Simulation Experiment

Gloria Q

2/1/2020

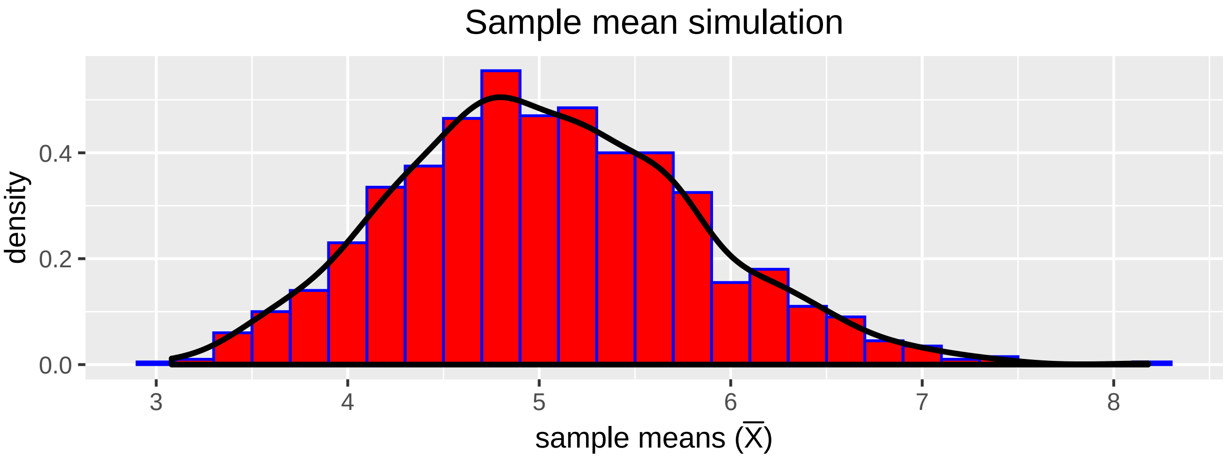
Introduction

This assignment was conducted as a requirement for the John Hopkins *Statistical Inference* course offered via Coursera, with the purpose of exploring expodential distribution and compare it to the Central Limit Theorem (CLT) in R environment. This will be tackled by: 1) simulating sample expodential data (expodential random uniform vs a large scale averages of 40 expodential random uniform), 2) comparing the simulation statistics versus the provided theoretical statistics (mean and variance), 3) determining the normality of the distribution. The following are pre-defined parameters of the assignment: $E[X] = \frac{1}{\lambda}$, $\sigma = \frac{1}{\lambda}$, $\lambda = 0.20$, n exponential variables=40, N_{sim} of simulations=1000.

Simulations

Based on the theoretical statistical distribution information provided in the instructions of this assignment, the average of 40 expodentials is calculated and resimulated 1000 times in R as follows:

```
#load needed packages
library(ggplot2); library(gridExtra); library(latex2exp)
#establish seed for reproducibility
set.seed(2352)
#define necessary variables
n <- 40; nsims <- 1000; lambda <- 0.20
#create empty vector, calculate the mean of 40 expodentials 1000 times and store in that vector
samples <- NULL; for(i in 1:nsims) samples = c(samples, mean(rexp(n, lambda)))
#stats of simulated data
simMean <- mean(samples); simVar <- var(samples);
simCI <- simMean + c(-1,1) * qnorm(0.975) * sd(samples)/100
#plot distribution of sample mean
ggplot(aes(x=samples), data=data.frame(samples))+
    geom_histogram(aes(y=..density..),binwidth=0.2, color="blue", fill="red") +
geom_density(lwd=1) + labs(title="Sample mean simulation", x=TeX("sample means ($\\bar{X}$)")) +
theme(plot.title = element_text(hjust = 0.5))
```

