Statistical Inference Project Pt I

R3M79

26 de Dezembro de 2017

Synopsis

This document pertains to Cousera's Statistical Inference model Project. The project is divided in two parts

1. A simulation exercise.
2. Basic inferential data analysis.

In this document we'll address part 1 of the project

Part 1: Simulation Exercise

Overview

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. It should

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

In point 3, 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.

Simulation

Define Variables and simulation data

Mean Comparison

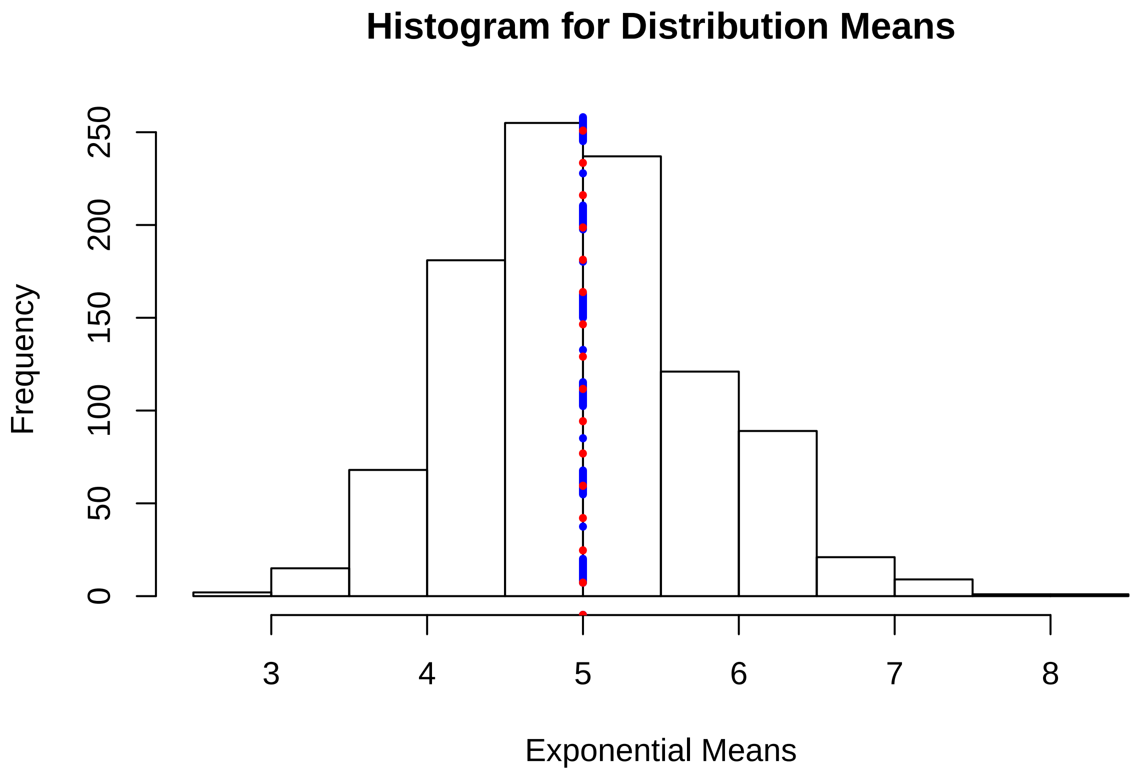
Prepare data and compare Theoretical and Sample means and variance.

```r
cat("Theory Mean:",theoryMean,"Sample Mean:",sampleMean)
```

```
## Theory Mean: 5 Sample Mean: 4.999702
```

As we can see from the output values above and the plot below the sample(red) and theoretical(blue) means almost overlap.

```
#plot data with comparison betweem Means
hist(simMeansVec,xlab = "Exponential Means",
     main="Histogram for Distribution Means")
abline(v=theoryMean,lty=4,lwd=4,col="blue")
abline(v=sampleMean,lty=3,lwd=4,col="red")
```



Variance Comparison

```
cat("Theory Variance:",theoryVariance,"Sample Variance:",sampleVariance)
```

