A Simulation Analysis that shows a Sample Distribution Estimating its Population Distribution.

Bamidele Tella

8/28/2020

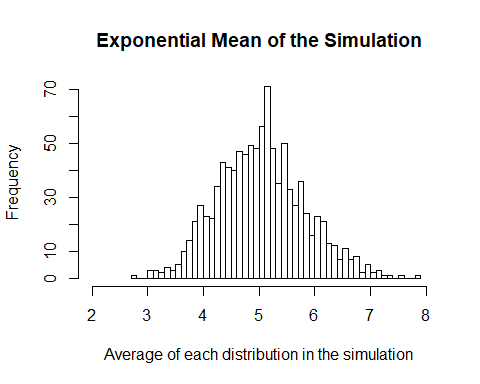
# OVERVIEW

This is the part 1 of the final project from Coursera’s statistical inference course. In this project we investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution is simulated using the R function rexp(n, lambda) where lambda is the rate parameter which is set to 0.2 and n is the number of exponential used in each sample distribution, which is set to 40. The theoretical mean of population distribution is calculated using the formula 1/lambda and the theoretical standard deviation is given as 1/lambda. A thousand simulations is used to estimate population distribution.

# SIMULATIONS

First, the seed is set to enable reproducibility, then the required packages are loaded into R. Next, The values of the parameters used (n,lambda,and number of simulations) are assigned to different variables. Then an exponential of 40 random values with a rate of 0.2 is created, and each distribution is repeated 1000 times and stored in the variable sample.sim (sample simulation). Next, the apply function in R is used to find the mean of each distribution and the result is stored in a variable exp.mn (exponential mean). Then we take a look at the frequency of the exponential mean with the aid of a histogram plot.

library(ggplot2)  
set.seed(2020)  
n <-40  
lambda <-0.2  
nsims <-1000  
sample.sim <-replicate(nsims,rexp(n,lambda))  
exp.mn <- apply(sample.sim,2,mean)  
hist(exp.mn,breaks = 40,xlab = "Average of each distribution in the simulation",xlim = c(2,8),main = "Exponential Mean of the Simulation")



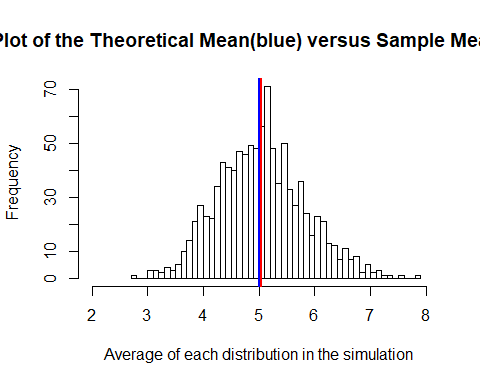
# Theoretical Mean versus Sample Mean

Now, we take a look at the sample mean and compare our result to the theoretical mean of our distribution. A histogram plot is also used to take a look at both theoretical mean(blue line) and sample mean(red line) on the exponential mean plot.

theoretical.mean<-1/lambda  
sample.mean<-mean(exp.mn)  
cbind(theoretical.mean,sample.mean)

## theoretical.mean sample.mean  
## [1,] 5 5.033948

hist(exp.mn,breaks = 40,xlab = "Average of each distribution in the simulation",xlim = c(2,8),main = "A Plot of the Theoretical Mean(blue) versus Sample Mean(red)")  
abline(v=theoretical.mean,col="blue",lwd=2)  
abline(v=sample.mean,col="red",lwd=2)

 As seen, our sample mean approximates our theoretical mean.

# Theoretical Variance versus Sample Variance

Then, we take a look at the variability between the theoretical variance and Sample variance.

theoretical.variance<-round((1/lambda/sqrt(n))^2,4)  
sample.variance<-round(var(exp.mn),4)  
theoretical.sd<-round((1/lambda/sqrt(n)),4)  
sample.sd<-round(sd(exp.mn),4)  
cbind(theoretical.variance,sample.variance)

## theoretical.variance sample.variance  
## [1,] 0.625 0.607

cbind(theoretical.sd,sample.sd)

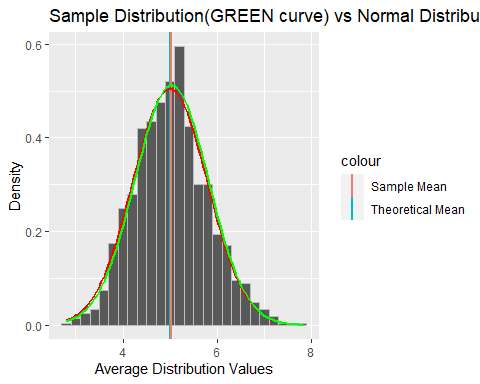
## theoretical.sd sample.sd  
## [1,] 0.7906 0.7791

Again, the sample variance approximates the theoretical variance. Also the sample standard deviation is closely approximated to the theoretical standard deviation.

# Distribution

Next, we take a look at the whole estimation of the exponential distribution and compare it to the visualisation of normal distribution to know how closely our sample distribution appears to the normal distribution.

plot3<-data.frame(exp.mn)  
gplot1<-ggplot(plot3,aes(x=exp.mn))  
gplot1<-gplot1 + geom\_histogram(aes(y=..density..,), colour="grey",binwidth = 0.2,show.legend = FALSE)  
gplot1<-gplot1 + geom\_vline(aes(xintercept = theoretical.mean, colour = "Theoretical Mean"), size=1)  
gplot1<-gplot1 + geom\_vline(aes(xintercept = sample.mean, colour = "Sample Mean"), size=1)  
gplot1<-gplot1 + stat\_function(fun = dnorm, args = list(mean=theoretical.mean,sd=theoretical.sd),colour="red",size=1)  
gplot1<-gplot1 + stat\_function(fun= dnorm,args = list(mean= sample.mean,sd=sample.sd),colour="green",size=1)  
gplot1<-gplot1 + labs(title="Sample Distribution(GREEN curve) vs Normal Distribution(RED curve)",x="Average Distribution Values",y="Density")  
gplot1



# Confidence Interval

Finally we take a look at the confidence level of both the theoretical distribution (theoretical.conf) and our sample distribution (sample.conf) to see the difference in range. The levels are represented as (x,y).

theoretical.conf<- theoretical.mean + c(-1,1) \*qnorm(.975) \* theoretical.sd/sqrt(n)  
  
sample.conf<- sample.mean + c(-1,1) \* qnorm(.975) \* sample.sd/sqrt(n)  
  
conf.int<-rbind(theoretical.conf,sample.conf)  
colnames(conf.int)<-c("x","y")  
conf.int

## x y  
## theoretical.conf 4.754995 5.245005  
## sample.conf 4.792507 5.275389

# Assumption

We assumed that according to the central limit theorem which states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

# Conclusion

As the analysis shows, the distribution of means of our sampled exponential distributions approximates a normal distribution. so we conclude that our assumption is valid, as stated by the Central Limit Theorem.