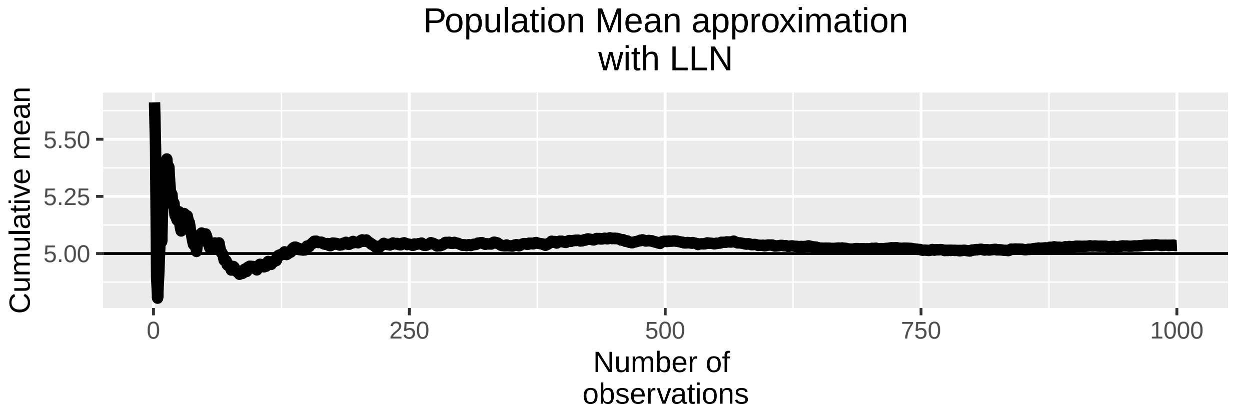


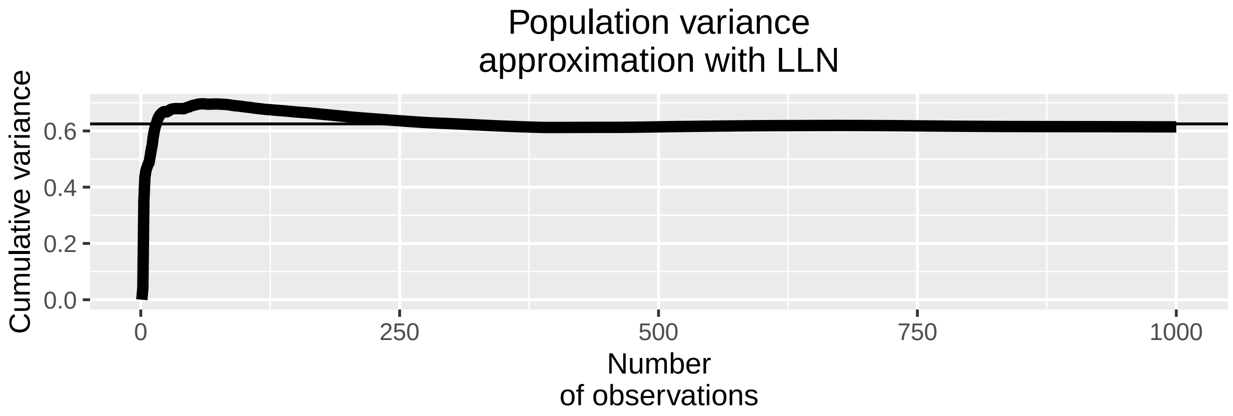
Based on this simulation, $\overline{X}$=5.0350275, s^2= 0.6049674 and the 95% confidence interval is 5.019783, 5.0502721.

Sample Statistics vs. Theoretical Statistics

The law of large numbers (LLN) will be considered due to the nature of this section and the large number of simulations conducted to acquire sample means. LLN states that with the increasing number of experimental repetitions, the average of these results should approximate/converge to the expected value. In other words, the sample stats (mean and variance) are estimators and are consistent with the population/theoretical stats. The following R code will be used to demonstrate this theory via cumulative means and variance:

```r
#calculate the cumultive sum of the means
expMeans <- cumsum(samples)/1:nsims
#graphing the cumulative sum of the means and demonstrating its convergence
#to the theoretical mean by indicating it on the y-axis (1/lambda)
llnMean<- ggplot(data.frame(x=1:nsims, y=expMeans), aes(x=x, y=y)) +
geom_hline(yintercept=1/lambda) + geom_line(size=2) + labs(x="Number of
observations", y="Cumulative mean", title="Population Mean approximation
with LLN") + theme(plot.title = element_text(hjust = 0.5))
#create empty vector, calculate the variance of 40 expodentials 1000 times
#and store in that vector. any NAs found in samVar converted to 0 for cumsum function
samVar <- NULL; for(i in 1:nsims) samVar <- c(samVar, var(samples[1:i]))
samVar[is.na(samVar)] <- 0
#calculate the cumultive sum of the means
expVar <- cumsum(samVar)/1:nsims
#graphing the cumulative sum of the variances and demonstrating its
#convergence to the theoretical variance by indicating it on the y-axis ((1/lambda^2)/40)
llnVar <- ggplot(data.frame(x=1:nsims, y=expVar), aes(x=x, y=y)) +
geom_hline(yintercept=(1/lambda^2)/n) + geom_line(size=2) + labs(x="Number
of observations", y="Cumulative variance", title="Population variance
approximation with LLN") + theme(plot.title = element_text(hjust = 0.5))
grid.arrange(llnMean, llnVar, nrow=2)
```