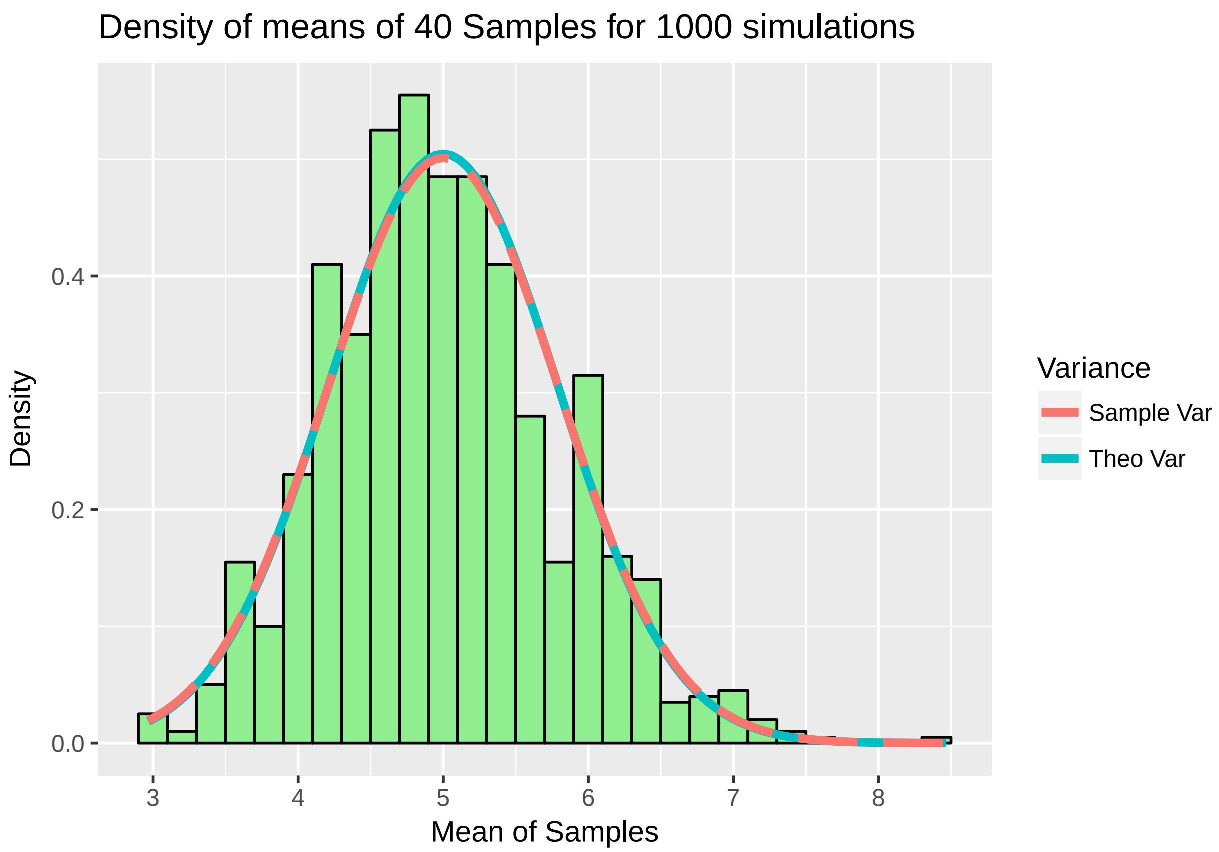
```
## Theory Variance: 0.625 Sample Variance: 0.6335302
```

As we can see the theoretical and sample Variance are quite close to each other as one would expect, after what we have observed with the mean.

Show distribution is approximately normal

We'll now demonstrate that the exponential distribution is approximately normal. For that we'll plot the histogram for the means density. We'll also present the curve distributions of the sample and theoretical standard deviation

```
plotdata
```



From the above histogram we see that the distribution is close to the normal distribution. We also see that the curves for the Sample and Theoretical Standard Deviations have very close values almost overlaping, as we've already seen for the Mean and Variance values.

Confidence Interval Comparison

We'll now check the Confidence Interval for our analysis

```
cat("Theory Conf. Int.:",theoCI,"Sample Conf. Int.:",sampleCI)
```

```
## Theory Conf. Int.: 4.753338 5.246662 Sample Conf. Int.: 4.753 5.246
```

As we can see the Confidence Intervals for both the Sample and Theoretical Means are very closely matched.

Conclusion

From the above outputed values and plots we can conclude that the distribution is close to a normal Distribution, as expected by the Central Limit Theorem (CLT)

Appendix

Code

Below follows all the code necessary for the displayed information and plots.

```
#load libraries
library(dplyr)
library(ggplot2)


## Part 1 Simulation

#Define Variables for simulation
set.seed(100) # set the seed value for reproducibility

nexp <- 40           #number of exponentials
lambda <- 0.2     #value for exp ditribution
nsim <- 1000      #number of simulation
quantile <- qnorm(.975) # 95th % quantile to be used in Confidence Interval
                 # ~1.696

#create matrix for simulation
simMatrix <- matrix(rexp(nexp * nsim, rate = lambda), nsim)

#create vector with means
simMeansVec <- rowMeans(simMatrix)

#theoretical mean
theoryMean <- 1/lambda

#calculate sample mean
sampleMean <- mean(simMeansVec)

##Mean Comparison

cat("Theory Mean:",theoryMean,"Sample Mean:",sampleMean)


#plot data with comparison betweem Means
hist(simMeansVec,xlab = "Exponential Means",
     main="Histogram for Distribution Means")
abline(v=theoryMean,lty=4,lwd=4,col="blue")
abline(v=sampleMean,lty=3,lwd=4,col="red")

##Variance Comparison

sampleVariance <- var(simMeansVec)
theoryVariance <- (1 / lambda)^2 / nexp

cat("Theory Variance:",theoryVariance,"Sample Variance:",sampleVariance)

##Show distribution is approximately normal
```

```r
simMeansDFrame <- data.frame(simMeansVec)
plotdata <- ggplot(simMeansDFrame, aes(x = simMeansVec))
plotdata <- plotdata +
    geom_histogram(aes(y=..density..),
                   colour="black", fill = "lightgreen",
                   binwidth = 0.2)
plotdata <- plotdata +
    stat_function(fun = dnorm, args = list(mean = theoryMean,
                                           sd = sqrt(theoryVariance)),
                  aes(colour = "Theo Var"), lwd = 1.5)
plotdata <- plotdata +
    stat_function(fun = dnorm, args = list(mean = sampleMean,
                                           sd = sqrt(sampleVariance)),
                  aes(colour = "Sample Var"), lwd = 1.5, lty = 5)
pltodata <- plotdata +
    scale_colour_manual(values = c("blue"="blue","red"="red") )
plotdata <- plotdata +
    labs(title = "Density of means of 40 Samples",
         x = "Mean of 40 Samples", y = "Density",colour="Variance")
plotdata


##Confidence Interval Comparison
theoCI <- theoryMean + c(-1,1)*quantile*sd(simMeansVec)/sqrt(nexp)
sampleCI <- round(mean(simMeansVec) + c(-1,1)*quantile*sd(simMeansVec)/sqrt(nexp),3)
cat("Theory Conf. Int.:",theoCI,"Sample Conf. Int.:",sampleCI)
```