

Exponential Distribution vs. CLT

Yanhui Zhou

October 22, 2015

Overview

This project is for the statistical inference class' course project of Coursera. We will investigate the exponential distribution in R and compare it with the Central Limit Theorem. We will use simulation to explore inference and do some simple inferential data analysis.

Simulations

Simulate by 40 exponentials with 1000 times, draw the distribution of the averages of 40 exponentials. Then investigate properites of this sample mean distribution, and compare with properites of theoretical distribution (Central Limit Theorem) .

```
# set lambda and other constants
lambda <- 0.2
n <- 40
exnum <- 1000

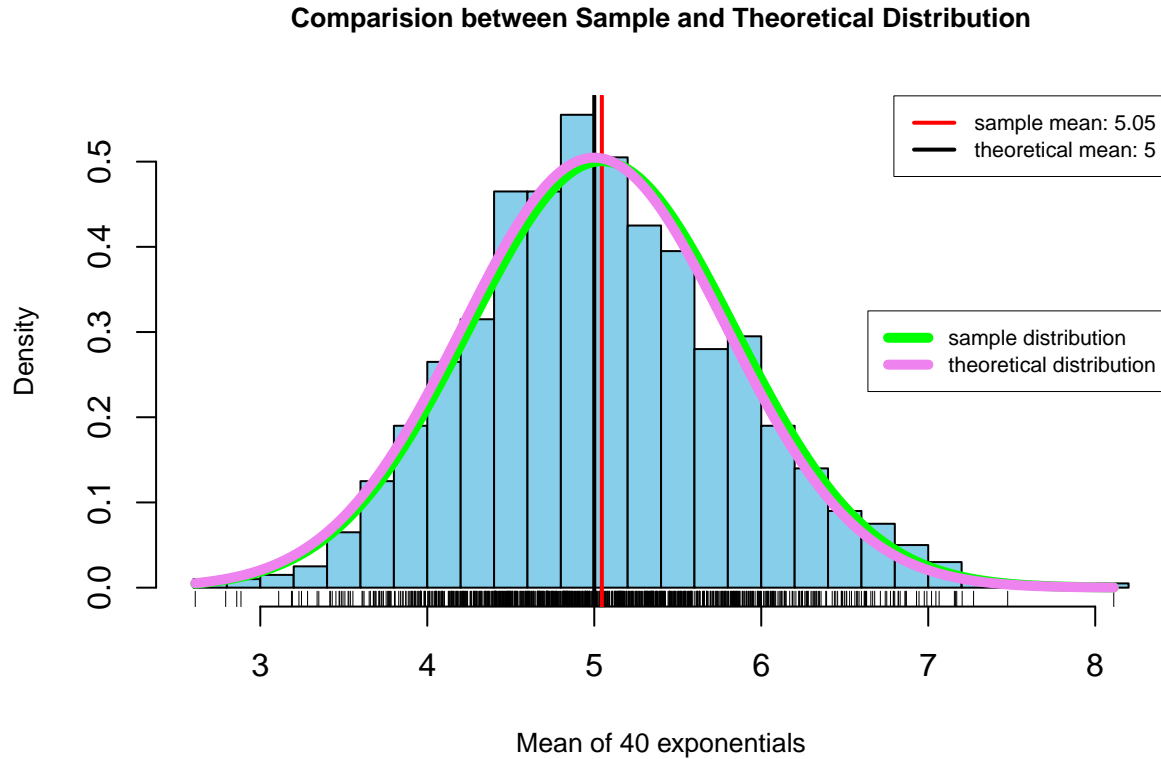
# begin simulation
set.seed(333)
dmatrix <- replicate(exnum, rexp(n, lambda))
mns <- apply(dmatrix, 2, mean)
h<-hist(mns, breaks = 30, prob=TRUE, col='skyblue',
        xlab='Mean of 40 exponentials', ylab = 'Density',
        main="Comparision between Sample and Theoretical Distribution",
        cex.main=0.8, cex.lab=0.8)
rug(mns)
# add the line of sample'mean
sample_mean <- mean(mns)
sample_sd <- sd(mns)
abline(v=sample_mean, col='red', lwd = 2)
# add the line of distribution'mean
theory_mean = 1/lambda
theory_sd = 1/lambda * 1/sqrt(n)
abline(v=theory_mean, col='black', lwd = 2)
legend("topright", col=c('red', 'black'),
       legend=c(paste('sample mean:', round(sample_mean, 2)),
                paste('theoretical mean:',theory_mean)),
       cex=0.7, lwd=2, box.lwd=0)

# simulate the distribution
xfit<-seq(min(mns),max(mns),length=length(mns))
yfit<-dnorm(xfit,mean=sample_mean,sd=sample_sd)
lines(xfit, yfit, col="green", lwd=5)

tyfit<-dnorm(xfit,mean=theory_mean,sd=theory_sd)
```

```
lines(xfit, tyfit, col="violet", lwd=5)

legend("right", col=c('green', 'violet'),
      legend=c('sample distribution', 'theoretical distribution'),
      cex=0.7, lwd=5, box.lwd=0)
```



Sample Mean versus Theoretical Mean, Sample Variance versus Theoretical Variance

In the figure “Comparison between Sample and Theoretical Distribution”, the vertical line of sample mean is colored by red, and the theoretical mean is black. The difference is very small. The green curve is the sample distribution and the violet curve is the theoretical distribution of mean of 40 exponentials. And the detail value of means and variances are in the table above.

From the analysis and simulation above, we know, the detail comparison are:

Variable	Theoretical	Sample
Mean	5	5.0453365
SD	0.7905694	0.7971902
Variance	0.625	0.6355122

Distribution

From simulation figure “Comparison between Sample and Theoretical Distribution”, we know the distribution of average of 40 exponentials is approximately normal, with mean=5.0453365 close to 5, variance=0.6355122 close to 0.625. According to Central Limit Theorem, the distribution of averages of iid variables becomes that of a normal with large sample size, approximately $\bar{X}_n \sim N(\mu, \sigma^2/n)$.