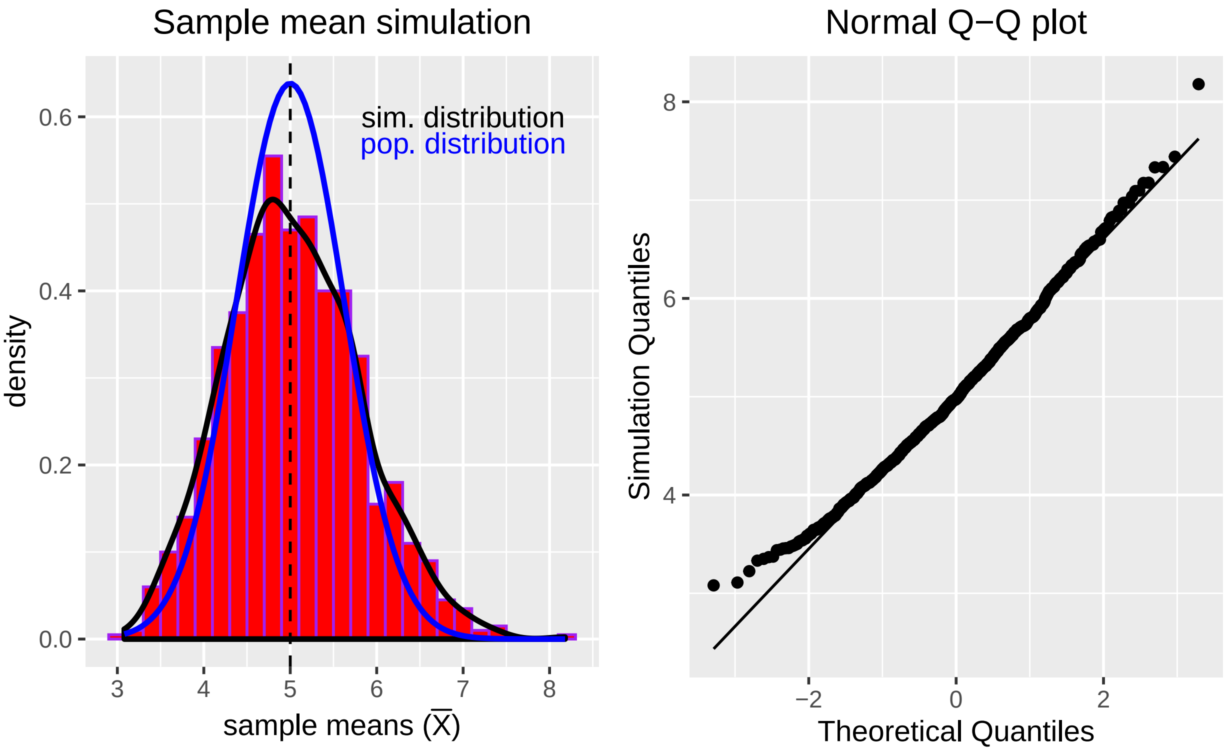


In the cumulative graphs above, the sample mean 5.0350275 converges onto the theoretical mean, $\frac{1}{\lambda} = \frac{1}{0.20} = 5$ and the sample variance 0.6049674 converges onto the theoretical mean, $\frac{\frac{1}{\lambda^2}}{n} = \frac{\frac{1}{0.04}}{40} = 0.625$ with the increase in experiment repeats.

Determining the Simulation is Approximately Normal

The final task of this assignment is to determine if the simulated exponential averages are a normal distribution. One clear way to do this is to compare the simulation statistics (mean and standard variation) to the theoretical statistics provided in the assignment instructions. This is demonstrated in the left graph below. A second way to very normal distribution is to examine the simulation data in a Quantil-Quantile plot (QQplot), right plot below, which compares the simulation quantiles against te theoretical quantiles. If the comparison shows a linear relationsip between the quantiles, it's indicative of the simulation and theoretical data deriving from a common distribution. For light background info on QQplots, refer to this stat site.

```r
#define theoretical/population stats
pMean <- 1/lambda; pVar <- (1/lambda^2)/n
#plot sample vs. theoretical distribution, with the pMean indicated by the dashed line in the graph
normDist <- ggplot(aes(x=samples), data=data.frame(samples))+
    geom_histogram(aes(y=..density..),binwidth=0.2, color="purple", fill="red") +
geom_density(lwd=1) + stat_function(fun=dnorm,
color="blue", lwd=1, args=list(mean=pMean, sd=pVar)) + labs(title="Sample mean simulation",
x=TeX("sample means ($\\bar{X}$)")) + theme(plot.title = element_text(hjust
=0.5)) + geom_vline(xintercept=pMean, linetype="dashed") + annotate("text", x=7,
y=0.6, label="sim. distribution", color="black") + annotate("text", x=7, y=0.57,
label="pop. distribution", color="blue")
#compare quantiles in sample exponential vs theoretical exponential to
#verify both originate from the same distribution
quantDist <- qplot(sample=samples, data=data.frame(samples)) +
labs(x="Theoretical Quantiles", y="Simulation Quantiles", title="Normal Q-Q plot") +
theme(plot.title = element_text(hjust=0.5)) + stat_qq_line()
grid.arrange(normDist, quantDist, nrow=1)
```



The left histogram demonstrates the simulated distribtuion with a delineations of simulated and theoretical gaussian distribution. This is indicative of the simulated data having normal distribution. This concures with CLT, that distribution of averages become standard normal with the increase in sample size (n). The linear relationship between the simulation and theoretical quantiles (right graph) demonstrate that they come from a common distribution and that the simulated data is again normally distributed.