Statistical Inference Project : Central Limit Theorem

M. Y. Cheong

December 27, 2015

Introduction

The project intends to investigate the distribution of averages of 40 exponentials and compare it with the Central Limit Theorem.

In the simulations, two thousand (nsim = 2000) avearages of 40 (n = 40) exponential random variables (r.v.) are generated in R. The rate of the exponential distribution is $\lambda = 0.2$. Thus, the mean and standard deviation of the exponential distribution are both $\frac{1}{\lambda} = 5$.

The Central Limit Theorem states that the distribution of averages of iid random variables with well-defined mean μ and variance σ^2 will be approximately normally distributed when the sample size is sufficiently large. The distribution of the averages will have a mean center around the population mean and its standard deviation can be approximated with the standard error given by $se = \frac{\sigma}{\sqrt{(n)}}$. In the following, we will try to examine the CLT with simulations.

R code

The R code that generates the exponentials, calculates the mean, se, sd, etc of the distribution and produces the plots is as follows:

```
##
## Exponential distribution of rate 1/lambda : mean = 1/lambda, std = 1/lambda
##

rate <- 0.2     # rate parameter of exponential distribution (lambda)
n <- 40         # number of exponential variables for averaging
nsim <- 2000    # number of simulations
mu <- 1/rate    # mean of the exponential variable
sd <- 1/rate    # standard deviation of the exponential variable

x1 <- matrix(rexp(n*nsim, rate), nrow = n)  # matrix of 40x1000 of exponentials
y <- colMeans(x1)   # vector containing 'nsim' average of 'n' exponentials
```

```
## Plot the histogram of the sample mean
## Overlay a normal curve with mean =5 and std. deviation = se for comparison
##
hist(y, xlab=bquote("Mean of"~.(n)~"exponentials,"~bar(X)),
     main=bquote("Histogram of mean of"~.(n)~ "exponentials,"~bar(X)),
     ylim=c(0, 0.5), xlim=c(2,8),
     probability = T, cex.main=1)
curve(dnorm(x, mu, sd/sqrt(n)), add=T, col="red", lwd=2)
lines(5-rep(se,3), c(0, 0.1, 0.26), col="blue", lwd=1, lty=2)
lines(5+rep(se,3), c(0, 0.1, 0.26), col="blue", lwd=1, lty=2)
lines(rep(mean(y), 2), c(0,0.5), col="blue", lwd=2)
text(5,-0.01, ~mu, cex=0.8)
```

