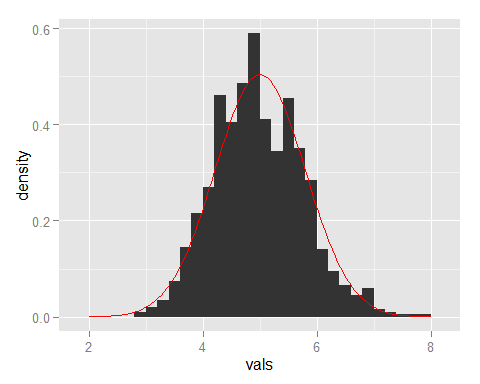
theoretical\_mean\_of\_means = 1/lambdas[1]  
print(theoretical\_mean\_of\_means)

## [1] 5

theoretical\_sd\_of\_means = (1/lambdas[1])/sqrt(sample\_size)  
print(theoretical\_sd\_of\_means)

## [1] 0.7905694

normPlot <- plotFunction(dnorm)  
p <- normPlot(means, list(mean=theoretical\_mean\_of\_means, sd=theoretical\_sd\_of\_means))  
print(p)



By inspection, the plot suggests the means are normally distributed with

and .

The 95% confidence interval for is

.

Which is:

mean(means) + c(-1,1) \* qnorm(0.975) \* sd(means)/sqrt(sample\_size)

## [1] 4.734303 5.219205

Given this, we can see the theoretical mean is within the 95% confidence interval.

## Sample Variance versus Theoretical Variance

*Instructions: Include figures (output from R) with titles. Highlight the variances you are comparing. Include text that explains your understanding of the differences of the variances.*

The CLT states the mean of the means is estimated by . It also states the sd of the means is estimated by

Therefore, given ~ ,

~

A confidence interval for is

.

In this case, 1000. Taking the square root of both sides to convert from the variance to the standard deviation, we get the confidence interval for :

alphas <- c(0.975,0.025)  
sigma\_ci <- sqrt((nsamples-1) \* sd(means)\*\*2/qchisq(alphas, nsamples-1))  
print(sigma\_ci)

## [1] 0.7495090 0.8182416

1/lambdas[1]/sqrt(sample\_size)

## [1] 0.7905694

is within this confidence interval.

Also, in this formula is equal to where .

Solving for by multiplying both sides by , the 95% confidence interval for is:

sqrt(40) \* sigma\_ci

## [1] 4.740311 5.175014

, which is within this confidence interval.

## Distribution

*Instructions: Via figures and text, explain how one can tell the distribution is approximately normal.*

There are statistical tests one can run to test for normality, but between the CLT and the visual appearance of the figure above where we overlaid the histogram of the data with a normal plot of and , we can tell the distribution is approximately normal.

We can show that the quantiles are similar:

quantile(rnorm(n=1000, mean=5, sd=1/lambdas[1]/sqrt(sample\_size)))

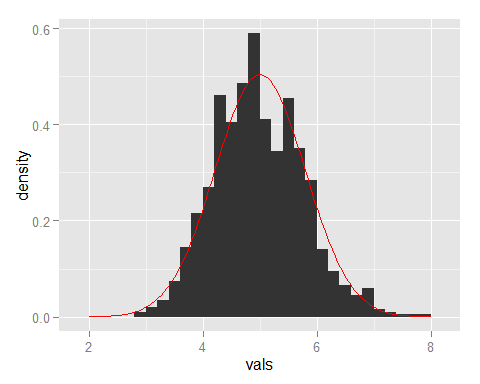
## 0% 25% 50% 75% 100%   
## 2.404706 4.472100 4.969785 5.540411 7.546106

quantile(means)

## 0% 25% 50% 75% 100%   
## 2.924724 4.406117 4.927258 5.513350 7.843717

The histogram of the means overlaid with the normal plot is shown again below. This is compared with a histogram of all 40000 negative exponentials as a single sample, overlaid with an exponential distribution with . This is clearly exponentially distributed.

print(p)



expPlot <- plotFunction(dexp, pl=seq(0,30))  
s <- unlist(samples)  
ep <- expPlot(s, list(rate=0.2), col="blue")  
print(ep)

