Statistical Inference Course Project - Part I

Pingping

7/11/2020

Introduction

This project investigated the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution is simulated in R with rexp(n, lambda) where lambda is the rate parameter. We investigated the distribution of averages of 40 exponentials with a thousand simulations.

Data simulation

A thousand of simulation data were create randomly and stored in a 1000*40 matrix. Each row stands for a simulation. Each simulation has 40 exponentials.

```
n <- 40 # sample size
lambda <- 0.2 # lambda for rexp
times <- 1000 # number of simulations
quantile <- 1.96 # 95th % quantile to be used in Confidence Interval
set.seed(1) # set the seed value for reproducibility
Data <- matrix(rexp(n * times, rate = lambda), times) # each row stands for a simulation
```

Statistical analysis

(1) Mean Comparison

```
Datamean <- rowMeans(Data) # get the row mean of the matrix
samplemean <- mean(Datamean) # get the mean value of the 1000 simulations
samplemean # samplemean
```

```
## [1] 4.990025
```

```
theomean <- 1 / lambda # get the theoretical mean
theomean # theoretical mean
```

```
## [1] 5
```

Conclusion: The distribution of the mean of the sample means is centered at **4.990025** and the theoretical mean is centered at **5**, which are very close.

(2) Variance comparison

```
samplevar <- var(Datamean) # get the sample variance
samplevar # sample variance
```

```
## [1] 0.6177072
```

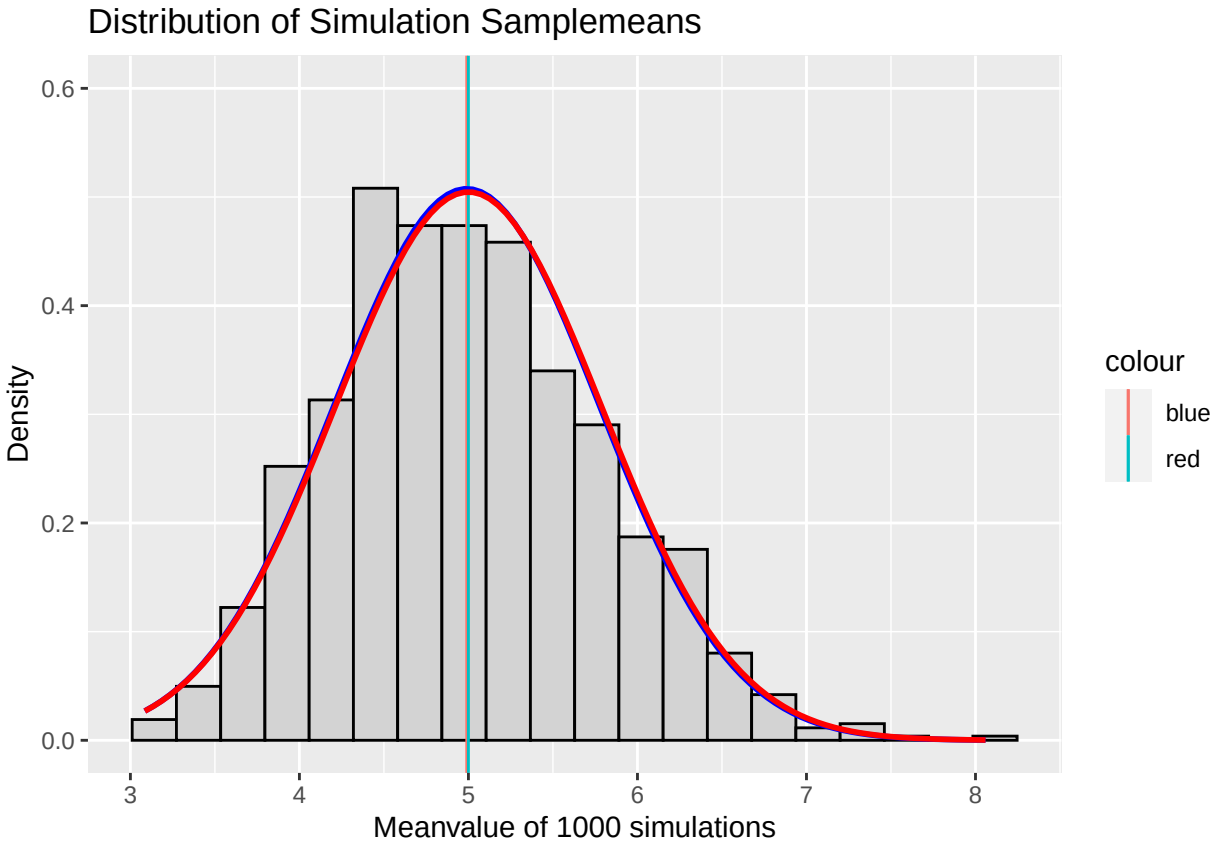
```
theovar  <- (1 / lambda)^2 / (n) # get the theoretical variance
theovar # theoretical variance
```

```
## [1] 0.625
```

Conclusion: The variance of the sample means is **0.6177072** and the thoeretical variance of the distribution is **0.625**, which are very close.

(3) Sample mean distribution

```
disdata <- as.data.frame(Datamean)
g <- ggplot(disdata, aes(x =Datamean))
g <- g + geom_histogram(aes(y=..density..), colour="black", fill = "light grey", bins = 20) + ylim(0, 0
g <- g + labs(title = "Distribution of Simulation Samplemeans", x = "Meanvalue of 1000 simulations", y =
g <- g + geom_vline(aes(xintercept = samplemean, colour = "blue"))
g <- g + geom_vline(aes(xintercept = theomean, colour = "red"))
g <- g + stat_function(fun = dnorm, args = list(mean = samplemean, sd = sqrt(samplevar)), color = "blue
g <- g + stat_function(fun = dnorm, args = list(mean = theomean, sd = sqrt(theovar)), colour = "red", s
g
```



Conclusion The distribution of the sample meanvalues of the 1000 sumulation was shown in the the figure. The blue vertical line stands for the samplemean of the simulations, while the red line stands for the theoretical mean. There two lines overlapped as the two values are very close to each other. The blue bell curve stands for the normal distribution with a mean value of the simulation mean, and sd value of the similation standard error. The red bell curve stands for the normal distribution with a mean value of the theoretical mean, and sd value of the theoretical standard error. There two lines overlapped as the values are very close to each other. The normal distribution bell curve fits the shape of the distribution of simulation samplemean, suggesting the distribution if approximately normal.