Statistical Inference Project - Part1

A. Swarup

May 19, 2018

Overview

This is the first part of the Coursera Statistical Inference Course project. Here we'd investigate the exponential distribution in R and compare it with the Central Limit Theorem. Using R function "rexp" we'd perform 1000 simulations of exponential distributions of 40 random values. We'd compare the mean and variance of the simulated exponential distribution to the theoretical mean and distribution and demonstrate that the distribution is approximately normal.

Simulations

The mathematical formula for exponential distribution is:

$$f(x;\lambda) = \lambda e^{\lambda x} \; with \; mean = \frac{1}{\lambda} \; and \; variance = \frac{1}{\lambda^2}$$

In R we can generate exponentially distributed random variables by using the function rexp(n,lambda) where lambda (λ) is the rate parameter. Our code for simulations is as follows. Please see the associated comments for explanations.

```
n <- 40 # no. of exponentials
lambda <- 0.2 # rate parameter lambda = 0.2
set.seed(123) # Set seed to get repeatable random numbers
simulations <- 1000 # no. of simulations
SimulatedData <- replicate(simulations, rexp(n, lambda))
```

The SimulatedData is a n (=40) x 1000 (no. of simulations) matrix.

Analysis

1. Sample Mean versus Theoretical Mean

The code to calculate mean of simulated data and the theoretical mean follows.

```
# Mean of the simulated data
meanExponential <- colMeans(SimulatedData)
# head(meanExponential)
meanOfSimulationMeans <- mean(meanExponential); meanOfSimulationMeans
```

```
## [1] 5.011911
```

```
# Theoretical mean
mu <- 1/lambda; mu
```

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

We see that simulations mean 5.01 is very close to the theoretical mean of 5.

Figure 1 in Appendix shows a histogram of the simulated means.

2. Sample Variance versus Theoretical Variance

R code to get standard variation and variance of simulated data and their theoretical values is as follow.

```
# Simulations standard deviation and variance
sd_Sim <- sd(meanExponential); sd_Sim
```

```
## [1] 0.7749147
```

```
var_Sim <- var(meanExponential); var_Sim
```

```
## [1] 0.6004928
```

```
# Theoretical standard deviation and variance
sd <- (1/lambda)/sqrt(n); sd
```

```
## [1] 0.7905694
```

```
Var <- sd^2; Var
```

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

We see here that the standard deviations are very close (.775 versus .790). Variance being the square of standard deviation, minor differences become much larger. Still, the simulations variance of .600 is close to the theoretical variance of .625.

3. Comparison to Normal Distribution

We plot the histogram again (with 40 breaks instead of 20 that were taken in Figure 1) and overlay it with the theoretical normal distribution line - please see Figure 2 in the Appendix. Figure 2 shows that the simulated exponential distribution closely matches a Normal Distribution.

4. Confidence Intervals Comparison

We have already seen that the mean and variance of the sample data is close to that of a normal distribution. We now look at the confidence intervals for the two cases.

```
sample_confinterval <- round (mean(meanExponential) + c(-1,1)*1.96*sd_Sim/sqrt(n),3)
sample_confinterval
```

```
## [1] 4.772 5.252
```

```
theo_confinterval <- mu + c(-1,1)* 1.96*sqrt(Var)/sqrt(n)
theo_confinterval
```

```
## [1] 4.755 5.245
```

The sample confidence interval is (4.772, 5.252). It closely matches the theoretical confidence level (4.755, 5.245), again confirming that the distribution is approximately normal.

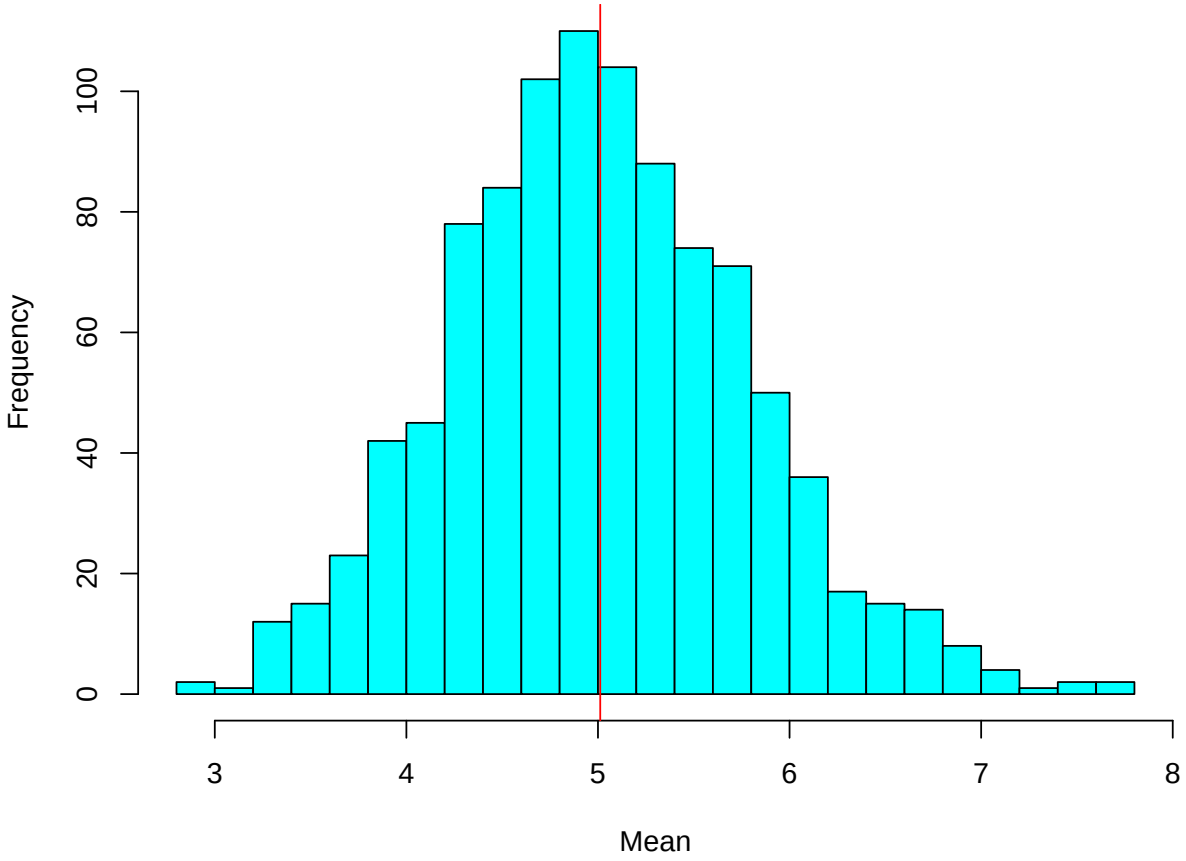
We can now plot the Q-Q for quantiles to show that the sample quantiles match the theoretical quantiles. This plot shown as Figure 3 in the Appendix also suggests normality.