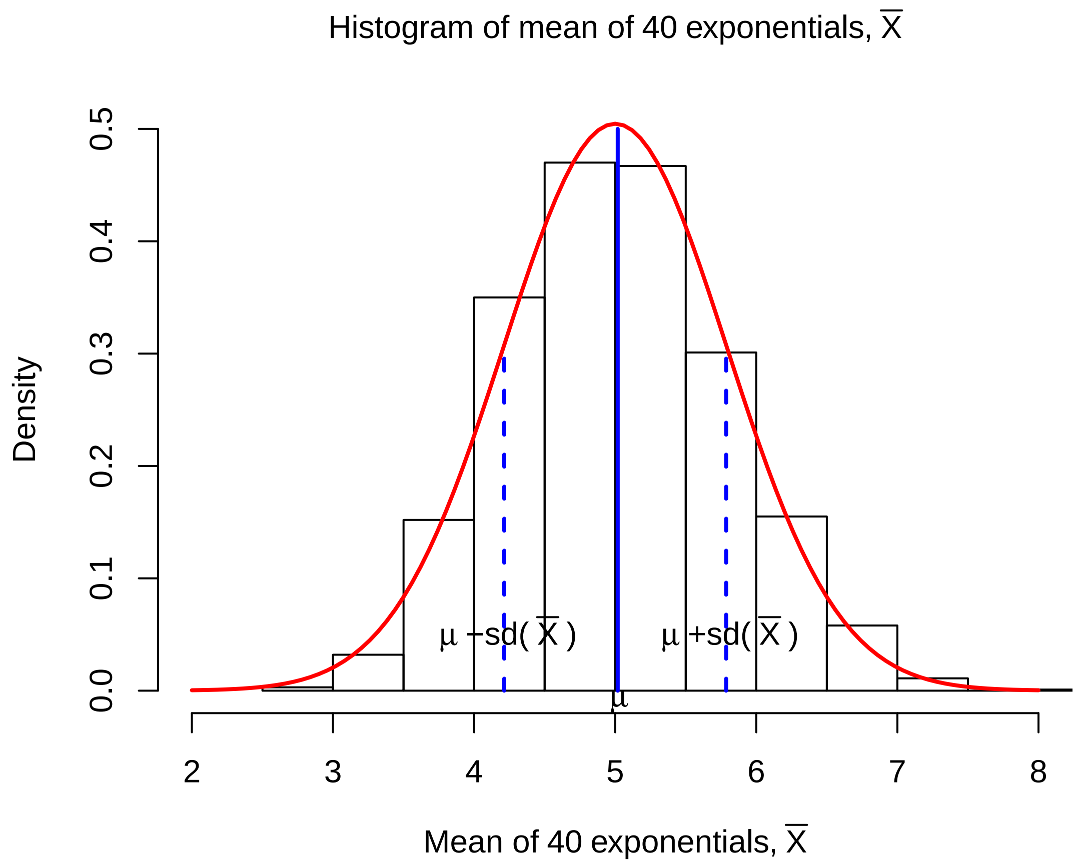
```r
text(3.5, 0.05, "5-sd("~bar(X)~")", cex = 0.6)
text(6.5, 0.05, "5+sd("~bar(X)~")", cex=0.6)

## Compare the standard deviation of the sample mean
## with the theoretical sample mean, i.e., sd/sqrt(n)
##
se <- sd(y)
se_t <- sd/sqrt(n)
```

```r
## Plot histogram of sample mean minus mean and divided by std error should give a std normal r.v
## Overlay a standard normal curve for comparison
##
y1 <- (y - mu)/se
hist(y, xlab=bquote("Mean of"~.(n)~"exponentials,"~bar(X)),
     main=bquote("Histogram of mean of"~.(n)~ "exponentials,"~bar(X)),
     ylim=c(0, 0.5), xlim=c(2,8),
     probability = T, cex.main=1)
curve(dnorm(x, mu, sd/sqrt(n)), add=T, col="red", lwd=2)
lines(c(5-se, 5-se, 5-se), c(0, 0.1, 0.3), col="blue", lwd=2, lty=2)
lines(c(5+se, 5+se, 5+se), c(0, 0.1, 0.3), col="blue", lwd=2, lty=2)
lines(rep(mean(y), 2), c(0,0.5), col="blue", lwd=2)
text(5,-0.01, ~mu, cex=1)
text(5-se, 0.05, ~mu~"-sd("~bar(X)~")", cex = 1)
text(5+se, 0.05, ~mu~"+sd("~bar(X)~")", cex=1)
```

Histogram of sample mean of 40 exponentials $\bar{X}$ approximates a normal distribution with mean $= \mu$ and standard dviation $= \frac{\sigma}{\sqrt{n}}$

The histogram of averages of 40 exponentials is shown in the plot below. For comparison, the normal curve with parameters mean$=\frac{1}{\lambda}$ and sd$=\frac{1}{\lambda\sqrt{n}}$ is also plotted for comparisson. It can be seen that the histogram approximates a normal distribution with mean around 5.



The mean of the distibution is indicated by the solid blue line, at 5.017983, which fall approximately around $\mu = 5$.

The dotted blue lines indicate one a shift of one standard deviation $\bar{X}$ from the mean. The standard deviation of sample mean $\bar{X}$ shows the variation of the sample mean from the population mean. From the simulation, the standard deviation of distribution of $\bar{X}$ is $\sigma_{\bar{X}} = 0.804$, which is close to the standard error which can be approximated with the population standard deviation as $\frac{\sigma}{\sqrt{n}} = 0.791$.