Part 1:Simulation Activities

Nirmal Ghimire

10/3/2020

Overview:

This is a class project for Statistical Inference Course in Data Science Certification Program through John Hopkins University. This project conducts thousands of simulations of data of exponential distribution in R and assess how their means and variance compare with the theoretical ones. The second half of the project analyzes the ToothGrowth data and performs some basic exploratory analyses.

Part 1-Simulation Exercise

This project investigates the exponential distribution in R and compares it with the Central Limit Theorem. The exponential distribution was simulated in R using ‘rexp (n, lambda)’. If you are not sure, exponential distribution has a mean and standard deviation of 1 over lambda (1/ʎ). Lambda was set to (0.2) for all the simulations. This project investigates the distribution of averages of 40 exponentials using a thousand of simulations. This project attempted to answer following questions: 1. Does the sample mean vary from the theoretical mean? 2. Calculate the sample variance and compare it with the theoretical variance of the distribution. 3. Check if the distribution is approximately normal. The answer to question 3 should focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

setting seed for generating data

```
set.seed(-123)
```

setting up 40 exponential

```
n<-40
```

setting lambda to 0.2 for all simulations

```
lambda<- 0.2
```

Need 1000 simulations

```
S<-1000
```

setting z value for 95% confidence interval (CI)

```
z<-1.96
```

A. Simulating data in R and calculating means

Creating dataframe with 40 columns and 1000 exponential simulations

```
mydata<-matrix(rexp(n*S,lambda),nrow=S)
```

Checking if the data frame is created

```
str(mydata)

##  num [1:1000, 1:40] 3.28 6.62 1.43 2.97 14.94 ...
```

B. Calculating Sample Mean and Theoretical Mean Using Simulated Data

Mean of simulated data per row

```
Row_Mean<-rowMeans(mydata)
```

Checking if that works

```
summary(Row_Mean)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.872   4.439   4.967   5.019   5.500   8.183
```

Single mean of all row means

```
MeanOfMean<-mean(Row_Mean)
```

Printing the Simulated Mean

```
MeanOfMean

## [1] 5.018855
```

Calculating Theoretical Mean

```
theoreticalMean<-1/lambda
```

Printing the Theoretical Mean

```
theoreticalMean

## [1] 5
```

C. Calculating Sample Variance and Theoretical Variance Using Simulated Data

```
sampleVar<-var(Row_Mean)
```

printing the sample variance

```
sampleVar

## [1] 0.7073296
```

Calculating the theoretical variance

```
theoreticalVar<-(1/lambda)^2/(n)
```

Printing the theoretical variance

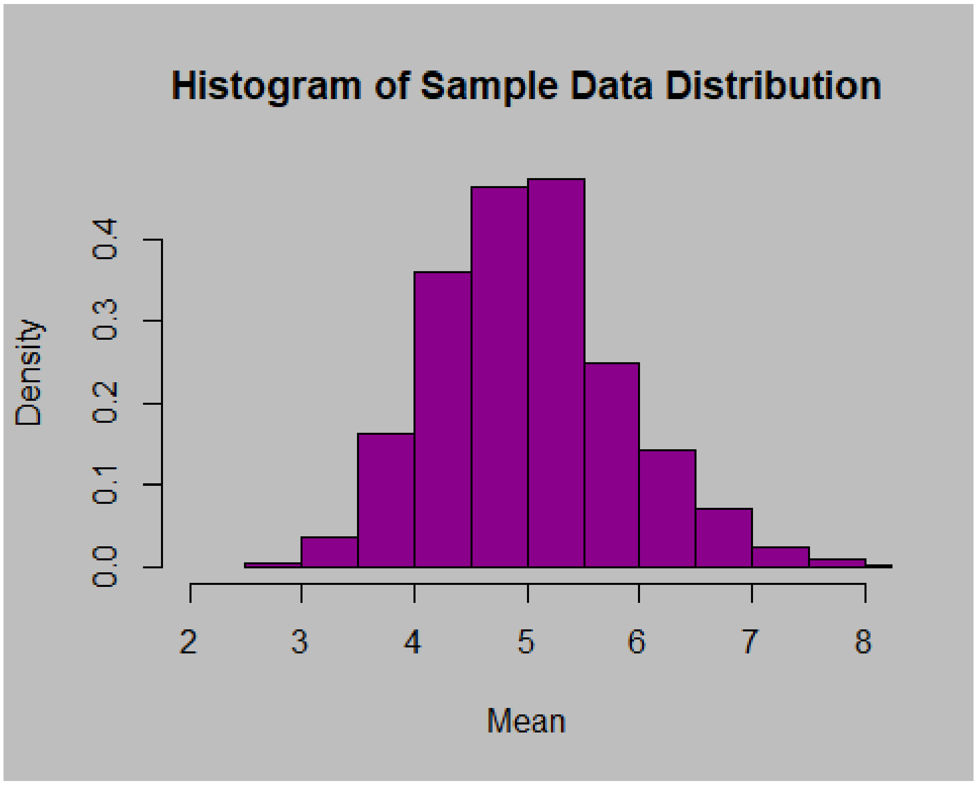
```
theoreticalVar

## [1] 0.625
```

D. Checking Data Distribution and Assessing if they are Approximately Normal

Histogram of the sample mean

```
par(bg='grey')
hist(Row_Mean,
     main="Histogram of Sample Data Distribution",
     xlab="Mean",
     xlim=c(2,8),
     col="darkmagenta",
     freq=FALSE
     )
```


Histogram of Sample Data Distribution
Density
Mean