Conclusion

We have demonstrated that the distribution of means of 40 exponential distributions is close to the normal distribution with the expected theoretical values based on the given value of lambda (λ=0.2).

APPENDIX

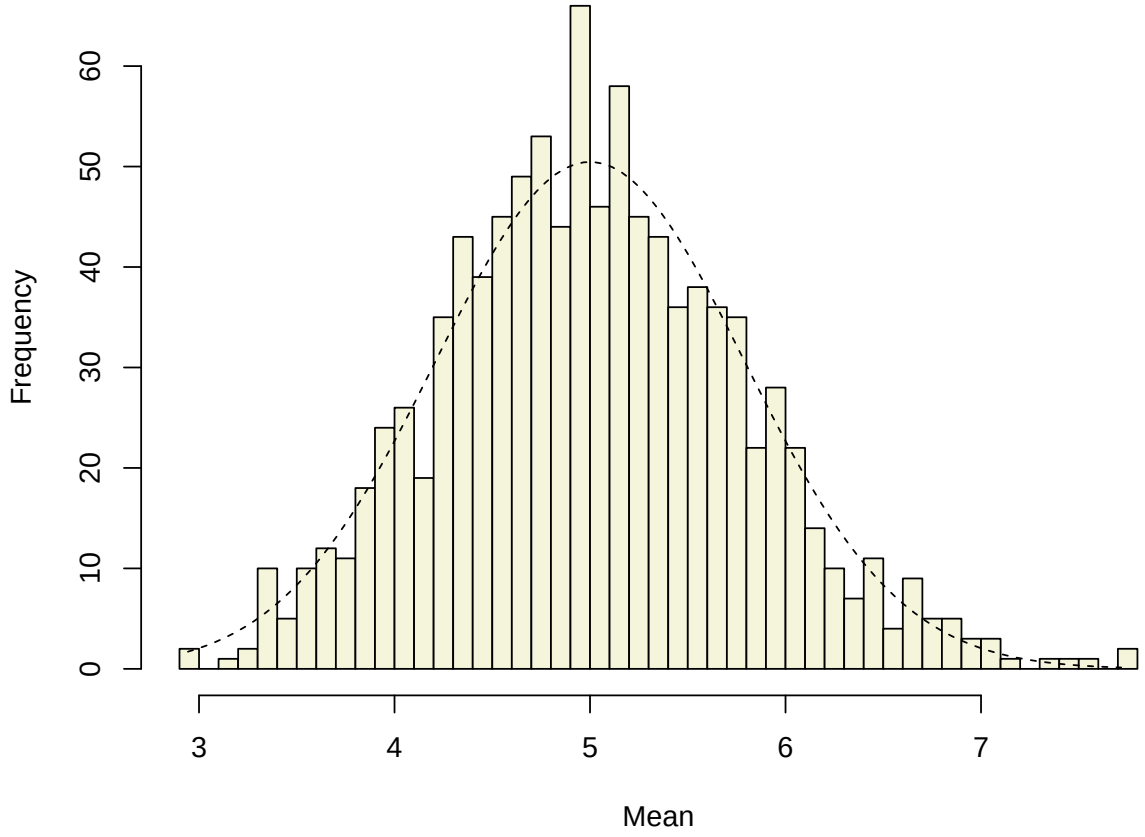
Figure 1. Histogram of 1000 Simulated Exponential Means



Code to overlay normal distribution curve on the histogram of simulated data

```
# Plot histogram
hist(meanExponential, breaks = 40, xlab = "Mean", col = "beige",
     main = "Figure 2. Comparison to a Normal Distribution")
# Add the theoretical Normal Distribution line
xfit <- seq(min(meanExponential), max(meanExponential), length = 100)
yfit <- dnorm(xfit, mean = 1/lambda, sd = 1/lambda/sqrt(n))
lines(xfit, yfit*100, lty=2)
```

Figure 2. Comparison to a Normal Distribution



Q-Q Plot Code

```
qqnorm(meanExponential, main="Figure 3. Normal Q-Q Plot",
       xlab="Theoretical Quantiles",
       ylab="Sample Quantiles")
qqline(meanExponential, col="red")
```